# ANALYSIS AND MODELLING OF SURFACE WATER QUALITY IN RIVER BASINS

**A**

**DISSERTATION**

**SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE**

**DEGREE OF**

**MASTER OF TECHNOLOGY (RESEARCH)**

IN

**CIVIL ENGINEERING**

WITH SPECIALIZATION IN

**WATER RESOURCES ENGINEERING**

By

**MRUNMAYEE MANJARI SAHOO**

Under the Supervision of

DR. K C PATRA

and DR. K K KHATUA


**DEPARMENT OF CIVIL ENGINEERING**

**NATIONAL INSTITUTE OF TECHNOLOGY**

**ROURKELA-769008**

**2014**

**NATIONAL INSTITUTE OF TECHNOLOGY**

**ROURKELA**

## CERTIFICATE

*This is to certify that the Dissertation entitled "***ANALYSIS AND MODELLING OF SURFACE WATER QUALITY IN RIVER BASINS***" submitted by* **MRUNMAYEE MANJARI SAHOO** *to the* **National Institute of Technology, Rourkela**, *in partial fulfillment of the requirements for the award of* ***Master of Technology (Research)*** *in* ***Civil Engineering*** *with specialization in* ***Water Resources Engineering*** *is a record of bonafide research work carried out by her under our supervision and guidance during the academic session 2012-14. To the best of our knowledge, the results contained in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.*

Date

Dr. Kanhu Charan Patra

and Dr. Kishanjit Kumar khatua

Professor, Department of Civil Engineering

National Institute of Technology, Rourkela

## ACKNOWLEDGEMENTS

A complete research work can never be the work of anyone alone. The contributions of many different people, in their different ways, have made this possible.

I would like to express my special appreciation and thanks to my supervisors Professor Dr. Kanhu Charan Patra and Professor Dr. Kishanjeet Kumar Khatua, you both have been tremendous mentors for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scholar. Your advices on both researches as well as on my career have been priceless.

I would also like to thank my committee members, Professor, Nagendra Roy; Head of the Civil Department, Professor Kishor Chandra Biswal, Professor Kali Pada Maity and Professor J. Srinivas for serving as my committee members even at hard times. I also want to thank you for letting my seminars be an enjoyable moment, and for your brilliant comments and suggestions.

I wish to express my sincere gratitude to Dr. S K Sarangi, Director, NIT, Rourkela for giving me the opportunities and facilities to carry out my research work.

In addition I would like to acknowledge the Central Water Commission, Bhubaneswar, Odisha and Odisha Pollution Control Board, Bhubaneswar, Odisha for providing the required data for my research work.

I would also thankful to my Husband, Mr. Janaki Ballav Swain, who supported me in writing, and inspired me to strive towards my ultimate goal of taking this thesis to a logical conclusion.

Above all, a special thanks to my family, words cannot express how grateful I am to my Father, Mother, Brother and Sister for all of the sacrifices they have made on my behalf. Your prayer for me was that sustained me thus far.

Mrunmayee Manjari Sahoo

## TABLE OF CONTENTS

**LIST OF FIGURES**

**LIST OF TABLES**

## ABSTRACT

Water is one of the prime elements responsible for life on the earth. India's surface water flows through 14 major river basins beyond innumerable medium/minor basins. The climate change is affecting the precipitation and ultimately affects the quantity of freshwater available, whereas, increasing waste water loads from point and non-point sources are deteriorating the quality of surface wateras well as ground water resources. The surface water quality is a very important and sensitive issue and is a great environmental concern worldwide. Surface water pollution by chemical, physical, microbial and biological contaminants can be considered as an epidemic all over the world. The Study area of research work is Brahmani River Basin in Odisha, India. The monthly water quality parameters are collected and analyzed from five selected gauging stations of Odisha during the months of January to December from 2003 to 2012. Eleven physical, chemical and biological water quality parameters viz.,pH, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Electrical Conductivity, Nitrogen as nitrate (Nitrate-N), Total Coli-form Bacteria(TC), Fecal Coli-form Bacteria(FC), Chemical Oxygen Demand (COD), Nitrogen as ammonia ($NH_4$-N), Total Alkali (TA) as $CaCO_3$, Total Hardness (TH) as $CaCO_3$ are selected for the analysis. Analysis of water quality for Brahmani River is done by techniques such as Spearman's Rank Correlation, Calculation of parts of water quality parameter, Overall Water Quality Index (WQI), Multivariate Analysis of variance (MANOVA) with Discriminant Analysis, Principal component Analysis and Factor Analysis, Canonical Correlation Analysis (CCA), Cluster Analysis (CA). Modelling is done by using Adaptive Neuro-Fuzzy Inference System (ANFIS) in MATLAB, Artificial Neural Network (ANN) and risk based analysis by Monte Carlo simulations (MCS). The Error analysis and performance evaluation of the applied models were also done to know the best fit model for the study. Regression plots between actual and predicted WQI via ANFIS revealed high values of coefficient of determination ($R^2$) of 0.994 and 0.995 for training and testing in summer season, 0.985 and 0.990 in monsoon season and 0.992 and 0.993 in winter season respectively. However, the coefficients of determination ($R^2$) for Artificial Neural Network (ANN) between actual and predicted values of WQI were 0.945, 0.941 and 0.965 for summer, monsoon and winter seasons respectively. Monte Carlo Simulations (MCSs) provide techniques for simulating the parameters having high degrees of freedom. There is least error in case of ANFIS when compared with ANN and MCS. Therefore, it can be stated that ANFIS predicted WQI with a far better accuracy than ANN and MCS. From the results of ANFIS, it can be concluded that if the present conditions can be considered to remain the future years could have most likely similar trend as from the trend observed during2003 to 2012.



Key Words: ANFIS, ANN, CCA, Discriminant Analysis, MCS, PCA, Brahmani River, WQI

# CHAPTER I

# INTRODUCTION

## 1.1 General

Water is one of the prime elements responsible for life on the earth. The six billion people on earth use nearly 30 percent of the world's total accessible renewal water supply. Yet billions of people are deprived of basic water availability. Among other countries in the world, India is one of the few selected countries endowed with reasonably good land as well as water resources. India is a country with vast geographic, biological and climatic diversity. Average annual precipitation including snowfall is approx. 4000 billion cubic meters (BCM) over the country. The average annual water resources in various river basins are estimated to be 1869 BCM, of which 1086 BCM is utilizable including 690 BCM of surface water and 396 BCM of ground water. The rest of the water is lost by evaporation or flows into the sea and goes unutilised.

India's surface water flows through 14 major river basins. In addition to major rivers, there are 44 medium and 55 minor river basins. These rivers are fast flowing and are mostly monsoon fed. Due to the spatial and temporal variations in precipitations as well as the rapid growth of population and improved living standards, the demand for supply of water resources in general and fresh water in practical is increasing. As a result of this, per capita availability of water is reducing day by day. However, surface water resources in the country are in much greater volume when compared to the groundwater resources. The climate change is affecting the precipitation and ultimately affects the quantity of water available, on the other hand, increasing loads from point and non-point sources are deteriorating the quality of surface as well as ground water resources. As the majority of the rivers in the country are not perennial, groundwater actually sustains much of the population during the lean months. There is a tremendous variation both in the quantity and quality of discharge from region to region in these river basins. With a few exceptions, all the medium and minor river basins originate in the mountains, and thus exhibit a common feature of fast flowing and monsoon-fed streams in the hilly regions. By the time they reach the plains they are mostly transferred as tidal streams. The treated or untreated discharges from such sources would always find a way into the rivers that oscillate like a pendulum due to the seasonal flow character of these rivers. During monsoon, when rainwater flows down the river the discharge in the pollutants, the flow rate and flow depth oscillate because of the tides in the tidal reaches. As the storm water moves

downstream, the flushing out time for the pollutants decreases substantially. All the major river basins are not perennial. Many of the major river basins also go dry during the summer leaving insufficient water for dilution of waste water discharged in them.

## 1.2 Water Quality in River Basins

Water is very vital for human beings and the health of its ecosystem. Thus quality of water is extremely important. The surface water quality is a very sensitive issue and is also a great environmental concern worldwide. Surface water pollution by chemical, physical, microbial and biological contaminants can cause epidemic problems, at times all over the world. Fish survival / growth and other biodiversity, conservation activities, recreational activities like swimming and boating, industrial / municipal water supply, agricultural uses such as irrigation and livestock watering, waste disposal and all other water uses are affected by the physical, chemical, microbial and biological conditions that exist in the water courses and also in subsurface aquifers. The surface water systems are naturally open to the atmosphere, such as lakes, rivers, estuaries, reservoirs and coastal waters. A natural process such as changes in erosion, precipitation, weathering of crustal material as well as any anthropogenic influences such as urban, industrial and agricultural activities, increasing rate of consumption of water resources, degrade in the quality and quantity of surface water and make it unsuitable for domestic uses. Industrial waste water, runoff over the agricultural lands and municipal sewage disposal are the most vulnerable for water pollution (Singh 2005). The concentration of biological available nutrients in excess and concentration of toxic chemicals leads to diverse problems such as toxic algal blooms, loss of oxygen in water, fish kill loss of biodiversity and loss of aquatic plants and coral reefs (Vousta et al., 2001).

## 1.3 Statistical and Multivariate Analysis of Water Quality

Statistical analysis is the study of collection, organization, analysis, interpretation and presentation of sample data. It also refers to a collection of methods used to process large amount of data and to report overall trends. It deals with all aspects of data, including planning of sample data collection. Statistical analysis can be broken into five discrete steps; such as (a) Describe the nature of data to be analyzed and presented, (b) Explore the relation of data with each other or underlying population, (c) Create a model to summarize or organize understanding of how the data relates to each other, (d) Prove or disprove the validity of the model, (e) Employ predictive analysis for the future trends.

Multivariate statistical analysis (MVA) is based on the statistical principle of multivariate statistical analysis, which involves observation and analysis of more than one variable at a particular time. Multivariate analysis concerns about different aims and background of each of

the different forms of multivariate analysis and how they relate with each other. Multivariate Statistics include univariate and multivariate analysis in order to understand the relationships between variable and their relevance to the actual problem being studied.

There are different statistical analysis methods, each with its own type of analysis according to the problem being selected. Those are:

(i) ANOVA (i.e. Analysis of variance) is a particular form of statistical hypothesis testing heavily used in the analysis of experimental sample data. A statistical hypothesis testing is the method of making decisions using sample data.

(ii) Multivariate analysis of variance (MANOVA), extends the analysis of variance to cover cases where there is more than one independent variable to be analyzed simultaneously. In order to explore the spatial variation among different stations and seasonal changes, MANOVA is used to group these on the basis of spatial similarities (Eneji et al., 2012).

(iii) Multivariate Analysis of Covariance (MANCOVA) is an extension of Analysis of Covariance (ANCOVA), which covers more than one independent variable and where the control of continuous independent variables required.

(iv) Principal component Analysis (PCA) is a mathematical tool that uses orthogonal transformation to convert a set of observations of possibly correlated variables called principal components.

(v) Factor Analysis (FA); factor analysis is closely related to PCA. It is a method used to describe variability among observed, correlated variables in terms of the lower number of variables called factors

(vi) Discriminate Analysis (DA) or Canonical Variate Analysis, is a statistical analysis to predict a categorical dependent variable or grouping variable by one or more binary variables independent of continuous variables called predictor variables. It attempts to establish whether a set of variables can be used to distinguish between two or more group classes.

(vii) Cluster Analysis (CA) or Clustering is the grouping of a set of variables in such a way that objects in the same group called cluster are more similar to each other than those of other groups, also called dissimilar clusters.

(viii) Discriminate Analysis (DA) is also like Cluster Analysis used to assess the temporal and spatial variations in the water quality parameters. DA and CA allow a reduction in the dimensionality of the large data set and indicate a few significant parameters that are responsible for most of the variation in water quality.

(ix) Canonical Correlation Analysis (CCA) is an analysis which finds a linear relationship between two or more sets of variables when it is used for two sets of variables it indicates the generalized or canonical version of bivariate correlation. It finds two bases, one for each

variable, that are optimal in terms of their correlations and at the same time it finds the corresponding correlations. CCA technique is applied to determine the relationship between both data sets like air pollution data and meteorological data (Statheropoulos et al., 1998).

## 1.4 Modelling and Monitoring of River Water Quality by ANFIS, ANN and MCS

Water quality models can be effective tools to predict and simulate pollutant flow in water environment, which saves the cost of labour and materials for a large number of chemical experiments to a certain extent. Depending upon the desired conclusion, a simple data based conceptual model or a very complex simulation model is used. The data set for the model need to be following the degree of complexity and accuracy of the flawed model.

If we go for a complex model with a large number of data, the additional data gathering and monitoring campaigns are necessary to run the model. Sometimes, due to the correlation and parameter dependencies, it is not possible to estimate required parameters from the collected data. Hence, some parameters are changed during calibration, validation, training and testing in the model. The aims of models are made to use in continuously changing society, climate, land use etc. and a lot of procedure and considerations are made to run a complex model. Every step of the modelling process should be done with precision and research.

Three models namely Adaptive Neuro Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN) and Monte Carlo Simulations (MCS) are used to describe the input and output relationships of the water quality data. In these studies, for each step, some important points related to model reliability are answered by discussing and applying the method and tools to analyze the behaviour of the model and to prepare actions that are to be taken to reduce error in outputs. The steps of model study are as under:

**Step 1: To Plan for the model study and to select the appropriate model for the study**

It is necessary to decide the type of model that can be fitted to the current state of river pollution, but there are some additional queries like: What model concept should be used in the changing scenario of water environment? / When the model should be used for environmental pollution analysis? To find the answers, to these queries different evaluations are done for different water quality concepts. Sometimes a water quality model is made with little or no available data. In such cases, it is very difficult to decide the processes to be included in the model.

**Step 2: To monitor and to gather the data**

It is important to know the type of input data needed for model calibration as well as for validation to minimize the error in the output results. Regular data are collected from the

selected gauging sites for better prediction and analysis of water quality. The monitored data are validated and tested before they are used as input to the model.

**Step 3: To set-up the model**

During the steps of the modelling process, different precisions and observations are taken to assure minimal error in output results of the model. There should be profound checks of input files and format, performance of the test runs and checking of mass balances. In this step, no additional researches are conducted for modelling approach.

**Step 4: To calibrate and validate the model**

The known data are compared with unknown data in the process of calibration. The calibration always contains the found and the left data after calibration of the model. By the process of validation, it is assured that the model is processed by the input correctly and effectively without performing much error.

**Step 5: To simulate and to evaluate the performance of the model**

Once the selected model is calibrated and validated, the model can be used for further analysis and comparisons between different studies. Simulation is the result of imitation of the proposed system over time. Performance evaluation is a periodic process by which the model performance as well as the output of the model is evaluated. Operational evaluation tests the ability of the model for estimation whereas diagnostic evaluation tests the ability to predict the visibility of the model.

## 1.5 Objectives for this Research

As already discussed in the introduction statistical analysis, modelling and monitoring of water quality parameters are the important approaches for water quality study in river basins. The Brahmani River Basin in Odisha, India is selected as the study area for the proposed research work. The river is found to be the most polluted in Odisha due to industrial effluents of nearby industries, agricultural waste disposal and municipal sewage effluents to the river. Five gauging stations are selected in the river and the monthly water quality data are collected from 2003 to 2012 for the analysis and modelling of water quality. The specific research objectives of the present study are:

1. To perform time series analysis, Trend Analysis and Correlation Analysis of water quality parameters.
2. To perform univariate and Multivariate Analysis of Variances (ANOVA and MANOVA) to investigate the spatial and temporal variations of water quality parameters at five gauging stations.

3. To propose and to study water Quality Index so as the complex dataset into a simplified index that is easily understandable by the general public.
4. To interpret the complex water quality data matrices as well as to identify the possible sources or factors that influence the water quality by using four different analysis methods viz., principal Component Analysis (PCA), Factor Analysis (FA), Discriminate Analysis (DA) and Cluster Analysis (CA).
5. To carry out correlation and Regression Analysis between physical and chemical parameters of water quality by Canonical Correlation Analysis (CCA).
6. To model and simulate as well as to predict the water quality parameters by Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN) and Monte-Carlo Simulations.
7. To compare all the employed models in terms of the predictive ability as well as to carry out the error and correlation analysis of the estimated water quality parameters so as to obtain the most suitable model.

## 1.6 Thesis Outline

Chapter I introduces about the water in the world, its quality, statistical analysis, modelling and monitoring techniques of different software used for further study. It also illustrates the significance and objectives for the proposed work.

Chapter II elaborates on the previous research work done related to water quality, statistical analysis, modelling and prediction of water quality.

Chapter III describes about the study area, its characteristics and available water quality data for the research work.

Chapter IV illustrates about the correlation analysis, statistical models, multivariate analyses, basic learning rules of ANFIS as well as ANN, risk based assessment by MCS and the error analysis for the models used.

Chapter V incorporates the results obtained from research work and analysis done for the water quality modelling.

Chapter VI concludes the research work by providing the summary, important conclusions derived from analysis and modelling of water quality in River basins.

# CHAPTER II

# REVIEW OF LITERATURE

Although the literature covers a wide variety of topics, this literature review presented here will focus on relevant topics which assist this study. The Chapter describes the past research work based on their relevance to the proposed study.

## 2.1 Water Quality and Water Quality Index

EI Kholy et al. (1997) proposed an assessment of the national water quality monitoring program of Egypt. They stated that, the first step towards water quality management was the establishment of a monitoring network. Monitoring in the logical sense, implied watching the ongoing water characteristics and activities in order to ensure that the laws and regulations were properly enforced besides detecting trends for modelling and prediction process. They also emphasized that the design of a network must clearly define the monitoring objectives, and accordingly the necessary simplifying assumptions have to be established. Based on the assumptions made, there were many levels of design that could be applied. Their research presented the process of redesigning the water quality monitoring network of Egypt to produce the national water quality monitoring network using the statistical approach proposed by Sanders and Adrian (1978) which would have the expected confidence interval for the mean value.

Singkran et al. (2010) used Dissolved oxygen, biochemical oxygen demand (BOD), nitrate-nitrogen, total phosphorus, faecal coliform bacteria, and suspended solids to evaluate water quality in the 5 north eastern rivers of Thailand viz., Lam Chi, Lam Pao, Lam Seaw, Loei, and Nam Oon. The mean observed values of the six water quality parameters in each river over a 5-year period (2003–2007) were used to compute the present water quality index ($WQI_{present}$) of each river in both the wet and dry seasons. The mean observed values of the study parameters of each river by season over a 14-year period (1994–2007) were used to build a set of time series models for predicting the values of the associated parameters of each river in the following 5-year period (2008–2012). These mean predicted values were used to compute the $WQI_{future}$ for every season for each river. According to the results, the water quality at many sampling stations was in good condition. This study also revealed that the time series models with the best predictions among the stations were often not of the same type. Several time series models were used and their prediction accuracy values were compared.

Akkaraboyina and Raju (2012) assembled different water quality parameters into a single number which would lead to an easy interpretation of an index, thus providing an important tool for management and decision making purposes. Water quality was represented as the overall water quality at a specific location and specific time based on several water quality parameters. The purpose of this index was to transform the complex water quality data into information that is easily understandable and usable by general public. Eight important water parameters viz., pH, Dissolved oxygen, Electrical Conductivity, Total Dissolved Solids, Total alkalinity, Total Hardness, Calcium and Magnesium were used to estimate WQI during the study period (2009-2012) and future period (2012-2015).

Mangukiya et al. (2012) re-confirmed that Groundwater is a natural resource for drinking water. Hence, like other natural resources, it should also be assessed regularly and people should be made aware of the quality of drinking water. The study was aimed at assessing the water quality index (WQI) for the groundwater of Surat city. For calculating the WQI, the following 13parameters were considered: pH, total hardness, calcium, magnesium, chloride, nitrate, sulphate, total dissolved solids, iron, boron, and fluorides, COD and DO. The calculation of Water Quality Index (WQI) was done by using the Weighted Arithmetic Index method. The statistical analysis in terms of mean, standard deviation (SD), correlation and regression of obtained data were carried out using Microsoft Office Excel 2007. The results of analyses were used to suggest models for predicting water quality. Their analysis revealed that the groundwater of the area needed some degree of treatment before consumption.

## 2.2 Multivariate Analysis of Variance

Chakrabarty and Sarma (2011) analysed drinking water quality with respect to parameters like Temperature, pH, Electrical conductivity, Total Solid (TS), Total Dissolved Solids(TDS), Total Suspended Solids(TSS), Turbidity, Dissolved Oxygen (DO), Total Hardness(TH), Calcium Hardness (CH), Magnesium Hardness(MH), Chloride (Cl), Sulphate(SO4), Sodium (Na) and Potassium(K) in Kamrup district of Assam, India. Forty six different sampling stations were selected for the study. Statistical analysis of the data was presented to determine the distribution pattern, localization of data and other related information. Statistical observations implied non-uniform distribution of the studied parameters with a long asymmetric tail either on the right or left side of the median. Descriptive statistics in the form of mean, variance (V), standard deviation (SD), standard error (SE),median, range of variation, and percentile at 95%, 75% and 25% (P95%, P75%, P25%) were calculated and summarized.

Eneji et al. (2012) investigated the spatial and temporal variation in water quality parameters at ten different locations along River Benue in Nigeria for twelve consecutive months. In order to explore the spatial variation among different stations and seasonal changes,

multivariate analysis of variance (MANOVA) was used to group these data on the basis of spatial similarities. Discriminate analysis used in the study identified all the parameters to discriminate between the three seasons of a year with 99.2% correct accuracy assignations. Discriminate function analysis would enable then to predict the likely season a water sample from the metropolitan area of Makurdi in Nigeria was collected given the values of the water quality parameters.

Saatsaz et al. (2013) evaluated spatio-temporal distributions of groundwater quality were evaluated for 23 different stations in the plain using multivariate statistical techniques. After descriptive analysis, Multivariate Analysis of Variance technique (MANOVA) and Cluster analysis (CA) were performed to measure significant effects of spatial, seasonal and annual differences on mean concentration of key hydro chemical parameters of groundwater. The MANOVA results explained that the interaction of location on seasonal variables was significant to increase the variations. In addition, the results of cluster analysis showed a 3-cluster dendrogram which reflects variations in natural and human activities.

## 2.3 Multivariate Statistical Analysis

Shlens(2003)confirmed that the Principal component analysis (PCA) was a mainstay of modern data analysis as a black box, that was widely used but poorly understood. The objective of this study was to dispel the magic behind this black box concept. This study also focused on building a solid intuition regarding how and why the principal component analysis would work. Furthermore, it also crystallized this knowledge of maths behind PCA by deriving it from the first principles. The hope was that by addressing both aspects, readers of all levels will be able to gain a better understanding of the power of PCA as well as the knowledge regarding when, how, where and why one can apply this technique.

Simeonov et al. (2003) applied different multivariate statistical approaches for the interpretation of a large and complex data matrix obtained during a monitoring program of surface waters in Northern Greece. The dataset was treated using cluster analysis (CA), principal component analysis and multiple regression analysis on principal components. CA showed four different groups of similarity between the sampling sites reflecting the different physicochemical characteristics and pollution levels of the studied water systems. A multivariate receptor model was also applied for source apportionment estimating the contribution of identified sources to the concentration of the physicochemical parameters. CA, principal component analysis (PCA) and source apportionment by multiple regression analysis on principal components (PC/MR) were employed in a dataset of almost twenty thousand values. Missing data were completed by mean values of the neighbour data. The STATISTICA 5.0 software package was employed for data treatment.

Boyacioglu et al. (2005) used the factor analysis technique to large water quality data sets in Buyuk Menderes River Basin, Turkey to analyse the surface water contamination and determining the correlations between water quality parameters. The correlation coefficients were also evaluated and presented in matrix format. Factor analysis explained in a better way, the structure of underlying system which produced the water quality data. On 2006 they proposed the application of factor analysis technique to surface water quality data sets obtained from the Buyuk Menderes River Basin, Turkey, during two different hydrological periods. Here the water quality was assessed separately for summer (low flow) and winter (high flow) periods in understanding the main pollutants, their sources and also determining priorities to improve water quality in two different hydrological periods. It was suggested that the high-flow period might have positive effects with dilution of surface water by rain and storm water. On the other hand, low flow runoff water had increased pollutant concentration leading to a decrease in the quality.

Ouyang et al. (2006) assessed seasonal changes in surface water quality for evaluating temporal variations of river pollution due to natural or anthropogenic inputs of point and non-point sources. In this study, surface water quality data for 16 physical and chemical parameters were collected from 22 monitoring stations in lower St. Johns River, Florida, USA during the years from 1998 to 2001 were analyzed. Principal component analysis technique was employed to evaluate the seasonal correlations of water quality parameters, while the principal factor analysis technique was used to extract the parameters that were most important in assessing seasonal variations of river water quality.

Zhou et al. (2006) used cluster analysis (CA) and discriminant analysis (DA) to assess temporal and spatial variations in the water quality of the watercourses in the North western New Territories of Hong Kong, over a period of five years (2000–2004) using 23 parameters at 23 different sites (31,740 observations). Hierarchical CA grouped the 12 months into two periods numbered as (the first and second periods) and classified the 23 monitoring sites into three groups viz., (Group A, Group B, and Group C) based on similarities of water quality characteristics. DA provided better results with great discriminatory ability for both temporal and spatial analysis. DA also proved to be an important tool for data reduction because it only used six parameters. DA allowed a reduction in the dimensionality of the large dataset and indicated a few significant parameters that were responsible for most of the variations in water quality.

Shrestha et al. (2008) applied multivariate statistical techniques, such as principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA), for the evaluation of temporal/spatial variations and the interpretation of a large complex water quality dataset of the Mekong River in Asia using datasets generated during 6 years (1995–2000) of monitoring

of 18 parameters (16,848 observations) at 13 different sites. Discriminant analysis showed the best results for data reduction and pattern recognition during both spatial and temporal analysis. Spatial DA revealed 8 parameters (total suspended solids, calcium, sodium, alkalinity, chloride, iron, nitrate nitrogen, total phosphorus) and 12 parameters (total suspended solids, calcium, sodium, potassium, alkalinity, chloride, sulphate, iron, nitrate nitrogen, total phosphorus, silicon, dissolved oxygen) were responsible for significant variations between monitoring regions and countries, respectively. Thus, this study illustrated the usefulness of principal component analysis, factor analysis and discriminant analysis for the analysis and interpretation of complex datasets and in water quality assessment, identification of pollution sources/factors, and understanding of temporal and spatial variations of water quality for effective river water quality management.

Li et al. (2009) also used cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA) to evaluate temporal and spatial variations and to interpret a large and complex water quality datasets collected from the Songhua River Basin, China. The data sets, which contained 14 parameters, were generated during the 7-year period (1998-2004) monitoring program at 14 different sites along the river. Three significant sampling locations viz., (less polluted sites, moderately polluted sites and highly polluted sites) were detected by CA and five latent factors viz., (organic, inorganic, petrochemical, physiochemical, and heavy metals) were identified by PCA and FA. The results of DA showed only five parameters and eight parameters were necessary in temporal and spatial variations analysis, respectively.

Zhao et al. (2009) analyzed the characteristics of surface water quality providing assistance in the reconstruction of an old water treatment plant using multivariate statistical techniques such as cluster analysis and factor analysis to the data of Yuqiao on the Luan River, China. The results of cluster analysis demonstrated that the months of a year were divided into 3 groups and the characteristic of clusters were matching with the seasonal characteristics in North China. Three factors were derived from the complicated dataset using factor analysis. The dataset was normalized for cluster analysis. Kaiser-Meyer-Olkin (KMO) and Bartlett's test were performed to examine the suitability of the data for principal component analysis/factor analysis. The hierarchical cluster analysis (HCA), which was performed on the normalized data matrix by the utilization of the Ward's method and this method, used the squared Euclidean distances as a measure of similarity that was reported as Dlink/Dmax, and was applied to the variables using Minitab 15 software (Minitab Inc.) to group the data in temporally suitable pattern.

Fan et al. (2010) used Principal component analysis (PCA) and cluster analysis (CA) to identify characteristics of water quality and to assess its spatial pattern in the delta of Pearl

River region, China. Hierarchical agglomerative CA was performed on the normalized data set by means of the Ward method, using Euclidean distances as a measure of similarity. The Euclidean distance gave the similarity between two samples and a distance which could be represented by the difference between analytical values from both the samples.

Jianqin et al. (2010) evaluated water quality were important because it could provide guidance when determining water utility. But many interacting impact factors were involved in water quality evaluation systems, making water quality evaluation difficult. Principal component analysis (PCA) was widely used in water quality evaluation because it could eliminate the correlation among factors. However, PCA ignored the degree of data dispersion, which was considered by information entropy (IE). To solve this problem, a model combined PCA and IE methods to obtain the weights of indicators was proposed and the proposed model was applied to assess the reused water quality of Jinshui River in Zhengzhou City, China in 2009.

Liping et al. (2010) assessed on water quality condition of Wen Yu River basin in Beijing, China. According to stationing principle and field investigation, the whole basin was divided in to 22 monitoring sections responsible for measuring 10 pollution indicators. By means of Statistical Package for the Social Sciences (SPSS) software applied along with principal component analysis method, they analysed on the main pollution indicators and the main pollution contributing sections. Then they attempted to solve the formula of the four extracted principal components viz., $F_1$, $F_2$, $F_3$, and $F_4$. According to the contribution percentage of variance, the function of comprehensive evaluation expression could be deduced as $F= 0.692F_1+0.125F_2+0.106F_3+0.077F_4$.

Mishra (2010) adopted the multivariate statistics approach for interpretation of large and complex data matrix obtained during the water quality monitoring of the River Ganga in Varanasi, India. 16 physicochemical and bacteriological variables were analysed. The dataset was treated using Principal Component Analysis (PCA) to extract the parameters that were the most important in assessing the variation in water quality. Four Principal Factors were identified as responsible for explaining 90% of the total variance of the dataset.

Noori et al. (2010) proposed the determination of principal and non-principal monitoring stations was carried out using principal component analysis (PCA) technique for the Karoon River, Iran. Also in this study a canonical correlation analysis (CCA) was used to determine relationship between physical and chemical water quality parameters. Water quality parameters including $BOD_5$, COD, EC, $NO_3^-$, $SO_4^{-2}$, temperature, $Cl^-$, DO, hardness, TDS, pH, and turbidity were measured in samples collected from 17 stations along Karoon River from 1999 to 2002. Four of these monitoring stations which proved less effective in

explaining the annual variation in the river water quality were removed. Further investigations indicated that all water quality parameters were important.

Kumar et al. (2011) applied multivariate statistical techniques to water quality dataset collected from Sarda Sagar Reservoir. The results revealed the usefulness of multivariate techniques for evaluation of large and complex water quality dataset for the effective management of water resources. The analysis results showed that the number of sampling sites and the sampling months could be reduced towards an optimum. This reduced set of sites and duration could be monitored over larger areas within the watershed to provide more detailed spatial information about sources and processes.

Salah et al. (2012)used multivariate statistical method including cluster analysis (CA) to assess temporal and spatial variations in the water quality of Euphrates River, Iraq, during a period from 2008-2009 using 16 parameters at 11 sampling sites. Hierarchical CA grouped the 8 months into three periods (I, II and III) and classified the 11 sampling sites into two groups (I and II) based on similarities of water quality characteristics. The temporal pattern of the dataset showed that April has higher pollution level relative to the other months. Spatially, sampling site no 7 ($S_7$) hada lower pollution level while the other sampling sites had higher pollution levels. Thus, this study showed usefulness of cluster analysis for analyzing and interpreting of surface water dataset to assess the temporal and spatial variations in the water quality parameters and the optimization of regional water quality sampling network.

## 2.4 Water Quality models

Basil et al. (2001) found that conventional uncertainty analysis by the Root Sum Square (RSS) method was often difficult in complex systems and required approximation at each stage of processing placing serious doubts on the validity of the results. They observed that recent developments in the analysis of uncertainty using Monte Carlo Simulation (MCS) had resolved many of the problems. These included non-symmetric uncertainty distributions, non-linearity within the measurement system, input dependency and systematic bias. Monte Carlo simulation was devised as an experimental probabilistic method to solve difficult deterministic problems since computers can easily simulate a large number of experimental trials that had random outcomes. When applied to uncertainty estimation, random numbers were used to randomly sample parameters' uncertainty space instead of point calculation carried out by conventional methods.

Juahir et al. (2004) discussed the development of Artificial Neural Network (ANN) model in estimating water quality index (WQI). An ANN model was developed and tested using data from 30 monitoring stations. The modelling data was divided into two sets. In the first

dataset, ANNs were trained, tested and validated using six independent water quality variables as input parameters. Consequently, Multiple Linear Regression (MLR) was applied to eliminate independent variables that exhibited the lowest contribution in variance. MLR was applied in the work to justify the relationship between water quality parameters and their impact on WQI. In second dataset, only four independent variables were used to train, test and validate the ANNs. ANN models were found to be capable of estimating WQI with acceptable accuracy when they were trained by eliminating the independent variables.

Khandelwal and Singh (2005) attempted to predict the chemical parameters like sulphate, chlorine, chemical oxygen demand, total dissolved solids and total suspended solids in mine water using artificial neural network (ANN) by incorporating the pH, temperature and hardness. The prediction by ANN was also compared with Multivariate Regression Analysis (MVRA). Feed forward network was adopted for the network architecture. Closer mapping gave better performance of the network. The purpose of MVRA was to learn about the relationship between several independent or predictor variables and dependent criterion variable.

Alexandridis (2007) discussed the usefulness of Monte Carlo simulation as an analysis tool aiming to capture the properties and patterns of change for sequences of events, and to generate scenarios and classifications of water quality change (WQC). For this analysis, the Crystal Ball Monte-Carlo simulation framework was used. Measuring and predicting probabilities of events allowed them to avoid many important pitfalls on modeling the dynamics of whole systems (e.g., a river system or a water distribution system) across various spatial and temporal scales. The heterogeneity of the water quality parameters used to assess the suitability and threshold values of the water, would otherwise require a very large variety of models which would use various stochastic and dynamic equations, mass conservation equation, momentum and energy conservation, thermodynamic equilibrium equations etc.

Najah et al. (2009) evaluated water quality in Johor River, Malaysia and discussed measures to develop better water resources management plan. They found that classical process-based modelling approach could provide relatively good prediction for water quality parameters. However, those models relied on large data and required lot of input data that were often unknown. New approaches such as Artificial Intelligence techniques had proven their ability and applicability for simulating and modelling various physical phenomena in the water resources engineering field.

Yan et al. (2010) used an adaptive neuro fuzzy inference system (ANFIS) for classifying water quality status of Rivers in China. A data set was collected from 100 monitoring stations in all major river basins in China and used for training and validating the model. ANFIS is a

multilayer feed-forward network that is generally used in neural network learning algorithms and fuzzy logic to map an input space to an output space. The performance was more competitive when compared with artificial neural networks. It was applied in evaluation and classification of water quality status.

Sahu et al. (2011) found that the groundwater near mines was contaminated heavily with acidity, alkalinity, toxicity, heavy minerals, and microbes. Evaluation of water quality index (WQI) of groundwater was done in urban areas close to mines to prepare for make remedial measures. They proposed an efficient methodology such as adaptive network fuzzy inference system (ANFIS) for the prediction of water quality. The parameters used to assess water quality were usually correlated and this made the assessment unreasonable. Therefore, the parameters were uncorrelated using principal component analysis with varimax rotation. The uncorrelated parameters values were fuzzified to take into account the uncertainty and impreciseness during data collection and experimentation. An efficient rule base and optimal distribution of membership function was constructed from the hybrid learning algorithm of ANFIS. The model performed quite satisfactorily with the actual and predicted water quality.

Areerachakul (2012) employed several techniques such as Fuzzy Inference System (FIS) and Neural Network (NN) for developing predictive models to quantify water quality. The main objective was to compare the predictive ability of the Adaptive Neuro-Fuzzy Inference System (ANFIS) model and Artificial Neural Network (ANN) model to estimate the Biochemical Oxygen Demand (BOD). The model performance was expressed in terms of observed and predicted values of the correlation coefficient and the root mean square error. ANFIS was trained with the help of MATLAB version 7.8 (2009).

Galavi et al. (2012) used Artificial intelligence (AI) based models in hydrological forecasting. Although, there was a common network structure among ANFIS models, there was no one-fits-all ANFIS architecture for every case. Moreover, it was discussed that in many application, theory did not guide in model building process by either suggesting the relevant model input variables or correct functional form and model configuration.

Khalil et al. (2012) studied the quality of a water body in the Nile Delta in Egypt usually characterized by sets of physical, chemical, and biological parameters, which were mutually interrelated. Correlation patterns were found between water quantity and water quality parameters at the same location, or among water quality parameters within a monitoring location or among locations. Serial correlation was also detected in water quality variables. Through their investigation of the level of information redundancy, assessment and redesign of water quality monitoring network they aimed to improve the overall network efficiency and cost effectiveness. The potential of the Artificial Neural Network (ANN) on simulating

interrelation between water quality parameters was also examined. Several ANN inputs, structures and training possibilities were assessed and the best ANN model and modelling procedure was selected. The prediction capabilities of the ANN were compared with the linear regression models with auto correlated residuals, usually used for this purpose. It was concluded that the ANN models were more accurate than the linear regression models having the same inputs and output.

Mahapatra et al. (2012) observed that ground water was contaminated heavily with due to population growth, urbanization and industrialization. Hence, evaluation of water quality of groundwater was done to prepare for remedial measures. An empirical approach was applied for classification of water samples based on 10 quality parameters. Q-mode principal component analysis was applied to classify the water samples into four different categories considering parameters such as pH, DO, turbidity, TDS, hardness, calcium ion ($Ca^{++}$), chloride ion ($Cl^{-}$), BOD, iron ($Fe^{++}$), sulphate ($SO^{--4}$ ). This classification was found to be useful for the planners and field engineers for taking ameliorative measures in advance for preventing the contamination of groundwater. The non-parametric method proposed was efficiently assessed the water quality index for classification of water quality. The model could also be used for estimating water quality on-line but the accuracy of the model would depend upon the judicious selection of parameters.

Jiang et al. (2013) observed that there was always presence of uncertainty in any water quality risk assessment. A Monte Carlo simulation (MCS) was found to be a flexible, efficient method for characterizing such uncertainties. However, the required computational effort for MCS-based risk assessment was great, particularly when the number of random variables was large and the complicated water quality models had to be calculated by a computationally expensive numerical method, such as the finite element method (FEM). To address this issue, this study presented an improved method that incorporated an artificial neural network (ANN) into the MCS to enhance the computational efficiency of conventional risk assessment. The conventional risk assessment used the FEM to create multiple water quality models, which could be time consuming or cumbersome. ANN model was used as a substitute for the iterative FEM runs, and thus, the number of water quality models that need to be calculated can be dramatically reduced. Chemical oxygen demand (COD) pollution risks in the Lanzhou section of the Yellow River in China was taken as a reference. Compared with the conventional risk assessment method, the ANN-MCS based method could save much computational effort without the loss of accuracy. The results showed that the proposed method was more applicable to assess water quality risks.

# CHAPTER III

# THE STUDY AREA AND DATA COLLECTION

## 3.1 General Description of the Study Area

The Study area of the proposed research work is Brahmani River Basin. The basin is an interstate basin which lies between latitudes of $20^0$ 28' North to $25^0$ 35' North and longitudes of $80^0$52' East to $82^0$ 30' East. The river is formed by two major tributaries namely Sankh and Koel, which originate in the state of Jharkhand and the Basin spreads across the states of Chhattisgarh, Jharkhand and Odisha. The river gets its name below the confluence point of Sankh and Koel at Vedvyas in Sunderagarh district of Odisha. The river splits into numerous channels, criss-crossing the channels of Baitarani River and finally merges to Bay of Bengal. The length of the river is 446 km. The map of Brahmani River is shown in Figure 3.1.

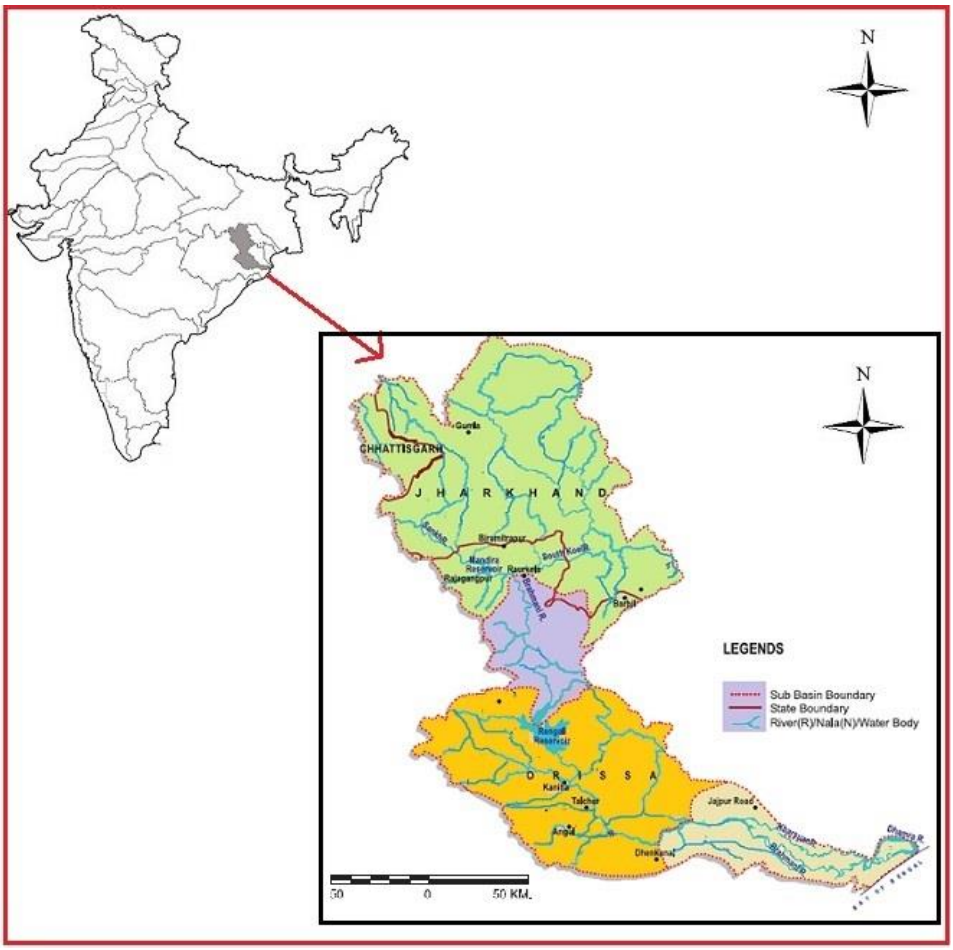


**Figure 3.1: Study Area showing the Brahmani River Basin**

## 3.1.1Brahmani River Basin

The River basin has a total drainage area of 39,268 km$^2$, out of which 22,516 km$^2$ is in Odisha state, 15,405 km$^2$ in Jharkhand state and 1,347 km$^2$ in Chhattisgarh state. In Odisha, eight districts are covered by the river basin viz., Sundergarh, Keonjhar, Sambalpur, Deogarh, Angul, Dhenkanal, Jajpur and Kendrapada. The river is formed by two principal tributaries viz., Koel and Sankh, which originates in the state of Jharkhand. The river referred as Brahmani River at the confluence point nearVedvyas, in Odisha at an elevation of 200m above mean sea level. Below the confluence point, the river heads its way to southeast direction up to Bay of Bengal and traverses a length of 461 km. Below Jenapur station, the Brahmani bifurcates into two major deltaic branches namely Kimiria and Kharsuan. The two deltaic branches join at the downstream almost at a distance of one hundred kilometres. The river receives flood flows from Baitarani River before discharging finally into Bay of Bengal near Dhamra.

Five gauging stations in the state of Odisha are selected for the proposed study in the Brahmani River Basin. Those are Panposh down-stream at Sundergarh, Talcher up-stream at Angul, Kamalanga downstream at Dhenkanal, Aul and Pottamundai stations are included in Kendrapada district in Odisha. Panposh down-stream includes 5,717.77 km$^2$, area Talcher up-stream includes 4,235.38 km$^2$, Kamalanga down-stream includes 3,968.66 km$^2$, and Aul and Pottamundai together include 1,114.41km 2 areas within the River basin. The synoptic view of the river flow is shown in Figure 3.2 and the river system along with five gauging stations of Brahmani River are shown in Figure 3.3.


**Figure 3.2: Synoptic View of the Brahmani River Basin (Source: Google Map)**

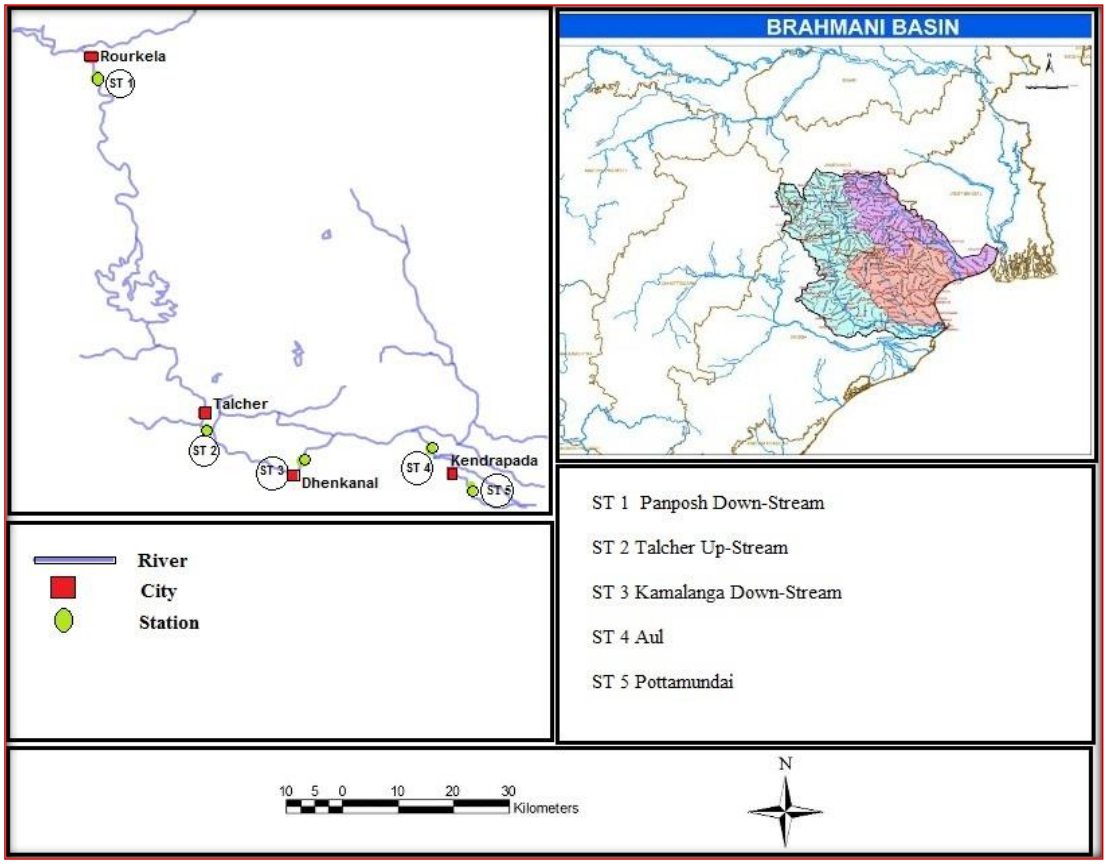


**Figure 3.3: Brahmani River Basin along with five Gauging Stations**

## 3.1.2 Climate and Rainfall

The river basin is under a tropical monsoon climate zone. The three well marked seasons in the river basin are winter, summer and rainy season. The climate close to the coastline, is somewhat affected by sea whereasin the elevated areas of hills and plateaus climate is cooler due to the altitude effect. The minimum annual temperature is $4^0$C in winter and maximum annual temperature is $47^0$C in summer.

The monsoon months include June to Septemberhaving normal annual rainfall of 1305 mm with a minimum of 969 mm and a maximum of 1574 mm. The water deficit in the basinis moderate from the months of November to February. However, during the months of March to May, the soil moisture deficit becomes large making it very difficult for shallow rooted crops and vegetation to survive.

## 3.1.3 Soils

Soil in the Brahmani river basin can be classified into two groups based on soil formation namely residual and transported soil. The upper Basin of river is grouped under red gravel, red earth and yellow soil. The central region of river basin goes under mixed red and black loams, where as the lower basin grouped under red loam lateritic and lateritic soils. The delta region

of Brahmani river basin goes under alluvial soil. The soil map of India showing the study area is shown in Figure 3.4.

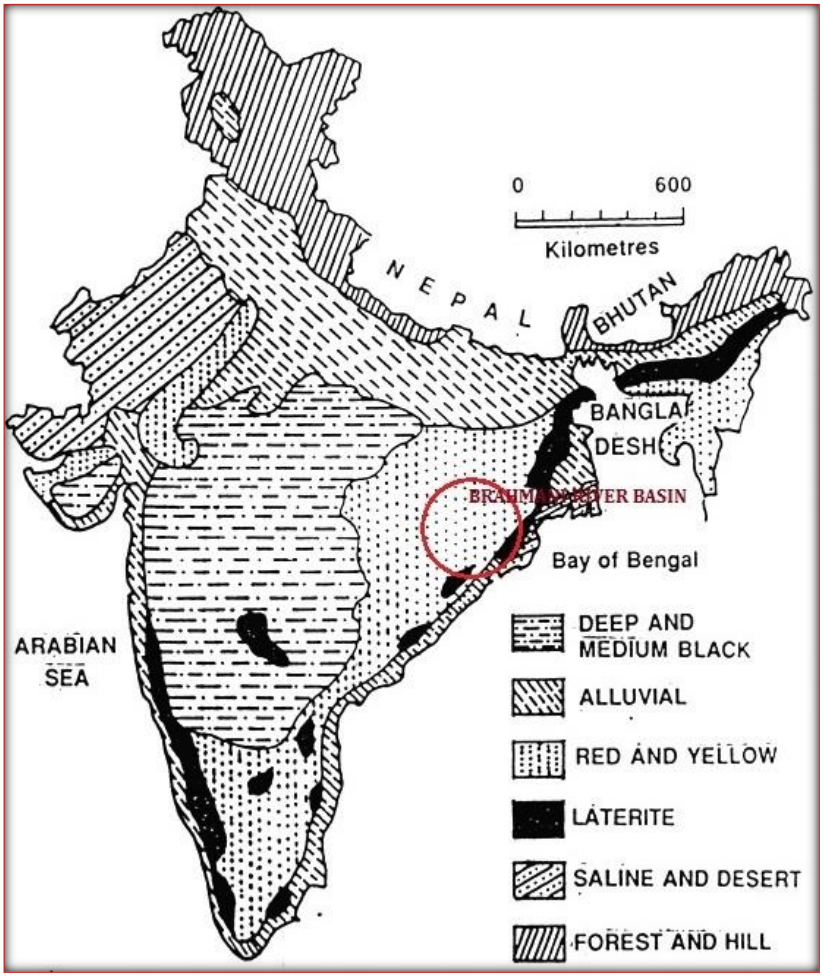


**Figure 3.4: Soil Map of India with area around the Brahmani River Basin**

## 3.1.4 Land Uses

The river basin is rich in forest cover with 38% of the total basin as forest area, whereas 55% of the total basin area is covered by agricultural fields constituting the main source of livelihood and income.The total geographical area of the river basin is 39,268 $km^2$, out of which the river basin forest covers 15,101 $km^2$; land under reservoirs cover an area of607 $km^2$, cultivable land includes 21,805 $km^2$.The agricultural landspread around each of the five gauging stations viz.,Panposh down-stream, Talcher up-stream, Kamalanga down-stream, Aul and Pottamundai are 2,589 $km^2$, 1,556 $km^2$, 1,815 $km^2$ and 322 $km^2$ and 324 $km^2$ respectively. The land use data is verified by Indian Remote Sensing (IRS)1D,Linear Image Self Scanner (LISS) III, satellite imagery maps collected from National Remote Sensing Agency, Hyderabad (NRSA). Figure 3.5 shows the imagery maps of Land use by LISS III data base.

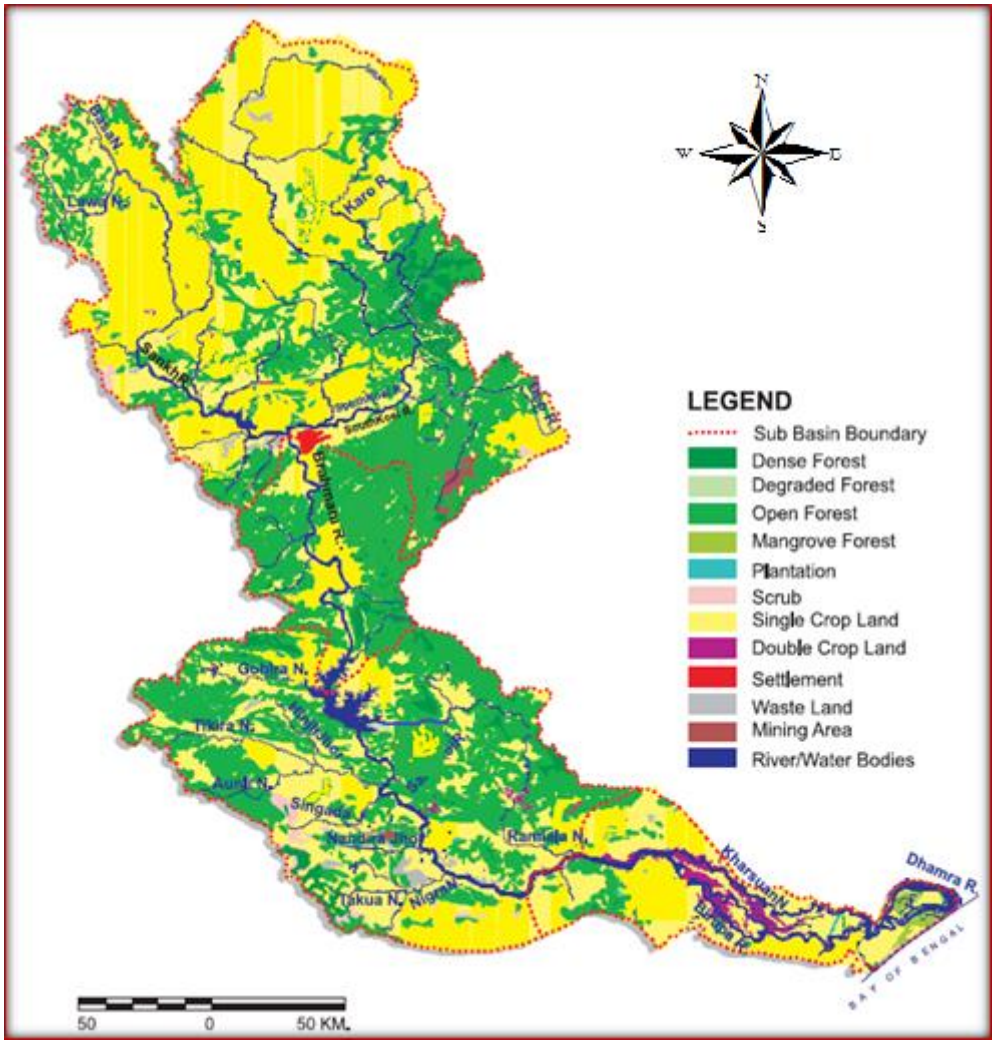


**Figure 3.5: Land Use Map of Brahmani Basin (Source: Central Water Commission, Bhubaneswar)**

## 3.1.5 Water Resources

The estimated annual renewable water resources in the river basin are 21,920 million cubic meters (MCM), which includes both surface and ground water resources. The ground water is mostly used for irrigation purpose from the total ground water recharge of 5,171 MCM.

## 3.1.6 Irrigation Uses

The accelerated irrigation development has led to a number of medium and major irrigation projects in the river basin including the Reangali Multipurpose Dam Project. Several other diversion structures for irrigation and water supply are made at Jenapur in the delta region of the river basin. Rice is the main crop grown in the basin along with other crops such as wheat, millets, pulses, groundnut, mustard, Ragi, maize etc.

### 3.1.7 Population and Urban Growth

The total population of the basin area is 9.547 million according to census 2011, of which 2.252 million is urban population and 7.295 million is rural population. The livestockpopulationis 3.928 million in the basin area. The basin has witnessed rapid growth of towns and total urban population which has led to extraction oflarge mineral wealth and consequent industrial growth. A number of towns including Rourkela Steel City, Talcher Thermal Complex, Angul, Rajgangpur, Dhenkanal, Birmitrapur have received a great impetus for the growth of industrial setups, which has led to the growthsocio economic activities.

### 3.1.8 Industries

An integrated steel plant at Rourkela in Sundergarh district includes associated mines, Ancillaries' by-products and downstream product units which has led to a wide scale industrialization of the area. The river basin with rich minerals and cheap labor offer an ideal ground for industrial growth. Sundergarh, Angul and Dhenkanal districts have a major share of industries. The Angul-Talcher complex has a significant water requirements, wastewater generation and environmental implications due to the consequent growth of industries in a few towns of the area. The growth of consumer goods and agricultural product processing units mostly forRice and Oil start growing fast due to growing demand for their product in this region.

### 3.1.9 Flood Management and Drainage

During floods, the River turns into a large turbulent channel with a potential threat to life and property of the population in the basin area. The worst flood was observed on 20$^{th}$ August 1975 with 24, 246 $m^3$/S with gauge level of 24.78 m at gauge site against the danger level of 23 m. The entire flood spill of the River Brahmani continued to the sea over a 10 to 20 km wide and 70 km long flat flood plain. Brahmani River bifurcates below Jenapur and consequently conveys 60 to 70 percent of the discharge to Baitarani and Mahanadi, The entire delta area of 3500 $km^2$ is significantly flood prone. But to protect the densely populated Aul area, a 70 km long ring bund is constructed blocking a part of the flood plain and protecting 25000 ha of agricultural land and 1,50,000 of population. The annual maximum flood discharge was recorded at Jenapur. The flood plain of 1500 $km^2$ is within the delta area experiences flooding of up to 1 to 2 m depth. Rengali Dam has moderated the high floods significantly and reduces the submersion of agricultural land during the kharif crop season. Flood inundation areain the delta region of river Brahmani is shown in Figure 3.6.


**Figure 3.6: Flood prone area in the Delta region of Brahmani River**

## 3.2 Data Collection and Analysis

The monthly water quality parameters are collected and analyzed from five selected gauging stations of Odisha during the months of January to December from 2003 to 2012. The complete data sets are divided into three seasons' viz., summer, monsoon and winter season. Summer season includes the month of March, April, May, June. The Monsoon season includes the month of July, August and September. The winter season Include the month of October, November, December, January and February. Eleven physical, chemical and biologicalwater quality parameters are selected for the analysis.The parameters are pH, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Electric Conductivity, Nitrogen as nitrate (Nitrate-N), Total Coliform (TC), Fecal Coliform (FC), Chemical Oxygen Demand (COD), Nitrogen as ammonia ($NH_4$-N), Total Alkali as $CaCO_3$ (TA as $CaCO_3$)and Total Hardness as $CaCO_3$ (TH as $CaCO_3$). The collected water quality parameterfrom five gauging stations viz., Panposh down-stream, Talcher up-stream, Kamalanga down-stream, Aul and Pottamundai data are tested, analyzed and validated with the data of Central Water Commission, Bhubaneswar, Odisha (CWC) and Odisha Pollution Control Board, Bhubaneswar, Odisha (OPSC) in the Environmental laboratory of National Institute of Technology, Rourkela.

### 3.2.1 pH

pH is a measure of acidity or alkalinity of water, pH is an important limiting factor for aquatic life as well as for domestic uses. Pure distilled water is neutral with a pH of 7. It is expressed in a scale which ranges from 1 to 14. The geology and soils of the catchment largely determine the pH of stream waters under base flow conditions. Photosynthesis by aquatic plants and algae can cause significant variations in pH.Excessive growth of algae and in-stream aquatic plants can lead to elevated pH at certain times of the day. Industrial wastewater or contaminated storm

water can cause significant changes to either acidic or alkaline conditions.A pH range of 6.5 – 8 is acceptable for freshwater. A range of 8 – 9 is optimal for estuary and sea water. The monthly average values of pH are analyzed from the five selected gauging stations of Brahmani River basin during 2003 to 2012 and are represented graphically in Figure 3.7.

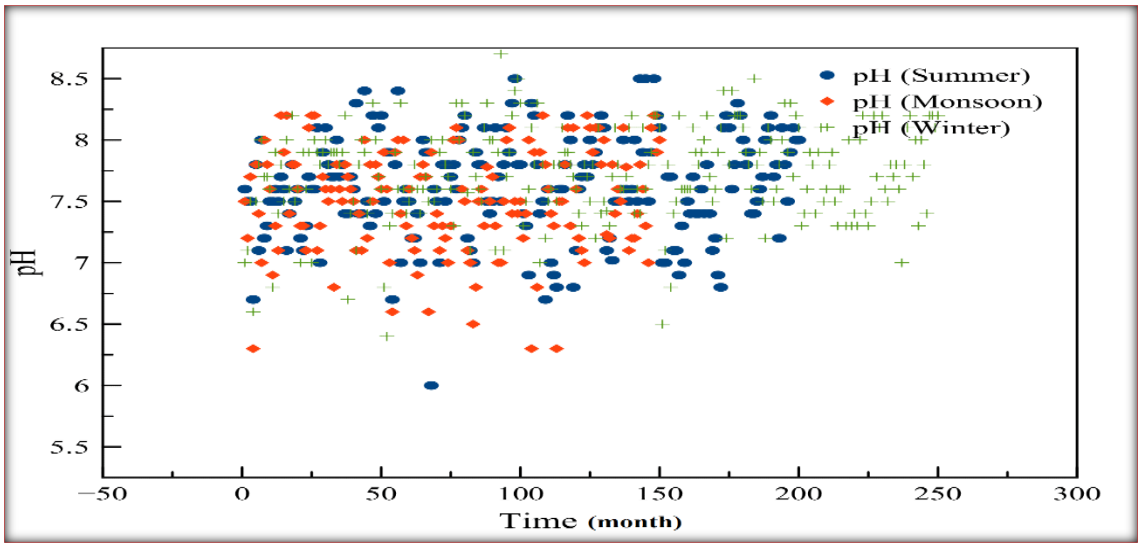


**Figure 3.7: Temporal Variations of monthly pH data from 2003to 2012**

## 3.2.2 Dissolved Oxygen (DO)

The amount of oxygen in water to a degree shows its overall health. Dissolved Oxygen (DO) is a measure of quantity oxygen in milligrams per liter present of water. Dissolved Oxygen saturation is vital for aquatic organisms. Sewage effluent, decaying aquatic vegetation, contaminated storm water discharges and wastewater from human activities all reduce DO levels as they are decomposed by micro-organisms present in river water. River water that has adequate levels of DO can usually sustain a diverse aquatic community.. In general, DO level of 3 mg/L are stressful to most aquatic organisms. Water with low DO from 0.5-2 mg/L is considered hypoxic and waters with less than 0.5 mg/L are anoxic. The temporal variation of DO values are graphically represented in Figure 3.8.

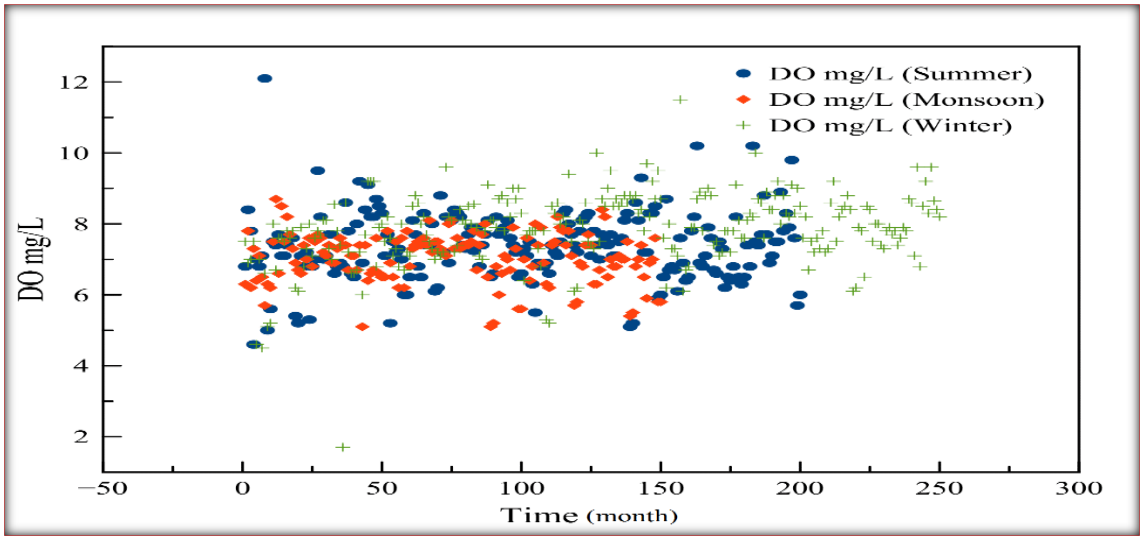


**Figure 3.8: Temporal Variations of DO values from 2003 to 2012**

### 3.2.3 Biochemical Oxygen Demand (BOD)

BOD is a measure of the amount of oxygen used by biological and chemical processes in a stream of water over a 5-day. BOD is calculated by measuring the oxygen level of the water on collection and then 5 days after storage in the dark at a constant temperature of $20^0$ C. The difference between DO and BOD is the demand or consumption of oxygen by chemical and biological process. The BOD is measured in milligram per liter of water. Unpolluted and natural waters should have a BOD of 5 mg/L or less. Raw sewage may have BOD levels ranging from 150-300 mg/L. The BOD values collected from 2003 to 2012 in five selected gauging stations of the Brahmani Riverare graphically shown in Figure 3.9.

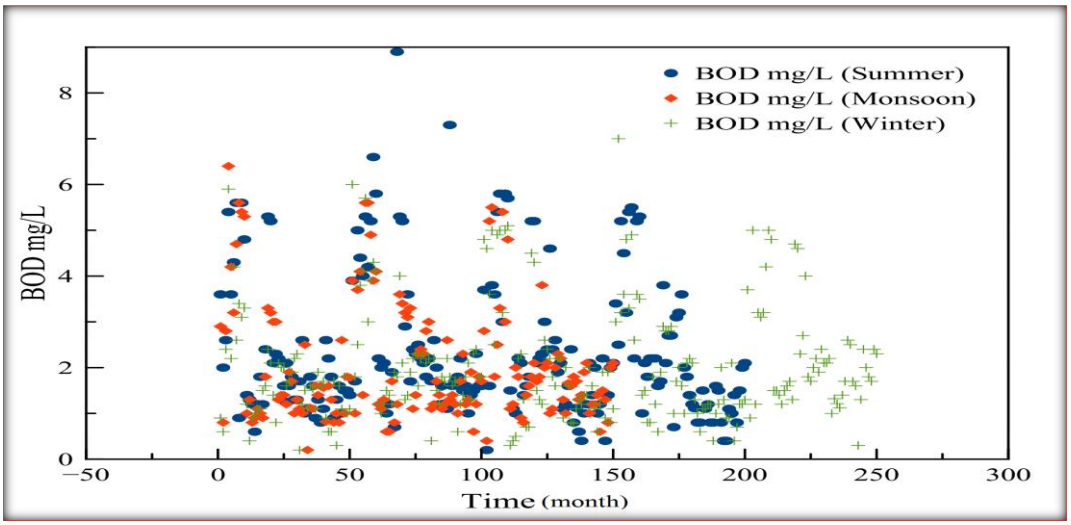


**Figure 3.9: Temporal Variations of BOD values from 2003 to 2012**

### 3.2.4 Electrical Conductivity (EC)

Electrical Conductivity (EC) is a measure of the ability of water to pass electric current through it in water is affected by the presence of dissolved solids such as chloride, nitrate, sulphide, phosphate, sodium, magnesium, calcium, iron and aluminium. Electrical Conductivity also is affected by water temperature with warmer the water, the higherwould be the EC. On the other hand streams that run through areas with clayey soils tend to have higher conductivity because of presence of materials that ionize when washed into water. The basic unit of measurement of EC is micro mho per centimetres or micro Siemens per centimetre. Distilled water has EC in a range of 0.5 to 3µmho/cm. The conductivity in rivers ranges from 100 to 1000 µmho/cm. The streams mixed with industrial waters can range as high as 10,000µmho/cm.Temporal variation of conductivity in Brahmani River is graphically represented in Figure 3.10.

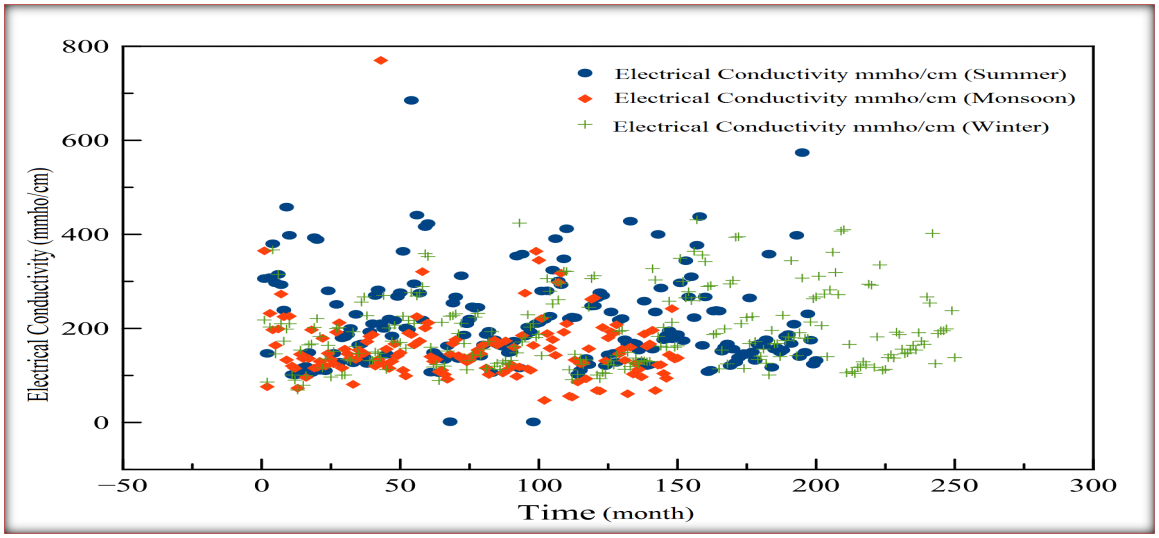


**Figure 3.10: Temporal Variation of Electrical Conductivity from 2003 to 2012.**

### 3.2.5 Nitrogen as Nitrate

Nitrate generally occurs in trace quantities in river water. Nitrate is a nutrient for plant growth that dissolves in water. Common sources of nitrate include commercial and manure based fertilizers, waste water treatment plants and faulty septic systems. Nitrate affects the health of fish, plants and other life in rivers. The drinking water should have a nitrate level of 45-100 mg/L. The nitrogen as nitrate (Nitrate-N) values from 2003 to 2012 is represented graphically in Figure 3.11.

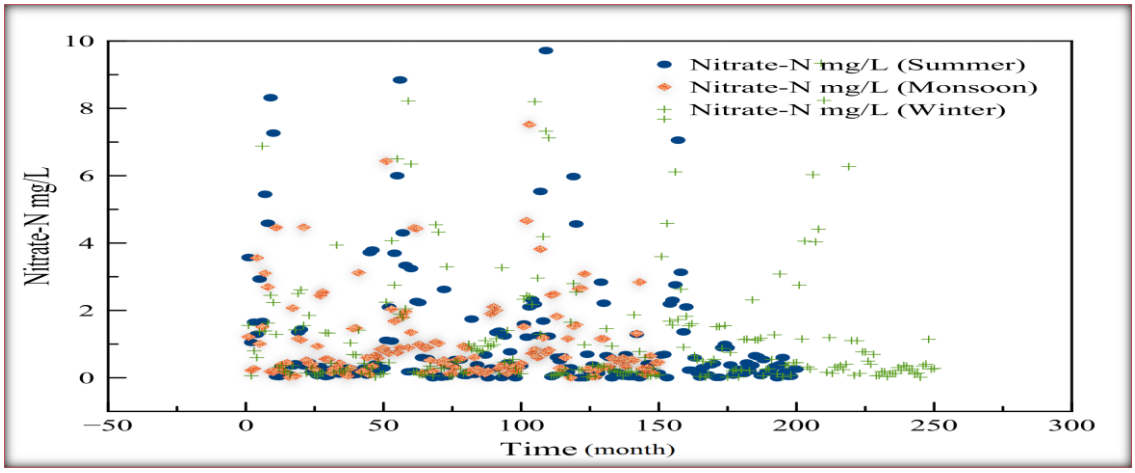


**Figure 3.11: Temporal Variation of Nitrate-N from 2003 to 2012**

## 3.2.6 Total Coli form Bacteria

Total Coli-forms (TC) are not likely to cause illness, but their presence indicates that your water supply may be vulnerable to contamination by more harmful microorganisms. The main sources of these pathogens are through improperly treated septic and sewage discharges, leaching of animal manure, storm water runoff, domestic animals or wildlife. The total coli-form count can be done by Most Probable Number (MPN) per 100 ml of water. The TC count should be in a range of 50-500 MPN/100 ml in drinking water. The TC values from 2003 to 2012 are shown graphically in Figure 3.12.

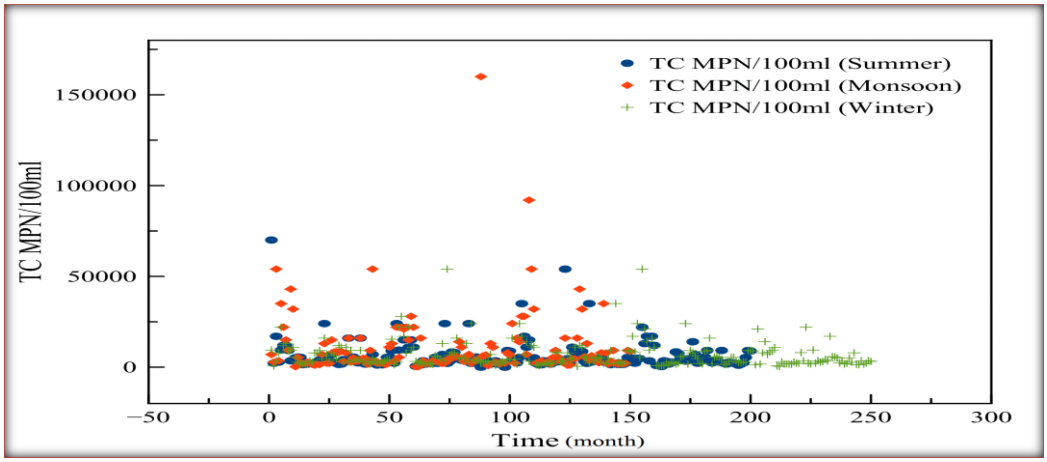


**Figure 3.12: Temporal Variation of Total Coliform Bacteria from 2003 to 2012**

## 3.2.7 Faecal Coliform Bacteria

Like Total Coli-form, Fecal Coli-forms (FC) are the sources of pathogenic or disease causing bacteria and viruses. The disease causing organisms are accompanied by other common types of non-pathogenic bacteria found in animal intestines, such as faecal coli-forms bacteria,

*Enterococci* bacteria and *Escherichia coli* (E.Coli) bacteria. The fecal coli-form count should be 150 MPN/100ml in primary contact like swimming. For drinking purposes the FC count should vary from 0-700 MPN/100ml. The temporal variation of FC in Brahmani River from 2003 to 2012 is represented graphically in Figure 3.13.

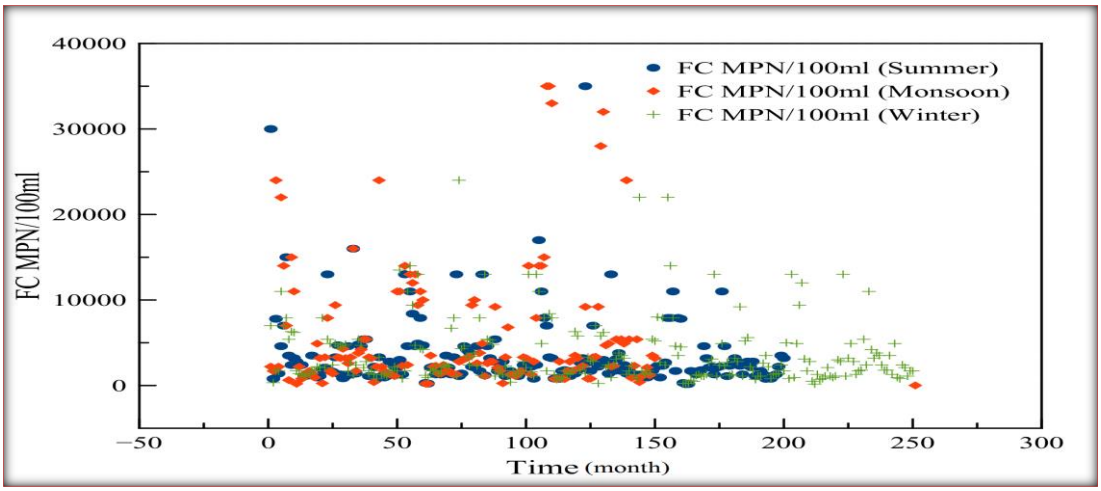


**Figure 3.13: Temporal Variation of Faecal Coliform Bacteria from 2003 to 2012**

## 3.2.8 Chemical Oxygen Demand

The Chemical Oxygen Demand (COD) is used to measure the amount of organic compounds in water. COD can be related empirically to BOD, organic carbon or organic matter. COD is an index of organic content of water because the most common substance oxidized by dissolved oxygen in water is organic matter having a biological origin; that is dead plant and animal waste. The concentration of COD is more in the bottom of water due to more organic matter in the bottom than the surface layer of water. The variations of Chemical Oxygen demand of the Brahmani River from 2003 to 2012 are shown graphically in Figure 3.14.

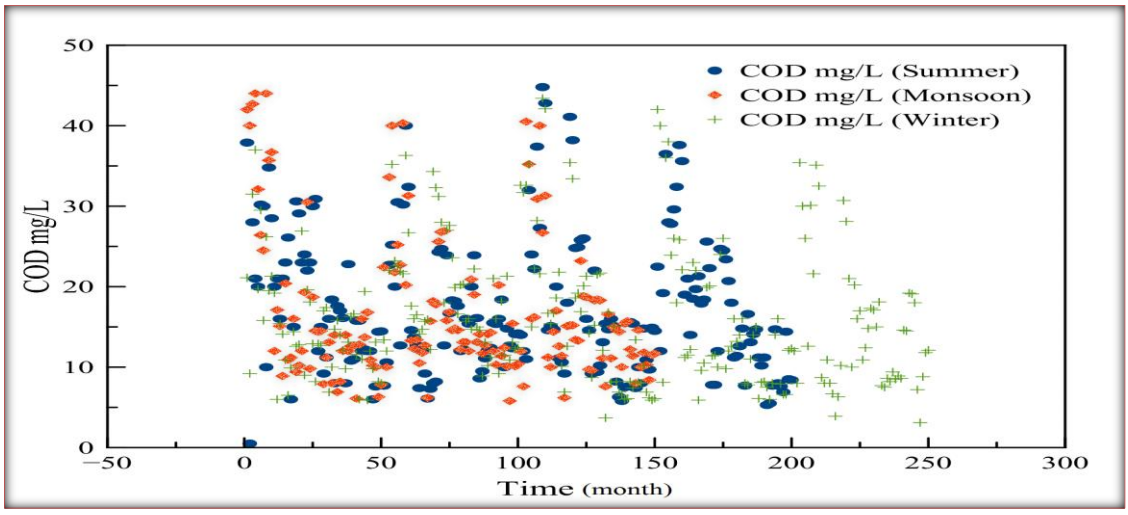


**Figure 3.14: Temporal Variation of COD from 2003 to 2012**

### 3.2.9 Nitrogen as Ammonia

Ammonia is one of the most important water pollutants in the aquatic environment because it is highly toxic and ubiquity in surface water. It is the result of microbiological activity which causes reduction of nitrogen containing compounds in water. It may be due to sewage and industrial pollution and the consequent possible presence of pathogenic microorganisms in the waters. In aqueous solution, ammonia can be of two chemical forms $NH_4^+$, which is ionized and less toxic and $NH_3$, which is unionized and more toxic. The presence of nitrogen as ammonia in drinking water should be in a range of 0-1.2 mg/L. The variations of Nitrogen as Ammonia from 2003 to 2012 are represented graphically as shown in Figure 3.15.

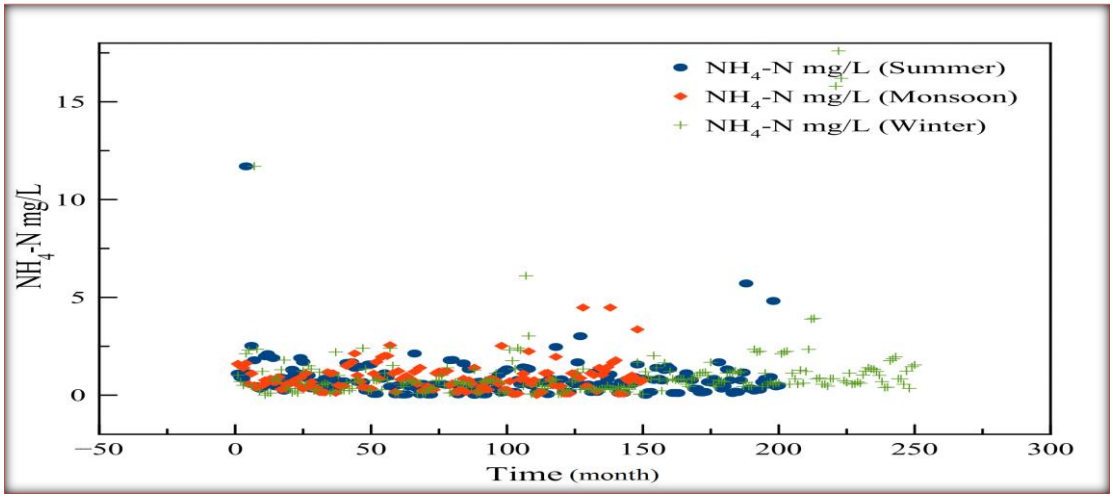


**Figure 3.15: Temporal Variation of Nitrogen as Ammonia from 2003 to 2012**

### 3.2.10 Total Alkalinity expressed as Calcium Carbonate

Total alkali is the measurement of all bases in the river water and can be thought of as the buffering capacity of water, or its ability to resist change in pH. The most common and important base is carbonate. Total alkalinity is expressed as milligrams per litre (mg/L)ofcalcium carbonate ($CaCO_3$). Waters that have moderate to high levels (50 mg/L or greater) of total alkali usually have a neutral to slightly basic pH. The pH is more stable and does not change greatly throughout the day because the presence of carbonates and bicarbonates neutralize, or "buffer," the carbon dioxide and other acids in the water. The TA as $CaCO_3$ should be in a range of 0-200 mg/L in drinking water. The variations of Total alkali as Calcium Carbonate (TA as $CaCO_3$) are shown graphically in Figure 3.16.

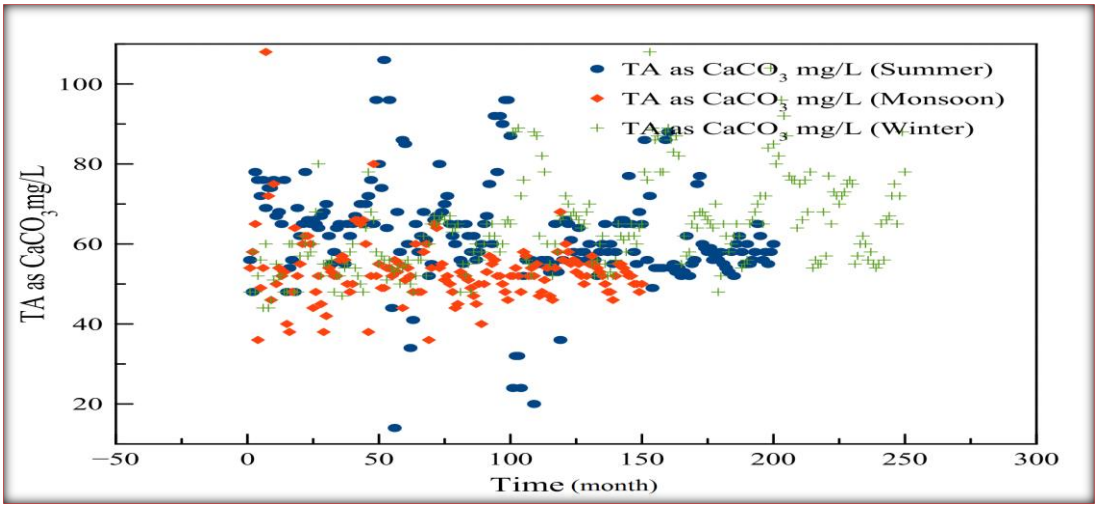


**Figure 3.16: Temporal Variation of TA as $CaCO_3$ from 2003 to 2012**

## 3.2.11 Total Hardness expressed as Calcium Carbonate

Total hardness is the measurement of divalent cations (+2 ions) in the water and, like total alkalinity, is expressed as milligrams per litre (mg/L) of calcium carbonate ($CaCO_3$).When limestone and dolomite dissolve in water, one half of the molecule is calcium (the "hardness") and the other half is the carbonate (the "alkalinity"), so most of the times they are equal. One of the most obvious signs of water hardness is a layer of white film left on the surface of showers. The concentration of calcium ions (Ca2+) in freshwater is found in a range of 4 to 100 mg/L (10–250 mg/L of calcium hardness as $CaCO_3$). Seawater contains calcium levels of 400 mg/L Ca2+ (1000 mg/L of calcium hardness as $CaCO_3$). The variations of TH as $CaCO_3$ in river water are represented graphically as shown in Figure 3.17.

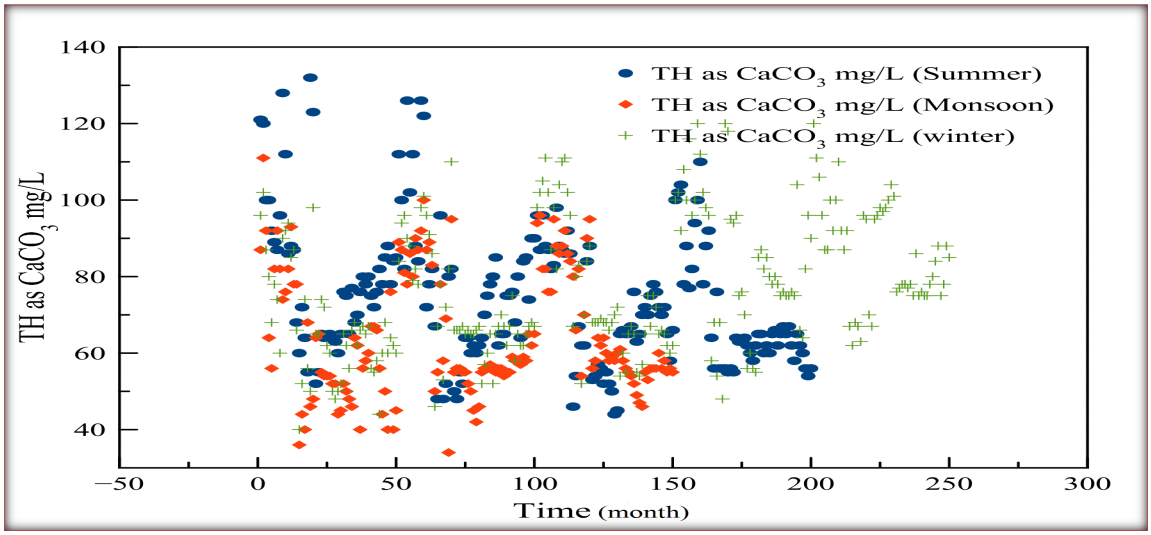


**Figure 3.17: Temporal Variation of TH as CaCO3 from 2003 to 2012.**

The heart of evaluation research is gathering information about the program or intervention. The evaluation and analysis is it to determine what it tells about the effectiveness of the research work, as well as about how effectiveness can be maintained and improved. Collecting quantitative data – information expressed in numbers – and subjecting it to a visual inspection or formal statistical analysis can tell whether the research work is having the desired effect, or may be able to tell about why or why not as well. It can also highlight connections (correlations) among variables, and call attention to factors that may not have considered. Collecting and analyzing qualitative data – interviews, descriptions of environmental factors, or events, and circumstances – can provide insight into the issue those are addressed during the analysis, the barriers and advantages are experienced, and the change or add to improve during analysis.

# CHAPTER IV
# METHODOLOGY

The Chapter illustrates the analysis of principles, methods, rules, and postulates employed in the proposed research work. The systematic study and description of methods used are emphasized below.

## 4.1 Time Series Trend and Correlation Analysis

Trend analysis determines whether the measured values of a water quality parameter increase or decrease over a period that may be temporally or spatially. There are several statistical techniques available for trend analysis depending upon the characteristics of water quality data. Spearman's Rank Correlation analysis was used for the trend and correlation analysis.

### 4.1.1 Spearman's Rank Correlation Analysis

In the test for trend, time series plots of all parameters at five selected consecutive gauging stations were produced temporally including water quality data about 12 months from 2003 to 2012. These plots provided a general indication of trend and supported observations were made later in statistical analysis.

Spearman's Rank Correlation coefficient is a non-parametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using monotonic function. For the second test for trend, the entire data series for each parameter, paired with cumulative season for temporal variation, were analyzed using non-parametric Spearman's criterion to detect the existence of trends in water quality parameters. The existence of a trend in temporal variation of data was checked for significance at a level of 5% from statistical tables of Student's t-distribution (Antonopoulous et al., 1998).

The Spearman's Rank Correlation coefficient ($R_{sp}$) can be described by as :

$$R_{sp} = 1 - \frac{6\sum_{i=1}^{n}(D_i D_i)}{n(n^2 - 1)}, \qquad (4.1)$$

where n = number of values in each set of water quality data, $D$= the difference, and $I$ = chronological order number. The difference between rankings can be computed as:

$$D_i = K_{xi} - K_{yi}, \quad (4.2)$$

where $K_{xi}$= Rank of measured variable in chronological order, $K_{yi}$= series of measurements transformed to its rank equivalents by assigning the chronological order number of the measured variable in the original series; $x$ to the corresponding order number in the ranked series, $y$.

The Spearman's Rank Correlation Coefficient; $R_{sp}$=0, at the null hypothesis, $H_O$, against the alternate hypothesis at $H_1$; there is a trend when $R_{sp}$< or > 0. The above condition was checked with the test statistic.

$$t_t = R_{sp}[\frac{n-2}{1-R_{sp}^{\ 2}}]^{0.5} \quad (4.3)$$

where $t_t$ = Student's t distribution, with $n$-2 degrees of freedom at a significance level of 5%, the time series had no trend if $t$ {$v$, 2.5% <$t_t$<$t$ {$v$, 97.5% }(Antonopoulous et al., 1998). The Spearman's Rank Correlation along with Student's $t$-distribution were estimated temporally for every parameter of five selected gauging stations from 2003 to 2012to know the positive and the negative trend of water quality parameters in Brahmani River.

## 4.2 Overall Water Quality Index (WQI)

In the calculation of water quality for river water, the importance of various water quality parameters depends on the intended use of water and from the point of view of suitability for domestic purposes. The standards (permissible values of various water quality parameters) for drinking water were recommended by Indian Council of Medical Research (ICMR). When ICMR standards for water quality were not available, the standards of United States Public Health Services (USPHS), World Health Organization (WHO), Indian Standard Institution (ISI) and European Economic Community (EEC) were considered.

### 4.2.1 WQI Development Procedure

The process of developing a WQI was followed by the steps as illustrated below:

**Step I:** The water quality parameters of interest were identified and were ranged according to the acceptability for their intended uses in a water body.

**Step II:** The measured values of parameters were calculated by the developed equations for every parameter in MS Excel and were compared with subjective rating curves which concluded on a dimensionless sub-index value ranging from 0-1 for every parameter.

**Step III:** The weighing factors or heuristics were defined for each parameter and were considered while building an overall WQI.

**Step IV:** The algorithm for calculation and formulation of WQI was selected with the available data and assumptions.

## 4.2.2 Rating Scale for Calculation of WQI

A rating scale was prepared as shown in Table 4.1 for a range of values of each parameter. The rating varies from 0 to 100 and divided into five intervals. The rating $Xr$ = 0 implied that the water quality parameter present in river water had the most desirable value. On the other hand $Xr$ = 100 implied that the parameter present in water exceeded the standard maximum permissible limits and the water was severely polluted. The other ratings were between these two extremes and were $Xr$ = 25, $Xr$ = 50, $Xr$ = 75; which meant for slightly polluted, moderately polluted and excessively polluted.

**Table 4.1: Rating Scale for Calculating WQI**

| **Water Quality Parameter** | **Ranges** | | | | |
|---|---|---|---|---|---|
| pH | 7.0-8.5 | 8.6-8.7<br>6.8-6.9 | 8.8-8.9<br>6.7-6.8 | 9.0-9.2<br>6.5-6.7 | > 9.2<br>< 6.5 |
| DO | 0- 5.0 | 5.1-7.0 | 4.1-5.0 | 3.1-4.0 | < 3.0 |
| BOD | 0-1.0 | 1.1-3.0 | 3.1-4.0 | 4.1-5.0 | > 5.0 |
| Conductivity | 0-75 | 75.1-150 | 150.1-225 | 225.1-300 | > 300 |
| Nitrate-N | 0-10 | 10.1-45 | 45.1-100 | 100.1-300 | > 300 |
| TC | 0-2000 | 2000.1-5000 | 5000.1-8000 | 8000.1-10000 | > 10000 |
| FC | 0-2000 | 2000.1-5000 | 5000.1-8000 | 8000.1-10000 | > 10000 |
| COD | 0-10 | 10.1-15 | 15.1-20 | 20.1-30 | > 30 |
| Ammonia-N | 0-10 | 10.1-45 | 45.1-100 | 100.1-300 | > 300 |
| TA as $CaCO_3$ | 21-50 | 50.1-70<br>15.1-20 | 70.1-90<br>10.1- 15 | 90.1-120<br>6.0-10 | > 120<br><6 |
| TH as $CaCO_3$ | 0-150 | 150.1-300 | 300.1-450 | 450.1-600 | > 600 |
| $Xr$ | 0 | 25 | 50 | 75 | 100 |
| Extent of Pollution | Clean | Slight Pollution | Moderate Pollution | Excess Pollution | Severe Pollution |

## 4.2.3 Formulation of WQI

The ranges of water quality parameters in drinking water according to its permit limits by IS 10500-1991and CPCB standards are given in Table 4.2.

The water quality rating $q_i$ for the $i$th water quality parameters is obtained from the relation:

$$q_i = 100(v_i/s_i), \quad (4.4)$$

**Table 4.2: Permissible Limits for Drinking Water Quality (IS 10500-1991, CPCB)**

| Water Quality Parameter | Permissible Ranges |
|---|---|
| pH | 7.0-8.5 |
| DO | 4.0-6.0 |
| BOD | 2.0-3.0 |
| Conductivity | 0-1000 |
| Nitrate-N | 45-100 |
| TC | 50-5000 |
| FC | 50-5000 |
| COD | 18-30 |
| Ammonia-N | 0-5.0 |
| TA as CaCO3 | 200-600 |
| TH as CaCO3 | 200-600 |

where $v_i$= value of the $i$th water quality parameter at a given sampling station and $s_i$= standard permissible value of $i$th water quality parameter. This equation ensures that $q_i$= 0 when a pollutant (the $i$th water quality parameter) is absent in the water while $q_i$= 100 if the value of this parameter is just equal to its permissible value for drinking water. Thus, the larger the value of $q_i$, the more polluted is the river water with the $i$th pollutant. However, water quality ratings for pH and DO require special handling and care. The permissible range of pH for the drinking water is 7.0 to 8.5. Water quality rating for pH can be written as:

$$q_{pH} = 100[(v_{pH} - 7)/(8.5 - 7.0)], \quad (4.5)$$

where $v_{pH}$= value of pH ~7, it means the numerical difference between $v_{pH}$ and 7.0 ignoring algebraic sign. Equation (5) ensures the $q_{pH}$ = 0 for pH = 7.0. The quality ratings of other water quality parameters were calculated like that of water quality rating of pH.

The more harmful a given water quality parameter is, the smaller is its permissible value of drinking water. So the 'weights' for parameters are various water quality assumed to be inversely proportional to the standards recommended by ICMR for the corresponding water quality parameters, that is

$$W_i = \frac{K}{S_i} \quad (4.6)$$

where $W_i$= unit weight for the $i$th water quality parameter ($i$ = 1, 2, 3, . . . , 11), $K$ = constant of proportionality which is determined from the condition and $K$ = 1 for sake of simplicity. The values of $k$ were calculated as:

$$k = \frac{1}{\sum_{i=1}^{11}(\frac{1}{x_i})}, \quad (4.7)$$

So the sum of unit weight of 11 water quality parameters can be given as:

$$\sum_{i=1}^{11} W_i = 1. \quad (4.8)$$

The weightage of all the factors were calculated by applying the above equation. The standard unit weights of water quality factors assigned by ICMR/ CPHEEO were given in Table 4.3. To calculate the WQI, first the sub index $(SI)_i$ corresponding the *i*th water quality parameter is calculated as the product of the quality rating $q_i$ and the unit weight $W_i$ of the *i*th parameter given as:

$$(SI)_i = q_i W_i \quad (4.9)$$

**Table 4.3: Water Quality factors: ICMR/CPHEEO Standards assigned unit Weights**

| Water Quality Factors | ICMR/CPHEEO Standards (xi) | Unit Weight (Wi) |
|---|---|---|
| pH | 7.0-8.5 ** | 0.322 |
| Dissolved Oxygen | > 5* | 0.548 |
| Biochemical Oxygen Demand | < 5* | 0.0055 |
| Conductivity | < 300* | 0.009 |
| Nitrate-N | < 45 | 0.1 |
| Total Coli-form | < 5000** | 0.18 |
| Faecal Coli-form | < 5000** | 0.18 |
| Chemical Oxygen Demand | 15-30* | 0.0006 |
| Ammonia-N | < 0.5* | 0.1 |
| TA as CaCO3 | < 120* | 0.023 |
| TH as CaCO3 | < 600** | 0.005 |

*ICMR Standards (1975) **CPHEEO Standards (1991)

The overall WQI of River Brahmani is then calculated by aggregating these sub indices (SI) linearly. Thus, WQI can be written as:

$$WQI = [\sum_{i-1}^{11} q_i W_i / \sum_{i=1}^{11} W_i] = \sum_{i=1}^{11} q_i W_i \quad (4.10)$$

where, $\sum_{i=1}^{11} W_i = 1$ as explained above in (10) Water quality can be categorized into five classifications depending on WQI values of the parameters. Water quality can be treated as excellent, good, poor, very poor, and unsuitable for drinking water and domestic purposes if WQI lies in the range of 0–25, 26–50, 51–75, 76–100, and 100 respectively.

## 4.3 Calculation of Parts of Water Quality Parameter in River Water

The ppm values of eleven water quality parameters from 2003 to 2012 were selected for the calculation of parts of each parameter in water in three respective seasons of summer,

monsoon and winter at the gauging station; Panposh down-stream. Along with the water quality parameter, the daily discharge values at the given station of the respective years were collected for the analysis. The parts of each parameter can be calculated as:

$$\text{ppm of water quality parameter} \times \text{Discharge of the Panposh down-stream station in } mm^3/sec = \text{parts of parameter in water}$$

From the above formulation, the increase or decrease of the parts of parameter in the river water was found out at the selected station i.e. Panposh down-stream, which was influenced mostly by industries and also by other effluents.

## 4.4 Multivariate Statistical Analysis

### 4.4.1 Multivariate Analysis of Variance (MANOVA)

In many ecological or biological analyses, the variables are not independent at all over each other. The variables have strong actual or potential interactions many times, inflating the error even higher than expected. In many such cases where the sequence of Analysis of Variance (ANOVA) was done, Multiple Analysis of Variance (MANOVA) was actual appropriate test for the study (Seber, 1983; Johnsonet al., 1988).

An experiment was conducted to compare '*v*' number of water quality parameters using Randomised Complete Block (RCB) design with '*r*' replications and data were collected on *p*-variables. Let $y_{ijk}$ be denoted as the observed value of $k^{th}$ response variable for the $i^{th}$ parameter in the $j^{th}$ replications, $i$=1,2,……$v$; $j$=1,2,……$r$; $k$=1,2,…..$p$. The example set of data for water quality parameters were arranged in Table 4.4. The table gives $y_{ij}$ = ($y_{ij1}$, $y_{ij2}$….$y_{ijk}$…$y_{ijp}$) is a *p*-variate vector of observations taken from the plot receiving the analysis of parameters *i* in replication *j*.

**Table 4.4: Water Quality Data Arrangement in Replications**

| | ← Replications → | | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | *1* | *2* | … | *j* | … | *r* | Parameter |
| | | | | | | | analysis |
| | | | | | | | means |
| *1* | $y_{11}$ | $y_{12}$ | … | $y_{1j}$ | … | $y_{1r}$ | $y_{1}$ |
| *2* | $y_{21}$ | $y_{22}$ | … | $y_{2j}$ | … | $y_{2r}$ | $y_{2}$ |
| … | … | … | … | … | … | … | … |
| *i* | $y_{i1}$ | $y_{i2}$ | … | $y_{ij}$ | … | $y_{ir}$ | $y_{i}$ |
| … | … | … | … | … | … | … | … |
| *v* | $y_{v1}$ | $y$ v1 | … | $y_{v1}$ | … | $y_{v1}$ | $y_{v}$ |
| Replication Means | $\overline{y_{1}}$ | $\overline{y_{2}}$ | … | $\overline{y_{i}}$ | … | $\overline{y_{r}}$ | $\overline{y..}$ |

The replication mean were represented as:

$$\overline{y_i} = \frac{1}{r}\sum_{j=1}^{r} y_{ij} \,, \tag{4.11}$$

$$\overline{y_{.j}} = \frac{1}{v}\sum_{i=1}^{v} y_{ij} \,, \tag{4.12}$$

$$\overline{y_{...}} = \frac{1}{vr}\sum_{i=1}^{v}\sum_{j=1}^{r} y_{ij} \,, \tag{4.13}$$

The observations of the parameters can be represented by two way multivariate classified model $\Omega$.

$$\Omega : y_{ij} = \mu + t_i + b_j + e_{ij} \,, \tag{4.14}$$

where $i$=1,2,….$v$;$j$=1,2,…$b$.

$\mu = (\mu_1, \mu_2, .....\mu_k .....\mu_p)'$ is the $p \times 1$ vector of general means.

$t_i = (t_{i1}, t_{i2} .....t_{ik} ....t_{ip})'$ are the result of analysis of parameters $i$ on $p$-characters.

$b_j = (b_{j1}, b_{j2} ...b_{jk} .....b_{jp})'$ are the results of replication $j$ on $p$-characters.

$e_{ij} = (e_{ij1}, e_{ij2} ....e_{ijk} ....e_{ijp})'$ is a $p$-variate random vector associated with $y_{ij}$ and are assumed to be distributed independently as $p$-variate normal distribution $N_p(0, \Sigma)$.

The equality of analysis of parameters was tested that can be represented as: $H_0 : (t_{i1}, t_{i2.} .....t_{ik} ....t_{ip})' = (t_1, t_2 ....t_k .....t_p)'$ can be assumed. So, $\forall i = 1,2,....p$ against the alternative $H_1$: at least two of the results of the analysis are unequal.

Again, under the null hypothesis, the model (4.14) can be reduced to

$$\Omega : y_{ij} = a + b_j + a_{ij} \,, \tag{4.15}$$

where $\alpha = (\mu_1 + t_1, \mu_2 + t_2 .....\mu_p + t_p)'$

The outline for testing the equality of the results of water quality parameter analysis and the effect of replication of the parameters was represented by Table 45. Here, *H*, *B*, *R* and *T* were the sum of squares and sum of cross product matrices (SSCPM) of analysis of parameters, replications and errors or residuals after the analysis and totals respectively.

The residual sum of squares and cross product matrix for the reduced applied model $\Omega_0$ was denoted as $R_0$ and the $R_0$ can be given as $R_0 = R + H$.

Again, the null hypothesis of equality of analysis of parameters means vectors was rejected if the ratio of generalised variance (Wilk's Lambda statistic) $\wedge = \frac{|R|}{|H+R|}$ was too small.

**Table 4.5: Dependent factors along with Sum of squares and Cross product matrix**

| Source | DF | SSCPM |
|---|---|---|
| Analysis | $v-1=h$ | $H=b\sum_{i=1}^{v}(\overline{y_{i.}}-\overline{y...})(\overline{y_{i.}}-\overline{y...})'$ |
| replication | $r-1=t$ | $B=v\sum_{j=1}^{b}(\overline{y_{.j}}-\overline{y...})(\overline{y_{.j}}-\overline{y...})'$ |
| Residual | $(v-1)(r-1)=s$ | $R=\sum_{i=1}^{v}\sum_{j=1}^{b}(y_{ij}-\overline{y_{i.}}-\overline{y_{.j}}+\overline{y...})(y_{ij}-\overline{y_{i.}}-\overline{y_{.j}}+\overline{y...})'$ |
| Total | $vr-1$ | $T=\sum_{i=1}^{v}\sum_{j=1}^{b}(y_{ij}-\overline{y...})(y_{ij}-\overline{y...})'=H+B+R$ |

As suggested by Rao at 1973, if the normal distribution was assumed, it showed that under null hypothesis $\wedge$ was distributed as the product of independent beta variables. There was a better but more complicated approximation of the distribution of $\wedge$ (Wilk's lambda statistic) was represented as:

$$\frac{1-\wedge^{1/b}}{\wedge^{1/b}}\frac{(ab-c)}{ph}\approx F(ph,ab-c), \tag{4.16}$$

where $a=(s-\frac{p-h+1}{2})$, $b=\sqrt{\{(p^2h^2-4)/(p^2+h^2-5)\}}$ and $c=\frac{ph-2}{2}$

For some particular values of *h* and *p*, the distribution of $\wedge$ was reduced to exact *F*-distribution. The selected special cases are illustrated below as:

For *h =1* and any *p*, the distribution of $\wedge$ was reduced to

$$\frac{(1-\wedge)}{\wedge}\frac{(s-p+1}{p}\approx F(p,s-p+1), \tag{4.17}$$

For *h =2* and any *p*, the distribution of $\wedge$ was reduced to

$$\frac{(1-\sqrt{\wedge})}{\sqrt{\wedge}}\frac{(s-p+1)}{p}\approx F(2p,2(s-p+1)), \tag{4.18}$$

For $p=2$ and any $h$:

$$\frac{(1-\sqrt{\wedge})}{\sqrt{\wedge}}\frac{(s-1)}{h} \approx F(2h,2(s-1)), \tag{4.19}$$

For p =1, the statistic was reduced to the usual variance ratio statistics. The null hypothesis depending upon the equality of replications of the parameter analysis and the results of the analysis was tested by replacing $\wedge$ by $\frac{|R|}{|B+R|}$ and $h$ by $t$ as in the sum of squares and cross product matrix.

## 4.4.2 Multivariate Parameter Contrast Analysis

In certain cases, the analyses of parameters were found to be significantly different through MANOVA. In such cases of multivariate parameter contrast analysis, the analysis was carried out using $\chi^2$ statistic. The $\chi^2$ statistic was based on the assumption that the error in variance-covariance matrix was known before the next step of analysis. The error in variance-covariance matrix was however generally unknown. Therefore, the estimated value of error in variance-covariance matrix was used. The error in variance-covariance matrix was estimated by sum of squares and cross product (SSCP) matrix for the error divided by the error degrees of freedom. The $\chi^2$ statistic and Wilk's Lambda were the approximate solution.

Let the hypothesis to be tested as: $H_0$: $t_i = t_{i'}$. The same hypothesis can be written as:

$$H_0: = (t_i - t_{i'}) = 0 \text{ against } H_1: = (t_i - t_{i'}) \neq 0, \tag{4.20}$$

where $(t_i - t_{i'})' = (t_{i1} - t_{i'1} \quad t_{i2} - t_{i'2} \quad .... \quad t_{ik} - t_{i'k} ... \quad t_{ip} - t_{i'p})$. Here, $t_{ik}$ be denoted as the effect of parameter analysis $i$ for the dependable variable $k$. The best linear unbiased estimate of $(t_i - t_{i'})$ was:

$$(\overline{y_{i.}} - \overline{y_{i.'}})' = (\overline{y_{i1}} - \overline{y_{i'1}} \ \overline{y_{i2}} - \overline{y_{i'2}} \quad ... \quad \overline{y_{ik}} - \overline{y_{i'k}} \quad ... \quad \overline{y_{ip}} - \overline{y_{i'p}}), \tag{4.21}$$

where $\overline{y_{ik}}$ is the mean of parameter analysis $i$ for variable $k$.

## 4.4.3 $\chi^2$ - Test

The covariance matrix of the contrast was based on the statistic $\chi^2$. The covariance matrix, in case of RCB for the parameter analysis contrast was obtained by the dividing the SSCP matrix for the errors obtained in MANOVA by half of the product of error degrees of freedom and the number of replications.

Let the variance-covariance matrix was denoted by $\sum_c$. Under the condition of null hypothesis, $x = \overline{y_{i.}} - \overline{y_{i.'}}$ was followed by *p*-variate normal distribution with mean vector 0 and variance-covariance matrix $\sum_c$. The Aitken's transformation was applied and it can be shown that $z = \sum_c^{-1/2} x$ which was followed by a p-variate normal distribution with mean vector 0 and variance-covariance matrix $I_g$, where $I_g$ was denoted by the identity matrix of order *g*. The results of the quadratic forms, it can easily be seen that $z'z = x'\sum^{-1} x$, was followed by a $\chi^2$ distribution with *p*-degrees of freedom.

## 4.4.4 Wilk's Lambda Criterion

By applying the null hypothesis in (4.20), sum of squares and products matrix was obtained for the above elementary parameter analysis contrast.

Let the SSCP matrix for the elementary parameter analysis contrast can be represented as $G_{p\times p}$. The diagonal elements of *G* were obtained by:

$$g_{kk} = (\frac{r}{2})(\overline{y_{ik}} - \overline{y_{i'k}})^2 \forall, \tag{4.22}$$

Where $k = 1,2,....p$; $i \neq i' = 1,2,....v$ and the off diagonal elements were obtained by:

$$g_{kk'} = \frac{r}{2}(\overline{y_{ik}} - \overline{y_{i'k}})(\overline{y_{ik'}} - \overline{y_{i'k'}}), \tag{4.23}$$

The null hypothesis was rejected if the value of Wilk's Lambda $\wedge^* = \frac{|R|}{|G+R|}$ was small, where R was the SSCP matrix due to residuals as obtained through MANOVA. The hypothesis was then tested using the following F-statistics based on Wilk's Lambda for *h=1*.

$$\frac{1-\wedge*}{\wedge*}\frac{edf - p + 1}{p} \approx F(p, s - p + 1), \tag{4.24}$$

## 4.5 Principal Component Analysis (PCA)

## 4.5.1 The Computational Approach for Defining Correlation

The correlation between a pair of variables measures to what extent their values co-vary with each other. The covariance between a pair of variables $(\overrightarrow{X_1}, \overrightarrow{X_2})$ computes the estimator for the covariance. The covariance of the two variables was represented as:

$$COV(\overrightarrow{X_1},\overrightarrow{X_2}) \equiv \frac{\sum_{i=1}^{m}(X_{1i}-\overline{X_1})(X_{2i}-\overline{X_2})}{m}, \tag{4.25}$$

$\overline{X}$ was denoted by the mean of $\overrightarrow{X}$ and *m* be denoted as the number of points in each variable, i.e. the number of samples in each variable. Standardizing the values of the variables $(X_{ki}^{s} = (X_{ki} - \overline{X_k})/\sigma_k$

where *k* can be denoted as the variable matrix, *i* can be denoted as the value index and *s* can be known as the stands for standardized value which gave the values in terms of standard deviation units from the variable's mean. The units were known as *Z*- scores.

Correlation between variables ($X_1$, $X_2$) was measured by a term named correlation coefficient can be denoted as $r_{x1x2}$or *r*. The coefficient can be termed as Pearson product moment correlation coefficient. It can be denoted as:

$$r_{x1,x2} = \frac{\sum_{i=1}^{m} Z_{1i}Z_{2i}}{m} = \frac{\sum_{i=1}^{m}(X_{1i}-\overline{X_1})(X_{2i}-\overline{X_2})}{m\sigma_1\sigma_2} = \frac{\sum_{i=1}^{m}(X_{1i}-\overline{X_1})(X_{2i}-\overline{X_2})}{\sqrt{\sum_{i=1}^{m}(X_{1i}-\overline{X_1})^2(X_{2i}-\overline{X_2})^2}}, \tag{4.26}$$

where $\sigma$ can be denoted as the standard deviation of $\overrightarrow{X}$.

## 4.5.2 The Visual Approach for Defining Correlation

A visual aspect of correlation can be obtained by representing each one of a pair of variables as an axis in Cartesian coordinate system as shown in Figure 4.1. The visual aspects of understanding may supply crucial insight about structures in the data being analysed and prevented the potential biases that may arise by directly interpreting numerical results yielded by running computational procedures.

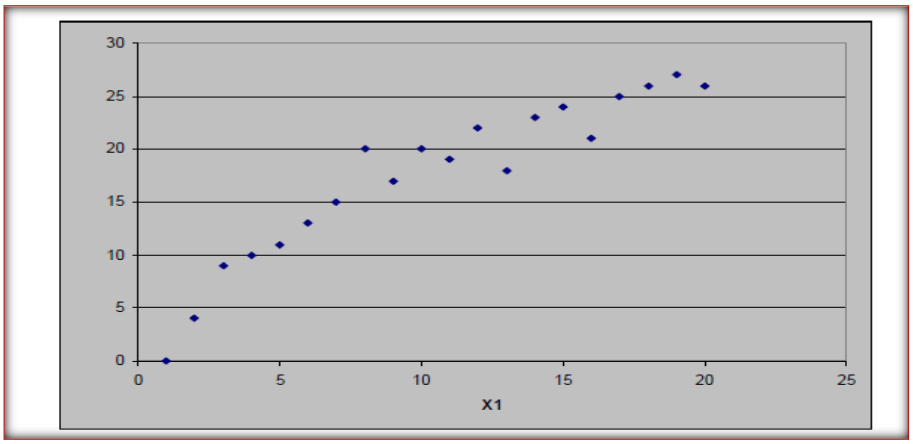


**Figure 4.1: Scatter plot of variables**

PCA deals with linear correlation focus on describing the connection between the points in the plot and a linear trend line. This linear trend line was constructed such that it minimized the perpendicular distances, from it to each point in the plot. It was known as the least squares fitted in line as shown in Figure 4.2.

Correlation might be confused with regression, thus it was important to stress out the differences between the two. In regression analysis; the dependable variable, *Y* and the variable designated a predictor variable called independent variable, *X* was tested.

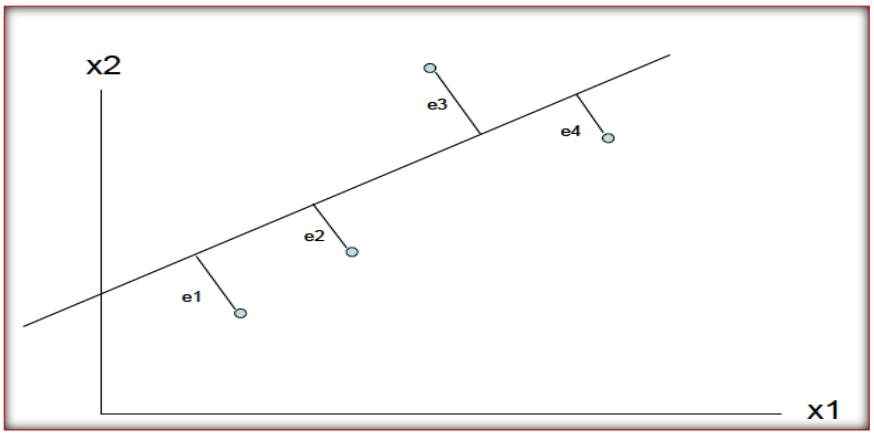


**Figure 4.2: The Best Fit trend line is the one that minimizes the sum** $e1^2 + e2^2 + e3^2 + e4^2$

In this case of testing an imperfect relation between *X* and *Y* was assumed, where an element of error, $\varepsilon_i$ for each value of $Y_i \in Y$ , but the error in *X* was assumed to be negligible. Therefore, to regress *Y* upon *X*, *X* was treated as error-free and hence only the squared vertical distances were minimised as shown in Figure 4.3.

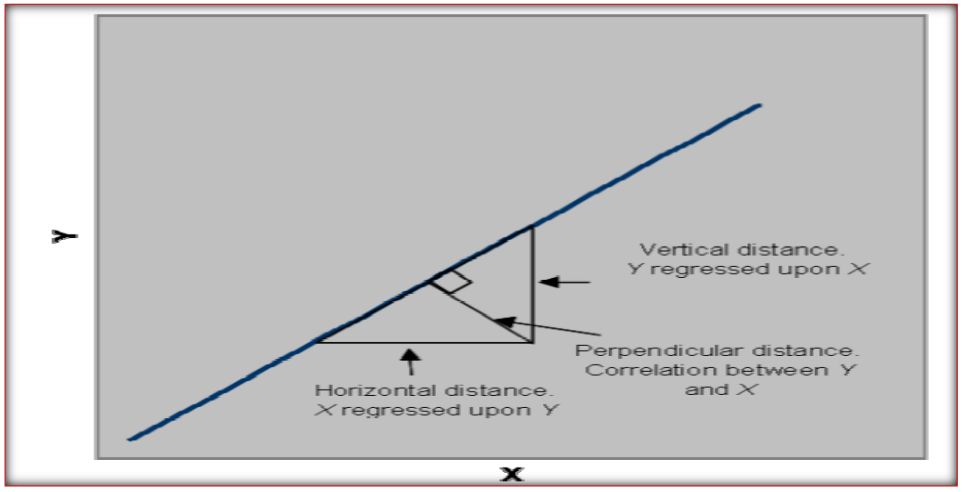


**Figure 4.3: Regression of Y upon X, Regression of X upon Y and Symmetric relation between X and Y.**

However, the symmetric relation between two random variables was measured by correlation. So, the correlation (*X, Y*) = correlation (*Y, X*), where both *X* and *Y* was contained by the amounts of errors; $\delta_i$ for each variable $X_i \in X$ and $\varepsilon_i$ for each value of $Y_i \in Y$ respectively and hence the fitted correlation line was the line minimising both vertical and horizontal distances. The ratio used for expressing correlation was brought upon by comparing the deviations from the least squares fitted line to the deviations from a hypothetical line that was fitted to the same points as they had no correlation with each other.

### 4.5.3 Extraction of Principal Components

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This was achieved by transforming to a new set of variables, the principal components (PCs), which were uncorrelated, and which were ordered so that the first few retain most of the variation present in all of the original variables. Principal component analysis was an analytical technique whereby a complex data set containing *p* variables was transformed to a smaller set of new variables, which maximized the variance of the original data set. All of the new variables were independent, i.e., were not correlated with each other (whereas the original, untransformed variables may have been correlated to a lesser or greater extent). The new principal component (PC) axes ($Y_1$, $Y_2$, . . . ,$Y_p$) were uncorrelated (e.g. $Y_1$ and $Y_2$ are perpendicular as shown in Fig. 4.4).

In principle, each of the principal components was a linear combination of the original *X* values for the *p* variables given as:

$$PC_1 = c_{11}X_1 + c_{12}X_2 + c_{13}X_3 + \cdots + c_{1p}X_p \text{ (axis } Y_1)$$

$$PC_2 = c_{21}X_1 + c_{22}X_2 + c_{23}X_3 + \cdots + c_{2p}X_p \text{ (axis } Y_2)$$

. . .

$$PC_p = c_{p1}X_1 + c_{p2}X_2 + c_{p3}X_3 + \cdots + c_{pp}X_p \text{ (axis } Y_p)$$

(4.27)

where $c_{a,b}$ is the component score coefficient for variable *b* on PC axis $Y_a$, and $X_b$ is the *X* score for variable *b*.

PCA converted a multivariate set of variables ($X_1$, $X_2$, . . . ,$X_p$) to new variables ($Y_1$,$Y_2$, . . . , $Y_p$), which were uncorrelated with each other. The first principal component consisted of a principal component coefficient ($\alpha_i$) for each variable (*p*) such that there was maximal

variance in the calculated score for each case ($n$); the factor score for each case was calculated as $\alpha_1X_1+ \alpha_2X_2+\cdots+\alpha_iX_i+\cdots+\alpha_pX_p$, where $X_i$ was the centre value for the $i$th variable ($X_i$: mean $X$ for the $i$th variable). The first principal component axis for the raw variables $X_1$ was now the new axis $Y_1$. The second principal component consisted of the next set of principal component coefficients ($\alpha_i$) such that the variance remaining was maximal in the calculated score for each case ($n$), and there was no correlation between first and second principal component. $Y_2$ was the second principal component axis for $X_2$. Further sets of principal components (third, fourth, etc.) can be calculated until no statistical significance can be attributed to that principal component (e.g. by $\chi^2$ test).

In principal component analysis, the PCs of Eigen value greater than unity were generally considered to be significant and to contain most of the variability of the original data set. Since the original loadings may not be readily interpretable, usually they were rotated until a 'simple structure' was achieved, that means each variable had very high factor loadings (as high as 1) on one of the PCs and very low factor loadings (as low as 0) on the other PCs. The communalities of each variable calculated before and after factor rotation will be the same. The varimax rotation method was popular among the researchers.

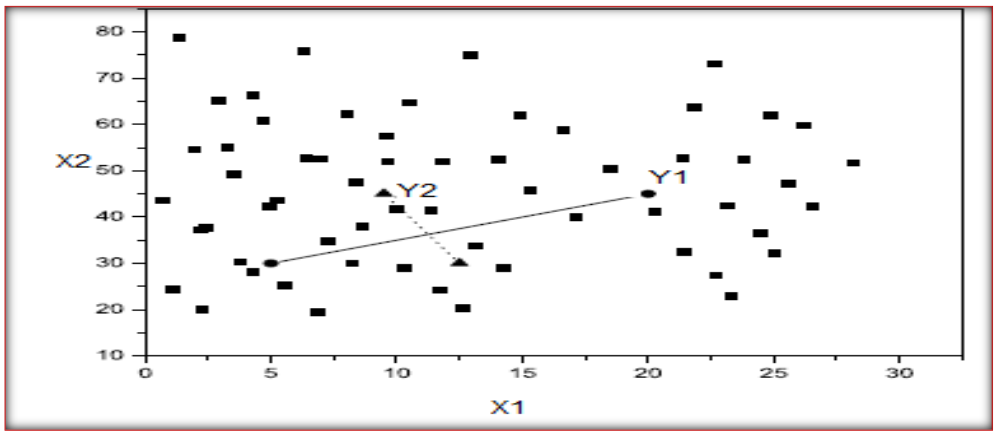


**Fig 4.4: Graph of uncorrelated principal component axis**

Principal component analysis with varimax rotation has been carried out on normalizes parameters data sets using SPSS 20. The number of PCs justified for the study can be judged from Scree plot.

## 4.6 Canonical Correlation Analysis

The multivariate sets of variables were divided naturally into two groups those were response data and predictor variable. A canonical correlation analysis can be used to investigate relationships between the two groups. As an exploratory tool, it was used as data reduction

method. The goal of CCA was to construct two new sets of canonical variates $U = \alpha X$ and $V = \beta Y$ that were linear combinations of the original variables such that the simple correlation between *U* and *V*was maximal, was subjected to the restriction that each canonical variate*U* and *V* had unit variance to ensure uniqueness and was uncorrelated with other constructed variates within the set. The $(p+q)\times(p+q)$ correlation matrix between the variables $X_1, X_2, ...., X_p$ and $Y_1, Y_2, ....., Y_q$ was assumed the following form when it was calculated from the sample for which the variables were recorded:

$X_1, X_2 .....X_p \; Y_1, Y_2, .....Y_q$

$$\left[\begin{array}{c|c} p\times p matrix & q\times q matrix \\ A & B \\ \hline q\times p matrix & q\times q matrix \\ C' & B \end{array}\right]$$

From this matrix a $q\times q$ matrix $B^{-1}C^{1}A^{-1}C$ can be calculated, and eigenvalue problem can be considered as:

$$(B^{-1}C^{1}A^{-1}C - \lambda I)b = 0 \quad , \tag{4.28}$$

It turned out that the eigenvalues $\lambda_1 > \lambda_2 > ....\lambda_r$ were the squares of the correlations between the canonical variates. The subscribed value of *r* was smaller value of *p* and *q*. The corresponding eigenvectors $b_1.b_2, b_{3,}.....b_r$ gave the coefficients of the *Y* variables for the canonical variates. The coefficients of linear combination of *X* variables were given by the elements of the $a_i$ vector.

$$a_i = A^{-1}Cb_i, \tag{4.29}$$

In these calculations it was assumed that the original *X* and *Y* variables were in a standardised form with means of zero and standard deviations of unity. The coefficients of the canonical variates were for these standardised *X* and *Y* variables.

## 4.7 Factor Analysis

Factor analysis is a statistical method used to describe variability among observed and correlated variables in terms of a lower number of unobserved variables called Factors. Factor analysis is related to principal component analysis (PCA), but the two analyses are not identical. Latent variable model which includes factor analysis, use of regression modelling techniques to test the hypotheses producing error terms, while PCA is a descriptive

technique. There has been significant controversy in the field over the equivalence or otherwise of the two techniques.

### 4.7.1 Statistical Model

Suppose we had a set of *p* observable random variables, $x_1,.....,x_p$ with means $\mu_1,.....,\mu_2$.

Suppose for some unknown constants $l_{ij}$ and *k* unobserved random variables $F_j$, where $i \in 1,....., p$ and $j \in 1,.....k,$ where $k < p$, we had

We had $x_i - \mu_i = l_{i1}F_1 + ..... + l_{ik}F_k + \varepsilon_i$.

Here, the number of $\varepsilon_i$ were independently distributed error terms with zero mean and finite variance, which may not be the same for all number of $i$.

Let $Var(\varepsilon_i) = \varphi_i$, so that we had

$Cov(\varepsilon) = Diag(\varphi_1,.....\varphi_p) = \psi$ And $E(\varepsilon) = 0$.

In matrix terms, we had

$x - \mu = LF + \varepsilon$

If we had n observations, then we will have the dimensions $x_{p\times n}, L_{p\times k}$, and $F_{k\times n}$. Each column of *x* and *F* values denoted foe one particular observations, and matrix *L* did not vary across the observations.

The assumptions of *F* were imposed as followed:

1. *F* and $\varepsilon$ were independent.
2. $E(F) = 0$.
3. $Cov(F) = I$, to make sure the uncorrelated nature of factors.

Any solution of the above set of equations following the constraints for *F* was defined as the factors, and *L* as the loading matrix.

Let $Cov(x - \mu) = \Sigma$. The form was noted the conditions just imposed on *F*, we had

$Cov(x - \mu) = Cov(LF + \varepsilon)$ , Or, $\Sigma = LCov(F)L^T + Cov(\varepsilon)$ , Or, $\Sigma = LL^T + \psi$.

For any orthogonal matrix *Q*, $L=LQ$ and $F=Q^TF$, the criteria for factors and factor loadings were in hold. Hence a set of factors and factor loadings was identical only up to orthogonal transformation.

### 4.7.2 Factor Loadings

The factor loadings were also called as the component loadings in PCA, were the correlation coefficients between the variables in rows and factors in columns. Analogous to Pearson's *r*, the squared factor loadings was the percent of variance in that indicator variable explained by the factor. The percent of variance was obtained in all variables accounted by each other, the sum of squared factor loadings were added for that factor in column and were divided by the number of variables.

By one rule of thumb in confirmatory factor analysis, loadings should be 0.7 or higher to make it confirm that independent variables identified a prior were represented by a particular factor, on that rationale the 0.7 was corresponded to about half of the variance in the indicator were being explained by the factor. The 0.7 standard was a high one and the real life data were not met the criterion. For exploratory purposes, 0.4 was used for the central factor and 0.25 for the factors called loadings above 0.6 that was high and 0.4 that was low.

In oblique rotation, the pattern matrix and structure matrix were obtained. The structure matrix was the factor loading matrix as in orthogonal rotation, represented by the variance in a measured variable explained by a factor on both unique and common contributions basis.

### 4.7.3 Communality

The sum of squared factor loadings for all factors for a given variable in row was considered as the variance in that variable accounted for by all the factors and this was called as the communality. The percent of variance was measured by the communality in a given variable explained by all the factors jointly and were interpreted as the reliability of the indicator.

### 4.7.4 Eigenvalues or Characteristic roots

The variance in all the variables was measured by the eigenvalue for a given factor which was accounted for by that factor. The ratio of eigenvalues was the ratio of explanatory importance of the factors with respect to the variables. There was a little contribution to the explanation of variances in the variables if the factor had low eigenvalue and was ignored as redundant with more important factors. Eigenvalues measured the amount of variation in the total sample accounted for by each of the factor.

### 4.7.5 Extraction Sums of Squared Loadings

Initial eigenvalues and eigenvalues after extraction were same for PCA extraction, but for other extraction methods, eigenvalues after extraction were lower than their counterparts.

Rotation Sums of Squared Loadings were extracted by SPSS also for PCA; these eigenvalues were differed from initial and extraction eigenvalues, though their total was same.

### 4.7.6 Factor Scores

Factor scores were also known as the component scores in PCA. Factor scores were the scores in each case in rows on each factor in column. For a given case and a given factor, the case's standardized score on each variable was multiplied by the corresponding loadings of the variable for a given factor, and the sum of these products was the computed factor score of the given factor. The computed factor scores were allowed one to look for factor outliers. Factor scores were used for subsequent modelling.

### 4.7.7 Kaiser Criterion

The Kaiser rule was to drop all components with eigenvalues under 1.0; this was being the eigenvalue equal to the information accounted for by an average single item. The Kaiser criterion was used as the sole cut-off criterion for estimating the number of factors as it was tended to over extract the factors.

### 4.7.8 Kaiser-Meyer-Olkin and Bartlett's Test

The KMO and Bartlett's test was used to test if *k* samples were from populations with equal variances.KMO and Bartlett's test was used to test the null hypothesis, $H_0$ that all *k* population variances were equal against the alternative that at least two were different.

If there were *k* samples with size $n_i$ and sample variances $S_i^2$ then Bartlett's test statistic was

$$X^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k+1)}(\sum_{i=1}^{k}\frac{1}{n_i - 1}) - \frac{1}{N-k}}, \tag{4.30}$$

where $N = \sum_{i=1}^{k} n_i$ and $S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)S_i^2$ was the pooled estimate for the variance. The test statistic had approximately a $\chi_{k-1}^2$ distribution. Thus the null hypothesis was rejected if $X^2 > \chi_{k-1,\alpha}^2$

where $\chi_{k-1,\alpha}^2$ was the upper tail critical value for the $\chi_{k-1}^2$ distribution. Bartlett's test was a modification of the corresponding likelihood ratio test designed to make the approximation to the $\chi_{k-1}^2$ distribution better (Bartlett, 1937).

### 4.7.9 Variance Explained Criteria

Enough factors were kept for 90% variation to explain the variance. The parsimony was emphasized;the criterion was as low as 50%.

### 4.7.10 Scree Plot

The components as the *X* axis and the corresponding eigenvalues as the *Y*-axis were plotted by the Cattell scree plots. The eigenvalues were dropped as one moved to the right, toward later components, when the drop was ceased and the curve made an elbow towards later components. After the starting of elbow, Cattell's scree test made the drop of all further components.

### 4.7.11 Rotation Method

The unrotated output maximised the variance accounted for by the first and subsequent factors, and forcing the factors to be orthogonal. Having many items load on the early factors, and usually, of having many items load substantially on more than one factor was came at data-compression. Rotation was served to make the output more understandable, by seeking so called 'simple structure'. Simple structure was a pattern of loadings where items were loaded most strongly on one factor, and much more weakly on the other factors.

Varimax rotation was an orthogonal rotation of the factor axes to maximise the variance of the squared loadings of a factor in column on all variables in rows in a factor matrix. Each factor had tended to have either small or large loadings of any particular variable. A varimax solution was yielded the results which made it as easy as possible to identify each variable with a single factor. This was the most common rotation option.

## 4.8 DiscriminantAnalysis

### 4.8.1 Discriminant Functions

Discriminant function analysis was a statistical analysis to predict a categorical dependent variable called as a grouping variable by one or more continuous or binary independent variables called as predictor variables.

Creating one more linear combinations of predictors, creating a new latent variable for each function was by discriminant analysis. These functions were called discriminant functions.

The number of functions possible was either $N_g - 1$ or p-1, (4.31)

where $N_g$= number of groups, or $p$ be the number of predictors, whichever was smaller. The first function was created and maximised the differences between groups on that function. The second function was maximised the differences on that function, but also must not be correlated with the previous function. This was continued with subsequent functions with the requirement that the new function cannot be correlated with any of the previous functions. Given group *j*, with $R_j$ was set of sample space, there was a discriminant rule such that $x \in R_j$ , then $x \in j$ .

## 4.8.2 Fisher's Linear Discriminant

Fisher's linear discriminant were the methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterised or separated two or more classes of objects or events.

The term Fisher's linear discriminant were often used interchangeably, although Fisher's original article actually was described a slightly different discriminant, which did not make some of the assumptions of fisher's linear discriminant analysis such as normally were distributed classes or equal class covariances.

Suppose two classes of observations had means $\overrightarrow{\mu_y} = 0, \overrightarrow{\mu_y} = 1$ and co-variances $\Sigma_{y=0}$ and $\Sigma_{y=1}$.Then the linear combination of features $\vec{\omega}.\vec{x}$ would have $\vec{\omega}.\vec{\mu}_{y=i}$ And variances $\vec{\omega}^T \Sigma_{y=i} \vec{\omega}$ for *i=0, 1*. Fisher was defined the separation between these two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma_{between}^2}{\sigma_{within}^2} = \frac{(\vec{\omega}.\overrightarrow{\mu}_{y=1} - \vec{\omega}.\vec{\mu}_{y=0})^2}{\vec{\omega}^T \Sigma_{y=1} \vec{\omega} + \vec{\omega}^T \Sigma_{y=0} \vec{\omega}} = \frac{(\vec{\omega}.(\vec{\mu}_{y=1} - \vec{\mu}_{y=0}))^2}{\vec{\omega}^T (\Sigma_{y=0} + \Sigma_{y=1}) \vec{\omega}} \quad , \tag{4.32}$$

This was a measure in some sense of signal to noise ratio for the class labelling. It can be shown that the maximum separation occurred when

$$\vec{\omega} \alpha (\Sigma_{y=0} + \Sigma_{y=1})^{-1} (\vec{\mu}_{y=1} - \vec{\mu}_{y=0}), \tag{4.33}$$

## 4.9 Hierarchical Clustering

In data mining, hierarchical clustering was a method of cluster analysis which seeks to build a hierarchy of clusters.

### 4.9.1 Cluster Dissimilarity

In order to decide which clusters should be combined or where a cluster should be split was a measured of dissimilarity between sets of observations was required. In most methods of hierarchical clustering, this was achieved by use of an appropriate metric, and a linkage criterion which specified the dissimilarity of sets as a function of the pair wise distances of observations in the sets.

### 4.9.2 Metric

The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.

Some commonly used metrics for hierarchical clustering were:

Euclidean distance $\|a-b\|_2 = \sqrt{\Sigma_i (a_i - b_i)^2}$

Squared Euclidean distance $\|a-b\|_2 = \Sigma_i (a_i - b_i)^2$

### 4.9.3 Linkage Criteria

The linkage criterion was determined by the distances between sets of observations as a function of the pair wise distances between observations.

Some commonly used linkage criteria between two sets of observations *A* and *B* were:

Maximum or complete linkage clustering = $\max\{d,(a,b) : a \in A, b \in B\}$.

Minimum or single-linkage clustering = $\min\{d,(a,b) : a \in A, b \in B\}$.

where *d* is the chosen metric.

## 4.10 Adaptive Neuro-Fuzzy Inference System (ANFIS) By MATLAB

Adaptive neuro-fuzzy inference system (ANFIS) was the result of coupled between artificial neural network (ANN) and fuzzy inference system (FIS) in MATLAB. A neural network and fuzzy logic were related and complementary technology to each other. The data and feedback can be learned by neural network, however understanding the knowledge or trend of data can be difficult. But fuzzy logic models and tool boxes were easy to execute because of the linguistic terms like IF-THEN rules. The neural network had the capabilities to learn the fuzzy decision rules by creating hybrid intelligent system. ANFIS model was first used symmetrically by Takagi and Sugeno at 1985 and they found numerous applications in the field of prediction and inference (Sugeno 1985;

Predrycz 1989). An Adaptive Neuro-Fuzzy Inference System consisted of five important functional building parts of the fuzzy logic tool box, those are (i) rule base, (ii) data base, (iii) decision making unit, (iv) fuzzification interface and (v) defuzzification interface.

The rule base and data base were known as knowledge base. The inference system of ANFIS was dependent upon logical rules, which made the input variables space to output variable spaces using IF-THEN rules and fuzzy logic decision making procedure (Jang and Gulley 1996; Dezfoli 2003). The fuzzification transition was used to transform deterministic value to fuzzy value, due to uncertainty of real field values. Likewise, defuzzification transition was used to convert fuzzy logic values to deterministic values as stated and demonstrated by Dezfoli (2003).

### 4.10.1 Architecture and Basic Learning Rules of ANFIS system

In a typical adaptive neural network, the network structures were consisting of number of nodes, characterized by node function with fixed or adjustable parameters. These nodes were connected through directional links. The basic learning rule for ANFIS was a back propagation method, which minimises the error; it was usually the sum of squared differences between network output and desired output for the data. Generally, Learning or training phase of ANFIS was a process to determine parameter values to best fit the training data given. The model performance can be checked by means of distinct data and best fit was expected in testing phase. Considering a first order Takagi, Sugeno and Kang (TSK) fuzzy inference system, a neoro-fuzzy model consisted of two rules, given by Sugeno and Kang (1988) as:

**Rule 1:** If $x$ is $A_1$ and $y$ is $B_1$ then $f_1 = p_1x + q_1y + r_1$

**Rule 2:** If $x$ is $A_2$ and $y$ is $B_2$ then $f_2 = p_2x + q_2y + r_2$

If $f_1$ and $f_2$ were constants instead of linear equations, we had zero order TSK fuzzy models. The node function in the same layer was of the same function family as described below. Here, $O_i^j$ can be denoted the output of the $i$th node in layer $j$.

**Layer 1:** Each node in this layer created a membership grade of a linguistic label. For instance, the node function of the $i$th node would be

$$O_i^j = \mu A_i(x) = \frac{1}{1+[(\frac{x-c_j}{a_i})]^{b_i}}, \qquad (4.34)$$

where x was the input to node *i* and $A_i$ was the linguistic label (small, large) associated with the node. The parameters that changed the shapes of the membership function were $\{a_i, b_i, ci\}$.The parameters in this layer were known as Premise parameters.

**Layer 2:** Each node in this layer finds the firing strength of each rule via multiplication, given as

$$O_i^2 = w_i = \mu A_i(x) \times \mu B_i(y) \quad \text{Where } i=1, 2. \tag{4.35}$$

**Layer 3:** Here, the ith node finds the ratio of ith rule's firing strength to the sum of all rule's firing strengths as

$$O_i^3 = \overline{w_i} = \frac{w_i}{w_1 + w_2} \quad \text{, where i=1, 2.} \tag{4.36}$$

The layer also called as normalised firing strengths.

**Layer 4:** Every node *i* in this layer was a squared node with a node function as given below:

$$O_i^4 = \overline{w_i} f_i = \overline{w_i}(p_i + q_{iy} + r_i), \tag{4.37}$$

where $\overline{w_i}$ was the parameter set as the output of layer 3.The parameters in this layer were known as Consequent parameters.

**Layer 5:** Here, the summation of all incoming signals is computed by the single circle node, and was given as:

$$O_i^5 = \text{Overall output} = \sum_i^n \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}, \tag{4.38}$$

The output layer backward to input nodes recursively stated by Werbos (1974). The back-propagation learning rule used here was exactly same as in common feed forward neural network (Rumelhart et al. 1986). An adaptive neural network structure presented in Figure 4.5 was functionally similar to fuzzy inference system. It was observed that the values of premise parameters and the overall output ' *f*'' were linear combination of the consequent parameters. The output *f* can be formulated as

$$f = \frac{w_1}{w_1 + w_2} f_2 + \frac{w_1}{w_1 + w_2} f_2 = \overline{w} f_1 + \overline{w} f_2$$

$$= (\overline{wx}) p_1 + (\overline{wy}) q_1 + (\overline{w_1}) r_1 + (\overline{w_2 x}) p_2 + (\overline{w_2 y}) q_2 + (\overline{w_2}) r_2$$

The output *f* is linear in the consequent parameters $p_1$, $q_1$, $r_1$, $p_2$, $q_2$, $r_2$.

Again, the Hybrid learning rule combined a gradient descent and the least squares method to find a feasible of antecedent and consequent parameters (Jang 1991a,1993).The details of the hybrid learning rule was described by Jang et al. (1997). In forward pass of hybrid learning algorithm, node outputs was went forward for identification of layer 4 and consequent parameters. Likewise, in backward pass, the error signal was propagated by thebackward and the premise parameters were checked and updated by gradient descent. Table 4.6 was shown by the two passes in hybrid learning algorithm of ANFIS.

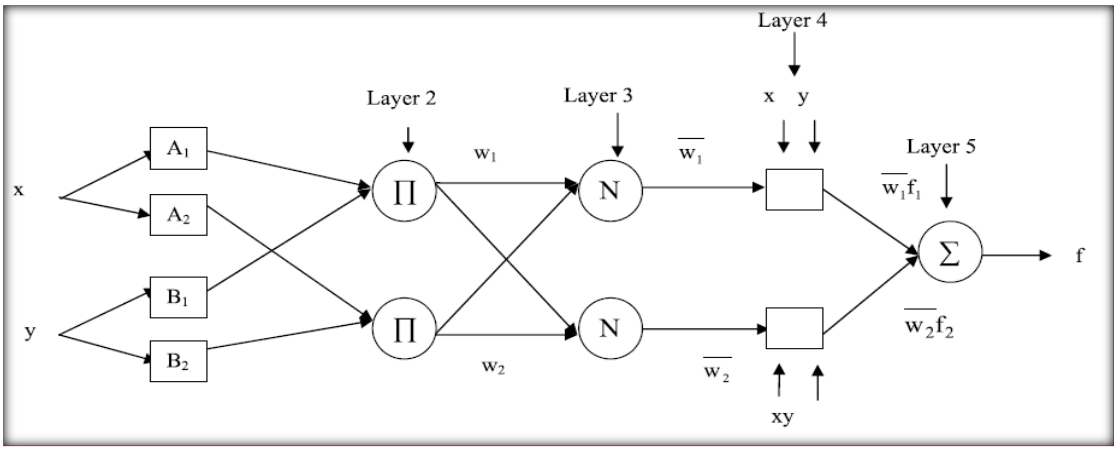


**Figure 4.5: A typical architecture of ANFIS system.**

The consequent parameters were identified as optimal under the condition that the premise parameters were fixed. According to these parameters, the hybrid learning rule converged much faster because it reduced the dimension of search space of the original back-propagation method. This network created, fixed the membership function and took only the consequent parameters, and then only the ANFIS were known as functional-linked network (Klassen and Pao, 1988).By this fine tuning membership functions and advantages of human knowledge, it was expressed as enhanced representation of neural network.

**Table 4.6 : Two passes in hybrid learning algorithm of ANFIS**

| | Forward Pass | Backward Pass |
|---|---|---|
| Premise Parameters | Fixed | Gradient Descent |
| Consequent Parameters | Least-squares estimator | Fixed |
| Signals | Node Outputs | Error Signals |

## 4.10.2 Training and Testing of data by ANFIS GUI Editor

The data were collected from five the gauging stations of River Brahmani during January to December from 2003 to 2011as illustrated in CHAPTER II. The data were normalised and

were used as input in Principal Component Analysis as described in section of Principal Component Analysis (PCA). The normalised data were used as an input to ANFIS. The output for each data was the WQI calculated as per procedure given in section of calculation and formulation of WQI.

The principal component data sets were divided into training set and testing data. Among total data sets; 66.67% of data were considered as training data and rest as testing data sets. The same procedure was repeated for three seasons as the analysis was done according to temporal variation of data. A five layered ANFIS model was created during training. Starting with two nodes the number of nodes in second layer was increased gradually during training of data. The error started decreasing by increasing the nodes up to three. Hence, number of nodes in second layer was fixed to three and further analysis of ANFIS model was carried out. The five layers were defined as, one input, three hidden layer and one output layer. The network was run in MATLAB 2012b Version 8.

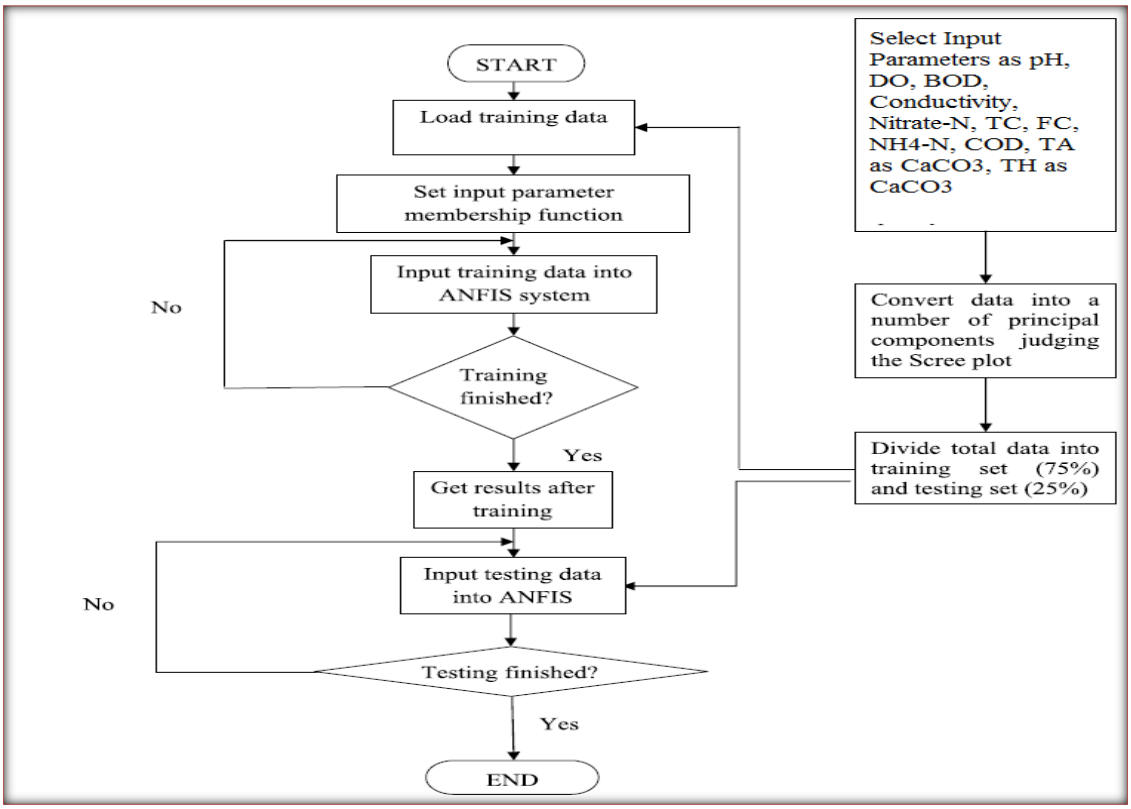


**Figure 4.6: Flow chart showing steps of ANFIS model.**

A membership function of Gaussian type (guessmf) was chosen for inputs and a membership function of constant type for output during generating fuzzy inference system. The flow chart for complete approach and ANFIS algorithm was shown in Figure 4.6. Due to

faster hybrid learning rule, the function ran steadily after 10 iterations, which was defined by that the model parameters matched in training and checking process. Then, the rest data sets leaving 66.67% of data for testing were used to verify the accuracy of the ANFIS model.

## 4.11 Artificial Neural Network (ANN)

The artificial neural network (ANN) has the capability to learn from the pattern known before. The artificial neural network can make predictions on the basis of its previous learning about the output related to new input data set; both set should be of same pattern. The prediction can go further once the network has been sufficiently trained. In the present study the water quality of Brahmani River was predicted depending upon the variation in water quality parameters. The paradigms in artificial neural network pattern were based on direct modelling of the human neuronal system. The network can be defined using three fundamental components: transfer function, network architecture and learning law.

Back Propagation algorithm was the most effective learning technique in multilayer neural network structure. The feed forward back propagation neural network (BPNN) was always consisted of at least three layers: input layer, hidden layer and output layer as given in Figure 4.7. A network was needed to be trained before interpreting new information for the next process. Each layer was consisted of neurons and each neuron was connected to the next layer through weights those were called neurons in the input layer which sent its output as input for neurons in the hidden layer and similar was the connection between hidden and output layer. Number of hidden layer and number of neurons in the hidden layer were changed according to the problem was to be solved. The number of input and output neuron was same as the number of input and output variables.

To differentiate between the different processing units, values were called as biases were introduced in the transfer functions and were referred as the temperature of a neuron. The bias was like a weight and has an input of 1, while the transfer function filtered the summed signals received from this neuron. The transfer functions were designed to map neurons or layers net output to its actual output and they were simple step functions either linear or non-linear functions. Except for the input layer, all neurons in BPNN were associated with a bias neuron and a transfer function. The application of transfer functions was depended on the purpose of the neural network. Output layer was produced and the vectors corresponding to the solution was computed.

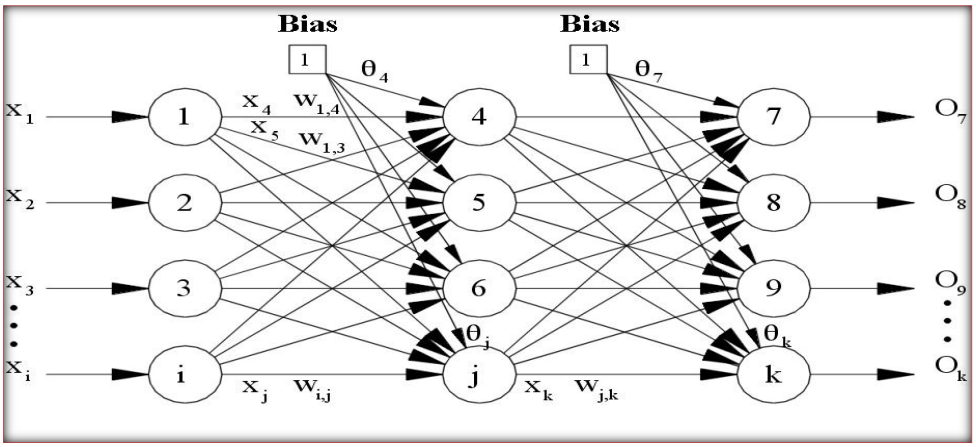


**Figure 4.7: Back Propagation Neural Network**

## 4.11.1 Components of Neuron

In components of neuron, most of the components were described and were contained in neural network. These components were valid even if the neuron was used like input, output and hidden layer. A single neuron with basic elements of an artificial neuron was described in Figure 4.8.

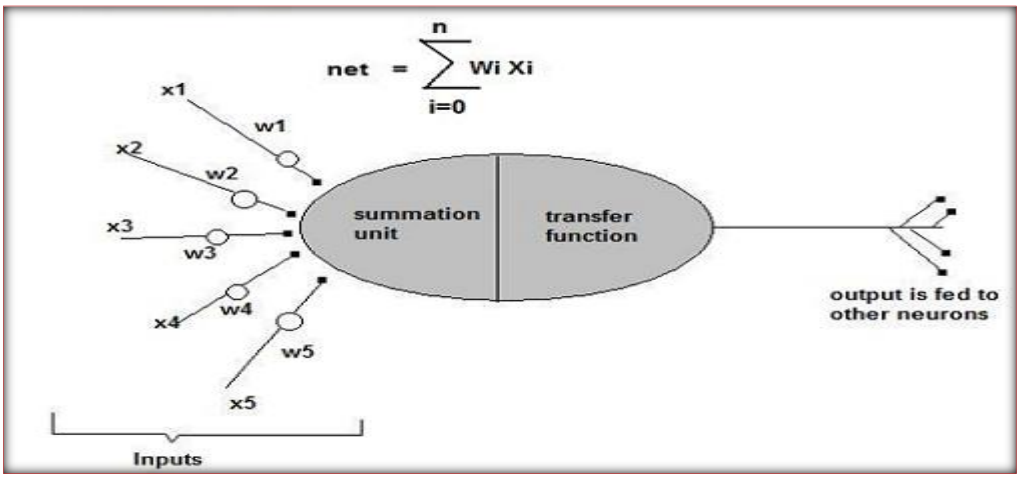


**Figure 4.8: Basic elements Artificial Neuron**

## 4.11.2 Weights

Typically a neuron was received by many simultaneous and multiple inputs. Each input of the neural network had its own relative weight which gave the importance of the input within the activation of the neuron. These weights were coefficients that can be adapted within the network to determine the intensity of input signal, was received by the artificial neuron. They were the measure of the strength of an input connection. These forces were modified in response to the training examples according to the specific topology or because of the training rules.

### 4.11.3 Activation or Transfer Function

The weighted sum was transferred input the actual neuron through an algorithm process known as activation function.

$$a_i(t) = f_i(a_i(t-1), h_i(t)), \quad (4.39)$$

In this case the activation function was depended on the postsynaptic potential $h_i(t)$and its previous state of activation. In many models of ANN was considered that the current state of neuron did not upon its previous state $a_i(t\text{-}1)$ but only the current.

$$a_i(t) = f_i(h_i(t)), \quad (4.40)$$

In activation function, the value of output combination can be compared with a threshold value for determining the output of the neuron. If the sum was greater than the threshold value, a neuron signal was generated. The threshold value or transfer function was linear.

### 4.11.4 Architecture and Basic Learning Rules of ANN

During the training of network in ANN, data was processed through the input layer to hidden layer until the output layer was obtained in the forward pass. The output values were compared with the input or true values. The differences or error between the input and output values were again processed back through the network and were called as backward pass. These were updated the individual weights of the connections and the biases of the individual neurons were get updated. The input and output data were mostly represented as vectors called training pairs. The process was repeated for all the training pairs of data until the network error was converged to a threshold minimum and was defined by a corresponding cost function; usually the root mean squared error (RMSE) or Summed Squared Error (SSE).

In Fig.7 the *j*th neuron was connected to number of inputs

$$x_i = (x_1, x_{2,}, x_3 \ldots\ldots x_n) \ , \quad (4.41)$$

The input values in the hidden layers were represented as:

$$Net_j = \sum_{i=1}^{n} x_i W_{ij} + \theta_j \ , \quad (4.42)$$

Where, $x_i$ =input values, $W_{ij}$ = weight on the connection point of $i^{th}$ and $j^{th}$ neuron, $\theta_j$ = bias neuron and $n$ = number of input units.

Net output from the hidden layer was calculated using a logarithmic sigmoid function given as:

$$O_j = f(Net_j) = 1/1 + e^{-(Netj+\theta j)}, \quad (4.43)$$

The total input to the $k^{th}$ unit is given as

$$Net_k = \sum_{j=1}^{n} W_{jk} O_j + \theta_k, \quad (4.44)$$

where, $\theta_k$ =bias neuron and $W_{jk}$ = weight between $j^{th}$ neuron and $k^{th}$output.

Total output from one unit were given as

$$\theta_k = f(Net_k), \quad (4.45)$$

The network was computed its own output pattern using its weights and thresholds from the given input pattern and corresponding output pattern. Now, actual output was compared with the desired output.

Hence, the error in any output layer *k* can be given as:

$$e_l = t_k - O_k, \quad (4.46)$$

where, $t_k$ = desired output and $O_k$ =actual output

The total error function was given as

$$E = 0.5\sum_{k=1}^{n} (t_k - O_k)^2, \quad (4.47)$$

Training of the network is the process to get an optimum weight space of the network. The descent down error surface is calculated by the following rule

$$\nabla W_{jk} = -\eta(\delta E / \delta W_{jk}), \quad (4.48)$$

where, $\eta$ =learning rate parameter and *E*=error function

The update of the weights for the $(n+1)^{th}$ pattern is given as

$$W_{jk}(n+1) = W_{jk}(n) + \nabla W_{jk}(n), \quad (4.49)$$

This network patterns were repeated for each pair of training network. Each pass through all training patterns was called cycle or epoch. The epochs were repeated to minimize the error. All input and output parameters were scaled between 0 to 1 to utilize the most sensitive part of neuron. As the output neuron was sigmoid, it can range its value from 0 to 1.

To test and validate the neural network model, a new data set was chosen. The results of the testing and validation were given bythe network performance, which was the correlation coefficient between predicted and observed values. Training was done using hidden layer.There was no danger of over fitting problems; hence network was trained with epochs.

## 4.12 Monte Carlo Simulation (MCS)

Simulation is generally defined as the process of replication of the real data based on a certain assumptions and the reality of the models are conceived (Kottegodaea al., 1998).

Monte Carlo Simulation was applied as an experimental probabilistic method to solve deterministic problems. Monte Carlo Simulation in computers can easily simulate a large number of experimental trials that have random outcomes. When MCS was applied to uncertainty estimation, random numbers were used to randomly sample parameters. The analysis was closer with the underlying physics of actual measurement processes that were probabilistic in nature. Nicolis et al., 1995 was pointed out in nature the process of measurement, by which the observer communicated with a physical system, and was limited by a finite precision.

### 4.12.1 MCS based Water Quality Model

Water quality model assessment was begun with a classical, deterministic water quality model that predicted pollutants, which overlaid probability theory in such a way that exceedance probability of the water quality indicator can be obtained. If the degradation reaction of the pollutant was a first order reaction, the water quality model equation was given as:

$$\frac{\partial C}{\partial t}+\frac{\partial(uC)}{\partial x}+\frac{\partial(vC)}{\partial y}+\frac{\partial(wC)}{\partial z}=E_x\frac{\partial^2 C}{\partial x^2}+E_y\frac{\partial^2 C}{\partial y^2}+E_z\frac{\partial^2 C}{\partial z^2}-K_1C, \qquad (4.50)$$

where $C$ is the concentration of the pollutant in mg/L, $t$ is the travel time in sec; $u$, $v$, $w$ are the longitudinal, lateral and vertical advective velocities in m/s respectively; $E_x$, $E_y$, $E_z$ are the longitudinal, lateral and vertical diffusion coefficients in $m^2/s$ respectively and $K_1$is the first order decay coefficient for a certain pollutant in $day^{-1}$. The $x$- and $y$-coordinates are in the horizontal plane, and $z$-coordinate is in the vertical plane.

In equation 4.50, the parameters $u$, $v$, $w$, $E_x$, $E_y$, $E_z$, $k_1$ are the main factors that influence the pollutant concentration $C$. The probability distribution of $C$ can more easily related to water quality by these parameters in the above equation as random variables.

To ensure that the simulation results were reasonable and that the sample distributions were sufficient, it is assumed that the $u$, $v$ and $w$ followed a uniform distribution, $U$ ($a$, $b$).

According to the study of Burn et al. 1985, the parameter values $E_x$, $E_y$, $E_z$and$k_1$ in the water quality model were distributed symmetrically and can be simulated as a normal distribution, $N(\mu,\sigma)$.

## 4.12.2 MCS-based Risk Assessment

In an MCS-based risk assessment, discrete values of the input random variables or model parameters were generated in a series of consistent with their probability distributions, and the water quality model was calculated for each generated input data set and also produced outputs in the form of statistical distribution. The input variables were sampled independently in MATLAB and the correlations among different water quality parameters were not considered. The process was repeated many times to evaluate the water quality risk by determining the regulatory limits.

The output from an MCS varied with the number of samples; that was, larger the number of sample was grater the output accuracy. According to the Kolmogorov Smirnov Test, the number of samples required for an MCS can be estimated using the following equation (Amstadter, 1971) was given as:

$$n \geq \lambda_\alpha^2 / D_n^2, \tag{4.51}$$

where $n$ is the sample number of the random variable, α is the confidence level, $\lambda_a$ is a constant and $D_n$ is the maximum desired error of the random variable.

In MCS, the attentions was paid to the water quality risk at different gauging stations in different seasons, i.e. temporally and were denoted as check points. Check points were important locations, such as water quality monitoring sections at three different stations. The water quality risk of a certain checkpoint was defined as followed:

$$risk_i = n_i / N \times 100\%, \tag{4.52}$$

where $risk_i$ is the risk at checkpoint $i$, $N$ is the total run number in the MCS, $n_i$ is the number of times that the deserved/predicted values of water quality indicator are beyond the regulatory limit at check point $i$.

As mentioned above, the mathematical formulations based on MCS was relatively simple, where its accuracy was primarily limited by the computational time. The assessment method had the capability of handling practically every possible case with its complexity; however it had not received overwhelming acceptance due to excessive computational effort was required.

## 4.13 Error Analysis

To know the percentage of error in input data, estimated data and predicted data by the above models of water quality parameters were done with the help of certain mathematical formulations and calculations like Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE).

### 4.13.1 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) was the quantity used to measure how close forecasts and predictions were to the eventual outcomes. The mean absolute error can be given as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|, \tag{4.53}$$

Where $e_i$ is the average absolute error, $e_i = |f_i - y_i|$, $f_i$ is the prediction and $y_i$ is the true value. The mean absolute error was a common measure of forecast error in time series analysis, where the terms 'mean absolute deviation' was sometimes used in confusion with the more standard definition of mean absolute deviation.

### 4.13.2 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error also known as Mean absolute Percentage Deviation, was a measure of accuracy of a method for constructing fitted time series values in statistics, specifically in trend estimation. It was usually expressed as a percentage, and was defined by the formula as given below:

$$M = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|, \tag{4.54}$$

Where $A_t$ is the actual value and $F_t$ is the forecasted value.

The absolute value in this formulation was summed for every fitted or forecasted point in time and was divided again by the number of fitted points; *n*, multiplying by 100, it gave percentage error.

### 4.13.3 Root Mean Squared Error (RMSE)

Root Mean Squared Error or Root Mean Squared Deviation was a measure of the differences between values predicted by model or an estimator and the actually observed values. These individual differences were called as residuals when the calculations were performed over the data sample that was used for estimation, and were known as estimation errors when computed out of the sample. The RMSE was served to aggregate the magnitudes of errors in

predictions for various times into single measure of predictive power. Root Mean Absolute Error was a good measure of accuracy, but was used only to compare forecasting errors of different models for a particular parameter and not between the parameter, as RMSE was a scale dependent error analysis.

The RMSE of an estimator, $\hat{\theta}$ with respect to an estimated parameter $\theta$ was defined as the square root of the mean square error, and was given as:

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E((\theta - \hat{\theta})^2)}, \tag{4.55}$$

For an unbiased estimator, the RMSE was the square root of variance, known as the standard error. The RMSE of predicted values $\hat{y}_t$ for times *t* of a regression's dependent variable *y* was computed for *n* different predictions as the square root of the mean of the squares of the deviations was given as:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}}, \tag{4.56}$$

## 4.14 Comparisons between the Models

The eleven water quality parameters in five selected gauging stations of Brahmani River were estimated and predicted temporally i.e. depending upon seasonal variation by different models such as Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN) and Monte Carlo Simulations (MCS) as discussed above. Comparisons were made with the help of graphical representation, correlation analysis and error analysis between the output values of the models and with the input values of the water quality parameters given to the applied models.

# CHAPTER V

# RESULTS AND DISCUSSIONS

The chapter describes the findings from the application of various tools as described in the preceeding chapter using the Time Series and Correlation analysis, Overall Water Quality Index calculation, Multivariate Statistical Analysis, Principal Component Analysis, Canonical Correlation Analysis, Factorial Analysis, Discriminant Analysis and Hierarchical Clustering. The modelling was done by Adaptive Neuro-Fuzzy Inference System by MATLAB, Artificial Neural Network and Monte Carlo Simulations.

## 5.1 Spearman's Rank Correlation Analysis

Correlation plots with Spearman's correlation coefficient value of all parameters at five selected consecutive gauging stations are made for summer, monsoon and winter for 12 months from 2003 to 2012. These plots provide a general indication of trend and supported observations were made later in statistical analysis. The correlation plots of the eleven water quality parameters are shown. The correlation plots of pH at three seasons are shown below in Figure 5.1, 5.2 and 5.3. At the right hand corner of the graph, it shows the numbers of data taken, mean, Standard deviation, maximum and minimum ranges of the data. Depending upon the maximum and minimum ranges of data, the graphs are plotted for Spearman's Rank Correlation Analysis.

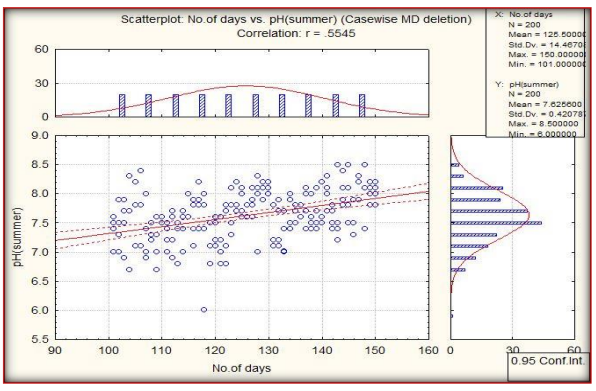


**Figure: 5.1 pH in Summer Season**

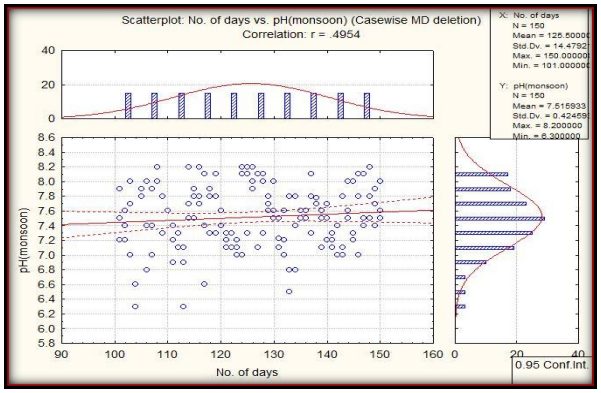


**Figure 5.2 pH in Monsoon Season**

The Spearman's rank correlation analysis of pH shows the moderate values of correlation coefficient ($R_{sp}$) of 0.55 for summer, 0.50 for monsoon and 0.55 for winter seasons respectively.

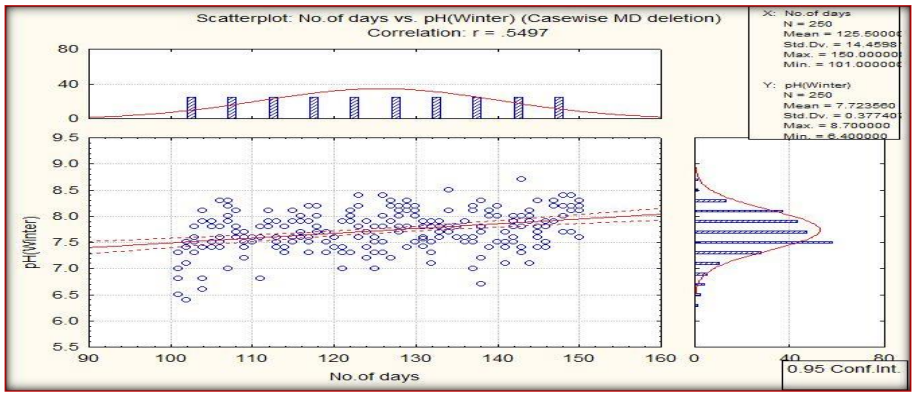


**Figure 5.3 pH in Winter Season**

The correlation plots of Dissolved oxygen (DO) are plotted below in Figure 5.4, 5.5 and 5.6 respectively and the correlation plots of Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) for the three seasons of summer, monsoon and winter are shown in Figure 5.7, 5.8 and . 5.9 & 5.10, 5.11 and 5.12 respectively.

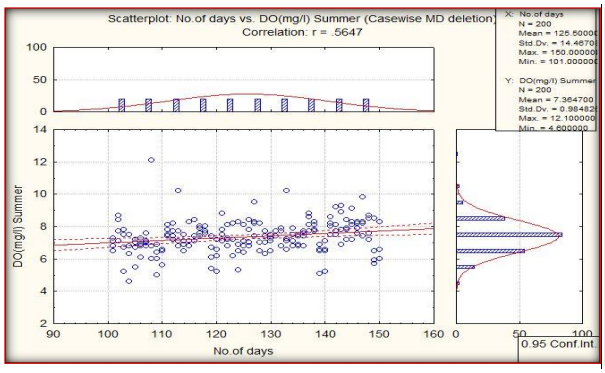


**Figure 5.4 DO in Summer Season**

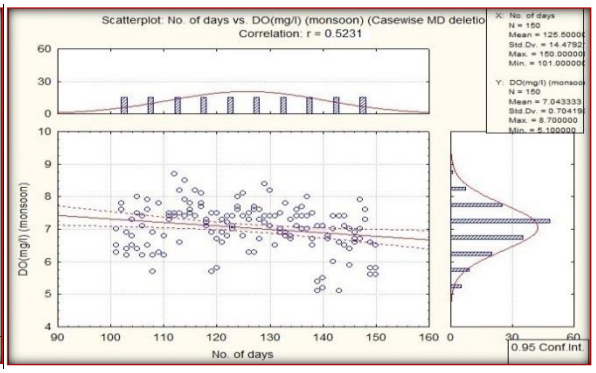


**Figure 5.5 DO in Monsoon Season**

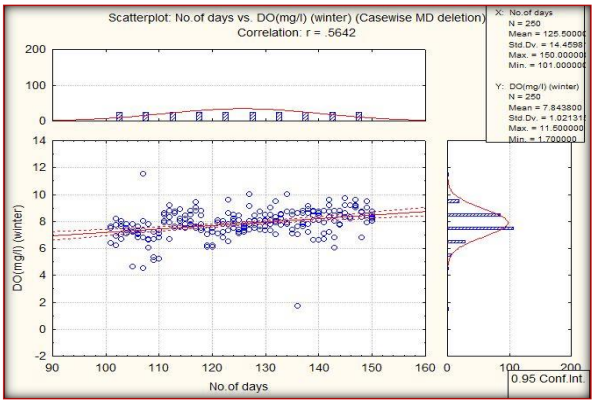


**Figure 5.6 DO in Winter Monsoon**

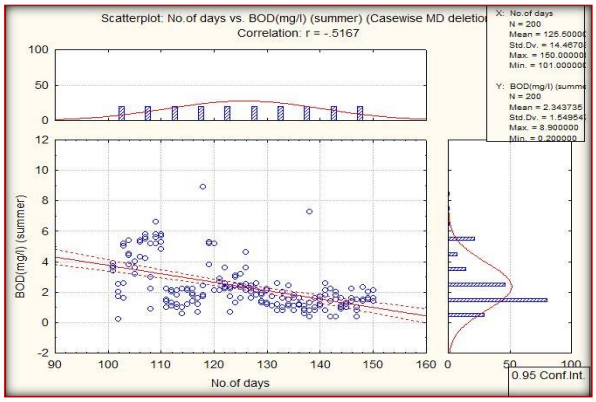


**Figure 5.7 BOD in Summer Season**

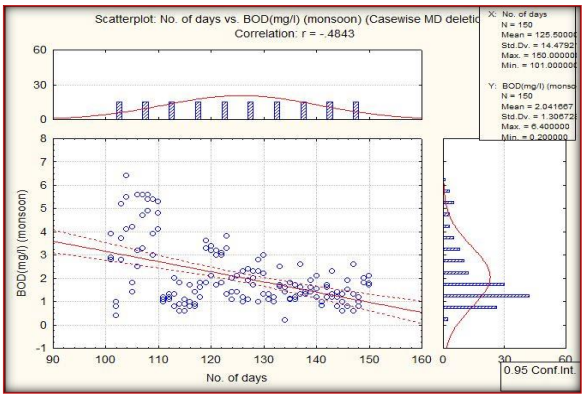


**Figure 5.8 BOD in Monsoon Season**

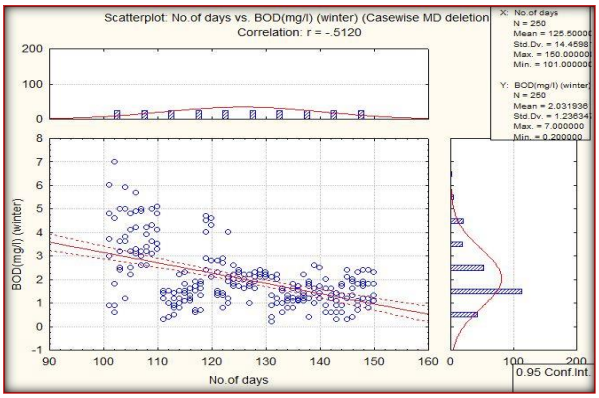


**Figure 5.9 BOD in Winter Season**

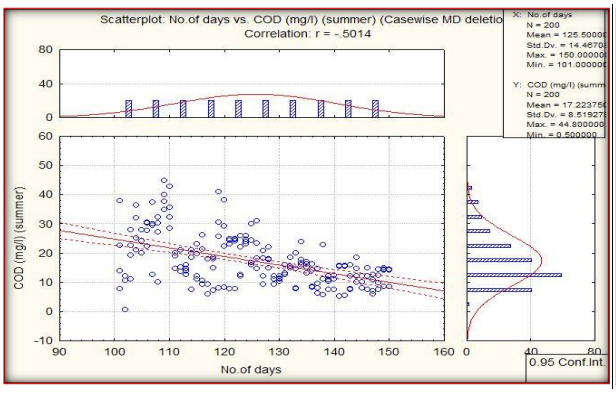


**Figure 5.10 COD in Summer Season**

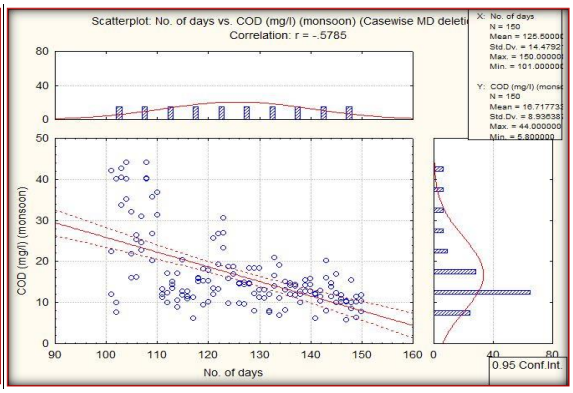


**Figure 5.11 COD in Monsoon season**

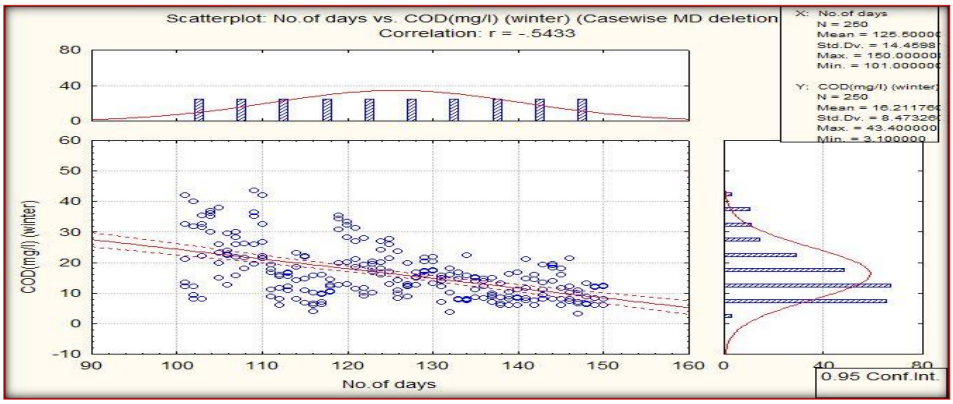


**Figure 5.12 COD in Winter Season**

The correlation coefficients ($R_{sp}$) of dissolved oxygen (DO) for summer, monsoon and winter seasons were 0.56, 0.52 and 0.56 respectively. The correlation coefficients ($R_{sp}$) for Biochemical Oxygen Demand (BOD) were -0.52, -0.48 and -0.51 for summer, monsoon and winter respectively. The negative correlations with moderate values are due to the decreasing values of BOD at the particular season. The values for COD were -0.50, -0.52 and -0.54 for the respective seasons as shown in the above figures. The correlation plots of Conductivity are shown in Figures 5.13, 5.14 and 5.15 respectively. The correlation coefficient for the

conductivity in three seasons of summer, monsoon and winter showed the negative moderate correlation values of -0.55, -0.51 and -0.54 respectively as extracted from the Spearman's correlation analysis.

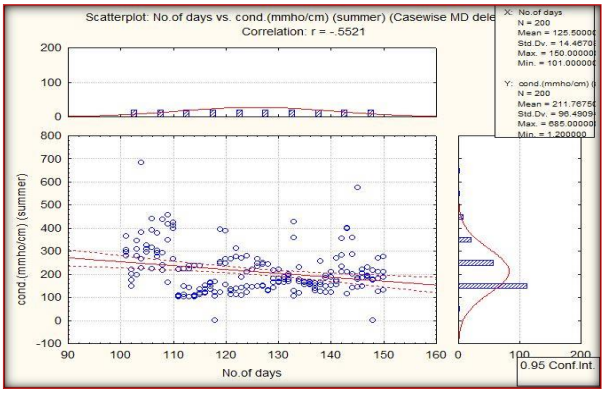


**Figure 5.13 EC in Summer Season**

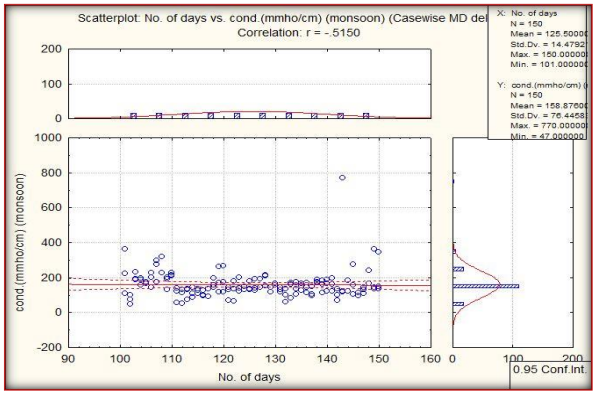


**Figure 5.14 EC in Monsoon Season**

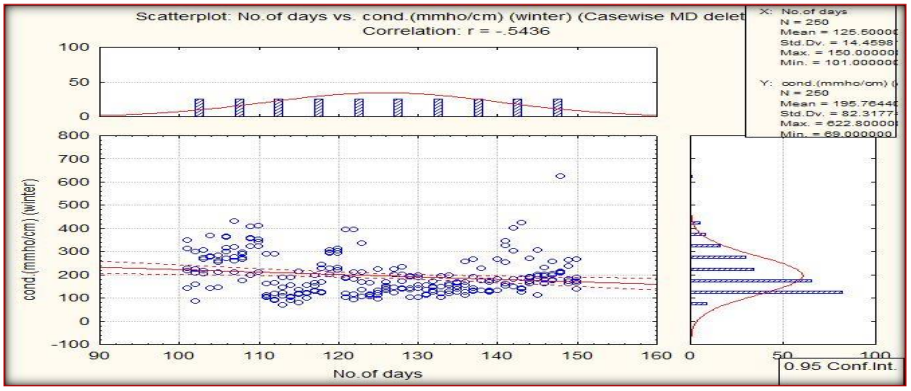


**Figure 5.15 EC in Winter Season**

The correlation plots of Nitrate-N for the three consecutive seasons are shown in Figures 5.16, 5.17 and 5.18 respectively and the correlation plots of Nitrogen as Ammonia ($NH_4$-N) are shown in Figure 5.19, 5.20 and 5.21 respectively. The correlation coefficients ($R_{sp}$) for Nitrate-N in the respective seasons are -0.52, -0.59 and -0.55 respectively. The values varied negatively with moderate variation. Likewise the negative correlation coefficients for $NH_4$-N are -0.62, -0.52 and -0.53 for respective seasons.

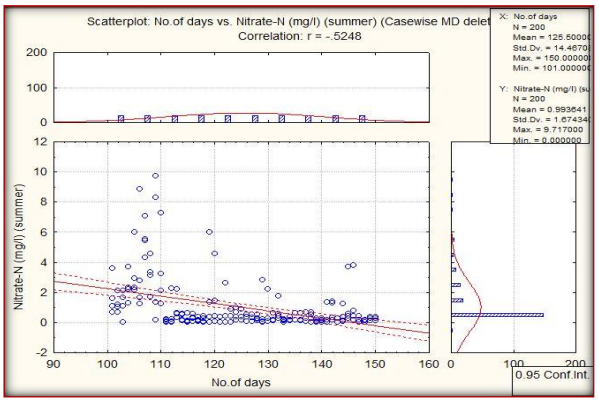


**Figure 5.16 Nitrate-N in Summer Season**

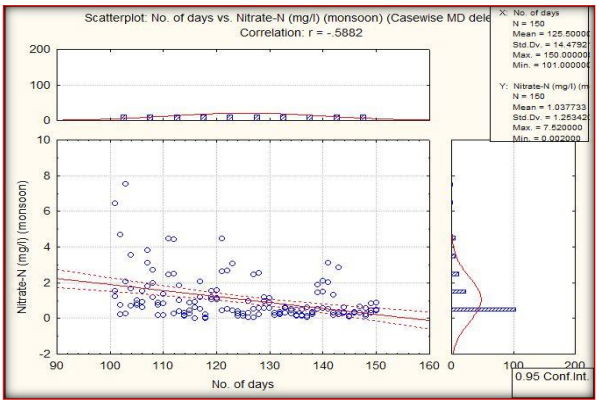


**Figure 5.17 Nitrate-N in Monsoon Season**

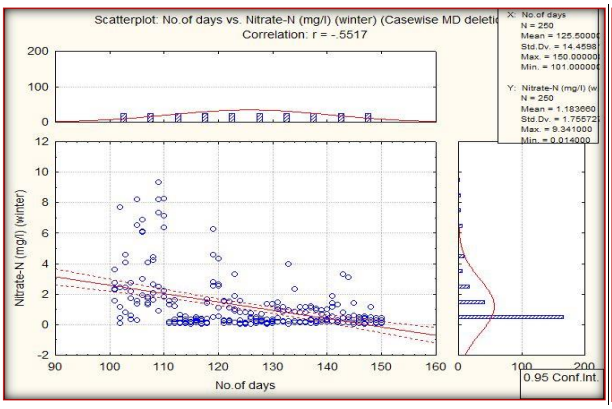


**Figure 5.18 Nitrate-N in Winter Season**

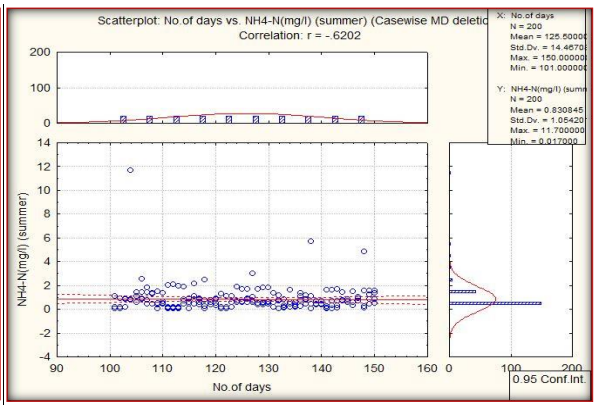


**Figure 5.19 $NH_4$-N in Summer Season**

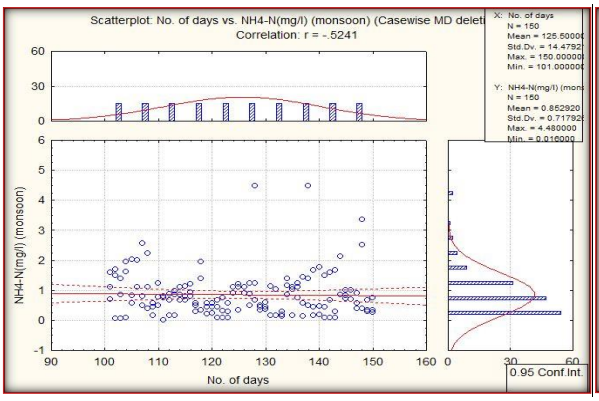


**Figure 5.20 $NH_4$-N in Monsoon Season**

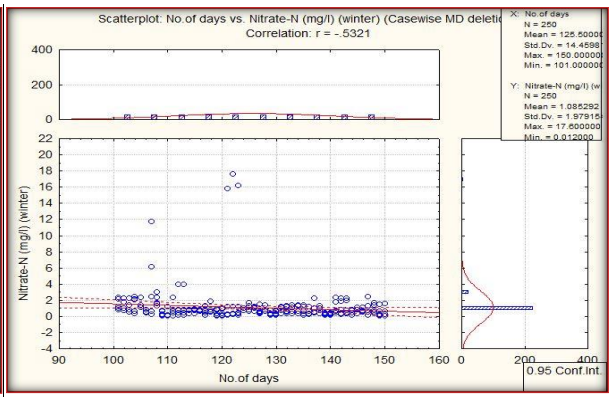


**Figure 5.21 $NH_4$-N in Winter Season**

The correlation plots for TC and FC are shown in Figures 5.22, 5.23, 5.24, 5.25, 5.26 and 5.27, which shows the spearman's rank correlation coefficient values of the TC and FC parameters. The correlation coefficients ($R_{sp}$) for TC in three respective seasons were -0.56, -0.56 and -0.56. The three correlation coefficients show the negative values and are moderately correlated. The correlation coefficients ($R_{sp}$) for FC in three seasons show the negative and moderate correlated values of -0.56, -0.53 and -0.57 respectively.

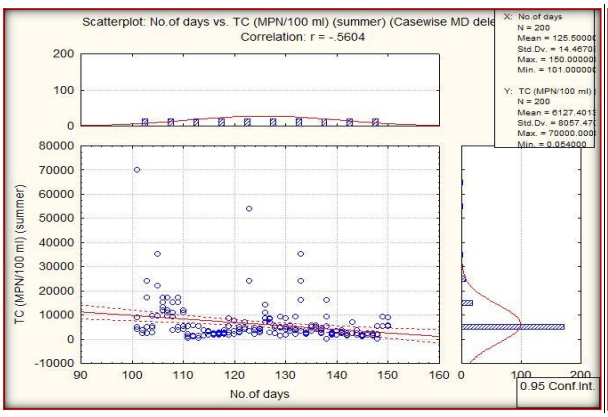


**Figure 5.22 TC in Summer Season**

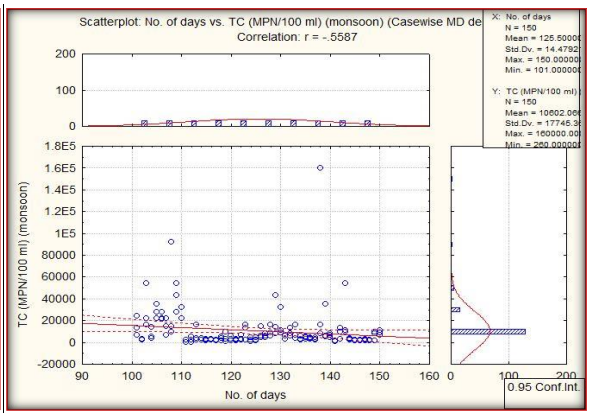


**Figure 5.23 TC in Monsoon Season**

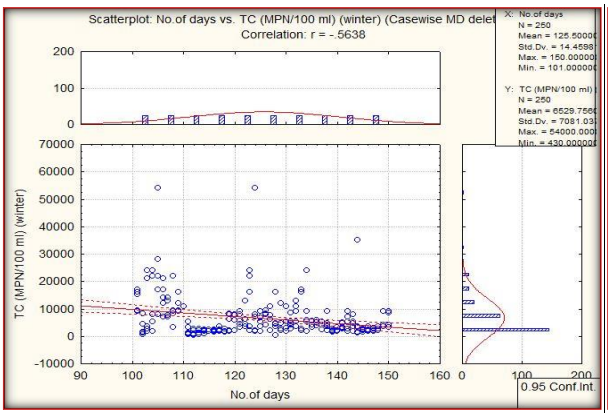


**Figure 5.24 TC in Winter Season**

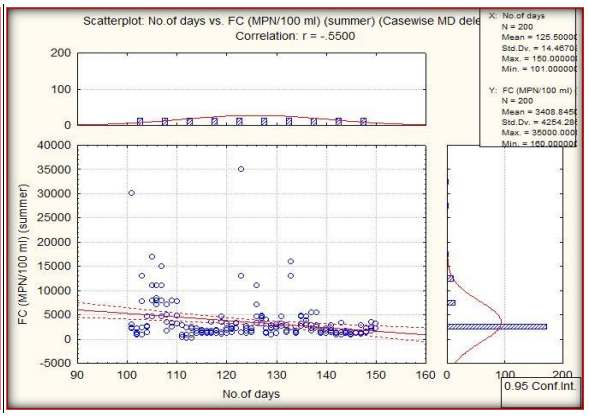


**Figure 5.25 FC in Summer Season**

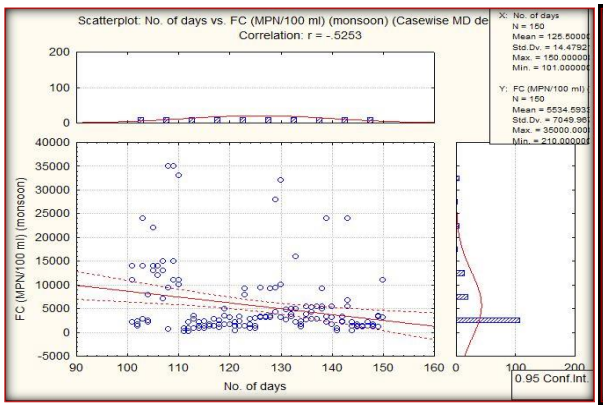


**Figure 5.26 FC in Monsoon Season**

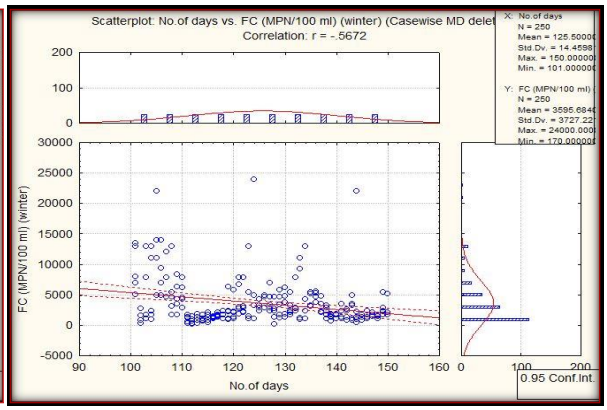


**Figure 5.27 FC in Winter Season**

The correlation graphs of TA as $CaCO_3$ and TH as $CaCO_3$ are shown in Figures 5.28, 5.29 and 5.30 & 5.31, 5.32 and 5.33 respectively. The Spearman's rank correlation coefficients ($R_{sp}$) for TA as $CaCO_3$ for three respective seasons are varied from positive to negative values, the values are 0.52, -0.57 and -0.54 respectively. The coefficients ($R_{sp}$) for TH as $CaCO_3$ are -0.53, -0.66 and -0.54 respectively for three seasons such as summer,

monsoon and winter. The positive values of correlation coeficients are due to the increasing trend of parameters with time and vice-versa for negatively correlated parameters.of the

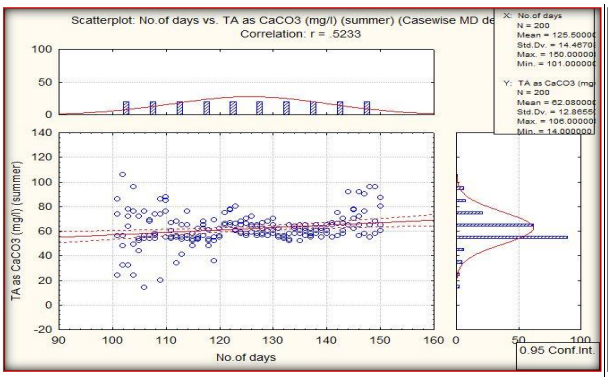


**Figure 5.28 TA as $CaCO_3$ in Summer Season**

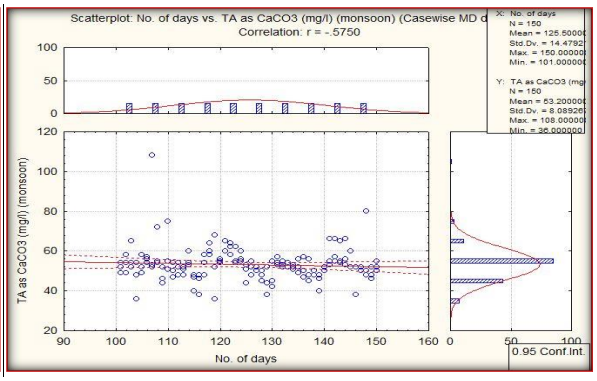


**Figure 5.29 TA as $CaCO_3$ in Monsoon Season**

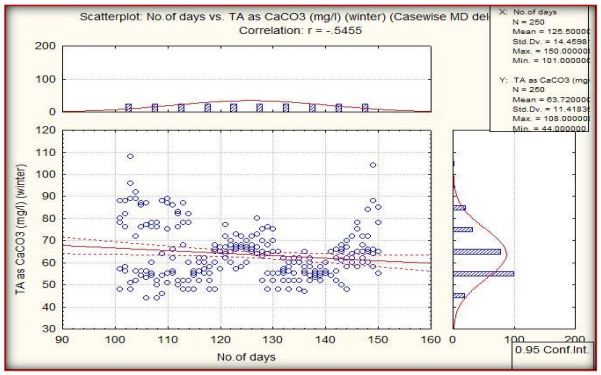


**Figure 5.30 TA as $CaCO_3$ in Winter Season**

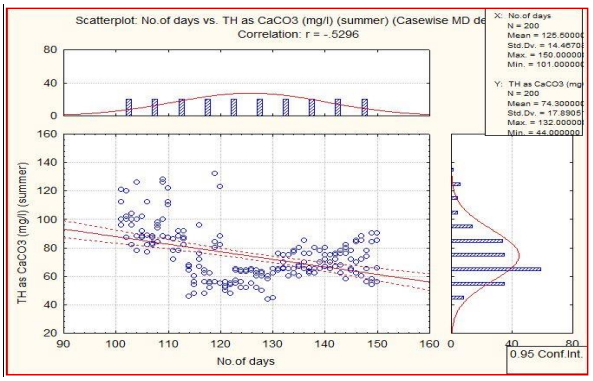


**Figure 5.31 TH as $CaCO_3$ in Summer Season**

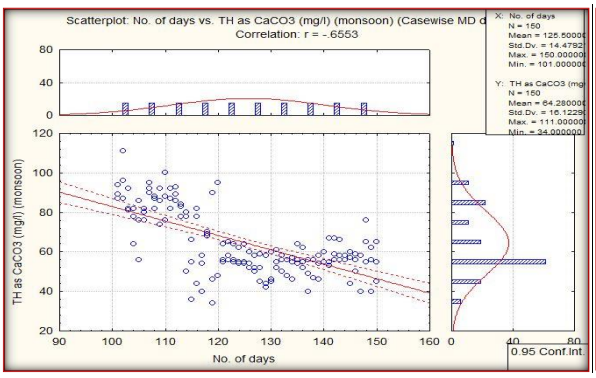


**Figure 5.32 TH as $CaCO_3$ in Monsoon Season**

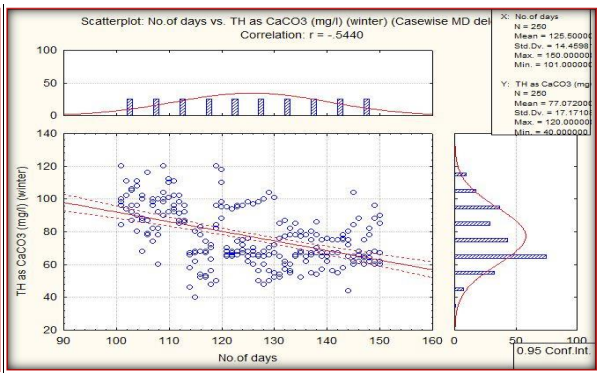


**Figure 5.33 TH as $CaCO_3$ in Winter Season**

## 5.2 Calculation of Parts of Water Quality Parameter in River Water

The parts of water quality parameter are calculated at Panposh down-stream to detect the increase or decrease of flow of pollution into the river water. Figures 5.34, 5.35 and 5.36 represented the parts of eleven selected parameters in water for summer, monsoon and winter respectively.

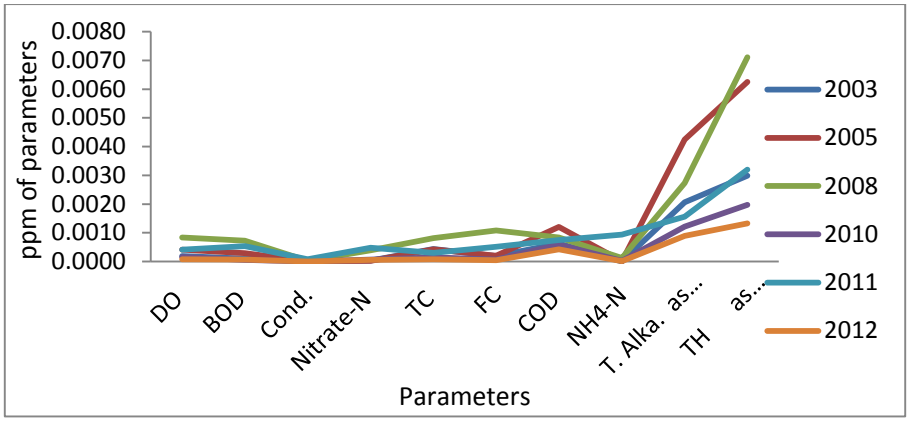


**Figure 5.34 Parts of parameters in water for summer season**

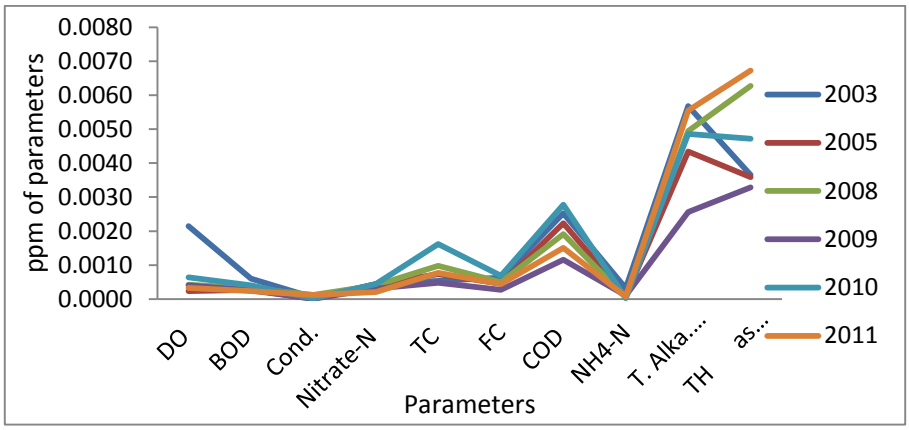


**Figure 5.35 Parts of parameters in water for monsoon season**

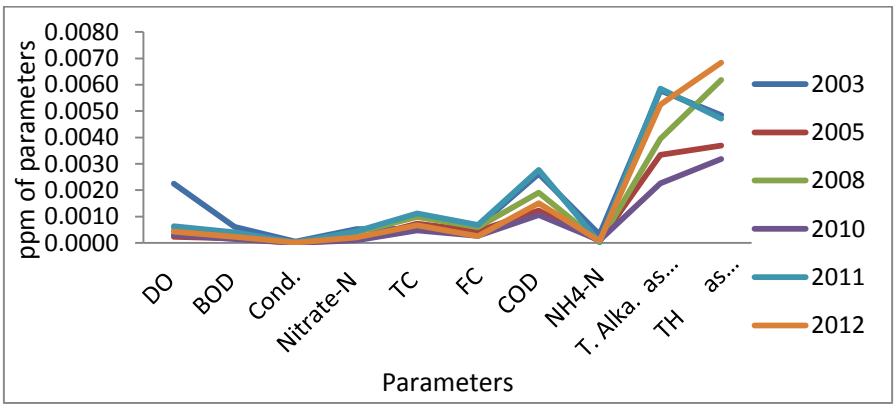


**Figure 5.36 Parts of parameters in water for winter season**

The graphical representations of the parts of parameters in three seasons are almost same, which concludes that the amounts of inflow of pollutants through the effluents into the river are almost same throughout the year**.**

## 5.3 Overall water Quality Index (WQI) Calculation

The water quality indices are calculated by applying the formulations and calculations as discussed in chapter IV. According to the water quality indices the level of pollution of water varies from excellent to unsuitable for use are known at five selected gauging stations in three seasons from January 2003 to December 2012. The mean, variance and standard deviation of the eleven water quality parameters are given in the Table 5.1.

Physio-chemical characteristics of surface water quality of Brahmani River depend upon the water quality parameters. These physio-chemical parameters show the temporal variations in water quality. Water is slightly alkaline with its $P^H$ ranging from 7.56 to 7.75. Dissolved Oxygen exceeds the permissible range and it varies from a mean value 8.60 mg/L in monsoon to a maximum value of 12.10 mg/L in summer. The values of Biochemical Oxygen Demand are within the permissible limits from 2.02 mg/L in monsoon to 2.35 mg/L in summer. Electrical Conductivity shows its maximum mean at summer season of 211.77 mmho/cm. The Nitrate-N in mg/L varies its value from 0.87 mg/L to 1.17 mg/L, which is the range with in the permissible limit. The total Coliform exceeds the permissible range in three consecutive seasons. Likewise the Faecal Coliform were within the permissible range in summer, exceeds slightly Conductivity of the surface water of Brahmani River Basin also within the permissible limits, which in monsoon and again within the range in winter season. The COD values are within the permissible range from 18-30 mg/L in three seasons. The Nitrogen as Ammonia exceeds slightly from its permissible range at summer and monsoon but exceeds a bit in winter season. The values of TA as $CaCO_3$ and TH as $CaCO_3$ are lower than their permissible limits in all the three seasons. The water quality Index shows the temporal variation of the given water quality parameters in Brahmani River Basin. For the easy interpretation of the water quality index values with temporal variation from 2003 to 2012 at five selected gauging stations, the variation in WQI values are graphically shown in Figure 5.37.

The Table 5.2 shown below includes the water quality index (WQI) as indicators for the level of pollution in three seasons at five gauging stations. From the above Table 5.2, it can be concluded that at most of the gauging stations in monsoon from 2003 to 2012 water quality index varies from excellent to good within the WQI range from 0-25 for excellent and 26-50, but the ranges were varied from good to poor in summer and winter. It is inferred from the results that overall quality of water is excellent for use at the sampling site in monsoon also in other seasons with some purification of water. The fluctuation can be attributed to the

industrial effluent discharge, sewage water disposal, agricultural waste disposal and other domestic effluents and hence the availability of water in the river.

**Table 5.1: Summary of Descriptive Statistics for water quality parameters**

| Summer Season | | | | | |
|---|---|---|---|---|---|
| Parameters | Minimum | Maximum | Mean | Std. Deviation | Variance |
| pH (s) | 6.00 | 8.50 | 7.66 | 0.44 | 0.20 |
| DO (mg/l) | 4.60 | 12.10 | 7.37 | 1.00 | 1.01 |
| BOD (mg/l) | 0.20 | 8.90 | 2.35 | 1.54 | 2.38 |
| Cond. mmho/cm | 1.20 | 685.00 | 211.77 | 96.49 | 9310.50 |
| Nitrate-N (mg/l) | 0.00 | 9.72 | 0.98 | 1.68 | 2.81 |
| TC (MPN/100 ml) | 0.00 | 70000.00 | 6127.40 | 8057.47 | 64922846.47 |
| FC (MPN/100 ml) | 160.00 | 35000.00 | 3408.85 | 4254.29 | 18098957.06 |
| COD (mg/l) | 0.50 | 44.80 | 17.21 | 8.51 | 72.49 |
| NH4-N (mg/l) | 0.00 | 11.70 | 0.83 | 1.08 | 1.16 |
| TA as CaCO3 (mg/l) | 14.00 | 106.00 | 62.08 | 12.87 | 165.52 |
| TH as CaCO3 (mg/l) | 44.00 | 132.00 | 74.30 | 17.89 | 320.07 |
| **Monsoon Season** | | | | | |
| pH | 6.00 | 8.20 | 7.56 | 0.48 | 0.23 |
| DO (mg/l) | 5.00 | 8.60 | 7.04 | 0.71 | 0.51 |
| BOD (mg/l) | 0.00 | 6.40 | 2.02 | 1.31 | 1.73 |
| Cond. mmho/cm | 47.00 | 770.00 | 158.88 | 76.45 | 5844.16 |
| Nitrate-N (mg/l) | 0.00 | 7.52 | 1.03 | 1.26 | 1.58 |
| TC (MPN/100 ml) | 260.00 | 160000.00 | 10602.07 | 17745.36 | 314897827.24 |
| FC (MPN/100 ml) | 210.00 | 35000.00 | 5534.59 | 7049.97 | 49702046.30 |
| COD (mg/l) | 5.80 | 44.00 | 16.73 | 8.94 | 79.88 |
| NH4-N (mg/l) | 0.00 | 4.48 | 0.83 | 0.72 | 0.53 |
| TA as CaCO3 (mg/l) | 36.00 | 108.00 | 53.20 | 8.09 | 65.44 |
| TH as CaCO3 (mg/l) | 34.00 | 111.00 | 64.28 | 16.12 | 259.95 |
| **Winter Season** | | | | | |
| pH | 6.40 | 8.70 | 7.75 | 0.38 | 0.14 |
| DO (mg/l) | 2.00 | 11.50 | 7.84 | 1.04 | 1.08 |
| BOD (mg/l) | 0.00 | 7.00 | 2.02 | 1.25 | 1.56 |
| Cond. mmho/cm | 69.00 | 622.80 | 195.76 | 82.32 | 6776.17 |
| Nitrate-N (mg/l) | 0.00 | 9.34 | 1.17 | 1.76 | 3.09 |
| TC (MPN/100 ml) | 430.00 | 54000.00 | 6529.76 | 7081.04 | 50141099.06 |
| FC (MPN/100 ml) | 170.00 | 24000.00 | 3595.68 | 3727.22 | 13892178.57 |
| COD (mg/l) | 3.10 | 43.40 | 16.21 | 8.47 | 71.74 |
| NH4-N (mg/l) | 0.00 | 17.60 | 1.06 | 1.99 | 3.96 |
| TA as CaCO3 (mg/l) | 44.00 | 108.00 | 63.72 | 11.42 | 130.38 |
| TH as CaCO3 (mg/l) | 40.00 | 120.00 | 77.07 | 17.17 | 294.85 |

**Table 5.2 Water Quality Index Values as Indicators**

| stations | R* | | R* | | R* | | R* | | | stations | R* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| pa.ma.03 | P | po.ma.12 | P | po.ap.11 | G | po.may.10 | G | po.jun.09 | G | po.jul.08 | Ex |
| pa.ma.04 | G | pa.ap.03 | G | po.ap.12 | G | po.may.11 | G | po.jun.10 | G | po.jul.09 | Ex |
| pa.ma.05 | P | pa.ap.04 | P | pa.may.03 | G | po.may.12 | G | po.jun.11 | G | po.jul.10 | Ex |
| pa.ma.06 | P | pa.ap.05 | P | pa.may.04 | G | pa.jun.03 | G | po.jun.12 | G | po.jul.11 | Ex |
| pa.ma.07 | G | pa.ap.06 | G | pa.may.05 | G | pa.jun.04 | G | pa.jul.03 | G | po.jul.12 | G |
| pa.ma.08 | G | pa.ap.07 | G | pa.may.06 | G | pa.jun.05 | G | pa.jul.04 | Ex | pa.aug.03 | G |
| pa.ma.09 | G | pa.ap.08 | G | pa.may.07 | P | pa.jun.06 | G | pa.jul.05 | G | pa.aug.04 | Ex |
| pa.ma.10 | G | pa.ap.09 | G | pa.may.08 | P | pa.jun.07 | G | pa.jul.06 | Ex | pa.aug.05 | G |
| pa.ma.11 | G | pa.ap.10 | G | pa.may.09 | G | pa.jun.08 | G | pa.jul.07 | G | pa.aug.06 | G |
| pa.ma.12 | G | pa.ap.11 | G | pa.may.10 | G | pa.jun.09 | P | pa.jul.08 | G | pa.aug.07 | G |
| ta.ma.03 | G | pa.ap.12 | G | pa.may.11 | G | pa.jun.10 | G | pa.jul.09 | G | pa.aug.08 | G |
| ta.ma.04 | G | ta.ap.03 | G | pa.may.12 | G | pa.jun.11 | G | pa.jul.10 | G | pa.aug.09 | G |
| ta.ma.05 | P | ta.ap.04 | P | ta.may.03 | G | pa.jun.12 | G | pa.jul.11 | G | pa.aug.10 | G |
| ta.ma.06 | G | ta.ap.05 | G | ta.may.04 | G | ta.jun.03 | G | pa.jul.12 | G | pa.aug.11 | G |
| ta.ma.07 | G | ta.ap.06 | G | ta.may.05 | G | ta.jun.04 | G | ta.jul.03 | Ex | pa.aug.12 | G |
| ta.ma.08 | G | ta.ap.07 | G | ta.may.06 | G | ta.jun.05 | G | ta.jul.04 | Ex | ta.aug.03 | Ex |
| ta.ma.09 | G | ta.ap.08 | P | ta.may.07 | G | ta.jun.06 | G | ta.jul.05 | Ex | ta.aug.04 | Ex |
| ta.ma.10 | P | ta.ap.09 | P | ta.may.08 | G | ta.jun.07 | G | ta.jul.06 | Ex | ta.aug.05 | G |
| ta.ma.11 | G | ta.ap.10 | P | ta.may.09 | G | ta.jun.08 | G | ta.jul.07 | Ex | ta.aug.06 | Ex |
| ta.ma.12 | G | ta.ap.11 | G | ta.may.10 | G | ta.jun.09 | G | ta.jul.08 | Ex | ta.aug.07 | Ex |
| ka.ma.03 | G | ta.ap.12 | G | ta.may.11 | G | ta.jun.10 | G | ta.jul.09 | Ex | ta.aug.08 | Ex |
| ka.ma.04 | P | ka.ap.03 | G | ta.may.12 | G | ta.jun.11 | G | ta.jul.10 | Ex | ta.aug.09 | G |
| ka.ma.05 | P | ka.ap.04 | G | ka.may.03 | G | ta.jun.12 | G | ta.jul.11 | Ex | ta.aug.10 | Ex |
| ka.ma.06 | G | ka.ap.05 | G | ka.may.04 | G | ka.jun.03 | G | ta.jul.12 | Ex | ta.aug.11 | Ex |
| ka.ma.07 | G | ka.ap.06 | G | ka.may.05 | G | ka.jun.04 | G | ka.jul.03 | Ex | ta.aug.12 | Ex |
| ka.ma.08 | G | ka.ap.07 | G | ka.may.06 | G | ka.jun.05 | G | ka.jul.04 | Ex | ka.aug.03 | Ex |
| ka.ma.09 | P | ka.ap.08 | G | ka.may.07 | G | ka.jun.06 | G | ka.jul.05 | G | ka.aug.04 | Ex |
| ka.ma.10 | P | ka.ap.09 | G | ka.may.08 | G | ka.jun.07 | G | ka.jul.06 | Ex | ka.aug.05 | Ex |
| ka.ma.11 | G | ka.ap.10 | G | ka.may.09 | G | ka.jun.08 | P | ka.jul.07 | Ex | ka.aug.06 | Ex |
| ka.ma.12 | G | ka.ap.11 | G | ka.may.10 | G | ka.jun.09 | G | ka.jul.08 | G | ka.aug.07 | Ex |
| au.ma.03 | G | ka.ap.12 | G | ka.may.11 | G | ka.jun.10 | G | ka.jul.09 | Ex | ka.aug.08 | Ex |
| au.ma.04 | G | au.ap.03 | G | ka.may.12 | G | ka.jun.11 | G | ka.jul.10 | Ex | ka.aug.09 | Ex |
| au.ma.05 | P | au.ap.04 | G | au.may.03 | G | ka.jun.12 | G | ka.jul.11 | Ex | ka.aug.10 | Ex |
| au.ma.06 | G | au.ap.05 | G | au.may.04 | G | au.jun.03 | G | ka.jul.12 | Ex | ka.aug.11 | G |
| au.ma.07 | G | au.ap.06 | G | au.may.05 | G | au.jun.04 | G | au.jul.03 | Ex | ka.aug.12 | G |
| au.ma.08 | G | au.ap.07 | P | au.may.06 | G | au.jun.05 | G | au.jul.04 | G | au.aug.03 | Ex |
| au.ma.09 | G | au.ap.08 | P | au.may.07 | P | au.jun.06 | G | au.jul.05 | P | au.aug.04 | Ex |
| au.ma.10 | G | au.ap.09 | P | au.may.08 | P | au.jun.07 | G | au.jul.06 | Ex | au.aug.05 | G |
| au.ma.11 | G | au.ap.10 | P | au.may.09 | G | au.jun.08 | G | au.jul.07 | Ex | au.aug.06 | G |
| au.ma.12 | G | au.ap.11 | G | au.may.10 | G | au.jun.09 | G | au.jul.08 | Ex | au.aug.07 | Ex |
| po.ma.03 | P | au.ap.12 | G | au.may.11 | G | au.jun.10 | G | au.jul.09 | Ex | au.aug.08 | Ex |
| po.ma.04 | P | po.ap.03 | G | au.may.12 | G | au.jun.11 | G | au.jul.10 | G | au.aug.09 | Ex |
| po.ma.05 | G | po.ap.04 | G | po.may.03 | G | au.jun.12 | G | au.jul.11 | Ex | au.aug.10 | G |
| po.ma.06 | P | po.ap.05 | G | po.may.04 | G | po.jun.03 | G | au.jul.12 | Ex | au.aug.11 | G |
| po.ma.07 | G | po.ap.06 | G | po.may.05 | G | po.jun.04 | G | po.jul.03 | Ex | au.aug.12 | Ex |
| po.ma.08 | G | po.ap.07 | G | po.may.06 | G | po.jun.05 | G | po.jul.04 | Ex | po.aug.03 | Ex |
| po.ma.09 | G | po.ap.08 | G | po.may.07 | P | po.jun.06 | G | po.jul.05 | G | po.aug.04 | G |
| po.ma.10 | G | po.ap.09 | G | po.may.08 | G | po.jun.07 | G | po.jul.06 | Ex | po.aug.05 | G |
| po.ma.11 | G | po.ap.10 | G | po.may.09 | G | po.jun.08 | G | po.jul.07 | Ex | po.aug.06 | Ex |

| | R* | | R* | stations | R* | | R* | | R* | | R* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| po.aug.07 | Ex | po.sep.06 | Ex | po.oc.05 | P | po.nov.04 | G | po.dec.03 | P | au.jan.12 | P |
| po.aug.08 | Ex | po.sep.07 | Ex | po.oc.06 | P | po.nov.05 | G | po.dec.04 | P | po.jan.03 | G |
| po.aug.09 | Ex | po.sep.08 | Ex | po.oc.07 | P | po.nov.06 | P | po.dec.05 | P | po.jan.04 | P |
| po.aug.10 | Ex | po.sep.09 | Ex | po.oc.08 | P | po.nov.07 | G | po.dec.06 | G | po.jan.05 | G |
| po.aug.11 | G | po.sep.10 | Ex | po.oc.09 | G | po.nov.08 | G | po.dec.07 | G | po.jan.06 | P |
| po.aug.12 | G | po.sep.11 | G | po.oc.10 | G | po.nov.09 | G | po.dec.08 | G | po.jan.07 | G |
| pa.sep.03 | P | po.sep.12 | G | po.oc.11 | P | po.nov.10 | G | po.dec.09 | G | po.jan.08 | G |
| pa.sep.04 | Ex | pa.oc.03 | G | po.oc.12 | G | po.nov.11 | G | po.dec.10 | G | po.jan.09 | G |
| pa.sep.05 | G | pa.oc.04 | G | pa.nov.03 | P | po.nov.12 | G | po.dec.11 | G | po.jan.10 | P |
| pa.sep.06 | G | pa.oc.05 | G | pa.nov.04 | P | pa.dec.03 | P | po.dec.12 | G | po.jan.11 | P |
| pa.sep.07 | G | pa.oc.06 | G | pa.nov.05 | G | pa.dec.04 | G | pa.jan.03 | P | po.jan.12 | G |
| pa.sep.08 | G | pa.oc.07 | P | pa.nov.06 | G | pa.dec.05 | P | pa.jan.04 | G | pa.feb.03 | P |
| pa.sep.09 | Ex | pa.oc.08 | G | pa.nov.07 | G | pa.dec.06 | G | pa.jan.05 | G | pa.feb.04 | P |
| pa.sep.10 | Ex | pa.oc.09 | P | pa.nov.08 | G | pa.dec.07 | G | pa.jan.06 | G | pa.feb.05 | P |
| pa.sep.11 | G | pa.oc.10 | G | pa.nov.09 | G | pa.dec.08 | Ex | pa.jan.07 | G | pa.feb.06 | P |
| pa.sep.12 | G | pa.oc.11 | G | pa.nov.10 | P | pa.dec.09 | G | pa.jan.08 | G | pa.feb.07 | G |
| ta.sep.03 | Ex | pa.oc.12 | G | pa.nov.11 | G | pa.dec.10 | Ex | pa.jan.09 | G | pa.feb.08 | G |
| ta.sep.04 | Ex | ta.oc.03 | G | pa.nov.12 | G | pa.dec.11 | G | pa.jan.10 | G | pa.feb.09 | G |
| ta.sep.05 | Ex | ta.oc.04 | G | ta.nov.03 | G | pa.dec.12 | G | pa.jan.11 | G | pa.feb.10 | P |
| ta.sep.06 | Ex | ta.oc.05 | G | ta.nov.04 | G | ta.dec.03 | P | pa.jan.12 | G | pa.feb.11 | G |
| ta.sep.07 | Ex | ta.oc.06 | P | ta.nov.05 | G | ta.dec.04 | P | ta.jan.03 | G | pa.feb.12 | G |
| ta.sep.08 | Ex | ta.oc.07 | G | ta.nov.06 | G | ta.dec.05 | G | ta.jan.04 | G | ta.feb.03 | G |
| ta.sep.09 | Ex | ta.oc.08 | G | ta.nov.07 | G | ta.dec.06 | G | ta.jan.05 | G | ta.feb.04 | G |
| ta.sep.10 | Ex | ta.oc.09 | G | ta.nov.08 | G | ta.dec.07 | G | ta.jan.06 | G | ta.feb.05 | G |
| ta.sep.11 | G | ta.oc.10 | G | ta.nov.09 | G | ta.dec.08 | G | ta.jan.07 | G | ta.feb.06 | G |
| ta.sep.12 | G | ta.oc.11 | G | ta.nov.10 | G | ta.dec.09 | G | ta.jan.08 | G | ta.feb.07 | G |
| ka.sep.03 | Ex | ta.oc.12 | G | ta.nov.11 | G | ta.dec.10 | G | ta.jan.09 | G | ta.feb.08 | G |
| ka.sep.04 | Ex | ka.oc.03 | G | ta.nov.12 | G | ta.dec.11 | G | ta.jan.10 | P | ta.feb.09 | G |
| ka.sep.05 | G | ka.oc.04 | G | ka.nov.03 | G | ta.dec.12 | G | ta.jan.11 | P | ta.feb.10 | G |
| ka.sep.06 | Ex | ka.oc.05 | G | ka.nov.04 | G | ka.dec.03 | G | ta.jan.12 | G | ta.feb.11 | G |
| ka.sep.07 | Ex | ka.oc.06 | G | ka.nov.05 | G | ka.dec.04 | G | ka.jan.03 | G | ta.feb.12 | G |
| ka.sep.08 | Ex | ka.oc.07 | G | ka.nov.06 | G | ka.dec.05 | G | ka.jan.04 | G | ka.feb.03 | G |
| ka.sep.09 | Ex | ka.oc.08 | G | ka.nov.07 | G | ka.dec.06 | G | ka.jan.05 | G | ka.feb.04 | G |
| ka.sep.10 | G | ka.oc.09 | G | ka.nov.08 | G | ka.dec.07 | P | ka.jan.06 | P | ka.feb.05 | P |
| ka.sep.11 | G | ka.oc.10 | G | ka.nov.09 | G | ka.dec.08 | G | ka.jan.07 | G | ka.feb.06 | G |
| ka.sep.12 | G | ka.oc.11 | G | ka.nov.10 | G | ka.dec.09 | G | ka.jan.08 | G | ka.feb.07 | G |
| au.sep.03 | G | ka.oc.12 | G | ka.nov.11 | G | ka.dec.10 | P | ka.jan.09 | G | ka.feb.08 | G |
| au.sep.04 | G | au.oc.03 | G | ka.nov.12 | G | ka.dec.11 | G | ka.jan.10 | P | ka.feb.09 | P |
| au.sep.05 | Ex | au.oc.04 | G | au.nov.03 | G | ka.dec.12 | G | ka.jan.11 | G | ka.feb.10 | G |
| au.sep.06 | G | au.oc.05 | G | au.nov.04 | G | au.dec.03 | P | ka.jan.12 | G | ka.feb.11 | G |
| au.sep.07 | Ex | au.oc.06 | G | au.nov.05 | G | au.dec.04 | P | au.jan.03 | G | ka.feb.12 | G |
| au.sep.08 | G | au.oc.07 | G | au.nov.06 | G | au.dec.05 | P | au.jan.04 | G | au.feb.03 | P |
| au.sep.09 | G | au.oc.08 | P | au.nov.07 | G | au.dec.06 | P | au.jan.05 | G | au.feb.04 | P |
| au.sep.10 | Ex | au.oc.09 | G | au.nov.08 | G | au.dec.07 | G | au.jan.06 | G | au.feb.05 | G |

NOTE: R*: Remarks, pa: Panposh D/S, ta: Talcher U/S, au: Aul, ka: Kamalanga D/S, po: Pottamundai.Ex: Excellent, G: Good, P: Poor, VP: Very poor

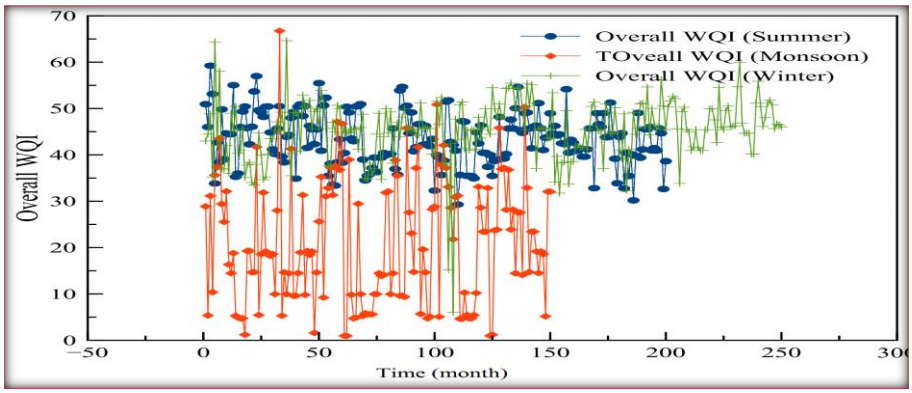


**Figure 5.37 Temporal variation of WQI**

## 5.4 Multivariate Analysis of Variance (MANOVA) with Discriminant Analysis

The summary of descriptive statistics of the results of the analysis is presented in Table 5.1, indicating the maximum, minimum, mean values of the parameters and the standard deviation. In summer season, TH as $CaCO_3$ has the high value of 132 mg/L, where as BOD has the highest value of 8.90 mg/L. Likewise in monsoon, TH as $CaCO_3$ has the high value of 111 mg/L and Ammonia-N has the maximum value of 4.48 mg/L. Similarly, in winter season again TH as $CaCO_3$ has the high value of 120 mg/L and BOD has a maximum value of 7 mg/L. The standard deviation around the means is substantially high and random. This may be the results of temporal as well as spatial changes and also the different anthropogenic activities surrounding the study area.

In order to explore the spatial variation among different gauging stations and seasonal changes, MANOVA is used to group these on the basis of spatial and temporal similarities as shown in Table 5.3. Analysis between and within the water quality parameters shows the small significant differences at a significant level of $\alpha$=0.05. From these results it can be concluded that the spatial sampling interval of 400 km is too a long distance to focus properly in Brahmani River Basin for the gauging stations situated.

The river is impacted by the same or similar influences over the spatial sampling interval which explain the lack of large significant variation in the properties in the study area in three consecutive seasons. This will be useful in monitoring protocol, but there should not be much change in the properties will be not be expected in the river. Any observed significant change could suggest a point source polluting it.

**Table 5.3: Model for Multivariate tests for all sample seasons on River Brahmani**

| Model | Value(S) | F(S) | p-value (Sig.)(S) | Value(M) | F(M) | p-value (Sig.)(M) | Value(W) | F(W) | p-value (Sig.)(W) |
|---|---|---|---|---|---|---|---|---|---|
| Pillai's Trace | 1.33 | 8.49 | 0 | 1.27 | 5.85 | 0 | 1.11 | 8.35 | 0 |
| Wilks' Lambda | 0.15 | 10.25 | 0 | 0.13 | 8.34 | 0 | 0.22 | 10.11 | 0 |
| Hotelling's Trace | 2.96 | 12.36 | 0 | 4.07 | 12.34 | 0 | 2.32 | 12.32 | 0 |
| Roy's Largest Root | 2.06 | 35.27 | 0 | 3.44 | 43.15 | 0 | 1.74 | 37.59 | 0 |

NOTE: S: Summer, M: Monsoon, W: Winter

**Table 5.4: Test of Equality of Group Means**

| Parameters | Wilks' Lambda (S) | F(S) | Sig. (S) | Wilks' Lambda (M) | F(M) | Sig.(M) | Wilks' Lambda (W) | F(W) | Sig. (W) |
|---|---|---|---|---|---|---|---|---|---|
| pH | 0.97 | 2.35 | 0.21 | 0.99 | 2.49 | 0.1 | 0.98 | 1.43 | 0.27 |
| DO | 0.98 | 1.52 | 0.22 | 0.97 | 2.29 | 0.11 | 0.99 | 1.31 | 0.27 |
| BOD | 0.93 | 7.51 | 0 | 0.98 | 1.55 | 0.22 | 0.95 | 6.33 | 0 |
| Conductivity | 0.97 | 3.3 | 0.04 | 0.98 | 1.49 | 0.23 | 0.99 | 1.43 | 0.24 |
| Nitrate-N | 0.93 | 7.68 | 0 | 0.97 | 2.17 | 0.12 | 0.97 | 3.7 | 0.03 |
| TC | 0.98 | 1.52 | 0.22 | 0.99 | 0.62 | 0.54 | 0.99 | 0.95 | 0.39 |
| FC | 0.99 | 0.59 | 0.56 | 1 | 0.34 | 0.71 | 1 | 0.33 | 0.72 |
| COD | 0.97 | 3.08 | 0.05 | 0.97 | 2 | 0.14 | 0.91 | 11.67 | 0 |
| $NH_4$-N | 0.98 | 2.2 | 0.11 | 0.98 | 1.29 | 0.28 | 1 | 0.02 | 0.98 |
| TA as $CaCO_3$ | 0.99 | 0.66 | 0.52 | 0.99 | 0.74 | 0.48 | 0.99 | 0.92 | 0.4 |
| TH as $CaCO_3$ | 0.98 | 2.17 | 0.12 | 0.97 | 2.4 | 0.09 | 0.97 | 4.18 | 0.02 |

The test of equality of group is as given in Table 5.4. It measures each parameter's potential before the discriminate model is created. Each test displays the results of one-way ANOVA for the parameter using season as the grouping variable. As shown by p-value (Sig.), all the water quality parameters significantly contributs to the model. The value of Wilk's lambda indicates the parameter is better at discriminating between groups. From the table 5.4 the parameters discriminating ability is ranked from the highest to the lowest in summer, as follows: Nitrate-N, BOD, Electrical Conductivity, COD, pH, $NH_4$-N, TH as $CaCO_3$, TC, DO, TA as $CaCO_3$ and FC. Likewise the parameters listed in monsoon are: pH, TH as $CaCO_3$, DO, Nitrate-N, COD, BOD, Conductivity, $NH_4$-N, TA as $CaCO_3$, TC and FC. For the season of winter the parameters are listed as: COD, BOD, TH as $CaCO_3$, Nitrate-N, Conductivity, pH, DO, TC, TA as $CaCO_3$, FC and $NH_4$-N.

The graphical representation in Figure 5.38 shows the Fisher's discriminate functions and Table 5.5 shows the Eigen values for discriminate functions for the three seasons. Two discriminate functions are obtained and the total variance cumulative was 100% between the seasonal months. The first function in the summer explained 63.8% of the total variance between the months while second function explained 36.12%. For the season of monsoon, the

first and second function explained 63.2% and 36.7% respectively. Likewise for the season of winter, the two functions explained 63.9% and 36.0% respectively. Fisher's linear discriminant function coefficient which can be used for predicting the likely season a particular water sample is collected in Brahmani River. Temporal variation of the River Brahmani is also analysed using discriminate functions as shown in Figure 5.38.

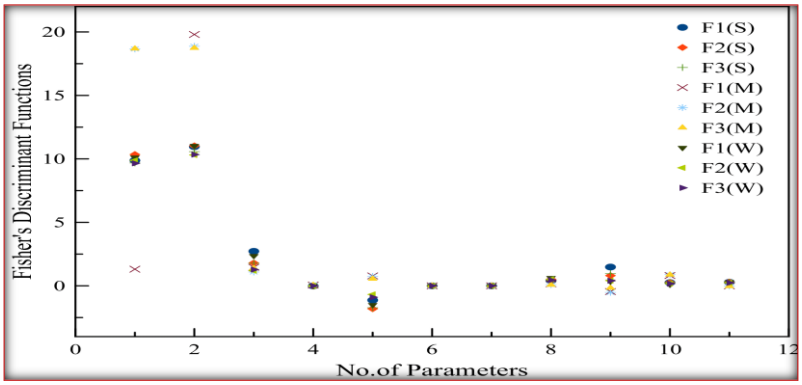


**Figure 5.38: Fisher's Discriminate Functions in Three Seasons**

**Table 5.5 Eigen values for Discriminate Functions for all the three Seasons**

| Summer | | | |
|---|---|---|---|
| Function | Eigen value | % of Variance | Cumulative % |
| 1 | 1.163[a] | 63.88 | 63.88 |
| 2 | 1.092[a] | 36.12 | 100 |
| Monsoon | | | |
| 1 | 1.118[a] | 63.26 | 63.26 |
| 2 | 1.068[a] | 36.74 | 100 |
| Winter | | | |
| 1 | 1.134[a] | 63.98 | 63.98 |
| 2 | 1.075[a] | 36.02 | 100 |
| a. First 2 canonical discriminant functions were used in the analysis. | | | |

Wilk's lambda is equal to the proportion of the total variance in the discriminate scores not explained by differences among groups. Smaller values of Wilk's lambda tests indicate greater discriminatory ability of the function. The small values of Wilk's lambda of 0.079 in summer to 0.082 in winter shows the discriminate functions, which expressed those very small portions of the discriminate scores, is not explained by the differences among seasonal months in three seasons. The two functions for three seasons are given in Table 5.6. The Chi-square test as given in the table is tested the hypothesis that all the means of the functions are listed as functions 1 and 2, and are equal across groups of seasonal months. The small p-value (sig.) of 0.00 to 0.40 indicates that discriminate functions separating the seasons from one another moderately.

**Table 5.6 Wilk's Lambda Test for Discriminate Function for Temporal Variation**

| Summer | | | |
|---|---|---|---|
| Test of Function(s) | Wilks' Lambda | Chi-square | p-Value (Sig.) |
| 1 through 2 | 0.079 | 46.1 | 0 |
| 2 | 0.092 | 16.99 | 0.05 |
| Monsoon | | | |
| 1 through 2 | 0.084 | 25.26 | 0.19 |
| 2 | 0.094 | 9.41 | 0.4 |
| Winter | | | |
| 1 through 2 | 0.082 | 48.04 | 0 |
| 2 | 0.093 | 17.61 | 0.04 |

Temporal variation of the River is also analysed using discriminate functions as shown in Table 5.7. The validity of using discriminate functions for the prediction is certain as 99.2% of original group cases correctly classified or predicted in Table 5.8.

**Table 5.7 Discriminate Functions for Temporal Variation**

| | Functions | | | | | |
|---|---|---|---|---|---|---|
| | S1 | S2 | M1 | M2 | W1 | W2 |
| pH | 0.02 | 0.69 | 0.65 | -0.55 | 0.11 | 0.35 |
| DO | 0.02 | 0.72 | 0.70 | 0.33 | 0.12 | 0.37 |
| BOD | 0.53 | 0.09 | -0.06 | 0.56 | 0.15 | 0.63 |
| Electrical Conductivity | 0.00 | 0.00 | 0.00 | 0.01 | -0.01 | 0.01 |
| Nitrate-N | 0.37 | -0.10 | 0.21 | -0.22 | 0.10 | -0.56 |
| TC | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FC | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| COD | -0.06 | 0.07 | 0.07 | 0.01 | 0.10 | 0.01 |
| NH4-N | 0.36 | -0.21 | -0.31 | 0.49 | -0.01 | -0.04 |
| TA as CaCO3 | -0.01 | -0.01 | -0.03 | -0.05 | 0.00 | -0.05 |
| TH as CaCO3 | -0.01 | 0.01 | 0.03 | -0.01 | 0.02 | 0.03 |

*$S_{1 \text{ and }} S_2$ represent summer season. $M_1$ and $M_2$ represents monsoon season. $W_1$ and $W_2$ represent winter seasons

The structure matrix as given in Table 5.9 shows the correlation of each parameter with the discriminate functions. The ordering of the discriminating ability of the parameters as indicated by the correlations is close to that indicated by using the Wilk's lambda criterion in Table 5.4. This is because the structure matrix is unaffected by the co-linearity among the parameters. The discrepancies are observed using the table of standardised discriminate functions as in Table 5.7. The $S_1$ and $S_2$ are extracted the discriminate functions by SPSS for summer season likewise, $M_1$, $M_2$, $W_1$ and $W_2$ are for monsoon and winter seasons. The discriminating ability of the parameter in the model are compared with those indicated by the

Wilk's lambda criteria in Table 5.4, this is obviously due to the existing co-linearity among parameters which could have inflated the discriminating ability of some of these water quality parameters.

**Table 5.8 Classification Results of Discriminate analysis for all the three seasons**

| Season | Predicted Group Membership | | | Percent Correct |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 1 | 40 | 50 | 10 | 100 |
| 2 | 25 | 37.9 | 37.1 | 100 |
| 3 | 20 | 42 | 38 | 100 |

**Table 5.9 Structure Matrix of each parameter with the discriminate functions**

| | Function | | | | | |
|---|---|---|---|---|---|---|
| | S1 | S2 | M1 | M2 | W1 | W2 |
| pH | 0.587* | 0.231 | 0.187 | -0.179 | 0.768* | 0.143 |
| DO | .659* | 0.243 | 0.219 | .475* | .530* | 0.425 |
| BOD | .652* | 0.306 | .477* | -0.201 | .452* | -0.188 |
| Electrical Conductivity | .386* | 0.315 | -0.298 | .379* | 0.034 | .389* |
| Nitrate-N | .335* | -0.208 | -0.194 | .438* | -0.014 | .039* |
| TC | -.185* | -0.109 | .215* | -0.130 | .215* | -0.130 |
| FC | 0.208 | .512* | 0.382 | .383* | .829* | 0.184 |
| COD | -0.116 | .379* | -.228* | 0.188 | .239* | 0.015 |
| NH4-N | -0.161 | .349* | .514* | 0.012 | -.268* | 0.116 |
| TA as CaCO3 | 0.290 | .300* | .520* | -0.113 | .460* | 0.270 |
| TH as CaCO3 | -0.106 | .211* | -.198* | -0.007 | .120* | 0.098 |
| *. Largest absolute correlation between each variable and any discriminant function | | | | | | |

*****S1 and S2 represent summer season. M1 and M2 represents monsoon season. W1 and W2 represent winter seasons

## 5.5 Principal Component Analysis and Factor Analysis

The data matrix is normalized for cluster analysis. Kaiser-Meyer-Olkin (KMO) and Bartlett's test are performed to examine the suitability of the data for principal component analysis/factor analysis. High value (close to 1) of KMO generally indicates that principal component analysis or factor analysis may be useful. KMO measures the sampling adequacy having values of 0.68, 0.69 and 0.75 for summer, monsoon and winter respectively. Bartlett's test of sphericity indicates whether correlation matrix is an identity matrix, which would indicate that variables are unrelated. The significance level is 0.06, 0.05 and 0.08 for three respective seasons in this study, demonstrating significant relationships among variables.

The Table 5.10 provides the correlation matrix of the water quality parameters obtained from the PCA. Only few parameters exhibited significant correlation with each other. High and positive correlation can be observed in summer between Nitrate-N, BOD, COD, Conductivity, TH as CaCO3, TC and FC, the strong correlations in monsoon are in between TC, FC, COD and BOD and in winter the positive and strong correlation are in between COD, BOD, TC, FC, TA as CaCO3, Nitrate-N, Electrical Conductivity and TH as $CaCO_3$ ($r$ = 50 to 0.93) which is responsible for faecal contamination in river. BOD and COD are positively correlated with each other in all the seasons which indicate contamination of organic matter. DO shows the negative correlation with pH, BOD and other parameters in summer, DO is correlated with some of the parameter in monsoon and winter because of solubility of oxygen as organic matter is partially oxidized by oxygen.

Table 5.11 represents the determined initial principal component (PC), its Eigen value and cumulative % of variance contributed in each PC. The Figures 5.39, 5.40 and 5.41 show the scree plot of the Eigen value for each component in which four principal components are obtained with Eigen value > 1 summing almost 70% of the total variance in the water quality dataset. The scree plot shows a pronounced change of slope after the third Eigen value (Cattel and Jas pers, 1967). Eigen values account that the first four PC is the most significant component which represents more than 60% of the variance in water quality of river Brahmani.

Component loading (correlation coefficients) as shown in Figure 5.42 measures the degree of closeness between the variables and the PC. The largest loading either positive or negative, suggests the meaning of dimensions; positive loading indicates that the contribution of the variables increased with the increasing loading in dimension, and negative loading indicates a decrease with decreased loading. Generally, component loadings larger than 0.45 are taken into consideration in the interpretation, in other words, the most significant variables in the components represent by high loadings has been taken into consideration in evaluation of the components (Mazlum et al., 1999). Component loadings and communalities for each variable in four selected component before varimax rotation are graphically shown by Figure 5.42 and after varimax rotation in Figure 5.43. Communalities provide an index to the efficiency of the reduced set of components and degree of contribution for each variable of the selected four principal components. The first PC in summer is accounting for 39.2% of the total variance of strong and positively correlated with BOD, Conductivity, Nitrate-N, TC, FC and COD where as DO shows negative correlation to the variance. The first PC in monsoon has a 28.6% of total variance. The PC is strong and positively correlated with BOD, Conductivity, TC, FC, COD and TH as $CaCO_3$ and is negatively correlated with DO.

Likewise the first PC in winter has 36.3% of total variance and is strongly correlated with BOD, Conductivity, Nitrate-N, TC, FC, COD and TH as $CaCO_3$. Here, also the PC is negatively correlated with DO like other first PCs in summer and monsoon. These correlations of first PC are explained that if large amount of dissolved oxygen consumed by large amount of organic matter in urban waste water, which are mainly consisted of carbohydrate, proteins and lipids. The organic reaction leads to anaerobic fermentation process leading to high ammonia and organic acid.

**Table 5.10 Correlation Matrix**

| Summer | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | pH | DO | BOD | Cond. | Nitrate-N | TC | FC | COD | NH4-N | TA as CaCO3 | TH as CaCO3 |
| pH | 1 | | | | | | | | | | |
| DO | -0.04 | 1 | | | | | | | | | |
| BOD | -0.23 | -0.37 | 1 | | | | | | | | |
| Cond. | -0.08 | -0.11 | 0.42 | 1 | | | | | | | |
| Nitrate-N | -0.2 | -0.12 | 0.52 | 0.48 | 1 | | | | | | |
| TC | -0.06 | -0.16 | 0.25 | 0.24 | 0.22 | 1 | | | | | |
| FC | -0.02 | -0.12 | 0.3 | 0.21 | 0.22 | 0.93 | 1 | | | | |
| COD | -0.19 | -0.34 | 0.61 | 0.37 | 0.51 | 0.34 | 0.33 | 1 | | | |
| NH4-N | 0.02 | -0.15 | 0.06 | 0.08 | 0.01 | 0.03 | 0.01 | 0.09 | 1 | | |
| TA as CaCO3 | 0.11 | 0 | -0.02 | 0.09 | -0.2 | 0 | -0.02 | -0.05 | 0.1 | 1 | |
| TH as CaCO3 | -0.14 | -0.23 | 0.42 | 0.54 | 0.48 | 0.19 | 0.15 | 0.37 | 0.08 | 0.16 | 1 |
| Monsoon | | | | | | | | | | | |
| pH | 1 | | | | | | | | | | |
| DO | -0.1 | 1 | | | | | | | | | |
| BOD | -0.07 | -0.22 | 1 | | | | | | | | |
| Cond. | 0.1 | -0.35 | 0.3 | 1 | | | | | | | |
| Nitrate-N | -0.16 | -0.04 | 0.3 | -0.01 | 1 | | | | | | |
| TC | 0.11 | -0.25 | 0.3 | 0.33 | -0.05 | 1 | | | | | |
| FC | 0.11 | -0.17 | 0.43 | 0.36 | -0.01 | 0.68 | 1 | | | | |
| COD | -0.1 | -0.19 | 0.75 | 0.32 | 0.21 | 0.32 | 0.4 | 1 | | | |
| NH4-N | 0.12 | 0.05 | 0.02 | 0.18 | -0.16 | 0.12 | 0.12 | 0.11 | 1 | | |
| TA as CaCO3 | -0.07 | -0.14 | 0.13 | 0.21 | 0.05 | 0.05 | 0.04 | 0.16 | -0.03 | 1 | |
| TH as CaCO3 | -0.13 | 0.03 | 0.34 | 0.2 | 0.37 | 0.18 | 0.23 | 0.46 | 0.08 | 0.28 | 1 |
| | | Winter | | | | | | | | | |
| pH | 1 | | | | | | | | | | |
| DO | 0.01 | 1 | | | | | | | | | |
| BOD | 0.01 | 0 | 1 | | | | | | | | |
| Cond. | 0.12 | 0.3 | 0.5 | 1 | | | | | | | |
| Nitrate-N | 0.15 | -0.07 | 0.36 | 0.51 | 1 | | | | | | |
| TC | 0.08 | -0.08 | 0.31 | 0.1 | 0.1 | 1 | | | | | |
| FC | 0.11 | -0.12 | 0.28 | 0.01 | 0.1 | 0.87 | 1 | | | | |
| COD | -0.09 | -0.12 | 0.54 | 0.38 | 0.28 | 0.16 | 0.11 | 1 | | | |
| NH4-N | 0.04 | -0.26 | -0.07 | -0.2 | -0.25 | 0.05 | 0.07 | -0.17 | 1 | | |
| TA as CaCO3 | 0.04 | 0.15 | 0.08 | 0.21 | 0.08 | 0.15 | 0.09 | 0.17 | -0.08 | 1 | |
| TH as CaCO3 | -0.12 | 0.16 | 0.03 | 0.32 | 0.21 | -0.03 | -0.08 | 0.18 | -0.23 | 0.63 | 1 |

The second PC in summer is highly loaded with FC (*Faecal Coliform*) and TC (*Total Coliform*), in monsoon it is highly loaded with Nitrate-N and is negatively correlated with pH. Likewise, in winter it is negative and highly loaded with TC and FC. Third PC in summer is highly loaded with pH, $NH_4$-N, TA as $CaCO_3$ and TH as $CaCO_3$. In monsoon, the PC is highly loaded with DO and likewise in winter, the PC is highly loaded with pH, DO and TA as $CaCO_3$. The third PC's correlation with these parameters represents the physicochemical source of the variability. Fourth PC in summer is highly loaded with DO, in monsoon; it is loaded with $NH_4$-N. These correlations indicate the waste water from domestic and industrial and its organic load disposed to the river.

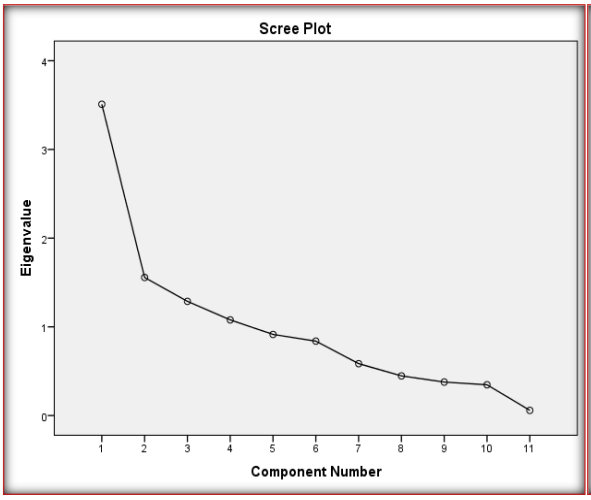


**Figure 5.39 Scree Plot in Summer Season**

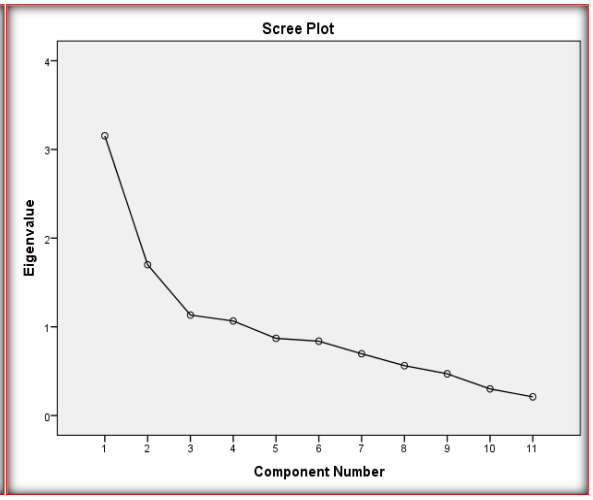


**Figure 5.40 Scree Plot in Monsoon Season**

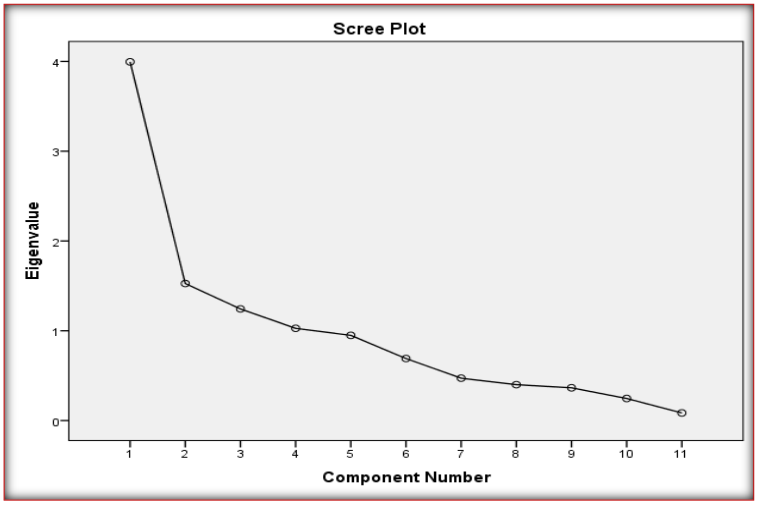


**Figure 5.41 Scree Plot in Winter Season**

**Table 5.11 Total Variance Explained**

| Component | Total (S) | Cumulative % of Var. (S) | Total (M) | Cumulative % of Var. (M) | Total(W) | Cumulative % of Var. (W) |
|---|---|---|---|---|---|---|
| 1 | 3.51 | 31.9 | 3.15 | 28.67 | 4 | 36.32 |
| 2 | 1.56 | 46.06 | 1.7 | 44.12 | 1.53 | 50.2 |
| 3 | 1.29 | 57.76 | 1.13 | 54.42 | 1.24 | 61.5 |
| 4 | 1.08 | 67.57 | 1.07 | 64.11 | 1.03 | 70.84 |
| 5 | 0.91 | 75.88 | 0.87 | 72.02 | 0.95 | 79.47 |
| 6 | 0.84 | 83.51 | 0.84 | 79.64 | 0.69 | 85.75 |
| 7 | 0.58 | 88.82 | 0.7 | 85.97 | 0.47 | 90.04 |
| 8 | 0.45 | 92.89 | 0.56 | 91.08 | 0.4 | 93.67 |
| 9 | 0.38 | 96.32 | 0.47 | 95.36 | 0.37 | 96.99 |
| 10 | 0.35 | 99.47 | 0.3 | 98.09 | 0.25 | 99.23 |
| 11 | 0.06 | 100 | 0.21 | 100 | 0.08 | 100 |

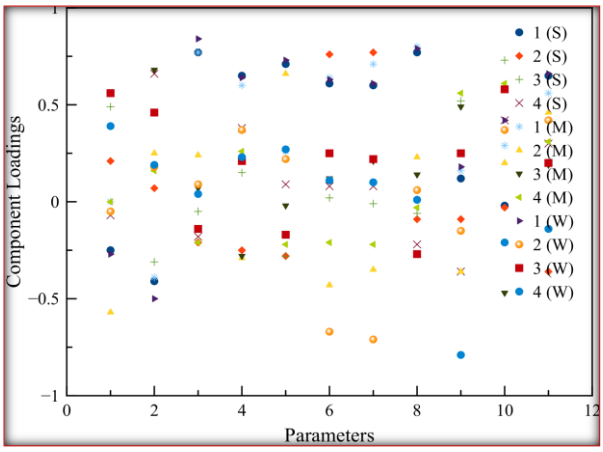


**Figure 5.42 Component Loading Factors in Three Seasons**

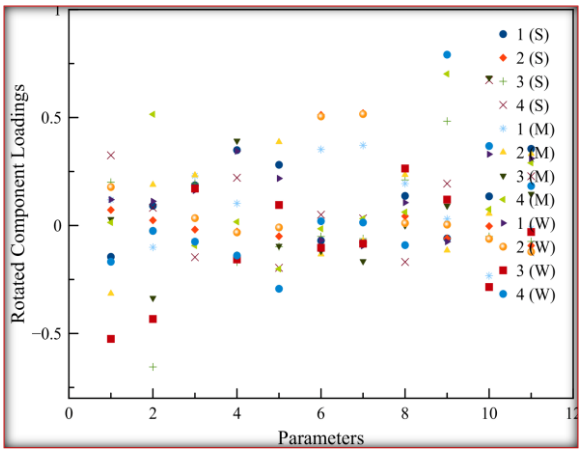


**Figure 5.43 Rotated Component Loading Factors in Three Seasons**

## 5.5.1 Determination of Principal Components for the Assessment of Water Quality

In the previous sections, interrelation of water quality parameters such as pH, DO, BOD, Conductivity, Nitrate-N, TC, FC, $NH_4$-N, COD, TA of $CaCO_3$ and TH of $CaCO_3$ have been established. As calculation of WQI considered an additive approach; the parameters considered in the study must be independent of each other for efficient forecasting of WQI of Brahmani River. Here, Principal component analysis with varimax rotation has been carried out on normalizes parameters data sets using SPSS 20.0.The number of PCs justified for the study in each season can be judged from scree plot shown in Figures 5.39, 5.40 and 5.41. It is observed that four principal components explaining 67.57% of total variation in summer, 64.11% in monsoon and 70.84% in winter are sufficient for

the study. The PCs in terms of actual parameters are given in equations (5.1) to (5.12). For the season of summer, the equations are:

$$PC1(S) = -0.145\times pH + 0.091\times DO + 0.186\times BOD + 0.349\times Cond. + 0.281\times Nitrate-N - 0.070\times TC - 0.079\times FC + 0.137\times COD - 0.060\times NH_4\text{-}N + 0.135\times TA \text{ as } CaCO_3 + 0.355\times TH \text{ as } CaCO_3 \quad (5.1)$$

$$PC2(S) = 0.072\times pH + 0.024\times DO - 0.019\times BOD - 0.028\times Cond. - 0.050\times Nitrate-N + 0.512\times TC + 0.520\times FC + 0.043\times COD - 0.060\times NH_4\text{-}N - 0.003\times TA \text{ as } CaCO_3 0.093\times TH \text{ as } CaCO_3 \quad (5.2)$$

$$PC3(S) = 0.200\times pH - 0.655\times DO + 0.193\times BOD - 0.169\times Cond. - 0.098\times Nitrate-N - 0.051\times TC - 0.061\times FC + 0.211\times COD + 0.483\times NH_4\text{-}N - 0.051\times TA \text{ as } CaCO_3 - 0.074\times TH \text{ as } CaCO_3 \quad (5.3)$$

$$PC4(S) = 0.323\times pH + 0.082\times DO - 0.147\times BOD + 0.221\times Cond. - 0.190\times Nitrate-N + 0.050\times TC + 0.033\times FC - 0.169\times COD + 0.194\times NH_4\text{-}N + 0.0673\times TA \text{ as } CaCO_3 + 0.228\times TH \text{ as } CaCO_3 \quad (5.4)$$

$$PC1(M) = 0.119\times pH - 0.101\times DO + 0.230\times BOD + 0.102\times Cond. - 0.012\times Nitrate-N + 0.352\times TC + 0.371\times FC + 0.194\times COD + 0.030\times NH_4\text{-}N - 0.233\times TA \text{ as } CaCO_3 + 0.228\times TH \text{ as } CaCO_3 \quad (5.5)$$

$$PC2(M) = -0.315\times pH + 0.189\times DO + 0.233\times BOD - 0.137\times Cond. + 0.387\times Nitrate-N - 0.133\times TC - 0.071\times FC + 0.235\times COD - 0.115\times NH_4\text{-}N + 0.055\times TA \text{ as } CaCO_3 + 0.014\times TH \text{ as } CaCO_3 \quad (5.6)$$

$$PC3(M) = -0.029\times pH - 0.336\times DO - 0.073\times BOD + 0.392\times Cond. - 0.095\times Nitrate-N - 0.118\times TC - 0.166\times FC + 0.0003\times COD + 0.091\times NH_4\text{-}N + 0.0684\times TA \text{ as } CaCO_3 + 0.146\times TH \text{ as } CaCO_3 \quad (5.7)$$

$$PC4(M) = 0.014\times pH + 0.515\times DO + 0.093\times BOD + 0.017\times Cond. - 0.201\times Nitrate-N + 0.015\times TC + 0.032\times FC + 0.062\times COD + 0.702\times NH_4\text{-}N + 0.075\times TA \text{ as } CaCO_3 + 0.290\times TH \text{ as } CaCO_3 \quad (5.8)$$

$$PC1(W) = 0.120\times pH + 0.112\times DO + 0.162\times BOD + 0.344\times Cond. + 0.218\times Nitrate-N - 0.068\times TC + 0.094\times FC + 0.106\times COD + 0.075\times NH_4\text{-}N + 0.330\times TA \text{ as } CaCO_3 + 0.310\times TH \text{ as } CaCO_3 \quad (5.9)$$

$$PC2(W) = 0.178\times pH - 0.025\times DO + 0.034\times BOD - 0.032\times Cond. - 0.009\times Nitrate-N + 0.505\times TC + 0.516\times FC + 0.011\times COD + 0.004\times NH_4\text{-}N - 0.062\times TA \text{ as } CaCO_3 - 0.122\times TH \text{ as } CaCO_3 \quad (5.10)$$

$$PC3(W) = -0.525\times pH - 0.433\times DO + 0.171\times BOD - 0.157\times Cond. + 0.095\times Nitrate-N - 0.103\times TC - 0.084\times FC + 0.264\times COD + 0.120\times NH_4\text{-}N - 0.285\times TA \text{ as } CaCO_3 - 0.030\times TH \text{ as } CaCO_3 \quad (5.11)$$

$$PC4(W) = -0.168\times pH + -0.025\times DO - 0.075\times BOD - 0.139\times Cond. - 0.293\times Nitrate-N + 0.020\times TC + 0.014\times FC - 0.091\times COD + 0.791\times NH_4\text{-}N + 0.368\times TA \text{ as } CaCO_3 + 0.183\times TH \text{ as } CaCO_3 \quad (5.12)$$

The extracted principal components from the above equations are graphically shown in Figures 5.44, 5.45 and 5.46 respectively for summer, monsoon and winter respectively.

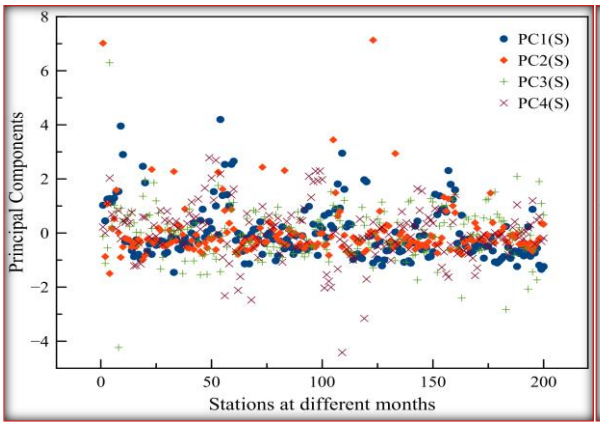


**Figure 5.44 Extracted Principal Components in summer season**

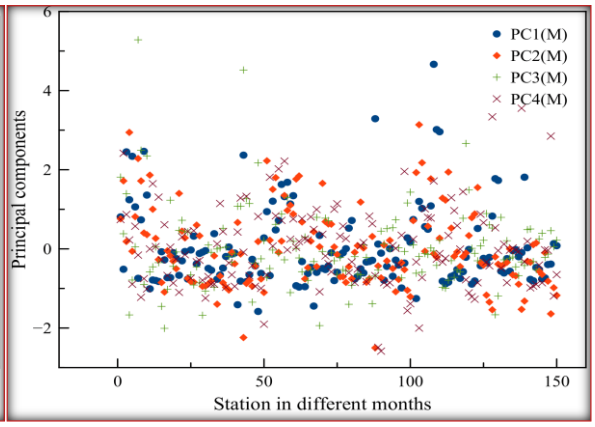


**Figure 5.45 Extracted Principal Components in monsoon season**

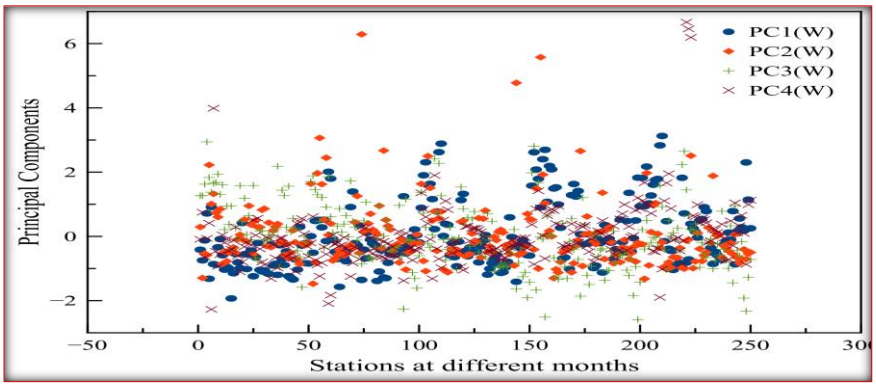


**Figure 5.46 Extracted Principal Components in winter season**

## 5.6 Canonical Correlation Analysis

The Eigen vectors are obtained by Table 5.11. So, for the Eigen values, Eigen vectors are calculated. Finally, the obtained Eigen vectors can be used for constructing the four principal components (PCs) in each season form input variables. The characteristics of PCs are presented in the Figures 5.44, 5.45 and 5.46. In Table 5.11, Eigen values of each season and cumulative variance proportion are given. Clearly, the components explain 67.57% of total variance in summer, 64.11% in monsoon and 70.84% in winter of the data sets respectively. Figure 5.43 represents the values of Eigen vectors, which are assessed the coefficients for formation of components. It should be noted that for retaining the PCs, a criterion equal $10^{-8}$ is used. Here, the correlation coefficients considered significant is one that greater than 0.60 (or > 60%). This conservative criterion is selected because of large study area and highly non-linear and dynamic Brahmani River system. The stations with rotated factor correlation coefficients less than this value are not considered as principal stations. The graphical representation of principal components in Figures 5.44, 5.45 and 5.46 indicats that some of stations at some particular months of the years from 2003 to 2012 have coefficient values less

than 0.60 for all the PCs. These stations are considered less important in explaining the annual variance of the river water quality, and here by could be the non-principal stations.

For validating above findings, in the first case, data from principal stations are compared. In this study, two cases are developed for comparisons. In first case, data from principal stations are used to formulate the relationships between parameters by regression. In the second case, data from all stations (principal and non-principal stations) are used to formulate the relationships between the parameters. These two cases are then compared to determine if the addition of data from non-principal stations improved the regression relationships. Comparison of the relationship between these parameters show that amount of $R^2$ increased in the case of principal monitoring stations. These increase in $R^2$ is different for each parameter as shown in Table 5.12.

**Table 5.12 Comparison of the relationship between the parameters in two cases (overall monitoring and principal monitoring stations)**

| | | BOD Cond. | BOD Nitrate-N | BOD COD | Cond. TH as CaCO3 | Nitrate-N COD | TC FC |
|---|---|---|---|---|---|---|---|
| $R^2$ (S) | All monitoring Stations | 0.42 | 0.52 | 0.61 | 0.54 | 0.51 | 0.93 |
| | Principal monitoring Stations | 0.51 | 0.57 | 0.69 | 0.58 | 0.55 | 0.94 |
| $R^2$ (M) | All monitoring Stations | 0.3 | 0.3 | 0.75 | 0.2 | 0.21 | 0.68 |
| | Principal monitoring Stations | 0.33 | 0.34 | 0.77 | 0.38 | 0.23 | 0.69 |
| $R^2$ (W) | All monitoring Stations | 0.5 | 0.36 | 0.54 | 0.32 | 0.28 | 0.87 |

In the present investigation, at first CCA is carried out on all complete sets of stations data. There are eight variables in the response data set i.e. chemical parameters including pH, DO, BOD, Nitrate-N, COD, $NH_4$-N, TA as $CaCO_3$ and TH as $CaCO_3$; the three variables in the predictor set i.e. physical and biological parameters including Conductivity, TC and FC. Table 5.13 represent the results of CCA for physical, chemical and biological variables. Correlation coefficient for canonical variates 1, 2 and 3 are given in Table 5.13 for summer, monsoon and winter respectively. Only the first canonical correlation is statistically significant ($p < 0.0001$). Although the second and third canonical correlations are large, they are not statistically significant by chi-square test. Therefore, there is no real evidence of any relationships between the physical, chemical and biological variables based on canonical variates 2 and 3.

The dominant variable in the first canonical variate for chemical parameters is TH as $CaCO_3$ and the dominant variables in the physical and biological parameters is conductivity for summer. For monsoon, the chemical variable is COD and physical-biological variable is conductivity. Likewise in winter, BOD and conductivity are the variable for chemical and

physical-biological parameters. The second canonical variate has high correlations of the response and predictor sets. In this canonical variate the predictor variables is TH as CaCO3; the response variables TC and FC for summer. Likewise in monsoon, the predictor variables are DO and BOD; response variable are conductivity and FC. For winter, the predictor variables for second canonical variate are DO, COD and TH as CaCO3; response variables are conductivity and FC. From the third canonical correlation, COD for chemical variables; TC and FC for physical-biological variable in the season of summer have high correlation. Nitrate-N and TA as caCO3 as chemical variables while TC and FC for physical-biological variable in the season of monsoon. DO, COD and TA as CaCO3 for chemical variables & TC and FC for physical-biological variables in winter. Considering the mentioned results, a regular pattern can be shown. TH as CaCO3 and DO are two dominant chemical parameters in all canonical variates. On the other hand, FC and conductivity are highly scored from physical-biological parameters. Verifying the ability of CCA, simple correlation factor between all parameters are given in Table 5.13. The correlation matrix reveals that a relationship existed between all physical, chemical and biological parameters.

**Table 5.13 Correlation Factors of all Parameters (Chemical, Physical and Biological)**

| Canonical variates | | 1(S) | 2(S) | 3(S) |
|---|---|---|---|---|
| Canonical Correlation | | 0.64 | 0.29 | 0.24 |
| Chi-square | | 128.45 | 27.82 | 11.52 |
| Degree of freedom | | 24 | 14 | 6 |
| Significant level | | <0.000 | <0.015 | <0.074 |
| Chemical parameters | pH | -0.14 | 0.29 | -0.04 |
| | DO | -0.13 | 0.31 | -0.42 |
| | BOD | -0.26 | 0.06 | -0.43 |
| | Nitrate-N | -0.35 | 0 | -0.33 |
| | COD | -0.19 | -0.03 | 1.08 |
| | NH4-N | -0.03 | -0.19 | 0 |
| | TA as CaCO3 | -0.1 | -0.07 | -0.1 |
| | TH as CaCO3 | **-0.53** | -0.71 | -0.27 |
| Physical and | Cond. | **-0.92** | -0.15 | -0.43 |
| Biological parameters | TC | -0.03 | -2.28 | 1.65 |
| | FC | -0.21 | 2.7 | -0.72 |
| **Canonical variates** | | **1(M)** | **2(M)** | **3(M)** |
| Canonical Correlation | | 0.59 | 0.32 | 0.09 |
| Chi-square | | 77.87 | 16.77 | 1.07 |
| Degree of freedom | | 24 | 14 | 6 |
| Significant level | | <0.000 | <0.269 | <0.983 |
| Chemical parameters | pH | -0.17 | 0.16 | 0.11 |
| | DO | 0.44 | 0.7 | -0.24 |
| | BOD | -0.43 | 0.76 | -0.4 |
| | Nitrate-N | 0.28 | -0.2 | -0.53 |
| | COD | -0.25 | -0.02 | 0.13 |
| | NH4-N | -0.22 | -0.29 | 0.18 |
| | TA as CaCO3 | -0.07 | -0.4 | -0.76 |
| | TH as CaCO3 | -0.27 | 0.09 | 0.37 |
| Physical and | Cond. | -0.63 | -0.77 | -0.42 |
| Biological parameters | TC | -0.23 | -0.3 | 1.32 |
| | FC | -0.4 | 1.11 | -0.74 |
| **Canonical variates** | | **1(W)** | **2(W)** | **3(W)** |
| Canonical Correlation | | 0.71 | 0.27 | 0.16 |
| Chi-square | | 198.06 | 24.49 | 6.14 |
| Degree of freedom | | 24 | 14 | 6 |
| Significant level | | <0.000 | <0.040 | <0.407 |
| Chemical parameters | pH | 0.22 | -0.21 | 0.32 |
| | DO | 0 | -0.57 | -0.68 |
| | BOD | 0.61 | 0.02 | 0.24 |
| | Nitrate-N | 0.35 | -0.4 | 0.09 |
| | COD | -0.02 | 0.62 | -0.77 |
| | NH4-N | 0.04 | 0.24 | -0.01 |
| | TA as CaCO3 | 0.13 | 0.08 | -0.6 |
| | TH as CaCO3 | 0.2 | -0.57 | -0.4 |
| Physical and | Cond. | 0.86 | -0.52 | 0.18 |
| Biological Parameters | TC | 0.15 | 0.06 | -2.25 |
| | FC | 0.22 | 0.89 | 2.26 |

## 5.7 Cluster Analysis

The hierarchical CA is involved on standardized log-transformed data sets sorted by season. CA generates a dendrogram for temporal similarity and period grouping as shown in Figure 5.47, 5.48 and 5.49.

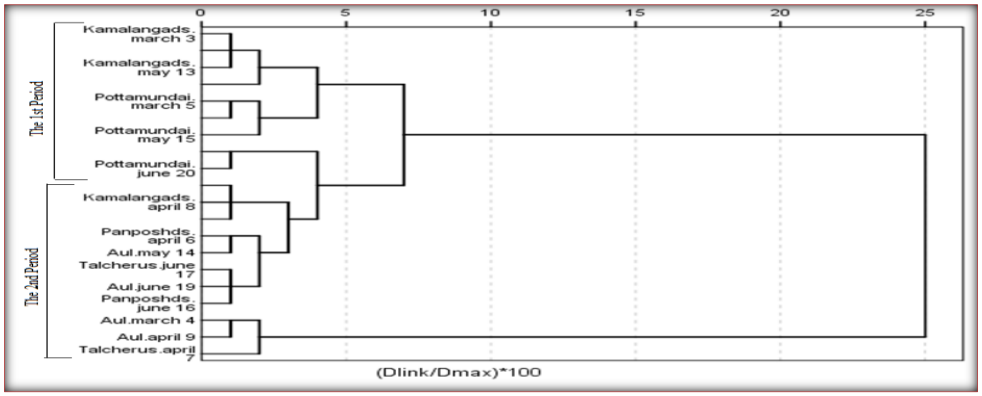


**Figure 5.47 Dendrogram for summer season**

The generate CA as a dendrogram, grouping the 12 months from 2003 to 2012 of five gauging stations into two clusters at $(D_{link}/ D_{max})\times 100 < 25$, and difference between the clusters are significant. For the season of summer, cluster I (the first period) includes the Kamalanga d/s and Pottamundai stations. Cluster II (the second period) includes the remaining stations.

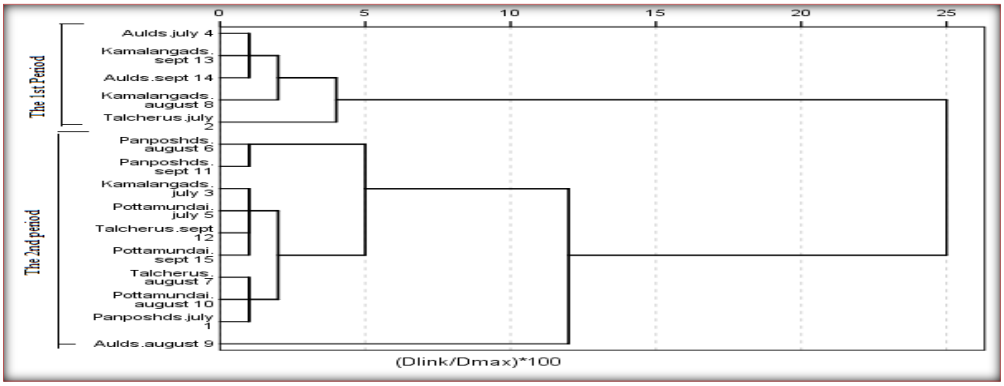


**Figure 5.48 Dendrogram for monsoon season**

Among the sampling sites, Cluster I shows the high pollution than Cluster II. In season monsoon, Cluster I (the 1st period) includes Aul, kamalanga d/s and Talcher u/s, where as Cluster II (the 2nd period) includes Kamalanga d/s, panposh d/s, Pottamunadi and Aul. Cluster II concludes high level of pollution than that of Cluster I. For the season of winter, Cluster I (the 1st period) include Talcher u/s, Aul, Pottamundai and Kamalanga d/s at different months from 2003 to 2012. Here, Cluster I concludes high pollution than Cluster II. From the

variation of parameter; in summer, Cluster II impacted more than Cluster I as shown in Figure 5.50.

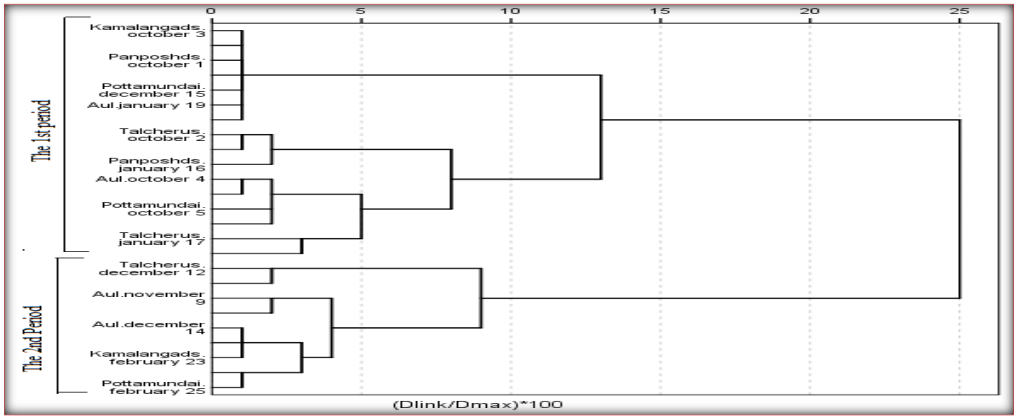


**Figure 5.49 Dendrogram for winter season**

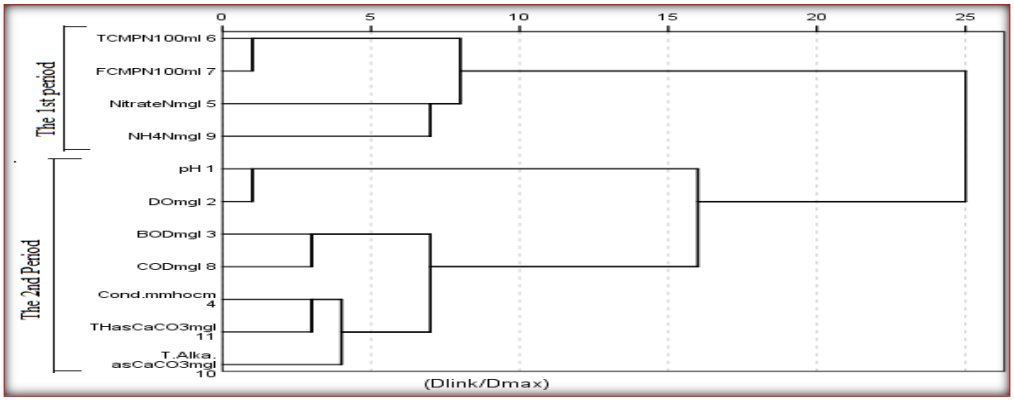


**Figure 5.50 Dendrogram in various parameters in summer season**

Likewise, in monsoon Cluster II has more impact in level of pollution than by Cluster I as shown in Figure 5.51. For the season of winter, Cluster II (the 2$^{nd}$ period) has more impact in level of pollution than by Cluster I (the 1$^{st}$ period) as shown by Figure 5.52.

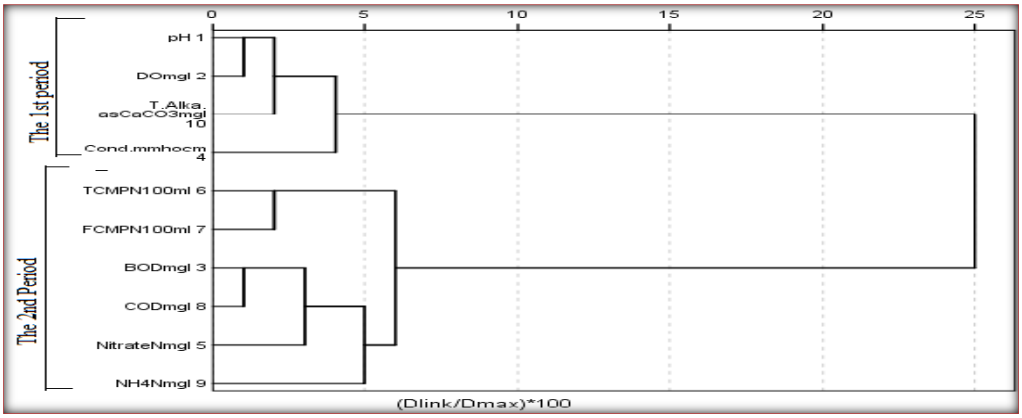


**Figure 5.51 Dendrogram in various parameters in monsoon season**

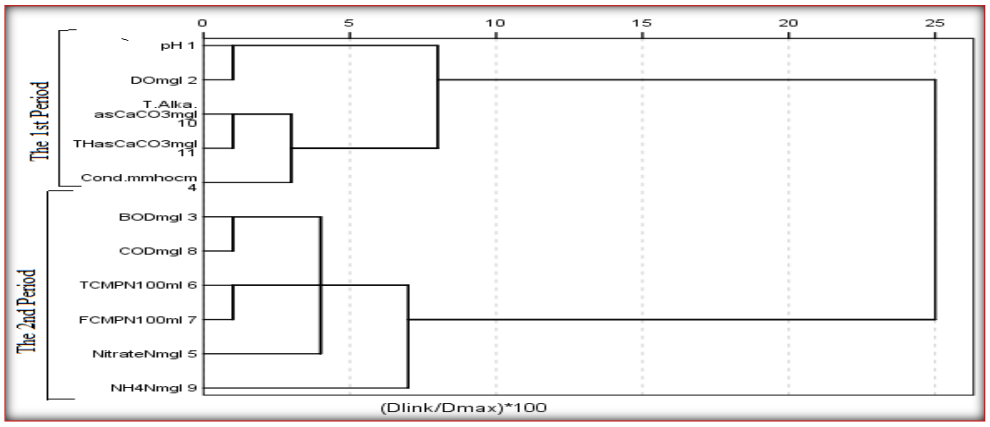


**Figure 5.52 Dendrogram in various parameters in winter season**

## 5.8 Adaptive Neuro-Fuzzy Inference System (ANFIS) by MATLAB

ANFIS predicted WQI of River Brahmani by adopting architecture of ANFIS network for creating a set of fuzzy IF-THEN rules and fuzzy inference system with the membership function to obtain the result. The building of fuzzy logic systems are initiated with the derivation of a set of IF-THEN fuzzy rules bearing the expertise and knowledge of modelling field (Dezfoli 2003). An appropriate rule is required for modelling. Hence a tool or predefined method and statistical analysis are done to achieve fuzzy logic rules. Fuzzy conditional statements are expressed as 'if DO is small then WQI is high' where these parameters are levels of fuzzy sets; those are characterised by membership functions. A Neuro-fuzzy rule plays an important role in human ability to make decisions. Hence, fuzzy IF-Then rules are used to make decisions in uncertainty analysis.

The data set from January 2003 to December 2012 of five selected gauging stations is divided into three sets that are of summer, monsoon and winter season data sets. The 200 data sets in the season of summer are used for training and testing in ANFIS model. The pattern of variation and distribution of actual and predicted WQI for training and testing data of River Brahmani are shown in Fig.5.53 (a) and (b) respectively. Likewise, for the season of monsoon and winter, 150 and 250 data sets are used for training and testing respectively. The pattern of variation and distribution of actual and predicted WQI for training and testing are shown in Figure 5.54 (a) and (b) for monsoon and Figure 5.55 (a) and (b) for winter season respectively. The plot of training and testing data along with FIS output show the coherence nature of data in the distribution.

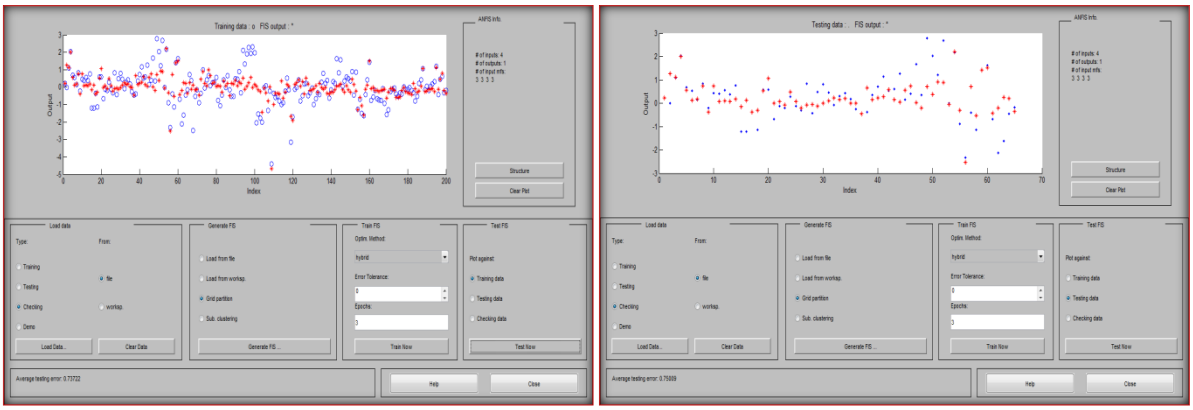

**Figure 5.53 (a) and (b) Distribution of actual and Predicted WQI for summer season**

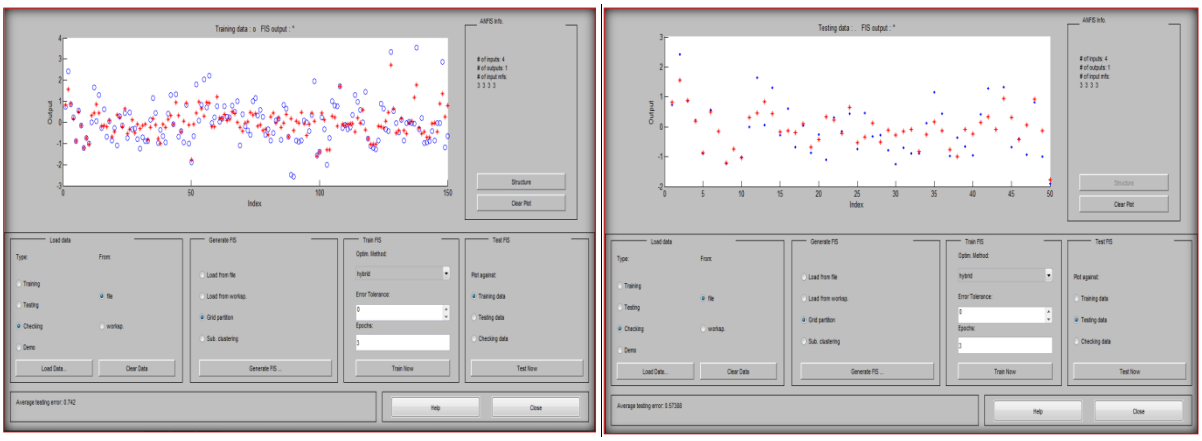

**Figure 5.54 (a) and (b) Distribution of actual and Predicted WQI for monsoon season**

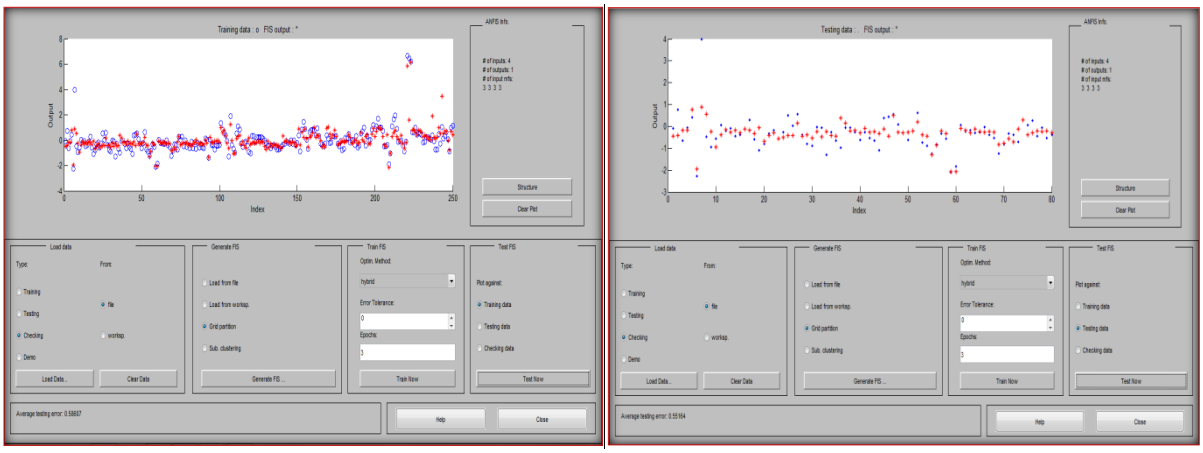

**Figure 5.55 (a) and (b) Distribution of actual and Predicted WQI for winter season**

Here, actual output is indicated by the blue dots and red dots presented predicted data of WQI. The surface plots for these data sets are shown in Figure 5.56, Figure 5.57 and Figure 5.58. It is shown that the surface covered the total landscape and decision space. A complete set of rule is generated by the Rule Editor in ANFIS GUI Editor for prediction of entrance length as shown in Figure 5.59, Figure 5.60 and Figure 5.61 respectively for summer,

monsoon and winter season. For assessing the model adequacy for prediction, the residual analysis is carried out by calculating the residuals from the actual and predicted WQI for training and testing data sets. It is observed that the residuals are distributed evenly along the centre line. Therefore, it can be concluded that the data are well trained and tested. It also incorporated the variation of patterns of actual and predicted WQI and showed non-linearity existed between WQI with quality parameters of Brahmani River.

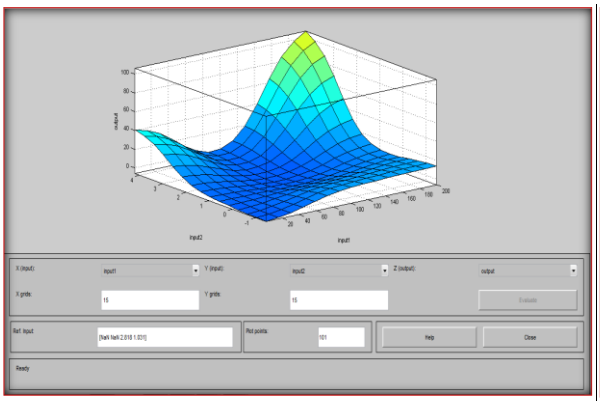

**Figure 5.56 The Surface Plot of WQI in summer season**

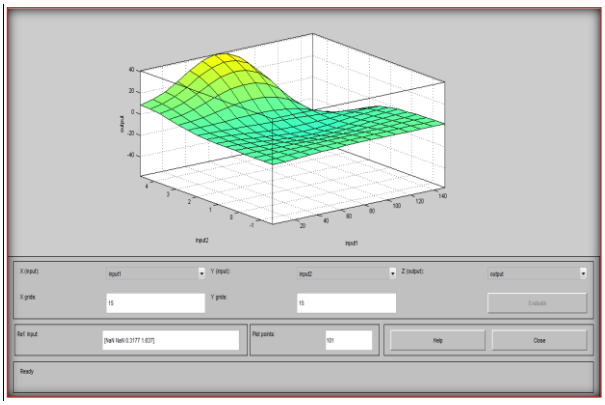

**Figure 5.57 The Surface of WQI Plot in monsoon season**

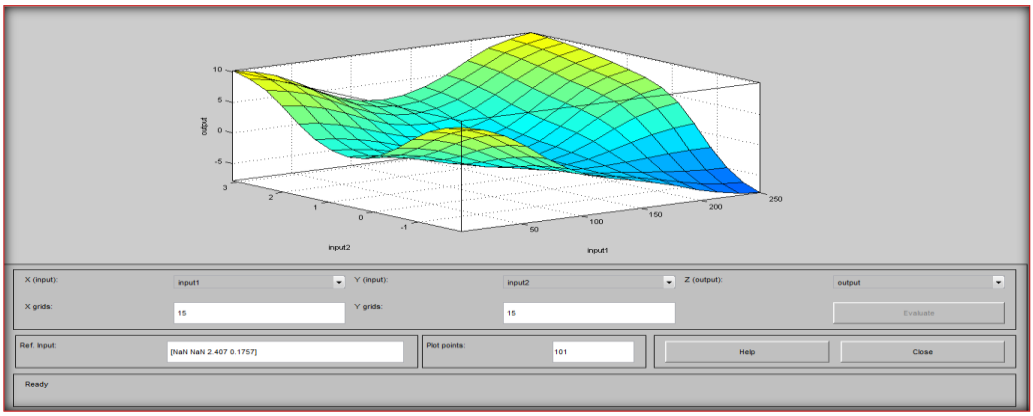


**Figure 5.58 The Surface Plot of WQI in winter season**

The regression relationship between actual WQI and predicted WQI via ANFIS model for training and testing data are plotted for the season of summer in Figure 5.62 (a) and (b) respectively. Similarly, for season of monsoon and winter, the plots for training and testing are shown in Figures 5.63 (a) & (b) and 5.64 (a) & (b) respectively. Degrees of coefficient of determination ($R^2$), for the season of summer are 0.994 and 0.995 for training and for testing respectively. For the season of monsoon, the coefficients are 0.985 and 0.990 respectively for training and testing. Likewise, for winter the coefficients were 0.992 and 0.993 respectively

for training and testing of data set. From, this higher degree of coefficient of determination, we came to know that the data are well fitted.

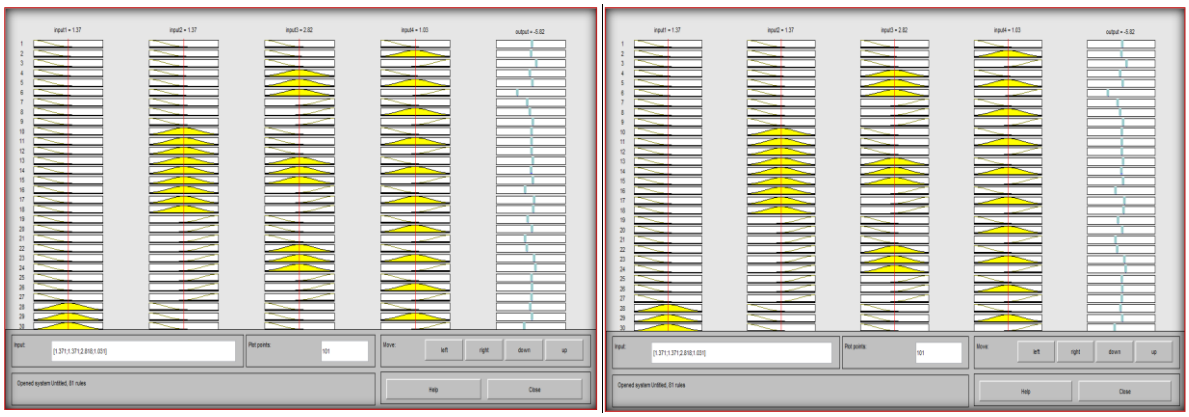

**Figure 5.59 and Figure 5.60 Sample set of rules by rule viewer for prediction of WQI entrance length for summer and monsoon respectively**

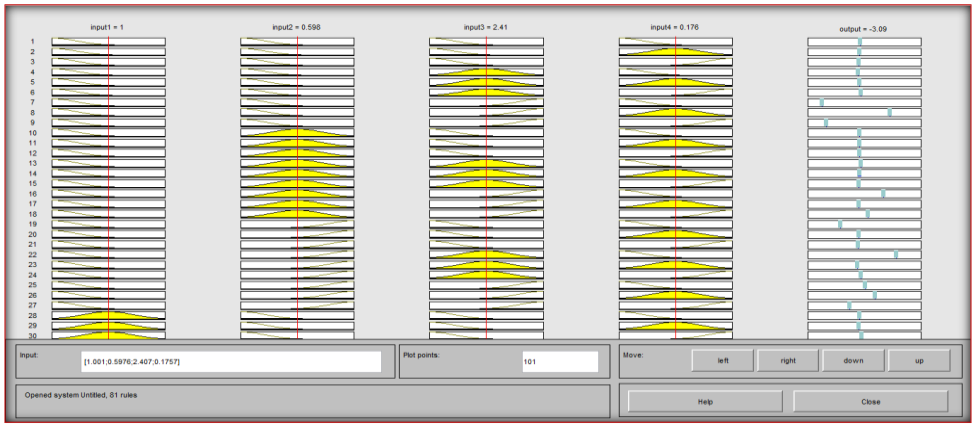


**Figure 5.61 A sample set of rules by rule viewer for prediction of WQI entrance length for winter**

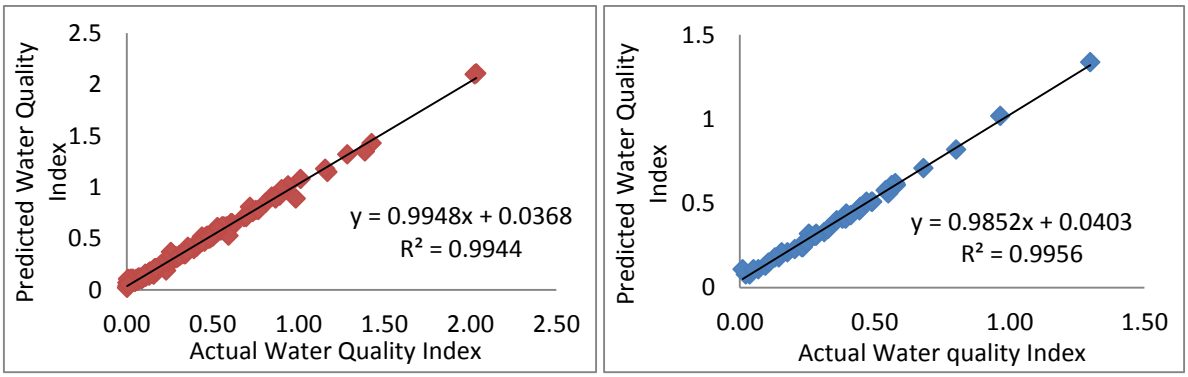


**Figure 5.62 (a) and (b) Correlation of Predicted and Actual WQI of Training and Testing data respectively in summer season by ANFIS**

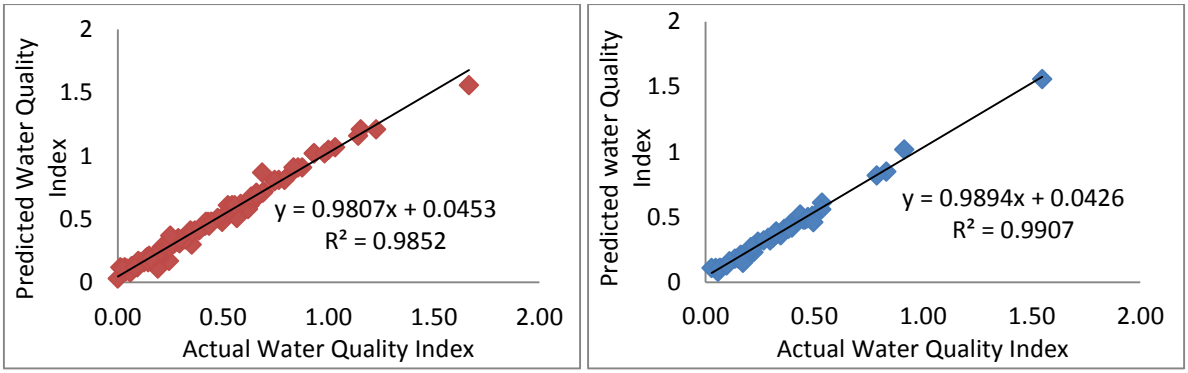


**Figure 5.63 (a) and (b) Correlation of Predicted and Actual WQI of Training and Testing data respectively in monsoon season by ANFIS**

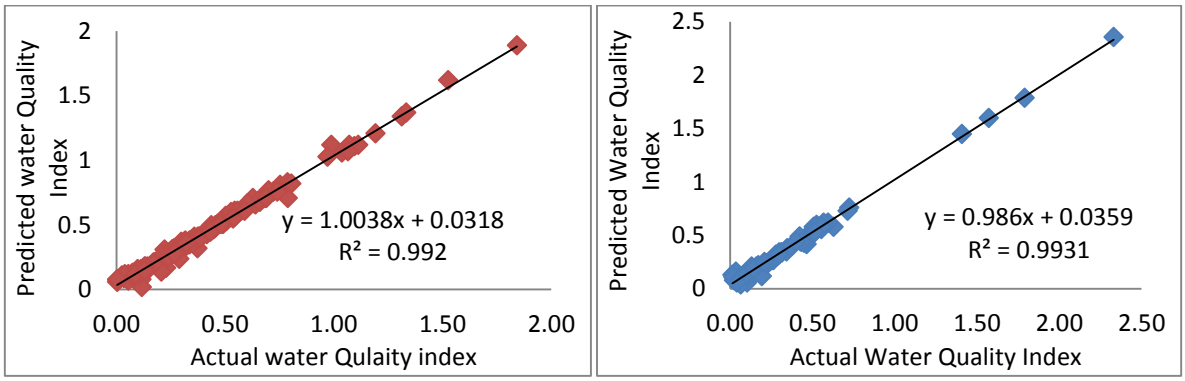


**Figure 5.64 (a) and (b) Correlation of Predicted and Actual WQI of Training and Testing data respectively in winter season by ANFIS**

The model performance is checked with different principal components as inputs. Two, three and four number of principal components is used as input accordingly to know the variation in the data. The variation in two principal components is used initially for every season. Then three and four principal components of total variation are used as input respectively to the ANFIS model. The error calculations for the principal components used as input as well as for the training and testing data are done to know the accuracy. The errors affected the normalisation and prediction of WQI by ANFIS model. An adaptive Neuro-Fuzzy Inference System of fuzzy logic model based on Mamdani system using fuzzy logic toolbox along with Gaussian type membership function is considered. These types of fuzzy model needed more than 1, 00,000 of rules to develop based on expert's advice and human knowledge for prediction and inferences.

## 5.9 Artificial Neural Network (ANN) By MATLAB

The data in neural networks are categorised into two sets; training or learning sets, and test or over fitting test sets. The learning set is used to determine the adjusted weights and biases of

a network. The test set is used for calibration, which prevents over training networks. The over fitting test set consists of a representative data set. It is important to divide the data set in such a way that both training and over fitting test data sets are statistically comparable. The 70% of data are used for validation and training i.e. each of 35% and 30% of the data are used for testing. The prediction by ANN determines the input vector to the network by two algorithms; those are back propagation network (BPN) and radial basis function (RBF). The input normalised data of water quality in the network algorithm are continued with adding one more data in the network input layer and the performance of input layer is examined on the basis of statistical indices i.e. goodness of fit statistics. As the number of hidden layer in network algorithm of the model is increased or decreased, there is a change in goodness of fit statistics. So it is clear that, the goodness of fit statistics had been done for training, validation and testing of time series data with fixing the hidden layers to 10 during modelling process. The target values are iterated by the model after the target time steps given. Simultaneously, the autocorrelation, input-output correlation and error correlation are calculated by the model. Root mean squared error (RMSE) is the error between output and given target values by the model. Here, R is the measure of regression between output and target data. Before iteration of the input layer by network algorithm, the regression coefficient (R) and RMSE of the given input data as shown in Figures 5.65, 5.66 and 5.67 for summer, monsoon and winter respectively.

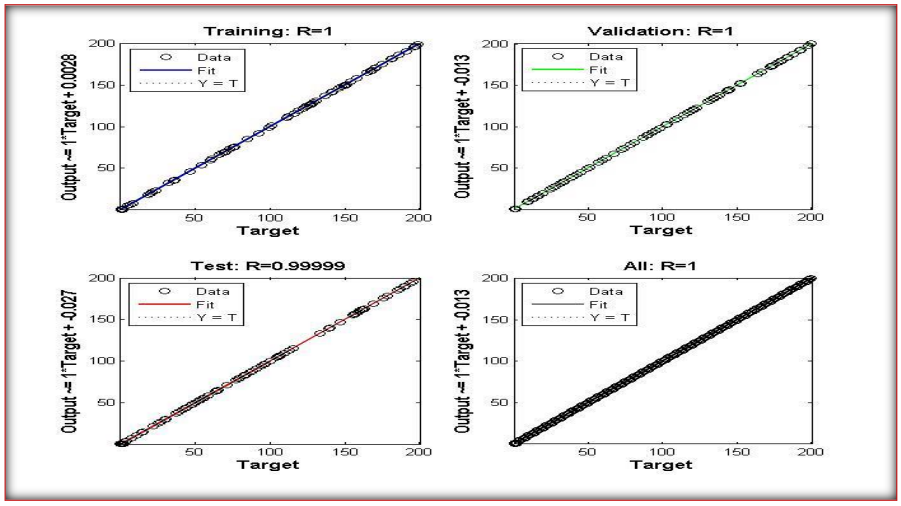


**Figure 5.65 Regression Output on ANN results for summer season**

The values of R and RMSE are given in Table 5.14. The figures shown below for the regression output of the three seasons with correlation coefficient (R) nearly equal to 1 can conclude for further analysis of prediction by ANN model.

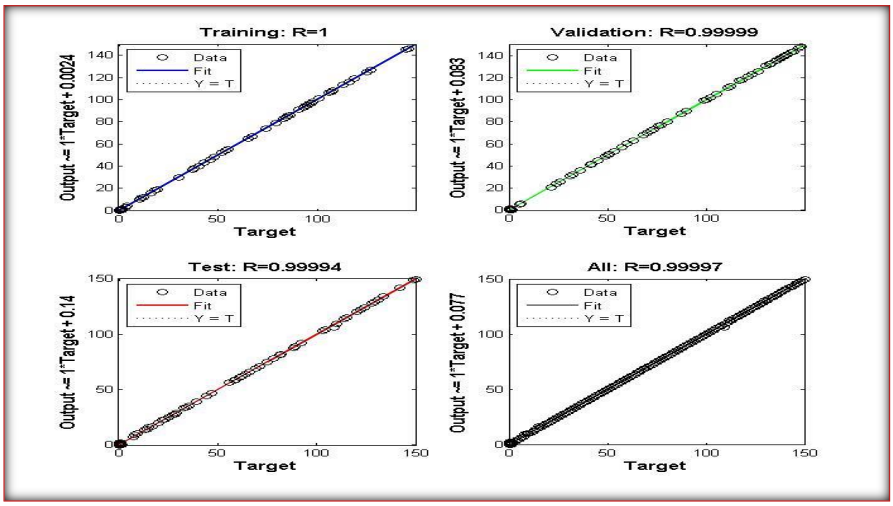


**Figure 5.66 Regression Output on ANN results for monsoon season**

The response output curves after the iterations in ANN are saved along with the predicted values as the output by ANN. The response output curves along with the error curve during response curve is generated and were shown in Figures 5.68, 5.69 and 5.70 respectively for summer, monsoon and winter, which shows the training , validation and testing targets as well as outputs.

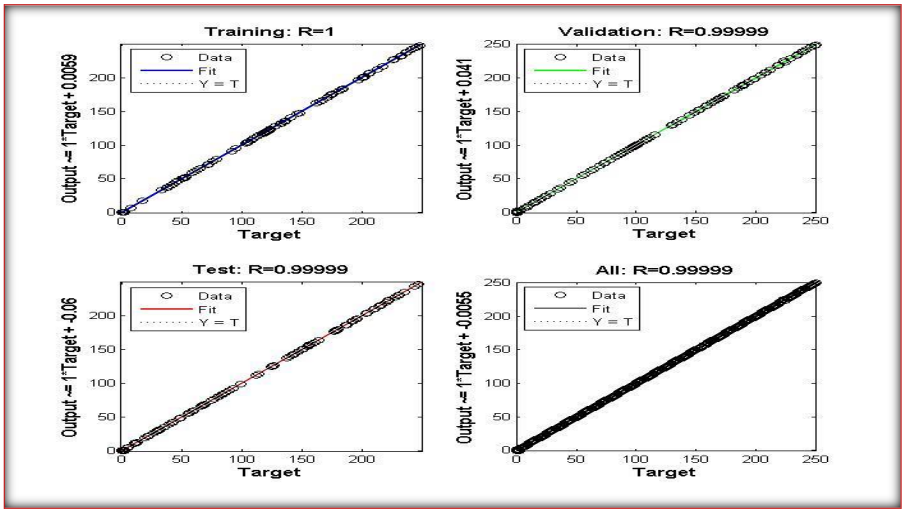


**Figure 5.67 Regression Output on ANN results for winter season**

The predicted outputs by ANN simulation are saved and the correlations are found out between the actual WQI and predicted WQI by ANN. The results are shown by the Figures 5.71, 5.72 and 5.73 for summer, monsoon and winter respectively. The degree of coefficient of determination or correlation coefficient between actual and predicted WQI by ANN in summer, winter and monsoon are 0.945, 0.941 and 0.965 respectively.

**Table 5.14 Results of the Goodness of fit statistics in three seasons**

| Data Set | Summer | Monsoon | Winter |
|---|---|---|---|
| Target Time Steps | | | |
| Validation | 35% | 35% | 35% |
| Training | 30% | 30% | 30% |
| Testing | 35% | 35% | 35% |
| Target Values | | | |
| Validation | 140 | 105 | 175 |
| Training | 120 | 90 | 150 |
| Testing | 140 | 105 | 175 |
| R (Correlation Coefficient) | | | |
| Validation | 1.00E+00 | 1.00E+00 | 1.00E+00 |
| Training | 1.00E+00 | 1.00E+00 | 1.00E+00 |
| Testing | 1.00E+00 | 1.00E+00 | 1.00E+00 |
| MSE | | | |
| Validation | 3.76E-01 | 1.05E-01 | 4.07E-02 |
| Training | 2.62E-02 | 8.33E-03 | 3.12E-01 |
| Testing | 5.87E-02 | 5.14E-01 | 1.56E-01 |

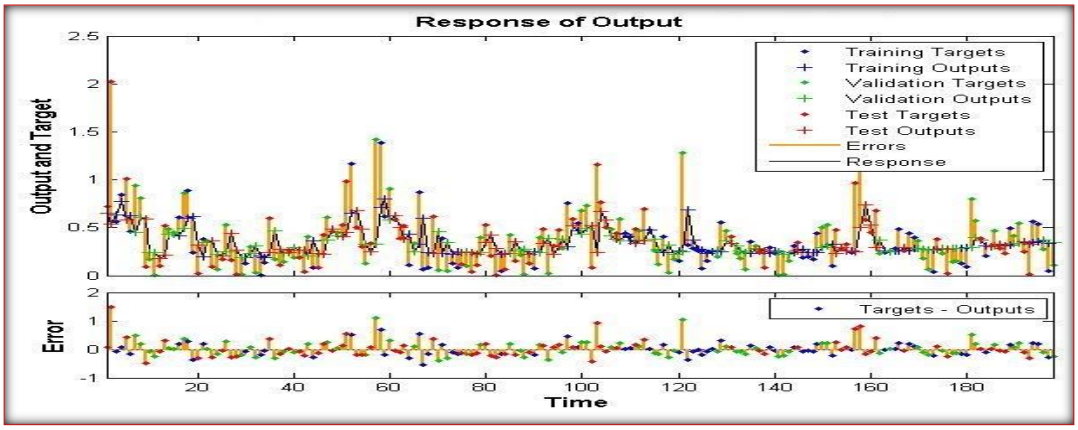


**Figure 5.68 Response Output Curve along with Error for summer season**

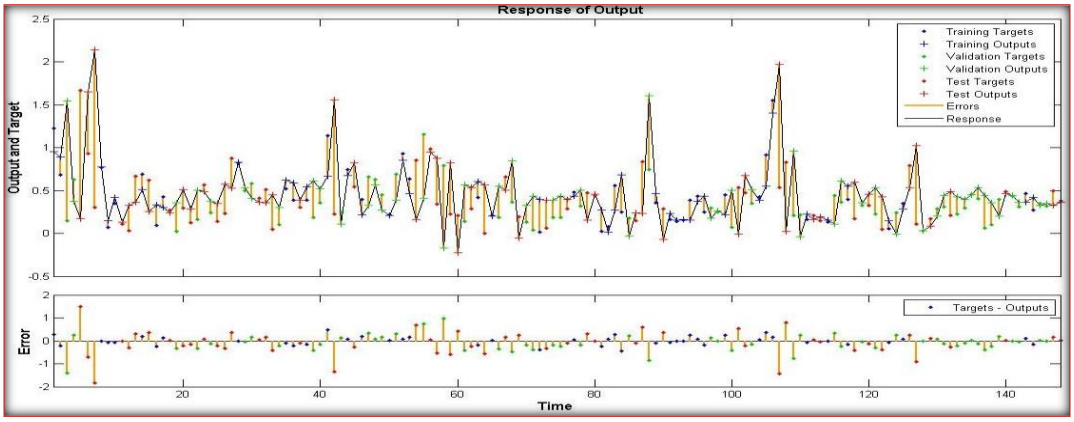


**Figure 5.69 Response Output Curve along with Error for monsoon season**

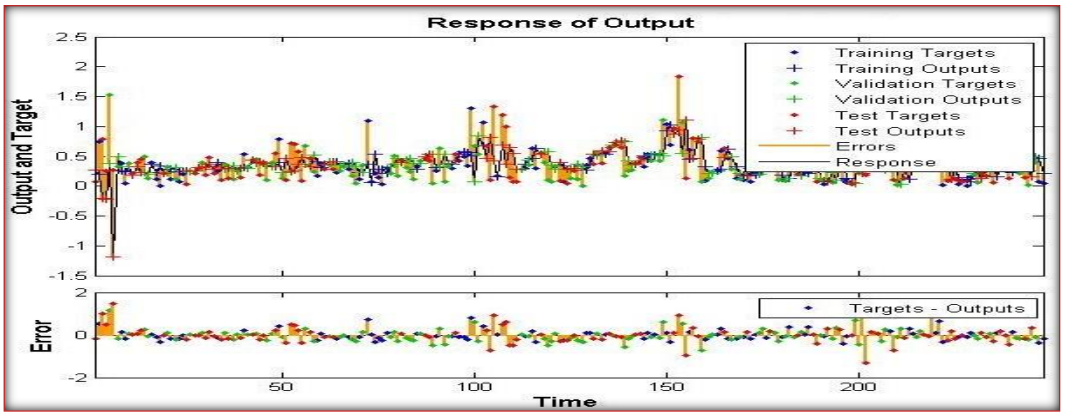


**Figure 5.70 Response Output Curve along with Error for winter season**

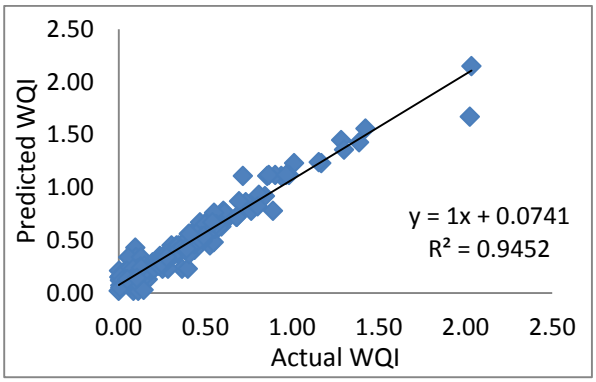


**Figure 5.71 Correlation of Actual and ANN Predicted WQI for summer season**

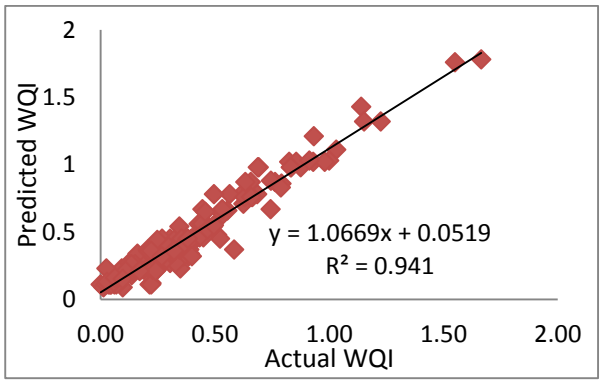


**Figure 5.72 Correlation of Actual and ANN Predicted WQI for monsoon season**

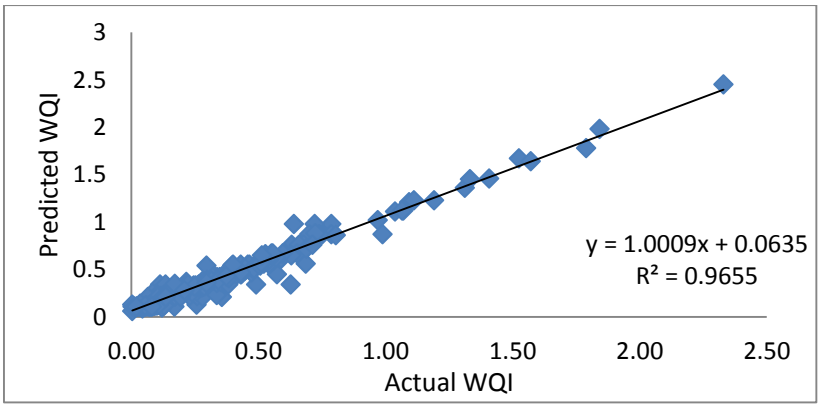


**Figure 5.73 Correlation of Actual and ANN Predicted WQI for winter season**

## 5.10 Monte Carlo Simulation (MCS)

As described earlier Monte Carlo Simulation is applied as a probabilistic method to solve deterministic problems. MCS can simulate a large number of experimental trials that have random outcomes. So, the statistical outcomes of the simulations of the parameters of three seasons are shown in Table 5.15.

**Table 5.15 Statistical Outcomes by MCS in respective seasons**

| season | Mean | Std.Dev | Variance | Skewness | Kurtosis | Mode | Min | Max | Range |
|---|---|---|---|---|---|---|---|---|---|
| Summer | 28.47 | 0.999 | 0.99 | -0.0005 | -0.008 | 28.48 | 24.6 | 32.23 | 7.63 |
| Monsoon | 17.12 | 0.999 | 0.99 | -0.0008 | -0.006 | 17.11 | 13.25 | 20.84 | 7.58 |
| Winter | 18.74 | 0.999 | 0.99 | -0.0004 | -0.046 | 18.84 | 15.46 | 21.84 | 6.38 |

In MCS based risk assessment, discrete values of input variables or model parameters are generated in a series of consistent with their probability distributions, and the water quality model is calculated for each generated input data and also produced outputs in the form of statistical distribution. The advanced statistics results of three consecutive stations are recorded and were shown in Table 5.16. The simulation results after the risk assessments are recorded as shown in Figure 5.74, 5.75 and 5.76 for summer, monsoon and winter respectively, the predicted results by the simulation are used further for correlation analysis as shown in Figure 5.77, 5.78 and 5.79 for three respective seasons.

**Table 5.16 Advanced Statistics by MCS for respective seasons**

| | Summer | Monsoon | Winter |
|---|---|---|---|
| Lower cutoff | 27.65 | 16.3 | 17.92 |
| Likelihood | 59% | 59% | 59% |
| Upper cutoff | 29.3 | 17.94 | 19.57 |
| Mean Abs. Dev | 0.79 | 0.79 | 0.79 |
| semi Variance | 0.5 | 0.5 | 0.49 |
| Semi deviation | 0.7 | 0.7 | 0.7 |
| Value at risk 95% | 30.12 | 18.76 | 20.39 |
| Cond. Value at risk 95% | 28.36 | 17.01 | 18.64 |
| mean confidence 95% | 0.01 | 0.02 | 0.06 |
| Std. Dev Confidence 95% | 0.01 | 0.01 | 0.04 |
| Coefficient of variation | 0.04 | 0.06 | 0.05 |
| Standard Error | 0.009 | 0.009 | 0.03 |
| Expected loss | 0 | 0 | 0 |
| Expected loss ratio | 0% | 0% | 0% |
| Expected gain | 28.47 | 17.12 | 18.74 |
| Expected gain ratio | 100% | 100% | 100% |

The graphical representations of simulated results along with the lower cut off, likelihood and upper cut off were shown below.

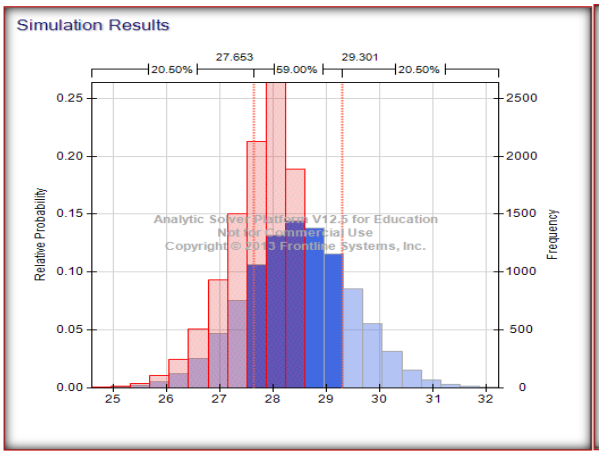


**Figure 5.74 Simulation results for Summer season**

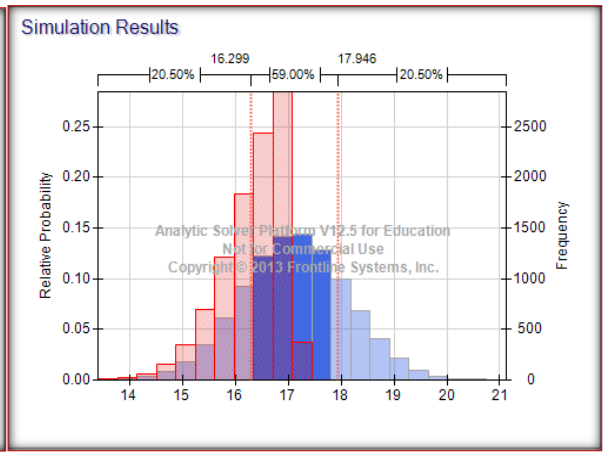


**Figure 5.75 Simulation results for Monsoon season**

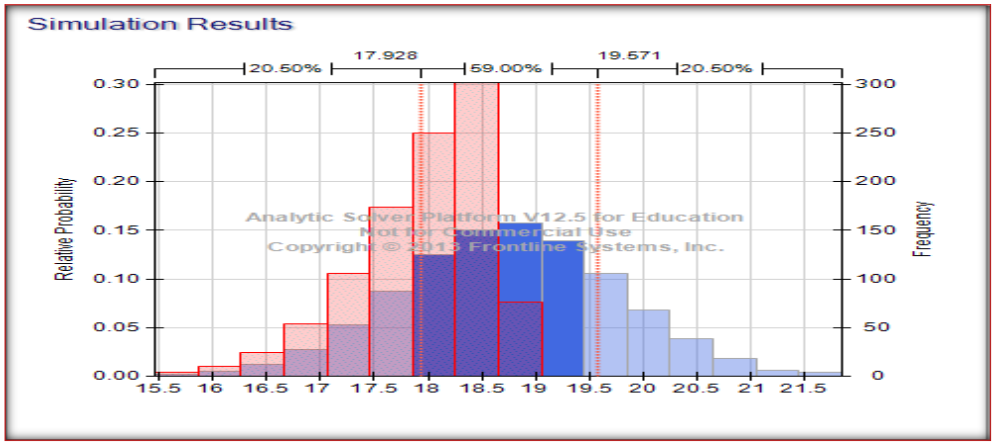


**Figure 5.76 Simulation results for winter season**

The predicted results after 1000 simulations in MCS were shown graphically along with the correlation coefficients between actual and predicted WQI. The correlation coefficients were 0.929, 0.953 and 0.970 in summer, winter and monsoon respectively.

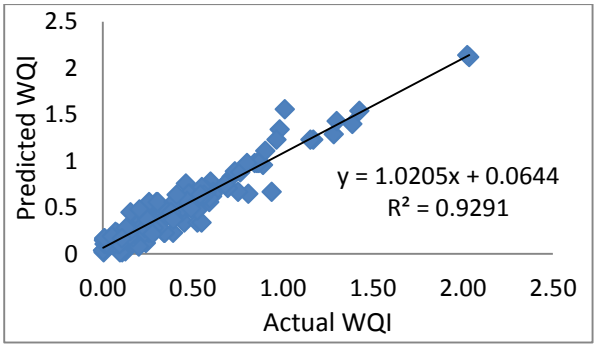


**Figure 5.77 Correlation of Actual and MCS Predicted WQI for summer season**

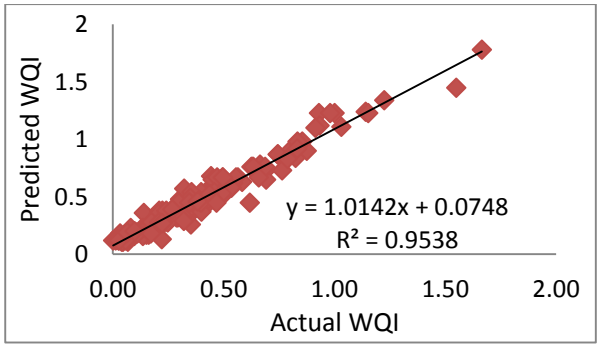


**Figure 5.78 Correlation of Actual and MCS Predicted WQI in monsoon season**

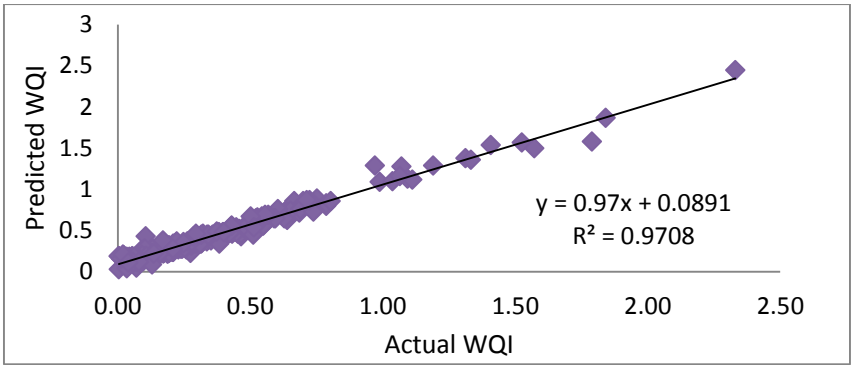


**Figure 5.79 Correlation of Actual and MCS Predicted WQI for winter season**

## 5.11 Performance Evaluation of Models

### 5.11.1 Adaptive Neuro-Fuzzy Inference System (ANFIS)

The mean absolute percentage error (MAPE) for training data and testing data are calculated in three seasons. The calculated MAPE is compared with the MAPE of same set of training and testing data used to forecast WQI without transforming into principal components by SPSS. The differences in percentage of improve were also calculated as shown in Table 5.17. The transformation of data sets into principal components reduced not also dimensionality problem but also improved the computational capability. So, up to a certain extent elimination of correlation between predictions of WQI in River Brahmani was possible by ANFIS model.

**Table 5.17 Mean Absolute Percentage Error calculation**

| | Summer | Monsoon | Winter |
|---|---|---|---|
| Without transforming into PCS (Training data) | 12.23 | 12.76 | 12.45 |
| Without transforming into PCS (Testing data) | 13.65 | 16.43 | 13.78 |
| Transforming into PCS (Training data) | 2.13 | 6.25 | 2.04 |
| Transforming into PCS (Testing data) | 1.99 | 0.8 | 1.78 |
| % of improvement (Training data) | 10.1 | 6.51 | 10.41 |
| % of improvement (Testing data) | 11.66 | 15.63 | 12 |

The model performance is checked with different principal components as inputs. Two, three and four number of principal components is used as input accordingly to know the variation in the data. The variations in two principal components are 46.06%, 44.12% and 50.20% in summer, monsoon and winter used initially. Then three principal components are having 57.76%, 54.42% and 61.50% for three respective months; where as 67.57%, 64.11% and 70.84% for four principal components in three seasons of total variation are used as input respectively to the ANFIS model. It is also observed that mean absolute percentage error (MAPE) for training data when two, three and four principal components are used as input parameters to ANFIS model are given in Table 5.18. It is found that as the input parameters

decreased the mean absolute percentage error is increased due to loss of information. It affected the normalisation and prediction of WQI by ANFIS model. Therefore, the ANFIS model performed well when four principal components explaining higher percentage of total variation are used as input components to ANFIS GUI Editor.

**Table 5.18 Mean Percentage Absolute Error of Training Data**

| | Summer | Monsoon | Winter |
|---|---|---|---|
| Two PCs | 0.98 | 0.93 | 0.94 |
| Three PCs | 0.67 | 0.72 | 0.63 |
| Four PCs | 0.32 | 0.36 | 0.31 |

## 5.11.2 Artificial neural Network (ANN)

Since the water quality parameters are collected from January 2003 to December 2012, the performance of the proposed ANN based architecture can be examined and evaluated. The performances of the models are evaluated using the gradient and μ values after the complete epochs by ANN. The self explaining graphs of the performance evaluation for summer, monsoon and winter are shown in Figure 5.80, 5.81 and 5.82 respectively.

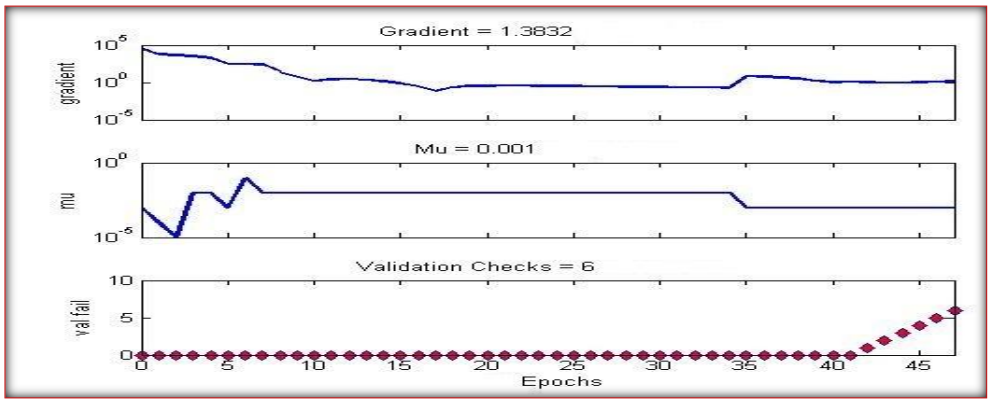


**Figure 5.80 Performance Evaluation Curve for summer season**

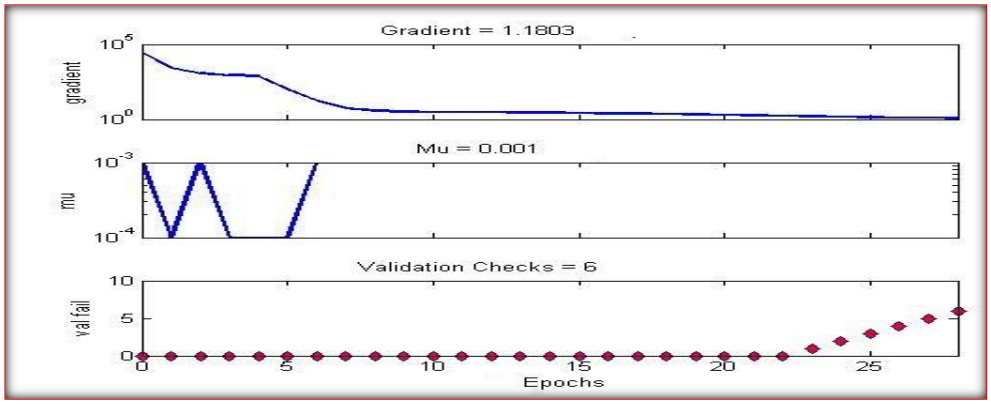


**Figure 5.81 Performance Evaluation Curve for monsoon season**

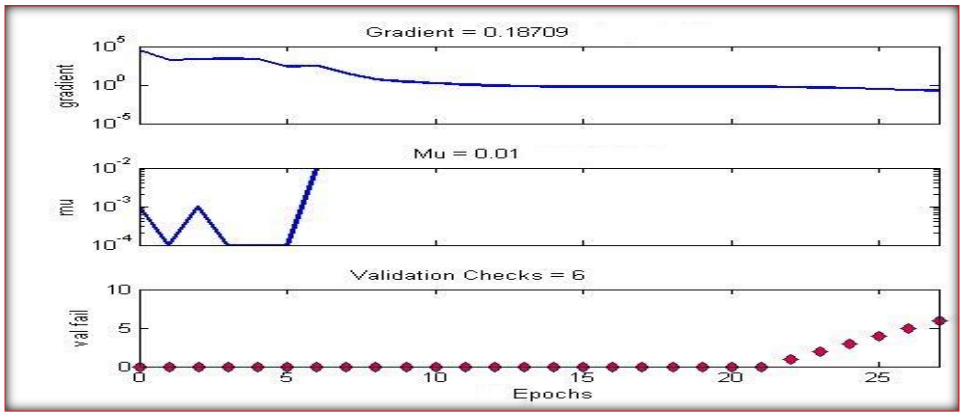


**Figure 5.82 Performance Evaluation Curve for monsoon season**

The figures shows the convergence at μ=0.01 after the completion of epochs in the model.

## 5.12 Error Calculation of the Models

The mean absolute error (MAE), root man square error (RMSE) and mean absolute percentage error (MAPE) between the predicted and actual WQI of the three models are calculated to conclude the best fit model for WQI prediction. The Table 5.19 represents the calculated errors in three respective seasons.

As described in Chapter IV, the error analysis measures how close forecasts and prediction are to eventual outcomes, measured mean absolute percentage deviation and also the differences between values predicted by the model and actual observed values.

**Table 5.19 Results of Error Analysis in Three Models**

| | ANFIS | | | ANN | | | MCS | | |
|---|---|---|---|---|---|---|---|---|---|
| | Summer | Monsoon | Winter | Summer | Monsoon | Winter | Summer | Monsoon | Winter |
| MAE | 0.037 | 0.045 | 0.038 | 0.088 | 0.095 | 0.076 | 0.094 | 0.088 | 0.082 |
| MAPE | 0.178 | 0.169 | 0.193 | 0.328 | 0.269 | 0.258 | 0.351 | 0.248 | 0.264 |
| RMSE | 0.043 | 0.051 | 0.044 | 0.109 | 0.114 | 0.09 | 0.118 | 0.105 | 0.097 |

According to the results of error analysis, the MAE, MAPE and RMSE are minimum in case of Adaptive Neuro-Fuzzy Inference System (ANFIS), where as the errors are less in monsoon of Monte Carlo Simulations then that of Artificial Neural Network Model. It can be concluded that ANFIS model was the best model in analysis and modelling of water quality.

# CHAPTER VI

# CONCLUSIONS

Analysis of water quality for Brahmani River is done by various techniques like Spearman's Rank Correlation, Calculation of Parts of water quality parameters, Overall Water Quality Index (WQI), Multivariate Analysis of variance (MANOVA) with Discriminant Analysis, Principal component Analysis and Factor Analysis, Canonical Correlation Analysis (CCA), Cluster Analysis (CA). Modelling is done using Adaptive Neuro-Fuzzy Inference System (ANFIS) in MATLAB, by Artificial Neural Network (ANN) and risk based analysis by Monte Carlo simulations (MCS). The Error analysis and performance evaluation of these models was also done to know the best fit model for this study. Conclusions from the present research work can be listed as follows:

- Spearman's Rank Correlation Analysis indicates season showing increasing trend. Biochemical Oxygen Demand, Chemical Oxygen Demand, Electrical Conductivity, Nitrate as Nitrogen, Total Coliform bacteria, Faecal Coliform bacteria, Nitrogen as Ammonia, Total Hardness as $CaCO_3$, Total Alkalinity as $CaCO_3$ in monsoon and winter seasons show decreasing trend. These variations are due to the temporal variations in gauging stations.
- From the calculation of parts of parameter in river water, it can be stated that the quantity of water parameters in three consecutive seasons follow an equal trend. It can be concluded that, the inflow of effluents to the river are constant throughout the year.
- The Water Quality Index (WQI) values for the gauging stations vary from excellent to good in monsoon season and from good to poor during summer and winter seasons. The ranges of water quality parameters were within the range as recommended by ICMR and the water can be used for domestic purposes. However, necessary preventive measures to maintain good water quality of Brahmani River Basin must be taken up to ensure the safety of the River Basin and to preserve this valuable resource to the future generations. Water Quality Index may be used as a tool to convey the useful information regarding the quality of water in an easy and understandable way to the public and policy makers.
- Multivariate statistical techniques are used to examine spatial and temporal variations in water quality. Discriminant analyses on the gauging stations show that three is small difference between the stations at three seasons investigated. This suggests that the anthropogenic activities, mainly the effluents of industries, runoff from agricultural lands and waste water from residential areas into the river account for the observed variability in the water quality (especially with respect to pH, Electrical Conductivity and Chemical Oxygen Demand).
- The surface water quality data for spatial variations and the relationship between physical, chemical and biological parameters are evaluated. Results of Principal Component Analysis (PCA) indicate that Panposh down-stream and Talcher up-

stream monitoring stations are the principal monitoring stations having more impact on quality of water, than other non-principal stations. These findings are supported by simple regression as well. Hardness and COD are important chemical parameters and Electrical Conductivity, TC and FC are the important physical and biological parameters. Canonical Correlation Analysis (CCA) is verified by simple correlation also. The methods used here can offer an effective solution to water quality management for the cases involving complexity in quality data.

- Hierarchical cluster analysis groups the sampling sites into two clusters for each season and also classifies 11 water quality parameters into two clusters based on similarity sites. The temporal pattern shows that January, March, June, October and December had high pollution level in comparison with the rest of months. The spatial pattern shows that Aul had the lowest level of pollution while other sampling sites have higher level of pollution. According to the water quality parameters; Electrical Conductivity, TC, FC and COD affect more than that of other parameters.
- The correlation analysis shows that there is moderate correlation between the parameters due to changes in land use, mining and improper effluent discharge in the river. When parameters exhibits strong or moderate correlation, explicit numerical representation of the input and output parameters is almost impossible and WQI may not effectively characterise quality of water. Therefore, it is vital to convert correlated parameters into uncorrelated parameters for efficient forecasting of water quality. PCA provides a suitable method to transform correlated parameters into uncorrelated parameters.
- Water Quality is a vague term that cannot easily be described using crisp data set. Instead, it made a sense when it is considered to be a fuzzy set that provides the mathematical foundation to express the term water quality in a linguistic way, e.g. excellent. Good, poor, very poor and unsuitable.
- Fuzzy reasoning technique does not rely on a crisp data set, rather uses linguistic terminology to process the output. To this end the Adaptive Neuro-Fuzzy Inference System (ANFIS), which integrate fuzzy logic with neural network, are proposed to predict WQI along with ANN model. The ANFIS and ANN model predicts the water quality with a reasonable accuracy. Regression plots between actual and predicted WQI through ANFIS model for training and testing data haverevealed a high degree of coefficient of determination ($R^2$) values of 0.994 and 0.995 for training and testing in summer, 0.985 and 0.990 in monsoon and 0.992 and 0.993 in winter respectively. However, the coefficients of determination ($R^2$) for Artificial Neural Network (ANN) between actual and predicted values of WQI are 0.945, 0.941 and 0.965 for summer, monsoon and winter respectively.
- Depending upon the degrees of freedom as calculated in canonical correlation analysis, Monte Carlo Simulations is applied, which provides technique for simulating the parameters having high degrees of freedom. The coefficients of determination ($R^2$) for the actual and simulated values of WQI are found to be 0.929, 0.953 and 0.970 for summer, monsoon and winter respectively.
- According to the performance evaluation and error analysis of the models, there is least error in case of ANFIS when compared with that of ANN and MCS. Therefore, it can be said that ANFIS predicted WQI with reasonable accuracy. From the results of ANFIS based analysis, it can be concluded that if the present conditions prevail the future years also WQI values will have the same trend as those from 2003 to 2012.

# CHAPTER VII

# REFERENCES


- ✓ Akkaraboyina, M.K. and Raju, B.S.N. (2012). Assessment of water Quality Index of River Godavari at Rajahmundry. Universal Journal of Environmental Research and Technology. 2(3), 161-167.
- ✓ Alexandridis, K. (2007). Monte Carlo Extreme Event Simulations for Understanding WaterQuality Change Classifications in the GBR Region. CSIRO Sustainable Ecosystems. Kostas.Alexandridis@csiro.au.1-18.
- ✓ Amstadter, B.L., (1971). Reliability Mathematics: Fundamentals; Practices; Procedures. McGraw Hill Book Company, New York, USA.
- ✓ Antonopoulous, V.Z, Papamichail, D.M., and Mitsiou, K.A. (1998). Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. Hydrology Earth System. Sc. 5 (4), 679-691.
- ✓ Areerachakul, S. (2012). Comparison of ANFIS and ANN for Estimation of Biochemical Oxygen Demand Parameter in Surface Water. International Journal of Chemical and Biological Engineering. 6, 286-290.
- ✓ Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. Proceedings of the royal statistical society. Series A 160. 268-282 JSTOR 96803.
- ✓ Basil, M., Papadopoulos, C., Sutherland, D. and Yeung, H. (2001). Application of Probabilistic Uncertainty Methods (Monte Carlo Simulation) in Flow Measurement Uncertainty Estimation. Flow Measurement 2001 – International Conference. 1- 21.
- ✓ Boyacioglu, H., Boyaciaoglu, H. And Gunduz, O. (2005). Application of Factor Analysis in the Assessment of water quality in Buyuk Menderes River Basin. European Water. 9(10), 43-49.
- ✓ Boyacioglu, H. (2006).Surface water quality assessment using factor analysis. Water SA. 32(3), 383-393.
- ✓ Burn, D.H. and McBean, E.A. (1985). Optimization modeling of water quality in an uncertain environment. Water Resource. 2 (2), 42- 44.
- ✓ Catell, R.B. and Jaspers, J. (1967). A general Plasmode (No. 30-10-5-2) for factor analytic exercises and research. Mult. Behav. Res. Monogr. 67, 1-212.
- ✓ Chakrabarty, S. and Sarma, H.P. (2011). A statistical approach to multivariate analysis of drinking water quality in Kamrup district, Assam, India. Archives of Applied Science Research. 3 (5), 258-264.
- ✓ Dezfoli, K.A. (2003). Principles of fuzzy theory and its application on water engineering problems. Iran Jihad Press, Tehran.
- ✓ EI Kholy, R.M.S., Khalil, B.M. and Gawad, S.T.A. (1997). Assessment of the National Water Quality monitoring program of Egypt. National Water research Centre NWRC, Qalyubia, Egypt.
- ✓ Eneji, I.S., Onuche, A.P., Sha'Ato,R. (2012). Spatial and Temporal Variation in Water Quality of River Benue, Nigeria. Journal of Environmental Protection. 3, 915-921.

- ✓ Fan, X., Cui, B., Zhao, H., Zhang, Z. and Zhang, H. (2010). Assessment of river water quality in Pearl River Delta using multivariate statistical techniques. Procedia Environmental Sciences. 2, 1220–1234.
- ✓ Galavi, H. and Shui, T. (2012). Neuro-fuzzy modelling and forecasting in water resources. Scientific Research and Essays. 7(24), 2112-2121.
- ✓ Jang, R.J. (1991a). Fuzzy modelling using generalized neural networks and Kalmman filter algorithm. In: Proceedings of ninth national conference on artificial intelligence, 762–767.
- ✓ Jang, R.J. (1991b). Rule extraction using generalized neural networks. In: Proceedings of 4th IFSA world congress, 82–86
- ✓ Jang, R.J. and Gulley, N. (1996). Fuzzy logic toolbox: Reference manual. The Math works Inc, Natick.
- ✓ Jang, J.S.R., Sun, C.T., Mizutani, E. (1997). Neuro-fuzzy and soft computing, A computational approach to learning and machine intelligence, 1st edn. Prentice Hall, Englewood Cliffs, NJ, USA.
- ✓ Jiang, Y., Nan, Z. and Yang, S. (2013). Risk assessment of water quality using Monte Carlo simulation and artificial neural network method. Journal of Environmental Management. 122, 130-136.
- ✓ Jianqin, M.A., Jingjing, G.O.U., and Xiaojie, L.I.U. (2010). Water Quality Evaluation Model Based on Principal Component Analysis and Information Entropy: Application in Jinshui River. J. Resour. Ecol. 1(3), 249-252.
- ✓ Johnson, R.A. and Wichern, D.W. (1988). Applied Multivariate Statistical Analysis, 2nd Edition. Prentice-Hall International, Inc., London.
- ✓ Juahir, H., Zain, S.M., Toriman, E., Mokhtar, M. and CheMan, H. (2004). Application of Artifiacial Neural Network Models for Predicting Water Quality Index. Jurnal Kejuruteraan Awam. 16(2), 42-55.
- ✓ Khalil, B.M., Awadallah, A.G., Karaman, H. and El-Sayed, A. (2012).Application of Artificial Neural Networks for the Prediction of Water Quality Variables in the Nile Delta. Journal of Water Resource and Protection. 4, 388-394.
- ✓ Khandelwal, M. and Singh, T. (2005). Prediction of mine water quality by physical parameters. Journal of Scientific and Industrial Research. 64, 564-570.
- ✓ Klassen, M.S. and Pao, Y.H. (1988). Characteristics of the functional link net: A higher order delta rule net. In IEEE proceedings of the international conference on neural networks, San Diego, CA, USA.
- ✓ Kottegoda, N.T., Rosso, R. (1998). Statistics, probability and reliability for civil and environmental engineers, McGraw-Hill, ISBN 0-07-035965-2.
- ✓ Kumar, P., Saxena, K.K., Singh, N. O., Nayak, A. K., Tyagi, B.C., Ali, S., Panney, N.N. and Mahanta, P.C. (2011). Application of multivariate statistical techniques for water quality characterisation of Sarada Sagar Reservoir, Ind. J. of Fisheries, 58 (4), 21-26.
- ✓ Li, Y., Linyu, X.U., Li, S. (2009).Water Quality Analysis of the Songhua River Basin Using Multivariate Techniques. Journal of Water Resource and Protection. 2, 110-121.
- ✓ Liping, Z., Jie, P., Yongchao, W., Muqi, Y., Yuanyuan, S. and Liu, Y. (2010). SPSS For Water Quality Assessment Of Beijing Typical RiverBased On Principal Component Analysis. International Conference on Digital Manufacturing & Automation. 396-398.
- ✓ Mahapatra, S.S., Sahu, M., Patel, R.K. and Panda, B.N. (2012). Prediction of Water Quality Using Principal Component Analysis. Water Qual Expo Health.4, 93–104.

- ✓ Mangukiya, R., Bhattacharya, T. and Chakraborty, S. (2012). Quality Characterization of Groundwater using Water Quality Index in Surat City, Gujarat, India. International Research Journal of Environment Sciences. 1(4), 14-23.
- ✓ Mazlum, N., Ozer, A. and Mazlum, S. (1999). Interpretation of Water Quality Data by Principal Components Analysis. Tran. J. Eng. Environ. Sci. 23, 19-26.
- ✓ Mishra, A. (2010). Assessment of water quality using principal component analysis: A case study of River Ganges. Химия и технология воды. ISSN 0204–3556. 32 (4), 415-427.
- ✓ Mohammad Salah, E.A., Turki, A.M. and Al-Othman, E.M. (2012).Assessment of Water Quality of Euphrates River Using Cluster Analysis. Journal of Environmental Protection.3, 1629-1633.
- ✓ Najah, A., Elshafie, A., Karima, O.A. and Jaffar, O. (2009). Prediction of Johor river water quality parameters using artificial neural networks. European Journal of Science Research. 28(3), 422-435.
- ✓ Nicolis, G. (1995). Introduction to non-linear scienc*e*, Cambridge University Press, ISBN 0. 521 46228 2, Cambridge, UK.
- ✓ Noori, R. Sabahi, M.S., Karbassi,A.R., Baghvand, A., Zadeh, H.T. (2010). Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. Desalination. Science Direct. 260, 129-136.
- ✓ Ouyang, Y., Nkedi-Kizza, P., Wu, Q.T., Shinde, D., Huang, C.H. (2006). Assessment of seasonal variations in surface water quality. Water Research. 40, 3800-3810.
- ✓ Pedrycz, W.B (1989). Fuzzy control and fuzzy systems. Wiley, New York.
- ✓ Rao, C.R. (1973). Linear Statistical Inference and Application. Wiley Eastern Ltd., New Delhi, India.
- ✓ Rumelhart, D.E., Hinton, G.E. and William, D.E. (1986). Learning internal representations by error propagation. In: Parallel distributed processing: Explorations in the microstructure of cognition. MIT Press, Cambridge, 1–8, 318–362.
- ✓ Saatsaz, M., Suliman, W. N. A. B., Ibrahim, S. and Mohammadi, K. (2013). Multivariate Statistical Techniques for the Evaluation of Spatial and Temporal Variations in Groundwater Quality of Astaneh- Kouchesfan Plain, Sefīd-Rūd Basin, North of Iran. 9th International River Engineering Conference. Shahid Chamran University, Ahwaz, Iran.
- ✓ Sahu, M., Mahapatra, S.S., Sahu, H.B. and Patel, R.K. (2011). Prediction of Water Quality Index Using Neuro Fuzzy Inference System. Water Qual Expo Health. 3, 175–191.
- ✓ Salah, E.A.M., Turki, A.M. and Al-Othman, E. M. (2011). Assessment of water quality of Euphrates River using cluster analysis. 3, 1629-1633.
- ✓ Sanders, T.G. and Adrian, D.D. (1978). Sampling frequency for river quality monitoring. Water resources research. 14, 569-576.
- ✓ Seber, G.A.F. (1983). Multivariate Observations. Wiley Series in Probability and Statistics.
- ✓ Shlens, J. (2003). A tutorial on principal component analysis (Derivation, Discussion and Singular Value Decomposition).jonshlens@ucsd.edu. 1-16.
- ✓ Shrestha, S., Kazama, F. and Nakamura, T. (2008). Use of principal component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the Mekong River. Journal of Hydro informatics. 10 (1), 43-56.
- ✓ Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsa, D. Anthemidis, A., Sofoniou, M. and Kouimtzis, T. (2003). Assessment of the surface water quality in Northern Greece. Water Research. 37, 4119–4124.

- ✓ Singh, K.P., Malik, A. And Sinha, S. (2005). Water quality assessment and apportionment of pollution sources of Gomti River (India) using,multi-variate statistical techniques- a case study. Analytica Chemica Acta, 538, 355-374.
- ✓ Singkran, N., Yenpiem, A. and Sasitorn, P. (2010).Determining Water Conditions in the North eastern Rivers of Thailand Using Time Series and Water Quality Index Models. Journal of Sustainable Energy & Environment.1, 47-48.
- ✓ Statheropoulos. M., Vassiliadis, N., Pappa, A. (1998). Principal and canonical correlation analysis for examining air pollution and meteorological data. Atmospheric Environment. 32, 1087-1095.
- ✓ Sugeno, M. (1985). Industrial applications of fuzzy control. Amsterdam, Elsevier.
- ✓ Sugeno, M. and Kang, G.T. (1988). Structure identification of fuzzy model. Fuzzy Sets Syst. 28, 15–33
- ✓ Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern. 15, 116–132.
- ✓ Vousta, D., Manoli, E., Samara, C., Sotoniou, M. And Stratis, I. (2001). A study of surface water quality in Macedonia, Greece: Speciation of nitrogen and phosphorous. Water Air Soil Pollution. 129, 13-32.
- ✓ Yan, H., Zou, Z. and wang, H. (2010). Adaptive neuro fuzzy inference system for classification of water quality status. Journal of Environmental Sciences. 22(12), 1891-1896.
- ✓ Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioural sciences. Unpublished Dissertation. Harvard University, Cambridge, MA, USA.
- ✓ Zhou, F., Liu, Y. and Guo, H. (2006). Application of Multivariate Statistical Methods to Water Quality Assessment of the Watercourses in North western New Territories, Hong Kong. Environ Monit Asses.132, 1–13.
- ✓ Zhao, Z.W. and CUI, F.Y. (2009). Multivariate statistical analysis for the surface water quality of the Luan River, China. Journal of Zhejiang University SCIENCE A. 10(1), 142-148.