

ANALYSIS OF WATER QUALITY DATA USING STATISTICAL AND ANN TECHNIQUE

A DISSERTATION

*Submitted in partial fulfillment of the
requirements for the award of the degree*

of

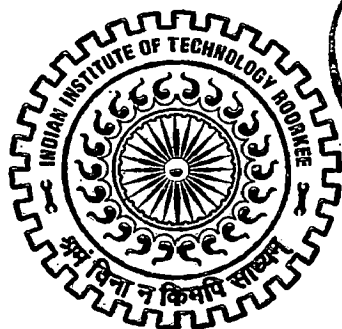
MASTER OF TECHNOLOGY

in

WATER RESOURCES DEVELOPMENT
(CIVIL)

By

JOYDEEP DUTTA



DEPARTMENT OF WATER RESOURCES DEVELOPMENT
AND MANAGEMENT

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE - 247 667 (INDIA)

JUNE, 2005

10

CANDIDATE'S DECLARATION

I hereby declare that the dissertation titled "ANALYSIS OF WATER QUALITY DATA USING STATISTICAL AND ANN TECHNIQUE", which is being submitted in partial fulfillment of the requirement for the award of the Degree of Master of Technology in Water Resources Development (Civil) at Department of Water Resources Development and Management (WRD&M), Indian Institute of Technology Roorkee (IITR), is an authentic record of my own work carried out during July, 2004 to June, 2005 under the supervision and guidance of Dr. S.K. Mishra, Assistant Professor, WRD&M, IIT Roorkee, (India) and Dr. M.K. Sharma, Scientist 'B', National Institute of Hydrology, Roorkee, (India).

I have not submitted the matter embodied in this dissertation for the award of any other degree.

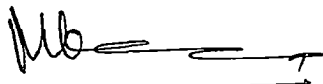
Place: IIT Roorkee,

Dated: 22 June, 2005

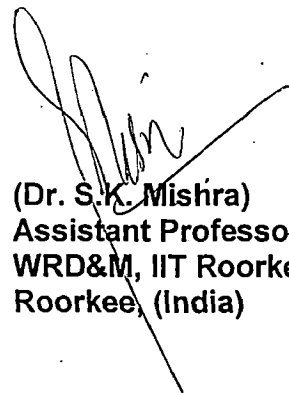
Joydeep Dutta
(Joydeep Dutta)

CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.



(Dr. M.K. Sharma)
Scientist - 'B',
National Institute of Hydrology,
Roorkee, (India)



(Dr. S.K. Mishra)
Assistant Professor,
WRD&M, IIT Roorkee,
Roorkee, (India)

ACKNOWLEDGEMENT

I take this opportunity to express my profound gratitude and sincere thanks to **Dr. S.K. Mishra, Assistant Professor, WRD&M, Indian Institute of Technology Roorkee, Roorkee,** and **Dr. M. K. Sharma, Scientist 'B', National Institute of Hydrology, Roorkee,** for their constant encouragement, inspiring guidance, persuasion and continued support for all aspects of this dissertation, without which this dissertation would have been incomplete.

I wish to express my deep sense of gratitude to **Dr. Avinash Agarwal, Scientist 'E', National Institute of Hydrology, Roorkee** for his kind cooperation, encouragement and guidance. I also extend my hearty thanks to all faculty members WRD&M, gratitude to all staff members of Water Resources Development and Management, especially those in computer lab and departmental library, for their co-operation.

I wish to express my gratitude to the **Secretary, CE, Addl. CE, and Deputy Secretary (1) to the Govt. of Assam, Water Resources Department,** for giving me an opportunity to under go this course.

I also express my sincere thanks to **Mr. Nitin Kumar** for his painstaking effort to type the dissertation report in due time.

It will be ungrateful, if I don't mention the co-operation and moral support from **my Wife Piyali and my Daughters Nikita and Ankita.**

Place: IIT Roorkee,

Dated: 22 June, 2005

Joydeep Dutta
(Joydeep Dutta)

ABSTRACT

In the present study, an effort has been made to develop statistical and ANN models for estimation of sodium concentration in pre-monsoon and post-monsoon seasons using routinely monitored water quality parameters of ground water wells in Jaipur district, Rajasthan (India). The Best Subset procedure based on R^2 (coefficient of determination) and F (Fisher's test) values was used in model dissemination. It was found that electrical conductivity, hardness, chloride, and sulphate could be used as surrogate parameters for the prediction of sodium. The model values of Na when compared with actual values (validation) showed a reasonably good matching. Further it is was noticed that there was not a single model which could be used to predict the Na levels. It is primarily attributed to the fact that sodium concentration not only varies from site to site but also varies from season to season. Secondly, Principal component analysis was used to predict the dominating water quality constituents and it was revealed that four principal components are accounted for the total chemical variability in the ground water quality for pre-monsoon season and three principal components for post-monsoon season, respectively. The common factors conductivity, fluoride, nitrate, alkalinity, and phosphate have perceptible influence on the quality of groundwater of Jaipur district, Rajasthan. Finally, Back Propagation, two layer feed forward ANN models for both pre-monsoon and post-monsoon season was developed for estimation of sodium using the steepest descent optimization technique. ANN models were developed considering a fixed number of iterations as 1000 and these were verified on the data not considered in calibration. The input variables considered for different model structures were identified through correlation analysis. Based on the statistical performance evaluation criteria such as root mean square error (RMSE), correlation coefficient (CC), and coefficient of efficiency (CE), the results indicated satisfactory performance of ANN based model.

CONTENTS

CHAPTER	PARTICULARS	PAGENO.
	CANDIDATE'S DECLARATION	i
	ACKNOWLEDGEMENT	li
	ABSTRACT	lii
	CONTENTS	iv
	LIST OF TABLES	viii
	LIST OF FIGURES	x
CHAPTER 1	INTRODUCTION	1
	1.1 General	1
	1.2 Sources of Pollution	1
	1.3 Occurrence of Groundwater	2
	1.4 Objective of Study	3
	1.5 Organization of Work	4
CHAPTER 2	LITERATURE REVIEW	6
	2.1 Best Subset Procedure of Regression Analysis	6
	2.2 Principal Component Analysis	10
	2.3 Artificial Neural Network (ANN)	17
CHAPTER 3	STUDY AREA DESCRIPTION	22
	3.1 General	22
	3.2 Location	22
	3.3 Climate	25
	3.4 Geology and Mineral	25
	3.5 Physiography and Soil	26
CHAPTER 4	STATISTICAL MODELS DEVELOPMENT	27
	4.1 Multiple Linear Regression	27
	4.1.1 Least squares	27
	4.1.2 The regression equation	27
	4.1.3 Unique prediction and partial correlation	28
	4.1.4 Predicted and residual scores	28

4.1.5	Residual variance and R-square	29
4.1.6	Interpreting the correlation coefficient R	29
4.1.7	Assumptions, limitations, and practical considerations	30
	(a) Assumption of linearity	
	(b) Normality assumption	
	(c) Limitations	
	(d) Choice of the number of variables	
	(e) Multi-collinearity and matrix III conditioning	
	(f) Fitting centered polynomial models	
	(g) The importance of residual analysis	
4.2	Formulation of Models	32
4.2.1	Preliminary analysis of data	33
4.2.2	Secondary analysis	33
4.2.3	Selection of independent variables for regression analysis	33
4.2.4	Best subset regression	34
	(a) R^2 - criterion	
	(b) F-value criterion	
4.3	Principal Component Analysis	37
4.3.1	Basic idea of factor analysis as a data reduction method	40
	(a) Combining two variables into a single factor	
	(b) Extracting principal component	
	(c) Generalizing to the case of multiple variables	
	(d) Multiple orthogonal factors	
	(e) How many factors to extract?	

4.3.2	Factor analysis as a classification method	42
	(a) Factor loadings	
	(b) Rotating the factor structure	
4.4.	Formulation of Model	43
CHAPTER 5	ANN PROCEDURE	44
5.1	General	44
5.2	The ANN Structure	46
	5.2.1 Biological neuron	46
	5.2.2. Artificial neuron	47
5.3	Gradient Descent Learning Algorithm	48
5.4	The Neural Network Topology	49
	5.4.1 Feed forward network	50
	5.4.2. Feed back network	50
5.5	Activation Function	50
	5.5.1 Sigmoid function	50
5.6	Architecture of ANN	51
5.7	Training of Artificial Neural Network	51
	5.7.1 Supervised training	52
	5.7.2 Un supervised training	52
5.8	Back Propagation Algorithm	52
5.9	Learning Factors of Back Propagation	57
	5.9.1 Initial weights	57
	5.9.2 Learning rate (η)	58
	5.9.3 Momentum factor (α)	58
	5.9.4 Data normalization	59
	5.9.5 Training data and generalization	60
5.10	Steps in Development of ANN Model	60
5.11	Performance Evaluation Criteria	61

CHAPTER 6	RESULTS AND DISCUSSION	63
6.1	Statistical Model Development	63
6.1.1	Best Subset Procedure	63
6.1.2	Principal Component Analysis	83
6.2	Artificial Neural Network Analysis	96
CHAPTER 7	SUMMARY AND CONCLUSIONS	107
7.1	Best Subset Procedure	107
7.2	Principal Component Analysis	107
7.3	Artificial Neural Network Analysis	108
7.4	Suggestion Proposed for Future Studies	108
	REFERENCES	110

LIST OF TABLES

TABLE	PARTICULARS	PAGE NO.
6.1.	Pearson correlation coefficient between water quality parameters for pre-monsoon, Jaipur District-(Rajasthan) with first 25 test data	64
6.2.	R ² of water quality of parameters with Na for pre-monsoon	66
6.3.	Various combinations of models and their statistics for pre-monsoon season	67-69
6.4.	Selected sets/subsets candidate for possible model independent variables for pre-monsoon season with sodium	70
6.5.	Selection of model variables on the basis of F-statistics for pre-monsoon season with sodium	71
6.6.	Validation for the model equation for pre-monsoon season with sodium	73
6.7.	Pearson correlation coefficient between water quality parameters for post-monsoon, Jaipur District (Rajasthan) with first 25 test data	76
6.8.	R ² of water quality of parameters with Na for post-monsoon	77
6.9.	Various combinations of models and their statistics for post-monsoon season	78-80
6.10.	Selected sets/subsets candidate for possible model independent variables for post-monsoon season with sodium	81
6.11.	Selection of model variables on the basis of F-statistics for post-monsoon season with sodium	82

6.12.	Validation for the model equation for post-monsoon season with sodium	83
6.13.	Eigen values based on correlation matrix for pre-monsoon	87
6.14.	Eigen values based on correlation matrix for post-monsoon	88
6.15.	Varimax rotated component loadings for pre-monsoon season	89
6.16.	Varimax rotated component loadings for post-monsoon season	90
6.17.	Description of various ANN models for training and testing of sodium levels for pre-monsoon season	98
6.18.	Description of various ANN models for training and testing of sodium levels for post-monsoon season	99
6.19.	Comparative performance of selected models for pre-monsoon season for sodium	100
6.20.	Comparative performance of selected models for post-monsoon season for sodium	101
6.21	Comparison of statistical and ANN Based models	106

LIST OF FIGURES

FIGURE	PARTICULARS	PAGE NO.
3.1.	Location of Rajasthan in India map	23
3.2.	Geographical map of Rajasthan (India)	24
3.3.	Study area, Jaipur District showing location of sampling sites	24
5.1	The building blocks of ANN	45
5.2	Anatomy of biological neuron	47
5.3	Anatomy of artificial neuron	48
5.4	Gradient descent in one dimension	49
5.5	The sigmoid function	51
5.6	A typical two layered back propagation feed forward neural network	53
6.1	Model validation of observed and computed values of sodium (mg/l) in the pre-monsoon season	74
6.2	Plot of observed and model values for pre-monsoon season	75
6.3	Model validation of observed and computed values of sodium (mg/l) in the post-monsoon season	84
6.4	Plot of observed and model values for post-monsoon season	85
6.5	Loading of variables for pre-monsoon season	92
6.6	Loading of variables for post-monsoon season	93
6.7	Calibration result of model ANN3 for pre-monsoon	102
6.8	Validation result of model ANN3 for pre-monsoon	103
6.9	Calibration result of model ANN11 for post-monsoon	104
6.10	Validation result of model ANN 11 for post-monsoon	105

INTRODUCTION

1.1 General

Water is indispensable for existence of life. It is an important component of hydrologic cycle. It can occur above the ground as surface water and can be hidden beneath as ground water. Groundwater was once considered to be free from pollution. But the rapid industrialization made a paradigm shift to this concept. The very uses for which the water is utilized are adding contaminants to ground water at an alarming rate. The various uses of ground water include industrial, agricultural, and human needs. The indiscriminate disposal of industrial wastes on mother earth slowly makes the ground water susceptible to pollution. Ground water when once gets polluted, its purification is hopelessly difficult. The quality of ground water is usually characterized in terms of certain water quality constituents according to its physical, chemical, and microbiological properties.

1.2. Sources of Pollution

The main sources of water pollution are:

- (a) Environmental: This type of pollution is due to the environment through which the flow of ground water takes place. Pollution caused due to the movement of groundwater through chemically active rocks, salt water intrusion etc., falls under this category.
- (b) Domestic: Domestic pollution is caused due to the accidental breakage of sewers, percolation from septic tanks, artificial recharge of aquifers by sewage water etc.

(c) Industrial: This is due to the indiscriminate disposal of industrial waste on land, rivers, etc. Effluents discharged from industries get infiltrated to the ground water and become polluted.

(d) Agricultural: This type of pollution occurs due to the infiltration of irrigation water and rainwater containing fertilizers, salts, pesticides, etc. The pollutant transport mechanisms are mainly advection and hydrodynamic dispersion.

1.3 Occurrence of Ground Water

Ground water is a precious and the most widely distributed resource of the earth and unlike any other mineral resource, it gets its annual replenishment from the meteoric precipitation. The world's total water resources are estimated as 1.37×10^8 million ha-m. Of these, global water resources (about 97.2 percent) is salt water, mainly in oceans, and only 2.8 percent is available as fresh water at any time on the planet earth. Out of this 2.8 percent, about 2.2 percent is available as surface water and 0.6 percent as ground water. Out of this 2.2 percent of surface water, 2.15 percent is fresh water in glaciers and icecaps and only of the order of 0.01 percent (1.36×10^4 Mha-m) is available in lakes and reservoirs, and 0.0001 percent in streams; the remaining being in other forms: 0.001 percent as water vapour in atmosphere and 0.002 percent as soil moisture in the top 0.6 meter. Out of the 0.6 percent of stored ground water, only about 0.3 percent (41.1×10^4 Mha-m) can be economically extracted with the present drilling technology, for the remaining is not available as it is situated below a depth of 800 meter.

The knowledge of the occurrence, replenishment and recovery of ground water assumes special significance in arid and semi-arid regions. Surface waters, except when brought in by rivers from elsewhere, are normally scarce, or even absent in such areas. The India Meteorological Department categorizes a 'year' as a 'drought year' in which rainfall deficiency is numerically equal to or lesser than 25 percent of normal (Tizro 1995). With a view to provide protection and control of pollution of water and matters connected to it, Indian Parliament has enacted 'Water (Prevention and Control of Pollution) Act 1974'.

1.4 Objective of Study

In Rajasthan, out of a total area of 1.1 million hectares under well irrigation about 57 percent of the area is affected by the problem of salinity and alkalinity (Paliwal, 1972). The salt affected area is about 70% of the irrigated area in the districts of Bikaner, Jaisalmer, Pali, Jodhpur, Bharatpur, Barmer, Nagpur, Jaipur and Bhilwara. On this basis a total area of more than 100,000 hectares of land is salt affected in the districts of Jaipur, Bhilwara and Bharatpur. The mean chemical composition of well waters in some of the districts of Rajasthan shows that, due to low rainfall, the ground waters of western region are more saline than those of eastern region.

The number of parameters needed to fully specify the water quality for a particular place of region is quite large. Moreover, due to lack of laboratory facilities and/or trained manpower it becomes difficult to determine all the constituents. Also routine chemical analysis of ground water is a lengthy, laborious and time-consuming process. Therefore, it would be worthwhile if an indirect approach is used to estimate water quality within the desired precision

using some easily measurable water quality constituents. Keeping this in view, this study's objective is to analyze water quality data for the development of a model predicting the concentration of sodium (one of the major constituents of salinity) using

- i. Best subset procedure of Regression analysis by developing suitable Regression models,
- ii. Principal component analysis (PCA) to investigate the chemical relationship between different water quality constituents and thereby predicting the dominating constituents.
- iii. Artificial neural network (ANN) analysis to compare the results obtained with the above statistical methods adopted.

District Jaipur is selected for the study, which is located in the northeastern part of Rajasthan, India. The District covers an area of 10878 km². Thirty-eight samples from district Jaipur were collected at different locations during May 2002 and November 2003 respectively i.e., for pre-monsoon as well as for post-monsoon season. Following standard methods, physico-chemical analysis was performed by National Institute of Hydrology, Roorkee.

1.5 Organization of Work

With the objective to determine the major water quality parameter "sodium" by indirect method i.e. by formulation of models, this dissertation is organized as follows:

Chapter 2: Deals with the Literature Review of Multilinear regression, Principal Component Analysis and Artificial Neural Network.

Chapter 3: Describes in brief the study area, Jaipur District, Rajasthan (India)

Chapter 4: Deals with the statistical formulation of models.

Chapter 5: Describes the artificial neural network models.

Chapter 6: Provides a discussion of the results of the study.

Chapter 7: Concludes and provides suggestions for future study.

LITERATURE REVIEW

2.1 Best Subset Procedure of Regression Analysis

As a result of increasing industrialization, urbanization, civilization and other developmental activities most of our water bodies, like ponds, lakes, streams, rivers as well as groundwater bodies have become polluted. The industrial effluents, sewage, domestic waste, agricultural and land drainage etc., are the major sources that cause water pollution. The necessity of rapid monitoring of water quality is being urgently felt. It is however, very difficult in developing countries, like India, where laboratory facilities and / or trained man power are inadequate.

Correlations among water quality parameters in a specific environmental condition have been shown to be useful and successful. When such correlations exist, measuring a few important parameters and then predicting others using these correlations and regression analysis would give some idea about the overall quality of water. Correlation analysis provides an excellent tool for rapid monitoring of the status of pollution of a water body and achieves economy in matters of collection and analysis of samples.

Kannan and Vallinuyagam (1992) carried out systematic study of correlation analysis of water quality parameters of industrial effluents, Match industry. Industrial effluent had been collected from different match units were analyzed bi-monthly for a period of 8 months. Physico-chemical water quality parameters were found to be well above the permissible levels. The computed water quality index (WQI) values indicated highly polluted nature of the effluents. Correlation analysis of water quality parameters were carried out

among all the possible pairs of quality parameters of match industry effluents, and correlation coefficients computed for all possible correlations. Significant correlations were noticed to exist between the following pairs of water quality parameters: TDS – EC, permanent hardness – Total hardness, WQI – K, and COD-PO₄. Correlation analysis of water quality data revealed existence of linear relationships between different pairs of parameters. Thus correlations provided an excellent tool for the prediction of physico-chemical water quality parameter values within reasonable degree of accuracy.

basis
for
these
correlations

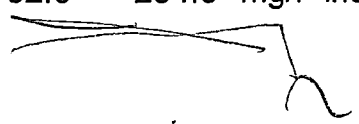
Singanen et al. (1995) carried out a correlation study on physico-chemical characteristics of groundwater in Rameswaram Island and discussed the usefulness of the correlations in predicting groundwater quality characteristics. Groundwater samples collected from bore wells and dug-wells of 5 villages of Rameswaram and observations clearly indicated that, most of the samples were alkaline in nature and had high salinity. A systematic calculation of correlation coefficient and regression analysis had been carried out among the various water quality parameters resulted in significant linear relationships between the following water quality parameters, namely, EC, Ca, Mg, Na, K Cl, SO₄, NO₃ and COD.

reason
for their
correlations
COD & EC

Krishna et al. (1995) carried out a study on well water of 7 villages of Reddigudam Mandal, Krishna district, Andhra Pradesh for physico-chemical and bacteriological examination. Well waters were found to be relatively harder, and sulphates and fluorides almost within permissible limits. The well water exhibited COD/BOD ratio of > 3 in all the villages. Correlation coefficients (R) among various water quality parameters were determined. The water quality index calculated from 11 physico-chemical and 1 biological parameter taken together varied from 82.6 – 254.6 mg/l indicating the

COD
well
water

Index
mg/l



pollution in the well water of Reddigudam Mandal. Thus, the water was unsafe for human use.

Jain & Sharma (1997) analyzed groundwater samples from different villages of Jammu district for various water quality parameters during pre-monsoon and post-monsoon seasons and established correlation coefficients among different parameters. The study revealed significant correlation between conductivity and total dissolved solids, alkalinity, hardness, chloride, nitrate, sulphate, sodium, and potassium, although the quality of groundwater varied significantly. Linear regression equations were also developed for the constituents having significant correlation coefficients.

Mary et al. (1998) analyzed the physico-chemical characteristics of wastewater samples collected at different points from carbonization plant, NLC (Nayveli Lignite corporation). No significant variation in water quality from point to point sample was found. The wastewater quality parameters were compared with the parameters of the raw water used. A correlation analysis carried out among the various parameters resulting in significant linear relationship between electrical conductivity and total dissolved solids.

Tyagi et al. (1998) developed statistical models for the estimation of conductivity for pre-monsoon and post-monsoon seasons using routinely monitored water quality parameter of groundwater wells in Saharanpur District (UP) and Haridwar district (Uttaranchal). Best subset procedure based on R^2 (coefficient of determination) and F (Fisher's test) values were used in model dissemination. The predicted values of conductivity were compared with observed (actual) values and reasonably good matching was obtained.

Rambabu et al. (1998) conducted a systematic study on water quality on the open well water sources of a thickly populated Chirale town, Prakasam

district, a municipal area. In addition, number of textile processing and small scale industries were present in its lower limit. Water samples were periodically bi-monthly collected and analyzed for various better quality parameters, like pH, EC, TDS, TAK, TH and metals like Na, K, Ca, and Mg. The anions Cl, SO₄, NO₃ and F and the other pollution parameter, like dissolved oxygen (DO) were also estimated. Pearson's correlation coefficient and regression analysis was carried out for all the water quality parameters. The study indicated that out of 13 open wells, 7 open wells were polluted. Hence, proper protection of these wells and good sanitary maintenance were recommended.

Jain et al. (1998) attempted to develop statistical models to find out the critical parameters responsible for the salinity in a coastal region of Andhra Pradesh. Best subset procedure based on R and F values was used in model formulation. It was found that chloride, alkalinity, magnesium and sodium could be used as surrogate parameters for the prediction of salinity / conductivity. The predicted and observed values were found to be in good agreement.

Jain and Sharma (2002) carried out a systematic calculation of correlation coefficients among water quality constituents for groundwater samples of Malprabha river basin, Karnataka. The regression equations were developed, and their utility discussed to predict the concentration of water quality constituents having significant correlations with electrical conductivity.

Neumann et al. (2003) developed an empirical model to predict daily maximum stream temperatures for the summer period. The model was developed using a stepwise linear regression procedure to select significant predictors. The predictive model includes a prediction confidence interval to

quantify the uncertainty. The methodology was applied to the Truckee River in California and Nevada. The stepwise procedure selected daily maximum air temperature and average daily flow as the variables to predict maximum daily stream temperature at Reno, Nev. The model was shown to work in a predictive mode by validation using three years of historical data. Using the uncertainty quantification, the amount of required additional flow to meet a target stream temperature with a desired level of confidence was determined.

2.2 Principal Component Analysis

Principal component analysis (PCA) is a multivariate statistical technique. This is a powerful tool used to investigate the chemical relationship between different water quality constituents, and finally dominating constituents may be predicted. This technique is used for reduction of data and decipher patterns within large sets. No constraints such as normality are imposed on data because principal component analysis is based solely upon eigen analysis of the correlation or covariance matrix. The ultimate target of principal component analysis is to describe the majority of the variance in the large data sets or in few principal components, with the remaining unexplained variance consisting of noise. Hidden patterns can then be amplified and the noise discarded.

Dawdy and Feth (1967) applied factor analysis to results of chemical analyses of 103 water samples from wells in the Upper and middle Mojave River valley, San Bernardino County, California. Chemical analyses showed that there were three principal chemical types of water, calcium bicarbonate, sodium sulphate, and sodium chloride, as well as many mixtures of the three. Data were studied by factor analysis to learn the relative importance of each principal ion in determining the variations among the samples, and to examine

the possibility of chemical equilibrium between aqueous and solid phases in the aquifers. Most of the covariance in the system might be accounted for by variances of Ca^{+2} , Mg^{+2} , Na^{+1} , SO_4^{-2} , and Cl^{-1} . There was almost identical loading on the constituents Na^{+1} and Cl^{-1} . The variance in chemical composition of the hydro chemical system was governed largely sources of sodium chloride. None of the components was controlled by equilibrium between ions by in the water and minerals in the aquifers. Concentrations of NO_3^{-1} , and F^{-1} varied independently of other constituents. Geographic distribution of statistical loadings of the principal constituents at individual wells did not reveal sources of the constituents, which must be deduced from geologic and hydrologic evidence. Factor analysis, however, furnished the critical information on chemical relationships basic to the deduction.

Reid et al. (1980) studied the chemistry of precipitation and river water for one year in Glendya, a 41 Km^2 moorland catchments in north east Scotland. The precipitation was very dilute, weekly acidic and highly variable in composition. River water was much less dilute, neutral, and less variable. Factor analysis was used to investigate the controls on water chemistry. This suggested three main processes affecting precipitation aerosols of oceanic spray, which affected sodium, magnesium, chloride and total organic carbon (TOC) concentrations, emission of gaseous sulphur and nitrogen oxides from industrial processes and the burning of fossil fuels, which affected pH, wind-below terrestrial dust. The factor affecting river water was quite different. The first factor represents cation exchange and weathering reactions in the soil and affected calcium, magnesium, sodium, bicarbonate and silicon concentration. The second factor affected the concentrations of the iron, TOC, Manganese and aluminum and represents the translocation of these elements

down the soil profile and into the river at times of the high discharge. The third factor affected the concentration of the chromium and nitrate and reflects nitrogen demand and mineralisation in the soil. Phosphate, sulphate, potassium and chloride appeared to vary independently, but low variability in river water compared with precipitation was apparent. The chemistry of river water from the catchments was also investigated during two storm events, and the results report the grouping of the variable produced by factor analysis. The chemistry of the river water was thus controlled by process in the soil, suggesting that nearly all the river water originates within the soil, and that direct surface runoff was of minor importance.

Puckett and Bricker (1992) studied the factor controlling the chemistry of 69 low-order streams in the Blue Ridge and Valley and Ridge physiographic provinces of Virginia and Maryland over a 13-month period. Principal component analysis was used to examine regional patterns in stream chemistry and to examine the degree to which the chemistry of low-order streams was controlled by the bedrock upon which they flow. Streams clustered into regionally isolated groups, strongly related to bedrock type, with SO_4^{2-} and HCO_3^- the chemical variables of most importance. Sulphate concentrations appear to be strongly controlled by climate and hydrology, and sorption in the soils within the watershed. Much of the atmospherically derived SO_4^{2-} accumulated in watersheds during the growing season and flushed out later. Weathering reactions were found to be particularly important in the production of HCO_3^- , accounting for 91 percent on an annual basis, and export of divalent cations from these watersheds, accounting for 48-50 percent on an annual basis. About half of non-anthropogenic Na^+ was derived from weathering of silicates, whereas nearly all K^+ was identified with leaching

by SO_4^{2-} . Water chemistry was strongly related to the rock type in the watershed and the weatherability of the component minerals. Rock type was not a randomly distributed function. Instead, it was controlled by geologic factors that result in clusters of similar rock type in a given region. When planning large synoptic studies, it is extremely important to consider that a sampling scheme based on random sampling of a non-randomly distributed function may not provide the most accurate representation of the variables of interest. A hierarchical sampling scheme may rather be indicated. This study suggested that although one sample in time might be sufficient to characterize the primary geochemical factors controlling stream chemistry throughout the year, it was sufficient to detect subtle, flow-related alterations in chemistry.

Chakrapani and Subramanian (1993) applied the multivariate analysis to the sediment composition and concluded that metals have been grouped into different factors depending upon their source of origin.

Subramanian and Balasubramanian (1994) applied the principal component analysis to identify the dominant ions which characterized groundwater chemistry of Tiruchandwe Coast, Tamil Nadu.

Vajrappa and Srinivas (1994) used factor analysis to identify dominant factors responsible for variation in the hydrochemistry of the Kabri river basin in Karnataka.

Tizro (1995) attempted principal component analysis and factor analysis for assessing the chemical characteristics of groundwater of Mahendragarh district, Haryana using a computer programme given by Davis, (1973). The experimental ionic values were utilized in these computations. Principal component analysis was carried out taking in account eight variables, namely Na^+ , K^+ , Mg^{++} , Ca^{++} , Cl^- , SO_4^- , HCO_3^- , and CO_3^- for the

shallow groundwater for April 1991, January 1992 and May 1993 and deep groundwater of May, 1993. Among the factors mentioned above, four factors accounted for over 97% percent of the total variance for the shallow groundwater of 1991. The first eigen value (3.47) corresponded to the largest factor, which accounted for 43.41% of total variance and was highly loaded with K^+ , HCO_3^{-2} , SO_4^{-2} , CO_3^{-2} and Ca^{+2} . The second higher eigen value (3.023) corresponded to 37.79% and was highly loaded with Cl^- , Mg^{+2} followed with Na^+ ions. The third eigen value (1.04) corresponding to 13.02% of trace, was loaded with Ca^{+2} only. Similarly, among the eight factors in groundwater sample of Jan, 1992, four factors accounted for over 95% percent of total variance. It was observed that the first eigen value (4.44) corresponded with 55.66 percentage of trace and was loaded with Cl^- , SO_4^{-2} , Ca^{-2} , Mg^{+2} , Na^+ , and K^+ . The second value (1.6673) corresponded to 20.84 percent of trace and was loaded with HCO_3^- only. The third eigen value (1.141) accounted for 14.26% and it was loaded with CO_3^{-2} . In case of the factor loading for the shallow groundwater in 1993, the first eigen value (4.9793, 62.24% of trace) was highly loaded with Cl^- , Na^+ , Ca^{+2} , SO_4^{-2} , and K^+ , and negatively loaded with CO_3^- . The second eigen value (1.5895 or 19.9%) was loaded with HCO_3^- and Mg^{+2} . For deep groundwater of 1993 the first eigen value (3.81 or 47.72% of trace) was highly loaded with K^+ , Mg^{+2} , Ca^{+2} , Na^{+2} , SO_4^{-2} and Cl^- . The second eigen value (1.6544 or 20.64% of trace) was negatively loaded with HCO_3^- , Cl^- , Na^+ , SO_4^{-2} .

Nolan et al. (1995) attempted to statistically verify the material risk map of the United States with groundwater nitrate data collected by National Water – Quality Assessment (NAWQA) Program during 1993-1995 and inferred mechanism by which nitrate concentration in groundwater of the Southeastern

United States was attenuated. A principal component analysis was performed and a "nitrate reduction" component explained 23% of the total variance and indicated that dissolved oxygen and nitrate were inversely related to ammonium, iron, manganese, and dissolved organic carbon. Additional component extracted by principal component analysis included "calcite – dissolution" (18 percent to variances explained) and "phosphate – dissolution" (9 percent of variance explained). Together, the three principal components explained 50% of the total variation in the data subset representing the South East.

Evans et al. (1995) used factor analysis to investigate processes controlling the chemical composition of four streams in the Adirondack Mountains, New York. Four streams were monitored intensively over a two-year period. Factor analysis was used to identify interrelationships between dissolved species during this period, and to determine physical processes controlling their behaviour. Analysis of the full data set identified species which varied predominantly on an episodic timescale, and species which were subject to seasonal cycles. Two-month subsets of data were defined to remove the influence of seasonal cycles, and factor analysis of individual subsets then allowed episodic behaviour to be examined for each 2-month period. Results showed that base cation dilution was a consistent cause of change in acid neutralization capacity (ANC) in all four streams. NO_3^- exhibits strong seasonality in concentration and also in episode behavior, increasing during winter – snowmelt episodes, but decreasing during some summer episodes. DOC concentration also varied seasonally, but 2-month analysis indicated episodic increases during all periods, SO_4^- did not exhibit consistent episodic behavior, as it was strongly influenced by antecedent conditions.

Behavior of Ca^{++} , Mg^{++} , was apparently influenced by a significant soil source in three of the streams.

St-Hilarie et al. (2004) used multivariate analysis of water quality in the Richibucto Drainage Basin (New Brunswick, Canada). Specific conductivity, pH, dissolved oxygen, carbon, phosphorus, and nitrogen species were measured at 36 stations in the Richibucto river drainage basin, including the estuary, in New Brunswick, Canada, over the six year period 1996 through 2001. Each station was sampled between 1 and 26 times (mean = 7.5, standard deviation = 6.0) during the ice free seasons without regard to tide. There was significant variance among stations in most parameters. Principal component analysis (PCA) was used to identify the processes explaining the observed variance in water quality. Because of the high variability in specific conductance, stations were first grouped in a fresh water subset and an estuarine (brackish water) subset. For fresh water stations, most of the variance in water quality was explained by pH and total organic carbon, as well as high nutrient concentrations. These high nutrient concentrations, along with water salinity, which varied with flow and tides, were also important in determining water quality variability in brackish water. It was recommended that water quality parameters that were found to explain most of the variance by principal component analysis be monitored more closely, as they formed key elements in understanding the variability in water quality in the Richibucto drainage basin. Cluster analysis showed that high phosphorus and nitrate concentrations were mostly found in areas of peak runoff, tributaries receiving treated municipal effluent, and lentic zones upstream of culverts. Peak runoff even from a harvested area was also shown to be acidic.

2.3 Artificial Neural Network (ANN)

On account of the unique structure in which the neurons are arranged and operate, humans are able to quickly recognize patterns, process data, and learn from past experiences. Artificial Neural Networks (ANNs) refer to computing systems whose central theme is borrowed from the analogy of biological neural networks. ANNs represent highly simplified mathematical models of our understanding of the biological neural networks. They include the ability to learn and generalize from examples to produce meaningful solutions to problems even when input data contains error or are incomplete, and to adapt solutions over time to compensate for changing circumstances and to process information rapidly.

During the last decade, ANNs have become very popular and have been applied to a wide range of problems. In the water sector also, ANNs have found applications in a range of problems dealing with surface water, groundwater, management of water resources systems, water quality, and so on. A large number of studies have been completed in which ANNs have been successfully applied to field problems related to water resources. The results of these studies confirm that ANNs are a versatile alternative to the conventional modeling techniques.

Agarwal and Singh (2003) developed multi-layer back propagation artificial neural network (BPANN) models to simulate rainfall runoff process for two sub-basins of Narmada river (India) viz. Banjar up to Hridnagar and Narmada upto Manot considering three time scales viz. Weekly, ten-daily and monthly with variable and uncertain data sets. The BPANN runoff models were developed using gradient descent optimization technique, and generalized through cross-validation. In almost all cases, the BPANN

developed with the data having relatively high variability and uncertainty learned in less number of iterations, with high generalization. Performance of BPANN models was compared with the developed linear transfer function (LTF) model and found to be superior.

Batisha (2004) attempted water quality sensing using multi-layer perception artificial neural networks. Traditional methods for classifying high volumes of such data into large numbers of classes based on statistical parametric methods often do not give sufficient descriptive accuracy for discriminating the numbers of classes required. The use of multiplayer perception neural networks as new method for solving this problem for realistic operational purposes had been established. The multiplayer perception offered a good classification method and completed well with the traditional techniques used in statistical methods. Induced by using reasonably large network architectures, the method seemed to work quite well with large number of classes that is where problems were normally encountered with the traditional parametric methods.

Bowden et al (2004) attempted forecasting chlorine residuals in a water distribution system using a general regression Neural Network. In a water distribution system (WDS), chlorine disinfections is important in preventing the spread of water borne diseases. By strictly controlling residual chlorine throughout the WDS, water quality managers could ensure the satisfaction and safety of their customers. However, due to the travel time of water between the chlorine dosing point and any strategic monitoring points, water treatment plant (WTP) operators often receive, information too late for their responses to be effective. Given the ability to forecast the chlorine residual at strategic points in or WDS, it would be possible to have superior control over

the chlorine dose, thereby preventing incidents of under-and over-chlorination. A general regression neural network (GRNN) was been developed for forecasting chlorine residuals in the Myponga Water Distribution System (WDS) to the South of Adelaide, South Australia, 24 hours in advance. A number of critical model issues were addressed including selection of an appropriate forecasting horizon; division of the available data into subsets for modeling, and the determination of the inputs there are relevant to the chlorine forecasts. In order to determine if the GRNN was able to capture any non-linear relationships that might be present in the data set, a comparison was made between the GRNN model and a multiple linear regression (MLR) model. When tested on independent validation set of data, the GRNN models were able to forecast chlorine levels to a high level of accuracy, up to 24 hours in advance the GRNN also significantly outperformed the MLR model, thereby providing evidence for the existence of non-linear relationships in the data set.

Jha and Jain (2005) investigated the use of ANN technique in modeling the complex rainfall – runoff process in a large watershed Kentucky River basin, USA. In addition, three different normalization methods were investigated as the pre-processing tools. The results obtained in this study indicated that the performance of an ANN rainfall – runoff model depended on the normalization method adopted. A normalization method that employs only one parameter was recommended for use in ANN model development due to its insensitiveness on the ANN model performance.

Raghuwanshi et al. (2005) developed ANN model to forecast stream discharge at Jamatara gauging site of Ajay river basin in Jharkhand at three different lead times of 3h, 6 h and 9h using hourly rainfall data at Jamtara

gauging site and hourly stream discharge data at Jamtara as well as three upstream gauging sites at Sweath, Dhakwa and Ghesko. The performance of the developed ANN model was evaluated using three different error functions, viz., root mean square error, Nash – Sutcliffe coefficient and percentage deviations in peak discharge. It was found that the developed ANN model for forecasting floods at Jamtara gauging site of Ajay river basin performed very well for 3 h and 6 h lead time.

Sarkar et al. (2005) developed back propagation artificial neural network (BPANN) runoff models using the steepest descent optimization technique to simulate and forecast daily runoff for a part of the Satluj basin of India. ANN models had been developed considering a fixed number of iterations as 5000 and verified on data not considered for calibration. The input variables considered for different model structures were identified through correlation analysis. Based on the statistical performance evaluation criteria such as root mean square error (RMSE), correlation coefficient (CC), coefficient of efficiency (E) and volumetric error (EV), it was observed that only rainfall and temperature, considered as inputs, were not adequate to develop a model for the simulation as well as forecasting of the catchment runoff resulting from rainfall and snowmelt contribution. In order to improve upon the performance of the models, the runoff of the upstream data was also included as an additional input to the model.

Kothyari and Jain (2005) made an approach for modeling monthly runoff using artificial neural network (ANN). The modeling was performed by coupling an auxiliary model for monthly runoff with an ANN. Data from different sub-catchments of the Barakar basin in India were stacked together for model application. The study demonstrated that the approach adopted

therein for modeling produced reasonably satisfactory results for data from catchments with varying characteristics.

Kumar et al. (2005) developed Artificial Neural Network (ANN) models for short term forecasting for Jamtara stream flow gauging site of Ajay river basin (lying in Jharkhand). Seven flood events were considered for model development to provide forecast for different lead times (6 hours, 9 hours, 12 hours) using previous runoff values. The computed and observed flood hydrographs of various lead times was evaluated in terms of Root Mean Square Error for each flood hydrograph. The study also used the alternative evaluation measures such as percentage errors in the peak flows and time to peak to examine specific performance of the ANN based flood forecasting models.

STUDY AREA DESCRIPTION

3.1 General

The study area selected is district Jaipur of Rajasthan, India (Fig 3.1-3.3). Thirty-eight samples from district Jaipur were collected during May 2002 and November 2003 respectively i.e., for pre-monsoon and post monsoon seasons. The samples were collected from different sources viz, handpumps, open wells, and tube-wells, which are being extensively used for drinking and other domestic purposes. The physico-chemical analysis was performed following standard methods ⁱⁿ National Institute of Hydrology, Roorkee (India).

18 months
38 samples
(sampling locations)
No.
?

3.2 Location

Jaipur district is situated in northeastern part of the state Rajasthan, India. It is located between 26° 25' and 27° 51' North latitude and 74° 55' and 76° 10' East longitude covering an area of 10878 sq. km. Jaipur district is bounded by Sikar district in north west, Alwar district in North east, Dausa in east, Tonk in south, Ajmer in south west and Nagpur in west

Administratively, Jaipur district is a part of Jaipur division and is also the capital of Rajasthan. The district is divided into 13 tehsils namely (i) Amer, (ii) Chomu, (iii) Jamwa Ramgarh (iv) Shahpura, (v) Viratnagar, (vi) Kotputli, (vii) Dudu, (viii) Phagi, (ix) Phulera, (x) Bassi, (xi) Chaksu, (xii) Sanganer, and (xiii) Jaipur. It comprises 13 panchayat samities namely :- (i) Kotputli, (ii) Viratnagar, (iii) Shahpura, (iv) Govindgarh, (v) Amer, (vi) Jamwa Ramgarh, (vii) Sambhar, (viii) Dudu, (ix) Sanganer, (x) Jhotwara, (xi) Bassi, (xii) Phagi, and (xiii) Chaksu.

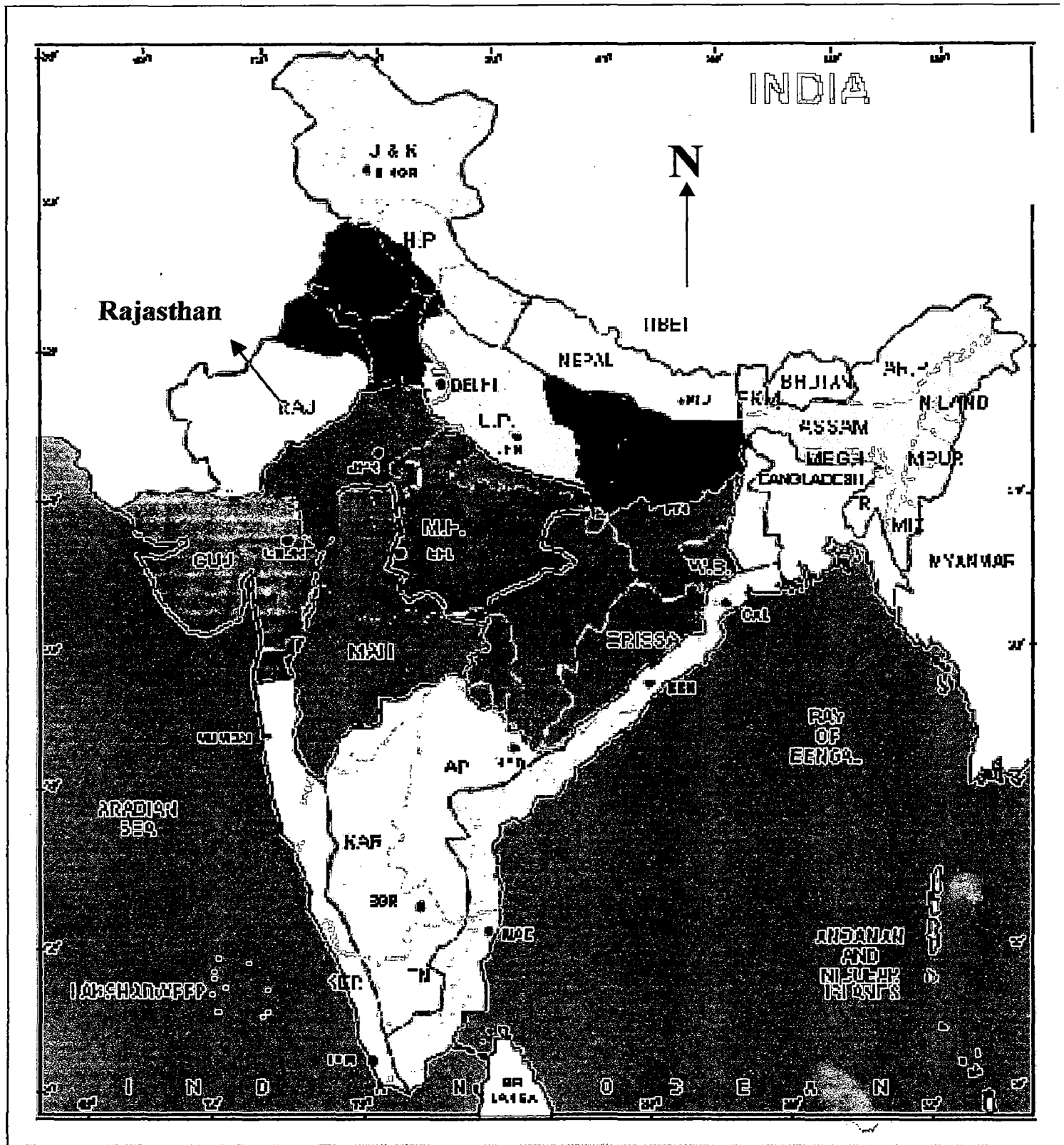


Fig.3.1 Location of Rajasthan in India Map.

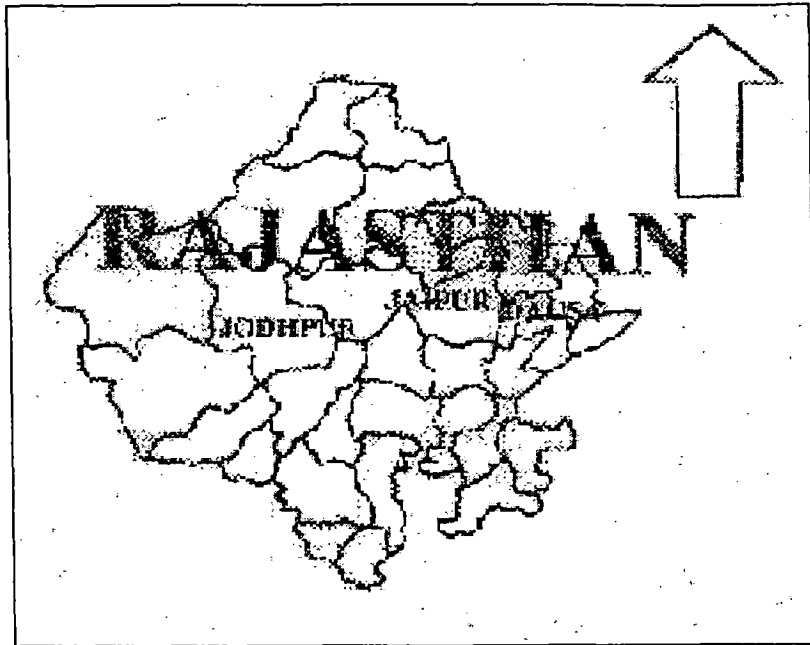


Fig. 3.2: Geographical Map of Rajasthan

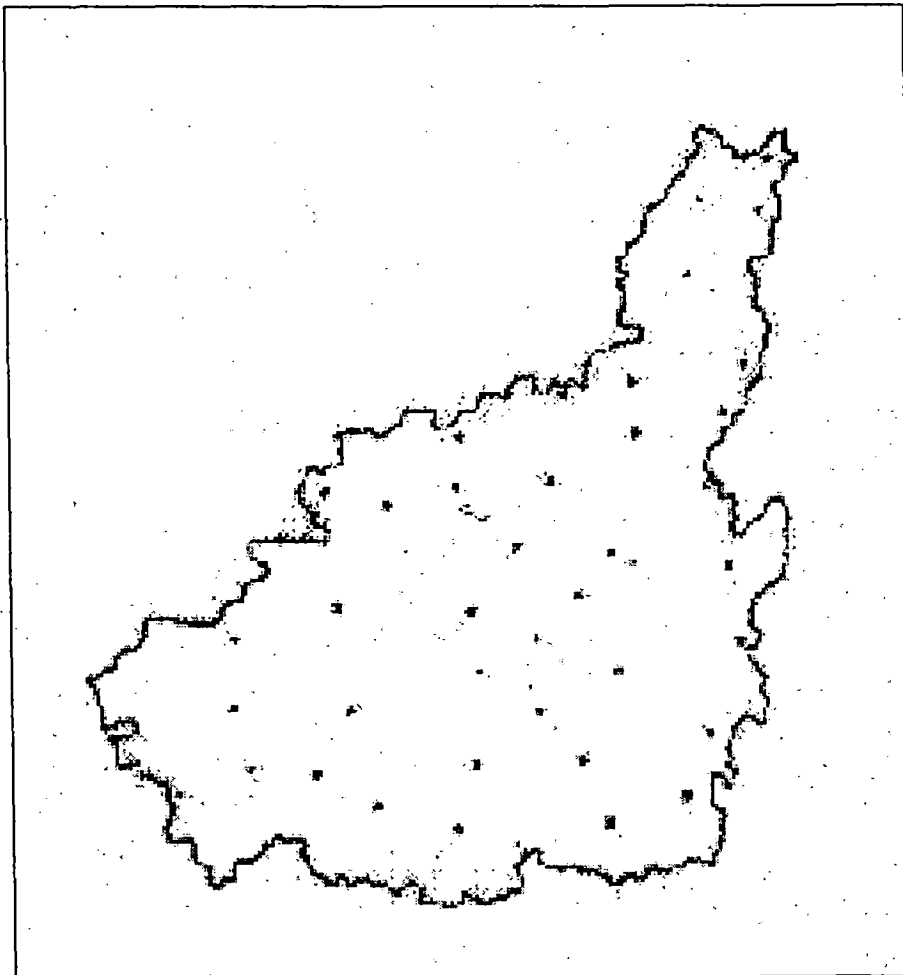


Fig. 3.3: Study area (Jaipur District) showing location of sampling sites.

3.3 Climate

The climate of the district is dry and healthy and is subjected to extremes of cold and heat at various places. The minimum and maximum temperatures are 3°C and 45°C, respectively while the mean temperature is 24°C. In the district rainy season usually from June to September, the normal annual rainfall is 548.2 mm.

3.4 Geology and Mineral

The oldest groups of rock, in the district are schist, gneisses, migmatite and quartzite of Pre-Aravalli, which are considered to be nearly 2,500 million year old. These rocks are covered under a mantle of sand and alluvium of recent to sub-recent age. Overlying these rocks with a major unconformity are the rocks of Delhi super group, which are made up of Rialo, Alwar, and Ajabgarh groups. The rocks of Rialo comprises mainly dolomitic marble and minor quartzite. The Alwar group consists of conglomeratic quartzites and schist which either lie unconfirmably over the Rialo or directly over the metamorphic of Pre-Aravalli. The Ajabgarh group is mainly made up of schist, phyllites, pegmatites and quartz veins.

A variety of mineral deposits found in the district are Chinaclay in Buchara & Torda, Copper near Gol, Badshahpur, Chanla and Chatigodlyana area, Iron ore is Moriza, Bonai etc. Cement grade limestone near Kotpulti and Maonda & impure lime stone at Nimala, Dabla etc. Silica sand is found at Banskhop and Jir hills. Soapstone occurs in Dogetha, Jharna, Geejgarh, Khawa etc.

3.5 Physiography and Soil

A large part of district is covered by thick mantle of soil, brown sand and alluvium in eastern and northern area is occupied by hills range and belong to Aravalli system and are known by different names at different places, the longest range starting from Sambhar lake in this district crosses over upto Singhana in the district of Jhunjhunun.

The district is drained by a number of largely non-perennial rivers of which Banganga and Sabi are important ones. The Banganga has been impounded near Jamwa Ramgarh which provides a major share of drinking water supplies to Jaipur city. A large area of the district has been affected by sand encroachment through wind gaps and river valleys.

The soils of the district are greyish brown-to-brown and yellowish brown, light to medium textured and deep to very deep. These soils can be classified in Entisols order by 7th approximation classification, some soils belong to Aridisols order.

About 4.06 percent of the total area of the district is under forest. Subsidiary edaphic type of dry tropical forests are found in the district. The total area under forest is about 44239 hectares.

STATISTICAL PROCEDURE FOR MODEL DEVELOPMENT

4.1 Multiple Linear Regression

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables as a dependent or criterion variable. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of" The general computational problem that needs to be solved in multiple regression analysis is to fit a straight line to a number of points. In the simplest case- one dependent and one independent variable one can visualize this in scatter plot.

4.1.1 Least squares

In the scatter plot, we have an independent or X variable, and a dependent or Y variable. The goal of linear regression procedures is to fit a straight line through the points. Specifically, a line is computed so that the squared deviations of the observed points from that line are minimized. Thus, this general procedure is sometimes also referred to as least squares estimation.

4.1.2 The regression equation

A line in a two dimensional or two- variable space is defined by the equation $Y=a+b \cdot X_j$, the Y variable can be expressed in terms of a constant (a) and a slope (b) times the X variable. The constant is also referred to as the intercept, and the slope (b) as the regression coefficient or B coefficient. In the multivariate case, when there are more than one independent variable, the

regression line cannot be visualized in the two dimensional space, that can be computed just as easily. In general, multiply regression procedures estimates a linear equation of the form:

$$y = a + b_1 * X_1 + b_2 * x_2 + \dots + b_n * X_n$$

4.1.3 Unique prediction and partial correlation

It may be noted from the above equation that the regression coefficients (or B coefficients) represent the independent contributions of each independent variable to the prediction of the dependent variable. Another way to express this fact is to say that, for example, variable x_1 is correlated with the Y variable, after controlling for all other independent variables. This type of correlation is also referred to as a partial correlation.

4.1.4 Predicted and residual scores

The regression line expresses the best prediction of the dependent variable (Y), given the independent variables (X). However, the nature is rarely (if ever) perfectly predictable, and usually there is substantial variation of the observed points around the fitted regression line. The deviation of a particular point from the regression line (its predicted value) is called the residual value. Residuals are calculated by working out 'expected' values of Y by applying the regression equation to the actual values of X and then subtracting each expected Y value from its corresponding actual Y value. Where the value of Y predicted by the equation is less than the actual value of Y, the residual is positive. Negative residuals result from cases where Y is predicted higher than it actually is.

4.1.5 Residual variance and R-square

The smaller the variability of the residual values around the regression line relative to the overall variability, the better is the prediction. For example, if there is no relationship between the X and Y variables, then the ratio of the residual variability of the Y variable to the original variance is equal to 1.0. If X and Y are perfectly related then there is no residual variance and the ratio of variance would be 0.0. In most cases, the ratio would fall somewhere between these extremes, that is, between 0.0 and 1.0. 1.0 minus this ratio is referred to as R-square or the coefficient of determination. This value is immediately interpretable in the following manner. If we have an R-square of 0.4 then we know that the variability of the Y-values around the regression line is 1.0-0.4 times the original variance. In other words, we have explained 40 percent of the original variability, and are left with 60 percent residual variability. Ideally, we would like to explain most part of it, if not all of the original variability. The R-square value is an indicator of how well the model fits the data (e.g., an R-square close to 1.0 indicates that we have accounted for almost all of the variability with the variables specified in the model.).

4.1.6 Interpreting the correlation coefficient R

Customarily, the degree to which two or more predictors (independent or X variables) are related to the dependent (Y) variable is expressed in the correlation coefficient R, which is the square root of R-square. To interpret the direction of the relationship between variables, one looks at the signs (plus or minus) of the regression or B coefficients. If a B-coefficient is positive, then the relationship of this variable with the dependent variable is positive; if the

B-coefficient is negative then the relationship is negative. Of course, if the B-coefficient is equal to zero then there is no relationship between the variables.

4.1.7 Assumptions, limitations, and practical considerations

(a) Assumption of linearity

First of all, as is evident in the name multiple linear regression, it is assumed that the relationship between variables are linear. In practice this assumption can virtually never be confirmed. Fortunately, multiple regression procedures are not greatly affected by minor deviations from this assumption. However, as a rule it is prudent to always look at bivariate scatter plot of the variables of interest. If curvature in the relationships is evident, one may consider either transforming the variables, or explicitly allowing for non-linear components.

(b) Normality assumption

It is assumed in multiple regression that the residuals (predicted minus observed values) are distributed normally (i.e., follow the normal distribution). Again, even though most tests (specifically the F-test) are quite robust with regard to violations of this assumption, it is always a good idea, before drawing final conclusions, to review the distributions of the major variables of interest.

(c) Limitations

The major conceptual limitation of all regression techniques is that one only ascertains relationships, but never be sure about underlying casual mechanism. For example, bronchitis rates correlate positively with population density, but there is no direct casual relationship between them. The correlation exists because they are both related to air pollution.

(d) Choice of the number of variables

Multiple regressions is a seductive technique: "Plug in" as many predictor variables as one can think of and usually at least a few of them will come out significant. This is because one is capitalizing on chance when simply including as many variables as one can think of as predictors of some other variables of interest. This problem is compounded when, in addition, the number of observations is relatively low. Most researchers recommend that there should be at least 10 to 20 times as many observations as one has variables, otherwise the estimates of the regression line are probably very unstable and unlikely to replicate if one were to do the study over.

(e) Multi-collinearity and matrix III-conditioning

This is a common problem in many correlation analyses. When there are many variables involved, it is often not immediately apparent that this problem exists, and it may only manifest itself after several variables have already been entered into the regression equation. Nevertheless, when this problem occurs it means that at least one of the predictor variables is (practically) completely redundant with other predictors. There are many statistical indicators of this type of redundancy as well as some remedies (e.g., Ridge regression).

(f) Fitting centered polynomial models

The fitting of higher-order polynomials of an independent variable with a mean not equal to zero can create difficult multicollinearity problems. With large numbers, this problem is very serious, and if proper protections are not put in place, can cause wrong results! The solution is to "center" the independent variable (sometimes, this procedure is referred to as "centered

polynomials”), i.e., to subtract the mean, and then to compute the polynomials.

(g) The importance of residual analysis

Even though most assumptions of multiple regressions cannot be tested explicitly, gross violations can be detected and should be dealt with appropriately. In particular outliers (i.e., extreme cases) can seriously bias the results by “pulling” or “pushing” the regression line in a particular direction, thereby leading to biased regression coefficients. Often, excluding just a single extreme case can yield a completely different set of results.

4.2 Formulation of Models:

The general representation of statistical models may be given by

$$Y_i = \sum_{j=0}^k \beta_j X_{ij} + \varepsilon \quad (4.1)$$

with $x_{i0} = 1$. Here, x_{ij} is the independent variable for the i th observation (various water quality constituents in the present study), Y_i is the dependent variable for the i th observation, β_j is unknown coefficients to be estimated, k is the number of coefficient (to be estimated) in the model, and ε is the error in the determination of Y_i which is generally assumed as having zero mean and constant standard deviation σ .

The unknown coefficients (β) are estimated by least squares method because here no assumption is necessary on the probability distribution of data due to its simplicity. Hence for finding out the coefficients of various water quality constituents in the model, least square method is used for prediction of sodium. This method has been used frequently by various authors (Draper and Smith, 1981; Weisberg 1980). Regression analysis was

performed on pre-monsoon and post-monsoon data sets. Initially preliminary analysis of data was carried out before starting actual statistical regression analysis.

4.2.1 Preliminary analysis of data

The preliminary analysis consists of

- (i) Initial filtration of data
- (ii) Partial visual inspection of the data files
- (iii) Creation of scatter plots.

If any outlier is detected from the scatter plot by taking all the water quality parameters as independent variables and sodium as a dependent variable, those were removed.

4.2.2 Secondary analysis

The filtered data for the two data sets obtained after preliminary analysis is used to find correlation matrices predicting correlation of each water quality constituent with sodium. To enhance the visualization of the correlation matrix the square of correlation coefficient (R^2) or coefficient of determination is calculated to indicate the contribution of individual water quality parameters in explaining the variation in the dependent variable respectively. Since pH, nitrate, phosphate and potassium had no significant correlation with sodium for pre-monsoon and post-monsoon data sets, respectively these parameters were excluded from model formulation.

4.2.3 Selection of independent variables for regression analysis

If more number of independent variables, as possible are used then in that case reliable fitted values can be determined and model prediction will be more accurate. Moreover, since R^2 gives the proportion of the variation in the

dependent variables that is explained by the fitted regression model, one obviously wants R^2 to be large. But on the other hand, because of the costs involved in obtaining information on a large number of independent variable and subsequently monitoring them, there is interest in including as few independent variables as possible. Thus one has to make compromise between these extremes i.e. for selecting the best regression variables and thereby the best model. There is no unique statistical procedure for doing this (Draper and Smith, 1981). Different researchers suggested different statistical procedures namely: backward elimination, all possible regression, ridge regression, forward elimination in stepwise regression, principal component regression, and stagewise regression which may help information of optimum model.

In the present study, attempt has been made to use the best subset regression approach to select the best set of independent variables.

4.2.4 Best subset regression

Using the R^2 information i.e., the proportion of variation explained in the dependent variable, different best subsets of independent variables could be selected. The regression was assessed for each subset according to:

The value of R^2 achieved,

The F value (given in equation 4.3), and

The number of observations used in developing the model.

The model obtained from large dataset and achieving higher values of R^2 and F value will always be preferred.

(a) R^2 criterion

The square of the multiple correlation coefficients R^2 is defined as

$$R^2 = \frac{SSR}{SS_Y} = 1 - \left(\frac{SSE}{SS_Y} \right) \quad (4.2)$$

with $SS_Y =$ sum of squares about the mean $= \sum (Y_i - \bar{y})^2$

$SSE =$ sum of squares about regression $= \sum (y_i - \hat{y}_i)^2$

$SSR =$ sum of squares due to regression $= \sum (\hat{y}_i - \bar{y})^2$

i.e., $SS_Y = SSE + SSR$

where, \bar{Y} is the average value of dependent variable, and

\hat{y}_i is the model computed values of the dependent variable.

The stronger the linear association between Y_i and \hat{y}_i , it will yield a large value of R^2 and vice-versa. Unfortunately, whenever comparing a subset model to a large model including the subset, R^2 provides an inadequate criterion for subset model selection because large model will always have an R^2 value larger than that for the subset model. However, for a fixed number of independent variables (equal to k) R^2 can be used to compare different models with a large value of R^2 indicating the preferred model.

(b) F value criterion

The value of F is mathematically expressed as

$$F = \left(\frac{N-K-1}{K} \right) \left(\frac{R^2}{1-R^2} \right) \quad (4.3)$$

where, $R^2 =$ explained variation of Y_i ,

$(1-R^2) =$ unexplained variation of Y_i ,

$N =$ number of data points, and

$K =$ number of independent variables.

From equation (4.3), it is evident that F value which is the ratio of the explained to the unexplained variation in Y_i , will be large when the proportion

of explained variation in Y_i will be large. The regression will be significant when the F value is large.

The F – statistic can also be used to compare any two models as long as all the independent variables in the smaller model are also included in the large model, that is small model is a subset model of the large model. The residual sum of squares describes unexplained variation in the dependent variable by the model. If the independent variables, which are important, then dropping these from the subset model, should result in a significant increase in unexplained variation of Y_i , that is SSE_r , should become considerably large than SSE_f . Using this idea, a simple test statistic proposed by Weisberg (1980) can be expressed as

$$F_{k-m, n-k-1} = \frac{(SSE_r - SSE_f) / (K - m)}{SSE_f / (n - k - 1)} \quad (4.4)$$

where, SSE_f = Residual error sum of squares of full model containing K independent variables;

SSE_r = Residual error sum of squares of the subset model containing (K-m) independent variables.

m = Number of independent variables dropped from the full model;

K = Number of independent variables.

The larger model will be preferred when the $F_{k-m, n-K-1}$ statistic is sufficiently large. One reasonable rule should be to prefer the full model if

$$F_{k-m, n-K-1} > F^*$$

where, F^* is the $\alpha \times 100\%$ point of the $F_{k-m, n-K-1}$ distribution. The choice of $\alpha = 0.05$ is typical.

4.3 Principal Component Analysis

Principal component analysis (PCA) is a classical statistical method. This linear transform has been widely used in data analysis and compression. Principal component analysis is based on the statistical representation of a random variable.

Suppose we have a random vector population X ,

where, $X = (x_1, x_2, x_3, \dots, x_n)^T$

and the mean of that population is denoted by

$$\mu_x = E(x)$$

and the co-variance matrix of the same dataset is

$$C_x = E \{ (x - \mu_x)(x - \mu_x)^T \}$$

The components of C_x , denoted by C_{ij} represent the covariances between the random variable components x_i and x_j . The component C_{ii} is the variance of the component x_i . The variance of a component indicates the spread of the component values around its mean value. If two components x_i and x_j of the data are uncorrelated, their covariance is zero ($C_{ij} = C_{ji} = 0$). The covariance matrix is, by definition, always symmetric.

From a symmetric matrix such as the covariance matrix, we can calculate an orthogonal basis by finding its eigen values and eigen vectors. The eigen vectors e_i and the corresponding eigen values λ_i are the solutions of the equation

$$C_x e_i = \lambda_i e_i, \quad i = 1, \dots, n$$

For simplicity we assume that the λ_i are distinct. These values can be found, for example, by finding the solutions of the characteristic equation:

$$|C_x - \lambda I| = 0 \tag{4.5}$$

where, I is the identity matrix having the same order as C_x , and the $||$ denotes the determinant of the matrix.

If the data vector has n components, the characteristic equation becomes of order n . By ordering the eigen vectors in the order of descending eigen values (largest first), one can create an ordered orthogonal basis with the first eigen vector having the direction of largest variance of the data. In this way, we can find directions in which the dataset has the most significant amounts of energy.

Suppose one has a dataset of which the sample mean and the covariance matrix have been calculated. Let A be a matrix consisting of eigen vectors of the covariance matrix as the row vectors. By transforming a data vector x , we get

$$y = A(x - \mu_x) \quad (4.6)$$

which is a point in the orthogonal coordinate system defined by the eigen vectors. Components of y can be seen as the coordinates in the orthogonal base.

We can construct the original data vector x from Y by

$$x = A^T y + \mu_x \quad (4.7)$$

Using the property of an orthogonal matrix,

$$A^{-1} = A^T$$

The A^T is the transpose of the matrix A . The original vector X was projected on the coordinate area defined by the orthogonal basis. The original vector was then reconstructed by a linear combination of the orthogonal basis vectors.

Instead of using all the eigen vectors of the covariance matrix, we may represent the data in terms of only a few basis vectors of the orthogonal basis. If we denote the matrix having the K first eigen vectors as rows by A_K , we can create a similar transformation as seen above.

$$\begin{aligned} y &= A_K (X - \mu_x) \\ \Rightarrow x &= A_K^T \cdot y + \mu_x \end{aligned} \tag{4.8}$$

This means that we project the original data vector on the coordinates areas having the dimension K and transforming the vector back by a linear combination of the basis vectors. This minimises the mean – square error between the data and this representation with given number of eigen vectors.

If the data is concentrated in a linear subspace, this provides a way to compress data without losing much information and simplifying the representation. By picking the eigen vectors having the largest eigen values we lose as little information as possible in the mean – square sense. One can choose a fixed number of eigen vectors and their respective eigen values and get a consistent representation, or abstraction of the data. This preserves a varying amount of energy of the original data. Alternatively, we can choose approximately the same amount of energy and a varying amount of eigen vectors and their respective eigen values. This would in turn give approximately consistent amount of information at the expense of varying representations with regard to the dimension of the subspace.

We are here faced with contradictory goals. On one hand, we should simplify the problem by reducing the dimension of the representation. On the other hand we want to preserve as much as possible of the original

information content. PCA offers a convenient way to control the trade-off between losing information and simplifying the problem at hand.

4.3.1 Basic idea of factor analysis as a data reduction method

The main applications of factor analytic techniques are: (1) to reduce the number of variables and (2) to detect structure in the relationships between variables, that is to classify variables. Therefore, factor analysis is applied as a data reduction or structure detection method.

(a) *Combining two variables into a single factor*

One can summarize the correlation between two variables in a scatter plot. A regression line can then be fitted that represents the "best" summary of the linear relationship between the variables. If we could define a variable that would approximate the regression line in such a plot, then that variable would capture most of the "essence" of the two items. Subject's single scores on that new factor, represented by the regression line, could then be used in future data analyses to represent that essence of the two items. In a sense we have reduced the two variables to one factor. Note that the new factor is actually a linear combination of the two variables.

(b) *Extracting principal components*

Basically, the extraction of principal components amounts to a variance maximizing (varimax) rotation of the original variable space. For example, in a scatter plot we can think of the regression line as the original X-axis, rotated so that it approximates the regression line. This type of rotation is called variance maximizing because the criterion for (goal of) the rotation is to maximize the variance (variability) of the "new" variable (factor), while minimizing the variance around the new variable.

(c) *Generalizing the case of multiple variables*

When there are more than two variables, we can think of them as defining a "space", just as two variables defined a plane. Thus, when we have three variables, we could plot a three-dimensional scatter plot, and, again we could fit a plane through the data, however, the logic of rotating the axes so as to maximize the variance of the new factor remains the same.

(d) *Multiple orthogonal factors*

After we have found the line on which the variance is maximal, there remains some variability around this line. In principal components analysis, after the first factor has been extracted, that is, after the first line has been drawn through the data, we continue and define another line that maximizes the remaining variability, and so on. In this manner, consecutive factors are extracted. Because each consecutive factor is defined to maximize the variability that is not captured by the preceding factor, consecutive factors are independent of each other. Put another way, consecutive factors are uncorrelated or are orthogonal to each other.

(e) *How many factors to extract?*

We extract consecutive factors, they account for less and less variability. The decision of when to stop extracting factors basically depends on when there is only very little "random" variability left. The nature of this decision is arbitrary; however, various guidelines have been developed, like the Kaiser criterion, the scree test etc.

4.3.2 Factor analysis as a classification method

(a) Factor loadings

Factors are produced through an eigen value analysis of the correlation matrix, the mathematics of which have been described by Davis (1973). Factors are vectors which lie orthogonal to one another within a multidimensional space defined by the number of variables in the analysis. Unlike the original variables, therefore, the factors are completely uncorrelated with each other. They are described by means of their correlations with (or 'loadings' on) the original variables, and ranked in order of the amount of the total variance they explain. A loading close ± 1.0 indicates a strong relationship between the factor and the variable, a zero loading indicates that two are unrelated.

(b) Rotating the factor structure

In most instances, the first few factors explain the bulk of total variance, and it is possible to exclude the remaining factors from further analysis without significant loss of information. A subjective decision must be made as to how many factors should be retained, for which several methods exist. In this study, factors, which explained at least as much of the total variance as one of the original variables, were retained. The retained factors were then 'rotated' using the Varimax method. Varimax rotation aims to attain 'simple structure', whereby factor loadings approach either ± 1.0 or zero. This aids interpretation, in that as far as possible a given factor either does or does not include a particular variable (Dawdy and Feth, 1967).

4.4 Formulation of Model

The methodology involves (Davis, 1973):

- (i) Standardization of the raw data in terms of zero mean and unit variance and computation of the linear correlation coefficient matrices.
- (ii) Computation of the eigen values and corresponding eigen vectors based on the correlation coefficient matrix.
- (iii) Computations of a set of mutually orthogonal unrotated factor matrix, the element in each factor are referred to as 'factor loading'.
- (iv) Idealization of the factor matrix by rotation of the factors around the origin, so that the loading on a particular factor is made as much close to + 1.0, 0.0 and -1.0 as possible.
- (v) Lastly a set of varimax rotated factor scores is computed for each sample.

ANN PROCEDURE

5.1 General

An artificial neural network (ANN) is a computing paradigm designed to mimic the human brain and nervous system. Neural network (NN) has a big role to play in the field of water sector where complex natural processes dominate. The high degree of empiricism and approximation in the analysis of water quality systems make the use of neural network highly suitable. In other words, when the possibility of representing the complex relationships between various aspects of the processes in terms of physical or conceptual modeling is very remote, the neural network plays an important role.

ANN is an information processing system that uses an approach entirely different from conventional algorithmic programming and roughly replicates the behaviour of a human brain by emulating the operations and connectivity of biological neurons. From a mathematical point of view, it is a complex non-linear function with many parameters that are trained in such a way that the ANN output becomes similar to the measured output on a known data set. ANNs are highly distributed interconnections of adaptive nonlinear processing-elements (PEs) (Figure 5.1). When implemented in digital hardware, the processing-element is a simple sum of products followed by non-linearity. The connection strengths, also called the network weights, can be adapted such that the network's output matches the desired response.

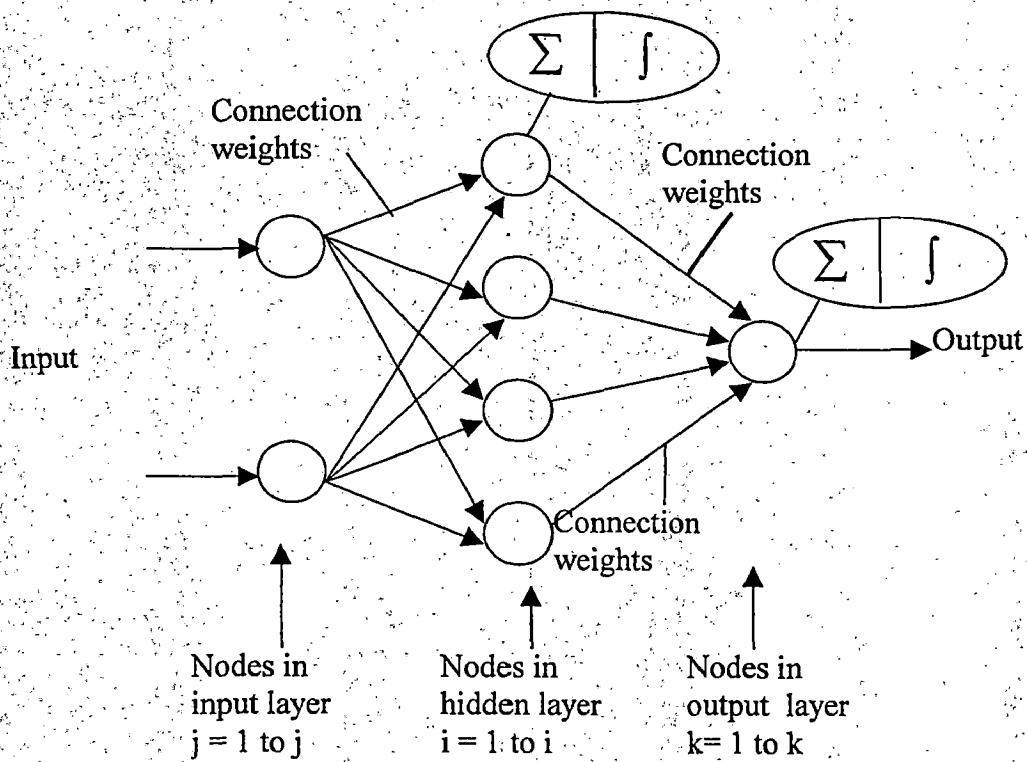


Fig. 5.1: The Building Blocks of ANN

In multi-layered perceptron, hidden layer means third layer of processing elements or units in between the input and output layers that increases computational power. In principle, the hidden layer can be more than one layer. In practice, the number of neurons in this layer is evaluated by trial and error. Hornik et al. (1989) proved that a single hidden layer containing a sufficient number of neurons could be used to approximate any measurable functional relationship between the input data and the output variable to any desired accuracy. In addition, De Villars and Barnard (1993) showed that an ANN comprising of two hidden layers tends to be less accurate than its single hidden layer counterpart.

The ANNs are not exactly the substitute of regression. General regression cannot solve the problems where the input dimension space is high and there is restriction on the number of input data. Regression imposes a priori variable selection, with all the inherent pitfalls, where one is limited to

a few inputs among hundreds available. Regressions are performed using simple dependency functions that are not very realistic. In regression there is only one dependency function over the whole data set, instead of many distinct niches, which is taken care of by ANNs. Where dependency between the input variables and the output are not well-defined, ANNs solve it better. The most important difference between ANN and regression is that the former maps the output by generalization whereas the later by memorization. Generalization refers to the neural network producing reasonable outputs for inputs not encountered during learning. To over-simplify, if an object is represented in a network as a pattern of activation of several units, and if a unit or two responds incorrectly, the overall patterns remain pretty well the same, and the network will still respond correctly to stimuli.

ANNs have been developed as a generalization of mathematical models of neural biology and are based on following rules:

- (i) Information processing occurs at many single elements called nodes, also referred to as units of neurons.
- (ii) Signals are passed between nodes through connection links.
- (iii) Each connection link has an associated weight that represents its connection strength.
- (iv) Each node typically applies a nonlinear transformation called activation function to its net input to determine its output signal.

5.2 The ANN Structure

5.2.1 Biological neuron

A typical biological neuron, comprises of Dendrites, Soma, Axon, and Synaptic Buttons, is shown in Figure (5.2). The dendrites form a very fine

filamentary brush surrounding the body of the neuron. The information is picked up at the Dendrite. The Soma is cell body whereas the Axon is long transmission line like structure and the tail end of the Axon is called Synaptic Buttons. These neurons are so powerful in processing the information, that even a small earthworm with only 302 neuron has a computing power around one thousand times the power of Pentium II. Processor. Thus the computation power of parallel processing in neuro biological system is very high.

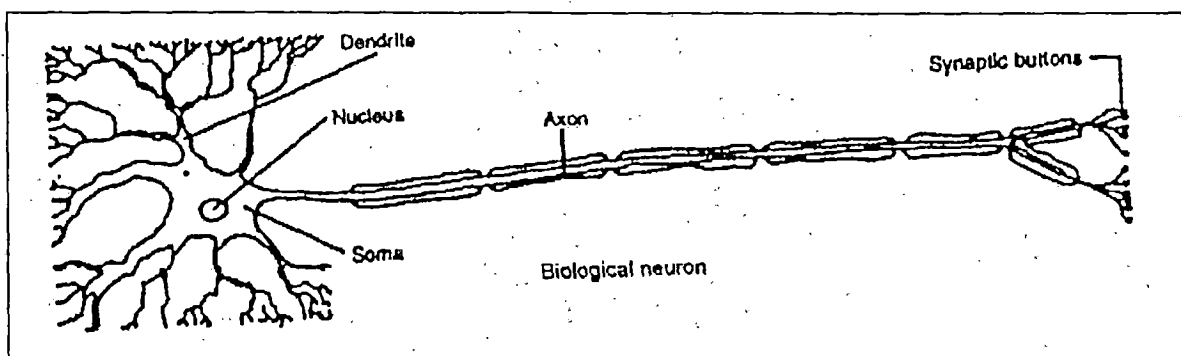


Fig. 5.2: Anatomy of Biological Neuron

5.2.2 Artificial neuron

Let neuron has a set of n inputs $x_1, x_2, x_3, \dots, x_n$ and $w_1, w_2, w_3, \dots, w_n$ are weights attached to the input link. The inputs to the neuron may come from the environment in which it is embedded or outputs to the other neuron are located in. The signals are passed to the cell body through the synapse, which may accelerate or retard. This acceleration or retardation of the input signal is modeled by the weights. Weights are multiplicative factors of the inputs to account for the strength of the synapse. The total output is

$$I = w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n$$

or,
$$I = \sum w_i x_i \quad (5.1)$$

To generate the final output y , the sum is passed through a non-linear fitter ϕ called activation function or transfer function, which releases the output y as.

$$Y = \phi (I) \quad (5.2)$$

The error (E) is calculated at the output as

$$E = \frac{1}{2} \sum [(y_{obs} - y_{est})^2] \quad (5.3)$$

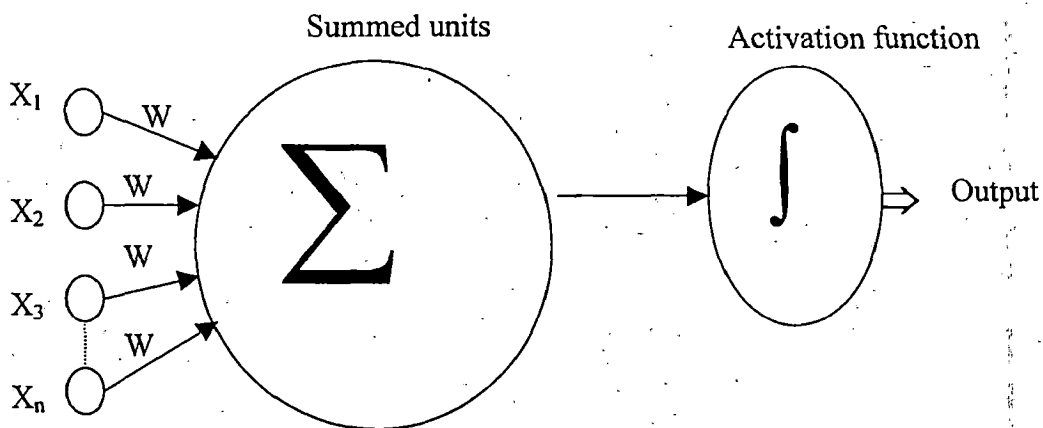


Figure 5.3: Anatomy of artificial neuron

5.3 Gradient Descent Learning Algorithm

Gradient descent learning is the mostly used principle of ANN training. The reason is that trial computation is required to implement this method, and the fact that the gradient can be computed with local information. The principle of gradient descent learning is very simple. The weights are moved in a direction opposite to the direction of the gradient. The gradient of a surface indicates to the direction of the maximum rate of change. Therefore, if the weights are moved in the opposite direction of the gradient, the system state will approach points where the surface is flatter.

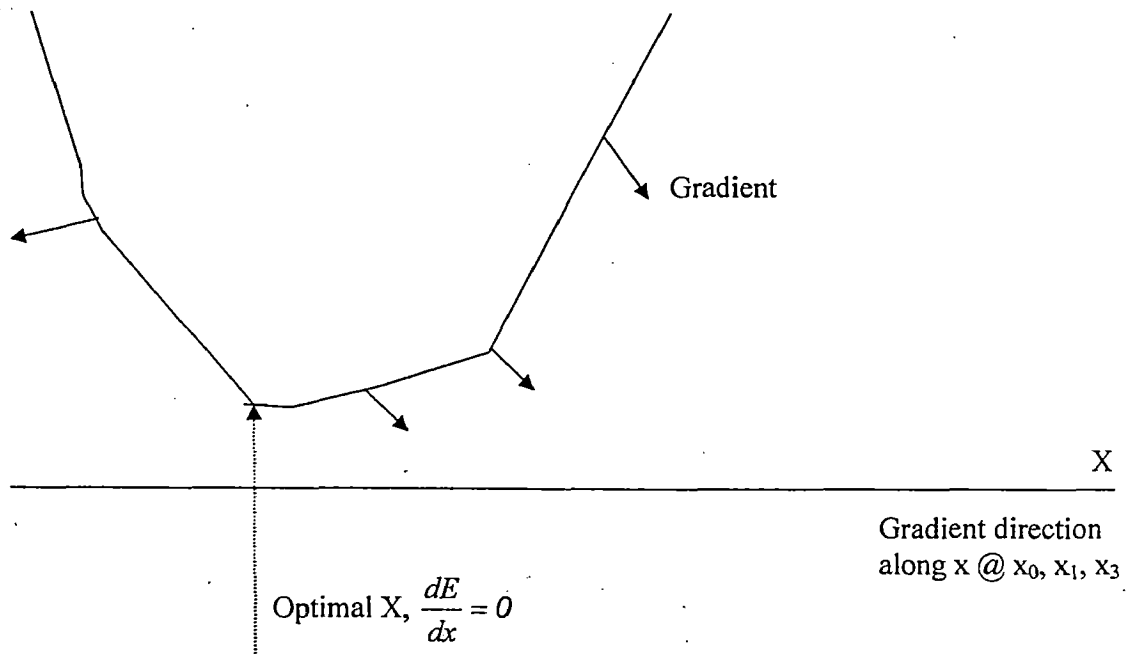


Fig. 5.4: Gradient Descent in One Dimension

5.4 The Neural Network Topology

The arrangement of the processing units, connections and pattern input / output in an ANN is referred to as topology. The processing units are arranged in three layers that are input, hidden and output. The units of a layer are similar in the sense that they all have the same activation dynamics and output function. The number of input and the number of output are problem specific. There are no fixed rules as to the how many units should be included in the hidden layer. If there are too less units in the hidden layer, the network may have difficulty in generalizing the problem. On the other hand, if there are too many units in the hidden layer, the network may take an unacceptably long time to learn. On the basis of direction of information flow and processing the ANNs are classified as feed forward and feed backward network.

5.4.1 Feed forward network

The nodes are generally arranged in layers, starting from first input layer and ending at the final output layer. There can be several hidden layers with each layer having one or more nodes. Information passes from the input to the output side. The neurons in one layer are connected to those in the next, but not to those in the same layer. Thus the output of a node in the one layer is only dependent on the input it receives from previous layers and the corresponding weights.

5.4.2 Feed backward network

Information flows through the nodes in both directions from the input to the output side and vice-versa. This is generally achieved by recycling previous network outputs as current inputs, thus allowing for feedback.

5.5 Activation Function

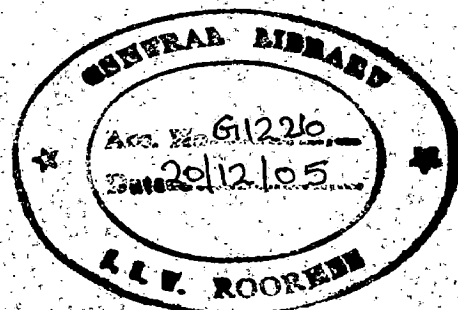
Sigmoid function is the mostly used function for solving ANN problems.

5.5.1 Sigmoid function

This function is a continuous function that varies gradually between asymptotic values 0 and 1 or -1 and +1 and is given by

$$\phi(x) = \frac{1}{1 + e^{-\beta x}} \quad (5.4)$$

where, β is the slope parameter, which adjusts the abruptness of the function as it changes between the two asymptotic values. Sigmoid functions are differentiable, which is an important feature of neural network theory. Experimental observations of biological neurons demonstrate that the firing is roughly sigmoid, when plotted.



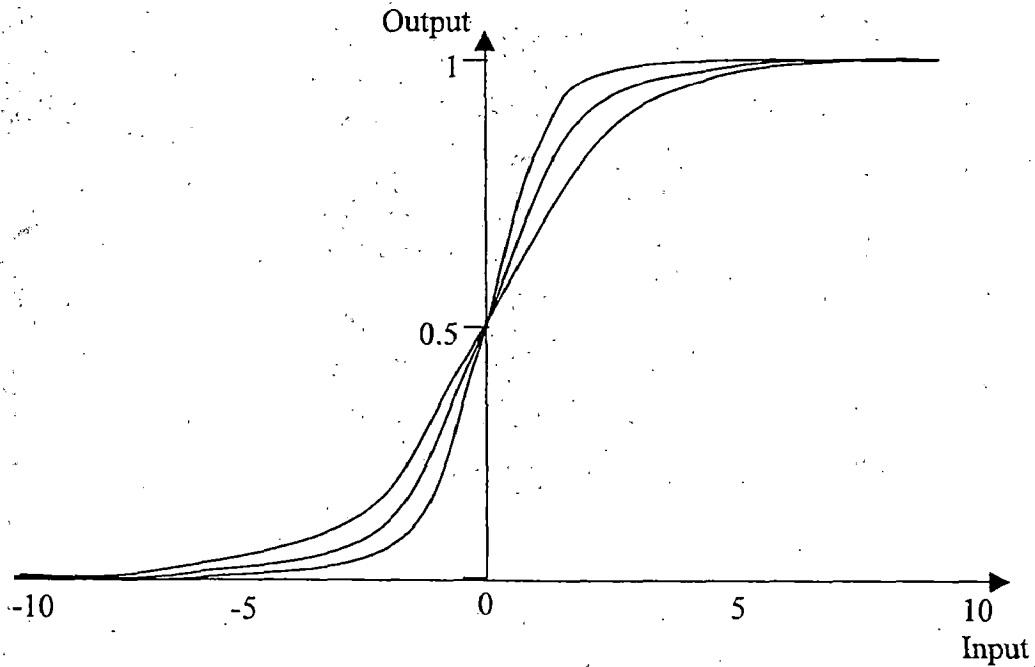


Fig 5.5: The Sigmoid Function

5.6 Architecture of ANN

The manner in which the neurons of a neural network are structurally and intimately linked with learning algorithm ceased to train the network. The optimal architecture is one, which yields the best performance in terms of error minimization, while training simple and compact structure. The numbers of input and output nodes are problem dependent. The flexibility lies in selecting number of hidden layers and in assigning the number of nodes to each of these layers.

5.7 Training of Artificial Neural Network

Once a network has been structured for a particular application, that network is ready to be trained. To start this process the initial weights are chosen randomly. Then, the training, or learning begins. Supervised and unsupervised are two methods used to train neural network.

5.7.1 Supervised training

In supervised training, both the inputs and the outputs are provided. The networks then process the inputs and compare its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights, which control the network. This process occurs over and over as the weights are continually tweaked. The set of data, which enables the training, is called the 'training set'. During the training of a network the same set of data is processed many times as the connection weights are ever refined. Sometimes, some networks never learn because the input data does not contain the specific information.

If a network simply cannot solve the problem, the designer then has to review the input and outputs, the number of layers, the number of elements per layer, the connections between the layers, the summation, transfer and training functions, and even the initial weight themselves, those changes create a successful network.

5.7.2 Unsupervised training

In unsupervised training, the network is provided with input but not with desired outputs. The system itself must then decide what features it will use to group the input data. At the present time, unsupervised learning is not well understood. Currently this field remains one that is still in the laboratory.

5.8 Back Propagation Algorithm

Back propagation is a system of method of training multi layer artificial neural networks. Scientist and Engineering community to the modeling has used it and processing of many quantitative phenomena using neural network has used it. This learning algorithm is applied to multi layer feed forwarded

network consisting of neurons with continuous differentiable activation functions. Such networks associated with the back propagation-learning algorithm are called back propagation networks. The back propagation algorithm is a generalization of the least mean square algorithm that modifies network weights to minimize the mean squared error between the desired and actual outputs of the network. Back propagation uses supervised learning in which the network is trained using data for which inputs as well as desired outputs are known. Once trained, the network weights are frozen and can be used to compute output values for new input samples.

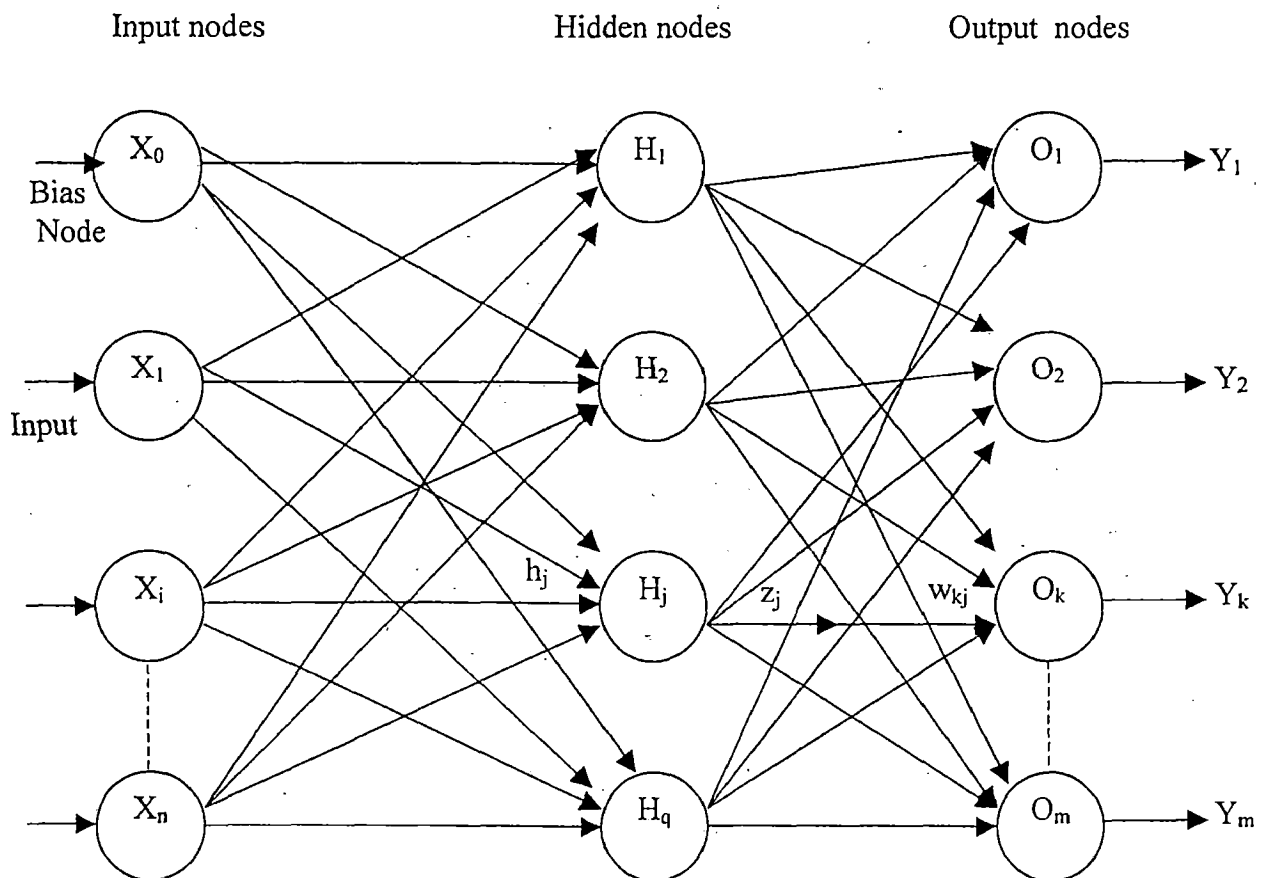


Fig 5.6: A typical two layered back propagation feed forward neural network

Considering a two layered feed forward neural network, the network has n nodes in the input layer, q nodes in the hidden layer and m nodes in the output layer. Bias node is connected to all the layers to take care of threshold values.

$$X = (x_1, x_2, \dots, x_n) \rightarrow \text{inputs}$$

$$T = (t_1, t_2, \dots, t_m) \rightarrow \text{target outputs}$$

$$y = (y_1, y_2, \dots, y_m) \rightarrow \text{actual output of ANN.}$$

Total input to j th hidden node

$$h_j = \sum_{i=0}^n w_{ji} x_i$$

Corresponding output

$$Z_j = S_{\beta}(h_j), \text{ where } S_{\beta} = \text{sigmoid function} = \frac{1}{1 + e^{-\beta h_j}}$$

$$\Rightarrow Z_j = S_{\beta} \left(\sum_{i=0}^n w_{ji} x_i \right)$$

Total input to K th output node

$$\begin{aligned} O_K &= \sum_{j=0}^q w_{kj} \cdot Z_j \\ &= w_{k0} x_0 + \sum_{j=1}^q w_{kj} \cdot S_{\beta} \left(\sum_{i=0}^n w_{ji} x_i \right) \end{aligned}$$

Final output will be

$$\begin{aligned} y_k &= S_{\beta}(O_K) \\ &= S_{\beta} \left[w_{k0} x_0 + \sum_{j=1}^q w_{kj} \cdot S_{\beta} \left(\sum_{i=0}^n w_{ji} x_i \right) \right] \end{aligned}$$

The error function,

$$E_p = \frac{1}{2} \sum_{k=1}^m (t_k - y_k)^2$$

$$= \frac{1}{2} \sum_{k=1}^m \left[t_k - S_\beta \left\{ w_{k0} x_0 + \sum_{j=1}^q w_{kj} S_\beta \left(\sum_{i=0}^n w_{ji} x_i \right) \right\} \right]^2$$

In steepest descent approach

Change in weight α negative gradient of E_p at the present location.

(i) Correction for weight between hidden to output layer:

$$\Delta w_{kj} = -\eta \frac{\partial E_p}{\partial w_{kj}}, \text{ where } \eta \text{ is a positive constant known as learning rate.}$$

$$E_p = \frac{1}{2} \sum_{k=1}^m [t_k - S_\beta(O_k)]^2$$

where, $O_k = \sum_{j=0}^q w_{kj} z_j$

$$\therefore \frac{\partial E_p}{\partial w_{kj}} = -[t_k - S_\beta(O_k)] S'_\beta(O_k) z_j$$

Again, $S'_\beta(O_k) = \beta y_k (1 - y_k)$

$$\therefore \frac{\partial E_p}{\partial w_{kj}} = -(t_k - y_k) \beta y_k (1 - y_k) z_j$$

$$\therefore \Delta w_{kj} = + \eta (t_k - y_k) \beta y_k (1 - y_k) z_j$$

$$\Rightarrow \Delta w_{kj} = \eta \delta_k \cdot z_j \quad (5.5)$$

where, $\delta_k = \beta (t_k - y_k) (1 - y_k) y_k$

ii. Correction of weights between input to hidden layer:

$$\Delta w_{ji} = -\eta \frac{\partial E_p}{\partial w_{ji}}$$

$$\begin{aligned}
\therefore \Delta w_{ji} &= -\eta \frac{\partial}{\partial w_{ji}} \left\{ \frac{1}{2} \sum_{k=1}^m (t_k - y_k)^2 \right\} \\
&= \sum_{k=1}^m \eta (t_k - y_k) \frac{\partial}{\partial w_{ji}} \cdot y_k \\
&= \sum_{k=1}^m \eta (t_k - y_k) \frac{\partial}{\partial w_{ji}} [S_{\beta}(O_k)] \\
&= \sum_{k=1}^m \eta (t_k - y_k) S'_{\beta}(O_k) \cdot \frac{\partial O_k}{\partial w_{ji}} \\
&= \sum_{k=1}^m \eta (t_k - y_k) S'_{\beta}(O_k) \cdot \frac{\partial O_k}{\partial z_j} \times \frac{\partial z_j}{\partial w_{ji}} \\
\therefore \Delta w_{ji} &= \sum_{k=1}^m \eta (t_k - y_k) S'_{\beta}(O_k) \cdot w_{kj} \cdot \frac{\partial}{\partial w_{ji}} \left[S_{\beta} \left(\sum_{i=0}^n w_{ji} x_i \right) \right] \\
&= \sum_{k=1}^m \eta (t_k - y_k) S'_{\beta}(O_k) \cdot w_{kj} \cdot S'_{\beta} \left(\sum_{i=0}^n w_{ji} x_i \right) \cdot x_i \\
&= \sum_{k=1}^m \eta (t_k - y_k) S'_{\beta}(O_k) \cdot w_{kj} \cdot S'_{\beta}(h_j) \cdot x_i \\
&= \sum_{k=1}^m \eta (t_k - y_k) \beta y_k (1 - y_k) \cdot w_{kj} \cdot S'_{\beta}(h_j) \cdot x_i \\
&= \sum_{k=1}^m \eta \cdot \delta_k \cdot w_{kj} \cdot S'_{\beta}(h_j) \cdot x_i \\
\therefore \Delta w_{ji} &= \sum_{k=1}^m \eta \delta_k \cdot x_i \tag{5.6}
\end{aligned}$$

where,

$$\delta_j = \delta_k w_{kj} \cdot S'_{\beta}(h_j)$$

$$= \delta_k w_{kj} \cdot \beta (1 - Z_j) Z_j$$

$$\therefore \Delta w_{ji} = \eta x_i \sum_{k=1}^m \delta_k$$

$$\Rightarrow \Delta w_{ji} = \eta \left[\beta (1 - Z_j) Z_j \cdot \sum_{k=1}^m \delta_k w_{kj} \right] x_i$$

$$\Rightarrow \Delta w_{ji} = \eta \delta_j x_i \quad (5.7)$$

where, $\delta_j = \beta(1 - Z_j)Z_j \sum_{k=1}^m \delta_k \cdot w_{kj}$

Hence,

(a) Change in weights

α (input to the node in the forward direction x Error term)

(b) δ_k values are calculated by using actual errors.

(c) δ_j values are calculated using the weighted sum of errors coming to the hidden node from the higher level nodes to which this node is connected.

5.9 Learning Factors of Back Propagation

One of the major issues concerning back propagation algorithm is its convergence. The convergence of back propagation is based on some important learning factors such as the initial weights, the learning rate, the nature of training set and the architecture of the network.

5.9.1 Initial weights

The initial weights of a multi layer feed forward network strongly affect the ultimate solution. They are typically initialized by small random values (between - 1.0 and 1.0 or -0.5 to +0.5). Equal weights values cannot train the network properly if the solution requires unequal weights to be developed. The initial weights cannot be large, otherwise the sigmoid will saturate, from the beginning and the system will stuck at a local minimum. The saturation is avoided by choosing the initial values of the synaptic weights to be uniformly distributed inside a small range of values. The range should not be too small as it can cause the learning to be very small.

5.9.2 Learning rate (η)

Weight vector change in back propagation are proportional to the negative gradient of the error, this guideline determines the relative changes that must occur in different weights when a training sample (or a set of samples) is presented, but does not fix the exact magnitudes of the desired weight changes. The magnitude change depends on the appropriate choice of the learning rate η . A large value of η will lead to rapid learning but the weight may then oscillate, while low values imply slow learning. This is typical of all gradient descent methods. The right value of η will depend on the application. Values between 0.1 and 0.9 have been used in many applications. The most efficient approach is to vary the learning rate as the training progresses, the effectiveness of learning rate may be checked as the training progresses and the value of the learning rate can be changed based on that.

5.9.3 Momentum factor (α)

Back propagation leads the weights in a neural network to a local minimum of the MSE, possibly substantially different from the global minimum that corresponds to the best choice of weights. This problem can be particularly bothersome if the "error surface" (plotting MSE against network weights) is highly uneven or jagged, with a large number of local minima.

We may prevent the network from getting stuck in some local minimum by making the weight changes depend on the average gradient of MSE in a small region rather than the precise gradient at a point average $\frac{\partial E}{\partial w}$ in a small neighbourhood can allow the network weights to be modified in the general direction of MSE decrease, without getting stuck in some local minima.

Calculating averages can be an expensive task. A shortcut, suggested by Rumelhart, et. al. (1986), is to make weight changes in the i th iteration of the back propagation algorithm depend on immediately preceding weight changes, made in the $(i-1)$ th iteration. This has an averaging effect, and diminishes the drastic fluctuations in weight changes over consecutive iterations. The implementation of this method is straight-forward, and is accomplished by adding a momentum term to the weight update rule,

$$\Delta w_{kj}(t-1) = \eta \delta_k x_j + \alpha \Delta w_{kj}(t) \quad (5.8)$$

where, $\Delta w_{kj}(t)$ is the weight required at time t , and

α is an additional parameter known as momentum factor.

Values for the momentum coefficient α can be obtained adaptively, as in the case of the learning rate parameter η . A well-chosen value of α can significantly reduce the number of iterations for convergence. A value close to 0 implies that the past history does not have much effect on the weight change, while a value closer to 1 suggests that the current error has little effect on the weight change.

5.9.4 Data normalization

The variables fall in the range of 0 to 1, because it smoothens the solution space and averages out some of the noise effects. Such process is called normalization or standardization. A typical variable, say electrical conductivity (EC), which can vary between zero to some maximum value EC_{max} can be standardized by the following formula:

$$EC_s = \frac{EC}{EC_{max}} \quad (5.9)$$

where EC_S is the standardized electrical conductivity. A different formula will be more suitable for a variable that varies within a certain range. There is, however, some danger of losing information in standardization.

5.9.5 Training data and generalization

The training data submitted to the network for it to learn and generalize the relation between input and output should be sufficient and proper. Networks with too many trainable parameters for a given amount of training data learn well but do not generalize well. This phenomenon is called over fitting with too few trainable parameters, the network fails to learn the training data. In estimation of parameter of a water quality model, the available data are divided into two parts. The first part is used to calibrate the model, and the second to validate it. This practice is known as 'Split-Sample' test. The length of calibration data depends upon the number of parameters to be estimated. The general practice is to use half to two-third of the data for calibration and the remaining for validation.

5.10 Steps in Development of ANN Model

The steps followed in the development of Artificial Neural Model are summarized as:

Step I: Identify parsimoniously all physically based input variables with their time memory that influence the output.

Step II: All inputs and output sets for the calibration (25 data sets) and verification (13 data sets) are normalized.

Step-III: Start with a three layered ANN model having only one hidden layer and the number of nodes in the hidden layer is approximately double of input models. The numbers of nodes in the input layer are equal to the

number of input variables, whereas, the number of nodes in output layer is equal to the number of output variables.

Step-IV: All the interconnecting weights are assigned a small value between -0.5 to +0.5 through a random numbers generation program.

Step-V: Select fixed or variable values of learning rate and / or momentum term depending upon the algorithm used for optimization.

Step-VI: Select the learning process that is either pattern learning or batch learning processes.

Step-VII: Execute the program, which performs:

- (a) feed forward calculation,
- (b) error back propagation in the network, and
- (c) finally change the weight.

Step-VIII: Estimate output for calibration and verification and apply performance evaluation criteria.

Step-IX: Perform whole operation for maximum desired iterations.

Step-X: Select the iteration that results in maximum generalization on the basis of performance evaluation criteria.

Step-XI: For required generalization repeat the learning process by assigning more numbers of nodes in the hidden layer or by increasing the numbers of hidden layers.

5.11 Performance Evaluation Criteria

The performance evaluation criteria used in the present study are RMSE, CC and CE.

Root Mean Square Error (RMSE):

It yields the residual errors in terms of mean square error, expressed as:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{\text{residual variance}}{n}} \\ &= \left(\sum_{j=1}^n (Y_j - \hat{Y}_j)^2 / n \right)^{1/2} \end{aligned}$$

where, Y and \hat{Y} are the observed and estimated values of sodium respectively and n is the number of observations.

Correlation coefficient (CC)

It is expressed as:

$$\text{CC} = \frac{\sum_{j=1}^n \{(Y_j - \bar{Y})(\hat{Y}_j - \bar{\hat{Y}})\}}{\left\{ \sum_{j=1}^n (Y_j - \bar{Y})^2 \sum_{j=1}^n (\hat{Y}_j - \bar{\hat{Y}})^2 \right\}^{1/2}} \times 100$$

where \bar{Y} and $\bar{\hat{Y}}$ are mean of observed and estimated values.

Coefficient of Efficiency (CE):

Based on the standardization of residual variance with initial variance, the coefficient of efficiency can be used to compare the relative performance.

It is expressed as:

$$\text{CE} = \left\{ 1 - \frac{\text{residual variance}}{\text{Initial variance}} \right\} \times 100\%$$

or,

$$\text{CE} = \left\{ 1 - \frac{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2} \right\} \times 100\%$$

RESULTS AND DISCUSSION

As discussed in the previous chapters, the present work aims at to develop (a) statistical models and (b) ANN-based models for determination of crucial ground water quality parameters for pre-monsoon and post-monsoon seasons of Jaipur District, Rajasthan, using the easily measurable quantities. Finally, the performance of these models is compared using the criteria discussed in Chapters 4 and 5. Thus, the following text discusses the results of the statistical and ANN-based models.

6.1 Statistical Model Development**6.1.1 Best subset procedure**

The best subset procedure primarily helps^{to} decide the sub-set which is the best in its performance, which is based on F , R^2 , and SSE-values. To this end, out of thirty eight water quality data values for both pre-monsoon and post-monsoon seasons, the first twenty five were used for model formulation, and the others for model validation. The results are discussed for pre- and post-monsoon seasons separately, in what follows.

(a) Pre-monsoon

To form various appropriate sub-sets, the Pearson correlation coefficient (R) between water quality parameters is computed, and the resulting coefficients are shown in Table 6.1. It is apparent from this table that Cl significantly correlates with EC ($R = 0.927$) and Hard ($R = 0.794$); SO_4 with EC ($R = 0.834$) and Cl ($R = 0.725$); NO_3 with Hard ($R = 0.752$); Na with EC ($R = 0.966$), Cl ($R = 0.855$), SO_4 ($R = 0.851$), and Alk ($R = 0.731$); K with NO_3 ($R = 0.885$), Hard ($R = 0.814$); and so on.

Table 6.1: Pearson correlation Coefficient between water quality parameters for pre-monsoon, Jaipur District (Rajasthan) with first 25 test data

	pH	EC	ALK	Hard	Cl	SO ₄	NO ₃	PO ₄	F	Na	K	Ca	Mg
pH	1.000												
EC	0.042	1.000											
ALK	0.435	0.647	1.000										
Hard	-0.524	0.650	-0.019	1.000									
Cl	-0.134	0.927	0.390	0.794	1.000								
SO ₄	0.053	0.834	0.409	0.466	0.725	1.000							
NO ₃	-0.336	0.434	0.022	0.752	0.449	0.195	1.000						
PO ₄	-0.134	-0.094	-0.279	0.212	0.014	-0.164	0.233	1.000					
F	0.452	0.556	0.867	-0.058	0.333	0.402	-0.083	-0.246	1.000				
Na	0.209	0.966	0.731	0.446	0.855	0.851	0.219	-0.188	0.642	1.000			
K	-0.384	0.475	-0.021	0.814	0.554	0.227	0.885	0.393	-0.089	0.260	1.000		
Ca	-0.555	0.569	-0.066	0.955	0.707	0.402	0.734	0.233	-0.087	0.359	0.803	1.000	
Mg	-0.448	0.674	0.030	0.957	0.811	0.489	0.705	0.174	-0.023	0.493	0.754	0.828	1.000

Since Na plays a greater role in determination of the level of salinity, which is the stipulated problem of the study, than Mg which correlates with other parameters such as Hard, Cl, NO₃, Ca, and K as well as does Na with other above-described parameters, the determination of Na is more crucial to the study than Mg. Therefore, taking the former element as a dependent variable, a model is developed for its determination.

Since R does not provide any literal interpretation, except for the strength of association with other variables, the coefficient of determination (R²) is computed for a better physical interpretation. It is computed and the resulting values are shown in Table 6.2. It is apparent from this table that Na is significantly correlated with EC, Cl, and SO₄, and least with PO₄, pH, NO₃, and K. Thus, a model can possibly be developed for Na using EC, Cl, and SO₄ as independent variables.

In the next step, Na was taken as a dependent variable; different combination of models was investigated for performance using R², F-value, and SSE as criteria. The computed values of these measures for different combinations are shown in Table 6.3, and its results are summarized in Table 6.4. Table 6.4 actually presents only those combinations, which show highest R² and F and lowest SSE. It leads to inferring that there exist eight possible combinations for model development. To further select an appropriate model containing least number of independent variables, F-criterion (Chapter 4) was applied, as shown in Table 6.5. If $F_{k-m, N-K-1} > F^*$ at significance level (α) equal to 5%, the model is more preferable than those showing otherwise trend. It is apparent from Table 6.5 that EC, Hard, Cl, and SO₄ form to be the most

Table 6.2: R² of water quality parameters with Na for pre-monsoon.

Water quality parameters	R ² Value for pre-monsoon
pH	0.044
EC	0.933
ALK	0.535
Hard	0.199
Cl	0.731
SO ₄	0.724
NO ₃	0.048
PO ₄	0.035
F	0.413
K	0.068
Ca	0.129
Mg	0.243

Table 6.3: Various combinations of models and their statistics for pre-monsoon season

No. Of. Variables	Variables	R ² Value	F- Value	SSE
1	EC*	0.933	3.19.913	654847
	Cl	0.731	62.489	2626707
	SO ₄	0.724	60.209	2698689
	ALK	0.535	26.434	4542529
	F	0.413	16.163	5733819
	Mg	0.243	7.401	7386332
	Hard	0.199	5.712	7821072
	Ca	0.129	3.397	8506971

No. Of. Variables	Variables	R ² Value	F- Value	SSE
2	EC, Cl	0.945	188.457	538441
	EC, SO ₄	0.940	171.397	588802
	EC, ALK	0.952	220.189	464538
	EC, F	0.949	205.065	497053
	EC, Mg	0.978	496.704	211532
	EC, Hard*	0.990	1115.338	95349
	EC, Ca	0.987	808.946	130979

No. Of. Variables	Variables	R ² Value	F- Value	SSE
3	EC, Hard, Cl*	0.995	1421	4.785 x 10 ⁴
	EC, Hard, SO ₄	0.992	847	7.999 x 10 ⁴
	EC, Hard, ALK	0.994	1237	5.493 x 10 ⁴
	EC, Hard, F	0.992	869	7.799 x 10 ⁴
	EC, Hard, Mg	0.991	743	9.104 x 10 ⁴
	EC, Hard, Ca	0.991	743	9.107 x 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
4	EC, Hard, Cl, SO ₄ *	0.998	2166	2.248 x 10 ⁴
	EC, Hard, Cl, ALK	0.996	1380	3.523 x 10 ⁴
	EC, Hard, Cl, F	0.996	1169	4.157 x 10 ⁴
	EC, Hard, Cl, Mg	0.995	1030	4.715 x 10 ⁴
	EC, Hard, Cl, Ca	0.995	1030	4.714 x 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
5	EC, Hard, Cl, SO ₄ , ALK*	0.998	1935	1.912 x 10 ⁴
	EC, Hard, Cl, SO ₄ , F	0.998	1667	2.219 x 10 ⁴
	EC, Hard, Cl, SO ₄ , Mg	0.998	1770	2.177 x 10 ⁴
	EC, Hard, Cl, SO ₄ , Ca	0.998	1700	2.177 x 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
6	EC, Hard, Cl, SO ₄ , ALK, F*	0.998	1807	1.617 x 10 ⁴
	EC, Hard, Cl, SO ₄ , ALK, Mg	0.998	1585	1.843 x 10 ⁴
	EC, Hard, Cl, SO ₄ , ALK, Ca	0.998	1585	1.843 x 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
7	EC, Hard, Cl, SO ₄ , ALK, F, Mg*	0.998	1503	1.574 x 10 ⁴
	EC, Hard, Cl, SO ₄ , ALK, F, Ca	0.998	1503	1.574 x 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
8	EC, Hard, Cl, SO ₄ , ALK, F, Mg, Ca*	0.998	1250	1.558 x 10 ⁴

Table 6.4: Selected Sets/ subsets candidate for possible model independent variables for pre-monsoon season with Sodium.

Variables	Set of independent variables	N	R ²	F-Value	SSE
8	EC, Hard, Cl, SO ₄ , ALK, F, Mg, Ca	25	0.998	1250	1.558 x 10 ⁴
7	EC, Hard, Cl, SO ₄ , ALK, F, Mg,	25	0.998	1503	1.574 x 10 ⁴
6	EC, Hard, Cl, SO ₄ , ALK, F,	25	0.998	1807	1.617 x 10 ⁴
5	EC, Hard, Cl, SO ₄ , ALK,	25	0.998	1935	1.912 x 10 ⁴
4	EC, Hard, Cl, SO ₄ ,	25	0.998	2166	2.248 x 10 ⁴
3	EC, Hard, Cl,	25	0.995	1421	4.785 x 10 ⁴
2	EC, Hard,	25	0.990	1115	9.534 x 10 ⁴
1	EC,	25	0.933	319	6.5484 x 10 ⁵

Table 6.5: Selection of model variables on the basis of F-statistics for pre-monsoon season with Sodium.

Full model with K parameters		Reduced model with (K-m) parameters			N-K-	F _{K-m, N-K-1}	F*, α=5%	Preferred model	
Model	N	SSE _f	Model	SSE _r	K-m	N-K-1			
EC+ Hard+Cl+ SO ₄ + ALK+F+Mg+Ca	25	1.558 x 10 ⁴	EC+ Hard+Cl+ SO ₄ + ALK+F+Mg	1.574 x 10 ⁴	7	16	0.023	2.66	Reduced
EC+ Hard+Cl+ SO ₄ + ALK+F+Mg	25	1.574 x 10 ⁴	EC+ Hard+Cl+ SO ₄ + ALK+F	1.617 x 10 ⁴	6	17	0.078	2.70	Reduced
EC+ Hard+Cl+ SO ₄ + ALK+F	25	1.617 x 10 ⁴	EC+ Hard+Cl+ SO ₄ + ALK	1.912 x 10 ⁴	5	18	0.657	2.77	Reduced
EC+ Hard+Cl+ SO ₄ + ALK	25	1.912 x 10 ⁴	EC+ Hard+Cl+ SO ₄	2.248 x 10 ⁴	4	19	0.833	2.90	Reduced
EC+ Hard+Cl+ SO ₄	25	2.248 x 10 ⁴	EC+ Hard+Cl	4.785 x 10 ⁴	3	20	7.523	3.10	Full
EC+ Hard+Cl+ SO ₄	25	2.248 x 10 ⁴	EC+ Hard	9.534 x 10 ⁴	2	20	32.413	3.49	Full
EC+ Hard+Cl+ SO ₄	25	2.248 x 10 ⁴	EC	6.548 x 10 ⁵	1	20	562.572	4.35	Full

optimal set of independent variables. It is noted that Na is taken as the dependent variable in this analysis. It follows that

$$\text{Na} = 27.082 + 0.173\text{EC} - 0.520\text{Hard} + 0.263\text{Cl} + 0.154\text{SO}_4 \quad (6.1)$$

$$(R^2 = 99.8\%, F = 2166)$$

The above proposed model (Eq. 6.1) is verified on the above remaining 13 data-points observed at the sites different from those used in model development. The validation results are shown in Table 6.6 and plotted in Figs. 6.1 and 6.2. Both these figures indicate more than satisfactory model performance in validation as the model-computed values fall quite close to the observed.

(b) Post Monsoon

Following the same steps as in the pre-monsoon, an analysis was repeated with the data set of post-monsoon season, and the results are shown in Tables 6.7-6.12. Finally, a model from Table 6.11 can be suggested as follows:

$$\text{Na} = 14.481 + 0.174\text{EC} - 0.519\text{Hard} + 0.301\text{Cl} \quad (6.2)$$

$$(R^2 = 99.8\% \text{ and } F = 4082)$$

This model (Eq. 6.2) is validated on the data set not used in model development and the results are shown in Table 6.12 and plotted in Figs. 6.3 and 6.4. The resulting R^2 value equal to 99.84% indicates almost a perfect match between the observed and computed values, as apparent from these figures.

Thus, the model (Eq. 6.2) suggested is appropriate for determination of Na from the indicated independent variables.

Table 6.6: Validation for the model equation for pre-monsoon season with Sodium.
Model equation is: $Na = 27.082 + 0.173 EC - 0.520 Hard + 0.263 Cl + 0.154 SO_4$

Site No.	Observed Value	Model Value
26	220	211
27	495	485
28	1302	1327
29	186	172
30	160	132
31	598	614
32	436	463
33	152	157
34	506	527
35	192	188
36	131	130
37	856	788
38	445	441

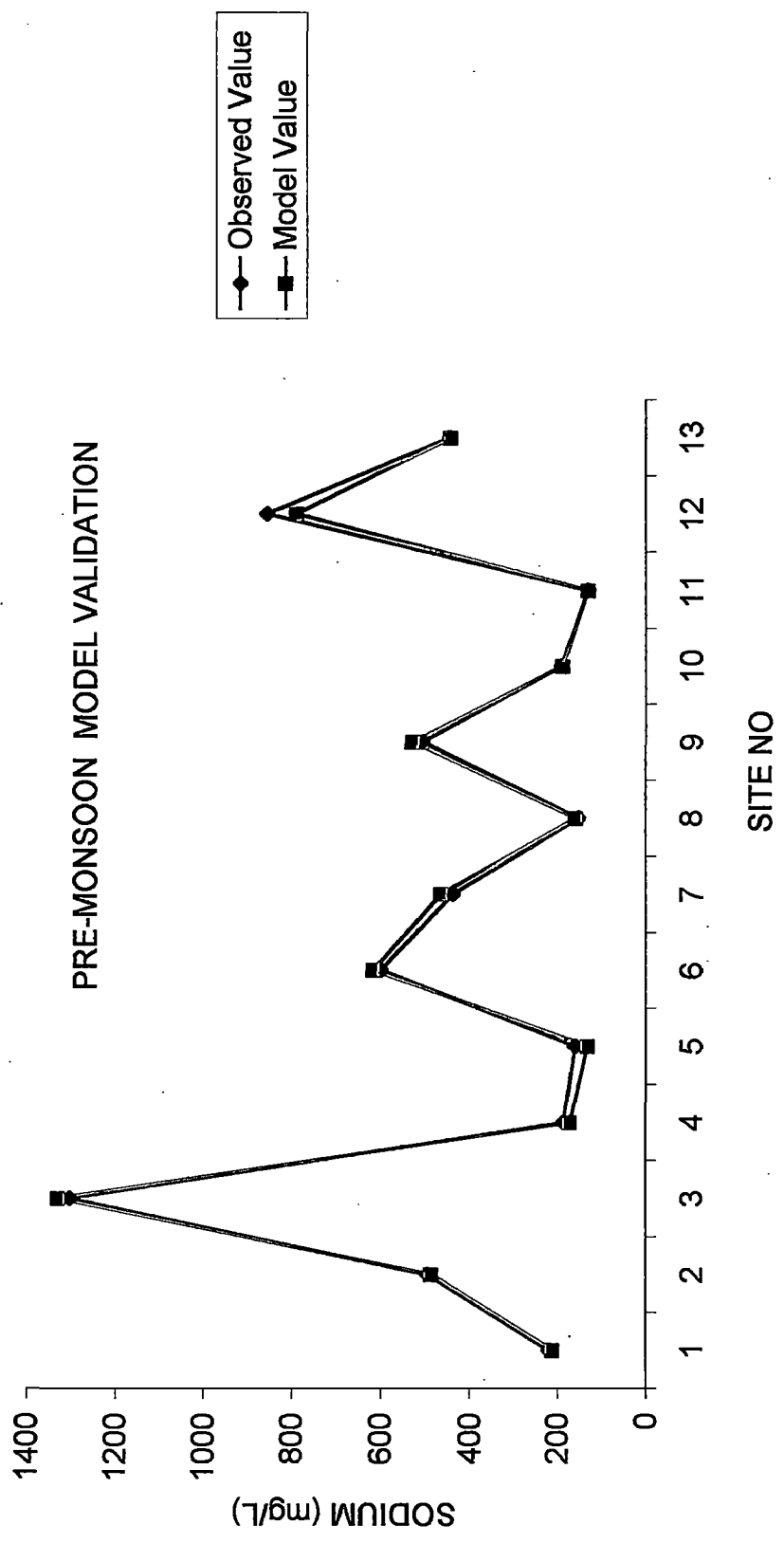


Fig.6.1: Model validation of observed and computed values of Sodium (mg/L) for pre monsoon season.

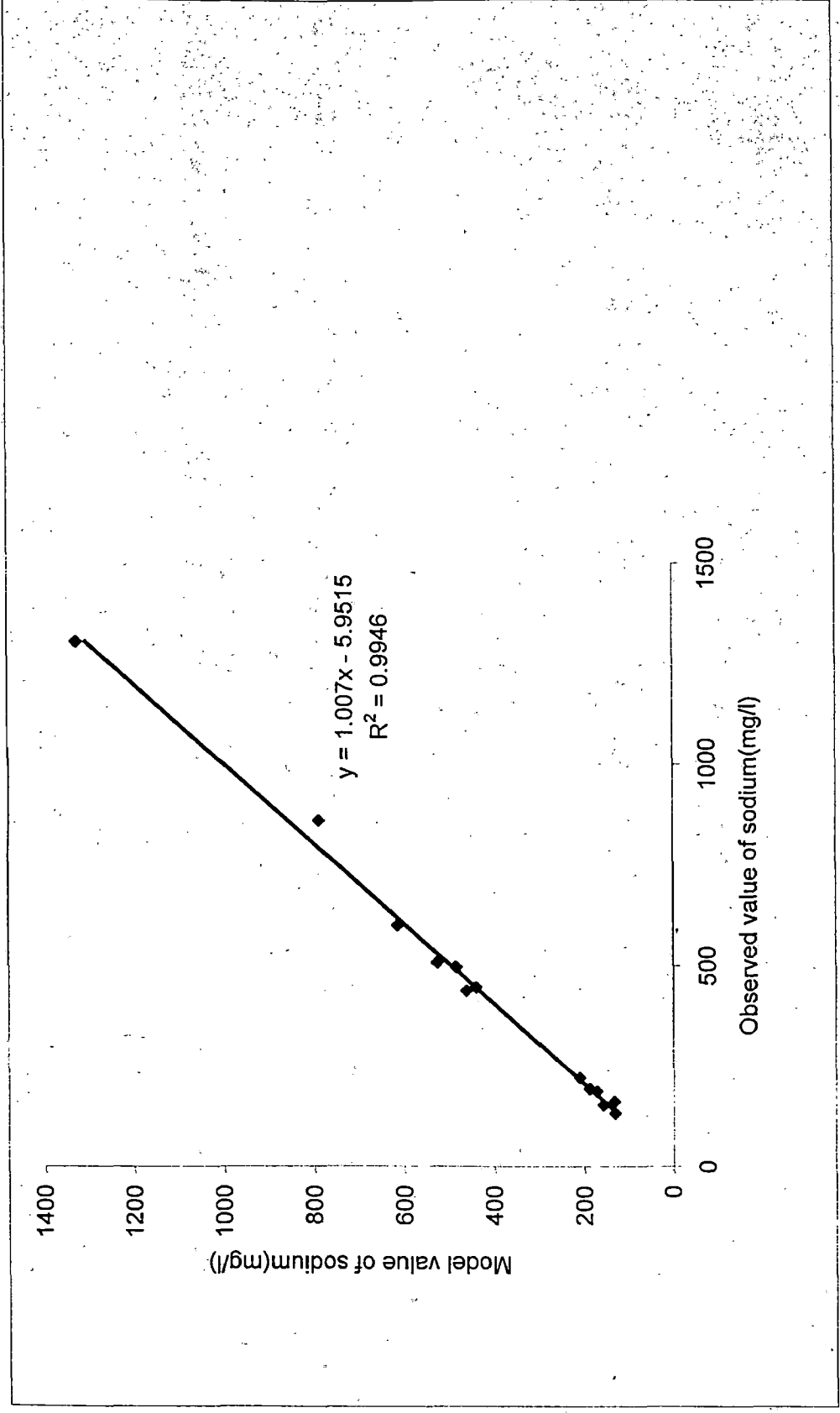


Fig. 6.2: Plot of observed and model value for pre-monsoon season

Table 6.7. Pearson correlation Coefficient between water quality parameters for post- monsoon, Jaipur District (Rajasthan) with first 25 test data

	pH	EC	ALK	Hard	Cl	SO ₄	NO ₃	PO ₄	F	Na	K	Ca	Mg
pH	1.000												
EC	0.011	1.000											
ALK	0.358	0.688	1.000										
Hard	-0.563	0.577	-0.067	1.000									
Cl	-0.189	0.929	0.419	0.765	1.000								
SO ₄	0.083	0.845	0.440	0.453	0.772	1.000							
NO ₃	-0.397	0.483	0.031	0.798	0.571	0.232	1.000						
PO ₄	0.006	-0.084	-0.303	-0.110	-0.047	0.290	-0.284	1.000					
F	0.399	0.516	0.807	-0.118	0.258	0.385	-0.028	-0.252	1.000				
Na	0.169	0.969	0.777	0.371	0.854	0.838	0.290	0.034	0.591	1.000			
K	-0.435	0.443	-0.039	0.819	0.569	0.253	0.858	-0.226	-0.113	0.240	1.000		
Ca	-0.577	0.551	-0.014	0.940	0.695	0.463	0.692	-0.160	-0.048	0.349	0.733	1.000	
Mg	-0.506	0.551	-0.014	0.965	0.757	0.408	0.815	-0.062	-0.164	0.356	0.818	0.817	1.000

Table 6.8: R² of water quality parameters with Na for post-monsoon.

Water quality parameters	R ² Value for post-monsoon
pH	0.029
EC	0.939
ALK	0.604
Hard	0.137
Cl	0.729
SO ₄	0.703
NO ₃	0.084
PO ₄	0.001
F	0.349
K	0.057
Ca	0.122
Mg	0.127

Table 6.9: Various combinations of models and their statistics for post-monsoon season

No. Of Variables	Variables	R ² Value	F- Value	SSE
1	EC*	0.939	355	3.920 X 10 ⁵
	Cl	0.729	61	1.750 X 10 ⁶
	SO ₄	0.703	54	1.917 X 10 ⁶
	ALK	0.604	35	2.559 X 10 ⁶
	F	0.349	12	4.201 X 10 ⁶
	Hard	0.137	3	5.568 X 10 ⁶
	Mg	0.127	3	5.635 X 10 ⁶
	Ca	0.122	3	5.666 X 10 ⁶

No. Of Variables	Variables	R ² Value	F- Value	SSE
2	EC, Cl	0.956	236	2.866 X 10 ⁵
	EC, SO ₄	0.941	174	3.837 X 10 ⁵
	EC, ALK	0.962	279	2.444 X 10 ⁵
	EC, F	0.951	211	3.190 X 10 ⁵
	EC, Hard*	0.993	1525	4.622 X 10 ⁴
	EC, Mg	0.984	693	1.008 X 10 ⁵
	EC, Ca	0.988	900	7.792 X 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
3	EC, Hard, Cl*	0.998	4082	1.105 X 10 ⁴
	EC, Hard, SO ₄	0.993	1011	4.435 X 10 ⁴
	EC, Hard, ALK	0.995	1472	3.053 X 10 ⁴
	EC, Hard, F	0.994	1234	3.638 X 10 ⁴
	EC, Hard, Mg	0.993	1047	4.285 X 10 ⁴
	EC, Hard, Ca	0.993	1047	4.286 X 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
4	EC, Hard, Cl, SO ₄ *	0.999	3615	8.914 X 10 ³
	EC, Hard, Cl, ALK	0.998	3045	1.057 X 10 ⁴
	EC, Hard, Cl, F	0.998	2948	1.092 X 10 ⁴
	EC, Hard, Cl, Mg	0.998	2959	1.088 X 10 ⁴
	EC, Hard, Cl, Ca	0.998	2960	1.088 X 10 ⁴

No. Of. Variables	Variables	R ² Value	F- Value	SSE
5	EC, Hard, Cl, SO ₄ , ALK*	0.999	3195	7.666 X 10 ³
	EC, Hard, Cl, SO ₄ , F	0.999	2747	8.914 X 10 ³
	EC, Hard, Cl, SO ₄ , Mg	0.999	2755	8.891 X 10 ³
	EC, Hard, Cl, SO ₄ , Ca	0.999	2755	8.890 X 10 ³

No. Of. Variables	Variables	R ² Value	F- Value	SSE
6	EC, Hard, Cl, SO ₄ , ALK, F	0.999	2561	7.552 X 10 ³
	EC, Hard, Cl, SO ₄ , ALK, Mg*	0.999	2572	7.519 X 10 ³
	EC, Hard, Cl, SO ₄ , ALK, Ca	0.999	2571	7.521 X 10 ³

No. Of. Variables	Variables	R ² Value	F- Value	SSE
7	EC, Hard, Cl, SO ₄ , ALK, Mg, F	0.999	2114	7.405 X 10 ³
	EC, Hard, Cl, SO ₄ , ALK, Mg, Ca*	0.999	2146	7.295 X 10 ³

No. Of. Variables	Variables	R ² Value	F- Value	SSE
8	EC, Hard, Cl, SO ₄ , ALK, F, Mg, Ca, F*	0.999	1804	7.014 X 10 ³

Table 6.10: Selected Sets/ subsets candidate for possible model independent variables for post-monsoon season with Sodium.

Variables	Set of independent variables	N	R ²	F-Value	SSE
8	EC, Hard, Cl, SO ₄ , ALK, Mg, Ca, F	25	0.999	1804	7.145 X 10 ³
7	EC, Hard, Cl, SO ₄ , ALK, Mg, Ca	25	0.999	2146	7.295 X 10 ³
6	EC, Hard, Cl, SO ₄ , ALK, Mg	25	0.999	2572	7.519 X 10 ³
5	EC, Hard, Cl, SO ₄ , ALK	25	0.999	3195	7.666 X 10 ³
4	EC, Hard, Cl, SO ₄	25	0.999	3615	8.914 X 10 ³
3	EC, Hard, Cl	25	0.998	4082	1.105 X 10 ⁴
2	EC, Hard	25	0.993	1525	4.622 X 10 ⁴
1	EC	25	0.939	355	3.920 X 10 ⁵

Table 6.11: Selection of model variables on the basis of F-statistics for post-monsoon season with Sodium.

Full model with K parameters			Reduced model with (K-m) parameters			K-m	N-K-1	$F_{K-m, N-K-1}$	F*, $\alpha=5\%$	Preferred model
Model	N	SSE _r	Model	SSE _r						
EC+ Hard+Cl+ SO ₄ + ALK +Mg+Ca+F	25	7.145 X 10 ³	EC+ Hard+Cl+ SO ₄ + ALK +Mg+Ca	7.295 X 10 ³	7	16	0.048	2.66	Reduced	
EC+ Hard+Cl+ SO ₄ + ALK+ Mg+Ca	25	7.295 X 10 ³	EC+ Hard+Cl+ SO ₄ + ALK +Mg	7.519 X 10 ³	6	17	0.08695	2.70	Reduced	
EC+ Hard+Cl+ SO ₄ + ALK+Mg	25	7.519 X 10 ³	EC+ Hard+Cl+ SO ₄ + ALK	7.666 X 10 ³	5	18	0.0701	2.77	Reduced	
EC+ Hard+Cl+ SO ₄ + ALK	25	7.666 X 10 ³	EC+ Hard+Cl+ SO ₄	8.914 X 10 ³	4	19	0.7736	2.90	Reduced	
EC+ Hard+Cl+ SO ₄	25	8.914 X 10 ³	EC+ Hard+Cl	1.105 X 10 ⁴	3	20	1.597	3.10	Reduced	
EC+ Hard+Cl	25	1.105 X 10 ⁴	EC+ Hard	4.622 X 10 ⁴	1	21	33.424	3.47	Full	
EC+ Hard+Cl	25	1.105 X 10 ⁴	EC	3.920 X 10 ⁵	1	21	363.903	4.32	Full	

Table 6.12: Validation for the model equation for post-monsoon season with Sodium.

Model equation is: $Na = 14.481 + 0.174EC - 0.519Hard + 0.301CI$

Site No.	Observed Value	Model Value
26	188	179
27	354	361
28	1235	1189
29	117	102
30	104	90
31	568	580
32	88	93
33	136	134
34	490	499
35	125	129
36	70	76
37	712	712
38	335	341

POST - MONSOON MODEL VALIDATION

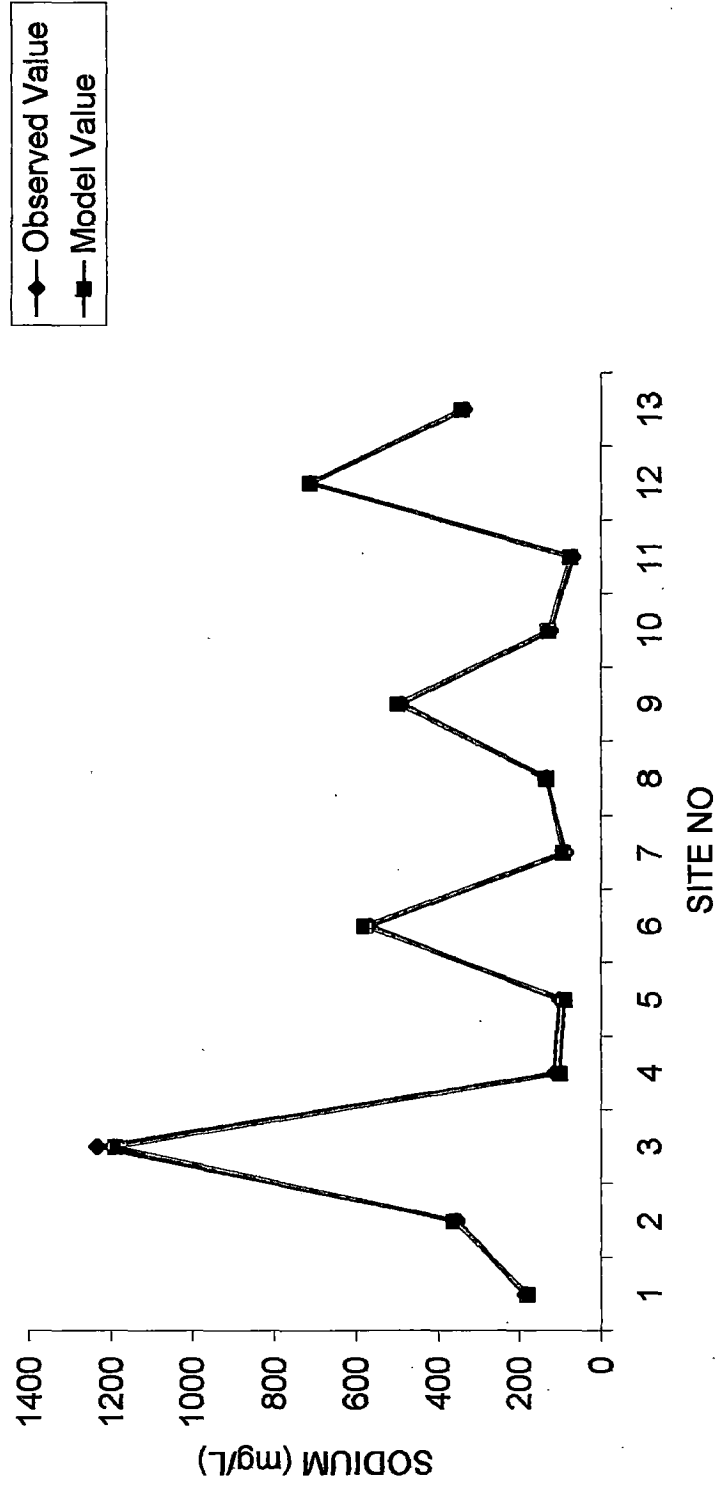


Fig. 6.3: Model validation of observed and computed values of Sodium (mg/L) for post -monsoon season.

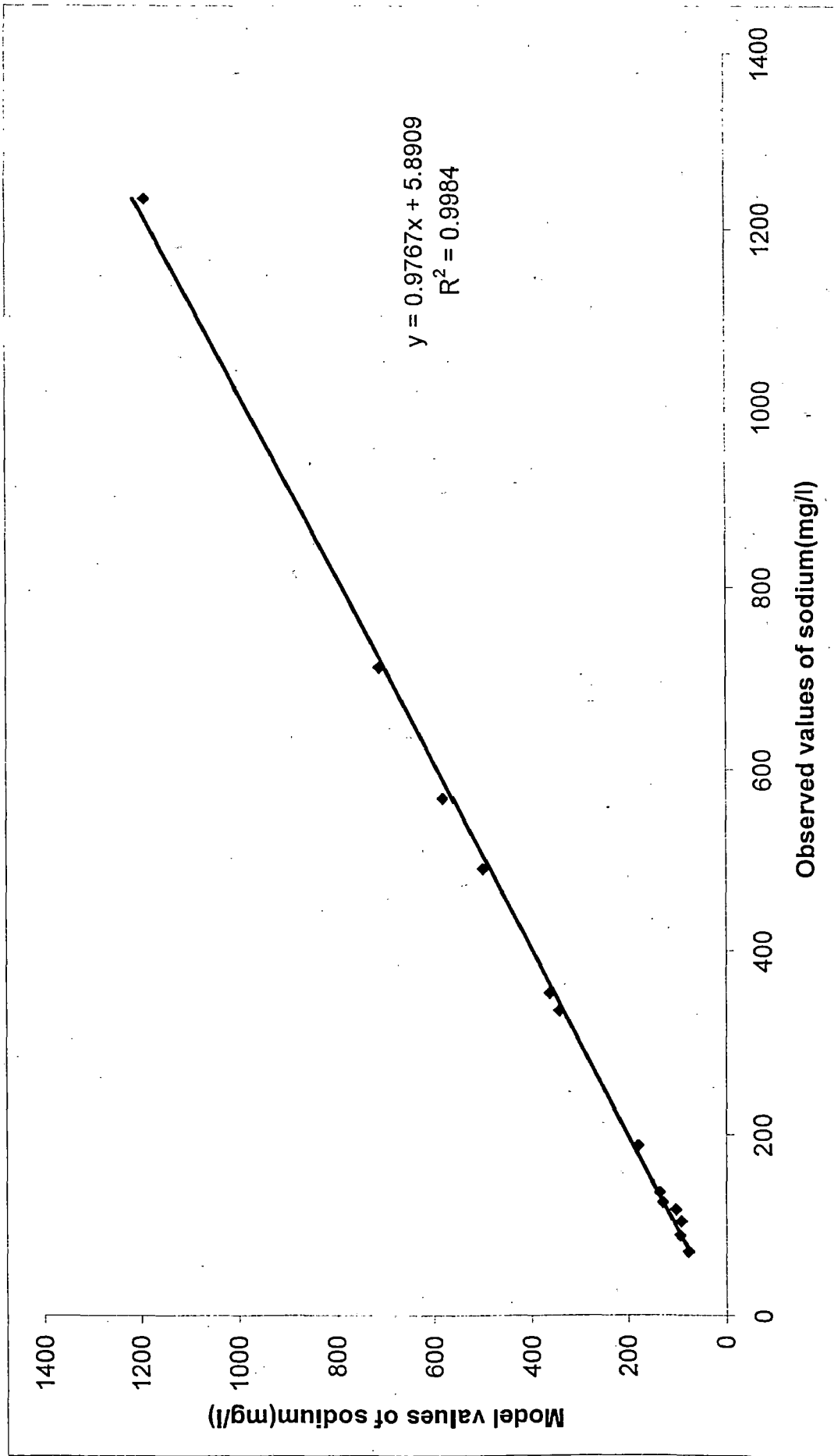


Fig. 6.4: Plot of observed and model value for post-monsoon season

It is to further emphasize that in both the models (Eqs. 6.1 and 6.2) developed for pre- and post monsoon seasons, EC is the most significant and dominating independent variable and Cl stands next to it. On the other hand hardness and SO_4 (valid for pre-monsoon season) are the less domination variables for use in the model. An elimination of these from the pre-monsoon model (Table 6.3) can be developed with $R^2 = 0.945$, and post-monsoon model (Table 6.9) with $R^2 = 0.956$. Both these R^2 -values are reasonably close to 1.00 to suggest appropriate models with consideration of only EC and Cl. The advantage of such a simplification is obvious for easy data collection and less expenditure reasons. Statistically, the most dominating variables can better be identified using the Principal component analysis and the results can be better interpreted, in what follows.

6.1.2 Principal Component Analysis

In the principal component analysis, as also discussed in Chapter 4, R-values are first determined, as shown in Tables 6.1 and 6.7 for above pre- and post-monsoon analyses, respectively. Then eigen values and associated percent of total variances are determined, as shown in Table 6.13 and 6.14 for both pre- and post-monsoon seasons, respectively. Afterwards, the principal component variables are extracted from the thirteen explanatory variables, as shown in Tables 6.15 and 6.16 for pre- and post-monsoon seasons, respectively. In Table 6.15, the nos. in the second row refer to 13 principal components and the resulting fractional values show the correlation of a particular variable with the principal component, If the value is close ± 1.00 , the variable is most correlated with the corresponding principal component.

Table 6.13: Eigen values based on correlation matrix for pre-monsoon

Sl. No.	Eigen values	Percent of total variance	cumulative percent of variance
1	6.159	47.374	47.374
2	3.407	26.210	73.584
3	0.978	7.521	81.105
4	0.793	6.102	87.207
5	0.558	4.289	91.496
6	0.387	2.976	94.472
7	0.246	1.889	96.361
8	0.206	1.583	97.944
9	0.166	1.275	99.219
10	0.100	0.768	99.987
11	0.002	0.013	100.000
12	0.000	0.000	100.000
13	0.000	0.000	100.000

Table 6.14: Eigen values based on correlation matrix for post-monsoon season

Sl. No.	Eigen values	Percent of total variance	cumulative percent of variance
1	6.188	47.599	47.599
2	3.412	26.247	73.846
3	1.076	8.275	82.121
4	0.643	4.949	87.07
5	0.604	4.647	91.717
6	0.401	3.084	94.801
7	0.268	2.059	96.86
8	0.212	1.632	98.492
9	0.134	1.033	99.525
10	0.061	0.469	99.994
11	0.001	0.005	100.000
12	0.000	0.000	100.000
13	0.000	0.000	100.000

Table 6.15: Varimax rotated component loadings for pre-monsoon season

Variable	Principal component loading												
	1	2	3	4	5	6	7	8	9	10	11	12	13
pH	-0.121	0.751	0.105	-0.389	0.386	0.301	0.041	0.027	0.105	0.071	0.000	0.000	0.000
EC	0.926	0.375	-0.027	0.025	0.053	-0.069	0.028	-0.064	-0.008	-0.028	0.001	-0.001	-0.001
ALK	0.413	0.774	-0.213	-0.138	-0.282	-0.041	-0.033	-0.254	-0.052	0.128	0.008	0.000	0.000
Hard	0.865	-0.457	0.072	0.076	-0.075	0.111	0.019	0.018	0.072	0.087	-0.004	0.000	-0.000
Cl	0.943	0.106	0.047	0.148	0.062	0.095	0.165	-0.063	0.064	-0.162	0.023	0.000	0.000
SO ₄	0.734	0.385	0.026	0.317	0.324	-0.160	-0.211	0.141	-0.096	0.083	0.010	0.000	0.000
NO ₃	0.600	-0.366	0.082	-0.583	0.050	-0.377	0.092	0.078	0.029	0.017	0.005	0.000	0.000
PO ₄	-0.029	-0.345	-0.912	0.038	0.193	-0.006	0.075	0.015	0.051	0.022	0.000	0.000	-0.000
F	0.329	0.771	-0.210	-0.078	-0.384	0.062	-0.026	0.302	0.054	-0.043	-0.000	-0.000	0.000
Na	0.805	0.552	-0.029	0.070	0.119	-0.103	0.064	-0.066	-0.017	-0.083	-0.031	0.000	0.000
K	0.693	-0.468	-0.134	-0.359	0.006	0.238	-0.214	-0.017	-0.206	-0.092	-0.001	0.000	0.000
Ca	0.811	-0.462	0.041	0.033	-0.070	0.043	-0.214	-0.053	0.263	0.030	-0.004	0.000	0.000
Mg	0.827	-0.404	0.093	0.109	-0.072	0.166	0.244	0.085	-0.122	0.133	-0.004	-0.000	0.000

Table 6.16: Varimax rotated component loading for post monsoon season

Variable	Principal component loading												
	1	2	3	4	5	6	7	8	9	10	11	12	13
pH	-0.252	0.716	-0.288	0.103	0.526	0.154	-0.027	0.013	0.172	0.017	0.000	-0.000	0.000
EC	0.900	0.416	0.056	0.036	0.013	-0.099	0.044	0.022	-0.007	-0.010	0.001	-0.002	0.000
ALK	0.431	0.788	0.020	-0.260	-0.194	-0.060	0.199	0.173	0.053	0.109	0.006	0.000	-0.000
Hard	0.857	-0.478	-0.053	0.059	-0.071	0.102	-0.003	-0.053	0.095	0.053	-0.002	0.000	0.000
Cl	0.949	0.127	0.018	0.176	0.009	-0.081	0.120	-0.098	0.053	-0.138	0.014	0.000	-0.000
SO ₄	0.723	0.409	0.266	0.378	0.102	0.100	-0.196	0.051	-0.170	0.079	0.005	0.000	-0.000
NO ₃	0.699	-0.289	-0.208	-0.373	0.297	-0.344	-0.178	-0.063	-0.045	0.037	0.003	0.000	-0.000
PO ₄	-0.141	-0.218	0.919	-0.143	0.238	-0.017	0.042	-0.035	0.088	0.011	0.000	-0.000	0.000
F	0.271	0.772	0.066	-0.379	-0.170	0.282	-0.154	-0.221	-0.019	-0.032	-0.000	0.000	0.000
Na	0.775	0.592	0.082	0.079	0.031	-0.163	0.062	0.007	-0.018	-0.054	-0.020	0.000	0.000
K	0.697	-0.469	-0.056	-0.265	0.239	0.304	0.112	0.203	-0.120	-0.066	-0.001	0.000	0.000
Ca	0.827	-0.382	0.030	0.017	-0.203	0.059	-0.250	0.159	0.190	-0.016	-0.002	0.000	-0.000
Mg	0.787	-0.501	-0.112	0.086	0.041	0.125	0.191	-0.214	0.010	0.101	-0.003	0.000	-0.000

Further, the number of principal components required to explain the variation in data is selected on the basis of eigen values (Tables 6.13 and 6.14). Components that explain 87.207 percent for pre-monsoon and 82.121 percent for post-monsoon of the total variance are chosen for further analysis. The eigen values, percent of total variance explained and cumulative percent of total variance are given in Tables 6.13 and 6.14 for pre-monsoon and post-monsoon, respectively. The rotated loading corresponding to each selected variable for components is given in Tables 6.15 and 6.16 for pre-monsoon and post-monsoon, respectively. The factors and their loadings are shown diagrammatically in Figures 6.5 and 6.6 for pre-monsoon and post-monsoon season, respectively. The rectangular boxes represent the factors and the horizontal central line represents zero loading for the variable. Lines near the top of boxes represent high positive loading, and the points near the bottom high negative loading.

(a) Pre-monsoon:

In case of pre-monsoon, the first eigen value is 6.59 and it explains 47.374% of total variance, second value (= 3.407) explains 26.10% of total variance, third value (= 0.978) explains 7.521%, and so on. Thus, the first four components account for a total of 87.207% of the total variance (Table 6.13).

The following four components have been interpreted as follows:

- | | |
|----------------|--|
| Component I : | Conductivity factor
Cl, EC, Hard, Mg, Ca, Na, SO ₄ |
| Component II: | Fluoride factor
Alk, F, pH |
| Component III: | Phosphate factor
PO ₄ |
| Component IV: | Nitrate factor
NO ₃ |

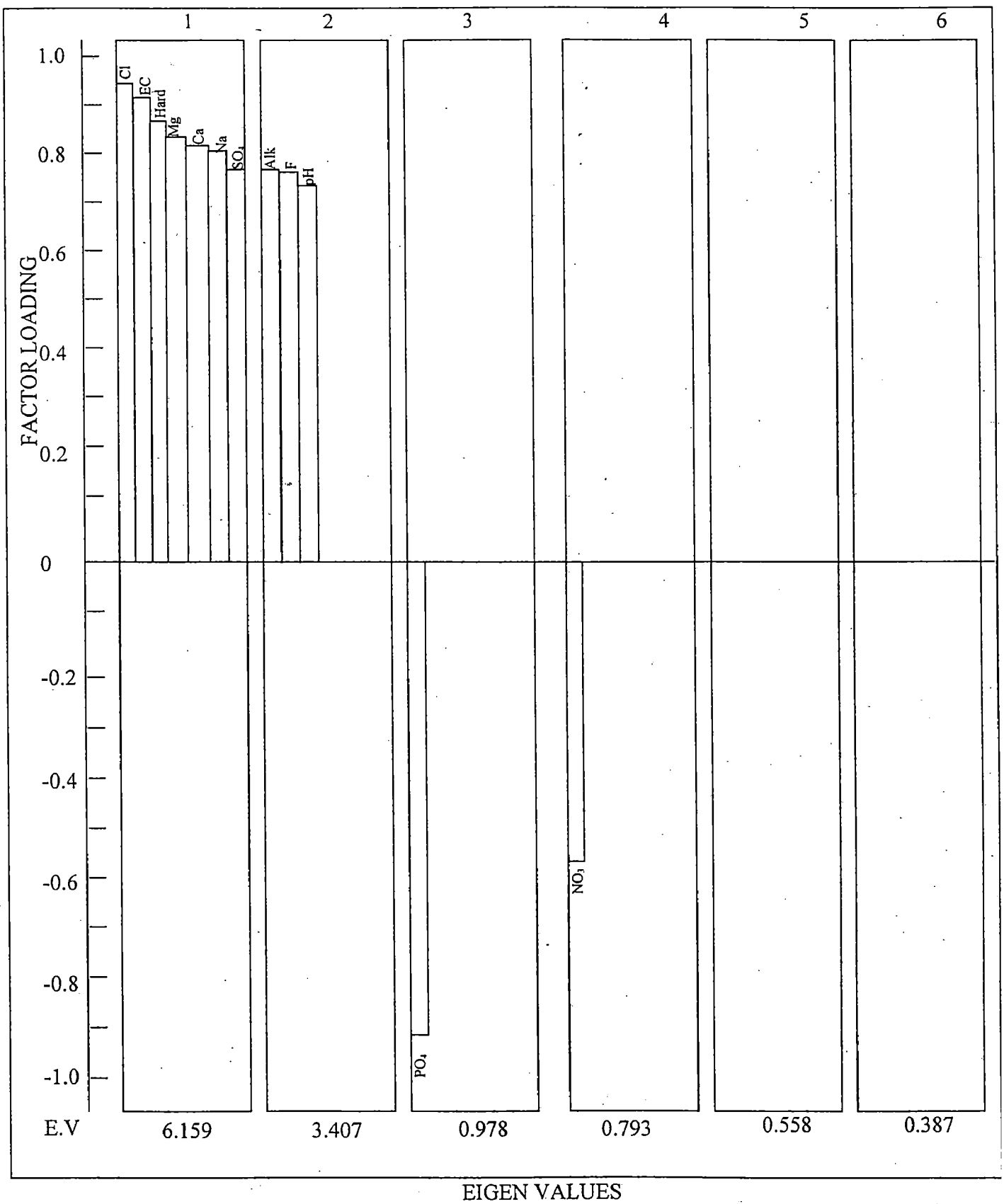


FIG. 6.5: LOADING OF VARIABLES PRE-MONSOON SEASON

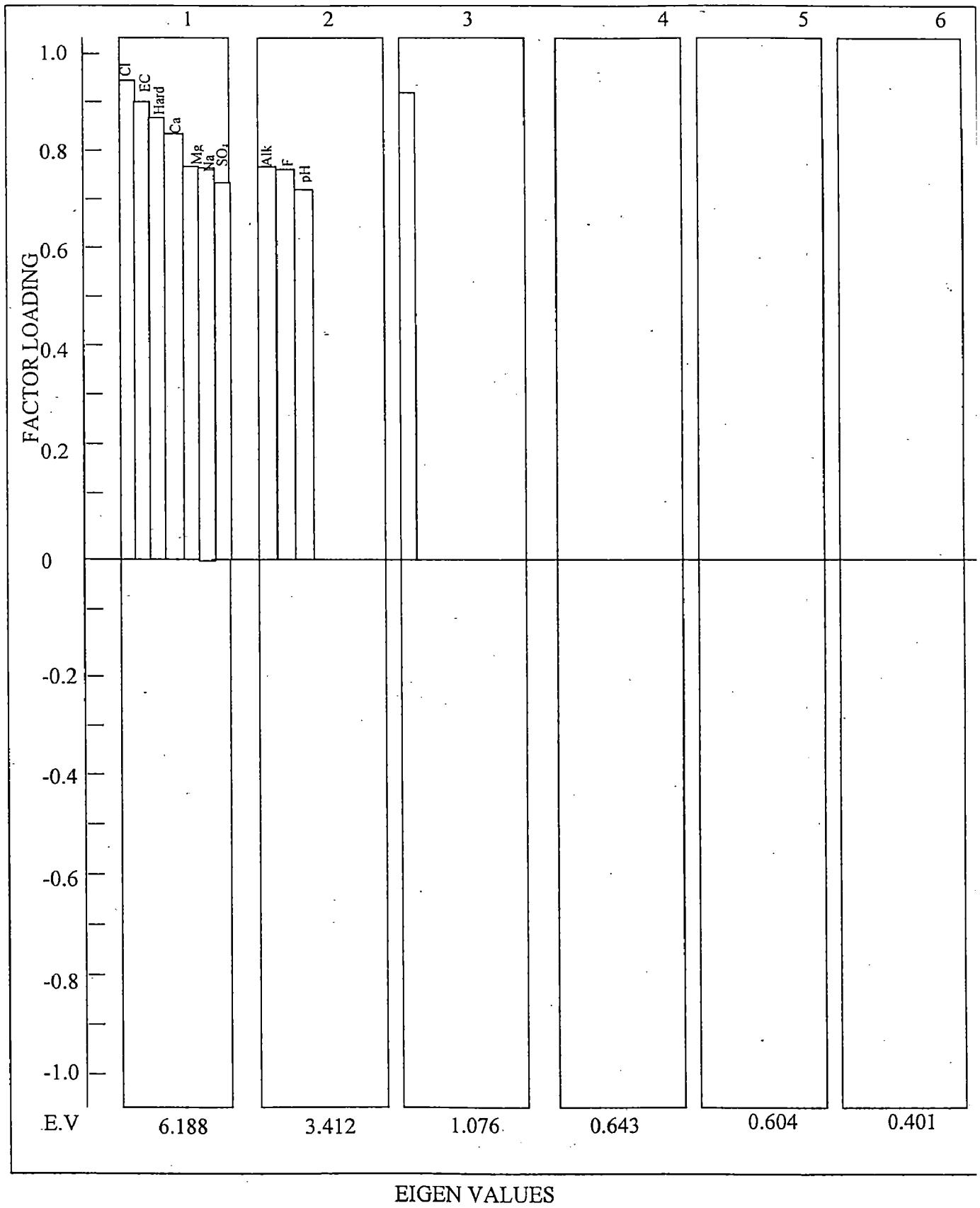


FIG. 6.6: LOADING OF VARIABLES POST-MONSOON SEASON

The first principal component accounts for 47.374% of total variance and is characterized by Cl, EC, Hard, Mg, Ca, Na, and SO₄. High loading of chloride and electrical conductivity were observed which may be called as conductivity factor. Moderate positive loading of Mg, Ca, Na, and sulphate ions indicate the contribution of these ions towards conductivity.

The second principal component which accounts for 26.10% of total variance, mainly loaded on alkalinity, fluoride, and pH ions. The loading for other cation and anion is observed to be negligible. Hence this factor is called fluoride factor. The influence of local lithology and soil added by other factors like very low fresh water exchange due to arid climate of the region is responsible for higher concentration of fluoride in these aquifers.

The third principal component explains 7.521 percent of total variance. This component is highly negatively loaded with phosphate. The loading for all other cation and anion is observed to be negligible. Therefore, this factor is called as phosphate ion factor.

The fourth principal component explains 6.102 percent of total variance is negatively loaded on nitrate with low loading of pH, sulphate and potassium. The loading for other cation and anion is again negligible. Therefore, this factor may be called nitrate factor.

Post –monsoon:

In case of post-monsoon, the first eigen value is 6.188 percent and explains 47.599 percent of total variance, second is 3.412 which explains 26.247 percent of total variance, third is 1.076 which explains 8.275 percent of total variance and so on. The first three components accounts for a total of

82.121 percent of total variance (Table 6.14). The following three components have been interpreted as follows:

Component I:	Conductivity factor
	Cl, EC, Hard, Ca, Mg, Na, SO ₄
Component II:	Fluoride factor
	Alkalinity, F, pH
Component III:	Phosphate factor
	PO ₄

The first principal component accounts for 47.599 percent of total variance and has more or less uniform loading of electrical conductivity and chloride, and may be called as conductivity factor. High positive loading of chloride, calcium, magnesium, sodium, and sulphate ions indicate the higher contribution of these ions towards conductivity in the post-monsoon of this area. As the salinity of ground water is measured in terms of electrical conductivity, the high salinity in these aquifers may be due to alluvial aquifers existing in the area.

The second principal component which accounts for 26.247 percent of total variance, mainly loaded with alkalinity, fluoride, and pH respectively. Moderate loading was observed on sodium. The loading for other cation and anion is negligible. Hence this factor is called fluoride factor. The presence of fluoride ion may be due to dissolution of fluoride bearing minerals present in the study area. Further arid climate of the region and long residence time of ground water in aquifers are also responsible for the higher concentration of fluoride.

The third principal component accounts for 8.275 percent of total variance. The highest loading on phosphate is observed in this component while almost negligible loading on all other cation and anion. Hence it is called phosphate factor.

Thus it can be inferred from the above discussion that EC, Hardness, chloride, nitrate, hydrogen ion, and phosphate may be responsible for the variation in the ground water quality in the pre-monsoon season whereas conductivity, fluoride, phosphate, hydrogen ion, and hardness in the post-monsoon season of the district Jaipur, Rajasthan. By and large the common factors fluoride, nitrate, phosphate have perceptible influence on the quality of ground water of both pre-monsoon and post-monsoon season.

6.2 Artificial Neural Network Analysis

The selected candidate sets/subsets used in statistical model development (Tables 6.4 and 6.10) were further used in development of a model based on the artificial neural network (ANN) theory. The possible sets of combinations for various ANN models are shown in Tables 6.17 and 6.18 for pre-monsoon and post-monsoon season, respectively. In these tables, the models are described as ANN1, ANN2, ANN3, and so on, and their structures shown in the second columns of Tables 6.19 and 6.20. In model development, the same, as above, twenty five data points were used, and the other thirteen data points in model validation. The root mean square error (RMSE), correlation coefficient (CC), and coefficient of efficiency (CE) were used as performance criteria for both training (calibration) and testing (validation). The resulting values of the performance indicators are given in Tables 6.19 and 6.20, and the computed water quality values are depicted in Figs. 6.7 and 6.9

for calibration and Figs. 6.8 and 6.10 for validation, for pre-monsoon and post-monsoon, respectively.

It is seen from Table 6.19 that, for pre-monsoon season, the resulting RMSE is minimum for ANN4 in calibration, whereas it is more than that for ANN3 in verification. In calibration, the values of coefficient of efficiency (CE) for both the models ANN3 and ANN4 are the same as 99%. However, in verification, CE is 97% for ANN3 and it is 95% for ANN4. Thus, both these criteria lead to prefer ANN3 to ANN4. A similar inference can also be drawn if coefficient of correlation (CC) values are considered.

Similar to the above analysis, ANN11 (Table 6.20) can be considered to be the most appropriate model, which indicates CE values equal to 99% in both calibration and validation. Since the CE values are quite high, the proposed models can be described to have performed satisfactorily. It can also be asserted from Figs. 6.7-6.10 showing satisfactory match of the computed and observed values, valid for both pre- and post-monsoon seasons.

Based on R^2 , Table 6.21 compares the performance of both the statistical and ANN models. It is apparent that using the observed concentration quantities of EC, Hard, Cl, and SO_4 , Na can be predicted by both the models reasonably with R^2 varying from 98.2-99.8% in both pre- and post monsoons. It is however noted here that the statistical model utilizes all these four constituents to predict Na in pre-monsoon season. On the other hand, all other models require only three constituents, viz., EC, Hard, and Cl, to predict Na in both the seasons. Thus, both the models perform equally well in Na-prediction, verifying the results derived from both the approaches.

Table 6.17: Description of various ANN models for training and testing of sodium levels for pre-monsoon season.

Model No.	ANN models	No. of training data	No. of verification data
ANN 1	EC	25	13
ANN 2	EC+Hard	25	13
ANN 3	EC+Hard+Cl	25	13
ANN 4	EC+Hard+Cl+SO ₄	25	13
ANN 5	EC+Hard+Cl+SO ₄ +Alk	25	13
ANN 6	EC+Hard+Cl+SO ₄ +Alk+F	25	13
ANN 7	EC+Hard+Cl+SO ₄ +Alk+F+Mg	25	13
ANN 8	EC+Hard+Cl+SO ₄ +Alk+F+Mg+Ca	25	13

Table. 6.18: Description of various ANN models for training and testing of sodium levels for post-monsoon season.

Model No.	ANN models	No. of training data	No. of verification data
ANN 9	EC	25	13
ANN 10	EC+Hard	25	13
ANN 11	EC+Hard+Cl	25	13
ANN 12	EC+Hard+Cl+SO ₄	25	13
ANN 13	EC+Hard+Cl+SO ₄ +Alk	25	13
ANN 14	EC+Hard+Cl+SO ₄ +Alk+Mg	25	13
ANN 15	EC+Hard+Cl+SO ₄ +Alk+Mg+Ca	25	13
ANN 16	EC+Hard+Cl+SO ₄ +Alk+Mg+Ca+F	25	13

Table. 6.19: Comparative performance of selected models for pre-monsoon season for sodium

Model No.	Nodes- input, hidden output	Performance evaluation of models					
		Calibration (Training)			Verification (testing)		
		RMSE	CC%	CE%	RMSE	CC%	CE%
ANN 1	1,2,1	363.02	97.07	66	133.64	91.86	83
ANN 2	2,4,1	73.89	99.37	99	72.62	98.55	95
ANN 3*	3,6,1	62.11	99.56	99	58.55	99.08	97
ANN 4	4,8,1	55.41	99.64	99	73.57	98.09	95
ANN 5	5,10,1	47.51	99.74	99	110.92	94.79	88
ANN 6	6,12,1	44.25	99.77	99	107.74	94.95	89
ANN 7	7,14,1	46.20	99.75	99	103.33	95.55	90
ANN 8	8,16,1	44.38	99.51	99	105.85	94.90	89

**Table. 6.20: Comparative performance of selected models for post-
monsoon season for sodium**

Model No.	Nodes- input, hidden output	Performance evaluation of models					
		Calibration (Training)			Verification (testing)		
		RMSE	CC%	CE%	RMSE	CC%	CE%
ANN 9	1,2,1	273.83	93.51	71	134.86	91.86	83
ANN 10	2,4,1	67.65	99.17	98	47.95	99.05	98
ANN 11*	3,6,1	46.01	99.62	99	33.05	99.51	99
ANN 12	4,8,1	45.05	99.63	99	38.04	99.37	99
ANN 13	5,10,1	42.77	99.67	99	47.38	99.03	98
ANN 14	6,12,1	43.89	99.65	99	50.41	98.86	98
ANN 15	7,14,1	41.74	99.68	99	47.00	99.28	98
ANN 16	8,16,1	40.76	99.70	99	58.26	99.09	97

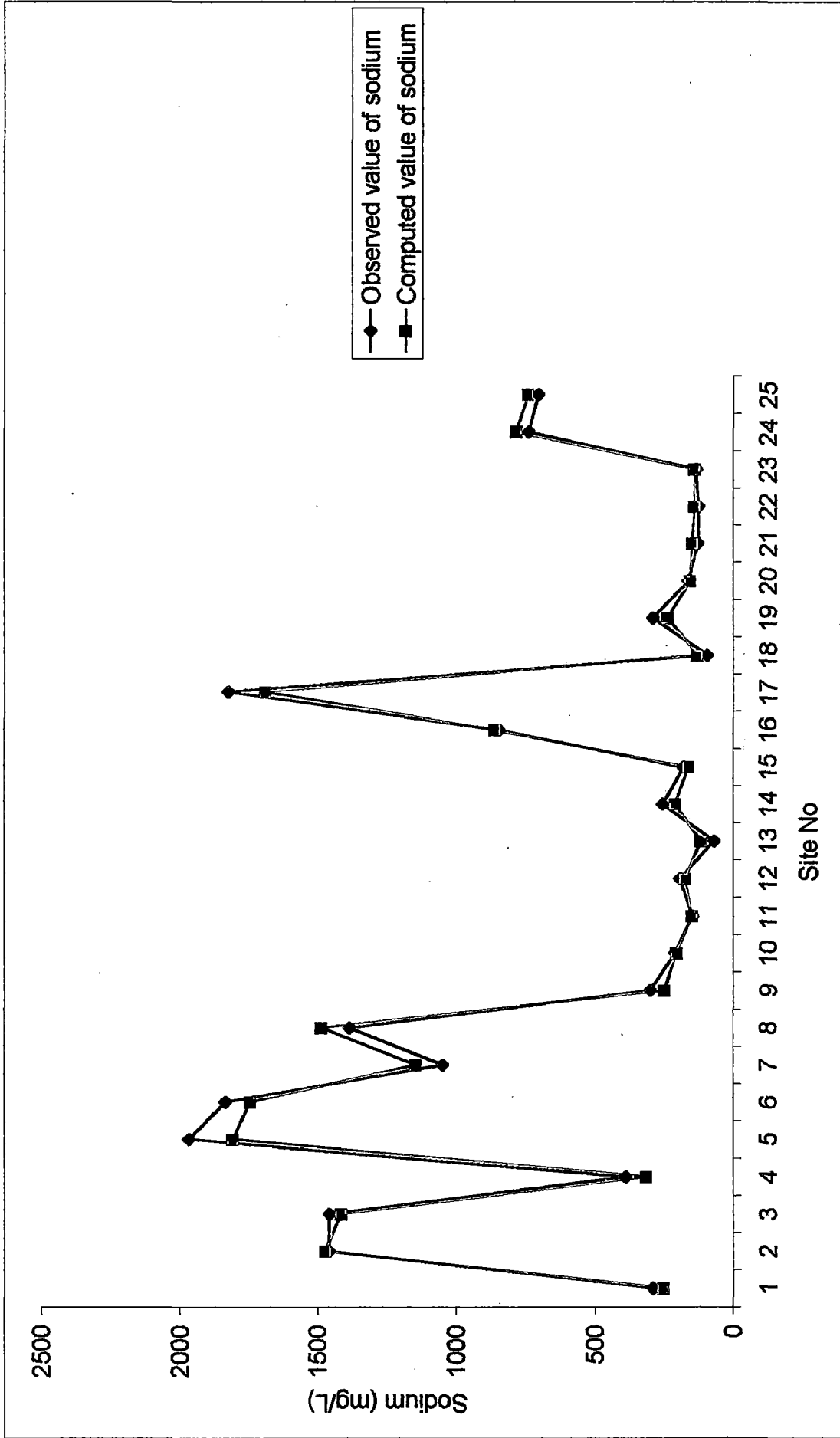


Fig. 6.7: Calibration result of model ANN3 for pre-monsoon.

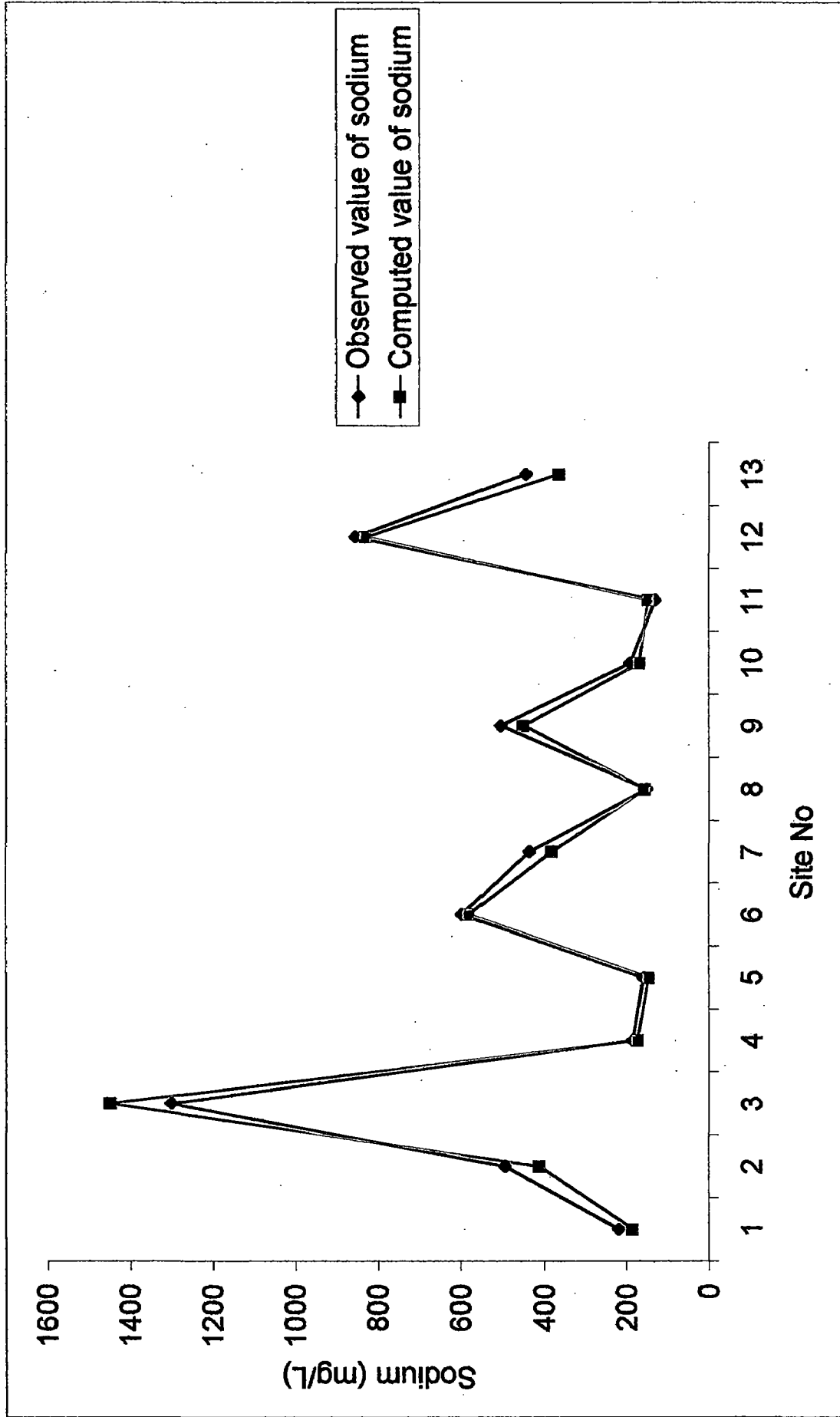


Fig. 6.8 : Validation result of model ANN3 for pre-monsoon.

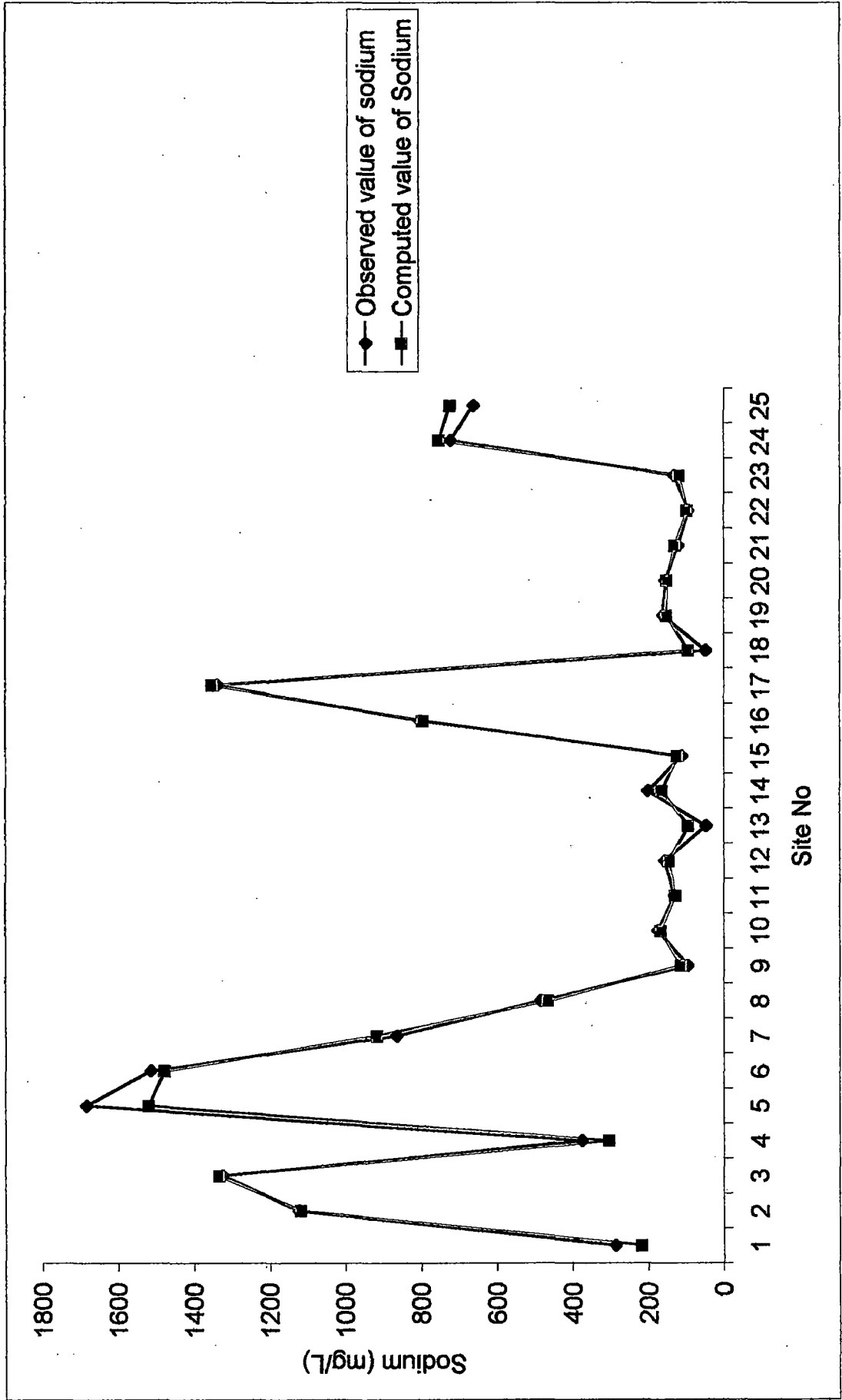


Fig. 6.9: Calibration result of model ANN11 for post-monsoon.

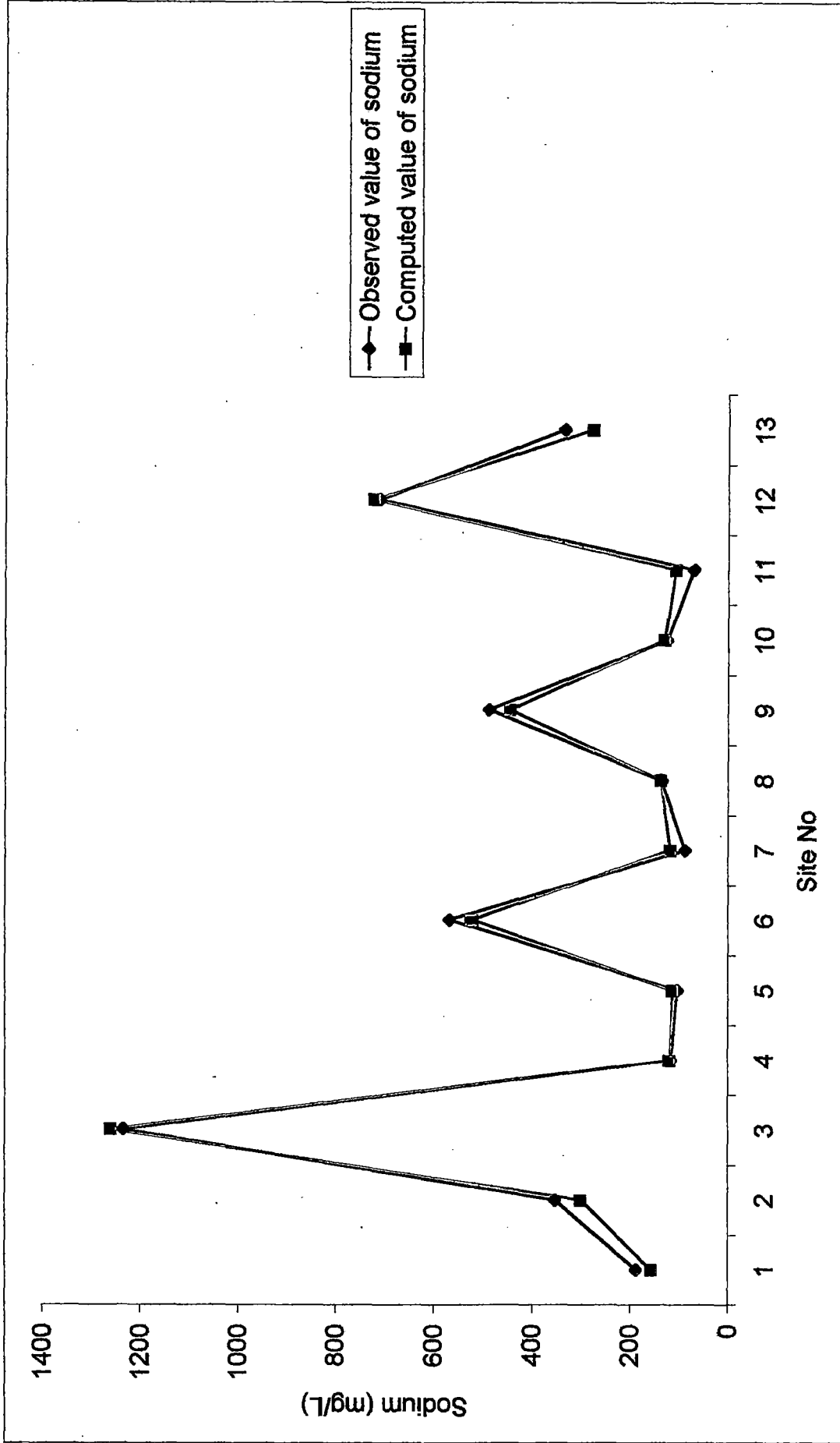


Fig. 6.10: Validation result of model ANN11 for post-monsoon.

Table 6. 21 Comparison of statistical and ANN-Based models

Model	Season	Mode	Resulting R ² -value (%)	
Statistical-based	Pre-monsoon (With 4 inputs: EC, Hard, Cl, SO ₄)	Calibration	99.8	
		Validation	99.5	
	Post-monsoon (With 3 inputs: EC, Hard, Cl)	Calibration	99.8	
		Validation	99.8	
	ANN-based	Pre-monsoon (with 3 inputs: EC, Hard, CL)	Calibration	99.1
			Validation	98.2
Post-monsoon (with 3 inputs: EC, Hard, CL)		Calibration	99.2	
		Validation	99.0	

SUMMARY AND CONCLUSIONS

7.1 Best Subset Procedure

Useful regression models for predicting sodium concentration using other ground water quality constituents were developed for both pre-monsoon and post-monsoon seasons of Jaipur district, Rajasthan. As both of the regression models for pre-monsoon and post-monsoon are successful in explaining about 99% of the variation in the sodium levels, the developed models may be used for the prediction of missing observed values. However, the variability of the results from one season to another indicates that a general model cannot be derived to predict the sodium concentration for both the seasons. Also the variation in sodium causing parameters varies from one site to another due to the regional geomorphic and hydrogeologic features.

7.2 Principal Component Analysis

Principal component analysis was used to predict the dominating water quality constituents and it was revealed that four principal components are accounted for the total chemical variability in the groundwater quality of district Jaipur for pre-monsoon season and three principal components for post-monsoon season, respectively. The common factors conductivity, fluoride, nitrate, alkalinity, and phosphate have perceptible influence on the quality of groundwater of district Jaipur. Higher conductivity represents the salinity of the groundwater, and higher fluoride content may be attributed to distribution of fluoride bearing minerals in the soil, their solubilization characteristics, nature of the product with soil and other environmental conditions. The parameters influencing the concentration of fluoride are observed to be pH

and alkalinity. The higher nitrate content may be attributed to the combined effect of contamination of domestic sewage and runoff from fertilized fields.

7.3 Artificial Neural Network Analysis

Using the steepest descent optimization technique and the sigmoid activation function, the back propagation two layered feed forward ANN models were developed for estimation of sodium for both pre- monsoon and post- monsoon seasons, respectively. The number of iterations was fixed at 1000, and the models were validated with data not used in calibration. The input variables considered for different model structure were identified using correlation analysis. The statistical performance evaluation criteria such as root mean square error (RMSE), correlation coefficient (CC), and coefficient of efficiency (CE) were used to demonstrate the model performance.

Based on the coefficient of determination (R^2), the performance evaluation of both the above statistical and ANN models revealed both of them to perform equally well in both pre- and post-monsoon seasons. In field application, the observations of EC, Hard, and Cl may be used to predict Na in both the seasons.

7.4 Suggestions Proposed For Future Studies

A suitable regression model may be estimated to predict the concentration of fluoride and nitrate in the Jaipur district, Rajasthan because fluoride and nitrate concentration in groundwater is a growing problem of Jaipur district. According to BIS (1991) permissible limit of fluoride is 1.0 – 1.5 mg/l depending on climate, whereas in the study area it varies from 0.07 to 22.4 mg/l with a mean or 2.48 mg/l in pre- monsoon and from 0 to 21.0 mg/l with mean value of 1.79 mg/l in post- monsoon season. Again, according to

BIS (1991), nitrate contamination above 45 mg/l may prove detrimental to human health. In district Jaipur, it varies from 2.4 to 986 mg/l with mean value of 135 mg/l in pre- monsoon and from 1.3 to 800 mg/l with mean value of 122 mg/l in post- monsoon, respectively. Using the same procedure as used in this study, an attempt was also made to develop statistical/ANN model for the prediction of F and NO₃, but to no avail. It is perhaps due to lack of observations in number and/or the constituent actually describing the concentration of F and NO₃.

REFERENCES

1. Agarwal, A. and Singh, R.D. (2003), *Runoff Modelling through back propagation artificial neural network with variable rainfall-runoff data. Water Resources Management 18: 285-300.*
2. Agarwal, A. (2004) *Training of artificial neural network models, In: Training workshop on Artificial Neural Network and its Applications in water resources-organized by National Institute of Hydrology, Roorkee – 247667 at Central Water Commission, Sewa Bhawan, R.K. Puram, New Delhi – 110066, June 09, 2004.*
3. Batisha, A.F. (2005), *Water quality sensing using multiplayer perception artificial neural networks. Researcher, Environment and Climate Research Institute, National Water Research Center, Cairo, Egypt.*
4. BIS, *Specification for Drinking Water, IS: 10500: 1991, New Delhi.*
5. Bowden, G.J., Nixon, J.B, Dandy G.C, Maier, H.R. and Holmes, M. (2005) *Forecasting chlorine residuals in a water distribution system using a general regression neural network. University of Adelaide, Australia (john.Nixon@uwi.com.au)*
6. Chakrapani, G.J and Subramaniam, V. (1993). *Heavy metals distribution and fraction in sediments of Mahanadi river basin, India, Environmental Geology, 22, 80-87.*
7. Davis, C. John. (1973) *Statistics and Data Analysis in Geology, John Wiley and Sons, Inc.*

8. Dawdy, D.R. and Feth, J.H. (1967). *Applications of factor analysis in study of chemistry of ground water quality, Mojave River valley, California. Water Resources Research, 3(2), 505-510.*
9. De Villars, J. and Barnard, E., (1993). *Back Propagation Neural Nets with one and two Hidden layers, IEEE Trans, Neural News, 4 (1), 136-141.*
10. Draper, N.R. and Smith, H. (1990). *Applied Regression Analysis, John Wiley and Sons, Inc.*
11. Evans, C.D, Davis, T.D, Wigington, Jr. P.J, Tranter, M, and Kretser, W.A. (1996) *Use of factor analysis to investigate processes controlling the chemical composition of four streams in the Adirondack Mountains, New York, Journal of Hydrology 185, 297-316.*
12. Garg, S.K. (1994) *Environmental Engineering (Vol. 1), Water Supply Engineering. Khanna Publishers, New Delhi.*
13. Gupta S.C. and Gupta, V.K.(2004). *Fundamental of Mathematical Statistics., Sultanchand Publishers, New Delhi.*
14. Haynes, R. (1980). *Environmental Science Methods, London, New York, Chapman and Hall.*
15. Hornik, K. Stinchcombe, M., and White, H., (1989). *Multilayer Feed Forward Networks are Universal Approximators, Neural Networks, 2 (5), 259-366.*
16. Jain, C.K. and Sharma; M.K. (1997), *Relationship among Water Quality Parameters of Groundwater of Jammu District, Poll Res. 16(4) : 241 – 246.*

17. Jain, C.K. and Sharma, M.K. (2002), *Regression Analysis of Groundwater Quality data of Malprabha River Basin, Karnataka, IWRS. 22 (1), 30-35.*
18. Jain, C.K., Ali Imran, Sharma, M.K. (1998). *Salinity Modelling of Groundwater of a Coastal Region Using Best Subset Procedure. Indian J. Environmental Protection, 18 (10), 762-768.*
19. Jha, S.K. and Jain, A (2005), *Evaluation of ANN technique for rainfall-runoff modeling in a large watershed. Proceedings of the International conference on Hydrological perspectives for sustainable development (HYPESD-2005), 23-25 February 2005, Roorkee, India.*
20. Johnson Richard, A (2001) *Miller & Freund's. Probability and Statistics for Engineers, Prentice Hall of India Private Limited, New Delhi.*
21. Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistic Analysis, Prentice Hall of India Private Limited, New Delhi.*
22. Kannan, N. and Vallinuyagam, P. (1992). *Correlation Analysis of Water Quality Parameters of Industrial Effluents : Match Industry, Indian J. Environmental Protection, 12 (7), 521-527.*
23. Kothiyari, U.C. and Jain, P. (2005). *Estimation of monthly runoff from catchments using artificial neural networks proceedings of the international conference on hydrological perspectives for sustainable development – (HYPESD-2005), 23-25 February, 2005, Roorkee, India.*
24. Krishna, J.S.R., Rambabu, K; and Rambabu, C. (1995). *Monitoring Correlations and Waterquality Index of Well Waters of Reddigudum Mandal, Indian J. Environmental Protection, 15 (12), 914-919.*

25. Kumar, S., Kumar, R., Lohani, A.K., Singh, R.D. (2005). *Short term forecasting using artificial neural networks. Proceedings of the International conference on hydrological perspectives for sustainable development – (HYPESD-2005), 23-25 February, 2005, Roorkee, India, 216-221.*
26. Mahodaya, M.M. (1999). *Compendium for Agricultural & Rural Development Officers and Water Resources Engineers, Krishna Printing, Bhopal.*
27. Mary, T.M. Usha, Nagarajan, S. and Swaminathan, M. A. (1998) *Correlation Study on Physico-Chemical Characteristics Carbonization Wastewater, Indian J. Environmental Protection, 18(9), 647-649.*
28. Mehrotra, K. Mohan, C.K., and Ranka, S. (1997) *Elements of Artificial Neural Networks. Penram International publishing, India.*
29. Montgomery, D. and Peck, E. (1982), *Linear Regression Analysis, John Wiley and Sons, New York.*
30. Neumann, W. David, Rajagopalan, B. and Zagona, A. (2003). *Empirical Regression Model for Daily Maximum Stream Temperature, 10.1061/(ASCE) 0733-9372 (2003) 129:7 (667). This paper is part of the journal of Environmental Engineering, 129 (7) ASCE, ISSN 0733-9372/2003/7-667-674 / S 18.00.*
31. Nolan, B.T., Ruddy, B.C., Hitt, K.J, and Helsel, D.R. (1995) *Nitrate in ground Waters of the United States.*

(www.nwqmc.org/98proceedings/papers/63-NOLAN.htm)
32. Paliwal, K.V. (1972), *Irrigation with saline water, IARI, Monograph-2, Water Technology Centre, IARI, New Delhi.*

33. Peavy Howard, S., Rowe Donald R. Tehobanoglous, G. (1990) *Environmental Engineering*. Mc Graw Hill Int. New York.
34. Puckett , Larry J. and Bricker, Owen P. (1992) *Factors controlling the major ion chemistry of streams in the blue ridge and valley and ridge physiographic provinces of Virginia and Maryland*, *Journal of hydrological processes*, Vol. 6, 79-98, 1992.
35. Raghunath, H.M. (1987) *Ground Water Hydrology*, New Age International (P) Ltd. Publishers, New Delhi.
36. Raghuwanshi, M. Chatterjee, C. Raghuwanshi, N.S. Singh, R, and Kumar, R. (2005). *Flood forecasting using artificial neural networks*. *Proceedings of the International conference on Hydrological perspectives for sustainable development (HYPESD-2005)*, 23-25 February, 2005, Roorkee, India, (187-195).
37. Rambabu, C., Rao, Srinivasa, B., Singanam, M., Ramachandran, D., and Rao Somasekhara, (1998). *Statistical Studies on the Water Quality Parameters of Chirala Town Open Wells, Prakasam District*, *Indian J. Environmental Protection*, 18 (3) 203-209.
38. Reid, J.M., Macleod, D.A. and Cressere, M.S. (1981). *Factors affecting the chemistry of precipitation and river water in an upland catchment*, *Journal of Hydrology*, 50 129-145.
39. Rumelhart, D.E., Hinton, G.E., and Mc Clelland, J.L. (1986). *A general framework for parallel distributed processing*. *Parallel Distributed Processing: Explorations in the Macrostructure of cognition*, 1.
40. Sarkar, A. Agarwal, A. Singh, R.D., and Jain, S.K. (2003) *Artificial neural networks for daily rainfall runoff modeling in satluj basin, India*.

Proceedings of the International conference on Hydrological perspectives for sustainable development (HYPESD-2005), 23-25 February, 2005, Roorkee, India, (196-205).

41. Singanan, M. Somasekhara, K. and Rambabu, C. (1995) *Correlation Study on Physico-Characteristics of Groundwater in Rameswaram Island Indian J. Environmental Protection, 15 (3) 213-217.*
42. St- Hilarie, A, Brun, G, Courtenary, S.C, Quarda Taha, B.M.J., Boghan, A.D, and Bob'ee, B. (2004). *Multivariate analysis of water quality in the Richibucto drainage basin (New Brunswick, Canada) AWRA, 40 (3), 691-703.*
43. Subramanian, S. and Balasubramanian, A. (1994), *Hydrogeochemical studies of Tiruchendur coast, Tamil Nadu, India, Regional workshop on Environment Aspects of groundwater Development, October 17-19, p III-26-III-32.*
44. Tirzo, A.T. (1995), *An integrated hydrogeological study of mahendragarh district, Haryana, India, Ph.D. Thesis. Department of Hydrology, University of Roorkee, Roorkee.*
45. Tyagi, A., Sharma, M.K., and Bhatia, K.K.S. (1998) *Salinity Modelling of Groundwater in Saharanpur and Haridwar District, CS (AR) – 4/97-98, National Institute of Hydrology, Jal Vigyan Bhawan, Roorkee.*
46. Vajarappa, H.C. and Srinvas, G. (1994), *'Hydrogeochemistry of Kabini river basin in Karnataka. Regional workshop on Environmental Aspects of ground water Development, October 17-19, p III-26-III-32.*
47. Weisberg, S. (1980), *Applied Linear Regression, John Wiley and Sons, New York.*