Data Analysis, Quality Indexing and Prediction of Water Quality for the Management of Rawal Watershed in Pakistan

Maqbool Ali, Ali Mustafa Qamar
Department of Computing
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
{*11msitmali, mustafa.qamar*}*@seecs.edu.pk*

***Abstract*—In contrast to managing the water quality only at the command level (where water is being consumed), one should also give importance to the water quality in the areas where water is being produced i.e. the watersheds. The failure to do so deteriorates the water quality for down streams and poses serious challenges for the water managers in order to meet the water quality requirements on sustainable basis. In order to have an effective water management in command areas, it is essential to assess different aspects of water quality. Rawal watershed is a relatively small watershed area which is being affected by the anthropogenic activities e.g. urbanization, deforestation etc. In this paper, we present the last four years (2009 – 2012) trends of water quality related parameters along with month-wise as well as source-wise parametric satisfactory analysis against *WHO* quality standards. Moreover, we applied regression models to check the seasonal water quality trends. The quality indices were analyzed by the combination of supervised and unsupervised machine learning techniques. Different sources of fecal coliforms contamination were also identified. Lastly the possible reasons for high contamination were identified by studying the watershed land covers. Our research suggests that in order to find the quality index of water, *Average Linkage (Within Groups)* method of *Hierarchical Clustering* using *Euclidean distance* is an accurate unsupervised learning technique. Similarly, for classifications, *Multi-Layer Perceptron (MLP)* has been found to be more accurate supervised learning technique. Higher values of fecal coliforms were found in the months of *March, June, July*, and *October*. Some of the possible reasons are land-covers especially *scrub forest* and *rain-fed agriculture* areas, *poultry farms*, and *population* settled around the streams.**

***Index Terms*—Classification, Cluster Analysis, Data Extrapolation, Regression, Statistical Analysis, Water Quality**

I. Introduction

Rawal watershed is a relatively small watershed area located near Islamabad, Pakistan. It has got 272.19 square kms catchment area. It is a strategic asset as it provides drinking as well as agricultural water supplies to the twin cities of Islamabad and Rawalpindi in Pakistan. The stake holders that are directly involved in the management of Rawal watershed are *Rawal Lake Filtration Plant, Water And Sanitation Authority (WASA), Water Resource Management, National Agricultural Research Center, Environmental Protection Agency, Capital Development Authority (CDA)* and *Islamabad Capital Territory (ICT)*.

Rawal Dam is the reservoir of Rawal watershed. It has two bank canals, right and left. The *Right bank canal*, having a capacity of 72 cusics, provides the raw water supply to *Rawalpindi Development Authority (RDA), WASA Rawalpindi* for drinking purpose whereas the *Left bank canal*, having a capacity of 40 cusics, provides water for irrigation purposes only.

In order to effectively measure the water quality, water samples were collected from 13 different locations of Rawal watershed as described in Table I and shown in Fig. 1. From 2009 to 2012, a total of 663 water samples have been collected from different locations (51 samples from each location) that are mentioned in Table I. In order to analyze the water quality, *Rawal Lake Filtration Plant* considered the parameters mentioned in Table II.

For effective management of Rawal Watershed, water quality issues have posed serious challenges for water managers: e.g. finding the areas where chemical and/or biological contamination exists. This, in turn greatly affects the aquatic life (especially fishes) and is also dangerous for public health. Similarly the cost of extensive sample collection prior to determining its physical, chemical and bacteriological characteristics makes it practically infeasible.

By considering all of these factors, this paper has been divided into five parts. In Part I, we present initial month-wise quality parameters trends. In Part II, parametric satisfactory analysis is given in which source-wise and month-wise analysis is performed against *World Health Organization* (WHO) quality standards. In Part III, we have pre-processed our data and removed all the outliers. This was followed by the development of regression models to check the seasonal water quality trends based on monthly and quarterly datasets. In Part IV, we found the best quality index using different clustering techniques. In the last phase, we have found the months in which fecal coliforms contamination is high, the streams / sources which have high contamination, and the possible reasons for this higher contamination. Moreover, we know of no previous work whereby data mining has been applied to Rawal watershed data.

978-1-4799-0615-4/13/$31.00 ©2013 IEEE

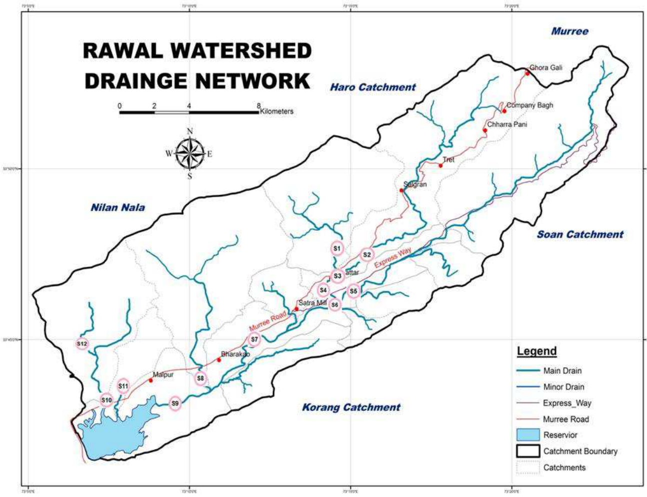

Fig. 1. Rawal Watershed Drainage Network

TABLE I
LOCATION OF STREAMS

Source	Location
S1	Stream coming from Bidhawa Village
S2	Stream coming from Ghora Gali
S3	At Junction of Bidhawa Village and Ghora Gali Streams
S4	At Junction near DESTO Lab and Upstream of Chattar Park
S5	Korang River before Chattar Park near Bahria Town
S6	At Junction of Korang River and Chattar Park Streams
S7	At Downstream Chattar Park and Upstream of Bara kahu
S8	Stream coming from Shahdra
S9	Korang River before entering Rawal Lake
S10	Stream coming from Bari Imam and Diplomatic Enclave
S11	Stream coming from Quaid-e-Azam University
S12	Stream coming from Bari Imam at Noor Pur Shahan
Reservoir	Rawal Lake

TABLE II
WATER QUALITY PARAMETERS

Category	Parameter	WHO Limits
Physical and Chemical	Appearance	Clear
	Temperature	°C
	Turbidity	5 NTU
	pH	6.5 - 8.5
	Alkalinity	500 mg/l
	Hardness as $CaCO_3$	500 mg/l
	Conductance	2000 μS/cm
	Calcium	200 mg/l
	Total Dissolved Solids	1000 mg/l
	Chlorides	200 mg/l
	Nitrite as NO_2^-	<1 mg/l
Bacteriological	Fecal Coliforms	Nil Colonies/100ml

This paper is organized as follows: Section II discusses the state of the art related to the current research. Section III presents the proposed methodology where as the results are given in Section IV along with detailed discussions. Section V concludes the paper along with future directions.

II. LITERATURE REVIEW

Related to Part I, IV and V; Kauffman et al. [2] reports on water quality trends in the Delaware River Basin, USA. They categorized the water quality on the basis of forest area that good water quality correlates with high amounts of forest area (greater than 50%) and poor water quality correlates with high amounts of cultivated land. They evaluated the water quality trends along 15 monitoring stations and compared water quality changes with watershed influences such as stream flow, seasonality, land use, and point source pollutants.

Related to Part III, Sessoms [3] discussed the application of linear regression for future prediction using *SPSS* (Statistical Package for the Social Sciences). He found the P-values, beta scores, R^2, mean and standard deviation parameters that helped to learn good models for future prediction. Gamble et al. [4] described different approaches for developing multivariate analysis models for assessing relationships between spatio-temporal physical attributes of the watershed and water quality conditions in monitored as well as unmonitored streams.

Related to Part IV, Karamouz et al. [5] presented a methodology for river water quality zoning using methods of *c-mean crisp classification* and a *fuzzy clustering* scheme to support decision-making in order to help river water quality management in the region. Robert et al. [6] applied *Artificial Neural Networks (ANNs)* to monitor the surface water quality and presented the literature review of water resources with respect to AI techniques. Govindaraju [7] reported that among machine learning techniques, *ANN* is the one that is widely used in various water-related researches, particularly in hydrology. Similarly, Roz [8] used *MLP* algorithm along with a data-driven approach for analyzing water quality. Regarding classifications. Serge et al. [9] and Neely [10] presented how predictive modeling based on training data can be carried out with learning techniques.

Related to Part V, Preis et al. [11] presented a new approach for identification of source contamination in water distribution systems through a coupled model trees-linear programming algorithm. Babin et al. [12] have proposed a novel solution to improve the early detection of water contamination events and their associated health effects using Belief Networks (BNs).

III. METHODOLOGY

In order to have a quality water data, data pre-processing is carried out. In our analysis, we filled the missing values by the attribute mean and found extreme outliers by box-plot analysis. Outliers from each parameter were removed and then replaced with their medians. Sometimes, there were more than one sample in a month. In that case, we took the mean of that particular month. After data pre-processing, different data mining techniques were applied on the data including correlation analysis, scattered plots for data distribution, regression models using Curve Estimation techniques, and clustering in order to find the quality index. For clustering, *K-Means* as well as *Hierarchical* techniques using different intervals were applied. For classifications, Nearest Neighbors using *Euclidean* and *City-block* distances; Neural Networks using *MLP* and Radial Basis Function, and Support Vector Machines techniques are used [14]-[15].

For finding water quality index, different techniques have been used in literature to categorize the water quality e.g. categorization of quality were made on the basis of forest area [2]. Similarly, water quality zoning was made using *Hard C-mean Crisp Classification* and *Fuzzy Clustering* [5]. *MLP* has also been used for analyzing water quality [6]. In all of these techniques, no one used a combination of unsupervised and supervised machine learning techniques. In our study, we focused on the combination of techniques to find the best possible quality indices.

For regression analysis, scatter plots with trend lines technique was used [13]. Regression is a type of *predictive calculation*, used to predict and forecast the relationships between dependent and independent variables. In our case, linear and cubic models were used. Linear and cubic regressions were found using response variable y and predictor variable x.

$$y = w_0 + w_1 * x \text{ (Linear)} \tag{1}$$

$$y = w_0 + w_1 * x + w_2 * x^2 + w_3 * x^3 \text{ (Cubic)} \tag{2}$$

where w_0, w_1, w_2,and w_3 are regression coefficients.

For measuring prediction accuracy, loss function i.e. *Relative Absolute Error (RAE)* has been used. Predictor accuracy measures how far off the predicted value is from the actual known value. Loss function measures the error between actual and predicted values. The value of *RAE* lies in between 0 to 1. Lesser the error value, more will be the accuracy of prediction [1].

$$\text{Relative Absolute Error} = \frac{\sum_{i=1}^{d} |y_i - y_i'|}{\sum_{i=1}^{d} |y_i - \bar{y}|} \tag{3}$$

where y_i represents the actual value, y_i' is the predicted value, $\bar{y}$ is the mean of actual values and d is the total number of values.

IV. Results and Discussions

This section presents detailed results along with relevant discussions.

A. Initial Data Trends

In order to perform extensive data analysis, water samples have been collected from *Rawal Lake Filtration Plant, WASA* and *National Agricultural Research Centre*. Here we have used line and area graphs for finding the initial trends. By drawing all parameters, we found that from 2009 to 2010, there are more variations in values as compared to 2011 and 2012. After-wards, in order to find the month-wise parametric trends, we grouped our data based on streams and months. Initially we have found that *S5, S7* and *S9* streams have high turbidity values, while *S4, S6* and *Reservoir* have moderate turbidity. Moreover, it was also found that in February, *S5, S7* and *S9*; in May, *S5*; in July, *S5, S6, S7* and *S9*; in August, *S7* and *S9*; in September, *S5* and *S6*; in October, *S4* streams have high turbidity values as shown in Fig. 2.

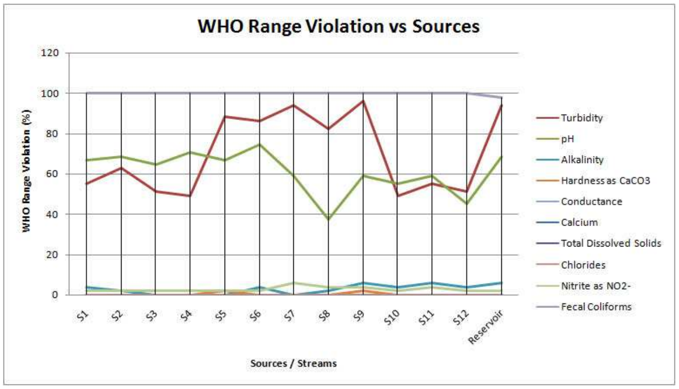

Fig. 3. WHO Range Violation trends of parameters against Sources

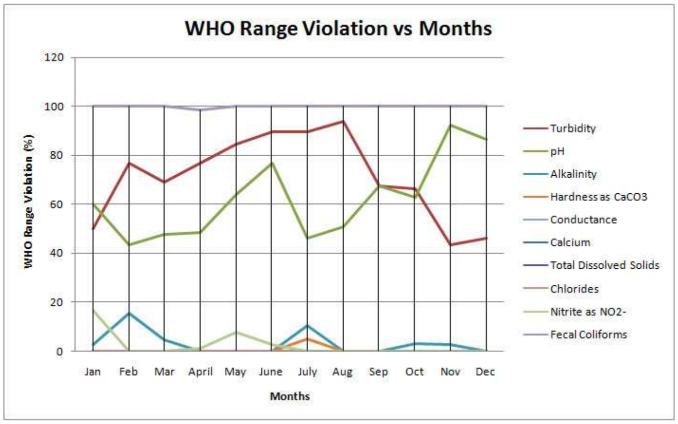

Fig. 4. WHO Range Violation trends of parameters against Months

Due to the limitation of space, the trends for the remaining parameters are briefly described in Table III.

B. Parametric Satisfactory Analysis

In order to determine how much a parameter satisfies the WHO limits, parameter violation percentages on the basis of sources and months are calculated as shown in Fig. 3 and Fig. 4 respectively. These figures show that the values of Alkalinity, Hardness as $CaCO_3$, Conductance, Calcium, Total Dissolved Solids, Chlorides and Nitrite as NO_2^- mostly lie within the WHO limits, while Turbidity, pH and Fecal Coliforms violate the limits. Fig. 3 shows that *S5, S6, S7, S8, S9* and *Reservoir*'s turbidity deviates more from the WHO limits as compared to the other streams. Similarly, Fig. 4 shows that the turbidity violation trend increases till August while decreases later till December. Similarly, the violation percentage of pH is high in November and December.

C. Data Pre-Processing

The missing data related to the streams is initially filled by their streams' mean for that particular month. After filling the missing values, outliers are found by box-plot analysis. In

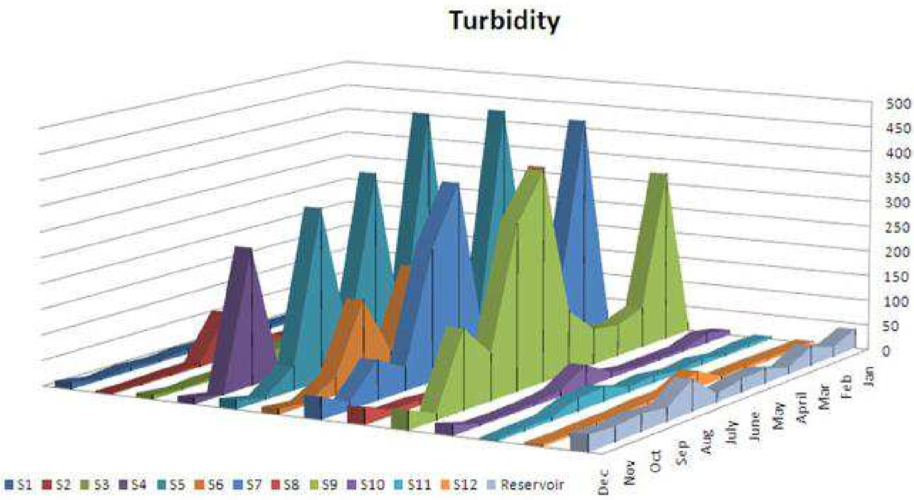

Fig. 2. Month-wise Turbidity Trend of Streams

TABLE III
MONTH-WISE PARAMETRIC TREND

Parameter	Trend
pH	High in June and November
Alkalinity	High values in February as compared to other months
Hardness as $CaCO_3$	High in July as compared to other months
Conductance	*S10, S11* and *S12* streams have high conductance in June; while *S4* stream in September and October
Calcium	*S12* stream has high Calcium contents in all months except January and February as compared to other streams
Total Dissolved Solids	*S10*, *S11* and *S12* streams have high dissolved solids in June; while *S4* stream in September, October and December
Chlorides	*S10*, *S11* and *S12* streams have high chlorides in June; while *S4* stream in September, October and December
Nitrite as NO_2^-	All streams have high Nitrite contents in January, April and May
Fecal Coliforms	*S5, S7, S8, S10, S11* and *S12* streams have high fecal coliforms especially in the last quarter of the year

box-plot analysis, we found the extreme outliers that are represented by asterisks in Fig. 5. The extreme outliers represent cases that have values more than three times the height of the boxes. Outliers from each parameter were removed and then replaced with their medians. In some instances, more than one sample were taken in a month. In that case, we considered the mean for that month to remove the biasness of data.

D. Regression Models

The correlation analysis helps in determining the regression models that will be useful for prediction. While generating regression models based on monthly and quarterly datasets, R^2 parameter is considered which represents the goodness of fitting the model. Its value ranges from 0 to 1. Large values of R^2 i.e. close to 1 represent that the chosen model fits the data well [1]. These models will help to find the relationship between parameters and to predict the quality parameter.

1) Monthly Dataset Regression Models: After finding the correlations based on monthly datasets, the best possible regression models showing their equations and goodness of fit are generated. One of them is shown in Fig. 6. It can be observed that *Total Dissolved Solids* and *Conductance* parameters are highly correlated, which concludes that these parameters are directly proportional to each other.

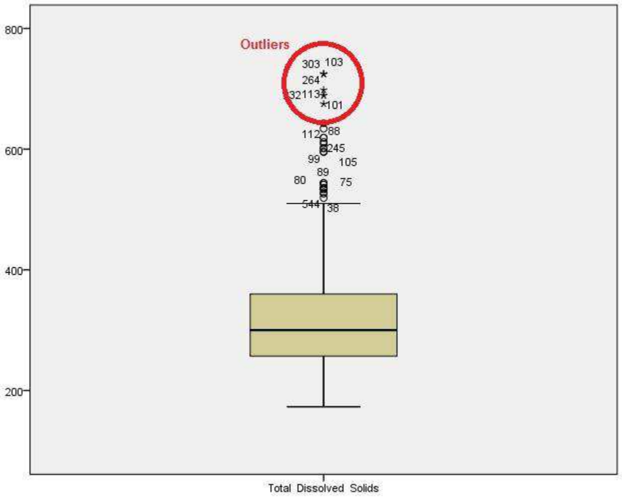

Fig. 5. Box-plot of Total Dissolved Solids Parameter

2) Quarterly Dataset Regression Models: In order to find the seasonal water quality conditions, quarterly datasets were prepared. After preparation, correlations were calculated and this time, it was observed that values of correlation coefficient are higher than monthly dataset values. Moreover, few more

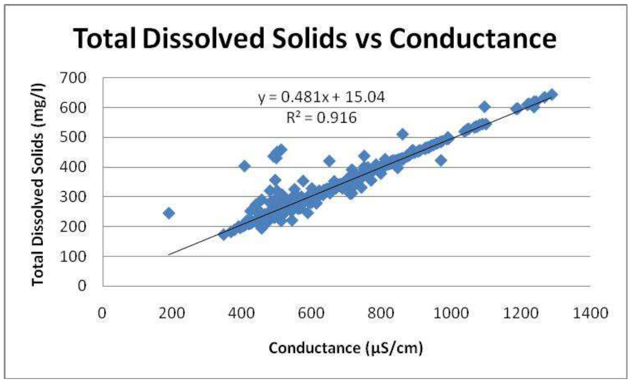

Fig. 6. Regression Model b/w Total Dissolved Solids and Conductance

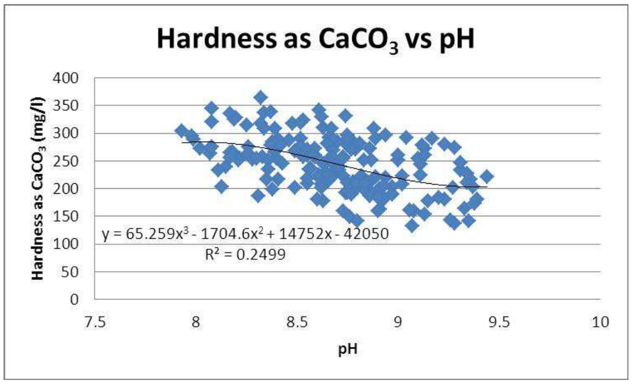

Fig. 7. Regression Model b/w Hardness as $CaCO_3$ and pH

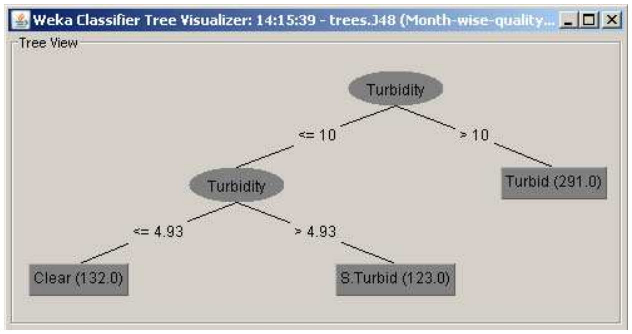

Fig. 8. Turbidity Classification Tree

significant correlations but with smaller values of R^2 were found like correlations with *pH* and *Temperature*. One of them is shown in Fig. 7, which shows that *Hardness as* $CaCO_3$ and *pH* parameters are negatively correlated.

E. Quality Indexing

There is no *Quality Index* in our data and moreover, it is quite difficult to find one by any equation. Initially we have assumed the category variable *Appearance* in our data as the quality index. This is followed by application of other classification techniques using 3, 5, and 10 fold cross-validation in order to find better accuracy. It was found that 10 fold cross-validation produces better results and only *Turbidity* attribute is involved as shown in Fig. 8. A decision tree is depicted that is generated by *J48* algorithm and it only describes the rules of *Turbidity*.

In the previous discussions of quality index, we found that all of the rules are based on *Turbidity* and *Appearance* only, while all of the remaining physical and chemical parameters were ignored. Therefore *Appearance* attribute cannot be an index variable. We tried to resolve this problem by using clustering technique, whereby clusters can represent the quality index. In our case, we have to categorize the water quality into *good, fair* and *bad*. To do that, we need three clusters, in which case each cluster will represent each quality category. We have used *K-Means* and *Hierarchical Clustering* algorithms by considering different methods and intervals to cluster the physical and chemical parameters.

In order to find which clustering technique produces the best quality index, we assumed cluster membership as a label and consider the variable; consisting of all cluster memberships; as a dependent variable, and all physical and chemical parameters as an independent variable. After this assumption, we applied different classification techniques, as shown in Table IV, in order to find the best clustering algorithm for determining the quality index. By comparing all classification techniques, we found that *MLP* produced more accurate results as compared to the others and *Average Linkage (Within Groups)* method of *Hierarchical Clustering* using *Euclidean* distance was the best algorithm to find the quality index of all physical and chemical parameters of water.

F. Prediction

WASA is interested in studying the behavior of *Bacteriological Parameters* especially *Fecal Coliform* which always violate the *WHO* range as discussed in *Parametric Satisfactory Analysis*. For that purpose, we grouped our fecal coliforms data by considering months and sources. These grouped data helped us to find the months and sources in which fecal coliforms are high as shown in Fig. 9. Then we found that in all streams, fecal coliforms are high in March, June, July and October, and *S1, S7, S9, S11* streams and *Reservoir* normally have high fecal coliforms in all months; while *S7* stream has the most fecal coliforms than any other stream.

We also studied the *Rawal Watershed Land Cover* so as to find any occurrence of high fecal coliforms. We found that *S1* stream lies in Scrub Forest area; *S7* stream lies in Rainfed Agriculture area while *S9* and *S11* streams lie in Rangelands area. Moreover, it was also found that there are more poultry farms around *S1* stream, and as far as the *S7* stream is concerned; there is more population settled around it. These land-covers, poultry farms, and population may be the possible reason for high fecal coliforms.

V. Conclusion

Rawal watershed area is a strategic asset as it feeds drinking and agricultural water supplies to the twin cities of Islamabad and Rawalpindi in Pakistan. For the effective management,

TABLE IV
CLASSIFICATION MODELS ACCURACY IN PERCENTAGE (10-FOLD CROSS-VALIDATION)

Clustering Scheme			**MLP**		**RBF**		**kNN**				**SVM**
			Training (70%)	***Testing (30%)***	***Training (70%)***	***Testing (30%)***	***Training (Euclidean)***	***Holdout (Euclidean)***	***Training (City-block)***	***Holdout (City-block)***	
K-Means	No Update		100.0	97.4	90.0	95.2	85.9	86.0	89.3	89.0	91.4
	Running Means		100.0	97.7	90.2	91.1	86.4	87.1	89.4	91.8	92.5
Hierarchical	**Average Linkage (Within Groups)**	**Euclidean**	**100.0**	**98.9**	**93.2**	**92.2**	**93.0**	**89.8**	**95.1**	**93.0**	**94.1**
		Squared Euclidean	97.9	95.9	89.3	85.2	88.7	87.3	87.0	88.1	92.5
		Cosine	99.2	94.8	82.6	83.3	85.5	86.7	87.7	87.8	87.4
	Complete Linkage	Euclidean	97.6	97.0	85.0	86.1	83.4	84.6	86.8	83.7	83.0
		Squared Euclidean	97.4	90.7	86.6	86.1	80.7	82.6	85.6	87.3	83.0
		Cosine	99.0	96.1	81.2	79.0	82.3	83.3	85.6	86.4	88.3
	Ward Linkage	Squared Euclidean	98.4	95.7	92.5	88.4	86.9	84.8	88.1	88.7	91.2

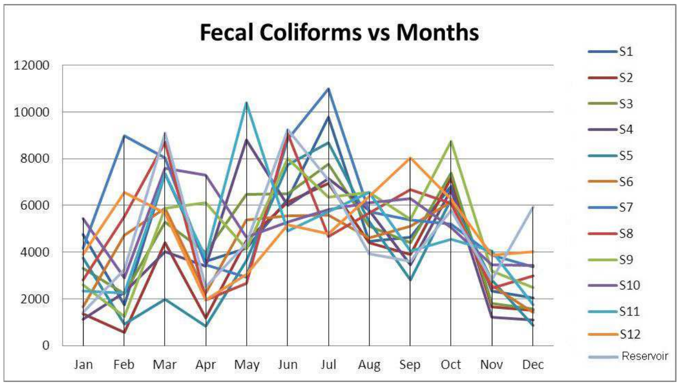

Fig. 9. Trends of Fecal Coliforms in Streams

modeling techniques will help for decision making for the improvement of water quality. In this regards, initial data trends, parametric satisfactory analysis, regression models, finding the quality index, and finally finding the source of contamination as well as the possible reasons for high contamination are found. For finding quality index of water, *Average Linkage (Within Groups)* method of *Hierarchical Clustering* using *Euclidean* distance performed better than other techniques. Similarly, for classification, *MLP* produces more accurate results as compared to its counterparts. In Rawal watershed streams, fecal coliforms are high in March, June, July and October. Similarly *S1, S7, S9, S11* Streams and *Reservoir* have high fecal coliforms in all months as compared to the other streams. Also, *S7* stream has the most fecal coliforms than any other stream. Land-covers especially *rain-fed agriculture area*, *poultry farms* and *population settled around streams* may be the primary reasons of these high fecal coliforms.

In future, the forecasting of fecal coliforms in *S1, S9, S11*, and especially *S7* streams will be done using different time series forecasting models.

ACKNOWLEDGMENT

The authors would like to thank the Principal Investigator, *Water Resource Research Institute - NARC*; Assistant Director, *Water Filtration Plant*, and Deputy Director, *WASA* for providing the data as well as fulfilling other requirements. We are also acknowledged to Mr. Bilal Ali for his valuable suggestions.

REFERENCES

[1] M. Ali, H. Qureshi, and M. S. Akhtar, *Analysis of growth in Students Intake and Degree Awarding Contribution: A Comparison of Stanford and MIT*, MLDM 2013: International Conference on Machine Learning and Data Mining, in press

[2] G. J. Kauffman, A. R. Homsey, A. C. Belden, and J. R. Sanche, *Water quality trends in the Delaware River Basin (USA) from 1980 to 2005*, Environ Monit Assess (2011) 177: 193-225

[3] C. Sessoms, *Statistical Analysis Research Papers*, Statistical Analysis SPEA Vol-506, 2010

[4] A. Gamble, and M. Babbar-Sebens, *On the use of multivariate statistical methods for combining in-stream monitoring data and spatial analysis to characterize water quality conditions in the White River Basin, Indiana, USA*, Environ Monit Assess (2012) 184:845-875

[5] M. Karamouz, N. Mahjouri, and R. Kerachian, *River Water Quality Zoning: A Case Study of Karoon and Dez River System*, Iranian Journal of Environmental Health Science & Eng., Vol. 1, No. 2, 2004, pg. 16-27

[6] R. O. Strobl, and P. D. Robillard, *Artificial Intelligence Technologies in Surface Water Quality Monitoring*, International Water Resources Association, Water International, Vol. 31, No. 2, 2006, pg. 198-209

[7] R. S. Govindaraju, *Artificial neural network in hydrology*, ASCE, Journal of Hydrologic Engineering, 5(2) (2000) 115-137

[8] E. P. Roz, *Water quality modeling and rainfall estimation: A data driven approach*, MS Thesis, University of Iowa, 2011

[9] S. Herzog, *Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis--Vis Regression*, New Directions for Institutional Research, Vol 2006, Issue 131, pg. 17-33

[10] R. Neely, *Discriminant Analysis for Prediction of College Graduation*, Educational and Psychological Measurement, Vol. 37, 1977

[11] A. Preis, and A. Ostfeld, *Contamination Source Identification in Water Systems: A Hybrid Model Trees-Linear Programming Scheme*, Journal of Water Resources Planning and Management 132, SPECIAL ISSUE: Drinking Water Distribution Systems Security, 263273, 2006

[12] S. M. Babin, H. S. Burkom, Z. R. Mnatsakanyan, L. C. R. Thomas, M. W. Thompson, R. A. Wojcik, S. H. Lewis, and C. Yund, *Drinking Water Security and Public Health Disease Outbreak Surveillance*, Johns Hopkins APL Technical Digest, Vol. 27, No. 4 (2008) 403-411

[13] J. Flowers, *Forecasting with Trend Lines using Microsoft Excel*, 2012. http://jcflowers1.iweb.bsu.edu/rlo/trendlines.htm

[14] 2011, *Curve Estimation, K-Means Clustering, Hierarchical Clustering, kNN, Multi-Layer Perceptron, and Radial Basis Function*, http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann Publishers, 2005