Sustainable Water Resources Management (2020) 6:29
https://doi.org/10.1007/s40899-020-00377-9

ORIGINAL ARTICLE

Development of multiple linear regression model for biochemical oxygen demand (BOD) removal efficiency of different sewage treatment technologies in Delhi, India

Prerna Sharma[1] · Smita Sood[2] · Sudipta K. Mishra[3]

Received: 9 September 2018 / Accepted: 31 January 2020 / Published online: 11 March 2020

Abstract
Among the various modeling techniques applied to dataset, multiple linear regression (MLR) analysis is the most efficient way to figure out the relationship between the response variable and the predictive variables. This study emphasizes on establishment of multiple linear regression models to analyze Biochemical Oxygen Demand (BOD) removal efficiency for technologies, namely Densadeck, Extended Aeration and Activated Sludge Process. Assumptions of multiple linear regression like linear relationship, multivariate normality, multicollinearity and Homoscedasticity were examined. The data that verify the assumptions were analyzed with multiple linear regression. Time series plots indicate drastic decline in BOD removal efficiency in the month of Feb and March during the years 2012 and 2013. This study was significant as it gives the technology having the best-fit regression equation based upon multiple correlation coefficient (R), coefficient of determination (R^2), standard error, residual and F-ratio value. Societal benefits include enhancement in the performance of sewage treatment plants.

Keywords Multiple linear regression (MLR) analysis · Biochemical oxygen demand (BOD) · Sewage treatment plants (STP's) · BOD model · Linear relationship · Multicollinearity

Introduction

Municipal Corporation usually takes care of the various sewage treatment plants (STPs). In Delhi, the same has been taken up by Delhi Pollution Control Committee (DPCC) as well as Delhi Jal Board (DJB). Study of the quality of effluents coming from these STPs is not only important as it is disposed of in inland surface water but also because it can be reutilised for the irrigation purposes (DPCC 2016). Biochemical oxygen demand (BOD) and chemical oxygen demand (COD) are commonly known as the potential representative parameters for sewer water quality and valuation of organic matter in sewage (Hur et al. 2010). Majorly predicting the COD value or developing model for the same is considered in industrial waste water rather than domestic waste water (Abyaneh 2014).

The quality of the effluent is dependent on the relations between the various physiochemical parameters interacting with one another. Positive or Negative relationship between the physiochemical parameters, directly trigger the impact on the effluent. Hence, in other words, we can say that it is important to examine which specific parameter has the maximum impact on the effluent; precisely, which independent parameter is more influential in determining the performance of the dependent variable (Najah et al. 2009). Hence, the various models have been established to predict the impact of explanatory variables (independent variables) on outcome variable, i.e., dependent variable (Dogan et al. 2008).

Development of models can be done by multiple linear regression as well as by various multivariate modeling technologies. Multivariate techniques are also used worldwide as they are efficient in assessing the potential parameters

✉ Prerna Sharma
prerna.sharma@gdgoenka.ac.in; prernaenv1701@mail.com

[1] Department of Basic and Applied Sciences, School of Engineering, G D Goenka University, Gurugram, Haryana 122103, India

[2] Department of Basic and Applied Sciences, School of Engineering, G D Goenka University, Gurugram, Haryana 122103, India

[3] Department of Civil Engineering, G D Goenka University, Gurugram, Haryana 122103, India

affecting the wastewater treatment technologies, and further help in deciding the performance and management related to wastewater/sewage or water quality (Vega et al. 1998; Yerel and Ankara 2012; Wang et al. 2014). Multiple linear regression (MLR) analysis is the most efficient tool which is utilized to determine the relationship between the explanatory variable and the outcome variable. Many researchers have used this tool in different educational fields (Fedotovai et al. 2013; Noller and Whitehouse 1982; Moustris et al. 2012). A study was conducted to verify the influence of STPs with respect to their working units through MLR model, and the results obtained revealed that the model was appropriately predicting the variances of the actual observed values, but the study did not focus on developing BOD model as the function of independent variables, i.e., predictive variables (Seung et al. 2014). In a similar study of Sfax STP, descriptive and multivariable analyses were performed on the parameters and it was concluded that the MLR model allows a more efficient process control (Belhaj et al. 2014). A study was also held to investigate the linear regression model of total coliform (TC), fecal coliform (FC) and enterococci (ENT) responses in the storage system of sewage effluents at different temperatures (room temperature 25 ± 2 °C, 55 and 65 °C); from the results obtained, it can be concluded that the storage system of sewage effluents has a significant potential for the reduction of indicator bacteria (Al-Gheethi et al. 2017). Researchers also conducted study on effectiveness of selected wastewater treatment plants in Yemen for reduction of fecal indicators and pathogenic bacteria in secondary effluents and sludge, and also on the elimination of enteric indicators and pathogenic bacteria in secondary effluents and lake water by solar disinfection (SODIS) (AI Gheethi et al. 2014; AI Gheethi et al. 2013).

Many researchers have also worked on evaluating the efficiency of various STPs in Delhi by primarily focussing on calculating the integrated efficiency and comparing the same with the standard integrated efficiency (Jamwal et al. 2009; Colmenarejo et al. 2006). STPs with different sewage treatment technologies were taken into consideration in those studies but the BOD model for the same has not been developed so far. A study was conducted emphasizing on the sensitive analysis of water quality for Delhi stretch of the River Yamuna which also focused on the development of certain model for BOD. Results of the same proved that parameters $K1$ (deoxygenation constant) and $K3$ (settling oxygen demand) are the most sensitive parameters for the considered river stretch. But BOD model in terms of effluent coming from various STPs using different sewage treatment technologies was not taken into consideration (Parmar and Keshari 2012). Similar kind of study was conducted in Korea which includes MLR analysis along with the use of the some of the multivariate tool (Zihan et al. 2018), but still in Delhi such kind of analysis all together is not been done so far with respect to the different sewage treatment technologies utilized in different STPs. The importance of the study lies in the fact that it will clarify which technology gives the best-fit regression equation based on multiple correlation coefficient (R), coefficient of determination (R^2), standard error, residual and F-ratio value. The study signifies which technology gives the best validation of the model obtained in terms of MLR and is helpful in identifying the technology which gives the maximum significant independent variable in predicting the dependent variable. In terms of the beneficial impacts to the society, the uniqueness of this study lies in bringing the robustness of the STP and the technology utilized. Hence, as a future scope, it can be inculcated in other STPs to get the optimum results.

The basic objective of the study is the establishment of most suitable MLR models relating BOD removal Efficiency (considered as the dependent variable) to independent variable like pH, BOD, COD, TSS, Oil and Grease, Ammoniacal Nitrogen and Phosphates (to the treated effluent) for all the three technologies. The best-fit regression equation will be developed based on the multiple correlation coefficient (R), coefficient of determination (R^2), standard error, residual and F-ratio value.

Materials and methods

Technologies covered in the study

The present study was carried out on three different sewage treatment technologies used in different STPs in Delhi, mainly Activated Sludge Process (ASP), Extended Aeration and Densadeck. Densadeck technology which is also known as the Biofor technology is an advanced aerobic process which is enhanced by the primary treatment with the use of the coagulants. It is also known as the two-stage filtration process. This technology is utilized at Dr. Sen Nursing Home STP of 20 MLD capacity which mainly follows physico-chemical treatment process for the sewage treatment. ASP is one of the aerobic sewage treatment technologies. It is generally utilized for the treatment of the raw sewage or the settled sewage and the return of the sludge to the primary sedimentation tank. It is used in Okhla Phase-VI STP whose design capacity is of 30 MLD. Extended Aeration is an aerobic technology utilized in Vasant Kunj STP of New Delhi. This includes pre-treatment viz. screening, degritting and Aeration, clarification and sludge dewatering on sludge drying beds. (DJB [Delhi Jal Board] 2015).The diagrammatic representation of the operational units of the STPs with different technologies is given in Figs. 1, 2, and 3

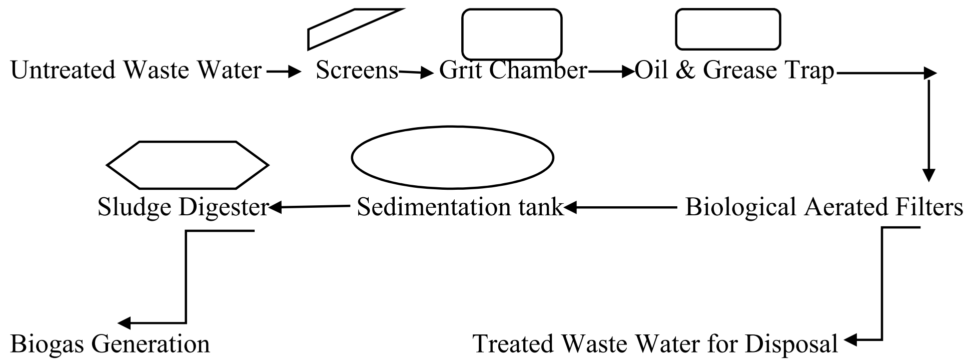


Fig. 1 Flow diagram for Densadeck technology

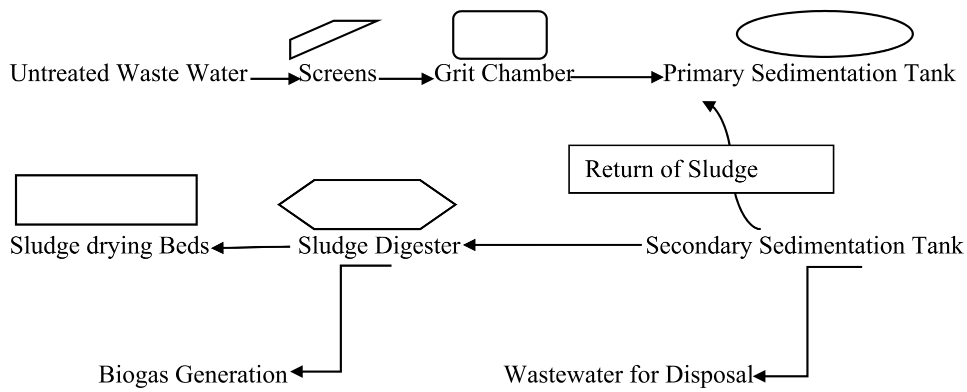


Fig. 2 Flow diagram for activated sludge process (ASP)

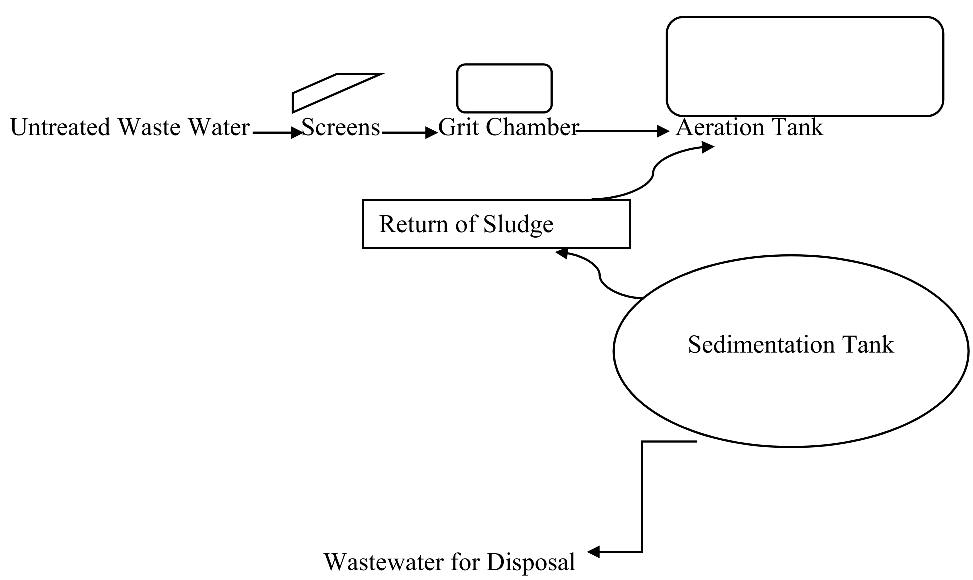


Fig. 3 Flow diagram for extended aeration system (EA)

Sampling points and frequency

The sampling points for the above-mentioned STPs in the study area were the outlet channel, i.e. it focussed on the effluents of each selected STP. Sampling was done every month from the year 2012 to 2017 (American public health association (APHA) 1998).

Physiochemical and biological parameters analyzed

The parameters considered for present study are pH, total suspended solids (TSS), biochemical oxygen demand (BOD), chemical oxygen demand (COD), oil and grease, ammoniacal nitrogen and phosphates. All the parameters were tested as per APHA standards (American Public Health Association

(APHA) 1998). Treated effluents' physio-chemical and biological parameters are being evaluated in the study. Selection of the above-mentioned independent variables is done by Delhi Pollution Control Committee (DPCC). Data collected from DPCC have been considered during the course of the study. Selection/testing process for all the dependent and independent variables is listed in Table 1. Moreover, BOD is considered as the potential parameter governing the performance of the STP; hence, it is important to foresee how other parameters influence BOD removal efficiency. Therefore, BOD is taken as dependent variable and others as independent variable.

Pearson correlation coefficients

It is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. Pearson correlation coefficients scale lies between −1 and 1. If the value lies between −1 and −0.50, it shows strong negative correlation; whereas, the value of −0.50 indicates a moderate negative correlation. If the value lies between −0.50 and 0, it means a weak negative correlation; whereas, at 0 it shows no correlation. For the value between 0 and 0.50, it indicates a weak positive correlation and at 0.50, it is of moderate positive correlation. Between 0.50 and 1 it is of strong positive correlation and if the value is 1, it represents perfect positive correlation.

Multiple linear regression analysis

The motivation of multiple regression analysis is to figure out an equation that can determine the response variable as a function of several explanatory variables (Coelho-Barros et al. 2008). The MLR equation, given n observations, is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon_i \qquad (1)$$

$i = 1, 2, \ldots n.$

Here, y is the dependent variable (BOD removal efficiency); x_1, x_2…, x_k are the independent variables (physico-chemical parameters); "n" sample observations; β_0 is the y intercept (the value of dependent variable "y" when all of the explanatory variables x_1, x_2…, $x_k = 0$); β_1, β_2, …, β_n are the estimated multiple regression coefficients; and the term ε is a random error (Agirre et al. 2006; Ferraro and Giordani 2012; Kovdienko et al. 2010).

In this study, MLR analysis was carried out for outcome/response variable (dependent variable) with respect to the predictive variables (independent variables). The dependent variable taken into consideration here is biological parameter, i.e., BOD removal efficiency and the rest of all the physio-chemical parameters are taken as independent variables. In this study, MLR analysis emphasizes on developing the model in terms of BOD removal efficiency as the function of independent variables. All the MLR analyses including the time series plots were carried out on SPSS.

Checking multiple linear regression (MLR) assumptions

Once the MLR analysis is done, it is followed by testing and verification of the proposed equation/model as per the MLR assumptions. The various assumptions of the MLR analysis include the following:

(a) Linear relationship: MLR assumes that there is a linear relation between the response variable and the predictive variables.
(b) Multivariate normality: MLR assumption also says that the residuals have normal distribution.
(c) No multicollinearity: this assumption indicates that the predictive variables are not having high correlation with each another.
(d) Homoscedasticity: MLR assumes that there is homogeneity in the variance, i.e., variance must be same for each variable.

Multiple linear regression (MLR) final outputs

F-Test: It is a statistical test in which the test statistics has *F*-distribution under the null hypothesis. It is generally used to make comparisons between the models that have been fitted

Table 1 Test methods adopted for various physiochemical parameters

S. No.	Name of the parameter	Test method adopted	Instruments used
1	pH	Electrometric	pH meter
2	Oil and grease	Soxhlet extraction	Soxhlet apparatus
3	Total Suspended Solids	Membrane filtration	Glass fiber apparatus
4	Biochemical oxygen demand	Winkler's titration	BOD incubator
5	Chemical oxygen demand	Closed reflux titrimetric	Titrimetric instruments
6	Ammonical–nitrogen	Distillation titrimetric	Titrimetric instruments
7	Phosphate	Ascorbic acid spectrophotometry	Spectrophotometer

to a data set to identify the model that best fits the population from which the data were sampled. The *F* value is always used along with the *p* value which decides whether the results obtained are significant enough to reject the null hypothesis or not. If we get a large *f* value (one that is bigger than the *F* critical value found in a table), it means something is significant; while a small *p* value means all the results are significant.

p value: The *p* value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. The *p* value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. The threshold value of acceptance of *p* value is $p < 0.05$.

Alpha level: It is also known as the significance level (denoted as $\propto$ level). It is known as the probability of rejecting the null hypothesis when it is true. Alpha levels are used in the hypothesis tests that run with an alpha level of 0.05 (5%) and also known as threshold of acceptance.

Variance of inflation (VIF): It is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. If the VIF is 5–10, the regression coefficients are poorly estimated, i.e., if the VIF values for each of the estimated regression coefficients are less than 10, then there is no multicollinearity in the model (Montgomery and Peck 1982).

R^2 and Adjusted R^2: It is measured using square of the multiple correlation coefficient R^2 (also called the coefficient of multiple determination). It is a statistical measure of how close the data are to the fitted regression line. The adjusted R^2 is another index that is often preferred as a measure of regression model quality. The value lies between 0 and 100%. More close is the value towards 100% which indicates that the model explains all the variability of the response data around its mean.

"Enter Method" of Multiple Regression Analysis on SPSS: Multiple Linear Regression using the 'Enter' method (default with the menu system) enters all variables into the equation at the beginning (one step), which is also called "forced entry".

Results and discussions

To foresee that how the BOD removal efficiency is varying with time, the time series plots have been prepared for all the three technologies from the period 2012–2017. The time series plots are mainly the data points listed in time order. Figure 4 depicting time series plots for Densadeck technology clearly shows that the BOD removal efficiency for majority of the time period lies between 94 and 98%; however, there was a sharp decline in it during the month of March in 2013. For the extended aeration technology, 84–87% of the BOD removal efficiency have been recorded for most of the cases, but the same declined drastically in the month of Feb and March during the year 2012. For ASP technology, 86–87% of the BOD removal efficiency was obtained for majority of the duration and there was a sharp decline with respect to the BOD removal efficiency in the month of March 2012 almost similar to extended aeration technology. The deep decline in the BOD removal efficiency is attributed towards heavy organic loading due to which the BOD values increases highly in that period resulting in less removal efficiency. On the other hand, sudden increase in organic loading of the STP indicates the more concentrated sewage waste entering the STP. Hence, the composition of the waste entering the STP with heavy organic load is also attributed towards high kitchen or domestic waste having little dilution from the Choe, drain, industrial effluent.

Multiple linear regression analysis for Densadeck technology

The relationship between various parameters is investigated using correlation coefficient values (*R*) which are listed in Table 2. A strong positive correlation of BOD removal efficiency with BOD and COD of effluent is observed as the Pearson Correlation value for the same is 0.594 and 0.425, respectively. However, weak positive correlation of BOD removal efficiency is depicted with ammonical nitrogen and TSS. Only, Oil and Grease has the negative correlation with BOD removal efficiency (-0.021).

After the above analysis, the predictive model was developed for target parameter using the multivariate regression analysis including multiple complex terms of variables. Several combination sets of predictor variables in conjunction with their interactions were considered for model generation. By defining the threshold *p* value of 0.05 and performing the forward method, predictors were added one at a time beginning with the predictor with the highest correlation with the dependent variable. The most significant of these variables is added to the model, as long as its *p* value is below 0.05.

The fitted model for the Densadeck technology is given in (Table 3):

$$\text{BOD Removal Efficiency} = 88.274 + 0.057 \times \text{BOD} - 0.352 \times \text{Phosphates} - 0.008 \times \text{TSS}, \tag{2}$$

where, 88.274 is the *y* intercept (the value of dependent variable "*y*") 0.057, -0.352 and -0.008 are the estimated multiple regression coefficients for BOD, phosphates and TSS. These are the monthly mean value from 2012 to 2017.

The above fitted models was tested for the overall ability to predict the response variable using an *F*-test, or equivalently, by an analysis of variance (ANOVA). From the analysis of

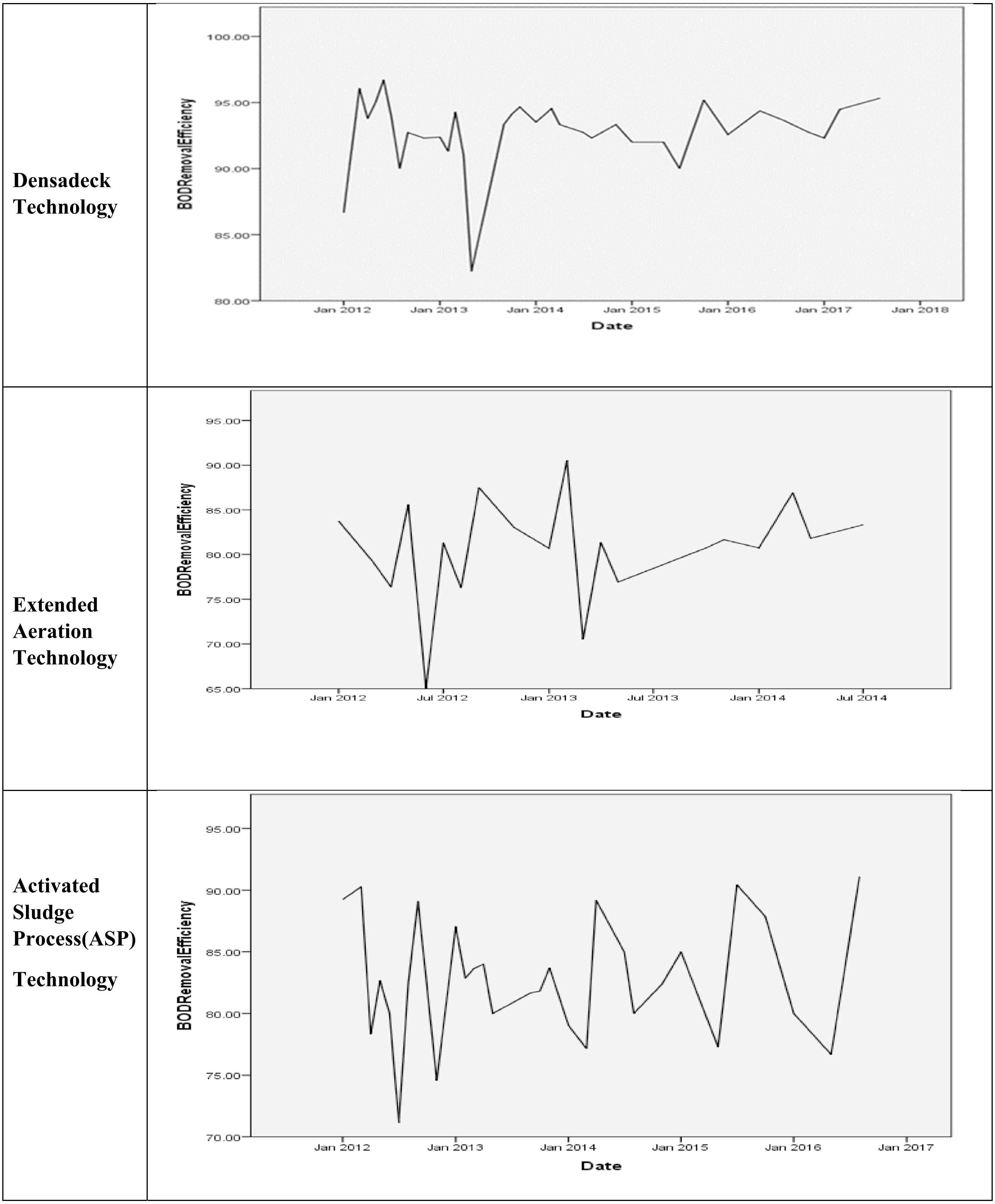


Fig. 4 Time series plots of three technologies

Table 2 Correlation analysis of three technologies

	BOD Removal Efficiency	pH	TSS	BOD	COD	Oil Grease	Ammonical Nitrogen	Phosphates
Densadeck technology								
BOD Removal Efficiency	1	0.052	0.032	0.594	0.425	−0.021	0.132	0.008
pH	0.052	1	0.227	0.337	0.159	−0.119	−0.353	0.337
TSS	0.032	0.227	1	0.46	0.498	0.027	−0.596	0.161
BOD	0.594	0.337	0.46	1	0.707	−0.103	−0.141	0.456
COD	0.425	0.159	0.498	0.707	1	0.014	−0.278	0.2
Oil Grease	−0.021	−0.119	0.027	−0.103	0.014	1	0.022	−0.106
Ammonical Nitrogen	0.132	−0.353	−0.596	−0.141	−0.278	0.022	1	0.022
Phosphates	0.008	0.337	0.161	0.456	0.2	−0.106	0.022	1
Extended aeration technology								
BOD Removal Efficiency	1	−0.011	0.086	0.713	0.519	0.052	0.235	−0.021
pH	−0.011	1	−0.347	0.117	−0.021	0.173	0.082	0.102
TSS	0.086	−0.347	1	−0.036	0.091	0.001	−0.163	−0.353
BOD	0.713	0.117	−0.036	1	0.797	0.279	0.218	−0.006
COD	0.519	−0.021	0.091	0.797	1	0.471	−0.187	−0.238
Oil Grease	0.052	0.173	0.001	0.279	0.471	1	−0.511	−0.316
Ammonical Nitrogen	0.235	0.082	−0.163	0.218	−0.187	−0.511	1	0.597
Phosphates	−0.021	0.102	−0.353	−0.006	−0.238	−0.316	0.597	1
Activated Sludge Process (ASP) technology								
BOD Removal Efficiency	1	0.021	0.094	0.634	0.207	−0.054	0.188	0.283
pH	0.021	1	−0.156	0.1	−0.157	−0.031	−0.472	−0.147
TSS	0.094	−0.156	1	0.212	−0.046	−0.004	0.571	0.263
BOD	0.634	0.1	0.212	1	0.01	−0.272	0.161	0.444
COD	0.207	−0.157	−0.046	0.01	1	−0.059	0.121	−0.159
Oil Grease	−0.054	−0.031	−0.004	−0.272	−0.059	1	−0.088	−0.337
Ammonical Nitrogen	0.188	−0.472	0.571	0.161	0.121	−0.088	1	0.301
Phosphates	0.283	−0.147	0.263	0.444	−0.159	−0.337	0.301	1

variance (ANOVA) statistics [F (3, 30) = 11.056, $p < 0.05$], it is observed that the p value is (0.000) which implies that the model estimated by the regression procedure is significant at $\propto$ level of 0.05 (Table 4). Hence, there exists one of the regression coefficients which is different from zero. The p values for the estimated coefficients of BOD, phosphates and TSS are, respectively, 0.00, 0.09 and 0.028, indicating that they are significantly related to BOD removal efficiency.

By multicollinearity, it means that the independent variables are correlated with one another. In this study, variance inflation factors (VIF) are examined, which is the measure of increase of variance in the estimated regression coefficient when the independent variables are correlated. If the VIF is 5–10, the regression coefficients are poorly estimated (Montgomery and Peck 1982). From Table 3, it can be seen that VIF for each of the estimated regression coefficient are less than 10; thus, there is no multicollinearity in the model.

The goodness of fit of the multiple regression model describes how well the regression model fits the data points. It is measured using square of the multiple correlation coefficient R^2 (also called the coefficient of multiple determination). The R^2 value obtained in Table 5 indicates that only 52.5% of the total variation of the BOD removal efficiency values about their mean can be explained by the independent variables used in the model. The R^2 statistic is to some extent problematic as a goodness-of-fit index because it constantly increases when an explanatory variable is added to the model. The adjusted R^2 is another index that is often preferred as a measure of regression model quality. The adjusted R^2 value in the study shows that 47.8% of the total variation of the BOD removal efficiency values about their mean can be explained by the predictor variables used in the model (Table 5). As the values of R^2 and adjusted R^2 are not very different, it appears that at least one of the predictor variables contributes information for the prediction of the response variable, i.e., BOD removal efficiency. Thus, both values indicate that the model fits the data well.

The goodness-of-fit model is also examined based on residual plots. From the normal probability plot, it is observed that there exists an approximately linear pattern

Table 3 Regression coefficients of three technologies

Coefficients[a]									
Model	Unstandardized coefficients		Standardized coefficients	*t*	Sig.	95.0% Confidence Interval for B		Collinearity statistics	
	B	Std. error	Beta			Lower bound	Upper bound	Tolerance	VIF
Densadeck technology									
1									
(Constant)	87.748	1.264		69.434	0	85.174	90.323		
BOD	0.037	0.009	0.594	4.178	0	0.019	0.056	1	1
2									
(Constant)	87.393	1.205		72.537	0	84.936	89.851		
BOD	0.047	0.01	0.746	4.94	0	0.028	0.066	0.792	1.263
Phosphates	−0.331	0.15	−0.333	−2.202	0.035	−0.638	−0.024	0.792	1.263
3									
(Constant)	88.274	1.191		74.121	0	85.842	90.706		
BOD	0.057	0.01	0.906	5.753	0	0.037	0.077	0.638	1.567
Phosphates	−0.352	0.141	−0.353	−2.49	0.019	−0.64	−0.063	0.789	1.268
TSS	−0.008	0.004	−0.328	−2.312	0.028	−0.015	−0.001	0.785	1.274
Extended aeration technology									
1									
(Constant)	71.179	27.692		2.57	0.023	11.354	131.004		
pH	−0.902	3.59	−0.052	−0.251	0.805	−8.658	6.853	0.8	1.25
TSS	0.003	0.007	0.092	0.436	0.67	−0.012	0.018	0.768	1.302
BOD	0.159	0.077	0.845	2.072	0.059	−0.007	0.324	0.208	4.815
COD	−0.005	0.018	−0.113	−0.28	0.784	−0.043	0.033	0.214	4.684
Oil Grease	−0.108	0.208	−0.132	−0.519	0.613	−0.556	0.341	0.535	1.87
Ammonical Nitrogen	0.006	0.131	0.015	0.046	0.964	−0.278	0.29	0.341	2.931
Phosphates	−0.217	1	−0.054	−0.217	0.831	−2.377	1.942	0.554	1.804
Activated Sludge Process (ASP) technology									
1									
(Constant)	49.844	20.137		2.475	0.021	8.083	91.605		
pH	0.803	2.438	0.06	0.329	0.745	−4.253	5.859	0.706	1.416
TSS	−0.011	0.015	−0.133	−0.684	0.501	−0.043	0.021	0.627	1.596
BOD	0.139	0.039	0.637	3.55	0.002	0.058	0.22	0.733	1.365
COD	0.014	0.011	0.211	1.299	0.207	−0.009	0.038	0.895	1.118
Oil Grease	0.261	0.247	0.177	1.057	0.302	−0.251	0.772	0.84	1.191
Ammonical Nitrogen	0.062	0.089	0.152	0.697	0.493	−0.123	0.247	0.496	2.017
Phosphates	0.261	0.548	0.091	0.477	0.638	−0.875	1.398	0.643	1.555

For Densadeck technology 1, 2 and 3 models have been given for MLR analysis obtained on SPSS as it has used iteration methods (step wise); hence, the best possible model obtained out of all is model 3

[a]Dependent Variable: BOD Removal Efficiency

(Fig. 5). This indicates the consistency of the data with a normal distribution, hence satisfying multivariate normality assumption. From the scatter plot of the residuals, it is evident that the variance around the regression line is the same for all values of the independent variables (Fig. 5). This may indicate that the residuals have constant variance showing homoscedasticity. The models are, therefore, considered valid for describing the dependent variable based on the data set.

Multiple linear regression analysis for extended aeration technology

A strong positive correlation was observed for BOD removal efficiency with BOD and COD giving the Pearson Correlation value as 0.713 and 0.519, respectively; weak positive correlation was detected between ammonical nitrogen, oil and grease, and TSS. However in this technology, BOD

Table 4 ANOVA analysis of three technologies

Model	Sum of squares	*df*	Mean square	*F*	Sig.
Densadeck technology					
1					
Regression	83.294	1	83.294	17.453	0.000[b]
Residual	152.719	32	4.772		
Total	236.012	33			
2					
Regression	103.955	2	51.978	12.202	0.000[c]
Residual	132.057	31	4.26		
Total	236.012	33			
3					
Regression	123.922	3	41.307	11.056	0.000[d]
Residual	112.09	30	3.736		
Total	236.012	33			
Extended aeration technology					
1					
Regression	350.075	7	50.011	2.275	0.095[e]
Residual	285.784	13	21.983		
Total	635.859	20			
Activated Sludge Process (ASP) technology					
1					
Regression	353.233	7	50.462	2.913	0.026[e]
Residual	381.041	22	17.32		
Total	734.274	29			

For Densadeck technology 1, 2 and 3 models have been given for MLR analysis obtained on SPSS as it has used iteration methods (step wise); hence, the best possible model obtained out of all is model 3

[a]Dependent Variable: BOD Removal Efficiency

[b]Predictors: (Constant), BOD

[c]Predictors: (Constant), BOD, Phosphates

[d]Predictors: (Constant), BOD, Phosphates, TSS

[e]Predictors: (Constant), Phosphates, BOD, pH, Oil Grease, TSS, Ammonical Nitrogen, COD

removal efficiency is having negative correlation with phosphates with the value, i.e., −0.21 (Table 2).

After the correlation analysis, predictive model was developed using MLR and the method used in this technology for fitting the model was "ENTER" method. The fitted model (Table 3) is given by:

$$\begin{aligned}\text{BOD Removal Efficiency} &= 71.179 - 902 \times \text{pH} + 0.003 \\ &\times \text{TSS} + 0.159 \times \text{BOD} \\ &- 0.005 \times \text{COD} \\ &- 1.08 \times \text{Oil Grease} + 0.006 \\ &\times \text{Ammonical Nitrogen} \\ &- 0.217 \times \text{Phosphates}.\end{aligned} \tag{3}$$

The model obtained above was tested for predicting the response variable. From the ANNOVA statistics [F (7, 13) = 2.275, $p < 0.05$], it is observed that p value is 0.095 which implies that the model estimated by the regression procedure is not significant at ∝ level of 0.05 (Table 4). The p values for the estimated coefficients of COD, Phosphates, Oil and Grease, Ammonical Nitrogen and TSS (0.784, 0.831, 0.613, 0.964 and 0.67) indicate that they are not significantly related to BOD removal efficiency.

For multicollinearity assumption, VIF values obtained indicate the range between 1.250 and 4.1815. As the VIF values for each of the estimated regression coefficient are less than 10, there is no multicollinearity in the model (Table 3).

The goodness of fit of the multiple regression model is given by R^2. As R^2 value in the regression output is 0.551, it depicts that 55.1% of the total variation of the BOD removal efficiency values about their mean can be explained by the predictor variables used in the model. The adjusted R^2 value indicates that only 30.9% of the total variation of the BOD removal efficiency values about their mean can be explained by the predictor variables used in the model (Table 5). As the values of R^2 and adjusted R^2 are a bit different, it appears that at least one of the predictor variables contributes information for the prediction of the response variable, i.e., BOD removal efficiency. Thus, the model obtained fits the data well.

Multivariate normality or goodness-of-fit model can be checked by normal P–P plot. Figure 5 clearly indicates that the residuals are normally distributed as they are showing linear pattern. Hence the assumption of multivariate normality is met. From the scatter plot of the residuals, it is evident that the variance around the regression line is the same for all values of the independent variables. This may indicate that the residuals have uniform variance showing homoscedasticity. Hence, this model is considered valid for describing the response variable based on the data set.

Multiple linear regression analysis for activate sludge process (ASP) technology

Pearson correlation coefficients indicates strong positive correlation of BOD removal efficiency with BOD (0.634), weak positive correlation with COD and phosphates (0.207 and 0.283), respectively. This technology is also showing negative correlation of BOD removal efficiency with oil and grease having the value −0.54 [Table 2].

Correlation analysis was followed by developing predictive model using MLR and the method used in this technology for fitting the model was also "ENTER" method. The fitted model (Table 3) obtained is given by:

Table 5 Model summary of three technologies

Model	R	R square	Adjusted R square	Std. error of the estimate	Change statistics					Durbin–Watson
					R square change	F change	$df1$	$df2$	Sig. F change	
Densadeck technology										
1	0.594[a]	0.353	0.333	2.1846	0.353	17.453	1	32	0	
2	0.664[b]	0.44	0.404	2.06395	0.088	4.85	1	31	0.035	
3	0.725[c]	0.525	0.478	1.93296	0.085	5.344	1	30	0.028	2.05
Model	R	R square	Adjusted R Square	Std. error of the estimate	Change statistics					Durbin–Watson
Extended aeration technology										
1	0.742[d]	0.551	0.309	4.68864	0.551	2.275	7	13	0.095	2.868
Activated Sludge Process (ASP) technology										
1	0.694[d]	0.481	0.316	4.16174	0.481	2.913	7	22	0.026	1.955

[a]Predictors: (Constant), BOD

[b]Predictors: (Constant), BOD, Phosphates

[c]Predictors: (Constant), BOD, Phosphates, TSS

[d]Predictors: (Constant), Phosphates, BOD, pH, Oil Grease, TSS, Ammonical Nitrogen, COD

[e]Dependent Variable: BOD Removal Efficiency

$$\begin{aligned}\text{BOD Removal Efficiency} = {} & 49.844 + 0.803 \times \text{pH} \\ & - 0.011 \times \text{TSS} + 0.139 \\ & \times \text{BOD} + 0.014 \times \text{COD} \\ & + 0.261 \times \text{Oil Grease} \\ & + 0.062 \times \text{Ammonical Nitrogen} \\ & + 0.261 \times \text{Phosphates}.\end{aligned} \tag{4}$$

Capacity of the model was tested for determining the dependent variable using analysis of variance ANOVA. ANOVA statistics [F (7, 30) = 2.913, $p < 0.05$] indicates p value as .026 which implies that the model estimated by the regression procedure is significant at $\propto$ level of 0.05. The p values for the estimated coefficients of COD, phosphates, oil and grease, ammonical nitrogen and TSS indicate that they are not significantly related to BOD removal efficiency as the p value for all of them is greater than 0.05. But in case of BOD as the p value is 0.002, only this parameter is significantly related to the BOD removal efficiency (Table 4).

For verifying the assumption of "no multicollinearity" Table 3 showing the coefficients of regression is used which depicts that VIF lies between 1.118 and 2.017 which shows that as VIF value for each of the estimated regression coefficient (which is less than 10), therefore no multicollinearity in the model.

From the Table 5, R^2 value in the regression output obtained is 0.481 which shows that 48% of the total variation of the BOD removal efficiency values about their mean can be explained by the predictor variables used in the model. The adjusted R^2 value indicates that only 31.6% of the total variation of the BOD removal efficiency values about their mean can be explained by the predictor variables used in the model. As the values of R^2 and adjusted R^2 are a bit different, it appears that at least one of the predictor variables contributes information for the prediction of the response variable, i.e., BOD removal efficiency. Thus, R^2 value indicates that the model fits the data well. Therefore, this proves the validation of this model in predicting the response variable.

Multivariate normality assumption or goodness-of-fit model can be checked by normal P–P plot and homoscedasticity from the scatter plot depicting in Fig. 5. It is observed that the residuals are having normal distribution as they are showing linear pattern; also it is evident that the variance around the regression line is the same for all values of the independent variables. This may indicate that the residuals have uniform variance, hence satisfying the homoscedasticity assumption.

Comparisons between different models

From the results obtained, a comparative account between different models is summed up and given as:

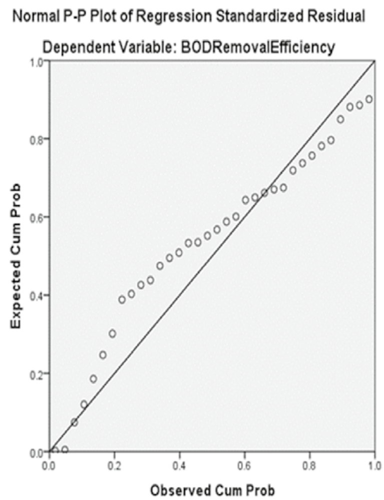
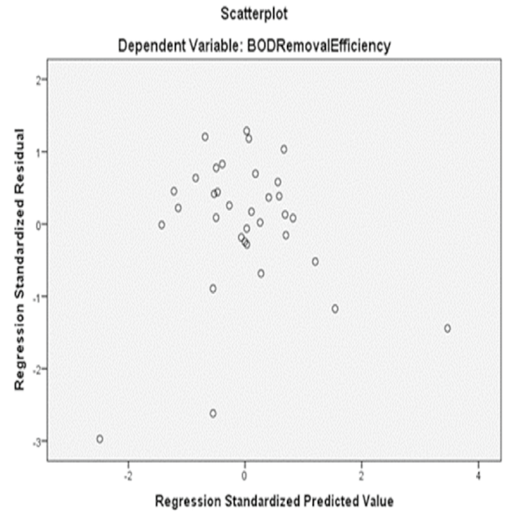
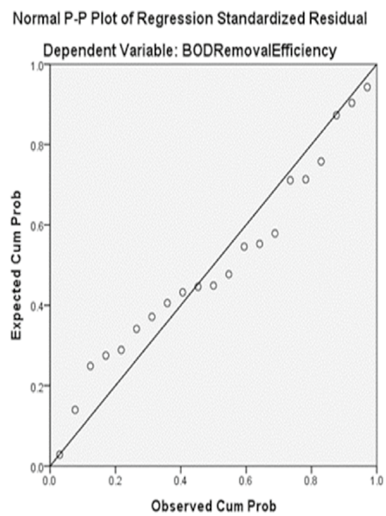
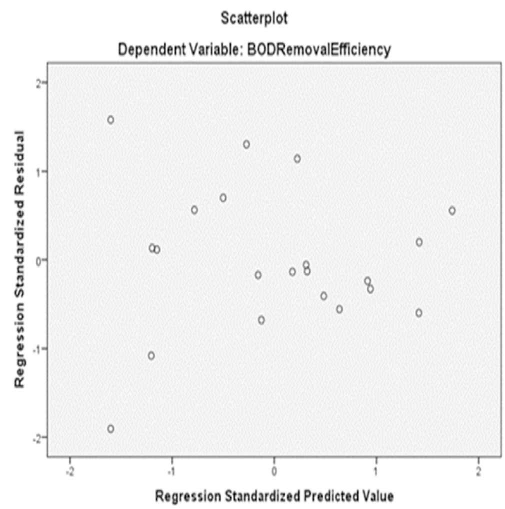
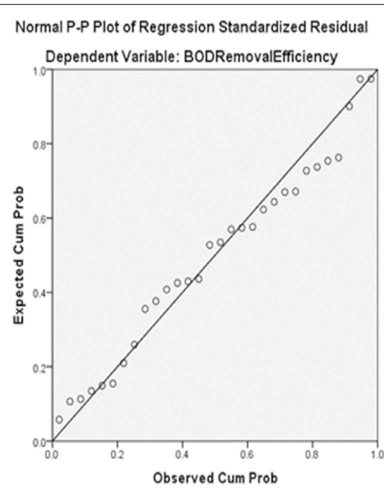
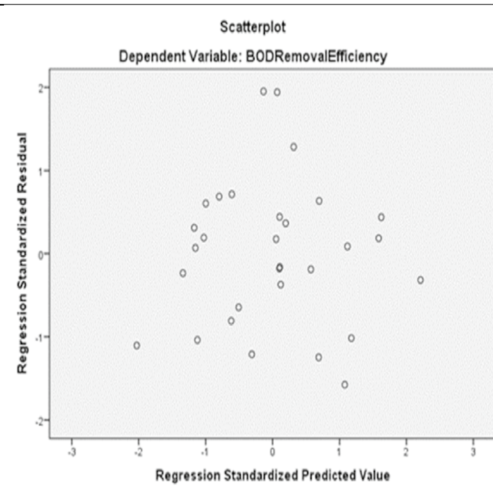
Technology	Normal P-P Plots	Scatter Plot
Densadeck Technology		
Extended Aeration Technology		
Activated Sludge Process (ASP) Technology		

Fig. 5 Normal P–P plots and scatter plots of three technologies

Technology	Results/model obtained
Densadeck	BOD Removal Efficiency = 88.274 + 0.057 × BOD − 0.352 × Phosphates − 0.008 × TSS The above model obtained was tested to predict the response variable. Here, 88.274 is the *y* intercept (the value of dependent variable "*y*") .057, -.352 and -.008 are the estimated multiple regression coefficients for BOD, phosphates and TSS. *p* values (0.009, 0.09 and 0.028) clearly indicated that the estimated coefficients of BOD, Phosphates and TSS respectively are significantly related to BOD removal efficiency
Extended Aeration	BOD Removal Efficiency = 71.179 − 902 × pH + 0.003 × TSS + 0.159 × BOD − 0.005 × COD − 1.08 ×Oil Grease + 0.006 × Ammonical Nitrogen − 0.217 × Phosphates The model obtained above shows that *p* values for the estimated coefficients of COD, phosphates, oil and grease, ammonical nitrogen and TSS (0.784, 0.831, 0.613, 0.964 and 0.67) indicates that they are not significantly related to BOD removal efficiency
Activated Sludge Process	BOD Removal Efficiency = 49.844 + 0.803 × pH − 0.011 × TSS + 0.139 × BOD + 0.014 × COD + 0.261 ×Oil Grease + 0.062 × Ammonical Nitrogen + 0.261 × Phosphates The model obtained as per ANOVA statistics indicates that *p* value is .026 which shows that the model estimated by the regression procedure is significant at ∝-level of 0.05. But in this case, only BOD is significantly related to the BOD removal efficiency as its *p* value is 0.002 else for rest of the parameters obtained in the model

Conclusion

Time series plots revealed that out of the three technologies taken into consideration ASP have proven to be the best, by giving R square value as .551 i.e., 55.1% of the variance in the dependent variable is explained by the predictive variables. If we consider the significant independent parameters, then Densadeck technology has given the maximum significant independent variables, i.e., BOD, TSS and Phosphates in predicting the dependent variable and the model is a good fit as it is contributing 52.5% in bringing the change in the variance of the dependent variable. Although both BOD and COD are considered as the pollution indicator of water body, but still BOD is widely taken as the prime most factor while assessing the performance of wastewater/STP than COD (Sharma et al. 2013; Singh et al. 2014; Kumar et al. 2017). It is, hence, recommended that the parameters affecting the performance of BOD in this Plant should be taken into consideration in future.

Acknowledgements We pay sincere thanks to Delhi Pollution Control Committee (DPCC) and Central Pollution Control Board (CPCB) for providing us relevant information regarding the various STPs in Delhi NCR. We also thank the anonymous reviewers for their constructive feedback and suggestions.

References

Abyaneh HZ (2014) Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. J Environ Health Sci Eng 12:1–8

Agirre BE, Ibarra BG, Madariaga I (2006) Regression and multilayer perceptron-based models to forecast hourly O-3 and NO_2 levels in the Bilbao area. Environ Modell Softw 21(4):430–446

AI Gheethi AAS, Norli I, Kadil MOAB (2013) Elimination of enteric indicators and pathogenic bacteria in secondary effluents and lake water by solar disinfection (SODIS). J Water Reuse Desalination 3:39–46

AI Gheethi AAS, Abdul-Moneum MO, AL-Zubirey AHS, Efaq NF, Shamar AM, AI Amery RMA (2014) Effectiveness of selected wastewater treatment plants in Yemen for reduction of faecal indicators and pathogenic bacteria in secondary effluents and sludge. Water Pract Technol 9:293–306

Al-Gheethi AAS, Mohamed MR, Efaq AN, Norli I, Adib MR, Amir HMK (2017) Reduction of bacteria in storage system of sewage effluents. Sustain Water Resour Manag 3:193–203

American public health association (APHA) (1998) Standard methods for the examination of waters and wastewaters', 20th edn. American Public Health Association (APHA), Washington, DC

Belhaj D, Jaabiri I, Turki N, Azri C, Kallel M, Ayadi H (2014) Descriptive and multivariable analysis of the water parameters quality of Sfax sewage treatment plant after rehabilitation. IOSR J Comput Eng 16:81–91

Coelho-Barros EA, Simoes PA, Achcar JA, Martinez EZ, Shimano AC (2008) Methods of estimation in multiple linear regression: application to clinical data. Rev Colomb Estad 31(1):111–129

Colmenarejo MF, Rubio A, Sanchez E, Vicente J, Gracia MG, Bojra R (2006) Evaluation of municipal wastewater treatment plants with different technologies at Las-Rozas, Madrid (Spain). J Environ Manag 81:399–404

DJB [Delhi Jal Board] (2015) Wastewater treatment technologies adopted at sewerage treatment plants, Retrieved June 2019 from https://elibrarywcl.files.wordpress.com/2015/02/sewage-treatment-and-technology.pdf

Dogan Ates, Yilmaz E, Eren B (2008) Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand. Environ Prog 27:439–445

DPCC (2016) [Delhi Pollution Control Committee Report 2016]. Status of sewerage treatment plants in Delhi. [Internet]. c2018 [Cited July 2018]. https://www.dpcc.delhigovt.nic.in/down/5th_meeting_II

Fedotovai O, Teixeira L, Alvelos H (2013) Software effort estimation with multiple linear regression: review and practical application. J Inf Sci Eng 29:925–945

Ferraro MB, Giordani P (2012) A multiple linear regression model for imprecise information. Metrika 75(8):1049–1068

Hur J, Lee BM, Lee TH, Park DH (2010) Estimation of biological oxygen demand and chemical oxygen demand for combined sewer systems using synchronous fluorescence spectra. Sens Basel 10:2460–2471

Jamwal P, Mittal AK, Mouchel J (2009) Efficiency evaluation of sewage treatment plants with different technologies in Delhi (India). Environ Monit Assess 153:293–305

Kovdienko NA, Polishchuk PG, Muratov EN, Artemenko AG, Kuzmin VE, Gorb L, Hill F, Leszczynski J (2010) Application of random forest and multiple linear regression techniques to QSPR prediction of an aqueous solubility for military compounds. Mol Inform 29(5):394–406

Kumar R, Vaid U, Mittal S (2017) Water crisis: issues and challenges in Punjab. Water Resour Manag 78:93–103

Montgomery DC, Peck EA (1982) Introduction to linear regression analysis. Wiley, New York

Moustris KP, Nastos PT, Larissi IK, Paliatsos AG (2012) Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens area, Greece. Adv Meteorol. https://doi.org/10.1155/2012/894714

Najah A, Elshafie A, Karim OA, Jaffar O (2009) Prediction of Johor River water quality parameters using artificial neural networks. Eur J Sci Res 28:422–435

Noller DG, Whitehouse GE (1982) Multiple linear-regression—a microcomputer application. Ind Eng 14:26

Parmar DL, Keshari A (2012) Sensitive analysis of water quality for Delhi stretch of the River Yamuna, India. Environ Monit Assess 184:1487–1508

Seung PL, Sang YM, Jin SK, Jong U, Man SK (2014) A study on the influence of a sewage treatment plant's operational parameters using the multiple regression analysis model. Environ Eng Res 19:31–36

Sharma P, Khitoliya RK, Kumar S (2013) A comparative study of sewerage treatment plants with different technologies in the vicinity of Chandigarh City. IOSR J Environ Sci Toxicol Food Technol 4(5):113–121

Singh S, Singh N, Kumar S (2014) Quality of water in and around chandigarh region—a review. J Chem Environ Sci Appl 1:33–34

Vega M, Pardo R, Barrato E, Deban L (1998) Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res 32:3581–3592

Wang ZM, Chen LD, Zhang HP, Sun RH (2014) Multivariate statistical analysis and risk assessment of heavy metals monitored in surface sediment of the Luan River and its tributaries, China. Hum Ecol Risk Assess 20:1521–1537

Yerel S, Ankara H (2012) Application of multivariate statistical techniques in the assessment of water quality in Sakarya River, Turkey. J Geol Soc India 79:89–93

Zihan L, Jin CJ, Sun HC, Namjoo H, Jungseok J, Jun WH (2018) Assessment of surface water quality in Geum River Basin, Korea using multivariate statistical techniques. Int J Appl Eng Res 13:6723–6732