International Journal of ChemTech Research
CODEN (USA): IJCRGG, ISSN: 0974-4290, ISSN(Online):2455-9555
Vol.14 No.01, pp 249-258, 2021

Multivariate statistics for soft sensing primary clarifier effluent quality in industrial wastewater treatment plant

Nital Patel[1*], Jayesh Ruparelia[2], Jayesh Barve[3]

[1]Instrumentation and Control Engineering Department, Institute of Technology, Nirma University, Ahmedabad, India
[2]Chemical Engineering Department, Institute of Technology, Nirma University, Ahmedabad,
[3]Ex-Professor, Nirma University, Ahmedabad (Now: GE Research, Bengaluru, India)

Abstract : In wastewater treatment plant clarification is a major step to remove the suspended solids. The performance of the primary clarifier is important as the effluent of primary clarifier subsequently treated further in downstream biological process. The main objective of primary clarifier is to remove the suspended solids present in influent wastewater. The monitoring of the primary clarifier operation is crucial in order to maintain the efficient performance. In this work, application of multivariate statastical techniques to predict or softsense the effluent quality of industrial primary clarifier is investigated. The industrial clariflocculator located at common effluent treatment plant (CETP), Vatva, Ahmedabad, India is considered. The Principal Component Analysis (PCA) is adopted to check and reveal the collinearity among influent COD, BOD, TDS and TOC. Three partial least square (PLS) models are developed to estimate effluent COD, BOD and TOC based on influent quality parameters. The PLS model of effluent TOC is found better than the PLS models for COD and BOD. It is observed that the fewer number of PLS components, that well explain the maximum variance in the effluent quality parameter (COD, BOD or TOC), gives better results. Hence, there is no need to consider all PLS components for effluent quality soft-sensor model development. The estimation of effluent COD, BOD and TOC can be done with two, three and four PLS components rather than all eight PLS components. These multivariate statistics based models are found effective and promising, hence can help avoid or reduce the need of sampling and experimental analysis for the effluent COD, BOD and TOC, because these can be estimated using soft sensors based on these PLS models using measured influent quality parameters.
Keywords : Primary clarifier, partial least square, principal component analysis..

DOI= http://dx.doi.org/10.20902/IJCTR.2021.140125

1. Introduction

Discharge of inadequately treated wastewater from domestic or industrial sources results in severe ecological issues. The overall performance of the whole wastewater treatment plant depends upon the performance of each process involved in wastewater. The primary clarifier is an important equipment in wastewater treatment plants and its efficiency affects the performance of the subsequent processes such as biochemical processes to remove the organic matters from the wastewater. The monitoring and control of the processes require on-line sensors which are expensive and requires maintenance. Various mathematical models have been reported to assess the behavior of the sedimentation process occurs in the primary and secondary clarifiers. A sedimentation model based on solid flux theory is reported[1]. The sludge settling velocity models were investigated and compared[2]. A generalized multi-layer model for sedimentation process and validated the model using field as well as pilot scale data. The model is capable of estimating solid profile, effluent and retentate suspended solid concentrations[3]. The one-dimensional model is applied to estimate the sludge blanket height in the secondary clarifier[4]. The simulation study of clarifiers with constant cross sectional area and varying cross sectional area, the effluent concentration was predicted in case of constant cross sectional area[5]. The cylindrical settler model was extended to conical settler. It has been shown that conical settler can handle a wide range of solids loading, an estimation of the characteristics of batch settling for conical settler was carried out and compared with experimental data[6]. The simulation of models with different values of fee flow rate and feed concentration[7]. It shows that at a low influent flux, the amount of solids transported to the effluent is negligible. A moderate increase of the influent solid flux prompts a higher steady state concentration in the underflow, while the effluent concentration remains unaffected. An increase in large influent flux overloads the settler, resulting in a non-negligible steady state effluent concentration. Also this study shows inconsistency of prediction with respect to the number of layers. The mathematical model for continuous sedimentation process of flocculated suspensions was simulated with stepwise change in feed concentration and it reveals that the model realistically describes the dynamics of flocculated suspensions in clarifier-thickner[8]. A fuzzy algorithm was developed for controlling sludge height in the secondary clarifier. The developed control strategy is based on on-line data of influent flow, removal and recycle flows, daily analytical values of sludge volume index. The developed controller has been applied to activated sludge wastewater treatment plant model located in spain and observed that the algorithm allow reduction in sludge height variations and thus increase the settling process efficiency[9]. The on-line instruments to determine the solid flux density function and the solid effective stress, the solid flux density function is used for flocculant selection and dosage optimization[10]. In this all studies the main objective is to assess the performance of clarifiers, estimation of effluent or underflow solids concentration. In one-dimensional model of the clarifier the knowledge of the system behavior is required. In industrial WWTP at the primary clarifier level many other pollutants such as pH, chemical oxygen demand (COD), total organic carbon (TOC), biochemical oxygen demand (BOD), total dissolved solids (TDS), total suspended solids (TSS), amoniacal nitrogen (NH_3-N) are also measured. In literature different techniques are reported to estimate the effluent pollutants based on measured influent pollutants. A large number of data is generally available based on daily analysis of samples collected at input and output side of the process or data acquisition system employed for monitoring. Generally statistical and artificial intelligent methods are used to develop the model based on available measured data.

Various multivariate statistics methods like PLS, NNPLS, KPLS, PCR, MPLS, MLR, APLS, RAPLS, QPLS are reported in the literature. In an early application for municipal activated sludge process, total phosphorus (TP) and COD and turbidity were estimated using conventional PLS[11]. A PLS model for estimating influent TP concentrations in a municipal WWTP was designed and the prediction accuracy of a model based on daily lab- oratory analyses was quite acceptable[12]. The Robust Adaptive PLS (RAPLS) for prediction of the total oxygen demand (TOD) in industrial wastewater[13]. The different data-driven modelling methods were compared for approximating COD concentration in the primary clarifier effluent and NH_4- N concentration in the bioreactor of a municipal ASP[14]. They observed Generalized Least Squares Regression (GLSR) estimates to be less accurate than the ANN estimates. The static models were not successful in approximating the output variables, whereas the use of the Kalman filter remarkably improved the predictions of the DSVI[15]. Besides the predictions of effluent pollutant variables of the WWTP processes the statistical approaches could be used for the abnormal operation and online sensor fault diagnostics. A dynamic concurrent kernel partial least squares (DCKPLS) method was proposed for process monitoring and the performance of the process was evaluated by simulated sensor faults of industrial WWTP data[16]. As reported in literature various data driven techniques for soft sensor are applied to activated sludge process, but the soft sensor development for primary or secondary

clarifiers is rarely attended. We have reported mechanistic model and fuzzy inference system for the prediction of total suspended solids present in the effluent of primary clarifier[17]. In this study the multivariate statistical techniques are applied for the prediction of effluent quality parameters (COD, BOD and TOC) of primary clarifier to investigate effectiveness of these soft sensing techniques. This can help save significant time and efforts otherwise required in sampling and lab analysis towards measurement of these effluent quality parameters, whereas another alternative approach of using online sensors is not commercially viable due to any such available online sensors usually being quite expensive.

1. Multivariate statistical methods

Partial least square regression (PLSR) and principal component regression (PCR) are the methods to model a response variable when there are a large number of predictor variables, and those predictor are highly correlated. Both methods construct new predictor variables, identified as components, as linear combinations of the original predictor variables. PCR builds components to describe observed variability in the predictor variables, without considering the response variable at all. PLSR creates components to explain observed variability in the predictor variable, considering the predictor as well as response variables.

1.1 Principal component analysis (PCA)

In a data set with many variables, group of variables often move together. The reason for this is that more than one variable might be measuring the same driving principle governing the behavior of the system. If there are a large number of groups of variables in a system than, it allows to take advantage of this redundancy of information. The group of variables can be replaced by single variable. PCA generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal component as a whole form an orthogonal basis for the space of the data[18]. The first principal component is a single axis in space, when each observation on that axis is projected the resulting values for a new variable and the variance of this variable is the maximum among all possible choices of the first axis. The second principal component is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis. The full set of principal components is as large as the original set of variables. But it is common place for the sum of the variance of the first few principal components to exceed 80% of the total variance of the original data.

1.2 Partial least square regression (PLSR)

Partial least squares (PLS) is a wide class of methods for modelling relations between sets of observed variables by means of latent variables. In its general form PLS creates orthogonal score vectors (latent vectors or components) by maximizing the covariance between different set of variables. PLS is dealing with two blocks of variables, one predictor variables' block and other response variables' block[19]. Let $x \in R^N$ is the N-dimensional space of predictor variables block and $y \in R^M$ is M-dimensional response variables block. The PLS decomposes the (n x N) matrix of zero-mean variables' block X and the (n x M) matrix of zero-mean variable Y in to the form

$$X = TP^T + E \qquad (1)$$

$$Y = UQ^T + F \qquad (2)$$

Where the T, U = (n x p) matrices represent latent vectors

P, Q = (N x p) and (M x p) matrices of loading vectors

E, F = (n x N) and (n x M) are matrices of residuals

2. Results and Discussion

2.1 Data collection and preprocessing

The study is based on the clariflocculator process at common effluent treatment plant (CETP), Vatva, Ahmedabad. The clariflocculator consists of two concentric tanks, the inner tank acts as flocculator and the outer tank acts as clarifier as shown in Fig. 1. For the same primary clariflocculator effluent total suspended solids (TSS_e) prediction based on influent flow rate (Q_f) and influent total suspended solids (TSS_{in}) were carried out by incorporating mechanistic and fuzzy inference system by us. The first principle based and fuzzy based soft sensors were reported for the CETP primary clarifier by us.[17]

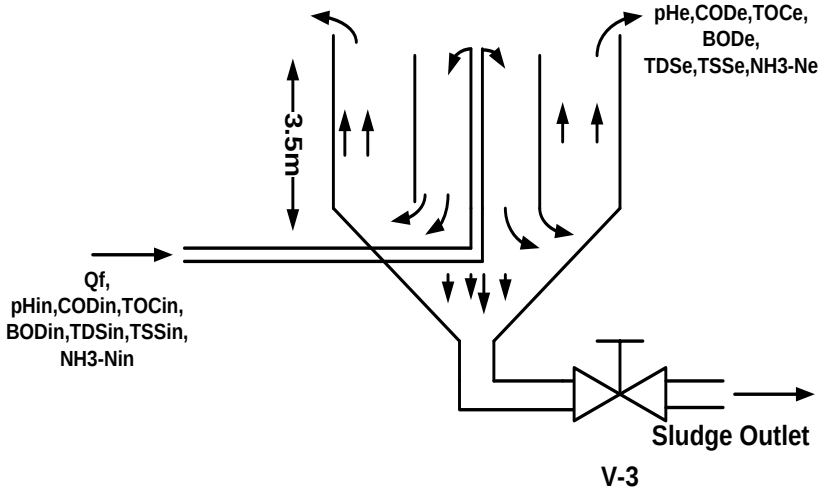


Fig. 1 Clariflocculator at CETP

The pollutants measured at inlet side of the clariflocculator are feed flow rate (Q_f), pH_{in}, chemical oxygen demand (COD_{in}), Total organic carbon (TOC_{in}), biochemical oxygen demand (BOD_{in}), total dissolved solids (TDS_{in}), total suspended solids (TSS_{in}) and ammoniacal nitrogen ($NH3\text{-}N_{in}$). The pollutants measured at the outlet side of the clariflocculator process are pH_e, chemical oxygen demand (COD_e), Total organic carbon (TOC_e), biochemical oxygen demand (BOD_e), total dissolved solids (TDS_e), total suspended solids (TSS_e) and ammoniacal nitrogen ($NH_3\text{-}N_e$). The measurement of COD, BOD takes more time and online sensors are costly. The dataset consists of different range for each pollutant variable, so first the dataset is normalized using Z score normalization, where x, μ and σ are score, mean and standard deviation.

$$z=((x-\mu))/\sigma \qquad (3)$$

The normalized data is shown in Fig. 2. The red dots represent outliers present in the dataset. The outliers are removed from the dataset and rest of the data is used for multivariate statistical models.

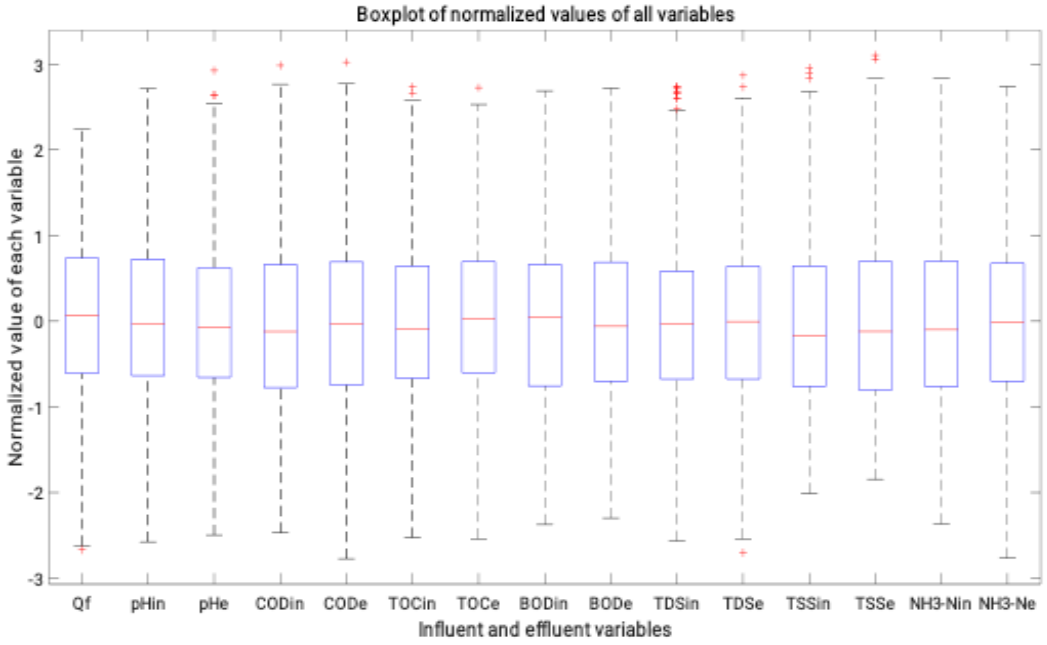


Fig.2 Normalized values of all influent and effluent variables

3.2 Correlation coefficient

A correlation coefficient is a numerical measure indicating statistical relationship between two variables. If the correlation coefficients lie between two variables in the range of 0.5 to 0.7, 0.8 to 1 than relationship is moderate and strong respectively. The correlation coefficients among all influent and effluent pollutants are represented in Table 1.

Table 1 Correlation coefficients among all influent and effluent pollutants[17]

	pH_e	TSS_e	TDS_e	COD_e	TOC_e	BOD_e	NH_3-N_e
Q_f	-0.23	-0.2	-0.28	-0.34	-0.38	-0.27	0.06
pH_{in}	0.7	0.02	-0.11	0.03	-0.03	-0.07	0.13
TSS_{in}	0.41	0.65	0.18	0.38	0.31	0.25	-0.02
TDS_{in}	0.17	0.31	0.87	0.65	0.58	0.25	0.05
COD_{in}	0.36	0.41	0.63	0.88	0.69	0.42	0.12
TOC_{in}	0.36	0.37	0.60	0.76	0.86	0.40	0.08
BOD_{in}	0.23	0.31	0.32	0.58	0.58	0.82	-0.16
NH_3-N_{in}	-0.07	-0.08	-0.12	0.07	-0.07	-0.12	0.88

As per the values represented in Table 1, the following relationship among effluent and influent pollutants exist.

$TSS_e = f(TSS_{in})$ (4)

$COD_e = f(COD_{in}, TOC_{in}, TDS_{in}, BOD_{in})$ (5)

$TDS_e = f(TDS_{in}, COD_{in}, BOD_{in})$ (6)

$TOC_e = f(TOC_{in}, COD_{in}, TDS_{in}, BOD_{in})$ (7)

$BOD_e = f(BOD_{in})$ (8)

$NH_3\text{-}N_e = f(NH_3\text{-}N_e)$ (9)

3.2 PCA of plant data

The principal component is applied to all influent variables of primary clariflocculator of the plant. The loading plot of principal component (PC1) and PC2 is shown in Fig.2, the influent pollutants COD_{in}, TOC_{in}, TDS_{in}, BOD_{in}, TSS_{in} lie in the same group. The loading of each component is represented in Table 2.

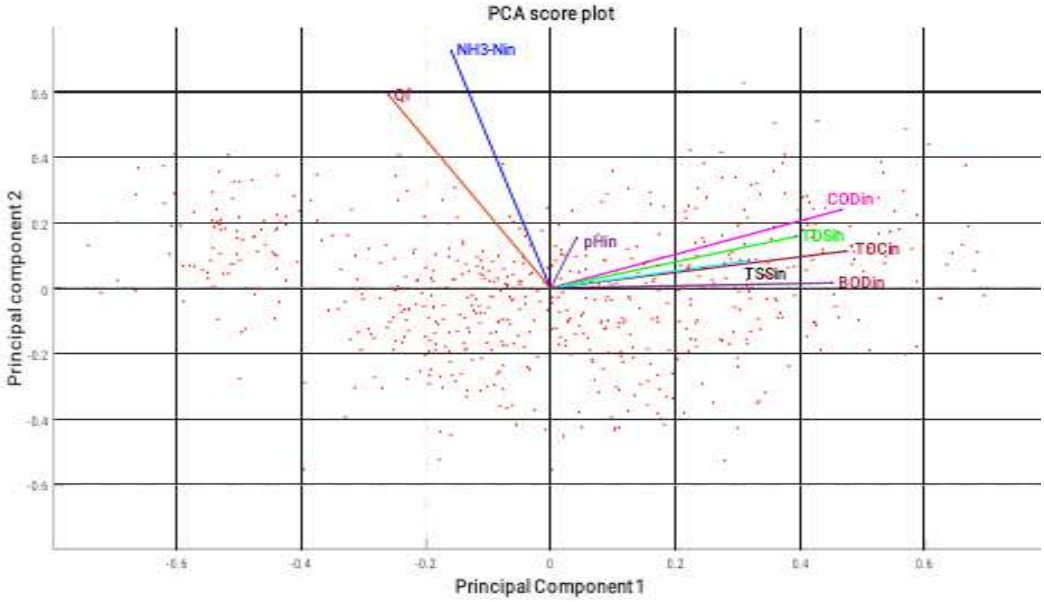


Fig.2 PCA loading plot

Table 2 Loading of each principal component

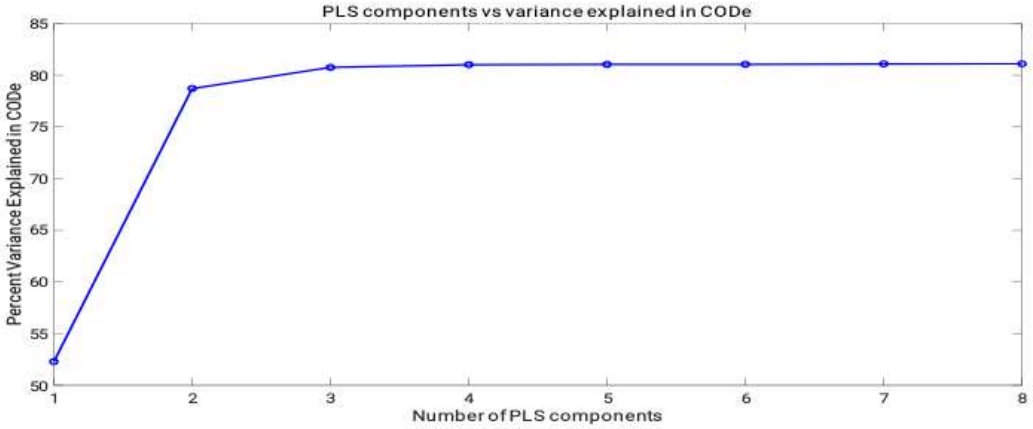
Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Q_f	-0.2601	0.5905	-0.2066	0.6610	-0.1755	0.2665	0.0447	-0.0171
pH_{in}	0.0418	0.1542	0.8919	-0.0160	-0.3571	0.1978	-0.0530	-0.0954
COD_{in}	0.4669	0.2401	-0.0113	-0.0247	-0.1730	-0.1142	0.1680	0.8077
TOC_{in}	0.4748	0.1133	-0.1076	-0.0367	-0.1494	-0.0267	0.6960	-0.4913
BOD_{in}	0.4523	0.0162	-0.1483	0.1301	0.1924	0.7879	-0.3087	-0.0555
TDS_{in}	0.3945	0.1580	-0.1706	0.0643	-0.3791	-0.4212	-0.6184	-0.2875
TSS_{in}	0.3187	0.0845	0.3146	0.4452	0.7226	-0.2549	-0.0566	-0.0608
NH_3-N_{in}	-0.1592	0.7244	0.0030	-0.5845	0.2939	-0.1134	-0.0413	-0.0856

As per the Table 2, principal component 1 (PC1) exhibits that COD_{in}, TOC_{in}, BOD_{in},TDS_{in} are group of variables with collinear relationship. The percentage variance explained by PC1, PC2, PC3, PC4, PC5, PC6, PC7 and PC8 are 37.3, 15.1, 13.2, 10.8, 8.6, 7.5, 3.8 and 3.3 respectively. The first six principal components out of eight components explain 85% variance. In place of considering all PCA variables the first six are representative of the influent variables.

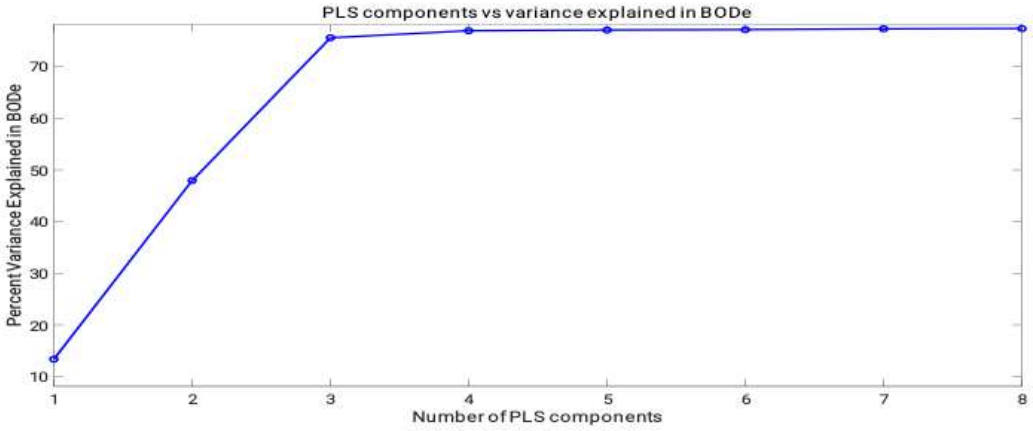
3.3 Partial least square models

The partial least square models for the estimation of COD_e, BOD_e and TOC_e as a function of other measured influent variables have been developed. The data within the range of influent variables Q_f (218-1080 m^3/h), pH_{in} (7.14-8.24), COD_{in} (1180-3800 mg/L), TOC_{in} (273-1270 mg/L), BOD_{in} (212-1333mg/L), TDS_{in} (10980-24820 mg/L), TSS_{in} (140-1196 mg/L), NH_3-N_{in} (31-175 mg/L) are considers for PLS models.

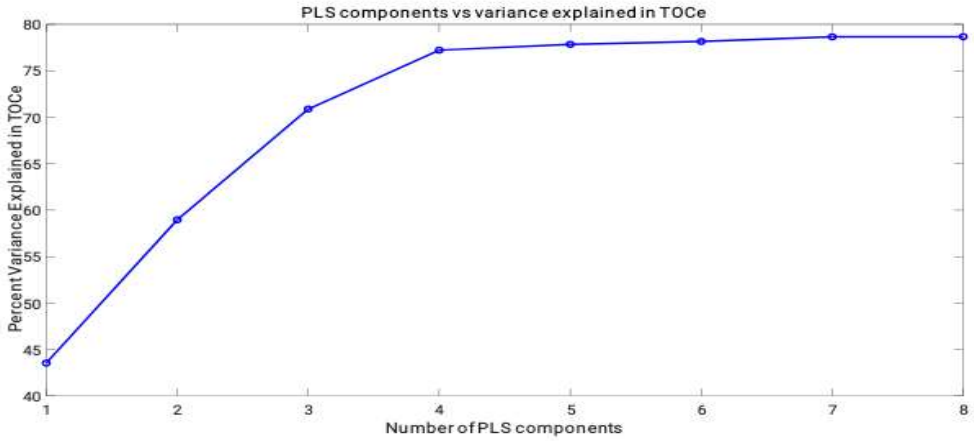
The partial least square technique converts the original dataset into number of PLS components. Fig.3 (a,b,c) represents the percentage variance explained by the eight components, in COD_e, BOD_e and TOC_e. The first three PLS components explain 80% variance in COD_e , in BOD_e first three PLS components explain 78% variance, and in TOC_e 78% variance is explained by first four PLS components.



(a) Variance explained by PLS components in COD_e



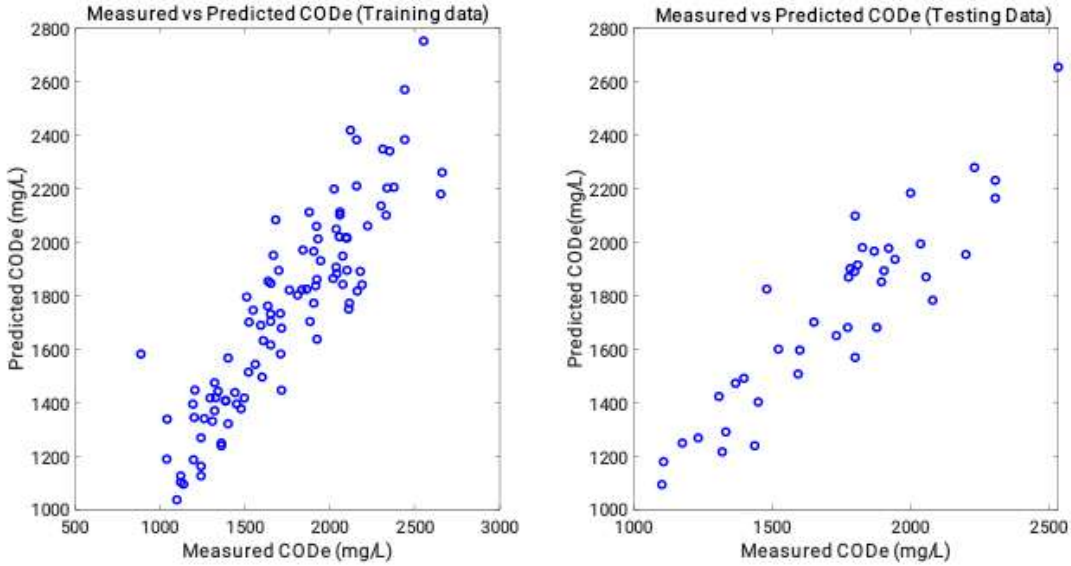
(b) Variance explained by PLS components in BODe



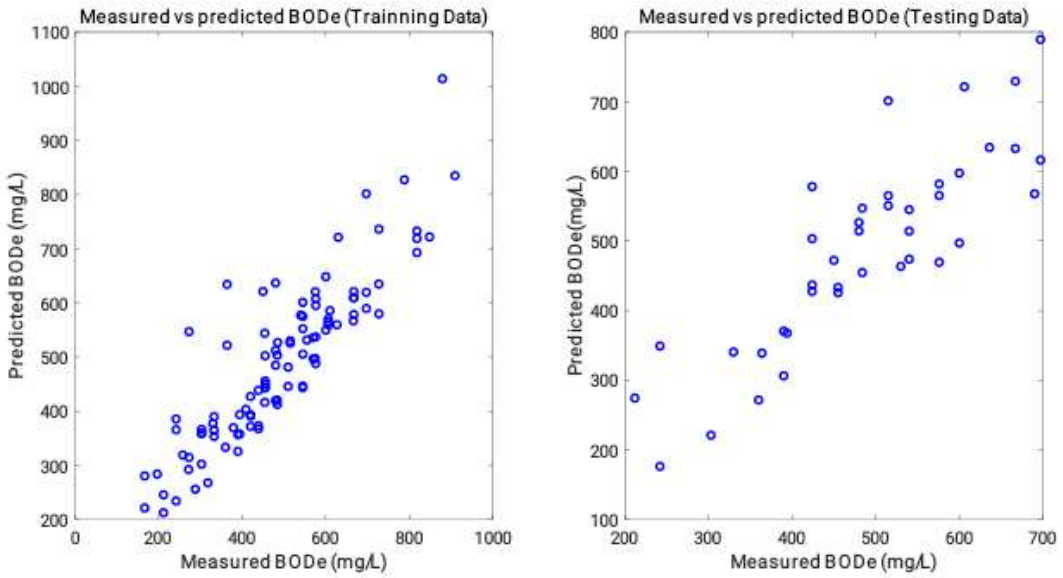
(c) Variance explained by PLS components in BODe

Fig. 3 Percentage variance explained by PLS components in COD_e, BOD_e and TOC_e

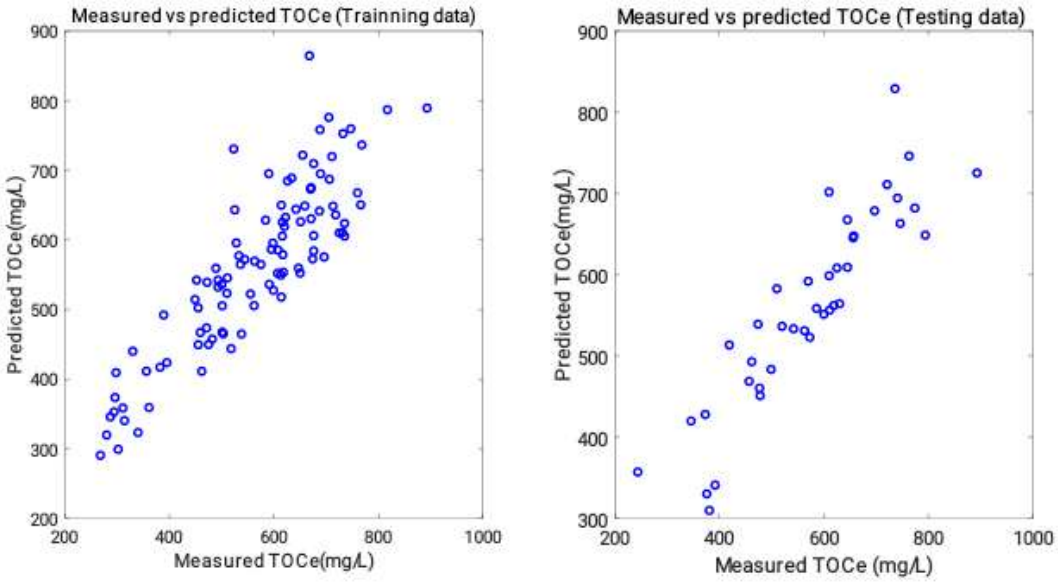
For development of PLS models 140 days data in the mentioned range of influent variables are considered, 70% of data is considered for model training and 30% data is considered for model testing. The results of the PLS models of COD_e, BOD_e and TOC_e for training and test data are shown in Fig. 4.



(a) Predicted vs measured CODe



(b) Predicted vs measured BODe



(c) Predicted vs measured TOCe

Fig.4 Measured vs predicted COD_e, BOD_e, TOC_e for training and test datasets

The error analysis of all the three PLS models for the estimation of COD_e, BOD_e and TOC_e are shown in the Table 3.

Table 3 Error analysis of PLS models

	COD_e		BOD_e		TOC_e	
	Training data sets	Testing data sets	Training data sets	Testing data sets	Training data sets	Testing data sets
R^2	0.80	0.84	0.78	0.68	0.77	0.8
Normalized RMSE (%)	10.3	9.7	10.4	14.6	10.7	9.6
MAPE (%)	8.3	6.4	13.9	12.3	9.8	9.2

The correlation coefficients among influent and effluent variables show that COD_e, BOD_e, TOC_e are related with influent COD_{in}, BOD_{in} and $TOC_{in.}$ Also the principal component analysis exhibits that COD_{in}, BOD_{in}, TDS_{in} and TOC_{in} are correlated. The PLS model of COD_e is promising as compared to PLS models of BOD_e and TOC_e with reference to R^2, normalized root mean squared error (NRMSE) and mean absolute percentage error (MAPE).

4 Conclusion

The relationship among influent and effluent quality parameters of plant primary clarifier is carried out along with collinearity among influent quality parameters COD_{in}, BOD_{in}, TDS_{in}, TOC_{in} using principal component analysis. In case of CODe, BOD_e, TOC_e, the selection of PLS components that explain up to 80% variance result in data size reduction with promising estimation.the explained variance is explained by three PLS components. The applied PLS models can estimate effluent COD_e, BOD_e, TOC_e based on other influent quality parameters. As the COD_e, BOD_e and TOC_e are estimated from the existing influent measured parameters, the time and efforts required for sampling and experimental laboratory analysis for these effluent parameters can be reduced.

Acknowledgement

The authors would like to thank the administration and team members of Common Effluent Treatment plant (CETP), The Green Environment Services Co-operative Society Ltd., Vatva, Ahmedabad, India for providing all necessary support.

References:

1 Vitasovic Z. Continuous Settler Operation: A Dynamic Model Dynamic Modeling and Expert Systems in Wastewater Engineering. Editors: Patry G G and Chapman D, Lewis Publishers, Michigan, 1989.
2 Koehne M, Hoen K, Schuhen M. Modelling and simulation of final clarifiers in wastewater treatment plants, Math and Comp in Simu. 1995; 39: 609-616.
3. Takacs I, Patry G, Nolasco D. A dynamic model of the clarification-thickening process. Wat. Res. 1991; 25: 1263-1271.
4. Chatellier P, Audic J M, A new model for wastewater treatment plant clarifier simulation. Wat. Res. 2000; 34: 690-693.
5. Diehl S, On boundary conditions and solutions for ideal clarifier–thickener units. Chem Engg J. 2000; 80: 119–133.
6 David A, White B, Verdone N. Numerical modelling of sedimentation processes. Chem Engg Sci. 2000; 55: 2213-2222.
7 Liesbeth B, Verdickt J F, Impe V. Simulation analysis of a one-dimensional sedimentation model. 2002; 15: 473-478.
8 Burger R, Karlsen K H, John D. Towers, Mathematical model and numerical simulation of the dynamics of flocculated suspensions in clarifier–thickener. Chem Engg J. 2005; 111: 119–134.
9 Traore A, Grieu S, Thiery F, Polit M, Colprim J, Control of sludge height in a secondary settler using fuzzy algorithms. Comps and Cheml Engg. 2006; 30: 1235–1242.
10 Concha F, Segovia J P, Vergara S, Pereira A, Elorza E, Leonelli P, Betancourt F, Audit industrial thickeners with new on-line instrumentation. Powder Tech. 2017; 314: 680-689.
11 Aarnio P, Minkkinen P, Application of partial least-squares modelling in the optimization of a waste-water treatment plant. Anal. Chim. Acta. 1986; 191: 457-460.
12 Blom, H A., Indirect measurement of key water quality parameters in sewage treatment plants. J. Chemometr. 1996; 10: 697-706.
13 Lee H W, Lee M W, Park J M, Robust adaptive partial least squares modeling of a full-scale industrial wastewater treatment process. Ind. Eng. Chem. Res. 2007; 46: 955-964.
14 Urrenmatt D, Gujer W, Data-driven modeling approaches to support wastewater treatment plant operation. Environ. Model. Softw. 2012, 30: 47-56.
15 Teppola P, Mujunen S P, Minkkinen P, Kalman filter for updating the co- efficients of regression models. A case study from an activated sludge waste- water treatment plant. Chemometr. Intell. Lab. 1999; 45: 371-384.
16 Liu H, Yang J, hang Y, Yang C, Monitoring of wastewater treatment processes using dynamic concurrent kernel partial least squares. Process Safety and Env Prote. 2021; 147: 274-282.
17 Patel N, Ruparelia J, Barve J. Prediction of total suspended solids present in effluent of primary clarifier of industrial common effluent treatment plant: Mechanistic and fuzzy approach, J. Wat. Pro. Engg. 2020; 34: 101146.
18 https://in.mathworks.com/help/stats/principal-component-analysis-pca.html, accessed on 21/12/2020.
19 https://link.springer.com/chapter/10.1007/11752790_2.

*****