Non-linear modeling using fuzzy principal component regression for Vidyaranyapuram sewage treatment plant, Mysore – India
Ayesha Sulthana, K. C. Latha, Mohammad Imran, Ramya Rathan, R. Sridhar and S. Balasubramanian

ABSTRACT
Fuzzy principal component regression (FPCR) is proposed to model the non-linear process of sewage treatment plant (STP) data matrix. The dimension reduction of voluminous data was done by principal component analysis (PCA). The PCA score values were partitioned by fuzzy-c-means (FCM) clustering, and a Takagi–Sugeno–Kang (TSK) fuzzy model was built based on the FCM functions. The FPCR approach was used to predict the reduction in chemical oxygen demand (COD) and biological oxygen demand (BOD) of treated wastewater of Vidyaranyapuram STP with respect to the relations modeled between fuzzy partitioned PCA scores and target output. The designed FPCR model showed the ability to capture the behavior of non-linear processes of STP. The predicted values of reduction in COD and BOD were analyzed by performing the linear regression analysis. The predicted values for COD and BOD reduction showed positive correlation with the observed data.

INTRODUCTION
To overcome the upsurge of water pollution there is a need for effective wastewater treatment; therefore competent modeling and efficient monitoring of wastewater treatment systems are very much essential. The wastewater treatment plant (WWTP) process is a combination of physical, chemical and biological non-linear processes; therefore the WWTP process cannot be modeled by linear statistical approaches (Oliveira-Esquerre et al. 2004). Understanding the behavior of complex non-linear relations between the process variables, the system parameters, the control inputs and the external perturbations of the WWTP is a difficult task. The non-linear behavioral processes are a requirement to develop a model which describes this real-life phenomenon; however, wastewater treatment parameters can be employed to construct the model for predicting its performance (Ruicheng & Xulei 2012). An explicit valid modeling and monitoring technique is an imperative requirement to retain the optimal functioning of wastewater treatment systems. An efficient model intensifies the interpretation of biological non-linear processes and it also constitutes a footing for improved process, operation and control.

Statistical data-based modeling approaches like ‘black-box’ do not require a specific mathematical structure of the process to be modeled; the black-box approach has been used to describe the input–output non-linear relationships of WWTPs (Lee et al. 2002; Motta et al. 2002). Artificial neural networks (ANNs) are non-linear black-box type models, which have been used to model the existing non-linear relationships between the influents and to predict the operational parameters of the WWTP (Mjalli et al. 2007; Mohammed et al. 2012). ANN models were developed to predict the conduct of a WWTP based on the past information (Hamed et al. 2004; Farouq et al. 2007). Optimal coagulant dosage in a drinking water plant was predicted, even in the unexpected conditions such as heavy rain and high turbidity, by ANN models (Sengul & Gormez 2013). Residual aluminum level in
drinking water was predicted by multiple linear regression and ANN models to be used in early warning systems (Tomperi et al. 2013). However, they limit the empirical understanding of operational controls of a process.

Multivariate statistical techniques like principal component analysis (PCA) are employed to reduce the multicollinearity and dimensionality of the data sets for process monitoring of WWTPs (Eriksson et al. 2001). A combination of PCA with subspace identification was used to procure a model which depicts the period-to-period multivariate performance of all the samples of a WWTP (Pan et al. 2004). A robust PCA was applied to remove the effect of outliers on the PCA model for process monitoring of a WWTP (Tharrault et al. 2009).

Nowadays, many studies have been conducted on the applications of computational intelligent techniques related to the predictions of output parameters of WWTPs. Among these, fuzzy systems and neural networks can be employed to create detailed predictive models from the given data. In fuzzy modeling, notably Takagi–Sugeno–Kang (TSK) fuzzy models serve as efficient tools for modeling complex non-linear systems (Takagi & Sugeno 1985; Yen et al. 1998; Nagy et al. 2010). Fuzzy control algorithms have been widely used to predict process parameters of aerobic biological treatment processes (Murnleitner et al. 2002).

Adaptive network-based fuzzy inference system was applied to predict carbon and nitrogen reduction in an aerobic biological treatment plant of the sugar industry (Civelekoglu et al. 2009). Application of neurofuzzy networks enhanced the flow rate forecast of wastewater (Fernandez et al. 2009). Prediction of reduction in effluent parameters of a hospital WWTP was modeled by employing a fuzzy inference system and neural network (Pai et al. 2009). A combination of PCA and TSK fuzzy modeling proved to be efficient in non-linear modeling and fault detection in a WWTP (Yoo et al. 2005). A hybrid fuzzy control path adequately accomplished the mandatory real-time control objectives and was exhibited as a cost-effective tool to accord with the unforeseen uncertainties in the WWTP processes (Chen et al. 2007).

In this study, the PCA of untreated wastewater parameters was carried out to process the data. Fuzzy-c-means (FCM) clustering method was applied to compress the PCA score values to generate the membership functions from the located stable points. A combination of TSK fuzzy model, fuzzy partitioned PCA score and output variables (reduction in biochemical oxygen demand (BOD) and chemical oxygen demand (COD)) were used to build the fuzzy principal component regression (FPCR model). Linear regression analysis was carried out to assess the prediction ability of the FPCR model (Figure 1).

**MATERIALS AND METHODS**

**Study area and data collection**

Mysore district (area 128.42 km², 11 45’ to 12 40’ N and 75 57’ to 77 15’ E) is located within South Karnataka, India. Based upon the existing undulating lie of the land the Mysore city is divided into five drainage districts, the drainage districts ‘A’ and ‘D’ oriented toward the south are assigned to the 60 ML (million litres per day) Rayankere sewage treatment plant (STP). The southwest drainage district ‘B’ is associated with the 67.65 MLD Vidyaranyapuram STP and the northern district ‘C’ is linked to the 30 MLD Kesare STP. The STP for the drainage district ‘E’ (11 skm²) is still under planning and progress. The Vidyaranyapuram STP of drainage district ‘B’ is a biological treatment plant, and was constructed in the year 2002. It covers an area of 27.21 km² with 7,000 m sewer length. Next to an area of a solid waste disposal site the Vidyaranyaapuram STP is located in the vicinity of the Chamundi foothills; the treated wastewater of the STP is discharge into the Dalvai Lake and then flows into the Kabini River, which is a drinking water source. Furthermore the Vidyaranyapuram STP receives more than 50% of sewage generated in Mysore city; on these grounds it was selected as a study area.

The Vidyaranyapuram STP has two facultative aerated lagoons; by every lagoon there are sedimentation basins each possessing a surface area of 50,544 m² (312 m length × 162 m width) and a volume of 176,904 m³ (312 m length × 162 m width × 3.5 m depth). Despite the capacity of the STP being 67.75 MLD, the inflow rate of wastewater fluctuates owing to numerous influencing elements like seasonal changes in rainfall.
changes and tourist inflow; however, there will be an approximate difference of 7 to 9 MLD between the raw wastewater received and the treated wastewater liberated due to seepage. The surface aeration is conducted with the aid of 36 blowers of 20 hp each which are proficiently operated for the assured minimization of the concentrated sludge and foul odor. Further, the STP acquires two maturation ponds of surface area 24,940 m² (172 m length × 145 m width) and a volume of 37,410 m³ (172 m length × 145 m width × 1.5 m depth). In each facultative lagoon the average detention time of wastewater is 11.8 days whereas it is 2.5 days in each maturation pond. The total savings on operational and maintenance cost is about 40%, considering the valid cost of power. The overall power usage is about 4% of that required to operate the conventional STP. Therefore Mysore City Corporation with this technology is able to save electricity and operates the STP with less labour.

To perform the non-linear modeling of the complex processes of a WWTP and to predict the reduction in COD and BOD, the daily basis data (October 2009 to June 2011) from Vidyaranyapuram STP (STP) (latitude of 12° 28′ and longitude of 76° 65′) was collected. The data comprised untreated wastewater parameters of COD, BOD, total solids (TS), total suspended solids (TSS), total dissolved solids (TDS) and chloride, and reduction in BOD and COD in effluent.

**Principal component analysis**

PCA was applied to a data matrix of 607 objects and 8 variables (pH, BOD, COD, TS, TDS, TSS, chloride and nitrate) (Figure 2). The descriptive statistics of untreated wastewater characteristics (eight variables), reduction in BOD and COD is shown in Table 1. The combinations of variables \(X_1, X_2, X_3, \ldots, X_p\), were found to develop uncorrelated orthogonal principal components (PC_1, PC_2, PC_3, \ldots, PC_p) of different dimensions in the dataset. The principal components (PC_3) represent the patterns of association in the data, through which the variables strongly driven by those patterns are grouped. The number of principal components is chosen based on the following equation (Kokot et al. 1998):

\[
PC_{jk} = a_{j1}X_{k1} + a_{j2}X_{k2} + a_{j3}X_{k3} + \cdots + a_{jn}X_{kn}
\]

(1)

where \(PC_{jk}\) is the score for object \(k\) on component \(j\), \(a_{ji}\) is the loading of a variable \(i\) on component \(k\), \(X_{ki}\) is the measured value of a variable \(i\) on object \(k\) and \(n\) is the original number of variable. The transformation of high dimension data generates a new data matrix, where the greatest variances of data always lie on the PC_1 and the consecutive components represent increasingly less variance. Here, the decrease in dimensionality is achieved without any loss of variance in data.

**Fuzzy principal component regression model**

The PCA model was constructed to reduce the dimension of data and to generate principal component scores; these scores were used to replace the original variables. FCM clustering was used to group the PCA scores and to create membership functions for each cluster. The TSK fuzzy inference system was built using the membership functions created by FCM to predict the reduction in COD and BOD.

**FCM clustering**

The FCM clustering (Bezdek 1981) algorithm was applied to the PCA score matrix, wherein each score is assigned to a centroid of cluster with different membership values. The fuzzy partitioning in membership matrix \((U)\) was randomly initialized conforming to Equation (2). The value of the centroid is calculated by Equation (3). The fuzzy partitions are dependent upon the exponent \(m\); a higher value of exponent builds an indistinct outer limit between the different clusters. The function of dissimilarity used in FCM is computed using Equation (4)

\[
\sum_{i=1}^{C} u_{ij} = 1, \quad \forall j = 1, \ldots, n
\]

(2)
Table 1 | Descriptive statistics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. error</th>
<th>Std. deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>607</td>
<td>9.80</td>
<td>9.80</td>
<td>9.80</td>
<td>7.7551</td>
<td>.03918</td>
<td>.96528</td>
<td>.932</td>
</tr>
<tr>
<td>BOD</td>
<td>607</td>
<td>252.00</td>
<td>98.00</td>
<td>350.00</td>
<td>268.8537</td>
<td>.98483</td>
<td>24.26351</td>
<td>588.718</td>
</tr>
<tr>
<td>COD</td>
<td>607</td>
<td>263.00</td>
<td>167.00</td>
<td>430.00</td>
<td>301.2622</td>
<td>1.90044</td>
<td>46.82188</td>
<td>2192.289</td>
</tr>
<tr>
<td>TS</td>
<td>607</td>
<td>1021.00</td>
<td>979.00</td>
<td>2000.00</td>
<td>1626.4959</td>
<td>5.04680</td>
<td>229.66549</td>
<td>52746.239</td>
</tr>
<tr>
<td>TDS</td>
<td>607</td>
<td>1935.92</td>
<td>340.00</td>
<td>2275.92</td>
<td>1606.2689</td>
<td>9.32183</td>
<td>229.66549</td>
<td>52746.239</td>
</tr>
<tr>
<td>Chloride</td>
<td>607</td>
<td>652.00</td>
<td>38.00</td>
<td>690.00</td>
<td>349.8151</td>
<td>8.63610</td>
<td>75.60242</td>
<td>9751702</td>
</tr>
<tr>
<td>TSS</td>
<td>607</td>
<td>250.00</td>
<td>30.00</td>
<td>280.00</td>
<td>59.0190</td>
<td>.83601</td>
<td>20.59704</td>
<td>424.238</td>
</tr>
<tr>
<td>Nitrate</td>
<td>607</td>
<td>56.00</td>
<td>24.00</td>
<td>80.00</td>
<td>45.8425</td>
<td>.44182</td>
<td>10.88530</td>
<td>118490</td>
</tr>
<tr>
<td>BOD Reduction</td>
<td>607</td>
<td>37.00</td>
<td>1.00</td>
<td>38.00</td>
<td>20.0917</td>
<td>.29796</td>
<td>7.34092</td>
<td>53889</td>
</tr>
<tr>
<td>COD Reduction</td>
<td>607</td>
<td>678.08</td>
<td>−8.08</td>
<td>670.00</td>
<td>149.6728</td>
<td>3.03580</td>
<td>74.79406</td>
<td>5594152</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>607</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where \( u_{ij} \) is between 0 and 1, \( C \) is the number of clusters and \( n \) is number of data points

\[
C_i = \frac{\sum_{i=1}^{n} u_{ij}^2 x_j}{\sum_{j=1}^{n} u_{ij}^m}
\]  (3)

\( u_{ij} \) is the degree of membership of \( x_j \); whether \( x_j \) is assigned to cluster \( i \) \((u_{ij} = 1)\) or not \((u_{ij} = 0)\); \( x_j \) is the \( j \)th component of measured data; \( C_i \) is the centroid of cluster \( i \); \( m \) is any real number greater than 1 \((1 < m < \infty)\)

\[
J(U, c_1, c_2, \ldots, c_C) = \sum_{i=1}^{C} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2
\]  (4)

\( J \) is the objective function: \( d_{ij} \) is the Euclidian distance between \( i \)th centroid \( (c_i) \) and \( j \)th data point; it is calculated by Equation (5)

\[
d_{ij} = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2}
\]  (5)

Here \( d_{ij} \) is the squared Euclidian distance, which is expressing the similarity between the measured data \((x_i)\) and center \((c_i)\).

The FCM algorithm is an iterative algorithm, and fuzzy partitioning is accomplished by an iterative optimization of the objective function (Equation (4)), by computing new membership \( u_{ij} \) (Equation (6)) and the cluster centroid \( c_i \) (Equation (3)). However, a maximum of 100 iterations was done to set the cluster centers at the right location.

\[
u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}
\]  (6)

**TSK fuzzy model**

Takagi, Sugeno and Kang (Takagi & Sugeno, 1985; Sugeno & Kang, 1988) introduced the TSK fuzzy model in an attempt to promote a systematic approach of generating fuzzy rules from a given input and output data set (Figure 1). If input1 \((x_1)\) is \( A_1 \) and input2 \((x_2)\) is \( A_2 \), then output \( y = f(x_1, x_2) \), where input1 and input2 are fuzzy sets in the antecedent characterized by membership function, whereas the output \( y = f(x_1, x_2) \) forms a crisp function in the consequence.

IF \( x_1 \) is \( A_{i1} \) and \( x_2 \) is \( A_v \) then output \( y = f(x_1 + x_2 + \ldots + x_{r}) = b_{i0} + b_{i1} x_1 + \ldots + b_{ir} x_r \) for \( i = 1, 2, \ldots, L \)  (7)

where \( L \) = number of rules; \( x_i = (x_1, x_2, x_3, \ldots, x_r)^T \) is input variables; \( y_i = \) output variables; \( f_i = \) linear model; \( A_v = \) fuzzy sets characterized by membership function \( uA_{i1}(x_i) \); \( b_l = (b_{i0}, b_{i1}, b_{i2}, \ldots, b_{ir})^T \) = real valued parameters.

Because of the crisp inputs (nonfuzzy members) of a TSK model the degree of the input matches \( i \)th rule is typically computed using the min operator (Equation (8))

\[
t_i = \min(x_1, x_2)(uA_{i1}(x_1), uA_{i2}(x_2))
\]  (8)

Here \( t_i \) is the matching degree of the rule \( R_i \), whereas \( uA_{i1} \) and \( uA_{i2} \) are the membership functions for input1.
Each rule constitutes a crisp output through weighted average (Equation (9))

\[ y = \frac{\sum_i r_i y_i}{\sum_i r_i} \tag{9} \]

The complete output of the model is computed by the following equation:

\[ y = \frac{\sum_{l=1}^{L} r_{il} y_i}{\sum_{l=1}^{L} r_{il}} \tag{10} \]

The TSK fuzzy inference system was built using FCM clustering by determining the membership functions (Figures 5 and 6). Three PCA scores were used as input variables; reduction in COD and BOD was used as output variables, so that the non-linear property of the treatment system is captured by the model. The following fuzzy rules were determined using the FCM function to identify the clustering nature of input data.

**Rule1**: If Input1 is Input1–cluster1 and Input2 is Input2–cluster1 and Input3 is Input3–cluster1, then output1 is output1–cluster1

**Rule2**: If Input1 is Input1–cluster2, Input2 is Input2–cluster2, Input3 is Input3–cluster2, then output1 is output1–cluster2

**Rule3**: If Input1 is Input1–cluster3 and Input2 is Input2–cluster3 and Input3 is Input3–cluster3, then output1 is output1–cluster3

The correlation studies of predicted values and observed data were carried out through linear regression analysis to analyze the prediction proficiency of the FPCR model.

## RESULTS

### Principal component analysis

The eight input variables were grouped into three principal components, PC1, PC2 and PC3, which explained 60.42% variance in the data set and hence they were selected for non-linear modeling (Table 2 and Figure 3). The results show the reduction in COD and BOD is strongly correlated to the variables (TDS, chloride, BOD, COD and pH) loaded in PC1 (Figure 4). This corresponds to the fact that both chloride ions in association with cations forming inorganic dissolved solids and organic matter subsidizing the organic dissolved solids will highly accelerate COD and BOD in the STP (Thirumalini & Kurian 2009). The variables loaded in PC2 (nitrate and TS) and PC3 (TSS) require oxygen for the decomposition process, interfering in the COD removal rate. However, the system monitoring stationed to the sole PCA is centered only to the static variation provided by the data set and may not satisfy the hypothesis of data normality.
Fuzzy principal component regression

The FPCR model was constructed with three PCA score cluster inputs (cluster 1, cluster 2 and cluster 3) and output variables (reduction in COD and BOD). The PCA score was fuzzy partitioned into three clusters through the FCM algorithm by analyzing the legitimate classification behavior of the inputs to form the number of rules required to best fit the model. The membership functions of PCA score clusters (Figures 5 and 6) and fuzzy rules of the TSK model determined by the FCM function could make a successful linear prediction by capturing the non-linear behavior of the STP influent (untreated wastewater parameters) and effluent (reduction in BOD and COD) data. The linear regression line fitted to the observed data and predicted values obtained through FPCR model gave the significant R-value of 0.967 for COD reduction (Table 3 and Figure 7), whereas the regression line fitted for the BOD reduction model scored a significant R-value of 0.959 (Table 4 and Figure 8).

Table 3 | Model summary (COD reduction)

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. error of the estimate</th>
<th>$R^2$ change</th>
<th>$F$ change</th>
<th>df1</th>
<th>df2</th>
<th>Sig. $F$ change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.967*</td>
<td>0.934</td>
<td>0.934</td>
<td>1.32314</td>
<td>0.934</td>
<td>3614.448</td>
<td>1</td>
<td>254</td>
<td>0.000</td>
</tr>
</tbody>
</table>

*Predictors: (constant), observed.
In this study, the proposed FPCR model is applicable to the STP under specific operating conditions and these conditions cannot be anticipated through this model. Application of this model is confined only to the process parameters of the STP and therefore it cannot be used to establish an entire depiction of the operation of a process. However, the model could detect the operating conditions in the data set of the STP, which includes the parameters involved with the real-life processes. The highly significant R-values for reduction in COD and BOD show the predicting capacity of the FPCR model.

CONCLUSION

The results concluded that collinearity remains in eight wastewater parameters and therefore only three ample principal components (PC1, PC2 and PC3) could successfully explain the static variations in the data and model the output instead of using all the variables. The FCM clustering grouped the PCA score into three partitions to avoid the overfitting complications of the PCA score matrix. The proposed FPCR model has the property of deciphering the PCA score clusters and fuzzy systems, as well as the non-linear modeling intelligence to capture the relation between fuzzy partitioned PCA score clusters and target output (reduction in BOD and COD).

The values of parameters such as BOD and COD are measured by kinetic equations with regard to the biomass or substrate assessed in the effluent of WWTP. Effluent parameter assessment is not adequate when considering the operational performance of the WWTP, whereas forecasting the effluent parameters by relying upon the FPCR modeling and influent water quality allows the operator to take the crucial preventive measures before the problem arises and helps to control the system. However, predicting reduction in BOD and COD based on intelligent techniques is restricted only to the static variation provided by the database, but not to the dynamic variations experienced through the kinetic variables influencing

Table 4 | Model summary (BOD reduction)

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R²</th>
<th>Adjusted R²</th>
<th>Std. error of the estimate</th>
<th>R² change</th>
<th>F change</th>
<th>df1</th>
<th>df2</th>
<th>Sig. F change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.959</td>
<td>0.919</td>
<td>0.919</td>
<td>17.12179</td>
<td>0.919</td>
<td>2704.311</td>
<td>1</td>
<td>238</td>
<td>0.000</td>
</tr>
</tbody>
</table>

*Predictors: (constant), observed.
the degree of effluent degradation. Since fluctuation and deviation in the input data set of the STP is unavoidable, the built-in model should be developed for a larger data set.

ACKNOWLEDGEMENTS

We would like to acknowledge the Mysore City Corporation and management authorities of Vidyaranyapuram STP, Mysore. This work has been supported by Maulana Azad National Fellowship (MANF) Program, University Grants Commission (UGC), New Delhi, India.

REFERENCES


