Desalination and Water Treatment
www.deswater.com

doi: 10.1080/19443994.2015.1108425

57 (2016) 20292–20297
September

Taylor & Francis
Taylor & Francis Group

Factorization of physicochemical parameters of activated sludge process using the principal component analysis

A. Lefkir[a,*], R. Maachou[b], A. Bermad[c], A. Khouider[b]

[a]*Laboratory of TPiTE, ENSTP, Algiers, Algeria, Tel. +213 21511261; email: a_lefkir06@yahoo.fr*
[b]*Laboratory of Faculty of Chemistry, Electrochemistry-Corrosion, Metallurgy and Mineral Chemistry, USTHB, BP 32 El-Allia, PC 16111 Algiers, Algeria*
[c]*Laboratory of Construction and Environment, ENP, Algiers, Algeria*

Received 4 June 2015; Accepted 8 October 2015

ABSTRACT

In order to reduce the complexity of wastewater treatment modeling, a principal component analysis (PCA) was introduced to allow reducing the dimensionality of the original historical data by projecting it into a lower dimensionality space. Indeed, an application of PCA on activated sludge treatment plant was effected to reduce the dimension of the problem described initially by several raw variables of the dominant parameters of the upstream and downstream pollution of the process, such as the physicochemical parameters necessary to describe organic and nitrogen pollutants (SS, COD, BOD, NH_4^+-N, NO_3^--N, NO_2^--N, PO_4^{3-}-P, and TKN, as well as the decision parameters like energy consumption and amount of recirculated sludge. The results show that the performance of the purification process on the energy consumption is primarily related to the excess removal of organic pollution and to excess nitrates product in the process.

Keywords: Wastewater treatment; Activated sludge; PCA

1. Introduction

Many different processes happen simultaneously in wastewater treatment plants (WWTP), which were originally designed to reduce the biological oxygen demand, total suspended solids (SS) and nitrogen and phosphorus pollution [1]. These processes leading to the complexity and the difficulty of understanding the whole system due to the variations in wastewater flow rate and its composition, combined with time-varying reactions in a mixed culture of microorganisms [2,3], to the random aspect of the polluted load injected at the input of the reactors increasing the difficulty of controlling such a process. Moreover, the nature of influents is continuously changing over time, leading to an important variability of the system. The performance of the biological treatment depends on the pollution degradation by the biomass and on the separation of the biomass from the treated water, the sludge settling, all those reasons make this process nonlinear [2] and complex.

The development of computer tools allowed to filter and to exploit data process to obtain the information and relevant knowledge. The extracting techniques of information and knowledge from data have evolved rapidly due to the necessity to reduce the complexity of phenomena.

*Corresponding author.

Recently, multivariate statistical process control, such as principal component analysis (PCA), has been used to monitor the chemical and biological treatment processes.

These multivariate statistical data have several applications, such as multilinear regression using principal components (PCR), reduction of number of variables, identification of structures that explain the most relevant variance of the data and for clustering analysis [4].

PCA reduces the number of variables in a data-set [5] by finding linear combinations of those variables that explain most of the variability and often generates components that have valuable biological meanings. Analysis via this technique produces easily interpretable results, and this method has been successfully applied to many industrial treatment processes [6].

PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e. uncorrelated with) the preceding components. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric.

This paper presents a statistical analysis by PCA in order to minimize and to define the dominant parameters for modeling the activated sludge process. Indeed, PCA includes, the raw water, purified water, elimination yields of pollution parameters such as: SS, organic matter (COD, BOD), nutrients (NH_4^+-N, NO_3^--N, NO_2^--N, and TKN) and phosphorus (PO_4^{3-}-P), recycle sludge (RS) as well as energy consumption.

2. Materials and methods

2.1. Activated sludge plant

Activated sludge treatment is a technically and economically feasible option to treat many types of wastewaters containing highly biodegradable organic matter [7,8]. It is a biological process in which microorganisms oxidize and mineralize organic matter in presence of oxygen by using agitators. The microorganisms grow in the aerated tank and are kept suspended either [9].

The activated sludge plant of this research located in Boumerdes (Algeria), it is within the "extended aeration activated sludge" category. This site has a processing capacity of 75,000 inhabitant equivalents with a low mass loading (of the order of 76 kg (BOD)/kg (VSS)/d). It is designed to treat domestic sewage, and the daily nominal flow is 15,000 m^3/d. It mainly consists of several biological reactors (aerated tanks), and solid–liquid separators (secondary clarifiers or settlers).

2.2. Description of PCA method

PCA is a multivariate statistical data analysis that uses projection into latent variables to reduce high-dimensional and strongly correlated data to a much smaller data-set that can then be interpreted. This approach is important for problems with a large number of input variables and features in chemical and biological processes [10]. PCA aims at finding and interpreting hidden complex and relationships between features in datasets [11,12]. Correlating features are converted to the so-called factors, which are themselves none correlated. PCA modeling, i.e. the approximation of a matrix by a model, defined by variables and a relatively small number of outer vector products, shows the correlation structure of a data matrix X, approximating it by a matrix product of lower dimension, called the principal components (PC), plus a matrix of residuals [12].

In PCA, the original data are projected onto principal component axes. Each of the principal components, PCs, captures as much as possible of the variation which has not been explained by the former PCs, i.e. the first PC maximizes the covariance in the original data and the subsequent PCs maximize the covariance in the residual matrices thatare left after extracting the former PCs [13]. This means that the first component will be correlated with at least some of the observed variables. It may be correlated with many.

In computational terms, the principal components are found by calculating the eigenvectors and eigenvalues of the data covariance matrix. This process is equivalent to finding the axis system in which the covariance matrix is diagonal. The eigenvector with the largest eigenvalue is the direction of greatest variation, the one with the second largest eigenvalue is the (orthogonal) direction with the next highest variation and so on.

The PCA considers "P" variables for which we arrange of "N" individuals. The individual "i" is described by the vector belonging to R^P:

$$X_i = \{X_{ij}/j = 1 \text{ to } P\} \quad (1)$$

The term X_{ij} is a real number that represents the measurement of the variable X_j on individual i. On an individual, there are a number of variables. The variable "j" is described by the vector R^n:

$$X_j = \{X_{ij}/i = 1 \text{ to } N\} \quad (2)$$

The matrix [X] resulting of the crossing "$N \times P$" constitutes the matrix of data.

The covariance matrix between X_j and X_k variables is given by:

$$\text{Cov}\,(X_j, X_k) = \frac{1}{N}\sum_{i=1}^{N}\left(X_{ij} - \overline{X_{ij}}\right) \times \left(X_{ik} - \overline{X_{ik}}\right) \quad j = 1, P; \quad k = 1, P \quad (3)$$

The initial variables undergo, sometimes, a change in reduced centered variables in order to reduce the distortion of valuable scales and to make dimensionless variables on the other hand. The matrix of covariance, in this case, describes the matrix of correlation between variables X_j and X_k and it is given by the following:

$$\text{Cor}\,(X_j, X_k) = \frac{\text{Cov}\,(X_j, X_k)}{S_j \times S_k} = \frac{\sum_{i=1}^{N}\left(X_{ij} - \overline{X_j}\right) \times \left(X_{ik} - \overline{X_k}\right)}{\left[\sum_{i=1}^{N}\left(X_{ij} - \overline{X_j}\right)^2 \times \sum_{i=1}^{N}\left(X_{ik} - \overline{X_k}\right)^2\right]^{1/2}} \quad j = 1, P; \quad k = 1, P \quad (4)$$

We note that:

$$[A] = \{\text{Cor}\,(X_j, X_k), \quad j = 1, P;\ k = 1, P\} \quad (5)$$

Note that the correlation matrix [A] is a symmetric matrix definite positive, it is therefore diagonalizable. The correlation matrix is replaced by a diagonal matrix noted [D] by reducing the number of variables necessary to describe individuals with a minimal loss of information.

The [D] matrix is obtained after resolution of the following polynomial equation:

$$\text{Det}\,(A - \lambda_i I) = 0 \quad (6)$$

where [I] is the Identity Matrix with ($P \times P$) dimension, λ_i are called the eigenvalues and represent the diagonal values of the diagonal matrix [D].

The (λ_i) values represent also the rates explanation of axis F_i.

These new variables are called principal components (PCs). PCs, Ratter F_j, represented as a linear combination of the X_j variables, which are calculated from the eigenvectors of the correlation matrix:

$$(A - \lambda_j I)F_j = 0 \quad (7)$$

The PCA consists to rigidly rotate the axes of this p-dimensional space to new positions (principal axes) that have the following properties:

Ordered such that principal axis $F1$ has the highest variance (λ_1) and the last axis F_p has the lowest variance (λ_p).

3. Results and interpretations

3.1. Application of PCA on input/output parameters

In order to reduce the number of the pollutants parameters, a PCA was applied. The parameters of the raw water and purified water used for this analysis are as follows: SS, COD, BOD, NH_4^+-N, NO_3^--N, NO_2^--N, PO_4^{3-}-P, and TKN. We symbolize the raw parameters by X_{raw} and purified parameters by X_{pur}. The excess parameters of nitrite and nitrate are symbolized by ENO_2 and ENO_3. The circle correlation is represented in Fig. 1.

A PCA analysis indicated two principal components. The first principal component ($F1$) explained 24.96% of the total variance and contains most of the information. The second principal component ($F2$) explains 11.75% of the total variance.

We note a provision of parameters pollution on two fictitious arcs, the first contains parameters of the raw water, the second those of purified water. The equidistant between these two fictitious arcs define the drawdown of pollution according to the vocation of the WWTP, in contrast to the parameters NO_3 and NO_2 which process is not carefully controlled considering NO_3 and NO_2 are in excess.

It is therefore more interesting to consider, in the following, corresponding to the drawdown's pollution parameters as analysis variables instead of raw variables at input and output of the step, which will further reduce the number of variables.

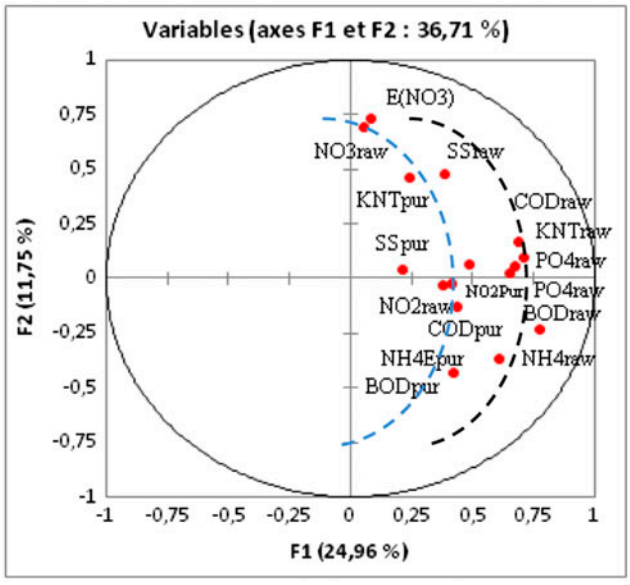

Fig. 1. Projection of the input/output variables on the first principal axis.

3.2. Application of PCA on elimination yields parameters and energy

It is appropriate to apply a PCA allowing that the elimination yields and energy. We symbolize the raw parameters by Y_x, its expression is as following:

$$Y_x[\%] = 1 - X_{pur}/X_{Raw}$$

There will therefore be:

(1) Energy [kWh/m^3] = energy consumption [kWh/d]/input flow [m^3/d].

(2) Explanatory variables: Reports elimination of pollution.

The elimination yield is presented in percentage of removal parameters. We symbolize it by $Y(X)$ of each parameter (X) and the energy is symbolized by En.

We obtain two circles of correlation by applying the PCA analysis. The first circle indicated two principal components explained 33.83% of the total variance and contains most of the information. The second explains 29.88% of the total variance.

By examining the correlation circle formed by the axes ($F1$ and $F2$), we note a provision of parameters pollution on fictitious arc that contains parameters En, Y(BOD), Y(COD), Y(SS), and Y(PO$_4$) show that the energy is related to the removal of organic matter expressed by elimination yield of BOD, COD, and SS as well as the elimination yield of NH$_4$ expressing the degree of nitrification.

By examining the correlation circle formed by the axis $F1$ and $F3$, Fig. 2(b), the same findings are made to Fig. 2(a) on energy variables (En), Y(BOD$_5$), Y(BOD), Y(COD), and Y(SS). We note also that the removal yields of the parameters (E(NO$_3$), Y(NH$_4$), Y(KNT), and E(NO$_2$) are arranged on a same fictitious arc which expressing the degree of nitrification of NH$_4$ to NO$_2$ and NO$_3$.

3.3. Application of PCA on elimination yields parameters and RS

In the following, application of PCA on the elimination yields and the quantity of the RS.

Recycle sludge [%] = recycle flow [m^3/d]/input flow [m^3/d]. We symbolize recirculated sludge by RS.

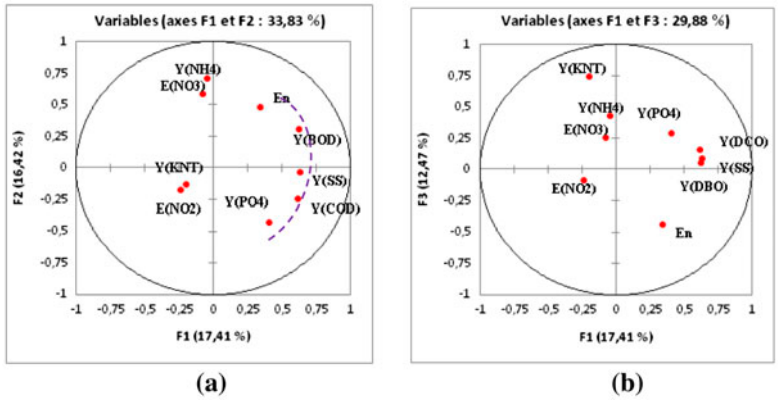

Fig. 2. Projection of the removal yields and energy on the first principal axis: (a) axis $F1$, $F2$ and (b) axis $F1$, $F3$.

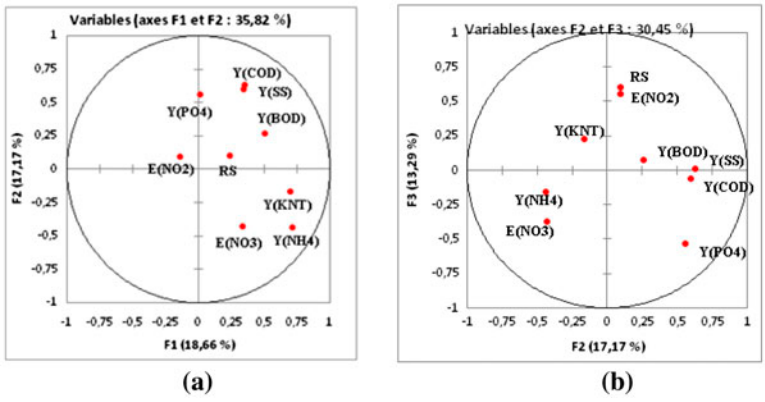

Fig. 3. Projection of the removal yields and RS on the first principal axis: (a) axis *F*1, *F*2 and (b) axis *F*1, *F*3.

We obtain two correlation circles by applying the PCA analysis. The first circle indicated two principal components explained 35.82%. The second circle explains 30.45% of the total variance illustrated by Fig. 3.

Correlation circle formed by the main axes (*F*2 and *F*3) shows that the removal yields (*Y*(COD), *Y*(BOD), *Y*(SS)), excess in nitrate E(NO_3), and the RS are arranged on a same axe, the second correlation circle formed by the main axes (2 and 3) shows that the removal yield (*Y*(NH_4), *Y*(KNT)), E(NO_3)), and the RS are arranged on a same axe, this confirms that the recirculation sludge is linked to the ammoniac elimination expressed by *Y*(NH_4) with production of nitrate expressing the degree of nitrification.

By analyzing the circle (2) formed by the main axes (2 and 3), there is a linear arrangement of variables R(TKN), R(NH_4), R(NO_2), and TRB.

Note that the cumulated explanation rate of correlation circles is low (35–49%), this is due to the nonlinearity of the process, while the PCA is a linear analysis approach, expressing the raw variables as a linear combination of the principal components, what gives a low reproductive capacity particularly the explained parameters such as the energy consumed and the amount of recirculated sludge. Nevertheless, the PCA was used successfully to identify the degree of relationship between the raw variables by projecting them into the new coordinate system generated by the principal components, and thus reduce the problem size.

4. Conclusion

Mastery of activated sludge process consists to determine the optimal values of decision parameters for removing the pollution load contained in the wastewater conforming to the standards discharge required by the environment. These parameters included energy deployed into the aeration basin.

The application of PCA allowed us to reduce the number of variables concluded initially determinants for the control of purification process, and consequently, simplification of the problem of optimization thus formulated.

The proposed approach constitutes the methodology for data processing contained in different recorded observations in the plant, in order to extract most information tool for understanding and optimizing of process.

References

[1] N. Topić Popović, I. Strunjak-Perović, R.S. Klobučar, J. Barišić, S. Babić, M. Jadan, S. Kepec, S.P. Kazazić, V. Matijatko, B. Beer Ljubić, I. Car, S. Repec, D. Stipaničev, G.I. Klobučar, R. Čož-Rakovac, Impact of treated wastewater on organismic biosensors at various levels of biological organization, Sci. Total Environ. 538 (2015) 23–37.

[2] A.M. Nagy-Kiss, G. Schutz, Estimation and diagnosis using multi-models with application to a wastewater treatment plant, J. Process Control 23 (2013) 1528–1544.

[3] A.C. Avella, T. Görner, J. Yvon, P. Chappe, P. Guinot-Thomas, P. de Donato, A combined approach for a better understanding of wastewater treatment plants operation: Statistical analysis of monitoring database and sludge physico-chemical characterization, Water Res. 45 (2011) 981–992.

[4] R.K. Tomita, S.W. Park, O.A.Z. Sotomayor, Analysis of activated sludge process using multivariate statistical tools—A PCA approach, Chem. Eng. J. 90 (2002) 283–290.

[5] P. Teppola, S.-P. Mujunen, P. Minkkinen, Partial least squares modeling of an activated sludge plant: A case study, Chemom. Intell. Lab. Syst. 38 (1997) 197–208.

[6] S. Jilanil, M.A. Khan, Pesticide removal in bioaugmented activated sludge using principal component analysis, J. Biodivers. Environ. Sci. (JBES). 11 (2013) 161–170.
[7] S. Tomida, T. Hanai, N. Ueda, H. Honda, T. Kobayashi, Construction of COD simulation model for activated sludge process by fuzzy neural network, J. Biosci. Bioeng. 88 (1999) 215–220.
[8] P. Samuelsson, B. Halvarsson, B. Carlsson, Cost efficient operation of a denitrifying activated sludge process, Water Res. 41 (2007) 2325–2332.
[9] R. Rustum, Modelling activated sludge wastewater treatment plants using artificial intelligence techniques (fuzzy logic and neural networks), PdD thesis, 2009.
[10] C.K. Yoo, A. Vanrolleghem, I.B. Lee, Nonlinear modeling and adaptive monitoring with fuzzy and multivariate statistical methods in biological wastewater treatment plants, J. Biotechnol. 105 (2003) 135–163.
[11] S.W. Choi, I.-B. Lee, Nonlinear dynamic process monitoring based on dynamic kernel PCA, Chem. Eng. Sci. 59 (2004) 5897–5908.
[12] J.C. Costa, M.M. Alves, E.C. Ferreira, Principal component analysis and quantitative image analysis to predict effects of toxics in anaerobic granular sludge, Bioresour. Technol. 100 (2009) 1180–1185.
[13] P. Teppola, S.P. Mujunen, P. Minkkinen, Adaptive Fuzzy C-Means clustering in process monitoring, Chemom. Intell. Lab. Syst. 45 (1999) 23.