Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Soft Sensor for Determining Municipal Wastewater BOD and COD from Excitation-Emission Matrix Coupled with Partial Least Square Regression

Raed AlJowder[1], Mohamed Bin Shams[2], Ibtisam Mohammad[3], Rashid Alhiddi[4], Marda Buali[5], Bassam Alhamad[6]
[1]*University of Bahrain*
[2]*University of Bahrain*
[1]*University of Bahrain*
[2]*University of Bahrain*
[1]*University of Bahrain*
[2]*University of Bahrain*

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) are critical indicators for assessing the organic pollution load in municipal wastewater. However, conventional BOD (BOD_5) and COD analytical methods are slow, hazardous, or require extensive laboratory preparation, limiting their use for real-time monitoring and rapid decision-making in wastewater treatment plants. This study proposes a soft-sensor approach based on fluorescence Excitation–Emission Matrix (EEM) spectroscopy coupled with multivariate analysis to rapidly estimate BOD and COD in municipal wastewater. Wastewater samples were collected from a treatment plant and pre-processed through filtration, pH adjustment, dilution, and scatter removal. The resulting three-dimensional EEM datasets were analyzed using Parallel Factor Analysis (PARAFAC) to identify dissolved organic matter (DOM) fluorophore components, and Partial Least Squares Regression (PLSR) was applied to develop predictive models for BOD and COD estimation. PARAFAC identified two dominant fluorophore groups: a tryptophan-like, protein-like component (Ex/Em ≈ 275/340 nm) and a humic-like component (Ex/Em ≈ 350/437 nm), consistent with known wastewater DOM profiles. While PARAFAC successfully characterized DOM, it showed limited predictive capability for BOD estimation when used alone. In contrast, PLSR demonstrated strong predictive performance, effectively reducing data dimensionality and correlating EEM spectral variations with measured BOD and COD values. The integration of EEM fluorescence spectroscopy with PLS regression shows high potential for developing rapid, reagent-free, and real-time soft sensors for wastewater monitoring, enabling timely operational decisions in wastewater treatment plants.

Keywords: Excitation–Emission Matrix (EEM); Fluorescence Spectroscopy; Partial Least Squares Regression (PLSR); PARAFAC; Soft Sensor; Biochemical Oxygen Demand (BOD); Chemical Oxygen Demand (COD); Dissolved Organic Matter (DOM); Wastewater Monitoring; Data-Driven Modeling; Multivariate Analysis; Real-Time Sensing.

INTRODUCTION

Modern urban society's rapid changes and developments have increased municipal wastewater pollutants loading. This increase has caused a growing need for continuous environmental monitoring to assess compliance with environmental agencies' standards and identify areas with continuing problems so that they can be rectified swiftly and effectively. Population lifestyles and temporal patterns are factors affecting the municipal sewage water pollution loading (Buerge et al., 2006; Hur et al., 2010; Willems, 2008; L. Yang et al., 2015). Because of the considerable human influence of treated and untreated discharges to surface water, wastewater quality assessment is vital for environmental monitoring (Cahoon & Mallin, 2013; Suthar et al., 2010). Municipal wastewater dissolved organic matter (DOM) plays a crucial role in surface water pollution loading. Its composition and concentration, directly and indirectly, affect the aquatic environment. DOM is complex and varied mixture of aromatic and aliphatic hydrocarbons with functional groups such as amides, carboxyl, hydroxyl, and ketones and others in smaller amounts

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

(Nebbioso & Piccolo, 2013). Thus, it is necessary to develop a sensitive, rapid, and real-time monitoring technique for tracking the dissolved organic matter concentrations and ensure the reliability of wastewater treatment performance. Biochemical oxygen demand (BOD), chemical oxygen demand (COD) and total organic carbon (TOC) can be used indirectly to quantify DOM (Bourgeois et al., 2001).

The objective of this study is to develop and validate a rapid soft-sensor model for estimating biochemical oxygen demand (BOD) and chemical oxygen demand (COD) in municipal wastewater using fluorescence Excitation–Emission Matrix (EEM) spectroscopy coupled with multivariate data analysis techniques. The study focuses on identifying dissolved organic matter (DOM) fluorophores present in wastewater through Parallel Factor Analysis (PARAFAC), and then using Partial Least Squares Regression (PLSR) to correlate the spectral signatures with laboratory-measured BOD and COD values. By converting high-dimensional fluorescence data into predictive models, the aim is to demonstrate that EEM-based soft sensing can effectively replace or supplement traditional BOD/COD laboratory tests, offering a faster, safer, and continuous monitoring alternative.

The significance of the study would be the accurate and timely measurement of BOD and COD, which is essential for wastewater treatment plants to monitor organic pollution loads and ensure compliance with discharge regulations. However, current standard analytical methods present significant limitations: the BOD_5 test requires a five-day incubation period, making it unsuitable for urgent operational decisions, while the COD test involves hazardous reagents such as mercury sulfate and dichromate, posing environmental and health risks. These challenges restrict real-time monitoring and hinder proactive response to sudden fluctuations in wastewater quality. The significance of this work lies in demonstrating that fluorescence EEM spectroscopy, combined with data-driven modeling, provides an immediate, reagent-free alternative capable of delivering near real-time BOD and COD estimates. The proposed soft sensor enables faster decision-making, improves process control efficiency, and supports a more sustainable and automated wastewater monitoring approach.

This paper is organized into four main sections. Section 1 presents the background and motivation for the study, highlighting the limitations of existing BOD and COD analysis methods and introducing fluorescence spectroscopy as a potential solution. Section 2 outlines the experimental methodology, including sample preparation procedures, fluorescence EEM data collection, pre-processing steps, and the application of PARAFAC and PLSR techniques. Section 3 discusses the results, including identification of DOM components through PARAFAC, development of the predictive soft-sensor model using PLSR, and evaluation of the model's performance. Finally, Section 4 summarizes the conclusions and emphasizes the potential of using EEM-based soft sensing as a rapid and reliable approach for continuous wastewater quality monitoring.

LITERATURE REVIEW

Monitoring the dissolved oxygen levels in treated and untreated water discharges is a large-scale approach to monitoring environmental quality. Dissolved oxygen is essential for aquatic life and is an indicator of water quality. Monitoring the levels of dissolved oxygen in water bodies can help identify areas where the levels are below acceptable levels due to excessive pollution from treated and untreated water discharges. This large-scale approach to monitoring helps to ensure that water bodies are adequately protected and that environmental hazards are identified promptly. By providing data on dissolved oxygen levels in water bodies over an extended period, this approach can aid in identifying trends and potential sources of pollution, which can help policymakers, managers, and the public at large to focus their attention on the most pressing issues related to environmental quality.

Dissolved organic matter (DOM) in municipal wastewater is a significant contributor to surface water pollution loading, with its composition and concentration having a direct and indirect impact on the aquatic environment. DOM is a complex and diverse mixture of organic compounds, including aromatic and aliphatic hydrocarbons, and various functional groups such as amides, carboxyl, hydroxyl, and ketones. Smaller amounts of other functional groups may also be present (Nebbioso & Piccolo, 2013).

Dissolved organic matter (DOM) is closely linked to other important water quality parameters, including biochemical oxygen demand (BOD), chemical oxygen demand (COD), and total organic carbon (TOC). BOD is a measure of the amount of dissolved oxygen that is required by microorganisms to decompose organic matter in water, including DOM. When BOD levels are high, it indicates a higher concentration of organic matter in the water and a

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

correspondingly lower level of dissolved oxygen. COD, on the other hand, measures the amount of oxygen required to oxidize organic and inorganic matter in water. Since DOM is composed of organic matter, it contributes to COD values. Finally, TOC is a measure of the amount of carbon that is present in organic compounds in water, including DOM. By measuring TOC, it is possible to estimate the amount of organic matter present in water, which is useful in evaluating the potential for water pollution and assessing the effectiveness of treatment processes. Overall, the relationship between DOM and these other water quality parameters highlights the importance of monitoring and managing organic matter in water to maintain healthy aquatic ecosystems and ensure that water resources are safe and suitable for human use.

Thus, it is necessary to develop a sensitive, rapid, and real-time monitoring technique for tracking the dissolved organic matter concentrations and ensure the reliability of wastewater treatment performance. Biochemical oxygen demand (BOD), chemical oxygen demand (COD) and total organic carbon (TOC) can be used indirectly to quantify DOM (Bourgeois et al., 2001).

COD is defined as a measure of the oxygen equivalent of the organic matter content of a sample susceptible to oxidation by strong chemical oxidants (Potassium permanganate [KMnO4], or potassium dichromate [K2Cr2O7] (Li et al., 2018). The standard COD determination method is commonly employed for environmental monitoring, drinking water treatment, and wastewater treatment, However, this method has some drawbacks, such as low detection sensitivity, high consumption of expansive agents (AgSO4), and highly corrosive, hazardous reagents (HgSO4 and Cr2O72-) (Zhang, 2019). The European Chemical Agency has classed potassium dichromate-one of the agents used in the slandered COD test- as carcinogenic, mutagenic, and reproductively harmful.

BOD is defined as the amount of oxygen, divided by the volume of the system, taken up through the respiratory activity of microorganisms growing on the organic compounds present in the sample when incubated at a specified temperature (usually 20 °C) for a fixed period (usually 5 days, BOD5) (Jouanneau et al., 2014). The BOD5 Standard test involves placing possibly contaminated water samples in specialized bottles, aerating them, and adding a microbial population. The bottles are then hermetically sealed and incubated at 20 °C in a dark environment. The dissolved residual oxygen is measured for all tested samples after an incubation time of 5 days to estimate the BOD (International Organization for Standardization, 2019). Although the BOD5 test do not involve reagents that are expensive of dangers as the case with the COD, yet the test is done over a course of five days which is a big disadvantage for municipal wastewater treatment plants (WWTPs) as it will be too late if a sudden increase in BOD was detected five days later. All those disadvantages associated with standard COD and BOD tests have motivated research for a safer, more relabel and continuous methods of detection. Several authors evaluated the development of local and global prediction models to predict the wastewater influent biochemical oxygen demand through supervised learning (Mekaoussi et al., 2023; Mongioví et al., 2024; Qambar & Al Khalidy, 2023; Qambar & Khalidy, 2022; L. Yang et al., 2026; T. Yang et al., 2026).

In recent years, fluorescence spectroscopy has become widely used as a predictable and reliable technique for monitoring BOD and COD (Carstea et al., 2016; Peng et al., 2025). Fluorescence spectroscopy is the emission of photons by a population of molecules to return from an excited state to a ground state. The excitation of molecules occurs because of the absorption of photons in the form of energy. The emission will have lower energy than the exciting light (Albani, 2007). Fluorescence spectroscopy has several advantages, such as being a quick and cost-effective method that does not require reagents or extensive sample preparation. It also has high sensitivity and can detect low concentrations of dissolved organic matter (DOM), making it useful for identifying fluorescent components of DOM in dilute solutions from multiple samples. Generating excitation-emission matrices (EEMs) by recording fluorescence intensity from excitations across a range of wavelengths can provide valuable information. Each EEM is a matrix, and combining EEMs from various samples produces a three-dimensional array (Gong et al., 2024; Sakr et al., 2025). By examining the EEM of a specific molecular configuration, the composition of organic substances can be determined, with the peak intensity detected in the sample being proportional to the concentration of the relevant fluorophore. These matrices contain a vast amount of data that can be utilized in various studies (Baker, 2001; Carstea et al., 2016; Chen et al., 2003; Hao et al., 2012; Kim & Kim, 2020; Murphy et al., 2011; L. Yang et al., 2015).

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Recent work has shown how multivariate data analysis methods can be applied to the study of EEMs in wastewater analysis. One of those methods is parallel factor analysis (PARAFAC).

Parallel factor analysis (PARAFAC) is a mathematical technique used to analyse multi-way data sets, such as those generated by fluorescence spectroscopy. It decomposes the data into a set of three-dimensional matrices, which represent the contributions of each component to the data. Each matrix is composed of scores, which represent the intensity of each component in each sample, and loadings, which represent the spectral characteristics of each component. PARAFAC assumes that the data can be represented as a sum of the outer products of three vectors, corresponding to the scores and loadings for each mode. This allows for the identification of the underlying factors that contribute to the data, even when these factors are not directly observable. PARAFAC has become a popular tool for analysing fluorescence excitation-emission matrices (EEMs) of dissolved organic matter, as it can effectively identify the different fluorescent components of DOM and track their changes over time or across different environmental conditions.

In wastewater analysis, PARAFAC is a powerful method for efficiently extracting information from EEMs by identifying individual components and following their behaviour in diverse contexts. This information extraction makes the identification and quantification of distinct underlying components easier (Murphy et al., 2011, 2013; Pokrovsky et al., 2018; Stedmon et al., 2003; Stedmon & Bro, 2008). The suitability of partial least squares (PLS) as an alternative method for estimating BOD and COD, along with PARAFAC, is examined in this study. section 1 of the paper highlights the necessity for a fast and reliable technique to predict BOD and COD and outlines the limitations of the traditional BOD5 method. The collection and pre-processing of raw EEM data are discussed in section 2, followed by the application of PARAFAC or PLS techniques. The paper then presents the results in section 3, comparing the effectiveness of PARAFAC and PLS in predicting BOD and COD. Furthermore, section 3 includes additional PLS model validation plots. Finally, section 4 summarizes the primary findings of the study.

DESIGN AND EXECUTION

Sample Collection and Preparation

The frozen wastewater samples obtained from the treatment plant were thawed to 20°C using a water bath heater before conducting fluorescence spectroscopy. To eliminate any solid particles present, the samples were filtered through 0.45 µm filter paper. To maintain the pH of the samples within the range of 6-8, 0.1 M HCl was used. Furthermore, a dilution ratio of 1:10 was employed, where 2 mL of the sample was mixed with 20 mL of water to acquire EEM data through fluorescence spectroscopy. This step was performed to avoid the impact of inner-filter effects on the samples.

Fluorescence Measurements

The PerkinElmer LS 55 Fluorescence Spectrometer was used to obtain the raw EEM data of the pre-treated wastewater samples. Connected to a PC running the FL WinLab software, the spectrometer had an excitation range of 240-425 nm and an emission wavelength range of 280-510 nm. The software recorded fluorescence intensities for all emission wavelengths for a single excitation value, with the excitation increment determining the next excitation value. However, the useful fluorescence data was difficult to detect due to the presence of very intense Rayleigh scatters, as depicted in **Figure 1**. Rayleigh scattering refers to the dispersion of light in all directions when it interacts with small particles in a medium, such as air or water. As a result, its intensity is directly proportional to the fourth power of the light frequency. In the context of fluorescence spectroscopy, the high intensity of Rayleigh scattering can interfere with the detection of useful fluorescence data, making pre-processing of the EEM necessary to filter out these scattering effects.

Data Pre-Processing

Pre-processing of raw EEM data is necessary before analysis, which involves removing Raman and Rayleigh scatters. The elimination of these scatters is crucial because they can obstruct the fluorescence signals and make them difficult to detect. As shown in **Figure 2**, the raw EEM data has intense Rayleigh scatters that have obscured the fluorescence signals while **Figure 3** displays the raw EEM data of all the samples that were imported into R. It is evident that the

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

fluorescence peaks are concealed by the Rayleigh scatters in almost all the samples, thus necessitating their elimination.

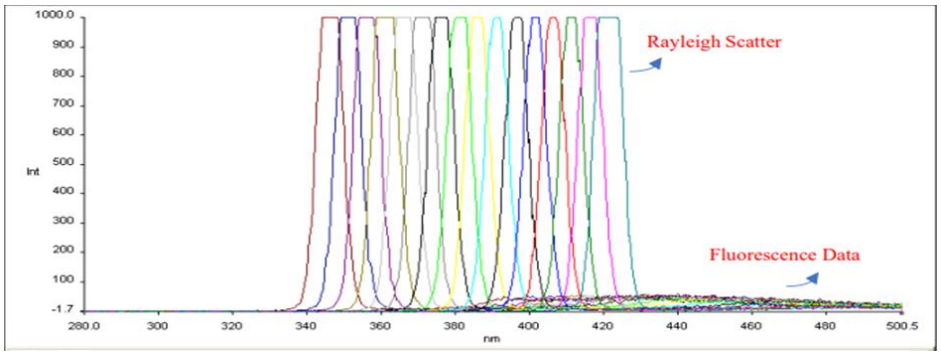

Figure 1. Comparison between useful fluorescence data and Rayleigh scattering

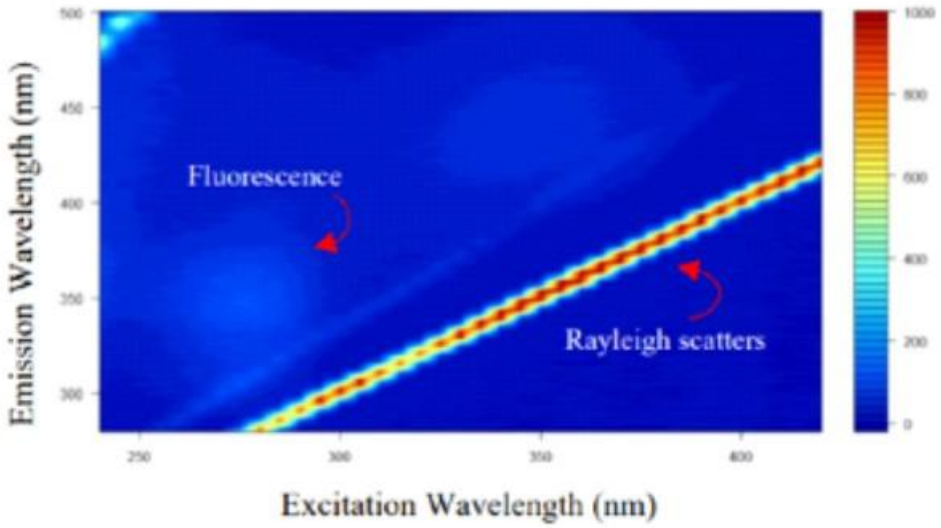

Figure 2. Fluorescence hidden by Rayleigh scatters

There are various pre-processing techniques available to address Raman and Rayleigh scatters in spectral data. One common method is to utilize mathematical algorithms that can subtract the scattered signal from the data. For instance, a blank spectrum that is obtained from the same medium but without the analyte can be subtracted to eliminate the Raman signal. Similarly, a Rayleigh scatter spectrum that can be acquired from a solution without fluorophores can be subtracted to remove the Rayleigh signal. Another technique involves interpolating over the affected data points or removing and replacing them with missing data. The latter method was utilized in the study being discussed.

R Programming language was used by the authors of the paper for conducting all data analyses. The EEM package in R is a powerful tool for the analysis of Excitation-Emission Matrix (EEM) fluorescence data, providing several functions for preprocessing, visualizing, and analyzing EEM data. The package includes various algorithms for removing Raman and Rayleigh scatter, smoothing, and normalizing data. The EEM package is widely used in environmental monitoring, water quality assessment, and other areas where fluorescence spectroscopy is employed.

The package "staRdom" was used to eliminate these scatters, and **Figure 4** shows the excitation-emission matrices of some of the samples after the removal of scatters. The fluorescence peaks that were previously hidden by the scatters are now visible. To prepare the data for further analysis, the 3-D EEMs were unfolded into 2-D data and normalized using the "EEM" package before applying PARAFAC or PLS analysis.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Normalization is an important step in accounting for differences in total fluorescence intensity between samples that can affect data interpretation. The EEM package includes various normalization methods such as total area normalization, sample-specific normalization, and Probabilistic Quotient Normalization (PQN). In this study, PQN was chosen because it can handle various sources of variation, does not assume data distribution, preserves the relative ratios between features, and is computationally efficient. PQN normalizes data based on the proportion of signal intensity in each sample, making it robust to outliers and other anomalies. PQN is suitable for downstream analyses such as differential expression analysis.

Parallel Factor (PARAFAC) Analysis

Parallel Factor Analysis (PARAFAC) is a widely used technique for decomposing trilinear data arrays, which helps in the identification and quantification of components. It has been shown that PARAFAC is most commonly applied to fluorescence excitation-emission matrices. The model created through PARAFAC analysis determines the number of components or fluorophores present in the samples, and provides quantitative information about them by decomposing the data into a set of trilinear terms and a residual term, as demonstrated in Equation 1.

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{if} c_{kf} + \varepsilon_{ijk} \quad (1)$$

In case of an EEM, x_ijk is the intensity of fluorescence for sample i at emission wavelength j and excitation wavelength k. Each f corresponds to a PARAFAC component. The loading a_if is directly proportional to the concentration of the component, b_if is an estimate of the emission spectrum of the component and c_kf is the estimate of the excitation spectrum of the component. If the obtained PARAFAC model is validated, these three loadings combined have a direct chemical interpretation of the sample components.

The process of performing PARAFAC analysis on fluorescence excitation-emission matrices involves several steps. Firstly, the raw EEM data files collected through fluorescence spectroscopy are imported. Secondly, the raw data is pre-processed, including correcting for "inner-filter effects" and removing Raman and Rayleigh scatters. Next, a PARAFAC model is created to best represent the samples and predict the number of components (fluorophores) in the samples. To validate the model, the number of components is confirmed via "split-half analysis". Finally, the results obtained from PARAFAC, such as fluorescence peaks and components' loadings, can be interpreted and combined with multiple linear regression to predict the samples' BOD (Murphy et al., 2013)

PARAFAC is a useful technique but has limitations. The algorithm is sensitive to initial conditions and can lead to errors if the parameters are not estimated correctly. Interpretation of factors can be difficult with complex data containing noise or other sources of variation, and the optimal number of factors may be difficult to determine, and the process of manually selecting and identifying peaks can be time-consuming and subjective. Estimating BOD/COD based on identified peaks may also not be precise.

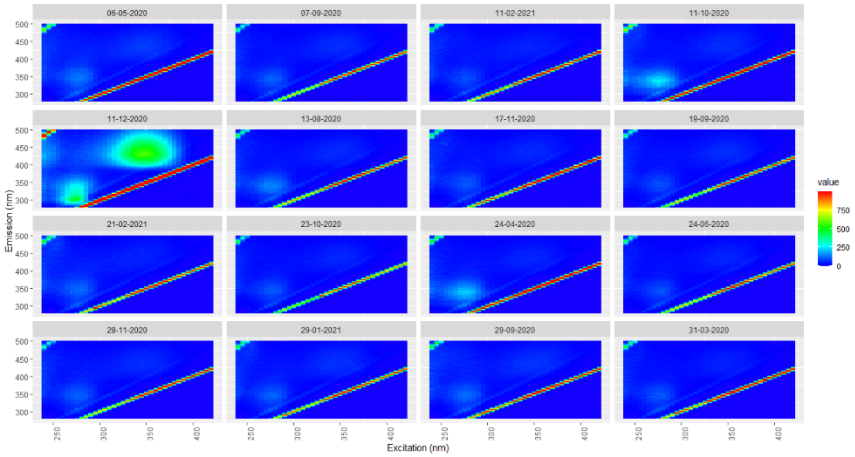

Figure 3. EEMs of raw data

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

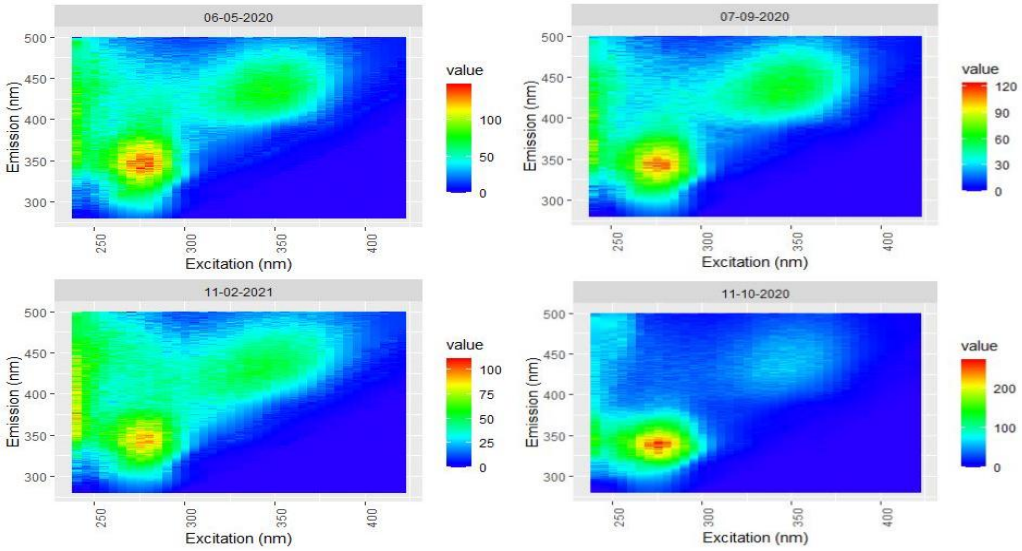

Figure 4. EEMs of samples after removing scatters

Partial Least Squares (PLS) Analysis and Statistical Tests

To address the limitations of PARAFAC analysis, partial least squares (PLS) was utilized to estimate the BOD of the wastewater samples. PLS is a modelling technique used for creating predictive models when there are many highly collinear factors affecting the responses, which may exceed the number of observations. Instead of focusing on understanding the underlying relationship between the factors, PLS aims to predict the responses. One of the key advantages of PLS is the ability to reduce the dimensionality of the data, which is particularly important when working with EEMs that cover a wide range of wavelengths. While multiple factors may affect the response, PLS extracts a few latent factors that explain most of the variation in the response while modelling the responses as well.

The overall process is illustrated in **Figure 5**, which outlines the objective of predicting responses from factors. In PLS analysis, latent variables T and U are extracted from the factors and responses, respectively, also known as X-scores and Y-scores. The X-scores are used to predict the Y-scores, which are then utilized to predict the responses. PLS extracts the X and Y-scores to capture maximum variance in the data and achieve maximum correlation between the predictor factors X and predicted responses Y. However, the extraction of X and Y-scores is dependent on the initial conditions of the model. The PLS algorithm used in this study was Geladi and Kowalski's PLS algorithm (Geladi and Kowalski, 1986).

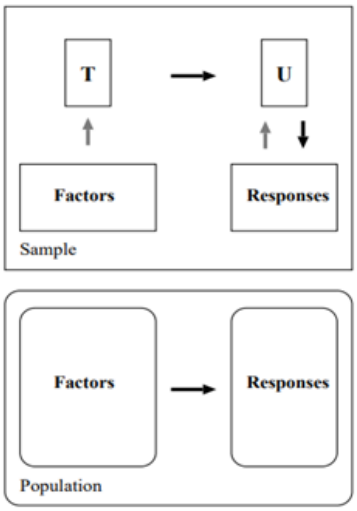

Figure 5. General schematic of PLS

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RESULTS AND DISCUSSIONS

PARAFAC Analysis and Peak-Picking Technique

Two components of dissolved organic matter (DOM) were detected in all 16 wastewater samples using PARAFAC analysis. The EEM contours representing these components are displayed in **Figure 6**.

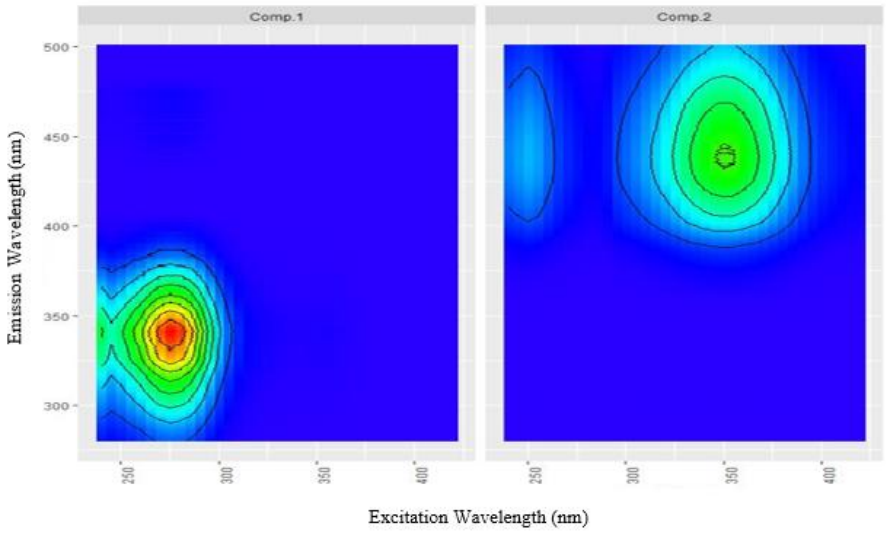

Figure 6. The EEM contours of the two identified components are as follows: Component 1 (figure on the left), which is Tryptophan-like and protein-like, and Component 2 (figure on the right), which is Humic-like..

To identify these components, the excitation-emission peaks of the contours were compared with those commonly found in literature(Coble et al., 2014) (**Table 1**). The components were identified as follows: Component 1, which is Tryptophan-like and protein-like, has an excitation peak at approximately 275 nm and an emission peak at approximately 340 nm. These substances are produced by the breakdown of biological matter and can be used as indicators of microbial activity in water. Component 2, which is Humic-like, has an excitation peak at approximately 350 nm and an emission peak at approximately 437 nm. These compounds are produced by the degradation of organic matter in soil and water and are often associated with the presence of humic acids and fulvic acids. The presence of these compounds in wastewater can have important implications for water treatment and environmental monitoring, as they can contribute to the formation of disinfection by-products and impact the overall quality of water resources.

Table 1: Commonly referenced excitation and emission peaks for aquatic humic substances and dissolved organic matter (DOM) (Coble, 1996)

Excitation Maximum (nm)	Emission Maximum (nm)	Description of Fluorophores
275	305	Tyrosine-like, protein-like
275	340	Tryptophan-like, protein-like
260	400-460	Humic-like
290-310	370-410	Marine humic-like
320-360	420-460	Humic-like
390	509	Soil fulvic acid
455	521	Soil fulvic acid
280	370	Plankton derived

The peak-picking technique was used to estimate BOD by noting the peaks of the two components in all samples. Multiple linear regression was then applied to derive Equation 2 for the estimation of BOD.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$BOD = 129.732 + 0.083F_1 - 0.058F_2 \quad (2)$$

Where F_1 and F_2 represent peak intensities of component 1 and component 2, respectively. As can be seen, the coefficients of peaks of components 1 and 2 are very low indicating a very low dependence of the BOD values on the peak intensities. In addition, the coefficient of multiple determination R^2 value obtained was 0.0594 representing a very bad regression fit.

Validation of the Model

The score loading plot in partial least squares (PLS) analysis is a graphical representation that displays the relationship between the sample scores and the component loadings. The score loading plot is a valuable tool for visualizing the variation in the data and identifying outliers or trends in the data set. In wastewater analysis, the score loading plot can be used to identify samples that deviate significantly from the norm, as well as to identify any correlations or patterns in the data set. By examining the score loading plot, shown in **Figure 7**, the following was inferred, it was discovered that the sample with BOD 119 appeared to be an outlier. Further investigation revealed that in comparison to the other samples, the one with BOD 119 (dated 11-12-2020) had a higher concentration of component 2, which was identified earlier in section 3.1 as humic-like. This was indicated by its EEM, which displayed a very fluorescent peak in the region of component 2. The presence of this higher concentration of humic-like material in the sample may explain why it appeared as an outlier in the score plot.

The elevated levels of humic-like material found in the wastewater sample could potentially be explained by the presence of organic matter, such as decomposing plant and animal material. The Muharraq wastewater plant (MWSP) receives water from two sources: the underground sewage network and tankers that gather wastewater from all over the country and then deposit it into the Muharraq wastewater pit, where it mixes with the sewage water from the network and is then treated. During Bahrain's rainy season in December, the tankers collect rainwater from across the kingdom and pour it into the pit, where it mixes with grey water and is then treated. This collected water may have a higher amount of decomposing plant and animal material, leading to the sudden increase in the concentration of humic-like material. Additionally, the increase could be due to changes in temperature, pH, or sunlight exposure. Further investigation is necessary to determine the source of the increase.

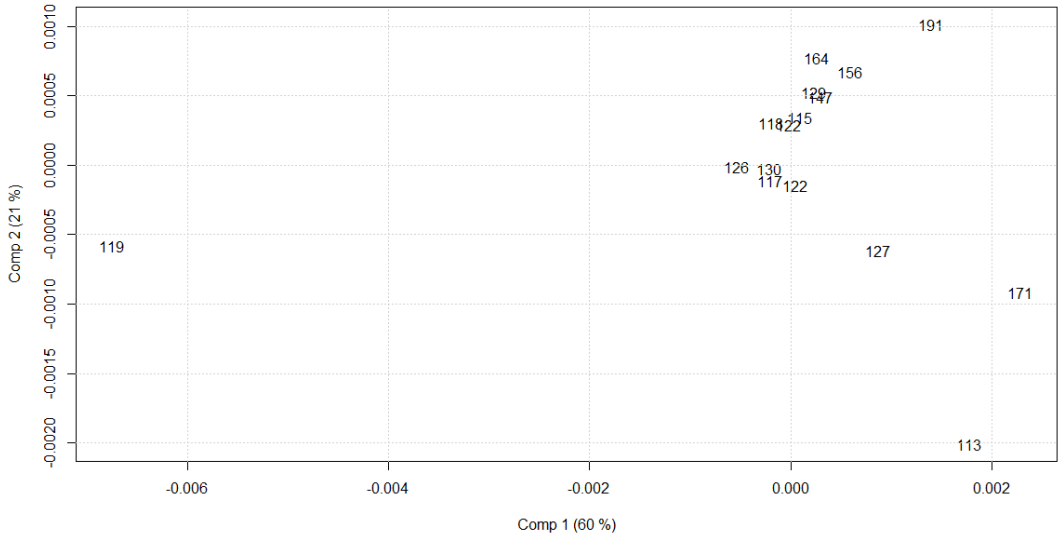

Figure 7. PLS Score plot

The variable importance plot (VIP) is a graphical representation that displays the relative importance of each variable in a multivariate model. In the context of wastewater analysis, the VIP plot is used to identify the most significant excitation-emission pairs in the estimation of BOD and COD. The VIP plot in **Figure 8** displays the 25 most important excitation-emission pairs in the model for estimation. The excitation-emission pair of Ex270-Em302 was identified as the most important in the model, and it closely corresponds to the excitation-emission peak of component 1 in wastewater, which is the tryptophan-like, protein-like component. This component is the primary

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

one found in all of the wastewater samples, except for the outlier sample with BOD 119. The importance of this excitation-emission pair suggests that it plays a crucial role in the estimation of the model and reinforces the significance of component 1 in the wastewater samples. The importance of this discovery lies in the fact that the specific excitation-emission wavelength identified can serve as a qualitative indicator of the increasing loading of wastewater. A higher intensity of this peak could signify a sudden surge in wastewater loading. Hence, this specific excitation-emission peak could potentially be utilized as a quick and easy method to determine wastewater loadings. These findings provide valuable insights into the characteristics of the wastewater samples and may be useful for further studies in this area.

The PLS loading plot is a useful method for displaying the correlation between the predictor variables and the response variable. This plot represents the predictor variables based on their association with each component, while also displaying the response variable. It is a valuable tool for identifying the most significant predictor variables that contribute to the response variable, which is important in many scientific fields. By visualizing the patterns and relationships between the predictor variables and the response variable, researchers can gain insights into the underlying mechanisms driving the system they are studying. The loading plot shown represents components 1 and 2, which together explain 81% of the variance in the data, as shown in **Figure 8**. This high level of explained variance suggests that the loading plot is an effective method for identifying the most significant predictor variables in the dataset.

PLS Analysis

In order to overcome the limitations of the peak-picking technique, partial least squares analysis was conducted on the wastewater sample data to assess its ability to estimate BOD and COD using their EEMs. The scree plot presented in **Figure 9** indicates that a satisfactory low RMSEP is achieved with five components. The model summary reveals that four components account for 89.3% of the variance in the data, and further increasing the number of components does not result in a significant improvement.

The PLS-predicted BOD5 values are compared to the measured values in **Figure 10**. The plot indicates that the model has achieved near-perfect prediction accuracy with the use of five components. **Table 2** shows that the mean squared error (MSE) between the measured and predicted BOD values was calculated to be 0.0554.

In the prediction of COD, it can be observed in **Figure 11** that the use of five components results in a low RMSEP that accounts for 89.15% of the variability in the dataset. **Figure 10** presents the comparison of predicted versus measured values of COD, and similarly to BOD, PLS proved to be a highly effective estimation technique. The MSE of the predicted values is reported in **Table 3**, which was calculated to be 0.0976. Prediction plot of COD in shown in **Figure 12.**

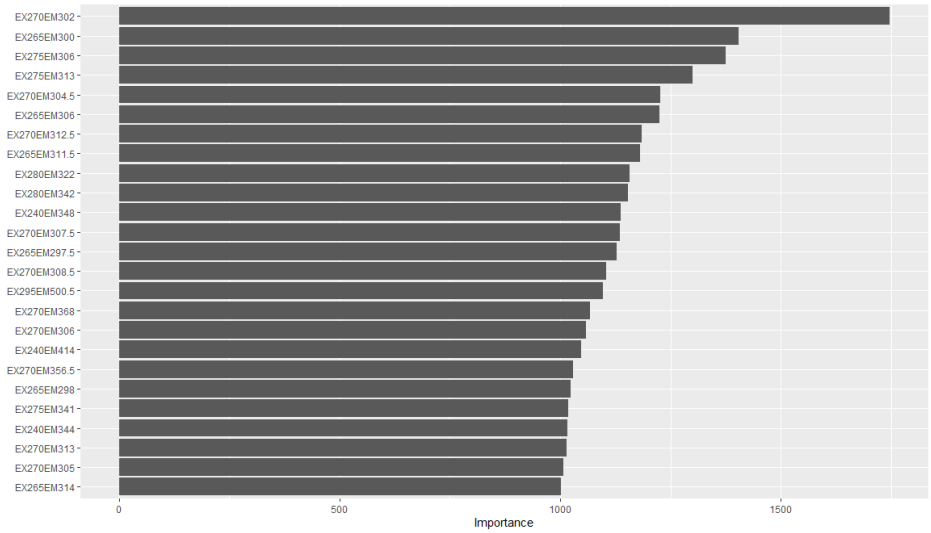

Figure 8. Variable Importance Plot (VIP)

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

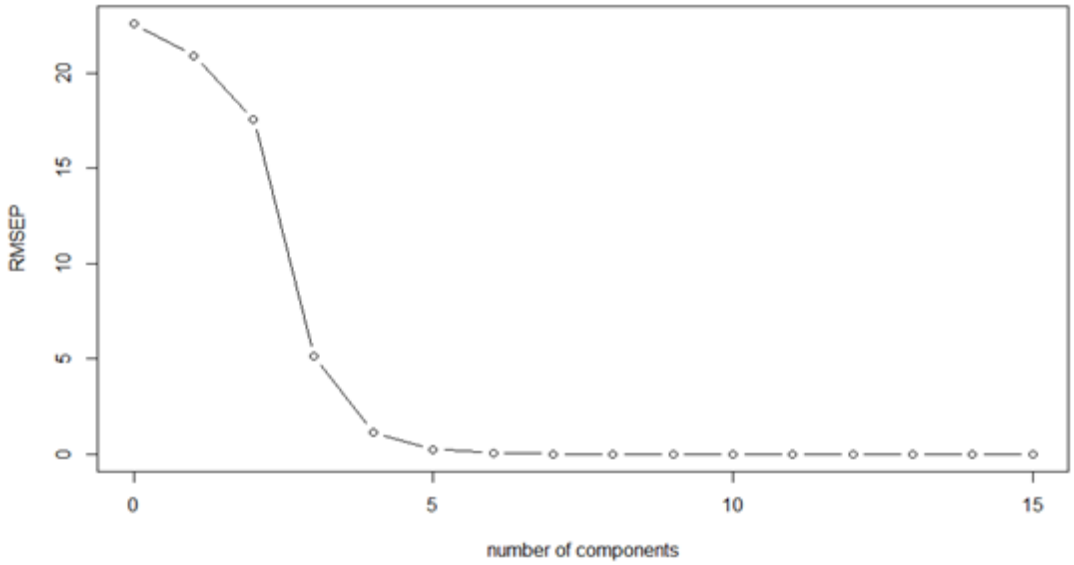

Figure 9. Scree plot of PLS model for predicting BOD

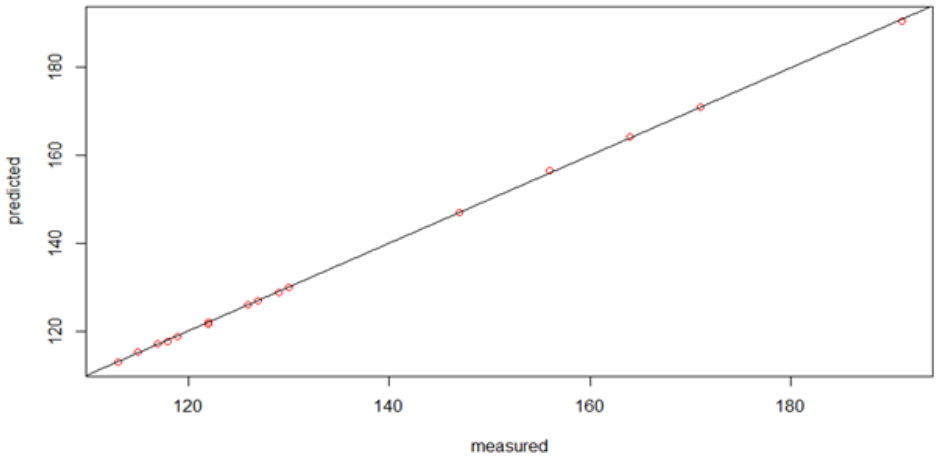

Figure 10. Prediction plot of BOD

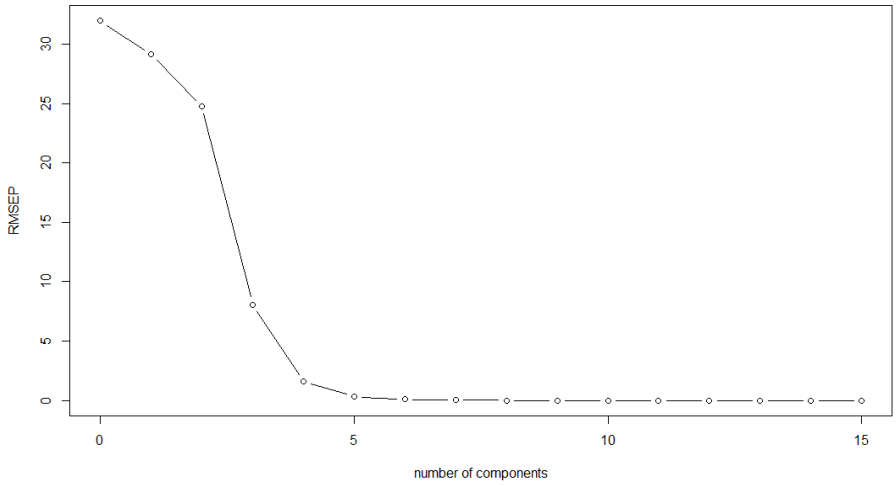

Figure 11. Scree plot of PLS model for predicting COD

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

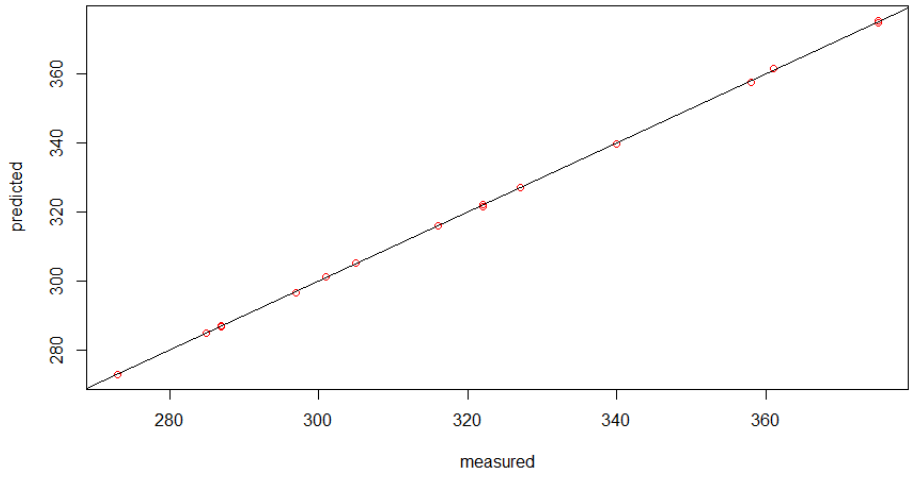

Figure 12. Prediction plot of COD

Table 2. Predicted vs. measured BOD values

Sample	Measured BOD5	Predicted BOD5	Sample	Measured BOD5	Predicted BOD5
1	164	164.2566	**9**	191	190.6549
2	130	130.0637	**10**	122	122.1038
3	147	147.0547	**11**	171	170.9447
4	113	112.9301	**12**	117	117.1890
5	119	118.9223	**13**	115	115.3025
6	127	126.9839	**14**	118	117.6886
7	129	128.7553	**15**	122	121.5874
8	126	126.1132	**16**	156	156.4493

Table 3. Predicted vs. measured COD values

Sample	Measured COD	Predicted COD	Sample	Measured COD	Predicted COD
1	361	361.5184	**9**	358	357.4010
2	305	305.1715	**10**	287	287.3032
3	322	322.0755	**11**	375	374.7954
4	273	272.9854	**12**	301	301.2355
5	287	286.8613	**13**	316	316.0587
6	327	327.0814	**14**	322	321.5271
7	340	339.7916	**15**	297	296.5688
8	285	285.1855	**16**	375	375.4398

The results demonstrate a major advancement toward real-time wastewater quality monitoring. The combination of fluorescence EEM spectroscopy with multivariate soft-sensor modeling effectively replaces the slow BOD_5 method (5-day incubation) and the hazardous COD method (dichromate and mercury reagents). The ability of PLS to extract latent factors directly from the unfolded EEM matrix provides a superior predictive capability compared to PARAFAC, which required manual peak selection and resulted in poor correlation with BOD values. PARAFAC successfully identified distinct dissolved organic matter (DOM) components-tryptophan-like (Ex/Em ≈ 275/340 nm) and humic-like (Ex/Em ≈ 350/437 nm), confirming that wastewater DOM originates from microbial activity and decomposed organic matter. However, when these fluorophore peaks were used alone to estimate BOD, the model

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

performance was weak ($R^2 \approx 0.06$). The analysis reveals that although fluorophore identification is valuable for characterizing wastewater composition, DOM peak intensity does not explain BOD/COD concentration linearly.

Partial Least Squares (PLS) overcame this limitation by using all fluorescence information simultaneously, not only individual peaks. By reducing thousands of spectral variables into 4–5 latent components that captured 89.3% of spectral variance, PLS achieved prediction errors of MSE = 0.0554 (BOD) and 0.0976 (COD) with near-perfect alignment between predicted and measured values. From a process control perspective, this means that a single EEM scan can reliably estimate regulatory discharge-critical parameters that previously required days of lab work.

The VIP plot and loading analysis further explained the predictive strength of PLS: the excitation-emission pair Ex270/Em302 was identified as the most influential predictor for BOD/COD. This wavelength pair corresponds to protein-like tryptophan fluorophores, which are known markers of fresh biological contamination. Therefore, tryptophan fluorescence acts as an early-warning biomarker of organic shock loading, enabling plant operators to detect upsets immediately rather than discovering them five days later.

Another important insight emerged from the PLS score plot. The wastewater sample outlier appeared due to a significantly elevated humic-like component concentration. This aligns with the operational reality of the treatment plant: during rain events, tanker trucks deliver surface runoff and decaying natural matter, increasing humic content in the influent. This confirms that the soft sensor does not merely fit data mathematically—it is chemically interpretable and capable of detecting real process abnormalities. Thus, PARAFAC aids process understanding, while PLS delivers operational predictive power (See **Table 4**).

Table 4. Value and Limitations between PARAFAC and PLS

Aspect	PARAFAC	PLS
Key value	Identifies DOM chemistry (what is present)	Predicts BOD/COD (how much is present)
Limitations	Weak regression; peak selection subjective	Captures all spectral variance; objective learning
Operational benefit	Understanding source of contamination	Enables real-time control decisions

CONCLUSIONS

In this study, the effectiveness of PARAFAC and PLS in estimating BOD and COD in wastewater samples was explored. PLS was found to be a more reliable technique than PARAFAC due to its ability to overcome the shortcomings associated with the latter method. In PARAFAC, the EEM needs to be split into components that must be identified before BOD is estimated, making the process time-consuming and challenging. In contrast, PLS can estimate BOD and COD using only five components, resulting in very low MSE values of 0.0554 and 0.0976, respectively.

The analysis of wastewater samples can be significantly enhanced by using powerful tools such as PARAFAC and PLS. PARAFAC can identify various components of dissolved organic matter (DOM), including Tryptophan-like, protein-like, and Humic-like components, which are essential in water treatment and environmental monitoring. Meanwhile, PLS can overcome the limitations of peak-picking techniques and achieve near-perfect prediction accuracy by using only a few components. Notably, the model summary indicates that only four components account for 89.3% of the variance in the data, underscoring the robustness of the PLS method.

However, despite the clear benefits of using these methods, their complexity can pose a challenge without appropriate guidance. It is crucial to have mentoring and training to ensure the methods' full potential is realized, and their results are accurately interpreted.

Also, the use of the excitation-emission pair of Ex270-Em302 in this study highlights the importance of selecting appropriate excitation-emission pairs to obtain accurate results. Careful selection of these pairs can lead to a better understanding of the various components present in wastewater samples and improve the accuracy of the analysis.

The scientific contribution confirms that municipal wastewater DOM is dominated by tryptophan-like and humic-like fluorescent components. It demonstrates that tryptophan fluorescence (Ex270/Em302) can serve as a rapid indicator of organic shock loading. In addition, it shows that PLS can model the nonlinear relationship between fluorescence signatures and oxygen demand.

Finally, the potential of using PLS to develop an online soft sensor for real-time monitoring of BOD and COD values is currently under investigation. This would allow for prompt detection and action to be taken when values exceed safe limits, ultimately contributing to the effective management of wastewater treatment processes.

The soft sensor developed in this study proves that fluorescence EEM spectroscopy combined with PLS modelling can accurately, rapidly, and safely estimate BOD and COD in municipal wastewater. By achieving near-perfect predictive accuracy using only five latent components, the proposed methodology presents a major step toward real-time wastewater quality surveillance, reduced laboratory dependency, and fully automated treatment plant control. This soft sensor shifts wastewater quality monitoring from reactive measurement to proactive control, enabling real-time environmental protection.

CONFLICT OF INTEREST

There are no conflicts of interest.

REFERENCES

[1] Albani, J. R. (2007). *Principles and Applications of Fluorescence Spectroscopy*. Blackwell Science.
[2] Baker, A. (2001). Fluorescence Excitation–Emission Matrix Characterization of Some Sewage-Impacted Rivers. *Environmental Science & Technology*, *35*(5), 948–953. https://doi.org/10.1021/es000177t
[3] Bourgeois, W., Burgess, J. E., & Stuetz, R. M. (2001). On-line Monitoring of Wastewater Quality: A Review. *Journal of Chemical Technology & Biotechnology*, *76*(4), 337–348. https://doi.org/10.1002/jctb.393
[4] Buerge, I. J., Poiger, T., Müller, M. D., & Buser, H.-R. (2006). Combined Sewer Overflows to Surface Waters Detected by the Anthropogenic Marker Caffeine. *Environmental Science & Technology*, *40*(13), 4096–4102. https://doi.org/10.1021/es052553l
[5] Cahoon, L. B., & Mallin, M. A. (2013). Water Quality Monitoring and Environmental Risk Assessment in a Developing Coastal Region, Southeastern North Carolina. In S. Ahuja (Ed.), *Monitoring Water Quality* (pp. 149–169). Elsevier. https://doi.org/10.1016/B978-0-444-59395-5.00006-6
[6] Carstea, E. M., Bridgeman, J., Baker, A., & Reynolds, D. M. (2016). Fluorescence Spectroscopy for Wastewater Monitoring: A Review. *Water Research*, *95*, 205–219. https://doi.org/10.1016/j.watres.2016.03.021
[7] Chen, W., Westerhoff, P., Leenheer, J. A., & Booksh, K. (2003). Fluorescence Excitation–Emission Matrix Regional Integration to Quantify Spectra for Dissolved Organic Matter. *Environmental Science & Technology*, *37*(24), 5701–5710. https://doi.org/10.1021/es034354c
[8] Coble, P. G. (1996). Characterization of marine and terrestrial DOM in seawater using excitation-emission matrix spectroscopy. *Marine Chemistry*, *51*(4), 325–346. https://doi.org/https://doi.org/10.1016/0304-4203(95)00062-3
[9] Gong, B., Chen, W., Qian, C., & Yu, H.-Q. (2024). Evaluating excitation-emission matrix for characterization of dissolved organic matter in natural and engineered water systems: Unlocking submerged secrets. *TrAC Trends in Analytical Chemistry*, *181*, 118045. https://doi.org/https://doi.org/10.1016/j.trac.2024.118045
[10] Hao, R., Ren, H., Li, J., Ma, Z., Wan, H., Zheng, X., & Liang, D. (2012). Use of three-dimensional excitation and emission matrix fluorescence spectroscopy for predicting the disinfection by-product formation potential of reclaimed water. *Water Research*, *46*(17), 5765–5776. https://doi.org/10.1016/j.watres.2012.08.003
[11] Hur, J., Lee, B.-M., Lee, T.-H., & Park, D.-H. (2010). Estimation of biological oxygen demand and chemical oxygen demand for combined sewer systems using synchronous fluorescence spectra. *Sensors*, *10*(4), 2460–2471. https://doi.org/10.3390/s100402460
[12] International Organization for Standardization. (2019). *ISO 5815-1:2019 — Water quality — Determination of biochemical oxygen demand after n days (BODn) — Part 1: Dilution and seeding method with allylthiourea addition*. https://www.iso.org/standard/69058.html

[13] Jouanneau, S., Recoules, L., Durand, M. J., Boukabache, A., Picot, V., Primault, Y., & Théron, M. (2014). Methods for assessing biochemical oxygen demand (BOD): A review. *Water Research*, *49*, 62–82. https://doi.org/10.1016/j.watres.2013.10.066

[14] Kim, J., & Kim, T.-H. (2020). Distribution of humic fluorescent dissolved organic matter in Lake Shihwa: The role of the redox condition. *Estuaries and Coasts*, *43*(3), 578–588. https://doi.org/10.1007/s12237-019-00635-z

[15] Li, J., Luo, G., He, L., Xu, J., & Lyu, J. (2018). Analytical approaches for determining chemical oxygen demand in water bodies: A review. *Critical Reviews in Analytical Chemistry*, *48*(1), 47–65. https://doi.org/10.1080/10408347.2017.1322552

[16] Mekaoussi, H., Heddam, S., Bouslimanni, N., Kim, S., & Zounemat-Kermani, M. (2023). Predicting biochemical oxygen demand in wastewater treatment plant using advance extreme learning machine optimized by Bat algorithm. *Heliyon*, *9*(11), e21351. https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e21351

[17] Mongioví, C., Morin-Crini, N., Lacalamita, D., & Crini, G. (2024). Impact of carbon technology on chemical and biochemical oxygen demand values as water quality indicators of physico-chemical treated laundry effluents. *Case Studies in Chemical and Environmental Engineering*, *10*, 101012. https://doi.org/https://doi.org/10.1016/j.cscee.2024.101012

[18] Murphy, K. R., Hambly, A., Singh, S., Henderson, R. K., Baker, A., Stuetz, R. M., & Khan, S. J. (2011). Organic matter fluorescence in municipal water recycling schemes: Toward a unified PARAFAC model. *Environmental Science & Technology*, *45*(7), 2909–2916. https://doi.org/10.1021/es103015e

[19] Murphy, K. R., Stedmon, C. A., Graeber, D., & Bro, R. (2013). Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Analytical Methods*, *5*(23), 6557–6566. https://doi.org/10.1039/c3ay41160e

[20] Nebbioso, A., & Piccolo, A. (2013). Molecular characterization of dissolved organic matter (DOM): A critical review. *Analytical and Bioanalytical Chemistry*, *405*(1), 109–124. https://doi.org/10.1007/s00216-012-6363-2

[21] Peng, N., Sun, Q., Zhao, Y., Ding, Y., Shi, J., Ping, Q., Wang, L., & Li, Y. (2025). Molecular signature evolution of dissolved organic matter in wastewater during far-ultraviolet/peracetic acid disinfection: Integrated characterization and machine learning. *Water Research*, 124880. https://doi.org/https://doi.org/10.1016/j.watres.2025.124880

[22] Pokrovsky, O. S., Bueno, M., Manasypov, R. M., Shirokova, L. S., Karlsson, J., & Amouroux, D. (2018). Dissolved organic matter controls seasonal and spatial selenium concentration variability in thaw lakes across a permafrost gradient. *Environmental Science & Technology*, *52*(18), 10254–10262. https://doi.org/10.1021/acs.est.8b02008

[23] Qambar, A. S., & Al Khalidy, M. M. M. (2023). Development of local and global wastewater biochemical oxygen demand real-time prediction models using supervised machine learning algorithms. *Engineering Applications of Artificial Intelligence*, *118*, 105709. https://doi.org/https://doi.org/10.1016/j.engappai.2022.105709

[24] Qambar, A. S., & Khalidy, M. M. Al. (2022). Prediction of municipal wastewater biochemical oxygen demand using machine learning techniques: A sustainable approach. *Process Safety and Environmental Protection*, *168*, 833–845. https://doi.org/https://doi.org/10.1016/j.psep.2022.10.033

[25] Sakr, M. E. M., Kandel, H. M., Abou Kana, M. T. H., Elwahy, A. H. M., Negm, N. A., & Khalil, A. A. A. (2025). Fluorescence and Photostability Studies of a Xanthenone-Based Dye via CdS Quantum Dot Complexation. *Physica B: Condensed Matter*, 417996. https://doi.org/https://doi.org/10.1016/j.physb.2025.417996

[26] Stedmon, C. A., & Bro, R. (2008). Characterizing dissolved organic matter fluorescence with parallel factor analysis: A tutorial. *Limnology and Oceanography: Methods*, *6*(11), 572–579. https://doi.org/10.4319/lom.2008.6.572

[27] Stedmon, C. A., Markager, S., & Bro, R. (2003). Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. *Marine Chemistry*, *82*(3–4), 239–254. https://doi.org/10.1016/S0304-4203(03)00072-0

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[28] Suthar, S., Sharma, J., Chabukdhara, M., & Nema, A. K. (2010). Water quality assessment of river Hindon at Ghaziabad, India: Impact of industrial and urban wastewater. *Environmental Monitoring and Assessment*, *165*(1–4), 103–112. https://doi.org/10.1007/s10661-009-0930-7

[29] Willems, P. (2008). Quantification and relative comparison of different types of uncertainties in sewer water quality modeling. *Water Research*, *42*(13), 3539–3551. https://doi.org/10.1016/j.watres.2008.05.013

[30] Yang, L., Hur, J., & Zhuang, W. (2015). Occurrence and behaviors of fluorescence EEM-PARAFAC components in drinking water and wastewater treatment systems and their applications: A review. *Environmental Science and Pollution Research*, *22*(9), 6500–6510. https://doi.org/10.1007/s11356-015-4204-2

[31] Yang, L., Su, J., Yu, Z., Chang, Y., Yu, D., Zhang, Z., & Dong, S. (2026). Development of a novel method of biochemical oxygen demand colorimetric detection and its application in actual water. *Talanta*, *297*, 128563. https://doi.org/https://doi.org/10.1016/j.talanta.2025.128563

[32] Yang, T., Han, Y., Zhang, M., Li, L., Chen, M., Li, N., & Wang, X. (2026). Quorum sensing enhances extracellular electron uptake of electrotrophic biofilm via metabolic cascade for aerobic biochemical oxygen demand sensing. *Bioresource Technology*, *439*, 133363. https://doi.org/https://doi.org/10.1016/j.biortech.2025.133363

[33] Zhang, X. et al. (2019). Dataset diversity in ML model training. *Chemical Engineering Research*, *48*(5), 405–415.