CrossMark

Development of variable pathlength UV–vis spectroscopy combined with partial-least-squares regression for wastewater chemical oxygen demand (COD) monitoring

Baisheng Chen [a], Huanan Wu [b], Sam Fong Yau Li [a,b,c,*]

[a] Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543, Singapore
[b] NUS Environmental Research Institute (NERI), #02-01, T-Lab Building (TL), 5A Engineering Drive 1, Singapore 117411, Singapore
[c] Shenzhen Engineering Laboratory for Eco-efficient Polysilicate Materials, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen 518055, PR China

ABSTRACT

To overcome the challenging task to select an appropriate pathlength for wastewater chemical oxygen demand (COD) monitoring with high accuracy by UV–vis spectroscopy in wastewater treatment process, a variable pathlength approach combined with partial-least squares regression (PLSR) was developed in this study. Two new strategies were proposed to extract relevant information of UV–vis spectral data from variable pathlength measurements. The first strategy was by data fusion with two data fusion levels: low-level data fusion (LLDF) and mid-level data fusion (MLDF). Predictive accuracy was found to improve, indicated by the lower root-mean-square errors of prediction (RMSEP) compared with those obtained for single pathlength measurements. Both fusion levels were found to deliver very robust PLSR models with residual predictive deviations (RPD) greater than 3 (i.e. 3.22 and 3.29, respectively). The second strategy involved calculating the slopes of absorbance against pathlength at each wavelength to generate slope-derived spectra. Without the requirement to select the optimal pathlength, the predictive accuracy (RMSEP) was improved by 20–43% as compared to single pathlength spectroscopy. Comparing to nine-factor models from fusion strategy, the PLSR model from slope-derived spectroscopy was found to be more parsimonious with only five factors and more robust with residual predictive deviation (RPD) of 3.72. It also offered excellent correlation of predicted and measured COD values with R^2 of 0.936. In sum, variable pathlength spectroscopy with the two proposed data analysis strategies proved to be successful in enhancing prediction performance of COD in wastewater and showed high potential to be applied in on-line water quality monitoring.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Water quality monitoring is of growing importance all over the world with increasing demand of the real-time water quality information. To monitor wastewater quality during different treatment processes is crucial to ensure the treatment efficiency and comply with increasingly stringent regulations [1–3].

Chemical oxygen demand (COD) is a common indicator of organic matter concentration to assess wastewater quality. The standard COD test involves adding toxic chemicals and is time-consuming, requiring 2–4 h until the result is obtained. With the development of optical techniques, spectroscopic analysis, including UV–vis spectroscopy [4–7], fluorescence spectroscopy [8–10]

and near-infrared spectroscopy [6,11], shows high potential in wastewater COD monitoring. They are fast, non-destructive and environment-friendly that requires no chemicals added. Among these spectroscopic techniques, UV–vis spectroscopy shows the most extensive application and exhibits the best correlation with COD by multivariate data analysis. In particular, partial least squares (PLS) regression is routinely adopted to generate regression model based on UV–vis spectral data to estimate the water quality parameters [12,13].

Conventionally, an optimal pathlength is required to be selected and fixed for UV–vis spectroscopic measurements. According to Beer–Lambert law, absorbance is proportional to pathlength. It is a crucial parameter for measurements as it defines the distance that light travels through a sample. However, the selection of the optimized pathlength is a difficult task because it varies from a few millimeters for wastewater influent to dozens of millimeters for wastewater effluent [4]. Furthermore, different treatment processes usually result in different matrixes and concentrations of wastewaters, thus

* Corresponding author at: Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543, Singapore. Tel.: +65 65162681; fax: +65 67791691.
E-mail address: chmlifys@nus.edu.sg (S.F.Y. Li).

requiring different pathlengths for different treatment processes. Even for the same treatment process, the wastewater concentrations vary significantly which would consequently compromise the performance of the monitoring technique. Therefore, UV–vis spectroscopic system with variable pathlength would be a more desirable solution.

A large quantity of spectral data, in this sense, would be generated thus requiring a suitable strategy to manage and analyze these data. Data fusion has become a popular method to deal with abundant amount of data in analytical chemistry recently [7,14–16]. It is the process to integrate multiple data from different sources with the goal of obtaining information of greater quality. However, simply fusing data does not generally deliver better results [17]. As far as UV–vis spectroscopy is concerned, data saturation occurs when too long a pathlength is implemented, whereas too short a pathlength would not be able to generate a sufficiently strong signal. Furthermore, not all the combinations of different pathlengths will yield better results than individual pathlength since data redundancy and poor information from some pathlengths might compromise the others. In order to overcome this problem, variable selection was implemented to select informative variables that contribute to multivariate regression. Two data fusion levels, low-level data fusion (LLDF) and mid-level data fusion (MLDF), combined with variable selection are presented and compared for analyzing variable pathlength spectral data.

Another novel strategy is also presented to handle informative features from different pathlengths by applying slope calculation to generate slope-derived spectroscopy. Slope spectroscopy™, developed by C Technologies, Inc. (Bridgewater, NJ, USA), is a spectroscopic technique that is applied in determining protein and antibody concentration with known extinction coefficients of proteins, particularly at high concentrations which traditionally requires a series of dilutions and is prone to preparation error and sample contamination [18,19]. This technique basically selects the peak wavelength for the slope calculation and successfully applies it in selecting the linear range for concentration calculation with a prior knowledge of the extinction coefficient of the analyte. But the slope information has not been reported for full-spectrum application or for regression purpose. Here, we propose a novel application of the slope calculation for each wavelength and incorporate calculated slope data for the corresponding wavelength to generate a slope-derived spectrum. As such, each sample measured at variable pathlength would be represented by a slope-derived spectrum. The slope-derived spectral data would be subsequently input for regression.

To the best of our knowledge, this is the first time to apply variable pathlength UV–vis spectroscopy to monitor wastewater quality. COD is chosen as the parameter to evaluate the application of the proposed variable pathlength spectroscopy. Two strategies, fusion analysis and slope-derived spectroscopy, are employed to analyze variable pathlength spectral data. This study focuses on investigating the prediction performance by the proposed variable pathlength spectroscopy compared with the conventional single pathlength spectroscopy.

2. Experimental

2.1. Sample preparation and UV–vis spectroscopic measurements

Wastewater samples were taken from four different treatment processes (clarification, flotation, activated sludge and effluent) of Singapore Airline Terminal Services (SATS) wastewater treatment plant. They are mostly composed of degraded food from airline catering and dishwashing detergents. All samples were filtered by 2.5 μm Whatman filter paper and kept at 4 °C before analysis. 98 wastewater samples were divided into two sets: 82 samples for

calibration set and 16 samples for validation set. The COD range in calibration set was 112 mg L^{-1} to 1872 mg L^{-1}, whereas in validation set, COD values were from 130 mg L^{-1} to 1792 mg L^{-1}.

UV–vis spectra were measured by HACH DR/5000 spectrophotometer with wavelength range from 200 to 650 nm ($\Delta = 2$ nm). A home-made sample cell (Fig. 1) was designed and fabricated to adjust pathlength (0.5–18 mm). The adjustment of pathlength, defined by the distance of two quartz windows, was subjected to the movement of a threaded cylindrical tube sealed with O-ring. One revolution of the cylindrical tube equals to 1 mm adjustment. Each wastewater sample was subjected to UV–vis spectroscopic measurements at 19 different pathlengths (0.5 mm and 1–18 mm with 1 mm increment). The pathlength accuracy was checked by comparing with standard cuvettes of 1 mm, 5 mm and 10 mm. The differences of absorbance values were found to be negligible. Absorbance values of different pathlengths were set to zero according to MilliQ deionized water (18.2 MΩ cm at 25 °C) at the corresponding pathlength. Triplicate measurements for each wastewater sample at each pathlength were taken and the average spectrum was generated for further data processing.

COD measurements were conducted according to standard methods using COD digestion reaction and direct reading spectrophotometer (DR/5000, Hach Company, USA).

2.2. Multivariate data analysis

All chemometric data analysis was made with Matlab 2010b Software (The MathWorks, Natick, MA, USA) and PLS Toolbox 7.0 (Eigenvector Research Inc., Wenatchee, WA, USA). Contiguous block with 10 splits was selected as the cross-validation method the relatively large number of samples analyzed in random order in this study [20]. During the cross-validation process, 82 samples were split into 10 blocks and each block would be left out and predicted using the model built by the rest of the data from the other nine blocks. Samples from each block would be tested out once accordingly. Regarding validation, samples from validation set are predicted by the model built with the calibration set.

Spectral data involve continuous information, so interval window instead of individual data is exploited [21]. Interval partial-least squares (iPLS) is suitable for this application. It involves selecting equidistant subintervals on spectra aiming to produce superior PLSR model compared to full spectral model [22]. The selection criterion of intervals is based on those that provide the lowest RMSECV generated by the selected variables. Two expansion methods of iPLS, designated as backward/forward iPLS (BiPLS/FiPLS), have been developed and performed sucessfully to optimize the infromative regions [23,24]. In FiPLS, PLSR models are built using successively improving intervals with respect to RMSECV measure [25]. If only one interval is desired, the algorithm would stop calculating at that point and produce the results. If, however, more than one interval is requsted, additional calculation cycles would be performed until no improvement of RMSECV is achieved.
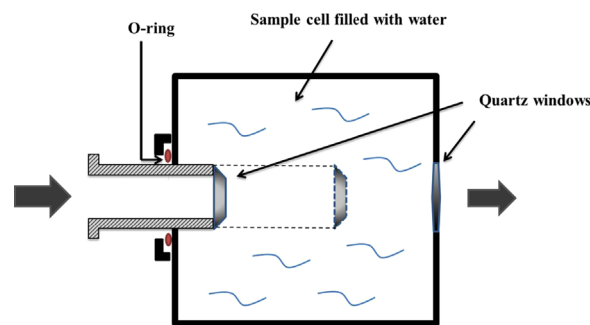

Fig. 1. Design of home-made sample cell with adjustable path length.