RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2020, **10**, 20691

Global calibration model of UV-Vis spectroscopy for COD estimation in the effluent of rural sewage treatment facilities†

Peng Li,‡[a] Jiangbei Qu,‡[a] Yiliang He, *[ab] Zhang Bo[a] and Mengke Pei[a]

In recent years, rural sewage treatment facilities have grown rapidly in China, and yet the water quality of the effluent has not been well monitored. The detection of chemical oxygen demand (COD) *via* ultraviolet-visible (UV-Vis) spectroscopy is an emerging technology with advantages of low cost and easy maintenance, which make it appropriate for the on-line monitoring of effluents from rural sewage treatment facilities. Because there are numerous sewage treatment devices in rural regions and as their locations are usually very scattered, it is difficult to calibrate the COD estimation model for each monitoring site. Hence, a COD estimation model with global calibration is a specific problem for application in rural regions. However, little research was performed on real rural sewage, yet much is desired in terms of the model accuracy and robustness. Consequently, a practical COD detection method with UV-Vis spectroscopy was established in this study. The COD estimation model was globally calibrated with effluents from rural sewage treatment devices. In order to avoid misleading data for evaluating the model performance caused by the differences in the COD concentration range of training sets, two new criteria, namely the Root Mean Square Relative Error (RMSRE) and Relative Error Variance (REV), were proposed to evaluate the model accuracy and robustness. Differences in the organic composition as characterized by excitation–emission matrix (EEM) fluorescence spectroscopy were shown to significantly affect the accuracy of the global calibration model. Through comparison among the methods of the partial least squares (PLS), support vector machine (SVM), and back-propagation neural network, PLS was verified to be able to attain sufficient accuracy and to be suitable for applying to the modeling with global calibration. A simplified modeling method was proposed to replace the absorption spectra at the full wavelength band with the absorbance at some specific wavelengths that were selected by interval partial least-squares regression (iPLSR) and synergy interval partial least-squares regression (siPLSR). In this study, the simplified model was proven to be reliable with three specific wavelengths: 251, 356, and 363 nm. An on-line COD monitor utilizing UV-Vis spectroscopy was thus developed for combining with the global calibration model.

Received 19th December 2019
Accepted 14th April 2020

DOI: 10.1039/c9ra10732k

rsc.li/rsc-advances

1 Introduction

As a developing country, China has more than 2.5 million natural villages and half of its total population lives in rural areas.[1,2] A decade ago, most rural sewage was discharged without treatment,[2] but with the promotion of the Chinese government, rural sewage treatment facilities have been rapidly constructed in recent years.[3] Nine local standards on the discharge of water pollutants for rural sewage treatment facilities have been promulgated until 2019. However, the water quality of the effluent from rural sewage treatment facilities has not achieved effective monitoring. On-line water quality monitors are widely used in tests for the discharge regulation and are regarded as the basis for the construction of an environmental information system.[4] Nevertheless, constrained by the high cost and long distance, it is a challenge to equip on-line water quality monitors in rural sewage treatment facilities, which are usually small in scale, large in quantity, and scattered widely over a specific geographical scope. In general, there would be several thousand treatment devices in the rural areas of one county, distributed over several hundred square kilometers. Hence, the ideal monitoring devices would be low cost, easy to maintain, and have stable operation.

Chemical oxygen demand (COD) is a key indicator of the concentration of reductive contaminants, and is considered to directly reflect the pollution level, and so on-line monitors for

[a]*School of Environmental Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: ylhe@sjtu.edu.cn*

[b]*China-UK Low Carbon College, Shanghai Jiao Tong University, Shanghai 200240, PR China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ra10732k

‡ Both authors contributed equally to this study.

COD are indispensable for the regulation of discharges.[5] The conventional COD on-line monitor is based on a method involving oxidization by potassium dichromate;[6] however, the device is very expensive and needs high maintenance costs. Taking HACH as an example, the device price is more than USD 17000, and the reagent cost is around USD 0.5 for each sample. Besides, the detection requires extra toxic chemicals and takes more than 2 hours.[5] Spectroscopic analysis, such as UV-Vis spectroscopy,[7] fluorescence spectroscopy[8] and near-infrared spectroscopy,[9] have been used to assess the wastewater quality, especially for predicting the COD concentration. These monitoring methods are rapid, non-destructive, and environmentally friendly. Meanwhile, the costs of the device utilizing optical techniques for COD monitoring are much cheaper than those of conventional ones.

The UV absorbance at a specific wavelength, such as 254 nm (A254),[10] or at multi-wavelengths in a narrow range,[11] were at first adopted to characterize the COD of wastewater. However, the COD concentration prediction was sometimes inaccurate because insufficient information was obtained from these limited wavelengths. Especially when the model were applied in different sampling sites, the relative deviation could reach as much as 45–50%. Many researchers attempted to use the UV-Vis absorption spectra of broader wavelengths, usually from 200 nm to 700 nm, for the construction of COD estimation models in wastewater quality monitoring.[7,12] On-line and *in situ* UV-Vis spectrophotometers have been extensively applied in recent years,[5,13,14] and could be a promising on-line COD monitoring technology too. Although the UV-Vis spectroscopic model, with increasing information available benefiting from much broader spectra, has been identified as a possible route to improve the performance of COD estimation in wastewater, the accuracy and robustness of the predictive results still leave much to be desired.

COD estimation models in previous studies were mainly calibrated by samples with a relatively constant composition, such as simulated wastewater prepared in a laboratory,[15] water samples from the same site as the wastewater treatment plant[16] or surface water,[17] which is called local calibration.[18] However, when a model set up with local calibration is applied to another monitoring site, the accuracy of the prediction results would decrease due to the differences in the organics composition. Therefore, the global calibration, which refers to calibrating a model with samples from different sampling sites in the target area,[18] should be adopted to promote the application of one model to many more monitoring sites. In terms of an on-line COD monitor for rural sewage treatment facilities, a global calibration model is appropriate due to the difficulty to calibrate the model for each treatment facility. Clearly, it is necessary to quantitatively assess the model accuracy used in global calibration and to verify the feasibility of the developed COD detection method. However, the feasibility of doing this has still rarely been investigated, and it might be impacted by differences in the organics composition when monitoring different treatment devices. Many statistic methods have been adopted to establish a COD estimation model with UV-Vis absorption spectra, including the traditional statistic methods (Linear Regression (LR), Partial Least Squares (PLS)), and the machine learning methods (Support Vector Machine (SVM), Back-Propagation Neural Network (BPNN)).[18] Some researchers proved that the linear methods, such as PLS and LR, are accurate enough to find a correlation between the spectral data and COD concentration according to the Lambert–Beer law.[19] While the machine learning methods were also found to offer good performance.[20] However, there is no consistent conclusion yet on the best modeling method, especially for the global calibration modeling of rural sewage. Given the differences in the organics composition in waters, a suitable modeling method might need to be specially investigated.

In previous studies, the quality of the COD estimation models were evaluated by a series of criteria, such as R^2, root-mean-square error (RMSE), root mean squared error standard deviation ratio of observations (RSR), and range to error ratio (RER).[21] However, these criteria can be influenced by the sample quantity or concentration range and lead to misleading conclusions in some conditions. For example, the frequently used coefficient of determination (R^2) may produce high values in poor models, and it is insensitive to additive and proportional differences and is also oversensitive to high extreme values.[22–24] No universal criteria have been established yet to fairly assess the predictive performance of different models.

The COD estimation method utilizing UV-Vis absorption spectroscopy usually uses a wavelength in the range of 200–700 nm,[25] so that it needs to be equipped with a broad wavelength light source, such as a xenon lamp. Meanwhile, some researchers used different wavelength regions, less than 200–700 nm, to construct a model for different wastewaters, while still guaranteeing the accuracy and robustness of their results.[5,12] The full-spectrum could be replaced by the absorption of a representative wavelength region, maintaining an adequate accuracy and robustness. Hence, a series of narrow band light sources could replace the broad band ones, and the cost of the device could decrease. However, few existing studies have focused on the selection of the optimal wavelength region to date.

The COD estimation method utilizing UV-Vis absorption spectroscopy needs comprehensive investigation. The aims of this study included: (1) to propose a practical COD detection method utilizing UV-Vis spectroscopy; (2) to evaluate the modeling methods in terms of accuracy and robustness with samples from rural sewage treatment facilities; (3) to determine the representative wavelength regions with interval partial least-squares regression (iPLSR) and variable selection; (4) to verify the feasibility of the developed COD detection method with a lab manufactured on-line COD monitor based on UV-Vis absorption spectroscopy. All the studies were performed with an aim to set up a global calibration for the optimization of COD measurement, which can then be applied to any wastewater, not just the effluent from biotreatment systems used for rural sewage.

2 Materials and methods

2.1 Water samples collection

A total of 150 water samples were collected from the rural sewage treatment devices dispersed around the countryside of Changshu City, Jiangsu province, China. In order to ensure that

each water sample was representative, the water sample collection sites were distributed in all rural areas in Changshu City. Besides, the sewage treatment devices at the sampling points were working normally, and the average daily water intake was 300–500 L. Of the samples, 110 marked as Group A were collected from different devices on the same day, March 20, 2018. While 40 samples marked as Group C were collected from one device on different days, from March to April in 2018.

2.2 COD analysis and optical measurements

The COD concentrations of the samples were detected according to Chinese standard methods (GB 11914-89) after filtration by 0.45 μm filters. UV-Vis absorption spectra were detected by a spectrophotometer (DR6000, HACH, USA) with the wavelength range from 190 to 1100 nm and the wavelength resolution of 1 nm. The spectra of all the water samples are shown in Fig. S1.† The absorption spectra from 200 to 700 nm were adopted as the input for the subsequent modeling because of the higher signal-to-noise ratio.[26] Fluorescence excitation–emission matrix (EEM) spectra were determined with a fluorescence spectrophotometer (F-7000, HITACHI, Japan). The excitation wavelength ranged from 200 to 500 nm, and the emission wavelength ranged from 250 to 550 nm.

2.3 Methods for the model construction for COD estimation

The raw UV-Vis absorption spectra were preprocessed by SG-smoothing, multiple scattering correction (MSC), and standard normal variate (SVN) with the software Unscrambler X 10.4.

Three methods for model construction for COD estimation, including partial least squares (PLS) regression, support-vector machines (SVM), and back-propagation neural network (BP-NN), were used and compared with the same data in this study. PLS regression, as a statistical method that bears some relation to principal components regression, includes a cross-verification procedure and identifying outliers. Also, the optimal number of components can be automatically confirmed by the software. SVM is a supervised learning model with associated learning algorithms that analyze the data used for classification and regression analysis. SVM regression and classification are very useful in order to detect patterns in complex and non-linear data,[27] and the method is used in conjunction with the leave-one-out cross-validation program (LOOCV), with the squared difference between the observed and SVM regression estimated data being the objective function to determine the most predictable SVM method. Both the PLS and SVM method were performed using Unscrambler X 10.4. The BP-NN is one of the artificial neural networks that adds the concept of backpropagation to artificial neural networks. It can create both a linear regression model and non-linear regression model with satisfying properties. The Sigmoid function was selected as the kernel function of the BP-NN, and the BP-NN model was conducted by the AMORE package in R software.

2.4 Model evaluation criteria

Many criteria, such as the R^2, root-mean-square error (RMSE), RSR, RER, have been used for model performance evaluation in previous studies.[21] However, these criteria might be influenced by the sample number or concentration range and can lead to misleading conclusions under some conditions. In this study, two new criteria were proposed to evaluate the performance of the COD estimation model. The Root Mean Square Relative Error (RMSRE) is defined by eqn (1) and (2), and the Relative Error Variance (REV) by eqn (1) and (3). RMSRE takes the observation value into consideration and it would not be influenced by the sample number and concentration range. The REV is the variance of the relative value between the observed value and the predicted value, which can be used to reflect the robustness of the model. The RMSRE, REV, together with R^2, and RMSE were calculated to evaluate the model performance in this study. The suitability of these criteria were compared.

$$u = \frac{y_i - \tilde{y_i}}{y_i} \quad (1)$$

$$\mathrm{RMSRE} = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n}} \quad (2)$$

$$\mathrm{REV} = \sqrt{\frac{\sum_{i=1}^{n} (u_i - \bar{u})^2}{n}} \quad (3)$$

where n: number of samples in the set. y_i: the observation value. $\tilde{y}_i$: the predicted value. u: the relative error. $\bar{u}$: average relative error.

2.5 Method for wavelength interval selection

The wavelength intervals were optimized by interval partial least-squares regression (iPLSR) and synergy interval partial least-squares regression (siPLSR). Both the methods used the software to split the UV-Vis absorption spectra of the data set into a number of intervals, and then found the best interval combinations according to the R^2 and the RMSE. The characteristic wavelengths were found with lasso regressions in the characteristic band screened with siPLSR. The goodness of fit of the model was evaluated by the Akaike information criterion (AIC), which is supposed to reach the smallest value when the model is optimized. Multiple Linear Regression algorithm (MLR) was used to establish the COD estimation model using the selected characteristic wavelengths. They were performed using the iToolbox[28] for Matlab.

3 Results and discussion

3.1 Influence of the COD concentration range on the model evaluation criteria

Evaluation criteria are crucial to choosing the most suitable COD estimation model among various candidates. Besides R^2 and RMSE, which are the most common criteria, REV and RMSRE were calculated and used to evaluate the performance of the predictive models. The samples of Group A were divided into three subgroups according to different COD concentration ranges, marked as data sets A1, A2, and A3, respectively. The

COD concentrations of each group were correspondingly in the range of 20 to 100 mg L^{-1}, 20 to 200 mg L^{-1}, and 70 to 150 mg L^{-1}. The COD estimation model was established by the PLS method, and 70% of the data in each data set were used for calibration while the rest were used for validation.

The criteria of the COD estimation model, including R^2, REV, RMSE, and RMSRE, were calculated with various data sets that had different COD concentration ranges. Through comparing the criteria values obtained from the different data sets, the influence of the COD concentration range on the criteria could be investigated. The results are shown in Table 1, in which data sets A1, A2, and A3 were collected from different sampling sites, and correspondingly, the COD concentrations ranged from 20 to 100 mg L^{-1}, 20 to 200 mg L^{-1}, and 70 to 150 mg L; Data set C was collected from one sampling site at different times. When the model was calibrated using Data set A1, the R^2 reached 0.843 and RMSE was 6.622. When calibrated using Data set A2, the R^2 increased to 0.949 and RMSE increased to 9.972. According to the criteria of R^2, the predicted performance of the model seemed to be optimized from A1 to A2, but when assessed by the criteria of RMSE, the opposite conclusion could be reached. However, for a reliable model, the performance, including accuracy and robustness, should be approximately uniform in each COD concentration region. The fact was that neither the measuring instruments nor the modeling methods were improved, but the calibration data were different based on the scale of the COD concentration range for A1 and A2. It was obvious that the scale of the COD concentration range could affect the R^2 and RMSE of the model and possibly lead to misleading conclusions. Similar results also could be found in previous reports. Charef *et al.*[29] established a COD estimation model for municipal sewage, and the R^2 of their model was 0.95 using the samples with COD concentration ranging from 112 to 422 mg L^{-1}, while Langergraber *et al.*[7] obtained an R^2 of 0.978 when the COD concentration ranged from 38 to 568 mg L^{-1} in similar studies. Due to the potential possibility of excessive fitting, it would be imprecise to assess the model based on R^2. Taking the influence of the data set range into account, REV and RMSRE were proposed as model evaluation criteria. As shown in Table 1, both REV and RMSRE were very close between the models calibrated using A1 and A2, indicating that the interference of the scale of the COD concentration range could be avoided when assessing the model with REV and RMSRE.

As shown in Table 1, the R^2 of the model calibrated using Data set A1 was larger than that calibrated using Data set A3, while the RMSE of the former was smaller than that of the latter. The results appeared to indicate that the performance of the model was worse in the COD concentration range of 70–150 mg L^{-1}. However, as previously mentioned, because the model has been confirmed to be reliable in the target range of COD concentration, the performance was supposed to be approximately uniform. Hence, R^2 would not be affected by the change in the observation value range according to the statistics. Comparing data sets A1 (20–100 mg L^{-1}) and A3 (70–150 mg L^{-1}), they were the same in the scale of COD concentration range, but different regions. Hence, besides the scale of the COD concentration range, the R^2 and RMSE were proved to be affected by the region too. In terms of the RMSER and the REV, both the models had the same accuracy and robustness. Consequently, the performance of the predictive models would not be affected when the COD concentration regions have the same width but different values.

Data set A2 were obtained from the water samples collected from different devices, while Data set C were from one device taken at different times. The organic composition of the samples in Group A2 was supposed to be more diverse than that of Group C. The performance of the model could be better when calibrated using samples with a constant organic composition. Zhao *et al.*[30] used a dilution of the potassium hydrogen phthalate and established the relationship between the COD concentration and A254–A546, and the R^2 was up to 0.995. It is easy to deduce that the performance of the model calibrated using Data set C would be better than that of Data set A2. The criteria of RMSE, REV, and RMSRE confirmed this hypothesis. However, the R^2 of the model calibrated using Data set A2 was very close to that calibrated using Data set C, indicating the misleading results when using the criterion of R^2.

When the COD estimation method by UV-Vis spectroscopy was applied in practical engineering, the influences of various factors, such as the COD concentration range and the organic components in the water, should be considered. Many researchers stated that the criterion of the coefficient of determination (R^2), which is frequently used, might reach to a very high value, even in a poor model. The RMSE is related to the COD concentration range. Hence, based on the criterion of RMSE, the accuracy could not be compared between the models that were calibrated using the data sets with different COD concentration ranges. It was proved that this problem could be solved by RMSRE. The REV has also been proved to be a suitable criterion for evaluating the robustness of the model. Hence, the

Table 1 Criteria of COD estimation models with different data sets

	A1		A2		A3		C	
Date set	Cal.	Val.	Cal.	Val.	Cal.	Val.	Cal.	Val.
RMSE	6.622	6.508	9.972	10.725	12.121	12.988	5.170	5.684
R^2	0.843	0.887	0.949	0.921	0.743	0.790	0.933	0.876
RMSRE	0.141	0.137	0.137	0.143	0.138	0.140	0.104	0.106
REV	0.0187	0.0199	0.0193	0.0178	0.0178	0.0183	0.011	0.0116

two new criteria can better reflect the predictive performance of the model. RMSRE is used to evaluate the prediction accuracy of the model. Compared with RMSE, it will not be affected by the size of the measured value. It can be used to evaluate the predictive performance among different models. REV is used to evaluate the predictive robustness of the model. The calculation formula is simple and the evaluation effect is real and reliable. It is thus suitable to choose the RMSRE and REV to evaluate the performance of the model.

3.2 Performances of models established by different methods

COD estimation models were established by various methods, including PLS, SVM, and BP-neural network, using the UV-Vis absorption spectra from Data set A. The performance of accuracy and robustness were evaluated by the criteria of R^2, REV, RMSE, and RMSRE and the are shown in Table 2. There was no significant difference among the R^2 of these three modeling methods. For RMSE, RMSRE, and REV, the results from the PLS model and SVM model were very close and slightly lower than those from the BP-neural network model. This result indicated that in terms of the accuracy and robustness, PLS and SVM would be the better modeling methods, although the differences were not very large among them. Considering the convenience of modeling at the same time, PLS, as a linear modeling method, was more suitable. In addition, although the evaluation finding based on RMSE was the same as that based on RMSRE, only the value of RMSRE could be used to compare the model performance with other models in different COD concentration ranges. According to previous reports, many modeling methods have been used for predicting the COD concentration by UV-Vis absorption spectra.[17,18,26] Some researchers made comparisons under the same experimental conditions. However, they did not reach agreement on the most suitable modeling method. Lepot *et al.*[18] considered that the methods of PLS and SVM performed very well for calibration. When comparing with the BP-neural network, the PLS model was more accurate, while the BP-neural network model was more robust. Brito *et al.*[21] found the PLS model was adequately accurate for COD estimation and it did not need more complex algorithms. The COD in the effluent of a biotreatment device is mainly contributed by organics, which might generate absorption spectra in the UV-Vis band.[31] According to the Lambert–Beer law, the concentrations of a certain organic component would be linearly dependent with some special absorption spectra in the range of the UV-Vis band.[32] On the other hand, if the compositions of organics in water samples were not uniform, or the detection of the spectra had interference by some external factors, such as the turbidity and the scattering, the performance of a linear model would decline, while a non-linear model might be more effective. For the BP-neural network method, a multi-layered neural network is trained, which can express the functional relation, including a linear or non-linear relationship.[33] However, a large quantity of calibration data is essential for the BP-neural network model, indeed, much more than for other modeling methods. Profiting from the development of on-line monitoring, it would be easy to obtain large amounts of data. It could be anticipated then that the BP-neural network would provide better performance.

The choice for the best modeling method needs to take both the mathematical complexity and the performance into consideration, and the model performance in terms of accuracy and robustness might be affected by the realistic practical conditions. There has not yet been a universal modeling method reported with acceptable performance for different kinds of wastewater. Establishing a specific model for each type of wastewater may thus be the most efficient way. In this study, the PLS method was found to be suitable for estimation modeling of the COD concentration in the effluent of biotreatment devices used for rural sewage.

Table 2 Performances of the COD estimation model according to PLS, SVM, and BP-neural network with the criteria of R^2, REV, RMSE, and RMSRE

	PLS		SVM		BP-neural network	
Methods	Cal.	Val.	Cal.	Val.	Cal.	Val.
RMSE	11.03	10.384	10.88	11.472	11.797	10.650
R^2	0.949	0.945	0.953	0.931	0.942	0.979
RMSRE	0.159	0.151	0.1625	0.1737	0.1696	0.1495
REV	0.023	0.0224	0.0237	0.025	0.0238	0.0213

3.3 Influence of the organics composition on the performance of a COD estimation model

The utilization of UV-Vis spectroscopy for COD measurement is a method for determining the amount of organic matter based on the absorption of ultraviolet-visible light. Because the UV-Vis absorption spectrum is specific for each organic compound, and the UV-Vis absorption spectrum observed is the accumulation curve contributed by multiple organics in water, the organics composition is supposed to affect the correlation between the COD concentration and UV-Vis absorption spectrum. The influence of the organics composition on the performance of COD estimation model was investigated by comparing the models calibrated using the different data sets. The case that only one kind of organic compound exists in water was simulated using potassium hydrogen phthalate (PHP) as the target contaminant. Next, 40 water samples were collected from different biotreatment devices, which were marked as Group A, while another 40 water samples with a relatively constant composition of organic compounds were collected from one biotreatment device at different times, which were marked as Group C. The COD estimation model for the PHP solution was established by unitary linear regression using absorbance at 254 nm. Meanwhile, for the samples of Group A and Group C, the models were established by the PLS algorithm using the UV-Vis absorption spectra. As shown in Fig. 1, the COD concentration of PHP solution was extremely well correlated with the absorbance at a wavelength of 254 nm. The performances of the models for Group A and Group C were

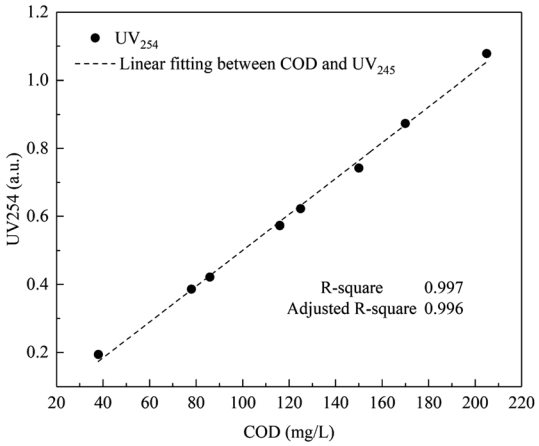

Fig. 1 Correlation between the UV_{254} and COD concentration (potassium hydrogen phthalate solution).

Table 3 Values of the applied criteria for each data set

	A		C	
Data set	Cal.	Val.	Cal.	Val.
RMSE	11.990	10.720	5.270	4.640
R^2	0.943	0.934	0.928	0.955
RMSRE	0.139	0.109	0.119	0.099
REV	0.019	0.011	0.014	0.008

evaluated by the criteria of RMSE, R^2, RMSRE, and REV, and the results are shown in Table 3. According to the RMSE and RMSRE, the accuracy of the model for Group C was higher than that for Group A; while in terms of the robustness, the model for Group C was also better than the model for Group A.

The organic compositions of the samples in Group A and Group C were investigated. Most organics in the effluent of sewage biotreatment devices would emit fluorescence under excitation conditions. The characteristics of excitation–emission fluorescence depend on the organic types. Hence, the fluorescence excitation–emission matrix regional integration (FRI) could be used to analyze the organics composition in water samples.[34] Excitation–emission matrix (EEM) fluorescence spectroscopy analysis of the water samples was divided into five regions, where every region represented a different type of compound. The proportions of the five types of organic compounds in Group A and Group C are shown in Fig. 2. The main substances in the samples include soluble microbial by-product-like substances (10–20%), fulvic acid-like substances (10–20%), and humic acid-like substances (60–70%). The variances of the proportion of humic acid-like, soluble microbial by-product-like, and fulvic acid-like substances were 0.0039, 0.0006, and 0.00147 for Group A, and correspondingly 0.0005, 0.00013, and 0.00017 for Group C. It was obvious that the differences in the organic composition in the samples from Group C were less than that from Group A. As the water samples of Group C were from the same biotreatment device that was operated in a stable condition, the organics composition was correspondingly stable. Therefore, in terms of the differences in organic composition among the water samples, that of the PHP solution was less than that of Group C, and the latter was less than that of Group A. Corresponding to the model performance for each group, it was thus obvious that the organic composition in the water samples would significantly affect the model performance. A higher accuracy and robustness of the COD estimation models would be obtained from a lower difference in organic composition in both the calibration and validation samples.

In the previous studies, the models were calibrated with simulation wastewater samples prepared in a laboratory or with samples from one sampling site,[15–17] which were similar with the research conditions of the PHP solution and Group A in this study. A method that is calibrated with a specific organic composition in samples is called local calibration.[7,18] According to the previous reports, local calibrated models could usually achieve satisfactory performances, benefitting from the relatively consistent organic composition.[35,36] Hu *et al.*[36] proposed a local calibration method with samples from four factories, and found the prediction accuracy was successfully improved. However, if on-line COD monitors were to be equipped on rural sewage treatment devices, it would be very difficult to calibrate each COD estimation model for each monitoring site, because in the rural

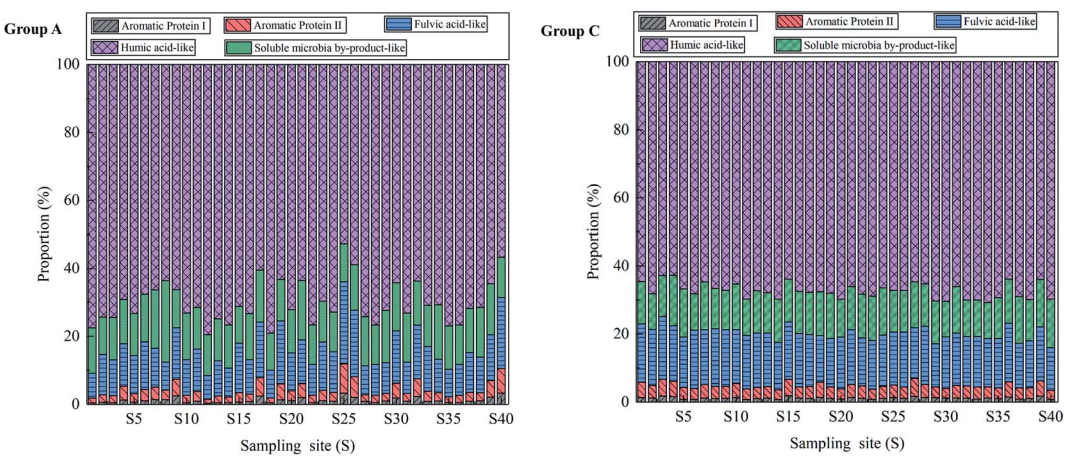

Fig. 2 Distribution of FRI in non-fractionated DOM from Group A and Group C water samples.

region, the amount of sewage treatment devices is very large and their locations are scattered. Therefore, the model can only be calibrated with samples from different treatment devices, which is called global calibration.[7,18] Because the organic composition of the samples for calibrating is inconsistent, the model accuracy is usually worse for global calibration than for local calibration. In this study, the COD estimation model was globally calibrated with samples from different treatment devices. The model accuracy was found to be acceptable with the RMSRE of 0.139. Although the accuracy was less than that using the chemical oxidation method, it was enough for monitoring the effluent of rural sewage treatment devices. Using a global calibration model to predict COD with UV-Vis spectroscopy has been verified as a reliable method in practice.

3.4 Model optimization by specific wavelength region

The COD estimation model with PLS needs the full-spectrum, which involves approximately hundreds of data points. It was precisely because of the complete data use that the prediction accuracy reached a relatively high level. On the other hand, it is difficult to obtain and treat excessive amounts of data. Hence, under the premise of the prediction accuracy finitely declining to an acceptable level, reducing the input data requirements of a model is a significant factor for reducing the manufacturing costs of on-line monitors.

The UV-Vis absorption spectra of all the samples with different COD concentrations are shown in Fig. 3. The difference in absorbance among these samples was very little at the wavelength beyond 400 nm. Furthermore, the absorbance at the wavelength beyond 600 nm dropped to small enough a level to be ignored for all the samples. There were obviously some non-COD-related wavelengths in the range from 200 to 700 nm. The absorbance at these wavelength should be eliminated during modeling so that the model would avoid such interference and to simplify the calculation.

The full UV-Vis absorption spectra of the water samples from Group A were equally split into 30 intervals in the wavelength range from 200 nm to 700 nm. The COD estimation models were established by PLS using the spectra of each interval and their combinations. Using the methods of iPLSR and siPLSR, the best interval combination could be screened from all possible combinations with the same amount of intervals. The optimal models using each amount of interval from one to thirty were established with the best interval combinations that were screened by iPLSR and siPLSR. The model performance was evaluated by the criteria of RMSRE and REV. The RMSRE and REV of the optimal models established by one, two, three, four, five, and all the thirty intervals are shown in Table 4. The optimal model was found to be established with the data of three interval combinations, in which the lowest REV and RMSRE were acquired. The optimal wavelength intervals were located in the near ultraviolet range from 200 nm to 400 nm, including 251–268 nm, 319–336 nm, and 353–370 nm, respectively. It has been confirmed by previous studies that the general saturated organic compounds would have no absorption in the near ultraviolet region, while those containing conjugated double bonds or a benzene ring would have obvious absorption in the ultraviolet region or a characteristic peak.[36] The main absorption wavelength of simple aromatic compounds containing a benzene ring is in the range of 250 nm to 260 nm.[12,36] Besides, the UV absorbance is positively correlated with the molecular weight of an organic compound.[37] Hence, the specific intervals identified by iPLSR and siPLSR could also bring information on the organic composition to some extent.

For the absorption spectra in a narrowband range, the absorbance at some wavelength was observed to be linearly correlated with those at its adjacent wavelength, and the R^2 was even more than 0.95. Therefore, a series of specific wavelengths could be selected, by which the absorbance could replace the spectral data to establish the model without significantly reducing the model accuracy.[38] In this study, lasso regressions and stepwise regression were used to select the specific wavelengths from the optimal wavelength intervals. Using the lasso regressions, the specific wavelengths were discovered to be 251, 356, 357, 362, and 363 nm. The stepwise regression was subsequently implemented, and the minimum AIC was 123.24. The specific wavelengths were finally concluded to be 251, 356, and 363 nm. Three peaks were observed at these three wavelengths in the UV-Vis absorption spectra, which might be from the organics with an aromatic structure or conjugated double bond.[39]

When using the absorbance at specific wavelengths for modeling, the variables involved sharply decrease, so that some

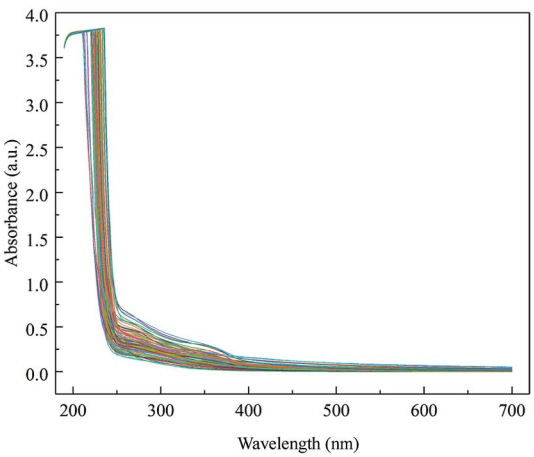

Fig. 3 UV-visible absorption spectra of all samples.

Table 4 REV and RMSRE of the optimal model with different amounts of interval combinations

	RMSRE		REV	
	Cal.	Val.	Cal.	Val.
One interval	0.17	0.16	0.025	0.022
Two intervals	0.13	0.135	0.018	0.0163
Three intervals	0.124	0.133	0.015	0.0165
Four intervals	0.146	0.14	0.021	0.02
Five intervals	0.152	0.148	0.022	0.024
All data	0.121	0.12	0.014	0.0135

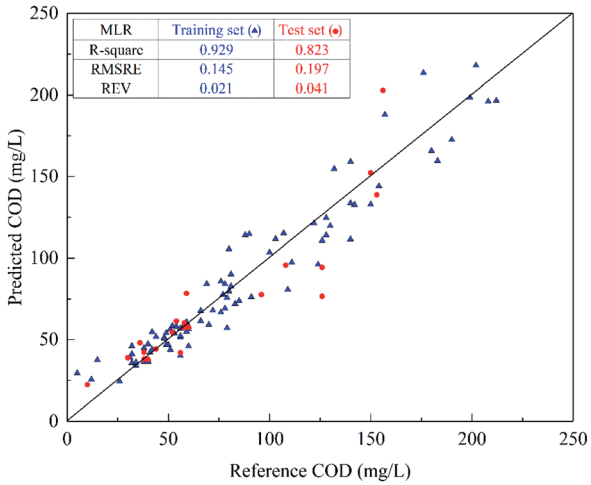

Fig. 4 Comparison between predicted COD concentrations and the true COD concentrations.

simple algorithms could become effective. Based on the three identified specific wavelengths, the Multiple Linear Regression (MLR) method was adopted for the COD estimation modeling. A comparison between the predicted COD concentrations and the true COD concentrations is shown in Fig. 4. The model performance was evaluated by the criteria of RMSRE and REV (Fig. 4). Compared with the value of these criteria of the model using the UV-Vis absorption spectrum (200–700 nm), the RMSRE only increased from 0.139 to 0.145 and the REV increased from 0.019 to 0.021, indicating that there was just a slight decline in the accuracy and robustness of the predictive results. Li *et al.*[17] optimized a COD determination model using a similar method with 144 samples collected from a lake. The predicted accuracy and model stability were improved by iPLS and siPLS, which was contrary with this study. In this study, the samples were collected from more than 50 rural sewage treatment devices, so the organic composition would be more diverse here. Therefore, the different findings may be ascribed to the differences in the organic composition, but this needs further study. Nevertheless, the specific wavelength selection can simplify the modeling process and guarantee the model accuracy and robustness to a reliable level. Benefitting from this optimization, the modeling process could be simplified a lot. Meanwhile, an even more important advantage could be achieved by this optimization. The narrowband light source, manufactured with a light emitting diode (LED), could take the place of the broadband light source due to only a few characteristic wavelengths being involved in the modeling. Therefore, the price of the monitors would sharply decrease. This would be significant for the practical application of COD measurement by UV-Vis spectroscopy.

3.5 Developing on-line COD monitors utilizing UV-Vis spectroscopy

Based on the above studies, an on-line COD monitor utilizing UV-Vis spectroscopy was developed for combining with the global calibration model. The schematic diagram of the internal structure and photos of prototype are shown in the ESI (Fig. S2 and S3†). The device was equipped with a spectrometer with 256 pixels PDA (MMS, Carl Zeiss, Germany) and a xenon lamp as the light source (L4642, Hamamatsu, Japan). The printed circuit board and power module for the light source were specifically designed in order to adapt to an immersed installation. The UV-Vis absorption spectra could be steadily acquired (Fig. S4†).

Table 5 Comparison between the predicted COD concentrations and the true COD concentrations

True COD (mg L^{-1})	Predicted COD (mg L^{-1})	Relative error
32.0	30.7	−4.1%
40.0	42.2	5.5%
66.0	83.6	26.7%
79.0	83.9	6.2%
90.0	93.8	4.2%
130.0	127.3	−2.1%
154.0	160.4	4.2%
59.0	61.1	3.6%
60.0	71.4	19%
199.0	193.3	−2.9%

The lab-manufactured monitor was installed in the field for monitoring the effluent, and real samples were collected and tested. The predictive results are shown in Table 5. When the true COD concentration ranged from approximately 30.0 to 200.0 mg L^{-1}, the relative error of the predicted COD was most acceptable in the range from −4.1% to 6.2%, with occasional outliers. The results indicated that the COD estimation method is effective. In addition, it should be noted that turbidity and the particles in the water will affect the UV-Vis spectrum, and the prediction accuracy would be reduced.[15,40,41] The on-line monitor was used to test the effluent of sewage treatment devices in practice. During the testing period, the turbidity and the particles were constant in the effluent due to the good operating conditions. Therefore, the interference would be not significant. Nevertheless, the influence of turbidity and the particles on the UV-Vis spectrum were very crucial for COD prediction. It is necessary to comprehensively investigate this further.

One feature of the monitor was adopting a default spectrum instead of the reference beam, and only the measuring beam was reserved. The device structure became much simpler and the interference due to instability of the reference spectra could be avoided. Another useful feature is configuring a wireless data transmission module based on General Packet Radio Service into the monitor. The real-time on-line data of the UV-Vis spectra could be transmitted to the cloud platform, and treated using the COD estimation model. The predicted COD concentration could thus be obtained and transmitted to the management platform or mobile termination of relevant people. The cost of this on-line COD monitor is competitive with the conventional ones. Accounting for the cost of the prototype, only USD 5000 was needed, which is less than 30% of the price of conventional COD monitors. Taking bulk production into account, there is still a lot of space to cut back the device costs.

Benefiting from these advantages, including the easier maintenance, lower price, little operating cost, and connectability to the Internet of things, the developed on-line COD monitor could be installed in the massive number of rural sewage treatment facilities. An information management system might be thereby supported based on the developed COD monitor.

4 Conclusion

Overall, in this study, a COD estimation model of UV-Vis spectroscopy with global calibration was established with various optimization measures, and was verified to be effective for monitoring the effluent from rural sewage treatment devices. In order to more accurately assess the model performance, two new evaluation criteria of REV and RMSRE were proposed to assess and compare the accuracy and robustness of models that were calibrated in different concentration ranges. The PLS was proved to be the optimal modeling approach, regarding convenience and accuracy at the same time. In the condition of global calibration, the model performance was found to significantly correlate with the identity of the organic composition. For the effluent from the rural sewage treatment device in this study, the predicted COD was credible, with an RMSRE of less than 0.14. A simplified model was established using a series of absorbance at specific wavelengths instead of the absorption spectra at the full waveband range, and the accuracy was proved to be reliable, with a slight increase in the RMSRE. The global calibration model of UV-Vis spectroscopy for COD estimation thus achieved feasible performance with sufficient accuracy and convenience. This provides a promising and practical strategy for the monitoring of effluent from scattered sewage treatment devices in rural areas.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Key R&D Program of China (grant numbers: 2016YFC0400801); National Science and Technology Major Projects of Water Pollution Control and Management of China (grant numbers: 2017ZX07206004).

References

1 Y. Han, J. Ma, B. Xiao, X. Huo and X. Guo, New integrated self-refluxing rotating biological contactor for rural sewage treatment, *J. Cleaner Prod.*, 2019, **217**, 324–334.
2 Z. Wang, China's wastewater treatment goals, *Science*, 2012, **338**, 604.
3 J. Chen, Y. Liu, W. Deng and G. Ying, Removal of steroid hormones and biocides from rural wastewater by an integrated constructed wetland, *Sci. Total Environ.*, 2019, **660**, 358–365.
4 C. B. Ojeda and F. S. Rojas, Process analytical chemistry: applications of ultraviolet/visible spectrometry in environmental analysis: an overview, *Appl. Spectrosc. Rev.*, 2009, **44**, 245–265.
5 B. Chen, H. Wu and S. F. Y. Li, Development of variable pathlength UV-Vis spectroscopy combined with partial-least-squares regression for wastewater chemical oxygen demand (COD) monitoring, *Talanta*, 2014, **120**, 325–330.
6 J. Bridgeman, A. Baker, C. Carliell-Marquet and E. Carstea, Determination of changes in wastewater quality through a treatment works using fluorescence spectroscopy, *Environ. Technol.*, 2013, **34**, 3069–3077.
7 G. Langergraber, N. Fleischmann and F. Hofstädter, A multivariate calibration procedure for UV/Vis spectrometric quantification of organic matter and nitrate in wastewater, *Water Sci. Technol.*, 2003, **47**, 63–71.
8 S. Lee and K. H. Ahn, Monitoring of COD as an organic indicator in waste water and treated effluent by fluorescence excitation–emission (FEEM) matrix characterization, *Water Sci. Technol.*, 2004, **50**, 57–63.
9 A. C. Sousa, M. M. L. M. Lucio, O. F. B. Neto, G. P. S. Marcone, A. F. C. Pereira, E. O. Dantas, W. D. Fragoso, M. C. U. Araujo and R. K. H. Galvão, A method for determination of COD in a domestic wastewater treatment plant by using near-infrared reflectance spectrometry of seston, *Anal. Chim. Acta*, 2007, **588**, 231–236.
10 J. Wu, M. N. Pons and O. Potier, Wastewater fingerprinting by UV-visible and synchronous fluorescence spectroscopy, *Water Sci. Technol.*, 2006, **53**, 449–456.
11 O. Thomas, F. Theraulaz, M. Domeizel and C. Massiani, UV spectral deconvolution: a valuable tool for waste water quality determination, *Environ. Technol.*, 1993, **14**, 1187–1192.
12 S. Fogelman, M. Blumenstein and H. J. Zhao, Estimation of chemical oxygen demand by ultraviolet spectroscopic profiling and artificial neural networks, *Neural Computing and Applications*, 2006, **15**, 197–203.
13 M. Zhao, S. Deng, B. Tang, J. Lu, P. Tang, H. Chen and S. Zhou, Research on Photoelectric Detection System Applying to UV-VIS Spectroscopy Water Quality Monitoring System, *DEStech Transactions on Computer Science and Engineering*, 2018, 403–406.
14 N. Bleyena, A. Albrecht, P. De Cannière, C. Wittebroodt and E. Valcke, Non-destructive on-line and long-term monitoring of in situ nitrate and nitrite reactivity in a clay environment at increasing turbidity, *Appl. Geochem.*, 2019, **100**, 131–142.
15 Y. Hu, Y. Wen and X. Wang, Novel method of turbidity compensation for chemical oxygen demand measurements by using UV-Vis spectrometry, *Sens. Actuators, B*, 2016, **227**, 393–398.
16 E. Carré, J. Pérot, V. Jauzein, L. Lin and M. Lopez-Ferber, Estimation of water quality by UV/Vis spectrometry in the framework of treated wastewater reuse, *Water Sci. Technol.*, 2017, **73**, 633–641.
17 J. Li, Y. Tong, L. Guan, S. Wu and D. Li, Optimization of COD determination by UV-Vis spectroscopy using PLS chemometrics algorithms, *Optik*, 2018, **174**, 591–599.

18 M. Lepot, A. Torres, T. Hofer, N. Caradot, G. Gruber, J. B. Aubin and J. L. Bertrand-Krajewski, Calibration of UV/Vis spectrophotometers: different methods to estimate TSS and total and dissolved COD concentrations in sewers, WWTPs and rivers, *Water Res.*, 2016, **101**, 519–534.

19 X. Qin, F. Gao and G. Chen, Wastewater quality monitoring system using sensor fusion and machine learning techniques, *Water Res.*, 2012, **46**, 1133–1144.

20 C. Wolf, D. Gaida, A. Stuhlsatz, T. Ludwig, S. Mcloone and M. Bongards, Predicting organic acid concentration from UV/Vis spectrometry measurements – a comparison of machine learning techniques, *Trans. Inst. Meas. Control*, 2013, **35**, 5–15.

21 R. S. Brito, H. M. Pinheiro, F. Ferreira, J. S. Matos and N. D. Lourenço, In situ UV-Vis spectroscopy to estimate COD and TSS in wastewater drainage systems, *Urban Water J.*, 2014, **11**, 261–273.

22 D. R. Legates and G. J. McCabe Jr, Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, 1999, **35**, 233–241.

23 P. Krause, D. Boyle and F. Bäse, Comparison of different efficiency criteria for hydrological model assessment, *Adv. Geosci.*, 2005, **5**, 89–97.

24 R. Harmel, R. Cooper, R. Slade, R. Haney and J. Arnold, Cumulative uncertainty in measured streamflow and water quality data for small watersheds, *Trans. ASABE*, 2006, **49**, 89–701.

25 D. P. Mesquita, C. Quintelas, A. L. Amaral and E. C. Ferreira, Monitoring biological wastewater treatment processes: recent advances in spectroscopy applications, *Rev. Environ. Sci. Bio/Technol.*, 2017, **16**, 395–424.

26 L. Guan, Y. Tong, J. Li, S. Wu and D. Li, An online surface water COD measurement method based on multi-source spectral feature-level fusion, *RSC Adv.*, 2019, **9**, 11296–11304.

27 B. Schölkopf, A. J. Smola and F. Bach, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, 2001.

28 L. Nørgaard and K. V. L. Denmark, *The iToolbox for MATLAB*, University of Copenhagen, Disponíisponsit: http://www.models.life.ku.dk/itoolbox, 2012, vol. 7.

29 A. Charef, A. Ghauch, P. Baussand and M. Martin-Bouyer, Water quality monitoring using a smart sensing system, *Measurement*, 2000, **28**, 219–224.

30 Y. Zhao, Y. Li, Y. Zhen and Y. Fang, A novel monitoring system for COD using optical ultraviolet absorption method, *Procedia Environ. Sci.*, 2011, **10**, 2348–2353.

31 H. Yu, F. Qu, X. Zhang, S. Shao, H. Rong, H. Liang, L. Bai and J. Ma, Development of correlation spectroscopy (COS) method for analyzing fluorescence excitation emission matrix (EEM): a case study of effluent organic matter (EfOM) ozonation, *Chemosphere*, 2019, **228**, 35–43.

32 J. Ma, F. Meng, Y. Zhou, Y. Wang and P. Shi, Distributed water pollution source localization with mobile UV-visible spectrometer probes in wireless sensor networks, *Sensors*, 2018, **18**, 606.

33 J. C. Cancilla, R. Aroca-Santos, K. Wierzchos and J. S. Torrecilla, Hazardous aromatic VOC quantification through spectroscopic analysis and intelligent modeling to assess drinking water quality, *Chemom. Intell. Lab. Syst.*, 2016, **156**, 102–107.

34 W. Chen, P. Westerhoff, J. A. Leenheer and K. Booksh, Fluorescence excitation–emission matrix regional integration to quantify spectra for dissolved organic matter, *Environ. Sci. Technol.*, 2015, **37**, 5701–5710.

35 G. Xin and J. L. Bertrand-Krajewski, A unified protocol for sensor calibration and verification in applications to WWTPs and sewer system monitoring, *Water Pollution*, 2012, vol. 164, pp. 391–402.

36 Y. Hu, C. Liu and X. Wang, Novel local calibration method for chemical oxygen demand measurements by using UV-Vis spectrometry, *IOP Conference Series: Earth and Environmental Science*, 2017, vol. 63.

37 J. R. Helms, A. Stubbins, J. D. Ritchie and E. C. Minor, Absorption spectral slopes and slope ratios as indicators of molecular weight, source, and photobleaching of chromophoric dissolved organic matter, *Limnol. Oceanogr.*, 2008, **53**, 955–969.

38 R. S. Brito, H. M. Pinheiro, F. Ferreira, J. S. Matos, A. Pinheiro and N. D. Lourenco, Calibration transfer between a bench scanning and a submersible diode array spectrophotometer for in situ wastewater quality monitoring in sewer systems, *Appl. Spectrosc.*, 2016, **70**, 443–454.

39 M. Wei, J. Zhang, X. Zhao, L. Zheng and Z. Xu, Research on a Spectral Recognition Method for On-line Measurement of COD in Dyeing Wastewater based on SIMCA, *RSC Adv.*, 2016, **6**, 110460–110465.

40 J. Agustsson, O. Akermann, D. A. Barry and L. Rossi, Non-contact assessment of COD and turbidity concentrations in water using diffuse reflectance UV-Vis spectroscopy, *Environ. Sci.: Processes Impacts*, 2014, **16**, 1897–1902.

41 A. Torres and J. L. Bertrand-Krajewski, Partial Least Squares local calibration of a UV-visible spectrometer used for in situ measurements of COD and TSS concentrations in urban drainage systems, *Water Sci. Technol.*, 2008, **57**(4), 581–588.

