PARTIAL LEAST SQUARES REGRESSION CALIBRATION OF AN ULTRAVIOLET-VISIBLE SPECTROPHOTOMETER FOR MEASUREMENTS OF CHEMICAL OXYGEN DEMAND IN DYE WASTEWATER

W. Mai, J.-F. Zhang, X.-M. Zhao*, Z. Li, Z.-W. Xu

School of Textiles, Tianjin Polytechnic University,
399 Binshuiwest Road, Xiqing District, Tianjin 300387, China; e-mail: tjpu_zhao@163.com

Wastewater from the dye industry is typically analyzed using a standard method for measurement of chemical oxygen demand (COD) or by a single-wavelength spectroscopic method. To overcome the disadvantages of these methods, ultraviolet-visible (UV-Vis) spectroscopy was combined with principal component regression (PCR) and partial least squares regression (PLSR) in this study. Unlike the standard method, this method does not require digestion of the samples for preparation. Experiments showed that the PLSR model offered high prediction performance for COD, with a mean relative error of about 5% for two dyes. This error is similar to that obtained with the standard method. In this study, the precision of the PLSR model decreased with the number of dye compounds present. It is likely that multiple models will be required in reality, and the complexity of a COD monitoring system would be greatly reduced if the PLSR model is used because it can include several dyes. UV-Vis spectroscopy with PLSR successfully enhanced the performance of COD prediction for dye wastewater and showed good potential for application in on-line water quality monitoring.

Keywords: *dye wastewater, UV-Vis spectrophotometer, chemical oxygen demand, partial least squares regression.*

ИСПОЛЬЗОВАНИЕ ЧАСТИЧНОЙ РЕГРЕССИИ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ КАЛИБРОВКИ СПЕКТРОФОТОМЕТРА, ИЗМЕРЯЮЩЕГО ХИМИЧЕСКУЮ ПОТРЕБНОСТЬ В КИСЛОРОДЕ СТОЧНЫХ ВОД ПРИ ПРОИЗВОДСТВЕ КРАСИТЕЛЕЙ

W. Mai, J.-F. Zhang, X.-M. Zhao*, Z. Li, Z.-W. Xu

УДК 535.853.673

Школа текстиля, Тяньцзиньский политехнический университет,
Тяньцзинь, 300387, Китай; e-mail: tjpu_zhao@163.com

(Поступила 8 июля 2016)

Сточные воды, образующиеся при производстве красителей, обычно анализируются стандартным методом измерения химической потребности в кислороде (ХПК) или одноволновым спектроскопическим методом. Чтобы преодолеть недостатки этих методов, в настоящей работе спектроскопия в УФ и видимой областях объединена с регрессией по основному компоненту и частичной регрессией методом наименьших квадратов (ЧРМНК). В отличие от стандартного метода этот метод не требует предварительной подготовки образцов. Показано, что модель ЧРМНК обеспечивает высокую точность прогнозирования ХПК со средней относительной погрешностью ~5 % для случая двух красителей. Эта ошибка близка к ошибке стандартного метода. В наших исследованиях точность модели ЧРМНК уменьшалась с числом присутствующих красителей. Вероятно, понадобятся несколько моделей, и сложность системы мониторинга ХПК будет значительно уменьшена, если использовать модель ЧРМНК, поскольку она может включать в себя несколько красителей. Спектроскопия в УФ и видимой областях в сочетании с ЧРМНК повышает эффективность прогнозирования ХПК для сточных вод красильной промышленности и перспективна для применения в мониторинге качества воды в режиме реального времени.

Ключевые слова: *сточные воды красильной промышленности, спектрофотометр для ультрафиолетовой и видимой областей спектра, химическая потребность в кислороде, частичная регрессия на основе метода наименьших квадратов.*

Introduction. Chemical oxygen demand (COD) is an important indicator of organic matter concentration in water quality assessments. Real-time detection of COD can be useful for control and analysis of the dye industry wastewater treatment process [1]. The standard COD test has low error and high precision but it involves a digestion process that uses toxic chemicals (e.g., mercuric salts and dichromate), is time consuming (2–4 h until the result is obtained), and is not applicable to on-line detection [2]. With the development of optical techniques, spectroscopic techniques, including ultraviolet-visible (UV-Vis) spectroscopy, have shown good potential for wastewater COD monitoring [3]. Research has shown that the UV absorbance is related to the COD of water samples [4]. In addition, spectroscopic analysis is rapid, nondestructive, environmentally friendly because it requires no additional chemicals, and can be applied to on-line detection [5].

In many studies, COD has been measured by UV-Vis spectroscopy at 254 nm, which is referred to as a single wavelength method. This method is simple and accurate for water samples that are of uniform composition and relatively constant over time but is unsuitable for water samples where the composition changes greatly with time [6]. Wastewater from industrial dyeing processes is highly colored and is complex in composition. The main ingredients of dye wastewater are dyes, with most being organic compounds, inorganic additives (e.g., NaCl and NaOH), and surfactants [7]. The COD of dye wastewater is mainly affected by the dye concentration [8]. There are thousands of compounds that are commonly used as dyes [9], and the dyes and additives used in different dyeing processes vary. Consequently, the components in wastewater can also display large differences, even after wastewater treatment, and thus the UV-Vis spectra can also vary [10]. Therefore, especially for the purpose of on-line effluent monitoring, new methods, such as those using mathematical modeling, are required for measurement of COD in dye wastewater. Full spectrum analysis combined with chemometrics, including principal component regression (PCR) [11], partial least squares regression (PLSR) [2, 12] and artificial neural networks (ANN) [13, 14], can effectively improve the precision of analytical prediction. Among these methods, ANN provides strong approximation ability, but a number of experiments are necessary to determine the appropriate neural network model and parameter settings, and its effectiveness on application depends on the user's experience. Therefore, ANN is not suitable for real-time spectral analysis. PCR and PLSR, which are widely utilized for spectral analysis of mixtures, can also greatly compress high-dimensional data and effectively remove multicollinearity. The aim of the present study was to develop a simplified model for dye wastewater using an aqueous dye solution. Models were constructed using both PCR and PLSR. The feasibility and the prediction accuracy of each method for measurement of COD in dye wastewater were evaluated.

Experimental. *Materials and instruments.* Two pure dye solutions (C.I. Direct Red 23 and C.I. Direct Yellow 11, Li'ang International Trading Co., Ltd., Tianjin, China) and a mixed solution of the two dyes (mass ratio of 1:1) were prepared in ultrapure water ($\rho > 18$ MΩ cm at 25°C). These solutions were used to simulate treated dye wastewater. A total of 45 solutions were prepared, with 15 different concentrations for each dye and the dye mixture. All solutions were kept at 4°C before analysis. UV-Vis spectra were measured using a UV-Vis spectrophotometer that was developed in our laboratory for the purpose of this investigation. This instrument had a xenon lamp as the light source, recorded light attenuation in the wavelength region between 200 and 800 nm, and communicated the results in real-time via a fiber optic probe.

Data collection. Data communication was via a USB interface. The band with $\lambda = 200$–240 nm was not suitable for modeling because there was strong short wavelength absorption and interference in this region. In solutions of appropriate dye concentrations to represent dye wastewater, the intensity of the color is relatively strong. Consequently, the absorbance in the visible wavelength region ($\lambda > 400$ nm) for these solutions will be outside the scope of the spectrophotometer. Therefore, COD was defined as the oxygen equivalents of organic dyes that could be detected in the UV range. For this reason, the band from $\lambda = 240$–400 nm was used for the mathematic modeling. Spectra were measured in this region for dye solutions of different concentrations (Fig. 1).

The COD results for the dye solutions were obtained according to a Chinese national standard method (Chinese Nation Standard "GB/T 11914-89 Water quality – Determination of the chemical oxygen demand – Dichromate method"). Taking into account possible interactions between different components in the dye mixture, we measured the COD and spectra of the dye mixture after 1 and 24 h. The 45 solutions were divided into two sets, with 30 samples (10 samples from each subset) as a training set (Table 1) to build the COD detection model and 15 samples as a validation set to analyze the error of the model.

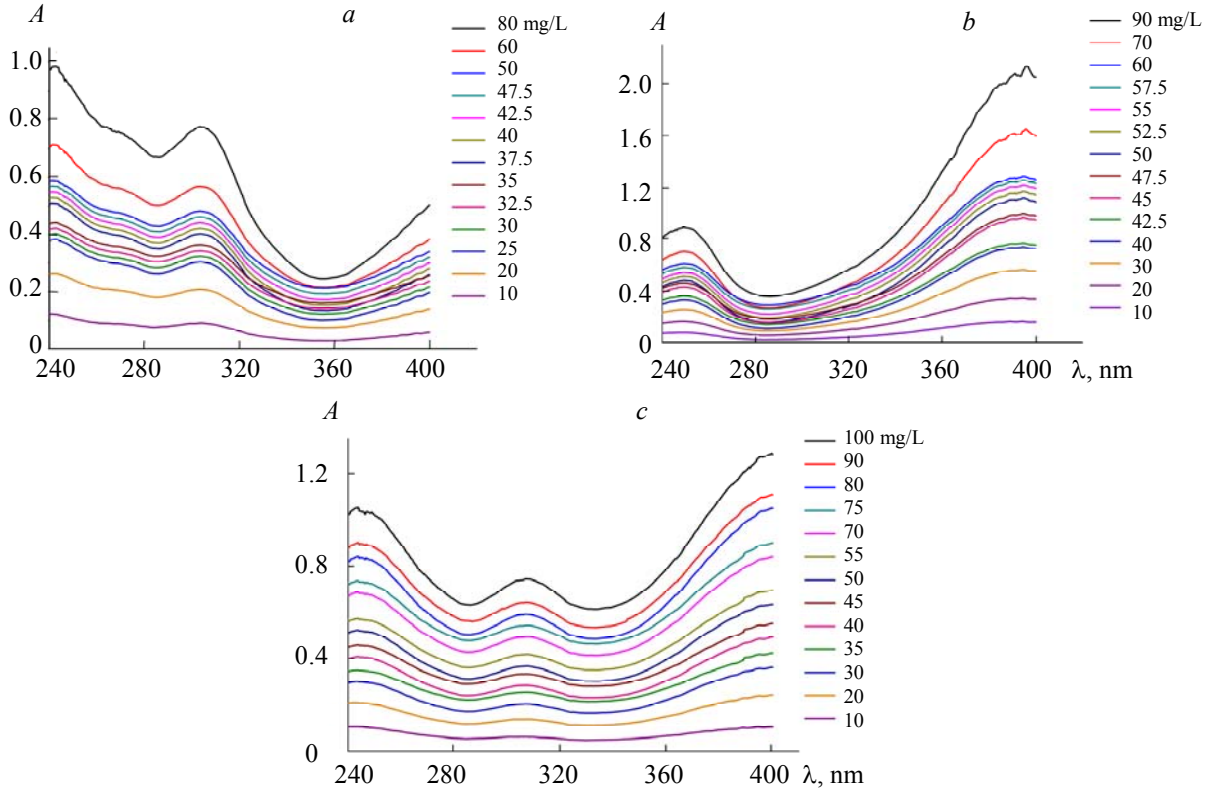

Fig. 1. UV-Vis spectra of dye solutions at a range of concentrations: a) Direct Red 4BS, b) Direct Yellow, c) two dyes (1:1) mixed solution after 1 h.

TABLE 1. COD Values (mg/L) of Each Dye Obtained by the Standard Method

No.	COD	No.	COD	No.	COD[1]	COD[2]
R1	52.5	Y1	70.0	M1	99.0	103.9
R2	42.0	Y2	56.0	M2	83.4	87.6
R3	37.8	Y3	44.8	M3	49.5	52.0
R4	21.0	Y4	28.0	M4	39.6	41.6
R5	16.8	Y5	22.4	M5	29.7	31.2
R6	12.6	Y6	16.8	M6	25.7	26.6
R7	10.5	Y7	14.0	M7	20.8	21.8
R8	9.4	Y8	11.2	M8	16.4	17.2
R9	7.3	Y9	7.0	M9	9.9	10.4
R10	4.2	Y10	5.6	M10	6.9	7.2

N o t e. R is C.I. Direct Red 23, Y is C.I. Direct Yellow 11, M is Mixed dyes (R/Y).
[1] 1 h after mixing.
[2] 24 h after mixing.

Performance evaluation. The precision of the model was assessed by leave-one-out cross-validation (LOOCV). Among n samples, LOOCV uses one randomly selected sample as the validation set and the remaining n–1 samples as the training set to build the model. This is repeated until all the samples are validated. The coefficients of determination (R^2) and prediction residual error sum of squares (PRESS) are then calculated [11].

Descriptions of the algorithms. There are two steps in PCR: principal component analysis (PCA) and multiple linear regression (MLR). PCA is a statistical method that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal com-

ponent has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. In the second step (MLR), important principal components are used to replace the original variables as the variables of MLR [11].

PLSR is a multivariate analysis method based on PCA. In PCA, the only variable (the X matrix) is decomposed orthogonally. In PLSR, both the X and Y matrices are decomposed at the same time. That is, a PLSR model will try to find the multi-dimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. The precision of the PLSR model is often higher than that for PCR, because PLSR can remove the noise in X and Y [2, 12].

Results and discussion. *Single wavelength method.* Generally, in application of the Beer–Lambert law, curve fitting should be based on the maximum absorption peak [15]. However, for the different solutions involved in this case, the wavelength of maximum absorption was different. In this investigation, the wavelength $\lambda = 254$ nm was chosen as a suitable standard with using the xenon lamp in the spectrophotometer. According to the Beer–Lambert law, for the three types of solutions (red dye, yellow dye, and the mixture), the three regression equations and their R^2 were obtained as follows:

$$y = 0.0682x - 0.1202, \quad R^2 = 0.9752,$$

$$y = 0.0342x - 0.0453, \quad R^2 = 0.9965,$$

$$y = 0.0427x - 0.0257, \quad R^2 = 0.9993.$$

Excellent fit (minimum R^2 of 0.9752) was obtained. The three equations were very different, which indicates that one model cannot be used for multiple dye compounds using the single wavelength method.

Comparison of the PLSR and PCR models for two pure dyes. Twenty samples of the pure dye solutions were chosen as the training set to build PCR and PLSR model for the pure dyes. For the PCR algorithm, more than 90% of the variance in the COD measurement was explained by the first two principal components (Fig. 2). The first two principal components were then used as the variables for MLR. For the PLSR algorithm, the factor (principal components) was calculated directly using Minitab 17.0. In the final regression equation, the number of variables was two for each algorithm (PCR and PLSR). Comparison of the prediction performance by PCR (R^2 = 0.989, PRESS = 5.07) and PLSR (R^2 = 0.994, PRESS = 2.61) for the training set showed that the PLSR model exhibited much better prediction performance. This could be attributed to dimension reduction of the spectral matrix (X) based on a comprehensive consideration of the interaction of both COD matrices (Y and X) in PLSR.

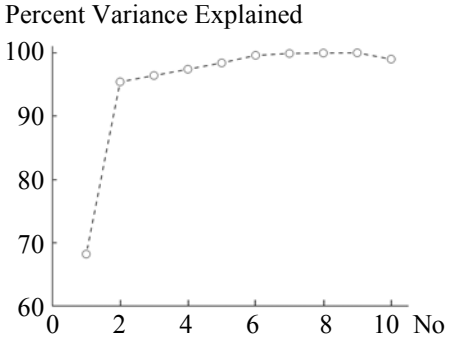

Fig. 2. The percent of the measurement variance explained by the PCR as a function of the number of components used in the PCR model.

One regression equation was established by PLSR or PCR for solutions of the two dyes, and the R^2 value was also close to one, which is the same as that for the single wavelength method. Thus, the PLSR and PCR models could potentially expand the scope of the application of COD analysis. To verify the actual prediction error of the PCR and PLSR models, red and yellow dye solutions were randomly selected from the validation set. The mean relative error of the PCR algorithm (6.61%) was slightly higher than that of the standard method (<4% [16]), and the PLSR algorithm (4.02%) (Table 2).

TABLE 2. Measured and Predicted COD (mg/L) Comparison for the Solutions of Two Dyes

No.	Measured	Predicted	
		PCR	PLSR
1	6.1	6.85	6.72
2	8.3	9.32	8.52
3	13.6	14.08	13.95
4	25	26.2	26.3
5	36.1	37.58	36.65
6	54.2	55.64	55.31
$\bar{\delta}$		6.61%	4.02%

The effect of mixtures of dyes on the precision of prediction. One hour after preparation of the mixed dye solution, the spectrum of the solution was similar to that expected for summation of the spectra of the pure red and yellow dyes in accordance with the Beer–Lambert law (Fig. 3). However, as the time after mixing increased, the absorbance increased and the wavelength of the maximum absorption peak changed slightly (Fig. 3). After 24 h, the COD prediction result from the PLSR model ($\bar{\delta}$ = 24.01%) was far higher than that after 1 h (Table 3). This was probably because interactions occurred between the dyes, which meant that the absorbance was no longer simply additive. Because the original PLSR model does not contain the spectral data resulting from this interaction, the prediction error is larger. According to the PLSR modeling process, a corrected PLSR model was established using the three dye solutions (i.e., two pure dye solutions and the dye mixture). The prediction accuracy of the corrected model was larger than that of the original (Tables 3 and 4). When the PLSR model contains the target dye sample, the relative error is greatly reduced. The LOOCV indicated that the prediction accuracy of the original PLSR model was better than that of the corrected model (original model R^2 = 0.951, PRESS = 10.1; corrected model R^2 = 0.994, PRESS = 2.6). The corrected model contains more dye compounds than the original. Consequently, the prediction accuracy of the PLSR model decreased as the number of compounds in the target sample increased.

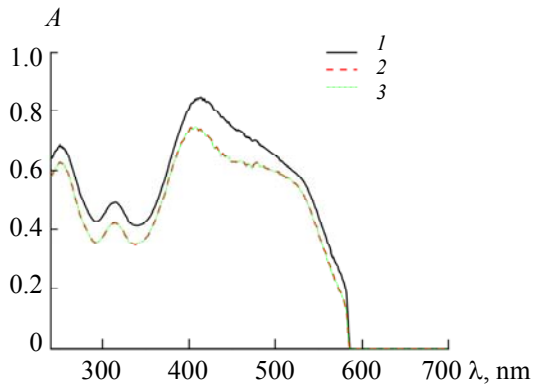

Fig. 3. Spectra of dye mixtures (30 mg/L) over time: 1 – mixed 24 h, 2 – mixed 1 h, 3 – Red+Yellow

TABLE 3. COD Comparison for the Two Dyes at Different Times after Mixing (mg/L)

No.	Measured	Predicted	Measured	Predicted
	1 h after mixing		24 h after mixing	
1	24.2	25.05	29.5	34.87
2	19.4	20.26	23.5	28.69
3	14.6	15.25	18.2	22.02
4	7.2	7.55	9.1	12.06
5	4.8	5.06	6.0	7.59
$\bar{\delta}$	24.01%			4.54%

TABLE 4. Results for the PLSR Model of "Three" Dyes Mixed 1 h

No.	Measured, mg/L	Precdict, mg/L
1	29.5	31.34
2	23.5	24.47
3	18.2	19.25
4	9.1	9.77
5	6.0	6.34
$\bar{\delta}$ 5.83%		

Results for validation for dye solutions. The reliability of each PLSR model (original and corrected) was verified using the prediction set (Table 5). The mean relative error of the corrected PLSR model (6.45%) was much smaller than that of the original PLSR model (12.17%). This shows that as much UV spectral data as possible should be used to establish the PLSR model to reduce the total prediction error.

The relative error in the prediction results was generally high for lower COD concentrations (Table 5). By contrast, when the standard method is applied to samples with low COD, the error is relatively large.

TABLE 5. Results of the Validation Set (mg/L)

No.	Measured	PLSR original model	δ, %	PLSR corrected model	δ, %
1	6.1	6.62	8.56%	6.69	9.68%
2	8.3	8.84	6.52%	8.92	7.52%
3	13.6	14.77	8.60%	14.93	9.78%
4	25	26.06	4.23%	26.58	6.32%
5	36.1	37.99	5.23%	38.07	5.45%
6	54.2	56.08	3.47%	55.95	3.22%
7	86.3	91.19	5.67%	90.03	4.32%
8	17.1	18.73	9.56%	18.60	8.75%
9	69.2	73.26	5.86%	73.97	6.89%
10	45.6	47.65	4.50%	48.17	5.64%
11	29.5	34.87	18.20%	31.34	6.24%
12	23.5	28.69	22.09%	24.47	4.13%
13	18.2	22.02	20.99%	19.25	5.77%
14	9.1	12.06	32.53%	9.77	7.36%
15	6	7.59	26.50%	6.34	5.67%
			$\bar{\delta}$ = 12.17%		$\bar{\delta}$ = 6.45%

Therefore, the use of UV spectral methods combined with PLS for online detection of COD is suitable for solutions with high COD. However, the detection range for COD should be studied further.

Conclusion. The single wavelength method and the PCR/PLSR method described in this paper use UV-Vis spectroscopy to measure COD directly, without the need for a chemical digestion process as is required for standard COD measurements. These methods can overcome many disadvantages of the standard method and can provide fast and on-line COD detection.

The single wavelength method cannot be applied to wastewater with a complex composition (e.g., a solution containing multiple dyes). Consequently, for each different dye, a new model must be established. By contrast, the PCR or PLSR models developed in this study can be used to measure COD for mixtures of two dyes. The precision of the PLSR model is higher than that of the PCR model. For on-line measurement of COD in dye wastewater, spectral data and COD for pure dyes and dye mixtures measured in advance can be used to construct a stable PLSR model to provide reliable COD results.

Acknowledgment. This work was financially supported by the National Natural Science Foundation of China (No. 11575126).

REFERENCES

1. S. K. A. Solmaz, A. Birgül, G. E. Üstün, T. Yonar, *Color. Technol*., **122**, 102–109 (2006).
2. G. Langergraber, N. Fleischmann, F. Hofstadter, *Water Sci. Technol*., **47**, 63–71 (2003).
3. F. Gul, A. M. Khan, S. S. Shah, M. F. Nazar, *Color. Technol*., **126**, 109–113 (2010).
4. H. Cao, W. Qu, X. Yang, *Anal. Methods*, **6**, 3799-3803 (2015).
5. G. Langergraber, N. Fleischmann, F. Hofstaedter, A. Weingartner, *Water Sci. Technol*., **49**, 9–14 (2004).
6. N.M. Mahmoodi, R. Salehi, M. Arami, *Desalination*, **272**, 187–195 (2011).
7. D. T. Phong, N. V. Thong, *Color. Technol*., **124**, 331–340 (2008).
8. N. Chu, S. H. Fan, *Spectrochim. Acta. A*, **74**, 1173–1181 (2009).
9. B. Shyla, Mahadevaiah, G. Nagendrappa, *Spectrochim. Acta*, **78**, 497–502 (2011).
10. L. Kröckel, G. Schwotzer, H. Lehmann, T. Wieduwilt, *Water Res*., **45**, 1423–1431 (2011).
11. C. P. Stemmet, J. C. Schouten, T. A. Nijhuis, *Chem. Eng. Sci*., **65**, 267–272 (2010).
12. A. Torres, J. L. Bertrand-Krajewski, *Water Sci. Technol*., **57**, 581–588 (2008).
13. M. V. Storey, B. V. D. Gaag, B. P. Burns, *Water Res.*, **45**,741–747 (2011).
14. P. Vanloot, C. Branger, A. Margaillan, C. Brach-Papa, J. L. Boudenne, B. Coulomb, *Anal. Bioanal. Chem*., **389**, 1595–1602 (2007).
15. A. Charef, A. Ghauch, P. Baussan, M. Martin-Bouyer, *Measurement*, **28**, 219–224 (2000).
16. A. Niazi, T. Momeni-Isfahani, Z. Ahmari, *J. Hazard. Mater*., **165**, 1200–1203 (2009).