DETERMINING THE CONCENTRATION OF POLYCYCLIC AROMATIC HYDROCARBONS IN WATER USING SURFACE ENHANCED RAMAN SPECTROSCOPY AND KERNEL PRINCIPAL COMPONENTS ANALYSIS COMBINED WITH SUPPORT VECTOR REGRESSION**

C. Jian*, J. Boyan, Zh. Ying, W. Zhenyu

School of Software and Microelectronics at Peking University,
Beijing, 102600, China; e-mail: caojian@ss.pku.edu.cn, boyan1212@qq.com,
351687656@qq.com, zw@ss.pku.edu.cn

Determining the concentration of polycyclic aromatic hydrocarbons (PAHs) in water is vital for reducing negative effects on human health, such as cancer and malformation. This study proposed an alternative analytical method based on surface enhanced Raman spectroscopy and kernel principal components analysis combined with support vector regression (SVR) for the determination of PAH concentration in water. For this, a dataset containing 300 Raman spectra of polycyclic aromatic hydrocarbon mixtures was made using naphthalene (NAP), pyrene (PYR), and phenanthrene (PHE) with concentrations ranging from 0 to 1000 ppb. In order to improve the effect of the model detection, different pre-processed methods were applied: normalization, multiplicative scatter correction, detrending, standart normal variate transformation, and Savitzky–Golay smoothing. For comparison, partial least squares (PLS) and SVR with the polynomial-kernel were also used. The pre-processing method with the best prediction effect was SNV for all the three substances. For NAP, the optimal correlation coefficient of cross-validation (R_{cv}), correlation coefficient of prediction (R_{pred}), RMSECV, RMSEP, and RPD are 0.90, 0.937, 138.9, 117.4, and 2.9 ppb, respectively, while for PYR the optimal R_{cv}, R_{pred}, RMSECV, RMSEP, and RPD are respectively 0.881, 0.897, 152.3, 142.8, and 2.3 ppb. For PHE, the optimal R_{cv}, R_{pred}, RMSECV, RMSEP, and RPD are 0.980, 0.982, 64.5, 62.9, and 5.3 ppb, respectively. This study provides a new method with a better prediction effect for quantitative analysis of low concentrations of polycyclic aromatic hydrocarbons in water by using surface enhanced Raman spectroscopy.

Keywords: *polycyclic aromatic hydrocarbons, surface enhanced Raman spectroscopy, kernel principal components analysis, kernel function, support vector regression.*

ОПРЕДЕЛЕНИЕ КОНЦЕНТРАЦИИ ПОЛИАРОМАТИЧЕСКИХ УГЛЕВОДОРОДОВ В ВОДЕ С ИСПОЛЬЗОВАНИЕМ ПОВЕРХНОСТНО-УСИЛЕННОЙ СПЕКТРОСКОПИИ КОМБИНАЦИОННОГО РАССЕЯНИЯ В СОЧЕТАНИИ С АНАЛИЗОМ ОСНОВНЫХ КОМПОНЕНТОВ ЯДРА И РЕГРЕССИЕЙ ОПОРНЫХ ВЕКТОРОВ

C. Jian*, J. Boyan, Zh. Ying, W. Zhenyu

УДК 535.375.5

Школа программного обеспечения и микроэлектроники Пекинского университета,
102600, Пекин, Китай; e-mail: caojian@ss.pku.edu.cn, boyan1212@qq.com,
351687656@qq.com, zw@ss.pku.edu.cn

(Поступила 12 августа 2019)

Для определения концентрации полициклических ароматических углеводородов в воде предложен альтернативный аналитический метод, основанный на использовании поверхностно-усиленной КР-спектроскопии и анализа основных компонентов ядра в сочетании с методом регрессии опорных

** Full text is published in JAS V. 88, No. 1 (http://springer.com/journal/10812) and in electronic version of ZhPS V. 88, No. 1 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

векторов. Составлен набор данных, содержащий 300 КР-спектров полициклических смесей ароматических углеводородов, содержащих нафталин (NAP), пирен (PYR) и фенантрен с концентрациями 0–1000 ppb. Для улучшения эффекта обнаружения применены методы предварительной обработки данных: нормализация, мультипликативная коррекция рассеяния, детрендирование, стандартное преобразование с нормальным отклонением и сглаживание Савицкого–Голея. Для сравнения использованы методы частичных наименьших квадратов и регрессии опорных векторов с полиномиальным ядром. Метод предварительной обработки с наилучшим прогнозирующим эффектом для трех веществ – преобразование с нормальным отклонением. Для NAP оптимальный коэффициент корреляции перекрестной проверки (R_{cv}), коэффициент корреляции предсказания (R_{pred}), RMSECV, RMSEP и RPD, соответственно, 0.90, 0.937, 138.9, 117.4 и 2.9 ppb, для PYR — R_{cv}, R_{pred}, RMSECV, RMSEP и RPD 0.881, 0.897, 152.3, 142.8 и 2.3 ppb соответственно, для PHE — R_{cv}, R_{pred}, RMSECV, RMSEP и RPD 0.980, 0.982, 64.5, 62.9 и 5.3 ppb соответственно.

Ключевые слова: *полициклические ароматические углеводороды, поверхностно-усиленная КР-спектроскопия, анализ основных компонентов ядра, функция ядра, регрессия опорных векторов.*

Introduction. Polycyclic aromatic hydrocarbons (PAHs) are a kind of aromatic compounds with two or more benzene rings, existing widely as persistent organic pollutants in the environment. Most of PAHs generated in industrial processes are released into the air directly, and a small amount of them gets into the soil and water. PAHs in the soil flow easily into rivers, groundwaters, etc., which has a considerable impact on the inland water and marine environment. Because of the high molecular weight, the volatility of most PAHs and their solubility in water are very low, so they easily get into water by adsorbing on particles and depositing on the bottom. PAHs are hardly degradable naturally. Owing to this, their content in water is much lower than that in the sediment, thus making it difficult to measure. PAHs in water may eventually accumulate in humans by biological accumulation in the food chain. However, PAHs have strong carcinogenicity and teratogenicity; therefore, it is important to detect their concentration [1].

At present, the most commonly used laboratory instruments for detecting organic pollutants are mainly the gas chromatograph-mass spectrometer and the high-performance liquid chromatograph, as they have high detection precision and stability [2, 3]. However, they require tedious chemical preprocessing to be conducted on the samples, and, besides, there exist other problems, such as long detection time, high consumption of the extracting solvent, and high detection cost. In recent years, surface enhanced Raman spectroscopy (SERS) [4, 5] has emerged as a trace analysis method that may be used for detecting low concentrations of organic compounds, rapidly analyzing organic pollutants in a complex system, and even realizing single molecule detection [6–8]. This method can be used for enhancing the Raman scattering spectrum signal of analyte molecules and studying the Raman scattering spectrum of the substance being detected based on the local surface plasma resonance effect of metal nanoparticles to realize the qualitative and quantitative analysis of the targeted substance. Compared with traditional methods for detecting organic compounds, the SERS method has the following characteristics: (1) fast detection speed (only tens of seconds needed from laser emission to signal collection); (2) non-invasiveness; (3) high sensitivity; (4) simultaneous analysis and detection of multiple organic pollutants.

For the analysis of spectroscopic data, the univariate data analysis method cannot always achieve desirable results, because it can only handle one variable at a time, while none of the spectroscopic measurements depend on a single variable [9]. In order to utilize the complete information of complex spectra, multivariate analysis able to process multiple variables at the same time is needed, such as principal component analysis (PCA), linear discriminant analysis (LDA), multiple linear regression (MLR), partial least squares regression (PLS), and support vector regression (SVR) [10–13]. The multivariate analysis method pays more attention to the statistical relationship between each variable.

The largest influence factor of multivariate data analysis for processing spectroscopic data is the random noise and baseline drift problem in the spectrum; thus, the spectrum is processed by utilizing the preprocessing technology before building the model [14]. The most frequently used pre-processing techniques are normalization, detrending (DT), Savitzky–Golay (SG) smoothing, multiplicative scatter correction (MSC), and standard normal variate transformation (SNV). These methods can remove data that have adverse effects on the concentration prediction in spectra and improve the performance of the model [15, 16].

In addition, due to the complexity and diversity of data in SERS, especially the spectrum of mixed substances, there are strong interactions between substances, so the actual spectroscopic data are always nonlinear. For basic PCA and PLS, only linear features can be extracted, so their precision is general in such ques-

tions with strong nonlinear features [17–19]. Meanwhile, SVR is proabably the best statistical analysis method for predicting regression problems with a small sample size, nonlinearity, and a high-dimensional data space [20]. Kernel PCA (KPCA), as an improvement on PCA for dealing with the nonlinear problems, solves the problem of PCA not being able to reduce the dimension of nonlinear features by introducing a kernel function into PCA, which is suitable for the analysis and processing of nonlinear spectroscopic data [21].

The aim of the present work is to find an alternative analytical method based on SERS and KPCA combined with SVR for the determination of the polycyclic aromatic hydrocarbon concentration in water. Introducing kernel PCA into SVR allows us to use KPCA in a high-dimensional data space to reduce the dimension of the original data, remove redundancy and noise, and then test the SVR model, which can speed up the training speed of the SVR model and improve the accuracy. For comparison, PLS and SVR with the polynomial-kernel were also used.

Materials and methods. Chloroauric acid hydrated ($HAuCl_4 \cdot 4H_2O$, Au ≥ 47.8%), trisodium citrate (Na_3Cit, ≥ 99.0%), nitric acid (HNO_3, 68.0%), and hydrochloric acid (HCl, 37.0%) were purchased from Sinopharm Chemical Reagent Co., Ltd. Absolute ethyl alcohol (chromatographically pure) and sodium chloride (NaCl, ≥99.5%) were purchased from Sinopharm Chemical Reagent Beijing Co., Ltd. Polycyclic aromatic hydrocarbon solid samples including naphthalene, phenanthrene, and pyrene were purchased from J&K Scientific Ltd. The water used for the experiment was ultra-pure water with a resistivity of 18.2 MΩ/cm. Unless otherwise specified, all the reagents used were analytical reagents (AR) not purified before use.

Gold nanoparticles were prepared using the reduction method with sodium citrate as a reducing agent and chloroauric acid solution as a metallic precursor [22].

The procedure was as follows. First, dissolve 1*g* $HAuCl_4 \cdot 4H_2O$ in 100 mL of water to obtain a chloroauric acid precursor solution with a mass concentration of 1%. Pipet 2.5 mL of the chloroauric acid precursor solution into a round-bottom flask, add 35.0 mL of ultrapure water, and heat it to boiling under stirring. Then quickly add 2.5 mL of the newly prepared sodium citrate solution (1%), after which the solution changes from a pale yellow to a reddish brown, then continue heating to boiling and keep the backflow for 30 min. Finally, cool down the prepared gold nanoparticles to room temperature naturally and store in 4°C refrigerators.

Preparation and SERS detection of the PAHs sample. Accurately weigh appropriate amounts of NAP, PYR, and PHE solids with a 0.1 mg analytical balance and dissolve them with absolute ethyl alcohol to prepare 10 mg/L PAH standard solutions. Then dilute a high concentration of PAH standard solutions to the required low concentration by adopting the stepwise dilution method, the concentration of which is between 0 and 1000 ppb. Thereafter, compound 100 mixed solutions of different concentrations by mixing the low concentration solutions of three substances. Then take 400 μL of the above prepared gold nanoparticles and 100 μL of the PAH solution, put them into a brown gas phase vial, and mix them completely. Then add an appropriate volume of the 1 mol/L NaCl solution, mix well, and let it stand for 5 min, then leave it for testing. Three samples were prepared for each concentration of the mixed solution.

We put the brown gaseous phase vial filled with a sample to be tested into the sample tank and collected the SERS signal with a QE Pro 65000 Raman spectrometer from Ocean Optics, Inc. The excitation light source was a 785 nm laser, and the power was set to 150 mW. SERS spectrograms of all samples were the Raman signal spectrum under average conditions of 5 s integration time and three measurements. All measurements were carried out at room temperature (23±1°C).

Chemometric analysis. Effective wavelength data in the wavelength range 350.8–1801.5 cm^{-1} in each spectroscopic data were intercepted to establish a dataset in matrix form, with each row representing a spectrum of mixtures and each column representing the Raman intensity of each wavelength in this spectrum. The first three columns of the dataset were respectively the true concentrations of NAP, PYR, and PHE of each data, regarded as the labels of model training and the answers to prediction. These 300 pieces of data were randomly divided into the training set including 240 pieces of data and the validation set including 60 pieces of data [23]. The reason for this division is that, in general, the more data in the training set, the better the model effect, and a large amount of data can reduce the over-fitting situation. At the same time, the validation set also needs some to ensure a good test effect on the model, so 240 data were divided into training sets, and 60 data were divided into test sets.

As there were noise and baseline drifts in spectroscopic data, different pre-processing methods were used to solve these problems, including normalization, DT, SG, filter based derivatives, MSC, and SNV. Then three chemometrics were used to build regression models: PLS, SVR with the polynomial-kernel, and KPCA combined with linear-kernel SVR. The optimal degree and the most appropriate parameter *C* of SVR

with the polynomial-kernel were found out with the grid-search method. In order to endow the training set with the model assessment ability, 5-fold cross-validation was used to obtain R of the cross-validation and RMSECV, which replaces the R and RMSE calculated by directly using all the data in the training set. Besides, and RMSEP in the validation set were used to assess the prediction effect of the model under the new data [24]. RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{n}}. \quad (1)$$

Moreover, the coefficient of correlation is

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}, \quad (2)$$

where n is the number of samples, y_i is the experimentally measured reference result for the sample i, and $\hat{y}_i$ are the estimated results of the model for the corresponding sample i (Eq. (1)). In Eq. (2), $\bar{y}$ is the mean of the reference measurement results for all samples in the training set and validation set. Low RMSE and high R values indicate the best model. The residual predictive deviation (RPD) is defined as the ratio of standard deviation (SD) of the reference data for predicting the RMSE of the validation set. For the performance ability of models, RPD>2 indicates a model with a good prediction ability; 1.5 < RPD < 2 is an intermediate model needing some improvement, and an RPD < 1.5 indicates that the model has a poor prediction ability. The dataset was made with Microsoft Office Excel 2016, and the algorithms were written with Python3.7.

Results and discussion. *Preprocessing procedure*. The spectrograms of such three single substances as NAP, PYR, and PHE with a concentration of 600 ppb respectively are shown in Fig. 1. As can be seen, the characteristic peaks of NAP are 508 cm^{-1} (C-C stretch and C-C-C bending), 758 (C-C stretch and C-C-C bending), 1020 (C-H rock and C-C stretch), 1381 (C-C stretch), and 1569 cm^{-1} (C-C stretch and C-H rock) [25, 26]. The characteristic peaks of PHE are 707 cm^{-1} (C-C-C bending), 1022 (C-C stretch), 1345 (C-C stretch), 1424 (C-H rock and C-C stretch), and 1602 cm^{-1} (C-C stretch) [27]. The characteristic peak with a high intensity of PYR is 588 cm^{-1} (C-C-C bending), and the rest of the characteristic peaks with a weak intensity are 1061, 1237, and 1621 cm^{-1}, which are generated by the C-C stretch [28]. The spectrogram of 100 PAH mixtures with different concentrations is shown in Fig. 2a. In order to attenuate noise in the spectrum and slow down the baseline drift, various pre-processing methods are adopted in this work, as shown in Figs. 2b-f.

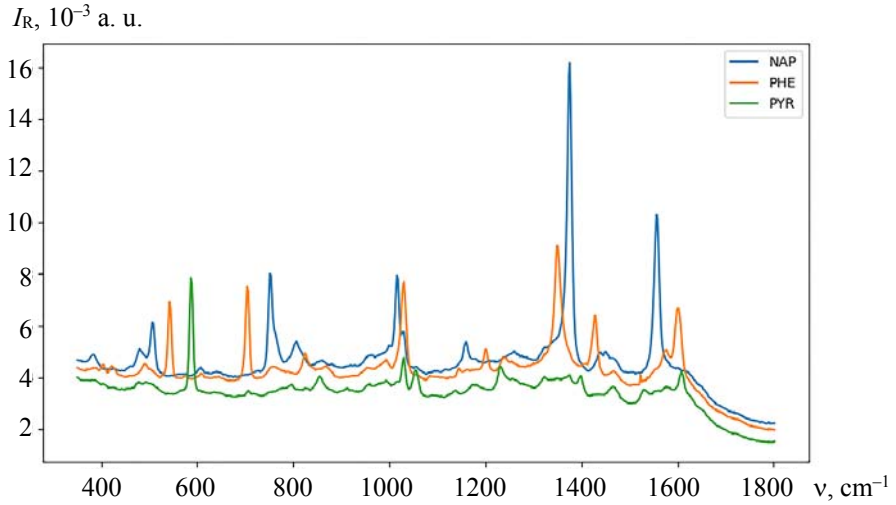

Fig. 1. The spectrograms of NAP, PYR, and PHE with a concentration of 600 ppb.

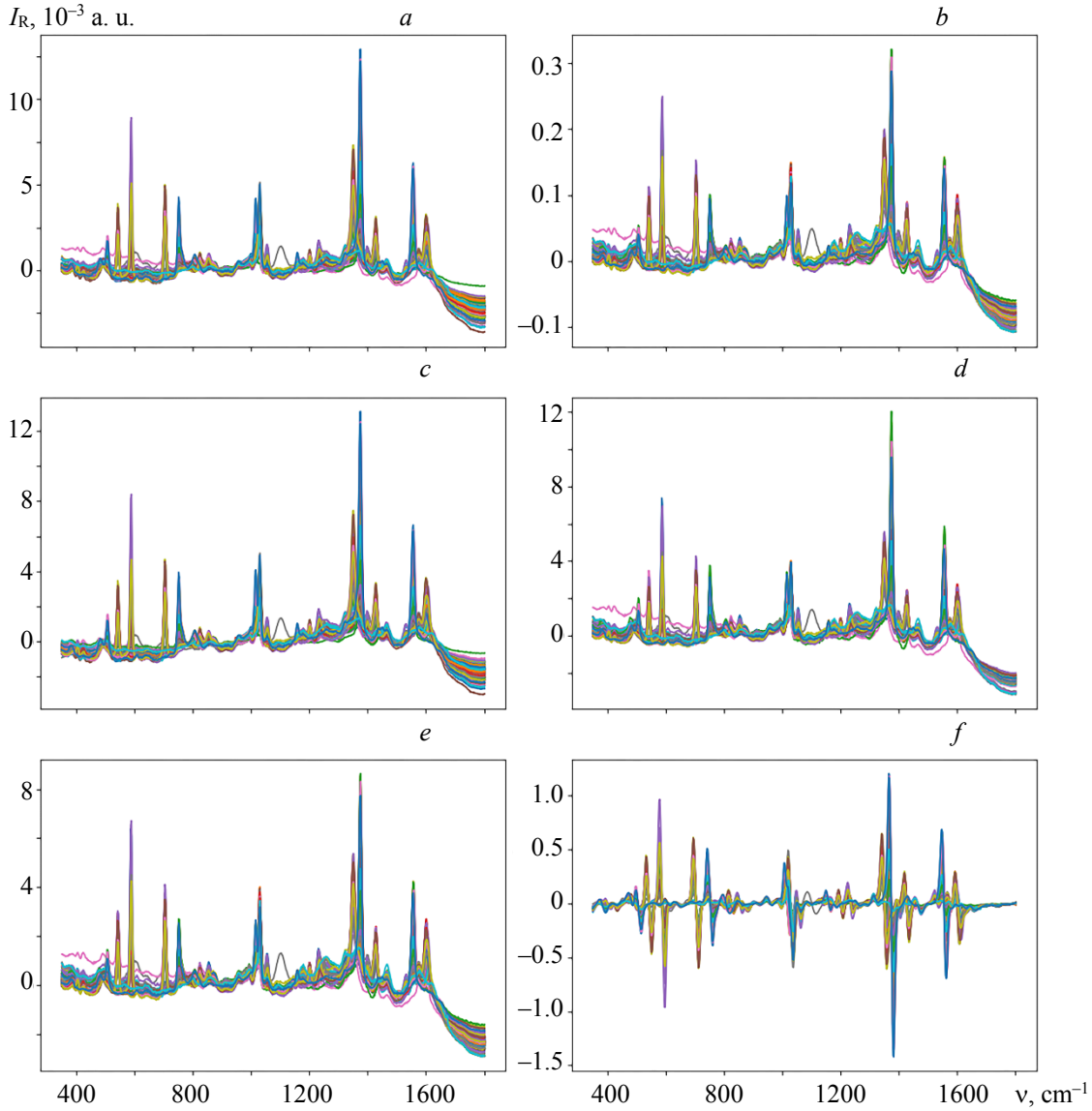

Fig. 2. The spectrogram of 100 PAH mixtures with different concentrations: raw (a) and preprocessed spectra with normalization (b), detrending (c), MSC (d), SNV (e), and first derivative Savitzky–Golay smoothing (f).

Determination of PAHs. The original spectrum and the spectrum processed with the above several preprocessing methods were adopted in this work to build regression models, including PLS, SVR, and KPCA-SVR. The best results of the determination of the concentrations of NAP, PYR, and PHE are summarized in Tables 1–3. The values in the tables are obtained from the optimal parameters in each algorithm. The parameters are the principal components of PLS in Table 1, the degree and *C* of poly-SVR in Table 2, the degree and principal components of poly-KPCA in KPCA-SVR, as well as *C* of linear-SVR in Table 3.

The results of the concentration prediction model of polycyclic aromatic hydrocarbon built with the PLS method are shown in Table 1. For PYR and PHE, the modeling effect is better when using preprocessing data, while for NAP, the modeling effect is better when using original data directly, indicating that the preprocessing of the whole spectrum of the mixture rather than a spectrum of a single substance has a different effect on different substances in the mixture. In building a PLS model of NAP using the original data, R_{cv} of 0.828, RMSECV of 178.9 ppb, R_{pred} of 0.814, RMSEP of 195.4 ppb, and RPD of 1.7 can be achieved. The best pre-processing method for PYR is SNV, achieving R_{cv} of 0.866, RMSECV of 161.3 ppb, R_{pred} of 0.882,

RMSEP of 152.4 ppb, and RPD of 2.1, and the effect is better than that when using the original data. However, the best pre-processing method for PHE is MSC, achieving R_{cv} of 0.967, RMSECV of 82.0 ppb, R_{pred} of 0.969, RMSEP of 81.7 ppb, and RPD of 4.0.

SVR with the polynomial-kernel is a nonlinear regression method, and its best prediction effect is shown in Table 2. Compared with the PLS method, the prediction effects of three substances are all improved, and the pre-processing method with the best overall effect is SNV. For NAP, the kernel function with the best effect is the 21-degree polynomial-kernel, achieving R_{cv} of 0.898, RMSECV of 140.2 ppb, R_{pred} of 0.934, RMSEP of 1 20.4 ppb, and RPD of 2.8. For PYR, it is the 2-degree polynomial-kernel, achieving R_{cv} of 0.880, RMSECV of 153.0 ppb, R_{pred} of 0.897, RMSEP of 142.9 ppb, and RPD of 2.3. For PHE, it is the 3-degree polynomial-kernel, achieving R_{cv} of 0.978, RMSECV of 67.0 ppb, R_{pred} of 0.984, RMSEP of 59.8 ppb, and RPD of 5.6. Compared with the PLS model, the effects of the three pollutants are all improved, among which NAP has the largest magnitude of improvement, and R_{cv} is increased from 0.828 to 0.898, indicating that the nonlinearity of NAP in the spectrum is strong, and accurate regression modeling cannot be conducted with a general linear method. In addition, as the intensity of the spectrum is too large to directly use the SVR algorithm, original data are not modeled directly in this work but using various pre-processing methods for comparison.

TABLE 1. Best Results for PLS Modeling with Different Pre-Processing Methods for PAHs Concentration Determination in Water

Pre-processing methods	NAP					PYR					PHE				
	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD
Raw	**0.828**	**178.9**	**0.814**	**195.4**	**1.7**	0.84	174.8	0.79	189.5	1.6	0.956	93.8	0.94	114.1	2.9
Normalization	0.805	189.1	0.805	199.8	1.7	0.862	163.2	0.867	161.4	2.0	0.955	95.1	0.963	88.6	3.7
DT	0.799	191.5	0.802	201.1	1.7	0.84	175.2	0.804	192.5	1.7	0.953	97.7	0.942	110.6	3.0
MSC	0.787	196.5	0.804	200.3	1.7	0.866	161.3	0.854	168.7	1.9	**0.967**	**82.0**	**0.969**	**81.7**	**4.0**
SNV	0.815	184.5	0.805	199.9	1.7	**0.866**	**161.3**	**0.882**	**152.4**	**2.1**	0.966	82.9	0.967	83.9	4.0
SG-1d	0.784	198.0	0.799	202.2	1.7	0.822	183.5	0.801	194.1	1.7	0.955	95.4	0.944	108.8	3.0

N o t e. The best result for each regression modeling (PLS, SVR, or KPCA-SVR) is indicated in bold.

TABLE 2. Best Results for SVR Modeling with Different Pre-Processing Methods for PAHs Concentration Determination in Water

Pre-processing methods	NAP					PYR					PHE				
	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD
Normalization	0.861	162.1	0.914	136.7	2.5	0.853	168.4	0.865	162.5	2.0	0.968	81.2	0.98	66.1	5.0
DT	0.853	166.2	0.899	147.4	2.3	0.862	163.5	0.888	149.1	2.2	0.968	80.2	0.98	65.9	5.0
MSC	0.881	150.5	0.913	137.2	2.5	**0.88**	**153.0**	**0.897**	**142.9**	**2.3**	**0.978**	**67.0**	**0.984**	**59.8**	**5.6**
SNV	**0.898**	**140.2**	**0.934**	**120.4**	**2.8**	**0.88**	**153.0**	**0.897**	**142.9**	**2.3**	**0.978**	**67.0**	**0.984**	**59.8**	**5.6**
SG-1d	0.891	144.9	0.864	169.6	2.0	0.859	165.2	0.872	158.4	2.0	0.978	67.1	0.981	64.8	5.2

N o t e. The best result for each regression modeling (PLS, SVR, or KPCA-SVR) is indicated in bold.

TABLE 3. Best Results for KPCA-SVR Modeling with Different Pre-Processing Methods for PAHs Concentration Determination in Water

Pre-processing methods	NAP					PYR					PHE				
	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD	R_{cv}	RMSECV, ppb	R_{pred}	RMSEP, ppb	RPD
Normalization	0.862	161.4	0.839	183.1	1.8	0.848	171.1	0.849	171.0	1.9	0.972	75.7	0.979	66.9	4.9
DT	0.881	150.9	0.873	163.9	2.0	0.859	165.3	0.88	153.7	2.1	0.969	79.5	0.976	71.4	4.6
MSC	0.878	152.4	0.86	171.7	2.0	0.86	164.4	0.878	154.8	2.1	0.976	70.1	0.981	64.1	5.2
SNV	**0.90**	**138.9**	**0.937**	**117.4**	**2.9**	**0.881**	**152.3**	**0.897**	**142.8**	**2.3**	**0.98**	**64.5**	**0.982**	**62.9**	**5.3**
SG-1d	0.9	138.8	0.889	154.3	2.2	0.857	166.4	0.869	160.2	2.0	0.977	68.9	0.982	63.2	5.3

N o t e. The best result for each regression modeling (PLS, SVR, or KPCA-SVR) is indicated in bold.

The modeling results of introducing the dimensionality reduction algorithm KPCA in the SVR method are shown in Table 3. The regression effects of three substances are all improved. The advantage of KPCA is that the original nonlinear data can be transformed into linear data in a high dimensional space by using polynomial-kernel, and some redundancy and noise in the original data can be eliminated to improve the model effect. Then the accurate regression of the dimension-reduced linear data can be achieved using SVR with the linear-kernel, which can greatly shorten the model training time. The best pre-processing methods for the three substances are all SNV. For NAP, when the principal component number is 143 and the polynomial-kernel is 23-degree, we achieve R_{cv} of 0.90, RMSECV of 138.9 ppb, R_{pred} of 0.937, RMSEP of 117.4 ppb, and RPD of 2.9. For PYR, when the principal components number is 162 and the polynomial-kernel is 3-degree, we achieve R_{cv} of 0.881, RMSECV of 152.3 ppb, R_{pred} of 0.897, RMSEP of 142.8 ppb, and RPD of 2.3. For PHE, the kernel function with the best effect is the 5-degree polynomial-kernel, and the number of its principal components is 101, achieving R_{cv} of 0.980, RMSECV of 64.5 ppb, R_{pred} of 0.982, RMSEP of 62.9 ppb, and RPD of 5.3. Figures 3 show the predicted vs reference plot for both training (yellow and square) and prediction (red and circle) samples of NAP, PYR, and PHE for the best results obtained by the KPCA-SVR model respectively. In Figures 3, the green straight line is the ideal error-free effect, and the blue straight line is the detection effect that the KPCA-SVR model actually achieves. From the evaluation metrics of the model, KPCA-SVR gives some improvement on all the three substances compared with SVR, which provides a better model choice for the detection of PAHs concentration by SERS.

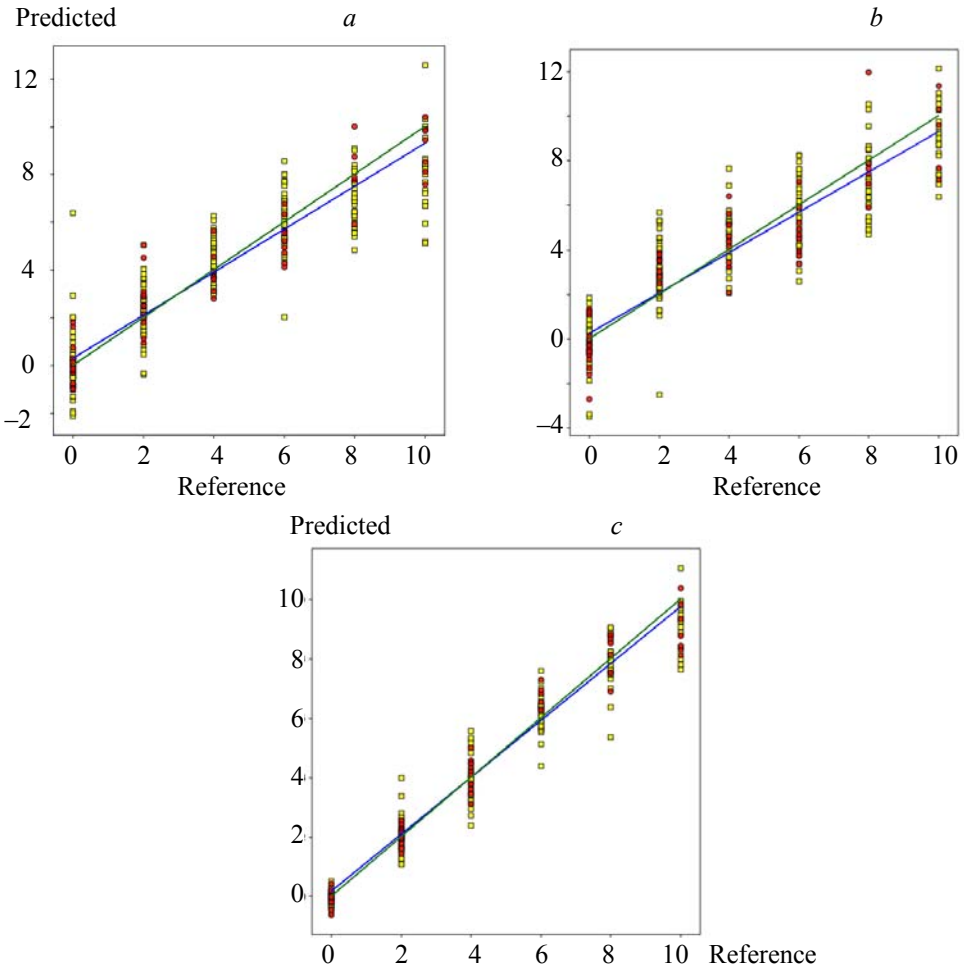

Fig. 3. Predicted vs reference plot for both training (yellow and square) and prediction (red and circle) samples of NAP (a), PYR (b), and PHE (c) for the best results obtained by KPCA-SVR model.

Conclusions. We proposed an alternative analytical method based on SERS and the KPCA-SVR algorithm for the determination of ppb-level PAHs concentrations in water, the results of which show that the model has a good prediction effect. For NAP, the optimal R_{cv}, R_{pred}, RMSECV, RMSEP, and RPD are 0.90, 0.937, 138.9 ppb, 117.4 ppb, 2.9, respectively, while for PYR, the optimal R_{cv}, R_{pred}, RMSECV, RMSEP, and RPD are 0.881, 0.897, 152.3 ppb, 142.8 ppb, and 2.3, respectively. For PHE, the optimal R_{cv}, R_{pred}, RMSECV, RMSEP, and RPD are 0.980, 0.982, 64.5 ppb, 62.9 ppb, and 5.3, respectively. Compared with the commonly used linear spectral processing methods PCA and PLS, the KPCA-SVR algorithm can not only build a good model for linear data but also fit the nonlinear data better, so it can be used as a new chemometric method for the quantitative analysis of spectral data. For the detection of the concentration of polycyclic aromatic hydrocarbons in water, this method can replace the chromatograph-mass spectrograph and other traditional methods, and its detection process is efficient and convenient.

Acknowledgments. The research was funded by the Chinese Academy of Sciences Scientific Equipment Research Project (Grant No. YJKYYQ20170003).

REFERENCES

1. I. Tongo, L. Ezemonye, K. A. Akpeh, *Environ. Monit.*, **189**, No. 6, 247 (2017).
2. N. Xiang, C. Jiang, T. Yang, P. Li, H. Wang, Y. Xie, et al. *Chin. Ecotoxicol. Environ. Saf.*, **152**, No. 8, 15 (2018).
3. M. L. Vestal, *Science*, **226**, No. 4672, 275–281 (1984).
4. B. Sharma, R. R. Frontiera, A.-I. Henry, E. Ringe, R. P. V. Duyne, SERS: Materials, Applications, and the Future. *Mater. Today*, **15**, No. 1-2, 16–25 (2012).
5. J. Kneipp, H. Kneipp, K. Kneipp, *Chem. Soc. Rev.*, **37**, No. 5, 1052 (2008).
6. H.-X. Gu, K. Hu, D.-W. Li, Y.-T. Long, *Analyst*, **141**, No. 14, 4359 (2016).
7. X. Gu, S. Tian, Q. Zhou, J. Adkins, Z. Gu, X Li, et al. *RSC Adv.*, **3**, No. 48, 25989–25996 (2013).
8. D. W. Li, W. L. Zhai, Y. T. Li, Y. T. Long, *Microchim. Acta*, **181**, No. 1-2, 23–43 (2014).
9. R. Gautam, S. Vanga, F. Ariese, S. Umapathy, *EPJ Tech. Instrum.*, **2**, No. 1, 8 (2015).
10. S. Wold, *Principal Component Anal.*, **2**, No. 1, 37–52 (1987).
11. A. C. Kak, A. M. Martínez, *PCA versus LDA*, **23**, No. 3-4, 228–233 (2001).
12. P. Geladi, B. R. Kowalski, *J. Anal. Chim. Acta*, **185**, No. 1, 1–17 (1986).
13. A. J. Smola, B. Schölkopf, *Stat. Comput.*, **14**, No. 3, 199–222 (2004).
14. L. A. Reisner, A. Cao, A. K. Pandya, *Chemometr. Intell. Lab. Syst.*, **105**, No. 1, 83–90 (2011).
15. T. Bocklitz, A. Walter, K. Hartmann, P. Rusch, J. Popp, *Anal. Chim. Acta*, **704**, No. 1, 47–56 (2011).
16. A. Rinnan, F. V. D. Berg, S. B. Engelsen, *Trends Anal. Chem.*, **28**, No. 10 1201–1222 (2009).
17. H. Chen, C. Tan, Z. Lin, T. Wu, *Spectrochim. Acta A: Mol. Biomol. Spectrosc.*, **189**, 183 (2017).
18. H. S. Tapp, M. Defernez, E. K. Kemsley, *J. Agric. Food Chem.*, **51**, No. 21, 6110–6115 (2003).
19. J. K. Holland, E. K. Kemsley, R. H. Wilson, *J. Sci. Food Agric.*, **76**, No. 2, 263–269 (2015).
20. X. L. Yang, Y. F. Li, X. W. Zhang, S. Q. Hu, *Appl. Mech. Mater.*, **494-495**, 964–967 (2014).
21. F. Kuang, W. Xu, S. Zhang, *Appl. Soft Comput. J.*, **18C**, 178–184 (2014).
22. P. C. Lee, D. J. J. Meisel, *J. Phys. Chem.*, **86**, No. 17, 3391–3395 (1982).
23. P. S. Sampaio, A. Soares, A. Castanho, A. S. Almeida, J. Oliveira, C. Brites, *Food Chem.*, **242**, 196–204 (2018).
24. G. Krepper, F. Romeo, D. D. S. Fernandes, P. H. G. Diniz, M. C. U. de Araújo, M. S. Di Nezio, et al. *Spectrochim. Acta A: Mol. Biomol. Spectrosc.*, **189**, 300–306 (2017).
25. O. Frank, J. Jehlika, H. G. M. Edwards, *Spectrochim. Acta A: Mol. Biomol. Spectrosc.*, **68**, No. 4, 1065–1069 (2007).
26. J. Neugebauer, E. J. Baerends, E. V. Efremov, F. Ariese, C. Gooijer, *J. Phys. Chem. A*, **109**, No. 10, 2100–2106 (2005).
27. I. Bandyopadhyay, S. Manogaran, *J. Mol. Struct. THEOCHEM*, **496**, No. 1, 107–119 (2000).
28. A. Bree, R. A. Kydd, T. N. Misra, V. V. B. Vilkos, *Spectrochim. Acta A: Mol. Spectrosc.*, **27**, No. 11, 2315–2332 (1971).