Measurement of Diesel Cetane Number Using Near Infrared Spectra and Multivariate Calibration

Bai-shao Zhan [1,2] and Jian-guo Yang [1*]

[1]College of Mechanical Engineering, Donghua University, Shanghai, 201620;

[2]Taizhou University, Zhejiang taizhou, 318000

Corresponding author:Yang Jianguo, Email:56445627@qq.com

Keywords: NIRS; Diesel Cetane; LS-SVM; LVs

Abstract. Near infrared spectroscopy (NIRS) was investigated for measurement of diesel cetane. Three types of pretreatments including standard normal variate (SNV), multiplicative scattering correction (MSC) and Savitzky-Golay smoothing combining first derivative were adopted to eliminate the system noises and external disturbances. Then, partial least squares (PLS) and least squares-support vector machine (LS-SVM) methods were implemented for calibration models. Simultaneously, the performance of least squares-support vector machine (LS-SVM) method was compared with three kinds of dimension reduction input, including principal components (PCs), latent variables (LVs), and effective wavelengths (EWs). The best predictions were obtained with LS-SVM-LVs model for diesel cetane number (R^2_{Pre}=0.557, RMSEP = 2.169 and residual prediction deviation (RPD) =1.61), which was deemed as good model predictions. It is recommended to adopt LS-SVM-LVs model technique for higher accuracy measurement of the selected diesel cetane number with NIR, in comparison with LS-SVM-PCs and LS-SVM-EWs model techniques.

Introduction

There is a widespread interest to assess diesel parameters with quick and cheap scanning methods instead of using more expensive standard diesel analysis procedures. They are highly sensitive organic components, making their use in the agricultural and environmental sciences particularly appropriate [1-3]. Diesel is one of the main power sources of vehicles. Using inferior or poor quality fuel could cause air pollutants and damage the engines. Rapid analysis of diesel fuel properties is of great importance. Recently, near infrared spectroscopy (NIRS) has become an effective method for rapid and real-time analysis of fuel properties [4-6]. The objectives of this study were (1) to investigate the feasibility of using NIR spectroscopy to predict the diesel cetane; (2) to obtain the optimal preprocessing methods for the diesel cetane; (3) to achieve the best calibration models after the comparison of PLS and LS-SVM models; and (4) three inputs of Least Squares Support Vector Machine(LS-SVM) calibration method are compared for measurement accuracy of diesel cetane number, namely principal components (PCs), latent variables (LVs), and effective wavelengths (EWs).

Materials and Methods

Diesel Samples and Spectra Pretreatment

In the present work, a total of 381 diesel samples were obtained from Taizhou Supervision and Inspection Institute, China. Sample statistics were shown in table 1.

Optical scanning was carried out on diesel samples to build calibration models. A NIRS spectroradiometer (Analytical Spectral Devices, Boulder, USA) covering the spectral range of

750-1550 nm was used to collect absorbance spectra of samples at 2 nm intervals. A spectral acquisition system was build for this investigation, and it consists of a portable spectrometer which has transmitting and receiving fiber bundles connected by a fiber optic probe, a halogen light source, and a computer. During the experiments, the distance of the probe to the Colorimetric plate kept exactly the same. Before collecting the spectral data, the instrument was calibrated by transmission through distilled water according to the manufacturer's instruction, which help to eliminate the influence of the background light. The transmission spectral data were collected at ambient temperature between 23°C and 26°C and in the range from 750 nm to 1550 nm at intervals of 2 nm. Each spectrum was recorded as average of 10 spectra with the integration time of 10 ms, and they were measured with fixed distance.

Table 1. Sample statistics of cross-validation and prediction data sets of samples sets

Property	Number of Samples	Min	Max	Mean	Standard deviation
Cross-validation set					
Cetane number,%	254	20.4	49.5	39.0	7.87
Prediction set					
Cetane number,%	127	22.8	48.8	35.7	7.55

In order to assess the prediction resultsunder different pretreatments , three types of pretreatments including standard normal variate (SNV), multiplicative scattering correction (MSC) and Savitzky-Golay smoothingcombining first derivative were applied[7]. The raw absorbance spectra and preprocessed spectra of diesel are shown in Fig. 1a–d.All pre-processing steps were carried out using the Unscrambler V9.7 (CAMO, PROCESS, AS, OSLO, Norway) software.

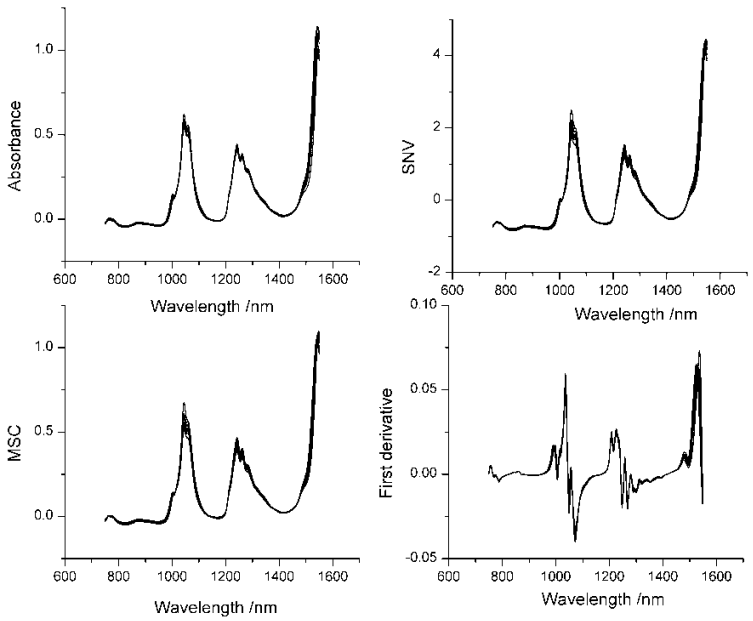

Figure.1. (a) raw spectra and preprocessed spectra by (b) SNV, (c) MSC and (d) first derivative

Partial Least Square

PLS regression could predict Y from X and describe their common structure. In the development of PLS model, full crossvalidation was used to validate its quality and to prevent overfitting of the calibration model, and centering was also adopted for calculating PLS models. The prediction performance was assessed by the samples in the validation set. PLS analysis was also used as a method to extract the latent variables (LVs) of the spectral data. The LVs could be used for further modeling instead of the original spectral data[8]. During PLS analysis, 20 latent variables were used for each property of diesel in the calibration set. Certain LVs were selected as the input data set of the LS-SVM according to their explained variance to Y-variable (chemical concentration). In order to reduce the computation time, certain selected latent variables were used as the input data according to their explained variance from PLS analysis. The LVs could explain most of the spectral variances and represent the main information relating the spectra to the chemical constituents. Another method for the selection of inputs for LS-SVM was based on the regression coefficients obtained by PLS analysis. The regression coefficients were calculated from the spectral data table and the response value Y-variables. It was quite helpful to find which variables were relevant and important for the prediction of Y-variables. The value of the coefficients gave an indication of which variable had an important impact on the response variables (Y). Large absolute regression coefficient values indicate their importance and the significance of the effect on the prediction of the Y-variable. Hence, certain relevant variables, named effective wavelengths(EWs)[9].

Least Squares-support Vector Machine

LS-SVM, a state-of-the-art learning algorithm, has a good theoretical foundation in statistical learning methods. The LS-SVM regression model can be expressed as:

$$y(x) = \sum_{k=1}^{n} \alpha_k K(x, x_k) + b$$

Where K(x, xk) is the kernel function, xk is the input vector, αk is the Lagrange multiplier (also called the support value), and b is the bias term. LS-SVM is capable of dealing with linear and nonlinear multivariate analysis and resolving these problems in a relatively fast way[10-12]. The details of LS-SVM algorithm can be found in the literature.

Prior to developing LS-SVM, one crucial problem needed to be addressed: the optimal input data set. Although the whole spectral wavelength region could be applied as the inputs, the training time using LS-SVM increased with the number of training variables. In this paper, the performance of least squares-support vector machine (LS-SVM) models was compared with three kinds of inputs, including principal components (PCs), latent variables (LVs), and effective wavelengths (EWs).

Results and Discussion

PLS Models

Partial least squares (PLS) models were developed using the preprocessed spectra are shown in Table 2. As can be seen, using the prediction performance evaluation indices (mainly R2 and RMSEP).The optimal determination coefficient (R2), RMSEP and RPD for the validation set were 0.62, 2.13 and 1.64 for diesel cetane number. The reason for the poor performance of standard normal variate (SNV) and multiplicative scattering correction (MSC) might be that these pretreatments introduced

significant noises to the spectra which can be seen in Figure. 1 b, c at the beginning and ending parts of spectra.

Table 2 The prediction results of diesel cetane number in calibration and validation sets by PLS models

Parameters	Pretreatments	Calibration			Validation		
%		R_{Cal}^2	RMSEC	RPD	R_{Pre}^2	RMSEP	RPD
Diesel cetane number	None	0.727	1.8452	1.91	0.624	2.1365	1.64
	SNV	0.552	2.379	1.48	0.501	2.479	1.41
	MSC	0.545	2.380	1.47	0.495	2.477	1.41
	SG+1st derivative	0.612	2.198	1.60	0.546	2.346	1.49

Selection of Inputs of LS-SVM

Several PCs were extracted from the spectra of diesel samples by the PCA model. PCA is often referred to as a data compression technique because it reduces the dimensionality of the data to fewer components that describe a large portion of the variance[13]. The first principal component accounts for the largest variance, while subsequent components account for decreasingly smaller portions. To provide a more general description of the diesel samples, the first three principal component scores (which accounted for 96% of the variance in the data) were clustered as shown in Figure.2. According to the explained variance, three PCs are selected as the inputs of LS-SVM to develop the calibration models.

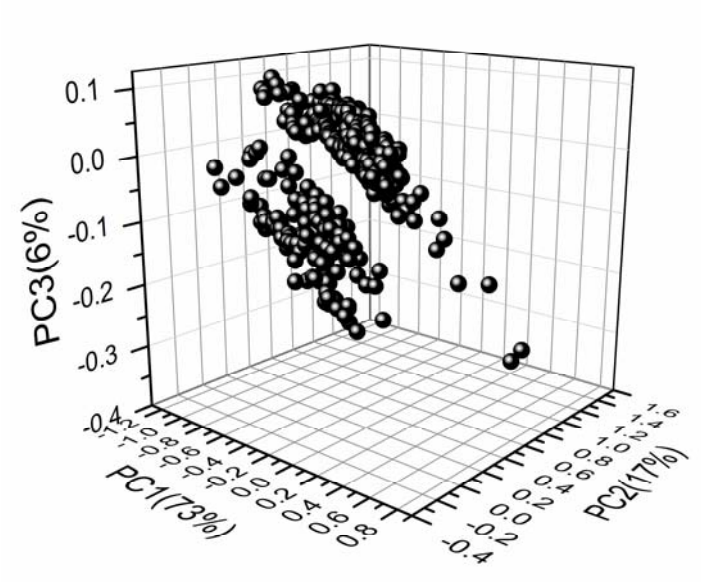

Figure. 2. Scatter plots of the first three principal components scores

The explained variance of the first 15 LVs by the best PLS models are shown in Table 3. The explained variance of the top 14 LVs could explain 98.5%of the total variance for diesel cetane number. The next LV only contributed less than 1% of total variance for diesel cetane number. Therefore, the top 14 LVs were selected as the input for LS-SVM for diesel cetane number.

Table 3.The explained variance of the first 15 LVs for diesel cetane number by the best PLS models

Latent variables (LVs)	1	2	3	4	6	8	10	14	15
Cetane numbers(%)	76.1	81.2	84.4	91.3	95.2	96.5	97.5	98.5	98.6

The regression coefficient of diesel cetane number was shown in Figure.3. The main criteria for selection were that the wavelength should have a large absolute regression coefficient value and should be at specific peaks, and valleys of the regression coefficient curve. An assumption was made that a wavelength with a large absolute regression coefficient value could represent useful information of wavelength at the peaks and valleys. Number of EWs should also be as little as possible and distance of the two EWs should not be too close, so that the model with EWs input can be as simple as possible. Thus, the above selected EWs were also used as inputs for LS-SVM.

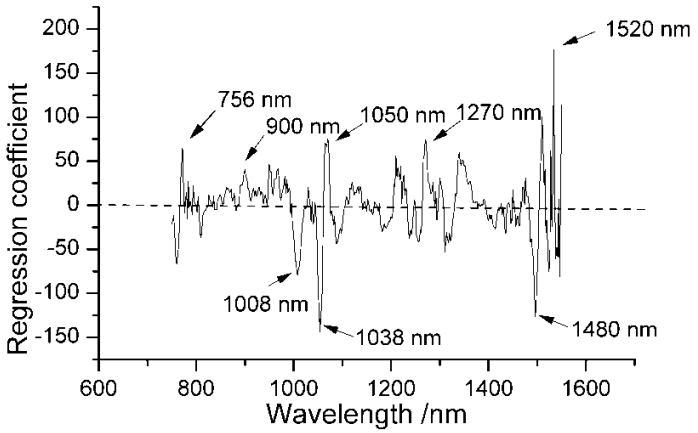

Figure.3. Regression coefficients distribution over the entire wavelength

With a carefully observation, eight wavebands were chosen for prediction of diesel cetane number, which were centered at around 756,900,1008,1038,1050,1270,1480 and 1520 nm, respectively. Figure.3 showed some strong peaks and valleys at certain wavelengths which were thought to be more significant for the prediction of diesel cetane number, such as 1038 and 1048 nm.

LS-SVM Models and Comparison

The aforementioned data of PCs, LVs, and EWs were used as the input matrix to develop LS-SVM models for the prediction of diesel cetane. The performance was confirmed using 127 independent samples in the validation set. The prediction results of calibration and validation set by LS-SVM-PCs (least squares-support vector machine with principal components), LS-SVM-LVs (least squares-support vector machine with latent variables) and LS-SVM-EWs (least squares-support vector machine with effective wavelengths) models are shown in Table 4. The performance of the model was evaluated by the determination coefficient (RPre2) and RMSEP. As shown in Table 4, the he optimal

LS-SVM-LVs models were achieved, and the determination coefficient (RPre2) and RMSEP were 0.6548 and 1.96. Li et al. reported similar results for the prediction of selected diesel cetane number using near infrared spectroscopy[14]. The results indicated that LS-SVM-LVs models performed slightly better than LS-SVM-PCs and LS-SVM-EWs model.

Table 4. Determination coefficient (R2) and root mean square error (RMSE) analyses for data sets

	R2	RMSE
LS-SVM-EWs-Cal	0.5575	2.33
LS-SVM-PCs-Cal	0.53497	2.49
LS-SVM-LVs-Cal	0.65813	1.83
LS-SVM -EWs-Pre	0.5375	2.50
LS-SVM- PCs-Pre	0.52563	2.55
LS-SVM -LVs-Pre	0.6548	1.96

All LS-SVM models performed slightly worse than PLS models and the reason might be that the LS-SVM models used the selected input variables, instead of original variables. Hence, the whole wavelengths used in the PLS model are more correlated with the diesel cetane than occurred with the LS-SVM model, which used the selected inputs to build models. Although the prediction results by LS-SVM and PLS was really similar and comparable in this specific case, the application of LS-SVM supplied a new way for systematic comparison and further potential applications in other fields. The performance of LS-SVM models indicated that PCs, LVs and EWs eigenvectors were very helpful for the dimensional compression and reduction. Seven or eight EWs could represent the most useful information of whole wavelength region 750–1550 nm. This result indicated that use of the PLS model was a powerful selection method for determining the LVs and EWs of chemical compositions. The selected LVs or EWs could be useful in the development of instruments or in situ detection of diesel cetane. The selected LVs or EWs could be useful in the development of instruments or in situ detection of diesel cetane. The LS-SVM-EWs models showed a little better performance than that of LS-SVM-PCs models, but LS-SVM-LVs model was the best one for detection of diesel cetane. All LS-SVM models had excellent prediction precision.

Conclusions

Near Infrared Spectroscopy combined with LS-SVM regression method was successfully utilized for the determination of diesel cetane. The PLS models were developed and compared using different spectral preprocessing methods. In the PLS models, the optimal prediction performance was achieved using raw spectra for the prediction of diesel cetane number. Moreover, LS-SVM models were developed using different input data set and performed only slightly worse than PLS models. Simultaneously, the performance of LS-SVM-LVs models was better than that of LS-SVM-EWs and LS-SVM-PCs models. The excellent performance of LVs showed a potential application using LVs to develop portable instruments or in situ detection of diesel cetane. These results were obtained under laboratory conditions and might be useful for the process and in situ monitoring of diesel cetane. Further parameters optimisation was needed to improve the calibration stability and robustness for situ measurement. In addition, the differences between results of PLS regression and LS-SVM are not significant. Therefore, the superiority of LS-SVM for handling nonlinear information was not very distinct in this work might be due to little nonlinear information in spectral data or small samples number used. In the future, more samples may need to be involved to further assess whether the

LS-SVM models was really better than other linear models such as PLS for prediction of quality in diesels.

Acknowledgements

This study was supported by the National Natural Science Fund of China (Project No: 61134011, 61565005).All computations were carried out with programm codes usingMatlab Version 2010b,Unscramble 9.7 and with Origin 8.5.

References

[1] M. Bampi, A.D.P. Scheer, F. de Castilhos, Application of near infrared spectroscopy to predict the average droplet size and water content in biodiesel emulsions, Fuel, 113 (2013) 546-552.

[2] S. Lee, H. Choi, K. Cha, H. Chung, Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha, Microchemical Journal, 110 (2013) 739-748.

[3] X. Ding, Y. Ni, S. Kokot, NIR spectroscopy and chemometrics for the discrimination of pure, powdered, purple sweet potatoes and their samples adulterated with the white sweet potato flour, Chemometrics and Intelligent Laboratory Systems, 144 (2015) 17-23.

[4] C. Kordulis, K. Bourikas, M. Gousi, E. Kordouli, A. Lycourghiotis, Development of nickel based catalysts for the transformation of natural triglycerides and related compounds into green diesel: a critical review, Applied Catalysis B: Environmental, 181 (2016) 156-196.

[5] J.C.L. Alves, R.J. Poppi, Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration, Fuel, 165 (2016) 379-388.

[6] L.D.A. Ribeiro, A.D.S. Soares, T.W.D. Lima, C.A.C. Jorge, R.M.D. Costa, R.L. Salvini, C.J. Coelho, F.M. Federson, P.H.R. Gabriel, Multi-objective Genetic Algorithm for Variable Selection in Multivariate Classification Problems: A Case Study in Verification of Biodiesel Adulteration, Procedia Computer Science, 51 (2015) 346-355.

[7] F. Zhu, S. Cheng, D. Wu, Y. He, Rapid Discrimination of Fish Feeds Brands Based on Visible and Short-Wave Near-Infrared Spectroscopy, Food and Bioprocess Technology, 4 (2010) 597-602.

[8] L. Xuemei, L. Jianshe, Measurement of soil properties using visible and short wave-near infrared spectroscopy and multivariate calibration, Measurement, 46 (2013) 3808-3814.

[9] F. Liu, Y. He, L. Wang, G. Sun, Detection of Organic Acids and pH of Fruit Vinegars Using Near-Infrared Spectroscopy and Multivariate Calibration, Food and Bioprocess Technology, 4 (2011) 1331-1340.

[10] Y.N. Shao, C.J. Zhao, Y.D. Bao, Y. He, Quantification of Nitrogen Status in Rice by Least Squares Support Vector Machines and Reflectance Spectroscopy, Food and Bioprocess Technology, 5 (2012) 100-107.

[11] D. Wu, Y. He, S. Feng, D.-W. Sun, Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM, Journal of Food Engineering, 84 (2008) 124-131.

[12] F. Chauchard, R. Cogdill, S. Roussel, J.M. Roger, V. Bellon-Maurel, Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity

prediction in grapes, Chemometrics and Intelligent Laboratory Systems, 71 (2004) 141-150.

[13] Z. Xiong, D.-W. Sun, H. Pu, Z. Zhu, M. Luo, Combination of spectra and texture data of hyperspectral imaging for differentiating between free-range and broiler chicken meats, LWT - Food Science and Technology, 60 (2015) 649-655.

[14] Li Jingyan, An Xiaochun, Tian Songbai, Yang Xingming, Research on Fast evaluation for Diesel Cetane Number by Near-Infrared Spectroscopy, Petroleum Processing and Petrochemicals, (2016) 101-107.