PAPER

Improved analysis of inorganic coal properties based on near-infrared reflectance spectroscopy

Ruifen Hu,[a] Yasheng Wang,[a] Meng Yang,[b] Xinmeng Li,[c] Zhiyuan Luo[b] and Guang Li*[a]

Near-infrared reflectance spectroscopy (NIRS) is a fast and convenient analytical tool, and it has become the preferred choice for online coal property analysis in recent years. Organic molecules have much stronger absorption ability than inorganic molecules. Therefore, better analysis accuracy can be achieved for organic properties of coal such as volatile matter and fixed carbon than that for inorganic properties such as ash and sulfur. This paper used a much better algorithm (least squares support vector machine) than a previous study and proposes a new method to improve the analysis accuracy of inorganic properties by utilizing the analysis results of volatile matter and fixed carbon to enhance the regression models for ash and sulfur. Four types of coal (*i.e.* fat, coking, lean and meager lean) were considered in our experiments. Individual models for each type of coal have been established and predicted values of volatile matter and fixed carbon based on NIRS were added with the relevant PCA components during the modeling. The experimental results have shown that our proposed method which utilizes information of the organic properties could improve the analysis results of the inorganic properties by around 35% using NIRS.

1. Introduction

Near-infrared reflectance spectroscopy (NIRS) is a fast and popular analysis technique and its advantages include time saving, safety and convenience.[1–3] Information obtained by NIRS mostly comes from the overtone and combination bands caused by the vibration of molecules. Therefore, NIRS is very appropriate to analyze organic compounds such as the nutrition compositions in food and biofuels.[4–8]

Although most inorganic compounds respond little to near-infrared waves, it is still possible to analyze inorganic properties with NIRS since the inorganic properties are highly related to organic properties in some particular materials. Thus the composition of organic compounds reflects the composition of inorganic compounds indirectly.[9] There are published studies on using NIRS to analyze the composition of trace elements in soil,[9–11] water,[12] animal bodies[13] and plants.[14] Existing analysis of inorganic properties with NIRS typically applies chemometric methods directly to the spectra to obtain the required regression results. A novel method taking advantage of the analysis of organic properties using NIRS to help the analysis of inorganic properties that are closely related to these organic properties is proposed in this paper. Organic compounds are firstly analyzed by NIRS and then these results are utilized in the regression models of inorganic compounds.

In this paper, coal is considered. Previous research showed that NIRS could be used to analyze coal properties such as ash, sulfur (S), fixed carbon (FC) and volatile matter (VM).[15–19] Since NIRS has advantages in analyzing organic compounds, organic properties such as volatile matter and fixed carbon were better predicted than inorganic properties such as ash and sulfur. Ash is the residue after the organic matters burn up and a large amount of sulfur exists in organic matters in many forms of coals. These two coal properties (ash and sulfur) are related to organic compounds. This paper investigates how to use the analyzed results of volatile matter and fixed carbon to improve the analysis of ash and sulfur. Experiments are performed on coal samples to evaluate our proposed method.

A step-by-step approach is used in this study. Firstly, prediction models were built to show that NIRS could be used to analyze inorganic properties of coal. Partial least squares regression (PLSR) models and least squares support vector machine (LS-SVM) models were compared and the better models were chosen for the next step. Secondly, the verification of our proposed method was carried out. The chemical reference data of volatile matter and fixed carbon obtained using the ISO standard method were added as extra features in the regression procedures for ash and sulfur. The analysis accuracy was evaluated and compared. Thirdly, the predicted values of volatile matter and fixed carbon based on NIRS were added instead of chemical reference data. The analysis accuracy was also evaluated. Finally, our previous study showed that individual models for different types of coal yielded better

[a]State Key Laboratory of Industrial Control Technology, Institute of Cyber Systems and Control, Zhejiang University, Hangzhou 310027, Zhejiang, China. E-mail: guangli@zju.edu.cn

[b]Computer Learning Research Centre, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, UK

[c]College of Computer Science, Sichuan University, Chengdu 610065, Sichuan, China

prediction results than the models using all samples together. The same idea was used for four types of coal (*i.e.* fat, coking, lean, meager lean) and predictions for both ash and sulfur were evaluated.[20] The experiments showed that the analysis results of organic properties could help the analysis of inorganic compounds.

2.   Materials and methods

2.1.   Coal samples

199 coal samples collected from 14 different coalmines in Shanxi Province, China, were used for the study discussed in this paper. Although they were all bituminous according to the ASTM standard, the coal samples were grouped into four types based on their commercial transactions and different uses in industry, including 50 fat coal samples, 50 coking coal samples, 49 lean coal samples and 50 meager lean coal samples. The particle size of all samples was less than 3 mm. Four properties (ash, S, VM and FC) of all samples were analyzed using ISO chemical method. For each sample, the four properties were analyzed 6 times under the same conditions by the same person and the mean value was calculated as the chemical reference value. The lab method had high accuracy and the maximum standard errors of ash, S, VM and FC using lab method were 0.0567%, 0.0103%, 0.126% and 0.184%. Table 1 shows the mean value and standard deviation of the properties of the 199 samples using ISO chemical method.

2.2.   Spectra recording

All spectra were collected with a Matrix-I Industrial Type Fourier transform NIR spectrometer manufactured by Bruker, Germany. The recording was controlled using the software package OPUS6.5 and the data were processed by MATLAB2010b.

The recording experiments were performed under the conditions of temperature of 20–22 °C and humidity of 50–60%. Each coal sample was held in a sample cup (97 mm in diameter and 45 mm in height). Each spectrum collected was the mean value of 64 scans covering the wavelength from 10 000 cm^{-1} to 4000 cm^{-1} (1000–2500 nm) with a digital resolution of 4 cm^{-1}. The spectral data were then analyzed with MATLAB2010b. Fig. 1 shows the spectra of all samples.

According to Fysh, the most obvious peaks were caused by water combination band (around 1900 nm), combination of C–H stretch of C–C stretch (around 2130 nm) and combination of CH$_3$, CH$_2$ asymmetric stretch band (around 2270 nm). Because of the an harmonicity and other coupling effects, the peaks usually did not appear at the calculated position and the
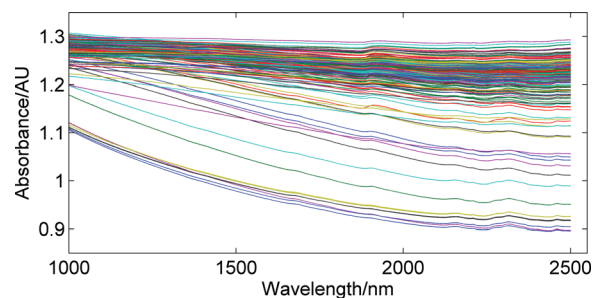

Fig. 1 Near-infrared reflectance spectra of the 199 coal samples.

composition of coal is very complicated; the specific compositions of properties such as ash and volatile matter were not clear. The best choice was to use the whole range of wavelength (1000–2500 nm) to build the regression models.[20,21]

2.3.   Data processing methods

2.3.1   Partial least squares regression (PLSR). PLSR is a popular and effective method to handle NIRS data. Since each coal spectrum contains more than 1000 highly related variables,[1,22] preprocessing of the spectra before applying PLSR can improve data quality. As reported in our previous study, the same set of preprocessing methods including smooth, differential, MSC (multiplicative scatter correction) was tried and the whole range of wavelength was used for each of the four coal properties (*i.e.* ash, sulfur, volatile matter and fixed carbon).[23]

2.3.2   Least squares support vector machine (LS-SVM). LS-SVM is an improvement of standard support vector machine. Instead of solving complicated quadratic optimization problems, LS-SVM finds the support vectors through a group of linear equations, thus the complexity is reduced and the calculation takes less time. LS-SVM is capable of dealing with regression problems in high-dimensional feature space and has proven to be very effective for dealing with NIRS data.[24–26] Like SVM, LS-SVM uses a kernel function but there is no agreed criterion for choosing a kernel function. Several popular kernel functions were experimented and RBF kernel function was chosen after comparison. The regularization parameter gamma (γ) of LS-SVM and the RBF kernel function parameter σ^2 were optimized by grid searching and decided according to the results of leave-one-out cross validation. To reduce the computation complexity, PCA (principal component analysis) was applied before LS-SVM to our NIRS data. Different numbers of principal components were tried for each model and the optimal number of principal components for modeling was determined based on the best prediction accuracy achieved for the cross validation.

2.3.3   Improvement calculation method. Because there is a close connection between inorganic matters and organic matters in coal, this paper investigates how to estimate properties of inorganic matters using the information of organic matters in NIRS data. A new method is proposed to improve the predictive capability of the models for inorganic properties by adding information of organic properties (*e.g.* volatile matter

Table 1 Mean value and standard deviation (S.D.) of the properties of the 199 samples

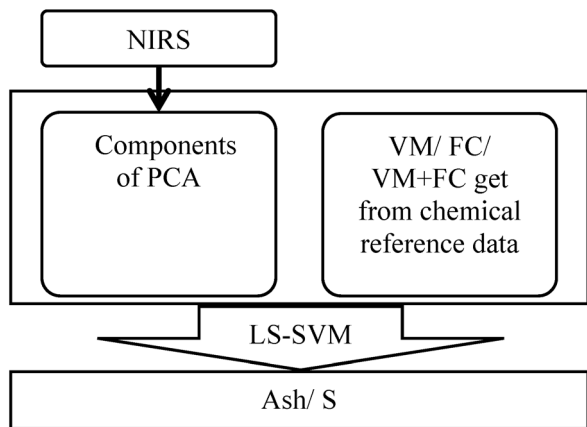| Properties	Ash (%)	S (%)	VM (%)	FC (%)
Mean value	10.76	1.10	18.46	70.18
S.D.	2.15	0.62	4.99	5.23

Fig. 2 Experiment 1 of the improvement method, adding chemical reference data to the regression procedure.

and fixed carbon) as extra components to the modeling procedure for inorganic properties (*e.g.* ash and sulfur). Thus, two parts of the data were combined and used in the modeling procedure of ash and sulfur. One part was the principal components obtained from the PCA of NIRS data and the other part was the information of organic matters. LS-SVM was used to process these two combined parts because it is easier to add extra features in LS-SVM in comparison with PLSR.

In this paper, 2 sets of experiments were designed and performed to illustrate this improved calculation method. The first set of experiments was designed as the verification of this new method. The chemical reference data of volatile matter and fixed carbon were used for the improvement (see Fig. 2). PCA was applied to the NIRS data to obtain part of the regression components, and then chemical reference values of volatile matter and fixed carbon were added. Three possible combinations of chemical values were considered: using chemical value of volatile matter only, using chemical value of fixed carbon only, and using chemical values of volatile matter and fixed carbon together. For the second set of experiments, instead of using chemical reference values, predicted values of volatile
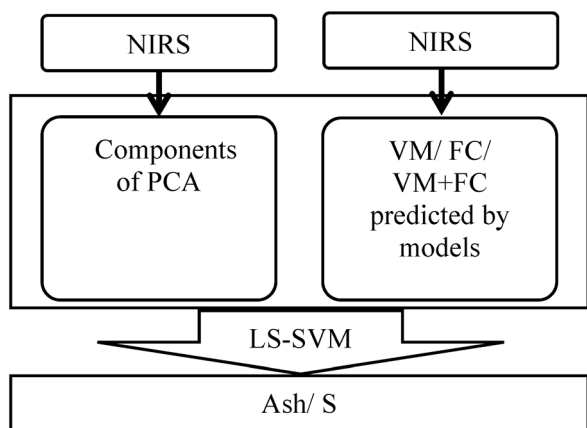

Fig. 3 Experiment 2 of the improvement method, adding predicted data to the regression procedure.

matter and fixed carbon were used for the improvement (see Fig. 3).

2.3.4 Evaluation of the models. The samples were divided into 2 sets, a calibration set to train the regression models and a validation set to test the models. Of all the 199 samples, 60 were randomly chosen as the validation set and the rest as the calibration set. For each model, a full leave-one-out cross validation was performed for the calibration set and the root mean square error of the cross validation (RMSECV) was calculated. Meanwhile, root mean square error of the prediction (RMSEP) was also calculated.

$$RMSECV = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

where N is the number of training samples, $\hat{y}_i$ represents the estimate value of the leave-one-out cross validation and y_i is the measured value.

$$RMSEP = \frac{1}{M} \sum_{i=1}^{M} (\hat{y}_i - y_i)^2$$

where M is the number of validation samples, $\hat{y}_i$ represents the estimate value and y_i is the measured value. For the convenience of comparison among different models, the percentage error around the mean was also calculated:

$$Percentage\ error = \frac{RMSEP}{Mean\ value} \times 100\%$$

3. Results and discussions

3.1. Results without improvements

A previous study used PLSR models to predict coal properties[20] and in this study, LS-SVM was also used. Models for ash, sulfur, volatile matter and fixed carbon were established based on LS-SVM and PLSR. 60 samples were randomly chosen as the validation set from the 199 samples and others as the calibration set.

Before building LS-SVM models, PCA was performed on the NIRS data to reduce the dimensions. Fig. 4 illustrates the first 2 components of PCA. Compared to the original NIRS, the samples varied a lot after PCA. Different numbers of PCA components were used in the LS-SVM and the number that
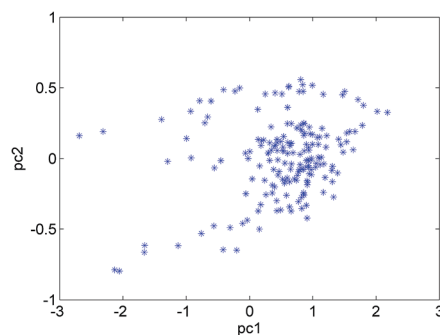

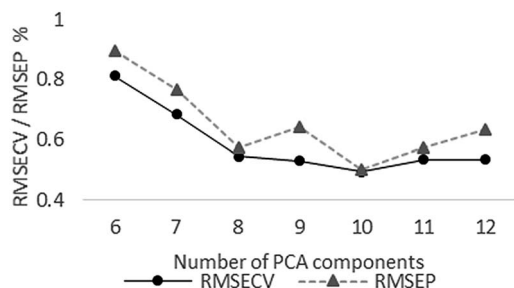Fig. 4 PCA plot of the first 2 components.

Paper

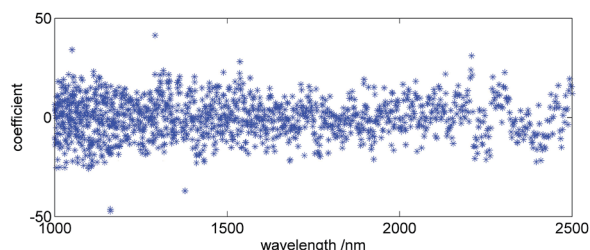**Fig. 5** RMSECV and RMSEP at different numbers of PCA components of LS-SVM models for VM.

Fig. 6 Coefficients of PLSR model for ash at each wavelength.

yielded the best RMSECV was chosen. For example, Fig. 5 displays the RMSECV and RMSEP at different numbers of PCA components of LS-SVM models for VM. The error of cross validation for the training set declined fast as the number increased and we got the best result at 10 components, so the model for

VM used 10 components. The RMSEP was of a similar value which means that the model is appropriate and not over fitted.

The PLSR model gives each wavelength a weight and Fig. 6 shows the coefficient of each wavelength from the model for ash. As the figure shows, the distribution of coefficients was almost the same from 1000 nm to 2500 nm. As we mentioned before, the composition of properties such as ash were very complicated, and the spectrum from every wavelength might contain information that contributed to the regression.

Corresponding RMSECV, RMSEP and percentage error around the mean were obtained and are given in Table 2. In general, the estimate of organic matters achieved better results, especially for fixed carbon. However, the prediction accuracy for properties of inorganic matters, especially sulfur, was not very satisfactory.

It has been reported that NIRS can be used to analyze inorganic compositions in complex materials such as soil and plants.[9–14] Previous studies of coal property analysis with NIRS also showed that the models for inorganic matters like ash and sulfur were highly related to their chemical references.[15–20] As Table 2 shows, although the models of ash and sulfur did not perform as well as those of volatile matter and fixed carbon, the models of ash and sulfur especially those by the LS-SVM method were actually related to their chemical references. Based on these results, it is feasible to analyze inorganic properties of coal such as ash and sulfur using NIRS.

The performances of the LS-SVM method were much better than those of the PLSR method for all coal properties. The RMSEP of ash, VM, FC and S had a reduction of 17.39%, 39.40%, 32.88%, and 27.27% using LS-SVM. RMSEP and RMSECV had similar values indicating that the models were stable and not over fitted. LS-SVM is a nonlinear method while PLSR is a linear method which implies that there is more of a nonlinear relationship between NIRS and coal properties. Therefore, volatile matter and fixed carbon values predicted by LS-SVM were added to the regression model of ash and sulfur in the improved method.

3.2. Verification of the improved method

In order to verify the improved method, the chemical reference data of volatile matter and fixed carbon were added to the LS-SVM regression models of ash and sulfur with the PCA components obtained from NIRS data as described in Section 2.3. Three different scenarios were considered: using volatile

Table 2 RMSECV, RMSEP and percentage error of PLSR and LS-SVM for coal properties

Properties	Methods	RMSECV (%)	RMSEP (%)	Percentage error (%)
Ash	LS-SVM	0.55	0.57	5.30
	PLSR	0.66	0.69	6.41
S	LS-SVM	0.21	0.24	21.18
	PLSR	0.37	0.33	30.00
VM	LS-SVM	0.49	0.50	2.65
	PLSR	0.74	0.73	3.95
FC	LS-SVM	0.72	0.78	1.11
	PLSR	0.91	1.03	1.47

Table 3 RMSECV, RMSEP and percentage error of the improved method with chemical reference data

Method	Ash			S		
	RMSECV (%)	RMSEP (%)	Percentage error (%)	RMSECV (%)	RMSEP (%)	Percentage error (%)
LS-SVM	0.55	0.57	5.30	0.21	0.24	21.18
LS-SVM + VM	0.46	0.45	4.18	0.18	0.18	16.36
LS-SVM + FC	0.29	0.31	2.88	0.15	0.16	14.45
LS-SVM + VM + FC	0.15	0.17	1.58	0.12	0.14	12.73

Table 4 Pearson correlation coefficient between ash, S and VM, FC

	VM	FC	Ash	S
Ash	−0.255	−0.358	1	0.0906
S	−0.245	0.279	0.0906	1

matter information only, using fixed carbon information only, and using volatile matter information together with fixed carbon information. The same calibration set and validation set were used in the experiments. RMSECV, RMSEP and percentage error obtained from these different experiments are listed in Table 3.

It is clear that the prediction error of ash decreased after the information of volatile matter and fixed carbon were added, especially that of fixed carbon, which might be due to the closer connection between ash and fixed carbon. When both volatile matter and fixed carbon information were added, the result was dramatically improved. The improvement of prediction of sulfur was not as good as that of ash, but it is evident that both volatile matter and fixed carbon contributed to the regression, and the accuracy was greatly improved by adding volatile matter measure together with fixed carbon measure. The results confirmed that it was feasible to use information about volatile matter and fixed carbon as additional features in the regression model of ash and sulfur to enhance the prediction performance. Thus, in an online application of NIRS setting, we could firstly predict the organic properties (*e.g.* volatile matter and fixed carbon) *via* NIRS regression method, and then use those predicted values of

organic properties to get better prediction for inorganic matters (*e.g.* ash and sulfur).

Table 4 lists the Pearson correlation coefficient between ash, S and VM, FC. Compared to the relationship between ash and S, the relationship between inorganic properties and organic properties was stronger and these two kinds of properties were not independent. For ash, more organic matter means less ash residue, so the coefficient was negative; for S, the situation was much more complicated since S exists in both organic and inorganic components of coal. The Pearson correlation coefficient is an index of linear relationship, so it is not unexpected that the coefficient is not a high value if the relationship is nonlinear. As discussed in Section 3.1, there is more of a nonlinear relationship between NIRS and coal properties. The experimental results indicated that as a nonlinear method, LS-SVM can take advantage of the relationship between organic properties and inorganic properties to improve the models.

3.3. Improvements with predicted data

Results in Table 3 demonstrated that information of fixed carbon and volatile matter actually had a positive effect on the regression of ash and sulfur. In this section, instead of using chemical reference data, predicted volatile matter and fixed carbon values were used in the same way as in experiment 2 described in Section 2.3. RMSECV, RMSEP and percentage error were obtained from the experiments and are listed in Table 5. Both VM and FC information helped to reduce the prediction error of ash and fixed carbon played a much more important role. Similarly, prediction error of sulfur was reduced after VM and FC were added respectively and even better results were obtained when both VM and FC were added together; the

Table 5 RMSECV, RMSEP and percentage error of improvement with predicted values by PLSR models

Method	Ash			S		
	RMSECV (%)	RMSEP (%)	Percentage error (%)	RMSECV (%)	RMSEP (%)	Percentage error (%)
LS-SVM	0.55	0.57	5.30	0.21	0.24	21.18
LS-SVM + VM	0.47	0.48	4.18	0.18	0.19	17.27
LS-SVM + FC	0.41	0.44	2.88	0.17	0.17	15.45
LS-SVM + VM + FC	0.36	0.37	1.58	0.14	0.15	13.64

Table 6 RMSECV and the percentage improvement of models for different types of coals (fat, coking, lean, meager lean) with improved method

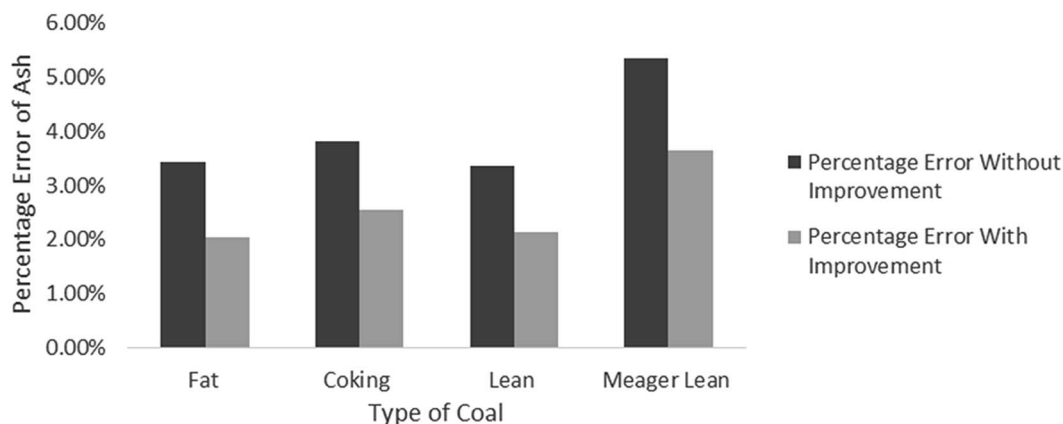| Coal type	Properties	RMSECV without improvement (%)	RMSECV with improvement (%)	Percentage improvement (%)
Fat	Ash	0.37	0.22	40.54
	Sulfur	0.07	0.05	28.57
Coking	Ash	0.39	0.26	33.33
	Sulfur	0.14	0.10	28.57
Lean	Ash	0.36	0.23	36.11
	Sulfur	0.07	0.05	28.57
Meager lean	Ash	0.60	0.41	31.67
	Sulfur	0.17	0.14	17.65

Fig. 7 Percentage errors of ash using models with and without improvement for individual types of coal (fat, coking, lean, meager lean).
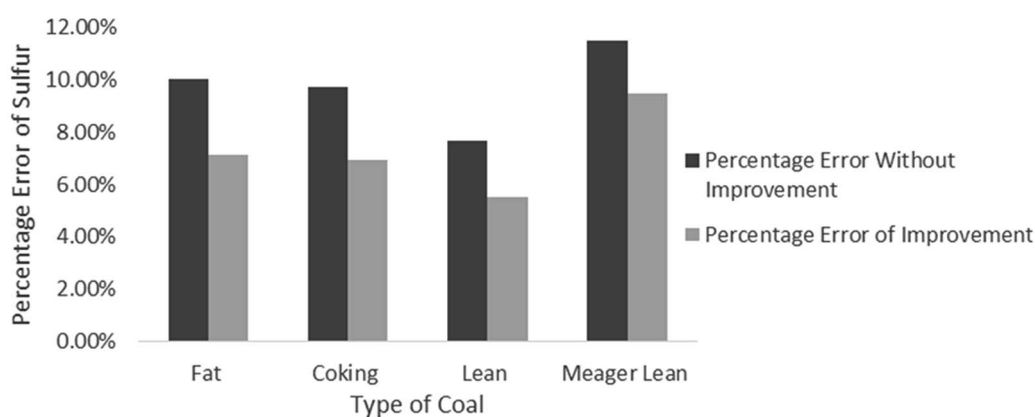

Fig. 8 Percentage errors of sulfur using models with and without the improved method for individual types of coal (fat, coking, lean, meager lean).

RMSEP of ash and S had a reduction of 35.09% and 37.50%. The alternative method by using predicted values of VM and FC achieved similar results to the method using chemical reference data in Section 3.2. Since predicted values by LS-SVM were not as accurate as the chemical reference data, it was not surprising that using predicted values yielded slightly greater errors. However, these improvements demonstrated that good prediction accuracy of inorganic matters such as ash and sulfur could be obtained based on NIRS data.

A previous study[20] showed that models for different types of coal had better efficiency than those for all samples. To test whether the improved method proposed in this paper could be applied to the models of individual types of coal or not, the same method as above (both volatile matter and fixed carbon information were added) was applied to the four types of coal separately. First, LS-SVM models for coal properties based on different types of coal were established; then improvement for ash and sulfur were performed using predicted values from the LS-SVM models. When building models for individual types of coal, it is not appropriate to separate the samples into calibration set and validation set since there were only about 50 samples of each type of coal. Leave-one-out cross validation

was performed in these models and RMSECV and percentage error were used to evaluate the models. The RMSEP in the percentage error was replaced by RMSECV of the leave-one-out validation.

Table 6 details the RMSECV and the percentage improvement for the different types of coal. Fig. 7 and 8 illustrate the improvement of percentage error for the four types of coal using our proposed method. The models yielded much more precise prediction using models for individual types of coal compared to models for all samples. Good improvements for both ash and sulfur for all four types of coal were achieved. Our proposed method also worked well for individual types of coal. To achieve the best result in an online coal analysis system, coal samples could firstly be grouped into different types, and then our proposed method could be applied to analyze inorganic properties such as ash and sulfur.

4. Conclusions

Inorganic matters and organic matters are highly related in many complicated materials such as coals. For example, the estimates of ash and sulfur in coals are closely connected to the

components of volatile matter and fixed carbon. This paper first utilized the LS-SVM method with the NIRS models which had a much more precise prediction than previous studies, then investigated how to use the organic properties to improve the analysis of inorganic properties with NIRS data and proposed a new method. Predicted values of organic properties by NIRS models combined with PCA components obtained from NIRS were used in the regression of ash and S. The experimental results have shown that organic properties were useful for analyzing inorganic properties and helped to reduce the prediction error by around 35%. The prediction performance for ash and sulfur could be further improved by modelling four different types of coal (*i.e.* fat, coking, lean, meager lean) individually as confirmed by the experimental results. Other properties of coal such as heat value will be investigated in a further study. This method can be tried when dealing with other inorganic properties in complicated materials such as soil.

Acknowledgements

References

1 E. Agelet and C. R. Hurburgh Jr, *Crit. Rev. Anal. Chem.*, 2010, **40**, 246–260.
2 D. Cozzolino, *Planta Med.*, 2009, **75**, 746–756.
3 C. Pasquini, *J. Braz. Chem. Soc.*, 2003, **14**, 198–219.
4 C. M. McGoverin, J. Weeranantanaphan, G. Downey and M. Manley, *J. Near Infrared Spectrosc.*, 2010, **18**, 87–111.
5 T. R. A. Lestander and R. Samuelsson, *Energy Fuels*, 2010, **24**, 5148–5152.
6 V. Gaydou, J. Kister and N. Dupuy, *Chemom. Intell. Lab. Syst.*, 2011, **106**, 190–197.
7 U. Siripatrawan, Y. Makino, Y. Kawagoe and S. Oshita, *Sens. Actuators, B*, 2010, **148**, 366–370.
8 H. Aleme and P. Barbeira, *Energy Fuels*, 2012, **26**, 5769–5774.
9 D. Malley and P. Williams, *Environ. Sci. Technol.*, 1997, **31**, 3461–3467.
10 D. Summers, M. Lewis, B. Ostendorf and D. Chittleborough, *Ecol. Indic.*, 2011, **11**, 123–131.
11 D. Cozzolino, W. U. Cynkar, R. G. Dambergs, N. Shah and P. Smith, *Commun. Soil Sci. Plant Anal.*, 2013, **44**, 1610–1619.
12 J. Li, Y. Zhang, W. Cai and X. Shao, *Talanta*, 2011, **84**, 679–683.
13 I. González-Martín, C. González-Pérez, J. Hernández-Méndez and N. Alvarez-Garcíia, *Anal. Chim. Acta*, 2002, **468**, 293–301.
14 D. Cozzolino, W. Cynkar, N. Shah and P. Smith, *Comput. Electron. Agr.*, 2011, **77**, 81–85.
15 M. Kaihara, T. Takahashi, T. Akazawa, T. Sato and S. Takahashi, *Spectrosc. Lett.*, 2002, **35**, 369–376.
16 J. Andres and M. Bona, *Anal. Chim. Acta*, 2005, **535**, 123–132.
17 J. Andres and M. Bona, *Talanta*, 2006, **70**, 711–719.
18 M. Bona and J. Andres, *Talanta*, 2007, **72**, 1423–1431.
19 D. W. Kim, J. M. Lee and J. S. Kim, *Korean J. Chem. Eng.*, 2009, **26**, 489–495.
20 Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo and G. Li, *Sens. Actuators, B*, 2014, **193**, 723–729.
21 S. A. Fysh, D. A. Swinkels and P. M. Fredericks, *Appl. Spectrosc.*, 1985, **39**, 354–357.
22 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
23 Å. Rinnan, F. V. D. Berg and S. B. Engelsen, *TrAC, Trends Anal. Chem.*, 2009, **28**, 1201–1222.
24 F. Chauchard, R. Cogdill, S. Roussel, J. Roger and V. Bellon-Maurel, *Chemom. Intell. Lab. Syst.*, 2004, **71**, 141–150.
25 F. Liu, Y. He, L. Wang and G. Sun, *Food Bioprocess Technol.*, 2011, **4**, 1331–1340.
26 D. Wu and D.-W. Sun, *Talanta*, 2013, **116**, 266–276.