JAAS

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

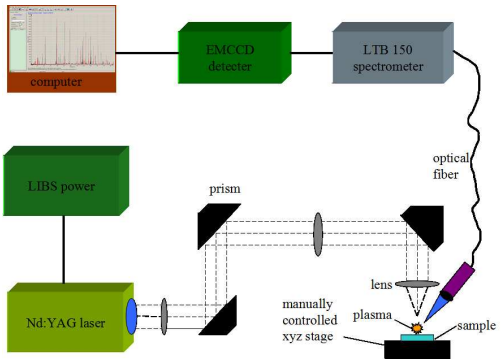
You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/jaas

A novel method based on laser induced breakdown spectroscopy(LIBS) combined with random forest regression(RFR) was proposed to quantitative analysis of multielement of fourteen steel samples. Normalized LIBS spectrum of steel in which characteristic line(Si, Mn, Cr, Ni and Cu) identified by NIST database was used as analysis spectrum. Then, two parameters of RFR were optimized by out-of-bag (OOB) error estimation. The performance of calibration model was investigated by different input variables(the whole spectral bands(220-800nm) and spectra feature bands(220-400nm), respectively). In order to validate the predictive ability of multielement calibration model in steels, we compared RFR with partial least-squares(PLS) and support vector machines(SVM) to predict the concentrations of multielement in steels. And, the three quantitative techniques are evaluated in terms of prediction accuracy and root mean square error(RMSE). Random forest is shown to correctly model nonlinear effects dues to self-absorption in the plasma and to provide the best results. It confirms that LIBS technique coupled with RFR has a good potential for the in situ rapid determination of multielement in steels and even metallurgy field.



Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

A novel approach for quantitative analysis of multi-elements in steels based on laser-induced breakdown spectroscopy(LIBS) and random forest regression(RFR)

Tianlong Zhang[a], Long Liang[a], Kang Wang[b], Hongsheng Tang[a], Xiaofeng Yang[c], Yixiang Duan[d], Hua Li[*a]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

A novel method based on laser induced breakdown spectroscopy(LIBS) and random forest regression(RFR) was proposed to quantitative analyze of multi-elements in fourteen steel samples. Normalized LIBS spectra of steel in which characteristic line(Si, Mn, Cr, Ni and Cu) identified by NIST database were used as analysis spectra. Then, two parameters of RFR were optimized by out-of-bag (OOB) error estimation. The performance of calibration model was investigated by different input variables(the whole spectral bands(220-800nm) and spectra feature bands(220-400nm), respectively). In order to validate the predictive ability of multi-elements calibration RFR model in steels, we compared RFR with partial least-squares(PLS) and support vector machines(SVM) by means of prediction accuracy and root mean square error(RMSE). Thus, RFR model can eliminate the influence of nonlinear factors dues to self-absorption in the plasma and provide a better predictive result. It confirms that LIBS technique coupled with RFR has a good potential for the in situ rapid determination of multi-elements in steels and even metallurgy field.

1. Introduction

Iron and steels is one of the most significant engineering and construction materials due to its low price and widely applicability. Some added elements(such as silicon, chromium and nickel) play an important role in improving its mechanical and chemical properties in steelmaking processes. Thus, the content of these elements must be strictly controlled, which contributes to improving the performance of the steel materials and a rapid and precise analytical method is desirable. It becomes particularly significant for accurately and sensitively quantitative analysis of steels in metallurgy and related fields. Conventional quantitative analysis technologies for steels mainly include:[1-6] chemical analysis, atomic emission spectrometry(AES), optical emission spectroscopy(OES), X-ray fluorescence(XRF), inductively coupled plasma mass spectrometry(ICP-MS), chromatography and so on. However, these techniques require complicated sample preparation and much analysis time, which hinders their application for in-situ, on-line and real-time analysis.

Laser induced breakdown spectroscopy(LIBS) is a new type of plasma spectral quantitative analysis technology with capable of rapid and real-time analysis. Compared with conventional analytical techniques, LIBS has many obvious advantages,[7-9] such as multi-elements simultaneous analysis, all types of the sample(solids, liquids, and gases) can be analyzed, less sample requirement and minimal sample preparation. Therefore, LIBS is considered to be one of the most valuable and prospect analysis tools. At present, the LIBS technology has become an international research focus on the metallurgical analysis.[10-13] Application of LIBS to metallurgical industry including iron ore selection,[14,15] process control[16,17] and iron slag analysis[18] has been widely studied by many groups. Several review works[16,19,20] have already been present.

Quantitative analysis methods on LIBS mainly refer to calibration method and calibration-free(CF) approach.[21] The first one is based on a set of calibration samples of known content, whereas CF-LIBS assumes local thermodynamic equilibrium(LTE) in the laser plasma to calculate its plasma temperature and its electron density, from which the composition of the sample is then derived, regardless of the matrix effect. The simplest and most widespread quantitative

analysis method is the standard calibration method, which construct the relationship between the integrated intensity of the analysis line or intensity ratio(analysis line vs reference line) of interest element and the known concentration of a set of calibration samples. The most common calibration curve is univariate, and its regression model is established by using the intensity of single feature line and the corresponding concentration of the element under test. However, the univariate calibration model can not often meet requirement of the quantitative analysis due to the fluctuation of laser energy, the inhomogeneity of samples and complex matrix effect.[22] The chemical composition of steel sample is affected by many matrix effects. There are serious overlapping spectral peak of the spectrum in the iron substrate, and the traditional univariate calibration model fails to eliminate the impact of these interference factors. Multivariate calibration method is an effective tool to overcome matrix effect for complex sample. At present, many multivariate calibration algorithms have been developed for quantitative analysis, such as partial least squares(PLS),[23-26] principal components analysis(PCA),[27-30] artificial neural network(ANN),[31-34] support vector machines (SVM),[35-37] and so on. However, Random Forest regression (RFR), a new regression algorithm based on multiple regression trees, was proposed by Leo Breiman in 2001.[38] It is based upon an ensemble of decision trees, from which the prediction of a continuous variable is proved as average of the predictions of all trees. In RF regression, an ensemble of regression trees is grown from separate bootstrap samples of the training data using the classification and regression tree(CART) algorithm. It has been proved that RFR has a good tolerance for the noise, as well as avoid over-fitting phenomenon by many researchers.

J. Remus proposed an approach of LIBS and RF to identify and classify five different materials(four rock samples and one pen ink sample).[39] However, in this article, we present a novel method for quantitative analysis of multi-elements in steels by means of integrating LIBS technology with RFR. Normalized LIBS spectra of steel were used as analysis spectrum. Two parameters (**ntree**-number of trees and **mtry**-random variables) of the RFR algorithm were optimized using out-of-bag (OOB) error estimation. The performance of calibration model was investigated by different input variable(the whole spectral bands(220-800nm) and spectra feature bands(220-400nm), respectively). Both the whole spectral bands and spectra feature bands are the peak intensity at each wavelengths. In order to validate the predictive ability of multi-elements calibration model in steels, we compared the result of RFR with partial least-squares(PLS) and support vector machines(SVM) by means of prediction accuracy and root mean square error(MSE).

2. Experimental

2.1 LIBS setup and acquisition conditions

The spectra for steel samples were recorded and collected by the LIBS system. A schematic diagram of the LIBS system on this work is presented in Figure 1. A dual-wavelength (the optional wavelength at 532nm and 1064nm) single pulse Q-switched Nd:YAG laser with the fundamental wavelength at 1064nm, the pulse laser energy of 80mJ(6 GW/cm^2), the pulse duration of 10 ns full width at half maximum (FWHM), and the repetition rate of 20Hz was used. The steel samples were placed directly on an X-Y-Z manual micrometric stage. The laser beam was focused onto the sample surface vertically by a 50mm focal-distance lens, producing a spot of about 2 mm diameter. The emission from the plasma created was collected with a 4-mm aperture, with a 7mm focus fused silica collimator placed at 45° angel with respect to the laser pules and a distance of 3 cm from the sample, and then focused into an optical fiber (with a 1000 nm core diameter and 0.22 numerical aper-ture), which was coupled to the entrance of the Echelle spectrometer (ARYELLE-UV-VIS, LTB150, German). The spectrometer provides a constant spectral resolution (CSR) of 6000 over a wavelength range 220-800 nm displayable in a single spectrum. An Electron-Multiplying CCD camera (QImaging, UV enhanced, 1004 ×1002 Pixels, USA), coupled to the spectrometer was used for detection of the dispersed light. The overall linear dispersion of the spectrometer camera system ranges from 37 pm (at 220 nm) to 133pm/pixel (at 800nm). To prevent the CCD from detecting the early plasma continuum, a mechanical chopper is used in front of the entrance slit. The experiments were carried out under atmosphere condition, and the gate width of spectrometer was set to 2 ms. The detector was set to 1.5 μs delay time between the laser pulse in order to prevent the detection of bremsstrahlung radiation.

2.2. Steel samples and LIBS measurements

A total of 14 typical steel samples were kindly provided by the China Xi-ning Special Steel CO., LTD(Xi-ning, Qing-hai, China). Table 1 lists the concentration of certain element of 14

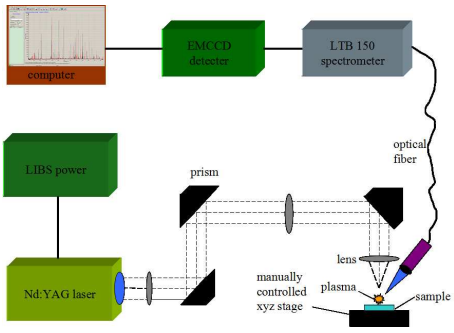


Fig. 1 Schematic setup of LIBS's experimental system

steel samples. The original size of steel samples is too long to facilitate the experiment, and then an cylinder(the height was 6 mm) was cut from random locations on each steel sample. LIBS spectra of 50 different position of each sample surface are gathered. In order to decrease the effects of shot to shot fluctuations, each measure spectrum was obtained by accumulating of 20 laser pulses. In this work, the analytical spectrum for each steel sample was the average of 50 LIBS spectra from different positions. The total of the spectra for steel sample was 14, ten samples were selected for the calibration the quantitative analysis(PLS, SVM and RFR) model, and the rest of samples were used for validation of the model. Background emission was subtracted from the spectral lines, and each LIBS spectra were normalized by the maximum integrated intensity. The data processing and quantitative analysis for steel samples by chemometrics methods were completed on Matlab(version 2007a, Mathworks).

2.3. Support Vector Machines

Support vector machine(SVM) is a new and promising classification and regression method proposed by Vapnik.[40] It was originally developed for classification problems, but can also be extended to solve non-linear regression problems by means of ε-insensitive loss function. In statistical learning theory, empirical risk is the error of prediction results by the model and real results, however, structure risk is the sum of empirical risk and confidence interval. Traditional learning method are based on empirical risk minimization criterion, and it only emphasized the empirical risk minimum error of the training sample, no minimum confidence limit value, so there is a poorer generalization ability. With regard to structural risk minimization, the training error is used as its optimization constraints, and the minimize of trust scope value is used as optimization target, its generalization ability is much better than traditional learning methods. Therefore, SVM method proposed is aimed at minimizing the structural risk rather than the empirical risk, and preserving a good generalization ability rather than optimizing the agreement with a given (limited) training set. In support vector regression, the input x is first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. The kernel functions used in SVM often include linear or polynomial functions, radial basis functions and sigmoid functions. Parameter C is a regularization constant which determines the trade-off between the model complexity and the degree to which deviations larger than ε are tolerated in optimization formulation. The generalization performance of SVR depends on a good setting of parameters: C, ε and the kernel type and corresponding kernel parameters. The selection of the kernel function and corresponding parameters is very important because they define the distribution of the training set samples in the high dimensional feature space. All SVM models in our present study were implemented using the shareware program LibSVM developed by Lin.[41] The radial basis function was used as kernel function in this work. For RBF kernel, the most important parameter is the width of the radial basis function.

Table 1 The elemental reference concentration of 14 steel samples(wt. %)

Sample number	Si	Mn	Cr	Ni	Cu
1#	0.238	0.420	0.083	0.011	0.024
2#	0.199	0.605	0.108	0.108	0.034
3#	0.201	0.733	0.754	0.030	0.102
4#*	0.277	0.622	0.939	0.018	0.016
5#	0.297	0.891	1.16	0.013	0.026
6#	0.200	0.542	0.972	0.046	0.062
7#	0.199	0.581	0.985	0.022	0.031
8#*	0.272	0.793	0.405	0.451	0.039
9#	0.360	0.920	1.12	0.015	0.017
10#	0.540	16.48	12.24	2.07	1.45
11#*	0.72	10.15	16.16	4.04	0.124
12#	0.63	1.76	17.10	8.58	0.41
13#	0.455	1.15	17.41	8.16	0.298
14#*	0.330	0.836	16.27	10.18	0.501

* were selected for validation sample

2.4. Random Forest

Random Forest(RF) as a new classification algorithm based on multiple classifier was proposed by Leo Breiman.[38] In RF method, combined model $\{h(X,\theta_k), k=1,\ldots,p\}$ consists of random vector(i.e. regression tree which is regard to the number and intensity of feature spectrum and consist of branching variables and nodes) by bootstrap resample method which is a

resample method with replacement, in which each tree casts an equal valued vote for the prediction at input vector X and where the $\{\theta_k\}$ is independent identically distributed random vectors. k is the index for the tree in the forest and p is the total number of trees in the forest. For the k-th regression tree, we generate is a random vector θ_k, independent of the previous random vectors $\theta_1,\ldots,\theta_{k-1}$ but with the same distribution; and we grow a tree using the training set and θ_k, resulting in the predictor $\{h(X,\theta_k)\}$ where X is an input vector. Predictive vector is numeric, and random forest generated is multivariate nonlinear regression analysis model. The prediction result of random forest was produced by the the average of $\{h(X,\theta_k)\}$ for k trees. The training set for random forest model were independent absolutely and selected from random vectors Y and X. The generalization mean square error of numeric predictive vector as follows:

$$E_{X,Y}(Y-h(X)) \quad (1)$$

In which, $E_{X,Y}$ is a function with regard to X and Y.

Random forest regression has the following characteristics:[38]

a. When the number of trees in the forest tending to infinity:

$$E_{X,Y}(Y-\alpha v_k h(X,\theta_k))^2 \rightarrow E_{X,Y}(Y-E_\theta h(X,\theta))^2 \quad (2)$$

In which, α and v_k are constant.

b. For all θ and $\theta, E(Y)=E_X h(X,\theta)$:

$$PE^*(forest) \le \bar{\rho} PE^*(tree) \quad (3)$$

In which, $PE^*(tree)=E_\theta E_{X,Y}(Y-h(X,\theta))^2$, $\bar{\rho}$ is weight related between relate $Y-h(X,\theta)$, θ is independent. It pinpoints the requirements for accurate regression forests--low correlation between residuals and low error trees. The random forest decreases the average error of the tree employed by the factor $\bar{\rho}$. The randomization employed needs to aim at low correlation.

The process of random forest regression algorithm as follows:[42,43]

(1) There are **n** samples in the original dataset. **b** bootstrap sample set were random drawn with replacement using bootstrap resampling method, and thus construct **b** regression trees. Hence, some of the samples will be repeated, while others will be "left out" from the dataset and form out-of-bag(OOB) samples(about 37% of the samples in the original dataset). This left out data, which is called OOB data, is used to calibrate the performance of each tree. The predictive set for random forest was consisted by **b** OOB date that were generated by no drawn samples for the each bootstrap sample.

(2) Suppose the variable number of original data is **p**, $\mathbf{m_{try}}$($\mathbf{m_{try}}$<**p**) variable as alternative branching variable were chosen randomly at each node in every regression tree, and then in which the optimal branch is selected according to branch optimum rule. In random forests regression, parameter $\mathbf{m_{try}}$ = **p**/3;[38]

(3) Since the beginning of each regression tree recursive branch of top-down, the smallest size set leaf node nodesize=5, as a regression tree growth termination conditions; the size and range of spectroscopy in the terminal nodes of the trees increases and the predictive accuracy decreases when the **nodesize** is increased above the optimum value.

(4) The **b** regression trees generated constitute the regression model of random forests, and the performance of regression model was evaluated using mean square error(MSE) and coefficients of determination(R^2) of OOB data:

$$MSE_{OOB}=n^{-1}\sum_{1}^{n}\left\{y_i-\hat{y}_i^{OOB}\right\}^2 \quad (4)$$

$$R^2_{RF}=1-\frac{MSE_{OOB}}{\hat{\sigma}_y^2} \quad (5)$$

Among them, y_i is the bag outside data of the actual value of the dependent variable, $\hat{y}_i$ is random forest predictive value for the data outside the bag, $\hat{\sigma}_y^2$ is random forest data predicted variance outside the bag.

3. Results and discussion

3.1 LIBS spectra and spectral normalization

Fig 2 shows the averaged normalization spectrum of 2# sample, which includes the emission lines of the major elements in steel. Steels are complex samples containing many chemical elements and thus related to LIBS spectra characterized by hundreds of atomic lines. Spectral lines of major element(Si, Mn, Cr, Ni and Cu) in steel sample were detected and identified based on NIST atomic database,[44] which are summarized in Table 2. Some of the stronger elemental emission lines were used for quantitative analysis of steels. There are relative rich for Fe emission lines in steels, and spectral intensity of Si, Mn, Cr, Ni and Cu were affected by matrix effect from steel samples and the rich iron emission lines. In order to get a better quantitative performance, some simple pre-processing methods(i.g smooth and de-noise) based on five points moving-

average were used to improve the signal-to-background ratio (SNR) of spectral line for specific element. There is a bit improvement for SNR of spectral line for specific elements and the analysis performance by moving-average method.

Table 2 Spectral line for quantitative analysis

Element	Spectral line based on NIST database (nm)
Si	250.690, 251.432, 251.611, 251.920, 252.411, 252.851
Mn	257.610, 259.372, 260.568, 279.827, 293.931, 294.921
Cr	357.868, 359.348, 360.532, 425.433, 427.481, 428.974, 520.451, 520.602, 520.842
Ni	231.604, 234.554, 300.249, 310.156
Cu	324.755, 327.396

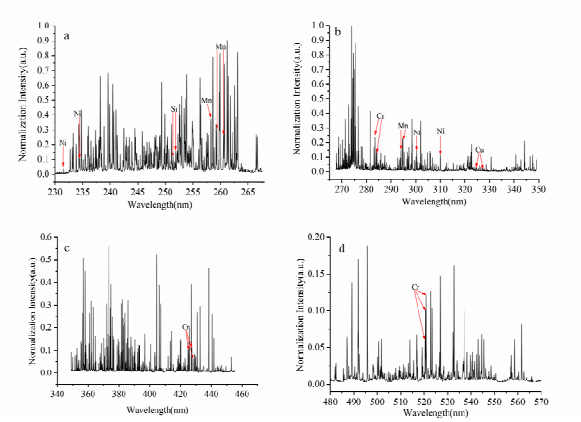


Fig. 2 The averaged normalization spectrum of 2# steel sample(a: 230-268nm; b: 270-350nm; c: 350-460nm; d: 480-570nm)

3.2 Calibration model of multi-elements in steels with RFR

Calibration model is an essential aspect for quantitative analysis of steels using LIBS and RFR. In general, the establishment of the calibration model mainly includes three parts: (1) The optimization of modeling conditions sample, including the selection of input variables and the pretreatment method; (2) The selection of training samples in modeling, enough training samples with rich feature information and minimized interference are the precondition for accurately model, which determine the adaptability and reliability of the calibration model; (3) The implementation of modeling algorithm.

Two important parameters in RFR are **ntree**-the number of the trees in the forest and $\mathbf{m_{try}}$-the number of the peaks randomly selected as the candidates for splitting at each node. Theoretically, the predictive error of the regression tree tends to a finite upper bound when **ntree** reaches a certain value. In other words, **ntree** increased is over the optimum value. There is a general increase in the computational expense, but the improvement of the predictive accuracy is minor. $\mathbf{m_{try}}$ is one of the most major characteristic through each division that introduces random nodes for randomly selected attributes. It was assumed that there were **P** attributes in the training sample, and $\mathbf{m_{try}}$ attributes were extracted randomly as candidate attribute between each of the internal nodes in the decision tree ($\mathbf{m_{try}} < \mathbf{P}$). The effects of different **ntree** and $\mathbf{m_{try}}$ for the calibration model were investigated by the OOB error estimate (as shown in Fig. 3). **ntree** were 1, 100, 200, 300, 400, 500, and 600, respectively. When the value of **ntree** is decreased too far, the results deteriorate significantly; if **ntree** reaches one, the random forest becomes a single unpruned regression tree. The OOB error of the RFR model is relatively high when **ntree** is below 300, and it reaches to a minimum when **ntree** reaches 300. In other words, the quantitative accuracy of the RF model was found to be the best. If the number of trees in the forest is increased above the optimum, there is a general increase in computational expense, but the results do not improve significantly, and as well as the OOB error tended to be limited by an upper bound. $\mathbf{m_{try}}$ used to test were 6404, 7319, 8539, 10247, 12809, 17078, 25617, and 51234, respectively. When $\mathbf{m_{try}}$ becomes very small, not enough peaks are considered at each split, and hence the predictive quality of each tree decreases. The exact value of $\mathbf{m_{try}}$ below which a decrease in predictive error is observed will depend on the number and relative importance of peaks present in the data set. Moreover, $\mathbf{m_{try}}$ =17078(namely **P**/3) was found to be the best choice based on the OOB error rate.[38] Therefore, the two optimized parameters of the random forest are as follows: **ntree** = 300 and $\mathbf{m_{try}}$ = 17078.

Input variables is significant for calibration model of steels. N-fold cross-validation is a method for model selection in terms of the predictive ability of the models. In machine learning, dataset **A** was divided into training set **B** and test set **C**. In the case of the amount of dataset is not enough large, in order to make full use of dataset to investigate the performance of the model algorithm, dataset **A** will be randomly divided into **n** package, one of **n** package is as test set, rest of the n-1 package are the train set each time. In this work, the whole spectral bands and feature spectral bands as input variables were investigated to improve the predictive accuracy of RFR, and then the RFR calibration model for multi-elements in steels was

validated by OOB estimation and 10-fold cross validation(CV).

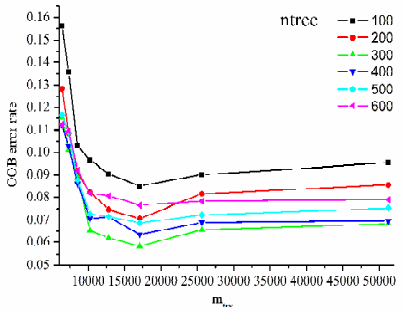


Fig. 3 Relationship of OOB error rate with **ntree** and **mtry**

Table 3 shows the correlation coefficients(R) and root mean square error(RMSE) with different input data of RFR model by means of OOB estimation and 10-fold CV. As seen as Fig. 2 and table 2, due to the most spectral line of major element (Si, Mn, Cr, Ni and Cu) in steels are most distributed in the range of 220-400nm, the spectral region of 220-400nm contains key features of specific element, hence, the spectra of 220-400nm are selected as the input data. The calibration model of multi-elements analysis in steels was constructed by the whole broadband(220-800) and the spectral feature bands(220-400nm) as input variables. For the calibration model with the whole LIBS spectra as input variables, RMSE and R were 1.5672 and 0.9285, respectively. Although the whole spectral band has a rich spectral information, there are many interference information from other element spectra and matrix effect. The feature spectrum lines for Si, Mn, Cr, Ni and Cu in steels are distributed in the range of 220-400nm. Therefore, feature spectra bands as input variables may improve the performance of calibration quantitative analysis model to some extent. For the spectral feature bands(220-400nm), RMSE and correlation coefficient were 0.4691 and 0.9735, respectively. Compared with the whole spectral bands as input variables to construct calibration model, it shows a better performance by using feature spectral band as input variables. In hence, the feature spectral bands(220-400nm) were selected as input variables to construct quantitative analysis calibration model of multi-elements in steel. OOB estimation is also a significant cross validation method in RFR. We compared OOB estimation with 5-fold cross-validation (CV) error rates for predictive ability of steel samples. The predictive error of the OOB estimation was lower than of 10-fold CV. Unlike cross-validation, the OOB estimates required no additional computing. The OOB validation is convenient for the random forest models owing to the utilization of the bootstrap method of data selection.

Table 3 Correlation coefficients(R) and root mean square error with different input data of RF models

Input data	Calibration		OOB estimation		10-fold cross validation	
	R	RMSE	R	RMSE	R	RMSE
the whole spectra (220-800nm)	0.9385	1.5672	0.9163	1.6542	0.9063	2.6542
spectra feature bands (220-400nm)	0.9735	0.4691	0.9619	0.5324	0.9523	0.6392

3.3 Validation of predictive model of for steel with RFR

In order to validate the predictive abilities of calibration RFR model of multi-elements in steels, we compared RFR method with PLSR and SVM method. Input variable of these three methods for calibration model are feature spectral bands(220-400nm). For the calibration model based on PLS, the best latent variables optimized by 5-fold cross-validation is 10. When the PLS model was trained upon the training set, the results were as follows: r^2=0.873 and RMSE=1.760. For the calibration model based on SVM, the best parameters selected by GA(genetic algorithm) were used as input for an epsilon regression SVM with a radial basis function(RBF) kernel. The optimum parameters were set as: penalty parameter $\boldsymbol{C}$ = 97.006 and kernel parameter of RBF $\boldsymbol{g}$ = 0.082. The statistics for 10-fold cross validation inside the training set were r^2 =0.880 and RMSE=0.726.

Based upon the cross-validation results for all three models, Random Forest has a better predictive performance than the PLS and SVM models of multi-elements in steels. The same is true for the prediction of the external test set. Table 4 shows the predicted results of multi-elements in steels with PLS, SVM and RFR model. The Random Forest model was able to predict percentage composition of multi-elements in steels for the test set with r^2=0.95 and RMSE=0.69. As we can see Table 4, there is a good linear relationship between predictive value and conference value of multi-elements upon test samples, and the r^2 of five elements in steels are above 0.9000. Due to the concentration of Mn element is relative greater than of Si, Cr, Ni and Cu in steels, the linear relationship of Mn element for test steel samples shows the best. While, the concentration of Cu in steel is minor and less than 1.45%, therefore the quantitative analysis result was affected by strong spectral line from other elements and matrix effect. The ability of the Random Forest to predict percentage composition of

multi-elements in steels not contained within the training set, in

Table 4 Predictive performance PLS, SVM and RFR model for quantitative analysis of muti-elements in steels

Component		Si	Mn	Cr	Ni	Cu
PLS	R^2	0.8526	0.8673	0.8695	0.8663	0.8512
	RMSEP(wt. %)	2.7562	2.0548	1.9629	2.0192	2.2681
SVM	R^2	0.8625	0.8775	0.8731	0.8792	0.8654
	RMSEP(wt. %)	2.5291	1.9562	1.7605	1.8243	2.2013
RFR	R^2	0.9248	0.9764	0.9681	0.9654	0.9462
	RMSEP(wt. %)	1.8657	0.8324	0.7395	0.6892	0.9468

conjunction with the 10-fold and out-of-bag cross-validation statistics, suggests that the model is useful for quantitative analysis of multi-elements in steels.

Conclusion

In this study, a novel method based on LIBS and RF was introduced for quantitative analysis of multi-elements in steel samples. The prediction results of both training and tested samples demonstrated that the developed RFR model is an effective approach for the multi-elements analysis of steel samples. 500 trees and 10172 random variables were optimized and selected as the best parameter for quantitative multi-elements analysis of steel samples. The predict model for the steel sample contains their chemical composition and percentage content of steel samples. The spectral feature bands(220-400nm) as input variable combined with RFR for LIBS calibration method proved to be an efficient approach for multi-elements analysis in steels. The RFR proposed presented a good accuracy(r^2=0.95 and MSE=0.69, respectively) for prediction of multi-elements(Si, Mn, Cr, Ni and Cu) in steels. Compared with predictive result using PLS and SVM, average predicted error rate of RF is lower than the results by PLS and SVM. Therefore, RFR will become a promising regression method for remote, real-time and in-situ analysis on quality supervision and process control in steel industry.

Acknowledgments

This research was financially supported by the National Special Fund for the Development of Major Scientific Instruments and Equipment(No. 2011YQ030113) of China, National Natural Science Foundation of China (No. 21175106 and No. 21375105), Research Fund for the Doctoral Program of Higher Education of China (No.20126101110019), and NWU Graduate Innovation and Creativity Funds(No. YZZ13020).

Notes and references

[a] *Institute of Analytical Science, School of Chemistry & Material Science, Northwest University, Xi'an 710069, P. R. China. Fax: Tel:86-27-88302635; E-mail: nwufxkx2012@126.com*

[b] *College of Science, Chang'an University, Xi'an, 710064, P. R. China*

[c] *College of Chemistry & Materials Science, Northwest University, Xi'an, 710069, P. R. China*

[d] *Research Center of Analytical Instrumentation, College of Chemistry, Sichuan University, Chengdu, 610064, P. R. China*

1 A. Varghese, L. George, *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy*, 2012, **95**, 46-52.

2 A. Sengupta, B. Rajeswari, R. M. Kadam, R. Acharya, *Atomic Spectroscopy*, 2011, **32**, 200-205.

3 M. Marotrao Pande, G. Muxing, R. Dumarey, S. Devisscher, B. Blanpain, *ISIJ international*, 2011, **51**, 1778-1787.

4 H. Ida, J. Kawai, *Analytical and Bioanalytical chemistry*, 2004, **379**, 735-738.

5 A. G. Coedo, T. Dorado, I. Padilla, J. C. Fariñas, *Talanta*, 2007, **71**, 2108-2120.

6 V. K. Ponnusamy, J. Jen, *Journal of Chromatography A*, 2011, **1218**, 6861-6868.

7 D. A. Cremers, R. C. Chinni. *Applied Spectroscopy Reviews*, 2009, **44**, 457-506.

8 F. J. Fortes, J. J. Laserna. *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2010, **65**, 975-990.

9 D. W. Hahn, N. Omenetto. *Applied Spectroscopy*, 2012, **66**, 347-419.

10 S. J. J. Tsai, S. Y. Chen, Y. S. Chung, P. C. Tseng, *Analytical Chemistry*, 2006, **78**, 7432-7439.

11 L. M. Cabalín, A. González, J. Ruiz, J. J. Laserna, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2010, **65**, 680-687.

12 J. Gurell, A. Bengtson, M. Falkenström, B. A. M. Hansson, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2012, **74-75**, 46-50.

13 R. Noll, V. Sturm, U. Aydin, D. Eilers, C. Gehlen, M. Höhne, A. Lamott, J. Makowe, J. Vrenegor, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2008, **63**, 1159-1166.

14 D. L. Death, A. P. Cunningham and L. J. Pollard, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2008, **63**, 763-769.

15 P. Yaroshchyk, D. L. Death and S. J. Spencer, *Journal of Analytical Atomic Spectrometry*, 2012, **27**, 92-98.

16 R. Noll, H. Bette, A. Brysch, M. Kraushaar, I. Monch, L. Peter and V. Sturm, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2001, **56**, 637-649.

17 L. Peter, V. Sturm and R. Noll, *Applied Optics*, 2003, **42**, 6199-6204.

18 M. A. Gondal, T. Hussain, Z. H. Yamani and A. H. Bakry, *Journal of Environmental Science and Health Part A- Toxic/Hazardous Substances and Environmental Engineering*, 2007, **42**, 767-775.

19 M. A. Khater, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2013, **81**, 1-10.

20 R. Noll, V. Sturm, U. Aydin, D. Eilers, C. Gehlen, M. Hohne, A. Lamott, J. Makowe and J. Vrenegor, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2008, **63**, 1159-1166.

21 A. Ciucci, M. Corsi, V. Palleschi, S. Rastelli, A. Salvetti, E. Tognoni, *Applied spectroscopy*, 1999, **53**, 960-964.

22 L. Xu, V. Bulatov, V. V. Gridin, I. Schechter, *Analytical Chemistry*, 1997, **69**, 2103-2108.

23 Z. Wang, J. Feng, L. Li, W. Ni, Z. Li, *Journal of Analytical Atomic Spectrometry*, 2011, **26**, 2175-2182.

24 S. Yao, J. Lu, J. Li, K. Chen, J. Li, M. Dong, *Journal of Analytical Atomic Spectroscopy*, 2010, **25**, 1733-1738.

25 P. Yaroshchyk, D. L. Death, S. J. Spencer, *Journal of Analytical Atomic Spectroscopy*, 2012, **27**, 92-98.

26 M. C. Ortiz, L. Sarabia, A. Jurado-Lopez, M. D. Luque de Castro, *Analytica Chimica Acta*, 2004, **515**, 151-157.

27 V. K. Unnikrishnan, K. S. Choudhari, S. D. Kulkarni, R. Nayak, V. B. Kartha, C. Santhosh, *RSC Advances*, 2013, **3**, 25872-25880.

28 D. L. Death, A. P. Cunningham, L. J. Pollard, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2008, **63**, 763-769.

29 M. R. Dong, J. D. Lu, S. C. Yao, J. Li, J. Y. Li, Z. M. Zhong, W. Y. Lu, *Journal of Analytical Atomic Spectroscopy*, 2011, **26**, 2183-2188.

30 M. M. Tripathi, K. E. Eseller, F. Y. Yueh, J. P. Singh, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2009, **64**, 1212-1218.

31 J. EI Haddad, M. Villot-Kadri, A. Ismaël, G. Gallou, K. Michel, D. Bruyere, V. Laperche, L. Canioni, B. Bousquet, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2013, **79**, 51-57.

32 J. B. Sirven, B. Bousquet, L. Canioni, L. Sarger, *Analytical Chemistry*, 2006, **78**, 1462-1469.

33 J. B. Sirven, B. Bousquet, L. Canioni, L. Sarger, S. Tellier, M. Potin-Gautier, I. L. Hecho, *Analytical and Bioanalytical Chemistry*, 2006, **385**, 256-262.

34 P. Inakollu, T. Philip, A. K. Rai, F. Y. Yueh, J. P. Singh, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2009, **64**, 99-104.

35 L. Liang, T. Zhang, K. Wang, H. Tang, X. Yang, X. Zhu, Y. Duan, H. Li, *Applied Optics*, 2014, **53**, 544-552.

36 N. C. Dingari, I. Barman, A. K. Myakalwar, S. P. Tewari, M. Kumar Gundawar, *Analytical Chemistry*, 2012, **84**, 2686-2694.

37 J. Cisewski, E. Snyder, J. Hannig, L. Oudejans, *Journal of Chemometrics*, 2012, **26**, 143-149.

38 L. Breiman, *Machine Learning*. 2001, **45**, 5-32.

39 J. Remus, K. S. Dunsin, *Applied Optics*, 2012, **51**, B49-B56.

40 V. Vapnik, *Statistical Learning Theory*,Wiley, New York, 1998.

41 C.C. Chang, C.J. Lin, *LIBSVM—A Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/□cjlin/libsvm/.

42 L. Breiman, A. Cutler, Random Forest. http://www.stat.berkeley.edu/~breiman/ RandomForests/cc_home.htm

43 A. Liaw, M. Wiener, *R news*, 2002, **2**, 18-22.

44 A. Kramida, Yu. Ralchenko, J. Reader, *NIST ASD Team, NIST Atomic Spectra Database* (ver. 5.0), [online], National Institute of Standards and Technology, Gaithersburg, MD, 2012. (Available: http://physics.nist.gov/asd).