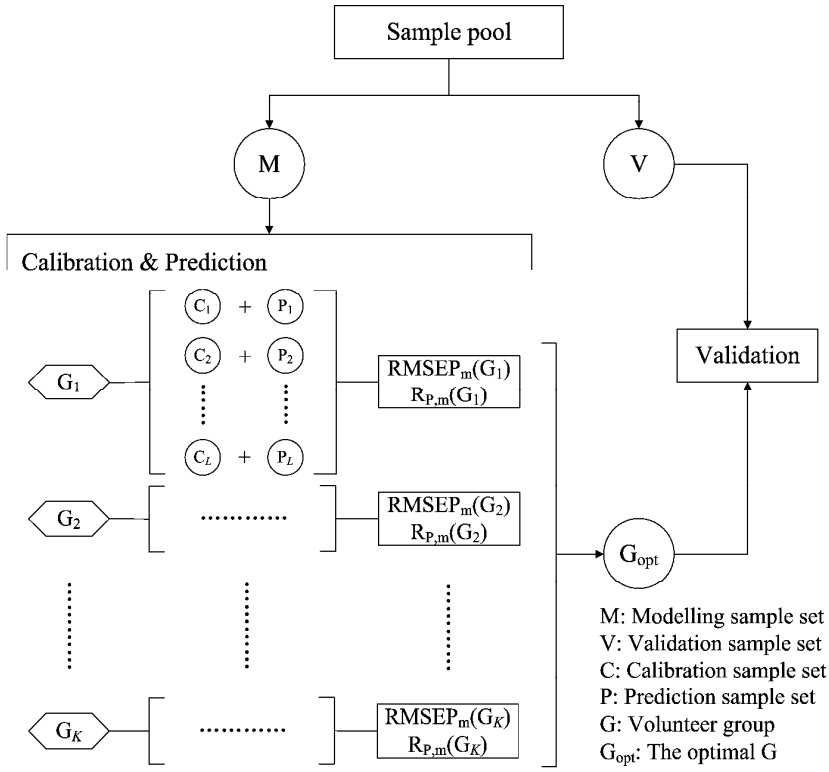
RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

Graphical Abstract



A framework for sample partitioning is proposed to take into account the tunable ratio of numbers of calibration and prediction samples, in consideration with the randomness, stability and robustness of calibration models. In the calibration-prediction-validation procedure, a fixed-number portion of validation samples (V set) is firstly extracted from the initial sample pool of the experimental data before the calibration-prediction partitioning. Then the remaining samples (M set) are partitioned into calibration set and prediction set. A fixed partitioning ratio generating a pair of calibration and prediction sets is marked as a volunteer data group (G sets). The partitioning ratio for calibration and prediction samples would influence the stability and robustness of this framework. The varied partitioning ratio corresponding to different calibration-prediction combinations result in different volunteer groups (G_1, G_2… G_K) for the parametric optimizational process. By comparing the modeling results, we can find out an optimized volunteer group (G_{opt}) to guarantee the stability and robustness of models.

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

Investigation of Sample Partitioning in Quantitative Near-Infrared Analysis of Soil Organic Carbon based on Parametric LS-SVR Modeling

Hua-Zhou Chen[a], Kai Shi[a], Ken Cai[b,*], Li-Li Xu[c], Quan-Xi Feng[a]

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

Soil organic carbon (SOC) can be quantitatively determined with enhanced stability of near-infrared (NIR) measurement. NIR analysis requires a modeling-validation division for real samples. The research of modeling robustness should be discussed in the modeling process, based on the investigation of the calibration-prediction sample partitioning. A framework for sample partitioning is proposed with the consideration of the tunable ratio of numbers of calibration and prediction samples. We addressed this issue in the multivariate calibration for NIR analysis of SOC, by using least squares support vector machine regression (LS-SVR) method with the interactive grid search of its two modeling parameters of γ and σ, where γ is the regularization parameter directly influencing the Lagrange multiplier in kernel transformation, and σ^2 represents the kernel width which is used to tune the degree of generalization. We created 7 volunteer groups for different ratios of calibration-prediction partitions. The calibration and prediction samples were re-produced for each volunteer group. LS-SVR models were established and parameters optimally selected by considering the stability and robustness based on the statistical theory of mean value and relative standard deviation. Furthermore, in all comparative partition ratios, the optimal volunteer group was selected, with the partition of 65 calibration samples and 35 prediction samples. Consequently, the optimized calibration model with correspondent optimal volunteer group was evaluated by the independent validation samples. The optimal LS-SVR parameters (γ, σ) were (110, 7), and the validation results observed a root mean square error of 0.302 and a correlation coefficient of 0.907. This validation effect was much satisfactory for the random validation samples because we have chosen an optimal volunteer group for calibration-prediction partition to guarantee the modeling stability and robustness in the process of model optimization.

1. Introduction

Near-infrared (NIR) spectroscopy is a rapid and reagent-less physical technique, requiring minimal or no sample preparation and, in contrast with traditional chemical analysis, does not require reagents, nor produces wastes [1-2]. This technology has been widely used in many industries including agriculture, environment, food processing, pharmaceutical and biomedicine [3-6]. NIR spectrometry permits the prediction of many soil properties from the measurements. It has become popular in field measurement for in situ prediction of various soil properties [7-9]. In particular, diffuse reflectance of NIR spectroscopy is sensitive to the composition of organic carbon in soil [10-12]. Soil organic carbon (SOC) is an important component in agro-ecological soil. It represents a key parameter in evaluating the fertility of soils. SOC can be commonly and successfully predicted by means of NIR diffuse reflectance spectroscopy under laboratory controlled conditions [13-14].

The use of NIR spectroscopy for prediction of soil organic carbon (SOC) content in the field is highly desirable for soil quality assessment and carbon accounting purposes [15-17]. It is demonstrated that SOC can be measured in the field with enhanced measurement stability, in the expense of a slight decrease of accuracy compared to laboratory experiments [18-20]. Once the spectra have been calibrated for SOC, the chemometric methods can provide rapid and inexpensive estimation of SOC in the field.

NIR analysis of SOC requires a modeling-validation division for real samples. Validation samples are utilized for model evaluation, and the strategy of modeling optimization should be discussed based on a calibration-prediction sample partitioning [21-22]. In multivariate calibration problems involving the complex analytes, it is difficult to reproduce the composition variability of samples by means of optimized experimental designs [23]. For the procedure of calibration-prediction modeling, differences in the partitioning of calibration and

prediction sets will lead to fluctuations in modeling parameters and thus yielding unstable results. Especially, changes of the numbers of real samples in calibration set and in the prediction set will influence the robustness of the calibration models, so that the prediction results will be unstable and the modeling optimization is hard to achieve. In such cases, a representative calibration set is intensively desired, which must be extracted from the sample pool by considering the randomness, similarity and stability of calibration-prediction sample partitioning. This refers not only to the samples but also to the partitioning ratio. Therefore, it is an important research hot spot that a tunable ratio of sample numbers for the partitioning of calibration and prediction sets.

Multivariate techniques take the spectrum into account and exploit the multi-channel nature of spectroscopic data to provide the signals of organic carbon from the spectral response of soil. Extraction of quantitative information requires use of a reliable multivariate calibration method [24-25]. Linear regression methods (e.g. principal component regression, PCR, and partial least-squares, PLS) showed their ability to output promising results in specific applications [26-27]. However, agricultural translation of the effective chemometrics in NIR analysis has been largely impeded by the variations in the measurement [28-29]. Linear approaches cannot meet the quantitative modeling accuracy because the spectroscopic analysis of a single component in complex systems (such as soil) is influenced by the responses of other components and noises. Several investigators have recently employed least squares support vector machine regression (LS-SVR), a nonlinear multivariate method that can handle ill-posed problems and lead to unique global models [30-32]. Several studies addressed the issue of improvement in prediction (or classification) accuracy arising from the use of LS-SVR in relation to conventional linear methods.

In this study, we emphasize the investigation of calibration-prediction sample partitioning in the NIR analysis of SOC with multivariate chemometrics. A framework for calibration-prediction partitioning is proposed with the consideration of the tunable ratio of numbers of calibration and prediction samples. First of all, a fixed-number portion of samples is randomly selected as the independent validation set, which should not be subjected to the modeling process. The remaining samples, with a dependently fixed number, were carried on for the process of modeling optimization. It is worth noting that the stable and robust calibration model depends on the partitioning ratio of the modeling samples divided into calibration and prediction sets by considering the randomness, similarity and robustness of the framework. We addressed this issue in the multivariate calibration problem involving NIR spectrometric analysis of organic carbon in soil. For illustration, the total number of modeling samples is fixed because, as abovementioned, the number of the independent validation samples is firstly identified. Therefore, the numbers of samples respectively in the calibration and the prediction sets can be changed when the tunable partitioning ratio varies. A group of calibration and prediction sets corresponding to an unchanged partitioning ratio is marked as a volunteer group. The study involved the comparison of the different volunteer groups of calibration and prediction samples. The models obtained in this manner can be compared in terms of their modeling performance. In each volunteer group, the calibration-prediction partitions can be performed for many times with random selection of calibration samples. Based on the varied calibration-prediction partitions, the LS-SVR models can be parametric optimized by considering the stability and robustness based on the fundamental statistical theory of mean value and standard deviation. Further the optimal volunteer group can be chosen in all comparative partition ratios. Consequently, the optimized calibration model with correspondent optimal volunteer group can be estimated and evaluated by the samples in the independent validation set.

2. Theories and Algorithms

2.1 The framework of Sample Partitioning and model optimization

A framework for sample partitioning is proposed to take into account the tunable ratio of numbers of calibration and prediction samples, in consideration with the randomness, stability and robustness of calibration models.

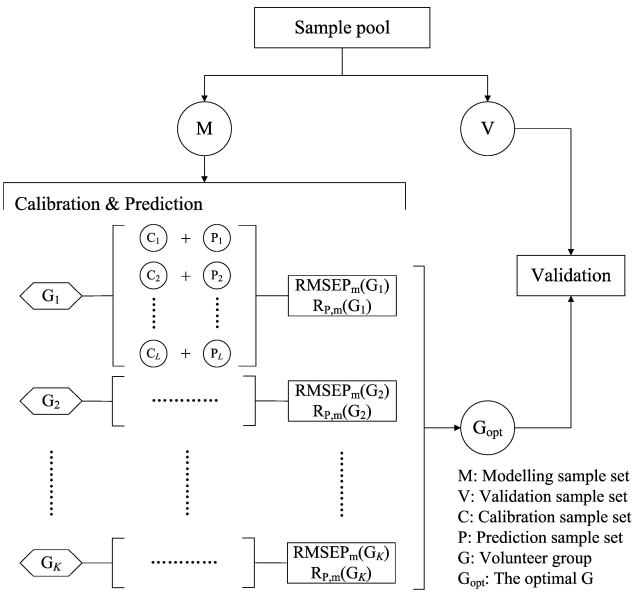


Fig 1 The algorithmic flow chart for the framework of the sample partitioning and model optimization

The algorithmic flowchart of this framework is showed in **Fig 1**. As can be seen in **Fig 1**, in the modeling-validation procedure, a fixed-number portion of validation samples (V set) is firstly extracted from the initial sample pool of the experimental data, before the calibration-prediction partitioning. For the purpose of ensuring the independence, the validation samples are randomly selected and totally not subjected to the modeling process. Then the remaining samples (M set), with a determined number, are partitioned into calibration set and prediction set, and further used for the model establishment and parameter optimization. It is worth noting that the change in the numbers of calibration samples will result in modeling differences, so as to affect the validating effects, thus a discussion of the partitioning ratio for calibration and prediction samples is quite necessary when

considering the stability and robustness of this framework. For illustration, a fixed partitioning ratio will generate a pair of calibration and prediction sets, which are marked as a volunteer data group (G sets). We try to change the partitioning ratio of calibration-prediction samples and perform model establishment and parametric optimizational process for each volunteer group (G_1, G_2… G_K). By comparing the modeling results, we can find out an optimized volunteer group (G_{opt}) to guarantee the stability and robustness of models.

The calibration-prediction process is carried out respectively for each volunteer group (G_k, k=1, 2… K). We noted that the volunteer data groups (G sets) are only related to a fixed partitioning ratio, which means that one volunteer group only determines how many samples are used for calibration and how many for prediction. There is still another issue for sample partitioning in each specific volunteer group: which sample for calibration can provide an improved modeling result? Several experimental evidences indicate that difference in sample partitioning of calibration and prediction sets will lead to fluctuations in predictive parameters and thus yielding unstable results [33-35]. On the level of stability and robustness, this issue requires this partitioning be randomly carried out for many times [36-37], resulting in many different pairs of calibration set and prediction set (C_l+P_l, l=1, 2… L). For the varied partitions of the calibration and prediction sets, the analytical models are established and the parameters are optimized by considering the modeling stability and robustness based on the mean value and standard deviation of the model indicators. For illustration, the root mean square error of prediction (RMSEP) and correlation coefficients of prediction (R_P) are taken as two important indicators for models. In one specific volunteer group (G_k), the modeling results for the pair of C_l+P_l are evaluated by RMSEP(l) and R_P(l), particularly. Going through all pairs of partitions in the fixed G_k, we have one RMSEP value and one R_P value for each pair of C+P. Based on all partitions in G_k, the mean value and standard deviation of all RMSEP's and R_P's are calculated and denoted as $RMSEP_m(G_k)$, $R_{P,m}(G_k)$, $RMSEP_{sd}(G_k)$ and $R_{P,sd}(G_k)$. As for statistical reasons, the relative standard deviation (RSD) was proposed here to evaluate the actual frustration accompanied with the mean values. The RSD values of RMSEP and R_P could be calculated and denoted as

$$RMSEP_{rsd}(G_k)=\frac{RMSEP_{sd}(G_k)}{RMSEP_m(G_k)},$$

$$R_{P,rsd}(G_k)=\frac{R_{P,sd}(G_k)}{R_{P,m}(G_k)}.$$

The $RMSEP_m(G_k)$ and $R_{P,m}(G_k)$ are used for evaluating the prediction accuracy of G_k and the $RMSEP_{rsd}(G_k)$ and $R_{P,rsd}(G_k)$ are for the modeling stability. According to this strategy, we can calculate $RMSEP_m$ and $R_{P,m}$ for each volunteer group (G_k, k=1, 2… K). All of the RMSEP (or R_P) values of each volunteer group will be located in the designated numerical region of $RMSEP_m \times (1 \pm RMSEP_{rsd})$ (or $R_{P,m} \times (1 \pm R_{P,rsd})$). We have the knowledge that a lower $RMSEP_m$ (or alternatively a higher $R_{P,m}$) indicates higher accuracy for the model and lower $RMSEP_{rsd}$ and $R_{P,rsd}$ reflect higher modeling stability. Therefore, by comparing the values of $RMSEP_m$ and $R_{P,m}$, we can select the optimal volunteer group (denoted as G_{opt}), which are expected to give a prospective promising result if utilized in validation process.

2.2 The theory of LS-SVR

LS-SVR algorithm employs a set of linear equations to reduce the complexity of optimization process associated with the SVR methodology [38]. For the NIR spectral data, the predictive concentration $\hat{c}_j$ of the j-th prediction sample is expressed in the following manner,

$$\hat{c}_j = \sum_{i=1}^{m} \alpha_i \varphi(A_j^P, A^C), \text{ and } \alpha_i = \left((A_i^C)^T A_i^C + \frac{1}{2\gamma} \right)^{-1},$$

where α_i is the Lagrange multiplier which depends on the regularization parameter γ [32], $\varphi(x_j, x_i)$ is the kernel function, A_j^P is the NIR spectrum of the j-th sample in the prediction set, and A^C is a linear combination of all the calibration spectra (the NIR spectra with m wavenumbers), weighted by the concentration values.

The distribution of the feature samples in high dimensional space depends on the selection of the kernel and corresponding parameters. The Gaussian radial basis function (RBF) kernel has moderate robustness and stability to enable nonlinear modeling for the acquired NIR dataset, and it is expressed as follows:

$$\varphi(A_j^P, A_i^C) = \exp\left(-\frac{(A_j^P - A_i^C)^2}{2\sigma^2} \right),$$

where σ^2 represents the kernel width and is used to tune the degree of generalization. When we select RBF as kernel, the performance of LS-SVR primarily depends on the selection of parameters γ and σ^2. The regularization parameter γ determines the trade-off between the training error (which can be thought of as the model accuracy in the calibration dataset) and the model robustness. To optimize these two parameters, we proposed a multi-scale interactive grid search is performed to enable the development of suitable calibration models. Careful selection of γ and σ^2 is quite necessary to search for a smooth subarea to obtain a low prediction error.

3 Samples and data

One hundred and thirty-five soil samples were collected from three farmlands in Guangxi (one autonomous region, China). In all cases the soils were under pure wheat or white rice or associated with other species, such as sweet potato. Approximately 10% of samples came from red soils and the rest of samples were the common yellow soils. The 135 sites were located depending on the area of each farmland. Based on the principle of homogeneous distribution, we chose 38, 45, 52 sites respectively from the small, the medium and the large farmland. The distances between each adjacent site were slightly different, ranging about 3 to 5 meters. At each site, 10-15 cores were extracted from 0-15 cm in depth. Each core was weighed about 2 grams and these cores were mixed together to comprise a sample. All samples were numbered successively from 1 to 135. The samples were firstly dried and finely ground in laboratory, and then passed through a 0.5-mm soil sifter, so as to ensure that the samples were refined to average small-size solid particles. Two equivalent sets weighing 10 grams were then extracted from each sample, with one set for biochemical measurement and the other for spectroscopic detection. The SOC content of each sample was

measured by using the routine biochemical method of potassium dichromate oxidation [39]. The measured values of all samples, in statistics, ranged from 1.10 to 6.42 (%), with an averaging value of 2.686 (%) and a standard deviation of 1.056 (%). These laboratorial values were used for spectroscopic analysis with the investigation of calibration-prediction-validation sample divisions.

The measurement of spectral data was performed by using Spectrum One NTS FT-NIR spectrometer (PerkinElmer Inc., USA). The inner part of a spectrometer is optical system composed by several devices. The NIR light is produced by a built-in tungsten halogen light source, going through a light-splitting unit so that the light is split into a series of point lights corresponding to each NIR wavelength. Then every point light one-by-one goes through the sample filled in a round sample cell. This is the key process of spectral scanning. In the sample cell, the light is absorbed and reflected by the sample particles and the out-coming light intensity becomes weakened. A diffuse reflectance accessory is equipped here to amplify the out-coming light. A pair of InGaAs detectors is monitoring the original and the weakened light information. The signal of NIR spectrum responses is generated by using the amplified original and weakened light information. Consequently, the spectrum goes through a Fourier transform amplitude analyzer and the signals are delivered to a computer for digital analysis.

The whole spectroscopic measurement should be conditioned throughout the spectral scanning process. The temperature was controlled at 25±1°C and the relative humidity was limited at the spot of 46±1%RH. The scanning range of the spectrum spanned 10000 cm^{-1} to 4000 cm^{-1} with a resolution of 8 cm^{-1}. Every sample was measured thrice and the average of the three measurements was further used for modeling. In this way, we had 135 average absorption spectra of soil (see **Fig 2**).

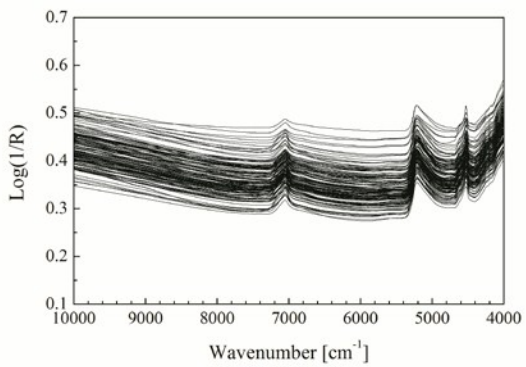


Fig 2 The near infrared spectra of 135 soil samples

4. Results and discussion

4.1. The NIR dataset

The whole scanning range of 10000-4000 cm^{-1} with the resolution of 8 cm^{-1} collected the NIR spectral responses at 1512 discrete wavenumbers per spectrum for each soil sample. The spectral absorbance includes the contributions of chemical components and also the noises arise from the light scattering and base-line drift, due to the sample particle factors (e.g. particle size and shape, thickness and tightness, etc.). Data pretreatment is quite indispensable for extracting the spectral signals. Multiplicative scatter correction was utilized for the pretreatment of the raw spectra of the calibration samples in each volunteer group.

The NIR dataset was constructed including the pretreated NIR data and the reference values of SOC. The whole sample pool was divided into calibration, prediction and validation sets by the framework of sample partitioning for model optimization and evaluation.

4.2. The Performance of Sample Partitioning on LS-SVR models

All the NIR data of 135 soil samples were prepared as the original sample pool. According to the calibration-prediction-validation procedure, 35 samples were randomly selected into the independent validation sample set (totally excluded in the modeling process), and the other 100 samples were remained. The statistic data of the validation samples and the remaining samples were showed in **Table 1**.

Table 1 The statistics for the randomly selected validation samples and the calibration-prediction samples

	Number of samples	SOC content (%)			
		Maximum value	Minimum value	Averaging value	Standard deviation
Validation	35	5.06	1.35	2.565	0.945
Calibration-prediction	100	6.42	1.10	2.728	1.093

The remaining 100 samples were further divided into calibration and prediction sets for modeling. We have to note that, for one thing, the samples for calibration should not be less than those for prediction, and for another, calibration samples should not be too much more than prediction samples to prevent over-fitting. Based on these concepts and on the framework of sample partitioning, we set the calibration-prediction partitioning ratio changed from 1:1 to 4:1. Practically, with the totally 100 modeling samples, we have the number for calibration changed from 50 to 80 with a step of 5, so that the number for prediction changed from 50 to 20 with a step of -5. Thus, we generated 7 different volunteer groups (i.e. K=7) and denoted them as G_1, G_2, G_3, G_4, G_5, G_6 and G_7. The detailed numbers for calibration and prediction samples in each volunteer group were listed in **Table 2**. For model establishment, we are planning to have the 100 modeling samples randomly divided into calibration set and prediction set according to the preset numbers designated in each volunteer group (see **Table 2**). Seven volunteer groups give seven different partitioning cases. We try to discuss which case will provide an optimal calibration model with the highest robustness.

As can be seen in **Table 2**, the volunteer groups are only related to the numbers of calibration and prediction samples, but a fixed number does not determine what samples for calibration and what samples for prediction because we used a random division strategy. This will raise the problem that different calibration samples influence the modeling results. To discuss this issue, we have the modeling sample set randomly partitioned for 30 times (i.e. L=30), obeying the preset partitioning numbers.

This operation would generate 30 different pairs of calibration and prediction sets (i.e. C_1+P_1, C_2+P_2… C_{30}+P_{30}) for each specific volunteer group (G_k, k=1, 2… 7). Calibration models were established for each pair of C_l+P_l (l=1, 2… 30) by using LS-SVR method with interactive grid search of the tunable parameters of γ and σ^2 (hereafter we discussed σ and successively easily get σ^2).

Table 2 The numbers of calibration and prediction samples in each volunteer group

Volunteer group	Number of calibration	Number of prediction
G_1	50	50
G_2	55	45
G_3	60	40
G_4	65	35
G_5	70	30
G_6	75	25
G_7	80	20

We have γ changing from 10 to 200 with a step of 10, and σ changing consecutively from 1 to 20. The LS-SVR models corresponding to each combination of (γ, σ) were established and the parameters of γ and σ were interactively optimized on a grid. Based on the 30 different pairs of C_l+P_l (l=1, 2… 30), the $RMSEP_m$ and $R_{P,m}$ were calculated as the stable modeling results corresponding to the interactive effect of γ and σ. The $RMSEP_{rsd}$ and $R_{P,rsd}$ were also calculated for evaluating the frustration of models. Successively, the optimal parameter combination can be found by searching for the minimum $RMSEP_m$ or the alternative maximum $R_{P,m}$. This optimal result was taken as the stable and robust modeling effect for the specific volunteer group G_k. Further, the optimal volunteer group can be selected by comparing the best values of $RMSEP_m(G_k)$ and $R_{P,m}(G_k)$ for the 7 volunteer groups (G_k, k=1, 2… 7). The LS-SVR models with the optimal parameter were further selected. **Table 3** showed the LS-SVR modeling results with the optimal parameters for the 7 designated volunteer groups. We can see from **Table 3** that G_4 output the minimum $RMSEP_m$ and a corresponding maximum $R_{P,m}$. And also the relatively low values of $RMSEP_{rsd}$ and $R_{P,rsd}$ demonstrated that the optimal stable LS-SVR model had little predicted frustrations. It could be concluded that the optimal volunteer group for NIR analysis of SOC is volunteer group G_4. The partition of 65 calibration samples and 35 prediction samples brought to the best prospective results.

Table 3 The optimal LS-SVR modeling results for the 7 volunteer groups

Volunteer group	γ	σ	$RMSEP_m$	$RMSEP_{rsd}$	$R_{P,m}$	$R_{P,rsd}$
G_1	120	8	0.283	0.197	0.900	0.159
G_2	110	6	0.261	0.187	0.916	0.154
G_3	100	7	0.258	0.190	0.923	0.148
G_4	110	7	0.247	0.185	0.937	0.147
G_5	130	8	0.254	0.205	0.932	0.155
G_6	120	10	0.269	0.214	0.909	0.161
G_7	100	9	0.285	0.217	0.885	0.175

For LS-SVR modeling, it is worth noting that the two parameters of γ and σ represent the regularization extension and the kernel width when using the RBF kernel. Particularly, we discussed the interactive grid searching of parameters based on the optimal volunteer group (G_4), with the projective insight of the influence from each separate tuning of γ and σ. The model predictive results corresponding to each value of γ were showed in **Fig 3** (**Fig 3(a)** distributes $RMSEP_m$ and $RMSEP_{rsd}$, and **Fig 3(b)** distributes $R_{P,m}$ and $R_{P,rsd}$). Similarly, the model predictive results corresponding to each value of σ were showed in **Fig 4** (**Fig 4(a)** distributes $RMSEP_m$ and $RMSEP_{rsd}$, and **Fig 4(b)** distributes $R_{P,m}$ and $R_{P,rsd}$).

It was seen from **Fig 3** that the $RMSEP_{rsd}$ and $R_{P,rsd}$ values varied on behave of γ, but most of them were smaller than 0.2, which demonstrated the modeling frustration was small enough and the models were taken as stable. The minimum $RMSEP_m$ was obtained when γ equals to 110, with a correspondingly highest $R_{P,m}$. And **Fig 4** showed that most $RMSEP_{rsd}$ and $R_{P,rsd}$ derived from every value of σ were also smaller than 0.2, which in another aspect revealed the modeling stability and robustness. And the minimum $RMSEP_m$ was obtained when σ equals to 7, with the correspondingly highest $R_{P,m}$. In summary, we have the optimal LS-SVR parameters (γ, σ) were (110, 7), and the optimal $RMSEP_m$ and $R_{P,m}$ were 0.247% and 0.937, respectively. This optimal modeling result was obtained by the nonlinear LS-SVR algorithm based on the calibration-prediction sample partitioning with different ratio. We concluded that the optimal model with (γ, σ) equaling to (110, 7) were presented stable and robust for calibrations in NIR analysis of SOC.

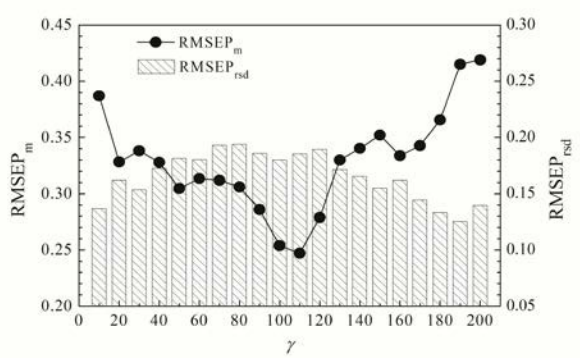


Fig 3 sub-figure (a)

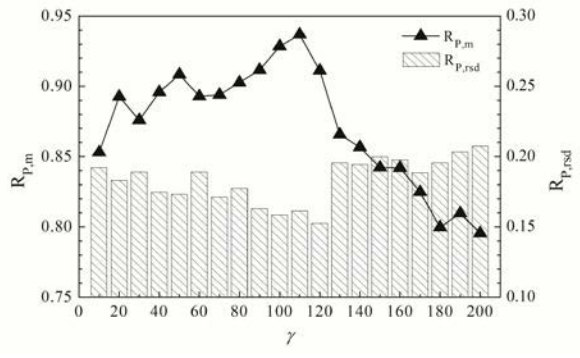


Fig 3 sub-figure (b)

Fig 3 The model predictive results corresponding to each value of γ in LS-SVR modeling (sub-figure(a) distributes RMSEPm and RMSEPrsd, and sub-figure(b) distributes RP,m and RP, rsd)

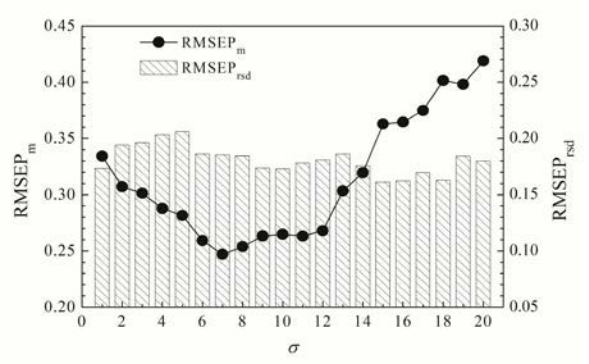


Fig 4 sub-figure (a)

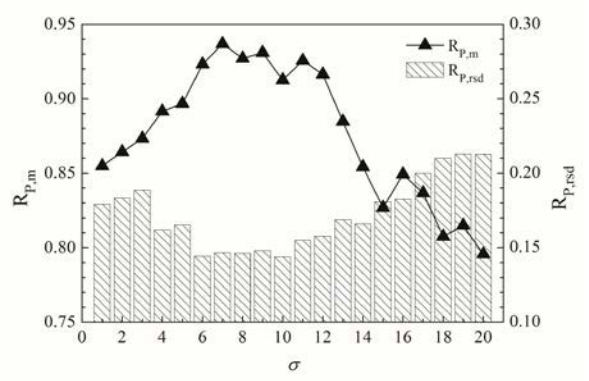


Fig 4 sub-figure (b)

Fig 4 The model predictive results corresponding to each value of σ in LS-SVR modeling (sub-figure(a) distributes $RMSEP_m$ and $RMSEP_{rsd}$, and sub-figure(b) distributes $R_{P,m}$ and $R_{P,rsd}$)

4.3 Validation of the optimal LS-SVR model

The randomly selected 35 independent validation samples were used to evaluate the LS-SVR models respectively on the 7 volunteer groups, using the corresponding optimal parameters. The LS-SVR models were established by using the spectral data and actual SOC contents (measured by potassium dichromate oxidation). We found out the optimal parameters and determined the model regressive coefficients in the calibration-prediction process. Further the NIR predicted values for the 35 validation samples can be estimated by fitting the NIR data into the model and using the coefficients. The NIR predicted values of SOC in each volunteer group were obtained and the RMSEV and R_V for the 35 validation samples were showed in **Table 4**. The validation process owns the objectiveness and representiveness as the validation samples were totally excluded in the modeling optimization process. We observed in **Table 4** that the optimal modeling volunteer group G_4 output the optimal validation results, The predicted values were close to the actual contents, with the minimum RMSEV of 0.302 (%) and the corresponding high R_V of 0.907. The correlation relationship between the predicted values and actual contents was showed in **Fig 5**. The results showed that the predicted values and the actual contents were highly correlated for SOC. The validation effect was much satisfactory for the random validation samples because we have achieved the modeling stability and robustness in the process of model optimization, with a prospective choice of the volunteer group for calibration-prediction partition.

Table 4 The LS-SVR modeling results for validation samples based on the optimal parameters in each volunteer group

Volunteer group	$RMSEV_m$	$RMSEV_{rsd}$	$R_{V,m}$	$R_{V,rsd}$
G_1	0.361	0.256	0.857	0.180
G_2	0.332	0.244	0.872	0.183
G_3	0.335	0.244	0.888	0.187
G_4	0.302	0.239	0.907	0.180
G_5	0.330	0.252	0.899	0.181
G_6	0.342	0.277	0.870	0.184
G_7	0.351	0.282	0.855	0.190

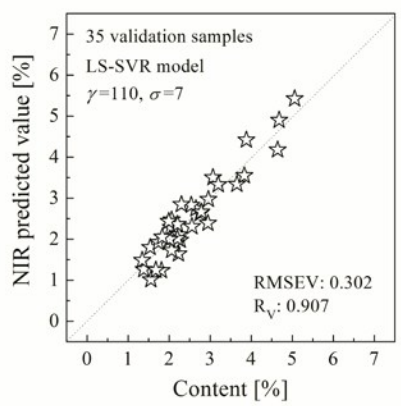


Fig 5 The correlation relationship between the NIR predicted values and actual contents of SOC

4 Conclusions

NIR analysis requires a modeling-validation division for real samples. Differences in the partitioning and changes of the numbers of real samples in calibration set and in the prediction set lead to fluctuations and influence the stability and robustness in modeling parameters and thus yielding unstable results. A representative calibration set must be extracted from the sample pool with the considerations referring to not only the samples but also the partitioning ratio. In our work, the strategy of modeling optimization is proposed based on a calibration-prediction sample partitioning. A framework for sample partitioning is built up with the consideration of the tunable ratio of numbers of calibration and prediction samples, aiming to confirm the modeling stability and robustness. We addressed this issue in the multivariate calibration involving NIR spectrometric analysis of SOC.

We created 7 volunteer groups (G_k, k=1, 2… 7) for different ratio of calibration-prediction partitioning. For each G_k, the calibration-prediction sample partition was randomly carried out for 30 times. The LS-SVR models were established and the optimal parameters were selected for each single partition. By considering the stability and robustness, we calculated the $RMSEP_m(G_k)$ and $R_{P,m}(G_k)$, as well as the $RMSEP_{rsd}(G_k)$ and $R_{P,rsd}(G_k)$ based on the 30 different calibration-prediction partitions in the specific G_k. Moreover, we optimized the LS-SVR modeling parameters of γ and σ in an interactive grid search way, and successively we found out the optimal volunteer group as G_4 by comparing all the 7 values of $RMSEP_m$. The optimal LS-SVR

parameters (γ, σ) were (110, 7), and the optimal $RMSEP_m$ and $R_{P,m}$ were 0.247% and 0.937, respectively. The values of $RMSEP_{rsd}$ and $R_{P,rsd}$ were small enough to confirm that the models were stable and robust.

Further, the optimized calibration model was evaluated by the independent validation samples, respectively for each of the 7 volunteer groups. The out-of-modeling validation effects were much satisfactory for the random validation samples, and the validation optimal volunteer group was also selected as G_4. We conclude that we have achieved the modeling stability and robustness in the process of model optimization base on the discussion of the tunable ratio of numbers of calibration and prediction samples.

Acknowledgment

This work was supported by the National Natural Scientific Foundation of China (61505037), the National Spark Program of China (2014GA780009), the Pearl River S&T Nova Program of Guangzhou (201506010035) and the University Scientific Research Project of Guangxi Education Office (KY2015ZL095).

Notes and references

[a] *College of Science, Guilin University of Technology, Guilin 541004, China.*
[b] *School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, 510225, China*
[c] *School of Ocean, Qinzhou University, Qinzhou, 535000, China*
[*] *Corresponding Author. E-mail: kencaizhku@foxmail.com (K. Cai)*
DOI: 10.1039/b000000x/

1. M.C. Sarraguça, A. Paulo, M.M. Alves, A.M. Dias, J.A. Lopes and E.C. Ferreira, *Anal. Bioanal. Chem.*, 2009, **395**, 1159-1166.
2. C. Collell, P. Gou, J. Arnau and J. Comaposada, *Food Chem.*, 2011, **129**, 601-607.
3. V.R. Sinija and H.N. Mishra, *LWT - Food Sci. Technol.*, 2009, **42**, 998-1002.
4. A. Saleem, C. Canal, D.A. Hutchins, L.A.J. Davis and R.J. Green, *Anal. Methods*, 2011, **3**, 2298-2306.
5. M. Soto-Barajas, I. Gonzalez-Martin, J. M. Hernandez-Hierro, B. Prado, C. Hidalgo and J. Etchevers, *Anal. Methods*, 2012, **4**, 2764-2771.
6. H. Chen, W. Ai, Q. Feng, Z. Jia and Q.Q. Song, *Spectrochim. Acta A*, 2014, **118**, 752-759.
7. R. Rinnan and A. Rinnan, *Soil Biol. Biochem.*, 2007, **39**, 1664-1673.
8. M. Nocita, A. Stevens, C. Noon and B. van Wesemael, *Geoderma*, 2013, **199**, 37-42.
9. B. Stenberg, R.A. Viscarra Rossel, A.M. Mouazen and J. Wetterlind, *Adv. Agron.*, 2010, **107**, 163-215.
10. A.M. Mouazen, M.R. Maleki, J. De Baerdemaeker and H. Ramon, *Soil Till. Res.*, 2007, **93**, 13-27.
11. R.A. Viscarra Rossel, S.R. Cattle, A. Ortega and Y. Fouad, *Geoderma*, 2009, **150**, 253-266.
12. B. Minasny, A.B. McBratney, V. Bellon-Maurel, J.M. Roger, A. Gobrecht, L. Ferrand and S. Joalland, *Geoderma*, 2011, **167-168**, 118-124.
13. L.K. Sorensen and S. Dalsgaard, *Soil Sci. Soc. Am. J.*, 2005, **69**, 159-167.
14. H.T. Xie, X.M. Yang, C.F. Drury, J.Y. Yang and X.D. Zhang, *Can. J. Soil Sci.*, 2011, **91**, 53-63.
15. R.S. Bricklemyer and D.J. Brown, *Comput. Electron. Agr.*, 2010, **70**, 209-216.
16. T.H. Waiser, C.L.S. Morgan, D.J. Brown and C.T. Hallmark, *Soil Sci. Soc. Am. J.*, 2007, **71**, 389-396.
17. C.D. Christy, *Comput. Electron. Agr.*, 2008, **61**, 10-19.
18. Stevens, B. Van Wesemael, H. Bartholomeus, D. Rosillon, B. Tychon and E. Ben-Dor, *Geoderma*, 2008, **144**, 395-404.
19. P.V. Ajayakumar, D. Chanda, A. Pal, M.P. Singh and A. Samad, *J. Pharmaceut. Biomed.*, 2012, **58**, 157-162.
20. H. Chen, Q. Feng, Z. Jia and Q.Q. Song, *Asian J. Chem.*, 2014, **26**, 4839-4844.
21. R.K.H. Galvao, M.C.U. Araujo, G.E. Jose, M.J.C. Pontes, E.C. Silva and T.C.B. Saldanha, *Talanta*, 2005, **67**, 736-740.
22. M. Silva, M.H. Ferreira, J.W.B. Braga and M.M. Sena, *Talanta*, 2012, **89**, 342-351.
23. M. Zeaiter, J. M. Roger and V. Bellon-Maurel, *TrAC Trends Analyt. Chem.*, 2005, **24**, 437-445.
24. X. Shao, X. Bian, J. Liu, M. Zhang and W. Cai, *Anal. Methods*, 2010, **2**, 1662-1666.
25. Z.P. Chen, L.J. Zhong, A. Nordon, D. Littlejohn, M. Holden, M. Fazenda, L. Harvey, B. McNeil, J. Faulkner and J. Morris, *Anal. Chem.*, 2011, **83**, 2655-2659.
26. S.R. Delwiche and J.B. Reeves III, The effect of spectral pre-treatments on the partial least squares modelling of agricultural products, *J. Near Infrared Spec.*, 2004, **12**, 177-182.
27. B. Igne and C.R. Hurburgh Jr., *J. Chemom.*, 2010, **24**, 75-86.
28. N.C. Dingari, I. Barman, G.P. Singh, J.W. Kang, R.R. Dasari and M.S. Feld, *Anal. Bioanal. Chem.*, 2011, **400**, 2871-2880.
29. H. Chen, G. Tang, Q. Song and W. Ai, *Anal. Lett.*, 2013, **46**, 2060-2074.
30. A. Borin, M. F. Ferrao, C. Mello, D.A. Maretto and R.J. Poppi, *Anal. Chim. Acta*, 2006, **579**, 25-32.
31. R.G. Brereton and G.R. Lloyd, *Analyst*, 2010, **135**, 230-267.
32. I. Barman, N.C. Dingari, G.P. Singh, J.S. Soares, R.R. Dasari, J.M. Smulko, *Anal. Chem.*, 2012, **84**, 8149-8156.
33. H. Chen, T. Pan, J. Chen and Q. Lu, *Chemometr. Intell. Lab.*, 2011, **107**, 139-146.
34. T. Pan, Z. Chen, J. Chen and Z. Liu, *Anal. Methods*, 2012, **4**, 1046-1052.
35. Z. Liu, B. Liu and T .Pan, *Spectrochim. Acta A*, 2013, **102**, 269-274.
36. H.Z. Chen, Q.Q. Song, G.Q. Tang and L.L. Xu, *J. Cereal Sci.*, 2014, **60**, 595-601.
37. H.Z. Chen W. Ai, Q.X. Feng and G.Q. Tang, *Anal. Methods*, 2015, **7**, 2869-2876.
38. N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, New York, USA, 2000.
39. R.K. Lu, *Methods for chemical analysis of soil agriculture*, China agricultural science and technology press, Beijing, China, 2000.