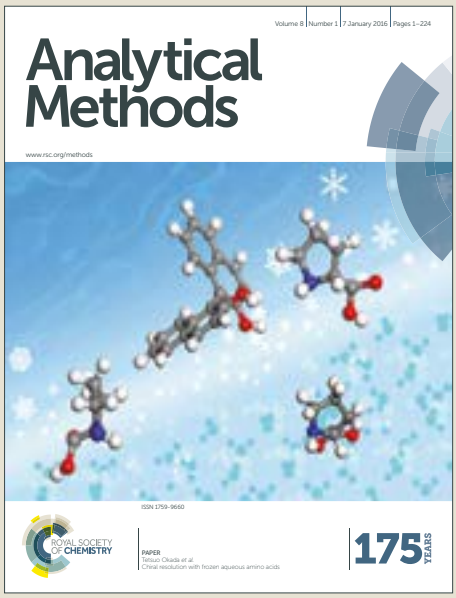
Analytical Methods

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: H. Chen, Z. Liu, K. Cai, J. Gu, W. Ai and J. Wen, *Anal. Methods*, 2018, DOI: 10.1039/C8AY01076E.



ROYAL SOCIETY OF CHEMISTRY

rsc.li/methods

View Article Online
DOI: 10.1039/C8AY01076E

Journal Name

ARTICLE

Quantitative Analysis of Soil Nutrition Based on FT-NIR Spectroscopy Integrated with BP Neural Deep Learning

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Huazhou Chen,[ab] Zhenyao Liu,[bc] Ken Cai,[*d] Jie Gu,[a] Wu Ai,[a] and Jiangbei Wen[b]

A framework of back propagation neural deep learning (BPN-DL) was constructed in this work for Fourier transform near-infrared spectroscopy (FT-NIR) to predict the nutrition components in soil samples. Characteristic wavenumbers were selected by competitive adaptive reweighted sampling (CARS) algorithm, to be the input variables to the BPN-DL framework. With the undergoing computer hard configuration, BPN-DL models were established and pre-set screening for up to 32 hidden layers and 50 nodes. The results were achieved in iteration and parameter identification. The best optimal BPN-DL model was constructed by 22 hidden layers and 30 neural nodes, with 91 input wavenumbers selected by CARS. The root mean square error of training was 0.104 and that of testing was 0.279. Another available optimal model was with 19 hidden layers and 46 nodes for 216 characteristic wavenumbers. The optimal results were fruther compared with the benchmark PCR, PLSR and conventional back propagation network models. This study indicated that the FT-NIR analytical model can be optimized integrated with appropriate chemometric methods and the prediction accuracy can be improved. The BPN-DL framework reveals its superiority in model training and testing processes.

1. Introduction

Soil is an essential part of agricultural ecological environment. With the development of modern bio-information, the nutrition of soil should be well detected for precise planting of crops. Organic carbon (OC) is one of the most important nutritive components measuring the fertility of soil.[1-2] In agricultural production process, the prediction of OC of soil can not only help people know the period of soil fertilization, but also can be used to more accurately select termination time of crop harvest to maximize production, and also can compare different regions of soil status to choose the best time of manure. Therefore, it has important guiding significance to the precise agriculture. At present, the main methods to measure the concentration of OC in soil are the routine biochemical measurements usually performed in the laboratory. However, these methods are tedious in the detection process, and need to use a certain number of chemical reagents, so that will lead to destruction of samples, environmental pollution and waste of time.[3] Therefore, it is of great significance in modern agriculture to develop a rapid, reliable and reagents-free analytical method to predict the OC concentration to assure the nutrition in soil for crop precise planting.

Fourier transform near infrared (FT-NIR) spectroscopy can possibly serve as a noninvasive technique for quantitative analysis of the concentration of OC in soil, as it interacts with some functional groups, such as biomass (C-H group), organic acid and moisture (O-H group) and scattering from microstructures.[4-5] Most of the absorptions in the near-infrared region associated with these groups are the first overtone, second overtone or combination bands, due to the vibrational and rotational transition of molecules.[6-7] The use of NIR technology for the analysis of soil components has been a significant research direction.[8-9] There have been many researches on NIR/FT-NIR spectroscopy analysis of soil in recent years. These works show that the FT-NIR spectroscopy technique has a high potential to analyze the nutrition of soil.[10-13]

However, the FT-NIR spectroscopy analysis technique is an indirect technology. Soil is a complex system with multiple components. The near infrared spectrum of soil has complicated backgrounds with peak overlapping and weak signal, and contains a lot of noise and interference. In addition, the FT-NIR spectral responses have hundreds of variables including the uninformative variables, the redundant variables and the variables with serious collinearity. If these spectral data wasn't dealt with, it will not only increase the amount of calculation, but also interfere with useful information in the process of modeling, so as to reduce the model prediction accuracy.[14-15] Therefore, it is the important issues asking for further study to select the appropriate spectral pretreatment method and to choose the effective chemometric method for wavelength selection, in order to reduce noise, and to improve the accuracy of FT-NIR analysis of soil nutrition.

[a] College of Science, Guilin University of Technology, Guilin 541004, China
[b] Guangdong Spectrastar Instruments Co. Ltd., Guangzhou 510663, China.
[c] Guangzhou Research Institute of O-M-E Technology, Guangzhou 510663, China.
[d] College of Automation, Zhongkai University of Agriculture and Engineering, Guangzhou, 510225, China.
* Correspondance at: College of Automation, Zhongkai University of Agriculture and Engineering, Zhongkai Road 501#, Guangzhou, 510225, China. E-mail: kencaizhku@foxmail.com (K. Cai).

View Article Online
DOI: 10.1039/C8AY01076E

Therefore, the application of a proper multivariate analysis method to model calibration has been proved be greatly beneficial in providing more reliable and parsimonious model. During the last few decades, many linear and nonlinear algorithms have been developed for model calibration, such as principle component regression (PCR),[16] partial least squares regression (PLSR),[17] extreme learning machine,[18] support vector machine,[19-20] and artificial neural network (ANN),[20-21] of which, ANN has the advantages of smaller calculation quantity and few parameters, it is suitable for dealing with the problem of non-normal distribution.[22-23] The criterion of optimization in the neural network is to make the error of the training set or the test set the smallest.[24] A back propagation neural network (BPN) would take the output errors backward to optimize the weights in the input layer and the hidden layers, so as to extract the comprehensive signal of the input variables.[25-26] With the concepts of deep learning and machine learning,[27-29] the BPN algorithm can be further optimized in the deep learning mode, by tuning and selecting the number of hidden layers and the number of nodes in each layer. Thus, an optimal back propagation network model can be determined for the evaluation of analyte, so that the predictive accuracy of the calibration model can be improved by deep learning. The test of model robustness is unavoidable for rapid analytical technology and it is associated with the inherent spectral noise in the prediction sample and the spectral makeup of the model.[30-31] Fortunately, model robustness can be evaluated by analysing model uncertainty. It is associated with repeating the measurement many times and analyzing the standard deviation of parameters extracted from each of these multiple measurements.[32] An approach was applied to determine the modeling uncertainty to provide physical insight into the sources of model uncertainty.[33-34]

In this work, the BPN deep learning (BPN-DL) technique was applied to build a regression model for prediction of the concentration of OC in soil samples. FT-NIR spectroscopy was used to quantitatively determine the concentration of OC in soil samples. Some proper multivariate data analysis methods were employed, with parameters optimized, to select informative variables to improve the prediction accuracy and to validate the model robustness. The specific objectives of this work were:

(1) to eliminate suspended particles, surface astigmatism and optical path change by multiplicative scatter correction (MSC).

(2) to select the informative variables from the raw spectral matrix and reduce the data dimension by the competitive adaptive reweighted sampling (CARS) algorithm.[35]

(3) to use optimally selected spectral variables to build calibration models by a back propagation neural deep learning (BPN-DL) platform, train it with the number of hidden layers and the number of nodes in each layer both tunable.

(4) to evaluate the model robustness by analysing the model uncertainty based on the testing part samples.

In order to highlight the superiority of the prediction precision of BPN-DL algorithm adopted in this work, the results of BPN-DL model were compared with the benchmark PLSR and PCR results. Simultaneously, the each step in process of the model calibration was discussed systematically, and parameters of the models were optimized by a cross-validation.

2. Materials and methods

2.1 Soil samples and nutrition measurement

One hundred and thirty-five soil samples were collected from three farmlands in Guangxi (one autonomous province in 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 OC concentration of each sample was measured by using the routine biochemical method of potassium dichromate oxidation.[36] 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 (%). The biochemically measured concentrations of OC were used as the reference for establishing spectroscopic calibration models applied with the use of BPN-DL framework.

2.2 FT-NIR spectra acquisition

FT-NIR spectral data were collected in reflectance mode with a Spectrum One NTS FT-NIR spectrometer (produced by PerkinElmer Inc. in USA). Each spectrum is an ensemble average of 32 scans in a round sample cell with a 6 cm diameter. The spectral data were acquired in the range from 10000 to 4000 cm^{-1}, which brought the spectral with 1512 variables (resolution: 8 cm^{-1}). In order to make the collected spectral data more accurate, each sample were scanned for three times and then take the mean of these three spectral data as the raw spectral data for the sample, which was employed to build analysis model. In FT-NIR spectra collection, the spectrometer was sensitive to the change of environment condition such as temperature and humidity. Therefore, the temperature was kept around 25°C and at a steady humidity level at 64±1% RH in the laboratory.

View Article Online
DOI: 10.1039/C8AY01076E

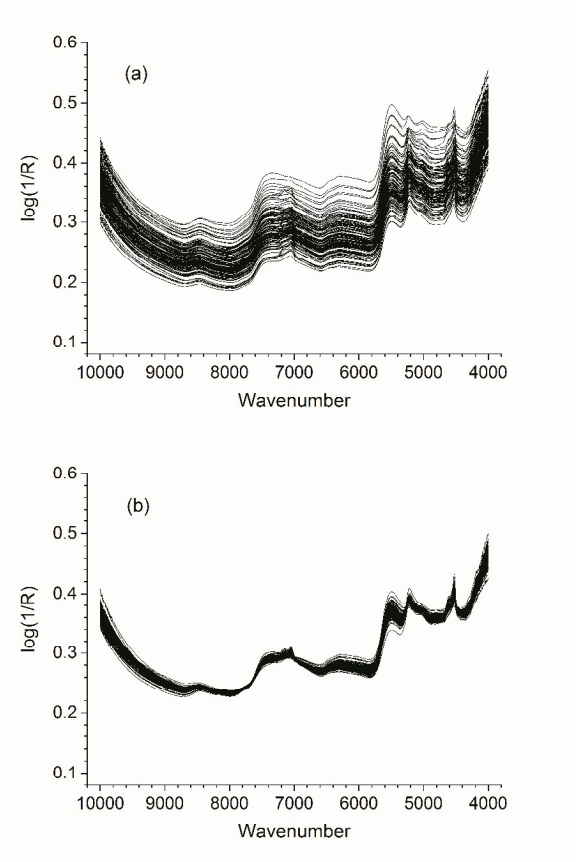


Fig.1 Raw spectra (a) and MSC preprocessed spectra (b) of 135 soil samples

Fig. 1(a) shows the raw FT-NIR spectra of the 135 soil samples. FT-NIR spectra are affected by multifarious conditions such as changes in temperature, diffusion of light, a baseline shift, or instrument noise.[37] In addition, FT-NIR spectra contain other chemical and physical information that should be taken as noise interference.[38] It is very important to select a suitable pretreatment method to weaken this interference to ensure the accuracy of the calibration model. Multiplicative scatter correction (MSC) performs in an averaging regression method to get rid of scattering signals.[39] It has advantages in correcting scatter light. Therefore, MSC was employed for light scatter correction and reducing the changes of light path length in this work. The MSC-preprocessed spectra are presented in **Fig. 1(b)**.

2.3 Methods for multivariate data analysis

(1) The CARS algorithm for variable selection

Recent studies on chemometrics indicated that variable selection methods are essential for multivariate data analysis.[40] The competitive adaptive reweighted sampling (CARS) algorithm is a well applicable algorithm for wavelength selection in fields of spectral analysis. It employs the principle "survival of the fittest" on which Darwin's Evolution Theory is based.[41] Studies show that CARS is very effective and fast adaptable in the selection of the spectral wavelengths.[42] Some weight values are generated for each wavelength variable and the weight values compete with each other instead of the wavelength variables. Some wavelength variables whose weights are relatively small can be removed. The detail procedure of CARS algorithm is shown as the following steps.[35]

Step 1: Sample division was executed by Monte Carlo randomness. Eighty percent of all samples were randomly selected for calibration. The common PLSR method was employed for training and the optimal model was determined. The regression coefficients were denoted as b_i for the i-th wavelength variable.

Step 2: the coefficients of the calibration model were retained and the weight values (w_i, i=1,2…p) were generated. The weight w_i corresponding to the i-th wavelength variable was defined as:

$$w_i = \frac{|b_i|}{\sum_{i=1}^{p}|b_i|}, \quad i = 1,2...p,$$

where p represents the total number of wavelength variables in the full scanning spectral range.

Step 3: An exponentially decreasing function was used to perform enforced wavelength selection. Wavelength retention rate (denoted as RATE) is calculated directly by the following exponential function:

$$\mathrm{RATE}_i = (\frac{p}{2})^{\frac{1}{n-1}} \times e^{-i\times\frac{\ln(\frac{p}{2})}{n-1}},$$

where p represents the total number of wavelength variables and n represents the number of training samples.

Step 4: Wavelength competitive selection was accomplished by adaptively re-weighting each variable. The wavelength variables with larger weights were selected to form the informative subsets. After iterating K times, CARS sequentially created K subsets of informative wavelengths to build PLSR models. Leave-one-out cross validation was utilized to evaluate each of the K subsets. The optimal subset was chosen aiming to reach the lowest value of root mean square error of cross validation ($\mathrm{RMSE_{CV}}$).

(2) The modeling framework of BPN-DL

There are increasing evidences showing that a three-layer back propagation neural network (i.e., including only one hidden layer) is enough to simulate almost all complicated nonlinear functions, and the instability of the network grows with the growth of hidden layers.[43] In this section, the basic procedure of BPN network was simply introduced, and the BPN-DL framework was constructed in the deep learning mode, by tuning and selecting the number of hidden layers and the number of nodes in each layer. The detail BPN-DL framework flow can be built as follows,

The BPN-DL dynamical framework is an algorithmic expansion of a basic BPN network. In the initial form of the network, the spectral data vectors (x_1, x_2… x_n) are introduced into the input layer, and processed in the hidden as well as the output layer. With the idea of deep learning, the number of

View Article Online
DOI: 10.1039/C8AY01076E
ARTICLE Journal Name

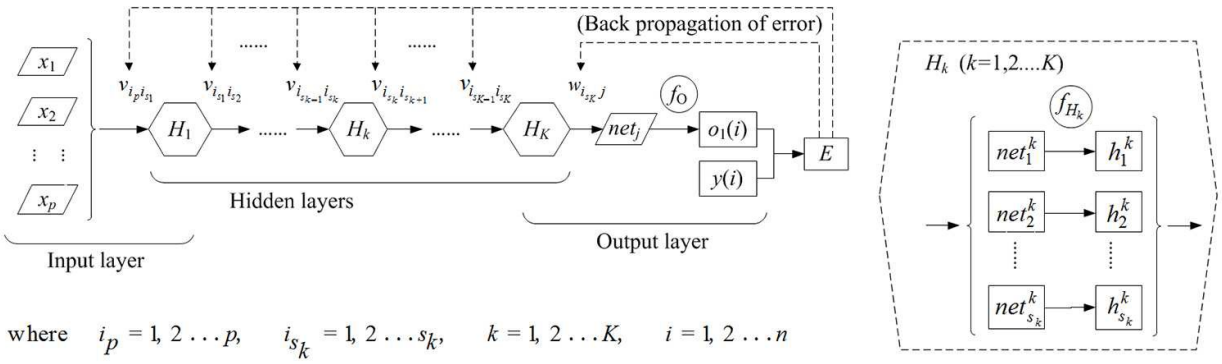


Fig. 2 The algorithmic scheme of BPN-DL framework (details of each H_k presented in the hexagon box)

hidden layers and back propagation circles can be unlimitedly expanded, and the number of nodes in each hidden layer can also be added to infinite, only in limited by the computational ability of current data processing unit. The framework of BPN-DL was constructed as shown in **Fig. 2**. The BPN-DL system was trained with different number of hidden nodes in multiple hidden layers as well as the corresponding number of back propagation cycles. At the beginning of a training run, the parameters of BPN network including learning coefficient, momentum and number of iterations were initialized with optional values. The network was trained in the way of tuning the weights and the transfer/activation functions. The training results would be obtained in the output layer, and further go to the back propagation circles. The deep learning mode would push the training mechanism in a long run to find the adjustable optimal values of weights, and establish an optimal calibration model so that the predictive accuracy could be eventually improved.

At the first hidden layer, the data input for each node is defined as follows:

$$net_{i_{s_k}}^k = \sum_{i_p=1}^{p} v_{i_p i_{s_k}} x_{i_p} + a_{i_{s_k}}, \quad i_{s_k} = 1,2...s_k, k = 1,2...K,$$

where $v_{i_p i_{s_k}}$ is the weight between the i_p-th node of the input layer and the i_{s_k}-th node of the hidden layer, and $a_{i_{s_k}}$ is the bias of the i_{s_k}-th node. The output of the i_{s_k}-th node is defined as follows:

$$h_{i_{s_k}}^k = f_{H_k}(net_{i_{s_k}}^k), \quad i_{s_k} = 1,2...s_k, k = 1,2...K,$$

where f_{H_k} is the transfer function of the k-th hidden layer.

Similarly, at each node of the output layer, the data input is defined as follows:

$$net_j = \sum_{i_{s_K}=1}^{s_K} w_{i_{s_K} j} h_{i_{s_K}}^K + b_j, \quad j = 1,2...J$$

where K represent the last hidden layer (the value of K can be initially designated), $w_{i_{s_K} j}$ is the weight between the i_{s_K}-th node of the hidden layer and the j-th node of the output layer, and b_j is the bias of the j-th node. The output of the j-th node is defined as follows:

$$o_j = f_O(net_j),$$

where f_O is the transfer function of the output layer. Since only one dependent variable is demanded for prediction in our study, there is only one output node, thus we have j=1, net_1 is the BPN-DL network output comprehensive variable and o_1 is the output data.

The output value (o_1) is largely affected by values of the weights in the hidden layers ($v_{i_{s_{k-1}} i_{s_{k+1}}}$) and the output weights ($w_{i_{s_K} j}$), which are automatically adjusted by the network itself in terms of the resulted error that is propagated back during the training process,

$$E = \sum_{i=1}^{n} (o_1(i) - y(i))^2,$$

where E is a self-defined error function, $o_1(i)$ is the output predictive values of the calibration/validation samples, $y(i)$ is the actual values, and n is the number of the target samples in the calibration-validation round or in the testing round.

For the back propagation, the E will be pushed back to refine the weight of each neuron. In the (n+1)-th iterative step, the change of weight $\Delta w_{i_{s_K} j}$ between the i_{s_K}-th hidden node and the j-th output node is modified as follows:

$$\Delta w_{i_{s_K} j}^{n+1} = -\alpha \frac{\partial E}{\partial w_{i_{s_K} j}} + \beta \Delta w_{i_{s_K} j}^{n},$$

where α and β are the learning rate and momentum, respectively. And the change of weight $\Delta v_{i_{s_{k-1}} i_{s_{k+1}}}$ between the $i_{s_{k-1}}$-th input/hidden node and the $i_{s_{k+1}}$-th hidden node is modified as follows:

$$\Delta v_{i_{s_{k-1}} i_{s_{k+1}}}^{n+1} = -\gamma \frac{\partial E}{\partial v_{i_{s_{k-1}} i_{s_{k+1}}}} + \delta \Delta v_{i_{s_{k-1}} i_{s_{k+1}}}^{n},$$

where γ and δ are the learning rate and momentum, just like the α and β. Given the learning rate, momentum and initial value of each weight, the algorithm could be trained step by step automatically until the error function converges at the minimum value.

2.4 Model uncertainty

To evaluate the robustness of the established BPN-DL model, the model uncertainty is investigated using the method proposed by Scepanovic.[32] The measured spectrum ($a_{p\times1}$) is functional expressed as

$$a = Tb + \varepsilon,$$

View Article Online
DOI: 10.1039/C8AY01076E

where the $T_{p\times n}$ contains the model constituent vectors, $b_{n\times 1}$ contains the underlying coefficients of the model constituents, and $\varepsilon_{p\times 1}$ represents the noise.

Uncertainty analysis is to estimate the best regression coefficient b by curve fitting, and the model uncertainty is quantitatively identified by the standard deviation. Assuming that the Gaussian noise was set up by the Cramér-Rao lower bound,[44] we decomposed the matrix T,

$$T=HD,$$

where D is a diagonal matrix. The j-th diagonal entry of D (denoted as d_j) is transferred to the Euclidean norm of the j-th component in T and thus matrix H was also normalized. The standard deviation of b_j can be estimated in a formula as,

$$\mathrm{std}(b_j)=\frac{\lambda}{d_j}\sqrt{(H^{\mathrm{T}}H)^{-1}},$$

where λ represents the measurement noise and d_j quantifies the signal strength of the j-th model component. The factor $\sqrt{(H^{\mathrm{T}}H)^{-1}}$ indicates the spectral overlap between the j-th component and the other (n-1) components.

Experiments showed that λ varies from sample to sample, whereas d_j and H are sample-independent. To estimate the model parameter uncertainty, we repeated the measurement for more than 20 times, extracting the robustness test from each individual measurement and calculate the standard deviation.

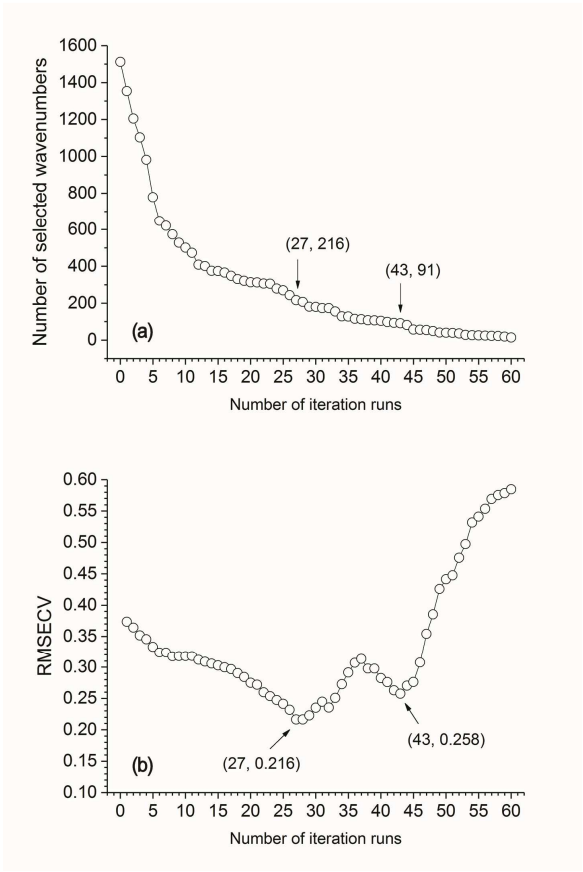


Fig. 3 The diagram of CARS variable selection

3. Results and discussion

3.1 Variable selection by CARS

Variable selection is a vital procedure for model simplification and result improvement. Effective wavelengths were selected by CARS algorithm before establishing calibration models. In the process of CARS parameter optimization, the number of Monte Carlo iteration runs was set up to 60, the optimal number of latent variables to be extracted was tested from 1 to 10, the cross validation was operated in 5 groups, and the pretreatment was accomplished by centralized algorithm.

The CARS optimizational results were interpreted in **Fig. 3**. Sub-figure (a) shows the relationship between the number of reserved characteristic wavenumbers and the number of Monte Carlo iteration runs, and sub-figure (b) shows the variation trend of $\mathrm{RMSE_{CV}}$. It can be seen from **Fig. 3(a)** that the selected wavenumber variables present a decreasing trend. This trend is firstly fast and then become slow, which reflects the CARS variable selection process firstly executed a rough selection and then gradually went to a detail selection. **Fig. 3(b)** shows that the $\mathrm{RMSE_{CV}}$ goes in a trend of firstly descending and then ascending. There appeared two minimum values when the numbers of iteration runs were 27 and 43, and the corresponding minimum $\mathrm{RMSE_{CV}}$ values reached as 0.216 and 0.258. The $\mathrm{RMSE_{CV}}$ goes larger after 43 runs, which means that the spectral data begins to remove some of the characteristic wavenumbers. As the minimum $\mathrm{RMSE_{CV}}$ is the aim of CARS variable selection, we took the wavenumbers selected at the 27th run and the 43rd run as the optimal result, thus we could determine from **Fig. 3** that there were 216 and 91 effective wavenumbers, respectively. Successively, the distribution of the 216 and 91 effective wavenumbers can be identified in the full-length spectral region (showed in **Fig. 4**). Next, we would apply these optimal 216 and 91 characteristic wavenumbers to establish calibration models and test the modeling results using the BPN-DL framework.

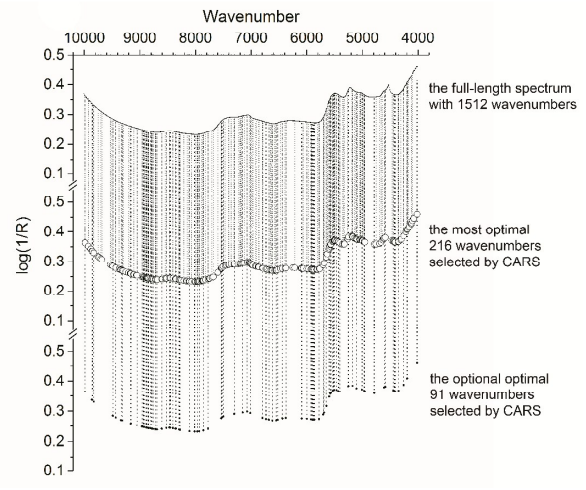


Fig. 4 Distribution of the wavenumbers selected by CARS method with 27 runs and 43 runs of iteration

View Article Online
DOI: 10.1039/C8AY01076E

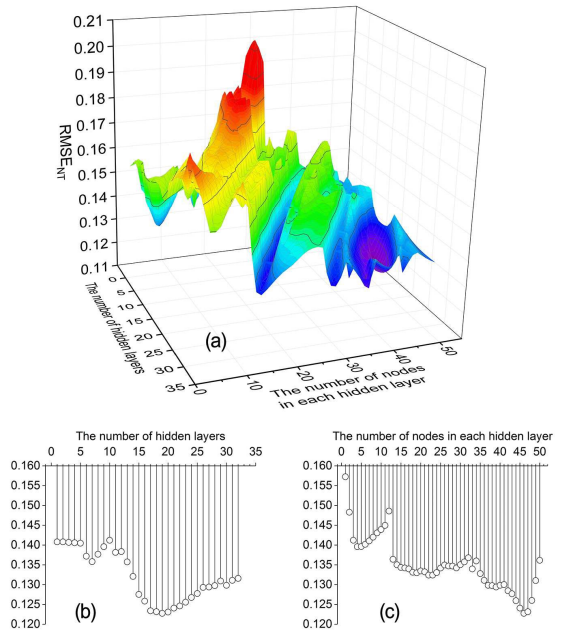


Fig. 5 Model prediction of soil organic carbon using the BPN-DL framework based on the 216 characteristic wavenumbers

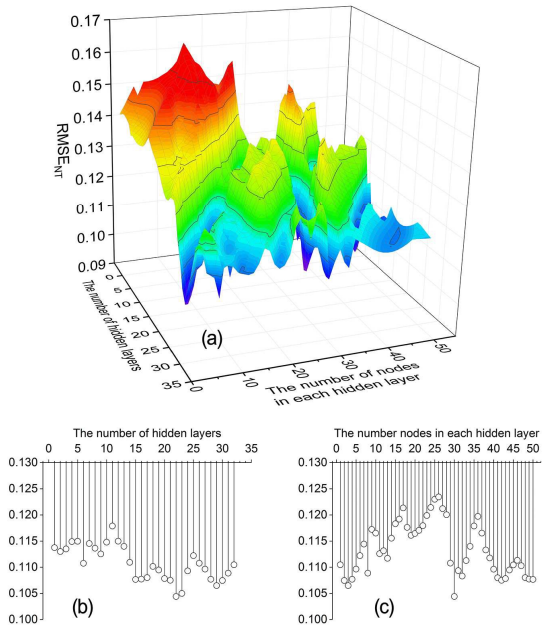


Fig. 6 Model prediction of soil organic carbon using the BPN-DL framework based on the 91 characteristic wavenumbers

3.2 The BPN-DL modeling and training results

For FT-NIR calibrations, the 135 soil samples ought to be divided in to the training part and the testing part. The pre-reserved test set comprised 35 samples, which were randomly chosen as independent of the training process. The remaining 100 samples were totally for model training. The focus of this study is to apply the BPN-DL framework to search for the improvement of the calibration model. BPN-DL models were developed using the specific input variables that were the characteristic wavenumbers preliminarily selected by CARS.

In the training part, the spectral data matrix with 216/91 wavenumbers for the 100 training samples was input to the BPN-DL framework. The framework will train the weights and the transfer functions by back-propagation circling the predictive error. The deep learning mode was activated by unlimitedly adding the number of hidden layers (K) and the number of the neural nodes in each hidden layer (s_k), until the computational complexity went out of the memory. With the undergoing computer hard configuration, we have established BPN-DL models for up to 32 hidden layers and 50 nodes (i.e. K=32 and s_k=50). The capabilities of each model were evaluated according to the root mean square error of neural training (RMSE$_{NT}$).

Figs. 5-6 show the training effect of all BPN-DL models under different values of K and s_k, respectively for 216 variables (observed by CARS with 27 iteration runs) and 91 variables (observed by CARS with 43 iteration runs). In **Fig. 5(a)** and **Fig. 6(a)**, the colored surfaces represent the predictive results from each BPN-DL model corresponding to a designated number of hidden layers and a designated number of neural nodes. To find the optimal model, we produced the projection on both of the parametric axes under the rule of minimizing the RMSE$_{NT}$ value. Thus we have the sub-figures (**Fig. 5(b)** and **Fig. 6(b)**) depicting minimum RMSE$_{NT}$ corresponding to each number of hidden layer (k, k=1,2…32) and the sub-figures (**Fig. 5(c)** and **Fig. 6(c)**) depicting minimum RMSE$_{NT}$ corresponding to each number of nodes (i_{sk} , i_{sk} =1,2…50).

It can be found from **Fig. 5** that, for the 216 input variables, the most optimal BPN model appeared when applying 19 hidden layers with 46 neural nodes. The minimum RMSE$_{NT}$ reached 0.123. Some available optimal BPN models should be built with deeper than 16 hidden layers and more than 35 nodes. Similarly, the model predictive effect for 91 input variables can be found from **Fig. 6**. The most optimal BPN model obtained the minimum RMSE$_{NT}$ of 0.104, when using 22 hidden layers and 30 neural nodes. Some available optimal models could have fewer number of nodes, but the hidden layers should be deeper than 14.

View Article Online
DOI: 10.1039/C8AY01076E

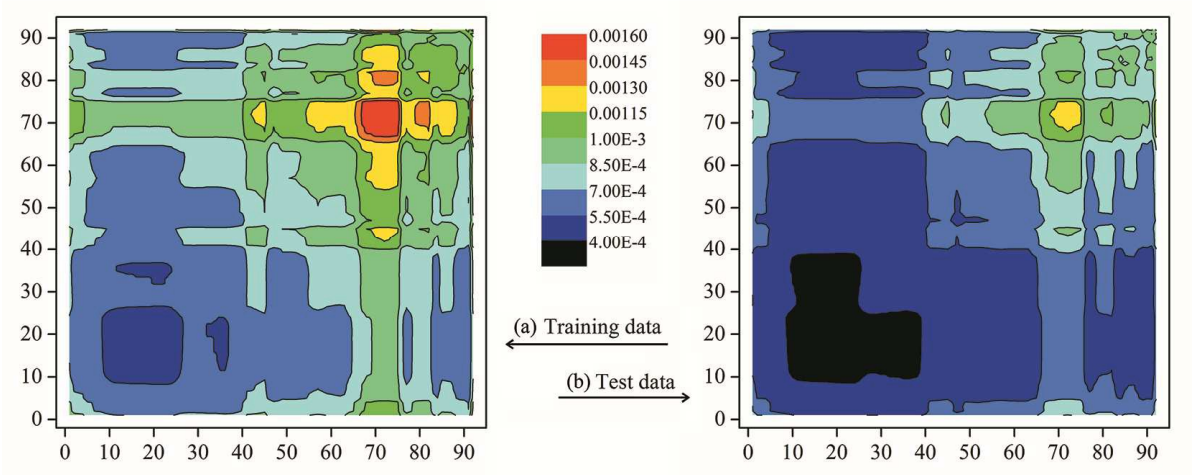


Fig. 7 The contour charts of regression vector covariance for the training data (a) and for the test data (b)

The selected input variables can be evaluated by regression vector analysis, to examine the informativeness and the independence of the variables. For instance, the regression vectors of the optimal BPN-DL model with CARS-selected 91 variables were found. The covariance matrix displaces the correlationship and the independence of the vectors. We perform the analysis respectively based on the cross validation training samples and on the independent test samples. The BPN-DL generated new comprehensive variable and the 91 selected informative variables are combined used for calibration, thus there are a total of 92 variables involved in the BPN-DL model. The covariance matrix was calculated to see whether there is any correlation between the 92 variables. The covariance maps for the training data and for the test data were presented as contour charts in **Fig. 7** (where subfigure (a) is for training samples and subfigure (b) for the test samples). We can found from **Fig. 7(a)** and **Fig. 7(b)** that the maximum covariance was lower than 1.6×10^{-3} both for the training part and the test part. Thus we conclude that the 92 variables do have little correlationship between any two of them. That is to say that these 92 variables contain little overlap information, which verifies that the variables generated from the BPN-DL model are quite informative.

3.3 The testing results of the optimal BPN-DL models

The randomly selected pre-reserved test set comprising 35 samples were used to evaluate the BPN-DL models on the optimal parameters determined by deep learning mode. The testing BPN-DL optimal models were established using the spectral data and actual OC contents (measured by potassium dichromate oxidation). We have found out the optimal parameters were 19 hidden layers with 46 neural nodes for 216 input wavenumbers, and 22 hidden layers and 30 neural nodes for 91 input wavenumbers. The model regressive coefficients were determined in the training process. Further the FT-NIR predicted values for the 35 validation samples can be estimated

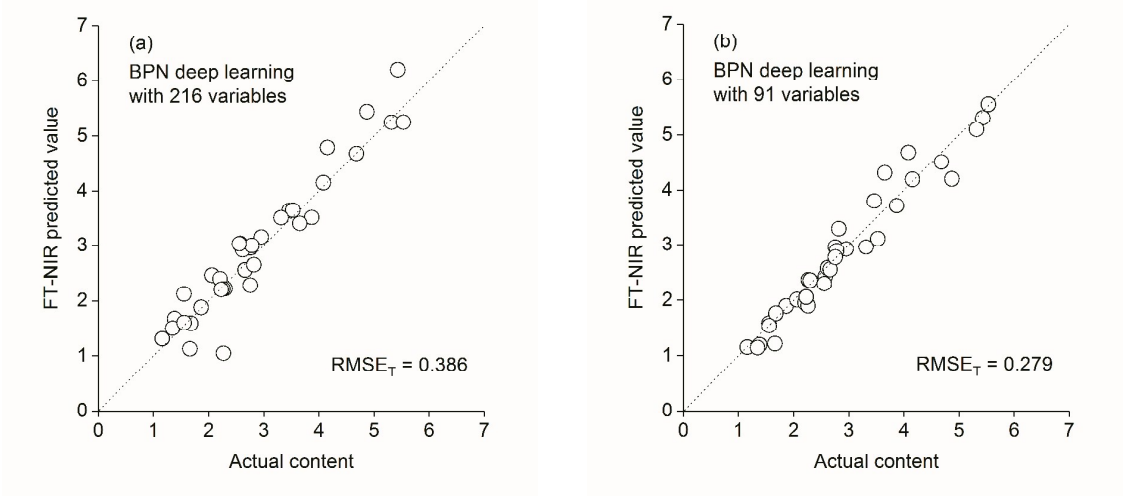


Fig. 8 The correlation relationship between the FT-NIR BPN-DL predicted values and actual OC contents (sub-figure (a) for 216 input variables and sub-figure (b) for 91 input variables)

View Article Online
DOI: 10.1039/C8AY01076E

by fitting the spectral data into the model and using the coefficients. The predicted OC values were obtained and the correlation relationship between the predicted values and actual contents were showed in **Fig. 8** (sub-figure (a) for 216 variables and sub-figure (b) for 91 variables). The root mean square error of testing ($RMSE_T$) was 0.386 and 0.279 respectively. The results showed that the predicted values and the actual contents were correlated acceptable high for the independent 35 samples. The predictive effect was satisfactory for the random testing part samples.

3.4 Comparison with benchmark models

In order to show that BNP-DL framework has a better predictive performance, the BNP-DL was compared with some benchmark methods such as PCR and PLSR. And, the BPN with only one hidden layer was regard as a counterpart method in this study. **Table 1** shows the optimal training results and the corresponding best predicted results obtained by PCR, PLSR, BPN and BPN-DL methods. As shown in **Table 1**, the BPN-DL models had a slightly smaller $RMSE_{NT}$ than other models, and the correlation coefficients (R_{NT}) are also higher in the training process. These advantages continuously appeared for the 35 test samples. These results indicated that the BPN-DL framework had a good generalization performance in both model training and testing processes. The relative predictive errors (relative $RMSE_{NT}$ and relative $RMSE_T$) in **Table 1** are also important indicators listed for detail comparison. They were calculated by dividing the root mean square error by the average of the actual OC contents of the undergoing samples. The residual predictive deviation for neural training samples and for testing samples (i.e. the RPD_{NT} and RPD_T) were also output so that the model reliability can be evaluated.

The predictive results of the common linear PCR and PLSR models were not as good as that of the BPN and BPN-DL models. That is because soil contains multiple nutritions and the raw spectral data are complex non-linear data set. The characteristic variables containing the information of OC content can hardly be identified by linear models. BPN model input the data in a neuron-feature node, stimulated and transferred by the activation function, which is the key to overcome the restraint of linear modeling. But BPN cannot be extentedly optimized because of the numbers of nodes and hidden layers are limited both in training and testing processes. Thus the application of the deep learning theory is quite necessary, and the results in **Table 1** confirmed that the best predictive results were observed at the optimal BPN-DL model.

In addition, CARS algorithm was considered to be applied for variable selection in the comparison. As the CARS parametric optimization was achieved by internal cross validation, the number of characteristic wavenumbers is unchanged and the selected informative 216 and 91 wavenumbers were available for test. All of the PCR, PLSR, BPN and BPN-DL models were re-established for the CARS-selected 216/91 characteristic wavenumbers. The training and testing results were also listed in **Table 1**. Results indicated that the models were further improved by CARS, which implied that CARS outperformed in the part of variable selection and the designated characteristic wavenumbers worked well for model improvement.

Therefore, the BPN-DL framework combined with the CARS algorithm reflects the excellent generalization in its theory, which brings a better prediction effect than the other regression methods.

3.5 Analysis of model uncertainty

The model uncertainty was estimated for the optimal BPN-DL models with the selected informative 216 and 91 wavenumbers based on the testing samples. The process was to repeat the FT-NIR measurement of the assigned testing sample for 20 times, re-establish the BPN-DL model and extract the estimation of the standard deviation of the coefficient b_j.

Fig.9 showed the model uncertainty analysis from the set of repeated measurements versus the lowest standard deviation for the analyte of soil OC content. It was seen in **Fig.9** that the model uncertainty for the 216-wavenumber BPN-DL model was within 2.1 times of the lowest, and the model uncertainty for the 91-wavenumber BPN-DL model was within 1.7 times of the lowest. The results indicated that the BPN-DL models with

Table 1 Results and comparisons of PCR, PLSR, BPN and BPN-DL models for the training and test sample sets

Models	No. of input variables	$RMSE_{NT}$	Relative $RMSE_{NT}$	R_{NT}	RPD_{NT}	$RMSE_T$	Relative $RMSE_T$	R_T	RPD_T
PCR	1050	0.383	14.7%	0.859	2.854	0.563	19.3%	0.790	2.665
PCR with CARS	216	0.291	11.2%	0.897	3.785	0.462	15.9%	0.814	4.630
	91	0.257	9.9%	0.923	4.683	0.427	14.7%	0.842	3.065
PLSR	1050	0.324	12.4%	0.876	3.352	0.515	17.7%	0.802	2.446
PLSR with CARS	216	0.273	10.5%	0.914	3.303	0.447	15.4%	0.838	4.789
	91	0.236	9.0%	0.928	5.371	0.414	14.2%	0.843	3.499
BPN	1050	0.296	11.3%	0.880	3.964	0.456	15.7%	0.821	3.138
BPN-DL with CARS	216	0.123	4.7%	0.935	4.373	0.386	13.3%	0.866	3.549
	91	0.104	4.0%	0.956	5.759	0.279	9.6%	0.912	5.375

View Article Online
DOI: 10.1039/C8AY01076E

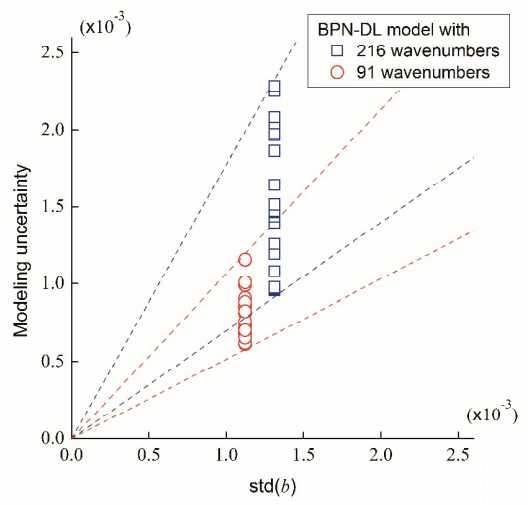


Fig. 9 The model uncertainty for the test of model robustness

the CARS-selected 216/91 characteristic wavenumbers were stable and robust in the testing round, with a relative small standard deviation.

4. Conclusions

In this study, we not only developed the CARS and BPN algorithm's application scope, but also constructed a deep learning framework for the rapid and non-destructive FT-NIR determination of the organic carbon content in soil samples.

The characteristic variables were extracted by CARS algorithm. The 216 and 91 characteristic wavenumbers at 27 and 43 iteration runs, respectively. These presumed optimal 216/91 characteristic wavenumbers were applied to establish calibration models and test the modeling results using the BPN-DL framework.

For model optimization, the BPN-DL framework was constructed from the BPN network with consideration of the deep learning mode. The number of hidden layers was extended to 32 and the number of the neural nodes to 50. The optimal BPN-DL system was identified including 19 hidden layers and 46 nodes in each layer for the case of 216 input variables. While for the case of 91 input variables, the optimal system included 22 hidden layers and 30 neural nodes. These selected training parameters were applied to test the independent 35 samples for evaluating the models. The models were tested robust and the results were better than other benchmark PLSR and PCR models, which indicated that the BPN-DL framework owns a higher generalization performance.

The idea of BPN-DL framework integrated with the CARS algorithm for variable selection makes a reference in the research of the improvement of FT-NIR analytical models and has broad application in the agricultural and tillage field.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Natural Scientific Foundation of China (No. 61505037, No. 61703117), the Guangdong Provincial Science and Technology Programs (No. 2017A040405054, No. 2017A040405051), China Postdoctoral Science Foundation (No. 2018T110880, No. 2017M620375), Natural Scientific Foundation of Guangxi (No. 2015GXNSFBA139259, No. 2017GXNSFBA198113) and the Project of Outstanding Young Teachers' Training in Colleges and Universities of Guangdong (No. YQ2015091).

Notes and references

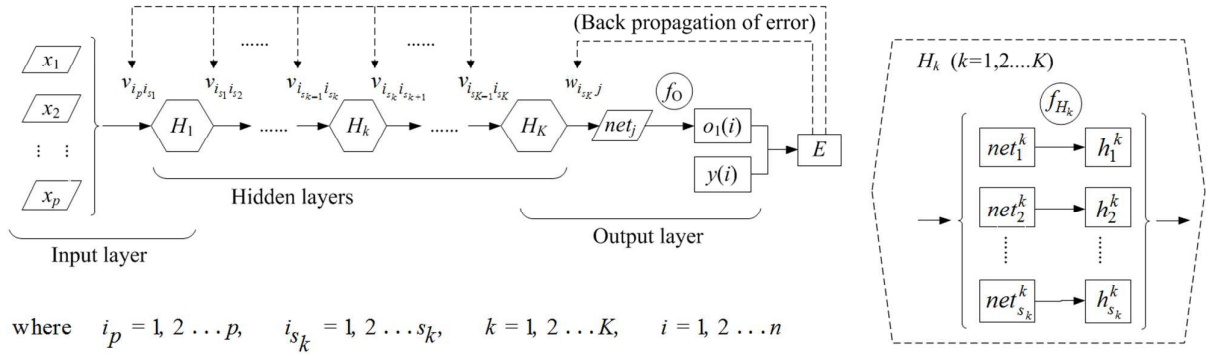
1 D. Cozzolino and A. Moron, *Soil Till. Res.*, 2006, **85**, 78-85.
2 T. Urselmans, K. Michel and M. Helfrich, *J. Plant Nutr. Soil Sc.* 2006, **169**, 168-174.
3 Y.J. Wu, Y. Jin, Y.R. Li, D. Sun, X.S. Liu and Y. Chen, *Vib. Spectrosc.*, 2012, **58**, 109-118.
4 Q.S. Chen, J.R. Cai, X.M. Wan and J.W. Zhao *LWT-Food Sci.Technol.*, 2011, **44**, 2053-2058.
5 H. Jiang, G. Liu, C.L. Mei and Q.S. Chen, *Anal. Methods*, 2013, **5**, 1872-1880.
6 E.D. Louw and K.I. Theron, *Postharvest Biol. Technol.*, 2010, **58**, 176-184.
7 H.Z. Chen, Q.Q. Song, G.Q. Tang and L.L. Xu, *J. Cereal Sci.*, 2014, **60**, 595-601.
8 R.A. Viscarra Rossel, D.J.J. Walvoort, A.B. McBratney, L.J. Janik and J.O. Skjemstad, *Geoderma*, 2006, **131**, 59-75.
9 R. Zornoza, C. Guerrero, J. Mataix-Solera, K.M. Scow, V. Arcenegui and J. Mataix-Beneyto, *Soil Biol. Biochem.*, 2008, **40**, 1923-1930.
10 T. Urselmans, K. Michel and M. Helfrich, *J. Plant Nutr. Soil Sc.*, 2006, **169**, 168-174.
11 J. Lee, P.K. Duy, J. Yoon, and H. Chung, *Analyst*, 2014, **139**, 3179-3187.
12 R.A. Viscarra Rossel and T. Behrens, *Geoderma*, 2010, **158**, 46-54.
13 H. Chen, Q. Feng, Z. Jia and Q. Song, *Asian J. Chem.*, 2014, **26**, 4839-4844.
14 A.X. Yang, J.L. Ding, H.L. Yan and K. Deng, *Spectrosc. Spect. Anal.*, 2016, **36**, 691-696.
15 H.Z. Chen, W. Ai, Q.X. Feng and G.Q. Tang, *Anal. Methods*, 2015, **7**, 2869-2876.
16 A. Watanabe, S. Morita and Yukihiro Ozaki, *Appl. Spectrosc.*, 2006, **60**, 1054-1061.
17 M.H.M Killner, J.J.R. Rohwedder and C. Pasquini, *Fuel*, 2011, **90**, 3268-3273.
18 G.B. Huang, H.M. Zhou, X.J. Ding and R. Zhang, *IEEE T. Syst. Man Cy. B*, 2012, **42**, 513-529.
19 X.D. Sun, X.L. Dong, L.J. Cai, Y. Hao, A.G. Ouyang and Y.D. Liu, *Sensor Lett.* 2012, **10**, 506-510.
20 R.M. Balabin and E.I. Lomakina, *Analyst*, 2011, **136**, 1703-1712.
21 A. Jiménez, G. Beltrán, M.P. Aguilera and M. Uceda, *Sens. Actuator B-Chem.*, 2008, **129**, 985-990.
22 Y. Allouche, E.F. López, G.B. Maz and A.J. Márquez, *J Near Infrared Spec.*, 2015, **23**, 111-121.
23 L.J. Janik, D. Cozzolino, R. Dambergs, W. Cynkar and M. Gishen, *Anal. Chim. Acta*, 2007, **594**, 107-118.
24 F. Douak, F. Melgani, N. Alajlan, E. Pasolli, Y. Bazi and N. Benoudjit, *J. Chemometr.*, 2012, **26**, 374-383.
25 A. Garrido-Varo, D. Pérezmarín, J.C. Gutiérrez-Estrada and J.E. Guerrero, *Appl. Spectrosc.*, 2006, **60**, 1062-1069.
26 B. Wang, G. Liu, Y. Dou, L. Liang, H. Zhang and Y. Ren, *J. Pharmaceut. Biomed.*, 2009, **50**, 158-163.

View Article Online
DOI: 10.1039/C8AY01076E

27 L. Deng and D. Yu, *Found. Trends Signal Proces.*, 2014, **7**, 1-19.
28 D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis, *Nature*, 2016, **529**, 484-489.
29 H.P. Lu, K.N. Plataniotis and A.N. Venetsanopoulos, *Pattern Recogn.*, 2011, **44**, 1540-1551.
30 B. Nadler and R.R. Coifman, *J. Chemometr.*, 2005, **19**, 45-54.
31 A.C. Olivieri, N.K.M. Faber, J. Ferre, R. Boque, J.H. Kalivas and H. Mark, *Pure Appl. Chem.*, 2006, **78**, 633-661.
32 O.R. Scepanovic, K.L. Bechtel, A.S. Haka, W.C. Shih, T.W. Koo, A.J. Berger and M.S. Feld, *J. Biomed. Opti.*, 2007, **12**, 1-10.
33 H. Chen, W. Ai, Q. Feng, Z. Jia and Q. Song, *Spectrochim. Acta A*, 2014, **118**, 752-759.
34 H. Chen, L. Xu, Z. Jia, K. Cai, K. Shi and J. Gu, *Anal. Lett.*, 2018, **51**, 1564-1577.
35 H. Li, Y. Liang, Q. Xu and D. Cao, *Anal. Chim. Acta*, 2009, **648**, 77-84.
36 R.K. Lu, *Methods for chemical analysis of soil agriculture*, China agricultural science and technology press, Beijing, China, 2000.
37 S. Türker-Kaya and C.W. Huck, *Molecules*, 2017, **22**, 168(1-20).
38 M.C.A. Marcelo, C.A. Martins, D. Pozebon and M.F. Ferrão, *Anal. Methods*, 2014, **6**, 7621-7627.
39 P. Geladi, D. Macdougall and H. Martens, *Appl. Spectrosc.*, 1985, **39**, 491-500.
40 H.L. Zhang and Y. He, *Spectrosc. Spect. Anal.*, 2016, **36**, 91-95.
41 W. Fan, Y. Shan, G.Y. Li, H.Y. Lv, H.D. Li and Y.Z. Liang, *Food Anal. Method*, 2012, **5**, 585-590.
42 C. Xie, X. Ning, Y. Shao and Y. He, *Spectrochim. ACTA. A*, 2015, **149**, 971-977.
43 C.B. Cai, H.W. Yang, B. Wang, Y.Y. Tao, M.Q. Wen and L. Xu, *Vib. Spectrosc.*, 2011, **56**, 202-209.
44 S.M. Kay, *Technometrics*, 2013, **37**: 465-466.

View Article Online
DOI: 10.1039/C8AY01076E

Table of Contents



The algorithmic scheme of BPN-DL framework (details of each H_k presented in the hexagon box)

A framework of back propagation neural deep learning (BPN-DL) was constructed in this work for Fourier transform near-infrared spectroscopy (FT-NIR) to predict the nutrition components in soil samples. Some characteristic wavenumbers were selected as the input variables to the BPN-DL framework based on competitive adaptive reweighted sampling (CARS) algorithm. The back propagation (BP) neural deep learning framework was employed to develop the calibration models for the determination of OC content. With the undergoing computer hard configuration, BPN-DL models were established and pre-set screening for up to 32 hidden layers and 50 nodes (i.e. K = 32 and s_K = 50). The results were achieved in iteration and parameter identification. The best optimal BP neural deep learning model and other available optimal models were found. Finally, the optimal results were compared with the benchmark PCA, PLS and the conventional BP network models. The BPN-DL framework showed its excellent performance of prediction and generalization. This study indicated that the FT-NIR spectroscopy integrated with appropriate chemometric methods could be utilized to quantitatively determine the target analyte, and BPN-DL reveals its superiority in model training and testing processes.