Water Science & Technology

Water Science & Technology Vol 89 No 7, 1613 doi: 10.2166/wst.2024.101

Determination of seawater COD spectra using double-loop contraction and sorted frog optimization

Shiwei Hou, Yingying Zhang*, Da Yuan, Xiandong Feng and Ying Zhang

Qilu University of Technology (Shandong Academy of Sciences), Institute of Oceanographic Instrumentation, Shandong Provincial Key Laboratory of Ocean Environmental Monitoring Technology, National Engineering and Technological Research Center of Marine Monitoring Equipment, No 37 Miaoling Road, 266061 Qingdao, China

*Corresponding author. E-mail: zhangyy@qlu.edu.cn

YZ, 0000-0001-7805-1703

ABSTRACT

This study develops a novel double-loop contraction and *C* value sorting selection-based shrinkage frog-leaping algorithm (double-contractive cognitive random field [DC-CRF]) to mitigate the interference of complex salts and ions in seawater on the ultraviolet–visible (UV–Vis) absorbance spectra for chemical oxygen demand (COD) quantification. The key innovations of DC-CRF are introducing variable importance evaluation via *C* value to guide wavelength selection and accelerate convergence; a double-loop structure integrating random frog (RF) leaping and contraction attenuation to dynamically balance convergence speed and efficiency. Utilizing seawater samples from Jiaozhou Bay, DC-CRF-partial least squares regression (PLSR) reduced the input variables by 97.5% after 1,600 iterations relative to full-spectrum PLSR, RF-PLSR, and CRF-PLSR. It achieved a test R^2 of 0.943 and root mean square error of 1.603, markedly improving prediction accuracy and efficiency. This work demonstrates the efficacy of DC-CRF-PLSR in enhancing UV–Vis spectroscopy for rapid COD analysis in intricate seawater matrices, providing an efficient solution for optimizing seawater spectra.

Key words: chemical oxygen demand, random frog, seawater monitoring, spectrum, UV–Vis spectrophotometry

HIGHLIGHT

- The double-contractive cognitive random field-partial least squares regression algorithm reduces input 97.5% with high seawater chemical oxygen demand (COD) prediction accuracy. Integrating variable selection and shrinkage frog leaping optimizes efficiency and robustness. An outer loop strategy and attenuation function balance convergence and efficiency. The *C* value criterion evaluates variables for seawater COD modeling. The algorithm improves ultraviolet–visible detection for rapidly determining seawater COD.

1. INTRODUCTION

In recent decades, the rapid development of industrialization and urbanization in China has led to a large discharge of wastewater into the ocean, exacerbating water pollution levels (Xu *et al.* 2021). Coastal water quality is not only affected by land-based pollutant inputs, but also closely related to the interaction with nearshore hydrodynamic processes such as waves and currents (Abolfathi & Pearson 2017). These hydrodynamic processes directly influence coastal water quality by driving water mixing and dilution (Abolfathi & Pearson 2014; Abolfathi *et al.* 2020). The chemical oxygen demand (COD), inorganic nitrogen, and active phosphates in wastewater from land-based sewage outlets have exceeded standards, resulting in eutrophication, frequent red tides and damage to marine ecosystems, and human health in coastal waters (Lin *et al.* 2020). COD is an important parameter for detecting water quality pollution (Fogelman *et al.* 2006), reflecting the relative content of organics and the pollution levels of reducing substances such as ferrous ions and sulfides in water bodies. Although traditional chemical detection methods such as potassium permanganate methods can achieve accurate detection results (Miller *et al.* 2001), they require 2–3 h for sample pretreatment, titration, and result calculation, which is time-consuming and may cause secondary pollution. In contrast, the ultraviolet–visible (UV–Vis) spectroscopy used in this study can rapidly scan samples for detection in 1–2 min (Guo *et al.* 2020). In addition, improper treatment of chemical reagents may also cause secondary pollution. Therefore, timely detection and protection of marine resources can be achieved by developing effective, rapid, and eco-friendly detection techniques. UV–Vis spectroscopy is a novel rapid non-destructive detection method (Picollo *et al.* 2018; Xu *et al.* 2021). It establishes regression models between the specific spectral curves of substance absorption and chemical measurement parameters, and is widely used in ecological, industrial, and environmental monitoring fields (Chen *et al.* 2014; Li *et al.* 2020).

Considering the deterioration of water environments, the composition of water quality has become increasingly complex, and the demand for model prediction accuracy has also increased (Cao *et al.* 2014). Therefore, research on the use of UV–Vis spectroscopy to determine the concentration of organic pollutants in water has developed from single- and double-wavelength modeling to the use of multi-wavelength and wide-band calibration modeling (Lee *et al.* 1999; Kim *et al.* 2001). Multivariate statistical methods such as partial least squares (PLS) (Kusnierek & Korsaeth 2015), principal component analysis (Sharifzadeh *et al.* 2017), and support vector machines (Devos *et al.* 2009) are commonly used for spectral analysis and modeling. The PLS algorithm can comprehensively filter the spectrum and play an important role in modeling and analyzing spectral information variables. However, a large amount of research has shown that water quality COD measurement is severely affected by various ion spectra. Considering the strong UV absorption spectrum band formed by a large number of ions in seawater, the use of PLS full-spectrum variables for modeling will greatly increase the complexity of the model and produce large measurement errors (Han *et al.* 2022). Therefore, the effective selection of spectral variables is a key issue in spectral detection. Population-based evolutionary algorithms can effectively solve this problem. These methods can fully explore the solution space and handle discrete data that traditional methods are difficult to handle, including genetic algorithms (GAs) (Feng *et al.* 2020), particle swarm optimization (Qi *et al.* 2018), ant colony optimization (Shamsipur *et al.* 2006), and gray wolf optimization (Gao *et al.* 2022). Among these methods, random frog (RF) (Li *et al.* 2012) is a feature wavelength selection algorithm proposed in recent years. It uses the reversible jump Markov chain Monte Carlo (RJMCMC) method to iteratively calculate the probability of each variable being selected in each iteration, evaluate the importance of variables, and select variables with high probabilities as feature variables. Given it is excellent performance in selecting explanatory feature variables from a wide range of complex datasets, this algorithm has been effectively applied in the field of spectral feature selection.

Although the RF algorithm has certain advantages in feature wavelength selection, when this method is applied to complex datasets, the randomness of the initial variable set leads to slow iteration speed and easy convergence to local optima. In response to this problem, some studies have introduced preprocessing or heuristic methods (Zhu *et al.* 2010; Fan *et al.* 2017) that can effectively remove irrelevant information, noise, and background interference from variables, optimize the initial variable set, and improve the convergence speed of the algorithm. However, when collinearity exists between variables, the association between the size of the coefficients and the importance of the variables decreases, thus decreasing the effectiveness of the initial variable set selected. The C value criterion proposed in recent years reasonably evaluates the importance of each variable (Zhang *et al.* 2019). By statistically calculating the contribution of different variables to model errors, it can effectively avoid the synergistic effects between variables (Kjeldahl & Bro 2010; Tran *et al.* 2015).

Moreover, although optimizing the initial variable set can improve the computational efficiency of the water quality spectral model, the high degree of randomness during the RF algorithm hinders the rapid and accurate extraction of feature variables. Relevant research has been conducted, but a general consensus has not been reached on the selection of probability-guided parameters and optimization methods. In response to the issue of individual fitness, Xu *et al.* (2014) proposed a least random shuffled frog-leaping algorithm, which optimizes the population's iteration efficiency based on a roulette wheel strategy. Yun *et al.* (2013) suggested that replacing a single wavelength point with a continuous window improves the optimization accuracy of the original RF algorithm. Sun *et al.* (2020) proposed interval selection based on RF, which optimizes wavelength selection and its width. However, balancing the search and optimization capabilities of the algorithm when applied to seawater COD spectra is crucial. Excessive searching leads to slow convergence, while focusing on finding the optimal solution may cause the algorithm to become trapped in a local optimum. The adaptive balance between search and optimization is essential for ensuring the robustness of the algorithm (Deng *et al.* 2017).

This paper proposes a double-contractive cognitive random field (DC-CRF) algorithm based on adaptive and C value sorting selection for the improved efficiency of the solution of seawater COD spectra to strengthen the balance in exploring global optimal solutions. The contributions and main findings of this paper are as follows: (1) the introduction of heuristic probability optimization and new contraction strategies, by nesting the RF algorithm in a dual-layer loop, to adapt to COD spectral optimization in seawater environments; (2) the validation of the rationality of the *C* value forward sorting parameter applied to seawater COD spectral optimization; (3) the evaluation of the RF, CRF, and DC-CRF wavelength selection methods using partial least squares regression (PLSR) modeling from the perspectives of speed, robustness, and accuracy; and (4) the determination of the optimal spectral region and sensitive bands for seawater COD.

2. MATERIALS AND METHODS

2.1. Study area and sample collection

Figure 1 displays the location of the study area. The research was conducted in Jiaozhou Bay, which is located in the central part of the Yellow Sea in China, and this area is affected by human activities such as industrial and agricultural waste discharge and domestic sewage. The COD content of the bay was investigated by randomly collecting 115 seawater samples. During the sampling period, the observed wind force was 3–4 grades and the wave height was less than 0.2 m, which were mild wave and breeze sea conditions. This ensured that the collected water samples were not affected by strong climate events in the short term and could represent the average status of the sea area. Although the collection conditions varied slightly, all samples were obtained at a depth of 50 cm and kept at 4°C. The samples were filtered through a 0.45 μm pore size membrane to remove suspended particles. The treated samples were then titrated using the alkaline potassium permanganate method to determine the actual COD value and the UV–Vis spectra were collected using a Cary5000 spectrophotometer. During spectral collection, the samples were placed in a 10 mm-long quartz cuvette, and the cuvette was washed with deionized water and dried before each sample spectrum was collected. Absorbance readings were taken at intervals of 1 nm in the wavelength range of 200–750 nm to obtain a spectral dataset. Each sample was scanned three times in the spectrophotometer, and the average of the three scans was obtained. The spectra of each sample were obtained at room temperature (15 ± 1°C) and relative humidity (30 ± 1%).

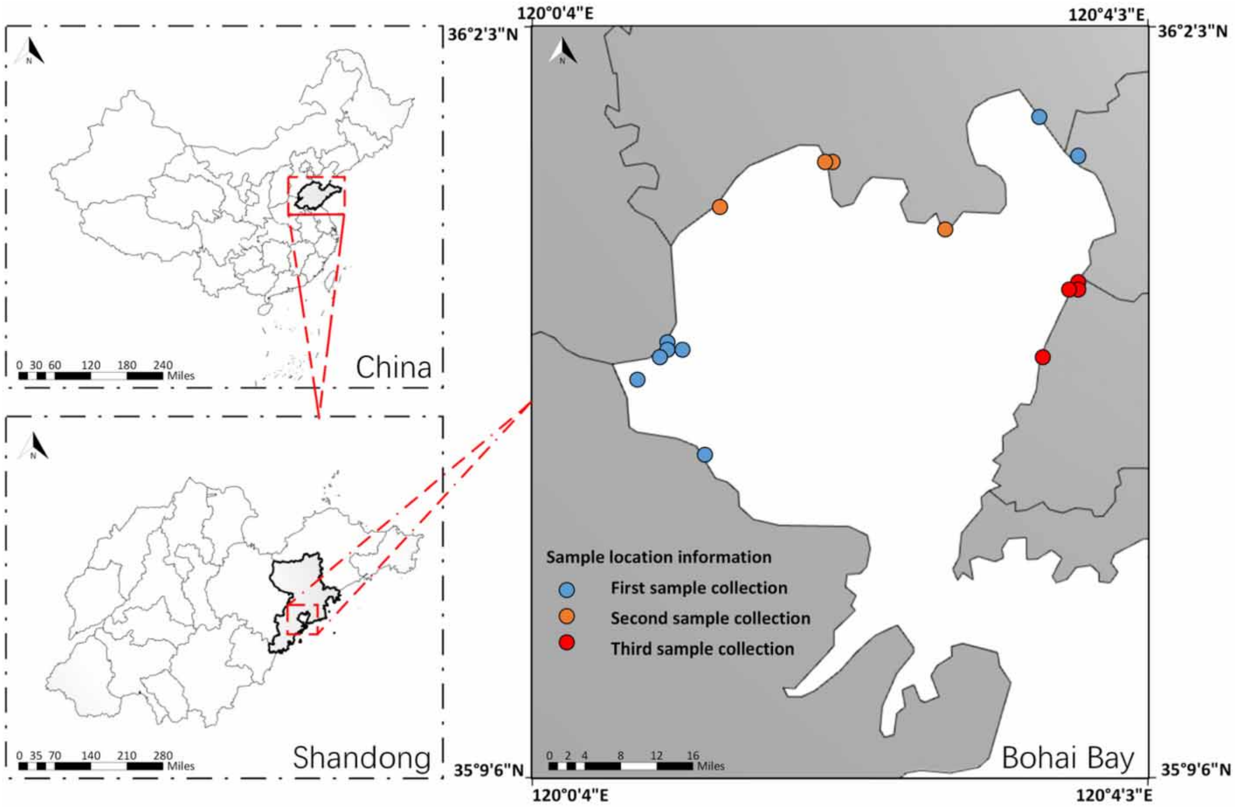

Figure 1 | Sampling sites for the seawater.

2.2. Division and preprocessing of spectral dataset

In this study, the sample set partitioning based on the joint *x*–*y* distance (SPXY) method was used to divide the original spectral dataset. This method calculates the Euclidean distance in both the feature dimension and the real concentration direction and then combines the distances in both directions through regularization to obtain a more comprehensive sample coverage (Galvao *et al.* 2005; Zhan *et al.* 2009). A total of 115 sample data were collected, including 71 samples in the training set, 22 samples in the validation set, and 22 samples in the test set. The influence of interfering factors was reduced by preprocessing the spectral data to enhance the absorption spectral information related to COD components. Five processing methods, including Savitzky–Golay (S–G) smoothing, multiplicative scatter correction (MSC), standard normal variate (SNV) transformation, first derivative (FD), and second derivative (SD), were used to process the original spectra. The results of the spectral transformations are shown in Figure 2. The correlation analysis and subsequent model building are discussed in the following sections.

2.3. Theory and algorithm

RF is an algorithm proposed by Li that bears resemblance to the RJMCMC algorithm. Its primary steps are as follows: (1) initialization: setting parameters and randomly selecting a variable subset Z_0 containing Q_0 variables; (2) probability-guided model search: based on Z_0, selecting a candidate variable subset Z^* containing Q^* (randomly generated) variables, accepting Z^* as Z_0 with a certain probability, and replacing Z_0 with Z^*, repeating this step until Lk iterations are completed (where Lk represents the number of iterations within the internal loop); (3) variable evaluation: calculating the probability of each selected variable, with higher probability indicating greater importance.

The DC-CRF algorithm proposed by the research institute adopts the new variable generation strategy in RF, but differs in three aspects. First, it uses the *C* value importance ranking criterion to determine the initial variable set instead of a random strategy. Then, it introduces a parameter adaptive contraction strategy to control the variable selection range and acceptance rate during the algorithm iteration process. Finally, it adds an outer contraction loop and metropolis acceptance criterion to preserve or eliminate new solutions of spectral wavelength variables, thus enhancing the robustness and facilitating escape from local optima. The specific method will be further elaborated.

2.4. The proposed DC-CRF algorithm

2.4.1. Determining initial subset with *C* value criterion

In RF, the initial variable subset Z_0 was randomly generated, which may result in the presence of uninformative or interfering variables, thereby increasing the number of iterations and runtime of the algorithm. The effectiveness of the initial subset Z_0 variables was improved, and the number of iterations was reduced by improving the generation of the Z_0 subset.

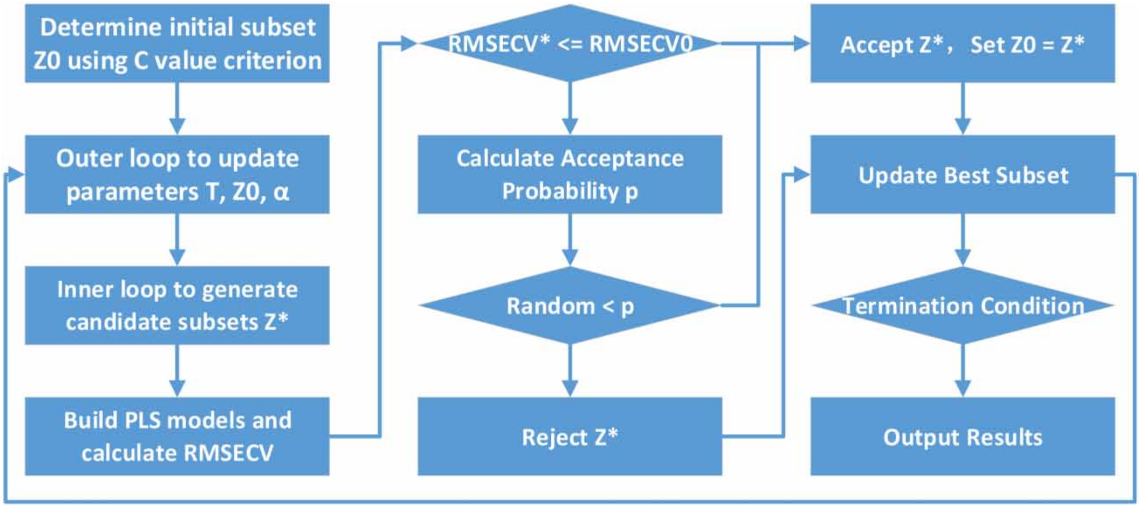
Determine initial subset Z0 using C value criterion
Outer loop to update parameters T, Z0, α
Inner loop to generate candidate subsets Z*
Build PLS models and calculate RMSECV
RMSECV* <= RMSECV0
Calculate Acceptance Probability p
Random < p
Reject Z*
Accept Z*， Set Z0 = Z*
Update Best Subset
Termination Condition
Output Results

Figure 2 | DC-CRF algorithm flowchart.

The C value criterion proposed by Xueguang Shao establishes a relationship between the contribution of variables to the model error and evaluates their importance. In a multivariate calibration model, the dependent variable y is obtained by weighting the spectral variables with a coefficient vector. After ignoring the non-fitting residuals, the model can be described as follows:

$$\sum_{j=1}^{N} x_{ij}\lambda_j = y_i \ \ (i = 1, 2, \ldots, M) \tag{1}$$

The element in the i-th row and the j-th column of the spectral matrix X, denoted as x_{ij} contains M samples and N spectral variables. The regression coefficient for the j-th variable is represented by λ_j and the measured concentration for the i-th sample is denoted as y_i. To establish a PLS model, we performed S random resampling iterations to select variables from the training set. Each iteration uses a value of 1 to indicate the selected variables and 0 to indicate the unselected variables. The binary vector representing the variable selection and the root mean square error cross-validation error (RMSECV) of the established PLS model was recorded for each iteration. Through this step, a binary matrix **b** ($S \times N$) and its corresponding PLS model's RMSECV vector **e** ($S \times 1$) were obtained. The model for the prediction error and binary matrix can be described as follows:

$$\sum_{j=1}^{N} b_{ij}c_j = e_i \ \ (i = 1, 2, \ldots, S) \tag{2}$$

The binary matrix **b** and error vector **e** were modeled using multiple linear regression (MLR), where the regression coefficients for the variables were defined as the C values to estimate the contribution of the variables to the model error. A smaller C value indicates a higher importance of the corresponding variable. The variables were sorted into positive, neutral, and negative categories, and the top Q_0 variables were selected as the initial subset Z_0.

2.4.2. Shrinkage strategy

2.4.2.1. Subset generation of variables. A strategy was set for generating new variables in RF as an inner loop with a loop count of Lk based on a normal distribution to control the number of variables and achieve variable selection and modification. In comparison with classical RF, an elastic shrinkage parameter k was introduced for the new variable set (the definition of k is introduced in 2.3). An integer β was randomly selected from a normal distribution with mean $k{\cdot}\alpha$ and variance 0.3α. Then, a candidate variable subset Z^* containing β variables was randomly generated using one of the following three constraint methods:

- If $\beta = \alpha$, let $Z_0 = Z^*$.
- If $\beta < \alpha$, a PLS model was first established for Z^*, and the regression coefficient values were recorded and compared for each variable in the model. The α–β variables with the smallest regression coefficients are then removed from Z^* and the remaining β variables form the candidate subset Z^*.
- If $\beta > \alpha$, δ (default value of 3) times the difference between β and α variables are randomly selected from V–Z^* (where V represents the set containing all N variables) to generate a variable subset T. A PLS model is established using the combination of Z^* and T, and δ variables with the largest regression coefficients in the model are retained and set as the candidate subset Z^*.

2.4.2.2. Update of the variable subset. The variable subset update strategy involves accepting suboptimal variable sets with a certain probability while updating the optimal solution. This elimination and update mechanism are beneficial for exploring inferior variables, enhancing variable set diversity, and avoiding algorithmic local optima. The update mechanism for the classical RF algorithm involves building PLS models for both the original variable set Z_0 and the new variable set Z^*, calculating the RMSECV, and obtaining RMSECV and RMSECV*. The objective function $R(i)$ is expressed as follows:

$$R(i) = \min \text{RMSECV} \tag{3}$$

The acceptance strategy for new solutions is as follows:

$$P(i \to j) = \begin{cases} 1 & R(i) \geq R(j) \\ 0.1 \cdot \dfrac{R(i)}{R(j)} & R(i) < R(j) \end{cases} \tag{4}$$

The probability of accepting new solutions is denoted by $P(i \to j)$, while $R(i)$ represents the RMSECV of the Z_0 subset and $R(j)$ represents the RMSECV of the Z^* subset.

However, the fixed threshold elimination method in RF can sometimes become a disadvantage, because it may lead increase in the computational burden and require a longer number of iterations to achieve convergence. This situation is particularly problematic for problems involving the evaluation of complex datasets, such as the optimization of seawater COD spectra. In connection with this, Li *et al.* have revealed a search mechanism that allows for elastic contraction strategies and is better suited for optimizing complex datasets. Accordingly, the elastic contraction operator in the simulated annealing (SA) algorithm was utilized to maintain the diversity of the initial population as much as possible while ensuring accuracy in the later stages, thereby reducing the number of iterations and adapting the termination threshold.

PLS models were established for both Z_0 and Z^* subsets and RMSECV was used as the evaluation metric. A smaller value of RMSECV corresponds to better predictive ability of the calibration model. The objective function $f(i_n)$ is represented as $f(i_n) = \min \text{RMSECV}$, and is transformed into a maximization problem. Therefore, the objective function is as follows:

$$F(i) = \frac{1}{1 + f(i_n)} \tag{5}$$

The metropolis criterion was applied to retain or reject spectral wavelength variables, and the acceptance probability from the original solution $F(i)$ to the new solution $F(j)$ for the newly generated variable set was determined using the following function:

$$P(T_k) = P(i \to j) = \begin{cases} 1 & F(i) \geq F(j) \\ \exp \dfrac{F(i) - F(j)}{T(n)} & F(i) < F(j) \end{cases} \tag{6}$$

In this equation, $F(i)$, $F(j)$ represents the objective functions of solutions i and j, respectively, while $T(n)$ represents the value of T during the nth iteration of the outer loop. If $F(i) \geq F(j)$, where the RMSECV of the original variable subset is not less than that of the new variable subset, then the new variable subset is accepted. If $F(i) < F(j)$, where the RMSECV of the original variable subset is less than that of the new variable subset, then the new subset is accepted with a corresponding probability determined by the contraction formula.

The improved non-optimal solution acceptance strategy was controlled by the parameter T, which gradually shrinks with the algorithm's iterations. The corresponding selection mechanism has a higher probability of accepting poor solutions in the early stages, thus expanding the search space. As the algorithm progressed, the probability of accepting poor solutions decreased, thus improving the search accuracy. The detailed setting of parameter T is described below. After each iteration of the inner loop, the subset and model error were saved. After Lk iterations, the optimal subset Z_1, the number of variables δ, and the model error were extracted. The initial subset Z_0 is updated by setting $Z_0 = Z_1$ and $\alpha = \delta$.

2.4.3. Outer loop and contraction variable range setting

The variable subset and parameters were optimized by implementing a new outer loop. Each outer loop updates the optimized variable subset and parameters by extracting the best variable set stored in Lk inner loops. The inner loops save the subset and model error after each iteration and after Lk iterations, the optimal subset Z_{best}, the number of variables δ in Z_{best}, and the model error were retrieved. A decay function k was utilized from the SA algorithm to contract the parameter T, thereby controlling the range of variable selection and the acceptance of new variable sets. The update function for T is as follows:

$$T(n+1) = k \cdot T(n) \tag{7}$$

where n represents the number of iterations in the outer loop and k is a constant, $k \in (0, 1)$.

After updating T, the initial subset Z_0 was updated by setting a Z_0 value that is equal to Z_{best} and Q_0 value that is equal to δ, and then entering the next round of Lk inner loops. The algorithm stops and outputs the final number of variable wavelengths, corresponding variable wavelengths, and model error when either the error acceptance threshold or the lower limit of T (T_f) is reached as the termination condition.

The process of the shrinking control algorithm and the acceptance probability p of the inner loop solution are controlled by the outer loop parameter. When the initial T is large, the selection range of variables is wide and the probability of accepting new solutions is high. As the outer loop iterates, the T parameter decreases, the selection range of variables narrows, and the probability of accepting new solutions decreases. The algorithm implements a wide range of initial selection to ensure that the search range includes the entire spectral dataset while avoiding local optima, and focuses on variable selection in the later stages of the algorithm. Therefore, the key to the algorithm is to choose a set of control parameters for the algorithm process that returns an optimal solution within a reasonable time. Such a set of control parameters is usually referred to as a schedule, which mainly includes the following parameters: (1) starting parameter T_0; (2) decay function k; (3) parameter lower limit T_f.

When selecting these parameters, multiple factors need to be considered. The size of the initial temperature T_0 determines the exploration range in the early stages. The decay coefficient k controls the cooling schedule and should be selected to gradually decrease the temperature. The final temperature T_f needs to be set small enough to allow convergence. The number of inner loops Lk affects optimization progress. In addition, the acceptance probability of new solutions depends on parameter T.

For the dataset used in this study, the parameters were optimized through extensive experiments. The initial temperature T_0 was set to 100 based on the search space size to ensure sufficient exploration. The decay coefficient k of 0.85 achieved a good balance between global and local search. The final temperature T_f of 0.01 allowed the algorithm to converge to a near-optimal solution. The number of inner loops Lk was set to 50 through trial and error. The parameter C was chosen as -0.016 for the acceptance probability, which led to the best results.

It should be noted that the parameters need to be individually tuned for different datasets to achieve optimal performance. We performed detailed sensitivity analysis and obtained robust value ranges for the parameters. The algorithm shows low sensitivity to small variations of the parameters within these ranges. The algorithm process is shown by the flow chart in Figure 2.

2.5. Model evaluation methods

PLSR is an analytical method that combines principal component regression and MLR, and this method was proposed by Herman Wold in 1966 and has been widely used in spectral data processing (Wold *et al.* 2001). PLSR mainly establishes a linear model of the independent variables (x_1, x_2, ..., x_{m}) in the case of a large number of highly linearly correlated variables, to solve the problem of having fewer samples than variables. PLSR involves the extraction of independent components T_h ($h = 1,2,...$) from the independent variables, which carried as much of the original components as possible, followed by the extraction of independent components U_h ($h = 1,2,...$) from the dependent variables (y_1, y_2, ..., y_m). These processes maximize the covariance between T_h and U_h and use MLR to establish a regression model of the dependent variables. The basic model of PLSR is follows:

$$\begin{aligned} X &= T_h P^T + E \\ Y &= U_h Q^T + F \end{aligned} \tag{8}$$

where P and Q are orthogonal loading matrices of size m*h, and E and F are error terms (random variables that follow a normal distribution).

In this study, the measured COD content of seawater was taken as the dependent variable, and a set of spectral variables was selected as the independent variables to establish four PLSR models, including a PLSR model based on the full spectrum, a PLSR model based on RF wavelength selection, a PLSR model based on CRF (C value sorting to determine the initial variable set of RF) wavelength selection, and a PLSR model based on DC-CRF wavelength selection. The effectiveness of DC-CRF-PLSR was verified by comparing the predictive abilities of the four models. The predictive ability of the models was mainly evaluated based on the calibration coefficient of determination, root mean square error of calibration (RMSEC), relative percent deviation of calibration, prediction coefficient of determination (Rp2), root mean square error of prediction (RMSEP), relative percent deviation of prediction, root mean squared logarithmic error (RMSLE) (Habib *et al.* 2023), mean absolute error (MAE) and relative absolute error (RAE) (Yeganeh-Bakhtiary *et al.* 2023). A value of *R* closer to 1 and RMSEC, RMSEP, RMSLE, and MAE closer to 0 indicate better model fitting and higher prediction accuracy (Rohman *et al.* 2010). An RPD value less than 1.4 indicates an unreliable model, an RPD value between 1.4 and 2.0 indicates a relatively reliable model, and an RPD value greater than 2.0 indicates a highly reliable model that can be used for model analysis (Fearn 2002). Data analysis and modeling software were carried out in MATLAB 2020A.

3. RESULTS AND DISCUSSION

3.1. Spectral analysis

Figure 3 shows the spectral scanning features of three seawater samples. The first sample was collected from the sea area near the industrial agglomeration area, and the second and third samples were collected from the sea areas near different residential areas. The three groups of spectra have some commonalities, namely an absorption valley around 240 nm, an absorption peak between 240 and 260 nm, and subtle fluctuations between 260 and 800 nm, which may be identified by algorithms as valid features but are hard to visually represent in the figure. In order to more comprehensively understand the differences and inherent characteristics of the three groups of data, we used t-stochastic neighbor embedding (t-SCN) dimensionality reduction visualization technology (Luo *et al.* 2021). t-SCN is a nonlinear dimensionality reduction method that can project high-dimensional data into two- or three-dimensional spaces while preserving the relative distance relationships between data points. We used the t-SCN algorithm to train the spectral data of the seawater samples and visualized it. The t-SCN visualization results found that the spectral clustering of the first sample group differed significantly from the second and third groups, which may be caused by emissions from different areas. The first sample group was near the industrial agglomeration area, where the types of pollutants were relatively single but the concentrations were higher. The second and third sample groups were located in residential areas, where the types and concentrations of pollutants varied. In addition, there was some overlap between the three sample groups. The overlapping areas corresponded to samples with similar spectral features, indicating that there were still certain similarities in pollution across different areas.

A PLSR prediction model for seawater COD was established based on the bands within the range of 200–800 nm, not only because seawater COD has characteristic absorption peaks between 200 and 800 nm which are closely related to COD values, but also there exist subtle fluctuations identifiable by the algorithm itself within this range that are useful for COD prediction. The spectral graphs after applying different preprocessing methods are shown in Figure 4, and the prediction

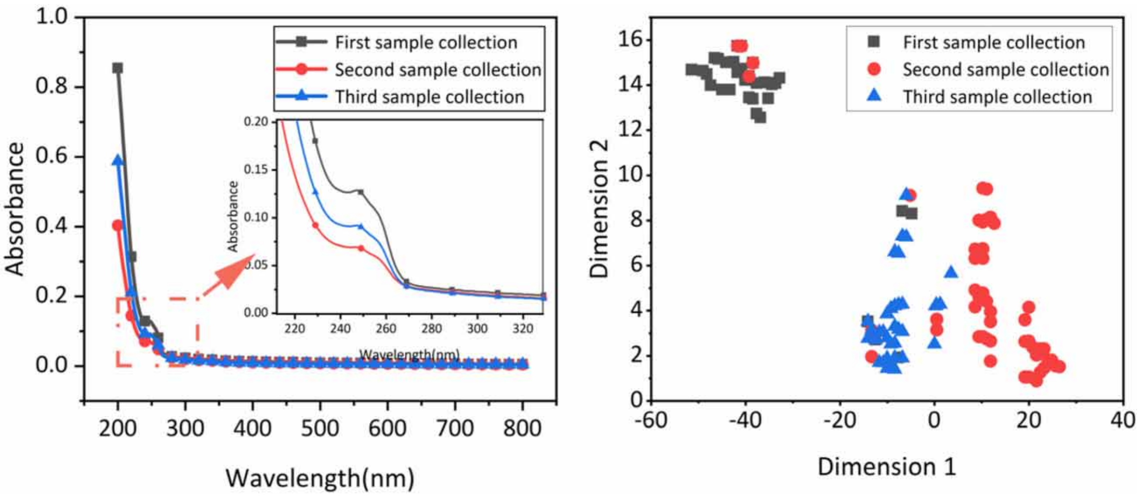

Figure 3 | Spectral features and t-SCN visualization of seawater samples.

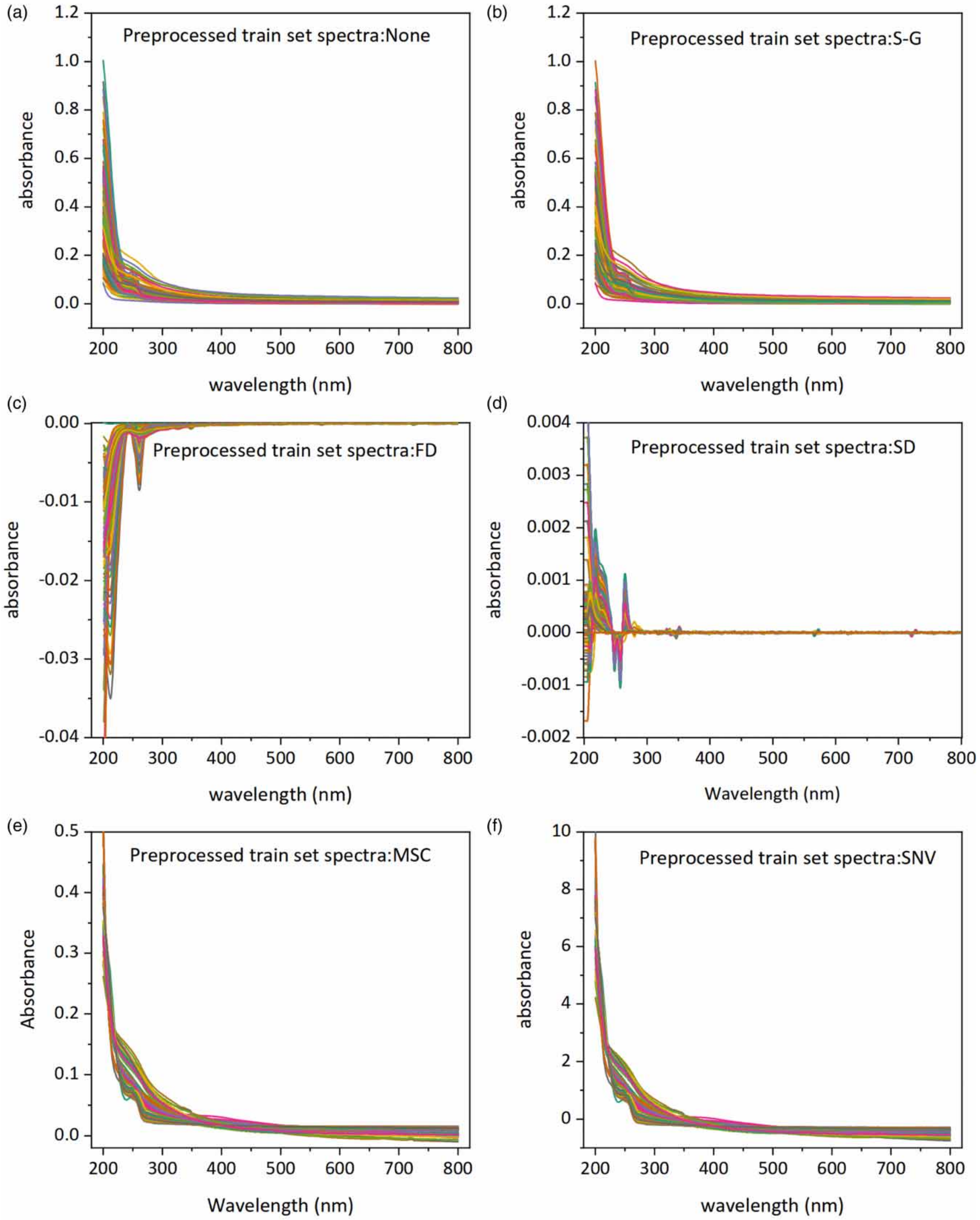

Figure 4 | Original and five preprocessed spectral graphs.

results are presented in Table 2. The model performance was poor when using the full-spectrum data, indicating that certain factors such as redundancy and interference information affect the performance and prediction results of the data. Therefore, an appropriate preprocessing method is needed to effectively optimize the data without affecting the spectral characteristics of the combined bands. For the seawater COD spectral model, after S-G preprocessing, the prediction results did not show significant improvement compared with the PLSR model established with the original spectra. The SNV preprocessing method had the best positive effect, with a prediction result of R^2_{pre} of 0.712, RMSEP of 4.787, and $\mathrm{RPD}_{\mathrm{pre}}$ of 1.863. The $\mathrm{RPD}_{\mathrm{pre}}$ value of the MSC-PLS model was 99.36% of the SNV-PLS model, the RPD value of the 1stDer-PLS model was 95.33% of the SNV-PLS model, and the RPD value of the 2ndDer-PLS model was 97.53% of the SNV-PLS model. Therefore, the SNV method was chosen as the preprocessing method for subsequent research.

3.2. Performance of *C* value

A total of 200 PLSR models were established by randomly selecting 20% of the variables from the population each time, and the *C* values of the seawater COD sample spectra variables were calculated. Figure 5 shows the distribution of *C* values for 601 variables, in which the upper quartile (Q1) is represented by the blue line and the lower quartile (Q3) is represented by the red line. The rationality of the *C* values was verified by examining three types of spectral variables, including positive variables below Q3, negative variables above Q1, and neutral variables in the middle range. According to the *C* value criterion, the bands in the positive variable interval should have a positive contribution to the model, while the bands in the negative variable interval should have a negative impact on the model. By using the full-spectrum model as a reference, Table 1 lists the optimal number of latent variables (nLV), the number of variables used in the model (nVar), RMSECV, and coefficient of determination (R^2). The nLV was optimized by the model and the RMSECV was calculated by 10-fold cross-validation.

Table 1 shows that removing positive variables leads to an increase in RMSECV and a decrease in R^2, indicating model degradation. On the other hand, removing the negative variable range results in a remarkable increase in RMSEP and a noticeable change in R^2, indicating model improvement. This result fully demonstrates the effectiveness of the *C*-criterion in evaluating the importance of spectral variables in modeling.

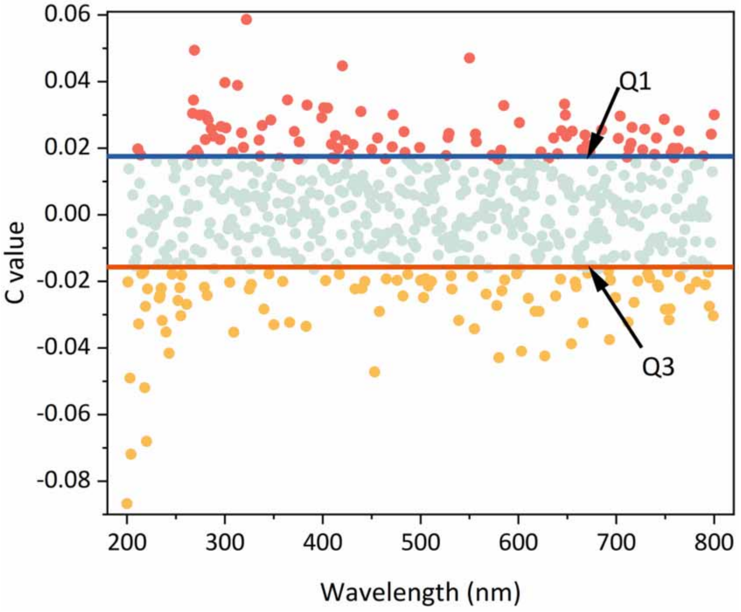

Figure 5 | The *C* value statistical chart of 601 variables. The blue line Q1 represents the upper quartile, while the red line Q3 represents the lower quartile.

Table 1 | The impact of variable sets with three different *C* value intervals on the PLSR model

Removed variables	nVar	nLV	RMSECV	R^2	RMSLE	MAE	RAE
None	601	6	4.86	0.699	0.039	1.245	0.134
Positive	450	6	5.03	0.626	0.042	1.372	0.149
Negative	450	6	4.61	0.734	0.037	1.092	0.118

nVar: Number of spectral variables used in the model and nLV: number of latent variables used in the model.

The effectiveness of using C values was further investigated to determine the initial subset instead of random variables by sorting the variables according to their C values. Figure 6 shows the changes in RMSECV as the selected variables increase by using both full-spectrum random variables and C value sorted variables to establish PLSR models. Randomness in the algorithm was avoided by building 200 models and plotting the standard deviation of the variables as vertical lines in the figure. The two curves have similar trends. Initially, both models selected too few variables, resulting in a lack of effective modeling information and large values of RMSECV and standard deviation. When all variables were included in the model, the two curves converged at the endpoint. The sharp decline followed by a gradual increase in the two curves is reasonable, because as the number of variables increases, the effective information between different variables will be repeated due to synergistic effects, and a large amount of irrelevant redundant information will lead to a decrease in model performance. However, the RMSECV is close to the minimum value, and a large number of variables are required for random variables, and the RMSECV is relatively large. This finding indicates that for this dataset, random variables cannot exclude variables that may deteriorate the model quality while carrying information variables. As for the curve obtained through C values, the minimum error with fewer variables was achieved, and the RMSECV gradually approached the minimum value when the number of variables was 33. The RMSECV values obtained based on the selected variables were subjected to the Friedman test, with a significance level of 0.15. Subsequently, the range of selected variables and RMSECV values steadily increased. In addition, as shown by the error bars in the figure, the standard deviation obtained from random variables was significantly larger than that obtained from C values. The results indicate that the stability and accuracy of initial variables determined by C values are significantly superior to those obtained through random methods.

3.3. Effect of outer loop shrinkage strategy

When a synergistic effect is presented between variables, uninformative variables may substantially affect the model performance. This situation can be avoided by increasing the outer loop shrinkage. The effect of adding an outer loop on the model was tested by conducting experiments on the COD dataset based on the DC-CRF algorithm with elastic shrinkage, and the parameters N, L, k, T_0, and Q_0 in DC-CRF were set to 40, 30, 0.85, 100, and 40, respectively. Figure 6 shows the changes in the number of selected variables and the mean RMSECV as the outer shrinkage iteration increases. The standard deviation was plotted as a vertical bar to show the random variation in the calculation.

The figure shows that the number of selected variables jumps significantly in the early and middle stages, indicating that the algorithm enters the search phase. As the number of shrinking iterations increases, the number of variables gradually shrinks

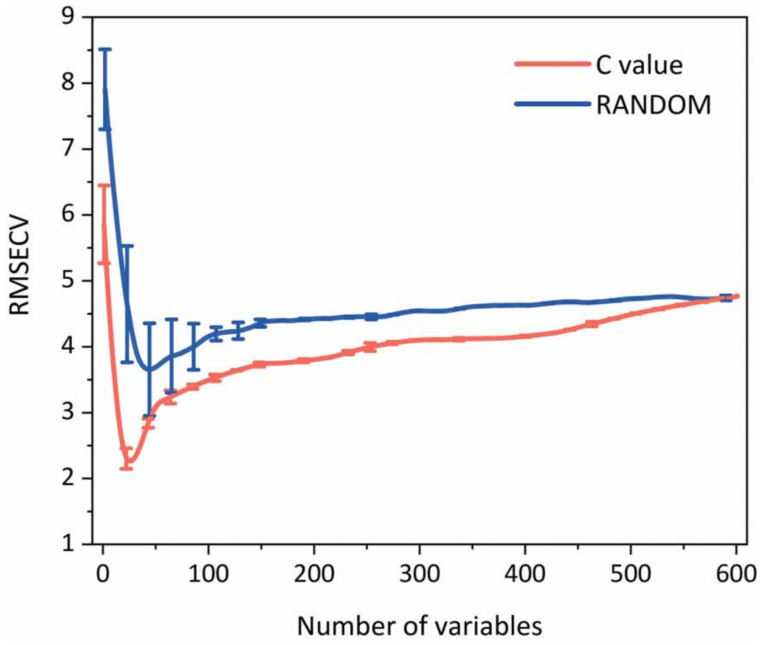

Figure 6 | Changes in RMSECV values with increasing number of variables using two methods: C-index ranking and random selection.

to the stable optimal value. In the early stages of the search, RMSECV is relatively large. As the shrinking progresses, the RMSECV of the model constructed by the selected variables exhibits severe fluctuations, with local valleys appearing in the curve, indicating that the algorithm converges to a local optimum. After this point, the curve slightly rises and then continues to decline, which is due to the non-optimal solution acceptance strategy during the algorithm iteration process, which helps the algorithm to jump out of local optima. At the end of the curve, RMSECV and the number of selected variables tend to stabilize at the global minimum. Compared with the minimum RMSECV (3.72) obtained by sorting the C-index in Figure 7, DC-CRF has an improvement in eliminating model errors, and the number of selected variables is reduced from 33 to 8, with RMSECV further reduced to 1.576. The difference is due to the selection sequence, and the shrinking strategy optimizes variable selection not only by considering the contribution of individual variables to model error but also by considering the synergistic effects and information coverage among variables on the basis of increasing the number of variables.

3.4. Comparison of models

This study explores the effect of four methods, namely RF, CRF, DC-CRF variable selection methods, and full-spectrum variables, on the accuracy and simplicity of PLSR models based on UV–Vis spectra transformed by standard normal transformation. All methods ran 200 times, and the mean values were taken. The prediction results of FULL-PLSR, RF-PLSR, CRF-PLSR, and DC-CRF-PLSR models for seawater are shown in Table 2, which shows that the three feature variable selection techniques improved the model performance compared with the full-spectrum variables. The number of wavelengths selected by the RF-PLSR model is less than that selected by the combined band PLSR model. The number of feature variables selected by the CRF-PLSR model is similar to that of the RF-PLSR model, as the variable selection methods of the two methods are the same. However, the number of feature variables selected by DC-CRF was further reduced, indicating that the variables selected by DC-CRF are smaller than the critical band range selected by RF. At the same time, the R^2, RMSE, and RPD scores of the three variable selection modeling systems are significantly higher than those of the full-spectrum PLSR. By using the RF-PLSR algorithm, the R^2 of the test set increased from 0.712 to 0.921, RMSEP decreased from 4.787 to 1.742, and RPD increased from 1.863 to 1.903. By using the CRF-PLSR algorithm, R^2 increased to 0.914, RMSEP decreased to 1.731, and RPD increased to 2.364. In comparison, the DC-CRF-PLSR model obtained a higher R^2 value of 0.943, a lower RMSEP value of 1.603, and a higher RPD value of 2.635. In addition, the RMSE, RAE, and RMSLE of the FUII-PLSR model were 4.787, 0.712, and 1.863 respectively; while for RF-PLSR they decreased to 1.742, 0.421, and

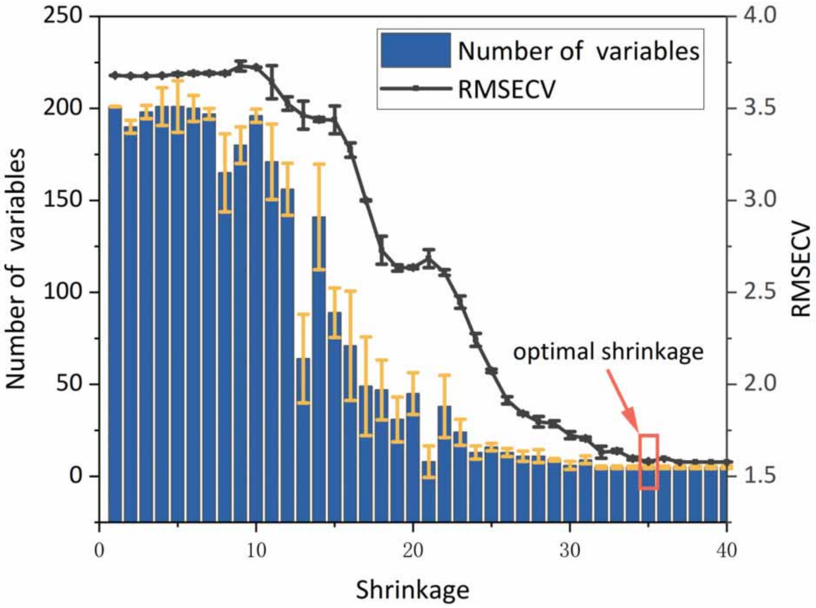

Figure 7 | Changes in the number of variables selected by DC-CRF and corresponding RMSECV of PLSR modeling with increasing iteration times.

Table 2 | Summary of PLSR, RF-PLSR, CRF-PLSR, and DC-CRF-PLSR prediction model results for seawater based on different preprocessing methods

Method	Pre.p	nVar	Testing RMSE	R^2_{coe}	RPD$_{coe}$	RAE	RMSLE	MAE
PLSR	RS	601	5.126	0.686	1.785	0.123	0.043	1.327
	FD	601	4.863	0.699	1.823	0.134	0.039	1.245
	SD	601	4.813	0.713	1.867	0.127	0.041	1.372
	S-G	601	4.854	0.706	1.844	0.115	0.038	1.201
	MSC	601	4.712	0.711	1.860	0.129	0.037	1.263
	SNV	601	4.722	0.720	1.890	0.124	0.036	1.192
RF-PLSR	SNV	38	1.728	0.935	3.922	0.116	0.028	0.921
CRF-PLSR	SNV	33	1.722	0.931	3.807	0.129	0.027	0.836
DC-CRF-PLSR	SNV	8	1.576	0.955	5.345	0.107	0.023	0.672

RS: raw spectra; Pre.p: preprocessing method; nVAR: the number of variables; FD: first derivative; SD: second derivative; S-G: Savitzky–Golay smoothing; MSC: multiple scattering correction; and SNV: standard normal variate.

1.903; for CRF-PLSR they decreased to 1.731, 0.407, and 2.364; and for DC-CRF-PLSR they further decreased to 1.603, 0.318, and 2.635. This indicates that the DC-CRF-PLSR model performs well in seawater prediction.

Compared to the study by Li *et al.* (2020), which used PLSR combined with successive projections algorithm for wavelength selection and obtained R^2 of 0.843 and RMSEP of 6.622 for COD prediction, our proposed DC-CRF-PLSR achieved better prediction performance in terms of both Rp2 and RMSEP. Another recent study by Cen *et al.* (2021) applied a multi-scale analysis method combined with GA to identify phytoplankton functional groups in coastal waters using UV–Vis spectroscopy, and attained R^2 of 0.993 and RMSEP of 1.29. In comparison, our DC-CRF-PLSR model achieved comparable results while using a simpler modeling approach. To further validate the differences in RMSEP values among these variable selection methods, a Wilcoxon signed rank test was performed, and the results indicated a significant difference between the DC-CRF method and other methods at a significance level of 0.05. Overall, for the performance of seawater COD spectra, the DC-CRF-PLSR is far more accurate than FULL-PLSR and slightly better than the other two methods.

The important wavelength distributions selected for RF, CRF, and DC-CRF are shown in Figure 8. Notably, the important wavelength distributions selected by RF and CRF have a similar location, in which 41 and 36 variables were selected, and they have a wide coverage range of the full-spectrum wavelength variables. This preserves the synergistic information of other wavelength variables, which can carry supplementary information but also cause redundant interference. Beyond the wavelength of 300 nm, the synergistic information cannot be explained by chemical bond absorption spectra, and the variables in the set, combined with key variables, have both positive and negative effects. The DC-CRF model selects eight variable points concentrated in the 240–280 nm region, which is also the concentrated region selected by RF and CRF variables, coinciding

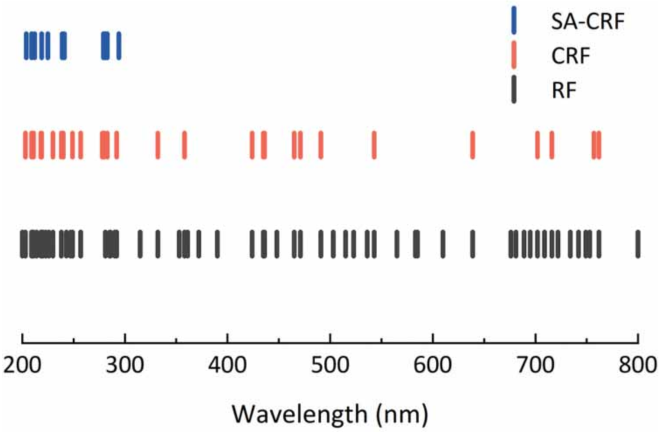

Figure 8 | Distribution of variables selected by three algorithms.

with the conjugated system of compounds or the absorption peaks of certain hydroxyl groups in the UV–Vis region. The simplicity of spectral variables is a key issue in model processing, which can greatly improve operational speed. However, one issue to note is the control of the DC-CRF variable selection range, which may selectively discard important variables due to a too small contraction range in the later stages, and fail to construct a satisfactory model. Therefore, the contraction function and initial value need to be controlled to avoid this situation.

Table 3 shows the RMSECV values and error fluctuations of the RF-PLSR, CRF-PLSR, and DC-CRF-PLSR algorithms as the number of iterations changes. Each case ran 20 times for statistical purposes, and the DC-CRF-PLSR iteration was calculated by multiplying the number of outer and inner loops. RF-PLSR and CRF-PLSR stabilized after 5,000 iterations, while DC-CRF-PLSR achieved stable performance in less than 2,000 iterations, with stability achieved at around 1,600 iterations. The reduction in the number of iterations indicates that a large amount of repeated random sampling was avoided, which improved the efficiency on the basis of algorithm stability. The RMSECV of the three algorithms all showed a downward trend with increasing iterations, with rapid initial decline and gradual flattening in the middle and later stages. The difference lies in the fact that at 50 iterations, the RMSECV of the random variables selected by RF for modeling was significantly higher than the other two because of the optimization of the *C* value ranking for the initial variable subset. After 2,000 iterations, the results of RF-PLSR and CRF-PLSR tended to be consistent, indicating that a large number of iterations had overwhelmed the advantages of the initial variable set optimization. In addition, after 2,000 iterations, the RMSECV of RF-PLSR and CRF-PLSR still showed a downward trend, although the change was small, while the RMSECV of DC-CRF-PLSR had solidified. Although the overall fluctuation error of DC-CRF-PLSR is smaller than that of the two other algorithms, this solidification of the variable set caused by the contraction function hindered the further optimization of the model, resulting in limitations in model optimization when facing different datasets.

3.5. Scatter plots of the four models

The prediction scatter plots of FULL-PLSR, RF-PLSR, CRF-PLSR, and DC-CRF-PLSR models based on UV–Vis spectra are shown in Figure 9, showing a good linear relationship between the four variable sets in the PLSR model. In comparison with the scatter plot predicted by the FULL-PLSR model, the scatter plots predicted by the other three models are clustered near the 1:1 line, indicating a high correlation between the wavelength variables selected by the algorithm and the COD of seawater spectra. However, after processing by the DC-CRF algorithm, the distribution of the scatter plot did not significantly differ from that of the scatter plots of the other two variable selection methods, which may be due to the limited range of samples covered.

The prediction scatter plots of FULL-PLSR (scatter index: 1.26), RF-PLSR (scatter index: 1.21), CRF-PLSR (scatter index: 1.23), and DC-CRF-PLSR (scatter index: 1.15) models based on UV–Vis spectra are shown in Figure 9. The plots exhibit a strong linear relationship between the predictor and response variables in each PLSR model. The scatter index was computed by taking the logarithmic transformation of the original data, fitting a linear regression model to the log-transformed data, calculating the mean squared error (MSE) of the regression model, and taking the square root of the MSE. Compared to the FULL-PLSR model predictions, the scatter points from the other three models cluster more tightly around the 1:1 line, indicating a high correlation between the selected wavelength features and the COD values of the seawater spectra. However, the distribution of points in the DC-CRF-PLSR scatter plot was not markedly different from that of the other two variable selection techniques. This suggests the limited diversity in the sample set may have constrained the performance of the DC-CRF algorithm.

Table 3 | Variation of RMSECV values for the three models with increasing number of iterations

nITE	RF-PLSR	CRF-PLSR	DC-CRF-PLSR
50	4.483 ± 2.131	3.802 ± 1.834	3.761 ± 1.345
200	3.105 ± 2.013	2.879 ± 1.637	3.686 ± 1.531
1,000	2.359 ± 0.831	2.351 ± 0.816	2.560 ± 1.653
2,000	2.133 ± 0.234	2.153 ± 0.315	1.575 ± 0.072
5,000	1.845 ± 0.213	1.834 ± 0.234	1.583 ± 0.046
10,000	1.722 ± 0.135	1.731 ± 0.101	1.577 ± 0.045

nITE: number of iterations.

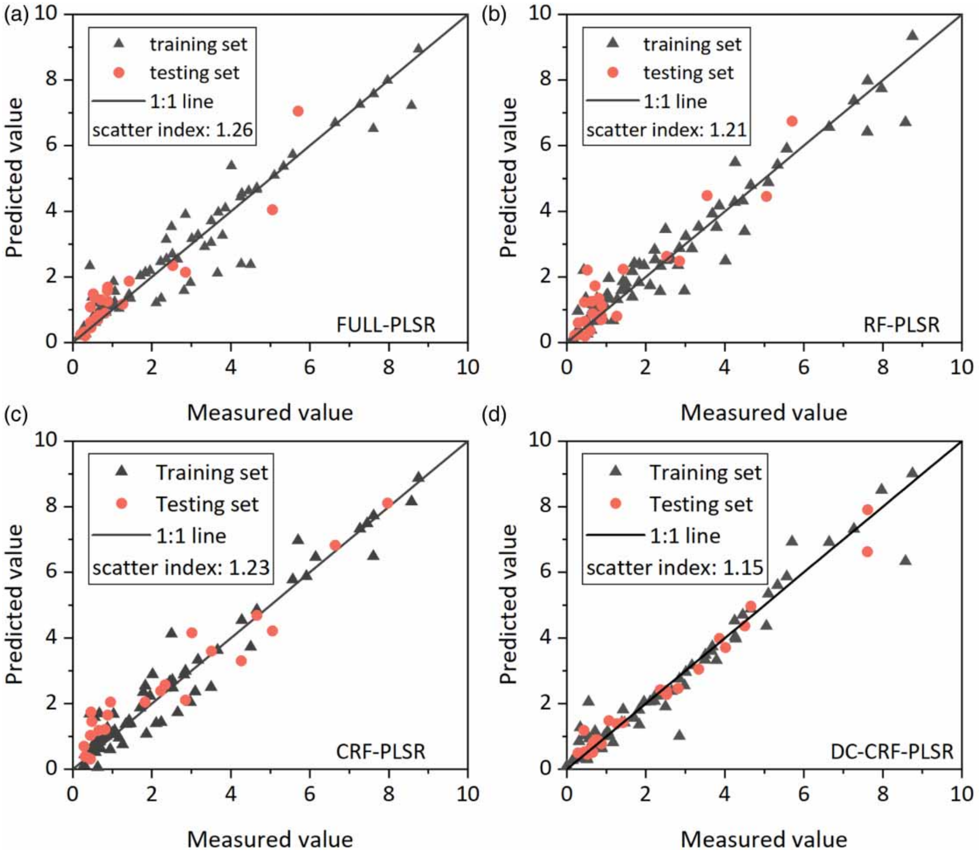

Figure 9 | Scatter plots of the four models.

4. CONCLUSION

This paper proposes a novel wavelength variable selection algorithm based on DC-CRF. The DC-CRF algorithm successfully addresses the issues of slow convergence, reduced exploratory power, and difficult parameter adjustment faced by previous methods for complex seawater spectral optimization. By analyzing five different spectral preprocessing methods and selecting SNV transformation as the optimal preprocessing approach, this study demonstrates the rationale behind selecting the *C* value and outer loop contraction settings in the DC-CRF algorithm. Through comparative analysis of the prediction performance of three models – RF-PLSR, CRF-PLSR, and DC-CRF-PLSR – for estimating seawater COD levels, it is shown that the DC-CRF-PLSR model achieves superior performance and maximizes simplification of model inputs.

The DC-CRF algorithm selects important spectral feature variables through a double-loop contraction mechanism, improving efficiency as well as identifying key wavelengths containing information relevant to COD variation. This helps gain insights into the spectral characteristics and variation patterns of COD and enhances the accuracy of COD detection. This study presents a preliminary investigation of the spectral variables related to seawater COD, providing useful improvements to guide further research. However, this research has limitations in only considering the internal factors in water and not accounting for external environmental factors in a more comprehensive way. Future studies will expand sample ranges, integrate other spectral models, and consider environmental factors such as waves and climate to improve model applicability. Expanding the range of low COD seawater samples and samples from diverse regions will also be crucial. Overall, this paper makes important contributions by proposing and validating a novel DC-CRF algorithm for optimizing complex seawater spectral analyses.

AUTHORS' CONTRIBUTIONS

All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by S. Hou. The first draft of the manuscript was written by S. Hou, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was financially supported by the Taishan Scholars Program, the National Natural Science Foundation of China (No. U2006222), the National Key Research and Development Program of China (No. 2020ZLHY04), and the Qilu University of Technology (Shandong Academy of Sciences) unveiling project (No. 2022JBZ01-02-02).

CODE AVAILABILITY

The code generated during the current study is available from the corresponding author on reasonable request.

DECLARATIONS

All authors have read, understood, and have complied as applicable with the statement on ‘Ethical responsibilities of authors’ as found in the Instructions for Authors.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abolfathi, S. & Pearson, J. M. 2014 Solute dispersion in the nearshore due to oblique waves. *Coastal Engineering Proceedings* **34**, 2156–1028.

Abolfathi, S. & Pearson, J. 2017 Application of smoothed particle hydrodynamics (SPH) in nearshore mixing: A comparison to laboratory data. *Coastal Engineering Proceedings* **35**, 35v.currents.16.

Abolfathi, S., Cook, S., Yeganeh-Bakhtiary, A., Borzooei, S. & Pearson, J. 2020 Microplastics transport and mixing mechanisms in the nearshore region. *Coastal Engineering Proceedings* **36v**, 63–63.

Cao, H., Qu, W. & Yang, X. 2014 A rapid determination method for chemical oxygen demand in aquaculture wastewater using the ultraviolet absorbance spectrum and chemometrics. *Analytical Methods* **6** (11), 3799–3803.

Chen, B., Wu, H. & Li, S. F. Y. 2014 Development of variable pathlength UV–Vis spectroscopy combined with partial-least-squares regression for wastewater chemical oxygen demand (COD) monitoring. *Talanta* **120**, 325–330.

Chen, X., Yin, G., Zhao, N., Gan, T., Yang, R., Xia, M. & Huang, Y. 2021 Simultaneous determination of nitrate, chemical oxygen demand and turbidity in water based on UV–Vis absorption spectrometry combined with interval analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **244**, 118827.

Deng, W., Zhao, H., Yang, X., Xiong, J., Sun, M. & Li, B. 2017 Study on an improved adaptive PSO algorithm for solving multi-objective gate assignment. *Applied Soft Computing* **59**, 288–302.

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. & Huvenne, J.-P. 2009 Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems* **96** (1), 27–33.

Fan, M., Liu, X., Yu, X., Cui, X., Cai, W. & Shao, X. 2017 Near-infrared spectroscopy and chemometric modelling for rapid diagnosis of kidney disease. *Science China Chemistry* **60**, 299–304.

Fearn, T. 2002 Assessing calibrations: SEP, RPD, RER and R2. *NIR News* **13** (6), 12–13.

Feng, Q., Chen, H., Xie, H., Cai, K., Lin, B. & Xu, L. 2020 A novel genetic algorithm-based optimization framework for the improvement of near-infrared quantitative calibration models. *Computational Intelligence and Neuroscience* **2020**, 7686724.

Fogelman, S., Zhao, H. & Blumenstein, M. 2006 A rapid analytical method for predicting the oxygen demand of wastewater. *Analytical and Bioanalytical Chemistry* **386**, 1773–1779.

Galvao, R. K. H., Araujo, M. C. U., José, G. E., Pontes, M. J. C., Silva, E. C. & Saldanha, T. C. B. 2005 A method for calibration and validation subset partitioning. *Talanta* **67** (4), 736–740.

Gao, Q., Wang, P., Niu, T., He, D., Wang, M., Yang, H. & Zhao, X. 2022 Soluble solid content and firmness index assessment and maturity discrimination of *Malus micromalus Makino* based on near-infrared hyperspectral imaging. *Food Chemistry* **370**, 131013.

Guo, Y., Liu, C., Ye, R. & Duan, Q. 2020 Advances on water quality detection by UV-Vis spectroscopy. *Applied Sciences* **10** (19), 6874.

Habib, M. A., O’Sullivan, J. J., Abolfathi, S. & Salauddin, M. 2023 Enhanced wave overtopping simulation at vertical breakwaters using machine learning algorithms. *Plos one* **18** (8), e0289318.

Han, X., Xie, D., Song, H., Ma, J., Zhou, Y., Chen, J. & Huang, F. 2022 Estimation of chemical oxygen demand in different water systems by near-infrared spectroscopy. *Ecotoxicology and Environmental Safety* **243**, 113964.

Kim, Y.-C., Sasaki, S., Yano, K., Ikebukuro, K., Hashimoto, K. & Karube, I. 2001 Photocatalytic sensor for the determination of chemical oxygen demand using flow injection analysis. *Analytica Chimica Acta* **432** (1), 59–66.

Kjeldahl, K. & Bro, R. 2010 Some common misunderstandings in chemometrics. *Journal of Chemometrics* **24** (7-8), 558–564.

Kusnierek, K. & Korsaeth, A. 2015 Simultaneous identification of spring wheat nitrogen and water status using visible and near infrared spectra and powered partial least squares regression. *Computers and Electronics in Agriculture* **117**, 200–213.

Lee, K.-H., Ishikawa, T., McNiven, S., Nomura, Y., Hiratsuka, A., Sasaki, S. & Karube, I. 1999 Evaluation of chemical oxygen demand (COD) based on coulometric determination of electrochemical oxygen demand (EOD) using a surface oxidized copper electrode. *Analytica Chimica Acta* **398** (2–3), 161–171.

Li, H.-D., Xu, Q.-S. & Liang, Y.-Z. 2012 Random frog: An efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. *Analytica Chimica Acta* **740**, 20–26.

Li, P., Qu, J., He, Y., Bo, Z. & Pei, M. 2020 Global calibration model of UV-Vis spectroscopy for COD estimation in the effluent of rural sewage treatment facilities. *RSC Advances* **10** (35), 20691–20700.

Lin, H., Li, H., Yang, X., Xu, Z., Tong, Y. & Yu, X. 2020 Comprehensive investigation and assessment of nutrient and heavy metal contamination in the surface water of coastal Bohai Sea in China. *Journal of Ocean University of China* **19**, 843–852.

Luo, N., Yang, X., Sun, C., Xing, B., Han, J. & Zhao, C. 2021 Visualization of vibrational spectroscopy for agro-food samples using t-distributed stochastic neighbor embedding. *Food Control* **126**, 107812.

Miller, D. G., Brayton, S. V. & Boyles, W. T. 2001 Chemical oxygen demand analysis of wastewater using trivalent manganese oxidant with chloride removal by sodium bismuthate pretreatment. *Water Environment Research* **73** (1), 63–71.

Picollo, M., Aceto, M. & Vitorino, T. 2018 UV-Vis spectroscopy. *Physical Sciences Reviews* **4** (4), 20180008.

Qi, Y., Li, C., Jiang, P., Jia, C., Liu, Y. & Zhang, Q. 2018 Research on demodulation of FBGs sensor network based on PSO-SA algorithm. *Optik* **164**, 647–653.

Rohman, A., Che Man, Y. B., Ismail, A. & Hashim, P. 2010 Application of FTIR spectroscopy for the determination of virgin coconut oil in binary mixtures with olive oil and palm oil. *Journal of the American Oil Chemists' Society* **87**, 601–606.

Shamsipur, M., Zare-Shahabadi, V., Hemmateenejad, B. & Akhond, M. 2006 Ant colony optimisation: A powerful tool for wavelength selection. *Journal of Chemometrics: A Journal of the Chemometrics Society* **20** (3-4), 146–157.

Sharifzadeh, S., Ghodsi, A., Clemmensen, L. H. & Ersbøll, B. K. 2017 Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection. *Engineering Applications of Artificial Intelligence* **65**, 168–177.

Sun, J., Yang, W., Feng, M., Liu, Q. & Kubar, M. S. 2020 An efficient variable selection method based on random frog for the multivariate calibration of NIR spectra. *RSC Advances* **10** (28), 16245–16253.

Tran, T. N., Blanchet, L., Afanador, N. L. & Buydens, L. M. 2015 Novel unified framework for latent modeling and its interpretation. *Chemometrics and Intelligent Laboratory Systems* **149**, 127–139.

Wold, S., Sjöström, M. & Eriksson, L. 2001 PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58** (2), 109–130.

Xu, H., Liu, G., Lu, M. & Mao, R. 2014 A Least Random Shuffled Frog-Leaping Algorithm. In *Foundations of Intelligent Systems: Proceedings of the Eighth International Conference on Intelligent Systems and Knowledge Engineering (ISKE 2013)*, November 2013, Shenzhen, China. Springer, Cham, Switzerland, pp. 417–425.

Xu, X., Xiong, G., Chen, G., Fu, T., Yu, H., Wu, J. & Shi, X. 2021 Characteristics of coastal aquifer contamination by seawater intrusion and anthropogenic activities in the coastal areas of the Bohai Sea, eastern China. *Journal of Asian Earth Sciences* **217**, 104830.

Yeganeh-Bakhtiary, A., EyvazOghli, H., Shabakhty, N. & Abolfathi, S. 2023 Machine learning prediction of wave characteristics: Comparison between semi-empirical approaches and DT model. *Ocean Engineering* **286**, 115583.

Yun, Y.-H., Li, H.-D., Wood, L. R., Fan, W., Wang, J.-J., Cao, D.-S. & Liang, Y. Z. 2013 An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **111**, 31–36.

Zhan, X.-r., Zhu, X.-r., Shi, X.-y., Zhang, Z.-y. & Qiao, Y.-j. 2009 Determination of hesperidin in tangerine leaf by near-infrared spectroscopy with SPXY algorithm for sample subset partitioning and Monte Carlo cross validation. *Spectroscopy and Spectral Analysis* **29** (4), 964–968.

Zhang, J., Cui, X., Cai, W. & Shao, X. 2019 A variable importance criterion for variable selection in near-infrared spectral analysis. *Science China Chemistry* **62**, 271–279.

Zhu, X., Li, S., Shan, Y., Zhang, Z., Li, G., Su, D. & Liu, F. 2010 Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics. *Journal of Food Engineering* **101** (1), 92–97.

First received 26 June 2023; accepted in revised form 2 February 2024. Available online 28 March 2024