Infrared Physics & Technology 133 (2023) 104827

Contents lists available at ScienceDirect

Infrared Physics and Technology

journal homepage: www.elsevier.com/locate/infrared

Determination of residual levels of procymidone in rapeseed oil using near-infrared spectroscopy combined with multivariate analysis

Mingxing Zhao [a], Hui Jiang [a,*], Quansheng Chen [b,*]

[a] School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China
[b] College of Ocean Food and Biological Engineering, Jimei University, Xiamen 361021, China

ARTICLE INFO

Keywords:
Rapeseed oil
Pesticide residues
Near infrared spectroscopy
Quantitative analysis

ABSTRACT

The issue of pesticide residues has always been a hot topic at home and abroad. A method for the quantitative detection of procymidone residues in grain and oil products using near-infrared (NIR) spectroscopy has been proposed. First, a NIR spectrometer was used to collect spectral data from rapeseed oil samples with different concentrations of procymidone residues. Based on full-spectrum data, the wavelength points selected by bootstrapping soft shrinkage (BOSS) algorithm, competitive adaptive reweighted sampling (CARS) algorithm, and variable combination population analysis (VCPA) algorithm then were compared and were quantified using support vector regression (SVR) model. Simultaneously, the prediction results of the SVR model optimized by dung beetle optimizer (DBO) algorithm and pigeon-inspired optimization (PIO) algorithm were compared using the full-spectrum data. Finally, the wavelength selection algorithms and parameter optimization algorithms with the best prediction results were selected for comparison and combination. In light of the outcomes, the three spectral characteristic wavelength selection algorithms and the two optimization algorithms can improve the coefficient of determination (R_P^2) and reduce the root mean square error of prediction (RMSEP). The SVR model that utilizing CARS and PIO algorithm demonstrates the best generalization performance among all models evaluated, and the R_P^2 is 0.9939 with a RMSEP of 2.3435 mg·kg^{-1}. The results indicate that the high-precision and rapid detection of procymidone in edible oil can be achieved using the SVR model optimized by input feature and parameter based on NIR spectral data. This has great significance in ensuring the safety of grain and oil food.

1. Introduction

Food safety is an important issue related to public health, and pesticide residues is an important factor affecting food quality and safety [1]. Long-term exposure or consumption of food containing pesticide residues will cause adverse effects on human health, and even lead to poisoning and immune system damage and chronic diseases [2]. Procymidone, a systemic fungicide, is mainly used to control the development of disease spots in crops, such as rape, watermelon and strawberry [3]. Rapeseed oil is pressed from rapeseed fruit and is popular for its high nutritional content [4]. In recent years, with the rising demand for rapeseed oil and the increasing expansion of planting area [5], the use of procymidone in the growth process of rapeseed has also been increasing. At the same time, consumers are increasingly concerned about pesticide residues in grain and oil products [6]. Hence, the efficient detection of procymidone residues in rapeseed oil has great practical significance and application value.

At present, the conventional methods for detecting pesticide residues in grain and oil products mainly include chromatography, biosensor technology and immunoassay [7]. Chromatographic methods, such as gas chromatography (GC) and high-performance liquid chromatography (HPLC), are commonly used for the separation and quantitative analysis of pesticide residues, providing high efficiency and accuracy in detection, but these methods are demanding and time-consuming [8]. Biosensor technology uses enzymes, microorganisms, cells, etc. to detect pesticide residues in vegetable oil, which has the characteristics of simple operation and high sensitivity, but the method has high cost and limitations in specific diversified detection [9]. The immunoassay is measured by preparing specific antibodies combined with pesticide residues, which has the characteristics of less time-consuming. However, the preparation process of antibodies in this method is cumbersome and expensive, which is not conducive to popularization [10]. The above

* Corresponding authors.
E-mail addresses: h.v.jiang@ujs.edu.cn (H. Jiang), chenqs@jmu.edu.cn (Q. Chen).

https://doi.org/10.1016/j.infrared.2023.104827
Received 3 May 2023; Received in revised form 13 June 2023; Accepted 12 July 2023
Available online 13 July 2023

methods cannot meet the needs of on-site testing of large numbers of samples. Therefore, an efficient, rapid and economical method of pesticide residue detection is required in the actual detection process.

Near-infrared (NIR) spectroscopy is a non-destructive analysis method for the structure and composition analysis of substances [11]. It measures the absorption and scattering of molecules in the NIR spectral range to determine relevant parameters such as the composition and mass of analyte, with a spectral range of 700–2526 nm [12]. This non-contact detection technology has several advantages, such as not requiring sample pretreatment, fast detection speed, no pollution, and no destruction. It also enables simultaneous analysis of multiple components, making it an efficient and cost-effective analytical method [13]. Recently, as progresses made in the field of modern electronics, spectral analysis and computers, the increasing perfection of NIR analysis technology has facilitated its application in detection of agricultural products [14–19]. In particular, in the detection of pesticide residues in edible oils, Xue et al. quickly detected chlorpyrifos residues in corn oil using a one-dimensional convolutional neural network structure based on a deep learning model of NIR spectroscopy [20]. This report confirms the enormous potential of NIR spectroscopy technology in the quality and safety testing of edible oils, but the detection model used in this study is relatively complex, and the detection accuracy still has room to improve.

Usually, there are two main types of multivariate correction models commonly used in NIR spectral data processing: linear modeling and nonlinear modeling. Currently, the partial least squares (PLS) method is widely used for qualitative and quantitative analysis of food products [21]. And PLS regression is extremely widely used in the field of spectral analysis because it can cope with the problem of multiple covariance of spectroscopy [22,23]. However, when dealing with some more complex sample sets, nonlinear regression methods have more unique advantages. For example, support vector regression (SVR) is a good choice for nonlinear regression problems with small sample sets. On the other hand, since most wavelength points may not be relevant to the target under study, feature selection of the full-spectrum data is needed to filter out the wavelength points that are relevant to the target. Doing so can improve model accuracy, reduce computational complexity and avoid overfitting. In addition, for SVR models, we can also optimize the model parameters using an intelligent optimization algorithm, which can further improve the model accuracy.

Accordingly, the following research objectives were proposed. (1) Acquisition and preprocessing of NIR spectral data. Rapeseed oil samples containing different concentrations of procymidone were configured and the original NIR spectra were acquired by spectrometer. (2) Feature extraction. Three characteristic variable selection algorithms were introduced, namely bootstrapping soft shrinkage (BOSS) algorithm, competitive adaptive reweighted sampling (CARS) algorithm and variable combination population analysis (VCPA), to screen and gain suitable characteristic wavelength points. (3) Parameter optimization of the model. Based on the full-spectrum data, the dung beetle optimizer (DBO) algorithm and the pigeon-inspired optimization (PIO) algorithm were employed to perfect the parameters in the SVR model. (4) Compared and analyzed the wavelength variable selection algorithms and parameter optimization algorithms with the best prediction effect.

2. Materials and methods

2.1. Preparation and acquisition of experimental samples

The procymidone standard (concentration greater than 99%) and n-hexane (chromatographic grade) were procured from Shanghai Aladdin Company, and nine brands of rapeseed oil were purchased from JD Mall for the experiment.

When preparing the sample, first weighed 200 mg of the procymidone standard with an electronic balance and dissolved it in a chromatography-grade n-hexane solvent to make standard solution of different concentrations, which were 1, 2, 3.5, 5, 6, 7, 8, 10, 20, 35, 50, 60, 70, 80, 100, 200, 350, 500, 600, 700, 800 and 1000 mg·kg^{-1}. A brand of rapeseed oil was then added to obtain samples of rapeseed oil with different concentrations of procymidone (i.e., mixing 18 g of oil with 2 g of standard solution). Finally, there were 22 concentration gradients of procymidone in the samples configured, which were 0.1, 0.2, 0.35, 0.5, 0.6, 0.7, 0.8, 1, 2, 3.5, 5, 6, 7, 8, 10, 20, 35, 50, 60, 70, 80 and 100 mg·kg^{-1}. For nine different brands of rapeseed oil, a total of 198 samples were obtained.

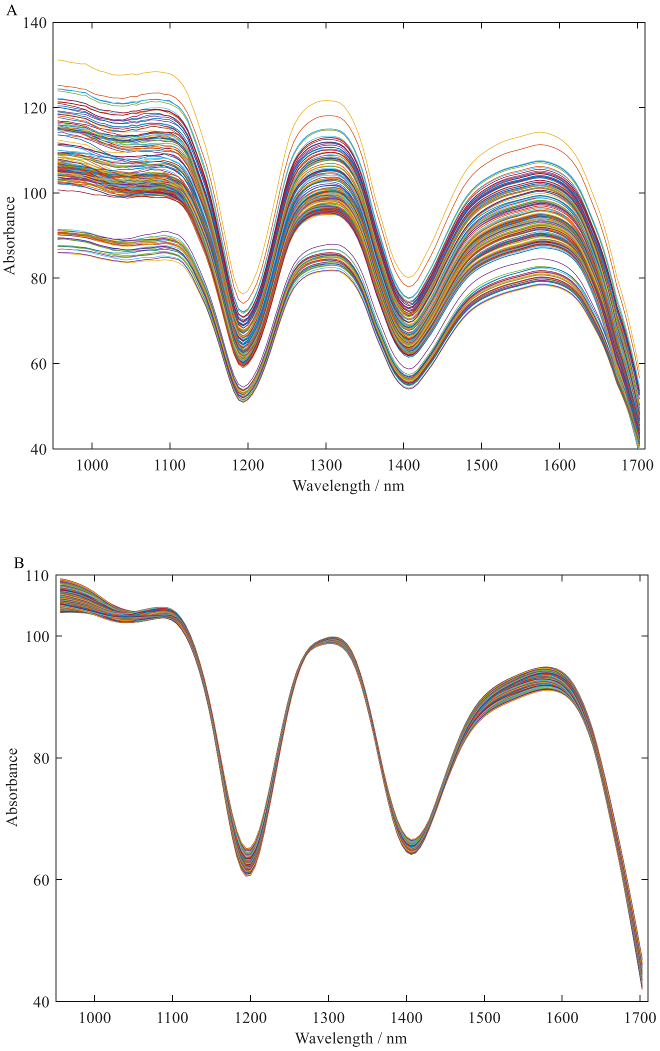

Fig. 1. Original NIR spectra and pretreated spectra by SG and MSC of all rapeseed oil samples.

2.2. Experimental apparatus and spectral sampling

Weighed using an electronic balance (Mettler Toledo Instruments Co., Ltd, Shanghai, China) with an actual index value of 0.01 mg. A pipette with a size of 1000 μL was used to aspirate 2 mL of the sample in a 5 mm width cuvette and a Flame-NIR spectrometer (Ocean Insights, USA) was employed to get the spectra of samples, and the acquired spectral data was recorded and stored with OceanView software (Ocean Insights, USA).

Prior to data acquisition, the following parameters were configured for the spectrometer: the integration time was 20 ms, the spectral scanning range was 950–1700 nm, the number of wavelength points measured was 128, and the empty cuvette was used as the spectral reference. During the spectral data acquisition process, samples were measured three times and the average of those measurements was taken as original date.

2.3. Spectral data preprocessing

The NIR spectrum can rapidly and precisely indicate the composition and structure of the substance, but it is also affected by the measured sample, the response of the spectroscopic instrument, the environment and other factors. Moreover, the interference information cannot be completely eliminated by relying on the equipment itself or improving the external environment. Therefore, prior to spectral data analysis, a series of preprocessing steps to clean and correct the data is necessary [24], so as to eliminate interference signals in the spectral data, improve the data quality and establish a reliable model.

In this study, Savitzky-Golay (SG) filtering method with a polynomial order of 2 and window size of 13 was utilized to reduce noise and enhance the peak signal-to-noise ratio of the signal. Multiplicative scatter correction (MSC) was used to eliminate the scattering influence on the surface of the material after filtering. This helped to enhance the peak information related to composition or component content in the spectra. As depicted in Fig. 1, it displays the original NIR spectra and pretreated spectra by SG and MSC of all rapeseed oil samples.

2.4. Data analyses methods

2.4.1. Bootstrapping soft shrinkage

The bootstrapping soft shrinkage (BOSS) algorithm, developed by Deng in 2016, is designed to screen out information variables with collinearity [25]. The algorithm combines bootstrap sampling (BSS) and weighted bootstrap sampling (WBS) to randomly combine variables and build sub-models. The model population analysis (MPA) method is applied to retrieve the information in the sub-models. Finally, the best set of variables is selected as a subset with the lowest root mean square error of cross-validation (RMSECV) values during the iteration. The algorithm streamlines the process of shrinking the feature variable space, while simultaneously minimizing the risk of discarding pertinent variables during optimization.

2.4.2. Competitive adaptive reweighted sampling

The competitive adaptive reweighted sampling (CARS) algorithm is a wavelength variable screening method that simulates the concept of “survival of the fittest” to select the optimal combination of wavelengths from the entire spectra [26]. The algorithm consists of 4 steps: Monte Carlo sampling (MCS), exponentially decreasing function (EDF) forced wavelength reduction, adaptive reweighted sampling (ARS) competing wavelength reduction, and calculation of subset RMSECV values. The final subset that achieves the smallest RMSECV value is considered as the optimal subset. The algorithm effectively eliminates unwanted information from spectral information while compressing collinear variables.

2.4.3. Variable combination population analysis

The variable combination population analysis (VCPA) algorithm is a wavelength selection method that considers the interaction between variables [27]. The algorithm is based on the exponentially decreasing function (EDF) and binary matrix sampling (BMS) to identify the optimal subset, and consists of the following 4 steps: BMS sampling, EDF forced variable reduction, wavelength variable selection according to the lowest 10% RMSECV based on modeling, and investigation of all possible combinations of final variables. The algorithm uses EDF to continuously narrow the variable space, with many parameters, few selected variables, low calculation amount and fast speed.

2.4.4. Pigeon-inspired optimization

The pigeon-inspired optimization (PIO) algorithm is a swarm intelligence optimization algorithm that takes inspiration by the homing behavior of pigeon [28]. The algorithm simulates the easy homing behavior of domestic pigeons through three guidance tools: geomagnetic field information, sun altitude information and landmark information, and obtains the optimal position through iterative update of map compass operator and landmark operator. PIO is known for its simplicity, ease of implementation, and strong robustness, making it suitable for a variety of optimization problems. It also has strong global search ability, enabling it to quickly find the global optimal solution.

In this study, the SVR model parameters were optimized using the PIO algorithm. The optimization algorithm parameters were set as follows: the maximum number of iterations was 50, the population number was 20, the lower limit was [0, 0], and the upper limit was [32, 32].

2.4.5. Dung beetle optimizer

The dung beetle optimizer (DBO) algorithm, proposed in 2022, is a novel swarm intelligence optimization technique that excels in both rapid convergence speed and high solution accuracy when used for global search and local utilization [29]. The inspiration for this algorithm comes from the behavior of dung beetles in daily life, such as rolling, dancing, foraging, stealing, and breeding. And 5 different update rules are designed accordingly. The DBO algorithm mainly includes four processes: rolling balls, reproduction, foraging and stealing, and the main idea is to treat each individual dung beetle as a possible feasible solution in a given search space. According to the different change rules of the design, it is continuously iterated towards the trend of smaller adaptation function values and updated positions in real time, and finally the best position is output.

In this study, the SVR model parameters were optimized by using the DBO algorithm. The optimization algorithm parameters were set as follows: the maximum number of iterations was 50, the population size was 20, the lower limit was [0.0001, 0.0001], and the upper limit was [200, 200].

2.4.6. Diagnosis of nonlinearity

In this study, it was verified whether there was a nonlinear relationship between NIR spectrum and the concentration of the pesticide procymidone. Assuming that the relationship between the two was linear, the PLS model was used to predict the content of procymidone in the sample. The partial residual plot (APaRP) method recommended by Mallows was then used to diagnose nonlinearity. A quantitative numerical tool was used in this study to determine the nonlinearity based on the APaRPs method [30]. When the expected randomness measure value is greater than 1.96, it can be determined that the relationship between the two is nonlinear.

2.4.7. Support vector regression

The support vector regression (SVR) is a common regression method based on support vector machine (SVM) [30]. The goal is to minimize the prediction error of the model and add penalty coefficients to the model to control model complexity. When predicting under nonlinear conditions, it is necessary to construct a mapping function to expand the spatial dimension, and its core idea is to find an optimal hyperplane in high-dimensional space to describe the relationship between characteristic variables, which has the advantages of excellent generalization performance and strong robustness. The kernel function in this study was radial basis function (RBF) selected from the libSVR toolkit. The optimization of the parameters was carried out using grid search (GS). Its search interval was [2^{-10}, 2^{-10}], and the search step size was 0.5.

2.5. Model evaluation

In this study, various indicators were employed to evaluate the performance of the feature variable selection method, algorithm optimization, and model generalization ability, including root mean square error of prediction (RMSEP) and coefficient of predictive determination

Table 1
Statistics of procymidone values in the training set and the prediction set for rapeseed oil.

Subsets	Number of samples	Units	Maximum	Minimum	Mean	Standard deviation
Training set	154	mg kg^{-1}	100	0.1	20.9432	29.9978
Prediction set	44	mg kg^{-1}	100	0.1	20.9432	30.2460

Table 2
The results of the runs test used to detect the non-linearity of NIR spectral signals and procymidone residue values by the APaRPs method.

n_+	n_-	u	μ	σ	$\lvert z\rvert$	Conclusion
23.8955	79.5993	1	37.7568	12.8231	10.1249	Nonlinearity

(R_P^2). Their calculations are published as follows:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_p}\left(y_i - y_i^{'}\right)^2}{n_p}} \tag{1}$$

$$R_P^2 = 1 - \frac{\sum_{i=1}^{n_p}\left(y_i - y_i^{'}\right)^2}{\sum_{i=1}^{n_p}\left(y_i - y_p^{''}\right)^2} \tag{2}$$

where, $y_i, y_i^{'}, y_p^{''}$ are the measured value, predicted value, and average value of the correction set, respectively. n_p is the sample size of the correction set.

2.6. Software

All algorithms were implemented on a computer with an Intel i5-11400H CPU, 16 GB RAM, running MATLAB R2021a (MathWorks, Natick, USA) under the Windows 10 operating system.

3. Results and discussion

3.1. Spectral dataset partitioning strategies

The experiment was conducted in 22 batches, each with nine samples. Among them, seven samples from each batch were selected randomly for model training, and the remaining samples were used for model prediction. In this way, there are 154 samples in the training set and 44 samples in the prediction set. Table 1 shows the statistics of procymidone values in the training set and prediction sets for rapeseed oil. Table 1 indicates that the mean and standard deviation (SD) of both the training set and the prediction set exhibit negligible disparities. Therefore, this division scheme is reasonable and reliable.

3.2. Results of diagnosis of nonlinearity

Table 2 presents the test results, the $|z|$ value is 10.1249, indicating a nonlinear relationship between NIR spectral signals and procymidone residue values. In order to accurately predict the residual amount of procymidone in rapeseed oil, this study employed nonlinear algorithms SVR to establish a regression model for predicting procymidone in rapeseed oil.

3.3. Analysis and comparison of BOSS, CARS and VCPA

When selecting the wavelength variables of the original spectra, there is a certain randomness at initialization due to the different selection strategies of BOSS, CARS and VCPA. To reduce the influence of initialization randomness, each of the three feature wavelength extraction algorithms was run 50 times, and the results of 50 runs were recorded. Fig. 2 displays the results of running BOSS, CARS, and VCPA separately 50 times, along with the distribution of wavelength points that correspond to the optimal outcome of three optimization algorithms. Among them, Fig. 2A displays the results of running the SVR model 50 times with three wavelength selection methods. In Fig. 2A, it can be obtained that BOSS achieves a mean RMSEP of 5.0392 mg·kg^{-1} with a SD of 0.6561 mg·kg^{-1}, CARS achieves a mean RMSEP of 4.8986 mg·kg^{-1} with a SD of 0.7160 mg·kg^{-1}, and VCPA achieves a mean RMSEP of 5.1867 mg·kg^{-1} with a SD of 1.5676 mg·kg^{-1}. Referring to the results in Fig. 2A, it is evident that the three algorithms exhibit some level of randomness, but the influence of this randomness on the performance of the SVR model is limited to minor fluctuations.

Fig. 2B displays the distribution of wavelength variables over the entire spectra for the optimal SVR model obtained from the three methods. As shown in Fig. 2B, the number of wavelength points selected by these three wavelength extraction algorithms varies greatly. Among them, VCPA selects the least number of wavelength points, 13, accounting for 10.2% of the entire spectrum; CARS picked the most, with 48, accounting for 37.5% of the entire spectra. The reason for this may be closely related to the difference in the selection strategy of each method. In addition, Fig. 2B reveals that the three different variable selection methods select many identical wavelength points as input feature variables to build the training model, which shows that the selected wavelength points are reasonable and targeted. Table 3 shows the best prediction results of SVR model combined with different variable selection algorithms. In Table 3, it is not difficult to see that three different wavelength screening methods can improve the prediction performance of the model and reduce the redundancy of the data. In particular, CARS performed best, screening 48 characteristic wavelengths. Compared to the model using the entire spectra, its RMSEP decreases from 4.2408 mg·kg^{-1} to 3.2223 mg·kg^{-1} and R_P^2 increases from 0.9799 to 0.9884. Based on comprehensive consideration, 48 feature variables selected by the CARS algorithm were finally selected as the final input feature variables of the model. According to the literature, the fourth overtone located near the 1160 nm wavelength belongs to the C=O stretching [31]. The band around 1400 nm is related to O—H absorption and absorption band related to C—H absorption are observed at about 1120, 1300 and 1360 nm [32]. They are related to the organic substances in procymidone.

3.4. Analysis and comparison of DBO-SVR, PIO-SVR and CARS-PIO-SVR

In this study, the parameters of the SVR model were optimized using the GS method. However, this method can only search a limited discrete parameter space, which has high discretization requirements and large computational requirements. The swarm intelligent optimization algorithm can not only search continuous space, but also has fast calculation speed, which can optimize high-dimensional, nonlinear and complex problems. Therefore, based on the full-spectrum data, the DBO and PIO algorithm with different parameter ranges due to performance effects were introduced to perfect the parameters of the model, and the prediction effect after parameter optimization was analyzed and compared. Fig. 3 shows the results of running different optimization strategies 50 times separately combined with SVR model. Fig. 3A shows the performances of the SVR model in the prediction set after running the SVR model 50 times separately under the premise of DBO and PIO optimization. Fig. 3A reveals that the RMSEP value of these two swarm intelligent optimization algorithms is lower than that of 4.2408 mg·kg^{-1} of the GS method, indicating that the swarm intelligent optimization algorithm has better results in optimizing SVR model parameters. In

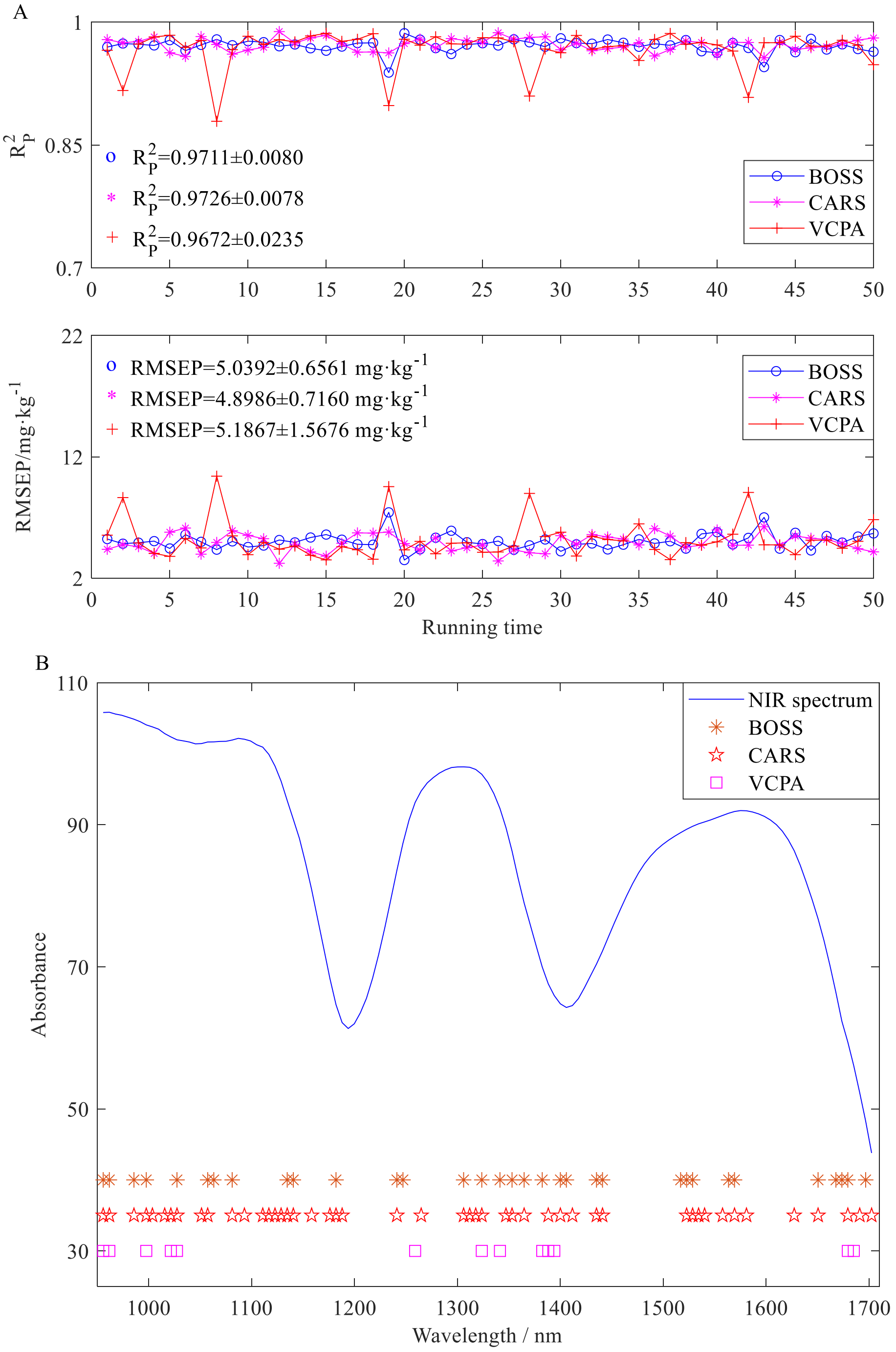

Fig. 2. The results of running BOSS, CARS, and VCPA separately 50 times, and the distribution of wavelength points that correspond to the best outcome of three optimization algorithms.

Table 3
The best prediction results of SVR model combined with different variable selection algorithms.

Methods	Number of input features	Parameters		Training set		Prediction set	
		C	g	R_C^2	RMSEC/ mg kg^{-1}	R_P^2	RMSEP/ mg kg^{-1}
Raw	128	1024	0.0055	0.9922	2.6246	0.9799	4.2408
BOSS	33	1024	0.0313	0.9947	2.1436	0.9864	3.4899
CARS	48	1024	0.0221	0.9958	1.9201	0.9884	3.2223
VCPA	13	1024	0.1250	0.9912	2.7379	0.9862	3.5032

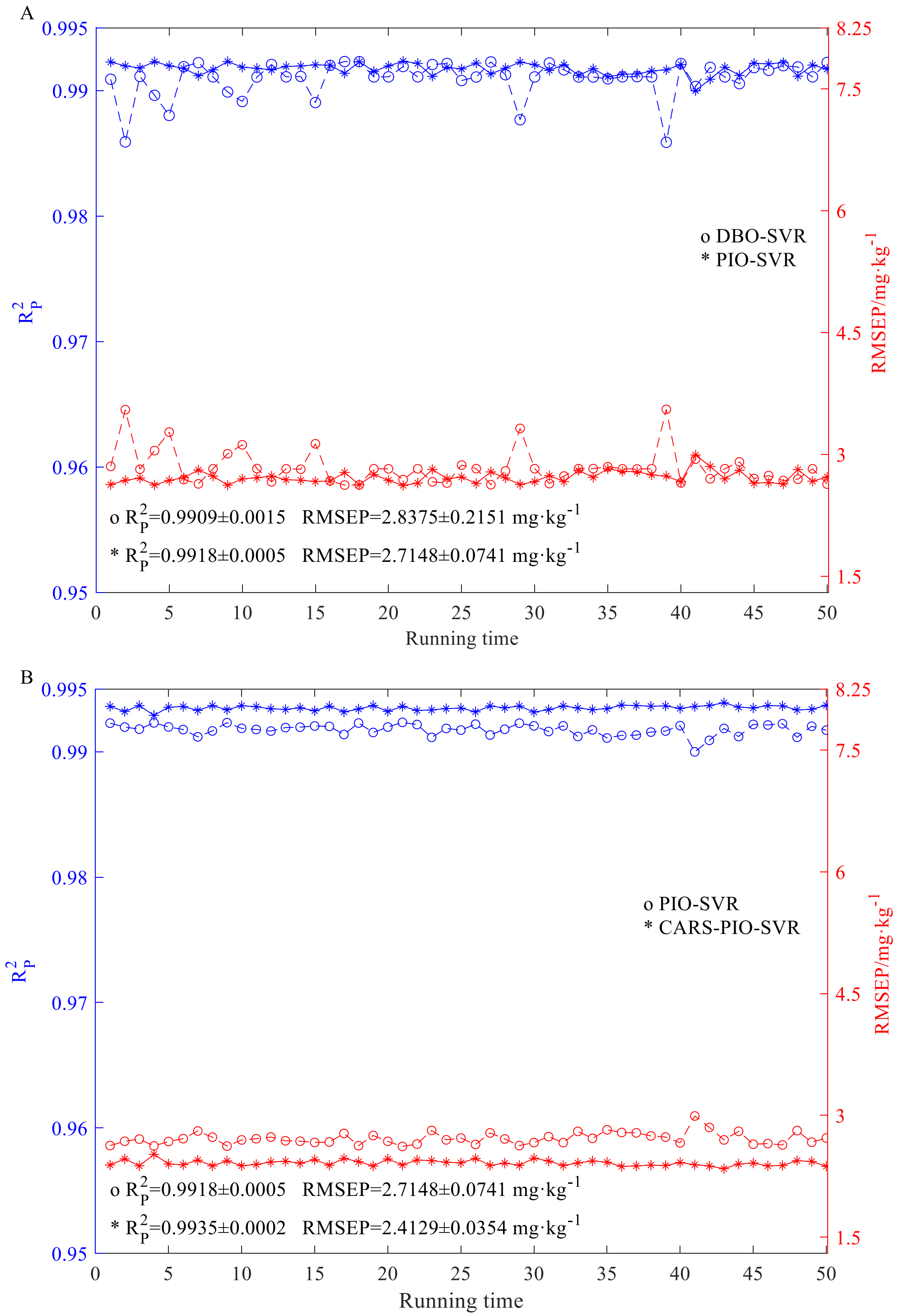

Fig. 3. The results of running different optimization strategies 50 times separately combined with SVR model.

addition, the mean RMSEP of the DBO-SVR model is 2.8375 mg·kg^{-1} with a SD of 0.2151 mg·kg^{-1}, and the R_P^2 is 0.9909 with a SD of 0.0015. The mean RMSEP of the PIO-SVR model is 2.7148 mg·kg^{-1} with a SD of 0.0741 mg·kg^{-1}, and the R_P^2 is 0.9918 with a SD of 0.0005. Therefore, compared with the DBO-SVR model, the PIO algorithm combined with the SVR model has better prediction effect, higher prediction accuracy and stronger stability. For this reason, the PIO algorithm was finally selected as the model parameter optimization method for subsequent analysis.

Through the above analysis, the best feature selection algorithm and model parameter optimization algorithm, namely CARS and PIO, were selected respectively, and the two methods were coupled in order to establish the best detection model, that was, the wavelength points selected by CARS were used as the input feature, and PIO was the parameter optimization algorithm of SVR. To verify the feasibility of this method, the PIO-SVR model established by 48 characteristic wavelength points optimized by CARS was compared with the prediction results of the PIO-SVR model established by the entire spectra. Fig. 3B shows the results of the prediction set after running 50 times independently of the PIO-SVR model with the entire spectra as the input feature, and the PIO-SVR model with 48 wavelength points optimized by CARS as the input

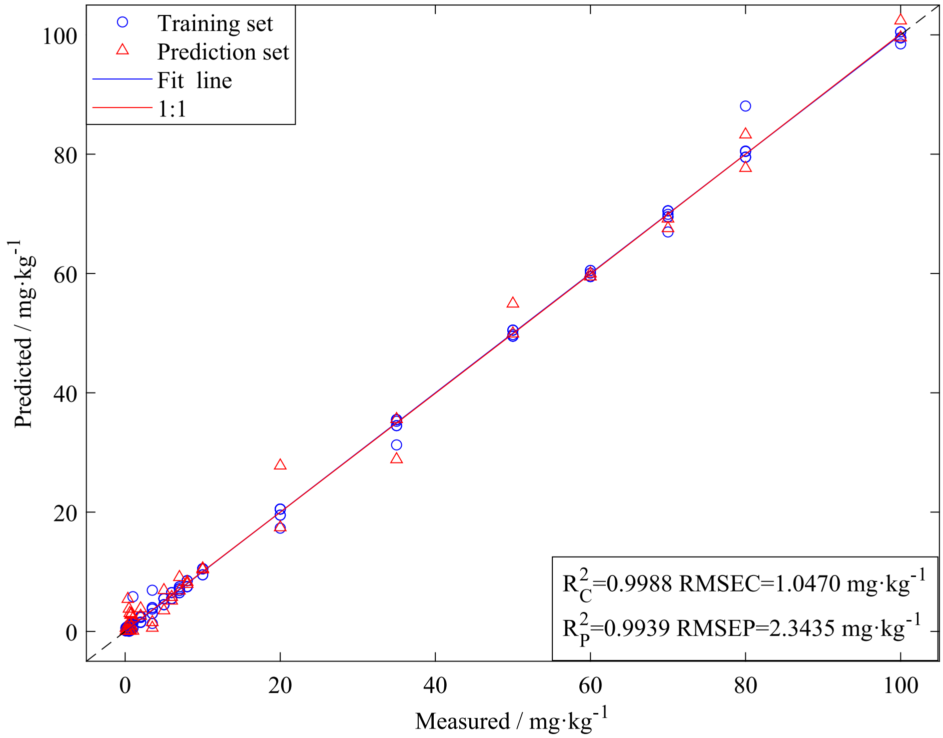

Fig. 4. Scatterplot between the model prediction value and the reference value based on the optimal CARS-PIO-SVR model for the content of procymidone in rapeseed oil samples.

Table 4
The best prediction results of SVR model under three different optimization strategies.

Models	Number of input features	Parameters				Prediction set	
		C	g	R_C^2	RMSEC/ mg kg^{-1}	R_P^2	RMSEP/ mg kg^{-1}
SVR	128	1024	0.0055	0.9922	2.6246	0.9799	4.2408
CARS-SVR	48	1024	0.0221	0.9958	1.9201	0.9884	3.2223
PIO-SVR	128	31.6688	0.1373	0.9987	0.4651	0.9923	2.6173
CARS-PIO-SVR	48	26.8422	0.1670	0.9988	1.0470	0.9939	2.3435

feature after running independently for 50 times. From Fig. 3B, it can be seen that the mean RMSEP of the CARS-PIO-SVR model is 2.4129 mg kg^{-1} with a SD of 0.0354 mg kg^{-1}, and the mean R_P^2 is 0.9935 with a SD of 0.0002. Compared with the PIO-SVR model, the mean R_P^2 of CARS-PIO-SVR is improved, and the mean RMSEP decreases by about 0.3 mg kg^{-1}. The results show that the PIO-SVR model established by screening the full-spectrum data by CARS is effective, which can improve the R_P^2 and reduce the RMSEP. Fig. 4 shows a scatterplot between the model prediction value and the reference value based on the optimal CARS-PIO-SVR model for the content of procymidone in rapeseed oil samples. From Fig. 4, the 1:1 line and the fitted line of the model are almost fitted, and the predicted scatterplots of both training set and prediction set are close to and distributed along the fitted line, which indicates that the model is successfully established.

3.5. Comparison of SVR model with different optimization strategies

In this study, three optimization strategies were proposed based on NIR spectral data, namely input feature variable optimization, model parameter optimization and model parameter optimization based on feature variable optimization under full-spectrum data. Table 4 illustrates the best prediction results of SVR models under three different optimization strategies. As Table 4 shown that both the feature wavelength selection of the entire spectra and the optimization of model parameters can improve the accuracy of model prediction. Compared with the SVR model, the CARS-PIO-SVR model based on C = 26.8422 and g = 0.1670 had the best prediction results, and its R_P^2 increased from 0.9799 to 0.9939, and the RMSEP decreased from 4.2408 mg kg^{-1} to 2.3435 mg kg^{-1}. As can be seen from Table 4, when only the optimization algorithm was utilized to perfect the model parameters, the model may have the risk of overfitting, indicating that the model is too adapted to the training data. However, if feature screening is performed on full-spectrum data first, and then model parameters are optimized, this situation can be effectively avoided. The reason for this is that the model is too complex in fitting the training data, and mistakenly learns all the features in the training set instead of the common features of the data, resulting in the model not generalizing well to the new data. Therefore, the optimization strategy combining the selection of characteristic wavelength points and the optimization of model parameters is the best solution, which not only ensures the accuracy of model prediction, but also avoids the situation of overfitting the training set.

4. Conclusions

In this study, NIR spectroscopy was used to quantify the procymidone residues in rapeseed oil. Three feature extraction algorithms (namely BOSS, CARS, VCPA) were used to screen the full-spectrum data, and combined with the SVR model, the feature extraction algorithm and feature variables with the best prediction results were selected. Using the full-spectrum data as the input of the SVR model, the stability and accuracy of the two optimization algorithms of DBO and PIO were compared. The results show that the SVR model optimized by PIO with 48 wavelength points selected by CARS as feature inputs exhibits the best generalization performance. The results show that NIR spectroscopy can be applied to the rapid and accurate detection of procymidone in grain and oil products, and also provides a powerful tool for the rapid detection of other pesticide residues and food safety monitoring.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the National Key Research and Development Program of China (Grant No. 2017YFC1600603).

References

[1] S. Van Boxstael, I. Habib, L. Jacxsens, M. De Vocht, L. Baert, E. Van De Perre, A. Rajkovic, F. Lopez-Galvez, I. Sampers, P. Spanoghe, B. De Meulenaer, M. Uyttendaele, Food safety issues in fresh produce: bacterial pathogens, viruses and pesticide residues indicated as major concerns by stakeholders in the fresh produce chain, Food Control 32 (2013) 190–197.

[2] A.M. Taiwo, A review of environmental and health effects of organochlorine pesticide residues in Africa, Chemosphere 220 (2019) 1126–1140.

[3] W. Wang, Z. Gao, C. Qiao, F. Liu, Q. Peng, Residue analysis and removal of procymidone in cucumber after field application, Food Control 128 (2021), 108168.

[4] Z. Ye, Y. Liu, Polyphenolic compounds from rapeseeds (Brassica napus L.): the major types, biofunctional roles, bioavailability, and the influences of rapeseed oil processing technologies on the content, Food Res Int 163 (2023), 112282.

[5] Q. Hu, W. Hua, Y. Yin, X. Zhang, L. Liu, J. Shi, Y. Zhao, L. Qin, C. Chen, H. Wang, Rapeseed research and production in China, The Crop Journal 5 (2017) 127–135.

[6] B.H. Jensen, A. Petersen, P.B. Petersen, T. Christensen, S. Fagt, E. Trolle, M. E. Poulsen, J. Hinge Andersen, Cumulative dietary risk assessment of pesticides in food for the Danish population for the period 2012–2017, Food Chem Toxicol 168 (2022), 113359.

[7] T. Thorat, B.K. Patle, M. Wakchaure, L. Parihar, Advancements in techniques used for identification of pesticide residue on crops, J. Nat. Pesticide Res. 4 (2023), 100031.

[8] N.A. Shad, A. Munawar, Y. Javed, A. Rakha, A. Riaz, S.U. Din, I. Zareef, M.M. Sajid, M.F. Khan, S. Akhtar, M. Salman, In-field deployable and facile nanosensor for the detection of pesticides residues, Anal. Chim. Acta 1259 (2023), 341204.

[9] U. Chadha, P. Bhardwaj, R. Agarwal, P. Rawat, R. Agarwal, I. Gupta, M. Panjwani, S. Singh, C. Ahuja, S.K. Selvaraj, M. Banavoth, P. Sonar, B. Badoni, A. Chakravorty, Recent progress and growth in biosensors technology: a critical review, J. Ind. Eng. Chem. 109 (2022) 21–51.

[10] X. Tang, Q. Zhang, Z. Zhang, X. Ding, J. Jiang, W. Zhang, P. Li, Rapid, on-site and quantitative paper-based immunoassay platform for concurrent determination of pesticide residues and mycotoxins, Anal Chim Acta 1078 (2019) 142–150.

[11] Q. Wu, M.M. Oliveira, E.M. Achata, M. Kamruzzaman, Reagent-free detection of multiple allergens in gluten-free flour using NIR spectroscopy and multivariate analysis, J. Food Compos. Anal. 120 (2023), 105324.

[12] J. Deng, H. Jiang, Q. Chen, Characteristic wavelengths optimization improved the predictive performance of near-infrared spectroscopy models for determination of aflatoxin B1 in maize, J. Cereal Sci. 105 (2022), 103474.

[13] H. Hao, S. Cheng, Z. Ren, L. Zhang, B. Wang, N. Li, Q. Bao, J. Feng, F. Hu, C. Liu, S. Zhang, X. Jian, Rapidly and accurately determining the resin and volatile content of CF/PPBESK thermoplastic prepreg by NIR spectroscopy, Compos. A Appl. Sci. Manuf. 169 (2023), 107517.

[14] S. Grassi, C. Alamprese, Advances in NIR spectroscopy applied to process analytical technology in food industries, Curr. Opinion in Food Science 22 (2018) 17–21.

[15] Q. Jiang, M. Zhang, A.S. Mujumdar, D. Wang, Non-destructive quality determination of frozen food using NIR spectroscopy-based machine learning and predictive modelling, J. Food Eng. 343 (2023), 111374.

[16] J.U. Porep, D.R. Kammerer, R. Carle, On-line application of near infrared (NIR) spectroscopy in food production, Trends Food Sci. Technol. 46 (2015) 211–230.

[17] H. Ning, J. Wang, H. Jiang, Q. Chen, Quantitative detection of zearalenone in wheat grains based on near-infrared spectroscopy, Spectrochim. Acta. A Mol. Biomol. Spectrosc. 280 (2022), 121545.

[18] H. Jiang, T. Liu, Q. Chen, Dynamic monitoring of fatty acid value in rice storage based on a portable near-infrared spectroscopy system, Spectrochim Acta. A Mol. Biomol. Spectrosc. 240 (2020), 118620.

[19] H. Jiang, J. Wang, Q. Chen, Comparison of wavelength selected methods for improving of prediction performance of PLS model to determine aflatoxin B1 (AFB1) in wheat samples during storage, Microchem. J. 170 (2021), 106642.

[20] Y. Xue, C. Zhu, H. Jiang, Comparison of the performance of different one-dimensional convolutional neural network models-based near-infrared spectra for determination of chlorpyrifos residues in corn oil, Infrared Phys. Technol. 132 (2023), 104734.

[21] X. Chen, Y. Xu, L. Meng, X. Chen, L. Yuan, Q. Cai, W. Shi, G. Huang, Non-parametric partial least squares–discriminant analysis model based on sum of ranking difference algorithm for tea grade identification using electronic tongue data, Sens. Actuators B 311 (2020), 127924.

[22] Z. Xie, X.a. Feng, X. Chen, Partial least trimmed squares regression, Chemometrics and Intelligent Laboratory Systems, 221 (2022) 104486.

[23] L. Su, W. Shi, X. Chen, L. Meng, L. Yuan, X. Chen, G. Huang, Simultaneously and quantitatively analyze the heavy metals in Sargassum fusiforme by laser-induced breakdown spectroscopy, Food Chem. 338 (2021), 127797.

[24] X. Zhang, J. Sun, P. Li, F. Zeng, H. Wang, Hyperspectral detection of salted sea cucumber adulteration using different spectral preprocessing techniques and SVM method, Lwt 152 (2021), 112295.

[25] H. Yan, X. Song, K. Tian, J. Gao, Q. Li, Y. Xiong, S. Min, A modification of the bootstrapping soft shrinkage approach for spectral variable selection in the issue of over-fitting, model accuracy and variable selection credibility, Spectrochim. Acta. A Mol. Biomol. Spectrosc. 210 (2019) 362–371.

[26] Y. Li, X. Yang, Quantitative analysis of near infrared spectroscopic data based on dual-band transformation and competitive adaptive reweighted sampling, Spectrochim. Acta A Mol. Biomol. Spectrosc. 285 (2023), 121924.

[27] H. Zhao, K.-W. Huan, X.-G. Shi, F. Zheng, L.-Y. Liu, W. Liu, C.-Y. Zhao, A Variable selection method of near infrared spectroscopy based on automatic weighting variable combination population analysis, Chin. J. Anal. Chem. 46 (2018) 136–142.

[28] H. Alazzam, A. Sharieh, K.E. Sabri, A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer, Expert Syst. Appl. 148 (2020), 113249.

[29] J. Xue, B. Shen, Dung beetle optimizer: a new meta-heuristic algorithm for global optimization, J. Supercomput. 79 (2022) 7305–7336.

[30] T. Liu, H. Jiang, Q. Chen, Input features and parameters optimization improved the prediction accuracy of support vector regression models based on colorimetric sensor data for detection of aflatoxin B1 in corn, Microchem. J. 178 (2022), 107407.

[31] D. Brunet, T. Woignier, M. Lesueur-Jannoyer, R. Achard, L. Rangon, B.G. Barthes, Determination of soil content in chlordecone (organochlorine pesticide) using near infrared reflectance spectroscopy (NIRS), Environ. Pollut. 157 (2009) 3120–3125.

[32] M.T. Sanchez, K. Flores-Rojas, J.E. Guerrero, A. Garrido-Varo, D. Perez-Marin, Measurement of pesticide residues in peppers by near-infrared reflectance spectroscopy, Pest. Manag. Sci. 66 (2010) 580–586.