Original Article
Food Science and Technology

ISSN 0101-2061 (Print)
ISSN 1678-457X (Online)

DOI: https://doi.org/10.1590/fst.113422

Rapid determination of cadmium residues in tomato leaves by Vis-NIR hyperspectral and Synergy interval PLS coupled Monte Carlo method

Shupeng ZENG[1], Xiaohong WU[1,2], Bin WU[3], Haoxiang ZHOU[4], Meng WANG[4*]

Abstract

Excessive heavy metal cadmium in tomatoes is harmful to human health. The detection of heavy metals in tomato leaves can determine whether the heavy metals in tomatoes exceed the standard. In order to quickly, non-destructively and efficiently detect whether heavy metals on the surface of tomato leaves exceed the standard, a new wavelength interval selection method called Synergy interval PLS couple with Monte Carlo method (MC-siPLS) was proposed. From the seedling stage, concentration of 0, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10 mg/L cadmium chloride (CdCl2) was used for irrigation under normal nutrient elements respectively. A total of 405 leaf samples were collected. Using the VIS-NIR hyperspectral instrument, the tomato leaves were set as the region of interest to obtain hyperspectral data, then used Atomic Absorption Spectrometry (AAS) method to detect heavy metals in tomato leaves. In addition, 5 different PLS algorithm were used to compare with MC-siPLS. Furthermore, the best model was given by MC-siPLS with RMSEP = 0.5378, R^2 = 0.9870. The results show that MC-siPLS was better than similar wavelength interval selection methods, which had great application potential in the nondestructive detection of heavy metals in tomato leaves. This method maybe determines whether tomatoes contain excess cadmium early.

Keywords: tomato leaves; heavy metal; visible/near infrared hyperspectral imaging; synergy interval partial least squares; Monte Carlo method.

Practical Application: Vis-NIR hyperspectral imaging technique is combined with partial least squares regression to detect the cadmium content on the tomato leaves. It has many advantages that traditional detection methods do not have, such as non-destructive, convenient and fast.

1 Introduction

Tomato is one of the most important crops in the world, which is very popular among consumers in China. Related research shows that the large amount of vitamin C contained in tomatoes which is beneficial to human health (Zhang et al., 2017). Excessive cadmium in soil is a major problem for agriculture (Hédiji et al., 2010). Although cadmium is not an essential element for the growth of tomatoes (Sanità di Toppi & Gabbrielli, 1999), it is easily absorbed and accumulated by tomatoes (Hédiji et al., 2010; Abdel-Latif, 2008). In addition, some scholars pointed out that cadmium affects not only tomato leaves, but also tomato fruits. The higher the heavy metal cadmium residue in the leaves, the higher the heavy metal content in the fruit (Hédiji et al., 2010). At the same time, leaves are organs closely related to photosynthesis and respiration, so it is necessary to detect and analyze the content of heavy metals in tomato leaves under cadmium stress. Through the detection of tomato leaves, it is possible to determine whether the tomato may be contaminated by cadmium in advance (Carvalho et al., 2018), and take relevant measures in time to avoid more harm to the human body and economic losses.

Up till the present moment, many scholars have studied the effect of heavy metals on tomatoes: Ramadan & Al-Ashkar (2007) investigated the effect of different fertilizers on heavy metals in soil and tomato plants. Yaqvob et al. (2011) studied two types of tomatoes under heavy metal stress and showed that some higher doses of heavy metals may lead to metabolic disturbance and growth inhibition in plant species. Piotto et al. (2018), used cadmium as a case study to study the tolerance of tomato to heavy metal toxicity, and the experiment showed that the dry weight of tomato decreased with the increase of CdCl2 concentration. Baruah et al. (2019), studied the effects of heavy metals on seed germination and seedling growth in wheat, pea and tomato, they found that tomato seeds were most sensitive during the germination stage and were also most sensitive to cadmium and copper. Bounar et al. (2020), used atomic absorption spectrometry to determine heavy metals in greenhouse-grown tomatoes and to assess human health risks. Most of the above studies on tomatoes and heavy metals have used chemical analysis methods, however, these physical and chemical indicators need to be carried out in the laboratory. The disadvantage is that the steps are cumbersome, the cost is

Received 10 Oct., 2022
Accepted 02 Dec., 2022
[1]*School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, China*
[2]*High-tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province, Jiangsu University, Zhenjiang, China*
[3]*Department of Information Engineering, Chuzhou Polytechnic, Chuzhou, China*
[4]*Department of Electrical and Control Engineering, Research Institute of Zhejiang University-Taizhou, Taizhou, China*
*Corresponding author: 403288077@qq.com

high, and the experiment time is long, and it is not suitable for large-scale tomato gardens.

Non-destructive testing technology was a hot topic in recent years. It was deeply loved by food safety scholars for its non-destructiveness, rapidity and accuracy. Many scholars used spectroscopy to carry out non-destructive testing, and near-infrared spectroscopy was popular in the field of food non-destructive testing. Khan et al. (2021), used NIR spectroscopy and partial least squares technique to evaluate the quality of milk powder. Silva et al. (2021) used near-infrared and mid-infrared spectroscopy and distinguish the origin of Coalho cheese. Tripaldi et al. (2022), used near-infrared technology to study the effect of frozen curd on the chemical properties and oxidative modification of Mozzarella cheese. Wang et al. (2022), used FT-NIR and LDA techniques to analyze green tea species. At the same time, hyperspectral technology was also widely used in non-destructive testing. Ma et al. (2022), used fractal theory and hyperspectral imaging technology to detect apple soluble solids. Zou et al. (2022), determined the moisture content in potato tubers by using hyperspectral and machine learning techniques. In the research on tomatoes, many scholars have also proposed various non-destructive testing methods. Zhang et al. (2021), used Vis/NIR technology and multivariate algorithm to evaluate the soluble solid content in tomatoes at different stages. Brito et al. (2022), used a portable Vis-NIR spectrometer to determine the color, titratable acidity and dry matter in whole tomatoes. Sun et al. (2021), used hyperspectral to study the effect of bruised tomato on drop and fruit size. These studies showed that non-destructive testing was fully applied in the field of tomato safety and quality, which has greatly improved the quality of human life. Meanwhile, it also showed the potential and advantages of non-destructive testing.

Partial least squares (PLS) has been developed for many years as an important algorithm in spectral regression analysis. The commonly used partial least squares regression method such as interval PLS (iPLS) (Norgaard. et al 2000), moving windows PLS (MWPLS) (Cheng et al., 2017), changeable size moving window PLS (CS-MWPLS) (Du et al., 2004), backward interval PLS (biPLS) (Leardi & Nørgaard, 2004), dynamic backward interval PLS (Song et al., 2020), and synergy interval PLS (siPLS) (Jiang et al., 2012), interval random frog (iRF) (Yun et al., 2013) and interval successive projection algorithm (iSPA) (De Araújo Gomes et al., 2013). A representative algorithm of the interval selection method is iPLS, this algorithm divides the spectrum into equal intervals then builds a PLS model on each sub-interval, choosing one or several intervals with the smallest root mean square error of cross validation (RMSECV) as the best calibration model (Wang et al., 2018). However, the problem of iPLS is that it ignores the relationship between intervals, and the calibration model established by interval with the smallest RMSECV may not be the best one (Yun et al., 2019). In order to solve the above problem backward interval PLS (biPLS) and synergy interval PLS (siPLS) were proposed to optimize the number of combined intervals by considering various interval combinations. These two algorithms are improved versions of iPLS, they select the best model through different interval combinations, which effectively improves the interpretability of the model (Yun et al 2013). MWPLS is another type of interval selection method that uses a fixed size window to move throughout the full spectrum, then choose the informative interval with low RMSECV and low model complexity as the final model. The difference between CS-MWPLS and MWPLS is using different numbers and size of windows in the process of moving windows (Wang et al., 2019). These methods do not optimize the interval size, the number of combinations, and the interval position at the same time. From the point of view of spectroscopy and chemistry, interval selection methods are more promising than wavelength selection methods due to the better interpretability and reliability of the model (Wang et al., 2018). In this study, we tried to optimize the interval division and interval selection by Monte Carlo method (MC). Then the optimized algorithm was used to detect the heavy metal cadmium on the surface of tomato leaves.

In this paper, NIR spectroscopy and six PLS algorithms were combined to detect the content of heavy metal cadmium on the tomato leaves, so as to achieve the identification of tomato food safety. The specific steps are as follows: (1) Obtain the hyperspectral image of tomato leaves using VIS-NIR hyperspectral spectrometer; (2) Obtain the cadmium content in tomato leaves using AAS atomic method; (3) Establishment of leaf surface hyperspectral and cadmium content models using six regression algorithms; (4) Compare the prediction results of six regression algorithms.

2 Materials and methods

2.1 Experiment materials

The tomato dataset consists of 405 tomato leaf samples. The seeds came from Hongwei Seed Industry, Shandong, China. The seeds were grown in the modern Agriculture laboratory of Jiangsu University. From the seedling stage (6–7 leaves), concentration of 0, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10 mg/L cadmium chloride (CdCl2) was used for irrigation under normal nutrient elements respectively. At the flowering stage, a leafless tomato leaf was randomly selected from each plant and had whole leaf mesophyll at the same leaf position. A total of 405 samples were collected, including 45 leaf samples at each concentration.

2.2 Vis-NIR hyperspectral collection

The visible-near infrared (Vis-NIR) hyperspectral data acquisition device consists of an imaging spectrometer with a spectral resolution of 2.8 nm (ImSpector V10E, Spectral Imaging Co. Ltd., Oulu, Finland), an illumination device consisting of two 150W fiber optical halogen lamps (2900-ER +9596-E, Illumination, USA), camera obscura (SC100, Beijing optical instrument factory, China) and electric displacement platform (MTS120, Beijing optical instrument factory, China). Selecting the whole leaf of tomato as the region of interest (ROI) to extract the hyperspectral data. Finally, the visible-near-infrared (Vis-NIR) hyperspectral data were collected by the above equipment with a spectral resolution of 2.9 nm and a wavenumber range of 431.05-962.45 nm (including 618 spectral channels).

After that, using Atomic Absorption Spectrometry (AAS) method to detect heavy metals in tomato leaves. The content of cadmium in the collected sample leaves was detected according to the cadmium heavy metal detection step in the Chinese

national standard GB5009.15-2014 (National Health and Family Planning Commission of the People's Republic of China, 2015).

2.3 Spectral preprocessing

Appropriate spectral preprocessing methods may improve model robustness and interpretability (Sun et al., 2021). In this section 4 pre-processing methods (MSC, SNV, SG-1, SG-2) have been applied to the raw spectrum. Figure 1 shows the raw hyperspectral image, each spectral curve represents the represents a leaf sample. The reference values of four preprocessing methods were the mean of the spectrum. MSC can effectively eliminate spectral differences caused by different scattering levels, enhancing the correlation between spectra and data. SNV eliminate the influence of solid particle size, surface scattering and optical path change on NIR. SG1 and SG2 can improve the smoothness of the spectrum and reduce the interference of noise. The hyperspectral data was preprocessed by 4 methods in Figure 2. We used partial least squares to decide which pre-processing method is best for this data. Table 1 shows the results of 4 different preprocessing methods. From the table, the raw spectrum works best for partial least squares. So, this dataset would not be preprocessed.

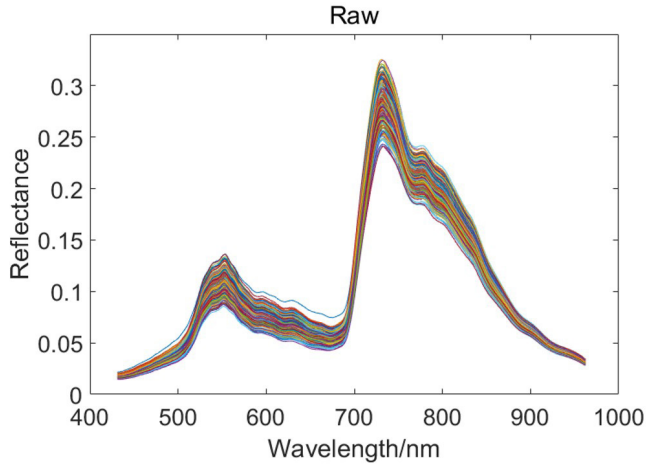

Figure 1. Raw spectral image.

2.4 Synergy interval PLS (siPLS)

Before introducing siPLS, we must introduce interval PLS (iPLS) first. IPLS was first proposed by L. Nørgaard et al., in 2000 (Norgaard et al., 2000). It is a basic sub-interval selection method for other interval selection methods. IPLS divides the full spectral into equal sub-intervals and builds PLS model on each sub-interval. The sub-interval which has the minimum RMSECV is used to build the calibration model. The idea of iPLS is to find a sub-interval with the smallest RMSECV to replace the full-band spectrum to build the calibration model. Although this method effectively avoids irrelevant variables,

Table 1. Performance of four spectral preprocessing methods on tomato data using the partial least squares algorithm.

Pre-processing	Optimal PLS Components	RMSE	R^2
Raw	9	0.6398	0.9900
MSC	15	0.8269	0.9825
SNV	18	0.7878	0.9842
SG1	5	0.6291	0.9898
SG2	7	0.7321	0.9863

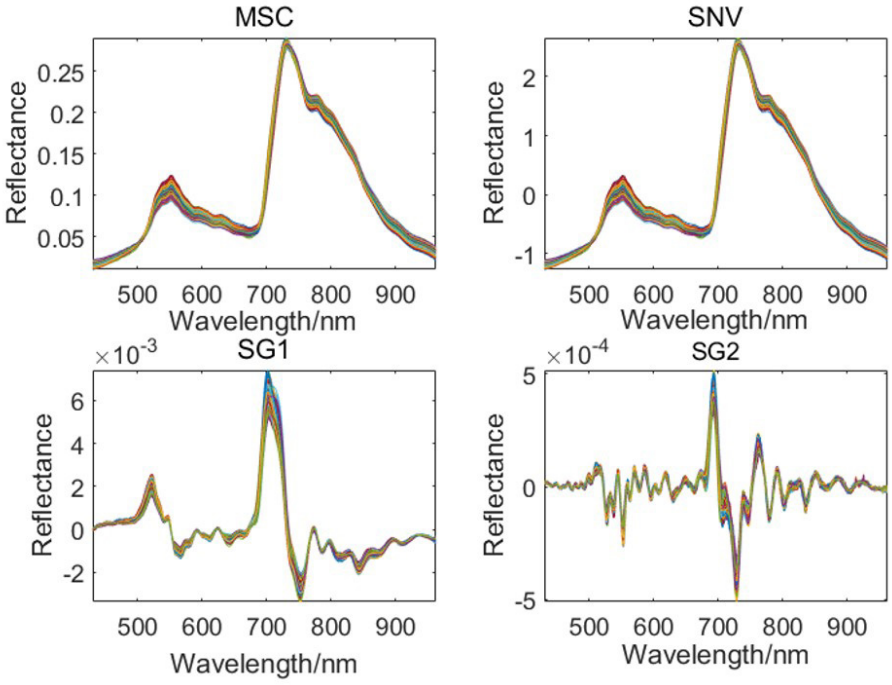

Figure 2. Spectral images after 4 different preprocessing method.

but whole spectrum is replaced by only a sub-interval, and the interpretation of single interval model is not strong enough.

SiPLS has been improved on the basis of iPLS (Yang et al., 2020). It divides the spectrum into equally sub-intervals then calculate all sub-interval PLS models on all combinations of 2, 3, or 4 intervals. For example, if siPLS considers combining n intervals of m intervals, the number of sub-intervals PLS models created by siPLS is $C_m^n = \frac{m!}{n!(m-n)!}$. Finally, the sub-interval combination with the lowest RMSECV is selected as the calibration model. Compared to the iPLS, siPLS adds an interval combination function on the basis of iPLS, and it takes into account the unequal distribution of variables on the spectrum (Norgaard et al., 2000; Jiang et al., 2012), so the calibration model is more interpretative.

In this study, siPLS was used to combine different numbers of intervals for building sub-PLS calibration models on each sub-interval combination and selecting the optimal sub-interval combination. The method of dividing the spectrum will be determined by the Monte Carlo method which is described as follow.

2.5 Monte Carlo method (MC method)

Monte Carlo method was first proposed by Nicholas Metropolis et al., in 1949 (Metropolis & Ulam, 1949). Due to the development of science and technology and the invention of electronic computers, a numerical calculation method guided by probability and statistics theory was proposed. It has been used in many fields such as Financial Engineering, and Macroeconomics, Computational Physics (Shapiro 2003).

In this study, MC method was used to divide the interval as required. A large number of siPLS models with different interval division methods determined by the MC method were produced, and then the best sub-interval combination was selected through the minimum RMSECV of the calibration model by synergy interval PLS. More detailed steps of the MC-siPLS are described below

2.6 Synergy interval PLS couple with Monte Carlo method (MC-siPLS)

The interval wavelength selection method MC-siPLS is based on siPLS and Monte Carlo method. In order to run the algorithm efficiently, some parameters need to be set before running the algorithm.

(1) M: the minimum number of variables in each interval. It must be larger than N (maximum number of components);

(2) N: the maximum number of components in each sub-interval PLS model. This parameter must be smaller than the number of variables in each interval, which means N<M;

(3) P: the number of intervals divided by MC method. This parameter can be determined according to the number of spectral variables. We recommend that the number of divisions is between 10 and 30, and the default value is 20;

(4) R: the number of combined intervals. The recommended number of combined intervals is 2, 3 or 4. The default value is 3. (We tried to combine a large number of intervals such as 5 and 6, and the result was unsatisfactory and the running time was too long);

(5) Q: the number of iterators. It needs to be set to a reasonable value, if Q is too large, the program may run for a long time; too small may not be able to search for the best interval combination. We suggest that the value of this parameter can be determined by the number of spectral independent variables and the number of intervals divided by MC method. The default value is 150, and the maximum is 300.

After setting the parameters, the detailed steps of the procedure are as follows:

Step 1: Randomly divided spectrum by MC method according to the above parameters;

Step 2: Determine whether the generated interval meets the requirements: the number of each sub-interval variable must larger than the maximum number of components in PLS;

Step 3: Calculate the RMSECV of the combined sub-intervals based on the requirement set in advance and intervals divided according to the Monte Carlo method;

Step 4: Select and save the combination with the smallest RMSECV among all combinations;

Step 5: Choose the smallest RMSEVC from all the saved models as the final calibration model.

It should be noted that in Step 1, P-1 points will be randomly generated in the whole spectrum by the Monte Carlo method. The number of variables between each point including start and end should be larger than the maximum number of components. If conditions are unsatisfied, the program will regenerate the points.

Figure 3 shows the flow of the conventional siPLS algorithm, and the flow of the MC-siPLS algorithm. It can be seen from Figure 1B that MC-siPLS optimizes siPLS in interval division and selection. This is one of the main reasons why MC-siPLS results are better than siPLS. Another reason for the better results of MC-siPLS is that the final model of MC-siPLS goes through two different rounds of screening: 1) Select the best sub-intervals combination in each divided. 2) Select the best calibration model among all division determined by the MC method. In other words, the first screening can be regarded as a local screening, which searches for the best interval combination in each interval division generated by the MC method. The second screening is based on the results of the first screening, which can be considered as a global screening, choosing the best calibration model from the first screening result.

Because of the two selection processes, the best sub-model and the best interval division can be found in MC-siPLS. The improvement of MC-siPLS is that it optimizes the number

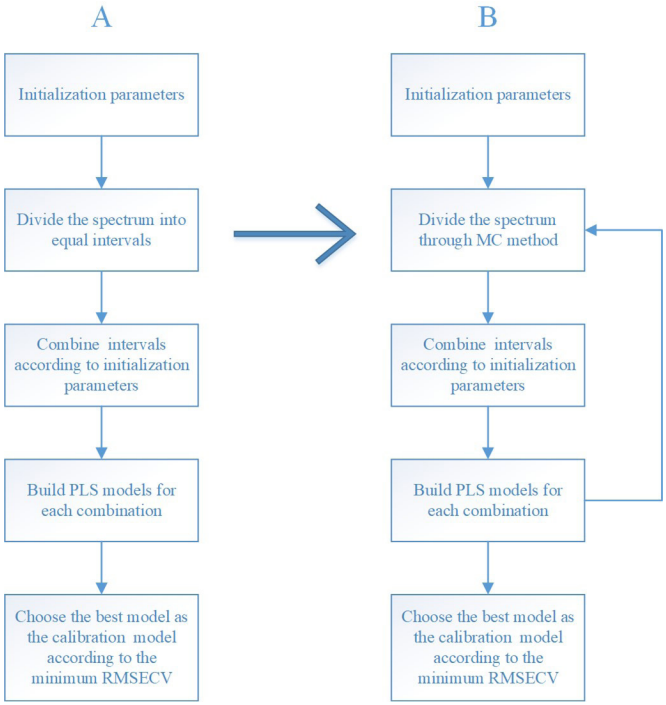

Figure 3. (A) a simple flowchart of siPLS method; and (B) A simple flowchart of MC-siPLS method.

of variables in each interval. Figure 4 shows the different ways of dividing intervals between two methods. In Figure 4A, siPLS divides the interval at equal interval which means that the number of variables in each interval is same, and in Figure 4B, MC-siPLS divides the interval by MC method. It is noteworthy that what Figure 4B shows is only one possible interval division determined by the MC method. In the next division, the number of variables contained in the first interval may be larger than that in the fifth interval or may be smaller than the sixth interval. This process of repeatedly dividing intervals and modeling will be repeated until the best interval division combination is found or the number of iterations reaches the upper limit.

Due to the introduction of the Monte Carlo method, the unequal interval division is used to replace the original division method, which makes the siPLS algorithm produce more results when enumerating intervals and building models. Consider the situation shown in Figure 4 where there are a large number of irrelevant variables in the red area and a large number of correlated variables in the green area. Under the method of dividing intervals at equal intervals, the fourth and fifth intervals are hardly selected because the red area contains a large number of irrelevant variables. However, as shown in Figure 4B, due to the introduction of the Monte Carlo method, the interval is divided into unequal intervals, and the fifth interval avoids the area where uncorrelated variables exist. The fifth interval is likely to be selected during the enumeration process because it contains a large number of correlated variables. The more relevant variables, the interpretability of the model, and the accuracy and robustness of the model will also be improved. Therefore, we believe that it is a feasible solution to introduce Monte Carlo method to optimize the interval division of siPLS.

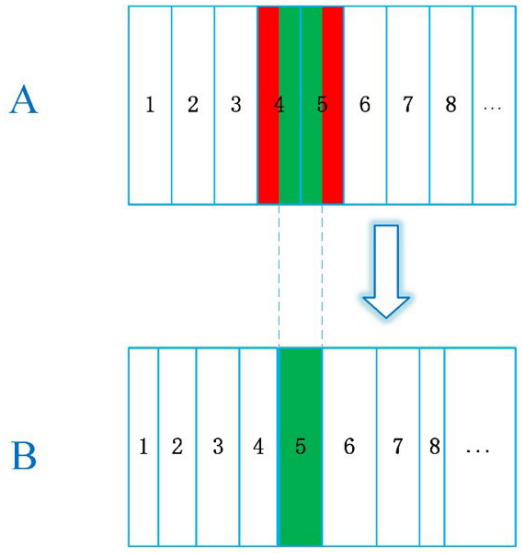

Figure 4. (A) Traditional interval selection methods dividing the spectrum at equal intervals; (B) MC-siPLS dividing the spectrum by MC method at unequal intervals.

3 Results and discussion

3.1 Estimation of model performance

Coefficient of determination R^2 and root mean square error (RMSE) are two common indicators to evaluate the ability of regression model. Coefficient of determination R^2 evaluates the model by measuring the linear correlation between independent and dependent variables. Root mean square error is not much different from variance and standard deviation in formula form, but it is obviously different in physical means. The difference is that there is a true value in the RMSE application scenario, which measures the deviation of each data from the true value. The formula of RMSE and R^2 are as (1) and (2):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \| y_i - \hat{y}_i \|_2^2} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (2)$$

where n is the number of samples; y_i and $\hat{y}_i$ are the observed value and the predicted value, respectively. $\bar{y}_i$ is mean value of observation values, The subscript i indicates the sample number from 1 to n.

The other important parameter in PLS model is PLS components. Therefore, we use the root mean squared error of cross-validation (RMSECV) to determine the best component numbers. The number of components corresponding to the smallest RMSECV is the best value. We also use root mean squared error of the prediction (RMSEP) to evaluate the performance of the calibration model on the test data sets.

In this paper, we evaluated the performance of MC-siPLS by comparing with other widely used interval selection methods, including iPLS, MWPLS, siPLS and biPLS. In some parameter selection, MC-siPLS is similar to siPLS, and they are greedy algorithm that investigates all possible combinations of intervals. Therefore, the same number of sub-intervals and the number of intervals combined are considered for both MC-siPLS and siPLS. Before running algorithms, all the datasets will be standardized, which means the mean is zero and the variance is one. The Kennard Stone (KS) method was used to divide the samples into calibration date set and test set. The calibration date set was used to select wavelength intervals and establish PLS calibration model.

The best number of PLS components was selected by 5-fold cross-validation method through the smallest RMSECV. Other detailed parameters such as the number of sub-intervals of biPLS, MWPLS and iPLS and the size of mowing windows of MWPLS will be described in different datasets.

3.2 Interval selection analysis

The tomato data was divided into calibration data set and test data set. The calibration set contains 304 samples and the test set contains 101 samples. In tomato dataset, the full spectrum was divided into 20 intervals for iPLS, biPLS, siPLS and MC-siPLS. The max PLS component was 10; size of mowing windows for MWPLS was 21; the number of combined intervals in siPLS and MC-siPLS was 3; the number of iterations in MC-siPLS was 150.

Table 2 shows the results of different algorithms on tomato dataset. From the RMSECV and R^2 of tomato dataset, the methods except MWPLS were better than the full spectrum PLS model. MC-siPLS was obviously better than other wavelength interval selection methods, and it has the smallest RMSEP (0.5378) and followed by siPLS (0.6521), iPLS (0.8281), biPLS (0.8944), PLS (0.9161) and MWPLS (0.9228) on test dataset. The intervals selected by MC-siPLS were 512.77-545.70 nm, 572.87-589.92 nm, 701.78-725.21nm and selected by siPLS were 534.70-560.12nm, 560.97-586.50 nm, 693.98-720.00 nm.

The reason for choosing these intervals may be attributed to the absorption of color and overtone of X-H (X=C, O, N, etc.) bonds. Meanwhile, the intervals selected by iPLS siPLS MWPLS and MC-siPLS had overlapping intervals and main overlapping interval range from 510nm to 589nm, which means that this band contained a large number of relevant variables and plays a key role in calibration model construction. It should be noted that the intervals selected by iPLS were 534.70-560.12 nm, and 560.97-586.50 nm. Compared to siPLS and MC-siPLS, it didn't choose the information interval approximately located at 700-720 nm. This may be one of the reasons why the results of iPLS are unsatisfactory. The general principle of the MWPLS is to generate a moving window over the entire band and all variables in the window are modeled using the PLS algorithm. Finally, the window with the smallest root mean square error was selected as the calibration model. According to this feature, MWPLS can always find the interval with the most relevant variables through the fixed window size. However, because it is only a fixed window, it lacks the function of interval combination, and selects an interval with the most relevant variables. This is one of the reasons why the results of MWPLS on calibration set and test set are quite different. BiPLS calculates the model performance after discarding one sub-interval. This way of choosing the interval is easy to fall into the local minimum situation. It finds the combination of the smallest interval, but did not try more possibilities, the result shows that RMSECV is the third smallest, however RMSEP is far larger than RMSECV.

Figure 5 shows the specified interval and overlapping interval selected by MC-siPLS and siPLS on the spectrum. The red wireframe is the spectrum selected by MC-siPLS and blue one is siPLS. The green area is the spectrum chosen by both siPLS and MC-siPLS. It is intuitive from the Figure 5 that the sub-intervals selected by siPLS and MC-siPLS were different, but the intervals selected by the two algorithms have overlapping areas.

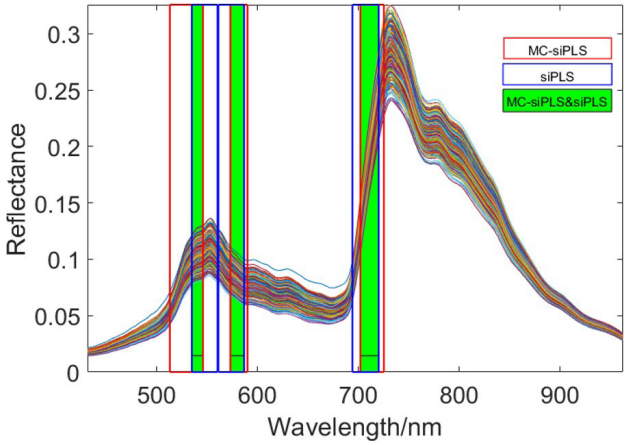

Figure 5. The selected intervals on the tomato dataset by MC-siPLS and siPLS.

Table 2. Results of different sub-intervals selection methods on the tomato dataset.

Method	Selected wavelength intervals(nm)	nVAR.	Optimal PLS comp.	RMSECV	RMSEP	R^2
PLS	431.05-962.00	618	6	0.6398	0.9161	0.9685
iPLS	534.70-560.12, 560.97- 586.50	62	5	0.6793	0.8281	0.9791
MWPLS	560.12-577.13	21	4	0.6308	0.9228	0.9675
biPLS	482.57-507.73, 587.36-613.00, 640.47-666.32, 667.18-693.12, 720.86-746.95, 747.82-773.97, 801.91-828.16, 910.61-936.09	247	8	0.5950	0.8944	0.9708
siPLS	534.70-560.12, 560.97- 586.50, 693.98-720.00	93	8	0.5249	0.6521	0.9825
MC-siPLS	512.77-545.70, 572.87-589.92, 701.78-725.21	88	8	0.4630	0.5378	0.9870

Combine the results in Table 1, the results of MC-siPLS were significantly better than siPLS. This means that MC-siPLS, as an improved version of siPLS, effectively optimizes the interval selected by siPLS and retains the advantages of the siPLS on selection interval method.

4 Conclusion

In this paper, a new interval selection method called MC-siPLS was proposed, which based on siPLS and MC method for rapid non-destructive detection of heavy metal content in tomato leaves under different cadmium stresses. Tomato leaves were selected as the region of interest (ROI) to collect hyperspectral data by the VIS-NIR hyperspectral instrument. The dataset was divided into calibration dataset and test dataset by KS method. Finally, three characteristic intervals containing 88 variables were selected to determine the cadmium content of tomato leaves, of which, RMSECV = 0.4630, RMSEP = 0.5378, R^2 = 0.9870. This study shows that MC-siPLS is an effective method to improve the accuracy of the calibration model, which had great application potential in the nondestructive detection of heavy metals in tomato leaves. We believe that in the future, if portable VIS-NIR hyperspectral equipment can be combined with this technology to predict in advance whether tomato fruit is contaminated with heavy metals, it will bring good benefits.

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.

Acknowledgements

This study was funded by National Natural Science Foundation of China (Grant No. 31471413), the Undergraduate Scientific Research Project of Jiangsu University (20AB0011), the Talent Program of Chuzhou Polytechnic (Grant No. YG2019026 and YG2019024) and Key Science Research Project of Chuzhou Polytechnic (Grant No. YJZ-2020-12).

References

Abdel-Latif, A. (2008). Cadmium induced changes in pigment content, ion uptake, proline content and phosphoenolpyruvate carboxylase activity in Triticum aestivum seedlings. *Australian Journal of Basic and Applied Sciences*, 2(1), 57-62.

Baruah, N., Mondal, S. C., Farooq, M., & Gogoi, N. (2019). Influence of heavy metals on seed germination and seedling growth of wheat, pea, and tomato. *Water, Air, and Soil Pollution*, 230(12), 1-15. http://dx.doi.org/10.1007/s11270-019-4329-0.

Bounar, A., Boukaka, K., & Leghouchi, E. (2020). Determination of heavy metals in tomatoes cultivated under green houses and human health risk assessment. *Quality Assurance and Safety of Crops & Foods*, 12(1), 76-86. http://dx.doi.org/10.15586/QAS2019.639.

Brito, A. A., Campos, F., dos Reis Nascimento, A., Damiani, C., da Silva, F. A., de Almeida Teixeira, G. H., & Cunha, L. C., Jr. (2022). Non-destructive determination of color, titratable acidity, and dry matter in intact tomatoes using a portable Vis-NIR spectrometer. *Journal of Food Composition and Analysis*, 107, 104288. http://dx.doi.org/10.1016/j.jfca.2021.104288.

Carvalho, M. E., Piotto, F. A., Gaziola, S. A., Jacomino, A. P., Jozefczak, M., Cuypers, A., & Azevedo, R. A. (2018). New insights about cadmium impacts on tomato: plant acclimation, nutritional changes, fruit quality and yield. *Food and Energy Security*, 7(2), e00131. http://dx.doi.org/10.1002/fes3.131.

Cheng, W., Sun, D. W., Pu, H., & Wei, Q. (2017). Chemical spoilage extent traceability of two kinds of processed pork meats using one multispectral system developed by hyperspectral imaging combined with effective variable selection methods. *Food Chemistry*, 221, 1989-1996. http://dx.doi.org/10.1016/j.foodchem.2016.11.093. PMid:27979190.

De Araújo Gomes, A., Galvão, R. K. H., de Araújo, M. C. U., Véras, G., & da Silva, E. C. (2013). The successive projections algorithm for interval selection in PLS. *Microchemical Journal*, 110, 202-208. http://dx.doi.org/10.1016/j.microc.2013.03.015.

Du, Y. P., Liang, Y. Z., Jiang, J. H., Berry, R. J., & Ozaki, Y. (2004). Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Analytica Chimica Acta*, 501(2), 183-191. http://dx.doi.org/10.1016/j.aca.2003.09.041.

Hédiji, H., Djebali, W., Cabasson, C., Maucourt, M., Baldet, P., Bertrand, A., Boulila Zoghlami, L., Deborde, C., Moing, A., Brouquisse, R., Chaïbi, W., & Gallusci, P. (2010). Effects of long-term cadmium exposure on growth and metabolomic profile of tomato plants. *Ecotoxicology and Environmental Safety*, 73(8), 1965-1974. http://dx.doi.org/10.1016/j.ecoenv.2010.08.014. PMid:20846723.

Jiang, H., Liu, G., Mei, C., Yu, S., Xiao, X., & Ding, Y. (2012). Measurement of process variables in solid-state fermentation of wheat straw using FT-NIR spectroscopy and synergy interval PLS algorithm. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 97, 277-283. http://dx.doi.org/10.1016/j.saa.2012.06.024. PMid:22771562.

Khan, A., Munir, M. T., Yu, W., & Young, B. R. (2021). Near-infrared spectroscopy and data analysis for predicting milk powder quality attributes. *International Journal of Dairy Technology*, 74(1), 235-245. http://dx.doi.org/10.1111/1471-0307.12734.

Leardi, R., & Nørgaard, L. (2004). Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(11), 486-497. https://doi.org/10.1002/cem.893.

Ma, X., Luo, H., Liao, J., Zhu, L., Zhao, J., & Gao, F. (2022). Study on the detection of apple soluble solids based on fractal theory and hyperspectral imaging technology. *Food Science and Technology (Campinas)*, 43, e96722. https://doi.org/10.1590/fst.96722.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247), 335-341. http://dx.doi.org/10.1080/01621459.1949.10483310. PMid:18139350.

National Health and Family Planning Commission of the People's Republic of China. (2015). *GB5009.15-2014. National standard for food safety: determination of cadmium in foods*. Beijing: National Health and Family Planning Commission of the People's Republic of China.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54(3), 413-419. http://dx.doi.org/10.1366/0003702001949500.

Piotto, F. A., Carvalho, M. E. A., Souza, L. A., Rabêlo, F. H. S., Franco, M. R., Batagin-Piotto, K. D., & Azevedo, R. A. (2018). Estimating tomato tolerance to heavy metal toxicity: cadmium as study case. *Environmental Science and Pollution Research International*, 25(27), 27535-27544. http://dx.doi.org/10.1007/s11356-018-2778-4. PMid:30051291.

Ramadan, M. A., & Al-Ashkar, E. A. (2007). The effect of different fertilizers on the heavy metals in soil and tomato plant. *Australian Journal of Basic and Applied Sciences*, 1(3), 300-306.

Sanità di Toppi, L., & Gabbrielli, R. (1999). Response to cadmium in higher plants. *Environmental and Experimental Botany*, 41(2), 105-130. http://dx.doi.org/10.1016/S0098-8472(98)00058-6.

Shapiro, A. (2003). Monte Carlo sampling methods. *Handbooks in Operations Research and Management Science*, 10, 353-425. http://dx.doi.org/10.1016/S0927-0507(03)10006-0.

Silva, L. K., Jesus, J. C., Onelli, R. R., Conceição, D. G., Santos, L. S., & Ferrão, S. P. (2021). Discriminating Coalho cheese by origin through near and middle infrared spectroscopy and analytical measures. Discrimination of Coalho cheese origin. *International Journal of Dairy Technology*, 74(2), 393-403. http://dx.doi.org/10.1111/1471-0307.12767.

Song, X., Du, G., Li, Q., Tang, G., & Huang, Y. (2020). Rapid spectral analysis of agro-products using an optimal strategy: dynamic backward interval PLS–competitive adaptive reweighted sampling. *Analytical and Bioanalytical Chemistry*, 412(12), 2795-2804. http://dx.doi.org/10.1007/s00216-020-02506-x. PMid:32090279.

Sun, Y., Pessane, I., Pan, L., & Wang, X. (2021). Hyperspectral characteristics of bruised tomatoes as affected by drop height and fruit size. *Lwt*, 141, 110863. http://dx.doi.org/10.1016/j.lwt.2021.110863.

Tripaldi, C., Palocci, G., Rinaldi, S., Di Giovanni, S., Cali, M., Renzi, G., & Costa, C. (2022). The multivariate effect of chemical and oxidative characteristics of Buffalo Mozzarella cheese produced with different contents of frozen curd. *International Journal of Dairy Technology*, 75(4), 850-863. http://dx.doi.org/10.1111/1471-0307.12888.

Wang, J., Wu, X., Zheng, J., & Wu, B. (2022). Rapid identification of green tea varieties based on FT-NIR spectroscopy and LDA/QR. *Food Science and Technology (Campinas)*, 42, 42. http://dx.doi.org/10.1590/fst.73022.

Wang, L. L., Lin, Y. W., Wang, X. F., Xiao, N., Xu, Y. D., Li, H. D., & Xu, Q. S. (2018). A selective review and comparison for interval variable selection in spectroscopic modeling. *Chemometrics and Intelligent Laboratory Systems*, 172, 229-240. http://dx.doi.org/10.1016/j.chemolab.2017.11.008.

Wang, S.-H., Zhao, Y., Hu, R., Zhang, Y.-Y., & Han, X.-H. (2019). Analysis of near-infrared spectra of coal using deep synergy adaptive moving window partial least square method based on genetic algorithm. *Chinese Journal of Analytical Chemistry*, 47(4), e19034-e19044. http://dx.doi.org/10.1016/S1872-2040(19)61150-3.

Yang, Z., Xiao, H., Zhang, L., Feng, D., Zhang, F., Jiang, M., Sui, Q., & Jia, L. (2020). Fast determination of oxides content in cement raw meal using NIR spectroscopy combined with synergy interval partial least square and different preprocessing methods. *Measurement*, 149, 106990. http://dx.doi.org/10.1016/j.measurement.2019.106990.

Yaqvob, M., Golale, A., Masoud, S., & Hamid, R. G. (2011). Influence of different concentration of heavy metals on the seed germination and growth of tomato. *African Journal of Environmental Science and Technology*, 5(6), 420-426.

Yun, Y. H., Li, H. D., Deng, B. C., & Cao, D. S. (2019). An overview of variable selection methods in multivariate analysis of near-infrared spectra. *Trends in Analytical Chemistry*, 113, 102-115. http://dx.doi.org/10.1016/j.trac.2019.01.018.

Yun, Y. H., Li, H. D., Wood, L. R., Fan, W., Wang, J. J., Cao, D. S., Xu, Q. 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. http://dx.doi.org/10.1016/j.saa.2013.03.083. PMid:23602956.

Zhang, D., Yang, Y., Chen, G., Tian, X., Wang, Z., Fan, S., & Xin, Z. (2021). Nondestructive evaluation of soluble solids content in tomato with different stage by using Vis/NIR technology and multivariate algorithms. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 248, 119139. http://dx.doi.org/10.1016/j.saa.2020.119139. PMid:33214104.

Zhang, L., Chen, F., Zhang, P., Lai, S., & Yang, H. (2017). Influence of rice bran wax coating on the physicochemical properties and pectin nanostructure of cherry tomatoes. *Food and Bioprocess Technology*, 10(2), 349-357. http://dx.doi.org/10.1007/s11947-016-1820-0.

Zou, Z., Wu, Q., Chen, J., Long, T., Wang, J., Zhou, M., Zhao, Y., Yu, T., Wang, Y., & Xu, L. (2022). Rapid determination of water content in potato tubers based on hyperspectral images and machine learning algorithms. *Food Science and Technology (Campinas)*, 42, e46522. http://dx.doi.org/10.1590/fst.46522.