Characterization of a Pesticide Formulation by Medium Wave Near-Infrared Spectroscopy with Uninformative Variable Elimination and Successive Projections Algorithm

8 authors, including:

Guo Tang
30 PUBLICATIONS **437** CITATIONS

SEE PROFILE

Xiangzhong Song
China Agricultural University
24 PUBLICATIONS **321** CITATIONS

SEE PROFILE

Hong Yan
Yale University
33 PUBLICATIONS **333** CITATIONS

SEE PROFILE

Kuangda Tian
University of Copenhagen
11 PUBLICATIONS **201** CITATIONS

SEE PROFILE

Analytical Letters

Characterization of a Pesticide Formulation by Medium Wave Near-Infrared Spectroscopy with Uninformative Variable Elimination and Successive Projections Algorithm

Guo Tang [a], Xiangzhong Song [a], Jing Hu [a], Hong Yan [a], Kaixian Qiu [a], Kuangda Tian [a], Yanmei Xiong [a] & Shungeng Min [a]

[a] College of Science , China Agricultural University , Beijing , China
Accepted author version posted online: 25 Jul 2014.Published
online: 08 Aug 2014.

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis
Taylor & Francis Group

Vibrational Spectroscopy

CHARACTERIZATION OF A PESTICIDE FORMULATION BY MEDIUM WAVE NEAR-INFRARED SPECTROSCOPY WITH UNINFORMATIVE VARIABLE ELIMINATION AND SUCCESSIVE PROJECTIONS ALGORITHM

Guo Tang, Xiangzhong Song, Jing Hu, Hong Yan, Kaixian Qiu, Kuangda Tian, Yanmei Xiong, and Shungeng Min
College of Science, China Agricultural University, Beijing, China

Near-infrared (NIR) spectroscopy, a rapid and nondestructive analytical method, has been widely used in many fields. In this paper, medium wave near-infrared (MWNIR) was used to determine the active ingredient of a deltamethrin formulation. An uninformative variable elimination-successive projections algorithm (UVE-SPA) was employed to investigate effective variables and was compared with UVE, SPA, and full-spectrum partial least squares (PLS) regression. The results indicate that MWNIR was able to determine the pesticide active ingredient and that UVE-SPA was an efficient variable selection approach by eliminating spectral redundancy and colinearity. The developed method is a meaningful exploration in the application of near-infrared spectroscopy and provides a valuable reference on pesticide quality control.

INTRODUCTION

It is necessary to develop rapid and reliable detection methods for pesticides to ensure their quality. Currently, chromatography is the most widely used technique (Lin and Hee 1998; Karasali et al. 2004; Phillips and Burns 2012). However, it requires large instruments, requires long analysis times, and considerable labor. Near-infrared (NIR) spectroscopy has been widely used as a rapid, noninvasive, and reliable technique in many fields including agriculture (Armenta, Garrigues, and Guardia 2007; Azzouz and Tauler 2008; Duan et al. 2012; Jamshidi et al. 2012), the petrochemical industry (Falla et al. 2006; Balabin, Safieva, and Lomakina 2010), and pharmacy (Porfire et al. 2012; Howland and Hoag 2013).

NIR radiation covers the range of the electromagnetic spectrum between 700 and 2500 nm. Medium wave near-infrared (MWNIR) ranges from 1100 to 1800 nm and contains the first overtone region of C-H, N-H, and O-H and the second combination of C-H (Silverstein, Basseler, and Morill 1981; Balabin and Safieva 2008; Zou et al. 2010). Generally, models are built using the whole range of NIR spectroscopy (Blanco and Peguero 2010; Dupuy et al. 2010; Sinelli et al. 2010; Puchert et al. 2010; Xiong et al. 2010). However, some of the regions contain little useful information. Furthermore, wider waveband collection sets higher demands on the instrument, which not only increases the cost but also limits the application. As MWNIR contains the most important spectral information (Silverstein et al. 1981) and a portable MWNIR spectrometer is available, this technique offers low cost and facilitates in situ determination.

Partial least squares (PLS) is a widely used multivariate technique in NIR modeling because it can analyze data that is strongly collinear (correlated), noisy, and contains numerous *X*-variables, and also simultaneously models several response variables, *Y*, i.e., profiles of performance (Wold, Sjöström, and Eriksson 2001). Uninformative and collinear variables are two important factors that affect the modeling efficiency. Uninformative variable elimination (UVE) can remove uninformative variables but fails in colinearity reduction while the successive projections algorithm (SPA) removes the collinear variables but cannot eliminate the uninformative influence (Centner and Massart 1996; Araújo et al. 2001). UVE–SPA, proposed by Ye, Wang, and Min (2008), is a method that combines UVE and SPA. In this technique, UVE is employed to select informative variables, and SPA follows to select variables that have minimum redundant information from the informative variables. After UVE-SPA, multiple linear regression (MLR) was employed to build the model. PLS employs latent variables instead of real variables while MLR models directly use the real variables that make them simpler and easier to interpret. However, MLR models usually suffer from the colinearity between variables (Næs and Mevik 2001). UVE–SPA solves this problem and has been reported in the food industry and agriculture (Wu et al. 2009; Wu et al. 2010).

Most of the research has used the whole spectrum of near-infrared spectroscopy (Balabin, Safieva, and Lomakina 2010; Dupuy et al. 2010; Puchert et al. 2010; Sinelli et al. 2010; Xiong et al. 2010), whereas only a few studied the application of MWNIR (Vincelette et al. 2008), but none have applied the technique to determine a pesticide active ingredient. Deltamethrin is a widely used insecticide used to protect crops from *Lepidoptera, Homoptera,* and *Hemiptera* (Mestres and Mestres 1992; Anadon et al. 1996). In this paper, MWNIR is used to develop a rapid and reliable method to determine the active ingredient in deltamethrin formulation instead of full-range NIR. PLS and MLR were used to establish quantitative models. UVE, SPA, and UVE–SPA were used to investigate the effective variables in modeling and their efficiencies were also compared.

MATERIALS AND METHODS

Samples

Three batches of the commercial formulation were used to prepare sixty samples with a certain amount of deltamethrin formulation, dimethylbenzene, and

technical deltamethrin for spectra collection. In order to avoid colinearity, the three reagents were added randomly. The gross mass of each sample was around 15 g and the concentration ranged from 0.11% to 5.39% (w/w). All samples were randomly divided into the calibration set (forty samples) and the prediction set (twenty samples) and the same divided sample sets were used in full-spectrum and UVE–SPA models. Cross validation was firstly performed on the calibration set to build the model and then prediction was made with the prediction set to determine the efficiency of the model. The exact content of deltamethrin in the commercial formulation was determined by HPLC.

Chemicals

The commercial deltamethrin formulation (25 g/L) was purchased from Bayer Crop Science, China. Technical deltamethrin (98.1%) was obtained from Jiangsu Huangma Agrochemicals, China. Dimethylbenzene (99.0%) and carbon tetrachloride (99.0%) were purchased from Beijing Chemical Works, China.

Instrumentation

NIR spectra were collected in transmission mode using vials of 2-mm path length with a Fourier transform spectrometer (Spectrum One NTS, PerkinElmer, USA). The spectral data were collected over the range 4000 cm^{-1}–12500 cm^{-1} (resolution 8 cm^{-1}, 64 scans) at room temperature. The cuvette was rinsed with carbon tetrachloride between the samples. Spectra acquisition and instrument control were performed by Spectrum (v 5.0, PerkinElmer, USA).

Datasets and Software

The spectra files were imported into Matlab (v7.11, MathWorks, USA) for data analysis. The scripts used in this study are based on Centner and Massart 1996, Ye et al. 2008, and Paiva et al. 2012. NIR data matrix were composed of 60 rows (samples) and 4251 columns (variables).

Theory

PLS considers a ($n \times p$) spectral matrix X including p predictor variables and ($n \times 1$) concentration vector y as response variables for n samples. PLS first decomposes the spectral matrix X and the concentration vector y as their own score matrix, loading matrix, and residual. The linear regression is made between the score matrix of X and $y \cdot b$ is the regression coefficient of the calibration set. In prediction, the score matrix of the unknown samples T_n is calculated from X_n based on the loading matrix of X and the prediction result y_n is obtained as Eq. 1.

$$y_n = T_n b Q \tag{1}$$

where Q is the loading matrix of y. In order to obtain a good estimates of b, the PLS model needs to be calibrated on samples that span the variation in y.

UVE is a method of variable selection based on stability analysis of the regression coefficient b. The details of UVE are described in Centner and Massart (1996), and its main principles are summarized as follows. First, PLS regression is performed on instrumental response data *Xcal* and the property values *y* of calibration set, and the optimal number of latent variables (LVs) is determined. Second, a noise matrix with the same size of the *Xcal* is generated and the elements are multiplied with a small constant to make their impact on the model negligible. The noise matrix is appended to the original one to form an extended matrix with twice as many variables as the original one. Third, PLS models are made on the extended matrix and *y* in manner of leave-one-out cross validation. This leads to a matrix of b values with as many rows as samples and one column for each variable, both original and random. The c value of each variable is calculated as the average of the b values of each column divided by the standard deviation of that column. The cut-off value is set as the maximum of absolute value c among the random variables. Every original variable with equal or lower absolute value of c is assumed to contain nothing but noise and is eliminated.

SPA employs simple projection operations in a vector space to obtain subsets of variables with small colinearity. A detailed descriptions of SPA is available in Araújo et al. (2001) and its main principles are summarized as follows. First, the maximum number of variables N to be selected was established. Then, starting from each variable, SPA yields K (the total number of variables) sets of selection of N variables. The optimal number of variables are determined on the performance in MLR calibration. After the SPA calculation, MLR was done with the retained variables to obtain the final regression results. In this study, the maximum number of retained variables in all SPA calculations was set as ten.

In UVE–SPA, UVE is first made with the raw spectral data and followed by SPA with the retained variables. In contrast to UVE, in UVE–SPA, as the following SPA can further eliminate the collinearity between the variables, MLR is able to perform with the retained variables instead of seeking latent variables with the help of PLS. Compared with SPA, as UVE-SPA first removes the uninformative variables before making the SPA calculation, the investigation of SPA calculation can be more efficient when less interference is introduced. After the selection, MLR is performed to obtain the final result.

RESULTS AND DISCUSSION

The collected NIR spectra of the samples are shown in Figure 1. The peaks from $4000\,\mathrm{cm}^{-1}$ to $5200\,\mathrm{cm}^{-1}$ are the first combination of stretching vibration of C-H and second overtone of C=O; peaks in the range of $5600\,\mathrm{cm}^{-1}$–$6200\,\mathrm{cm}^{-1}$ are attributed to the first overtone of C-H; peaks in the range of $7000\,\mathrm{cm}^{-1}$–$7400\,\mathrm{cm}^{-1}$ and $8300\,\mathrm{cm}^{-1}$–$8900\,\mathrm{cm}^{-1}$ are the second combination and second overtone of C-H, respectively (Kelly and Gallis 1990).

Full-Range NIR Models

Spectral pretreatments were usually employed in modeling to eliminate influences such as baseline drift. Sophisticated models were employed using spectral
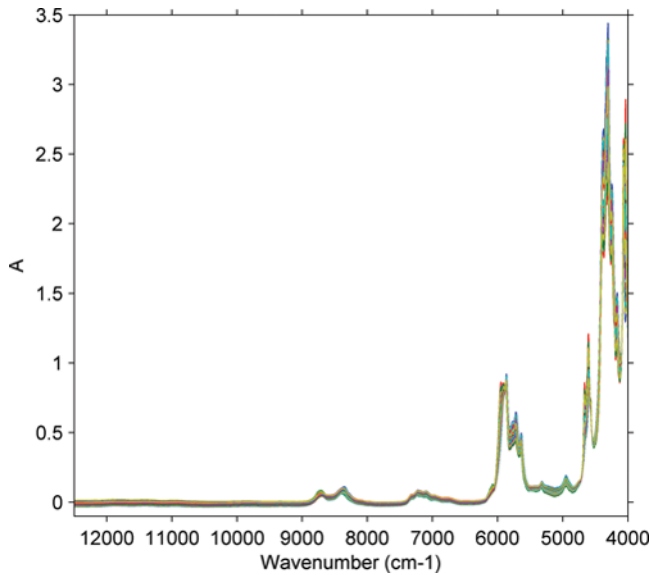
Figure 1. Near-infrared spectra of the deltamethrin formulation.

pretreatments including derivative, smoothing, and normalization, but they did not make improvements compared to the non-preprocessed model. Therefore, no pre-treatments were used. The results of the full-range NIR models are shown in Table 1, and models with variable selection techniques all outperformed the full-range NIR PLS model which indicates the redundancy of the full-range NIR and its impact on modeling. With UVE, 60% variables were eliminated and better results were obtained by the following PLS model. However, MLR was unable to perform after UVE as redundancy still existed. On the contrary, just several variables were retained after SPA, and MLR could run successfully as redundancy was almost eliminated. The SPA models also achieved better results than the UVE–PLS model.

Table 1. Results of full-range near-infrared models in determination of active ingredient in deltamethrin formulation

Method	Partial least squares factors	Variables	Root mean standard error of cross-validation (%)	Root mean standard error of prediction (%)
Partial least squares	6	4251	0.1250	0.1320
Uninformative variable elimination-partial least squares	6	1558	0.0942	0.0854
Successive projections algorithm-multiple linear regression	–	8	0.0854	0.0512
Uninformative variable elimination-successive projections algorithm-multiple linear regression	–	9	0.0432	0.0396
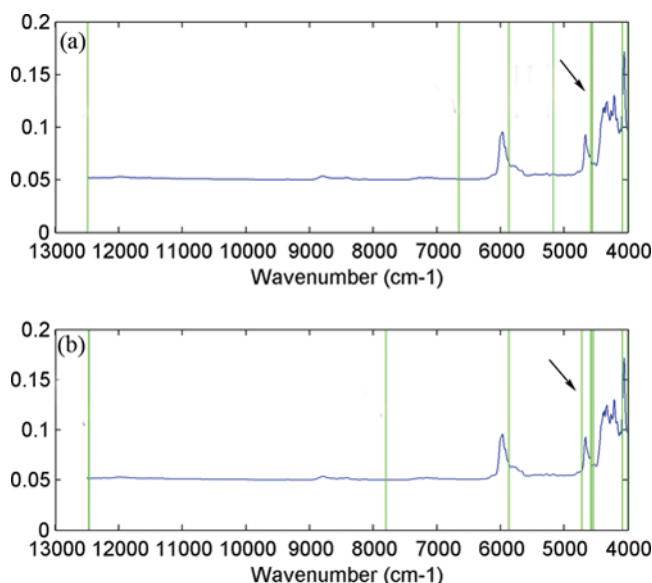
Figure 2. Selected variables by successive projections algorithm and uninformative variable elimination-successive projections algorithm near-infrared models under contrast with deltamethrin technical spectrum (a) successive projections algorithm and (b) uninformative variable elimination-successive projections algorithm.

As SPA just removed the collinearity between the variables, but failed to deal with the uninformative influence, the SPA model was not as efficient as the UVE–SPA model. It should be noted that although the UVE-SPA model gave the best results, more variables were employed in the UVE–SPA model than the SPA model. It can be interpreted as some useful information was recognized as collinearity and removed with just SPA selection. However, with UVE–SPA, as the uninformative ones were initially eliminated by UVE, those useful variables were reserved after SPA. Figure 2 illustrates the variables selected by SPA and UVE–SPA with the del-tamethrin technical spectrum (dissolved in carbon tetrachloride). Some ''baseline'' variables were reserved by SPA and UVE–SPA. Meanwhile, variables in the two main absorption wavebands of deltamethrin, i.e., $4000 \, \text{cm}^{-1}$–$4900 \, \text{cm}^{-1}$ (the first combination of stretching vibration of C-H and second overtone of C=O) and $5700 \, \text{cm}^{-1}$–$6200 \, \text{cm}^{-1}$ (the first overtone of C-H) were also employed. The differences occurred because UVE-SPA selected more informative variables than SPA (narrowed in the range from $4500 \, \text{cm}^{-1}$ to $4700 \, \text{cm}^{-1}$). The better results demonstrated that UVE–SPA was more efficient than SPA.

MWNIR Models

As significant differences were obtained, no spectral pretreatments were used. As shown in Table 2, the MWNIR PLS model far surpassed the full-range NIR PLS and its results were almost comparable to the NIR–UVE–SPA model, which further

Table 2. Results of medium near-infrared models in determination of active ingredient in deltamethrin formulation

Method	Partial least squares factors	Variables	Root mean standard error of cross-validation (%)	Root mean standard error of prediction (%)
Partial least squares	6	1751	0.0572	0.0558
Uninformative variable elimination-partial least squares	4	414	0.0384	0.0349
Successive projections algorithm-multiple linear regression	–	5	0.0404	0.0274
Uninformative variable elimination-successive projections algorithm-multiple linear regression	–	5	0.0439	0.0284

indicated the redundancy of the full-range NIR and efficiency of MWNIR in analysis of deltamethrin. After UVE, 75% variables were eliminated and better regression results were obtained. In contrast to the NIR–UVE–PLS model, fewer PLS factors were used in the MWNIR-UVE-PLS model for a leaner variable subset was introduced to the following PLS regression. In contrast to the full-range NIR models, promotion of SPA was not so visible as the MWNIR PLS and UVE–PLS models were approaching optimal. Nevertheless, the SPA and UVE–SPA models of MWNIR achieved better results than the full-range NIR approaches even with fewer variables in MLR calculation. It should be noted that the SPA model had a slim advantage over the UVE–SPA model. Since the MWNIR waveband contained the important information of the interest and uninformative variables were fewer than the full-range NIR, SPA was sufficient to deal with this problem. The variables selected by SPA and UVE–SPA, in contrast with the deltamethrin technical spectrum in carbon tetrachloride, are presented in Figure 3. The two methods selected the variables in the range of $5700 \, cm^{-1}$–$6000 \, cm^{-1}$ (first overtone of C-H), whereas differences between the two methods were the selection of the narrowed variable. The variable selected by the SPA model lay in the range of the second combination of C-H while the one retained by UVE–SPA was attributed to the second overtone of C-H. On the other hand, variables selected by MWNIR SPA and UVE–SPA models were different from those in the full-range NIR models. Variables selected in full-range NIR model were scattered across the spectrum while just one variable was in the range of MWNIR. As UVE and the following SPA just focused on eliminating the variables with low signal-to-noise ratio and dealt with the collinearity between the variables retained by UVE, the variables in the other parts of spectrum had equal probability to be selected when full-range NIR was used to model. In addition, MWNIR MLR models surpassed those of full-range NIR both in the number of variables selected and the regression results. Fewer variables were imported for variable investigation, fewer interferences were introduced into the MWNIR model, and the performance was enhanced.
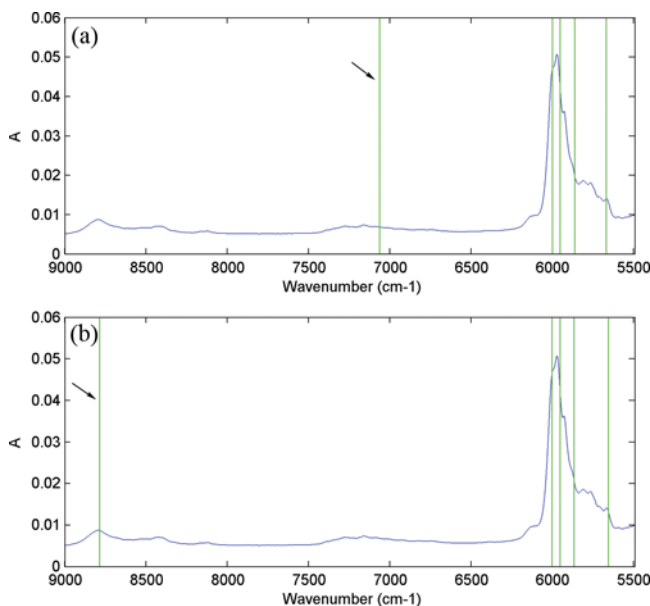
Figure 3. Selected variables by successive projections algorithm and uninformative variable elimination-successive projections algorithm medium near-infrared models in contrast with deltamethrin technical spectrum (a) successive projections algorithm and (b) uninformative variable elimination-successive projections algorithm.

These results demonstrate that MWNIR was capable of establishing an efficient model to quantify the active ingredient because it included the important information of the functional groups in NIR and its signal intensity was well suited for chemometric analysis. This analysis may be performed using in situ determination by portable MWNIR instrumentation. UVE–SPA was demonstrated to be efficient in dealing with the redundancy that not only further simplified but also improved the model. In addition, UVE–SPA was more efficient when redundancy was severe (e.g., full-range NIR). When collinearity was the primary contradiction (e.g., MWNIR), the advantages of UVE–SPA were less obvious and SPA was sufficient for the analysis.

CONCLUSIONS

MWNIR was successfully applied to the determination of a pesticide active ingredient by conventional PLS regression. Results indicated that MWNIR was well suited for this analysis and could substitute for full-range NIR modeling. UVE–SPA was shown to be a powerful variable selection approach that not only eliminated uninformative but also collinear variables, which made the model much simpler and more efficient. The feasibility of MWNIR in pesticide active ingredient determination reduces the cost and instrument requirements, allows in situ analysis, expands applications of NIR, and provides a valuable reference for pesticide quality control.

FUNDING

REFERENCES

Anadon, A., M. R. Martinez-Larranaga, M. L. Fernandez-Cruz, M. J. Diaz, M. C. Fernandez, and M. A. Martinez. 1996. Toxicokinetics of deltamethrin and its 4′-HO-metabolite in the rat. *Toxicol. Appl. Pharmacol.* 141: 8–16.

Araújo, M. C. U., T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, and V. Visani. 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* 57: 65–73.

Armenta, S., S. Garrigues, and M. Guardia. 2007. Partial least squares-near infrared determination of pesticides in commercial formulations. *Vib. Spectrosc.* 44: 273–278.

Azzouz, T., and R. Tauler. 2008. Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples. *Talanta.* 74: 1201–1210.

Balabin, R. M., and R. Z. Safieva. 2008. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. *Fuel.* 87: 1096–1101.

Balabin, R. M., R. Z. Safieva, and E. I. Lomakina. 2010. Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. *Anal. Chim. Acta.* 671: 27–35.

Blanco, M., and A. Peguero. 2010. Analysis of pharmaceuticals by NIR spectroscopy without a reference method. *Trends Anal. Chem.* 29: 1127–1136.

Centner, V., and D. L. Massart. 1996. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* 68: 3851–3858.

Duan, J., Y. Huang, Z. Li, B. Zheng, Q. Li, Y. Xiong, L. Wu, and S. Min. 2012. Determination of 27 chemical constituents in Chinese southwest tobacco by FT-NIR spectroscopy. *Ind. Crops Prod.* 40: 21–26.

Dupuy, N., O. Galtier, D. Ollivier, P. Vanloot, and J. Artaud. Comparison between NIR, MIR, concatenated NIR and MIR analysis and hierarchical PLS model. Application to virgin olive oil analysis. *Anal. Chim. Acta.* 666: 23–31.

Falla, F. S., C. Larini, G. A. C. Le Roux, F. H. Quina, L. F. L. Moro, and C. A. O. Nascimento. 2006. Characterization of crude petroleum by NIR. *J. Petrol. Sci. Eng.* 51: 127–137.

Howland, C. H., and S. W. Hoag. 2013. Analysis of curing of a sustained release coating formulation by application of NIR spectroscopy to monitor changes physical-mechanical properties. *Int. J. Pharm.* 452: 82–91.

Jamshidi, B., S. Minaei, E. Mohajerani, and H. Ghassemian. 2012. Reflectance Vis/NIR spectroscopy for nondestructive taste characterization of Valencia oranges. *Comput. Electron. Agr.* 85: 64–69.

Karasali, H., H. Anagnostopoulos, H. Ekonomopoulou, and A. Hourdakis. 2004. Quality control data of fenthion and trifluralin determination in pesticide formulations. *Int. J. Environ. Anal. Chem.* 84: 55–63.

Kelly, J. J., and J. B. Gallis. 1990. Nondestructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. *Anal. Chem.* 62: 1444–1451.

Lin, Y. W., and S. S. Q. Hee. 1998. Simultaneous gas-chromatographic mass-spectrometric quantitation of the alkylbenzene inert components, pesticide manufacturing by-products and active ingredient in 2 malathion formulations. *J. Chromatogr. A.* 814: 181–186.

Mestres, R., and G. Mestres. 1992. Deltamethrin: Uses and environmental safety. *Rev. Environ. Contam. Toxicol.* 124: 1–18.

Næs, T., and B. Mevik. 2001. Understanding the collinearity problem in regression and discriminant analysis. *J. Chemometr.* 15: 413–426.

Paiva, H. M., S. F. C. Soares, R. K. H. Galvão, and M. C. U. Araújo. 2012. A graphical user interface for variable selection employing the successive projections algorithm. *Chemom. Intell. Lab. Syst.* 118: 260–266.

Phillips, T., and A. Burns. 2012. Determination of phenols and phenates in disinfectant formulations by liquid chromatography with UV detection: Collaborative study. *J. AOAC Int.* 95: 411–418.

Porfire, A., L. Rus, A. L. Vonica, and I. Tomuta. 2012. High-throughput NIR-chemometric methods for determination of drug content and pharmaceutical properties of indapamide tablets. *J. Pharmaceut. Biomed. Anal.* 70: 301–309.

Puchert, T., D. Lochmann, J. C. Menezes, and G. Reich. 2010. Near-infrared chemical imaging (NIR-CI) for counterfeit drug identification—a four-stage concept with a novel approach of data processing (Linear Image Signature). *J. Phamacuet. Biomed. Anal.* 51: 138–145.

Silverstein, M., G. Clayton Basseler, and C. Morill. 1981. *Spectrometric Identification of Organic Compounds.* New York, NY: Wiley.

Sinelli, N., L. Cerretani, V. D. Egidio, A. Bendini, and E. Casiraghi. Application of near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity. *Food Sci. Technol. Int.* 43: 369–375.

Vincelette, R. L., A. J. Welch, R. J. Thomas, B. A. Rockwell, and D. J. Lund. 2008. Thermal lensing in ocular media exposed to continuous-wave near-infrared radiation: the 1150–1350-nm region. *J. Biomed. Opt.* 13: Art. No. 054005.

Wold, S., M. Sjöström, and L. Eriksson. 2001. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58: 109–130.

Wu, D., Y. He, P. Nie, F. Cao, and Y. Bao. 2010. Hybrid variable selection in visible and near-infrared spectral analysis for non-invasive quality determination of grape juice. *Anal. Chim. Acta.* 659: 229–237.

Wu, D., Y. He, J. H. Shi, and S. J. Feng. Exploring near and mid-infrared spectroscopy to predict trace iron and zinc contents in powdered milk. *J. Agric. Food Chem.* 57: 1697–1704.

Xiong, Y. M., G. Tang, J. Duan, C. Z. Li, D. Wang, and S. G. Min. 2010. Quantitative analysis of the content of deltamethrin in agrochemicals by near-infrared attenuated total reflectance infrared and Raman spectroscopy. *Spectrosc. Spectr. Anal.* 30: 2936–2940.

Ye, S. F., D. Wang, and S. G. Min. 2008. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemom. Intell. Lab. Syst.* 91: 194–199.

Zou, X. B., J. W. Zhao, M. J. W. Povey, M. Holmes, and M. Hanpin. 2010. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta.* 667: 14–32.