Talanta 198 (2019) 215–223

Contents lists available at ScienceDirect

Talanta

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

Rapid assessment of total MCPD esters in palm-based cooking oil using ATR-FTIR application and chemometric analysis

Kok Ming Goh[a], M. Maulidiani[b], R. Rudiyanto[c], Yu Hua Wong[a], May Yen Ang[d], Wooi Meng Yew[d], Faridah Abas[e], Oi Ming Lai[f], Yonghua Wang[g], Chin Ping Tan[a,*]

[a] Department of Food Technology, Faculty of Food Science and Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[b] School of Fundamental Science, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
[c] School of Food Science and Technology, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
[d] Shimadzu Malaysia Sdn Bhd, No.6 Lorong Teknologi 3/4 A, Nouvelle Industrial Park 2, Taman Sains Selangor 1, Kota Damansara, 47810 Petaling Jaya, Selangor, Malaysia
[e] Department of Food Science, Faculty of Food Science and Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[f] Department of Bioprocess Technology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[g] Guangdong Research Center of Lipid Science Applied Engineering Technology, School of Food Science and Engineering, South China University of Technology, Guangzhou 510640, China

ARTICLE INFO

Keywords:
MCPD prediction
GC-MS
FTIR
Chemometric analysis
Consensus model

ABSTRACT

The technique of Fourier transform infrared spectroscopy is widely used to generate spectral data for use in the detection of food contaminants. Monochloropropanediol (MCPD) is a refining process-induced contaminant that is found in palm-based fats and oils. In this study, a chemometric approach was used to evaluate the relationship between the FTIR spectra and the total MCPD content of a palm-based cooking oil. A total of 156 samples were used to develop partial least squares regression (PLSR), artificial neural network (nnet), average artificial neural network (avNNET), random forest (RF) and cubist models. In addition, a consensus approach was used to generate fusion result consisted from all the model mentioned above. All the models were evaluated based on validation performed using training and testing datasets. In addition, the box plot of coefficient of determination (R^2), root mean square error (RMSE), slopes and intercepts by 100 times randomization was also compared. Evaluation of performance based on the testing R^2 and RMSE suggested that the cubist model predicted total MCPD content with the highest accuracy, followed by the RF, avNNET, nnet and PLSR models. The overfitting tendency was assessed based on differences in R^2 and RMSE in the training and testing calibrations. The observations showed that the cubist and avNNET models possessed a certain degree of overfitting. However, the accuracy of these models in predicting the total MCPD content was high. Results of the consensus model showed that it slightly improved the accuracy of prediction as well as significantly reduced its uncertainty. The important variables derived from the cubist and RF models suggested that the wavenumbers corresponding to the MCPDs originated from the $-CH=CH_2$ or $CH=CH$ (990–900 cm^{-1}) and C-Cl stretch (800–700 cm^{-1}) regions of the FTIR spectrum data. In short, chemometrics in combination with FTIR analysis especially for the consensus model represent a potential and flexible technique for estimating the total MCPD content of refined vegetable oils.

1. Introduction

Monochloropropanediol (MCPD) is a contaminant that is detected extensively in refined vegetable oils. MCPD is present in refined vegetable oils in the ester form. According to a report published in the EFSA Journal in 2016, palm-based fats and oils have the highest content of MCPD of all available refined vegetable oils. The average MCPD content of palm oils and fats was reported to be 1.5 ppm for 2-MCPD and 3.9 ppm for 3-MCPD [10].

In terms of toxicology, 3-MCPD and its dipalmitate fatty acid ester are readily absorbed by the gastrointestinal tract after de-esterification and can cause effects in the liver or kidney. Although there are insufficient supporting data to suggest toxicological effects of 2-MCPD, an isomer of 3-MCPD, the potential hazards caused by 2-MCPD are con-

[*] Corresponding author.
E-mail address: tancp@upm.edu.my (C.P. Tan).

https://doi.org/10.1016/j.talanta.2019.01.111
Received 26 October 2018; Received in revised form 29 January 2019; Accepted 31 January 2019
Available online 31 January 2019
0039-9140/ © 2019 Elsevier B.V. All rights reserved.

sidered to be equal to those of 3-MCPD [5].

The common methods for determining the MCPD content of foods require the application of gas chromatography with mass spectrometry detection (GC-MS). These methods have been established as official by the AOCS, and there are three approved methods. For each of these methods, intensive sample preparation is required. Given the potential hazards imposed by MCPD, there is a need to develop a rapid method for screening and quantitating the MCPD content of edible oils as an alternative to the GC-MS approach.

Chemometric analysis is considered a technique that is able to perform pattern recognition and computational learning theory. It is closely related to predictive modeling and mathematical optimization. It is a study that can evaluate input patterns and make predictions or decisions based on algorithms. In short, chemometric analysis makes it possible to use computers to learn and predict difficult tasks (normally from big data) in ways that are simply not possible to perform manually. This technique has been widely applied in science to make predictions involving diseases such as Parkinson's [11], leukemia [18] and heart disease [14]. Therefore, we are interested in exploring the potential of the chemometric analysis to determine the relationship between the FTIR spectra of palm-based cooking oils and their respective total MCPD content.

The available literature shows that FTIR spectroscopy is a commonly applied analytical tool in studies of edible fats and oils. The relevant studies can be extended to determine the adulteration of oils [20,32] and to predict characteristics such as antioxidant activity [6] and free fatty acid content [34]. In addition, the screening of contaminants such as acrylamide can be successfully accomplished using FTIR in combination with chemometric analysis [2,3].

Partial least squares regression (PLSR) modeling is a widely used linear modeling method that can explain the relationship between two datasets. As mentioned earlier, PLSR, together with the FTIR technique, was extensively used by scientists in the prediction of the physicochemical properties of food compounds. However, as an alternative to PLSR to predict the relationship if nonlinear behavior is observed, modeling such as artificial neural network (nnet) modeling can be applied. The use of nnet modeling in the prediction of antioxidant activities in plant extracts was also reported [25]. Neural network modeling can be further improved to avNNET (average neural network) modeling, which utilizes several segregated neural networks [15].

Most biochemical properties can be explained by using decision tree modeling, namely, random forest (RF) and cubist models. Decision tree modeling uses a specific technique to partition a dataset into subsets in a binary splitting manner. The partitioning process is continued until no available option remains. Then, the RF model establishes multiple decision trees, similar to a forest. The observation and variables randomly become the subsets of the trees, representing the branching operation. The cubist model utilizes an advanced regression tree model that uses the rules of "if-then" conditions to make predictions. The RF model has been used to predict the antioxidant properties of plant extracts [19]; cubist modeling is still not widely used in biochemical-related fields but it has been proven to be a powerful and cost-saving machine learning tool for digital mapping [33].

Ultimately, a consensus strategy can be applied to improve the weakness (for example overfit or underfit behavior) of using conventional multivariate calibration techniques based on single model. The member models are trained individually, and then their predictions ability can be evaluated. Then, these prediction powers are combined by simple averaging or weightage averaging. In consensus or fusion of model, it is expected to increase the prediction accuracy and robustness [22].

The objective of this study is to develop and compare the performance of the selected single model (PLSR, nnet, avNNet, RF and cubist modeling) and a consensus regression model method in predicting the total MCPD content of palm-based cooking oils based on their FTIR spectra as a rapid assessment method.

2. Materials and methods

2.1. Chemicals and materials

PP-3-MCPD (1,2-dipalmitoyl-3-chloropropanediol, purity > 95%), PP-2-MCPD (1,3-dipalmitoyl-2-chloropropanediol, purity > 95%), and pentadeuterated forms of PP-3MCPD-d5 (purity > 95%) were purchased from Toronto Research Chemical Inc., North York, ON, Canada. All the palm-based cooking oils were purchased at local grocery stores by random selection.

2.2. Preparation of standards

Stock solutions of PP-3-MCPD (1,2-dipalmitoyl-3-chloropropanediol, purity > 95%) and PP-2-MCPD (1,3-dipalmitoyl-2-chloropropanediol, purity > 95%) standards were prepared at a concentration of 1 mg/mL. An internal standard of PP-3-MCPD-d5 was also prepared at a final concentration of 40 μg/mL. Standard curves of 2- and 3-MCPD at concentrations ranging from 0.10 to 7.2 mg/kg were plotted based on area ratios and used to quantitate the content of the compounds in the selected palm-based cooking oils.

2.3. Measurement of total MCPD by GC-MS

All the oil samples were derivatized based on the AOCS Official Method Cd 29a-13. Briefly, 100 mg of sample was measured and spiked with a known amount (50 μL) of internal standard of PP-3-MCPD-d5 solution. Bromination of glycidyl esters (GE) was then performed using an acidic bromide solution to differentiate GE from MCPD in a 15-min incubation at 50 °C. Bromination was terminated by the addition of a solution of 0.6% sodium hydrogen carbonate. The oil was extracted in n-heptane solution and collected by evaporating the excessive solvent. De-esterification of the MCPD ester was then conducted under acidic conditions with a 16-hr incubation at 40 °C. The de-esterification reaction was stopped by adding 3 mL of saturated sodium hydrogen carbonate solution, and the unwanted fatty acid portion was discarded by n-heptane solution. Then, free MCPD was collected and derivatized with a saturated phenylboronic acid (PBA) in a sonicator bath for 5 min. Finally, the derivatized MCPD was extracted with n-heptane, evaporated to complete dryness under a nitrogen stream and reconstituted in 400 μL of n-heptane. The supernatant was collected as the prepared analyte and stored at 4 °C prior to GC-MS analysis.

Injection of 1 μL of the analyte into a GS-MS system (Shimadzu GCMS-TQ 8040) was performed using an autosampler (AOC-20i); selected ion mode (SIM) detection was used as described in AOCS official method Cd 29a-13.

2.4. FTIR measurements

The FTIR spectra of the palm-based cooking oils were acquired using a MIRacle attenuated total reflection ATR accessory (Pike technologies, Germany) in a Fourier transform infrared spectrometer (Shimadzu, IRTracer-100). Spectra were collected in the wavelength range of 4000–600 cm^{-1} with 40 interferograms at a resolution of 4 cm^{-1}. Each sample scan was started by scanning a blank background (air background) followed by dropping a 40μL oil sample onto the surface of a diamond/ZnSe plate. After each sample scan, the ATR plate was cleaned using a dust-free tissue and hexane solution.

2.5. Datasets

The total MCPD content of the selected palm-based cooking oil was determined by the AOCS method as described in Section 2.3; a calibration curve was used as described in Section 2.2. The MCPD content was determined separately in terms of the isomers of 2- and 3-MCPD esters. The combined concentration of the two isomers was calculated by adding the concentrations of the two compounds. The total MCPD content was used as the Y-variable for modeling in this study.

A total of 156 samples (n = 156), each consisting of 1546 observations (in nm wavenumber), was obtained using ATR-FTIR measurement as described in Section 2.4. The spectral data were further processed and were used in the modeling as the X-variables.

The purchase of samples, the GC-MS measurements of MCPD content and the collection of FTIR spectral data on the selected palm-based cooking oils were conducted between September 2016 and May 2017.

2.6. Data preprocessing

All the collected spectra were converted from absorbance wavenumber in the units of nm (represented by R) to the form of log (1/R). The converted spectra were filtered by the smoothing method Savitzky-Golay (SG) algorithm with a window size of 21 and a polynomial of order 2. The data processing was followed by Standard Normal Variate (SNV) transformation. By applying the described data processing, instrumental noise within the spectra can be removed using the SG algorithm with polynomial regression, whereas SNV is a method of normalizing the absorbance data (normally an FTIR spectrum) that scales the spectrum to zero mean and corrects the signal. The combination of the SG algorithm and SNV normalization was based on a previous study [26,30]. The prospectr R package [35] was used for Savitzky-Golay (SG) filtering. Fig. 1 shows an example of the spectral filtering.

Each FTIR spectrum consisted of 1546 variables from wavelength number 599.86–4000.36 nm with an interval of 2 nm. The total number of datasets used in the study was 156 samples (n = 156).

2.7. Variable selection

The use of a large number of variables (= 1546) can lead to non-optimal solutions for predictive regression models, since some variables might be correlated. Thus, only relevant variables were selected for prediction of MCPD in palm-based cooking oil. In this study, selection of variables was done by the R software package Boruta: Wrapper Algorithm for all Relevant Feature Selection [17] as a variable selection method. This method is based on a wrapper algorithm around Random Forest [39]; it searches relevant variables by comparing the importance of the original variable with that of randomly permuted copies of the variables [17]. Advantages of Boruta are, it works well with regression and classification problem, multi-variable relationships are taken into account, and it is an all-relevant variable selection method, which considers the features relevant to outcome variable [17]. Of the 1546 variables, 51 variables were relevant, while 92 variables are tentative and the remaining variables is rejected. Only the relevant variables were selected for input into all six models which will be described in Section 2.8.

2.8. Model development

Six regression models were generated for prediction of the total MCPD content of palm-based cooking oil; these were the PLS regression (PLSR) [38], artificial neural network (nnet) [13], average artificial neural network (avNNET) [15], random forest (RF) [24], cubist models [16] and a consensus regression model comprised of all the mentioned models

2.8.1. Partial Least Squares Regression (PLSR)

PLSR has been widely used in the field of near-infrared (NIR) spectroscopy with considerable success. It is a multivariate regression method that decomposes X-variables into orthogonal scores, T, and loadings, P, and regresses Y (dependent variables) not on X itself but on the first column of the T scores [38]. Here, the PLSR model was applied using the PLS package in R software; the cross-validation approach was used to determine the number of optimum components and it was found equal to 9 components.

2.8.2. nnet and avNNET

The artificial neural network (nnet) model is a commonly used chemometric analysis that performs classification, pattern recognition and prediction modeling. The model assumes that there is a nonlinear relationship between each layer, and the layers are connected by a weightage [1]. In avNNET, some neural network models were fitted using different random number seeds. All the resulting models were used for prediction, and the average results were then calculated. The nnet [37] and neural networks using model averaging (avNNET) R packages were used in this study. In this study, both nnet and avNNET

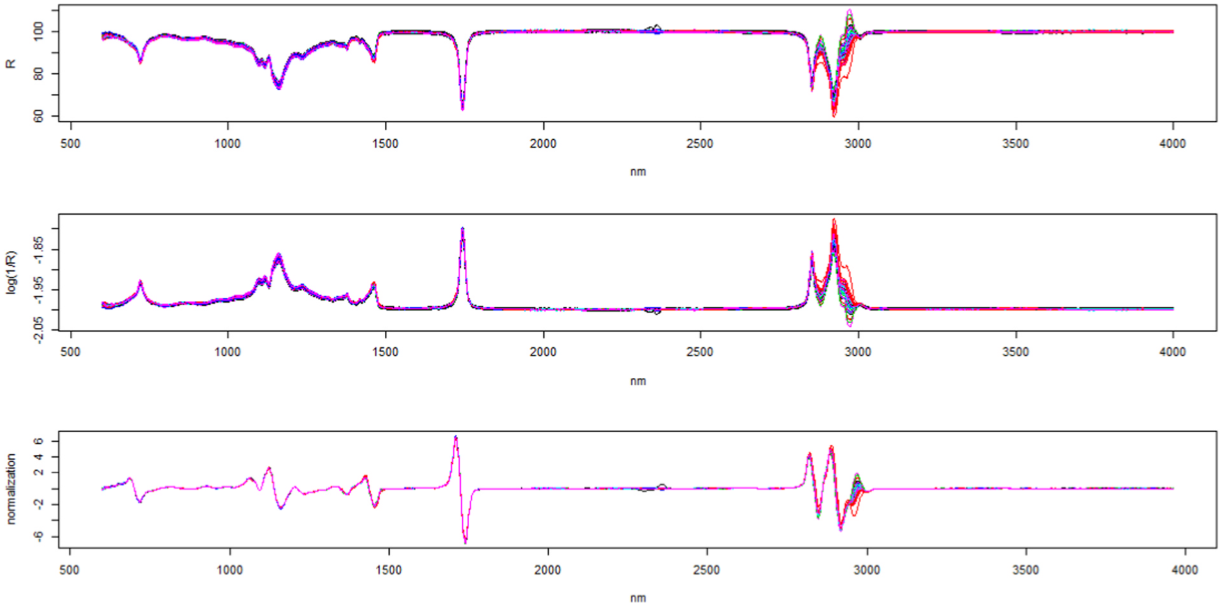

Fig. 1. Top: The original FTIR spectrum without preprocessing. Middle: The spectrum was converted from its absorbance wavelength (represented by R) to units of nm using log (1/R) and filtered using the Savitzky-Golay (SG) smoothing method algorithm with a window size of 21 and a polynomial of order 2. Bottom: The spectrum was further normalized by Standard Normal Variate (SNV) transformation.

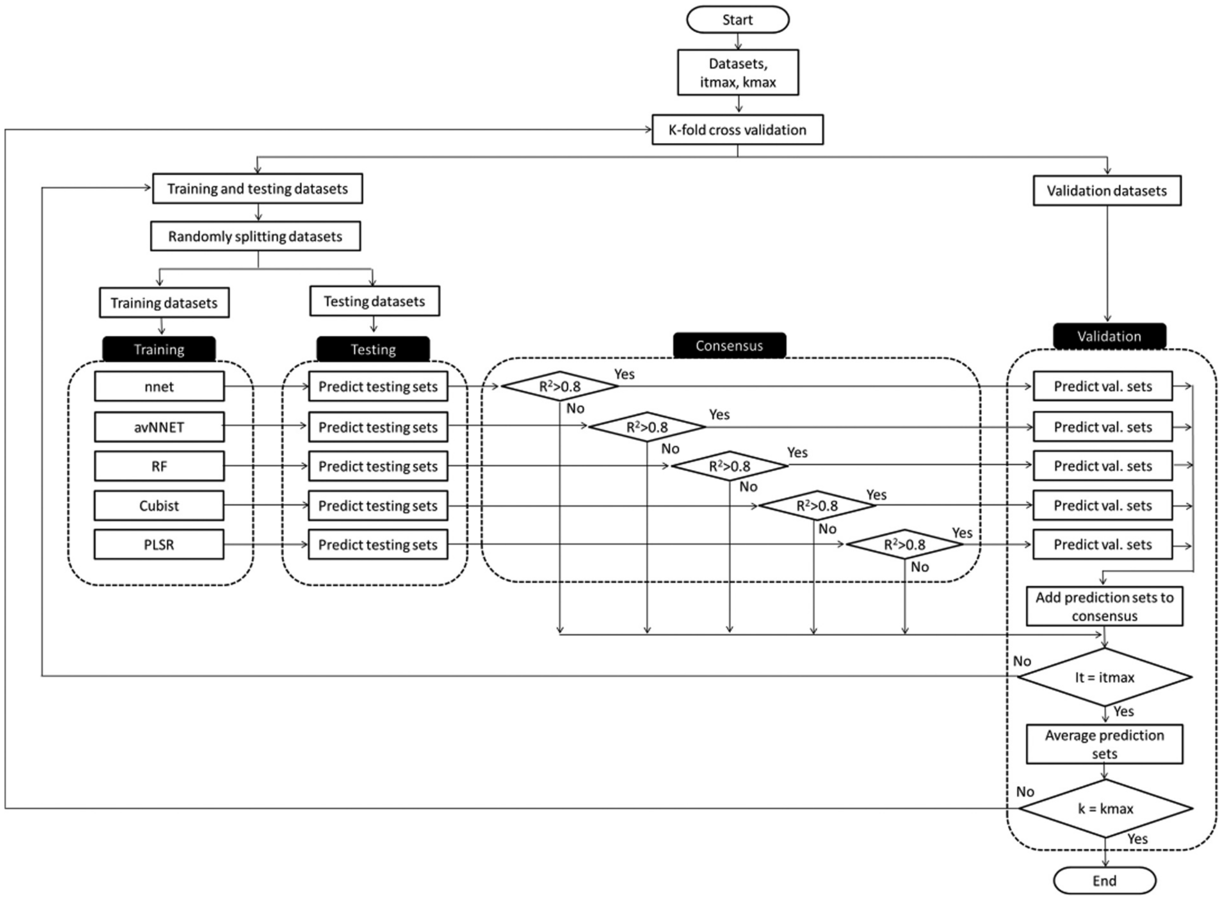

Fig. 2. Flow chart of consensus regression model from fusion of Cubist, random forest, nnet, avNNET and PLSR models. R^2 value from testing is used for evaluating model acceptance as consensus.

consisted of three layers: input, hidden and output layers with 51 nodes for inputs, 5 hidden nodes, and a single output for MCPD in the sample. Decay parameter for both models was also same, 0.01. In training process, both nnet and avNNET were iterated until converged or to reach maximum iterations (= 1000). Relatively small values for numbers of hidden node and maximum number of iterations were addressed to avoid overfitting in both models.

2.8.3. Random forest

Random forest (RF) is a typical tree-based ensemble method that was first proposed by Breiman [4]. RF ensemble a number of regression (known as trees). RF considered to have multiple decision tree. A bootstrapping technique is used to subset the data randomly from the observations and variables. In the case of regression, prediction is based on the average output from each tree. Oshiro et al. [27] recommended that between 64 and 128 trees be used as a compromise between accuracy and processing time; in this study, the number of trees was fixed at 100, as recommended. The random forest model was performed using the Random Forest R package [23,24,27].

2.8.4. Cubist

Cubist is a further development of Quinlan's M5 model tree [28]. Cubist results in a series of “if–then” rules in terms of the multivariate linear model. Whenever a set of variables matches a rule's conditions, the prediction value of the variables is calculated using the corresponding model. The cubist model was conducted using the Cubist R package [16]. In the Cubist, two parameters: committees and neighbors were set equal to 5 and 2, respectively. One of the conditional rules established by the cubist model in this study is shown below.

Model 1:

Rule 1/1: [94 cases, mean 4.6003, range 1.277–9.455, est. err 0.9483]

if

$$X956.692736\ < =\ 0.2028248$$

$$X991.411424\ >\ 0.09607272$$

then

$$\begin{aligned}\text{outcome}\ =\ &-65.1398\ +\ 369\ X993.34024\ -\ 283\ X991.411424\\ &-\ 199\ X1261.445664\ +\\ 233\ X964.408\ -\ &102\ X979.838528\ -\ 77\ X981.767344\\ &-\ 83\ X956.692736\ +\ 19.5\ X1163.0760\\ 48\ +\ 87\ X1265.303296\ -\ &26\ X1689.642816\ +\ 48\ X948.977472\\ &-\ 69\ X806.245088\end{aligned}$$

where X956 and X991 represent the wavenumbers from X-variables with conditional rules followed by the outcome quantitation model.

2.8.5. Model validation

The validations of those five models above were repeated 100 times; the dataset (n = 156) was split randomly at a ratio of 7:3 for training and testing dataset. The performance of the models during training and testing was evaluated using the coefficient of determination (R^2), the root mean square error (RMSE), slopes and intercepts of linear regression between observed and predicted MCPD. These repetitions resulted in 100 realisation models for each model in both training and testing. From these realisations, the difference between mean R^2 (or RMSE) from training and testing would be used to evaluate degree of overfit of the model.

2.8.6. Consensus regression model

The consensus modeling combines the results of multiple individual models. These individual models are often known as the member models. Consensus modeling is based on the idea whereby the multiple models can identify and encode more aspects of the relationship between the independent and dependent variables as compared to use a single model. It is believed consensus model can solve overfit and underfit problem due to small training set and enhance the prediction stability (reduce uncertainty) [22]. Consensus model presents the prediction results from multiple member models in average and therefore the relationship between dependent and independent variables will be estimated more effectively [21]. A simplify operational flow chart of consensus regression model is showed in Fig. 2. The models consisted from all the member model evaluated earlier in the study (PLSR, nnet, avNNet, RF and cubist modeling). The operation based on K-fold cross validation (K = 5), with 80% data as training and 20% data as prediction. Among the 80% training datasets, it was further divided into training and testing subsets at ratio 7: 3 for each model training purpose. The acceptance criteria of R^2 value from testing was 0.8 and above. If the particular dataset generates a $R^2 > 0.8$, it goes to the prediction testing. These training-testing steps were repeated up to 100 iterations. As a result, total iteration was 500.

3. Results and discussion

3.1. Analysis of the MCPD spectra of refined vegetable oils

The composition of refined vegetable oil has been well established by FTIR analysis. In this study, the FTIR spectrum obtained from palm-based cooking oils spiked with MCPD did not differ from the spectra obtained in previous studies. Notably, the overall spectra showed characteristic bands at 1743–1744 cm^{-1}, 2852 and 2922 cm^{-1} resulting from C⁼O stretching of esters (carboxylic acid or triglycerides), C-H asymmetric stretching of CH_2 and C-H symmetric stretching of CH_2 [29]. In addition, bands in the 1300–1150 cm^{-1} region that corresponded to CH_2 wagging, bands of 1470–1450 cm^{-1} resulting from CH_3 bending, and bands of 3100–3000 cm^{-1} due to the CH stretching of cis double bonds were also seen [9].

A typical FTIR spectrum of a palm -based cooking oil with Boruta selection is shown in Fig. 3. The figure shows the confirmed (but not limited to) variables were fall between some of the described functional bands of an oil FTIR spectrum, Evidently, Boruta selection was able to target on important bands, for example, CH_2 wagging (1300–1150 cm^{-1}). Several wavenumbers between bands 990–900 cm^{-1} (CH = CH), were selected. Besides, wavenumbers between 800 and 700 cm^{-1} (702, 704, 705, 802, 804, and 806) corresponding to C-Cl bond [8], whereby a MCPD functional band located. This finding provided strong evident that a hypothetical region of C-Cl bond is important to MCPD prediction.

When MCPD exists in the free form, its chemical structure is rather simple because one of the OH groups of the glycerol backbone is replaced by a chloride ion at either the 2- or 3- position. However, in nature, MCPD is commonly found in an ester form, the structure of which is more complicated [40]. The ester group or the fatty acid attached to the glycerol can vary according to the nature of the oil; for instance, esterified palmitic acid is commonly found in palm oil [41]. The complexity of the fatty acids contributes to variation in the structure of the MCPD ester. Furthermore, MCPD esters can be present in mono- or di-ester forms. The detection of MCPD or MCPD ester from the FTIR spectrum of the oil alone appeared to be extremely difficult due to the presence of long-chain fatty acids that could be attached to the MCPD ester.

3.2. Model comparison

The results of performance were evaluated in terms of R^2, RMSE, slopes and intercepts of linear regression between observed and predicted MCPD from all five models Supplementary Tables 1–4 show the numerical data in detail. Also, the consensus model was evaluated by R^2, RMSE, slope and intercept and compared with the member models. Comparison among the models are summarized as box plots in Fig. 4(a)–(d).

3.2.1. Accuracy of testing dataset

Based on the testing R^2 mean, the accuracy of the testing model can be evaluated. Notably, the cubist model achieved the highest accuracy in predicting the total MCPD content, with an R^2 = 0.78, followed by the RF model (R^2 = 0.76). The two neural network models presented similar performances, with R^2 differences of approximately 0.1. The nnet model was considered slightly weaker than the avNNET model in

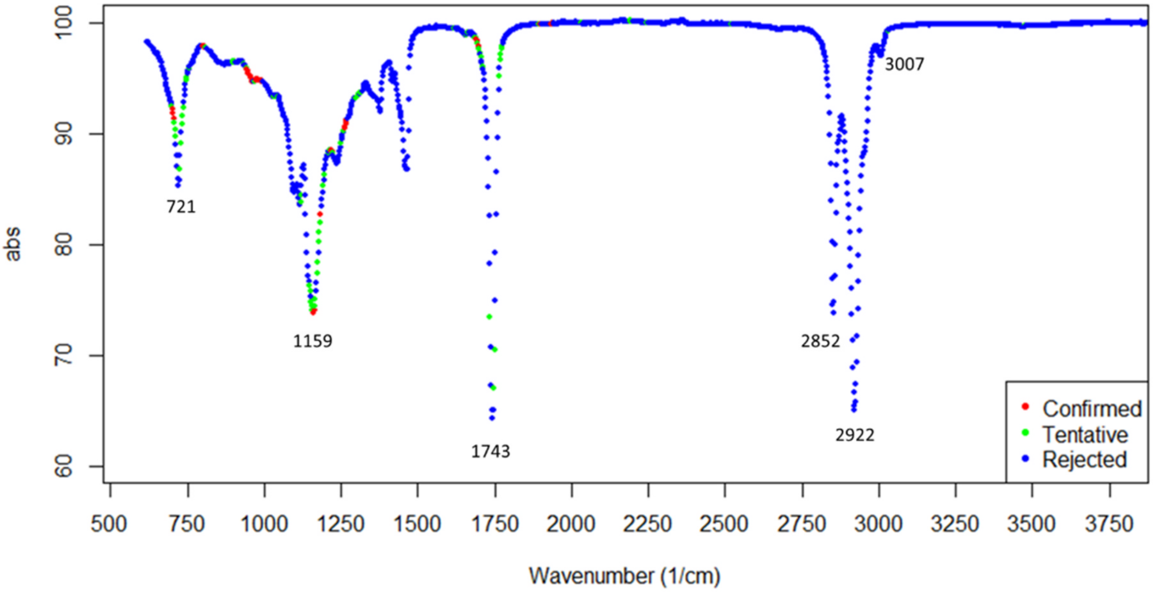

Fig. 3. Representative FTIR spectra obtained by spiking palm-based cooking oil with PP-3-MCPD. The numbers indicate the wavenumbers of the peaks corresponding to the functional groups. The red dots are confirmed and the ones in green and blue are tentative and rejected, respectively based on variable selection results from the Boruta algorithm.

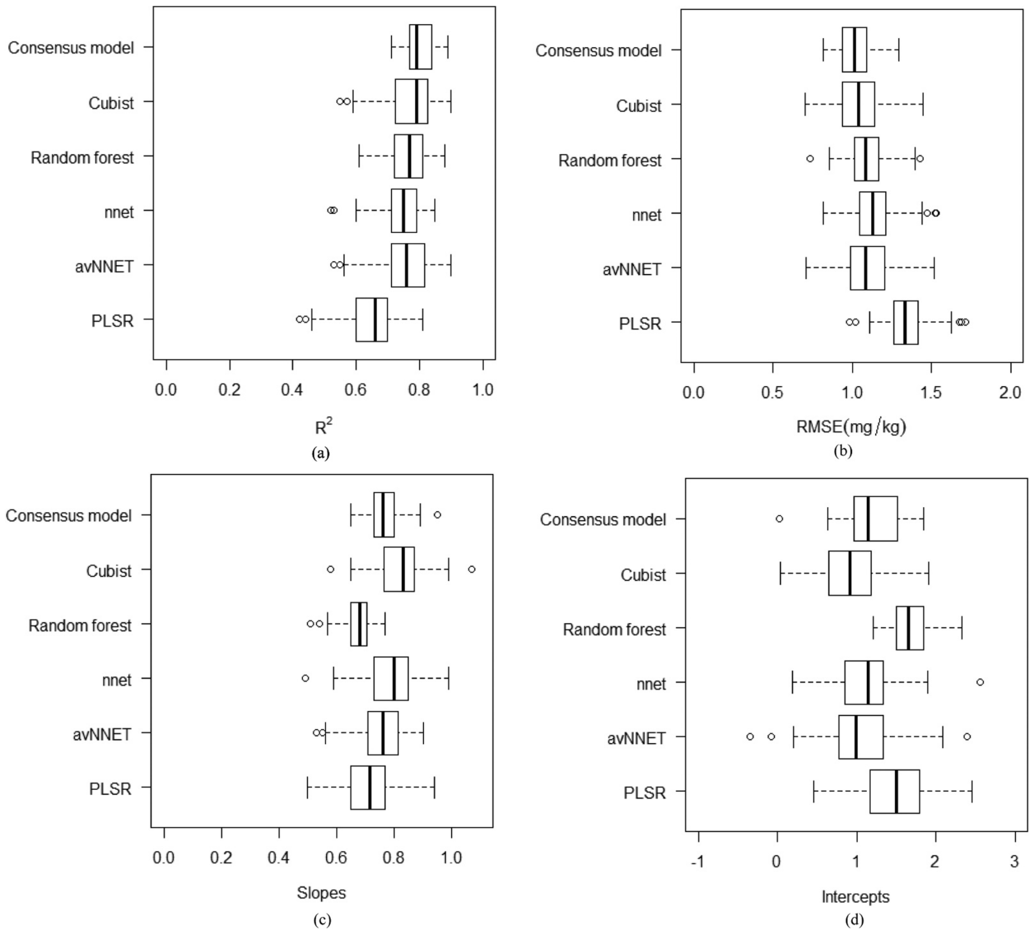

Fig. 4. (a): Box plots of R^2 of the models in total MCPD predictions. (b): Box plots of RMSE of the models in total MCPD predictions. (c): Box plots of slope values of the linear regression between observed and predicted total MCPD by the models. (d): Box plots of intercept values of the linear regression between observed and predicted total MCPD by the models.

total MCPD content prediction. Lastly, the testing accuracy by the PLSR model with R^2 = 0.65 was the lowest of the tested models.

The accuracy of the testing models can be further evaluated by mean RMSE. The results showed the same trend as their R^2 performance. The cubist model achieved the prediction of total MCPD content at the lowest RMSE = 1.046, followed by avNNET (RMSE =1.093), RF (RMSE = 1.095), nnet (RMSE = 1.141) and PLSR (RMSE = 1.346). Therefore, the sequential performance of the models based on their R^2 and RMSE box plots was, in decreasing order, cubist, RF, avNNET, nnet and PLSR. Evidently, consensus model was having comparable R^2 value with cubist model, but with much lower standard deviations. Besides, consensus model presented a prediction of total MCPD at the lowest RMSE with low standard deviation of RMSE as well. This observation clearly showed that a consensus model made up from high level fusion of cubist, RF, nnet, avNNET and PLSR models successfully improve the prediction of total MCPD in palm-based cooking oil to a higher level accuracy with low uncertainty.

3.3. Overfitting behavior

Generally, an overfitted model can provide an optimistic model with high R^2 (usually close to 1.0) and low RMSE in the training (calibration). However, an overfitted model will not perform optimally for prediction when a new dataset is introduced [31]. In other words, the overfitting phenomenon indicates that the specific algorithm simply learns all the given parameters and is able to predict the outcome within the training dataset without failure, but the predictive power for unseen data decreases. This feature should be avoided in the prediction of total MCPD content in palm-based cooking oils simply because the presence of MCPD in fats and oils is a critical parameter in terms of hazard potential and legislative issues.

Overfitting properties can be observed when there are drastic changes in the R^2 and RMSE of the models between their training and testing datasets. In our study, the difference between the training and testing R^2 means was approximately 0.13–0.23 for all the models. The training set RMSE ranged from 0.081 to 1.041. In the testing set, the RMSE mean range increased to between 1.004 and 1.346. Based on the observed R^2 and RMSE differences, the cubist model showed the largest differences in RMSE between training and testing among all the models generated. In addition, it showed a higher tendency to exhibit overfitting compared to its similar RF model.

For the neural network models, The behavior of reduction in R^2 and RMSE was similar, decreases of R^2 was about 13% and RMSE of both models were increased near 1.1 from 0.73 and 0.76 in the training set. This result showed nnet and avNNET model did possess certain level of overfitting but lower than a cubist model.

Although PLSR was considered to have the least accuracy and least overfitting tendency, we do not reject the potential application of PLSR in assessing biochemical properties. In a previous study of the relationship between the antioxidant properties of pegaga extract and its NMR spectra evaluated by PLS and neural network modeling, PLS was a preferred model compared to the neural network model because PLS provided better generalization and was safe from overfitting [25].

Based on the high accuracy discussed in the previous section, the results showed that all the models were relatively useful in predicting the total MCPD content of palm-based cooking oil based on the FTIR spectrum data. However, if the results of these 5 models are compared with a consensus model, a consensus model was a better alternative than using single model in the prediction. The overfit tendency was lowest with high accuracy. As discussed in the literature, consensus model is capable to reduce the overfit or underfit problem when the sample size is considered small. In this case, at fixed amount of sample size, (n = 156), consensus model showed better generalization and it should be practically used in the total MCPD prediction from FTIR spectrum.

3.4. Comparison between slopes, intercepts and predicted total MCPD content

Interestingly, all of the models showed overestimation of the predicted total MCPD content when the measured total MCPD content increased, but underestimation was observed after the interception between the perfect and actual fitted line. The interception of the perfect fitted line and the actual fitted line was recorded at approximately 5 ppm MCPD.

The observations suggested to us that the potential use of FTIR and multivariate analysis to determine total MCPD content during prescreening prior to tedious GCMS analysis is highly feasible. The slight overestimation at total MCPD content values below approximately 5 ppm provides a margin of safety for the estimation of the level of MCPDs in palm-based cooking oil.

Comparison of the slope and interception values among the models can access the accuracy of the models other than evaluation of R^2 and RMSE values. The evaluation as shown in Fig. 4(c) and (d). A slope value of 1.0 and an intercept value of 0 considered a model is having perfect accuracy, and perform well in predicting the Y-variables. From Fig. 4(c), the slope values of cubist and nnet models were consider high, average slope values were 0.81 and 0.79, respectively. Contrary, PLSR model was having low slope value (0.71) among the member models suggested the predictions were deviated from a perfect fitted line.

On the other hand, intercept values also evaluate the model which can overestimate the predicted result. One model interception at Y-axis is preferred to be as close to 0. From Fig. 4(d), PLSR model showed high intercept values with large standard deviation. RF model was having intercept at 1.66 by average, slightly higher compared to PLSR but the uncertainty was much narrow (low standard deviation).

Notably, a consensus model made from the mentioned member models showed that it was able to comprise the weakness of using single model. The slope value of consensus model was still slightly lower than cubist, but the slope values were always consistent among the models. Similarly, intercept values of a consensus model were least deviated, although it was not as close to 0 like cubist model.

3.5. Contribution of member models to the consensus model and feasibility of consensus model use

The contribution portion of each member model to the consensus model is shown in Fig. 5. Each member model was tested among 100 iterations (and repeated 5 times with K-fold cross-validation) and the testing R^2 acceptance criteria or threshold was 0.8. When a testing test was fulfilled a $R^2 > 0.8$, it is accepted as one of the weightage to contribute for average R^2 in consensus model. From the figure, it is not surprised the highest acceptance percentage was cubist model, followed by RF, avNNET, nnet and PLSR. The acceptance sequence agreed with the discussion above. Cubist model possessed the properties of higher R^2, lower RMSE, slope value closer to 1, and interception close to 0. Cubist model performed better as a single model among other tested model, therefore, the higher chance that cubist testing was contribute larger portion to the established consensus model. Contrary, the acceptance percentage of a PLSR was as low as 2.4%. Obviously, PLSR model was least preferred in MCPD prediction but the result was expected due to lower average R^2 of PLSR in the training if access individually.

The ultimate consensus result derived from every member models prediction. The comparisons between consensus model and single model clearly showed that consensus model was able to reduce redundancy and guarantee complementarity. Therefore, the prediction power of the systems was improved based on a fusion result [36].

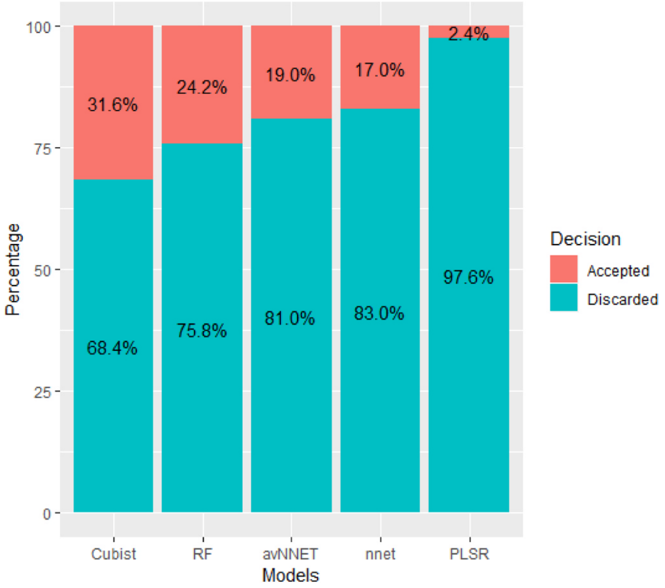

Fig. 5. Accepted and discarded percentages of member models to the final consensus regression model among 500 iterations.

3.6. Importance of variables for total MCPD content prediction

Unlike neural network models, the tree-based model is able to provide the user with a list of variables that are important in the prediction. In this study, the top five important variables (among the selected 51 wavenumbers) derived from the cubist model were at wavenumbers 950, 956, 1269, 991, and 952. These variables were furthered refined from important variables by Boruta selection as described in Section 3.1. From the available literature, it is suggested that compounds with C-O bonding from esters or ether functionality can appear as overlapping peaks in the region between 1350 and 950 cm^{-1} in the FTIR spectrum. In addition, a terminal or vinyl–CH=CH_2 and trans unsaturated CH=CH bonds can appear in the 990–900 cm^{-1} region [8]. This information suggested that the cubist model could estimate the total MCPD content based on the ester and unsaturated carbon-carbon bonds in the fatty acid chain. Although the exact relationships between the important variables and the content of MCPD are not known, unsaturated carbon bonds are found in abundance in palm oil, which consists of approximately 40% of the monounsaturated fatty acid oleic acid and 10% of the polyunsaturated fatty acid linoleic acid [7]. MCPDs in ester form may possess at least one unsaturated fatty acid chain [12] and could function as a potential indicator for determining their concentration based on analysis of the FTIR spectrum using the cubist model.

Similarly, the top few important variables contributing to the prediction by the RF model were found at wavenumbers 954, 956, 987, 950, 989, 991, 945 and 975 cm^{-1}. In addition, the wavenumbers 802 and 702 cm^{-1} were considered important variables in the RF model. The wavenumber in the region 800–700 cm^{-1} is well defined as the region corresponding to the C-Cl stretch bond in aliphatic chloro compounds [8]. These findings agreed with our initial hypothesis that the C-Cl bond should serve as an indicator for predicting the total MCPD content of palm-based cooking oils. We believe that processing of the FTIR spectrum using the SG algorithm and SNV normalization can improve the signal corresponding to this unique region.

4. Conclusions

It is necessary to compromise between the advantages and disadvantages of applying chemometrics to the prediction of total MCPD content from the FTIR spectrum. Based on our study, one shortcoming of using chemometrics such as the cubist, RF, nnet and avNNET models was the need for preprocessing of the FTIR spectrum. Furthermore, the demonstration performed in this study showed that there is an extensive need for data processing by coding software.

FTIR spectra include a large amount of data and require a tremendous amount of computational processing. Fortunately, R software and the processing library needed are available as open source materials. Although PLSR was developed by R software in the current study, PLSR is also commonly available as part of some commercial software packages such as SIMCA, which requires no capability in coding technique and offers a better graphical user interface (GUI).

Some of the main findings of the current study on predicting the total MCPD content from FTIR spectrum via chemometric analysis are as follows:

- The most accurate model in term of testing R^2 was the tree decision model, especially the cubist model, which recorded a testing $R^2 = 0.78$.
- The models that displayed a low degree of overfitting were the RF, avNNET, and PLSR models.
- The overall performance of the generated models in predicting the total MCPD content from the FTIR spectrum was good and showed high R^2 and similar RMSE values.
- The evaluation of slopes and intercepts of linear regression between observed and predicted MCPD showed cubist model was able to predict the data with least overestimation. PLSR model had the tendency to overestimate the total MCPD in prediction due to a steeper slope (away from 1.0) and high intercept value (1.48).
- A consensus modeling made from member models, namely, cubits, RF, nnet, avNNET, and PLSR was able to establish a stronger prediction tools for total MCPD prediction from palm-based cooking oil.
- The most accepted member model with a pre-set R^2 criteria above 0.8 was cubist model, followed by RF, avNNET, nnet, and PLSR.
- The main important variables suggested by the cubist and RF that ensured the detection of MCPD functional groups were $-CH=CH_2$ or $CH=CH$ (990–900 cm^{-1}) and C-Cl stretch (800–700 cm^{-1}).

Finally, the current study successfully demonstrated the application of chemometric analysis in the prediction of total MCPD in palm-based cooking oil. The use of chemometric analysis provides a cost-effective and reliable method for the rapid screening of MCPD in palm-based cooking oil. It is a highly flexible and repeatable method that extends the potential for estimation of the total MCPD content of other refined vegetable oils.

Acknowledgments

The work was supported by the Putra Grant, Universiti Putra Malaysia (Project no. UPM/700-2/1/GPB/2018/9629500). The authors want to thank Shimadzu (Malaysia) Sdn. Bhd for providing FTIR and GC-MS instruments.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.talanta.2019.01.111.

References

[1] S. Agatonovic-Kustrin, R. Beresford, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, J. Pharm. Biomed. Anal. 22 (5) (2000) 717–727.

[2] H. Ayvaz, M. Plans, K.M. Riedl, S.J. Schwartz, L.E. Rodriguez-Saona, Application of infrared microspectroscopy and chemometric analysis for screening the acrylamide content in potato chips, Anal. Methods 5 (8) (2013) 2020.

[3] H. Ayvaz, L.E. Rodriguez-Saona, Application of handheld and portable spectrometers for screening acrylamide content in commercial potato chips, Food Chem. 174 (2015) 154–162.

[4] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[5] T. Buhrke, R. Weisshaar, A. Lampen, Absorption and metabolism of the food contaminant 3-chloro-1,2-propanediol (3-MCPD) and its fatty acid esters by human intestinal Caco-2 cells, Arch. Toxicol. 85 (10) (2011) 1201–1208.

[6] L. Cerretani, A. Giuliani, R.M. Maggio, A. Bendini, T. Gallina Toschi, A. Cichelli, Rapid FTIR determination of water, phenolics and antioxidant activity of olive oil, Eur. J. Lipid Sci. Technol. 112 (10) (2010) 1150–1157.

[7] R. Clemens, A.W. Hayes, K. Sundram, P. Pressman, Palm oil and threats to a critically important food source, Toxicol. Res. Appl. 1 (2017) (239784731769984).

[8] J. Coates, Interpretation of infrared spectra, a practical approach, Encyclopedia of Analytical Chemistry, 2000.

[9] P. de la Mata, A. Dominguez-Vidal, J.M. Bosque-Sendra, A. Ruiz-Medina, L. Cuadros-Rodríguez, M.J. Ayora-Cañada, Olive oil assessment in edible oil blends by means of ATR-FTIR and chemometrics, Food Control 23 (2) (2012) 449–455.

[10] EFSA, Risks for human health related to the presence of 3-and 2-monochloropropanediol (MCPD), and their fatty acid esters, and glycidyl fatty acid esters in food, EFSA J. 14 (5) (2016) e04426.

[11] D. Gupta, A. Julka, S. Jain, T. Aggarwal, A. Khanna, N. Arunkumar, V.H.C. de Albuquerque, Optimized cuttlefish algorithm for diagnosis of Parkinson's disease, Cogn. Syst. Res. 52 (2018) 36–48.

[12] T.D. Haines, K.J. Adlaf, R.M. Pierceall, I. Lee, P. Venkitasubramanian, M.W. Collison, Direct determination of MCPD fatty acid esters and glycidyl fatty acid esters in vegetable oils by LC-TOFMS, J. Am. Oil Chem. Soc. 88 (1) (2011) 1–14.

[13] K. Kafadar, J.R. Koehler, Am. Stat. 53 (1) (1999) 86–87.

[14] R. Kannan, V. Vasanthi, Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease, (2019), pp. 63–72.

[15] M. Kuhn, Contributions from Jed Wing SW, Andre Williams, Chris Keefer and Allan Engelhardt. caret: Classification and Regression Training, R package version 5.15-023, 2012.

[16] M. Kuhn, S. Weston, C. Keefer, N. Coulter, R. Quinlan, Cubist: rule-and instance-based regression modeling, R package version 0.0, 18, 2014.

[17] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (11) (2010) 1–13.

[18] S.I. Lee, S. Celik, B.A. Logsdon, S.M. Lundberg, T.J. Martins, V.G. Oehler, E.H. Estey, C.P. Miller, S. Chien, J. Dai, A. Saxena, C.A. Blau, P.S. Becker, A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia, Nat. Commun. 9 (1) (2018) 42.

[19] S.Y. Lee, A. Mediani, M. Maulidiani, A. Khatib, I.S. Ismail, N. Zawawi, F. Abas, Comparison of partial least squares and random forests for evaluating relationship between phenolics and bioactivities of Neptunia oleracea, J. Sci. Food Agric. 98 (1) (2018) 240–252.

[20] B. Li, H. Wang, Q. Zhao, J. Ouyang, Y. Wu, Rapid detection of authenticity and adulteration of walnut oil by FTIR and fluorescence spectroscopy: a comparative study, Food Chem. 181 (2015) 25–30.

[21] Y. Li, J. Jing, A consensus PLS method based on diverse wavelength variables models for analysis of near-infrared spectra, Chemom. Intell. Lab. Syst. 130 (2014) 45–49.

[22] Y. Li, X. Shao, W. Cai, A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples, Talanta 72 (1) (2007) 217–222.

[23] A. Liaw, M. Wiener, Classification and regression by random, Forest R News 2 (3) (2002) 18–22.

[24] A. Liaw, M. Wiener. RandomForest: Breiman and Cutler's random forests for classification and regression, R package version 4.5-25. URL: ⟨http://CRAN.R-project.org/package=randomForest⟩.

[25] Abas, F. Maulidiani, A. Khatib, M. Shitan, K. Shaari, N.H. Lajis, Comparison of partial least squares and artificial neural network for the prediction of antioxidant activity in extract of Pegaga (Centella) varieties from 1H nuclear magnetic resonance spectroscopy, Food Res. Int. 54 (1) (2013) 852–860.

[26] W. Ng, B.P. Malone, B. Minasny, Rapid assessment of petroleum-contaminated soils with infrared spectroscopy, Geoderma 289 (2017) 150–160.

[27] T.M. Oshiro, P.S. Perez, J.A. Baranauskas. How many trees in a random forest? In: Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition, Springer, 2012, pp. 154–168.

[28] J.R. Quinlan. Combining instance-based and model-based learning, in: Proceedings of the Tenth International Conference on Machine Learning, 1993, pp. 236–243.

[29] M.M. Radhi, E.A. Jaffar Al-Mulla, W.H. Hoiwdy, Effect of temperature on frying oils: infrared spectroscopic studies, Res. Chem. Intermed. 39 (7) (2012) 3173–3179.

[30] Å Rinnan, F.V.d. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends Anal. Chem. 28 (10) (2009) 1201–1222.

[31] A. Rohman, Y.B. Che Man, Quantification and classification of corn and sunflower oils as adulterants in olive oil using chemometrics and FTIR spectra, ScientificWorldJournal 2012 (2012) 250795.

[32] A. Rohman, Y.B.C. Man, Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil, Food Res. Int. 43 (3) (2010) 886–892.

[33] B. Rudiyanto, Minasny, B.I. Setiawan, S.K. Saptomo, A.B. McBratney, Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands, Geoderma 313 (2018) 25–40.

[34] Y. Shen, S. Chen, R. Du, Z. Xiao, Y. Huang, B.A. Rasco, K. Lai, Rapid assessment of the quality of deep frying oils used by street vendors with Fourier transform infrared spectroscopy, J. Food Meas. Charact. 8 (4) (2014) 336–342.

[35] A. Stevens, L. Ramirez–Lopez, An introduction to the prospectr package, R Package

Vignette, Report No.: R Package Version 0.1, 3, 2014.

[36] Z. Su, W. Tong, L. Shi, X. Shao, W. Cai, A partial least squares-based consensus regression method for the analysis of near-infrared complex spectral data of plant samples, Anal. Lett. 39 (9) (2006) 2073–2083.

[37] W.N. Venables, B.D. Ripley. Tree-based methods, in: Modern Applied Statistics with S, Springer, pp. 251–269.

[38] R. Wehrens, B.-H. Mevik, The Pls Package: Principal Component and Partial Least Squares Regression in R, 2007.

[39] M.N. Wright, A. Ziegler, Ranger: a fast implementation of random forests for high dimensional data in C+ + and R, 2017 77 (1) (2017) 17.

[40] H. Zhang, P. Jin, M. Zhang, L.Z. Cheong, P. Hu, Y. Zhao, L. Yu, Y. Wang, Y. Jiang, X. Xu, Mitigation of 3-monochloro-1,2-propanediol ester formation by radical scavengers, J. Agric. Food Chem. 64 (29) (2016) 5887–5892.

[41] X. Zhang, B. Gao, F. Qin, H. Shi, Y. Jiang, X. Xu, L.L. Yu, Free radical mediated formation of 3-monochloropropanediol (3-MCPD) fatty acid diesters, J. Agric. Food Chem. 61 (10) (2013) 2548–2555.