

Article

Robust Wavelength Selection Using Filter-Wrapper Method and Input Scaling on Near Infrared Spectral Data [†]

Divo Dharma Silalahi ¹, Habshah Midi ^{2,3,*}, Jayanthi Arasan ^{2,3}, Mohd Shafie Mustafa ^{2,3} and Jean-Pierre Caliman ¹

¹ SMART Research Institute, PT. SMART TBK, Pekanbaru 28289, Riau, Indonesia; divosilalahi@yahoo.co.id (D.D.S.); caliman@indo.net.id (J.-P.C.)

² Institute for Mathematical Research, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia; jayanthi@upm.edu.my (J.A.); mshafie@upm.edu.my (M.S.M.)

³ Department of Mathematics, Faculty of Science, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia

* Correspondence: habshah@upm.edu.my

[†] The results of this study were presented at the 62nd International Statistical Institute-World Statistics Congress (ISI-WSC 2019), held in Kuala Lumpur (Malaysia), on 19 August 2019.

Received: 4 August 2020; Accepted: 30 August 2020; Published: 3 September 2020



Abstract: The extraction of relevant wavelengths from a large dataset of Near Infrared Spectroscopy (NIRS) is a significant challenge in vibrational spectroscopy research. Nonetheless, this process allows the improvement in the chemical interpretability by emphasizing the chemical entities related to the chemical parameters of samples. With the complexity in the dataset, it may be possible that irrelevant wavelengths are still included in the multivariate calibration. This yields the computational process to become unnecessary complex and decreases the accuracy and robustness of the model. In multivariate analysis, Partial Least Square Regression (PLSR) is a method commonly used to build a predictive model from NIR spectral data. However, in the PLSR method and common commercial chemometrics software, there is no standard wavelength selection procedure applied to screen the irrelevant wavelengths. In this study, a new robust wavelength selection procedure called the modified VIP-MCUVE (mod-VIP-MCUVE) using Filter-Wrapper method and input scaling strategy is introduced. The proposed method combines the modified Variable Importance in Projection (VIP) and modified Monte Carlo Uninformative Variable Elimination (MCUVE) to calculate the scale matrix of the input variable. The modified VIP uses the orthogonal components of Partial Least Square (PLS) in investigating the informative variable in the model by applying the amount of variation both in X and $y\{SSX,SSY\}$, simultaneously. The modified MCUVE uses a robust reliability coefficient and a robust tolerance interval in the selection procedure. To evaluate the superiority of the proposed method, the classical VIP, MCUVE, and autoscaling procedure in classical PLSR were also included in the evaluation. Using artificial data with Monte Carlo simulation and NIR spectral data of oil palm (*Elaeis guineensis* Jacq.) fruit mesocarp, the study shows that the proposed method offers advantages to improve model interpretability, to be computationally extensive, and to produce better model accuracy.

Keywords: near infrared spectral data; robust statistics; partial least squares; scaling; variable selection; variable importance in projection; uninformative variable eliminations

1. Introduction

The selection of relevant wavelengths in chemometrics analysis in Near Infrared (NIR) spectral data is crucial to prevent the number of relevant variables to be removed in the analysis. Partial Least Square Regression (PLSR), a conventional method used in chemometrics, has no standard procedure. There are related studies of its application in oil and fat assessment in oil palm (*Elaeis guineensis* Jacq.). However, many researchers are still using the band partition experiments (see [1–3]) by manually segmenting the wavelengths into several bands: visible (400–700 nm), NIR (701–1100 nm), Shortwave Infrared 1 (SWIR1: 1101–1351 nm), SWIR2 (1400–1800 nm), and SWIR3 (2000–2500 nm). The selection procedure is commonly done based on trial and error [1] through the improvement in model accuracy attained. This observation is inefficient and requires advanced experience. Moreover, an in-depth analysis needs to be carried out to understand the NIR spectra signature based on its chemical information. A wavelength selection method is therefore required to assess the contribution of each wavelength. This selection will reduce the number of variables used in the model. In the interpretation, the selection method may give a better understanding of the underlying process of the sample studied.

In the review papers (see [4–7]), several recommended wavelength selection methods have been discussed. The researchers have highlighted the limitations and properties of each method presented. No one has suggested which method is better than the other. A convenient approach is to make a comparison between the methods and examine their superiority using the experimental simulation and real given problems. There are three main categories of variable selection methods: filter, wrapper, and embedded. The main differences between these categories are based on their processing steps. In the filter methods, relevant variables are ranked and selected according to the threshold on the relevancy index calculated from the fitted model [8,9]. The filter methods are considered fast and straightforward because no learning algorithm is required in the computational process. However, the filter methods do not take into account biasness in the learning model and neglect any conditional dependence (or independence) that might probably exist [10]. In the wrapper methods, the supervised learning approach adopts the search algorithm iteratively with either the deterministic or randomization [5]. The wrapper methods are known to be costly because in the evaluation criterion, a predefined learning algorithm and cross-validation procedure is performed [5]. In the embedded methods, the advantages of both the filter and wrapper methods are seized [7]. The embedded methods evaluate the quality of selected relevant variables during the model building without performing an evaluation process on the learning model [11]. This embedded strategy has motivated the current study to integrate the combination steps in the filter and wrapper methods with the objective to improve the performance of the PLSR model.

There are several methods associated with both the filter and the wrapper methods. In the filter methods, some selection procedures have been reported, such as the Stepwise Regression Coefficients [12], Loading Weights [13], Correlation Coefficient [14], and Variable Importance in Projection (VIP) [15]. Among these procedures, the VIP has earned considerable attention because of its stability and consistency to select the relevant wavelengths and its relatively low computational cost. Using different dataset matrix as simulation, the VIP outperforms the other selection methods (see [6,16]). In the classical VIP [17], the score is calculated by using the weight combination of the overall component variables of the squared PLSR weight. However, the score highlights the variation only in the predictive components without including the orthogonal components. An upgrade of the VIP score based on the Orthogonal Projections to Latent Structures (OPLS) [18] was then introduced. The OPLS-VIP score systematically includes and differentiates the variation both in the predictive and orthogonal components. In some studies (see [18,19]), the OPLS-VIP has shown its superiority to the classical VIP. In the wrapper methods, some selection methods have also been discussed, such as the Genetic Algorithm (GA) [20], Uninformative Variable Elimination (UVE) [21], Iterative Predictor Weighting [22], and Backward Variable Elimination [23]. Among these methods, the UVE method is the most consistent method to improve the PLSR performance [21,24].

Nonetheless, the leave-one-out validation procedure is still applied in the classical UVE. This procedure leads to overfitting and is time-consuming while acquiring the stability values for a large dataset. As an improvement, the Monte Carlo Uninformative Variable Elimination (MCUVE) method was proposed by Cai [25]. The MCUVE adopts the principle of Monte Carlo to evaluate the stability of the corresponding coefficients. However, the reliability coefficient and the cut-off criterion performed in the method to date is not robust enough. Therefore, this has motivated the current study to propose a slight improvement to make the MCUVE procedure more robust.

In practice, it is difficult to eliminate all the irrelevant wavelengths in spectra processing. A smaller number of wavelengths (as predictor variables) used in the model calibration will result in overfitting or underfitting. To overcome this, a new robust procedure to highlight the relevant wavelengths and to downgrade the influence of the irrelevant wavelengths in the PLSR model is needed. It has been investigated that the scaling method in the PLS model is also essential to improve the convergence speed of the algorithm [26,27]. In addition, the auto-scaling method using the mean centering and standard deviation, is a common scaling procedure in the data pre-processing step. Hence, another scaling strategy should be considered in the improvement.

The main objectives of this study are (1) to establish a new procedure for wavelengths selection called the modified VIP-MCUVE (mod-VIP-MCUVE) with input scaling strategy in the PLSR model; (2) to evaluate the performance of the proposed method with the standard auto-scaling procedure in the PLSR and the input scaling strategy using the classical VIP and MCUVE methods; (3) to apply the proposed method on the artificial data and NIR spectra of oil palm fruit mesocarp (fresh and dried ground).

2. Materials and Methods

2.1. Partial Least Square Regression

Partial Least Square Regression (PLSR) was firstly initiated by Wold [28] as the generalized statistical method and standard method used in the spectroscopy analysis. Let us define a multiple regression model that relates several m predictors \mathbf{X} to a response variable \mathbf{y} . In matrix form this can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (1)$$

where \mathbf{y} is an $n \times 1$ vector of the response variable, \mathbf{X} is $n \times m$ matrix of predictors, \mathbf{b} is a $m \times 1$ vector of unknown parameters, and \mathbf{e} is a $n \times 1$ vector of random errors. The solution for the estimator \mathbf{b} using the least-squares method is given as

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (2)$$

Here the data set problem is in condition with a large number of m predictors. Hence, there will be an infinite number of solutions for estimating \mathbf{b} as $\mathbf{X}^T\mathbf{X}$ is singular, which does not meet the usual trivial theorem on rank in the regression. In this case, it is necessary to extract the new latent variables by maximizing a covariance criterion between predictor \mathbf{X} and response \mathbf{y} that link the central values of these two sets [29].

Initialize a starting $n \times 1$ score vector of \mathbf{u} from any single \mathbf{y} as in Equation (1), there exists an outer relation for predictor \mathbf{X} as

$$\mathbf{X} = \mathbf{V}\mathbf{P}^T + \mathbf{E} \quad (3)$$

where $\mathbf{P} \left\{ \mathbf{p}_g = (\mathbf{X}^T \mathbf{v}_g) / (\mathbf{v}_g^T \mathbf{v}_g) \right\}_{g=1}^l$ is the matrix $m \times l$ consists of loading vector $m \times 1$, $\mathbf{v}_g \left\{ \mathbf{v}_g = (\mathbf{X} \mathbf{w}_j) / (\mathbf{w}_j^T \mathbf{w}_j) \right\}_{g=1}^l$ is the $n \times 1$ column vector of scores \mathbf{x}_j in \mathbf{X} involves $\mathbf{w}_j \left\{ \mathbf{w}_j = (\mathbf{X}^T \mathbf{u}) / (\mathbf{u}^T \mathbf{u}) \right\}_{j=1}^m$ as $m \times 1$ vector of weight for \mathbf{X} , \mathbf{V} is a $n \times l$ matrix of $n \times 1$ vector \mathbf{v}_g , and \mathbf{E} is a $n \times m$ matrix of residual in outer relation for predictor \mathbf{X} . Following these, the outer relation for the response \mathbf{y} also can be defined as

$$\mathbf{y} = \mathbf{u} \mathbf{q}^T + \mathbf{f} \quad (4)$$

where \mathbf{q} is the loading $l \times 1$ vector $\{\mathbf{q}_g = (\mathbf{y}^T \mathbf{v}_g) / (\mathbf{v}_g^T \mathbf{v}_g)\}_{g=1}^l$ and \mathbf{f} is a $n \times 1$ vector of residual in \mathbf{y} . $\mathbf{u} \{\mathbf{u} = \mathbf{b}_g \mathbf{v}_g\}_{g=1}^l$ is also called as linear inner relation between \mathbf{X} and \mathbf{y} block score, which simply can be written as

$$\mathbf{u} = \mathbf{V} \mathbf{b}_{inner} + \mathbf{g} \quad (5)$$

where $\mathbf{b}_{inner} \{\mathbf{b}_g = \mathbf{u}^T \mathbf{v}_g / (\mathbf{v}_g^T \mathbf{v}_g)\}_{g=1}^l$ is a $l \times 1$ vector of regression coefficient as Least Square solution on the decomposition of vector \mathbf{u} , and \mathbf{g} is a $n \times 1$ vector of residual in the inner relation. Applying the normalization in \mathbf{P} , \mathbf{W} , and \mathbf{q} as the process to improve the inner relation, the mixed relation of PLSR model by integrating Equations (4) and (5) results as

$$\begin{aligned} \mathbf{y} &= \mathbf{u} \mathbf{q}^T + \mathbf{f} \\ \mathbf{y} &= \mathbf{V} \mathbf{b}_{inner} \mathbf{q}^T + \underbrace{\mathbf{g} \mathbf{q}^T}_{\tilde{\mathbf{f}}} + \mathbf{f} \\ \mathbf{y} &= \mathbf{V} \mathbf{b}_{inner} \mathbf{q}^T + \tilde{\mathbf{f}} \\ \mathbf{y} &= \mathbf{V} \mathbf{a}^T + \tilde{\mathbf{f}} \end{aligned} \quad (6)$$

where $\mathbf{a} \{\mathbf{a}^T = \mathbf{b}_{inner} \mathbf{q}^T\}$ is the $l \times 1$ vector coefficient and $\tilde{\mathbf{f}} \{\tilde{\mathbf{f}} = \mathbf{g} \mathbf{q}^T + \mathbf{f}\}$ denotes $n \times 1$ vector of residual in the mixed relation. Equation (6) holds $\mathbf{a} = \mathbf{V}^T \mathbf{y}$, and without loss of generality $\mathbf{X} = \mathbf{V} \mathbf{P}^T$ as in Equation (3), the formulation in Equation (6) can be reconstructed by multiplying the two sides with weight matrix of \mathbf{W} which is

$$\begin{aligned} \mathbf{y} &= \mathbf{X} \mathbf{W}^* \mathbf{a} + \tilde{\mathbf{f}} \\ \mathbf{y} &= \mathbf{X} \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{a} + \tilde{\mathbf{f}} \end{aligned} \quad (7)$$

with $\mathbf{V} = \mathbf{X} \mathbf{W}^*$ and $\mathbf{W}^* = \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1}$.

Let us define $\mathbf{b}_{PLSR} = \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{a}$ as $m \times 1$ vector coefficient of mixed relation in the PLSR, then Equation (7) is equivalent to

$$\mathbf{y} = \mathbf{X} \mathbf{b}_{PLSR} + \tilde{\mathbf{f}} \quad (8)$$

where $\tilde{\mathbf{f}}$ has to be minimized. Applying the relation in Equations (3) and (4), so that $\mathbf{W} = \mathbf{X}^T \mathbf{u}$, $\mathbf{P} = \mathbf{X}^T \mathbf{V} (\mathbf{V}^T \mathbf{V})^{-1}$. Therefore, the estimator for the parameter \mathbf{b}_{PLSR} can be calculated as

$$\hat{\mathbf{b}}_{PLSR} = \mathbf{X}^T \mathbf{u} (\mathbf{V}^T \mathbf{X} \mathbf{X}^T \mathbf{u})^{-1} \mathbf{V}^T \mathbf{y}, \quad \hat{\mathbf{b}}_{PLSR} \in \mathfrak{R}^{m \times 1} \quad (9)$$

$\hat{\mathbf{b}}_{PLSR}$ denotes the m dimensional vector of regression coefficient in the PLSR model.

2.2. Variable Selection Methods

2.2.1. Variable Importance in Projection

In the classical VIP [17], the VIP score measures the contribution of each j th wavelength in the multivariate models based on the projection to the PLS components. The method becomes popular because of its simple procedure and less computational complexity [16]. The VIP score is formulized through the normalized loading weights $\mathbf{v}_g \{\mathbf{v}_g = \mathbf{w}_g / \|\mathbf{w}_g\|\}$ and the explained sum of squares for the predictive component \mathbf{y} . Mathematically, the VIP score for each j th wavelength in the PLS model with l components can be calculated as

$$VIP_j^2 = \frac{m \sum_{g=1}^l (v_g^2 SSY_{comp;g})}{\sum_{g=1}^l SSY_{comp;g}}$$

where m is the number of predictors, $SSY_{comp;g}$ is the variance of \mathbf{y} explained by the g th PLS component, and $\sum_{g=1}^l SSY_{comp;g}$ is the total variance summarized by the PLS model over l components. The criteria are the j th wavelength with VIP score > 1 is considered as the most relevant, while the VIP score < 0.5 is considered as irrelevant wavelength.

2.2.2. Uninformative Variable Elimination

The classical UVE method [21] uses the leave-one-out jackknife method and artificial random noise variables denoted as $n \times m$ matrix \mathbf{N} to compute the statistic parameters. The reliability of each wavelength through variable selection criterion then is calculated based on the PLSR coefficient $\mathbf{b}_{PLSR} = [b_1, b_2, \dots, b_m]$. However, when handling a large dataset, this procedure becomes costly [25,30]. As a solution, the Monte Carlo method, which is based on random selection and probability statistics, is applied in the UVE and so-called as MCUVE [25]. In the MCUVE, some specific number of a subsample N_t from the training set are randomly selected to build the r PLS sub-model. Then, it produces the number of set of PLSR coefficient $\hat{\mathbf{b}}_{PLSR}$ as $r \times m$ matrix. The reliability c_j is computed based on the fraction between the mean and standard deviation of $m \times 1$ column vector of PLSR coefficient b_{i^*j} in each j th wavelength from the i^* vectors of coefficients. The highest c_j represents the most reliable wavelengths; otherwise, the class is as a less reliable wavelength.

$$c_j = \frac{\bar{b}_j}{s(b_j)}$$

$$\bar{b}_j = \frac{\sum_{i^*=1}^r b_{i^*j}}{r}$$

$$s(b_j) = \left(\sum_{i^*=1}^r \frac{(b_{i^*j} - \bar{b}_j)^2}{r-1} \right)^{1/2}$$
(10)

The cut-off threshold criterion in MCUVE is defined through the maximum absolute value of the reliability c_{artif} of the artificial random noise variables matrix \mathbf{N} . The wavelength with c_j less than the artificial random noise c_{artif} is removed from the PLSR model.

$$|(c_j)| < |\max(c_{artif})|$$
(11)

3. Input Scaling of Filter-Wrapper Method

Auto-scaling is a common input scale method used to standardize a dataset in the modeling process of PLSR. This auto-scaling transforms each numeric in the input variable into the same variance through its mean and standard deviation [31,32]. However, this method is observed disadvantageous when the original input variables are measured on the same scale. Moreover, it removes the interpretability related to the loadings [33]. Taking a particular concern in the intensity power of Near Infrared region, auto-scaling fails to keep useful interpretive information about the wavelength contribution since the low-intensity regions are enhanced to the same magnitude as like in the high intensity [34]. The scaling method is very crucial to correct for wavelengths-dependent scattering effects in the NIR spectra dataset [35,36]. To overcome this, a new alternative input scaling method based on the Filter and Wrapper methods is proposed. This method is then simply denoted as mod-VIP-MCUVE. Besides correcting the wavelengths-dependent scattering effects and preserving the chemical interpretive information in each wavelength, the proposed method will eliminate the influence of irrelevant

wavelengths during the modeling process. In general, the three main computational steps of the mod-VIP-MCUVE method are the following.

- Step 1: Calculate the OPLS-VIP [19] score as initial input variable scaling matrix.
- Step 2: Run the modified MCUVE procedure using the scaled input matrix of OPLS-VIP to get the reliability scores.
- Step 3: Re-scale the input matrix using the reliability scores as final scaled input matrix in the PLSR model.

In the OPLS-VIP [19], the VIP score is measured not only based on the projections to the PLS components but also included the orthogonal components. This score considers variations both in the predictor variable \mathbf{X} (SSX) and the response variable \mathbf{y} (SSY). Here, the fourth variant of four versions of OPLS-VIP is preferred due to its interpretative information ability. The OPLS-VIP score uses the combinations {SSX,SSY} in the weighting parameters and normalized loadings \mathbf{v}_g . The total OPLS-VIP score then is used as the final VIP score as it is calculated based on the VIP_{pred} (predictive components) and VIP_{ortho} (orthogonal components).

Let us redefine g as the predictive component and g_o as the orthogonal component, then l stands for the total number of predictive components, and l_o stands for the total number of orthogonal components with m and m_o are the total number of variables used in the predictive and orthogonal components, respectively. The formulation for OPLS-VIP score both in predictive and orthogonal can be written as

$$VIP_{pred} = \sqrt{\frac{m}{2} \left(\frac{\sum_{g=1}^l (\mathbf{v}_g^2 SSX_{comp;g})}{SSX_{cum}} + \frac{\sum_{g=1}^l (\mathbf{v}_g^2 SSY_{comp;g})}{SSY_{cum}} \right)} \quad (12)$$

$$VIP_{ortho} = \sqrt{\frac{m_o}{2} \left(\frac{\sum_{g_o=1}^{l_o} (\mathbf{v}_{o_{g_o}}^2 SSX_{comp;g_o})}{SSX_{cum}} + \frac{\sum_{g_o=1}^{l_o} (\mathbf{v}_{o_{g_o}}^2 SSY_{comp;g_o})}{SSY_{cum}} \right)} \quad (13)$$

where, as in Equations (12) and (13), the sum of square (SS) both in variable \mathbf{y} and variable \mathbf{X} has subscript $comp; g$ and $comp; g_o$. The subscript $comp; g$ refers to the explained SS of g th component in the predictive, while the subscript $comp; g_o$ refers to the explained SS of g_o th component in the orthogonal. The SS with subscript cum refers to the cumulative explained SS overall components in the model. The total of OPLS-VIP score (denotes as VIP-total) is then formulized as

$$VIP_{total} = \sqrt{\frac{M}{2} \left(\frac{\sum_{g_o=1}^{l_o} (\mathbf{v}_{o_{g_o}}^2 SSX_{comp;g_o})}{SSX_{cum}} + \frac{\sum_{g=1}^l (\mathbf{v}_g^2 SSX_{comp;g})}{SSX_{cum}} + \frac{\sum_{g_o=1}^{l_o} (\mathbf{v}_{o_{g_o}}^2 SSY_{comp;g_o})}{SSY_{cum}} + \frac{\sum_{g=1}^l (\mathbf{v}_g^2 SSY_{comp;g})}{SSY_{cum}} \right)} \quad (14)$$

where M is the sum of variables used both in the predictive and orthogonal components

$$\left\{ m = M / \left(\frac{SSX_{cum;g}}{SSX_{cum}} + \frac{SSY_{cum;g}}{SSY_{cum}} \right) \right\}; \left\{ m_o = M / \left(\frac{SSX_{cum;g_o}}{SSX_{cum}} + \frac{SSY_{cum;g_o}}{SSY_{cum}} \right) \right\}$$

The above total OPLS-VIP score is used to scale the original wavelength variables as the new input matrix. Let us define $\tilde{\mathbf{X}}$ as the scaled input variable that is calculated by using the OPLS-VIP score on predictor variable \mathbf{X} which are initially not scaled. Mathematically it can be written as

$$\begin{aligned}\tilde{\mathbf{X}} &= \mathbf{X}\Omega \\ \Omega &= \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)\end{aligned}\quad (15)$$

where $\Omega \in \mathfrak{R}$ is said to be a diagonal weight matrix with size $m \times m$. The element λ_j in the diagonal matrix Ω is a non-negative variable scaling factor for the j th input wavelength. The new scaled $\tilde{\mathbf{X}}$ is used as a new input matrix in the modified MCUVE.

In the modified MCUVE, the drawback of the classical cut-off threshold criterion in Equation (11) has been discussed by Centner (see [21]). As an alternative, a new modified robust cut-off criterion based on a one-sided tolerance interval from Natrella [37] is proposed with a better stable elimination on the irrelevant wavelengths. The cut-off value is calculated using the robust location and scale of the reliability coefficients obtained from the added artificial uninformative random variable. In addition, it includes the value of k factor as a function of the γ desired proportions, α as a level of error, and r number of repetition used in MC random subsample selection. Using the c_{artif} in Equation (11), then the new proposed cut-off criterion can be defined as

$$\text{cut-off value} = \text{median}((c_j)_{artif}) + k(\text{MAD}(c_j)_{artif}) \quad (16)$$

where k can be calculated as

$$k = \frac{z_\gamma + \sqrt{z_\gamma^2 - ab}}{a}$$

With constant parameters $\left\{a = 1 - \frac{z_\alpha^2}{2(r-1)}\right\}$ and $\left\{b = z_\gamma^2 - \frac{z_\alpha^2}{r}\right\}$. This new cut-off criterion benefits from classifying the most ($c_j > \text{cut-off value}$) and less relevant variable for further interpretation. Applying the reliability c_j of modified MCUVE as the element λ_j in Ω , the new scaled input variable $\tilde{\mathbf{X}}^*$ for the PLSR model is then updated.

4. Monte Carlo Simulation Study

A simulation study was carried out to evaluate the performance of our proposed method and to compare its performance with some existing methods discussed in this study. Following the simulation study of Kim [26], the artificial dataset was generated randomly using the Uniform distribution (0,1) and included the added noise that follows the normal distribution. This dataset was applied in the linear combination equation with different scenarios. Five sample sizes ($n = 40, 60, 150, 400$, and 600), three levels of number of predictor variables ($m = 41, 101$, and 201), and five levels of number of important variables ($IV = 0.10, 0.20, 0.40, 0.60$, and 0.80) were considered. The 100 (IV) % of predictor variables were selected as important variables, and the remaining 100 ($1 - IV$) % were considered as less important. The formulation of this simulation can be defined as follows

$$\begin{aligned}m &= m_o + m_e \\ \mathbf{c}_{j^o} &\sim U(1, 10) \quad (j^o = 1, 2, \dots, m_o) \\ \mathbf{c}_{j^e} &\sim U(5, 20) \quad (j^e = 1, 2, \dots, m_e) \\ \mathbf{e}_j &\sim N(0, 1) \quad (j = 0, 1, 2, \dots, m) \\ \mathbf{b} &\sim U(0, 7) \\ \mathbf{X} &= \{\mathbf{c}_{j^o}, \mathbf{c}_{j^e}\} + \mathbf{e}_j \quad (j = 1, 2, \dots, m; j^o = 1, 2, \dots, m_o; j^e = 1, 2, \dots, m_e) \\ \mathbf{iv} &= \{iv_1, iv_2, \dots, iv_{100(IV)\% * m}\} \\ \mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{e}_0 \quad (i = 1, 2, \dots, n; j = iv_1, iv_2, \dots, iv_{100(IV)\% * m})\end{aligned}\quad (17)$$

where m is the total number of predictors, m_o is the number of observable variables and the m_e ($m_e = (m - 100(IV)\% * m)/2$) is the number of artificial noise variable. These artificial variables are classified as less important variables in the dataset. In Equation (17), the \mathbf{c}_{j^o} , \mathbf{c}_{j^e} , and \mathbf{e}_j are

independent of each other while \mathbf{X} and \mathbf{y} are illustrated as observable variable. The \mathbf{c}_j follows the Uniform distribution (1,10) with size n . The artificial noise variables \mathbf{ce}_j are added to the predictor and follow the Uniform distribution (5,20) with size n . This \mathbf{ce}_j is classified as a less important variable. The \mathbf{e}_j follows the standard normal distribution with size n and \mathbf{b} represents a vector coefficient for selected important variables which follows the Uniform distribution (0,7) with size m . The \mathbf{iv} as the set of selected important variables in m_0 and \mathbf{e}_0 is added error in the linear combination of \mathbf{y} . In the PLSR model, the number of PLS components is a principal indicator in the modeling since it is always viewed to be subjective.

In Figure 1, a re-sampling procedure called cross-validation, showing the lowest Root Mean Square Error of Prediction (RMSEP) is used to select the optimum number of PLS components. ‘Selection’ means the selected optimum number of PLS components suggested by cross-validation techniques (highlighted with the blue dashed line). While, the ‘Abs.minimum’ refers to the lowest RMSEP (highlighted with the gray dashed line). As the number of PLS components used in the PLSR model increases, the mean of RMSEP also decreases. The optimum number of PLS components depends on how well the specific number of original variables have contribution to the model. In the experiment using different levels of n , m , and IV , the proposed mod-VIP-MCUVE requires a smaller amount of PLS components to fit minimum RMSEP than the other methods (classical PLSR method with no input scaling applied, VIP and MCUVE). The results are consistent and satisfying. The MCUVE input scaling method is comparable to the proposed method, but it still produces higher RMSEP values. When accommodating fewer variables used as predictor, a faster computational speed will be attained. Based on the global minimum cross-validation, the proposed mod-VIP-MCUVE has succeeded in reducing the RMSEP and improving the accuracy of the PLSR model.

Several statistical measures as evaluation indices are used to assess the goodness of the methods: Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Ratio of Performance to Deviation (RPD), and Standard Error (SE). The RMSE indicates the absolute measure of fit, R^2 measures the proportion of variation in the data explained by the model, RPD assess the reliability of the goodness of fit for model, and SE measures of the uncertainty in the NIRS prediction. In this section, the Monte Carlo simulation was run 10,000 repeated times, and the results are based on the average of statistical measures (see Table 1). Some scenarios using different treatments are applied to evaluate the PLSR model. According to the results, comparing the RMSE, R^2 , and SE values in all scenarios, the proposed mod-VIP-MCUVE produced better accuracy than the other methods. The reliability of proposed method based on its RPD is still outperformed. Downgrading the irrelevant variables in the fitting process, the performance of the proposed method is comparable to the classical PLSR method with full variables involved. This shows that the proposed mod-VIP-MCUVE with fewer numbers of variables is more efficient than the traditional PLSR model since it could obtain a similar accuracy.

The most relevant variables selected by the methods are classified based on their cut-off threshold criterion on the score values. To evaluate the interpretability, the calculations are plotted in Figure 2.

In Figure 2, the selection of variables in each method uses different cut-off criteria. The selection includes the use of the VIP total score because this procedure is also included in the proposed mod-VIP-MCUVE method. For selection, the classical VIP and the VIP total method use VIP score >1 , while the MCUVE and the mod-VIP-MCUVE use cut-off threshold criterion. The VIP-total suggests a greater number of relevant variables than the classical VIP. The MCUVE uses standard cut-off threshold criterion then classifies a greater number of relevant variables than the classical VIP and the VIP-total. The proposed mod-VIP-MCUVE uses robust cut-off threshold (red line threshold) then collects a higher number of relevant variables than the MCUVE (green line threshold). The mod-VIP-MCUVE succeeds to downgrade (close to 0) the irrelevant variables and highlight the pertinent variables of the computational process. Using the proposed mod-VIP-MCUVE, the final subset of selected relevant variables guarantees the best prediction capabilities with better accuracy than the other methods.

The computing time performance during the fitting process was recorded to evaluate the efficiency of the proposed mod-VIP-MCUVE method. In Figure 3, the proposed mod-VIP-MCUVE method

outperformed the others. Using different sample sizes and numbers of predictor use, the proposed method is still consistent in expediting the convergence speed. The PLSR has the worst performance due to its inefficient computing time even the auto-scaling is naturally applied using its mean and standard deviation in the procedure.

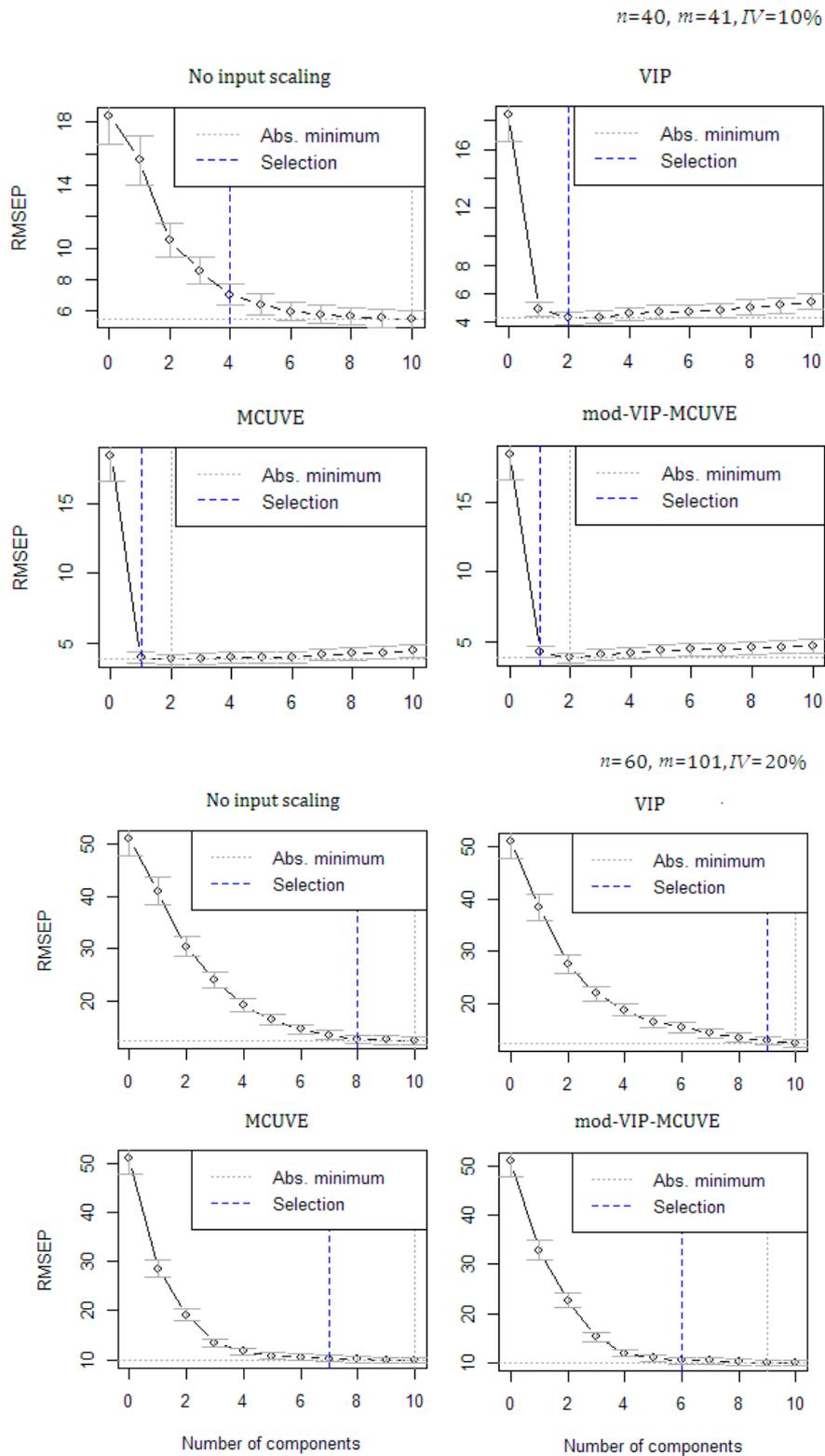


Figure 1. Cont.

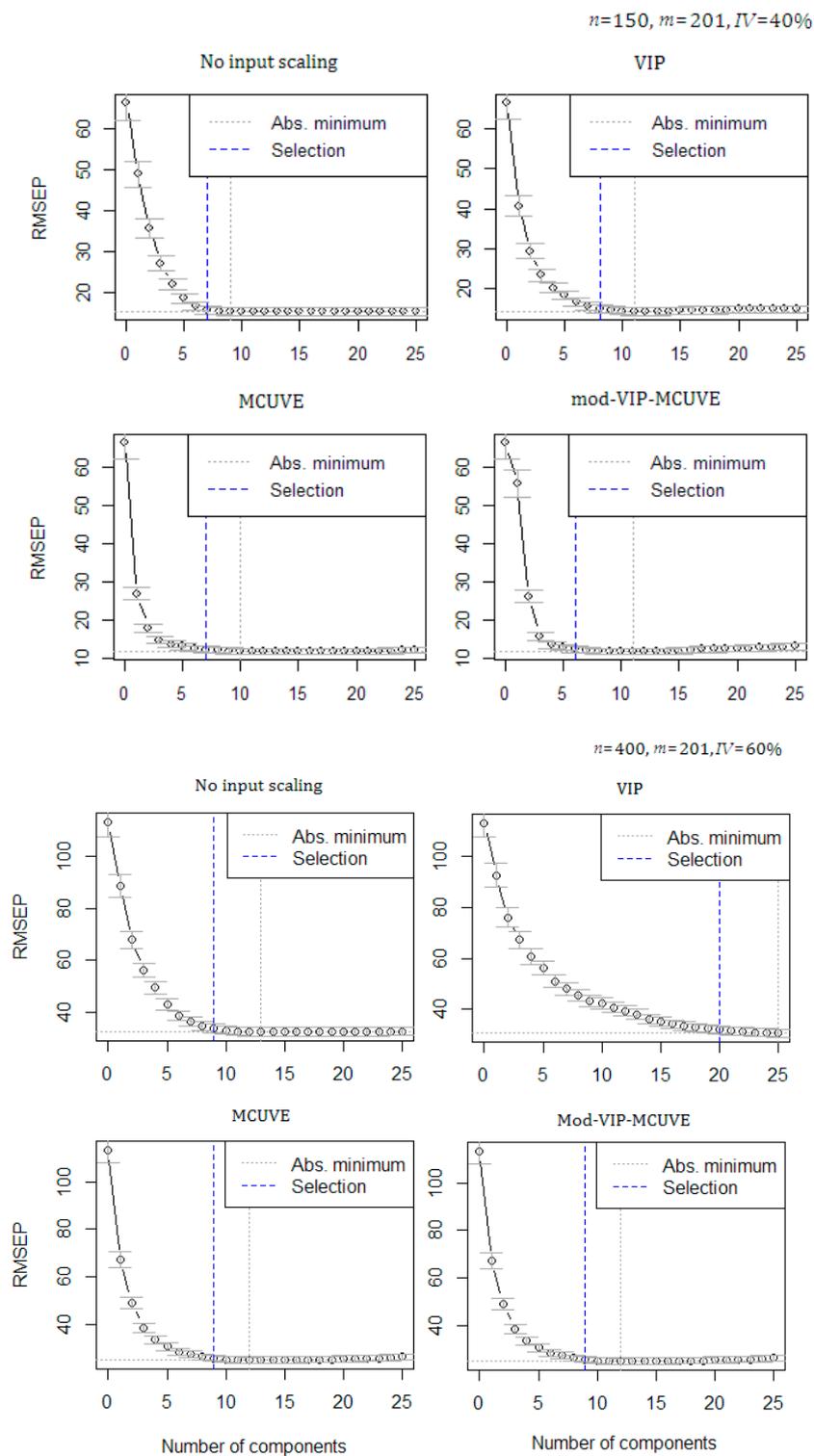


Figure 1. Global minimum cross-validation for the optimum number of PLS components on different dataset scenarios.

Table 1. The RMSE, R^2 , RPD, and SE in variable selection methods using Monte Carlo Simulation with different dataset scenarios.

<i>n</i>	<i>m</i>	<i>IV</i>	Method	<i>n</i> PLS	RMSE	R^2	RPD	SE
40	41	10%	PLSR	5	2.130	0.994	12.824	2.164
			VIP	4	2.524	0.991	10.488	2.567
			MCUVE	2	2.016	0.994	13.133	2.050
			mod-VIP-MCUVE	2	2.025	0.994	13.072	2.060
40	101	10%	PLSR	4	9.715	0.940	4.130	9.839
			VIP	5	9.368	0.944	4.283	9.487
			MCUVE	4	10.705	0.927	3.748	10.841
			mod-VIP-MCUVE	4	4.302	0.988	9.327	4.356
40	201	10%	PLSR	N/A	N/A	N/A	N/A	N/A
			VIP	N/A	N/A	N/A	N/A	N/A
			MCUVE	2	15.843	0.834	2.797	16.044
			mod-VIP-MCUVE	2	6.363	0.970	8.353	6.444
60	41	20%	PLSR	8	4.448	0.974	7.005	4.486
			VIP	6	4.848	0.973	6.322	4.889
			MCUVE	5	5.266	0.967	5.812	5.311
			mod-VIP-MCUVE	5	4.443	0.976	7.873	4.447
60	101	20%	PLSR	5	10.248	0.968	5.601	10.335
			VIP	4	16.726	0.914	3.432	16.867
			MCUVE	7	13.579	0.943	4.227	13.693
			mod-VIP-MCUVE	4	5.143	0.992	11.160	5.187
60	201	20%	PLSR	N/A	N/A	N/A	N/A	N/A
			VIP	N/A	N/A	N/A	N/A	N/A
			MCUVE	2	33.192	0.735	2.544	33.472
			mod-VIP-MCUVE	3	10.678	0.962	11.028	10.768
150	41	40%	PLSR	6	6.359	0.971	5.931	6.381
			VIP	7	7.183	0.963	5.242	7.207
			MCUVE	5	7.097	0.964	5.306	7.121
			mod-VIP-MCUVE	4	6.291	0.971	5.933	6.297
150	101	40%	PLSR	9	8.949	0.979	7.379	8.979
			VIP	8	12.947	0.958	5.047	12.991
			MCUVE	6	11.023	0.970	5.850	11.060
			mod-VIP-MCUVE	5	8.858	0.987	7.761	8.861
150	201	40%	PLSR	5	14.798	0.975	6.406	14.847
			VIP	5	27.199	0.917	3.485	27.290
			MCUVE	5	23.706	0.937	3.999	23.785
			mod-VIP-MCUVE	2	14.809	0.975	5.347	14.852
400	41	60%	PLSR	5	9.424	0.967	5.515	9.436
			VIP	9	10.057	0.962	5.157	10.070
			MCUVE	7	9.991	0.963	5.200	10.003
			mod-VIP-MCUVE	7	9.423	0.969	5.549	9.435
400	101	60%	PLSR	6	14.258	0.972	6.025	14.275
			VIP	9	16.240	0.964	5.290	16.260
			MCUVE	7	15.424	0.968	5.571	15.443
			mod-VIP-MCUVE	7	14.310	0.968	5.611	14.329
400	201	60%	PLSR	10	14.258	0.972	6.025	14.275
			VIP	14	16.240	0.964	5.290	16.260
			MCUVE	10	15.424	0.968	5.571	15.443
			mod-VIP-MCUVE	7	14.310	0.968	5.611	15.319
600	41	80%	PLSR	10	10.789	0.966	5.469	10.798
			VIP	8	11.072	0.965	5.328	11.081
			MCUVE	8	11.139	0.964	5.297	11.147
			mod-VIP-MCUVE	8	10.985	0.967	5.471	10.994
600	101	80%	PLSR	12	16.264	0.969	5.737	16.278
			VIP	10	16.998	0.967	5.495	17.012
			MCUVE	9	17.019	0.967	5.483	17.033
			mod-VIP-MCUVE	9	16.210	0.970	5.781	16.223
600	201	80%	PLSR	12	20.988	0.975	6.305	21.005
			VIP	12	22.276	0.971	5.942	22.295
			MCUVE	8	22.588	0.971	5.855	22.607
			mod-VIP-MCUVE	8	20.295	0.976	5.940	20.314

Note: *n*PLS is a number of optimum PLS components used in the PLSR model.

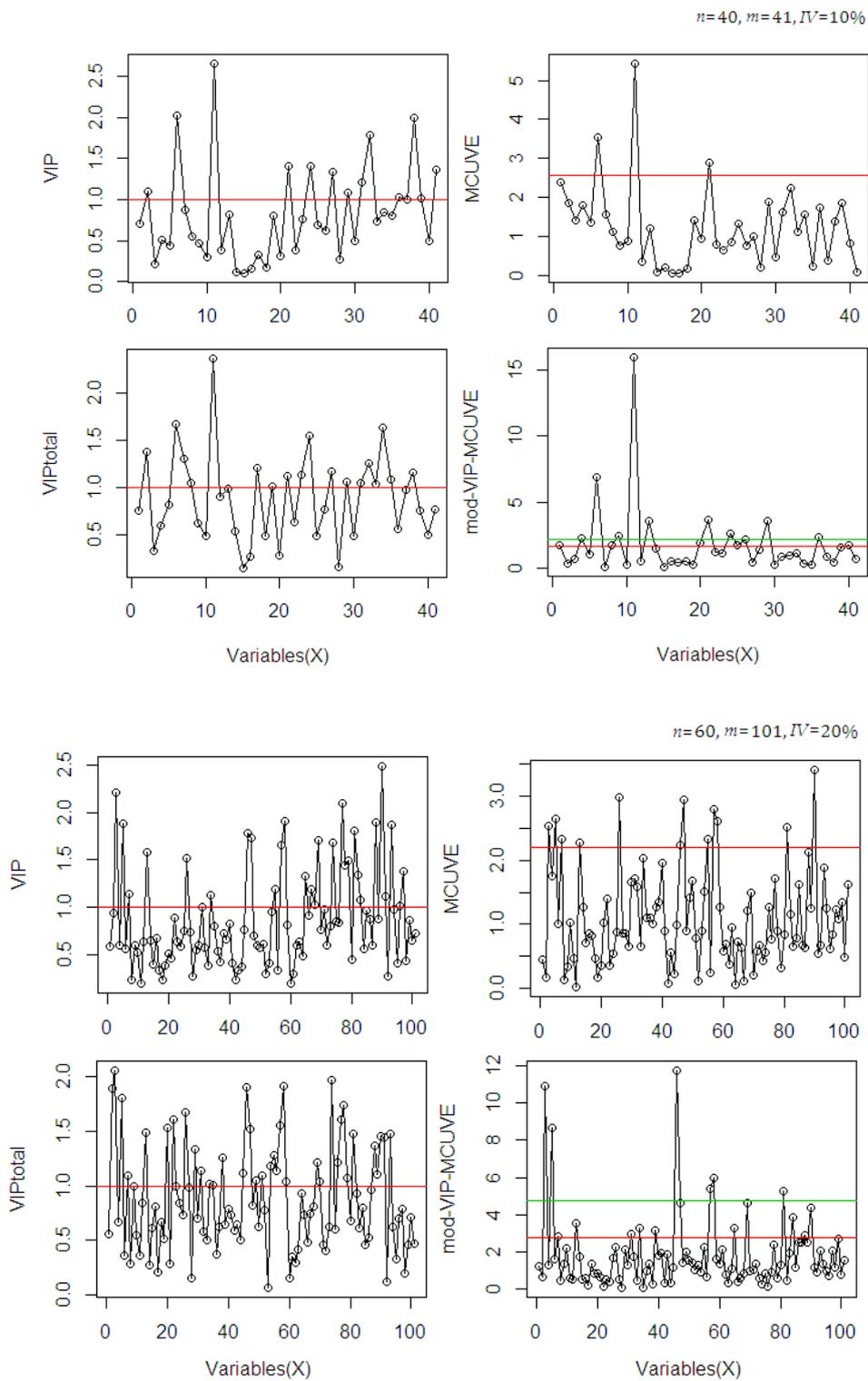


Figure 2. Cont.

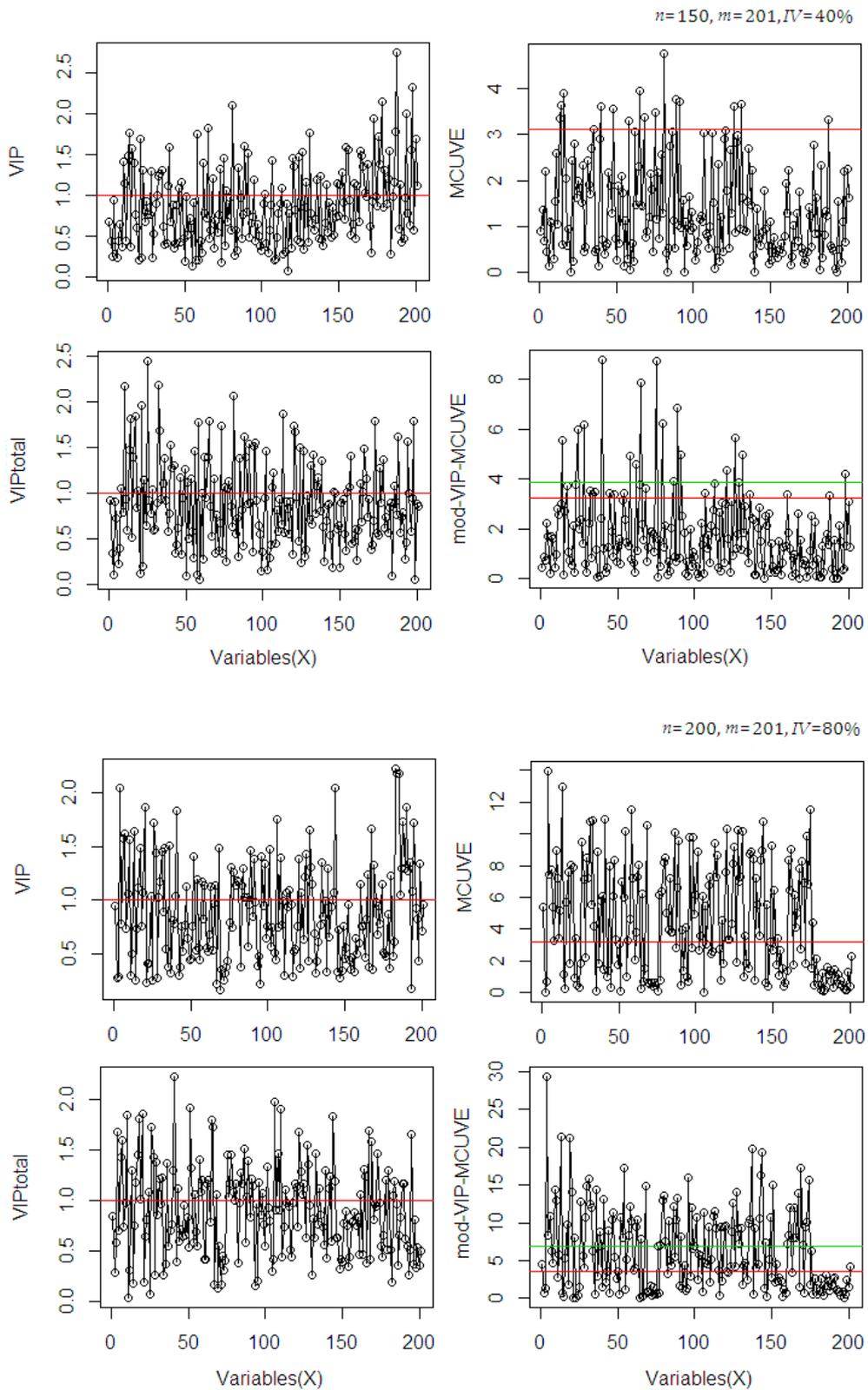


Figure 2. Comparison of the selected relevant variables based on the cut-off criteria in variable selection methods using different dataset scenarios.

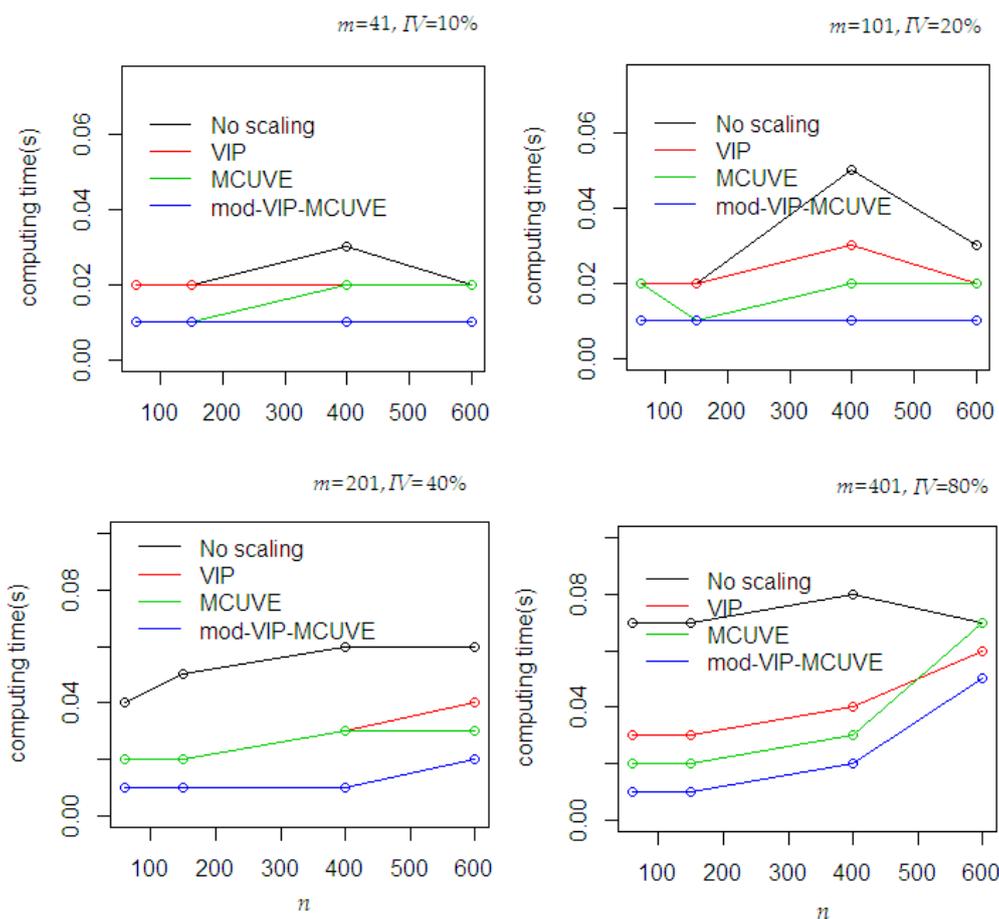


Figure 3. Time-consuming performances between methods during the fitting process using different dataset scenarios (n = number of samples, m = number of predictors, IV = number of important variables).

5. NIR Spectral Dataset

A total of 80 fruit bunches were collected from the site of breeding trial in Palapa Estate, PT. Ivomas Tunggal, Riau Province, Indonesia. The source of variability such as planting material (Dami Mas, Clone, Benin, Cameroon, Angola, Colombia), planting year (2010–2012) and ripeness level (unripe, under ripe, ripe, over ripe) were considered for covering as much as possible of the whole range of potential variation in the palm population. Right after harvest, the bunch samples were sent immediately to the laboratory for spectral measurement and wet chemistry analysis. The fruit mesocarp samples were collected from 12 sampling positions by considering the vertical and horizontal lines in a bunch (see Figure 4): bottom-front, bottom-left, bottom-back, bottom-right, equator-front, equator-left, equator-back, equator-right, top-front, top-left, top-back, and top-right. The spectral measurement was done by scanning (in contact) the oil palm fruit mesocarp using a Portable Handheld NIR spectrometer, QualitySpec Trek, from Analytical Spectral Devices (ASD Inc., Boulder, CO, USA). A dataset of NIR spectral data is shown in Figure 5 then was used in evaluating the proposed method. The spectral data as a result of the light absorbance in each j wavelength bands were adopted from Beer–Lambert Law [38] and presented in $m \times 1$ column vector \mathbf{x}_j using the log base 10.

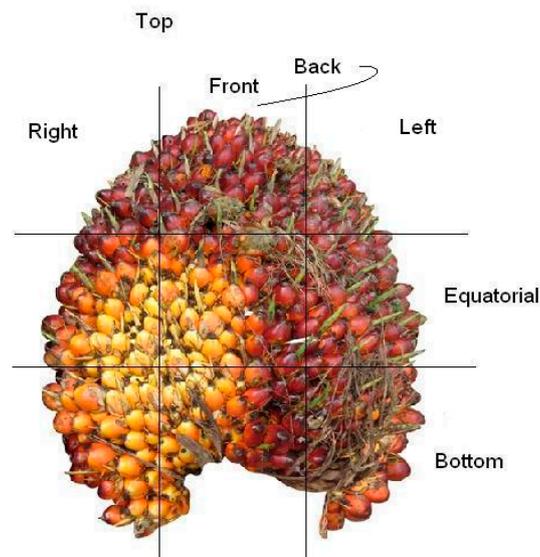


Figure 4. Twelve sampling positions for fruit mesocarp samples of an oil palm fresh fruit bunch.

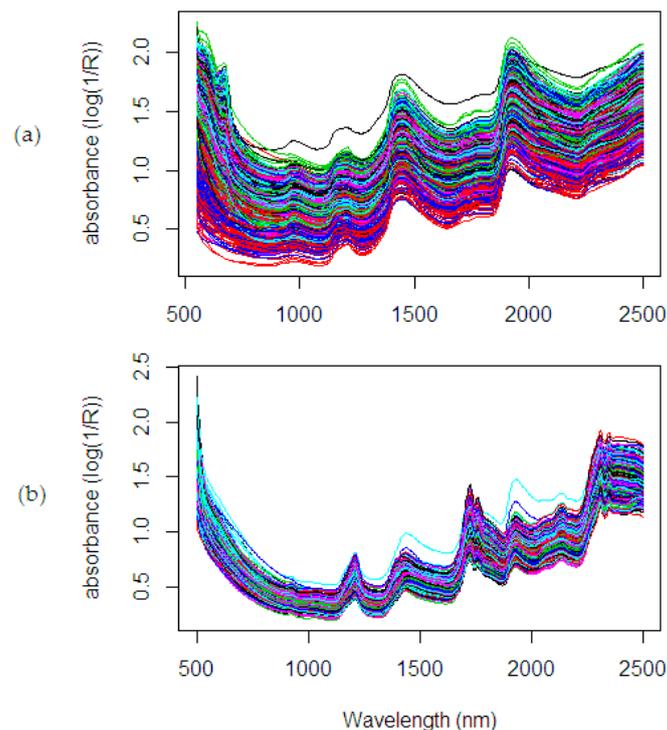


Figure 5. NIR spectra on oil palm fruit mesocarp: (a) fresh mesocarp, (b) dried ground mesocarp.

The spectra collection was measured three times in each fruit mesocarp sample. The averaged spectra were used in the computation (see Figure 5). There are two sample conditions with different parameters observed in this study: fresh fruit mesocarp and dried ground mesocarp. The fresh fruit mesocarp is used to estimate the percentage of Oil to Dry Mesocarp (%ODM) and Oil to Wet Mesocarp (%OWM), while the dried ground mesocarp is used to estimate the percentage of Fat Fatty Acids (%FFA). These parameters were analyzed through wet chemistry analysis using the standard test methods from the Palm Oil Research Institute of Malaysia (PORIM) [39,40]. The %ODM is calculated in dry matter basis, which removes the weight of water content, while the %OWM uses wet matter basis. As seen in Figure 6, the distribution of the %ODM is 56.38–86.9%, the %OWM is 19.75–64.81%,

and the %FFA is 0.17–6.3%. These wide ranges of the distribution showed the possible actual range variation covered in the analysis.

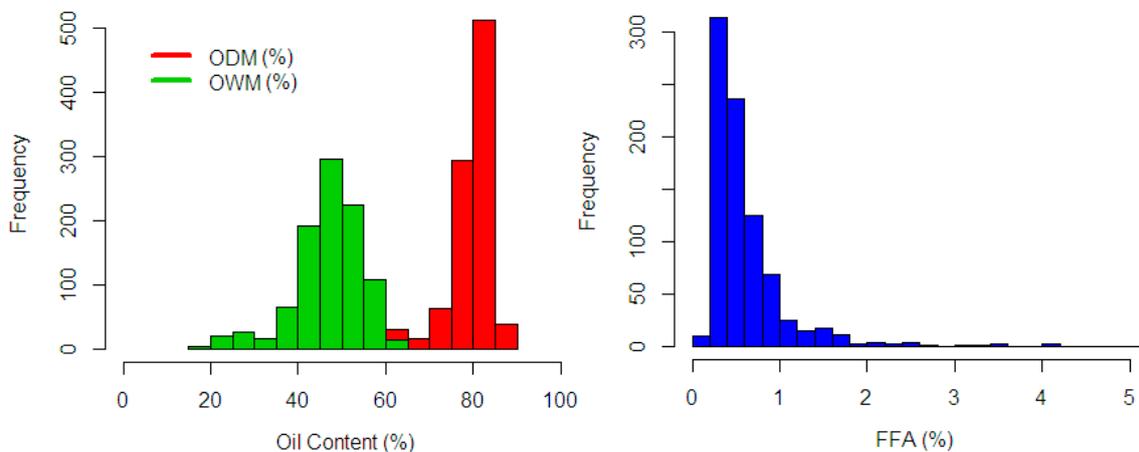


Figure 6. Frequency distribution on response variable: %ODM (red), %OWM (green), and %FFA blue).

5.1. Oil to Dry Mesocarp

In the spectral measurement on fresh fruit mesocarp sample (Figure 5a), each spectrum is composed of 489 wavelengths as data points (range 550–2500 nm: 4 nm interval) with a total spectrum collecting about 960 observations. Here, the importance of the wavelengths is generally unknown, and it needs to be investigated. The selection of the most informative wavelengths in the NIR spectral related to the %ODM in the fresh fruit mesocarp is crucial for further data interpretation. The summary of the fitting performance on the dataset using the calibration model with wavelength selection methods is presented in Table 2.

Table 2. The RMSE, R^2 , RPD, and SE in variable selection methods using %ODM data.

Dataset	Methods	n PLS	RMSE	R^2	RPD	SE
%ODM	PLSR	29	3.267	0.652	1.607	3.269
	VIP	29	3.011	0.657	1.702	3.013
	MCUVE	26	3.107	0.633	1.650	3.108
	mod-VIP-MCUVE	26	3.009	0.659	1.725	3.011

Note: n PLS is a number of optimum PLS components used in the PLSR model.

As seen in Table 2, the proposed mod-VIP-MCUVE is superior to the other methods. Using the auto-scaling method, the classical PLSR shows the worst performance compared to the other methods with wavelength selection and input scaling applied. The proposed mod-VIP-MCUVE and MCUVE use fewer PLS components than the classical PLSR and VIP method in the fitting process. The classical PLSR suffers overfitting due to the higher number of PLS components used in the model. The VIP method has low accuracy since there are many variables removed in the computation (see Figure 7). Unlike in the MCUVE method, the proposed mod-VIP-MCUVE has better performance and more efficient computationally because of a lower number of PLS components (26 PLS) in the model. The proposed method succeeds in highlighting the most relevant wavelengths and downgrades the influence of irrelevant wavelengths. This result confirms the usefulness of the wavelengths selection and input scaling applied in the input variables which leads to faster convergence speed.

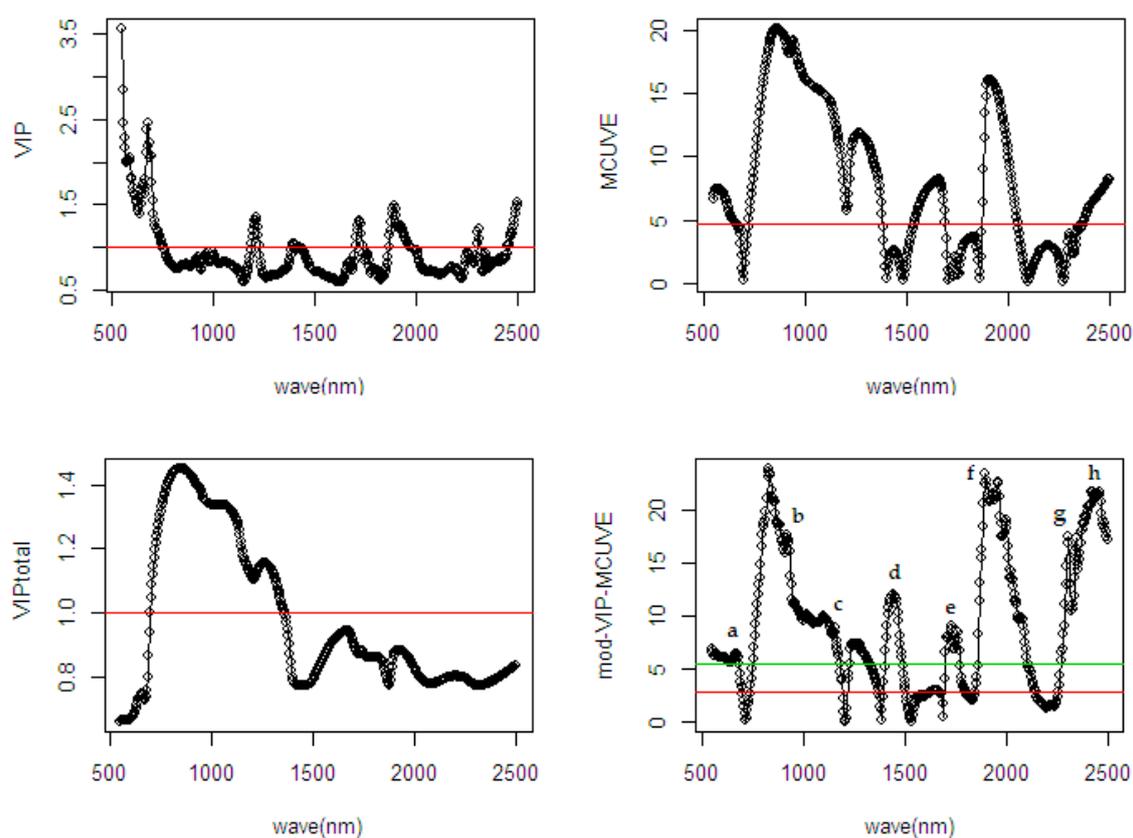


Figure 7. Comparison of selected wavelengths from different wavelength selection methods using spectral data of fresh fruit mesocarp on the %ODM.

In Figure 7, information regarding the relevant wavelengths related to the %ODM is presented. All the variable selection methods selected the same spectral region, which has the most relevant contribution to the response variable. The methods show a different cut-off threshold to remove the irrelevant wavelengths in the regions. It can be observed that the VIP, MCUVE, and VIP-total removed many irrelevant wavelengths in the model. As assumed earlier, the more wavelengths are excluded in the model the lower the accuracy in the prediction result. In the fourth plot of mod-VIP-MCUVE, the green line shows the old cut-off threshold using previous MCUVE (threshold = 5.486), while the red line is the new proposed robust cut-off threshold (threshold = 2.916). The proposed mod-VIP-MCUVE method shows better cut-off threshold since there is only a smaller number of wavelengths indicated as the most irrelevant wavelengths.

The diffuse selected reflectance [38] is important to identify the relevant wavelengths related to the %ODM. This exhibits their fundamental attribute to the overtone or combination bands involving the molecular stretching and bending absorption over a wide spectral range. The main absorption in the NIR spectral range is produced by the combination and overtone of C-H, O-H, N-H, and C=O groups. The relevant wavelength ranges are indicated through lowercase alphabet notation in the graphic. Based on Figure 7, it is feasible to observe that the well-defined absorption bands are from visible red color (a: 668–684 nm), CH₂ of oil and O-H of water (b: 936–961 nm), C-H absorption by stretching-bending (c: 1232–1344 nm), first overtone (d: 1404–1444) of C-H stretch O-H of water and C-H of oil and its combinations (e: 1700–1776) of C-O oil and C-H stretching by first overtone, C=O absorption by stretching-bending (f: 1888–2008 nm), C-H second overtone of protein and oil (g: 2296–2360 nm), and corresponding to absorption which associated with the second overtone in C-H of oil (h: 2364–2496 nm).

5.2. Oil to Wet Mesocarp

Using a similar set of NIR spectral data from fresh mesocarp as in Section 5.1, the %OWM is used as a response variable. As indicated in Figure 6, the variability of the water content has impacted the shifting of distribution in the response variable. The summary of the fitting performance on the prediction results using different variable selection treatment is presented in Table 3. The comparison shows that the proposed mod-VIP-MCUVE could achieve superior performance than the other methods. It offers the accuracy improvement of RMSE and R^2 to the VIP, MCUVE, and PLSR methods. Moreover, the number of selected optimal PLS components used in the calibrated model of the proposed mod-VIP-MCUVE is the smallest (17 PLS) which indicates that important information could still be attained even by using fewer variables. This also has proven the necessity to accomplish the wavelength selection before fitting the calibration model.

Table 3. The RMSE, R^2 , RPD, and SE in variable selection methods using %OWM data.

Dataset	Methods	nPLS	RMSE	R^2	RPD	SE
%OWM	PLSR	20	4.558	0.651	1.693	−0.067
	VIP	20	4.506	0.659	1.713	0.062
	MCUVE	18	4.461	0.666	1.730	−0.016
	mod-VIP-MCUVE	17	4.400	0.675	1.754	0.060

Note: nPLS is a number of optimum PLS components used in the PLSR model.

In Figure 8, the cut-off threshold in VIP and MCUVE has minimized many wavelengths which forces much important information to become lost. The proposed mod-VIP-MCUVE and new proposed cut-off (red line) succeeds to minimize only the most irrelevant variable and to retain rich information in the spectra. Using old cut-off (green line = 2.881), it has minimized a greater number of wavelengths than the new cut-off (red line = 4.270). From the highlighted wavelength information (mod-VIP-MCUVE) in Figure 8, it is possible to identify the essential molecules in raw NIR spectral and its chemical association to the parameter %OWM. The most relevant wavelengths are O-H of water and CH_2 of oil (a: 890–950 nm), C-H second overtone and C=O stretch fourth overtone (b: 1140–1166 nm), first and second overtone of the C-H stretching mode (c: 1326–1370 nm), N-H stretch/C-H stretch first overtone (d: 1558–1650 nm), C-H stretch O-H of water and CH of oil and its combinations (e: 1782–1802 nm), O-H stretch/C-O stretch second overtone combination (f: 1830–1854 nm), and C-H stretch/C=O stretch combination of protein and oil (g: 2126–2184 nm).

5.3. Fat Fatty Acids

Another NIR spectral dataset (Figure 5b) using a total of 839 observations and 500 wavelengths (in the range 500–2500 nm: 4 nm interval) were collected from dried ground mesocarp sample. The importance of the wavelengths related to the %FFA is unknown. The summary of the fitting performance on the dataset using the calibration model with wavelength selection methods is presented in Table 4.

Table 4. The RMSE, R^2 , RPD, and SE in variable selection methods using %FFA data.

Dataset	Methods	nPLS	RMSE	R^2	RPD	SE
%FFA	PLSR	28	0.270	0.730	1.924	0.271
	VIP	28	0.266	0.734	1.932	0.267
	MCUVE	26	0.264	0.736	1.946	0.265
	mod-VIP-MCUVE	25	0.265	0.735	1.940	0.266

Note: nPLS is a number of optimum PLS components used in the PLSR model.

Similar to previous results, the proposed mod-VIP-MCUVE is superior to the classical VIP and the classical PLSR. The performance of the MCUVE is comparable to the mod-VIP-MCUVE. However,

the proposed mod-VIP-MCUVE uses less PLS components (25 PLS) than MCUVE, the classical PLSR, and the VIP method in the fitting process. These have proven its efficiency in the computation. The classical PLSR is still suffering overfitting due to a higher number of PLS components used in the model. The VIP method also has low accuracy since there are many variables removed in the computation (see Figure 9). The MCUVE shows its performance is better than the classical PLSR and VIP; however, it is a non-robust method. Again, the proposed mod-VIP-MCUVE method has succeeded in highlighting the most relevant wavelengths and downgraded the influence of irrelevant wavelengths. These results confirmed the usefulness of the wavelengths selection method in the input scaling strategy to improve the classical PLSR model.

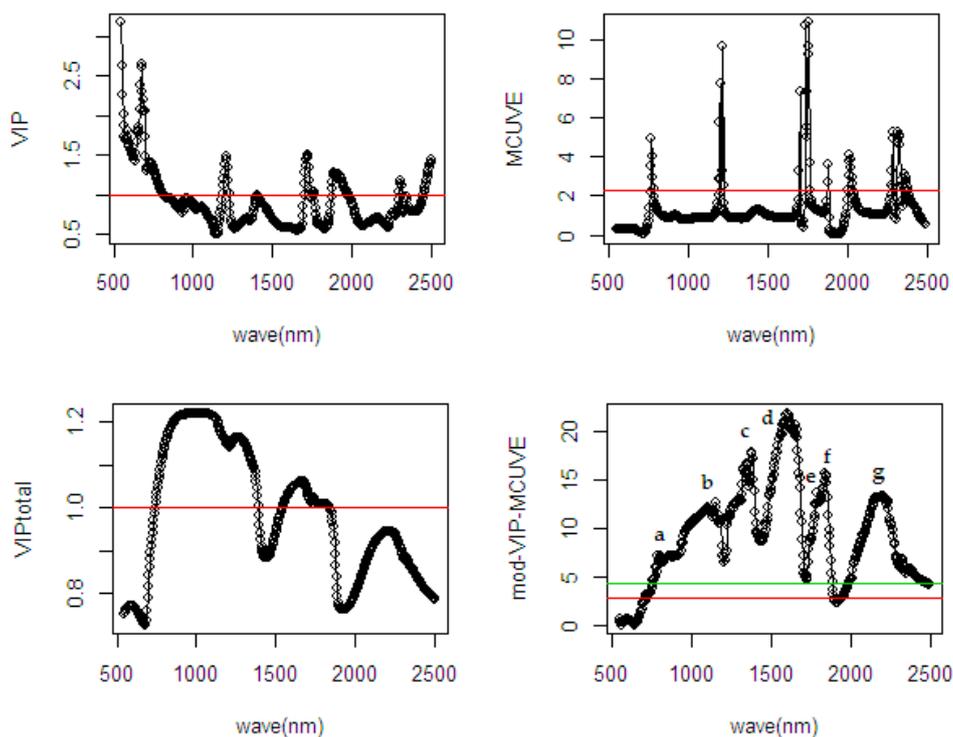


Figure 8. Comparison of selected wavelengths from different wavelength selection methods using spectral data of fresh fruit mesocarp on the %OWM.

As shown in Figure 9, the cut-off threshold in the mod-VIP-MCUVE and MCUVE has succeeded in removing only the most irrelevant wavelengths and keeping the remaining of the relevant variables in the model. The VIP and the VIP-total has removed many variables in the model, and this yields the less relevant variables also included in the removal process. Hence, the less important variables were reduced in the fitting process. As shown in the fourth plot of the mod-VIP-MCUVE method, the old cut-off threshold using previous MCUVE represented as the green line is 8.071, while the new proposed cut-off threshold represented as the red line is 3.039. Based on these thresholds, the number of the irrelevant variables indicated in the mod-VIP-MCUVE is still less than the MCUVE.

The fundamental attribute of diffuse selected reflectance related to the %FFA is crucial to be investigated. The mod-VIP-MCUVE method is used to interpret the selected wavelength, which has well-defined absorption both in the visible and NIR regions. In Figure 9, the most relevant wavelengths are C=O stretch fourth overtone and C-H second overtone (a: 1176–1220 nm), O-H first overtone and C-H of oil (b: 1408–1436 nm), absorption by stretching and bending of C-H stretch first overtone (c: 1652–1688 nm), O-H stretch/C-O second overtone and C=O stretch second overtone (d: 1820–1896 nm), N-H second overtone of protein, absorption by stretching-bending N-H/Amide and C=O stretch second overtone (e: 1996–2040 nm), O-H bend/C-O stretch and C-O-O stretch third overtone, C-H stretch/C=O oil (f: 2116–2288 nm), C-H stretch and C-O of oil (g: 2324–2344 nm), and corresponding

to absorption which associated with the second overtone in C-H and C-N-C stretch first overtone of protein (h: 2456–2500 nm).

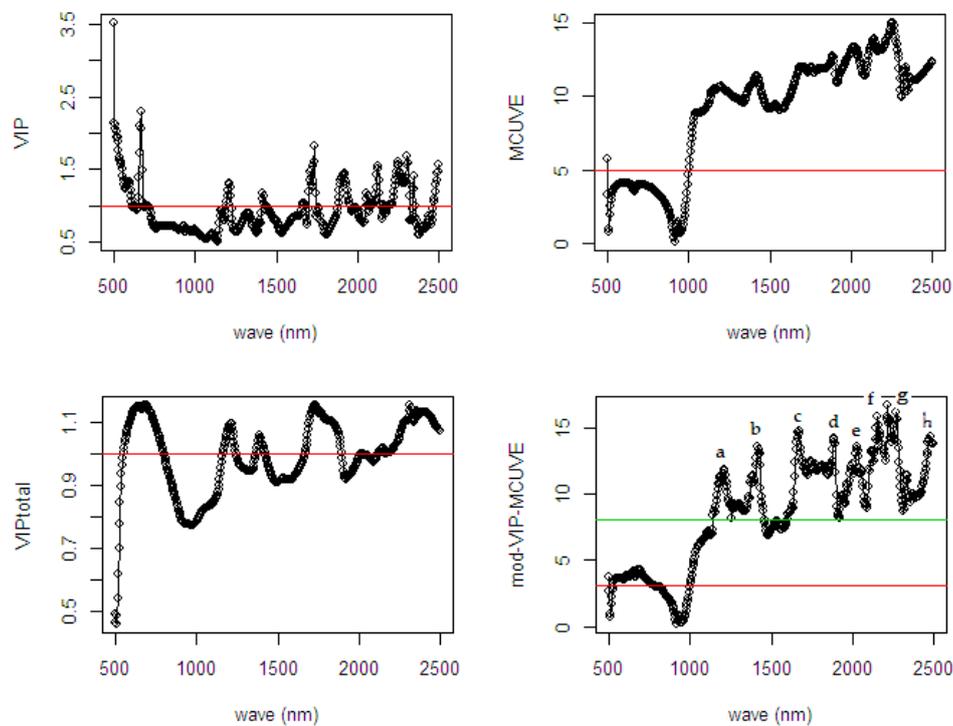


Figure 9. Comparison of selected wavelengths from different wavelength selection methods using spectral data of ground dried mesocarp on the %FFA.

6. Conclusions

In summary, the NIR spectral region contains rich and abundant information that warrants further interpretation using advanced chemometric techniques. The study has shown that the wavelength selection using input scaling method strategy to be promising, particularly with the application on a high dimension dataset such as the NIRS spectral data. The proposed method is robust since it uses robust reliability weight procedure to rescale the original input matrix. According to the evaluation indices, the proposed mod-VIP-MCUVE method confirmed its superiority to the classical PLSR, the VIP method, and the MCUVE method. Using the modified robust cut-off threshold, the proposed method succeeds to highlight the irrelevant wavelengths in the model. Moreover, the proposed method also has the benefit to reduce the data dimension and to improve the model accuracy and computational complexity. The proposed method has even investigated successfully the fundamental attribute of diffuse selected reflectance of the NIRS spectral absorption. This is essential in the improvement of the chemical interpretability. Further, the input scaling procedure using robust selection procedure on optimum number of PLS component is expected to get better improvement in the computational complexity.

Author Contributions: Conceptualization and methodology: D.D.S., H.M., J.A., M.S.M., J.-P.C.; Data Collection: D.D.S., H.M., J.-P.C.; Computational and Validation: H.M., J.A., M.S.M.; First draft preparation: D.D.S., H.M.; Writing up to review and editing: D.D.S., H.M., J.A., M.S.M., J.-P.C. All authors have read and agreed to the published version of the manuscript.

Funding: The present research was partially supported by the Universiti Putra Malaysia Grant under Putra Grant (GPB) with project number GPB/2018/9629700.

Acknowledgments: This work was supported by a research grant and scholarship from the Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA). We are also grateful to SMARTRI, PT. SMART TBK for providing the portable handheld NIRS instrument, research site, and analytical laboratory services. We would like to thank Universiti Putra Malaysia for the journal publication fund support. Special thanks are also extended to all research staff and operators of SMARTRI for their cooperation and outstanding help with data collection.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schowengerdt, R.A. *Remote Sensing Models and Methods for Image Processing*; Academic Press: Cambridge, MA, USA, 1997.
2. Hourant, P.; Baeten, V.; Morales, M.T.; Meurens, M.; Aparicio, R. Oil and Fat Classification by Selected Bands of Near-Infrared Spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 1168–1174. [[CrossRef](#)]
3. Kasemsumran, S.; Thanapase, W.; Punsuvon, V.; Ozaki, Y. A Feasibility Study on Non-Destructive Determination of Oil Content in Palm Fruits by Visible–Near Infrared Spectroscopy. *J. Near Infrared Spectrosc.* **2012**, *20*, 687–694. [[CrossRef](#)]
4. Chong, I.-G.; Jun, C.-H. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103–112. [[CrossRef](#)]
5. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [[CrossRef](#)]
6. Mehmood, T.; Martens, H.; Sæbø, S.; Warringer, J.; Snipen, L. A Partial Least Squares based algorithm for parsimonious variable selection. *Algorithms Mol. Biol.* **2011**, *6*, 27. [[CrossRef](#)]
7. Wang, S.; Tang, J.; Liu, H. Feature selection. In *Encyclopedia of Machine Learning and Data Mining*; Springer Science + Business Media: New York, NY, USA, 2016; pp. 1–9.
8. Reunanen, J. Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res.* **2003**, *3*, 1371–1382.
9. Andersen, C.M.; Bro, R. Variable selection in regression—A tutorial. *J. Chemom.* **2010**, *24*, 728–737. [[CrossRef](#)]
10. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
11. Hira, Z.M.; Gillies, D.F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv. Bioinform.* **2015**, *2015*. [[CrossRef](#)]
12. Kokaly, R.F.; Clark, R.N. Spectroscopic Determination of Leaf Biochemistry Using Band-Depth Analysis of Absorption Features and Stepwise Multiple Linear Regression. *Remote Sens. Environ.* **1999**, *67*, 267–287. [[CrossRef](#)]
13. Gidskehaug, L.; Anderssen, E.; Flatberg, A.; Alsberg, B.K. A framework for significance analysis of gene expression data using dimension reduction methods. *BMC Bioinform.* **2007**, *8*, 346. [[CrossRef](#)] [[PubMed](#)]
14. Wu, W.; Walczak, B.; Massart, D.; Prebble, K.; Last, I. Spectral transformation and wavelength selection in near-infrared spectra classification. *Anal. Chim. Acta* **1995**, *315*, 243–255. [[CrossRef](#)]
15. Oussama, A.; Elabadi, F.; Platikanov, S.; Kzaiber, F.; Tauler, R. Detection of Olive Oil Adulteration Using FT-IR Spectroscopy and PLS with Variable Importance of Projection (VIP) Scores. *J. Am. Oil Chem. Soc.* **2012**, *89*, 1807–1812. [[CrossRef](#)]
16. Palermo, G.; Piraino, P.; Zucht, H.-D. Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Adv. Appl. Bioinform. Chem.* **2009**, *2*, 57–70. [[CrossRef](#)] [[PubMed](#)]
17. Wold, S.; Johansson, E.; Cocchi, M. PLS—Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers. B. V.: Leiden, The Netherlands, 1993; pp. 523–550.
18. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119–128. [[CrossRef](#)]
19. Galindo-Prieto, B.; Eriksson, L.; Trygg, J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *J. Chemom.* **2014**, *28*, 623–632. [[CrossRef](#)]
20. Leardi, R.; González, A.L. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195–207. [[CrossRef](#)]

21. Centner, V.; Massart, D.-L.; De Noord, O.E.; De Jong, S.; Vandeginste, B.M.; Sterna, C. Elimination of Uninformative Variables for Multivariate Calibration. *Anal. Chem.* **1996**, *68*, 3851–3858. [[CrossRef](#)]
22. Forina, M.; Casolino, C.; Pizarro Millan, C. Iterative predictor weighting (IPW) PLS: A technique for the elimination of useless predictors in regression problems. *J. Chemom. Soc.* **1999**, *13*, 165–184. [[CrossRef](#)]
23. Guzmán, E.C.; Baeten, V.; Pierna, J.A.F.; García-Mesa, J.A. Application of low-resolution Raman spectroscopy for the analysis of oxidized olive oil. *Food Control.* **2011**, *22*, 2036–2040. [[CrossRef](#)]
24. Wang, Z.X.; He, Q.P.; Wang, J. Comparison of variable selection methods for PLS-based soft sensor modeling. *J. Process. Control.* **2015**, *26*, 56–72. [[CrossRef](#)]
25. Cai, W.; Li, Y.; Shao, X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 188–194. [[CrossRef](#)]
26. Kim, S.; Okajima, R.; Kano, M.; Hasebe, S. Development of soft-sensor using locally weighted PLS with adaptive similarity measure. *Chemom. Intell. Lab. Syst.* **2013**, *124*, 43–49. [[CrossRef](#)]
27. Kim, J.; Kiss, B.; Lee, D. An adaptive unscented Kalman filtering approach using selective scaling. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016; pp. 784–789.
28. Wold, H. *Multivariate Analysis*; Krishnaiah, P.R., Ed.; Academic Press: New York, NY, USA, 1973; Volume 3, pp. 383–407.
29. Martens, H.; Naes, T. Multivariate calibration. *Math. Soc.* **1992**, *68*, 337–404.
30. Han, Q.-J.; Wu, H.-L.; Cai, C.-B.; Xu, L.; Yu, R.-Q. An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. *Anal. Chim. Acta* **2008**, *612*, 121–125. [[CrossRef](#)]
31. Gelman, A. Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* **2008**, *27*, 2865–2873. [[CrossRef](#)]
32. Saccenti, E.; Westerhuis, J.A.; Smilde, A.K.; Van Der Werf, M.J.; Hageman, J.A.; Hendriks, M.M.W.B. Simplivariate Models: Uncovering the Underlying Biology in Functional Genomics Data. *PLoS ONE* **2011**, *6*, e20747. [[CrossRef](#)]
33. Vandeginste, B.G.M.; Massart, D.L.; De Jong, S.; Massaart, D.L.; Buydens, L.M.C. *Handbook of Chemometrics and Qualimetrics: Part B*; Elsevier BV: Amsterdam, The Netherlands, 1998.
34. Bakeev, K.A. *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
35. Orso, A.; Shi, N.; Harrold, M.J. Scaling regression testing to large software systems. *ACM SIGSOFT Softw. Eng. Notes* **2004**, *29*, 241. [[CrossRef](#)]
36. Berg, R.A.V.D.; Hoefslot, H.C.J.; Westerhuis, J.A.; Smilde, A.K.; Van Der Werf, M.J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genom.* **2006**, *7*, 142. [[CrossRef](#)]
37. Natrella, M.G. *Experimental Statistics Handbook 91*; US Government Printing Office: Washington, DC, USA, 1963.
38. Stuart, B. *Infrared Spectroscopy: Fundamentals and Applications*; Wiley: Hoboken, NJ, USA, 2004; p. 78.
39. Siew, W.L.; Tan, Y.A.; Tang, T.S. *Methods of Test for Palm Oil and Palm Oil Products: Compiled*; Lin, S.W., Sue, T.T., Ai, T.Y., Eds.; Palm Oil Research Institute of Malaysia: Selangor, Malaysia, 1995.
40. Rao, V.; Soh, A.C.; Corley, R.H.V.; Lee, C.H.; Rajanaidu, N. Critical Reexamination of the Method of Bunch Quality Analysis in Oil Palm Breeding. PORIM Occasional Paper. 1983. Available online: <https://agris.fao.org/agris-search/search.do?recordID=US201302543052> (accessed on 1 September 2020).

