OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Establish thecalibrationmodelfora portable near-infrared (NIR) spectroscopydevicetodeterminethe total carbohydratecontent in food

Ngoc Lieu Le[1*]

[1]*Lecturer of Department of Food Technology, International University, Vietnam National University inHoChiMinh City, Quarter 6, LinhTrung Ward, Thu Duc District, HoChiMinh City, Vietnam.*
[*]*CorrespondingAuthor: Ngoc Lieu Le*

ABSTRACT:*This study was todevelop a calibrationmodelfor a portable near-infrared (NIR) spectroscopydevicetodeterminethe total carbohydratecontent in processedfoods. NIR spectra in the 740 - 1070 nmregionof 62 foodproductswereanalyzed. The resultsprovidednewinsightsintotheshort-wavelength NIR spectroscopy. The modelbetweenthe total carbohydratecontentmeasuredby traditional methodsandthespectraldatacollectedfromthe NIR device was builtwiththe partial least square (PLS) algorithmcombinedwiththecross-validation techniquebytheSciOlabsoftware (Consumer Physics, USA). The effectsofnumberof latent variables (LVs) on themodelfitting was also investigated. The resultsshowedthatwith LVs = 12 a high correlationcoefficient $R^2$of 0.952 and a lowrootmeansquarederror (RMSE) of 2.230 wereobtained, indicatingthegoodfittingofthemodel. The performancevaluesachieved in thisstudyindicatedthatthepossibilityofthe NIR methodtobecome a popularoneforthe fast determinationof total carbohydratecontent in foodisfully promising.*

KEYWORDS: *Near-infrared spectroscopy, NIR, SCiO, partial least square, total carbohydrate content, calibration*

Date of Submission:08-03-2019 Date of acceptance: 28-03-2019

I. INTRODUCTION

Near-infrared (NIR) spectrum is a type of electromagnetic spectrum which corresponds to the wavelength range of 750 to 2500 nm. This is the spectral region that exists between the visible spectral range and the mid infrared region. Near-infrared spectroscopy (NIRS) is based on emission, reflection and diffuse-reflection of light across its entire wavelength range [1].

The near-infrared spectroscopy has been proven to be a great method for the proximate composition analysis due to its many advantages such as fast, accurate, reliable, low-cost, sample non-destructive method and small requirement of sample preparation. The combination of these characteristics with the development of its advanced devices and data treatment methods has made the analytical methods resulted from the application of the NIRS become more popular and have found its widespread use in quick analyses for bulk components such as water, protein, lipid, carbohydrate, etc. [2]. Moreover, this technique has shown its analytical potential of NIR in food quality analysis, food engineering, medical science, environmental science and on-line monitoring [3].

Many studies have yielded promising results of applying the NIR method for qualitative and quantitative purposes in the food industry. In the study by Büning-Pfaue et al. (1997) [4], the performance values achieved by the NIR spectroscopy indicated that the accuracy of NIRS analysis was comparable to those of reference methods. They suggested that in the quantitative determination of the fat, crude protein and carbohydrate contents, the NIR method may be used as a replacement for conventional and time-consuming wet chemical analyses. Potatoes and "consumable meals" were used as samples in this research and the spectra were recorded in the near infrared region within 1100-2500 nm. Data were analyzed using the ISI software and the building model with a very low standard error of prediction (SEP) value for carbohydrate content in “consumable meals” confirmed the quality of this analysis.

Moreover, the short-wave NIR region (wavelengths ranging from 700 to 1100 nm) was used by Sasic& Ozaki (2001) [5] in quantitative analysis of raw milk's main components like fat and protein content by partial least squares regression. A total of 100 milk samples were used in this research. The fat and protein contents were analyzed by Milko-scan 134A/B (N Foss Electric). The NIR spectra were recorded by a NIR system 6500 spectrometer in the 800 - 1100 nm region with a step size of 2 nm. After that, all set of data were evaluated by the Unscrambler software version 6.1 (CAMO AS, Trondheim, Norway). The optimum number of components to be used in the regression was automatically determined by the software.

Later in 2013, Chen et al. [6] used the NIR spectroscopy to determine the protein, total carbohydrate and crude fat contents of foxtail millet. By using the Unscrambler software with PLS algorithm for spectral preprocessing, the building model for total carbohydrate content prediction was accepted with an R^2 of 0.93 and a low root mean square error of cross-validation value (RMSECV) (0.67g/100g). Consequently, the model was considered as the optimal model and chosen for total carbohydrate analysis. The authors also stated that the models of protein and total carbohydrate concentrations illustrated good accuracies and concluded that NIR spectroscopy was successfully utilized for the nondestructive determination of protein, total carbohydrate and crude fat contents in foxtail millet.

This project aims to build a calibration model for a NIR device that can perform quantitative analysis of carbohydrate content in food. With the model integrated in the NIR device, its users can quickly analyze the food in the market to evaluate their carbohydrate content. Previous studies often focused on the analysis of individual products (e.g. yogurt, milk, beef trimmings, etc.). In this research, a wide range of possibly applicable materials were covered, including noodles, rice vermicelli, *pho*and so on.

II. MATERIAL AND METHODS

2.1 Materials and sample preparation

Different processed foods which contained various carbohydrate contents were obtained from the local supermarket (Ho Chi Minh city). They included variety of noodles from different brands (BinhTay, Bich Chi, Coop Mart, Safoco and Phu Huong). A set of 8 different products (as shown in Table 1) with the total number of 62 samples were chosen to build the calibration models.

Table 1.Number of samples per product for building the calibration models

PRODUCTS	NUMBER OF SAMPLES
Rice vermicelli (BINH TAY)	8
Unpolished rice vermicelli (BICH CHI)	11
Pandan leaf rice vermicelli (COOP MART)	10
Vegan noodle (COOP MART)	5
Cellophane noodle (COOP MART)	9
Egg noodle (SAFOCO)	6
Mung bean vermicelli (PHU HUONG)	6
Pho (BICH CHI)	7
Total	**62**

2.2Chemicals

All reagents and solvents including hexane, sodium hydroxide (NaOH), sulfuric acid (H_2SO_4), boric acid (H_3BO_3), potassium sulfate (K_2SO_4), copper sulfate ($CuSO_4$) and Tashiro indicator used in the experiment were of analytical grades.

2.2 NIR device

A SCiO NIR device (Consumer Physics – USA) with the wavelength from 740 to 1070 nm was used. This wavelength range belongs to the short-wave near infrared region. The short-wave NIRS technique was used in this research because this region is suitable for nondestructive analyses of biological materials. The transmittance of light in this region is high, so this region is available for designing an excellent detector [7].

2.4Experimental design

2.4.1. Numerical data collection

In order to obtain the numerical data for the calibration models and testing models, the food samples were analyzed by the traditional methods. For the preparation of powdered samples, the food samples were firstly dried, ground in a grinder and then sieved by a sieving machine to obtain the powder with an average diameter of 500 μm. After that, all powdered samples were carefully stored in airtight poly-ethylene plastic bags and put in a desiccator until further analyses. All experiments were done in triplicate.

2.4.2. Spectral data collection

In order to obtain the spectral data for the calibration models and testing models, the food samples were analyzed by the NIR methods by using a SCiO device (Consumer Physics). Each type of food was divided into small portions. Each sample was weighed and recorded. Subsequently, the samples were boiled at different intervals of time, which were randomly generated. The sample was then placed in a plane surface and the absorbance of each sample was measured at a wavelength in the near-infrared range of 740 – 1070 nm. The SCiO sensor was placed at the distance of 1 cm away from the samples. The primary spectral data of each sample after scanned 5 times were averaged. All spectral data were stored in a computer and analyzed using the SCiO Lab software (Consumer Physics, USA). The weights of samples were measured one more time. All experiments were conducted at room temperature.

2.5Analytical methods

Protein content was determined by the method of AACC International Approved method 46-10.01. Lipid content was determined by the Soxhlet method. Ash content was measured by the method of AACC International Approved 08-01.01. Moisture content was determined by the method of AACC International Approved Method 44-15.02. Total carbohydrate content is the subtraction of 100 percent and the sum of percent of moisture, ash, protein and lipid.

2.6Statistical analysis

All experiments for obtaining numerical data were performed in triplicate. Data was expressed as the mean values derived from triplicate determinations. The results were significant differences for $p < 0.05$. The statistical analysis of the spectral data was done by using the SCiO Lab Software.

2.7NIR analysis

2.7.1. Spectral preprocessing

In this study, the absorbance data were stored as log (1/R) (R = reflectance). Before developing predictive equations, the scatter correction methods which include first derivative with a 2^{nd}-order polynomial and a 35-point window [8] and subtract average method were performed to smooth the obtained spectra and subsequently differentiate them.

2.7.1. Calibration model establishing

Spectral prediction models were built using partial least square regression (PLSR) [9]. PLS is a multivariate method used to establish the relationship between the independent variable X and the dependent variable Y, from which it is possible to predict the Y result when the information of X is known and vice versa. According to Haalan& Thomas (1988) [10], partial least squares method could be based on the relationship between the signal matrix (spectral data X) and the properties of the sample (variable Y) (in this study Y is the total carbohydrate content).

Before modeling, spectral data were randomly divided into calibration, validation and test subsets using repeated 10-fold cross-validation technique. Cross-validation was employed to avoid overfitting of the model [11].

In order to improve predictive power, in addition, determination the number of latent variables was also carried out. This process was used to shorten a set of many variables that measure interdependence into a less variable set (called latent variables) to make them more meaningful but still contain most of the original variable information content. Based on statistical theory and simulation studies, there is a rule that fewer than three items per latent variable is inadequate [12].

The root mean square error (RMSE) and coefficient of determination for calibration R^2 were used to decide the number of LVs. Ziegel (2004) [13] stated that the critical criteria for evaluating a predictive regression model are based on the RMSE value, where the lower the RMSE value, the higher the accuracy of the model. Root mean square error (RMSE) measures how much error there is between two data sets. In other words, it compares a predicted value and a reference value. The RMSE of calibration was calculated as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{1_i})^2}{n - LVs - 1}}$$

where n, LVs, y_i, y_{1i} are the number of samples and latent variables, the predicted values and the reference values, respectively.

III. RESULTS AND DISCUSSIONS

3.1Distribution of total carbohydrate content in studied samples

The distribution of total carbohydrate content in calibration samples is revealed in Figure 1 and Table 2. Figure 1 shows that total carbohydrate contents of all samples cover a wide range of 20.59% to 68.99%. Table 2 summarizes the maximum and minimum values, mean and standard deviation for the value traits of the calibration samples. The high standard deviation (10.40%) demonstrates the suitability of the sample set in building the calibration models.

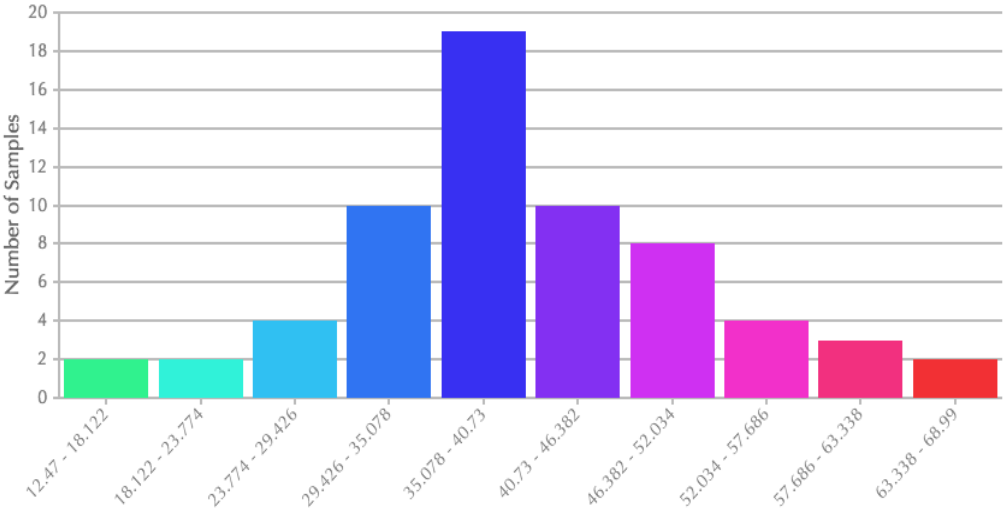

Figure 1. Total carbohydrate content distribution of 62 samples

Table 2. Statistical parameters of total carbohydrate content of samples

Min	20.59
Max	68.99
Mean ± SD	41.05 ± 10.40

3.2Near-infrared spectroscopy analysis

The raw spectra of 62 samples obtained from the portable NIR device, which was ranged from the wavelength of 740 to 1070 nm, are shown in Figure 2.

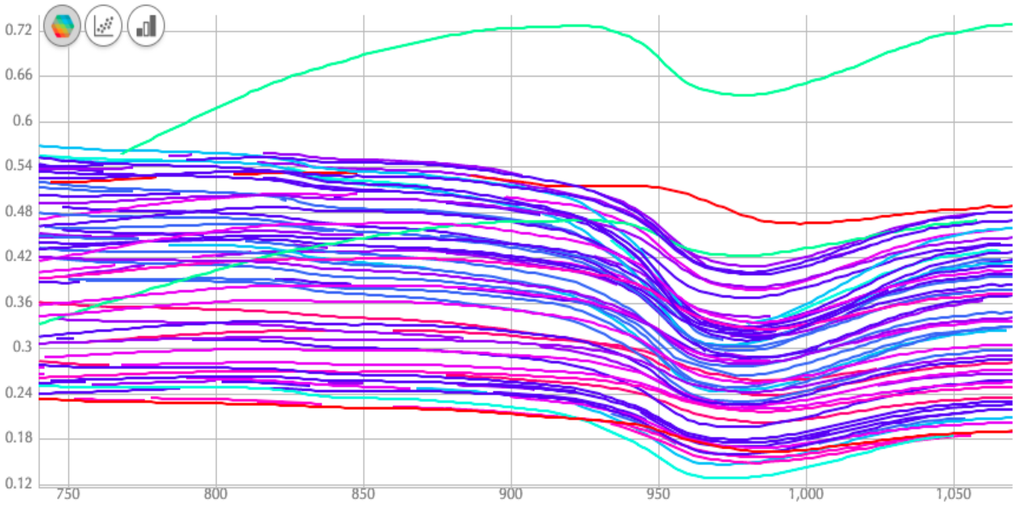

Figure 2. Raw near-infrared spectra of 62 samples. The abscissa represents the wavelength and the ordinate represents spectral reflectance

The scans could also be view under Principle Component Analysis (PCA). PCA is a common tool in machine learning in general, and even more in chemometrics. It is used to emphasize variation and bring out strong patterns in the spectra. In SCiO Lab, PCA is used to make data easier to explore and visualize by reducing the whole spectrum from a vector of 330 values (one per wavelength) to a shorter vector (typical 3-6 values), without losing too much of the information stored in the original spectrum. Using the PCA view, each spectrum is visualized as a point in 3D space. As shown in Figure 3, we can clearly see the red point was well separated from others which indicates it may have different properties. In specific, it had the highest total carbohydrate content (68.99%).

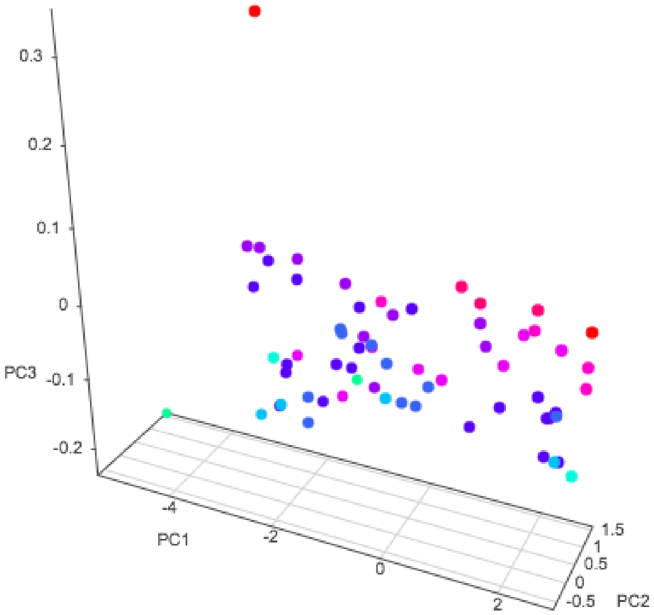

Figure 3.PCA view of total carbohydrate content of 62 food samples

The spectra were treated with certain pre-processing methods to reduce the effects not related to the chemical absorption of light. A variety of different mathematical treatments were tried, and it was found that the best results were obtained with a first-derivative math treatment (Savitzky – Golay) in combination with subtract average. The first derivative was commonly used to eliminate baseline offset variations within a set of spectra [14]. Subtract average was meant to subtract the average-over-wavelength from each point of the spectrum to eliminate remaining trends after log and derivative process, or to eliminate the gain (lambda independent gain) after log. Finally, the obtained spectra of samples were shown in Figure 4.

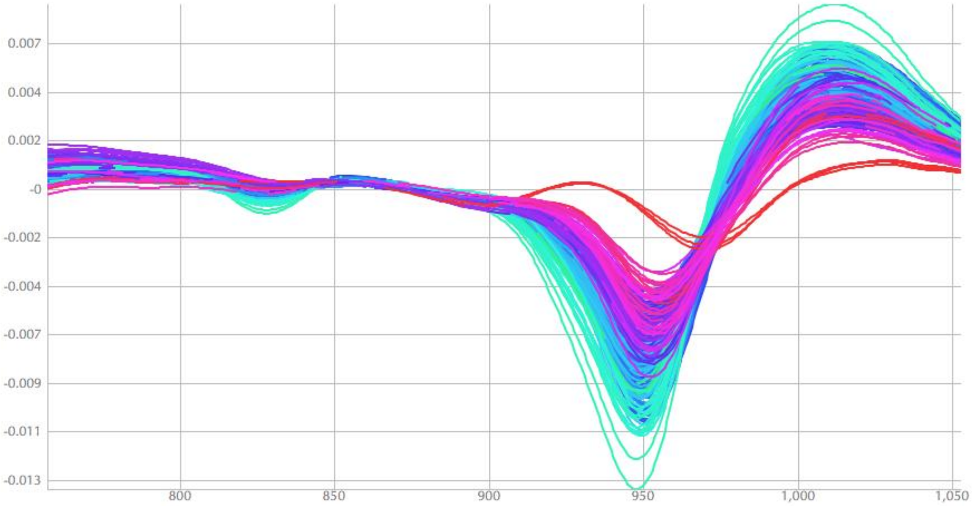

Figure 4. Final near-infrared spectra of 62 samples after filtering and using pre-processing methods. The abscissa represents the wavelength and the ordinate represents processed spectral reflectance

3.3Building calibration models

Table 3 presents R^2 and RMSE of the calibration models at different numbers of latent variables. To be more specific, when the number of latent variables = 1, the R^2 and RMSE values were 0.638 and 6.101 respectively. This means that the appropriateness of the predicted content with the initial total carbohydrate content in the standard matrix was not adequate. The RMSE value decreased when the number of latent variables increased from 1 to 12 and increased slightly when the number of latent variables was 13 and 14. Moreover, RMSE values then fluctuated when the latent variables ranged from 15 to 20. Similarly, the value of R^2 increased as the number of latent variables varied from 1 to 13 and fluctuated steadily between 0.951 and 0.952 when latent variables were between 14 and 19. This situation may be explained that if only 11 latent variables were used to build the calibration model of total carbohydrate content, the model was called under-fitting which was explained by Ziegel (2004) [13]. In their study, the authors stated that the saved number of latent variables was too small would leading to the loss of useful information from the original data set. As a result, the obtained regression model was distorted. In contrast, the RMSE value of the total carbohydrate content fluctuated gradually when the number of latent variables changed from 13 to 20. In this case, the model was encounter overfitting. This was because there were more latent variables stored than necessary and those factors contained the interference signal and thus, the regression model result had a large error.

Kilaji and Ferreira (2009) suggested that the evaluation and the optimal number of latent variables should be based on the smallest RMSE value [15]. Therefore, the PLS testing model can be chosen with the number of latent variables by 12 to build a calibration model to predict the total carbohydrate content in food samples. When the number of latent variables = 12 for the R^2 and RMSE values are 0.952 and 2.230 respectively.

Table 3 Effects of chosen latent variable number on R2 and RMSE of calibration models

Number of latent variables	R^2	RMSE	Number of latent variables	R^2	RMSE
1	0.638	6.101	**11**	0.948	2.304
2	0.803	4.501	**12**	0.952	2.230
3	0.844	4.002	**13**	0.952	2.231
4	0.895	3.294	**14**	0.951	2.244
5	0.919	2.879	**15**	0.953	2.204
6	0.937	2.542	**16**	0.952	2.220
7	0.940	2.488	**17**	0.953	2.196
8	0.943	2.431	**18**	0.952	2.220
9	0.947	2.343	**19**	0.952	2.225
10	0.948	2.310	**20**	0.949	2.284

The results of the PLSR model built on the sample set with 12 latent variables were shown in Figure 5. In this figure, the abscissa represents the known values (or the experimental data) and the ordinate represents the estimated values (or the predicted data) of total carbohydrate content in samples. The right column shows the distribution of the values in the order from low to high and each range of the values corresponding to different colors. Moreover, it can be seen that most of the points are located between the upper and lower limit lines. This means that the accuracy of the regression equation is high in determining the total carbohydrate content. On the other hand, Figure 5 also indicates that the model may has less accuracy in determining total carbohydrate content in the low range close to 20% where a few data points are located outside of the error margin. This implies the limitation of NIR devices in determining low amounts of total carbohydrate content.

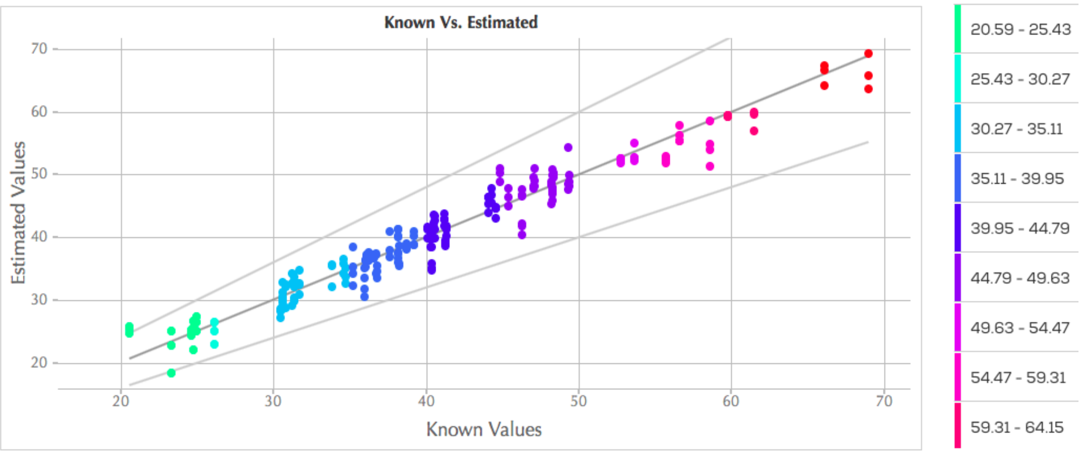

Figure 5. PLS modeling of total carbohydrate content of 62 food samples

IV. CONCLUSIONS AND RECOMMENDATIONS

In the current study, near-infrared spectroscopy was shown to provide an accurate, reliable and rapid method for the determination of total carbohydrate content in foods, which was demonstrated by good accuracy of the calibration model. The accuracy of the model still could be further improved to achieve even higher accuracy by increasing the number of samples and/or expanding the range of the total carbohydrate content. However, the device may have the limitation in determining total carbohydrate content with small amounts (e.g. <20%), which needs further investigation.

Acknowledgements. This research is funded by International University - VNUHCM under grant number T2017-01-BT.

REFERENCES

[1]. Y. Ozaki, Near-infrared spectroscopy—its versatility in analytical chemistry, Analytical Sciences, 28 (2012) 545-563.
[2]. P. Williams, K. Norris, Near-infrared technology in the agricultural and food industries, American Association of Cereal Chemists, Inc., 1987.
[3]. B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR spectroscopy with applications in food and beverage analysis, Longman scientific and technical, 1993.
[4]. H. Büning-Pfaue, R. Hartmann, J. Harder, S. Kehraus, C. Urban, NIR-spectrometric analysis of food. Methodical development and achievable performance values, Fresenius' journal of analytical chemistry, 360 (1998) 832-835.
[5]. S. Šašić, Y. Ozaki, Short-wave near-infrared spectroscopy of biological fluids. 1. Quantitative analysis of fat, protein, and lactose in raw milk by partial least-squares regression and band assignment, Analytical Chemistry, 73 (2001) 64-71.
[6]. J. Chen, X. Ren, Q. Zhang, X. Diao, Q. Shen, Determination of protein, total carbohydrates and crude fat contents of foxtail millet using effective wavelengths in NIR spectroscopy, Journal of Cereal Science, 58 (2013) 241-247.
[7]. D. Wu, S. Feng, Y. He, Short-wave near-infrared spectroscopy of milk powder for brand identification and component analysis, Journal of dairy science, 91 (2008) 939-949.
[8]. Å. Rinnan, F. Van Den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry, 28 (2009) 1201-1222.
[9]. H. Martens, T. Naes, T. Naes, Multivariate calibration, John Wiley & Sons, 1992.
[10]. D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Analytical chemistry, 60 (1988) 1193-1202.
[11]. M. Kuhn, K. Johnson, Applied predictive modeling, Springer, 2013.
[12]. L. Ding, W.F. Velicer, L.L. Harlow, Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices, Structural Equation Modeling: A Multidisciplinary Journal, 2 (1995) 119-143.
[13]. E.R. Ziegel, A user-friendly guide to multivariate calibration and classification, in, Taylor & Francis, 2004.
[14]. B.M. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review, Postharvest biology and technology, 46 (2007) 99-118.
[15]. R. Kiralj, M. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, Journal of the Brazilian Chemical Society, 20 (2009) 770-787.

Ngoc Lieu Le"Establish thecalibrationmodelfora portable near-infrared (NIR) spectroscopydevicetodeterminethe total carbohydratecontent in food"International Journal of Modern Engineering Research (IJMER), vol. 09, no. 2, 2019, pp 01-07