Multivariate analysis and artificial neural network approaches of near infrared spectroscopic data for non-destructive quality attributes prediction of Mango (*Mangifera indica* L.)

Dissertation

to obtain the Ph.D degree
at the Division of Agricultural Engineering, Faculty of Agricultural Sciences,
Georg-August-Universität Göttingen, Germany

Presented by
Agus Arip Munawar
Born in Bandung, Indonesia

Göttingen, February 2014

D7

1. Name of supervisor : Prof. Dr. Wolfgang Lücke
2. Name of co-supervisor : Prof. Dr. Elke Pawelzik

Date of disputation : 10 February 2014

Acknowledgements

I am fully indebted to many colleagues, family and friends in preparing this PhD thesis and I should start my sincere gratitude to both my academic supervisors Prof. Dr. Wolfgang Lücke and Prof. Dr. Elke Pawelzik for their willingness and guidance during my study. Many thanks also go to Dr. Dieter von Hörsten, whom I spent many hours together discussing many matters not only my progress study but also other things which are related to my daily live in Germany. I appreciate all the good laughs and help I had at work from him.

I am also most grateful to Dr. Daniel Mörlein for being my third examiner and external supervisor. I benefited much from him for being a living tutorial related to near infrared (NIR) and chemometrics. He always gave me a good advice when I need help in NIR analysis. Thanks also to Dr. Jens Wegener for his valuable comments and suggestions during my study. I also thanks to Dr. Christian Möllers for willingly accepted me for the first time and sharing his knowledge when I faced many various challenges within my first period of study.

Many thanks go to all my colleagues, former colleagues and staff at the division of Agricultural Engineering: Adnan, Maren, Christoph, Christian, Amro, Rita, Marco, Holger and Anna Maria for their friendship and cheerful discussion about many things. My deepest sincere thanks also go to Mrs.Martina, Harry, Siggi, Carsten for their valuable administrative and technical support during my work.

I would like to express my gratitude to the Directorate General for Higher Education (DIKTI), Department of National Education, Indonesia for providing scholarship through BLN-DIKTI 2010. Also thanks to Georg-August-Universität Göttingen for last five months financial support. Finally, I would like to thank all my family, especially my parents, and friends whom I could not name them one by one for all theirs great supports, understanding and patience.

Göttingen, 9 August 2013
Agus Arip Munawar

Table of content

List of figures

List of tables

List of abbreviations

ANN	:	Artificial neural network
BPNN	:	Back propagation neural network
C	:	Carbon
CV	:	Cross validation
DM	:	Dry matter
DT	:	De-trending
FM	:	Fresh mass
FT	:	Fourier transform
g	:	gram
GRNN	:	Generalized regression neural network
H	:	Hydrogen
LVs	:	Latent variables
MC	:	Mean centering
MDS	:	Method development sampling
mg	:	milligram
MLR	:	Multiple linear regression
MN	:	Mean normalization
MSC	:	Multiplicative scatter correction
N	:	Nitrogen
n	:	Number of samples
NIPALS	:	Non-iterative partial least square
NIR	:	Near infrared reflectance
NIRS	:	Near infrared reflectance spectroscopy
nm	:	nanometer
NMR	:	Nuclear magnetic resonance
O	:	Oxygen
OSC	:	Orthogonal signal correction
PCA	:	Principal component analysis
PCR	:	Principal component regression
PCs	:	Principal components
PLSR	:	Partial least square regression
r	:	Coefficient of correlation
R	:	Reflectance
R^2	:	Coefficient of determination
RMSEC	:	Root mean square error calibration
RMSECV	:	Root mean square error cross validation
RMSEP	:	Root mean square error prediction
RPD	:	Residual predictive deviation
SD	:	Standard deviation
SNV	:	Standard normal variate
SSC	:	Soluble solid content
SVMR	:	Support vector machine regression
TA	:	Titratable acidity
W	:	Weight
X	:	Independent variable, input signal
Y	:	Dependent variable, output signal

Chapter outline

Chapter 1 describes some basic concepts and features of near infrared spectroscopy (NIRS) technique, along with its recent application in agriculture especially for quality attributes assessment. The chapter also describes a brief introduction into chemometrics focused on multivariate data analysis for NIRS spectral data. This analysis is including; principal component analysis (PCA) and outlier detection, NIRS spectra pre-treatment, linear calibration method using principal component (PCR) and partial least squares (PLSR), and non-linear regression method with support vector machine regression (SVMR) and artificial neural networks (ANN).

Chapter 2 addresses to evaluate the feasibility of NIRS method in determining quality attributes of intact commercial mango purchased in local markets. These quality attributes here refer to sweetness (soluble solids content (SSC)) and acidity (titratable acidity (TA) and ascorbic acid (AA)) of mango. The calibration or prediction models for these quality attributes were developed by principal component regression and partial least squares regression method.

Chapter 3 focuses on the comparisons among different NIR spectra pre-processing methods (mean centering (MC), mean normalization (MN), multiplicative scatter correction (MSC), standard normal variate (SNV), de-trending (DT), orthogonal signal correction (OSC)) prior to prediction model development. The model accuracy and robustness obtained from these NIR spectra treatments are then compared.

In Chapter 4, nonlinear regression methods (SVMR and ANN) were applied to develop NIR prediction models for TA and SSC in which MSC and SNV spectra treatment are also used during calibration. The results obtained from these both methods are compared with ordinary linear PLS regression method. The optimal models for TA and SSC prediction were then selected from these there regression methods based on their accuracy and robustness performance.

Chapter 5 presents the general conclusion and summary for the whole contents of all works answering the objectives of the present studies.

Chapter 1. General Introduction

1.1. Basic concept of near infrared spectroscopy and its application in agriculture

In foods and agricultural products processing, the quality evaluation of these products is an important issue. Consumers are gradually looking for quality seals and trust marks on food and agricultural products, and expect producers and retailers to provide products with high quality. In order to ensure and maintain the chain supply of acceptable agricultural products, it is important to sort and grade products based on their quality. Thus, quality control plays a major important role in in every phase of the agricultural products processing (Cen and He, 2007; Jha et al., 2012). To determine quality parameters in food and agricultural products, several methods are already widely used whereby most of them are based on solvent extraction followed by other laboratory procedures. However, these methods often require laborious and complicated processing for samples. Also, they are time consuming and destructive. Therefore, a rapid and non-destructive method is required as an alternative method in determining quality parameters of foods and agricultural products.

During the last few decades, near infrared spectroscopy (NIRS) has become one of the most promising and used non-destructive methods of analysis in many field areas including in agriculture due to its advantage; simple sample preparation, rapid, and environmental friendly since no chemicals are used. More importantly, it has the potential ability to determine multiple quality parameters simultaneously (Liu et al., 2010). Numerous studies have been carried out to investigate and apply NIRS in quality assessment of foods and agricultural products (Vesela et al., 2007; Gomez et al., 2006; Jaiswal et al., 2012; Liu et al., 2008; Curda and Kukackova, 2004; Kavdir et al., 2007; Liu et al., 2007; Cen et al., 2007; Chen et al., 2011; Fan et al., 2009; Bobelyn et al., 2010; Penchaiya et al., 2009). The increasing importance of NIRS in agriculture is obvious from the recent increase in numbers of publications, as well as from the fact that many manufacturers and agricultural industries (e.g., grains, beverage, milk and dairy, and fruits and vegetables) have now implemented NIRS systems to measure and determine various quality parameters (Nicolai et al., 2007; Cozzolino et al., 2011).

The NIRS is a technique or method which uses near infrared radiation (780 – 2500 nm) of the electromagnetic spectrum to analyze the chemical composition of organic matter. It provides information through spectra signatures and patterns, regarding with the intrinsic organic bonds of the molecules and thus the primary chemical constituents of the object can be determined (Strang, 2004; Workman and Shenk, 2004; Nicolai et al., 2007). The term spectroscopy as defined by Clark (1999) is the study of electromagnetic radiation as a function of wavelength, which has been reflected, absorbed or transmitted from a solid, liquid or gas material. Spectroscopy generates a unique spectral pattern of the material monitored. Each biological object has its own special optical properties, which means it has a different spectra pattern or signatures indicated its chemical compositions. The spectral patterns of different matter are defined by their reflectance or absorbance as a function of wavelength (Siesler et al., 2002). These special signatures were then used to describe and predict the chemical constituents of biological matter.

In NIRS, the object is irradiated with near infrared radiation and the reaction (reflection, absorption or transmission) is captured. While the radiation penetrates the object, its spectral characteristics changes through wavelength dependent scattering and absorption process. The contribution of each reaction depends on the chemical composition, cell structure and physical properties of the object (Clark, 1999; Cozzolino et al., 2006; Nicolai et al., 2007). A captured NIR spectra of biological object consists the response of the molecular bonds O-H, C-H, C-O and N-H. These bonds are subject to vibrational energy changes when irradiated by NIR frequencies (Cen and He, 2007).

The primary information that can be gathered from the interaction of the near-infrared radiation with the biological object is its physical, optical and chemical properties. Fruit, grain and forage material have shown to have identifiable C-H, N-H, and O-H absorption bands in the near-infrared region whereas each have a specific vibrational frequency and it is different between one object and the others (Workman and Shenk, 2004). Figure 1.1 shows typical diffuse reflectance spectra of NIR for some selected agricultural products. NIR spectra of fruits and vegetables is essentially composed of a large set of overtones and combination bands and further may be complicated since the spectra is influenced by wavelength dependant scattering effects, tissue heterogeneities, instrumental noise, ambient effects and other source of variability (Nicolai et al., 2007; Cozzolino et al., 2011). These factors may generate spectra noise and influence NIR prediction performance. Several methods are introduced as spectra pre-treatment to overcome these factor effects such as spectra smoothing, standardization, normalization and transformation (Pontes et al., 2006).

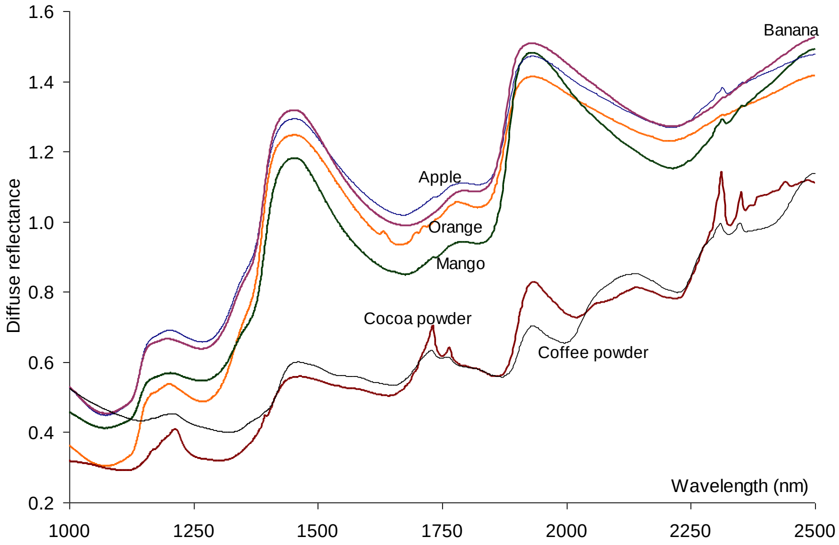

Figure 1.1.Typical diffuse reflectance spectra of some agricultural products. The near infrared reflectance spectra were acquired using a FT-NIR instrument: Nicolet Antaris *(Source: Own data analysis).*

The whole measurement processing in NIRS generally consists of the following: (1) NIR spectra data acquisitions, (2) spectra pre-processing to eliminate noises and baseline shift from the instrument and background, (3) develop calibration models using a set of samples with known analyzed concentration obtained by suitable and standard laboratory procedures, and (4) validate the prediction models using another set of independent samples. Since NIRS itself cannot reveal chemical information in the spectra, chemometrics is required to extract the information about quality attributes buried on NIR spectra through a process called multivariate calibration from which a mathematical relationship between NIR spectra and the measured quality parameter will be revealed to determine desired quality attributes.

1.2. Chemometrics

According to Naes et al. (2004) chemometrics is the use of statistical and mathematical procedures to extract information from chemical and physical data. It has been used to extend and improve the potential application of NIRS technique in many fields including food and agricultural industries. In NIRS analysis, this method is includes three facets: (1) spectral data pre-processing to eliminate noise and enhance spectra prior to models development, (2) building calibration models for quantitative and qualitative analysis and (3) model transfer for real-time and in-line prediction (Cen and He, 2007).

1.2.1. Spectra pre-processing

The spectra data acquired from NIR instrument contain spectra background information and noises which are interfered desired relevant quality attributes information. Interfering spectral parameters, such as light scattering, path length variations and random noise resulted from variable physical sample properties or instrumental effects need to be eliminated or reduced in order to obtain reliable, accurate and stable calibration models (Reich, 2005; Cen and He, 2007). Thus, it is very necessary to pre-process spectral data prior to modeling. The most commonly used spectra pre-processing methods are briefly discussed with respect to the effect they are able to correct.

The first stage in spectra pre-processing is *mean centering*. This is often the simply preferred pre-processing method prior calibration development as it focuses on differences between observations rather than their absolute values. It ensures that the resulting data or model may be interpreted in terms of variation around the mean (Naes et al., 2004; Nicolai et al., 2007). Another common spectra pre-processing is smoothing from which improves the visual aspect of the NIR spectra. Spectra *standardisation* is also commonly used when variables are measured in different units or have different ranges (Cozzolino et al., 2011). Standardisation means dividing spectrum at every wavelength by the standard deviation of the spectrum at this wavelength. Typically variances of all wavelengths are standardised to 1, which results in an equal influence of the variables in the model (Naes et al., 2004).

The other spectra pre-processing method is *normalization*. Multiplicative scatter correction (MSC) and standard normal variate (SNV) are the most popular normalization technique. MSC is used to compensate for additive (baseline shift) and multiplicative effects in the spectral data, which are induced by physical effects, such as the non-uniform scattering throughout the spectrum. The degree of scattering is dependent on the wavelength of the radiation, the particle size and the refractive index. This method attempts to remove the effects of scattering by linearizing each spectrum to an '*ideal*' spectrum of the sample, which is normally corresponds to the average spectrum. On the other hand, in SNV each individual spectrum is normalized to zero mean and unit variance. Apart from the different scaling, the result is more-less similar to that of MSC (Naes et al., 2004; Nicolai et al., 2007; Cozzolino et al., 2011). Figures 1.2 show a visual example result after spectra pre-processing which is in this case after MSC treatment.

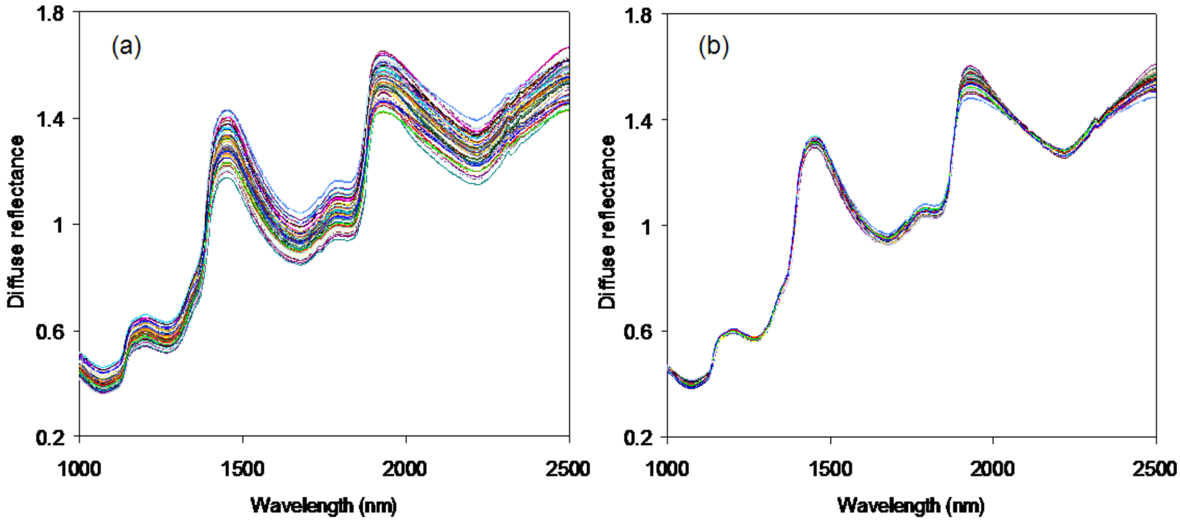

Figure 1.2. Raw untreated near infrared spectra (a) and after multiplicative scatter correction (b) *(Source: Own data analysis).*

Spectra *transformation* into its derivatives is also used as a spectra pre-processing since this method also has the ability for correcting both additive and multiplicative effects (like MSC or SNV). Derivation is usually calculated according to the Savitzky-Golay algorithm (Naes et al., 2004). The smoothing parameters of the algorithm, interval width and polynomial order, should be considered carefully in order to avoid spectral noise amplification. These smoothing factors determine how many adjacent variables will be used to estimate the polynomial approximation used for derivatives (Mouazen et al., 2010). The latest spectra pre-processing is a method called orthogonal signal correction (OSC). It is the method developed to reduce the data variance in the spectra due to light scatter effects and to more general types of interferences that have no correlation with the measured property i.e quality attributes or chemical constituents of the object (Azzouz et al., 2003; Felizardo et al., 2007).

1.2.2. Principal component analysis and outlier detection

Prior to calibration model development, normally original NIR spectra acquired from the NIR instrument were analyzed through principal component analysis (PCA). It employs a mathematical procedure that transforms a set of possibly correlated response variables into a new set of non-correlated variables, called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. PCA is used as a tool for screening, extracting, compressing and discriminating samples based on their similarities or dissimilarities of multivariate data. Figure 1.3 shows PCA discrimination result of some agricultural products based on NIR diffuse reflectance spectral data.

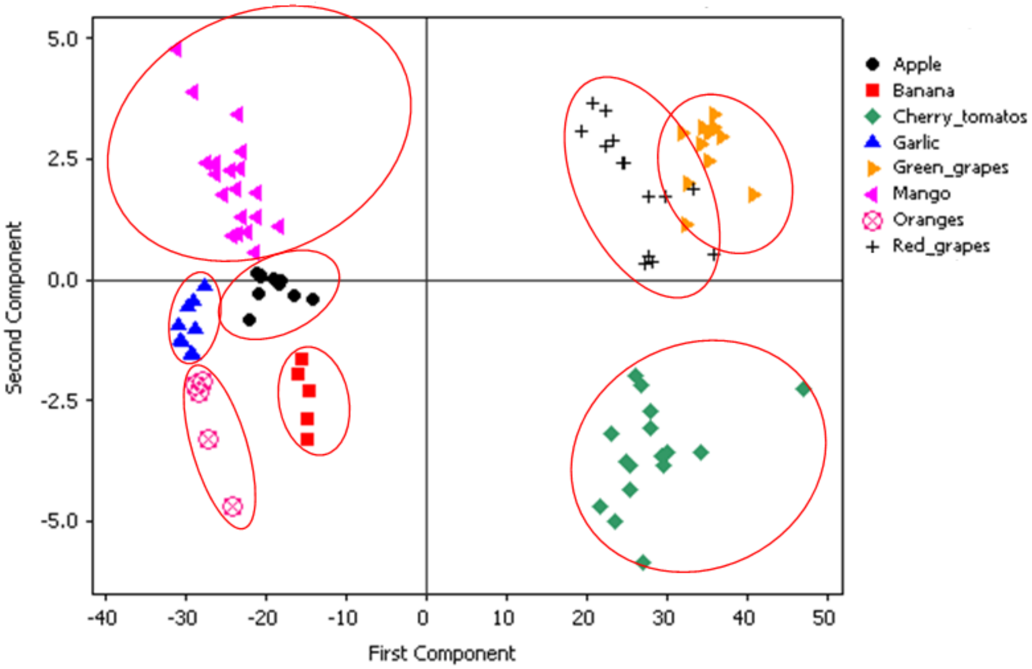

Figure 1.3. Principal component analysis of some agricultural products based on near infrared reflectance diffuse reflectance spectra *(Source: Own data analysis).*

Sample or variable outliers may be induced by typing errors, file transfer, interface errors, sensor malfunctions and fouling, poor sensor calibration, poor sampling or poor sample presentation. A sample can be considered as an outlier according to the spectra (X-variables) only, to the reference or measured property (Y-variables), or to both. It might also not be an outlier for either separate sets of variables, but become an outlier when the X–Y relationship is considered. The detection of outliers was carried out by subjecting the 95% confidence ellipse (Hotelling T^2 ellipse analysis) onto PCA map. Outliers related to the spectra show up in the first two principal components of PCA scores plot as points outside the ellipse and should be removed before building a calibration model (Constantinou et al., 2004; Naes et al., 2004; Nicolai et al., 2007; Dardenne, 2010; Cozzolino et al., 2011).

1.2.3. Calibration models

The main part of NIR - chemometrics is building a calibration model used to predict quality attributes or chemical constituents of agricultural products to be observed. This model consists of the relationship between the observed response variable y (Y-variables: quality attributes such as total soluble solids, acidity, protein content or fat content) and the independent variable x (X-variables: NIR spectra matrices). The common linear regression methods for building this model such as multiple linear regression (MLR), principal component regression (PCR) and partial least squares regression (PLSR). Meanwhile, as nonlinear regression, supporting vector machine regression (SVMR) and artificial neural networks (ANN) are two common examples.

In *multiple linear regressions*, the response (Y variable) is approximated by a linear combination of the spectral values at every single wavelength. The regression coefficients are estimated by minimizing the error between predicted and observed response values based on least squares. A stepwise or backward elimination method may be applied to select a number of optimum variables for the equation. MLR models typically do not perform well because of the often high co-linearity of the spectra and easily lead to over-fitting and loss of robustness of the calibration models (Saranwong et al., 2001; Naes et al., 2004; Nicolai et al., 2007).

On the other hand, *principal component regression* is a two-step procedure; first decomposes the X-variables by a principal component analysis (PCA) and then fits a MLR model, using a small number of principal components (PCs) or latent variables (LVs) instead of the original variables (NIR spectra) as predictors (Naes et al., 2004). The advantage with respect to MLR is that the X-variables (principal components) are uncorrelated, and that the noise is filtered. Also, usually a small number of principal components are preferable and sufficient for the models. The main disadvantage of PCR is that the principal components are ordered according to decreasing explained variance of the spectral matrix, where the first principal component which is used for the regression model is not necessarily the most informative with respect to the response variable (Wold et al., 2001; Naes et al., 2004; Nicolai et al., 2007; Cozzolino et al., 2011).

Partial least squares regression is a regression method with close likely to principal component regression. The main difference of the PLSR method is that both the input X-variables (NIR spectra) and response variables Y are projected to new spaces (Balabin et al., 2011). PLS regression is of particular interest because, unlike MLR, it can analyze data with strongly collinear (correlated), noisy and redundant variables (X variables or wavelengths) and also model several characteristics (Y values) at the same time. In PLS regression an orthogonal basis of latent variables is constructed one by one in such a way that they are oriented along the directions of maximal covariance between the spectral matrix and the response vector (Wold et al., 2001). In this way it is ensured that the latent variables are ordered according to their relevance for predicting the Y-variable. Interpretation of the relationship between X-data and Y-data (the regression model) is then simplified as this relationship is concentrated on the smallest possible number of latent variables (Wold et al., 2001; Naes et al., 2004).

This method performs particularly well when the various X-variables express common information, for example, when there is a large amount of correlation, or even co-linearity, which is the case for spectral data of biological materials such as fruits or vegetables. The required number of latent variables is typically smaller than that in a PCR calibration model for a similar model performance (Wold et al., 2001; Brereton, 2000; Naes et al., 2004).

In many current and potential applications of NIR spectroscopic measurement, the relationship between NIR spectra and targeted constituents to be modeled is not always linear. The source of nonlinearity may vary widely, and is difficult to identify. In NIRS, as in other spectroscopic techniques, some deviations from linearity are of known origin (breakdown of the Lambert–Beer law at high analyte concentrations, nonlinear detector response, light source scatter), whilst others are intrinsic to the parameter to be measured. This means that classical linear regression such as MLR, PCR or PLSR methods are not always the most suitable option. Extrinsic deviations from linearity may be corrected by mathematical pretreatment of the signal prior to using linear calibration techniques. Intrinsic source of non-linearity in NIR spectral data and target chemical components cannot be corrected by spectral pretreatments and require the use of special nonlinear adjustment approaches (Perez-Marin et al., 2007).

Besides with linear regression technique (PCR and PLSR), NIR calibration models were also can be constructed using nonlinear regression like supporting vector machine regression (SVMR) and artificial neural networks (ANN). Based on some previous studies, SVMR and ANN are more flexible methods since they can handle both linear and nonlinear relationship between the NIR spectra and corresponding chemical constituents (Blanco and Peguero, 2008; Cozzolino et al., 2011; Blanco et al., 2000; Zou et al., 2010).

Support vector machine is very specific class of algorithm, characterized by usage of kernels (kernel based). In earlier development, this method was applied for classification problems, but nowadays it also has been used to the case of regression. In kernel-based methods, the calibration is carried out in a space of non-linearly transformed input data (so called feature space) without actually carrying out the transformation. The feature space is defined by the kernel function, a measure of similarity between spectra. The most popular kernel functions are the Gaussian and polynomial functions (Nicolai et al., 2007). A number of studies has been reported the use of ANN or SVM method and comparison between both of them for NIR calibration to solve various regression and classification problems in agriculture (Janik et al., 2007; Xiaoying et al., 2012; Cao et al., 2010; Wu et al., 2008; Borin et al., 2006; Huang et al., 2011).

On the other hand, *artificial neural network* is a machine learning algorithm inspired to mimic human brain that is characterized by its analogy with a biological neuron (Sima and Orponen, 2003; Naes et al., 2004). In the biological neuron the input signal from the dendrites travels through the axons to the synapse (Janik, et al., 2007). There the information is transformed and sent across the synapse to the dendrites of the next neuron forming part of a highly complex network.

Artificial neural network typically consists of three layers so called input layer, hidden layer and output layer. Like our brains, each input is connected with cells called neurons. Every neuron of the input layer is connected to every neuron of the hidden layer, and every neuron of the hidden layer is connected to the output layer. In agreement with Naes et al. (2004), Brereton (2000) stated that the multivariate techniques based on ANN simulates the biological neuron by multiplication of the input signal (X) with the synaptic weight (W) to derive the output signal (Y). A neuron is acted as a computational device that calculates the weighted sum of its input and calculates the output signal from this using a non-linear function (Kim, et al., 2000; Hahn et al., 2004; Nicolai et al., 2007). In NIR cases, the spectral value at every wavelength is fed to the input layer, while the output layers delivers the prediction of the quality attributes observed. To simplify the input and reduce calculation times, the spectral value of NIR were subjected firstly to the PCA, then, five to seven PCs were used as input instead of all spectral value. This combination method is called PCA-NN (principal component analysis-neural network).

1.2.4. Model validation and accuracy

During calibration model development, cross validation procedures have to be applied in order to assess the accuracy of the model and to avoid over-fitting. In NIR feasibility studies, cross validation is a practical method to demonstrate that NIRS can predict something, but the actual accuracy must be estimated with an appropriate test dataset or validation set (Dardenne, 2010). Therefore in such studies different cross validation techniques can be used. For example, in leave one out cross validation, one sample is removed from the dataset, and a calibration model is constructed for the remaining subset. The removed samples are then used to calculate the prediction residual (Brereton, 2000; Naes et al., 2004). The process is repeated with other subsets until every sample has been left out once, and in the end the variance of all prediction residuals is estimated. In multifold cross validation, a well-defined number of samples (‘*segment*’) are left out instead of one. In internal validation, the dataset is split into a calibration and a validation set. The calibration model is constructed using the calibration set, and the prediction residuals are then calculated by applying the calibration model to the validation dataset (Naes et al., 2004; Nicolai et al., 2007; Cozzolino et al., 2011).

The predictive ability and accuracy of the NIRS method needs to be demonstrated using an independent validation set. Independent means that samples need to come from different experiments, harvest times, or batches with spectra taken at a time different from the calibration spectra (Norris, 2007; Dardenne, 2010). For example, samples obtained from a different orchard, different season or different region or environment. Many statistics are reported in the literature to interpret a calibration such as the coefficient of determination (R^2) of calibration and validation which essentially represents the proportion of explained variance of the response in the calibration or validation data set (Nicolai et al., 2007), coefficient of correlation (r) between predicted and measured quality attributes, prediction error which is defined as the root mean square error of calibration (RMSEC), standard error of calibration (SEC), root mean square error of cross validation prediction (RMSECV), root mean square error prediction (RMSEP), and the difference between RMSEC and RMSECV or RMSEP (Jha, et al., 2006; Flores et al., 2009).

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - y_m)^2} \qquad \text{(eq. 1.1)}$$

$$RMSEC, RMSECV, RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad \text{(eq. 1.2)}$$

Where $\hat{y}_i$ is the predicted value of the *i*-th observation, y_i is the measured value of the *i*-th observation from desired quality attributes, *n* is the number of observations in the calibration, validation or prediction set, and y_m is the mean value of the calibration or validation data set. The prediction error of a calibration model is defined as the RMSECV when cross validation is used or the RMSEP when independent validation is used (Naes et al., 2004; Zeaiter et al., 2004; Walsh and Kawano, 2009). As defined by Golic and Walsh (2006), The RMSECV describes total error for samples within the calibration dataset while the RMSEP is an estimate of total prediction error for an independent validation dataset. Another useful statistic commonly used to interpret NIRS calibrations is the residual predictive deviation or RPD (Williams, 2001; Fearn 2002). It shows the ratio between the standard deviation (SD) of the original reference data to the root mean square error of cross validation (RMSECV) or to the root mean square error of prediction (RMSEP).

$$RPD = \frac{SD_{ref}}{RMSECV} \text{ or } RPD = \frac{SD_{ref}}{RMSEP} \qquad \text{(eq. 1.3)}$$

Based on literature, An RPD between 1.5 – 1.9 means that coarse quantitative prediction are possible, but still need some improvement in calibration. A value between 2 and 2.5 indicates that prediction model is sufficient. Meanwhile, an RPD value between 2.5 and 3 or above corresponds to good and excellent prediction accuracy respectively (Williams, 2001; Fearn 2002; Nicolai et al., 2007).

1.3. Objectives

The present study was addressed to achieve the following objectives:

a. Evaluate the feasibility of NIRS method in determining quality attributes in term of soluble solids content (SSC), titratable acidity (TA) and ascorbic acid (AA) of intact mango as example for agricultural product through multivariate calibration model followed by cross validation.
b. Compare different spectra pre-processing methods prior to calibration models development and evaluate their impact to the model prediction accuracy and robustness.
c. Investigate the use of nonlinear regression method (supporting vector machine regression, SVMR and artificial neural networks, ANN) to predict quality attributes of mango and compare the results obtained by linear partial least square regression (PLSR) method.

References

Azzouz, T., Puigdomenech, A., Aragay, M., & Tauler, R. (2003). Comparison between different pre-treatment methods in the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method. *Analytica Chimica Acta*, 484, 121-134.

Balabin, R. M., Lomakina, E. I., & Safieva, R. Z. (2011). Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. *Fuel*, 90, 2007–2015.

Blanco, M., & Peguero, A. (2008). An expeditious method for determining particle size distribution by near infrared spectroscopy: Comparison of PLS2 and ANN models. *Talanta*, 70, 647-651.

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., & Pages, J. (2000). NIR calibration in non-linear systems: Different PLS approaches and artificial neural networks. *Chemometrics and Intelligent Laboratory System*, 50, 75-82.

Bobelyn, E., Serban, A. S., Nicu, M., Lammertyn, J., Nicolai, B. M., & Saeys, W. (2010). Postharvest quality of apple predicted by NIR-spectroscopy: Study of the effect of biological variability on spectra and model performance. *Postharvest Biology and Technology*, 55, 133-143.

Borin, A., Ferrao, M. F., Mello, C., Maretto, D. A., & Poppi, R. J. (2006). Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Analytica Chimica Acta*, 579, 25–32.

Brereton, R. G. (2000). Introduction to multivariate calibration in analytical chemistry. *The Analyst*, 125, 2125−2154.

Cao, F., Wu, D., & He, Y. (2010). Soluble solids content and pH prediction and varieties discrimination of grapes based on visible-near infrared spectroscopy. *Computers and Electronics in Agriculture*, 715, 515–518.

Cen, H. Y., Bao, Y. D., He, Y., & Sun, D. W. (2007). Visible and near infrared spectroscopy for rapid detection of citric and tartaric acids in orange juice. *Journal of Food Engineering*, 82, 253–260.

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology*, 18, 72-83.

Chen, L., Xue, X., Ye, Z., Zhou, J., Chen, F., & Zhao, J. (2011). Determination of Chinese honey adulterated with high fructose corn syrup by near infrared spectroscopy. *Food Chemistry*, 128, 1110–1114.

Clark, R. N. (1999). Spectroscopy of rocks and minerals and principles of spectroscopy. In Rencz, A. N (Ed.), *Manual of remote sensing*. (*Volume 3): Remote sensing for the earth sciences*, (pp. 3-58). New York: John Wiley and Sons, (Chapter 1).

Constantinou, M. A., Papakonstantinou, E., Benaki, D., Spraul, M., Shulpis, K., Koupparis, M. A., & Mikros, A. (2004). Application of nuclear magnetic resonance spectroscopy combined with principal component analysis in detecting inborn errors of metabolism using blood spots: a metabonomic approach. *Analytica Cimica Acta*, 511, 303-312.

Cozzolino, D., Cynkar, W. U., Dambergs, R. G., Shah, N., & Smith, P. (2009). Multivariate methods in grape andwine analysis. *International Journal of Wine Research*, 1, 123−130.

Cozzolino, D., Cynkar, W. U., Shah, N., & Smith, P. (2011). Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Research International*, 44, 1888-1896.

Curda, L., & Kukackova, O. (2004). NIR spectroscopy: a useful tool for rapid monitoring of processed cheeses manufacture. *Journal of Food Engineering*, 61, 557–560.

Dardenne, P. (2010). Some considerations about NIR spectroscopy: Closing speech at NIR-2009. *NIR News*, 21, 8-9.

Fan, G., Zha, J., Du, R., & Gao, L. (2009). Determination of soluble solids and firmness of apples by Vis/NIR transmittance. *Journal of Food Engineering*, 93, 416-420.

Fearn, T. (2002). Assessing calibrations: SEP, RPD, RER and R^2. *NIR News*, 13, 12−14.

Felizardo, P., Baptista, P., Menezes, J.C., & Correia, M.J.N. (2007). Multivariate near infrared spectroscopy models for predicting methanol and water content in biodiesel. *Analytica Chimica Acta*, 595, 107-113.

Flores, K., Sanchez, M. T., Perez-Marin, D., Guerrero, J. E., & Garrido-Varo, A. (2009). Feasibility in NIRS instruments for predicting internal quality in intact tomato. *Journal of Food Engineering*, 91, 311-318.

Golic, M., & Walsh, K. B. (2006). Robustness of calibration models based on near infrared spectroscopy for the in-line grading of stone fruit for total soluble solids content. *Analytica Chimica Acta*, 555, 286–291.

Gomez, A. H., He, Y., & Pereira, A. G. (2006). Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using Vis/NIR-spectroscopy techniques. *Journal of Food Engineering*, 77, 313–319.

Hahn, F., Lopez, I., & Hernandez, G. (2004). Spectral detection and neural network discrimination of *Rhizopus stolonifer* spores on red tomatoes. *Biosystem Engineering*, 89, 93–99.

Huang, L., Wu, D., Jin, H., Zhang, J., He, Y., & Lou, C. (2011). Internal quality determination of fruit with bumpy surface using visible and near infrared spectroscopy and chemometrics: A case study with mulberry fruit. *Biosystem Engineering*, 109, 377–384.

Jaiswal, P., Jha, S. N., & Bharadwaj, R. (2012). Non-destructive prediction of quality of banana using spectroscopy. *Scientia Horticulturae*, 135, 14-22.

Janik, L. J., Cozzolino, D., Dambergs, R., Cynkar, W., & Gishen, M. (2007). The prediction of total anthocyanin concentration in red-grape homogenates using visible-near infrared spectroscopy and artificial neural networks. *Analytica Chimica Acta*, 594, 107–118.

Jha, S. N., Jaiswal, P., Narsaiah, K., Gupta, M., Bhardwaj, R., & Singh, A. K. (2012). Non-destructive prediction of sweetness of intact mango using near infrared spectroscopy. *Scientia Horticulturae*, 138, 171-175.

Jha, S. N., Kingsly, A. R. P., & Chopra, S. (2006). Non-destructive determination of firmness and yellowness of mango during growth and storage using visual spectroscopy. *Biosystems Engineering*, 94, 397-402.

Kavdir, I., Lu, R., Ariana, D., & Ngouajio, M. (2007). Visible and near–infrared spectroscopy for nondestructive quality assessment of pickling cucumbers. *Postharvest Biology and Technology*, 44, 165–174.

Kim, J., Mowat, A., Poole, P., & Kasabov, N. (2000). Linear and non-linear pattern recognition models for classification of fruit from visible-near infrared spectra. *Chemometrics and Intelligent Laboratory System*, 51, 201-216.

Liu, Y. D., Ying, Y. B., Fu, X. P., & Lu, H. S. (2007). Experiments on predicting sugar content in apples by FT-NIR Technique. *Journal of Food Engineering*, 83, 986–989.

Liu, Y., Chen, X., & Ouyang, A. (2008). Nondestructive determination of pear internal quality indices by visible and near-infrared spectrometry. *LWT - Food Science and Technology*, 41, 1720–1725.

Liu, Y., Sun, X., & Ouyang, A. (2010). Nondestructive measurement of soluble solid content of navel orange fruit by visible–NIR spectrometric technique with PLSR and PCA-BPNN. *LWT - Food Science and Technology*, 43, 602–607.

Mouazen, A. M., Kuang, B., De Baerdemaeker, J., & Ramon, H. (2010). Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma*, 158, 23–31.

Naes, T., Isaksson, T., Fearn, T., & Davies, T. (2004). A user-friendly guide to multivariate calibration and classification. Chichester, UK: NIR publications.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lamertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. *Postharvest Biology and Technology*, 46, 99-118.

Norris, K. (2007). Hazards with near infrared spectroscopy in detecting contaminants. *Journal of Near Infrared Spectroscopy*, 17, 165−166.

Penchaiya, P., Bobelyn, E., Verlinden, B. E., Nicolai, B. M., & Saeys, W. (2009). Non-destructive measurement of firmness and soluble solids content in bell pepper using NIR spectroscopy. *Journal of Food Engineering*, 94, 267-273.

Perez-Marin, D., Garrido-Varo, A., & Guerrero, J. E. (2007). Non-linear regression methods in NIRS quantitative analysis. *Talanta*, 72, 28-42.

Pontes, M.J.C., Santos, S.R.B., Araújo, M.C.U., Almeida, L.F., Lima, R.A.C., Gaião, E.N., & Souto, U.T.C.P. (2006). Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometry. *Food Research International*, 39, 182-189.

Reich, G. (2005). Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Advanced Drug Delivery Reviews*, 57, 1109-1143.

Saranwong, S., Sornsrivichai, J., & Kawano, S. (2001). Improvement of PLS calibration for Brix value and dry matter of mango using information from MLR calibration. *Journal of Near Infrared Spectroscopy*, 9, 287−295.

Siesler, H.W., Ozaki, Y., Kawata, S., & Heise, H. M. (2002). Near Infrared Reflectance Spectroscopy : Principles, Instrument and Application. Wiley VHC Verlag, GmbH, Weinheim.

Sima, J., & Orponen, P. (2003). General purpose computation with neural networks: A survey of complexity theoretic result. *Neural Computation*, 15, 2727−2778.

Strang, G.C. (2004). Near Infrared Reflectance Spectroscopy and its Specific Applications in Livestock Agriculture. School of Bioresources Engineering and Environmental Hydrology. University of Kwazulu-Natal, Pietermaritzburg.

Vesela, A., Barros, A. S., Synytsya, A., Delgadillo, I., Copikova, J., & Coimbra, M. A. (2007). Infrared spectroscopy and outer product analysis for quantification of fat, nitrogen, and moisture of cocoa powder. *Analytica Chimica Acta*, 601, 77–86.

Walsh, K.B. & Kawano, S. (2009). Near infrared spectroscopy, In *Optical monitoring of fresh and processed agricultural products*, M. Zude, (eds). CRC Press, Taylor and Francis Group, Boca Raton, Fl. 192–239.

Williams, P. C. (2001). Implementation of near-infrared technology. In P. C. Williams, & K. H. Norris (Eds.), *Near Infrared Technology in the Agricultural and Food Industries*. St. Paul, Minnesota, USA: American Association of Cereal Chemist. 145-169.

Wold, S., Sjöstrom,M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109−130.

Workman, J., & Shenk, J. (2004). Understanding and using the near-infrared spectrum as an analytical method. In: *Near-infrared spectroscopy in agriculture*. Roberts, C.A., J. Workman, and J.B. Reeves III (Eds). ASA, CSSA and SSSA publications, Madison, Wisconsin, 3-10.

Wu, D., He, Y., Feng, S., & Sun, D.W. (2008). Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM. *Journal of Food Engineering,* 84, 124-131.

Xiaoying, N., Zhilei, Z., Kejun, J., & Xiaoting, L. (2012). A feasibility study on quantitative analysis of glucose and fructose in lotus root powder by FT-NIR spectroscopy and chemometrics. *Food Chemistry,* 133, 592-597.

Zeaiter, M., Roger, J. M., Bellon-Maurel, V., & Rutledge, D. N. (2004). Robustness of models developed by multivariate calibration. Part I. The assessment of robustness. *Trends in Analytical Chemistry*, 23, 157−170.

Zou, X. B., Zhao, J. W., Poveyb-Malcolm, J. W., Mel, H., & Mao, H. P. (2010). Variables selection methods in near infrared spectroscopy. *Analytica Chimica Acta,* 667, 14-32.

Chapter 2. Non-destructive prediction of mango quality attributes using near infrared spectroscopy

2.1. Abstract

To establish a non-destructive method for prediction of mango quality attributes, near infrared reflectance spectroscopy (NIRS) combined with chemometrics was studied. NIR spectra were recorded on intact mangos (cv. *Kent*, n = 58) in the wavelength range of 1000 to 2500 nm using a Fourier transform near infrared spectroscopy, followed by its quality attributes measurement. Partial least squares regression (PLSR) and principal component regression (PCR) based on various spectral pre-treatment (multiplicative scatter correction (MSC), standard normal variate (SNV) and first derivative) were used to develop prediction models for quality attributes such as soluble solids content (SSC), titratable acidity (TA) and ascorbic acid (AA). The models yielded satisfactory results with coefficient of determination of calibration ranging from 0.66 to 0.91 (SSC), 0.94 to 0.98 (TA) and 0.62 to 0.92 (AA) while standard error of calibration and cross validation were low. It is concluded that NIRS and chemometrics is feasible for rapid and non-destructive prediction of mango quality.

Keywords: NIRS, chemometrics, soluble solids, ascorbic acid, acidity, rapid methods.

2.2. Introduction

Mango (*Mangifera indica*) is one of the most important and popular tropical fruits for people around the world due to its taste, appearance and excellent overall nutritional source from which lead to a heavy demand in world fruit market. To ensure and maintain the chain supply of good quality fruit, it is important to sort and grade mango based on its quality. Thus, quality control plays a major important role in deciding the export value of the fruit (Jha et al., 2012). The quality of mango fruit is judged by external parameters such as surface colour, size and shapes , and internal attributes such soluble solids content (SSC), vitamin A, vitamin C, acidity, pH and dry matter indicating sensory properties (Xudong et al., 2009). The internal quality is commonly determined by wet laboratory methods which are destructive in nature, involve chemicals, laborious and time consuming (Jha et al., 2012). Hence, research is being performed worldwide to develop rapid and non-destructive techniques to predict the quality attributes simultaneously. These techniques include ultrasound, tapping method, nuclear magnetic resonance (NMR), machine vision including image processing, and near infrared spectroscopy (Khalaf et al., 2004).

Recently, the application of near infrared reflectance spectroscopy (NIRS) as non-destructive technique in food and agricultural product sector is gaining more attentions both in term of instrumental design and spectra data analysis. By definition, NIRS covers of spectra wavelength range from 780 to 2500 nm (Cen & He, 2007; Nicolai et al., 2007), works based on the principle of electromagnetic radiation interaction with biological objects. Interacting with biological objects, the incident radiation is partially reflected (diffuse and specular reflection), absorbed, and transmitted. The contribution of each reaction depends on the chemical composition, cell structure and physical properties of the object (Clark, 1997; Nicolai et al., 2007).

NIRS is considered to be suitable for determining inner quality of food and agricultural product since it is characterized by low labour costs, non-destructive, pollution free and rapid. NIRS also allows several constituents to be evaluated at the same time (Moghimi et al., 2010). This technique has been gaining widespread acceptance for analyzing foods and agricultural products since its development in 1964 (Khalaf et al., 2004; Nicolai et al., 2007). The increasing numbers of research and publications in NIRS showed obviously that this method is becoming more important and promising technique in quality control as well as from the fact that many manufacturers have now implemented NIRS system to measure various physical and chemical quality attributes (Cen and He, 2007: Nicolai et al., 2007).

A number of studies have been investigating the feasibility of NIRS to determine various quality attributes in horticultural products as shown in Table 2.1. For mango quality, there are, however, few studies applying NIRS (Jha, et al., 2010). Schmilovitch et al. (2000) used

NIR spectra with wavelength range of 1200-2400 nm to evaluate major physiological properties and quality indices of mango fruit (cv. *Tommy Atkins*) included softening of the flesh, soluble solid content and acidity. Mahayothee et al. (2004) evaluated firmness on two Thai mango cultivars in the visible to NIR range (650-2500 nm). Mango eating qualities, i.e. soluble solid content, dry matter (DM) and flesh colour have been predicted using short wave NIR of 700-1100 nm (Saranwong et al., 2004) and 300-1150 nm (Subedi et al., 2006). Firmness and yellowness of mango during growth and storage has also been determined using NIRS (Jha et al., 2006; Valente et al., 2009).

Table 2.1. Overview of NIRS application to predict some quality attributes of horticultural products

Fruit	Quality attributes	Reference
Apples	SSC and firmness	Ventura et al. (1998); Lammertyn et al. (1998); Xiaobo et al. (2007); Fan et al. (2009); Mendoza et al. (2012)
Strawberries	SSC, TA and firmness	Sánchez et al. (2012)
Orange	SSC	Cayuela (2008); Liu et al. (2010)
Mandarin	SSC and TA	Gomez et al. (2006)
Tomato	SSC, TA and firmness	Shao et al. (2007); Flores et al. (2009)
Banana	Dry matter content, pH, SSC and acid to brix ratio	Jaiswal et al. (2012)
Kiwifruit	SSC and acidity	Moghimi et al. (2010)
Apricot	SSC, firmness and TA	Camps and Christen (2009)
Bayberry	TA, malic acid, and citric acid	Xie et al. (2011)

SSC: soluble solids content, TA: titratable acidity

As NIRS itself cannot predict quality, chemometrics is required to extract the information provided by near infrared spectra. Multivariate calibration is used to establish mathematical and statistical relationships between NIR spectra and target quality parameters (Naes et al., 2004). In NIRS analysis, this chemometrics includes three facets: (i) spectral data pre-processing, (ii) building calibration models for quantitative and qualitative analysis (including model validation) and (iii) model transfer (Cen and He, 2007). As NIR spectra of fruits and vegetables are characterized by a large set of overtones and combination bands, and further may be complicated since the spectra is influenced by wavelength dependant scattering effects, tissue heterogeneities, instrumental noise, ambient effects and other source of variability. Several pre-processing methods have been introduced to overcome these effects such as spectra smoothing, standardisation, normalisation, derivation, wavelet transforms, Fourier transform, orthogonal signal correction, net analyte signal and combination among them (Pontes et al., 2006; Wang et al., 2006; Nicolai et al., 2007; Cen and He, 2007; Cozzolino et al., 2011).

Therefore, the main objective of this study is to develop multivariate calibration models using PLS and PCR based on scatter corrected spectra, namely multivariate scatter correction (MSC) and standard normal variate (SNV) and their transformed spectra into first derivative to predict soluble solid content (SSC), titratable acidity (TA), and ascorbic acid (AA) representing sweetness and acidity of mango respectively in a non-destructive manner. Calibration model performance was evaluated through ten segments cross validation.

2.3. Materials and methods

2.3.1. Samples

A total of 58 mango samples (cv. *Kent*) selected from three different origins (Brazil, Spain, and Israel) were purchased at local market in Göttingen, Germany. These samples were stored at ambient temperature of 25°C and measured every 2 days (0, 2, 4, 6, 8 and 10) in order to have samples with varied SSC, TA and AA. Day 0th was the day when all samples mango were purchased and initial day of measurement for NIR spectra, SSC, TA and AA. For each measurement day, ten mango samples were analyzed except for the day of 10th, it was eight remaining samples. NIR spectra of samples were recorded before SSC, TA and AA measurement and they were performed in the same day or maximum one day after.

2.3.2. Spectra acquisition

NIR spectra data of all samples were acquired using a benchtop Fourier transform near infrared (FT-NIR) instrument (Thermo Nicolet, Antaris model MDS-method development sampling). High resolution (2 nm interval) sample measurement with integrating sphere was chosen as a basic measurement in this study. Background spectra correction was performed every hour automatically. Sample mangoes were placed manually upon the measurement window of the integrating sphere (1 cm of diameter) of the light source to ensure direct contact and eliminate noise. Diffuse reflectance (Log 1/R) spectra in wavelength range of 1000 – 2500 nm with 2 nm resolution were acquired 64 times and averaged (Figure 2.1). Taking a potential variation of the quality attributes within the fruit into account, spectra were collected in six different points of each sample (two in the left and right edge, and four in the center). The mean values of these measurements were noted as single spectrum with a total of 1557 data points for one sample mango that were used for further analysis.

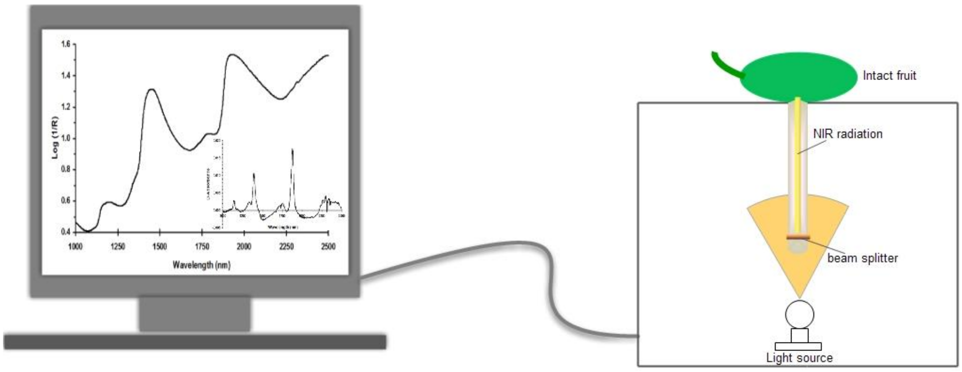

Figure 2.1. Experimental setup for near infrared spectroscopy on intact mango.

2.3.3. Soluble solids content, titratable acidity and ascorbic acid measurement

After collecting and recording the spectra, each sample fruit was sliced at the same marked location of the NIR acquisition and the pulp was taken. The ascorbic acid (AA) was analyzed firstly since this quality attribute is susceptible to oxidation by atmospheric oxygen after slicing (Cozzolino et al., 2011). Titration method was used to determine ascorbic acid using

2.6 Dichlorophenolindophenol solution (Arya et al., 2000). Five grams of pulp sample was macerated and mixed with 20 ml of 5% meta-phosphoric acid (Roth, Germany) into a beaker to prevent oxidation. It was then homogenized using the ultra-turrax (IKA T 18B, Germany) for about two minutes. Distilled water was added to the solution until 50 ml of volume was reached, then filtered through filter paper (MN $615_{1/4}$ with diameter of 150 mm, Macherey-Nagel, Germany). Ten ml of the filtrate was taken and transferred into a 25 ml beaker glass and was titrated with 0.064 M 2.6 Dichlorophenolindophenol. The ascorbic acid, expressed in $mg \cdot 100g^{-1}$ fresh mass (FM), was quantified based on its reaction with this solution as an indicator in titration method. The titration was stopped when light red (pink) color is appeared.

Soluble solids content (SSC) and titratable acidity (TA) measurement were carried out simultaneously by making another juice from 20 grams of pulp sample and maximum 100 ml distilled water. In order to obtain clarified sample juice and separate suspended solids, the centrifuge (20^oC, 10 000 g) was applied for about 10 minutes (Schmilovitch et al., 2000). A single drop filtered supernatant juice was squeezed and dropped onto a hand-held analog refractometer (model HRO32, Krüss Optronic GmbH) to record SSC as ^{o}Brix (Xiaobo et al., 2007) whilst automatic titration (Titroline 96, Schott) with 0.1 N NaOH to an end point of pH 8.1 was used to measure TA expressed as $mg \cdot 100g^{-1}$ fresh mass (Flores et al., 2009). All these three quality attributes were measured in duplicate and averaged.

2.3.4. Spectra calibration and validation

Spectra were analyzed using The Unscrambler® X version 10.2 Network Client (CAMO software AS, Oslo-Norway). Prior to further analysis, spectra were visually inspected for typical spectra features. Principal component analysis (PCA) was then applied to the untreated (raw) spectra in order to explore spectral similarities among samples and to detect outliers by subjecting a Hotelling T^2 ellipse as shown in Figure 2.2. Data points (representing samples) outside this ellipse were marked as spectral outliers and deleted (Constantinou et al., 2004; Mouazen et al., 2010).

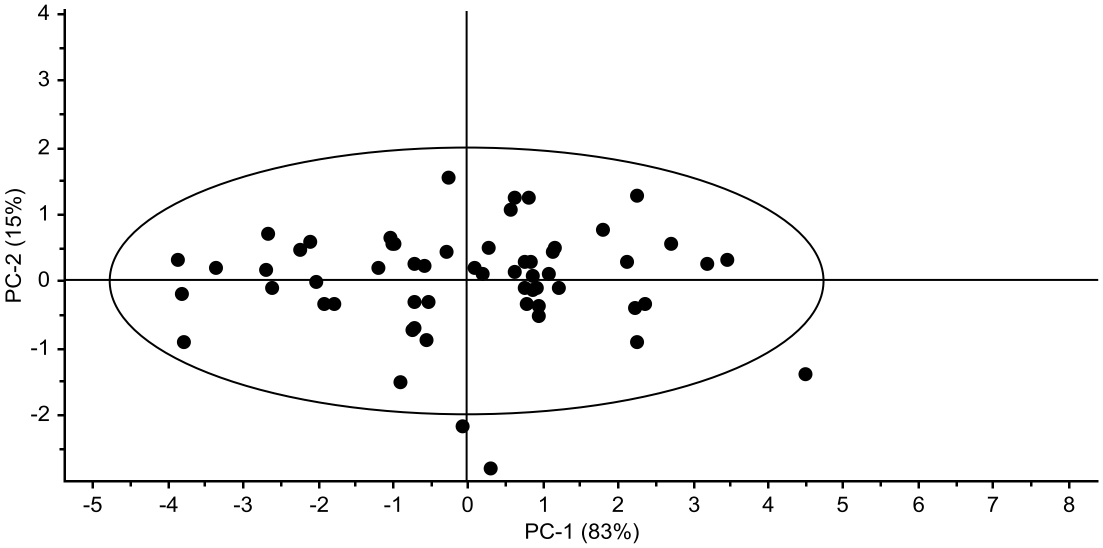

Figure 2.2. Hotelling T^2 ellipse applied to the raw spectra of 58 samples after principal component analysis for outlier detection.

Prior to calibration model development, spectra pre-processing was performed to eliminate noise and scattering. Multiplicative scatter correction (MSC) and standard normal variate transformation (SNV) algorithms followed by Savitzky-Golay smoothing (three smoothing points) were applied to correct additive (baseline shift) and multiplicative scatter effects (Liu, et al., 2010; Cozzolino, et al., 2011). First derivative spectra were obtained using Savitzky-Golay algorithm (three smoothing points, 2nd polynomial order) for both MSC and SNV spectra. It was also used as spectra pre-treatment since spectra transformation into its derivative also has the ability for correcting both additive and multiplicative effects (Naes, et al., 2004). Thus, four different spectra pre-treatments (MSC, SNV, D_1+MSC and D_1+SNV) were used prior to model development.

Calibration models were established to predict SSC, TA and AA using principal component regression (PCR) and partial least squares regression (PLSR). Full cross validation with ten random segments was applied during model development to quantify the model performance and to prevent over fitting. Predictive capabilities of these calibration models and their validation were evaluated by using several statistical parameters, i.e. (i) the coefficient of determination (R^2) of calibration and validation representing the proportion of variance (fluctuation) of the response variable that is explained by the spectral features in the calibration or validation model. It also measure how certain one can be in making predictions from a certain models (Nicolai et al., 2007), (ii) the prediction error which is defined as the root mean square error of calibration (RMSEC), standard error of calibration (SEC), root mean square error of cross validation prediction (RMSECV), (iii) the error difference between RMSEC and RMSECV (Jha, et al., 2006; Flores, et al., 2009), and (iv) the residual predictive deviation (RPD) providing the ratio between the standard deviation of the target variable and the standard error of prediction performance RMSECV or RMSEP. RPD is a commonly used to interpret and compare NIR calibration models (Fearn, 2002; Kapper, et al., 2012). The higher the RPD, the greater is the probability of the model to predict desired chemical constituent in samples set accurately (Sinelli et al., 2008).

Finally, the number of factors or latent variables used in the prediction models was also taken into account since they represent the main spectral variance. Fewer latent variables are preferable to avoid modeling noise signal (Schmilovitch et al., 2000; Nicolai et al., 2007). Apparently, the ideal model should have a high R^2 using a few latent variables, a high RPD, a low error prediction value (RMSEC, RMSECV or RMSEP) and small difference between RMSEC and RMSECV or RMSEP.

2.4. Results and discussion

2.4.1. Spectra features

A typical diffuse reflectance spectrum and its first derivative for intact mango in the NIR region (1000-2500 nm) are shown in Figure 2.3. The NIR spectrum indicates the presence of organic materials as derived from the bands that result from the interaction of molecular bonds of O-H, C-H, C-O and N-H with the incident radiation (Cen and He, 2007). These bonds are subject to vibrational energy changes in which two vibration patterns exist in these bonds including stretch vibration and bend vibration. Here, the presence of strong water absorbance bands was observed at around 1460 nm and 1930 nm because of O-H tone combination and its first overtone. Absorption bands at around 1400 nm and 1900 nm were previously assigned to water absorption (Workman and Weyer, 2008). Moreover, the absorption bands in the range of 2200 - 2300 nm are suggested to be related to C-H-O structures such as glucose, fructose, vitamin A and C; whilst absorption bands at around 1400, 1800 and 2100 nm are associated with organic acids (Cen and He, 2007).

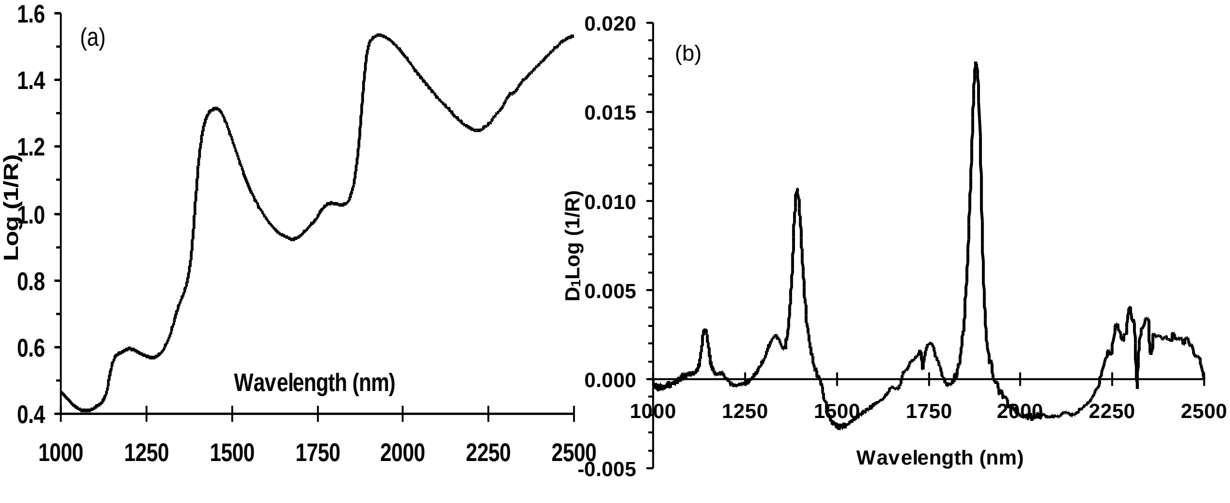

Figure 2.3. Typical diffuse reflectance spectra (a) and first derivative (b) of one intact mango after multiplicative scatter correction.

SSC, TA and AA are organic molecules, and contain bonds of C-H, O-H, C-O and C-C. Thus, it appears feasible to apply NIR methods to predict these quality attributes in intact mango. In overall, the spectral pattern of mango is similar to that for mandarin (Gomez et al., 2006), apple (Xiaobo et al., 2007), blueberries (Sinelli et al., 2008), apricot (Bureau et al., 2009) and tomato (Flores et al., 2009). Important wavelengths of near infrared spectra for desired quality attributes predictions can be identified by inspecting regression coefficients of the respective PCR or PLSR models. The size of the coefficients indicates which independent variables (wavelengths) significantly impact on the response variables (referred quality attributes). Wavelengths having higher absolute value of regression coefficient than others, located at the peaks or valleys of the spectrum, were noted as the important or effective wavelengths that contributed more to the models (Liu et al., 2008).

2.4.2. Calibration and prediction of quality attributes

Descriptive statistics for actual measured SSC, TA and AA are given in Table 2.2. Since the number of samples in this study was limited, no independent prediction using external samples was performed. Instead, the robustness of calibration models was evaluated using internal samples through cross validation. Two different calibration models (PCR and PLSR) using the whole spectral range for each pre-processing method (MSC, SNV and first derivative) were developed for predicting SSC, TA and AA of intact mango.

Table 2.2. Descriptive statistics for measured soluble solids content (SSC), titratable acidity (TA) and ascorbic acid (AA)

	SSC (^{o}Brix)	TA (mg·100g^{-1} FM)	AA (mg·100g^{-1} FM)
number of samples	58	58	58
min	9	189.72	28.93
max	20	772.77	35.66
mean	13.39	495.48	32.21
range	11	583.05	6.73
SD	2.43	131.07	1.31

SD: standard deviation, FM: fresh mass.

During model development, PCA was firstly applied to detect sample spectra that can affect the model performance, i.e. spectral outliers. Two samples for SSC, one sample for TA and three samples for AA were left out due to their potential bad influences over the models. In agreement with Xie et al. (2011), sample outliers may contain valuable information but they may also be non-representative samples that could introduce errors to a model.

2.4.2.1. Soluble solids content

Calibration and validation results for determining soluble solids content (SSC) using NIR spectra are shown in Table 2.3. In general, both PCR and PLSR regressions provide better and robust models using smoothed MSC or SNV treated spectra compared to using first derivative spectra. Though given higher coefficient of determination (R^2) and lower errors, derivative spectra involved more latent variables (LVs) and larger error differences than non-derived spectra in all cases. The maximum error difference for non-derived spectra was 0.13 while derivative spectra produced maximum error difference of 0.69. Liu et al. (2008) also reported that non-derived log (1/R) spectra perform better than derived spectra (first and second derivative) to predict SSC in pears with coefficient of correlation and root mean square error of prediction (RMSEP) values were 0.91 and 0.66 ^{o}Brix respectively.

Table 2.3. Calibration statistics for the prediction of soluble solids content, ^{o}Brix

Method	Spectra treatment	LVs	Calibration			Cross validation			Error difference
			R^2	RMSEC	SEC	R^2	RMSECV	RPD	
PCR	MSC	3	0.67	1.36	1.37	0.65	1.43	1.70	0.07
	SNV	3	0.66	1.36	1.37	0.65	1.43	1.70	0.07
	D_1+MSC	7	0.69	1.30	1.31	0.51	1.67	1.46	0.37
	D_1+SNV	7	0.68	1.34	1.36	0.54	1.68	1.45	0.34
PLSR	MSC	3	0.67	1.34	1.36	0.62	1.47	1.65	0.13
	SNV	3	0.67	1.34	1.36	0.65	1.42	1.71	0.08
	D_1+MSC	5	0.90	0.75	0.75	0.64	1.44	1.69	0.69
	D_1+SNV	5	0.91	0.71	0.72	0.66	1.35	1.80	0.64

LVs: number of latent variables, R^2: coefficient of determination, RMSEC: root mean square error of calibration, RMSECV: root mean square error of cross validation, SEC: standard error of calibration, RPD: ratio prediction to deviation, Error difference: absolute value of (RMSECV-RMSEC), PCR: principal component regression, PLSR: partial least square regression, MSC: multiplicative scatter correction, SEC: standard error calibration, SNV: standard normal variate.

Partial least square regression using SNV treated spectra yielded the best model for SSC prediction with R^2 of calibration 0.67, root mean square error cross validation (RMSECV) of 1.42 ^{o}Brix, error difference of 0.08 ^{o}Brix and residual predictive deviation (RPD) of 1.71 (Figure 2.4a). Similar finding also reported by Jha et al. (2012) where SSC of seven major mango varieties in India (*Chausa, Langra, Kesar, Neelam, Dashehri, Mallika and Malda*) could be predicted with reasonable accuracy by PLSR using second derivative NIR transmittance spectra in which maximum correlation coefficient of calibration and validation was 0.782 and 0.762, respectively.

The standard error of calibration (SEC) in the present study was slightly better than that obtained by Schmilovitch et al. (2000) in mango (cv. *Tommy Atkins*) with SEC of 1.81 ^{o}Brix using 4 latent variables, yet the R^2 values of studies (0.86) was better than in the present study (0.67). SSC is correlated with soluble sugar; it is expected to be represented by spectral features related to C-H-O absorption bands. Based on the regression coefficient curve, the observed important or relevant wavelengths for the SSC calibration model were 1081, 1940, 2250, 2315 and 2355 nm (Figure 2.4b). The highest relevant wavelength proportion was found at wavelength range from 2200 to 2400 nm. This is an agreement with Cen and He (2007) that near infrared absorption bands at 2000 – 2300 nm were associated with C-H-O compositions like sugars or carbohydrates.

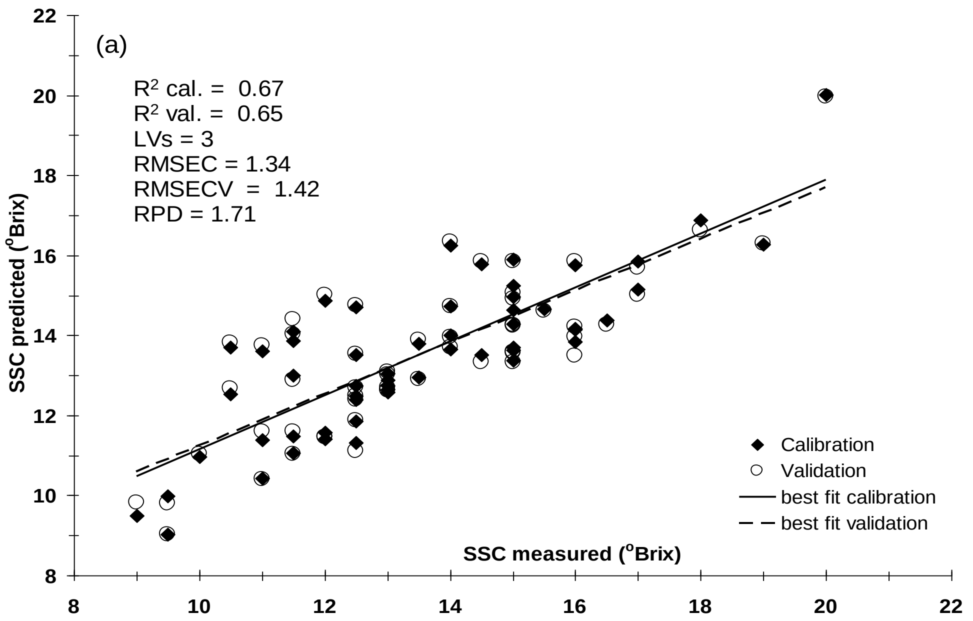

Figure 2.4a. Scatter plot of measured versus predicted soluble solids content based on partial least square regression-standard normal variate model.

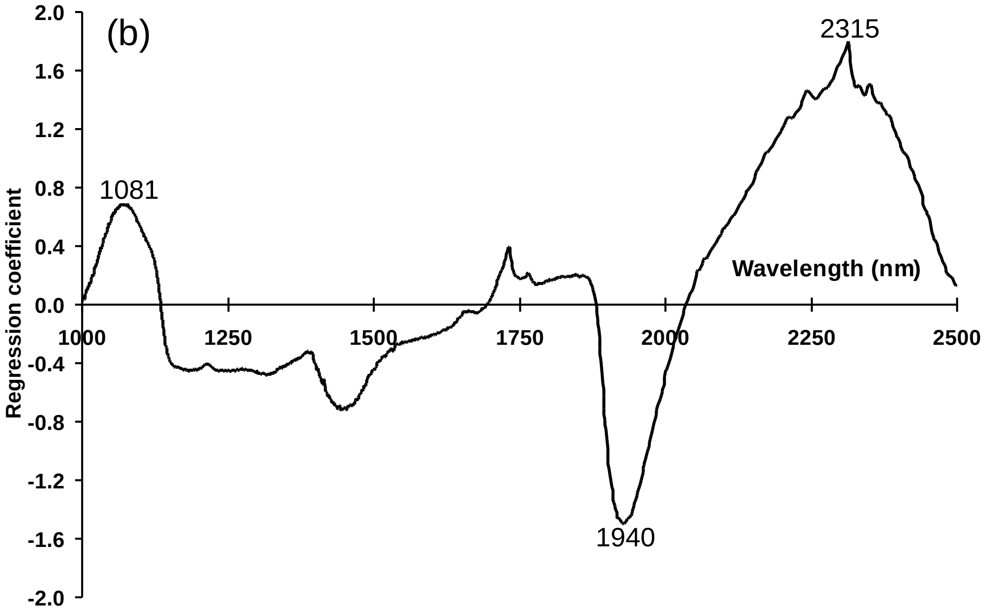

Figure 2.4b. Important wavelengths for soluble solids content predictions derived from regression coefficients plot based on partial least square regression-standard normal variate model.

2.4.2.2. Titratable acidity

Titratable acidity (TA) calibration and cross validation results are presented in Table 2.4. PLSR based on MSC spectra calibration was found best model to predict TA of intact mango fruit. The model required only 3 latent variables of PLS and achieved R^2 coefficient of cross validation of 0.95 and yielded the lowest error (RMSECV) of 26.94. The RPD index based on this model was also the most robust (4.87).

Table 2.4. Calibration statistics for the prediction of titratable acidity, mg·100g^{-1} FM

Method	Spectra treatment	LVs	Calibration			Cross validation			Error
			R^2	RMSEC	SEC	R^2	RMSECV	RPD	difference
PCR	MSC	3	0.95	27.57	27.82	0.94	29.89	4.39	2.32
	SNV	3	0.95	27.58	27.83	0.95	29.36	4.46	1.78
	D_1+MSC	7	0.94	31.07	31.35	0.91	39.04	3.36	7.97
	D_1+SNV	7	0.94	31.37	34.65	0.90	40.77	3.21	9.40
PLSR	MSC	3	0.96	25.83	26.05	0.95	26.94	4.87	1.11
	SNV	3	0.96	25.82	26.05	0.95	27.44	4.78	1.62
	D_1+MSC	5	0.98	14.95	15.08	0.95	28.84	4.54	13.89
	D_1+SNV	5	0.98	14.98	15.11	0.95	28.57	4.59	13.59

FM: fresh mass, LVs: number of latent variables, R^2: coefficient of determination, RMSEC: root mean square error of calibration, RMSECV: root mean square error of cross validation, SEC: standard error of calibration, RPD: ratio prediction to deviation, Error difference: absolute value of (RMSECV-RMSEC), PCR: principal component regression, PLSR: partial least square regression, MSC: multiplicative scatter correction, SEC: standard error calibration, SNV: standard normal variate.

Coefficients of determination in all cases (MSC and first derivative MSC) of PLSR were found to be 0.96 and 0.98 for calibration and 0.95 for cross validation. Like SSC, calibration models based on non-derived spectra were preferable providing fewer numbers of LVs and smaller error difference compared to derivative spectra. Calibration models based on MSC and SNV required 3 latent variables while derivatives spectra required 5 (PLSR) and 7 (PCR) latent variables to predict TA. Scatter plot drawn from this best selected model (Figure 2.5a) indicate that slope of the curve is near to ideal of 45^o and imply that predicted TA is nearly to that in measured TA. Based on the regression coefficient curve of the best TA calibration model (Figure 2.5b), relevant wavelengths were found at around 1470, 1733, 1895, 2308 and 2355 nm. The last two mentioned wavelengths contributed more to the TA prediction model. Cen and He (2007) also found absorption bands of NIR associated with acid compound in the wavelength region of 1400 to 1500 nm, 1850 to 1900 nm and, since many acids are formed in

C-O-H structures, the higher peak of the absorption bands were found mostly in the region 2200 – 2300 nm.

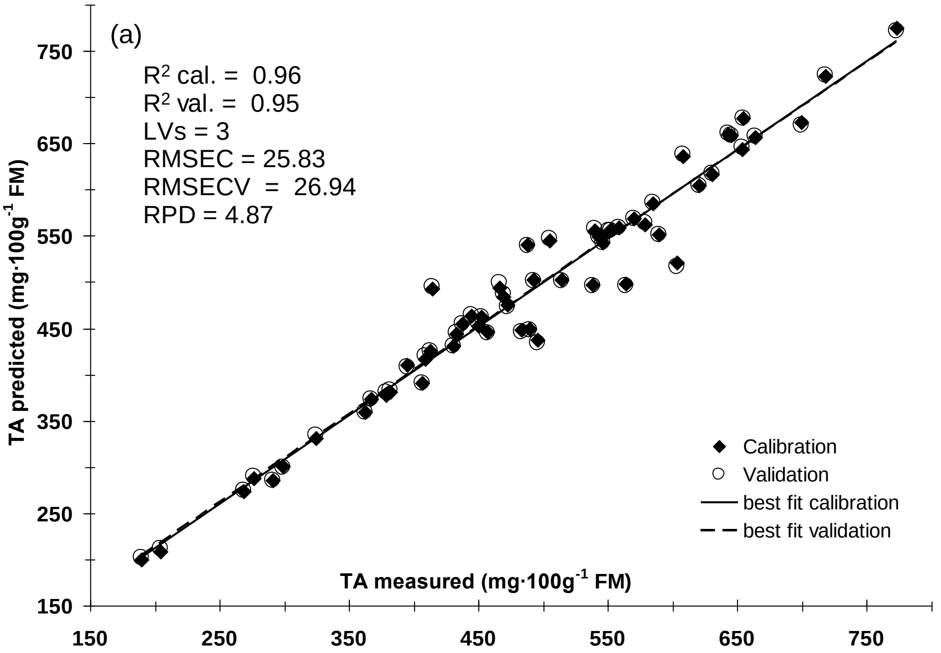

Figure 2.5a. Scatter plot of measured versus predicted titratable acidity based on partial least square regression - multiplicative scatter correction model.

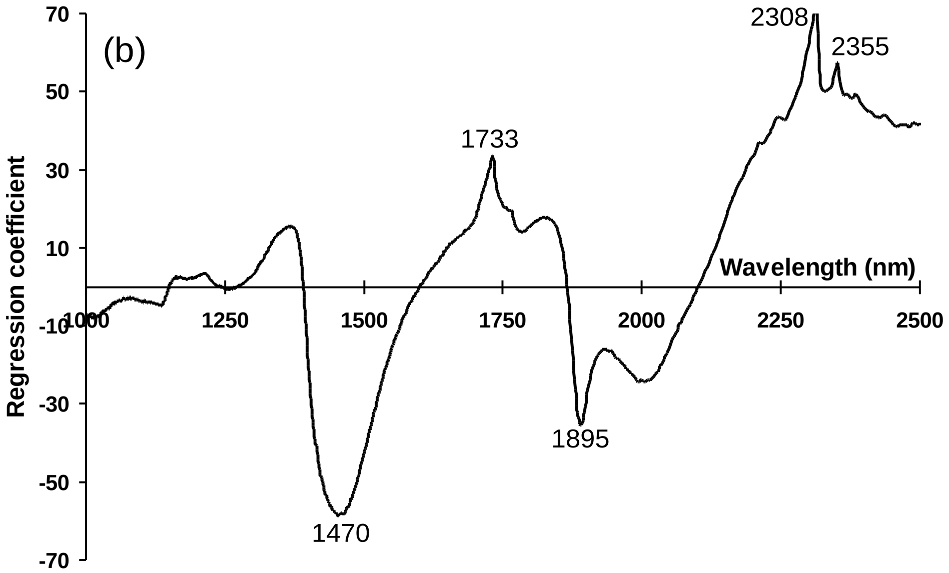

Figure 2.5b. Important wavelengths for titratable acidity predictions derived from regression coefficients plot based on partial least square regression - multiplicative scatter correction model.

A wide range of actual TA in this study was observed and perhaps contributed positively to the model performance since a better model performance may be achieved with a higher sample range.

2.4.2.3. Ascorbic acid

As shown in Table 2.5, PLSR based on MSC spectra with six latent variables was found to be the best model to predict ascorbic acid (AA) content of intact mango with R^2 coefficient of 0.70 and 0.60 for calibration and cross validation respectively. The first derivative of MSC spectra also achieved a good result from which the R^2 of calibration and cross validation were 0.82 and 0.61 respectively. The highest RPD index resulted from this model was 1.70 which means that coarse quantitative prediction are still possible, but the model need some improvements in calibration. Predicted versus measured AA derived from PLSR-MSC model is shown in Figure 2.6a. Furthermore, based on regression coefficients curve from this model, the important relevant wavelengths for ascorbic acid prediction were found at 1930, 2015, 2027, 2300, 2418, 2430 and 2446 nm as presented in Figure 2.6b.

Table 2.5. Calibration statistics for the prediction of ascorbic acid, mg·100g^{-1} FM

Method	Spectra treatment	LVs	Calibration			Cross validation			Error
			R^2	RMSEC	SEC	R^2	RMSECV	RPD	difference
PCR	MSC	7	0.62	0.75	0.75	0.43	0.93	1.41	0.18
	SNV	7	0.63	0.74	0.75	0.53	0.87	1.51	0.13
	D_1+MSC	6	0.67	0.70	0.70	0.58	0.82	1.60	0.12
	D_1+SNV	7	0.69	0.68	0.69	0.59	0.80	1.64	0.12
PLSR	MSC	6	0.70	0.65	0.66	0.60	0.80	1.64	0.15
	SNV	6	0.71	0.65	0.66	0.55	0.84	1.56	0.19
	D_1+MSC	4	0.82	0.51	0.52	0.61	0.77	1.70	0.26
	D_1+SNV	5	0.92	0.33	0.34	0.60	0.80	1.64	0.47

FM: fresh mass, LVs: number of latent variables, R^2: coefficient of determination, RMSEC: root mean square error of calibration, RMSECV: root mean square error of cross validation, SEC: standard error of calibration, RPD: ratio prediction to deviation, Error difference: absolute value of (RMSECV-RMSEC), PCR: principal component regression, PLSR: partial least square regression, MSC: multiplicative scatter correction, SEC: standard error calibration, SNV: standard normal variate.

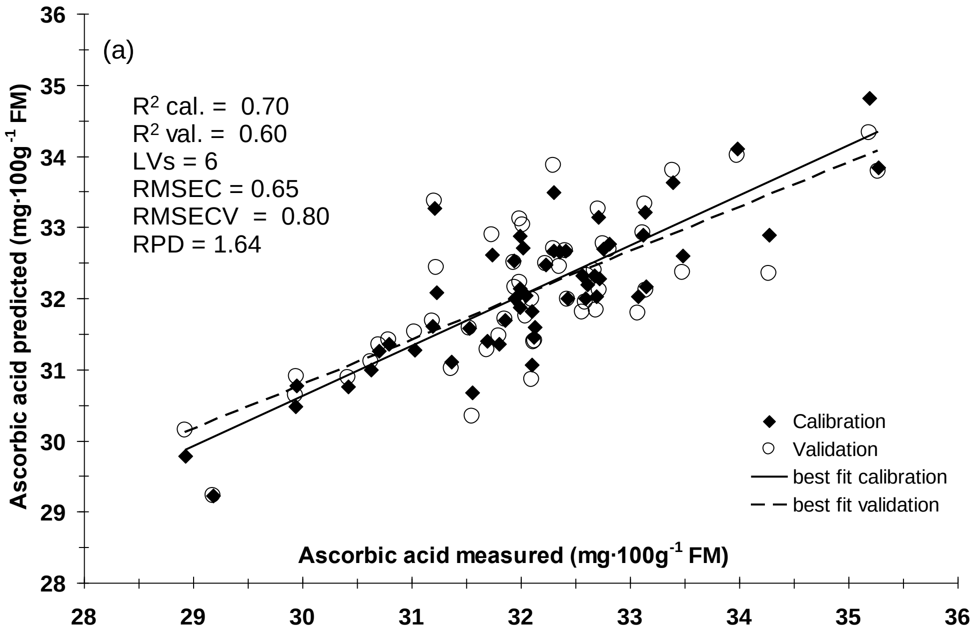

Figure 2.6a. Scatter plot of measured versus predicted ascorbic acid based on partial least square regression - multiplicative scatter correction model.

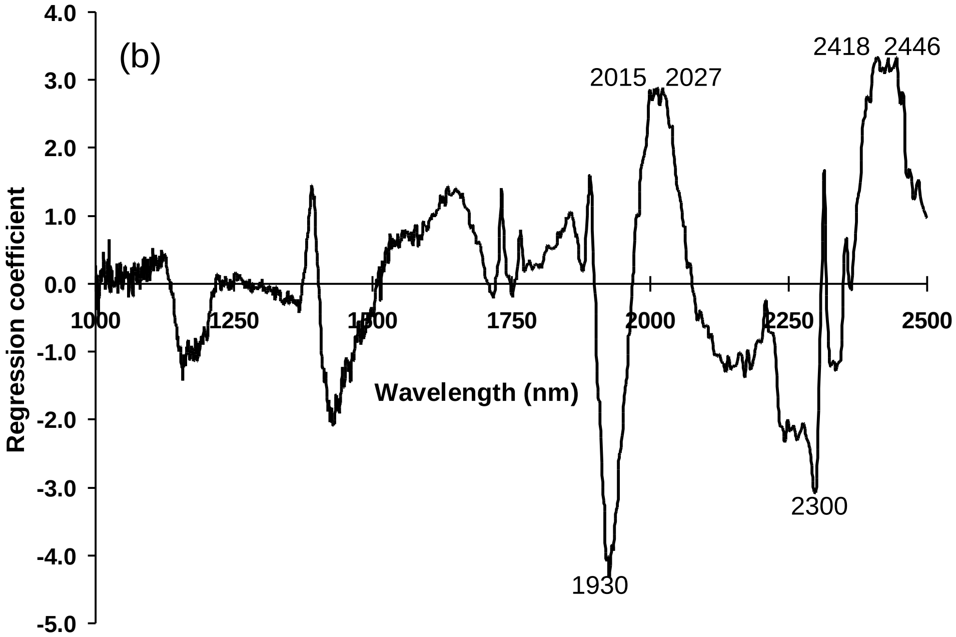

Figure 2.6b. Important wavelengths for ascorbic acid predictions derived from regression coefficients plot based on partial least square regression - multiplicative scatter correction model.

Compared to SSC and TA, prediction of AA yielded unsatisfactory result as to be seen from R^2 coefficient; even R^2 resulted from AA (0.70) prediction was slightly better than that in

SSC prediction (0.67). However, the difference between R^2 calibration and R^2 cross validation was higher in AA (0.10) prediction compare to the R^2 calibration-cross validation difference in SSC prediction (0.05). Furthermore, RPD for AA is the lowest (1.64) indicating that the model provides rather a rough screening and low prediction capability for AA. This might be due to spectral noise: Based on the regression coefficients plot, it is observed that noise is present in the model from the beginning of wavelength 1000 to 1400 nm and 1450 to 1700 nm. Another possible reason is the low AA range of the studied mangos (6.34 mg·100g^{-1}FM). Nevertheless, this prediction model was reasonably reliable to predict ascorbic acid of intact mango.

In general, judging from calibration and cross validation performance, NIRS can predict TA of intact mango with higher accuracy compare to SSC and AA (R^2 = 0.67, RPD = 1.71 for SSC, R^2 = 0.96, RPD = 4.87 for TA and R^2 = 0.70, RPD = 1.64 for AA). An RPD between 1.5 and 1.9 means that the model can discriminate low from high values of the response variable and still need improvement. A value between 2 and 2.5 indicates that quantitative predictions are possible, and can be seen as promising prediction accuracy. An RPD value between 2.5 and 3 or above 3 corresponds to good and excellent prediction accuracy respectively. These values give the average uncertainty that can be expected for predictions of future samples (Williams, 2001; Fearn 2002; Nicolai et al., 2007). In addition, non-derived spectra were found to be preferable than first derivative spectra, even they produce higher R^2 value in calibration and validation, there is a big difference between RMSEC and RMSECV. This indicates that prediction model based on derived spectra is over-fitted, generating good R^2 and RMSEC values, which fits parts of the noise of the calibration set but performing poorly in the future validation set.

The optimal number of latent variables in PCR or PLSR was determined by the lowest value of predicted residual error sum of squares. In non-derived spectra, 3 latent variables were found to be sufficient to explain variance in the model whilst derived spectra required five and 7 latent variables in PCR and PLSR method respectively. A relatively low number of latent variables are desired to avoid the modeling of spectral noise. Within this study non derivative NIR spectral calibration models are preferable. From the regression method point of view, even PCR provided lower error differences between calibration and validation; PLSR appeared to be more robust where coefficient of R^2 and RPD were higher whilst RMSEC, RMSECV were lower compared to PCR.

2.5. Conclusion

Near infrared spectroscopy appears feasible to predict main quality attributes of intact mango namely soluble solids content (SSC), titratable acidity (TA) and ascorbic acid (AA). Spectral pre-processing such as derivative, SNV and MSC, affects the quality of the prediction models using PCR and PLSR. Best prediction models for all quality attributes were achieved using PLSR method combined with non-derived spectra. TA could be predicted with the highest R^2 and lowest prediction error compared to SSC and AA content that could be modeled with reasonable performance. Despite the fact that determination coefficient observed in SSC and AA prediction models are not very high and the RPD for both models are lower than two, It is assumed that NIRS still could be used for sorting and grading of mango fruit. The results are based on samples from commercially purchased mangos and can be considered as a promising technique for the use of sorting and grading mango quality. However improvement should be made to generate a better SSC and AA prediction model performance. Thus, NIRS combined with appropriate multivariate analysis could become a rapid and non-destructive alternative method for determining inner quality attributes of mango.

References

Arya, S. P., Mahajan, M., & Jain, P. (2000). Non-spectrophotometric methods for the determination of vitamin C. *Analytica Cimica Acta,* 417, 1-14.

Brereton, R. G. (2000). Introduction to multivariate calibration in analytical chemistry. *The Analyst*, 125, 2125−2154.

Bureau, S., Ruiz, D., Reich, M., Gouble, B, Bertrand, D., Audergon, J. M., Renard, C. M. (2009). Rapid and non-destructive analysis of apricot fruit quality using FT-near-infrared spectroscopy. *Food Chemistry,* 113, 1323-1328.

Camps, C., & Christen, D. (2009). Non-destructive assessment of apricot fruit quality by portable visible-near infrared spectroscopy. *LWT - Food Science and Technology,* 42, 1125–1131.

Cayuela, J. A. (2008). Vis/NIR soluble solids prediction in intact oranges (*Citrus sinensis* L.) cv. Valencia Late by reflectance. *Postharvest Biology and Technology,* 47, 75–80.

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology*, 18, 72-83.

Clark, R. N. (1999). Spectroscopy of rocks and minerals and principles of spectroscopy. In Rencz, A. N (Ed.), *Manual of remote sensing*. (*Volume 3): Remote sensing for the earth sciences*, (pp. 3-58). New York: John Wiley and Sons, (Chapter 1).

Constantinou, M. A., Papakonstantinou, E., Benaki, D., Spraul, M., Shulpis, K., Koupparis, M. A., & Mikros, A. (2004). Application of nuclear magnetic resonance spectroscopy combined with principal component analysis in detecting inborn errors of metabolism using blood spots: a metabonomic approach. *Analytica Cimica Acta*, 511, 303-312.

Cozzolino, D., Cynkar, W. U., Dambergs, R. G., Shah, N., & Smith, P. (2009). Multivariate methods in grape andwine analysis. *International Journal of Wine Research*, 1, 123−130.

Cozzolino, D., Cynkar, W. U., Shah, N., & Smith, P. (2011). Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Research International*, 44, 1888-1896.

Fan, G., Zha, J., Du, R., & Gao, L. (2009). Determination of soluble solids and firmness of apples by Vis/NIR transmittance. *Journal of Food Engineering*, 93, 416-420.

Fearn, T. (2002). Assessing calibrations: SEP, RPD, RER and R2. *NIR News*, 13, 12−14.

Flores, K., Sanchez, M. T., Perez-Marin, D., Guerrero, J. E., & Garrido-Varo, A. (2009). Feasibility in NIRS instruments for predicting internal quality in intact tomato. *Journal of Food Engineering*, 91, 311-318.

Gomez, A. H., He, Y., & Pereira, A. G. (2006). Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using Vis/NIR-spectroscopy techniques. *Journal of Food Engineering*, 77, 313–319.

Jaiswal, P., Jha, S. N., & Bharadwaj, R. (2012). Non-destructive prediction of quality of banana using spectroscopy. *Scientia Horticulturae*, 135, 14-22.

Jha, S. N., Kingsly, A. R. P., & Chopra, S. (2006). Non-destructive determination of firmness and yellowness of mango during growth and storage using visual spectroscopy. *Biosystems Engineering*, 94, 397-402.

Jha, S. N., Narsaiah, K., Sharma, A. D., Singh, M., Bansal, S., & Kumar, R. (2010). Quality parameters of mango and potential of non-destructive techniques for their measurement-a review. *J Food Sci Technol*, 47, 1-14.

Jha, S. N., Jaiswal, P., Narsaiah, K., Gupta, M., Bhardwaj, R., & Singh, A. K. (2012). Non-destructive prediction of sweetness of intact mango using near infrared spectroscopy. *Scientia Horticulturae*, 138, 171-175.

Kapper, C., Klont, R.E., Verdonk, J.M.A.J., & Urlings, H.A.P. (2012). Prediction of pork quality with near infrared spectroscopy (NIRS) 1.Feasibility and robustness of NIRS measurements at laboratory scale. *Meat Science*, 91, 294-299.

Khalaf, A., Bennedsen, N. B., & Bjorn, B. (2004). Distinguishing carrot’s characteristics by near infrared reflectance and multivariate data analysis. *CIGR Journal of Scientific Research and Development*, 6, 342-359.

Lamertyn, J., Nicolai, B. M., Ooms, K., de Smedt., & Baerdemaeker, J. (1998). Non-destructive measurement of acidity, soluble solids and firmness of Jonagold apples using NIR spectroscopy. *Trans. ASAE*, 41, 1089-1094.

Liu, F., He, Y., & Wang, L. (2008). Determination of effective wavelengths for discrimination of fruit vinegars using near infrared spectroscopy and multivariate analysis. *Analytica Cmimica Acta*, 615, 10-17.

Liu, Y., Chen, X., & Ouyang, A. (2008). Nondestructive determination of pear internal quality indices by visible and near-infrared spectrometry. *LWT - Food Science and Technology*, 41, 1720–1725.

Liu, Y., Sun, X., & Ouyang, A. (2010). Nondestructive measurement of soluble solid content of navel orange fruit by visible–NIR spectrometric technique with PLSR and PCA-BPNN. *LWT - Food Science and Technology*, 43, 602–607.

Mahayothee, B., Neidhart, S., Leitenberger, M., Mühlbauer, W., & Carle R. (2004). Non-destructive determination of maturity of Thai mangoes by near infrared spectroscopy (NIR). *Acta Horticulturae*, 645, 581-588.

Mendoza, F., Lu, R., & Cen, H. (2012). Comparison and fusion of four nondestructive sensors for predicting apple fruit firmness and soluble solids content. *Postharvest Biology and Technology*, 73, 89-98.

Mouazen, A. M., Kuang, B., De Baerdemaeker, J., & Ramon, H. (2010). Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma*, 158, 23–31.

Moghimi, A., Aghkhani, M. H., Sazgarnia, A., & Sarmad, M. (2010). Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of kiwifruit. *Biosystems Engineering*, 106, 295-302.

Naes, T., Isaksson, T., Fearn, T., & Davies, T. (2004). A user-friendly guide to multivariate calibration and classification. Chichester, UK: NIR publications.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lamertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. *Postharvest Biology and Technology*, 46, 99-118.

Pontes, M.J.C., Santos, S.R.B., Araújo, M.C.U., Almeida, L.F., Lima, R.A.C., Gaião, E.N., & Souto, U.T.C.P. (2006). Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometry. *Food Research International*, 39, 182-189.

Sánchez, M. T., De la Haba, M. J., Benítez-López, M., Fernández-Novales, J., Garrido-Varo, A., & Pérez-Marín, D. (2012). Non-destructive characterization and quality control of intact strawberries based on NIR spectral data. *Journal of Food Engineering*, 110, 102-108.

Saranwong, S., Sornsrivichai, J., & Kawano, S. (2004). Prediction of ripe-stage eating quality of mango fruit from its harvest quality measured nondestructively by near infrared spectroscopy. *Postharvest Biology and Technology*, 31, 137-145.

Schmilovitch, Z., Mizrach, A., Hoffman, A., Egozi, H., & Fuchs, Y. (2000). Determination of mango physiological indices by near-infrared spectrometry. *Postharvest Biology and Technology*, 19, 245-252.

Shao, Y., He, Y., Gomez, A. H., Pereira, A. G., Qiu, Z., & Zhang, Y. (2007). Visible/near infrared spectrometric technique for nondestructive assessment of tomato 'Heatwave' (*Lycopersicum esculentum*) quality characteristics. *Journal of Food Engineering*, 81, 672–678.

Sinelli, N., Spinardi, A., Di Egidio, V., Mignani, I., & Casiraghi, E. (2008). Evaluation of quality and nutraceutical content of blueberries (*Vaccinium corymbosum* L.) by near and mid-infrared spectroscopy. *Postharvest Biology and Technology*, 50, 31-36.

Subedi, P. P., Walsh, K. B., & Owens, G. (2007). Prediction of mango eating quality at harvest using short-wave near infrared spectrometry. *Postharvest Biology and Technology*, 43, 326-334.

Valente, M., Leardi, R., Self, G., Luciano, G., & Pain, J. P. (2009). Multivariate calibration of mango firmness using vis/NIR spectroscopy and acoustic impulse method. *Journal of Food Engineering*, 94, 7-13.

Venture, M., de Jager, A., de Putter, F. H., & Roelofs, F. P. (1998). Non-destructive determination of soluble solids in apple fruit by near infrared spectroscopy (NIRS). *Postharvest Biology and Technology*, 14, 21-27.

Wang, L., Lee, F.S.C., Wang, X., & He, Y. (2006). Feasibility study of quantifying and discriminating soybean oil adulteration in camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR. *Food Chemistry*, 95, 529-536.

Workman, J., & Weyer, L. (2008). Practical guide to interpretative near infrared Spectroscopy. Taylor and Francis, Boca Raton, USA: CRC Press 332.

Xiaobo, Z., Jiewen, Z., Xingyi, H., & Yanxiao, L. (2007). Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji' apple based on different PLS models. *Chemometrics and intelligent laboratory systems*, 87, 43-51.

Xie, L., Ye, X., Liu, D., & Ying, Y. (2011). Prediction of titratable acidity, malic acid, and citric acid in bayberry fruit by near-infrared spectroscopy. *Food Research International*, 44, 2198-2204.

Xudong, S., Hailiang, Z., & Yande, L. (2009). Nondestructive assessment of quality of nanfeng mandarin fruit by a portable near infrared spectroscopy. *Int J Agric & Biol Eng*, 2, 65-71.

Chapter 3. Comparisons of different spectra pre-processing to predict quality attributes of intact mango using near infrared spectroscopy and partial least square regression

3.1. Abstract

Spectra pre-processing is one of the main factors affecting model robustness and prediction capabilities with near-infrared spectroscopy (NIRS). The objective of the present study is to compare the performance of untreated and six different treated spectra methods, namely mean centering (MC), normalization, de-trending (DT), multiplicative scatter correction (MSC), standard normal variate (SNV) and orthogonal signal correction (OSC), through partial least square (PLS) regression for the prediction accuracy of quality attributes in form of titratable acidity (TA) and soluble solids content (SSC) of intact mango. A total of 99 samples mango (cv. *Kent*) were used as data-set for calibration, cross validation (CV) and prediction procedures. Diffuse reflectance spectra (log1/R) were acquired and recorded in wavelength range of 1000 – 2500 nm by Antaris Fourier transform NIR instrument whilst TA and SSC were measured by titration and refraction index method respectively. Judging from calibration and prediction performance, OSC found to be the best pre-processing method prior to PLS model development (R^2 CV = 0.923, R^2 prediction = 0.718, RMSECV = 38.869 mg·100g^{-1}, RMSEP = 52.363 mg·100g^{-1}) for TA and (R^2 CV = 0.828, R^2 prediction = 0.605, RMSECV = 0.533 ^{o}Brix, RMSEP = 0.716 ^{o}Brix) for SSC. It may conclude that proper NIR spectra pre-processing combine with PLS multivariate calibration method may lead to good prediction accuracy.

Keywords: NIR, quality parameter, spectroscopy, spectra pre-processing.

3.2. Introduction

In last two decades, near infrared reflectance spectroscopy (NIRS) technique have become more attractive and widely used for quality control analysis in many field sectors including agro-food industries due to its advantage characteristics; nearly without sample preparation, rapid, low labor and non-destructive (Perez-Marin et al., 2007). It also allows several quality parameters can be analyzed simultaneously in the same time. NIR spectra of fruits and vegetables are characterized by a large set of overtones and combination bands representing light absorption by organic compound of specific quality parameter (i.e sugar, acid or amino) of the biological object (Nicolai et al., 2007). Therefore, multivariate analysis is required to extract the information about quality attributes buried on near infrared spectra through a process called calibration modeling from which a mathematical relationship between NIR spectra and the measured quality parameter will be revealed to determine desired quality attributes.

Partial least squares regression (PLSR) is one of the most popular methods for multivariate calibration of NIR spectra data. Unlike principal component regression (PCR), the PLSR takes both X (NIR spectra data) and Y (desired quality attributes) matrices into account when developing the models to find the *latent (or hidden)* variables in X that will best predict the latent variables in Y. PLSR maximizes the covariance between X and Y. In this case, convergence of the system to a minimum residual error is often achieved in fewer factors than using PCR. This is in contrast to PCR, which first performs principal component analysis (PCA) on X and then regresses the PCA scores with the Y data. PLSR also leading to the reduction of the number of latent variable or principal component (in PCR) that are needed to describe the model compared to PCR (Felizardo et al., 2007; CAMO The Unscrambler, 2012).

The NIR spectra of biological objects is influenced by wavelength dependent scattering effects, tissue heterogeneities, instrumental noise, ambient effects and other source of variability. In order to eliminate these variations unrelated to desired chemical concentrations, NIR spectra should be conveniently pre-processed prior to calibration model development. Several pre-processing methods were applied to overcome these effects such as spectra smoothing, standardization, normalization, derivation, wavelet transforms, Fourier transform, orthogonal signal correction, net analyte signal and combination among them

(Pontes et al., 2006; Wang et al., Cen and He, 2007; Cozzolino et al., 2011; Blanco et al., 2001; Azzouz et al., 2003; Chen et al., 2003). These pre-processing methods may produce good and accurate prediction model.

The main aim of this study is to compare the prediction model performance of untreated spectra and six different treated spectra methods, namely mean centering (MC), mean normalization (MN), de-trending (DT), multiplicative scatter correction (MSC), standard normal variate (SNV) and orthogonal signal correction (OSC). The models were developed through partial least square regression (PLSR) for quality attributes prediction in form of titratable acidity (TA) and soluble solids content (SSC) of intact mango. PLSR was chosen based on our previous study results on the feasibility of NIRS for SSC, TA and ascorbic acid prediction of mango fruit concluded that PLSR was slightly better than PCR (Munawar et al., 2013).

3.3. Materials and methods

3.3.1. Experimental

Laboratory works were performed in order to acquire NIR spectra data, SSC and TA from a total of 99 mangos (cv. *Kent*) obtained from local market. The NIR diffuse reflectance of mango samples were recorded in wavelength range of 1000 – 2500 nm using a benchtop Fourier transform (FT) NIR instrument (Thermo Nicolet, Antaris model MDS-method development sampling) with the aid of the Thermo® integrated software: Thermo Integration and Thermo Operation. The workflow has been set up to run specified tasks of the instrument. High resolution sample measurement with integrating sphere was chosen as a basic measurement. Background spectra correction was set to be performed every hour automatically. Samples were placed manually upon the measurement window of the integrating sphere (one cm of diameter) of the light source to ensure direct contact and eliminate noise. The spectra of each sample were collected in six different points and averaged (two in the left and right edge, and four in the center) representing acidity and sweetness distribution of mango fruit.

After spectra acquisition, each sample fruit was sliced and the pulp was taken. 20 grams of pulp sample was macerated, mixed and homogenized with maximum 100 ml distilled water. To obtain clarified sample juice and separate suspended solids, the centrifuge with 20^{o}C and 10 000 g was applied for about 10 minutes (Schmilovitch et al., 2000). A little filtered juice was squeezed and dropped into a hand-held analog refractometer (model HRO32, Krüss Optronic GmbH) to record SSC in form of ^{o}Brix (Xiaobo et al., 2007) whilst automatic titration method (Titroline 96, Schott) with 0.1 N NaOH to an end point of pH 8.1 (Flores et al., 2009) was used to measure TA expressed mg·100g^{-1} FM. These both quality attributes measurement were performed in duplicate and averaged.

3.3.2. Outlier detection

Sample outliers were inspected by plotting all 99 original diffuse reflectance spectra onto principal component analysis and subjecting Hotelling T^{2} ellipse. Data points lying outside this ellipse were marked as outliers and deleted (Constantinou et al., 2004; Mouazen et al., 2010). This process was applied until no more outliers were detected. Thus, remaining samples after outlier removal were used to establish calibration models using different spectra pre-processing.

3.3.3. Spectra data pre-processing

Prior to PLSR model development, different spectra data pre-processing were applied to the data. This work presents PLSR calibration obtained from untreated spectra (identified as none) and from six different pre-processing methods as follows:

3.3.3.1. Mean centering (MC)

This is often the simply preferred pre-processing method prior calibration development as it focuses on differences between observations rather than their absolute values. MC ensures that the resulting data or model may be interpreted in terms of variation around the mean. It was applied to all spectra (calibration and prediction).

3.3.3.2. Mean normalization (MN)

The purpose of this pre-treatment is to scale samples in order to get all data on approximately the same scale based on area, mean, range, maximum, peaks and unit vector. All spectra data were also normalized as mean normalization.

3.3.3.3. De-trending (DT)

This pre-treatment method tends to remove nonlinear trends in spectroscopic data. It calculates a baseline function as a least squares fit of a polynomial to the sample spectral data. DT was applied to individual spectra. As the polynomial order of the DT increases, additional baseline effects are removed. Zero-order: offset; first-order: offset and slope; second-order: offset, slope and curvature (CAMO The Unscrambler, 2012). In this work, second-order of DT was applied to all spectra prior to PLSR calibration.

3.3.3.4. Multiplicative scatter correction (MSC)

Multiplicative scatter correction is one of the approaches to reduce *amplification* (multiplicative, scattering) and *offset* (additive, chemical) effects in NIR spectra (Chen et al., 2002). It rotates each spectrum so that it fits as closely as possible to a standard spectrum that may often be the mean spectrum (r_m). Every spectrum is then corrected using the linear equation (Azzouz et al., 2003):

$$r = a + br_\mathrm{m} + v \qquad \text{(eq. 3.1)}$$

whereas *a* and *b* are the correction coefficients computed from a regression of each individual spectrum onto the mean spectrum. Coefficient *a* is the intercept of the regression line indicating a constant linear absorption additive effect, coefficient *b* is the slope indicating the influence of absorption multiplicative effect and *v* is the residuals vector giving the differences between original spectrum (*r*) and mean spectrum (r_m). This residual vector is assumed to contain the chemical variance in r. The new corrected NIR spectrum (r_1) is obtained by means of the expression (Azzouz et al., 2003):

$$\frac{r-a}{b} = r_m + \frac{v}{b} \qquad \text{(eq. 3.2)}$$

$$r_1 = r_m + \frac{v}{b} \qquad \text{(eq. 3.3)}$$

MSC was applied to all spectra with an option full MSC (common amplification and offset removal) prior to calibration and prediction procedures.

3.3.3.5. Standard normal variate (SNV)

Standard normal variate was also applied to all spectral data as a spectra pre-processing method prior to calibration and prediction. It is a transformation which removes scatter effects from spectra by centering and scaling individual spectra. Like MSC, the practical result of SNV is that it removes multiplicative interferences of scatter effects from spectral data. An effect of SNV is that on the vertical scale, each spectrum is centered on zero and varies roughly from –2 to +2. Apart from the different scaling, the result is more-less similar to that of MSC. The practical difference is that SNV standardizes each spectrum using only the data from that spectra, it does not use the mean spectra of any set (CAMO The Unscrambler, 2012).

3.3.3.6. Orthogonal signal correction (OSC)

Orthogonal scatter correction is a relatively new pre-processing method applied to the NIR spectra. The OSC seek to correct X data matrix; NIR spectra data by removing the information from the spectra that is orthogonally uncorrelated to the Y data matrix; reference quality attributes data. This is was done in order to avoid the removal of useful information that is important for modeling, and removes only the irrelevant variation that creates problems for the regression model (Cen and He, 2007; Blanco et al., 2001). This treatment is applied jointly to all the spectra in the calibration set. Later, the correction on the *X* matrix can be applied to an external prediction set to evaluate the prediction ability of the calibration model built with the treated data. The algorithm used in this type of correction is similar to the non-iterative partial least square (NIPALS) algorithm, commonly used in PCA and PLS. In each step of the algorithm, the weight vector (*w*) is modified, imposing the condition that $t = X \cdot w$ is orthogonal to the *Y* matrix, and where *t* is the corresponding score vector. In PLS the condition that weights would be calculated to maximize the covariance among *X* and *Y* is imposed, but in OSC just the opposite is attempted, to minimize this covariance, making *t* as close as possible to the orthogonality with *Y*. The result of this calculation are scores and loadings matrices that contain the information not related to the concentration. Each internal latent variable (score by loading product) removes a part of the *X* matrix variance (Blanco et al., 2001).

3.3.4. Model calibration and prediction

Once spectra pre-processing was completed, calibration models were developed for each quality attributes (SSC and TA) based on raw and pre-processed spectra. Samples were split into calibration datasets and prediction datasets. PLSR calibration with ten-segments full cross validation was applied to establish these models. The capability of calibration models were quantified by predicting quality attributes of external prediction samples.

All of developed models, obtained using different spectra pre-processing, were compared and evaluated by using the following statistical parameters: the coefficient of determination (R^2) of calibration, cross validation and prediction which essentially represents the proportion of explained variance of the response in the calibration or validation data set. Prediction error which is defined as the root mean square error of calibration (RMSEC), root mean square error of cross validation prediction (RMSECV) and root mean square error prediction (RMSEP), the difference between RMSEC and RMSEP (Jha et al., 2006; Flores et al., 2009).

Another statistical parameter commonly used to interpret and compare NIR calibration models is the RPD, defined as the ratio between standard deviation of the reference or actual value of SSC and TA and the RMSEP. The higher the value of RPD, the greater is the probability of the model to predict desired chemical constituent in samples set accurately (Sinelli et al., 2000; Naes et al., 2004). The number of factors or latent variables was also taken into account since they could represent the main features of the spectra, reduce the number of variables and express most of the variance in the data set. Fewer latent variables are preferable to avoid modeling noise signal.

3.3.5. Software

All spectra data pre-processing and calibration model development were carried out using software packages namely The Unscrambler® X version 10.2 network client (CAMO software AS, Oslo-Norway).

3.4. Results and discussion

At first, after obtaining spectra data of all 99 samples, PCA with Hotelling T^2 ellipse were applied to the original log (1/R) spectra for outlier detection. Eight samples were detected as outliers and removed as shown in Figure 3.1. Thus, there were 91 samples used for model development through PLSR method. From these samples, 56 were used for calibration and cross validation whilst remaining 35 samples were used for external prediction.

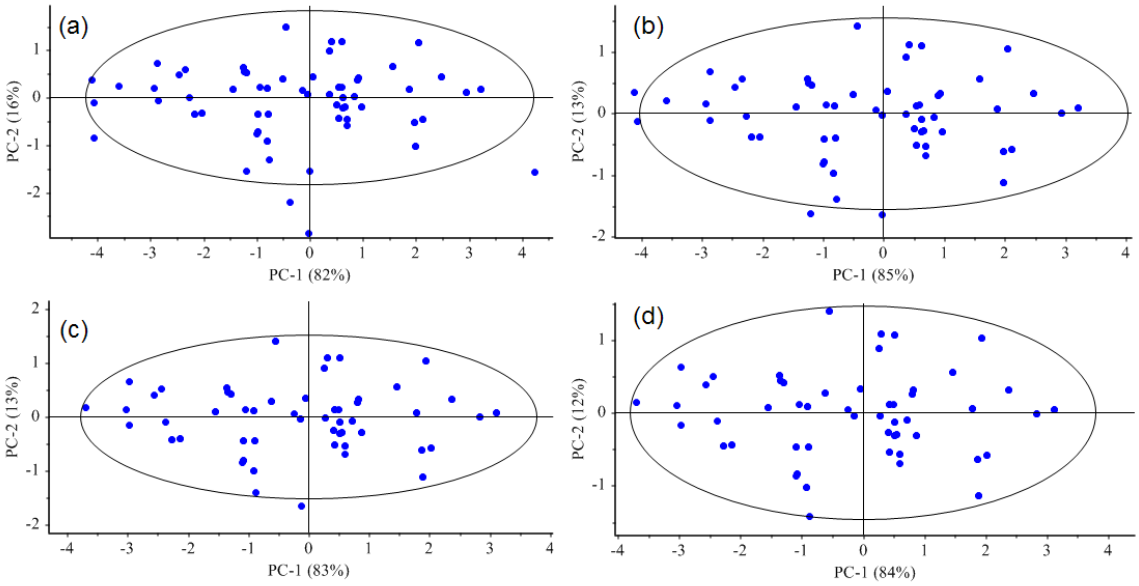

Figure 3.1. Principal component analysis with Hotelling T^2 ellipse for outliers detection; 4 outliers detected (a), 3 outliers (b), 1 outlier (c) and no outlier (d).

In order to observe samples set distribution of these two datasets after splitting, PCA then again was performed to the calibration samples data set and prediction samples one. As expected, prediction samples data sets distributed and located in one cluster within the calibration samples data. Original raw diffuse reflectance spectra of all calibration and prediction samples are shown in Figure 3.2a. The presence of strong water absorbance bands were observed at around 1460 nm and 1930 nm because of O-H tone combination and first overtone of water. Absorption bands at around 1400 nm and 1900 nm were noted to be associated with water absorption.

3.4.1. Spectra features of different data pre-processing methods

Figure 3.2b and Figure 3.2c give a multiplicative scatter correction (MSC) and standard normal variate (SNV) pre-processing. Overall improvements of baseline shift and signal overlap are clearly apparent in these two figures. Multiplicative interferences of scatter and additive effects from spectral data are also drastically reduced from raw untreated diffuse reflectance spectra. MSC and SNV spectra provide quite similar spectra results. However, when these two spectra are visually compared, the slope of the SNV spectra decreases compared to the slope of MSC spectra. SNV standardizes each spectrum using only the data from that spectra, it does not use the mean spectra of any set whilst MSC rotates each spectrum so that it fits as closely as possible to a standard spectrum that may often be the mean spectrum. Figure 3.2d shows second order de-trending (DT) spectra of all samples. Spectra shift changes, additional baseline and curvatures effects are also removed compare to raw spectra.

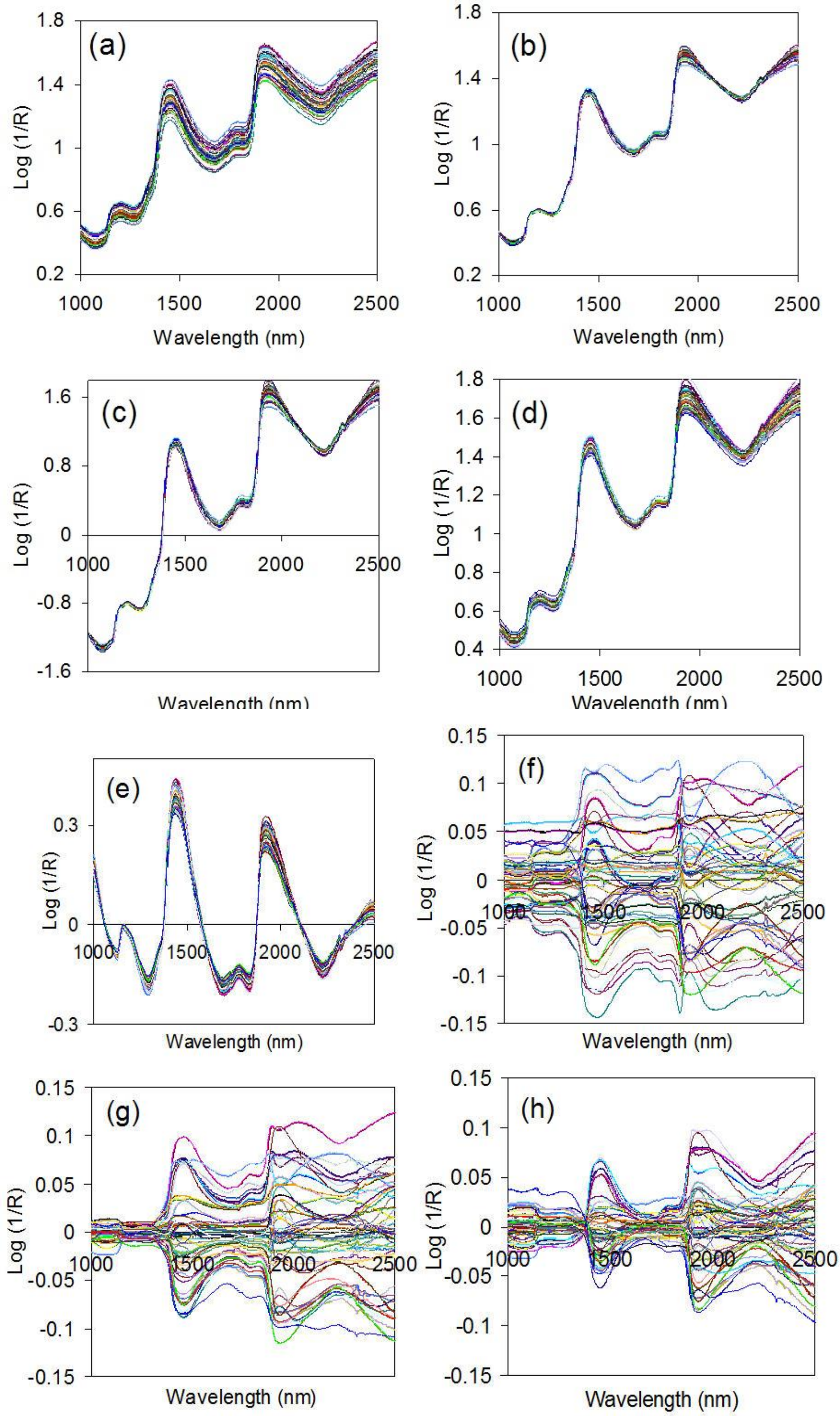

Figure 3.2. Raw spectra (a), multiplicative scatter correction (MSC) spectra (b), standard normal variate (SNV) spectra (c), mean normalization (MN) spectra (d), de-trending (DT) spectra (e), mean centered (MC) spectra (f), orthogonal signal correction (OSC) for titratable acidity spectra (g), OSC for soluble solids content spectra (h).

In Figure 3e, mean normalization (MN) spectra are given from which all spectra data were normalized in order to obtain spectra on approximately the same scale based on mean. On the other hand, mean-centered (MC) spectra of samples are shown in Figure 3f; and in Figure 3g and 3h, orthogonal scatter correction (OSC) spectra for titratable acidity (TA) and soluble solids content (SSC) are respectively given. For both OSC cases, only one OSC component was extracted to removes spectra data variance which is not correlated with the respective quality attributes either of titratable acidity or soluble solids content. When visually compared, OSC removes more spectral variance in the case of soluble solids content than in the case of titratable acidity. Yet, the OSC corrected spectra for titratable acidity are better than mean-centered original spectra with no OSC pre-processing. In agreement with previous work (Azzouz et al., 2003) and also found in this work that one or two OSC components are should be enough to represent spectra data set with respective quality attributes.

3.4.2. Comparison of calibration and prediction performance of different data pre-processing methods

Partial least squares regression (PLSR) models were built based on untreated and treated spectra using 56 calibration samples datasets with the wavelength range of 1000 - 2500 nm. Then, 35 prediction samples datasets were used to evaluate the models. The standard error for calibration and prediction as well as correlation coefficient and number of PLS factors, were compared. The obtained results of these models are discussed below.

3.4.2.1. Calibration model for titratable acidity

The content of titratable acidity (TA) in all samples ranged from 189.72 to 757.02 mg·100g^{-1} fresh mass for calibration, and from 190.04 to 632.03 mg·100g^{-1} fresh mass for prediction. The partial least square regression (PLSR) models were developed by applying the above mentioned spectra pre-processing methods to predict TA contents. The comparison among spectra pre-processing methods prior to PLSR on the prediction of TA is presented in Table 3.1. As shown in this table, the untreated spectra which is identified as none and mean centering (MC) provided quite similar values for R^2 calibration, prediction and error values. Four numbers of latent variables of PLSR were required to build the models. In calibration step, both none and MC produced R^2 of 0.9123 whilst root mean square errors of cross validation (RMSECV) were 43.8665 and 42.0972 mg·100g^{-1} for none and MC respectively.

Table 3.1. Partial least square calibration and prediction results of titratable acidity based on different spectra pre-processing

Statistical Parameter	**none**	**MC**	**MN**	**DT**	**MSC**	**SNV**	**OSC**
LVs	4	4	4	7	3	3	2
R^2 calibration	0.912	0.912	0.921	0.910	0.925	0.925	0.925
R^2 cross validation	0.892	0.903	0.908	0.844	0.919	0.916	0.923
R^2 prediction	0.691	0.691	0.675	0.686	0.682	0.682	0.708
RMSEC	39.494	39.483	37.526	39.998	36.508	36.503	36.642
RMSECV	43.867	42.097	41.916	53.017	39.072	39.376	38.869
RMSEP	53.879	53.795	55.283	54.355	54.628	54.629	52.363
RPD	2.498	2.502	2.435	2.476	2.464	2.464	2.570
Error difference	4.372	2.614	4.390	13.019	2.564	2.874	2.228

MC: mean centering, MN: mean normalization, DT: de-trending, MSC: multiplicative scatter correction, SNV: standard normal variate, OSC: orthogonal signal correction, LVs: number of latent variables in PLSR, RMSE: root mean square error; C: calibration, CV: cross validation, P: prediction respectively, RPD: residual predictive deviation, error difference: RMSEC-RMSECV.

When these models were tested using 35 samples data set for prediction, they produced R^2 value of 0.6911 and root mean square error of prediction (RMSEP) is 53.84 mg·100g^{-1}. On the other hand, mean normalization (MN) pre-processing provided slightly better result in calibration and cross validation with the increasing R^2 calibration to 0.9208 and cross validation to 0.9077 with the same number of LVs (four LVs) to develop PLSR model compared to none and MC. However, this pre-processing method produced poor external prediction result (R^2 = 0.6747 and RMSEP = 55.2834 mg·100g^{-1}). This indicates that prediction model based on MN pre-processing method is over-fitted; generating good R^2 and RMSE in calibration but performing poorly in the external prediction data set.

Furthermore, as presented also in Table 1, the use of MSC and SNV as a pre-processing method produced a better result in calibration. R^2 calibration was increased to 0.9251 and RMSECV was decreased to 39.3763 mg·100g^{-1} with fewer number of latent variables (three LVs) required to establish PLSR model. MSC and SNV provided similar results both in calibration and prediction. Then again, when both MSC and SNV were tested using 35 external samples, these models produced slightly lower R^2 value (0.6824) and higher RMSEP (54.6292 mg·100g^{-1}) compared to none and MC. De-trending (DT) spectra pre-processing appeared to be unworthy method to apply in case of solids sample like mango. It produced highest RMSECV (53.017 mg·100g^{-1}) and seven latent variable were required to establish PLSR model. Based on literature, DT pre-processing was suitable to be applied on bulk samples like flour or soil samples (Azzouz et al., 2003).

The best calibration and prediction results was obtained when OSC was applied as a pre-processing method from which highest R^2 (0.925 and 0.708 for calibration and external prediction respectively) and lowest error (RMSECV = 38.869 and RMSEP = 52.363 mg·100g^{-1}) were produced. The number of latent variable required was also decreased to two for PLSR with only one principal component (PC) or factor of OSC. The number of OSC factors is the number of times that OSC is applied to the NIR spectra. One OSC factor is usually enough, since a second factor may lead to a great over-fitting. Figure 3.3 presented a scatter plot resulted from the OSC-PLSR model for TA calibration and prediction.

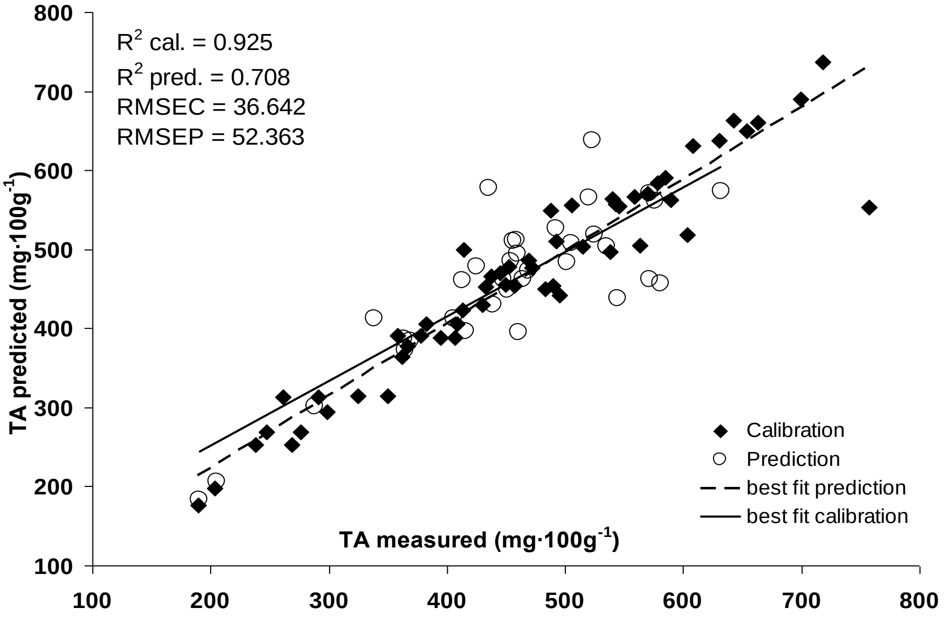

Figure 3.3. Scatter plots of predicted and measured titratable acidity based on orthogonal signal correction – partial least square regression model.

3.4.2.2. Calibration model for soluble solids content

Like in TA predictions, all pre-processing methods were also applied before PLSR to predict soluble solids content (SSC) of intact mango fruit. The models were developed using 56 samples and tested using 35 external mango samples. These samples contain between 7 and 18 ^{o}Brix of SSC for calibration and between 6 and 17 ^{o}Brix for prediction. Calibration results for SSC based on different spectra pre-processing are presented in Table 3.2.

Table 3.2. Partial least square calibration and prediction results of soluble solids content based on different spectra pre-processing

Statistical Parameter	**none**	**MC**	**MN**	**DT**	**MSC**	**SNV**	**OSC**
LVs	5	5	4	9	3	3	2
R^2 calibration	0.8537	0.8541	0.8369	0.8231	0.8349	0.8349	0.8433
R^2 cross validation	0.8235	0.8258	0.8079	0.6357	0.8280	0.8124	0.8285
R^2 prediction	0.5407	0.5417	0.5402	0.5261	0.5754	0.5753	0.6050
RMSEC	0.4858	0.4847	0.5129	0.5343	0.5116	0.5160	0.5029
RMSECV	0.5401	0.5411	0.5760	0.7780	0.5505	0.5517	0.5330
RMSEP	0.7681	0.7653	0.7677	0.7803	0.7368	0.7382	0.7164
RPD	1.6691	1.6752	1.6699	1.6430	1.7400	1.7367	1.7895
Error difference	0.0543	0.0564	0.0631	0.2437	0.0389	0.0357	0.0301

MC: mean centering, MN: mean normalization, DT: de-trending, MSC: multiplicative scatter correction, SNV: standard normal variate, OSC: orthogonal signal correction, LVs: number of latent variables in PLSR, RMSE: root mean square error; C: calibration, CV: cross validation, P: prediction respectively, RPD: residual predictive deviation, error difference: RMSEC-RMSECV.

As shown in Table 3.2, five latent variables were required to capture all variance of the data and develop the model when the NIR spectra were directly modeled without spectra pre-processing. Even so, the calibration and prediction results were quite sufficient with R^2 of 0.8235 and 0.5407 for cross validation and prediction respectively. Table 2 also shows that for SSC calibration and prediction, the application of MC, MN and DT did not significantly influence the models performance. The coefficient determination and error values were still quite similar to that for none (original spectra). As also shown in TA calibration, the use of DT as pre-processing method did not suitable in this case and, even so, the R^2 coefficient of cross validation was decreased to 0.6357 and the errors of prediction (RMSEP) was increased to 0.7803 ^{o}Brix.

The use of MSC and SNV spectra data pre-processing appeared to improve the calibration models. Beside the reduction number of latent variables required (three LVs), the use of these both pre-processing methods has been increased the R^2 prediction value to 0.5753 and decreased the errors in prediction (RMSEP) to 0.73 ^{o}Brix. The results presented in Table 2 shows the significant influence of OSC prior to PLSR model from which the calibration and prediction performance has been improved. Figure 3.4 presents the performance of PLSR model after OSC pre-processing of the spectra data, developed for the calibration and prediction of SSC.

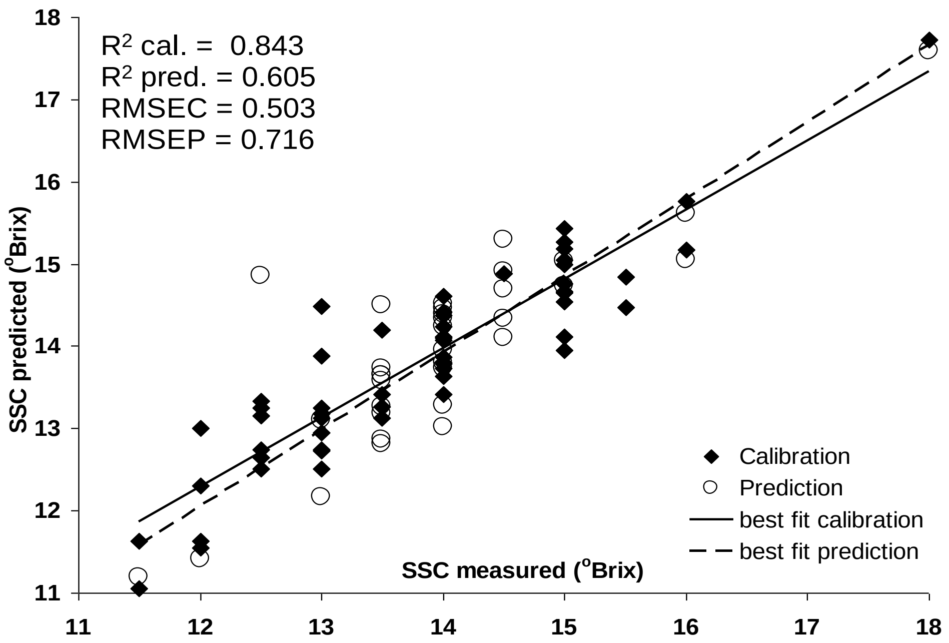

Figure 3.4. Scatter plots of predicted and measured soluble solids content based on orthogonal signal correction – partial least square regression model.

The use of orthogonal signal correction allowed a decrease of the number of latent variables. Concerning with the R^2 and errors of prediction, better results were obtained using OSC. On top of that, the model uses only two latent variables of PLS with one principal component of OSC to obtain R^2 prediction of 0.605 and prediction errors (RMSEP) of 0.71 °Brix. This may argue that OSC is more superiors than other spectra pre-preprocessing methods both for TA and SSC predictions.

3.5. Conclusion

The main objective of this study was to compare spectra pre-processing methods prior to partial least square regression in the ability performance of the calibration and prediction models developed to relate the near infrared spectra of mango fruit samples and their quality attributes in form of titratable acidity (TA) and soluble solids content (SSC). The presented results reinforce that the use of orthogonal signal correction (OSC) pre-processing method particularly allow a significant improvement both for TA and SSC prediction compared to other methods. Compared to untreated original spectra, the use of OSC found to be the best spectra pretreatment prior to PLSR model, allowing the reduction of the number of latent variables from four to two for TA and from five to two for SSC. It also allowed the reduction of prediction errors from 53.879 to 42.362 for TA and from 0.768 to 0.641 for SSC ^{o}Brix.

References

Azzouz, T., Puigdomenech, A., Aragay, M., & Tauler, R. (2003). Comparison between different pre-treatment methods in the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method. *Analytica Chimica Acta*, 484, 121-134.

Blanco, M., Coello, J., Montoliu, I., & Romero, M. A. (2001). Orthogonal signal correction in near infrared calibration. *Analytica Chimica Acta*, 434, 125-132.

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology*, 18, 72-83.

Chen, J. Y., Iyo, C., Terada, F., & Kawano, S. (2002). Effect of multiplicative scatter correction on wavelength selection for near infrared calibration to determine fat content in raw milk, *J. Near Infrared Spectrosc*. 10, 301-306.

Constantinou, M. A., Papakonstantinou, E., Benaki, D., Spraul, M., Shulpis, K., Koupparis, M. A., & Mikros, A. (2004). Application of nuclear magnetic resonance spectroscopy combined with principal component analysis in detecting inborn errors of metabolism using blood spots: a metabonomic approach. *Analytica Cimica Acta*, 511, 303-312.

Cozzolino, D., Cynkar, W. U., Shah, N., & Smith, P. (2011). Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Research International*, 44, 1888-1896.

Felizardo, P., Baptista, P., Menezes, J.C., & Correia, M.J.N. (2007). Multivariate near infrared spectroscopy models for predicting methanol and water content in biodiesel. *Analytica Chimica Acta*, 595, 107-113.

Flores, K., Sanchez, M. T., Perez-Marin, D., Guerrero, J. E., & Garrido-Varo, A. (2009). Feasibility in NIRS instruments for predicting internal quality in intact tomato. *Journal of Food Engineering*, 91, 311-318.

Jha, S. N., Kingsly, A. R. P., & Chopra, S. (2006). Non-destructive determination of firmness and yellowness of mango during growth and storage using visual spectroscopy. *Biosystems Engineering*, 94, 397-402.

Mouazen, A. M., Kuang, B., De Baerdemaeker, J., & Ramon, H. (2010). Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma*, 158, 23–31.

Munawar, A. A., Hörsten, D. v., Mörlein, D., Pawelzik, E., & Wegener, J. K. (2013). Rapid and non-destructive prediction of mango sweetness and acidity using near infrared spectroscopy. In *Gesellschaft für Informatics lecture notes*: Proceeding of the GIL Jahrestagung.

Naes, T., Isaksson, T., Fearn, T., & Davies, T. (2004). A user-friendly guide to multivariate calibration and classification. Chichester, UK: NIR publications.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lamertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. *Postharvest Biology and Technology*, 46, 99-118.

Perez-Marin, D., Garrido-Varo, A., & Guerrero, J. E. (2007). Non-linear regression methods in NIRS quantitative analysis. *Talanta*, 72, 28-42.

Pontes, M.J.C., Santos, S.R.B., Araújo, M.C.U., Almeida, L.F., Lima, R.A.C., Gaião, E.N., & Souto, U.T.C.P. (2006). Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometry. *Food Research International*, 39, 182-189.

Schmilovitch, Z., Mizrach, A., Hoffman, A., Egozi, H., & Fuchs, Y. (2000). Determination of mango physiological indices by near-infrared spectrometry. *Postharvest Biology and Technology*, 19, 245-252.

Sinelli, N., Spinardi, A., Di Egidio, V., Mignani, I., & Casiraghi, E. (2008). Evaluation of quality and nutraceutical content of blueberries (*Vaccinium corymbosum* L.) by near and mid-infrared spectroscopy. *Postharvest Biology and Technology*, 50, 31-36.

The Unscrambler® X 10.2 method reference, (2012). CAMO, Oslo.

Wang, L., Lee, F.S.C., Wang, X., & He, Y. (2006). Feasibility study of quantifying and discriminating soybean oil adulteration in camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR. *Food Chemistry*, 95, 529-536.

Xiaobo, Z., Jiewen, Z., Xingyi, H., & Yanxiao, L. (2007). Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of ‘Fuji’ apple based on different PLS models. *Chemometrics and intelligent laboratory systems*, 87, 43-51.

Chapter 4. Comparisons of different regression approaches in prediction of quality attributes in intact mango using near infrared spectroscopy

4.1. Abstract

In the present study, three different linear and non-linear regression techniques namely partial leas squares regression (PLSR), support vector machine regression (SVMR), and artificial neural network (ANN) were studied and compared in predicting titratable acidity (TA) and soluble solids content (SSC) of intact mangoes. TA and SSC prediction models were established based on near infrared diffuse reflectance spectra acquired between 1000 and 2500 nm. The results obtained show that ANN and SVMR are more powerful than PLSR for TA and SSC prediction. The optimal prediction models for both quality attributes were obtained by ANN with the first 4 principal components (PCs) as input. For TA prediction, the coefficient determination of calibration (R^2_{cal}) and prediction (R^2_{pred}), the root-mean square error of calibration (RMSEC) and prediction (RMSEP), and the residual predictive deviation (RPD) were 0.97, 0.89, 25.29 mg·100g^{-1}, 28.42 mg·100g^{-1} and 4.02, respectively. For SSC prediction, the accuracy indexes were 0.92, 0.86, 0.65 ^{o}Brix, 0.69 ^{o}Brix and 2.52. The overall results sufficiently demonstrate that NIR spectroscopy coupled with the ANN regression approach has the optimal results to determine TA and SSC of intact mango.

Keywords: regression tool, NIR spectroscopic data, non-linear approach.

4.2. Introduction

Mango which is known as king of fruits is one of the most popular fruits in tropical horticultures and very popular worldwide due to its high nutritional value, delicious taste and excellent overall flavor. It has very high demand and fetches a good price in the world market (Schmilovitch et al., 2000). With the increasing demand and consumption of mango, quality control becomes more and more important nowadays. Many national and international authorities are setting such criteria for quality factors to ensure good chain supply of mangoes. In general, consumers purchase fresh fruits on the basis of quality which is an important subject to those engaged in horticultural industries. Their acceptance depends on highly subjective factors including appearance, touch, smell and taste. Titratable acidity (TA) and soluble solid content (SSC) are two main quality attributes among others for mango fruit. The TA is represents mango total acidities whilst the SSC is normally responsible for mango's sweetness.

To determine acidity and sweetness of mango or other agricultural products, several methods were already widely used in which most of them are based on solvent extraction followed by other laboratory procedures. However, these methods often require laborious and complicated processing for samples. They are time consuming and destructive, therefore unsuitable for the quality control of fresh agricultural products which requires real time, rapid, on-line and non-destructive measurements. Nowadays, several rapid and non-destructive methods are available such as ultrasound, microwaves absorption, nuclear magnetic resonance and infrared spectroscopy (Xie et al., 2011). Near infrared spectroscopy (NIRS) is among these methods has been proved to be a fast, simple, chemical free and non-destructive method and widely used for quality control assessment of various agricultural products (Chen et al., 2012; Agelet et al., 2012; Xiaoying et al., 2012; Niu et al., 2008; Guthrie et al., 2006; Liu et al., 2006; Han et al., 2006; Tewari et al., 2008).

Recently, our own studies and some authors attempted to determine quality parameters such as soluble solids content, yellowness, acidity, pH, ascorbic acid and firmness of mango using NIR spectroscopy technique (Jha et al., 2012; Schmilovitch et al., 2000; Jha et al., 2006; Valente et al., 2009; Subedi et al., 2007). However, the most commonly used multivariate calibrations in these previous studies are based on linear approaches such as multiple linear regression (MLR), principal component regression (PCR), and partial least square regression (PLSR). NIRS is based on overtones and combinations of fundamental vibrations from the hydrogenous bonds. Most absorption bands in the near infrared region are overtone or combination bands of the fundamental absorption bands in the infrared region of the electromagnetic spectrum which are due to vibrational and rotational transitions (Nicolai et al., 2007; Cozzolino et al., 2011; Chen et al., 2012). In large molecules and in complex mixtures, such as foods and agricultural products, the multiple bands and the effect of peak-broadening result in NIR spectra that have a broad envelope with few sharp peaks (Nicolai et al., 2007). Mango is the biological object that contains a great quantity of hydrogenous bonds (i.e C-H, O-H and N-H). As water (O-H) highly absorbs near infrared radiation, the near infrared spectrum of mango is dominated by water. This, in combination with the complex chemical constituents of a typical fruit causes the near infrared spectrum to be highly convoluted. Finally, the spectrum may further be complicated by wavelength dependent scattering effects, tissue heterogeneities, instrumental noise, ambient effects and other sources of variability (Nicolai et al., 2007). As a consequence, these overtones and combinations bands in the NIR region are typically very broad, leading to complex spectra, and it could be difficult to assign specific features to specific chemical constituents (Chen et al., 2012; Nicolai et al., 2007). Hence, the correlation between the NIR spectra and SSC and TA in mango fruit would incline to nonlinearity.

In many spectroscopic applications, PLSR are often used to make regression models because of its simplicity to use, speed and good performances. However, as described by Perez-Marin et al. (2007), in many current and potential applications of NIRS measurement, the relationship to be modeled is not always linear. This means that the classical linear regression methods (i.e MLR, PCR or PLSR) alone are not always the most suitable option and may not provide a complete solution to the regression problem in this work. Therefore, it is necessary that different linear and nonlinear tools should be attempted, and that is the aim of our present study. The use of support vector machine regression (SVMR) approach as non-linear method for NIR spectroscopic regression purposes may be an option. Another possibly method to overcome non-linear regression of NIR spectroscopic data is artificial neural network (ANN). These both two mentioned approaches have now recently applied for NIR spectroscopic data calibration in addition to classical linear regression method.

Therefore, the objective of the present work is to systematically study and compare three different linear and non-linear regression techniques namely partial leas squares regression (PLSR), supporting vector machine regression (SVMR), and artificial neural networks (ANN) to develop calibration models for titratable acidity (TA) and soluble solids content (SSC) prediction of intact mangos. The models were established based on NIR spectroscopic data after multiplicative scatter correction (MSC) and standard normal variate (SNV) spectra pre-processing, and the optimal model was selected from these three models.

4.3. Materials and methods

4.3.1. Spectra acquisition

In this experiment, NIR diffuse reflectance spectra were acquired for a total of 90 samples using a benchtop Nicolet-Antaris Method Development Sampling (MDS) system (Thermo Scientific Inc., Madison, WI, USA). Samples mango (cv. *Kent*) were purchased in local markets and measured every two days after the day purchased with around 10 samples per measurement. This was done by means to obtain varied distributions of SSC and TA of mango. Diffuse reflectance spectra were recorded on wavelength range between 1000 and 2500 nm, by co-adding 64 scans per sample using an integrating sphere as a basic setting measurement. Spectra for each sample were captured 6 times at different point (two in the left and right edge, and four in the center) and the average of the six spectra was stored and used for further analysis.

4.3.2. Reference measurement

The reference measurement of mango quality attributes (TA and SSC) was taken directly after spectra acquisitions. Each sample fruit was sliced at the same marked point of the NIR acquisition and the pulp was taken. TA and SSC measurement were carried out simultaneously by making sample juice from 20 grams of pulp sample and adding maximum 100 ml distilled water. To ensure clarified sample juice and separate suspended solids were obtained, the centrifuge (20^{o}C, 10 000 g) was applied for about 10 minutes (Schmilovitch et al., 2000). For TA measurement, automatic titration (Titroline 96, Schott) with 0.1 N NaOH to an end point of pH 8.1 was used to obtain TA expressed as mg·100g^{-1} Fresh Mass (Flores et al., 2009). Meanwhile for SSC measurement, a single drop of filtered supernatant juice was squeezed and dropped onto a hand-held analog refractometer (model HRO32, Krüss Optronic GmbH) to record SSC as ^{o}Brix (Xiaobo et al., 2007). Both these two reference quality attributes were measured in duplicate and averaged. The data were then used as an actual or measured value of TA and SSC for calibration models development.

4.3.3. Spectra pre-processing

In order to achieve a reliable, accurate and stable prediction model, NIR spectra of all samples were pre-processed using multiplicative scatter correction (MSC) and standard normal variate (SNV). Figure 4.1a presents the raw original spectra of all samples which contained not only specific chemical constituent information, but also background information and noises. From our previous studies, MSC and SNV spectra pre-processing were found to be effective to eliminate noises and scatter effects prior to TA and SSC models development. Therefore, MSC and SNV were attempted in this work. First of all, MSC spectra pre-processing separates scatter and multiplicative, and additive and chemical effects on measurement by linear fitting each individual spectrum to a reference spectrum that is usually a mean or average spectrum (Chen et al., 2002; Azzouz et al., 2003). The NIR spectra of all samples after MSC preprocessing are presented in Figure 4.1b. On the other hand, SNV is used to remove slope variation, and to correct scatter light effects. Each spectrum was first centered individually, and then scaled by the standard deviation calculated from the individual spectrum (Chen et al., 2012). The NIR spectra corrected by SNV pre-processing are shown in Figure 4.1c. As a result after MSC and SNV pre-processing, overall improvements of baseline shift and signal overlap are clearly apparent. Multiplicative interferences of scatter and additive effects from spectral data are also drastically reduced from raw NIR spectra.

4.3.4. Software

Integrated software, Thermo Integration® and Thermo Operation® (Thermo Inc.) were used to develop workflow and run specified tasks of the NIR instrument for spectra acquisition. The Unscrambler X version 10.2 network clients (CAMO Software AS, Oslo) was used for spectra MSC and SNV pre-processing, and building calibration models with PLSR and SVMR algorithms. Meanwhile, NeuralTools 6.0 (Palisade Corp., NY) was used for developing ANN calibration models using generalized regression neural networks (GRNN) algorithm.

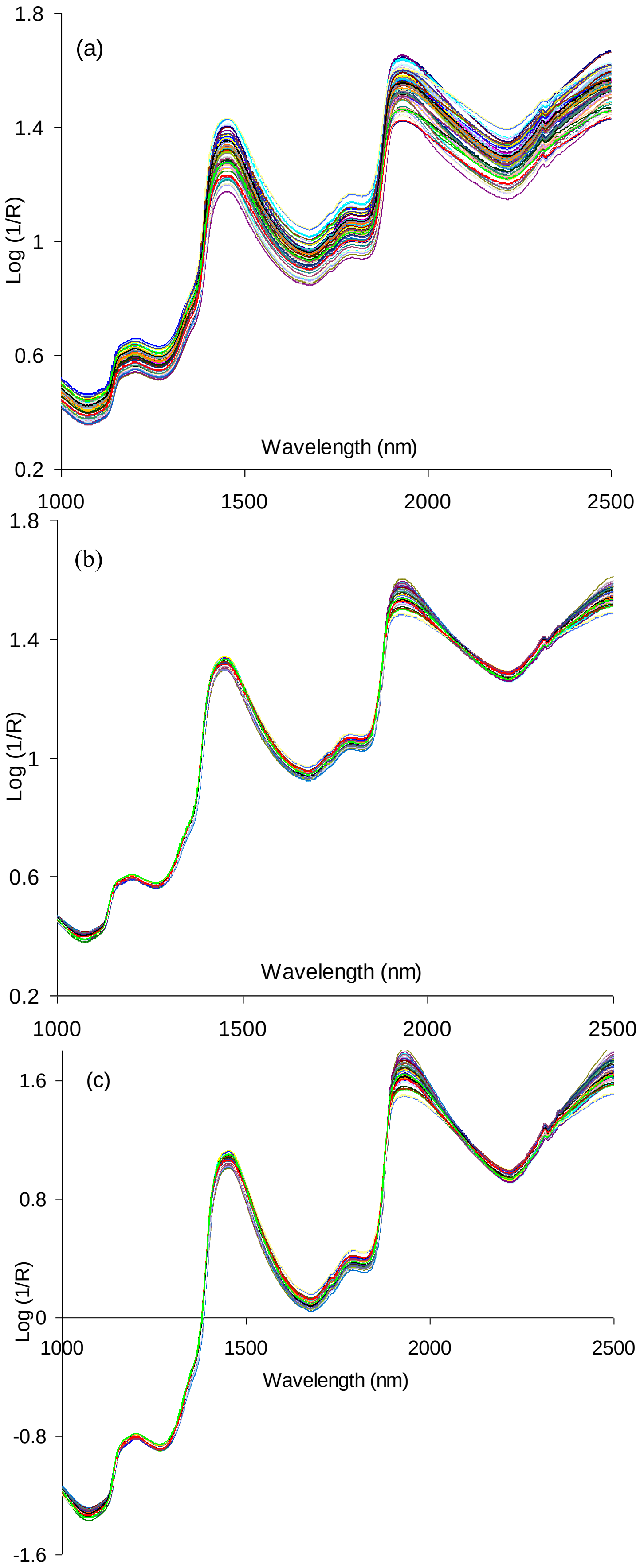

Figure 4.1. Near infrared raw spectra (a), spectra after multiplicative scatter correction (b), spectra after standard normal variate (c).

4.4. Results and discussion

4.4.1. Calibration models and prediction

Calibration models for tiratable acidity (TA) and soluble solids content (SSC) prediction were established using multiplicative scatter correction (MSC) and standard normal variate (SNV) spectral data respectively. For each quality attribute, three different regression approaches were applied: partial least square regression (PLSR), support vector machine regression (SVMR) and artificial neural network (ANN) algorithm. In this present study, 90 samples were divided into two subsets; one called calibration dataset containing 55 samples in which these samples were used for establishing calibration models. Other 35 remaining samples called prediction dataset were used for external prediction to evaluate the performance of the models. In order to achieve a robust calibration model, the selection samples for calibration and prediction datasets were treated carefully. We ensure that the range of actual reference values (*y*-value) for TA and SSC in calibration dataset covers their range in the prediction set as shown in table 4.1. Also, the distributions of *y*-values are close both in the calibration and prediction datasets.

Table 4.1. Descriptive statistics of reference titratable acidity and soluble solids content

QA	Datasets	n	mean	ranges	SD
TA (mg 100g^{-1})	Calibration	55	457.71	189.72 - 717.59	129.70
	Prediction	35	453.08	190.04 - 632.03	114.35
SSC (^{o}Brix)	Calibration	55	13.56	9 - 20	2.31
	Prediction	35	13.97	12 - 18	1.74

QA: quality attributes, n: number of samples, SD: standard deviation, TA: titratable acidity, SSC: soluble solids content.

Calibration models were established between reference measurement data as *y*-variable and NIR spectra data as *x*-variable in the calibration dataset. For TA prediction, calibration models were built based on MSC spectra while for SSC predictions, all models were built based on SNV spectral data. The models were then tested by independent samples in the prediction dataset. The performance of the final model was quantified based on calibration and prediction results according to the coefficient of determination of calibration (R^2_{cal}) and prediction (R^2_{pred}), the root mean square error of calibration (RMSEC) and prediction (RMSEP), the error difference between RMSEC and RMSEP, and the residual predictive deviation (RPD) indexes obtained by dividing standard deviation of reference data with the RMSEP value. Finally, the number of principal components or latent variables required to build the model was also took into account for optimal model criteria selections.

The optimal model here refers to its prediction accuracy and robustness, and commonly, an accurate and robust model should have a higher R^2 coefficient both in calibration and prediction, lower RMSEC and RMSEP, small difference between RMSEC and RMSEP, higher RPD index, and fewer numbers of latent variables. Some literatures mentioned NIR calibration model robustness criteria as follows: An RPD between 1.5 – 1.9 means that coarse quantitative prediction are possible, but still need some improvement in calibration. A value between 2 and 2.5 indicates that prediction model is sufficient. Meanwhile, an RPD value between 2.5 and 3 or above corresponds to good and excellent prediction accuracy respectively (Williams, 2001; Fearn 2002; Nicolai et al., 2007).

4.4.2. Calibration and prediction results of PLSR, SVMR and ANN regression approaches

First of all, the PLSR method was attempted to predict TA and SSC. PLSR is one of the most widely used regression methods for NIR calibration. In this present study, during calibration, PLSR was optimized by a multifold cross validation with 10 segments and determined according to the lowest minimum root mean square error cross validation on the respective latent of variable. Based on the best calibration result, the optimum PLSR model was achieved when five latent variables are included both for TA and SSC prediction as presented in Table 4.2.

For TA calibration, the optimal PLSR model produced the coefficient of determination (R^2) and the root mean square error of calibration (RMSEC) is 0.95 and 28.51 mg $100g^{-1}$ respectively. When the model tested by using samples on the prediction dataset, it yields the R^2 prediction and the root mean square error of prediction (RMSEP) is 0.83 and 33.08 mg $100g^{-1}$ respectively. Thus, the error difference between calibration and prediction is 4.57, and the residual predictive deviation is 3.46. This means, PLSR was accurate and robust for TA prediction. Scatter drawn from the PLSR model for predicted TA versus reference measured TA is presented in Figure 4.2a.

Table 4.2. Calibration and prediction of various regression methods for titratable acidity and soluble solids content

Method	QA	PCs / LVs	Calibration		Prediction		Error difference	RPD
			R^2	RMSEC	R^2	RMSEP		
PLSR	TA	5	0.95	28.51	0.83	33.08	4.57	3.46
SVMR		4	0.96	26.94	0.87	30.13	3.19	3.80
ANN		4	0.97	25.29	0.89	28.42	3.13	4.02
PLSR	SSC	5	0.84	0.81	0.76	0.92	0.11	1.89
SVMR		4	0.90	0.70	0.83	0.75	0.05	2.32
ANN		4	0.92	0.65	0.86	0.69	0.04	2.52

QA: quality attributes, TA: titratable acidity, SSC: soluble solids content, PCs: number of principal components, LVs: number of latent variables, R^2: coefficient of determination, RMSEC: root mean square error calibration, RMSEP: root mean square error prediction, error difference: RMSEP – RMSEC, RPD: residual predictive deviation, PLSR: partial least square regression, SD: standard deviation, SVMR: support vector machine regression, ANN: artificial neural networks.

For SSC calibration and prediction using PLSR approach, the optimal PLSR produced R2 calibration and prediction of 0.84 and 0.76 respectively; while the RMSEC and RMSEP for SSC prediction is 0.81 ^{o}Brix and 0.92 ^{o}Brix. The RPD index for SSC prediction by PLSR is 1.89. This may imply in this case that the model is quite sufficient for SSC prediction, but need some improvements. The scatter plot of predicted and measured SSC based on PLSR model is presented in Figure 4.2b.

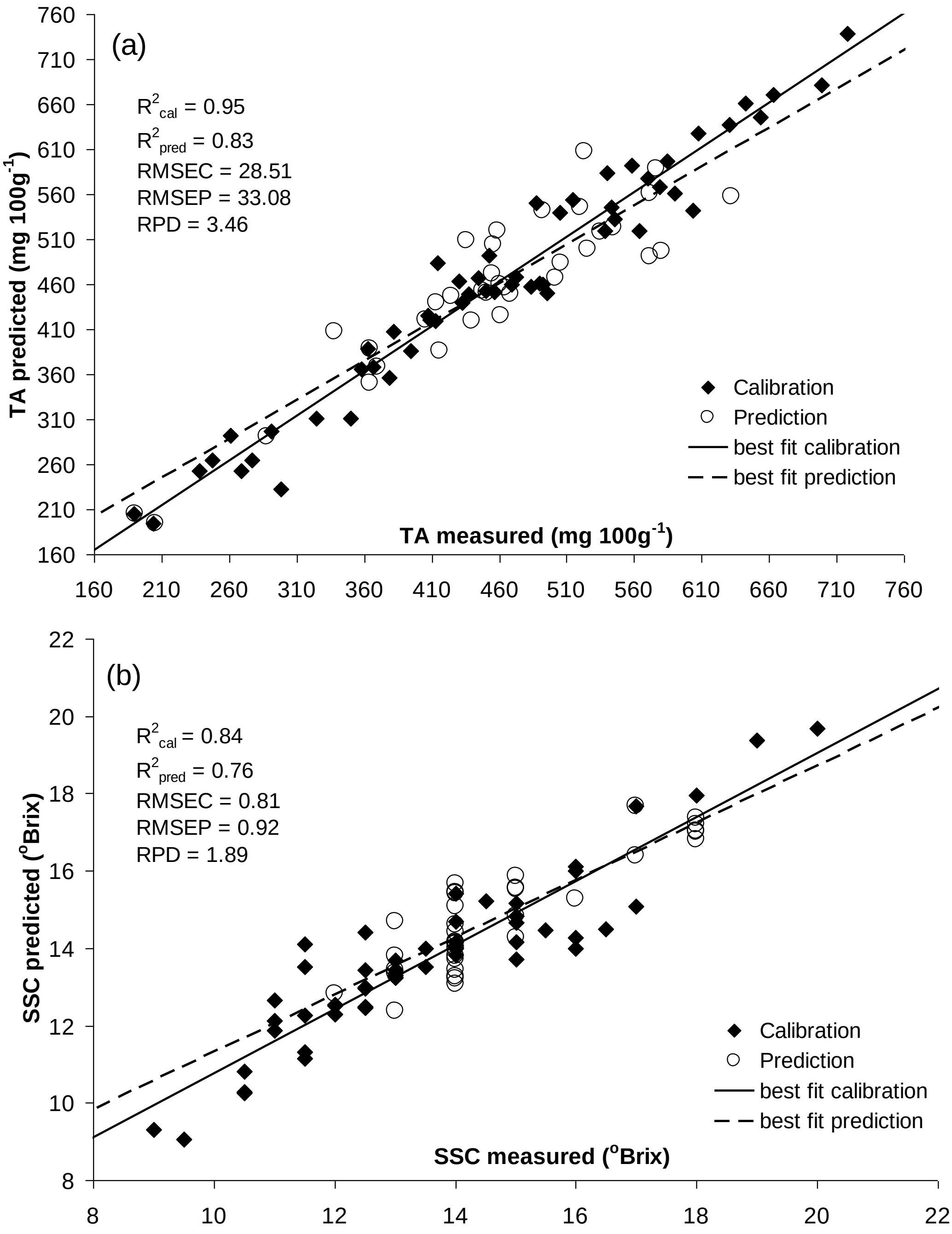

Figure 4.2. Scatter plot of predicted versus measured titratable acidity (a) and soluble solids content (b) based on partial least square regression model.

A nonlinear regression approach for TA and SSC calibration was attempted by support vector machine regression (SVMR) method. It can map the complex and nonlinear data into a higher dimensional feature space, where the nonlinear problem could be solved. Mapping data into a higher dimensional space is often implemented through a kernel function (Chen et al., 2007).

In general, SVMR has three classical kernel functions: polynomial, radial basic function (RBF), and sigmoid. Kernel functions have a great influence on the performance of SVMR model. Among three kernel functions, RBF is simpler and faster to compute in contrast to the others (Labbe et al., 2008). Thus, RBF as the kernel function was applied in this experiment. To obtain a good performance, two parameters in SVMR model had to be optimized by cross-validation. These parameters include: (1) Regularization parameter (γ), it determines the balance between minimizing the training error and minimizing model complexity; and (2) constraint parameter (C), it is used to decide a trade-off between the training error and the margin and can affect the number of support vectors used and loss function (ε) to construct the regression function (Chen et al., 2012; Eskidere et al., 2012). Based on grid search optimization, the parameter values for SVMR approach are as follows: C=100, γ=0.01, ε=0.1 for TA, and C=10, γ=0.1, ε=0.1 for SSC.

Prior to calibration models development using SVMR, the independent variables or *x*-variable (wavelength) based on MSC and SNV spectral data were compressed using principal component analysis (PCA). The PCA is extensively used for data synthesis, since the principal components (PCs) successively represent the sources of maximum variance in the NIR spectroscopic data and may replace wavelength as input (Perez-Marin et al., 2007). To determine the optimum number of principal components to be used as inputs for SVMR, we use the number of PCs that can explain the all variance nearly to 99% or 100%. In this study, it is found that four principal components (4 PCs) of PCA are sufficient to represent maximum variance (99%) of both MSC and SNV spectroscopic data as shown in Figure 4.3. The score results of these four PCs were used as inputs for SVMR algorithm.

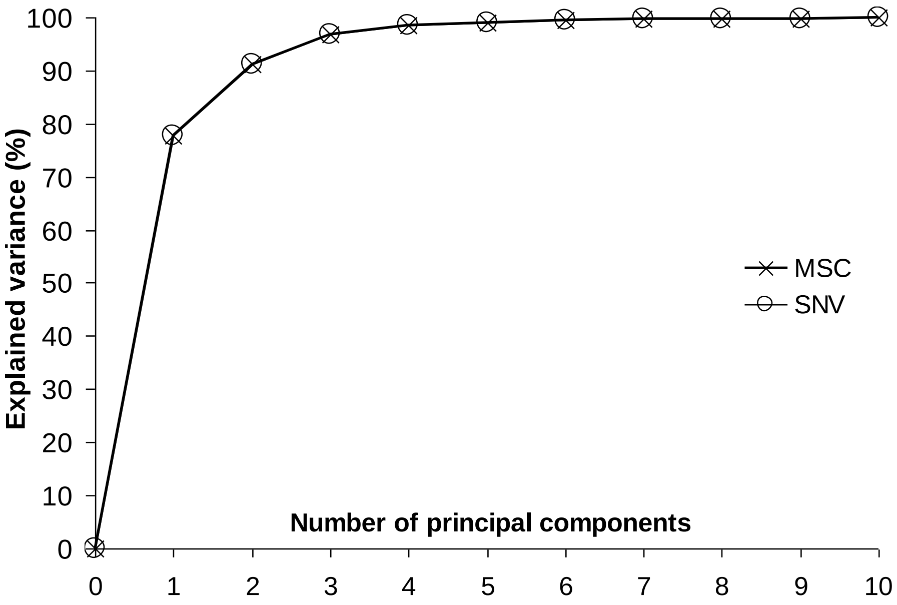

Figure 4.3. Explained variance of principal component analysis based on multiplicative scatter correction and standard normal variate spectroscopic data.

As can be seen in Table 4.2, SVMR provided superior results compared to PLSR both for TA and SSC predictions. For TA calibration and prediction, the R^2 coefficient indexes are 0.96 and 0.87 respectively whilst error coefficient are 26.94 mg $100g^{-1}$ (RMSEC) and 30.13 mg $100g^{-1}$ (RMSEP). The use of SVMR improves the R^2 coefficient the error both in calibration and prediction phases. As a consequence, the RPD index by SVMR for TA is also better (3.80) than by PLSR (3.46). Nevertheless, since the RPD is above 3, we may argue that TA can be predicted well either by SVMR or PLSR. Furthermore, for SSC calibration and prediction, SVMR also provided better results than PLSR as shown in Table 4.2. The R^2 coefficients are increase to 0.90 for calibration and 0.83 for prediction whilst error coefficients are decrease to 0.70 and 0.75 for RMSEC and RMSEP respectively. Scatter plot for TA and SSC based on SVMR calibration and prediction are presented in Figure 4.4a and Figure 4.4b respectively.

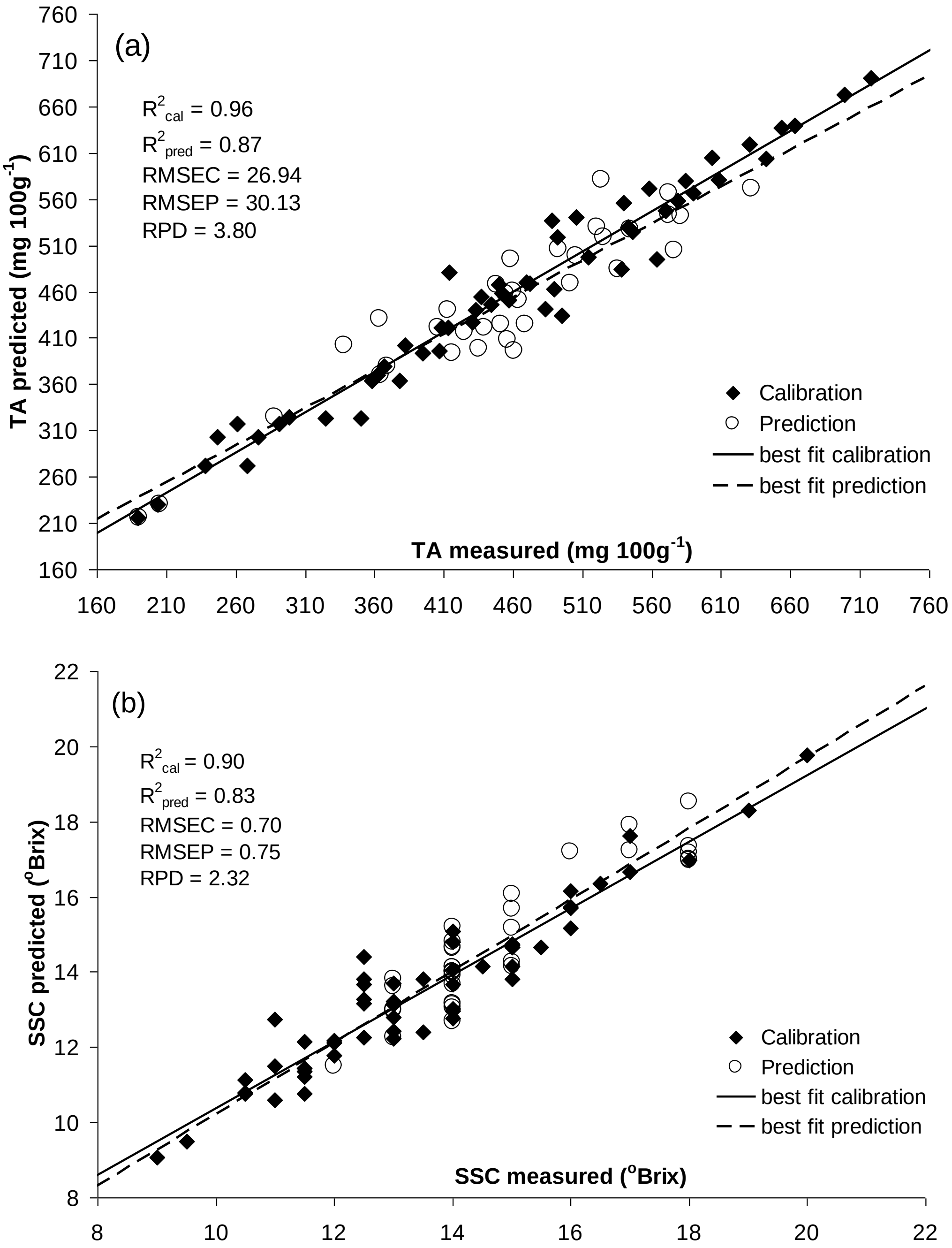

Figure 4.4. Scatter plot of predicted versus measured titratable acidity (a) and soluble solids content (b) based on supporting vector machine regression model.

In addition to SVMR approach as nonlinear regression method, calibration models for TA and SSC prediction were also established using artificial neural network (ANN) algorithm approach. In this present study, a generalized regression neural network (GRNN) was used as an ANN method to establish the models. The GRNN does not require an iterative training procedure as in other neural network algorithm such back propagation neural network (BPNN). It trains itself in a significantly shorter time as compared with the back propagation based training. The GRNN topology consists of four layers; the input layer, hidden (pattern) layer, the summation layer, and the output layer. The input layer is fully connected to the hidden layer. Kernel function is computed in the hidden layer and outputs of the hidden layer are passed on to the summation layer in order to calculate the sum of the weighted outputs of the hidden layer (Kim et al., 2004). As a result, the neurons in the output layer take results from summation units and then compute the output. The main advantage of GRNN is that it is easily trained and requires only one free parameter (Celikoglu and Cigizoglu, 2007). The input data subjected into this GRNN approach was also based on the PCA results of MSC and SNV as shown in Figure 4.3. Score results of first 4 PCs were used as inputs for the ANN structure both in TA and SSC calibration.

If all NIR spectra data, recorded at several hundred wavelengths are used as inputs for training the neural network, the number of weights to be estimated will be too high. Network training will take a long time (hours, depending on the complexity of the network architecture). But the main disadvantage of using all wavelengths as inputs is one of dimensionality, since there will be too many parameters in the model and too few samples in the training (calibration) set to enable fitting (Perez-Marin et al., 2007). This could pose problems, particularly where there is excess noise in the calibration set. Hence, the use of PCA as data compression tools is also enables to eliminate irrelevant information such as noise or redundancies present in NIR spectra data matrix. This also in agreement with other previous works (Bertran et al., 1999; Udelhoven and Schütt, 2000; Moros et al., 2007; Chen et al., 2012) that most applications of ANN to the processing of spectroscopic data for quantitative analysis use PCs as input variables instead of original NIR spectral data.

Calibration and prediction results for TA and SSC using ANN approach are presented in table 4.2. Figure 4.5 is the scatter plot between reference measurements and NIR predicted results for TA and SSC. For TA prediction, the R^2 index is increase to 0.97 in calibration and 0.89 in prediction while RMSEC and RMSEP are decreased to 25.29 and 28.42 respectively. Meanwhile, for SSC prediction, ANN achieved highest R^2: 0.92 for calibration and 0.86 for prediction and lowest error indexes: 0.65 for RMSEC and 0.69 for RMSEP.

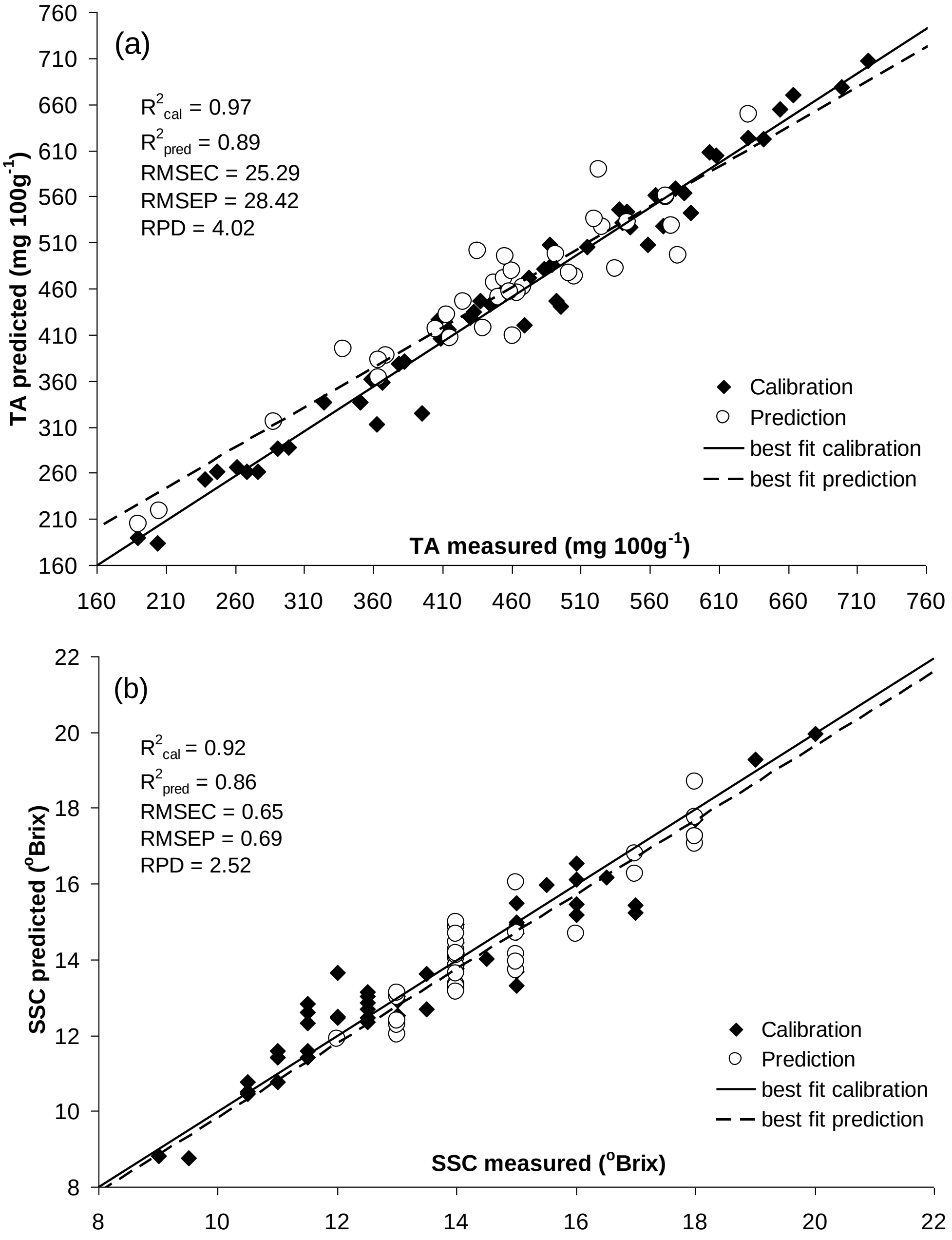

Figure 4.5. Scatter plot of predicted versus measured titratable acidity (a) and soluble solids content (b) based on artificial neural network model.

In general, judging from calibration and prediction performances, ANN and SVMR methods provided comparable results, and both of them are robust. They used the same number of PCs as inputs obtained after PCA. In term of error and RPD as shown in Figure 4.6 and Figure 4.7, both of them are also achieved excellent results for TA prediction and good enough results for SSC prediction. ANN and SVMR are superior over PLSR for TA and SSC prediction, and may imply that nonlinear regression approach is better that linear regression one. The optimal model for quality attributes prediction was achieved by using ANN approach as regression method because ANN provided smallest error index (RMSEC and RMSEP) and difference between each other: 3.13 for TA and 0.04 for SSC. ANN also provided the highest RPD index for both quality attributes: 4.02 for TA and 2.52 for SSC.

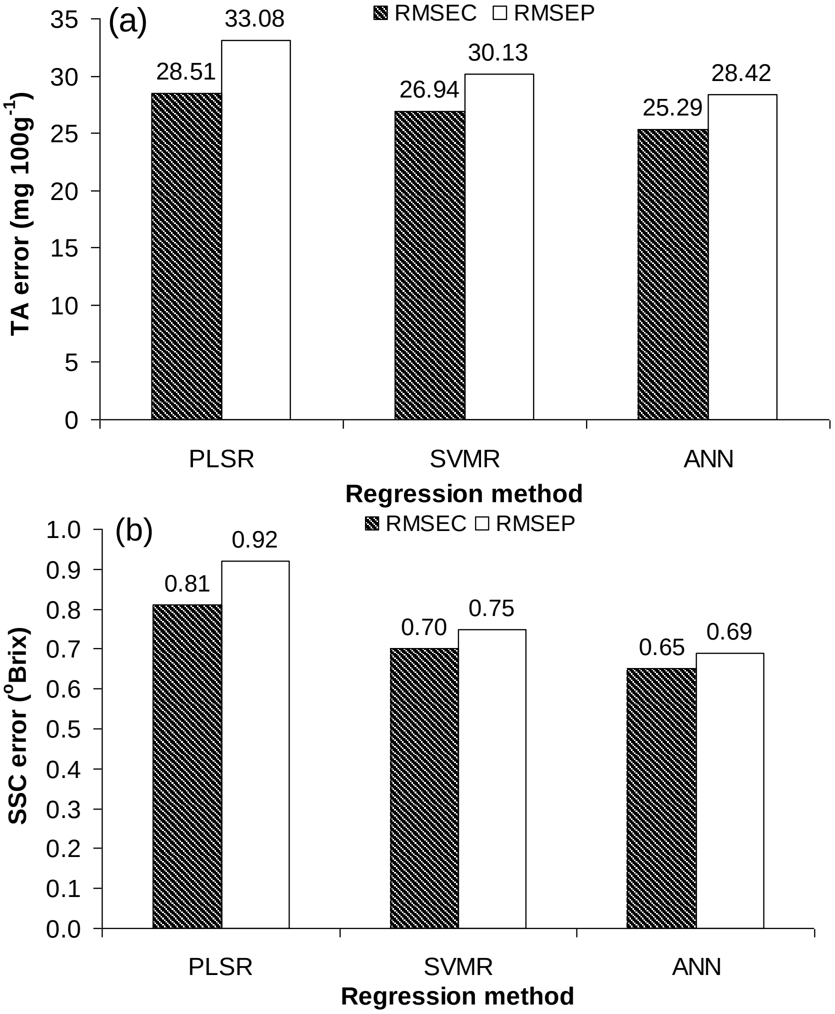

Figure 4.6. Summary of the results for titratable acidity (a) and soluble solids content (b) calibration and prediction in term of error using different regression approaches.

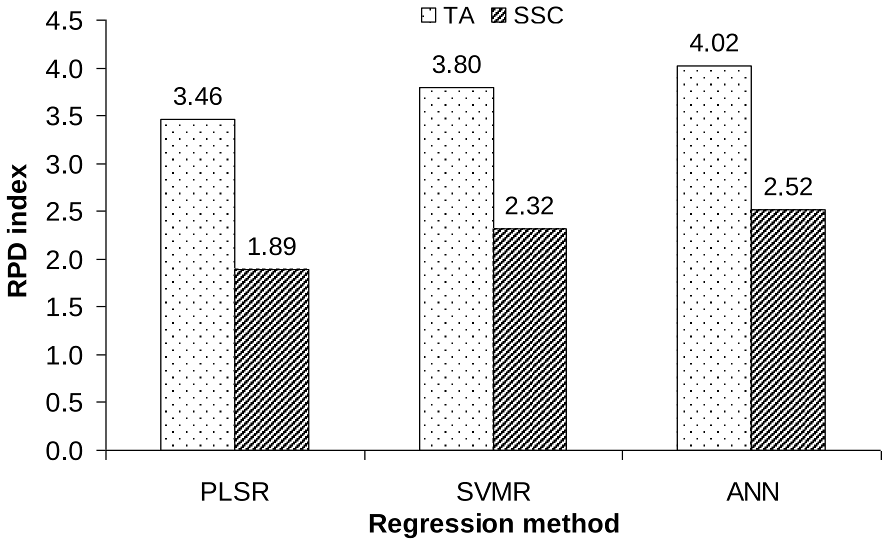

Figure 4.7. Summary of the results for titratable acidity and soluble solids content prediction in term of residual predictive deviation index using different regression approaches.

As mentioned previously, it is obvious that mango is a biological object containing a great quantity of hydrogenous bonds. NIR spectroscopy is based on overtone and combination of fundamental vibration from these bonds. The Overtone and combination bands in the near infrared region typically very broad, and leading to a complex spectra from which difficult to assign specific bands to specific chemical components. Hence, the correlation between the spectra and mango quality attributes could be nonlinear. When handling NIR spectroscopic data for calibration, PLSR is the most widely used regression method because it is traditionally assumed that Lambert-Beer’s law can be applied. This law states that absorption values are linearly related to the concentration concerned in each sample. PLSR also proved better than other linear regression approach. However, the use of linear regression including PLSR may hinder chemical interpretation (Perez-Marin et al., 2007), and the relationship to be modeled is not always linear.

The source of nonlinearity may vary widely, and is difficult to identify. Some source of this nonlinearity may be corrected by spectra pre-processing like MSC, SNV or others prior to using linear calibration technique. However, by contrast, some others source of nonlinearity require the use of special nonlinear calibration approach. This means that classical regression methods are not always the most suitable option. Similar finding also reported by others (Pierna et al., 2011; Chen et al., 2012; Eskidere et al., 2012; Xiaoying et al., 2012) that nonlinear regression approach achieved optimal results compare to linear regression. It is clearly found from our present study that nonlinear regression approach, either SVMR or ANN achieved a better accuracy and more robust than linear regression method for quality attributes (TA and SSC) prediction of intact mango.

4.5. Conclusion

In this present work, we systematically studied and compared the performance of three different regression approaches namely PLSR, SVMR and ANN for near infrared spectroscopic data calibration. These regression approaches were applied to predict TA and SSC of intact mango non-destructively by NIR spectra data corrected after MSC and SNV pre-processing. The overall calibration and prediction results sufficiently demonstrate that NIR spectroscopy with the help of these various regression approaches can be successfully used in determination of TA and SSC in mango. The evaluation performance results of PLSR, SVMR and ANN with respect to TA and SSC prediction accuracy and robustness could be arranged in the following order: ANN (4 PCs of PCA as input) > SVMR (4 PCs of PCA as input) > PLSR (5 LVs). The optimal models were obtained using ANN (GRNN) with the first 4 PCs extracted by PCA in spectral range of 1000-2500 nm: the result for TA with accuracy and robustness index of $R^2_{cal.}$, $R^2_{pred.}$, RMSEC, RMSEP and RPD of 0.97, 0.89, 25.29 mg 100g^{-1}, 28.42 mg 100g^{-1} and 4.02, respectively had excellent predictive accuracy and robustness; the result for SSC with the same respective indexes of 0.92, 0.86, 0.65 ^{o}Brix, 0.69 ^{o}Brix and 2.52 gave a good precision and its robustness was acceptable. The obtained results in this present study show the high potential of NIRS with nonlinear regression approach for determination of TA and SSC of intact mango rapidly and non-destructively. Further study with more amounts of samples, varieties, and wide range of maturity is needed to enhance the expandability of TA prediction model and improve the robustness of SSC model.

References

Agelet, L. E., Ellis, D. D., Duvick, S., Goggi, S., Hurburgh, C. R., & Gardner, C. A. (2012). Feasibility of near infrared spectroscopy for analyzing corn kernel damage and viability of soybean and corn kernels, *Journal of Cereal Science*, 55, 160-165.

Azzouz, T., Puigdomenech, A., Aragay, M., & Tauler, R. (2003). Comparison between different pre-treatment methods in the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method. *Analytica Chimica Acta*, 484, 121-134.

Bertran, E., Blanco, M., Maspoch, S., Ortiz, M. C., Sanchez, M. S., Sarabia, L. A. (1999). Handling intrinsic non-linearity in near-infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 49, 215–224.

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., & Pages, J. (2000). NIR calibration in non-linear systems: Different PLS approaches and artificial neural networks. *Chemometrics and Intelligent Laboratory Systems*, 50, 75–82.

Celikoglu, H. B., & Cigizoglu, H. K. (2007). Public transport trip flow modeling with generalized regression neural networks. *Advances in Engineering Software*, 38, 71-79.

Chen, J. Y., Iyo, C., Terada, F., & Kawano, S. (2002). Effect of multiplicative scatter correction on wavelength selection for near infrared calibration to determine fat content in raw milk, *J. Near Infrared Spectroscopy*. 10, 301-306.

Chen, Q. S., Zhao, J. W., Fang, C. H., Wang, D. M. (2007). Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM), *Spectrochimica Acta Part A*, 66, 568–574.

Chen, Q., Guo, Z., Zhao, J., & Ouyang, Q. (2012). Comparisons of different regressions tools in measurement of antioxidant activity in green tea using near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 60, 92-97.

Cozzolino, D., Cynkar, W. U., Shah, N., & Smith, P. (2011). Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Research International*, 44, 1888-1896.

Eskidere, Ö., Ertas, F., & Hanilci, C. (2012). A comparison of regression methods for remote tracking of Parkinson's disease progression. *Expert Systems with Application*, 39, 5523-5528.

Fearn, T. (2002). Assessing calibrations: SEP, RPD, RER and R^2. *NIR News*, 13, 12−14.

Flores, K., Sanchez, M. T., Perez-Marin, D., Guerrero, J. E., & Garrido-Varo, A. (2009). Feasibility in NIRS instruments for predicting internal quality in intact tomato. *Journal of Food Engineering*, 91, 311-318.

Guthrie, J.A., Liebenberg, C.J., & Walsh, K.B. (2006). NIR model development and robustness in prediction of melon fruit total soluble solids. *Aust. J. Agric. Res*., 57, 1–8.

Han, D.,Tu, R., Lu, C., Liu, C., & Wen, Z. (2006). Nondestructive detection of brown core in the Chines pear '*Yali*' by transmission visible-NIR spectroscopy. *Food Control*, 17, 604–608.

Jha, S. N., Jaiswal, P., Narsaiah, K., Gupta, M., Bhardwaj, R., & Singh, A. K. (2012). Non-destructive prediction of sweetness of intact mango using near infrared spectroscopy. *Scientia Horticulturae*, 138, 171-175.

Jha, S. N., Kingsly, A. R. P., & Chopra, S. (2006). Non-destructive determination of firmness and yellowness of mango during growth and storage using visual spectroscopy. *Biosystem Engineering*, 94, 397-402.

Kim, B., Lee, D. W., Parka, K. Y., Choi, S. R., & Choi, S. (2004). Prediction of plasma etching using a randomized generalized regression neural network. *Vacuum*, 76, 37–43.

Liu, Y. D., Ying, Y. B., Yu, H. Y., & Fu, X. P. (2006). Comparison of the HPLC method and FT-NIR analysis for quantification of glucose, fructose, and sucrose in intact apple fruits. *Journal of Agricultural and Food Chemistry*, 54, 2810–2815.

Moros, J., Inon, F. A., Garrigues, S., & de la Guardia, M. (2007). Near-infrared diffuse reflectance spectroscopy and neural networks for measuring nutritional parameters in chocolate samples, *Analytica Chimica Acta*, 584, 215–222.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lamertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. *Postharvest Biology and Technology*, 46, 99-118.

Niu, X. Y., Shen, F., Yu, Y. F., Yan, Z. K., Xu, K., & Yu, H. Y. (2008). Analysis of sugars in Chinese rice wine by Fourier transform near-infrared spectroscopy with partial least-squares regression. *Journal of Agricultural and Food Chemistry*, 56, 7271–7278.

Perez-Marin, D., Garrido-Varo, A., & Guerrero, J. E. (2007). Non-linear regression methods in NIRS quantitative analysis. *Talanta*, 72, 28-42.

Pierna, J. A. F., Lecler, B., Conzen, J. P., Niemoeller, A., Baeten, V., & Dardenne, P. (2011). Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products. *Analytica Chimica Acta*, 705, 30-34.

Schmilovitch, Z., Mizrach, A., Hoffman, A., Egozi, H., & Fuchs, Y. (2000). Determination of mango physiological indices by near-infrared spectrometry. *Postharvest Biology and Technology*, 19, 245-252.

Subedi, P. P., Walsh, K. B., & Owens, G. (2007). Prediction of mango eating quality at harvest using short-wave near infrared spectrometry. *Postharvest Biology and Technology,* 43, 326-334.

Tewari, J. C., Dixit, V., Byoung-Kwan, C., & Malik, K. A. (2008). Determination of origin and sugars of citrus fruits using genetic algorithm, correspondence analysis and partial least square combined with fiber optic NIR spectroscopy. *Spectrochimica Acta Part A*, 71, 1119–1127.

Udelhoven, T., & Schütt, B. (2000). Capability of feed-forward neural networks for a chemical evaluation of sediments with diffuse reflectance spectroscopy, *Chemometrics and Intelligent Laboratory Systems*, 51, 9–22.

Valente, M., Leardi, R., Self, G., Luciano, G., & Pain, J. P. (2009). Multivariate calibration of mango firmness using vis/NIR spectroscopy and acoustic impulse method. *Journal of Food Engineering*, 94, 7-13.

Williams, P. C. (2001). Implementation of near-infrared technology. In P. C. Williams, & K. H. Norris (Eds.), *Near Infrared Technology in the Agricultural and Food Industries*. St. Paul, Minnesota, USA: American Association of Cereal Chemist. 145-169.

Xiaobo, Z., Jiewen, Z., Xingyi, H., & Yanxiao, L. (2007). Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji' apple based on different PLS models. *Chemometrics and intelligent laboratory systems*, 87, 43-51.

Xiaoying, N., Zhilei, Z., Kejun, J., & Xiaoting, L. (2012). A feasibility study on quantitative analysis of glucose and fructose in lotus root powder by FT- NIR spectroscopy and chemometrics. *Food Chemistry*, 133, 592-597.

Xie, L., Ye, X., Liu, D., & Ying, Y. (2011). Prediction of titratable acidity, malic acid, and citric acid in bayberry fruit by near-infrared spectroscopy. *Food Research International*, 44, 2198-2204.

Chapter 5. Conclusion and summary

5.1. Conclusion

The present studies obviously show that near infrared reflectance spectroscopy (NIRS) combined with chemometrics has the ability to determine quality attributes of intact mango satisfactorily. The first study attempted to apply principal component regression (PCR) and partial least square regression (PLSR) to predict soluble solids content (SSC), titratable acidity (TA) and ascorbic acid (AA). The result shows that PLSR is preferable than PCR in all prediction cases providing the R^2 coefficient range from 0.67 to 0.96, and RPD index from 1.56 to 4.87 with maximum number of latent variables is 6.

The second study mainly focuses to the application of different spectra pre-processing prior to PLSR calibration. Based on prediction results with different spectra pre-processing, orthogonal signal correction (OSC) is the most optimum pre-processing method for TA and SSC prediction. The RPD index for TA and SSC are 2.57 and 1.79 respectively with only 2 LVs required to develop the model. Multiplicative scatter correction (MSC) and standard normal variate (SNV) are also provided robust results for TA and SSC prediction. They required maximum 3 LVs with the RPD index for TA and SSC are 2.46 and 1.74 respectively.

Non-linear regression method for NIR calibration was attempted in the third study. The machine learning methods: supporting vector machine regression (SVMR) and artificial neural networks (ANN) were applied to the first 4 principal components of PCA result derived from MSC and SNV spectra. The results show that non-linear regression method (SVMR and ANN) was superior to linear regression (PLSR) for both TA and SSC prediction.

In general, it may conclude that NIRS combined with proper chemometrics approaches may be used as an alternative for quality attributes measurement of intact mango.

5.2. Summary

Mango is one of the most important and popular tropical fruits for people around the world due to its taste, appearance and excellent overall nutritional source from which lead to a heavy demand in world fruit market. With the increasing demand and consumption of mango, quality control becomes more and more important nowadays. Many relevant authorities are setting such criteria for quality factors to ensure good chain supply of mangoes. Therefore, to ensure the chain supply of good quality fruit, it is important to sort and grade mango based on its quality. To determine quality parameters in mango, several methods were already widely used in which most of them are based on solvent extraction followed by other laboratory procedures. However, these methods often require laborious and complicated processing for samples. Also, they are time consuming and destructive. Hence, a rapid and non-destructive method is required as an alternative method in determining quality parameters of mangos.

Near infrared spectroscopy (NIRS) has become one of the most promising and used non-destructive methods of analysis in many field areas including in agriculture due to its advantage; simple sample preparation, rapid, and environmental friendly since no chemical materials are used. More importantly, it has the potential ability to determine multiple quality parameters simultaneously. Since NIRS itself cannot reveal chemical information in the spectra, chemometrics is required to extract the information about quality attributes buried on NIR spectra through a process called multivariate calibration from which a mathematical relationship between NIR spectra and the measured quality parameter will be revealed to determine desired quality attributes. Thus, the main objective of this study is to investigate the use of NIRS as non-destructive method combined by chemometrics for quality attributes in term of soluble solids content (SSC), titratable acidity (TA) and ascorbic acid (AA) predictions of intact mango.

A total of 99 mangos were used as samples in the study from which NIR spectra were acquired at wavelength range of 1000-2500 nm. A reference measurement for desired quality attributes were obtained by standard laboratory procedures: solvent extraction, refractive index by refractomter (for SSC), and titration method (for TA and AA). Chemometrics, which are include principal component analysis (PCA), outlier detections, spectra pre-processing (mean centering (MC), mean normalization (MN), de-trending (DT), multiplicative scatter correction (MSC), standard normal variate (SNV) and orthogonal signal correction (OSC)), linear calibration models by principal component regression (PCR) and partial least square regression (PLSR), and non-linear regression by supporting vector machine regression (SVMR) and artificial neural networks (ANN) were applied to reveal chemical information buried in the NIR spectra by creating calibration models followed by validation or prediction for models evaluation.

The results show that for linear regression method, PLSR seems to be more accurate and robust than PCR. From the spectra pre-processing point of view, the use of MSC, SNV and OSC prior to PLSR models development, significantly has an impact to the model accuracy and robustness. This can be seen from the reduction of latent variables used in PLSR (3 LVs after MSC and SNV, and 2 LVs after OSC), and increasing coefficient of determination (R^2) and residual predictive deviation (RPD) index in calibration and validation. Based on accuracy (R^2, RMSEC and RMSEP) and robustness index (RPD), the non-linear regression method (SVMR or ANN) was found to be better than linear regression (PLSR). The most optimal models for mango quality attributes prediction were achieved when ANN is used in combination with PCA as input. Thus, it may conclude that NIRS coupled with proper spectra pre-processing and regression method may be used as non-destructive technique for quality attributes measurement of intact mango and replace laboratory measurement method.

CURICULUM VITAE

Personal details

Full name	:	Agus Arip Munawar
Place, date of birth	:	Bandung, 9 August 1980
Nationality	:	Indonesian
Current address	:	• Albrecht Thaer Weg 14 a / 004, Göttingen 37075, Germany • Gutenberg Str. 33 Göttingen 37075, Germany
Permanent address	:	Jl. Rengasdengklok 4 No.29 Antapani, Bandung 40291 Jawa Barat - Indonesia.
E-mail address	:	amunawa@uni-goettingen.de // agusarif.munawar@yahoo.com

Formal higher education

2010 – 2014	PhD student, Division of Agricultural Engineering, Department of Crop Sciences, Georg-August Universität, Göttingen-Germany.

Publication during PhD study

Munawar, A. A., Hörsten, D.v., Pawelzik, E., Wegener, J. K., & Mörlein, D. (2012). Rapid and non-destructive assessment of SSC, TA and Ascorbic Acid in intact mango by NIRS. The Tropentag International conference on Research on Food Security, Natural Resource Management and Rural Development, 19-21 September 2012 in Göttingen, Germany.

Munawar, A. A., Hörsten, D.v., Pawelzik, E., Wegener, J. K., & Mörlein, D. (2013). Prediction of soluble solids content and acidity of intact mango by NIRS and Multivariate Analysis. *Gesellschaft für Informatik* (GIL) 2013 *Jahrestagung*, 20-22 February 2013 in Potsdam, Germany.

Munawar, A. A., Hörsten, D.v., Pawelzik, E., Wegener, J. K., & Mörlein, D. (2013). The application of near-infrared reflectance spectroscopy for quality attributes prediction of intact mango. *Deutsche Gesellschaft für Qualitätsforschung* (DGQ) 2013 *Jahrestagung* 2013, 18-19 March 2013 in Göttingen, Germany.

Research experiences

Year	Research topic
2002	Soluble solids content and firmness prediction in star fruit (*Averrhoa carambola*) with Near Infrared Reflectance technique.
2005	Simulation of Drying Rate and Conductivity Constanta in *Bandeng* fish Drying.
2008	Non-destructive Inner Quality Prediction in intact Mango with Near Infrared Reflectance Spectroscopy.
2009	Decision Support Software Design for Analyzing Head Loss and Pump Power Requirement in Irrigation Hydrodynamics System
2009	Inner Quality and Maturity Prediction in Mango using Near Infrared Sensor
2010	Computer Simulation and Modeling to Predict Ground Beef Thermal Properties During Freezing
2010	Planck Model Simulation to Predict Optimum Freezing Time and Temperature of Ground beef Freezing
2010	Sprinkle Irrigation Automation System for *Oyster* Mushroom Cultivation Using Soil Moisture and Temperature Sensor