DETECTION AND QUANTIFICATION OF ADULTERATION IN KHAO DAWK MALI 105 RICE USING NEAR-INFRARED SPECTROSCOPY COUPLED WITH CHEMOMETRICS

SAKUNNA WONGSAIPUN

DOCTOR OF PHILOSOPHY
IN CHEMISTRY

GRADUATE SCHOOL
CHIANG MAI UNIVERSITY
AUGUST 2019

DETECTION AND QUANTIFICATION OF ADULTERATION IN KHAO DAWK MALI 105 RICE USING NEAR-INFRARED SPECTROSCOPY COUPLED WITH CHEMOMETRICS

SAKUNNA WONGSAIPUN

A THESIS SUBMITTED TO CHIANG MAI UNVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN CHEMISTRY

GRADUATED SCHOOL, CHIANG MAI UNIVERSITY

AUGUST 2019

DETECTION AND QUANTIFICATION OF ADULTERATION IN KHAO DAWK MALI 105 RICE USING NEAR-INFRARED SPECTROSCOPY COUPLED WITH CHEMOMETRICS

SAKUNNA WONGSAIPUN

THIS THESIS HAS BEEN APPROVED TO BE A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN CHEMISTRY

Examination Committee:

R. Chantiwas ..... Chairman
(Assoc. Prof. Dr. Rattikan Chantiwas)

Parichat Theanjumpol ..... Member
(Dr. Parichat Theanjumpol)

Jaroon Jakmunee ..... Member
(Assoc. Prof. Dr. Jaroon Jakmunee)

Kongkiat Trisuwan ..... Member
(Asst. Prof. Dr. Kongkiat Trisuwan)

Sulawan Kaowphong ..... Member
(Asst. Prof. Dr. Sulawan Kaowphong)

Sila Kittiwachana ..... Member
(Asst. Prof. Dr. Sila Kittiwachana)

Advisory Committee:

Sila Kittiwachana ..... Advisor
(Asst. Prof. Dr. Sila Kittiwachana)

Parichat Theanjumpol ..... Co-advisor
(Dr. Parichat Theanjumpol)

Kongkiat Trisuwan ..... Co-advisor
(Asst. Prof. Dr. Kongkiat Trisuwan)

Sulawan Kaowphong ..... Co-advisor
(Asst. Prof. Dr. Sulawan Kaowphong)

1 August 2019

ACKNOWLEDGEMENT

I would first like to express my sincere thankfulness to my advisor, Asst. Prof. Dr. Sila Kittiwachana for the continuous support of my Ph.D. study and related research, for his supervisor, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Beside my advisor, I would like to thank my co-advisor Asst. Prof. Dr. Sulawan Kaowphong, Asst. Prof. Dr. Kongkiat Trisuwan and Dr. Parichat Theanjumpol for the kind help and useful recommendations. In addition, I would like to express my sincere appreciation to Assoc. Prof. Dr. Rattikan Chantiwas and Assoc. Prof. Dr. Jaroon Jakmunee for kindly giving valuable comments on my thesis.

I am very grateful to Dr. Parichat Theanjumpol from Postharvest Technology Research Center, Faculty of Agriculture, Chiang Mai University for supporting the NIR instrument and the beneficial information. I am also many thanks to technical staffs and officers of Department of Chemistry, Faculty of Science, Chiang Mai University, Thailand for improvement of my valuable skills, experiences and every facility.

Importantly, my appreciation is also expressed to the Science Achievement Scholarship of Thailand (SAST) and Graduate School of Chiang Mai University for financial support throughout my study.

I would like to take this special opportunity to express my sincere to all of my beloved friends for sharing their kindness, friendship and advices.

Finally, the special thanks are address to my lovely family and other members in my laboratory for their support, kind encouragement and suggestions all the time.

Sakunna Wongsaipun

หัวข้อดุษฎีนิพนธ์	การตรวจวัดและการวิเคราะห์หาปริมาณการปลอมปนในข้าวขาวดอกมะลิ 105 โดยใช้เทคนิคเนียร์อินฟราเรดสเปกโทรสโกปีร่วมกับเคโมเมตริกซ์
ผู้เขียน	นางสาวสกุลนา วงค์สายปัน
ปริญญา	ปรัชญาดุษฎีบัณฑิต (เคมี)
คณะกรรมการที่ปรึกษา	ผศ. ดร. ศิลา กิตติวัชนะ อาจารย์ที่ปรึกษาหลัก ดร. ปาริชาติ เทียนจุมพล อาจารย์ที่ปรึกษาร่วม ผศ. ดร. ก้องเกียรติ ไตรสุวรรณ อาจารย์ที่ปรึกษาร่วม ผศ. ดร. สุลาวัลย์ ขาวผ่อง อาจารย์ที่ปรึกษาร่วม

บทคัดย่อ

ในอุตสาหกรรมข้าวต้องการวิธีการวิเคราะห์ที่รวดเร็ว เป็นมิตรต่อสิ่งแวดล้อม ประหยัด และไม่ทำลายตัวอย่าง เพื่อประเมินคุณภาพของผลิตภัณฑ์ข้าว สำหรับข้าวขาวดอกมะลิ 105 (Khao Dawk Mali 105; KDML105) จำเป็นที่จะต้องทำการคัดกรองผลิตภัณฑ์ข้าวว่ามีการผสมหรือเจือปนโดยข้าวขาว (white rice; WR) ที่มีคุณภาพต่ำ และราคาถูกอื่น ๆ หรือไม่ หากมีการเจือปน ระดับการปลอมปนควรถูกวิเคราะห์เพื่อให้สามารถปฏิบัติได้ถูกต้องตามกฎหมายการติดฉลาก การใช้เทคนิคเนียร์อินฟราเรดสเปกโทรสโกปีร่วมกับเคโมเมตริกซ์ประสบความสำเร็จในการพัฒนามาสำหรับกระบวนการควบคุมคุณภาพ วัตถุประสงค์ในงานวิจัยนี้เพื่อตรวจวัดหาตัวอย่างข้าวขาวดอกมะลิ 105 ที่มีการปลอมปน และเพื่อหาปริมาณการปลอมปนที่แน่นอน โดยใช้เทคนิคการจำแนกประเภทแบบ one class เช่น เทคนิค D และ Q statistics และวิธี soft independent modeling of class analogy (SIMCA) เพื่อระบุการปลอมปนของข้าว การประเมินประสิทธิภาพการจำแนกประเภทจะพิจารณาค่าความไว ความจำเพาะ ความแม่นยำ และความถูกต้อง เป็นต้น แบบจำลองการจำแนกประเภทถูกสร้างขึ้นโดยอาศัยข้อมูลของข้าวขาวดอกมะลิ 105 บริสุทธิ์ และข้าวขาวบริสุทธิ์เป็นชุด training set ผลจากการวิเคราะห์พบว่า การใช้แบบจำลอง SIMCA ที่สร้างจากขาวดอกมะลิ

105 บริสุทธิ์ที่ผ่านกระบวนการปรับข้อมูลก่อนการวิเคราะห์ด้วยวิธีปรับความแปรปรวนให้เป็นมาตรฐาน (standard normal variate; SNV) ตามด้วยวิธีปรับข้อมูลเข้าสู่ศูนย์กลางโดยใช้ค่าเฉลี่ย (mean centring) ให้ผลการทำนายดีที่สุด โดยมีค่า ความไว ความจำเพาะ ความแม่นยำ และความถูกต้องเท่ากับ 68.42 99.15 90.70 และ 95.83 % ตามลำดับ สำหรับการวิเคราะห์หาปริมาณการปลอมปน เทคนิค orthogonal projection (OP) เทคนิค piecewise direct standardization (PDS) และเทคนิค partial least squares (PLS) ถูกนำมาใช้เพื่อให้ได้ผลการทำนายที่ดีที่สุด แบบจำลอง PLS ปกติ แบบจำลอง OP-PLS แบบจำลอง PDS-PLS และแบบจำลอง OP-PDS-PLS ถูกสร้างขึ้นเพื่อเปรียบเทียบประสิทธิภาพทำนายของแบบจำลองแต่ละแบบ จากผลการวิเคราะห์โดยพิจารณาจากแบบจำลองที่มีข้อมูลของข้าวขาวดอกมะลิ 105 และข้าวขาวที่ต่างชนิดกัน พบว่า แบบจำลอง PDS-PLS ปรับข้อมูลด้วย SNV ที่มีการลบข้อมูลของข้าวขาวออก ให้ผลการทำนายที่ดีโดยให้ค่า Q^2 และ RMSEP เท่ากับ 0.9887 และ 2.21 ตามลำดับ ในทางกลับกันข้อมูลที่มีการลบความแปรปรวนของข้าวขาวดอกมะลิ 105 ออก พบว่า แบบจำลอง OP-PDS-PLS ที่ปรับข้อมูลโดย SNV ให้ผลการทำนายที่ดีที่สุด โดยให้ค่า Q^2 และ RMSEP เท่ากับ 0.9889 และ 2.19 ตามลำดับ

Dissertation Title	Detection and Quantification of Adulteration in Khao Dawk Mali 105 Rice Using Near-infrared Spectroscopy Coupled with Chemometrics
Author	Miss Sakunna Wongsaipun
Degree	Doctor of Philosophy (Chemistry)

Advisory Committee		
	Asst. Prof. Dr. Sila Kittiwachana	Advisor
	Dr. Parichat Theanjumpol	Co-advisor
	Asst. Prof. Dr. Kongkiat Trisuwan	Co-advisor
	Asst. Prof. Dr. Sulawan Kaowphong	Co-advisor

ABSTRACT

Rice industry normally request rapid, environmentally-friendly, economical and non-destructive methodologies for evaluating the quality of rice products. Khao Dawk Mali 105 (KDML105) rice is necessary that the rice product should be screened if it has been mixed or adulterated by some other low quality and cheaper white rice (WR). The adulteration level should be quantified so that the rice quality can be complied with the regulation by law. The use of near-infrared (NIR) spectroscopy in conjunction with chemometrics has been recently developed for the quality control process. The aims of this research were to detect the adulterated KDML105 rice sample and to quantify the exact amount of adulteration. One class classification techniques such as D, Q statistics and soft independent modeling of class analogy (SIMCA) was used for identifying the rice samples. The values of sensitivity, specificity, precision and accuracy were used to evaluate the classification performance. The classification models were generated based on pure KDML105 and pure WR as a training set. The results indicated that using SIMCA model based on pure KDML105 with standard normal variate (SNV) followed mean

centring preprocessed provided the best class prediction with sensitivity, specificity, precision and accuracy of 68.42, 99.15, 90.70 and 95.83 %, respectively. For quantitative analysis, the methods of orthogonal projection (OP), piecewise direct standardization (PDS) and partial least squares (PLS) were utilized to obtain the best predictive ability. Four calibration models including ordinary PLS, OP-PLS, PDS-PLS and OP-PDS-PLS were established to compare the model efficiency. Considering all results based on the model having both difference of KDML105 and WR data, by removing the variation of WR, the model of PDS-PLS with SNV resulted in the best prediction having Q^2 and RMSEP of 0.9887 and 2.21. On the other hand, by removing the variation of KDML105, the established OP-PDS-PLS with SNV model provided the Q^2 and RMSEP values of 0.9889 and 2.19, respectively.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AOTF	Acousto-optic tunable filter
CCDs	Charged-coupled devices
cm	Centimeter
CT	Calibration transfer
DNA	Deoxyribonucleic acid
DO	Direct orthogonalization
DS	Direct standardization
EDA	Exploratory data analysis
FT-NIR	Fourier transform near infrared spectroscopy
g	Gram weight
IIR	Independent interference reduction
IR	Infrared
KDML105	Khao Dawk Mali 105
LEDs	Light-emitting diodes
LV	Latent variable
MLR	Multiple linear regression
MSC	Multiplicative scatter correction
NIPALS	Non-linear iterative partial least squares
NIR	Near infrared
nm	Nanometer

NOC	Normal operating condition
ºC	Degree celsius
OP	Orthogonal projection
OSC	Orthogonal signal correction
PCA	Principal component analysis
PCR	Principal component regression
PCs	Principal components
PDS	Piecewise direct standardization
PLS	Partial least squares
PLS-VIP	Partial least squares-variable influence on projection
PRESS	Predictive error sum of squares
PT1	Pathum Thani 1
QDA	Quadratic discriminate analysis
R^2 and Q^2	Coefficient of determination
RD15	Rice Department 15
RD49	Rice Department 49
RD57	Rice Department 57
RMSEC	Root mean square error of calibration
RMSEP	Root mean square error of prediction
RSS	Residual sum of squares
SIMCA	Soft independent modeling of class analogy
SNV	Standard normal variate
SOM	Self organizing map

SPE	Squared prediction error or Q statistic
SVD	Singular value decomposition
SVDD	Support domain data description
SVM	Support vector machine
$\boldsymbol{T}^2$	Hotelling's $\boldsymbol{T}^2$ statistic
TSS	Total sum of squares
UV	Ultraviolet
WR	White rice

LIST OF SYMBOLS

InGaAs	Indium gallium arsenide
PbS	Lead sulfide
%w/w	Percent weight by weight
$(\boldsymbol{X}_s^T\boldsymbol{X}_s)^{-1}$	Pseudoinverse of $\boldsymbol{X}_s$
$\overline{\omega}_{ij}^{d}$	Contribution of the variable j^{th} in $\boldsymbol{x}_i$ to the D statistic
$\bar{x}_j$	Average for variable j^{th} calculated over all I samples in $\boldsymbol{X}$
$\hat{y}_i$	Predicted response value of sample i^{th}
$\hat{\boldsymbol{p}}$	Predicted loadings vector
$\boldsymbol{p_r}$	Loadings vector of r component
$q_{lim\propto}$	Confidence level for the Q statistic at α
$\hat{\boldsymbol{t}}$	Predicted score vector
$\boldsymbol{t_r}$	Scores vector of r component
$\bar{\boldsymbol{x}}$	Vector of the mean value corresponding to each variable
$x_{i,obs}$	Observed data
$\bar{y}$	Average of all observed response values
y_i	Response value of sample i^{th}
|| ||	Normalization

´	Transpose
A	Number of PCs
$\boldsymbol{c}$	Response vector
d_i	D statistic value of sample i^{th}
$\boldsymbol{E}$	Residual matrix
e_{ij}	Residual error at the row i^{th} and the column j^{th}
F	F-distribution
$\boldsymbol{F}$	Transformation matrix
g_r	Eigenvalue of the PC r^{th}
h_i	Unwanted data
I	Number of samples in data matrix
J	Number of variables in data matrix
$\boldsymbol{P}$	Loadings matrix
$\boldsymbol{P}_G$	Loading matrix of $\boldsymbol{X}_G$
$\boldsymbol{q}$	$\boldsymbol{c}$-loadings
q_i	Q statistic value of sample i^{th}
r	A number that is subset of R
R	Optimum number of PCs
$\boldsymbol{S}$	Diagonal matrix of the standard deviation of each variable
$\boldsymbol{T}$	Scores matrix
T	Transpose
$\boldsymbol{T}_G$	Score matrix of $\boldsymbol{X}_G$
$\boldsymbol{V}$	Covariance matrix of the residual matrix $\boldsymbol{E}$

V_r	Percentage eigenvalue for the component number r
$\boldsymbol{W}$	Normalized PLS weights
$\boldsymbol{X}$	Data matrix X
$\boldsymbol{X}_{\text{exp}}$	Data matrix without orthogonal variation
$\boldsymbol{x}_i$	Sample i^{th} of data matrix X
x_{ij}	Value at the row i^{th} and the column j^{th} of data matrix X
$\boldsymbol{X}_{\text{m}}$	Data matrix measured on the master instrument
$\boldsymbol{X}_{\text{mix}}$	Data matrix of all variation (wanted and unwanted)
$\boldsymbol{X}_{\text{s}}$	Data matrix measured on the slave instrument
$\boldsymbol{X}_{\text{WR}}$	Data matrix contains variation of white rice (unwanted)
$\boldsymbol{X}_{\text{WR in mix}}$	Data matrix of variation of white rice in mixed sample
$\boldsymbol{y}_{\text{obs}}$	Observed response or percentage of w/w of KDML 105
$\boldsymbol{Z}$	Pre-processed data matrix Z
z_{ij}	Value at the row i^{th} and the column j^{th} of data matrix Z
z_{α}	Standardized normal variate with $(1-\alpha)$ confidence limit
α	Confidence level
θ_1	Trace of $\boldsymbol{V}$
θ_2	Trace of $\boldsymbol{V^2}$
θ_3	Trace of $\boldsymbol{V^3}$

ข้อความแห่งการริเริ่ม

1. วิทยานิพนธ์นี้นำเสนอวิธีการทางเคโมเมตริกซ์ที่รวดเร็ว ถูกต้อง แม่นยำ ราคาไม่แพง และไม่ทำลายตัวอย่างที่สามารถใช้เพื่อตรวจวัดและหาปริมาณการปลอมปนในข้าวขาวดอกมะลิ 105 ได้ แม้ว่าข้าวขาวหรือข้าวที่นำมาปลอมปนไม่ถูกระบุว่าเป็นชนิดใด โดยอาศัยการตรวจวัดด้วยเทคนิคเนียร์อินฟราเรด (เอ็นไออาร์) สเปกโทรสโคปี

2. วิธีการวิเคราะห์ด้วยเทคนิคเคโมเมตริกซ์ร่วมกับเอ็นไออาร์สามารถรับมือกับอิทธิพลจากความแปรปรวนที่เกิดจากความแตกต่างของชนิดข้าวขาวดอกมะลิ 105 และข้าวขาวได้ ความท้าทายที่สำคัญคือ การพัฒนาแบบจำลองที่คงทนและแม่นยำสามารถใช้กับข้าวขาวที่นำมาปลอมปนได้ทุกชนิด แทนการใช้แบบจำลองที่มีข้อมูลข้าวขาวเพียงชนิดเดียว หรือข้าวขาวที่มีความจำเพาะเท่านั้นมาใช้ในการสร้างแบบจำลองในการทำนาย

3. การประยุกต์ใช้วิธีการทางเคโมเมตริกซ์ รวมไปถึงเทคนิค orthogonal projection (OP) เทคนิค calibration transfer (CT) และเทคนิค partial least squares (PLS) regression สามารถปรับปรุงการทำนายระดับการปลอมปนในข้าวผสมได้ทุกชนิด แม้ว่าตัวอย่างข้าวผสมจะมาจากข้าวขาวดอกมะลิ 105 และข้าวขาวต่างชนิดกัน ผลการทดลองนี้แสดงให้เห็นว่าวิธีการที่พัฒนาขึ้นสามารถนำไปใช้ประโยชน์กับตัวอย่างจริงในอุตหกรรมข้าวได้

STATEMENTS OF ORIGINALITY

1. This thesis proposes a rapid, accuracy, precision, cost-effective and non-destructive chemometric methodology that can be used to detect and quantify the adulteration in Khao Dawk Mali 105 (KDML105) rice grain although the white rice or substituent rice are not identified based on the detection of near-infrared (NIR) spectroscopy.

2. The chemometric-NIR methodology can cope with the influent variations from the different types of KDML105 and white rice. The major challenge is to develop the robust and accurate universal models can be applicable to all substituent white rice rather than models that can only be applied successfully to single or specific substituent white rice that is used to establish the prediction model.

3. The application of chemometric methods, including orthogonal projection (OP), calibration transfer (CT) and partial least squares (PLS) regression, can improve the prediction of adulteration levels in all mixed sample types even though the mixture samples are from the different types of KDML105 and white rice. This result implies that the developed method can universally utilize with real sample in rice industry.

CHAPTER 1

Introduction

1.1 Introduction to the research problem and its significance

Khao Dawk Mali 105 (KDML105) is traditional Thai Hom Mali rice which is well known for its fragrance and taste. This aromatic rice is not only popular in Thailand but also around the world. Since KDML105 rice is more expensive than the other white rice or non-glutinous rice, some manufactures add some substituent rice into KDML105 to gain more profit in rice trade causing adulteration. According to the notification of the Ministry of Commerce of the Kingdom of Thailand, subject: prescribing Thai Hom Mali rice as a standardized commodity and the standard of Thai Hom Mali rice (third edition) B.E. 2559 [1], the purity of Thai Hom Mali rice should be minimum at 92.0% by quantity so that the product can be labeled as "THAI HOM MALI RICE". In addition, the purity should be greater than 98.0% so the product can be labeled as "PRIME QUALITY THAI HOM MALI RICE". The presence of the substitution more than the regulation reveals deliberate substitution, which is unless stated, would be illegal under food labelling rules. Therefore, it can be seen that the quality control of KDML105 is a unique and complicate task. Firstly, it is important to precisely detect if the rice is adulterated and, secondly, the level of the rice substituent or the adulteration level should be accurately quantified.

Several methods can be used to detect adulteration in rice. A test panel based on human sensory is a common practice in milling industry, but this procedure could be subjective, and much relied on personal experience. Recently, some protein and DNA based techniques have been developed and proved successful over the conventional methods [2]. Nevertheless, these techniques deal with complex chemicals and complicated sample pretreatments are required for these techniques. In addition, it is not convenient to incorporate the test procedures for continuous

online measurement in manufacturing plants where there are a large number of samples. Near-infrared (NIR) spectroscopy which is a vibrational technique offers a straightforward, non-destructive, rapid and cost-effective alternative detection [3]. Using a reflectance mode, NIR spectroscopy allows measurements without or less sample preparation for solid samples suitable for rice grain. Hence, this analytical technique could meet the requirements of industrial applications as well as for continuous quality control and process monitoring.

Recently, NIR spectroscopy was applied to identify the difference between the rice varieties. However, there is a problem when employing NIR spectroscopy to quantitatively evaluate the mixing levels between KDML105 and some other blending rice. A calibration model can provide good predicting performance only if calibration (training) and unknown (test) samples are parts of the same population. Simply, variation of unknowns should be matched or within the variation of the samples used for the training set. Since, rice cultivars differ significantly in their morphological and chemical properties. These factors strongly depend on the wide diversity of rice germplasm as well as starch and protein compositions. For this reason, NIR with traditional chemometric models such as partial least squares (PLS) regression hardly gives accurate prediction results if unknown samples are provided with limited information or the variety of the blended rice was not known. For example, PLS model, built using KDML105 adulterated with a substituent rice, called Chai Nat 1, can be specifically used only for predicting the adulteration in the rice samples adulterated with Chai Nat 1. The developed model cannot provide accurate predictive results to KDML105 adulterated with other rice varieties, for example, RD 49, RD 57 and Phitsanulok 2. Therefore, a critical challenge is to develop a robust and accurate universal model that is applicable to all types of the substituent rice rather than the model that can be successfully applied only to a single mixed rice sample which is used for establishing the model.

This research proposed a methodology to improve the calibration performance of PLS model applied to NIR spectroscopy when the model was confused by variation from unknown (adulterated) rice. The purpose was to remove the variation from non-aromatic rice in the NIR spectra prior to the modeling using some data

filtration methods called orthogonal projection (OP) methods. Using OP, a structure of non-aromatic rice is modeled at the first step. After that, the variation from the non-aromatic rice can be discarded from the dataset based on the previous structural model to obtain pure variation of the KDML105 rice. Then, a unimodal calibration can be created using the pure variation and standardization methods in particular calibration transfer methods. It is expected that the regression is solely based on KDML105 rice as the variation of the non-aromatic rice is modeled and removed prior to the prediction.

1.2 Literature review

Rice is a staple food for over one-third of the world's population [4]. Rice is not only consumed as cooked rice but also used as a raw material for many products such as noodle, snacks, instant noodle and cosmetics. Currently, there are a number of rice varieties but only a few varieties are known worldwide. Khao Dawk Mali 105 (KDML105) is among the most popular rice varieties in Thailand and is often called as "Thai Hom Mali rice" or "Thai jasmine rice". KDML105 has desirable characteristics such as the good cooking quality and aroma, and therefore it is of great demand not only in Thailand but also to the other countries [5]. The demand of aromatic rice has steadily increased while it has a low yield because its sensitivity to pest and diseases [6]. Therefore, the aromatic rice is insufficient to market demand and leads to the high price. Due to its higher price, KDML105 is often blended with some other rice grains to get more profit in rice trade causing adulteration or blending in rice product.

To comply with the notification of the Ministry of Commerce of the Kingdom of Thailand, subject: prescribing Thai Hom Mali rice as a standardized commodity and the standard of Thai Hom Mali rice (third edition) B.E. 2559 [1], the blending level should be restricted to maintain the rice products with the desirable properties such as color, volume after cooking, taste, aroma and texture. For a rice product labeled as "THAI HOM MALI RICE", the concentration of KDML105 should be greater than 92.0% by quantity. Also, this blended rice should be cheaper than the rice product labeled as "PRIME QUALITY THAI HOM MALI RICE" or 98.0% purity. The presence of the substitution more than the regulation reveals deliberate substitution, which unless stated, would be illegal under food labelling rules.

Nowadays, the adulteration and fraudulent mislabeling is the main problem in food industry [7]. There are several researches have studied the adulteration in food products, for example, milk powder [8], honey [9], vegetable oils [10] and rice samples [7, 11, 12]. At the present, various methodologies have been developed for detection the adulteration of rice. Table 1.1 summaries the previous methods used for detecting adulteration in rice.

Table 1.1 The properties used for rice adulteration

Properties	Aims	Samples	Methods	Conclusion
DNA [12]	Detect the quantity of non-Basmati rice present in Basmati rice	Basmati and non-Basmati rice	High resolution melting (HRM)	DNV based method and HRM protocol can detect and quantify adulterations in rice varieties.
Pasting properties [13]	Jasmine rice with other varieties based on their pasting properties	KDML105, PT1 and CNT1	Texture analysis PCA and PLS	Pasting properties can be used for classification and prediction the adulteration in rice.
Texture properties [13]	Jasmine rice with other varieties based on their pasting properties	KDML105, PT1 and CNT1	RVA analyser PCA and PLS	Texture properties can be used for classification and prediction the adulteration in rice.
Shape and size [14]	Identify the difference among three rice cultivars	Coarse (IR-8), fine (PR-106) and superfine (Basmati-386) rice cultivars	Vernier caliper	Shape and size can be used for distinguishing the difference of rice samples.
Amylose and protien contents [15]	Compare the chemical properties of rice cultivars with different origins	Rough rice samples of Bengal and Medark	Iodine colorimetry Micro-Kjeldah	The variations in rice chemical characteristic can identity the genetics,location and crop year of rice.

In Table 1.1, the reported methods usually applied to distinguish each rice variety based on their specific properties. For examples, the distinctness between varieties was identified by determination of physico-chemical properties [13], genotype (DNA-based method) [12], physical characteristics (length, breadth, weight, density) [14] pasting properties [15], and chemical properties (amylose, protein, aroma compositions) [11, 15]. It is possible that the previous developed methods could be used to identify adulteration of rice. However, they provide unsatisfactory result due to their low accuracy [11]. Especially when Thai jasmine rice was blended with other white rice having similar amylose content, it was difficult to estimate the adulteration based on their physico-chemical properties [13].

(A) (B)

Figure 1.1 (A) KDML105 and (B) WR rice grains

The difference between KDML105 and some other white rice (WR) can be seen in Figure 1.1. It is noted that all of the rice samples are considered as non-waxy rice or non-glutinous rice. Although the non-waxy rice usually has high concentration of amylose starch meaning that all of them has a low amount of amylopectin. From the physical properties such as size, shape and color, most of them possess the similar characteristics. When the rice grain was mixed as shown in Figure 1.2, it is not easy to identify if the rice is a combination or it is the pure rice. In addition, when the rice was mixed by the different concentration levels, the difference was hardly recognized using naked eyes.

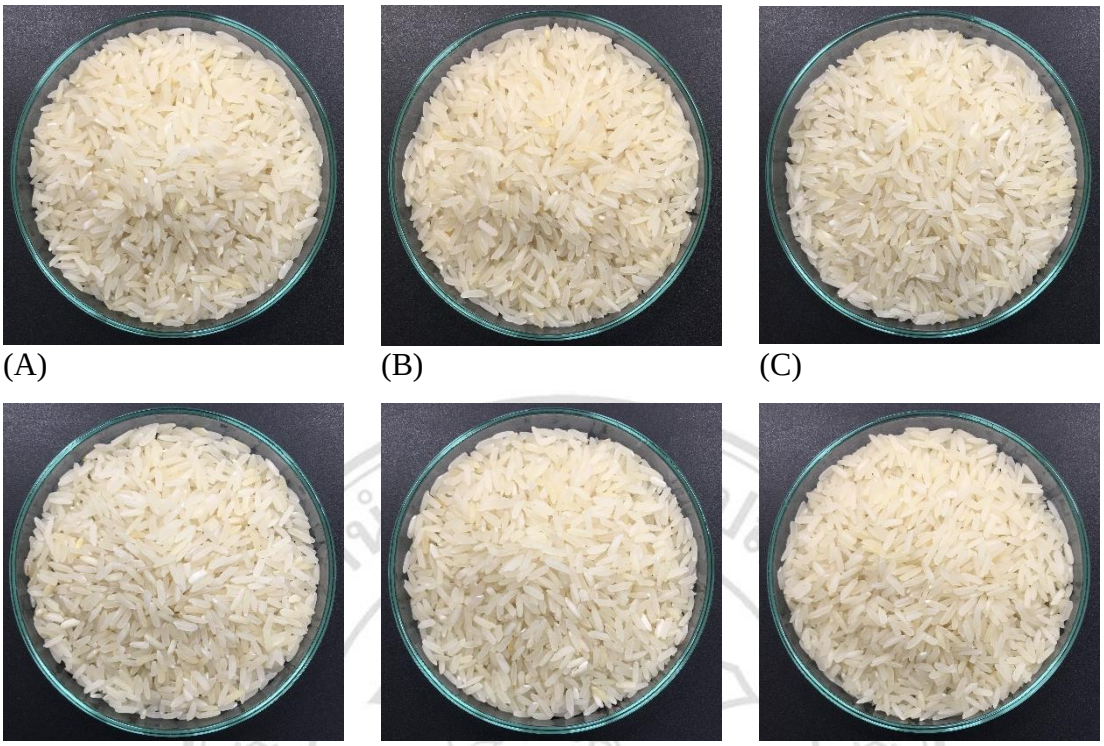


Figure 1.2 The 50 %w/w mixing between grain between (A) KDML105 and WR1, (B) KDML105 and WR2 and (C) KDML105 and WR3. (D) – (F) The KDML105 rice grain mixed WR1 with different concentration levels from left to right (20, 60 and 95 %w/w)

Apart from that, some of the previous techniques especially the DNA-based methods were time-consuming and inconvenient methods if used in industry. NIR spectroscopy is among the vibrational techniques that covers wavelengths from 780 nm up to 2500 nm. This technique offers a straightforward, non-destructive, rapid and cost-effective detection [16]. It also allows non-complicated measurements without or less sample preparation. Hence, this analytical technique could meet the requirements of many industrial applications as well as for continuous quality control and process monitoring [17]. However, NIR is the detection of overtone and combination of the characteristic IR region so that the recorded signals are usually broad with highly convoluted peaks [18]. Consequently, it is not easy to interpret NIR spectra using unaided eye with conventional statistic methods. Therefore, a more sophisticated algorithm with mathematical and statistical tools or “chemometrics” is required to extract analytical information from the corresponding spectra.

Chemometrics could be regarded as the combination uses of mathematics, statistics and computer science for ease the interpretation of complicated scientific data especially chemical data [19]. Chemometric methods can be categorized into groups according to their purpose such as data exploratory, classification, calibration, variable selection and design of experiment. This thesis reported the use of the exploratory, classification and calibration methods to deal with the analysis of rice samples. Therefore, they will be introduced in this review.

Data exploratory or exploratory data analysis (EDA) [19] aims to investigate the relationship between the studied samples. This relationship information could be used to evaluate if the samples are similar or different. In the other word, EDA looks for the pattern in the data. Principal component analysis (PCA) and self-organizing map (SOM) are some the most popular EDA methods. PCA simply creates new variables or factors that are usually called principal components (PCs) based on the assumption that the first few PCs could contain as much as possible the variation in the data. Therefore, the plots of PC1 against PC2 (or so on) could represent the main characteristics of data. For example, Figure 1.3(B) shows the score plot between PC1 and PC2. This result demonstrates the use of PCA to exploratorily analyze the raw NIR spectra in Figure 1.3(A). It can be seen that PCA visualize the data using a newly generated PC space. On the contrary, SOM represents the data using a generated map composing of a number of grid units as shown in Figure 1.3(C). Ideally, PCA allocates the samples into the linear regression space so it expects that the data follows multivariate normal distribution. On the other hand, SOM adopts the grid units to be similar to the data structure itself so that it does not need that the data can be explained based on the linear regression equations.

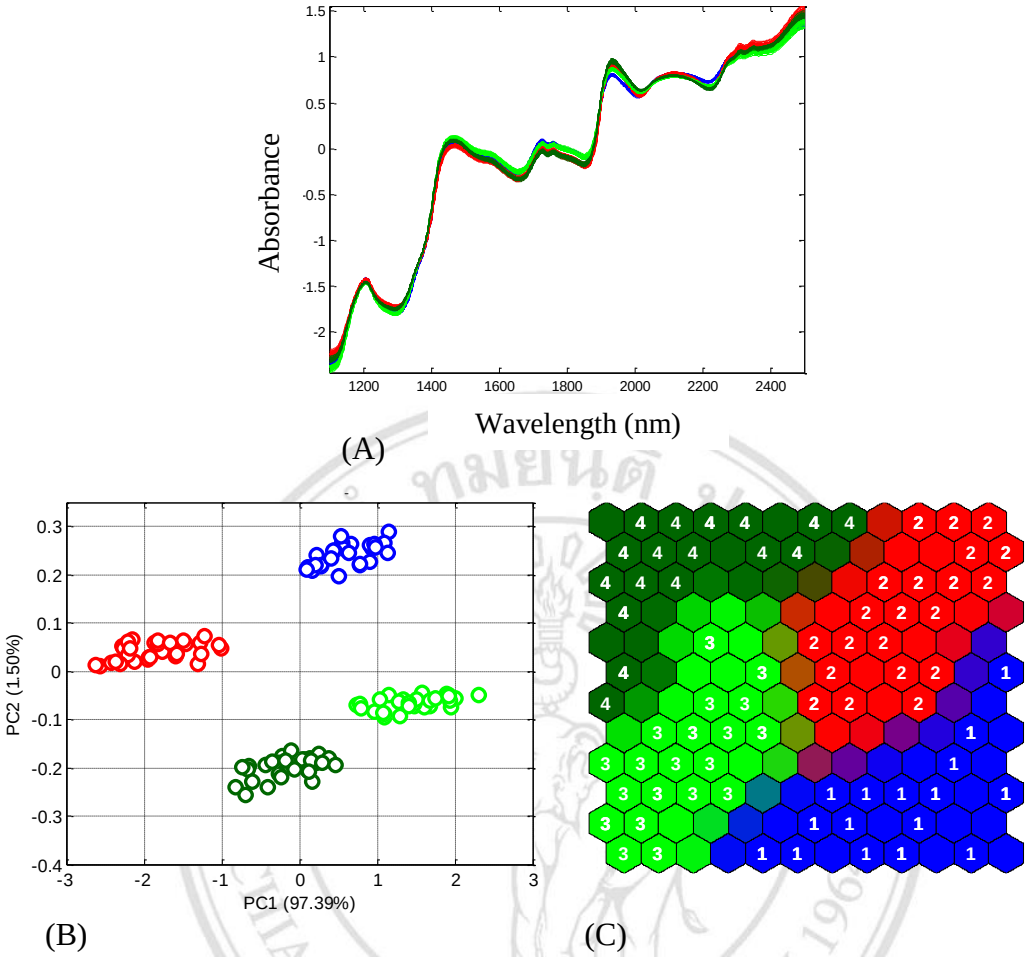
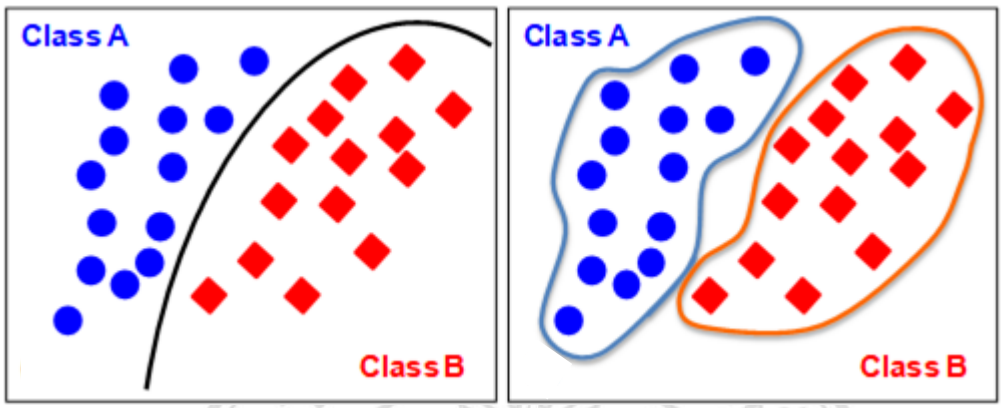


Figure 1.3 (A) NIR spectra of coffee been having different moisture contents, (B) PCA scores plot of PC1 against PC2 and (C) supervised color shading SOM of the NIR spectra where the different colors indicate the different moisture contents

Multivariate classification builds a boundary around sample classes and identifies the class region to which an unknow sample should belong to. Normally, classification methods can be categorized into two groups, one class and multiclass classification [20]. In multiclass classification, classifiers divide hyperspace into regions according to the number of classes whether an unknown sample falls in the region corresponding to a particular category. The number of sample classes should be predefined, and an unknown sample is classified into a group according to its position in the data space. Unlike multiclass classification, one class classifiers, sometimes called untargeted technique, focus on identifying the difference among the sample classes and model each class independently [21]. For example, if there are three classes, three different models are formed for each class and any unknowns are independently assigned to each of the

three models. Therefore, the classification criteria for one class modellings are probabilistic and their class boundaries can be overlapped in the data space [22].



(A) (B)

Figure 1.4 Difference between (A) a single support vector machine (SVM) model separating two classes and (B) two support vector data description (SVDD) models for each class [24]

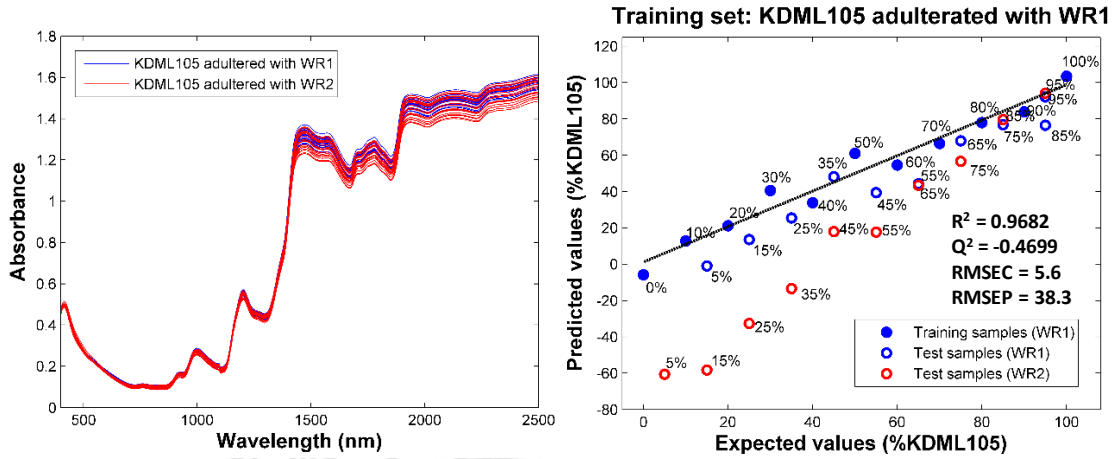
Figure 1.4 illustrates the difference between one class and multiclass classifications. Using multiclass classification, the model space is divided into regions according to the class membership number. In Figure 1.4(A), there are two regions belonging to class A and class B. The prediction is corresponding to the location of the unknown sample on the model space. Using one class classifiers such as support domain data description (SVDD) in Figure 1.4(B), D and Q statistics, a model is formed just using one group of samples [23]. Unknowns (or test samples) are then assigned to this group according to a predefined confidence limit. If they are outside the limit, they are not considered to be part of this group. This can be the advantage of one class classification over the multiclass classification for the specific problem in the detection of adulteration in rice. In addition to SVDD, soft independent modeling of class analogy (SIMCA) [24] is commonly used for classification. SIMCA makes classification decision based in the distributions of systematic and non-systematic variation in the data. Usually, the systematic variation and non-systematic variation can be evaluated using D and Q statistics, respectively. Although the calculation of SIMCA is quite less complicate, this one class classification method could provide satisfactory prediction in many applications. For example,

Zábrodská and Vorlová (2015) reported the use of SIMCA for identifying the honey samples adulterated by man-made sugar [25].

Multivariate calibration, known as targeted technique, involves the building of a model using the relationship between two sets of variables, usually defined as 'predictive' variables (independent or '$\boldsymbol{X}$' block) and 'response' variables (dependent or '$\boldsymbol{c}$' block). Examples of predictive variables include physical measurements, such as spectra, temperature, pressure and flow rate. Responses are properties of interest, such as concentrations of compounds in mixtures which are difficult to measure directly. Multivariate calibration always involves two stages: (1) modelling where a calibration model is constructed using samples with known properties or 'training samples'; and (2) prediction which entails the prediction of properties of unknown samples or 'test samples' based on the relationship model built in the first stage. Multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) are among the most common methods in multivariate calibration [26]. Among these methods, PLS regression may be regarded as the most powerful approach, in contrast to MLR and PCR, as it assumes that the variations in both predictive and the response variables are equal, and hence the errors in both directions are considered in the model equally.

NIR spectroscopy in the combination to PLS could be used to estimate the adulteration level. However, there is a problem when employing NIR spectroscopy to quantitatively evaluate the mixing levels between KDML105 and some other blending rice. Because, to obtain a calibration model with satisfy predicting performance, calibration (training) and unknown (test) samples need to be part of the same population. Simply, variation of unknowns should be matched or within the variation of the samples used for the training set. Since, rice cultivars differ significantly in their morphological and chemical properties. These factors strongly depend on the wide diversity of rice germplasm as well as starch and protein compositions. For this reason, NIR with traditional chemometric models such as partial least squares (PLS) hardly gives accurate prediction results if unknown samples are provided with limited information or the variety of the blended rice is not known. For example, PLS model is built using KDML105 adulterated with a substituent white rice, called WR1, can be specifically used only for predicting the adulteration in the rice samples adulterated with WR1. The developed

model cannot provide accurate predictive results to KDML105 adulterated with other rice varieties, for example, white rice 2 (WR2). The result shows in Figure 1.5.



(A) (B)

Figure 1.5 (A) NIR spectra of two mixtures of KDML105 + WR1 and KDML105 + WR2 and (B) a correlation graph of the PLS models established using KDML105 blended with WR1

Figure 1.5 shows results from a preliminary study using NIR spectroscopy for predicting the concentrations of KDML105 adulterated by two different Thai white rice, WR1 and WR2. The NIR spectra of the KDML105 adulterated with WR1 and WR2 are shown in Figure 1.5(A) labeled using blue and red lines, respectively. Figure 1.5(B) is the correlation graph showing the expected and predicted concentrations of the studied KDML105 rice using conventional PLS1. The PLS model established using the WR1 training set could provide significant better results when it is used to predict the contamination levels of its own variety (the contamination of WR1 in the test samples labeled using the open-faced blue circles). Whereas the prediction of the test samples contaminated with WR2 are deviated from the regression line implying poor predictive results in the open-faced red circles. Considering the model statistics, root mean square error of prediction (RMSEP) of the PLS models trained using the WR1 contamination was significantly increased from 5.6 to 38.3 when it is tested with the rice samples adulterated with WR2. This implies that the predictive results are dramatically deteriorated when it is used to predict the samples of different types.

One possible solution is to remove the variation from non-aromatic rice in the NIR spectra prior to the modeling using some data filtration methods such as orthogonal

projection (OP). Ideally, OP was invented the separate the variation in data, for example, systematic and non-systematic variations. Based on the similar idea, OP could be used to identify the variation for aromatic and non-aromatic rice. Using OP, a structure of non-aromatic rice is modeled at the first step. After that, the variation from the non-aromatic rice can be discarded from the data set based on the previous structural model to obtain pure variation of the KDML105 rice. Then, a unimodal calibration can be created using the pure variation. It is expected that the regression is solely based on KDML105 rice as the variation of the non-aromatic rice is modeled and removed prior to the prediction. Several OP methods have been proposed recently [27]. However, independent interference reduction (IRR) could be regarded as the most successful methods for projecting the target variation in calibration and classification analyses [28].

The other solution is calibration transfer (CT). CT is a chemometric method that can be used to standardize the variation between analytical instruments so that the calibration models between instruments can be transferred [29]. This allows the signals from a new (slave) instrument can be matched to the old (primary or master) instrument. This CT method has been widely used coupled with NIR spectra for standardization the model to remove the interfered variation from using different instrument, detector instability, time and temperature [29]. The common calibration transfer based on standardization consist of direct standardization (DS) and piecewise direct standardization (PDS). The principal of DS and PDS is the same as both create an effective transformation matrix F which is directly related to the samples of transfer set. Then the transformation matrix is used to modify the spectra of test sample to predict in the calibration model. The different between DS and PDS is that the DS uses whole spectra for establishing the F. Whereas, the PDS firstly uses a small subset limited by window size k of the spectra measured on the first instrument called master instrument to relate with the spectra measured on the second instrument called slave instrument and then calculate the regression coefficient based on PLS method. After the coefficient of all subsets are obtained, the transform matrix will be occurred. The steps of PDS was thoroughly described by Chen [30]. The PDS usually provides the better result than DS due to the use of a small subset. Because the spectral variations are often limited to much smaller region. Therefore, each spectral data of master would relate to the spectra

measurements at nearby wavelengths than the full spectrum of slave. Consequently, the PDS will be used as a calibration transfer in this task.

In this research, a methodology based on the use of near infrared (NIR) spectroscopy and chemometrics will be developed to detect or identify the adulteration in KDML105, untargeted analysis based on SIMCA will be used. If the rice sample is classified as adulterated, then this will be used as a test set for quantitative analysis. The next step is to quantify the adulteration level of KDML105 in the rice sample. To perform this quantification, a novel calibration method will be proposed. In the first step, an OP method will be used to remove the unwanted variation from training set. Then, a CT method will be used to standardize the preprocessed spectra to improve the predictive performance. Finally, the treated spectra will be applied with PLS to establish the calibration model to predict an amount of KDML (%w/w) in adulterated samples. This developed method can rapidly detect and accurately quantify the rice adulteration, although the white rice or substituent rice are not identified.

1.3 Research objectives

1.3.1 To detect the adulteration in KDML105 rice using NIRS and chemometrics

1.3.2 To quantify the adulteration level (%w/w) of KDML105 in adulteration rice samples

1.3.3 To apply the novel chemometric method for determination of the adulteration in rice grain samples

CHAPTER 2

Theory and methods

2.1 Characteristics of Khao Dawk Mali 105 rice

Khao Dawk Mali 105 (KDML105) rice has been discovered since 1959 by a small group of researchers from Thailand [31]. This rice variety has a unique quality which is especially matched for eating with Thai cuisine. This name “Khao Dawk Mali” is from its grain color which is as white and pure as a jasmine flower. Therefore, this rice often called as “Thai Jasmine rice”. In addition, the other names “Hom Mali” or “Thai Hom Mali” are widely used due to its present fragrance after cooked. In fact, Thai Hom Mali rice consists of two dominant cultivars; KDML105 and the other which is called “Rice Departments 105 (RD15)”. Ideally, the characteristics of RD15 rice grain is very similar to that of KDML105 since the RD15 rice plant was genetically developed from KDML105 rice plant so that the rice requires a slightly shorter time for cultivation [32].

(A) (B)

Figure 2.1 Grains of KDML105; (A) paddy and (B) milled rice

Figure 2.1 shows the KDML105 grain before and after milling. In general, KDML105 is a rice with high amylose content so it is classified as non-waxy or non-glutinous rice. This implies that the cooked rice has a softer and less sticky texture than the high amylopectin or waxy or glutinous rice.

2.2 Adulteration in rice

Due to their good properties after cook, the KDML105 is the most requested among others. The demand of KDML105 in the world market is increasing but the annual production does not fulfill the demand of international market [33]. Therefore, the price of KDML105 is relatively higher than other rice especially some other white rice such as RD 49, RD 57, Chai Nat 1, Chai Nat 2 and Phitsanulok 2. This leads to the blending of rice, in other words, rice adulteration. The adulteration or fraudulent mislabelling or rice is the major problem in several areas.

It is important to note that the words “adulteration”, “fraudulent”, “falsification” and “contamination” could have similar meanings and often they are interchangeably used. Ideally, the terms contamination and adulteration are usually used when some addition substances are mixed into the products and they could cause the changes which may be harmful to the consumers. For example, the addition of melamine which is not eatable chemical compound in milk powder [34]. In this case, the contaminated products could have a serious adverse effect on the baby. On the contrary, adulteration implies that the addition should not cause any adverse effects on the consumer. For example, the addition of some WR in the KDML105 rice only alters the quality of the rice product and the rice still should be eatable safely. To prove the adulteration of a product can be closely related to prove its authentication. The prove of authentication implies that the quality of a product is specifically according to its label. For example, if a rice product is labeled as KDML105 from Chiang Mai province, the rice should be only the KDML105 that was cultivated in Chiang Mai province or else the rice is not authentic. Therefore, to limit the scope and avoid confusion, the term adulteration is used in this thesis.

2.3 Near-infrared (NIR) spectroscopy

2.3.1 Background of NIR

Historically, the NIR radiation was discovered in year 1800 by a musician and very successful amateur astronomer, William Herschel [35]. He found that the heat maximum was beyond the red end of the spectrum after separated the electromagnetic spectrum with a prism that is now called the near-infrared. Over eighty years later, Abney and Festing investigated the liquid samples using near-infrared and obtained the first infrared spectra in year 1881. They suggested that the absorption was related with the chemical composition of samples. Near-infrared (NIR) is the electromagnetic radiation in wavelength region of 800 - 2,500 nm corresponding to the wave number range at 12,500-4,000 cm^{-1} [36]. This region is between visible and IR regions that shown in Figure 2.2.

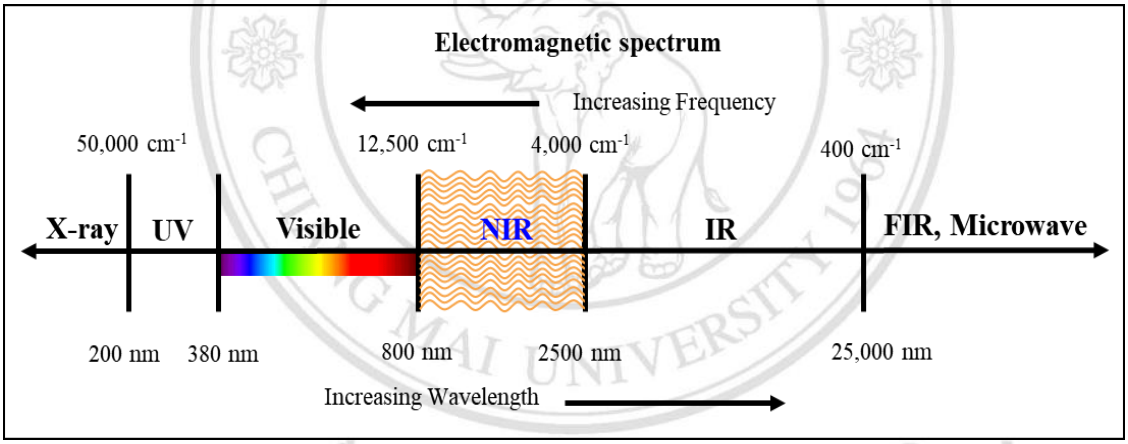


Figure 2.2 Electromagnetic spectrum [36]

NIR radiation is absorbed by the molecules in samples. If the radiation matched with the frequency of vibration, it would be absorbed, and they can produce excitation to a higher vibrational energy level. This energy stimulates the vibrations of the molecules in two different vibrations (stretching and bending) of functional groups (OH, CH, NH and C=O). The absorptions in the NIR region are generated from fundamental vibration of the functional groups in samples are overtones and combinations. Overtones are the harmonics of fundamental absorptions at multiples of the frequency. Combinations arise from sharing of NIR energy between two or more fundamental absorptions. The vibrations of functional group which relate to NIR wavelength is illustrated in Figure 2.3.

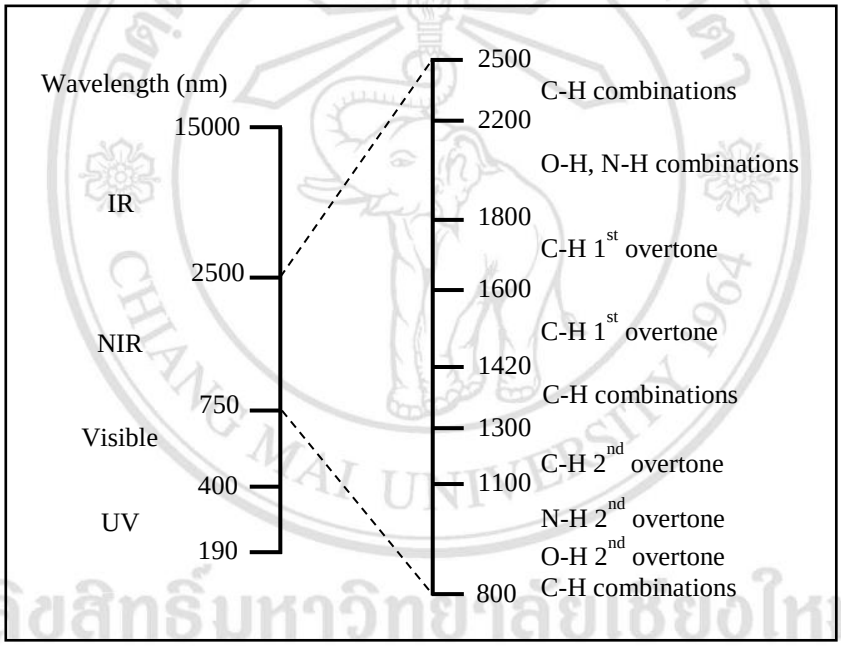


Figure 2.3 The NIR range and the vibration modes of their corresponding functional groups [37]

The effect of these absorptions trend to make the NIR spectra look complicate and consist of only a few rather broad peaks. For example, NIR spectrum of biscuit dough is shown in Figure 2.4. It contains of various molecules corresponding the ingredients. Normally, there should have many absorption peaks, however, the NIR spectrum expressed in the integration of absorption band which are low and broad peaks.

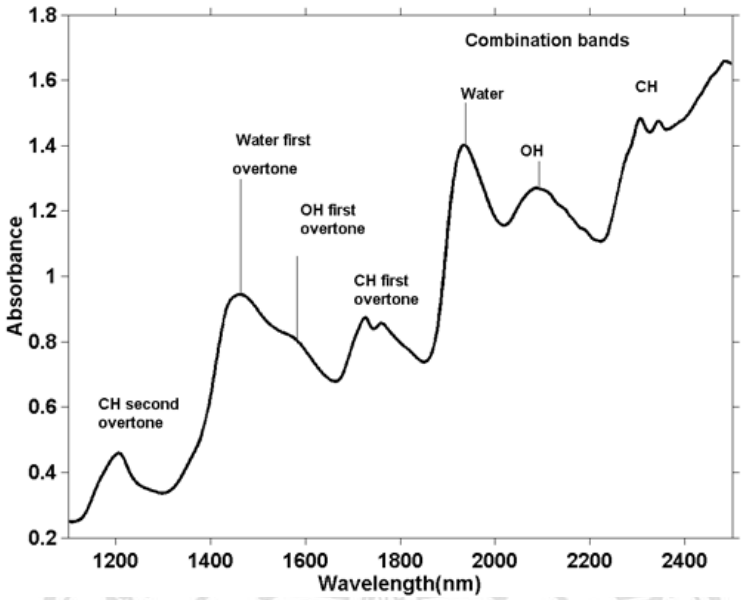


Figure 2.4 Spectrum of biscuit dough with the main areas of absorption identified [35]

Since the NIR spectra are sophisticated, seems to be uninteresting, thus there are not many researchers utilize that until in 1940's. Several researches about NIR spectra were published in scientific journals [36]. Karl Norris was called a 'Father of NIR spectroscopy', succeeded to employ NIR spectroscopy in quality assessments of agricultural products [38]. He applied statistical methods to develop calibrations using NIR data. His unique idea and the development of computers in the 1960 to 1970 opened the way for practical applications of NIR spectroscopy. Nowadays, NIR has gained wide acceptance with many fields especially in agricultural and pharmaceutical industries [39].

2.3.2 NIR instrument

Generally, NIR instrument consists of light source, monochromator, sample holder and detector. A simple diagram of NIR instrument is displayed in Figure 2.5. The light source could be a tungsten halogen lamp since in small and rugged [39]. The detectors are normally are silicon and lead sulfide detectors [40]. The light is distributed and controlled to be in the desired wavelength range with an entrance slit and exit slit light by grating in the monochromator. The detector is used to measure the amount of light that passes through or reflect from a sample. The position of detector can be based on modes of measurement.

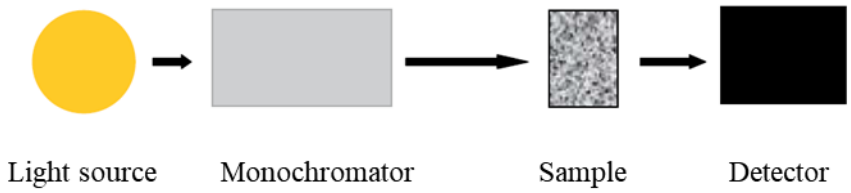


Figure 2.5 Configuration of NIR spectrometer [39]

Up to date, the NIR instrument has been considerably developed to enhance the performance of analysis and the components of NIR spectrometer could be vary based on the objective used. In Figure 2.6, the different components of NIR instrumentation was presented [41].

I.) Radiation or light source

There are two types of radiation; thermal and nonthermal light sources. The thermal source, such as tungsten-halogen and Nernst filament, can produce the thermal radiation, which span either narrow or wide range of light energy. The nonthermal type could be a more effective light source but usually provides much narrower bands of radiation. This source could consist of discharge lamps, light-emitting diodes (LEDs), laser diodes and lasers.

II.) Wavelength selector or monochromator

Wavelength selector could be used to classify the type of NIR spectrophotometers; discontinuous wavelength or continuous spectrum. The former one, a discrete wavelength, irradiates a sample using filters or LEDS. On the other hand, the later, continuous spectrum, allows multiple wavelengths to be exposed to the samples. The continuous spectra could use grating, diode array, acousto-optic tunable filter (AOTF) or Fourier transform-near-infrared (FT-NIR). The diffraction grating works in the purpose for spreading out the radiation according to the wavelength. Differently, for diode array, the whole wavelength is measured and detected simultaneously by a specific detector. In addition, AOTF generates discrete wavelengths across an extended range by using radio frequency signals to change the refractive index of a crystal, usually tellurium dioxide (TeO_2). The crystal behaves as a longitudinal diffraction grating with a periodicity equal to the wavelength of sound across the material [42]. The main advantage of AOTF

instruments over grating instruments is their mechanical simplicity and has no moving parts.

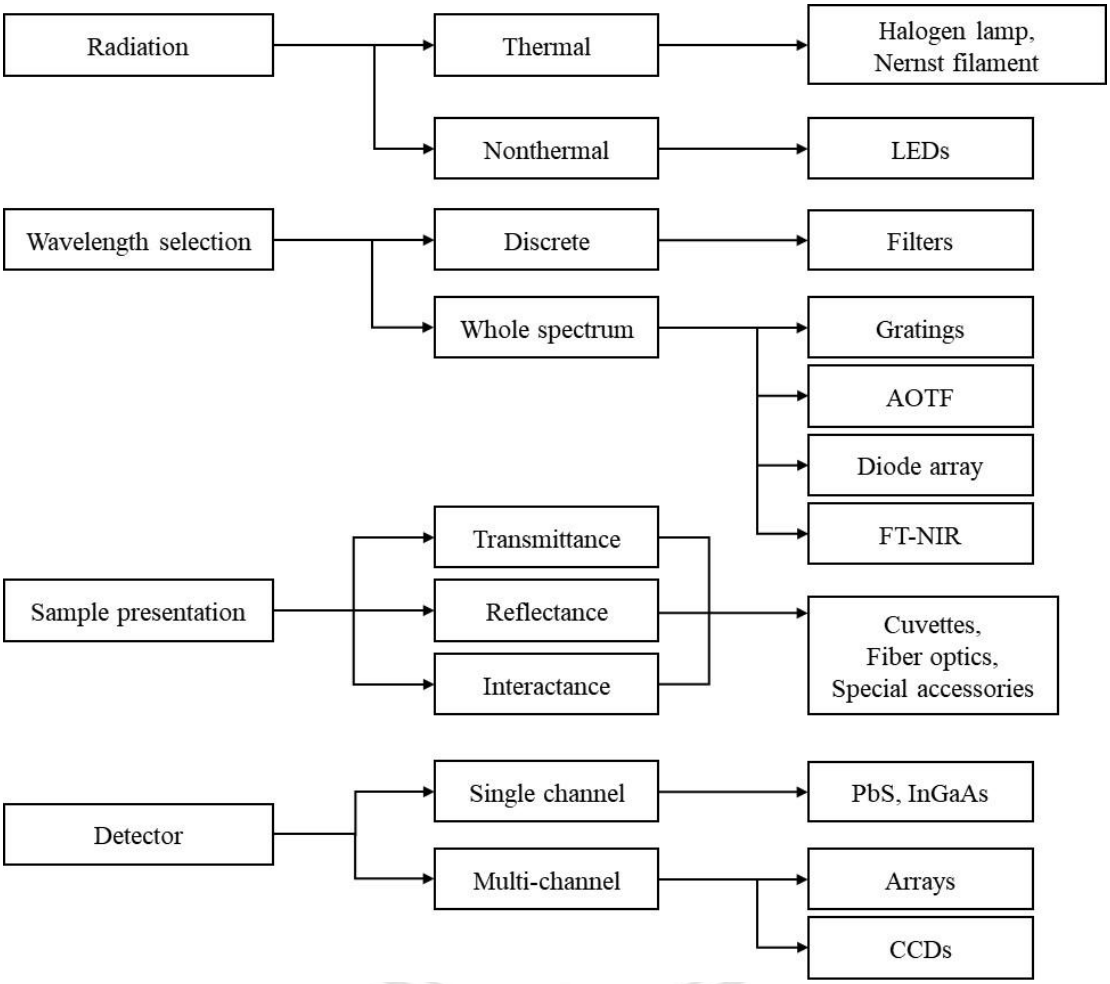
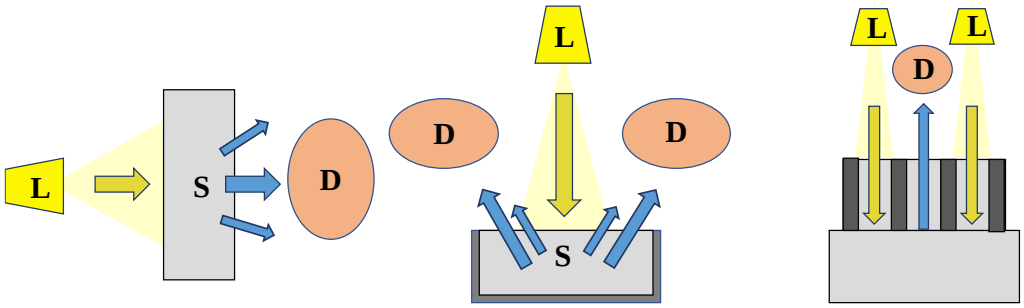


Figure 2.6 Principal features of NIR spectroscopy instrumentation [41]

III.) Sample presentation modes

The advantages of NIR are that the samples can be analyzed rapidly, non-destructively without or less sample preparation. In addition, NIR can be applied for analyzing samples with different states, shapes and thickness since it has different detection modes such as transmittance, reflectance and interactance. Figure 2.7 shows the three common detection modes of NIR.



(A) Transmittance (B) Reflectance (C) Interactance

Figure 2.7 Modes for the acquisition of the NIR spectra. L = light source, S = sample, D = detector [43]

For the transmittance mode in Figure 2.7 (A), light incident directly to the sample and exits the sample at the point opposite to the light source into the detector. The concentration of the absorbing component and the sample path-length can be described by the Beer-Lambert law. The sample that utilizes this detection mode usually allows the light to propagate through it such as water, oil, wine and fuel [42]. For the reflectance mode, the light source and detector are set on the same side of sample shown in Figure 2.7(B). The detector is placed at an angle of 45 degree to the sample for avoiding the surface reflection. This mode is normally used with the solid and granular samples [42, 44]. The main problem of this mode is the NIR spectra could be disturbed by light scattering. Therefore, the spectra could be preprocessed using multiplicative scatter correction (MSC) or standard normal variate (SNV) [32, 45] before analysis. In Figure 2.7(C), interactance mode is a combination between transmittance and reflectance modes applied for solid sample. In this case, the sample is illuminated using fiber-optic cable in direct contact with the sample surface. The incident light passes through the sample and interacts with the interior with reflection, transmission and absorption. Fiber bundles are placed on the same side as the source subsequently return the light that propagates inside the sample to the detector. The direct contact between the fiber bundles and the sample removes the effect of surface reflection and increases the penetration depth.

IV.) Detectors

The detector which universally use in NIR analysis are single and multichannel detectors. The single detector consists of silicon detector in a region of 400-1100, indium gallium arsenide (InGaAs) which is used for the detection in a region of 800-1700 nm and lead sulfide (PbS) could cover the range of 1100-2500 [46-47]. For multichannel detectors, diode arrays and charged-coupled devices (CCDs) which several detection elements are used. Therefore, the multichannel detectors can record many wavelengths at once.

It is well known that, although the NIR analysis can be easily and rapidly measured, the NIR spectra normally broad and overlapped. Therefore, its spectra cannot be interpreted in a straightforward manner. It is the reason that, chemometrics is widely utilized and applied with NIR spectroscopy to gain the worth information.

2.4 Chemometric analysis

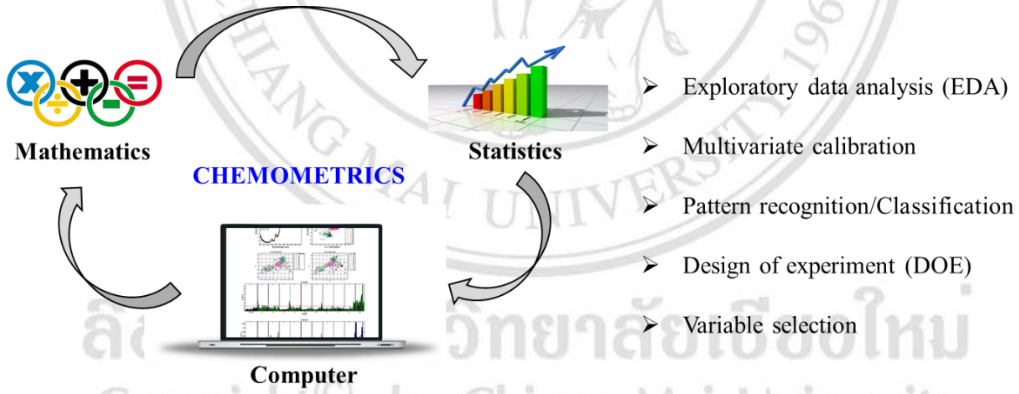


Figure 2.8 Introduction to chemometrics and five chemometric techniques

Chemometrics combines the knowledge of mathematics, statistics and computer for analysis the complicated data in several fields such as chemistry [48], biology [49], pharmaceutical [50] and agricultural products [51]. This technique can improve the understanding of chemical information and to correlate quality parameters to analytical instrument data. Chemometrics can be classified into several groups (shown in Figure 2.8) according to their applications are exploratory data analysis (EDA) [52], multivariate calibrations [53], multivariate classifications/pattern recognition [54], design of

experiment (DOE) [55] and variable selection [56]. In this research, a method that is normally used to explore the data, principal component analysis (PCA) is used to investigate all data.

2.4.1. Data preprocessing

Data preprocessing is to use a mathematical modification to the values of raw data before performing the data analysis methods. This is to ensure that the aspects wanted from the data analysis or to improve the interpretation of the data and the right trends are being studied. In addition, the analysis methods do not get confused by non-essential information such as noise or outlier. For NIR data analysis, mean centring, normalization and standard normal variate (SNV) are among the most common preprocessing methods.

Supposing a raw data matrix $\boldsymbol{X}$ is available, x_{ij} is the i^{th} row and the j^{th} element of the data matrix and z_{ij} is the i^{th} row and the j^{th} element of the preprocessed data matrix $\boldsymbol{Z}$. The following transformations can then be applied.

I) Mean centring

Mean centring can be calculated by subtracting the mean of each variable.

$$^{mc}z_{ij} = x_{ij} - \bar{x}_j ;$$

$$\bar{x}_j = \sum_{i=I}^{I} x_{ij} / I$$

where $\bar{x}_j$ is the mean for variable j calculated over all I samples. After performing mean centring, the sum of each column is zero. This has the effect of focusing the analysis upon the deviations from the mean and therefore investigating variance from the data mean rather than the absolute values.

II) Normalization

Normalization can be done by dividing each element of a row vector by a constant. In this thesis, a sample is scaled to a constant total which means that the sum over all variables after the normalization is equal to 1.

$$^{rs}z_{ij} = \frac{x_{ij}}{\sum_{j=1}^{J} x_{ij}}$$

Normalization can be useful if it is difficult to control the absolute concentration of samples precisely. This could be due to the problem from the injection volume in the chromatography where the precise amount of sample may be not easy to control but the relative proportion of each chemical can be measured. After normalization, each variable now contains information about the relative proportion of that compound in the mixture, and a comparison between samples reveals information about how the ratios between compounds differ.

III) Standard normal variate

Standard normal variate (SNV) pretreatment is used quite often in near-infrared to remove the scatter. It is applied to every spectrum individually. The average and standard deviation of all the data points for that spectrum is calculated. The average value is substracted from the absorbance for every data point and the result is divided by the standard deviation. The SNV processed data can be calculated by following equation.

$$^{SNV}z_{ij} = (x_{ij} - \bar{x}) \sqrt{\frac{\sum_{j=1}^{J} (x_j - \bar{x})^2}{J-1}}$$

2.4.2 Principal component analysis

Principal component analysis (PCA) is among the most common EDA methods which can be used for representing the variation in data by establishing some new variables, called principal components (PCs). The way to compute the PCs is illustrated in Figure 2.9. The first PC is, generated to contain the most of data, called PC1. The distance value from origin point to each sample on PC1 is called variation. The next PC is, generated in the direction that is an orthogonal with PC1, called PC2. It is expected that most of the major variation in the original data could be visualized using only the first few numbers of PCs [57]. Using PCA, a pattern in multivariate data can be assessed. Some basic questions may be answered by PCA such as "which samples behave dissimilarly?", "which samples can be identified into a similar cluster?" and "if there are

any outliers in the dataset?". At this moment, several algorithms can be used for PCA modelling such as singular value decomposition (SVD) [58] and non-linear iterative partial least squares (NIPALS) [59]. However, the results obtained should be the same. In this thesis, NIPALS algorithm will be used since it could be considered a less computationally intensive algorithm.

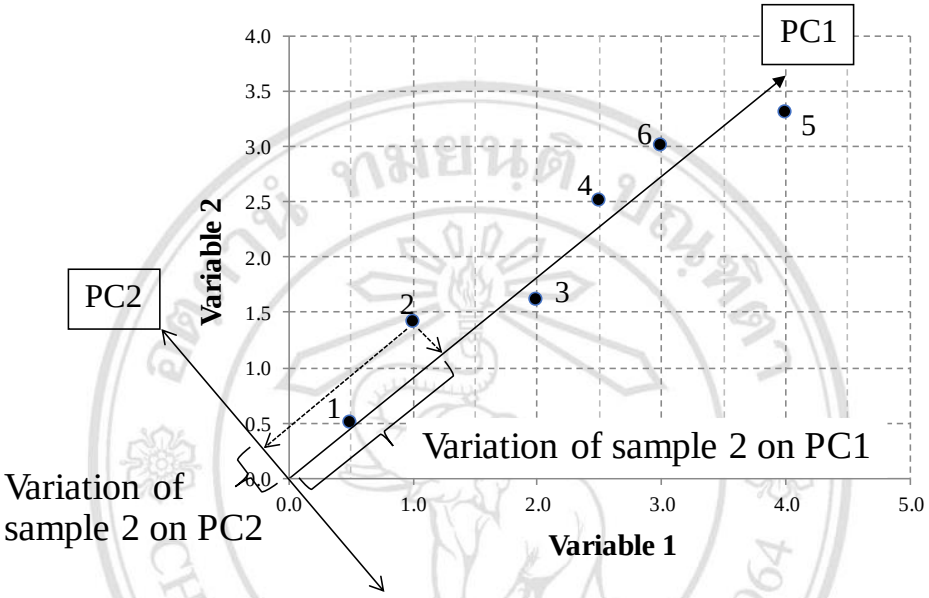


Figure 2.9 Computation of PC1 and PC2 in PCA algorithm

I) Scores and loadings

PCA captures a structure of dataset by extracting the variation into two parts: systematic variation containing certain properties that can be related to chemical factors in a data matrix and non-systematic variation. The mathematical transformation of a data matrix ***X*** by PCA is:

$$\boldsymbol{X} = \boldsymbol{TP} + \boldsymbol{E}$$

where ***X*** is a data matrix, ***T*** is a scores matrix, ***P*** is a loadings matrix, and ***E*** is an error or a residual matrix. The size of each data is illustrated in Figure 2.10.

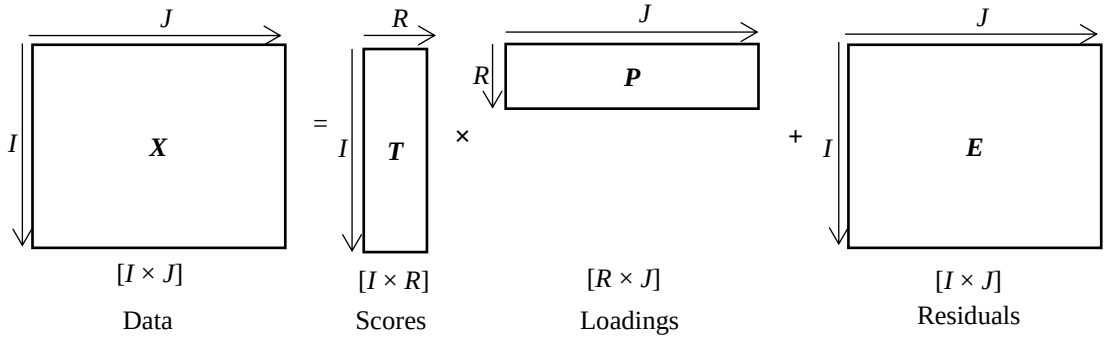


Figure 2.10 Diagram of PCA algorithm [19]

where $\boldsymbol{X}$ is the data matrix with the size of [$I \times J$], $\boldsymbol{T}$ is a score having a dimension of [$I \times R$], $\boldsymbol{P}$ is loadings with size of [$R \times J$] and $\boldsymbol{E}$ is a residual that should have the same size of $\boldsymbol{X}$. In this diagram, the number of samples, parameters and PCs are represented in I, J and R, respectively. The PC model is a product matrix of $\boldsymbol{TP}$ which in this case is considered systematic part of the variation. In particular, the relationship between each sample can be reviewed using the scores and the relationship between each variable can be reviewed using the loadings. The error matrix contains non-systematic variation or the variation that cannot be modelled by the PC model and has the same dimension as the original data matrix. The NIPALS algorithm for calculating the scores and loading matrices is as follows:

- An initial start of the score $\hat{\boldsymbol{t}}$ is generated. In this thesis, this initial vector was taken as the column of the data matrix $\boldsymbol{X}$ with the largest sum of squares.
- A loading vector is obtained using the estimated score:

$$\hat{\boldsymbol{p}} = \frac{\hat{\boldsymbol{t}}'\boldsymbol{X}}{\hat{\boldsymbol{t}}'\hat{\boldsymbol{t}}}$$

- The estimated loading is then normalized:

$$\hat{\boldsymbol{p}} = \frac{\hat{\boldsymbol{p}}}{\|\hat{\boldsymbol{p}}\|}$$

- A new estimate of the score is made using the normalized loading:

$$^{new}\hat{\boldsymbol{t}} = \boldsymbol{X}\hat{\boldsymbol{p}}'$$

- If the difference between the new score and the prior score is small.

For example;

$$({}^{new}\hat{\boldsymbol{t}} - \hat{\boldsymbol{t}})'({}^{new}\hat{\boldsymbol{t}} - \hat{\boldsymbol{t}}) < 0.001)$$

then the algorithm has converged and the scores $\boldsymbol{t}_r$ and loading $\boldsymbol{p}_r$ for component r are set to $\hat{\boldsymbol{t}}$ and $\hat{\boldsymbol{p}}$, respectively. If not, the initial score is replaced by the new score.

- The calculated component is then subtracted from the data matrix to give the residual matrix $\boldsymbol{E}$.

$$\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{t}_r \boldsymbol{p}_r$$

- The matrix of residual $\boldsymbol{E}$ is used as a new data matrix $\boldsymbol{X}$ and the algorithm repeated from step 1 until R components have been calculated where are in the desirable number of PCs.

In general, the sizes of scores and loadings vectors can be as large as desired, provided that the common dimension is not larger than the smallest dimension of the original data matrix. However, normally, most of the information or variation of the original data can be represented using the first few numbers of PCs e.g. PCs 1-3. The rests contain only unnecessary information or noise. The critical issue is that the optimum number of PCs used in the data modelling should be carefully identified and it can be chosen by a variety of approaches such as bootstrap [60] or cross-validation [61]. If a data matrix $\boldsymbol{X}$ consists of I samples with J measurements (or variables) and the optimum number of PCs is denoted by R, then the dimensions of $\boldsymbol{T}$ will be $I \times R$ and the dimensions of $\boldsymbol{P}$ will be $R \times J$ as seen in Figure 2.8.

As described, each scores matrix consists of a series of column vectors $\boldsymbol{t}_r$, and each loading matrix a series of row vector $\boldsymbol{p}_r$ where r is the number of PC (1, 2, 3,..., R). Therefore, using R PCs, it is possible to establish a model for each element x_{ij} of the original data matrix $\boldsymbol{X}$ by:

$$x_{ij} = \sum_{r=1}^{R} t_{ir} p_{rj} + e_{ij}$$

where t_{ir} and p_{rj} are the score of sample i and the loading of variable j for principal component r, respectively. e_{ij} is the residual error of the PCA model using R PCs.

II) Ranks and eigenvalues

After performing PCA, it is possible to measure the size of each PC. This is often called as an eigenvalue [62]. The earlier (and more significant) the components, the larger size of their eigenvalues will be. The eigenvalue of each calculated PC can be defined as the sum of squares of the scores as follows:

$$g_r = \sum_{i=1}^{I} t_{ir}^2$$

where g_r is the eigenvalue of the r^{th} PC. In the other words, the sum of all non-zero eigenvalues for a data matrix $\boldsymbol{X}$ equals the sum of squares of the entire data matrix:

$$\sum_{r=1}^{R} g_r = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}^2$$

where R is always equal or smaller than I or J. It is important to note here that if the data are preprocessed prior to PCA, the same preprocessing must be performed for $\boldsymbol{X}$. It is also possible to present the eigenvalues as percentages:

$$V_r = 100 \frac{g_r}{\sum_{i=1}^{I} \sum_{j=1}^{J} x^2{}_{ij}}$$

where V_r is the percentage eigenvalue for the component number r. Theoretically, the successive eigenvalues should correspond to smaller percentages of eigenvalue. The cumulative percentages of the eigenvalues can be used for indicating if the data has been well modelled by the PCA model. If it is closer to 100%, the more faithful is the model. Or, most of the variation will be included in the prediction. However, it is not necessary that as many as PCs should be used for the modelling in order to reach closer to the 100% of the cumulative percentage eigenvalue because using too many PCs

may lead to an overfitted model as noise or unwanted variation will be included in the model.

III) Conjoint and disjoint PC models

PCA is not only useful for the data exploration but it is also often performed prior to data analysis. The reason could be that there are often more variables than samples and many methods such as those based on the Mahalanobis distance [63] require that there to be a smaller number of variables than samples. In other cases, it is to avoid redundancy in original measurements and noise. It is possible to perform PCA either on the overall dataset (a conjoint model), or on each individual class (disjoint models) [64].

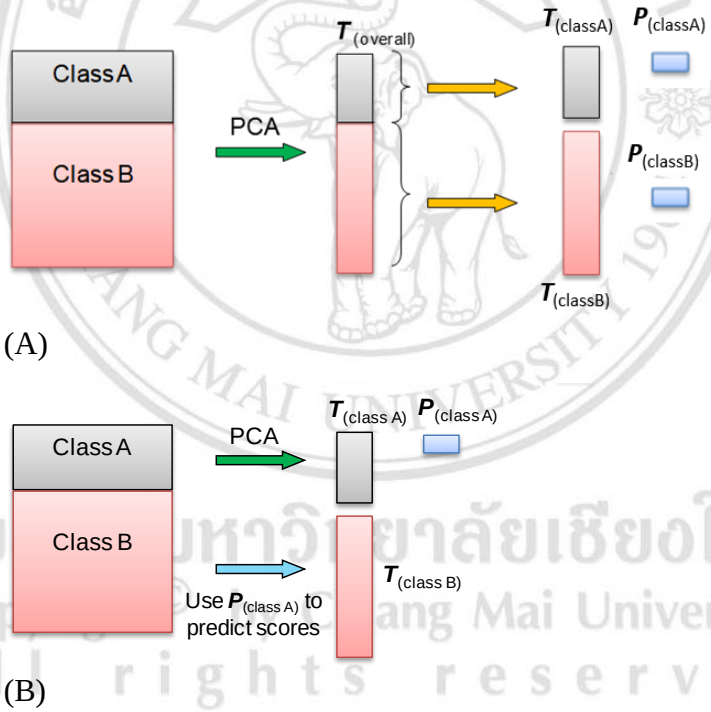


Figure 2.11 Diagram showing (A) conjoint PC model and (B) disjoint PC model. In this case, two classes, A and B are decomposed. For the disjoint PC model, the PCA was modelled using class A [21]

A scheme for a conjoint PC model can be seen in Figure 2.11(A). For a conjoint PC model, the same PC space is used to represent all samples in all classes. This means that all of the samples are participating for the modeling. The calculation could be

less simple but this could be a disadvantage because all of the samples need to be included for the modeling. Consequently, the model will have to be regenerated every time if samples are removed or new samples are introduced. Moreover, the model could not deal well with outliers since these samples belong to none of the predefined groups. In fact, a conjoint PC model is only preferable for EDA or data visualization. It only visualizes if the dataset is complicated and there is a useful trend in the modeled dataset.

An alternative to handle the situation is to use disjoint PC models. In this case, PCA is performed separately on each class. A scheme for a disjoint PC model can be seen in Figure 2.9(B). The loadings of a class modeling (class A) are used as the estimator of the scores for test samples (class B). This means that the model does not have to be reformed if new samples or groups are introduced. Using disjoint PC models, there is only one class needed to define a model and this could be ideal for process monitoring where there is no information from future samples available and the only the information obtained is the predefined normal operating condition (NOC) samples. This idea is often called one class classification which will be discussed in detail in section 2.4.3. In this thesis, for the methods where PCA is performed prior to the analysis in the prediction mode, only disjoint PCA will be used.

2.4.3 One class classification

The basic concept of classification is to identify the class membership of an unknown sample. Classification methods can be generally categorized into two groups, one class and multiclass classification that mentioned in chapter 1. One class classification, a supervised soft modeling method, aims to generate the classification model when the unknown sample could not be identified. Using one class classification, the class of samples are known beforehand. The model can be established separately in each group of samples, therefore, the unknow sample could be classified more than one group. There are various methods use in this thesis for one class classification such as Hotelling's T^2, D and Q statistics and soft independent modeling of class analogy (SIMCA).

I) Hotelling's T^2 and D statistics

The basic idea of these methods is to determine the Mahalanobis distance of a sample to the centre of the model. In process monitoring, the model can be constructed using the defined NOC samples. This distance can be converted to a confidence limit or probability, assuming that the underlying data is distributed normally. The greater this distance, the lower the chance of a sample belonging to the corresponding class. This is similar to quadratic discriminate analysis (QDA) [65-66] but using a cut-off threshold instead of the boundary between the groups of samples

When raw data is used (no PCA prior the analysis), the method is usually called Hotelling's T^2 statistic. In case of comparing a sample $\boldsymbol{x}_i$ to a group of I samples, the Hotelling's T^2 distance can be given by:

$$T_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$$

where $\bar{x}$ is a vector containing the mean value corresponding to each variable. $\boldsymbol{S}$ is a diagonal matrix corresponding to the standard deviation of each variable. The Hotelling's T^2 statistic or the confidence limit α for the distribution can be calculated by:

$$T^2 \approx \frac{J(I-1)(I+1)}{I(I-J)} F(J, I-J, \alpha)$$

where J is the number of variables being measured. It is assumed that the statistic follows an F-distribution and so $F(J, I\text{-}J, \alpha)$ represents the F-distribution with J and I-J degrees of freedom at confidence level α.

However, the Mahalanobis distance can only be computed if the number of variables is less than the number of samples. If this is not so, it is necessary to reduce the dimensionality of the data. One possibility is to use PCA. An extension of Hotelling's T^2 for use after PCA is called D statistic. Using the loadings matrix $\boldsymbol{P}$ obtained from the NOC samples decomposed by PCA, a predicted scores vector $\hat{\boldsymbol{t}}_i$ can be calculated for sample $\boldsymbol{x}_i$:

$$\boldsymbol{x}_i = \hat{\boldsymbol{t}}_i\boldsymbol{P} + \boldsymbol{e}_i \text{ and so } \hat{\boldsymbol{t}}_i = \boldsymbol{x}_i\boldsymbol{P}'$$

Provided that PCA was performed on centred data within the NOC region, the values of the mean variables for the NOC region must be subtracted from the data to give $\boldsymbol{x}_i$. The D statistic can be obtained as follows:

$$d_i = \hat{\boldsymbol{t}}_i \left(\frac{\boldsymbol{T}'\boldsymbol{T}}{I-1} \right)^{-1} \hat{\boldsymbol{t}}_i' \cong \frac{R(I-1)(I+1)}{I(I-R)} F(R, I-R, \alpha)$$

A confidence level α can be calculated using the F-distribution with R and I-R degrees of freedom at confidence limit α. Note that there are some other equations to calculate this statistic which are described elsewhere [67-68] but it is not expected that there will be much difference when different approaches are used. The bar chart for demonstration of D value is shown in Figure 2.12. If a value of d exceeds a limit given by α, this indicates that the corresponding sample does not fall within the class boundaries at a given confidence limit, and it will be identified as an out-of-control sample. Typically, 95% or 99% confidence limits are computed but this can be adjusted based on the process engineer's experience.

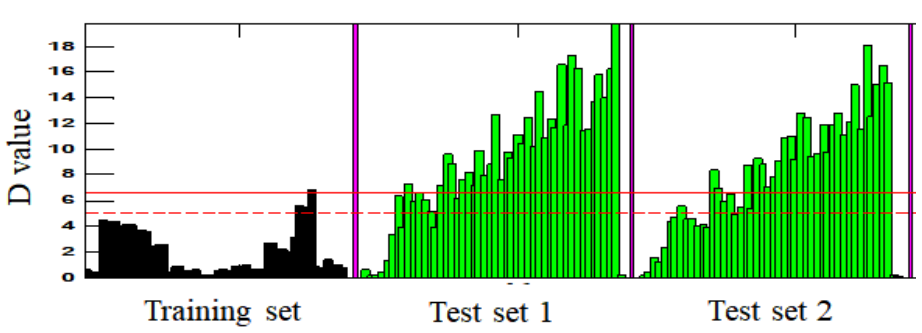


Figure 2.12 Bar chart of D value of training set and two test set data where the dashed and solid red lines represented 95% and 99% confidence limits

It is possible to look at the contribution of each variable (either a chromatographic peak or an elution region) to the D statistic. The contribution of the j^{th} variable in sample $\boldsymbol{x}_i$ to the D statistic can be obtained as follows:

$$\varpi_{ij}^{d} = \hat{\boldsymbol{t}}_i \left(\frac{\boldsymbol{T}'\boldsymbol{T}}{I-1} \right)^{-1} \left[x_{ij} \boldsymbol{P}_j \right]$$

The contribution of each variable to the D statistic as a function of process time can be viewed when the values of the contribution against sample numbers are plotted. In addition, the contribution of each variable to the D statistic for each sample can be viewed when the contribution of each variable in a specific sample is plotted. These plots are commonly called contribution plots. The D statistic can be illustrated together with their contribution plots to help investigate which variable is most responsible for the process deviating from the NOC.

II) Q statistics

Instead of considering the variation on the PC space, it is also possible to look at the variation which is not modelled by PCA or the residuals. This is to determine how well a sample fits into the PC model and it can be calculated by computing the Euclidean distance of its residuals to the PC model. This method is called Q statistic or squared prediction error (SPE) and sometimes it is embedded in other methods such as soft independent modeling of class analogy (SIMCA) [69]. Using a PC model of the NOC samples, the Q statistic can be calculated as follows:

$$q_i = \sum_{j=1}^{J} (e_{ij})^2$$

where e_{ij} is the error for reconstruction of the j^{th} variable for sample $\boldsymbol{x}_i$ given by:

$$e_{ij} = \boldsymbol{x}_{ij} - \boldsymbol{x}_i \boldsymbol{P}_j' \boldsymbol{P}_j$$

The confidence level for the Q statistic used in this thesis can be calculated by:

$$q_{\lim,\alpha} = \theta_1 \left[1 - \theta_2 h_0 \left(\frac{1-h_0}{\theta_1^2} \right) + \frac{z_\alpha \sqrt{(2\theta_2 h_0^2)}}{\theta_1} \right]^{1/h_0}$$

where $\boldsymbol{V}$ is the covariance matrix of the residual matrix $\boldsymbol{E}$ (after performing PCA on the NOC region). θ_1 is the trace of $\boldsymbol{V}$, θ_2 is the trace of $\boldsymbol{V^2}$, θ_3 is the trace of $\boldsymbol{V^3}$, $h_0 = 1 - (2\theta_1\theta_3 / 3\theta_2^2)$ and z_α is the standardized normal variate with $(1-\alpha)$

confidence limit. An alternative equation was proposed by Box [70] in the 1950s where the confidence level is given by a weighted chi-squared distribution $g\chi^2$ of ν degrees of freedom where $g = \theta_2/\theta_1$ and $\nu = \theta_1^2/\theta_2$. However, it can be shown that these methods provide very similar results [71]. The bar chart which demonstrates in-group and out-group of classification model considering with Q statistic is shown in Figure 2.13. Same as the D value, if the sample that has Q value higher than that confidence limit, it is classified as an out-group sample.

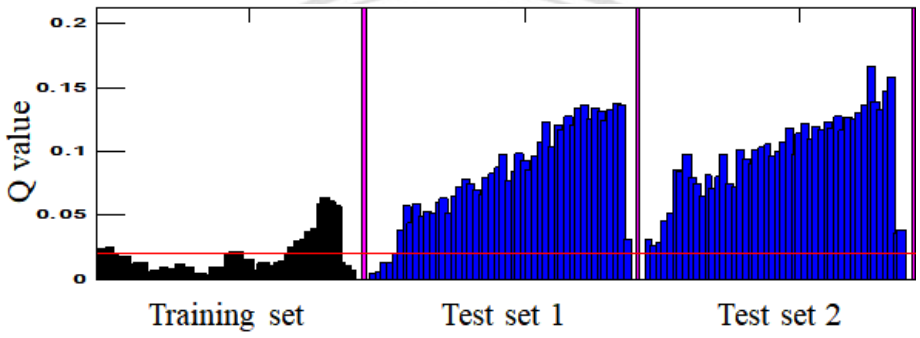


Figure 2.13 Bar chart of Q value of training set and two test set data where a horizontal red line represents 95% confidence limit

Similar to the D statistic, a contribution of each variable can be calculated for the Q statistic. The contribution of the j^{th} variable in sample $\boldsymbol{x}_i$ can be obtained as follows:

$$\varpi_{ij}^{q} = (e_{ij})^2$$

This Q contribution chart can be used together with the Q statistic to understand which variable or compound is most responsible for the fault in the process.

III) Soft independent modeling of class analogy

Soft independent modeling of class analogy (SIMCA) is an untargeted prediction that combines the classification decision from D and Q statistics [24, 72]. The term “soft” means that no hypothesis on the distribution of variables is made and can be flexibly. This concept makes that overlapping of classes is not problematic [73]. SIMCA has been used for the non-targeted detection of adulterants in several products [74]. This method uses both D and Q together in order to identify local models for possible groups

and to predict a possible class membership for unknown sample. At first, this approach runs a global PCA on the whole dataset in order to identify groups of observations. Local models are then estimated for each class. Finally, unknown sample are classified to one of the established class models on the basis of their best fit to the respective model.

Normally, two datasets of sample are used in the classification; training (tr) and test (ts) sets, for example, class A and class B. Figure 2.14 presents the definition of each sample in SIMCA model that is established based on the sample of class A (the sample in class A is used to generate the classification model). In the prediction of the training samples (class A) if the sample is classified as training/in-group (class A), this is counted as a true positive (TP). On the other hand, if it is classified as out-group of training (class B), this training sample is considered as a false negative (FN). Meanwhile, if a test sample (class B) is tested and classified as test/out-group (class B), it is considered as true negative (TN). Or if it is classified as in-group (class A), it is counted as false positive (FP).

Model of A		Prediction	
		A	B
Actual	A	TP	FN
	B	FP	TN

Figure 2.14 Definition of four group of samples in SIMCA

To estimate the classification performance of SIMCA, there are four statistical values that are usually used including sensitivity, specificity, precision and accuracy. The calculations of these statistic values are as follows;

- Sensitivity (SE)

$$SE = \frac{TP}{TP+FN}$$

- Sensitivity (SP)

$$SP = \frac{TN}{TN+FP}$$

- Precision (Pre)

$$Pre = \frac{TP}{TP+FP}$$

- Accuracy (Acc)

$$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}$$

The sensitivity value is used to verify whether the modelled samples are classified properly into a group. While the specificity is normally used to test the robustness of a classifier with respect to the samples from the other groups. The precision shows whether the model can classify the training sample correctly. In case of the accuracy, this value expresses the correct classification in both in-group and out-group when comparison with all predictive classification.

2.4.4 Orthogonal projection

Since the variation in an adulterant have strongly affect to the prediction, which could make a model wrongly accept more adulterated rice. Therefore, the variation from non-aromatic rice (normal white rice) should be remove from NIR spectra prior to the modeling using some data filtration methods such as orthogonal projection (OP). The concept of orthogonal projection was to deal with the interferences and scatter effects of spectroscopic data [75]. In this case, the interference is the variation of white rice in whole data which has the orthogonal relation with amount of KDML 105.

The interference in spectroscopic data are obtained from temperature, moisture, scattering and other orthogonal variation [76]. The major objective of OP is to remove an influence from the wanted data for the modelling. In general, the observed data ($x_{i,obs}$) always contains of expected data (x_i) and unwanted data (h_i) [77]:

$$x_{i,obs} = x_i + h_i$$

If a good estimation of h_i is obtainable, the first possibility would be to perform a subtraction, and so x_i is estimated as:

$$x_{i,exp} = x_{i,obs} - h_i$$

Normally, the unwanted variation is not well estimate and is not easy to subtract it from each spectrum *i* directly. Therefore, the OP is performed to get rid of the unnecessary variation from the observed data. The OP methods used in this research is independent interference reduction (IIR). The diagram for explanation of IIR algorithm is presented in Figure 2.15.

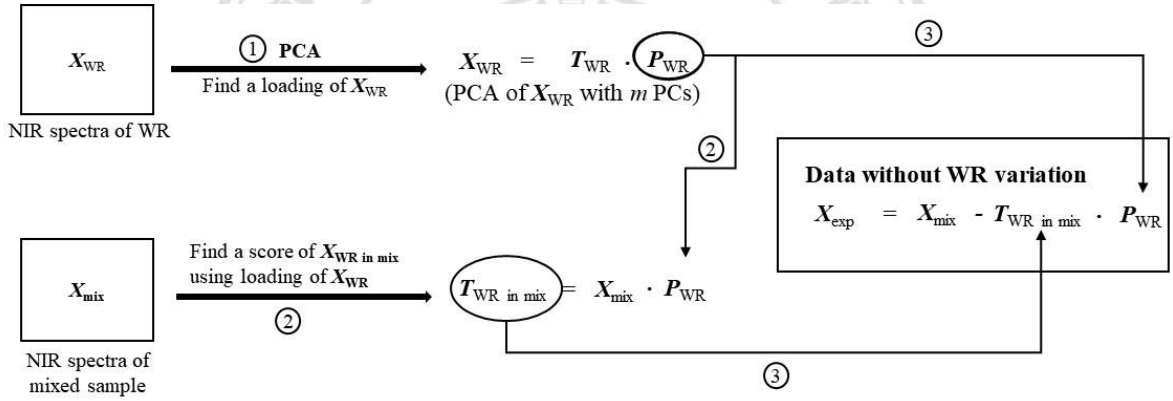


Figure 2.15 Diagram of IIR algorithm

The algorithm of IIR is similar to two methods proposed recently, orthogonal signal correction (OSC) [78] and direct orthogonalization (DO) [79]. The IIR is better than the two methods in case of the analysis is disturbed with different conditions (temperature and time) and this IIR method is not required all reference analyzed in the sample [75]. Therefore, the IIR method is suitable to determine the amount of adulterated white rice in KDML 105 rice samples. Performing the IIR, two datasets are required:

(1) The samples containing of large samples as all variation except the interest data (unwanted variation), $\boldsymbol{X}_{WR}$

(2) The observed data from all samples consist of NIR spectra ($\boldsymbol{X}_{mix}$) and the percentage of KDML 105, ($\boldsymbol{y}_{obs}$)

These two datasets are used in the OP approach to remove the orthogonal information from all data following equations.

Firstly, a data matrix $\boldsymbol{X}_{WR}$ represents all variation except the interest information is carried out using principal component analysis (PCA) to obtain two data which are score ($\boldsymbol{T}_{WR}$) and loading ($\boldsymbol{P}_{WR}$) with *m* principal components (PCs).

$$\boldsymbol{X}_{WR} = \boldsymbol{T}_{WR} \cdot \boldsymbol{P}_{WR} \quad \text{(PCA of } \boldsymbol{X}_{WR} \text{ with } m \text{ PCs)}$$

After getting $\boldsymbol{P}_{WR}$, a score of observed data in $\boldsymbol{X}_G$ ($\boldsymbol{T}_{WR\ in\ mix}$) is obtained by modelling $\boldsymbol{X}_{mix}$ on $\boldsymbol{P}_{WR}$.

$$\boldsymbol{T}_{WR\ in\ mix} = \boldsymbol{X}_{mix} \cdot \boldsymbol{P}_{WR}$$

The $\boldsymbol{T}_{WR\ in\ mix} \cdot \boldsymbol{P}_{WR}$ represents an interference part that is useless to generate the calibration model. Therefore, the interference part should be subtracted from the original data for gaining the expected information ($\boldsymbol{X}_{exp}$).

$$\boldsymbol{X}_{exp} = \boldsymbol{X}_{mix} - (\boldsymbol{T}_{WR\ in\ mix} \cdot \boldsymbol{P}_{WR})$$

From all steps, it can notice that the appropriate number of *m* PCs should be concerned if the suitable information is collected to get rid of the orthogonal data. After that, the data without orthogonal variation ($\boldsymbol{X}_{exp}$) would be employed with the response ($\boldsymbol{y}_{obs}$) to establish the calibration model using partial least squares (PLS) regression.

2.4.5 Calibration transfers

The calibration transfer (CT) is used in this research for enhancing the analysis and for reducing the error of baseline and wavelength drifts [80] according to the differences condition during spectroscopic measurement such as temperature, time and different instrument [81]. The CT used in this task is piecewise direct standardization (PDS) that regularly provides more superior results than other CT methods [80]. The PDS is extended from direct standardization thus the mathematics of PDS are derived based on the general DS methods [74, 78].

DS algorithm:

In CT, there are the main two datasets of spectra measured on different instruments of different times. The first dataset called master data and the second one called slave data. Assuming the two datasets are from different instrument, the spectra measured on the slave instrument are corrected to match the spectra measured on the master instrument while the calibration model remains unchanged. This correlation is described by the transformation matrix ***F***. The algorithm for establishing the transformation metric is shown in Figure 2.16.

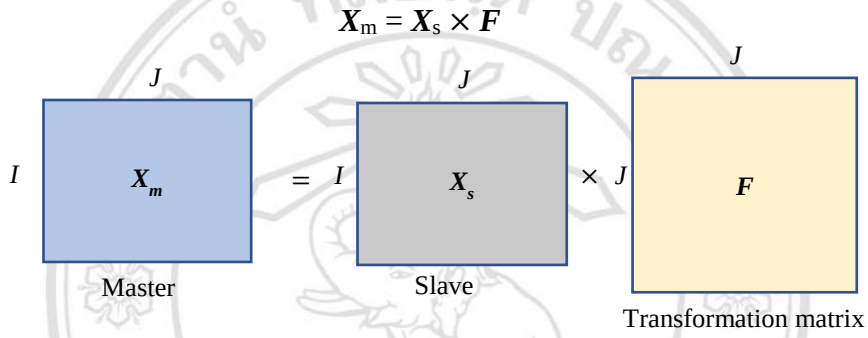


Figure 2.16 Data matrix of DS algorithm [80]

where $\boldsymbol{X}_m$ and $\boldsymbol{X}_s$ are the spectra measured on the master instrument and the slave instrument, respectively. Both data matrix represented in dimension of sample (*I*) by wavelength (*J*) and ***F*** is a transformation matrix (dimensioned *J* by *J*) calculated as follow:

$$\boldsymbol{F} = (\boldsymbol{X}_s^T \boldsymbol{X}_s)^{-1} \cdot \boldsymbol{X}_s^T \boldsymbol{X}_m$$

where the superscript "T" denotes transpose matrix and "$(\boldsymbol{X}_s^T \boldsymbol{X}_s)^{-1}$" is the matrix pseudoinverse of $\boldsymbol{X}_s$

However, the calculation of the transfer matrix ***F*** in PDS is different form DS in that it is not required the whole spectrum of all transfer samples to correct each wavelength on the slave instrument.

PDS algorithm:

In spectroscopic data, each spectral point master would be related to the neighbor spectral data than the full spectra of slave due to the limitation of spectra variations [76]. Therefore, to improve the accuracy of calibration transfer, the PDS could be used for regenerating each spectral point of master ($\boldsymbol{x}_{m,j}$) to relate with the small spectral subset of slave ($\boldsymbol{X}_{s,j}$) at nearby wavelengths. The PDS algorithm are as following steps:

Step 1: Selecting the spectral point of master ($\boldsymbol{x}_{m,j}$) at wavelength *j*

Step 2: Defining the subset spectra of slave $\boldsymbol{X}_{s,j}$ nearby wavelength *j* form index *j*- *k* to *j* + *k*

$$\boldsymbol{X}_{s,j} = [\boldsymbol{x}_{s,j-k}, \boldsymbol{x}_{s,j-k+1}, \cdots, \boldsymbol{x}_{s,j+k-1}, \boldsymbol{x}_{s,j+k}]$$

Step 3: Establishing the regression coefficient

$$\boldsymbol{x}_{m,j} = \boldsymbol{X}_{s,j} \times \boldsymbol{b}_j$$

Each $\boldsymbol{b}_i$ can be calculated using any multivariate regression such as principal component regression (PCR) and partial least squares (PLS).

Step 4: Generating the transformation matrix $\boldsymbol{F}$ by arrangement the $\boldsymbol{b}_j$ to a banded diagonal matrix

$$\boldsymbol{F} = \text{diag}(\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, \cdots, \boldsymbol{b}_j^T, \cdots, \boldsymbol{b}_n^T)$$

where *n* is the number of spectra channels included. The variations in terms of off-orthogonal elements would be set to zero. The diagram of step 1 - 4 represents in Figure 2.17.

Step 5: Standardization the spectra of unknown sample $\boldsymbol{x}_{s,un}$ using $\boldsymbol{F}$ to obtain the modified spectrum $\boldsymbol{x}_{s,std}$

$$\boldsymbol{x}_{s,std} = \boldsymbol{F} \times \boldsymbol{x}_{s,un}$$

Then the $\boldsymbol{x}_{s,std}$ could be tested with the calibration model established from the master samples to gain a desirable predictive result.

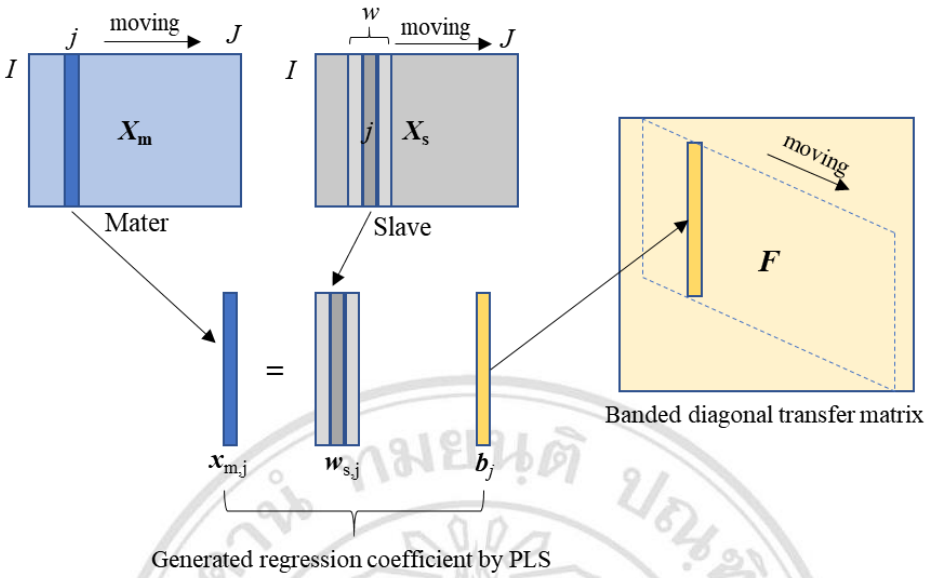


Figure 2.17 Computation of the transfer matrix in PDS algorithm [82]

In this research, piecewise direct standardization (PDS) was used. PDS is an extension algorithm of a simple direct standardization (DS) and it regularly provides more superior results [76]. A method for generating the calibration model, in this task, is partial least squares (PLS) regression.

2.4.6 Partial least squares regression

The multivariate calibration methods such as partial least squares (PLS) regression could is widely used supervised linear modelling technique in chemometrics [83]. Moreover, this technique can form a linear relationship between data matrix (***X***) and response vector (***c***) establishing a calibration model to predict the response of test sample [84]. The PLS algorithm used to estimate the adulteration in this research was modified from the ordinary PLS process following [85]. The equations of PLS algorithm are as follow:

$$\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E}$$

$$\boldsymbol{y} = \boldsymbol{Tq}' + \boldsymbol{F}$$

Using the non-linear iterative partial least squares (NIPALS) algorithm, ***X*** is decomposed into ***X***-scores (***T***) and ***X***-loadings (***P***). As the same time, ***c*** is the product

approximation of $\boldsymbol{T}$ and $\boldsymbol{c}$-loadings ($\boldsymbol{q}$). The aim of the PLS algorithm is to minimize the norm of $\boldsymbol{F}$. To predict the response of unknown sample $\boldsymbol{X}_{\text{test}}$, the following equation is used:

$$\hat{\boldsymbol{c}}_{test} = \boldsymbol{X}_{\boldsymbol{test}}\boldsymbol{W}\boldsymbol{q}'$$

where $\boldsymbol{W}$ refers to normalized PLS weights. PLS, in most cases, could satisfactorily provide predictive results if the data follows a multivariate normal distribution. In addition to the predictive model, the importance of the predictive parameters corresponding to the studied response can be evaluated using partial least squares-variable influence on projection (PLS-VIP) [83] The PLS-VIP values represent the influence of $\boldsymbol{X}$ variables on $\boldsymbol{c}$ variance and can be calculated from the PLS weights obtained from the PLS model [86]. The VIP score greater than one is typically used as a criterion for variable selection [87].

2.4.7 Model statistics (quality of calibration model)

After the calibration model is performed, the accuracy of prediction should be estimated. The most popular model statistics are root mean square error of calibration (RMSEC), root mean square error of prediction (RMSEP) and coefficient of determination (R^2 and Q^2) were used in this research.

I) Root mean square errors of calibration (RMSEC) and prediction (RMSEP)

RMSEC and RMSEP values are commonly used to estimate the difference between two values, in this case, the observed value and predicted value [84]. Moreover, these values can be indicated a predictive ability of model calculate by this equation:

$$\text{RMSEC, RMSEP} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N}}$$

where $y_i, \hat{y}_i$ are the observed value and the predicted value, respectively and N is the number of samples.

For interpretation, high value represents high performance of model prediction. The values of RMSEC and RMSEP have the same usability. The difference between them was the response of test set. The RMSEC was the error of training set in case of autoprediction, meanwhile, the RMSEP was the error of test set. After considering the error of prediction, the value indicates a correlation between actual and predicted values could be measured in terms of coefficient of determination.

II) Coefficient of determination

The coefficient of determination, generally denoted r^2 (simple linear regression) or R^2 (multiple correlation), a number that indicates the proportion of the variance in the dependent variable (*x*) that is predictable from the independent variable (*y*). Moreover, the R^2 can represent the linear relationship between *x* and *y*. For example, if then $R^2 = 0.89$, which means that 89% of the total variation in *y* can be explained by the linear relationship between **x** and **y** in positive trend. The other 19% of the total variation in **y** remains unexplained. The R^2 can be calculated following the equation below:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \qquad 0 \le R^2 \le 1$$

where the *'RSS'*, a residual sum of squares, is a sum of the square differences between *y* response and *y* calculated by the regression model. And the '*TSS*', a total sum of squares, is the sum of the squared differences between the observed value and their average value.

For prediction, an improved validation method, in this case, leave-one-out cross validation is compatible used with R^2 to evaluate the predictive prediction power of model and denoted in Q^2 value. The Q^2 value, a cross-validated R^2, is used frequently to measure the ability of model [24] and can be calculated by the equation below:

$$Q^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_{i/i})^2}{\sum(y_i - \bar{y})^2}, \qquad 0 \le Q^2 \le 1$$

where the '*PRESS*', a predictive error sum of squares, is the sum of the square differences between observed value and the predicted value which provides from the leave-one-out cross validation technique. The y_i, $\hat{y}_{i/i}$ are observed value not used to establish the model and predicted value, respectively and $\bar{y}$ is an average of all observed values. The Q^2 value of the best prediction model should get close to 1.

CHAPTER 3

Experimental procedure

3.1 Sample preparation and NIR measurement

A total of thirty-nine rice samples was used in this research comprising of *Oryza sativa* L. ssp. *indica* cv. KDML105 (19 samples), *Oryza sativa* L. ssp. *indica* cv. Pathum Thani 1 (PT1) (4 samples) and Thai white rice (16 samples). The KDML105 and PT1 rice has been known as Thai Jasmine rice which has specific aromatic quality with characteristic textures when cooked. The rice samples were purchased from local markets in Chiang Mai, Thailand in year 2018. The rice grain was attached by GMP, HACCP and ISO 9001 so that the purity of the rice grain could be guaranteed and each of the rice samples was from the same lot number. The samples were separated into two sets, which were the pure samples and the adulterated samples. For a pure sample set, it was prepared in the weight of 250.00 g for NIR measurement. To generate the adulterated samples with different adulteration levels, three aromatic rice (KDML1 – KDML3) were, collected as a representative, mixed with three white rice (WR1 – WR3). There are nine-mixture dataset that was the combination among the three KDML samples and three WR samples. The code name of nine datasets presented in Table 3.1. For example, an A1 was the mixture between KMDL 1 and WR 1 with different adulteration levels in a total of 47 samples.

Table 3.1 Nine-mixture datasets of rice samples

Mixture between	WR 1	WR 2	WR 3
KDML 1	A1	A2	A3
KDML 2	B1	B2	B3
KDML 3	C1	C2	C3

Each of mixture dataset having different levels of 0, 10, 20, 30, 40, 50, 60, 61, 62,…and 100 %w/w of KDML105 with a total weight of 250.00 g for each sample. The mixed rice sample was thoroughly stirred for 5 min. Table 3.2 shows the ratio of rice grain samples for one set. All sample were directly used and measured without any washing, cleaning or heating treatment. The mixture of rice samples had 47 samples per set, thus there were totally 423 adulterated samples.

Table 3.2 Weight of KDML105 and WR with different %w/w of KDML105 in each mixed sample

No.	Weight (g)		% w/w of KDML105	No.	Weight (g)		% w/w of KDML105
	KDML105	WR			KDML105	WR	
1	0.00	250.00	0.0	25	195.00	55.00	78.0
2	25.00	225.00	10.0	26	197.50	52.50	79.0
3	50.00	200.00	20.0	27	200.00	50.00	80.0
4	75.00	175.00	30.0	28	202.50	47.50	81.0
5	100.00	150.00	40.0	29	205.00	45.00	82.0
6	125.00	125.00	50.0	30	207.50	42.50	83.0
7	150.00	100.00	60.0	31	210.00	40.00	84.0
8	152.50	97.50	61.0	32	212.50	37.50	85.0
9	155.00	95.00	62.0	33	215.00	35.00	86.0
10	157.50	92.50	63.0	34	217.50	32.50	87.0
11	160.00	90.00	64.0	35	220.00	30.00	88.0
12	162.50	87.50	65.0	36	222.50	27.50	89.0
13	165.00	85.00	66.0	37	225.00	25.00	90.0
14	167.50	82.50	67.0	38	227.50	22.50	91.0
15	170.00	80.00	68.0	39	230.00	20.00	92.0
16	172.50	77.50	69.0	40	232.50	17.50	93.0
17	175.00	75.00	70.0	41	235.00	15.00	94.0
18	177.50	72.50	71.0	42	237.50	12.50	95.0
19	180.00	70.00	72.0	43	240.00	10.00	96.0
20	182.50	67.50	73.0	44	242.50	7.50	97.0
21	185.00	65.00	74.0	45	245.00	5.00	98.0
22	187.50	62.50	75.0	46	247.50	2.50	99.0
23	190.00	60.00	76.0	47	250.00	0.00	100.0
24	192.50	57.50	77.0				

After sample preparation, all samples were stored in a sealed polyethylene bag at room temperature (25 ºC) for one day before NIR measurement.

3.2 NIRS analysis

3.2.1 Sample packing

Before NIR measurement, 250.00 g of sample would be contained in a rectangular container called coarse sample cell (width × length × depth, 5.7 × 29.4 × 2.0 cm) that shown in Figure 3.1. One side of the container is the glass that allows the NIR light pass through to the samples.

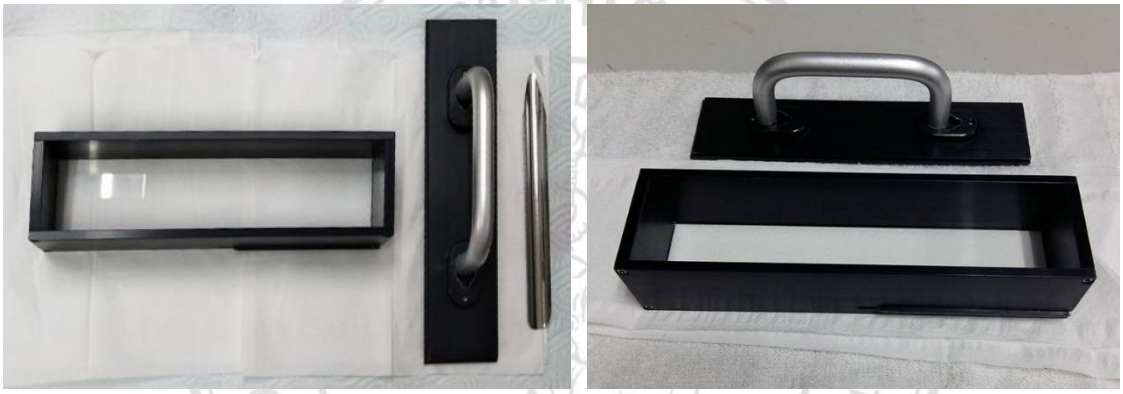
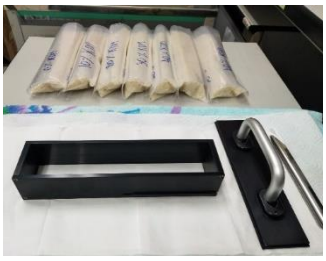


Figure 3.1 Coarse sample cell (width × length × depth, 5.7 × 29.4 × 2.0 cm)

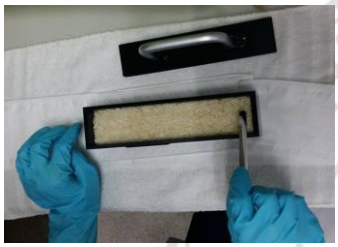
The steps of the experiment shown in Figure 3.2. Firstly, the rice grain sample was placed into a container. After that, the surface of rice grain was equally spread by a spatula to ensure that the grain should be close-packed without the flowing of rice grain. The container was then closed and compressed by the cover. The last step was checking the packing sample that it has no the grain can flow in the container.

(A) The rice grain samples in a sealed polyethylene bag



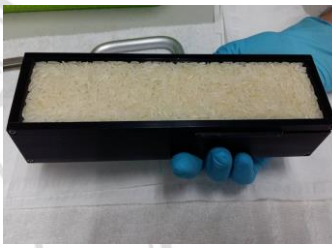
(B) Coarse sample cell used in this experiment

(C) place the sample in a container



(D) Equalize the surface of sample by a spatula

(E) Close and compress the container cover



(F) Flip over the container to check if it tightly packed

Figure 3.2 The steps (A – F) of sample packing

Figure 3.3 NIR spectrometer

3.2.2 NIR measurement

NIR spectrometer, a NIRSystem 6500 (Multi-Mode^{TM} Analyzer, Foss, USA), was used in this research for measuring the rice grain sample spectra and presented in Figure 3.3. The NIR spectrometer was kindly supported from Postharvest Technology Research Center, Faculty of Agriculture, Chiang Mai University. Before measurement, the NIR spectrometer was warmed up and calibrated for one to three hours until the reference spectrum was constant. The steps of NIR measurement were shown in Figure 3.4. The coarse sample cell contained with rice grain sample was placed into the NIR transportation module. After that, setting the position of container and checking if there was not any flowing of rice grain. The sample was measured using NIR transportation module and the reflectance spectra were recorded as an absorbance (log (1/*R*)) in wavelength region from 400 to 2500 nm with 2-nm intervals giving a total of 1050 data points per spectrum. Each spectrum is a mean of 32 scans. The sample was scanned three times and the spectra were averaged to provide a mean spectrum for each sample.

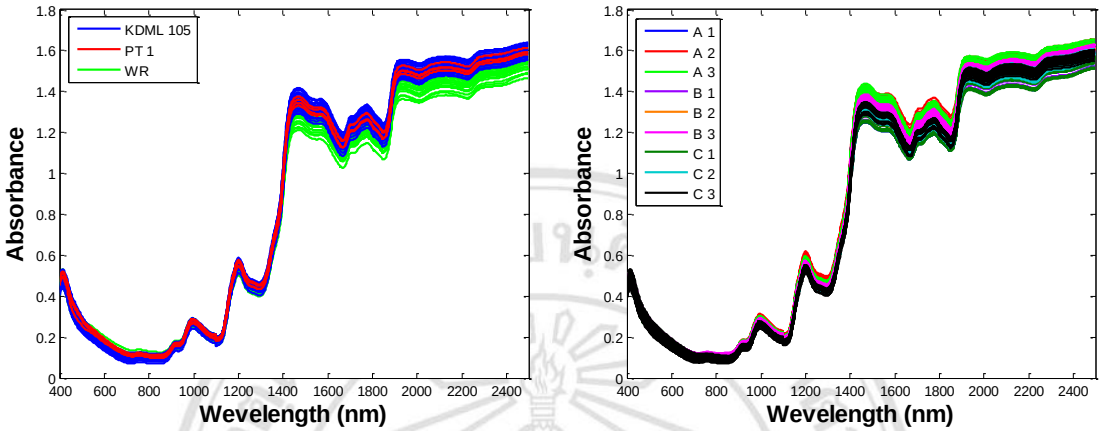
(A) Place the sample container into the instrument

(B) Set the position of container

(C) Check the position and then close the cover

Figure 3.4 The steps (A – C) of placing the coarse sample cell into NIRSystem 6500 with transportation module

NIR spectral measurement took around five minutes per one sample. Then the reflectance was converted to the absorbance spectra in wavelength range of 400 – 2500 nm appeared in Figure 3.5.



(A) (B)

Figure 3.5 NIR spectra of (A) pure rice samples and (B) mixture samples

Figure 3.5(A) presents the NIR spectra of pure rice sample including KDML105, PT1 and WR represented in blue, red and green solid lines, respectively. From the spectra, the difference between KDML105 and WR obviously observed at the wavelength over 1400 nm. While the PT1 spectra was not different from KDML105 according their overlap spectra. Considering the mixture samples between KDML105 and WR in Figure 3.5(B), there were nine different mixture set represent in nine different colors spectra. The trend of each mixture set was complicated due to the overlapped spectra. Therefore, those spectra should be investigated using chemometric technique to gain the significant information.

CHAPTER 4

Results and discussion

4.1 Exploratory data analysis of NIR spectra

A total of thirty-nine pure rice samples, including KDML105 (19 samples), WR (16 samples) and PT1 (4 samples), was measured using NIRSystem 6500 in wavelength of 400-2500 nm and the NIR spectra presented in Figure 4.1(A). Indicated using different colors, the NIR spectra of each KDML105, WR and PT1 were labelled in blue, red and green solid lines, respectively. A PCA scores plot against PC1 and PC2 is presented in Figure 4.1(B). The difference among the rice varieties can be observed. The spectrum profiles of the aromatic and non-aromatic rice can be separated. Most of KDML105 samples were projected on the top right corner and most of WR were projected on the bottom left corner. Although PT1 is categorized into an aromatic-rice grain, the fragrant quality of PT1 is usually less than that of the KDML105. In addition, the grain texture of PT1 is rather similar that of the white rice. This could be the reason that the samples of PT 1 were projected on the space overlapping between KDML105 and WR.

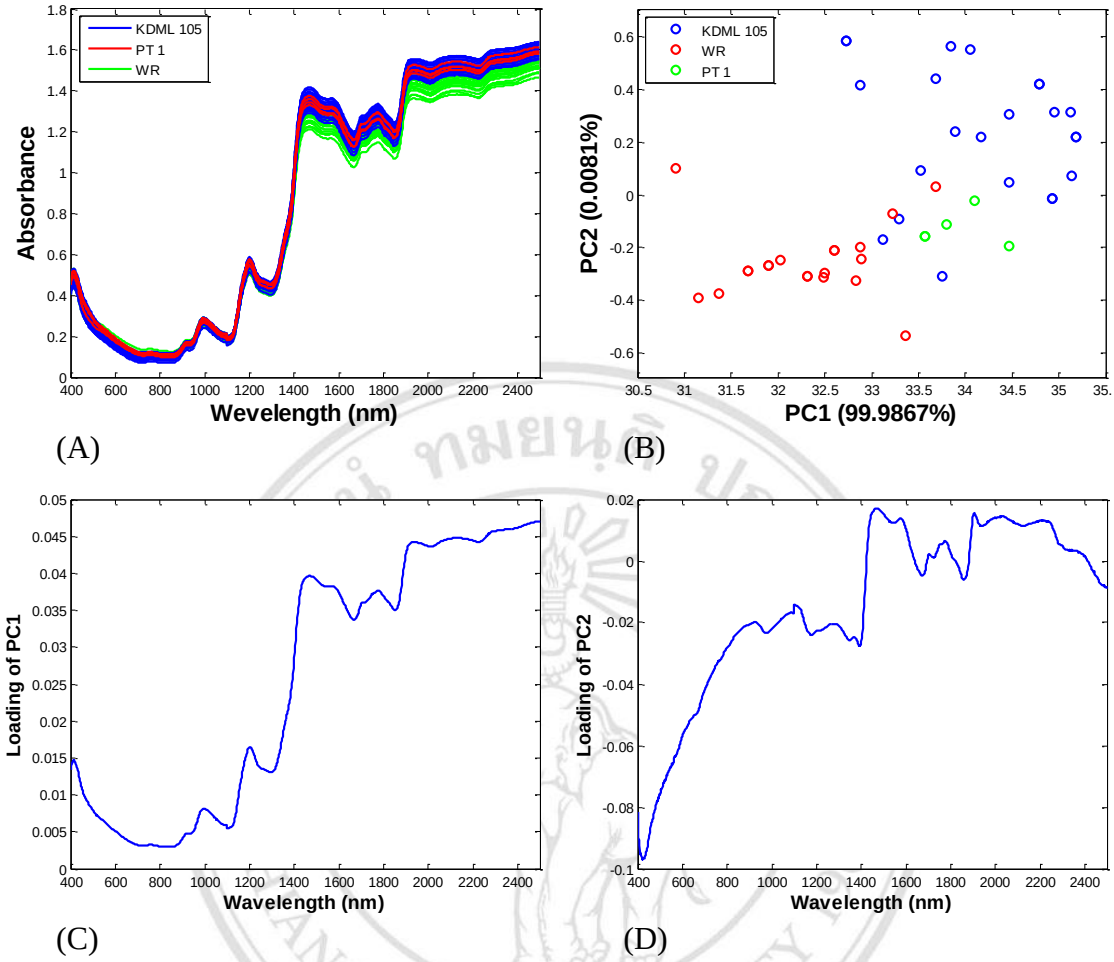


Figure 4.1 (A) NIR spectra of the studied rice, (B) PCA score plot of the rice samples and (C) and (D) are loading of PC1 and PC2, respectively with none data preprocessing

Loading of each parameter (nm) is visualized in Figure 4.1(C) and 4.1(D), respectively, for PC1 and PC2. In Figure 4.1(C), the loading profile of the PCA model is nearly identical to overall shape of the NIR spectra in Figure 4.1(A). This could be that the loading of PC1 (using non-data preprocessing) contributed 99.98% of the whole variation. In other words, PC1, in this case, is very important and characterized almost the entire data. Although the variation expressing using PC1 could be information and it is possible to conduce some conclusion in relation to the NIR-active functional bands, the size of the variation is relatively low.

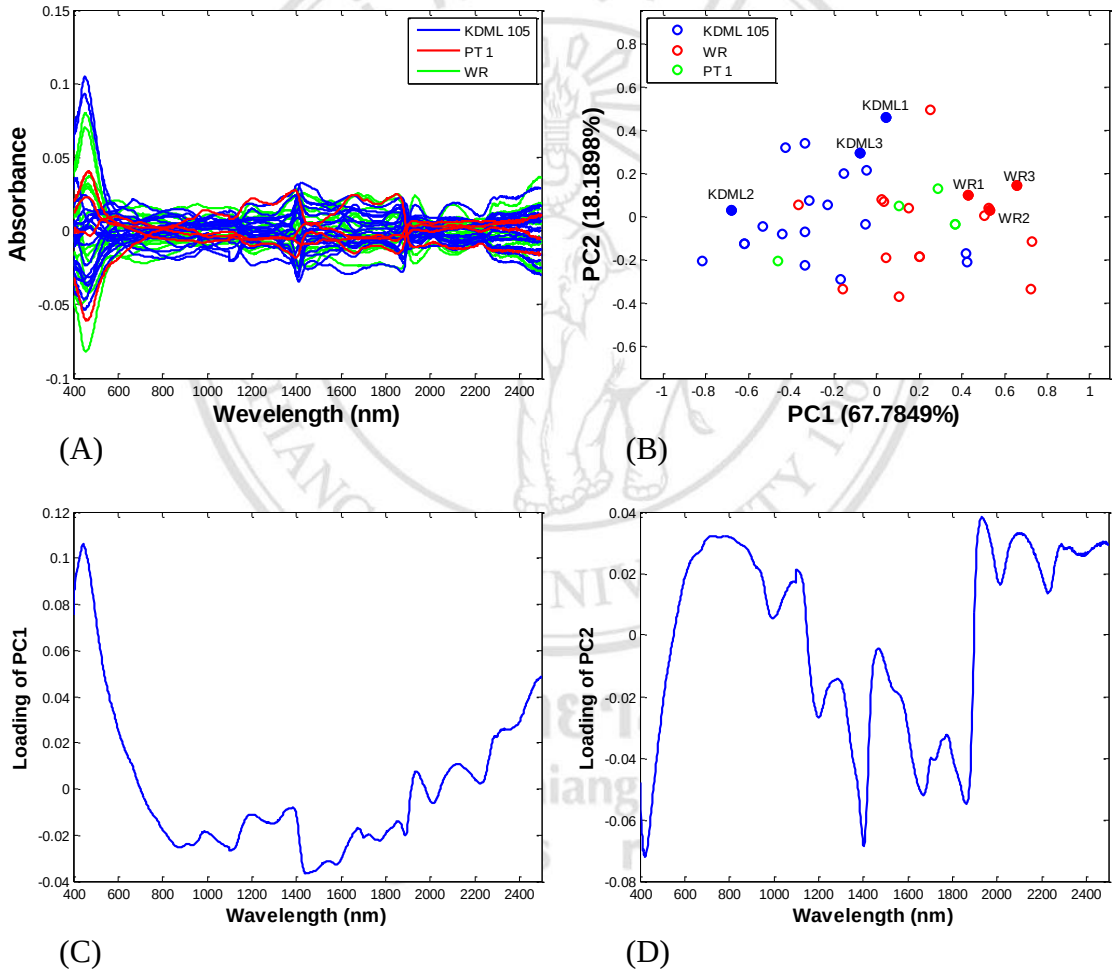


Figure 4.2 (A) NIR spectra of the studied rice, (B) PCA score plot of the rice samples and (C) and (D) are loading of PC1 and PC2, respectively, after processed by SNV and mean centring data preprocessing

In Figure 4.2, the same NIR spectra were preprocessed using SNV and followed by mean centring. After SNV and mean centring preprocessed, the shape of the NIR spectra was dramatically changed (Figure 4.2(A)). The organization of the rice samples after PCA was also different. However, the variation (%) of PC1 is decreased to 67.78%, whereas of PC2 is increased to 18.18%. In addition, the characteristic structures of the loading profiles are relatively clearer. The interpretation of the loading based in the NIR-active functional groups is summarized in Table 4.1

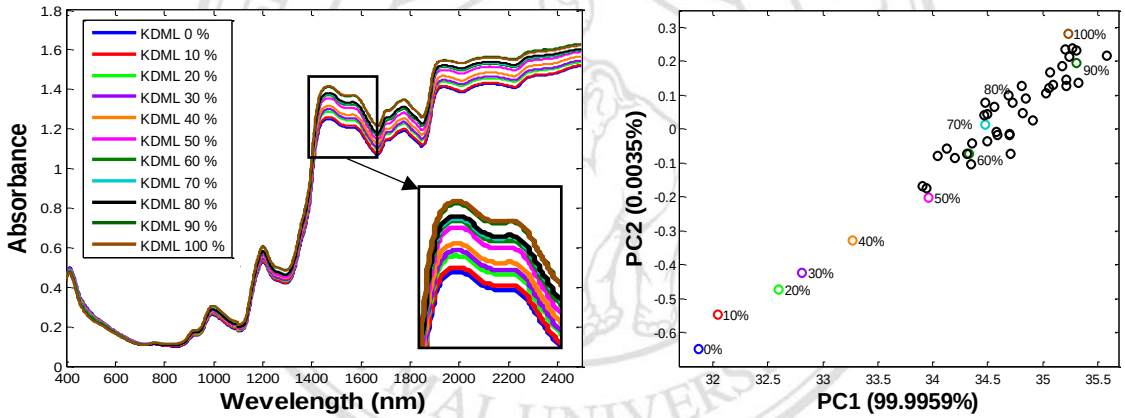
Table 4.1 The vibration of functional groups corresponding in wavelength

Wavelength (nm)	Vibration in functional group
1200	2^{nd} overtone of amylose
1490	1^{st} overtone of water
1780	1^{st} overtone of CH-streching of starch
1940	combination of OH-stretching of water and CH-stretching of amylose
2130 2300	CH-stretching of starch and protein

4.2 Generation of adulterated rice samples

From the EDA results, the samples presented using closed-face symbols were selected as a representative of two rice types (KDML105 and WR) for generating the adulterated samples. This includes three KDML105 (KDML1, KDML2 and KDML3) and three white rice (WR1, WR2 and WR3) as shown in Figure 4.2(B). Based on the selected samples, nine different datasets were generated to demonstrate the adulteration in the KDML105 rice. The nine-mixture datasets of rice samples, three different types of KDML105 adulterated with three different WR, were shown in Table 3.1. The capital letters A, B and C represents the different samples of KDML1, KDML2, and KDML3, respectively. Likewise, the numbers 1, 2 and 3 presents the WR1, WR2 and WR3, respectively. For example, the dataset A1 refers that samples are mixtures between KDML1 and WR1. In each dataset, the adulterated levels were 0, 10, 20, 30, 40, 50, 60, 61, 62,…, 100 %w/w of KDML105 resulting in a total of 47 samples. Figure 4.3

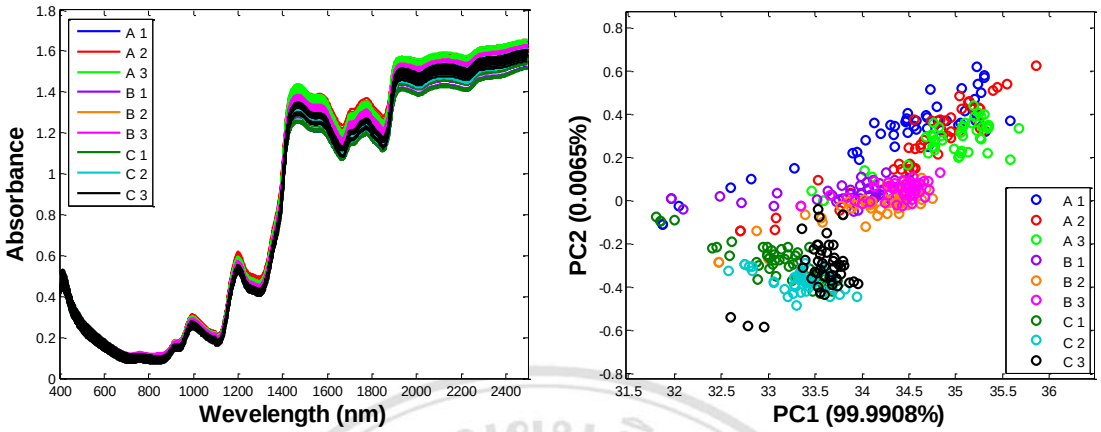
demonstrates the effect of adulterated level on NIR spectra in dataset A1. NIR spectra of each %w/w of KDML105 were presented in different color according their adulterated levels (0-100 %). In a region of 1250-1650 nm, the lower absorbance is observed if the rice samples have lower amount of KDML105. The absorbance increased when the percentage of KDML105 was higher. The PCA score plots of adulterated samples presented in Figure 4.3 (B). This result, from each PCA scores, highlights a linear trend in accordance with the amount of adulterated level where the samples having low amount of WR are located on the left-hand side. In addition, the concentration of KDML105 seems to increase along with the increase of PC1 and PC2 score values. This demonstrate the possibility that NIR can be used to distinguish the different between of the adulterated samples.



(A) (B)

Figure 4.3 (A) NIR spectra and (B) PCA scores plot of mixture samples with different adulterated levels

NIR spectra of nine-mixture datasets are illustrated in Figure 4.4(A) where the different color line indicates the different rice mixtures. The absorbance of each dataset was slightly varied due to its rice property used. To illustrate the feature in this dataset, a PCA scores plot was generated and shown in Figure 4.4(B). In this case, each dataset tended to be separated into three clusters according to the KDML105 samples where the groups of the samples using the same white rice, for example A1-A3, B1-B3 and C1-C3, were projected on their nearby the PCA space.



(A) (B)

Figure 4.4 (A) NIR spectra and (B) PCA scores plot of nine-mixture datasets

To study the feature of samples, the PCA scores plots of the mixture samples are presented in Figure 4.5. The mixture of KDML1, KDML2 and KDML3 with three different WR illustrated in Figure 4.5(A), 4.5(B) and 4.5(C), respectively. From those results, it obviously notices that the feature of mixture with the same KDML105 did not much distinguish although they were mixed with different WR types. In contrast, the mixture of WR1, WR2 and WR3 mixed with three different KDML105 displayed in Figure 4.5(D), 4.5(E) and 4.5(D), respectively. The result indicates that the mixture having different KDML105 tended to appear a different feature presented on PCA scores plot. The sample with the identical KDML105 was projected on nearby their groups. It could be that the variation in KDML105 was more significant row in NIR data of the mixture datasets.

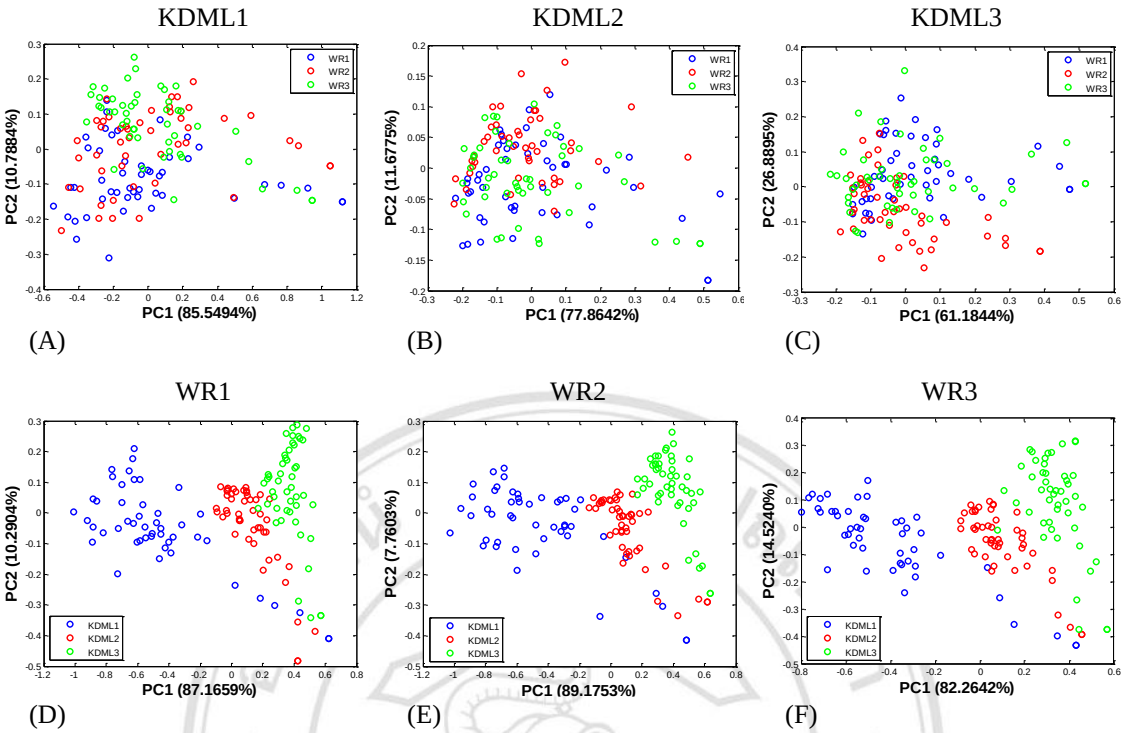


Figure 4.5 PCA score plots of mixture samples where (A)-(C) and (E)-(F) are the mixtures using the same KDML105 and WR, respectively

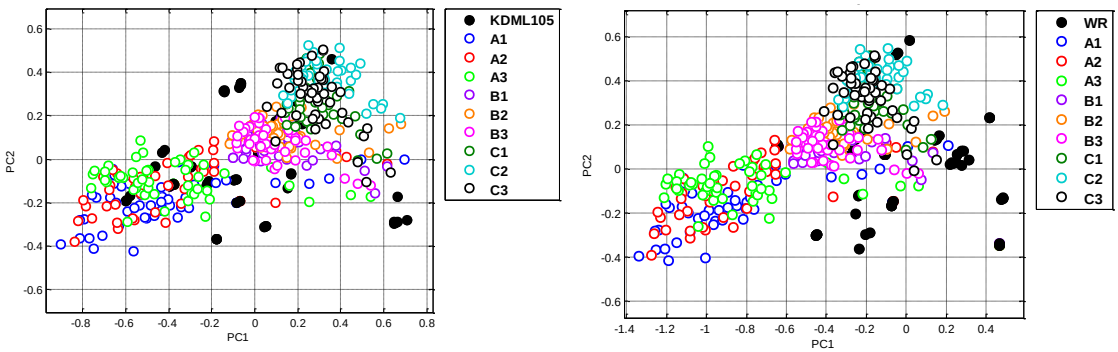
4.3 Untargeted detection using one class classification

The aim of this research section is to apply untargeted methods using D and Q statistics and SIMCA for identifying if the rice grain samples are the mixtures between KDML105 and WR.

4.3.1 Disjoint PCA of adulterated rice samples

Using one class classification, the model is established based on only the samples having specific quality. In this research, the target of the classification is KDML105 rice. However, the classification decision could be based on two different models. The first model is based on the variation of pure KDML105 samples. In this case, if an unknown is identified as out-group, this implies that the sample has been adulterated or it is a mixture sample. On the other hand, it is possible to generate the model using the variation from the pure WR samples. In this case, it is expected that the test samples should be predicted as out-group samples. In other words, the out-group samples are the WR samples that have been mixed by KDML105 rice.

To investigate the characteristic difference among the pure and the mixed rice samples, disjoint PCA plots modeled based on the pure rice samples were generated. The disjoint PCA modeled based on the pure KDML105 and the pure WR rice samples are presented in Figure 4.6(A) and (B), respectively. In Figure 4.6(A), although the samples are not completely separated according to their labelling classes, it is possible to observe the different between the samples from the different datasets. For example, The A1 dataset can be separated from the B3 dataset. Whereas, samples that are located nearby each other are from the samples having the same ingredients. For instance, samples from A1, A2 and A3 dataset which are clustered on the left-below side of the PCA space. For the disjoint KDML105 PCA model, the scores of the mixture samples are rather scattered around the KDML105 samples (labelled using the closed-face black circle). Differently, the scores of the mixture samples can be slightly separated from the mixture samples in the disjoint WR PCA model.



(A) (B)

Figure 4.6 Disjoint PCA modelled using loadings of (A) KDML105 and (B) WR

4.3.2 Identification of the adulterated rice samples using D and Q statistics

Using the KDML105 and WR as the reference samples, D and Q statistic were established. For this work, D and Q statistics were calculated (using the equation in section 2.4.3) and compared with their control limits. These control limits are the critical values estimated from the distribution of the reference systematic variation (***TP***) for the D statistic, and the distribution of the reference residual variation (***E***) for the Q statistic. The explanation of how to calculate the D and Q values and their confidence limit can be seen in Chapter 2. Unknown samples are of the same variation and could be considered as in-group if the calculated D and Q statistic values are small and fall within these limits. In this study, the statistic value is calculated for each sample and plotted in a bar chart together with the critical values. The 90% and 95% confidence limits for the D and Q charts are used for demonstration.

Figure 4.7 presents the D and Q charts for the one class modelling using KDML105 as the reference. In each bar chart, the first block presents the statistic values for the training samples, in Figure 4.7 (A), the KDML105 samples. The prediction results of the mixture samples are presented later from A1 to C3. The prediction of the last block is for the WR samples. As expected, most of pure samples from the different class are predicted as out-group. For example, the prediction of WR in the last block of Figure 4.7(A) predicted using KDML105 samples. However, in this case, it appears that the D statistic using the pure KDML105 samples as the reference variation could not result in good predictive classification as only a few mixture samples are identified as out-group according to the 95% confidence limit.

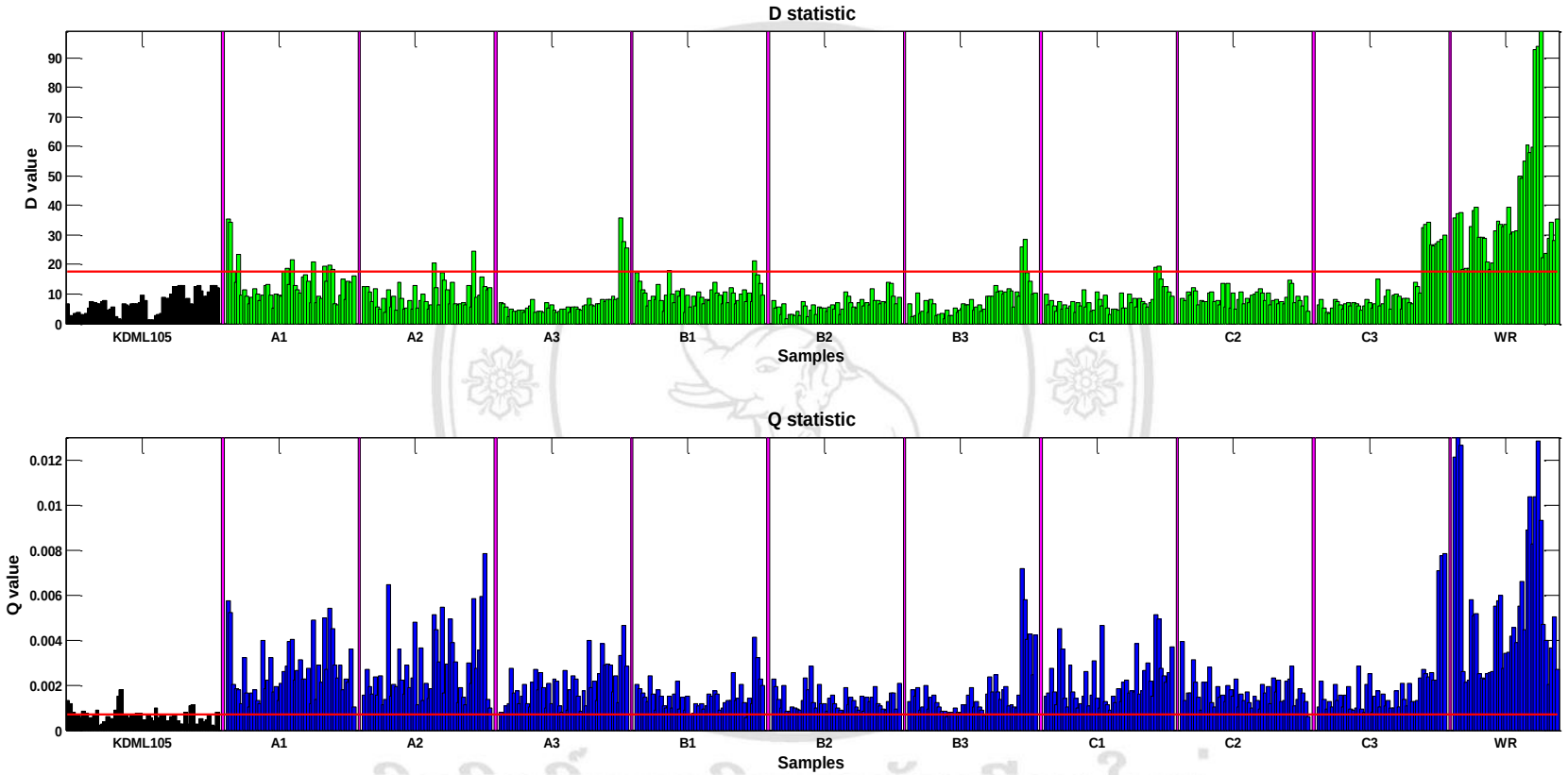


Figure 4.7 Bar charts for (A) D and (B) Q statistics using 95% confidence limit where the models were established using pure KDML105 as reference samples; the horizontal red line are 95% confidence limit

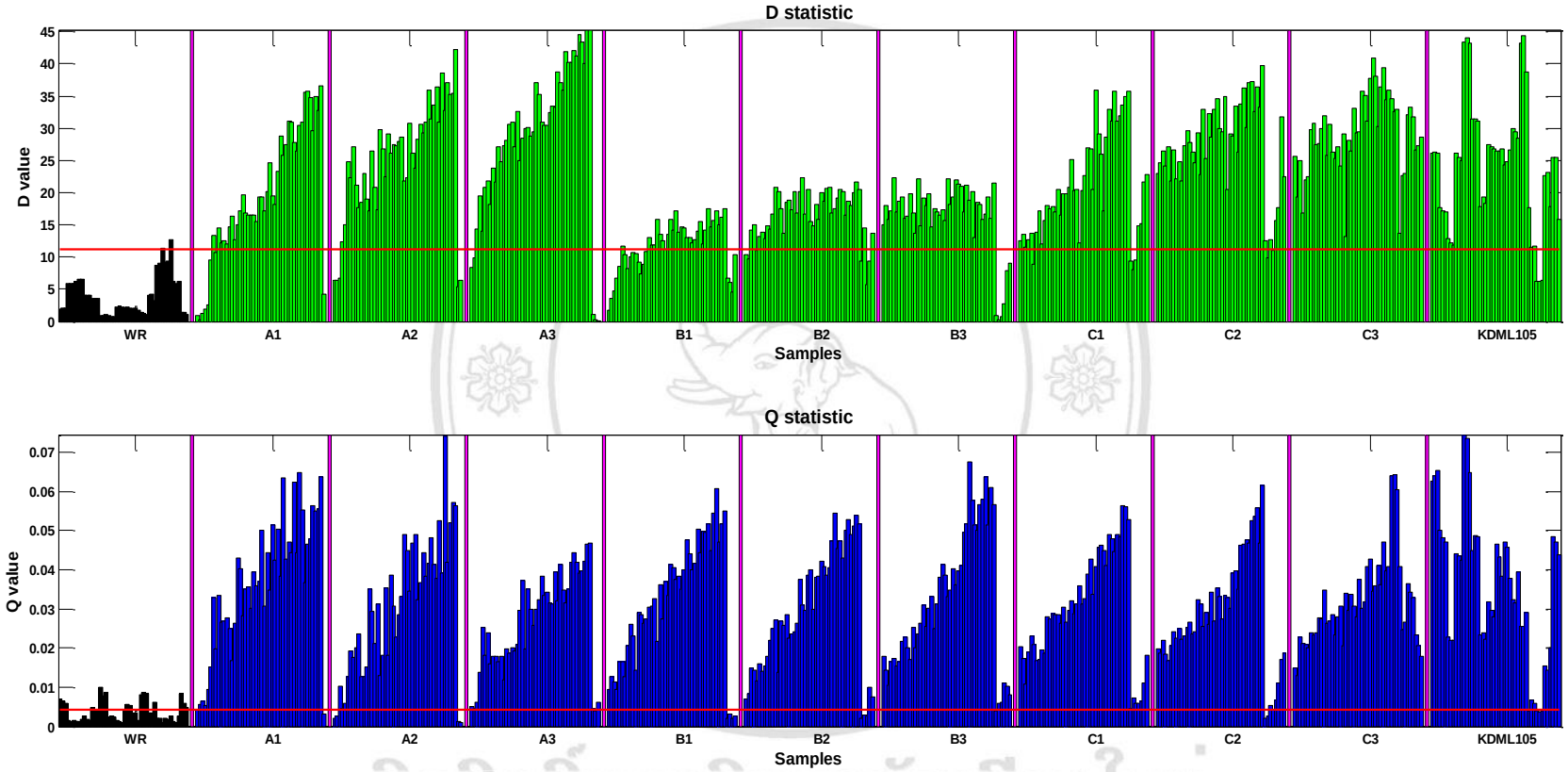


Figure 4.8 Bar charts for (A) D and (B) Q statistics using 95% confidence limit where the models were established using pure WR as reference samples; the horizontal red line are 95% confidence limit

When considering the Q statistics presented in Figure 4.7(B), the bar chart could provide more faithful predictive results as this one class classification model could successfully separate the mixture samples from the pure MDKL105 samples. This could be that in Q charts the statistic looks at the variation that cannot be modelled in a sample and saw if this non-systematic variation is significantly different from the non-systematic variation in the reference rice samples. Interestingly, the prediction models using the pure WR as reference could provide relatively better results. In Figure 4.8, using both D and Q statistics illustrated in Figure 4.8(A) and 4.8(B), respectively, most of the impure samples could be identified successfully. The increase of the D and Q values in each sample block are agreed well with the increase of the concentration of the KDML105. It is noted here that the adulteration rate of the rice product sold in market should be higher than 50%.

4.3.3 Identification of the adulterated rice samples using SIMCA

SIMCA combines the decision results from D and Q statistics. In this work, the SIMCA prediction results are visualized using a 2-dimensional graph where x and y axes represents the calculated D and Q values, respectively. The confidence limits are indicated using vertical and horizontal red lines.

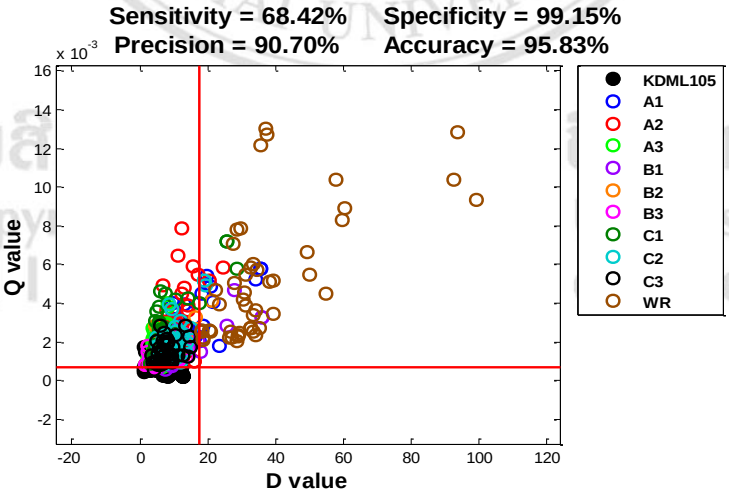


Figure 4.9 SIMCA model using pure KDML105 as reference where the red lines present 95% confidence limits for D (horizontal line) and Q (vertical line) statistics, respectively

Figure 4.9 demonstrates the SIMCA prediction using pure KDML105 as the reference samples. This identification results are corresponding to the D and Q results as discussed in the previous section. For KDML105 SIMCA model using 95% confidence limit, most of the samples are identified as out-group for only the prediction using Q statistics. In contrast, not many samples are identified as out-group for D statistics. However, using the criterion that in-group should only within both D and Q statistics. This means that the in-group samples can be only in the square box on the left-below if the SIMCA space. In Figure 4.9, the in-group area is highlighted using a light-yellow color. In this case, the specificity of the model is relatively high (99.15%). This implies that the SIMCA model could successfully capture most of the mixture samples. In addition, the model could satisfactorily provide classification results for both pure samples and mixture samples when considering the accuracy of 95.83%. A slightly lower of the accuracy could be due to that the low sensitivity of the classification model (68.42%).

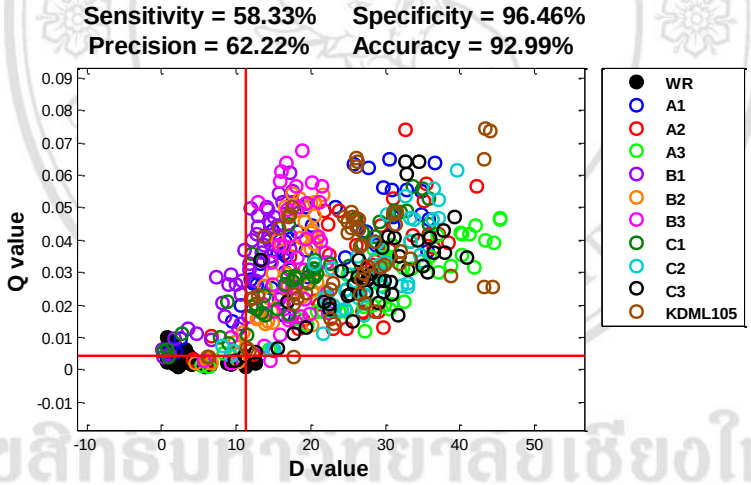


Figure 4.10 SIMCA model using pure WR as reference where the red lines present 95% confidence limits for D (horizontal line) and Q (vertical line) statistics, respectively

Although the prediction using D and Q statistics based on the WR samples could identify most of the mixture samples, the SIMCA prediction based on the WR samples shows the specificity of 92.99% as presented in Figure 4.10. This could be that in this research, the aim is to identify the adulteration in KDML105. Using the disjoint PCA modeling, the non-systematic variation part (the variation that was not included in the D statistics) could be well related to the variation of the WR. Therefore, most of the mixture

samples which are adulterated by WR could be identified using the Q statistics (Figure 4.8(B)). This support the classification ability of the SIMCA model.

4.4 Calibration for quantification of adulteration in KDML105

This research section aims to develop a NIR-chemometric model that can be used to quantify the adulteration level of the rice mixtures. The pure variation of KDML105 in the test samples will be extracted using OP and PDS so that the PLS calibration can be exclusively established without the variation of WR. Figure 4.11 is a diagram explaining the research procedure in this section. A variation of four different PLS models will be tested including PLS, OP-PLS, PDS-PLS and OP-PDS-PLS.

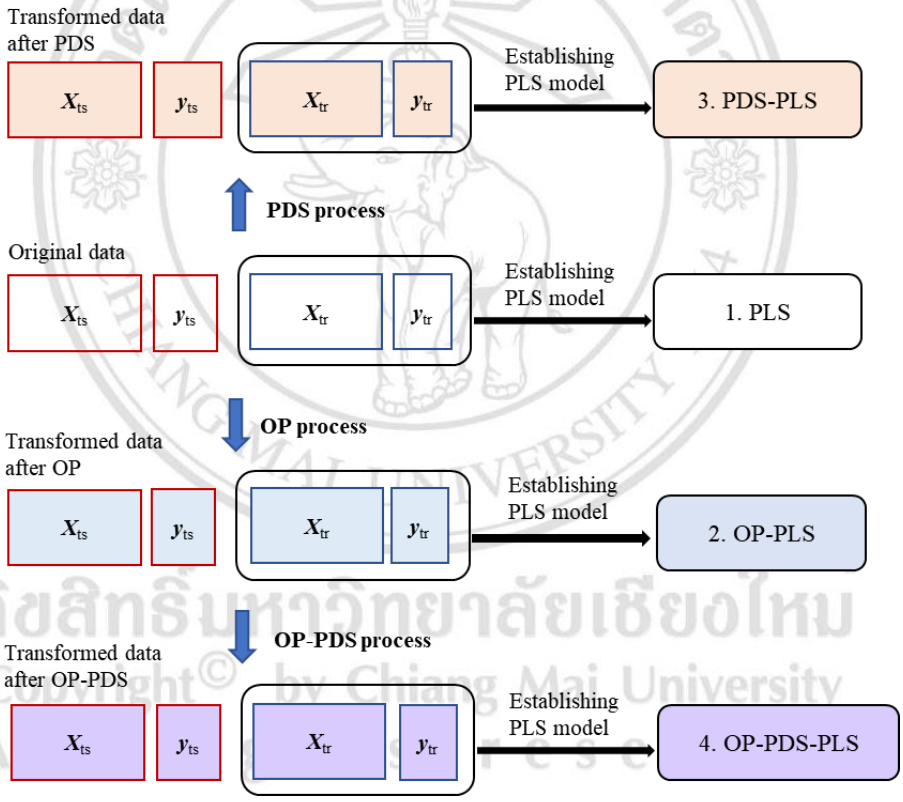


Figure 4.11 Diagram of the developed calibration models; PLS, OP-PLS, PDS-PLS and OP-PDS-PLS

For quantification of an amount of KDML105 in mixture samples, two significant methods which are extensively utilized with calibration analysis, orthogonal projection (OP) and calibration transfer (using PDS), are applied. The effect of these two methods are concerned and investigated in the next section.

4.4.1 The effect of OP and PDS methods on NIR data

To investigate the effect of OP and PDS methods on the variation in sample data, the NIR spectra have been analyzed in terms of original data, OP, PDS and the combination of OP and PDS methods using PCA. Two datasets of A1 and C3 were selected as a representative of the data having different KDML105 and WR types to investigate that effect. Figure 4.12 presents the PCA scores of the data that were removed the variation of WR from mixture data in OP process.

The data was carried out with none data preprocessing to ensure that any changes was based on the method used without any effects of data pretreatment. Considerate the PCA scores plot of original data in Figure 4.12(A), two mixture data of A1 and C3 represented two types of KDML1 and KDML3, respectively, were projected on different space according to their different characteristics. It indicated that the calibration analysis between the original sample having different kinds of KDML105 and WR could poor. After applied OP process, two groups of samples tended to be projected nearby each other that shown in Figure 4.12(B). This result shows that after removing the variation of WR, these two datasets have more similar characteristic than that original data. In case of PDS process, Figure 4.12(C), there were the overlapping of some parts of two groups after the PDS was carried out. This confirmed that using PDS can standardize two different data to analyze in the same calibration model. However, there were some parts of the data that were not overlap in the bottom left area. It might be caused the error of analysis results. Therefore, the combination of OP and PDS was performed with the NIR spectra shown in Figure 4.12(D). This meant that the original data have been removed the WR variations and then standardized between two different datasets. Comparing with the original data in Figure 4.12(A), most of samples in two groups obviously moved closer and overlapped in the same space of PCA scores plot. It indicated that using OP coupled with PDS can modify the variations of two different dataset for calibration analysis.

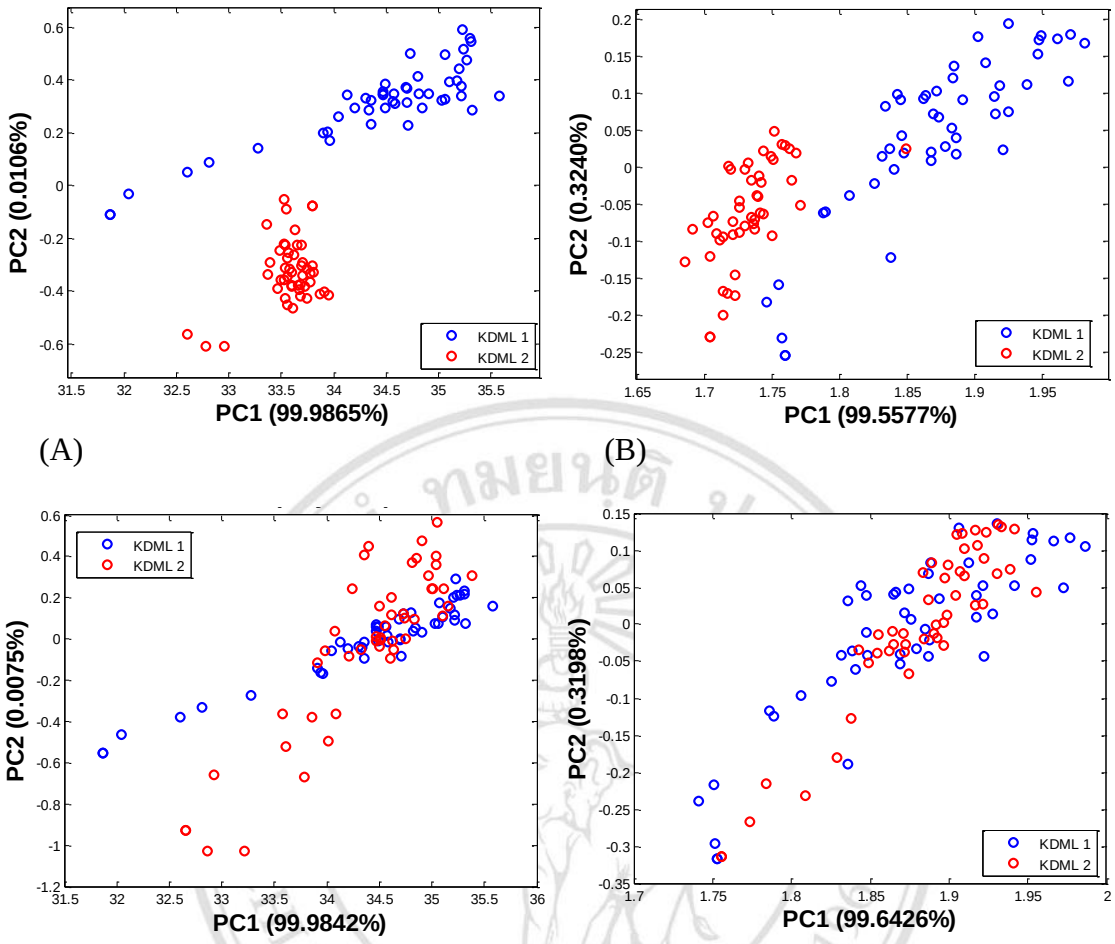


Figure 4.12 PCA scores plots of two mixture datasets (A1 and C3) with (A) original data, (B) after OP, (C) after PDS and (D) after OP coupled with PDS processes; the removed data was WR

The effect of OP and PDS on variation in data was also performed with the data removing KDML105 variation. Figure 4.13(A)- 4.13(D) illustrate the PCA scores plots of NIR data with original data, after OP, after PDS and the combination of OP and PDS processes, respectively. Likewise, the results of the data removing KDML105 variation were as similar as the WR removed data. In conclude, both OP and PDS processes have the effect on NIR data. After applying OP, the unwanted variations, in this case, WR or KDML, were removed from the original data, the characteristics of two datasets could be adjusted. Therefore, the space between two groups could decrease that was noticeable in PCA scores plot. For PDS process, the scores of two datasets overlapped in over half of all samples. It demonstrated that using calibration transfer

could transform two different datasets to be suitable for calibration analysis. However, the combination of OP and PDS can overcome other methods.

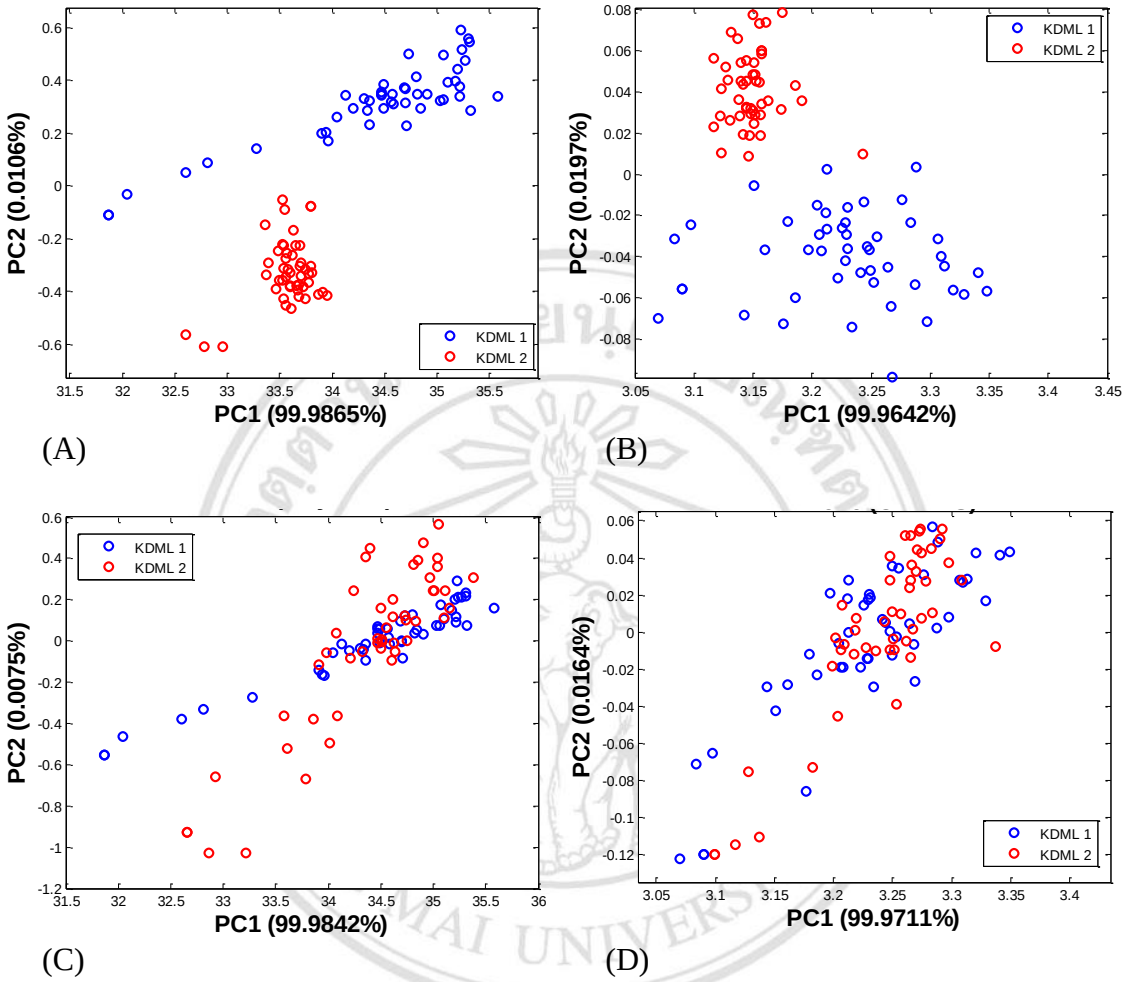


Figure 4.13 PCA scores plots of two mixture data sets (A1 and C3) with (A) original data, (B) after OP process, (C) after PDS process and (D) after OP coupled with PDS process; the removed data was KDML105

In this research, the calibration model of four methods including (1) only PLS, (2) OP-PLS, (3) PDS-PLS and (4) OP-PDS-PLS were established to quantify the amount of KDML105 in adulterated sample. To estimate these methods, these four datasets (A1, A2, B1, C3) were selected for constructing the calibration models to investigate the types of rice data as they affected to the predictive performance including of the identical or different type of KDML105 and the identical or different type of WR (the rest datasets were used to generate the calibration model and shown in Appendix I). Six calibration

models, in Table 4.2, were established for investigation which method provide the best predictive performance.

Table 4.2 The datasets for establishing six calibration models

No.	Dataset	Training set	Test set	Notes
1	A1-A2	A1	A2	Identical KDML105 and different WR
2	A2-A1	A2	A1	Identical KDML105 and different WR
3	A1-B1	A1	B1	Different KDML105 and identical WR
4	B1-A1	B1	A1	Different KDML105 and identical WR
5	A1-C3	A1	C3	Different KDML105 and WR
6	C3-A1	C3	A1	Different KDML105 and WR

In addition, to obtain the best calibration model, the significant parameters of each method should be concerned such as the number of PC for OP (*m*), the number LV for PLS (*n*) and the window size for PDS (*w*). These significant parameters were optimized where *m* and *n* were defined from 0 to 12 and *w* was 5, 13, 15, 31 and 35.

4.4.2 Optimization of the window size of PDS from the predictive performance of only PLS and PDS-PLS methods

Those six models were used to optimize the subset or window size (*w*) in PDS process. In each model, the dataset of sample was divided into four groups, master (training and test sets) and slave (training and test sets). Since the data preprocessing have more effect to the spectroscopic analysis, in this case, there were three preprocessing applied with all data include SNV, mean centring and the combined of both methods. The samples were collected with different window size (w = 5, 13, 15, 31, 35) on training of slave to relate with training of master for computing the transformation matrix ***F*** by PDS. Then the test of slave was modified by ***F*** to obtain the transformed data of slave. The training set of master was used to establish the PLS model and two forms of test set of slaves (before and after PDS process) were then predict with that calibration model to evaluate the capability of PDS. The predictive performance with different window size of the SNV preprocessed shown in Table 4.3.

The model of A1-A2 and A2-A1 represent the model using an identical type of KDML105 adulterated with different types of WR. These models provided the dominant predictive abilities when comparing with other models. The predictive ability before (Q^2_{be} and $RMSEP_{be}$) and after (Q^2_{af} and $RMSEP_{af}$) PDS process was not much different, thus it revealed that the sample having identical KDML105 mixed with any type of WR could be well estimated the adulteration using only PLS process. Considering other models, the model using different type of KDML105 mixed with identical WR (A1-B1 and B1-A1) and the models using the data having both different KDML105 and WR provided the decreased predictive result in comparison with those former models. Additionally, using PDS did not improve the prediction if the samples have different types of KDML105 with SNV preprocessing. The optimum *w* and *n* were 31 and 4, respectively, with A2-A1 model.

Table 4.3 The optimization of window size (*w*) of six datasets with SNV where *w* = window size, *n* = number of LV, R^2 = coefficient of determination for calibration, RMSEC = root mean square error of calibration, Q^2_{be} and the Q^2_{af} = coefficient of determination for prediction before and after PDS process, $RMSEP_{be}$ and $RMSEP_{af}$ = RMSEP before and after PDS process

Model	*w*	*n*	R^2	RMSEC	Q^2_{be}	$RMSEP_{be}$	Q^2_{af}	$RMSEP_{af}$
	5	4	0.9936	1.8833	0.9787	3.0405	0.9852	2.5325
	13	4	0.9936	1.8833	0.9818	2.8141	0.9852	2.5325
A1-A2	15	4	0.9936	1.8833	0.9817	2.8158	0.9852	2.5325
	31	**4**	**0.9936**	**1.8833**	**0.9833**	**2.6905**	**0.9852**	**2.5325**
	35	4	0.9936	1.8833	0.9829	2.7208	0.9852	2.5325
	5	**6**	**0.9961**	**1.4645**	**0.9851**	**2.5448**	**0.9782**	**3.0734**
	13	6	0.9961	1.4645	0.9844	2.6015	0.9782	3.0734
A2-A1	15	6	0.9961	1.4645	0.9844	2.5998	0.9782	3.0734
	31	6	0.9961	1.4645	0.9830	2.7147	0.9782	3.0734
	35	6	0.9961	1.4645	0.9826	2.7443	0.9782	3.0734
	5	4	0.9936	1.8833	0.9421	5.0146	0.9852	2.5325
	13	4	0.9936	1.8833	0.9459	4.8458	0.9852	2.5325
A1-B1	15	4	0.9936	1.8833	0.9539	4.4718	0.9852	2.5325
	31	4	0.9936	1.8833	0.9739	3.3678	0.9852	2.5325
	35	**4**	**0.9936**	**1.8833**	**0.9740**	**3.3623**	**0.9852**	**2.5325**
	5	**13**	**1.0000**	**0.1625**	**0.8989**	**6.6226**	**0.9491**	**4.7005**
	13	13	1.0000	0.1625	0.8731	7.4216	0.9491	4.7005
B1-A1	15	13	1.0000	0.1625	0.8574	7.8657	0.9491	4.7005
	31	13	1.0000	0.1625	0.7543	10.3262	0.9491	4.7005
	35	13	1.0000	0.1625	0.7750	9.8822	0.9491	4.7005
	5	4	0.9936	1.8833	0.9770	3.1605	0.9852	2.5325
	13	4	0.9936	1.8833	0.9794	2.9928	0.9852	2.5325
A1-C3	15	4	0.9936	1.8833	0.9799	2.9519	0.9852	2.5325
	31	**4**	**0.9936**	**1.8833**	**0.9817**	**2.8158**	**0.9852**	**2.5325**
	35	4	0.9936	1.8833	0.9814	2.8396	0.9852	2.5325
	5	9	0.9988	0.8186	0.8894	6.9282	0.9759	3.2355
	13	**9**	**0.9988**	**0.8186**	**0.9179**	**5.9675**	**0.9759**	**3.2355**
C3-A1	15	9	0.9988	0.8186	0.9165	6.0196	0.9759	3.2355
	31	9	0.9988	0.8186	0.9017	6.5332	0.9759	3.2355
	35	9	0.9988	0.8186	0.9083	6.3088	0.9759	3.2355

Table 4.4 The optimization of window size (*w*) of six datasets with mean centring where *w* = window size, *n* = number of LV, R^2 = coefficient of determination for calibration, RMSEC = root mean square error of calibration, Q^2_{be} and the Q^2_{af} = coefficient of determination for prediction before and after PDS process, $RMSEP_{be}$ and $RMSEP_{af}$ = RMSEP before and after PDS process

Model	*w*	*n*	R^2	RMSEC	Q^2_{be}	$RMSEP_{be}$	Q^2_{af}	$RMSEP_{af}$
	5	**4**	**0.9933**	**1.9269**	**0.9540**	**4.4694**	**0.9828**	**2.7315**
	13	4	0.9933	1.9269	0.9459	4.8460	0.9828	2.7315
A1-A2	15	4	0.9933	1.9269	0.9463	4.8268	0.9828	2.7315
	31	4	0.9933	1.9269	0.9474	4.7784	0.9828	2.7315
	35	4	0.9933	1.9269	0.9424	5.0010	0.9828	2.7315
	5	**5**	**0.9951**	**1.6455**	**0.9825**	**2.7577**	**0.9688**	**3.6816**
	13	5	0.9951	1.6455	0.9818	2.8130	0.9688	3.6816
A2-A1	15	5	0.9951	1.6455	0.9820	2.7965	0.9688	3.6816
	31	5	0.9951	1.6455	0.9573	4.3068	0.9688	3.6816
	35	5	0.9951	1.6455	0.9548	4.4304	0.9688	3.6816
	5	4	0.9933	1.9269	0.8468	8.1541	0.9828	2.7315
	13	4	0.9933	1.9269	0.9359	5.2733	0.9828	2.7315
A1-B1	15	4	0.9933	1.9269	0.9461	4.8382	0.9828	2.7315
	31	4	0.9933	1.9269	0.9738	3.3699	0.9828	2.7315
	35	**4**	**0.9933**	**1.9269**	**0.9765**	**3.1905**	**0.9828**	**2.7315**
	5	**14**	**1.0000**	**0.0937**	**0.7579**	**10.2507**	**0.9436**	**4.9453**
	13	14	1.0000	0.0937	0.3264	17.0978	0.9436	4.9453
B1-A1	15	14	1.0000	0.0937	-0.2725	23.5004	0.9436	4.9453
	31	14	1.0000	0.0937	-1.0447	29.7888	0.9436	4.9453
	35	14	1.0000	0.0937	-0.9333	28.9657	0.9436	4.9453
	5	**4**	**0.9933**	**1.9269**	**0.7507**	**10.4020**	**0.9828**	**2.7315**
	13	4	0.9933	1.9269	0.7297	10.8302	0.9828	2.7315
A1-C3	15	4	0.9933	1.9269	0.7452	10.5162	0.9828	2.7315
	31	4	0.9933	1.9269	0.7466	10.4871	0.9828	2.7315
	35	4	0.9933	1.9269	0.7503	10.4107	0.9828	2.7315
	5	**9**	**0.9983**	**0.9842**	**-0.2183**	**22.9942**	**0.9683**	**3.7100**
	13	9	0.9983	0.9842	-0.9041	28.7461	0.9683	3.7100
C3-A1	15	9	0.9983	0.9842	-1.4225	32.4244	0.9683	3.7100
	31	9	0.9983	0.9842	-2.0358	36.2976	0.9683	3.7100
	35	9	0.9983	0.9842	-1.2139	30.9971	0.9683	3.7100

The optimization of PDS process for the data with mean centring shown in Table 4.4. The results were quite similar to the preprocessed SNV that the best predictive abilities occurred with the model using data having identical type of KDML105. Also, the PDS process did not enhance the results. It indicated that all predictive abilities were not well when using mean centring. The optimum *w* and *n* were 5 and 5, respectively, with A2-A1 model.

Table 4.5 The optimization of window size (w) of six datasets with SNV + mean centring where *w* = window size, *n* = number of LV, R^2 = coefficient of determination for calibration, RMSEC = root mean square error of calibration, Q^2_{be} and the Q^2_{af} are coefficient of determination for prediction before and after PDS process, $RMSEP_{be}$ and $RMSEP_{af}$ = RMSEP before and after PDS process

Model	*w*	*n*	R^2	RMSEC	Q^2_{be}	$RMSEP_{be}$	Q^2_{af}	$RMSEP_{af}$
	5	3	0.99361	1.88335	0.97853	3.05280	0.98520	2.53415
	13	3	0.99361	1.88335	0.98119	2.85728	0.98520	2.53415
A1-A2	15	3	0.99361	1.88335	0.98104	2.86829	0.98520	2.53415
	31	**3**	**0.99361**	**1.88335**	**0.98120**	**2.85650**	**0.98520**	**2.53415**
	35	3	0.99361	1.88335	0.97968	2.96967	0.98520	2.53415
	5	**5**	**0.99615**	**1.46227**	**0.98297**	**2.71844**	**0.97830**	**3.06871**
	13	5	0.99615	1.46227	0.98091	2.87824	0.97830	3.06871
A2-A1	15	5	0.99615	1.46227	0.98090	2.87942	0.97830	3.06871
	31	5	0.99615	1.46227	0.98200	2.79477	0.97830	3.06871
	35	5	0.99615	1.46227	0.98170	2.81789	0.97830	3.06871
	5	3	0.99361	1.88335	0.92512	5.70059	0.98520	2.53415
	13	3	0.99361	1.88335	0.94547	4.86453	0.98520	2.53415
A1-B1	15	3	0.99361	1.88335	0.95273	4.52919	0.98520	2.53415
	31	**3**	**0.99361**	**1.88335**	**0.97392**	**3.36401**	**0.98520**	**2.53415**
	35	3	0.99361	1.88335	0.97387	3.36721	0.98520	2.53415
	5	12	0.99995	0.16232	0.71064	11.20621	0.94910	4.70006
	13	12	0.99995	0.16232	0.78438	9.67345	0.94910	4.70006
B1-A1	**15**	**12**	**0.99995**	**0.16232**	**0.84775**	**8.12851**	**0.94910**	**4.70006**
	31	12	0.99995	0.16232	0.81613	8.93287	0.94910	4.70006
	35	12	0.99995	0.16232	0.83694	8.41230	0.94910	4.70006
	5	3	0.99361	1.88335	0.97428	3.34099	0.98520	2.53415
	13	3	0.99361	1.88335	0.97641	3.19975	0.98520	2.53415
A1-C3	15	3	0.99361	1.88335	0.97678	3.17422	0.98520	2.53415
	31	3	0.99361	1.88335	0.98016	2.93417	0.98520	2.53415
	35	**3**	**0.99361**	**1.88335**	**0.98058**	**2.90325**	**0.98520**	**2.53415**
	5	8	0.99880	0.81765	0.87076	7.48911	0.97589	3.23445
	13	8	0.99880	0.81765	0.95331	4.50155	0.97589	3.23445
C3-A1	**15**	**8**	**0.99880**	**0.81765**	**0.96808**	**3.72179**	**0.97589**	**3.23445**
	31	8	0.99880	0.81765	0.95198	4.56513	0.97589	3.23445
	35	8	0.99880	0.81765	0.93503	5.31019	0.97589	3.23445

Table 4.5 presents the results of all models using the combined SNV and mean centring. In overall, using the combined preprocessing provided the better results in all models comparing with those former preprocessing. It could be assumed that some errors were get rid of from NIR spectra including the scattering error and the influence of the mean. However, the PDS process did not still enhance the predictive performance except a little better in the A2-A1 model. The optimum *w* and *n* were 5 and 5, respectively, with A2-A1 model.

The predictive performance of all calibration models before PDS process meant that the models were generated without passed the PDS, thus those models represented only PLS performance. Most of the model presented the best predictive results using only PLS aside from the A2-A1 model that gave a better result with using the PDS-PLS method. Considering the window size, the best size of *w* was 5, however, that did not lead to the satisfactory predictive result for all models. Therefore, the $w = 31$ was selected in a case of compromising the good results in all models.

4.4.3 Optimization of number of PCs (*m*) and LVs (*n*) and the comparison of the predictive performance between OP-PLS and OP-PDS-PLS methods

Using OP and PLS methods should be aware of dimension of PC used since it has influence on the prediction. In this research, these two parameters (*m* and *n*), therefore, were performed in various values to optimize the appropriate PC and LV considering by Q^2 and RMSEP. The predictive performances were exhibited in colormap where the x and y axes represented the number of *m* and *n*, respectively. The value of Q^2 and RMSEP represented in various shade of pink where the lighter color was the higher values and the darker color was the lower values. The red text in the color map shows the optimum values of OP-PDS-PLS method. The white and black text shows the optimum values of OP-PLS method. The optimization of *m* and *n* were carried out through the six models with different three preprocessed methods like the previous section. Only the colormap of A1-C3 results with three preprocessing were shown in this case as a representative for discussion.

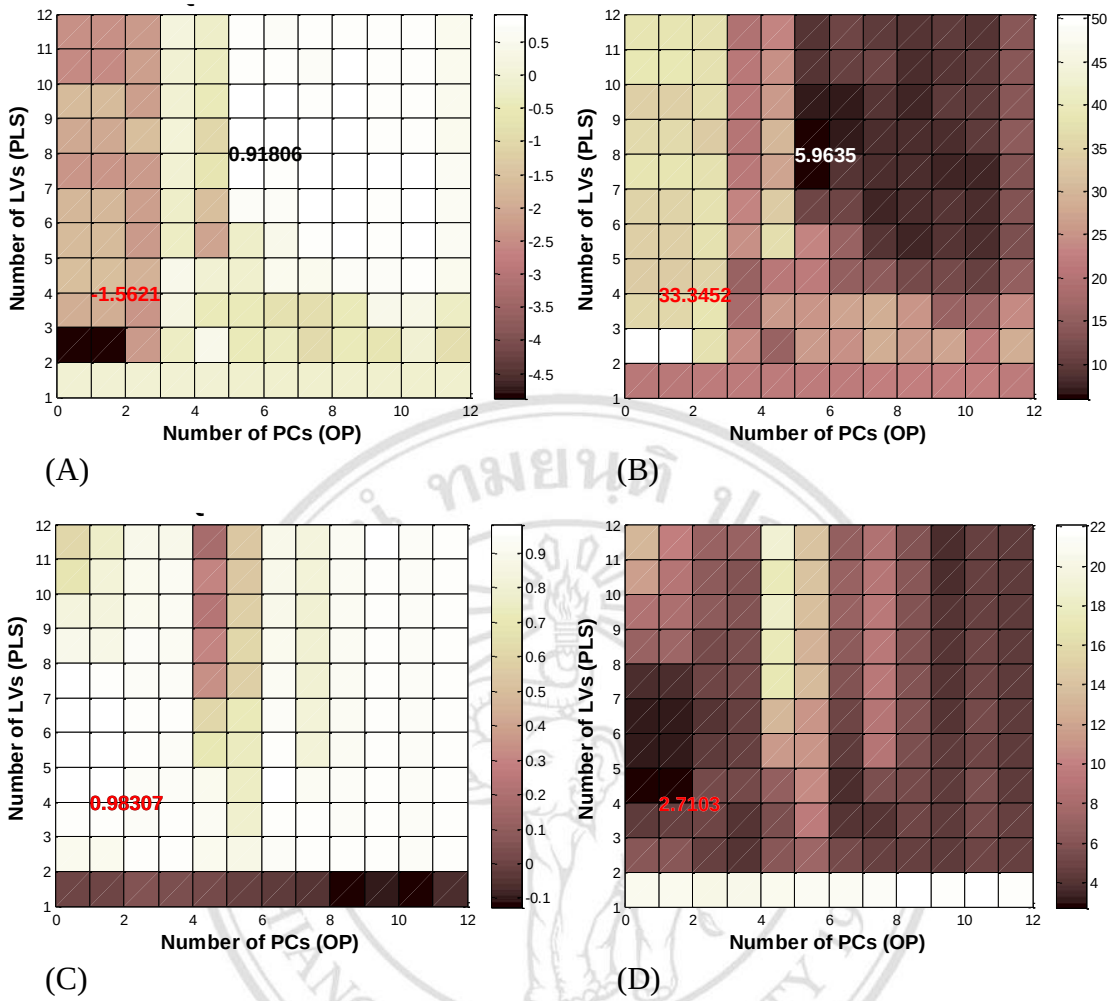


Figure 4.14 The colormaps of (A) Q^2 and (B) RMSEP values of OP-PLS, (C) Q^2 and (D) RMSEP values of OP-PDS-PLS methods with SNV prepressing; $m = 0 - 12$, $n = 1 - 12$, $w = 31$

Figure 4.14 is the colormap of the SNV preprocessing. These results show the colormap of Q^2 and RMSEP for two methods (OP-PLS and OP-PDS-PLS) which the position is (m, n). the darker color represents the lower value and the lighter color represents the higher value. Considering on the method OP-PLS in Figure 4.14(A) and 4.14(B), the optimum values are (5,8) that the Q^2 and RMSEP are 0.91806 and 5.9635, respectively. The trend of those colormap showed that using low value of m and n leaded to the poor predictive results. In case of the OP-PDS-PLS in Figure 4.14(C) and 4.14(D), the optimum values are (1,4) with Q^2 and RMSEP are 0.9831 and 2.7103, respectively. Overall result of the preprocessed SNV indicated that using OP with PDS can reduce the number of LV from 8 to 4. The colormap of OP-PLS shows that good result occurred

when the selected PC > 5 and LV > 3. For OP-PDS-PLS, all good predictions obtained in almost selected PC and LV except the number of n =1.

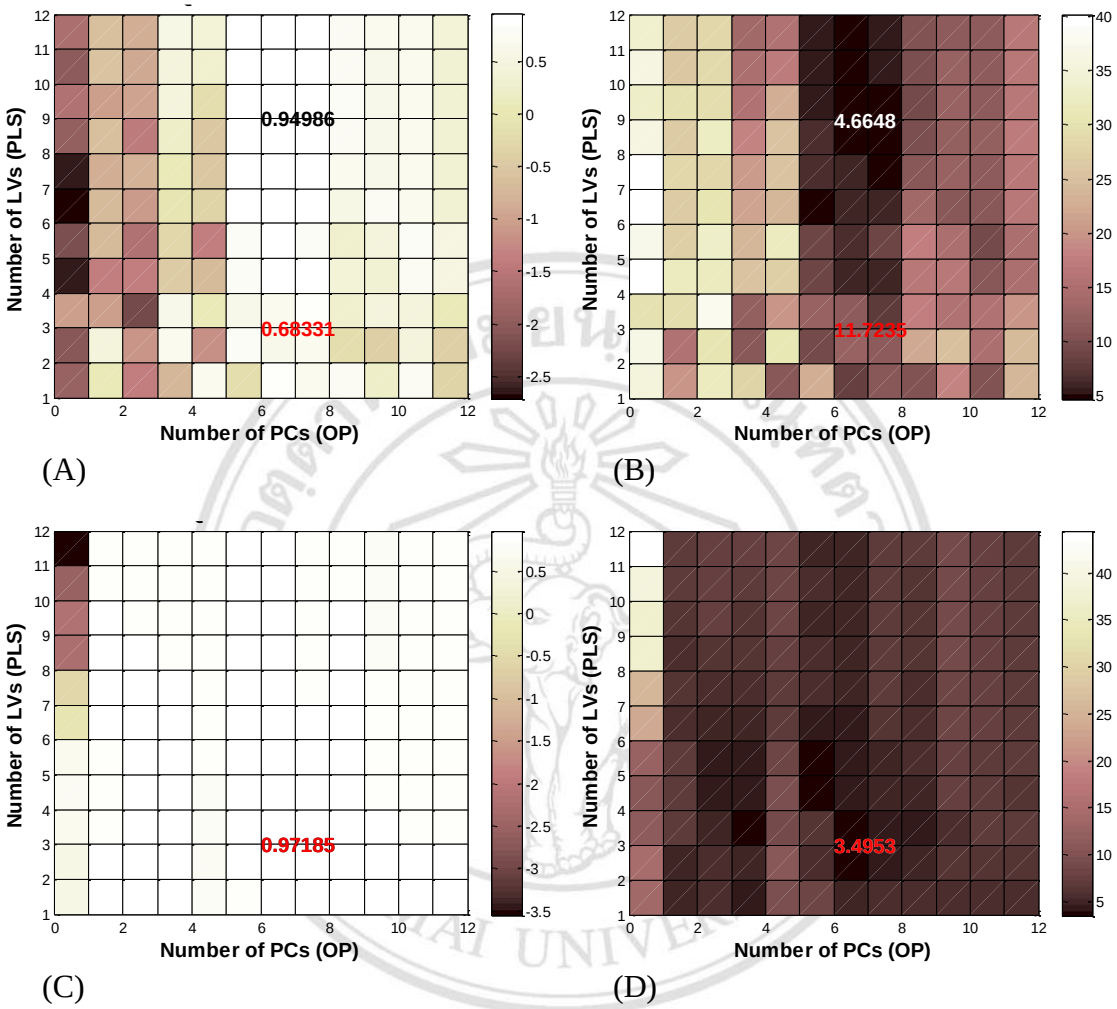


Figure 4.15 The colormaps of (A) Q^2 and (B) RMSEP values of OP-PLS, (C) Q^2 and (D) RMSEP values of OP-PDS-PLS methods with mean centring prepressing; m = 0 – 12, n = 1 – 12, w = 31

The effect of mean centring on prediction is exhibited in Figure 4.15. The best O^2 and RMSEP values occurrs at the position of (6,9) for using OP-PLS with 0.94986 and 4.6648, respectively. In OP-PDL-PLS, the best predive performance (Q^2 = 0.97185, RMSEP = 3.4953) shown at the position (6,3). It is obvious that using OP couple with PDS can reduce the number of LVs and can improve the predictive performance. Since the OP could remove some noise in the data before PLS process. Thus, the required number of LV containing the important information could be decreased.

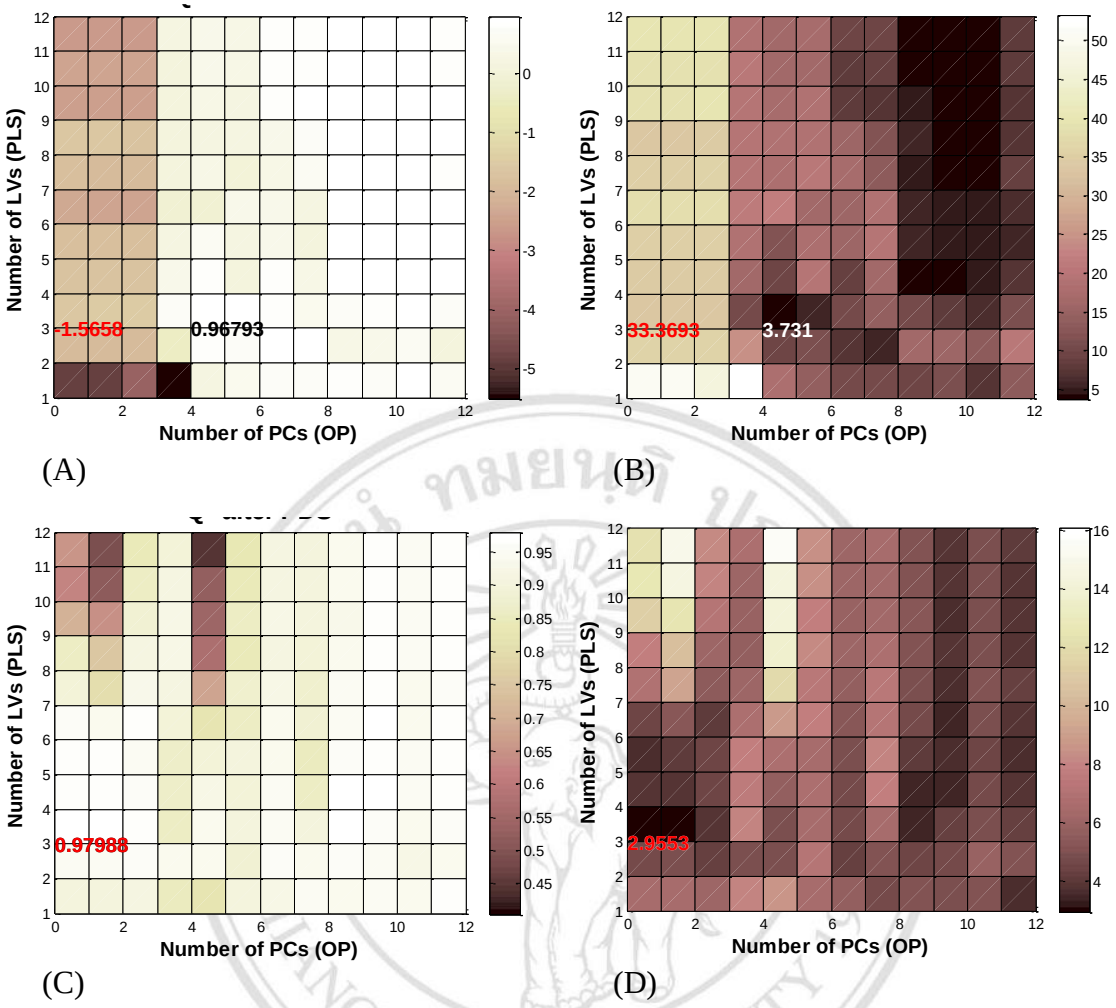


Figure 4.16 The colormaps of (A) Q^2 and (B) RMSEP values of OP-PLS, (C) Q^2 and (D) RMSEP values of OP-PDS-PLS methods with SNV + mean centring prepressing; m = 0 – 12, n = 1 – 12, w = 31

Considering the result from the combined preprocessing (SNV + mean centring) in Figure 4.16. This preprocessing shows more potential than other preprocessing in both OP-PLS and OP-PDS-PLS methods. The optimum results are at the position (4,3) with Q^2 = 0.96793 and RMSEP = 3.731. For the OP-PDS-PLS method, the optimum values showed at (0,3) with Q^2 = 0.97988 and RMSEP = 3.731. From colormap of OP, the predictive result would be improved if the selected number of PC > 3 and LV > 1. For OP-PDS results, the OP was unnecessary if the PDS was performed (m = 0).

4.4.4 The predictive performance of four methods (PLS, OP-PLS, PDS-PLS and OP-PDS-PLS) with different data preprocessing

To carefully investigate the potential of four methods, the best predictive performance of all methods with different data preprocessing were summarized in Table 4.7 and Table 4.8 in terms of removing WR and KDML105, respectively.

Removing WR variation in OP process

Considering the influence of removing WR variation in Table 4.6, the identical KDML105 models (A1-A2 and A2-A1), the results from four methods are equally good. The best method is OP-PLS with SNV + mean centring and only SNV in the models of A1-A2 and A2-A1, respectively. This ensures that using OP can remove the variation of different WR from the interested variation.

If the sample having different KDML105, A1-B1 and B1-A1, using only PDS with normalization and OP-PDS with SNV + centring can improve the prediction more than other methods. It indicates that the PDS should be performed if the samples have different KDML105.

In case of the different KDML105 and WR, A1-C3 and C3-A1, the best predictive results occurred when performed the OP-PDS-PLS with SNV and only PDS with SNV, respectively. The results obviously reveal if the samples having different KDML105 and WR, the necessary methods, OP and PDS should be carried out together to ensure that the variation used was investigated without unwanted variations. Among those methods, the OP-PDS-PLS with SNV often shows the best result in several models.

Table 4.6 The predictive performance of four methods with five preprocessing for predicting the six models (removing WR variation in OP process)

No.	Dataset	Data preprocessing	PLS		OP-PLS		PDS-PLS		OP-PDS-PLS	
			Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP
1	A1-A2 (identical KDML, different WR)	None	0.9756	3.26	0.9732	3.41	0.9390	5.15	0.9753	3.27
		SNV	0.9832	2.70	0.9852	2.54	0.9830	2.72	0.9813	2.85
		SNV + centring	0.9833	2.69	**0.9865**	**2.42**	0.9799	2.95	0.9809	2.88
		Mean centring	0.9742	3.34	0.9818	2.81	0.9534	4.5	0.972	3.49
		Normalization	0.9789	3.03	0.9729	3.43	0.9592	4.21	0.9666	3.81
2	A2-A1 (identical KDML, different WR)	None	0.9793	3.00	0.9752	3.28	0.9731	3.42	0.9838	2.65
		SNV	0.9836	2.67	**0.9907**	**2.01**	0.9845	2.59	0.9816	2.83
		SNV + centring	0.9836	2.67	0.9860	2.47	0.9837	2.66	0.9837	2.66
		Mean centring	0.9778	3.10	0.9781	3.09	0.9700	3.61	0.9784	3.06
		Normalization	0.9843	2.61	0.9784	3.06	0.9855	2.51	0.9819	2.80
3	A1-B1 (different KDML, identical WR)	None	0.0942	19.83	0.9364	5.25	0.9761	3.22	0.9750	3.29
		SNV	-0.0015	20.85	0.8167	8.92	0.9734	3.40	0.9760	3.23
		SNV + centring	-0.7346	27.44	0.9130	6.15	0.9725	3.45	0.9748	3.31
		Mean centring	-0.0743	21.59	0.9400	5.10	0.9740	3.36	0.9740	3.36
		Normalization	0.0059	20.77	0.9454	4.87	**0.9778**	**3.11**	0.9761	3.22
4	B1-A1 (different KDML, identical WR)	None	0.1495	19.21	0.9626	4.03	0.9562	4.36	0.9707	3.57
		SNV	-0.0029	20.86	0.9352	5.30	0.9813	2.85	0.9792	3.00
		SNV + centring	-0.1055	21.90	0.9429	4.98	0.9765	3.20	**0.9833**	**2.69**
		Mean centring	-0.0743	21.59	0.9627	4.03	0.9740	3.36	0.9740	3.36
		Normalization	0.0043	20.79	0.9664	3.82	0.9789	3.02	0.9740	3.36
5	A1-C3 (different KDML and WR)	None	-0.0176	21.01	0.9629	4.01	0.9016	6.54	0.9703	3.59
		SNV	-0.0020	20.85	0.9181	5.96	0.9817	2.82	**0.9831**	**2.71**
		SNV + centring	-1.5658	33.37	0.9679	3.73	0.9799	2.96	0.9798	2.96
		Mean centring	-1.0436	29.78	0.9499	4.66	0.7442	10.54	0.9719	3.50
		Normalization	-0.0022	20.86	0.9770	3.16	0.9820	2.79	0.9695	3.64
6	C3-A1 (different KDML and WR)	None	0.9235	5.76	0.9337	5.36	0.8564	7.90	0.9749	3.30
		SNV	0.9409	5.07	0.9484	4.73	**0.9887**	**2.21**	0.9881	2.27
		SNV + centring	0.9399	5.11	0.9408	5.07	0.9791	3.01	0.9850	2.55
		Mean centring	0.9018	6.53	0.9292	5.54	0.8198	8.84	0.9694	3.64
		Normalization	0.9307	5.48	0.9496	4.68	0.9700	3.61	0.9816	2.83

Removing KDML105 variation in OP process

Considering the effect of removing KDML105 variation on the prediction performances in Table 4.7, the model of A1-A2, identical KDML105 and different WR, obtained the best prediction with only PLS method with SNV + centring. While the OP-PLS method with SNV provided the best predictive result in the model of A2-A1. Regarding the predictive performance, it was obviously that using only PLS can cope with the influence of different WR variation presented with satisfactory prediction in all preprocessing.

In case of the data having different KDML105 and identical WR, A1-B1 and B1-A1, the best prediction occurred with the methods of PDS and OP-PDS, respectively. It was the same result with removing WR, if the adulterated sample having different type of KDML105, PDS is the best alternative method for data pretreatment before calibration analysis.

For the sample having both different of KDML105 and WR, the models of A1-C3 and C3-A1, the best method for prediction of adulterated level is OP-PDS-PLS with SNV. Since the influent variation based on the different of KDML105 and WR were removed, this meant that two datasets can be calibrated in the same model effectively.

From all results in Table 4.6 and 4.7, the best predictive performances are always appeared in the data preprocessing of SNV and the combined of SNV and mean centring. The results indicate that the spectroscopic data should be pretreated with SNV and mean centring to gain the correct prediction in this research. The effect of four studied methods on six datasets will be revealed and discussed in section 4.3.5.

Table 4.7 The predictive performance of four methods with five preprocessing for predicting the six models (removing KDML105 variation in OP process)

No.	Dataset	Data preprocessing	PLS		OP-PLS		PDS-PLS		OP-PDS-PLS	
			Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP
1	A1-A2 (identical KDML, different WR)	None	0.9737	3.38	0.9815	2.83	0.9331	5.39	0.9640	3.95
		SNV	0.9832	2.70	0.9832	2.70	0.9828	2.73	0.9809	2.88
		SNV + centring	**0.9833**	**2.69**	0.9832	2.70	0.9799	2.95	0.9810	2.87
		Mean centring	0.9742	3.34	0.9798	2.96	0.9534	4.50	0.9787	3.04
		Normalization	0.9789	3.02	0.9812	2.85	0.9660	3.84	0.9784	3.06
2	A2-A1 (identical KDML, different WR)	None	0.9808	2.88	0.9865	2.42	0.9735	3.39	0.9736	3.38
		SNV	0.9836	2.67	**0.9872**	**2.35**	0.9845	2.59	0.9819	2.80
		SNV + centring	0.9836	2.67	0.9835	2.67	0.9837	2.66	0.9841	2.63
		Mean centring	0.9778	3.10	0.9844	2.61	0.9700	3.61	0.9630	4.01
		Normalization	0.9843	2.61	0.9770	3.16	0.9737	3.38	0.9674	3.76
3	A1-B1 (different KDML, identical WR)	None	-0.0391	21.24	0.9487	4.72	0.9733	3.40	0.9725	3.46
		SNV	-0.0023	20.86	0.7613	10.18	0.9736	3.38	0.9758	3.24
		SNV + centring	-0.7346	27.44	0.9284	5.57	0.9725	3.45	0.9754	3.27
		Mean centring	-0.0743	21.59	0.9029	6.49	0.9740	3.36	0.9487	4.72
		Normalization	-0.0051	20.89	0.9468	4.81	**0.9777**	**3.11**	0.9687	3.69
4	B1-A1 (different KDML, identical WR)	None	-0.0614	21.46	0.9548	4.43	0.9529	4.52	0.9633	3.99
		SNV	-0.0018	20.85	0.9600	4.17	0.9809	2.88	0.9771	3.15
		SNV + centring	-0.1055	21.90	0.9800	2.95	0.9765	3.20	**0.9882**	**2.26**
		Mean centring	0.4089	16.02	0.9571	4.32	0.9317	5.44	0.9631	4.00
		Normalization	-0.0044	20.88	0.9659	3.84	0.9561	4.36	0.9623	4.05
5	A1-C3 (different KDML and WR)	None	0.0008	20.82	0.9629	4.01	0.8889	6.94	0.9703	3.59
		SNV	-0.0022	20.86	0.7799	9.77	0.9820	2.79	**0.9833**	**2.69**
		SNV + centring	-1.5658	33.37	0.8932	6.81	0.9799	2.96	0.9792	3.01
		Mean centring	-1.0436	29.78	0.9302	5.50	0.7442	10.54	0.9575	4.29
		Normalization	-0.0122	20.96	0.8511	8.04	0.8636	7.69	0.9701	3.60
6	C3-A1 (different KDML and WR)	None	0.8981	6.65	0.9337	5.36	0.8675	7.58	0.9749	3.30
		SNV	0.9409	5.07	0.9409	5.06	0.9884	2.24	**0.9889**	**2.19**
		SNV + centring	0.9399	5.11	0.9428	4.98	0.9791	3.01	0.9851	2.55
		Mean centring	0.9018	6.53	0.8606	7.78	0.8198	8.84	0.9724	3.46
		Normalization	0.9308	5.48	0.9126	6.16	0.6965	11.48	0.9684	3.71

4.4.5 Comparison of four methodologies of six calibration models

The effect of four methods on six models with removing WR and KDML105 were investigated and demonstrated in Figure 4.17 and 4.18, respectively. In one bar chart, there were four groups of method used including (1) PLS, (2) OP-PLS, (3) PDS-PLS and (4) OP-PDS-PLS represented in x-axis. The y-axis represents RMSEP values. Each bar represents the RMSEP values of each calibration models. The relation between each method and each model could be estimated when considering these charts. Figure 4.17(A), 4.17(B), 4.17(C), 4.17(D) and 4.17(E) display none preprocessing, SNV, SNV + centring, mean centring and normalization, respectively. Overall, using ordinary PLS, the first group of bar charts, did not cope with all types of adulterated rice sample. Using only PLS provided RMSEP values over 20 in the data having different KDML105 and WR types. It meant that the PLS could provide good result in only the data having the same KDML105 type. After OP process, the second group of bar charts, the predictive results of all models clearly improved in all data preprocessing. The RMSEP values decreased from 20 to approximately 5. The results of OP-PLS show that, using OP process for removing the influence of WR variations can enhance the prediction in the models of A1-B1, B1-A1 and A1-C3. Likewise, the predictive ability of prediction model after PDS, the third groups of bar charts, are better than using only PLS. The errors of prediction dropped to just under 5. However, those good results did not appear in all data preprocessing. The RMSEP values were higher than 5 in none preprocessing and mean centring in Figure 4.17(A) and 4.17(D), respectively. The unsatisfied results obtained from dataset having different types of both KDML105 and WR (A1-C3 and C3-A1 models). Considering the fourth group represented the OP-PDS-PLS, it was surprised that the RMSEP values decreased to lower 5 in all calibration models and all data preprocessing.

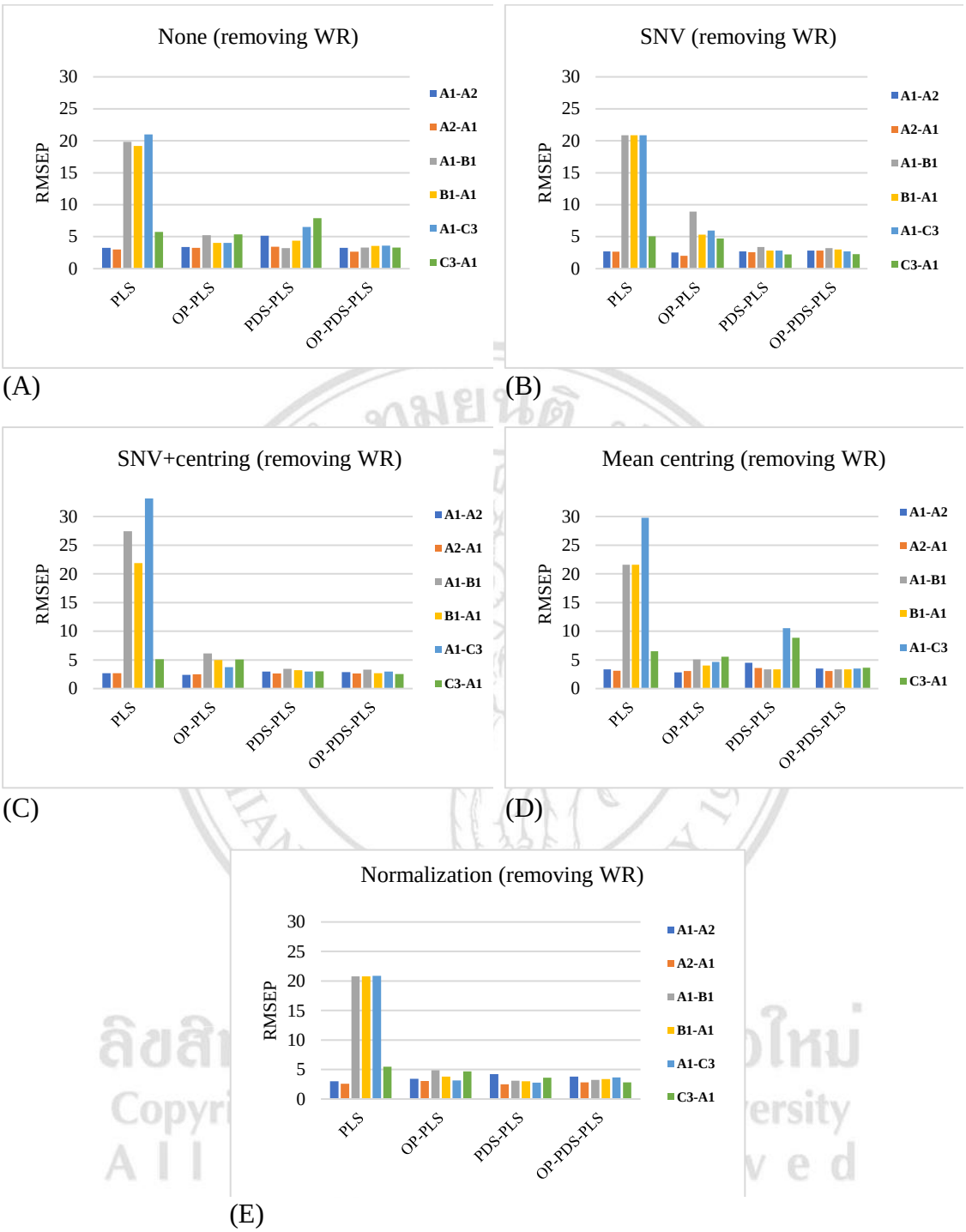


Figure 4.17 Comparison of four methods and six models removed WR variation with different pretreatments. The predictive results represented in RMSEP of (A) none data preprocessing, (B) SNV, (C) SNV+ mean centring, (D) mean centring and (E) normalization

To consider the effect of removing KDML105 variation, the RMSEP values of different data preprocessing were shown in Figure 4.18(A) to Figure 4.18(E). This result indicated that using only PLS did not good enough for predicting an amount of KDML105 (%w/w) in all mixed rice samples. The errors were high around 20-30 %. Using OP, the model contained only WR variation, could improve the prediction result in almost all models except the model having different types of KDML105 which preprocessed with SNV in Figure 4.18(B). In case of the model having the different KDML105, the PDS could play an important role to enhance the predictive result more than OP in the preprocessing of SNV and SNV + centring (Figure 4.18(B) and Figure 4.18(C). However, using only PDS did not deal with the samples having different KDML105 and WR when the none preprocessing, mean centring and normalization were used (Figure 4.18(A), Figure 4.18(D) and Figure 4.18(E)). In addition, using OP and PDS separately could not be gained the best result with the samples have both different types of KDML105 and WR. To enhance the predictive performance, the method using OP cooperate with PDS is performed to ensure that the prediction model in this research can cope with all types of adulterated rice sample.

In summary, using ordinary PLS could not effective with the adulterated sample has different KDML105. OP process, the method aimed to get rid of unwanted variations, could improve the predictive results in case of the data having different KDML. In this case, removing WR could decrease the RMSEP value better than removing KDML105. Remarkably, using OP coupled with PDS, removed unwanted variation and standardization, could overcome other methods. This method provided the best predictive performance not only in all types of sample but also all data preprocessing.

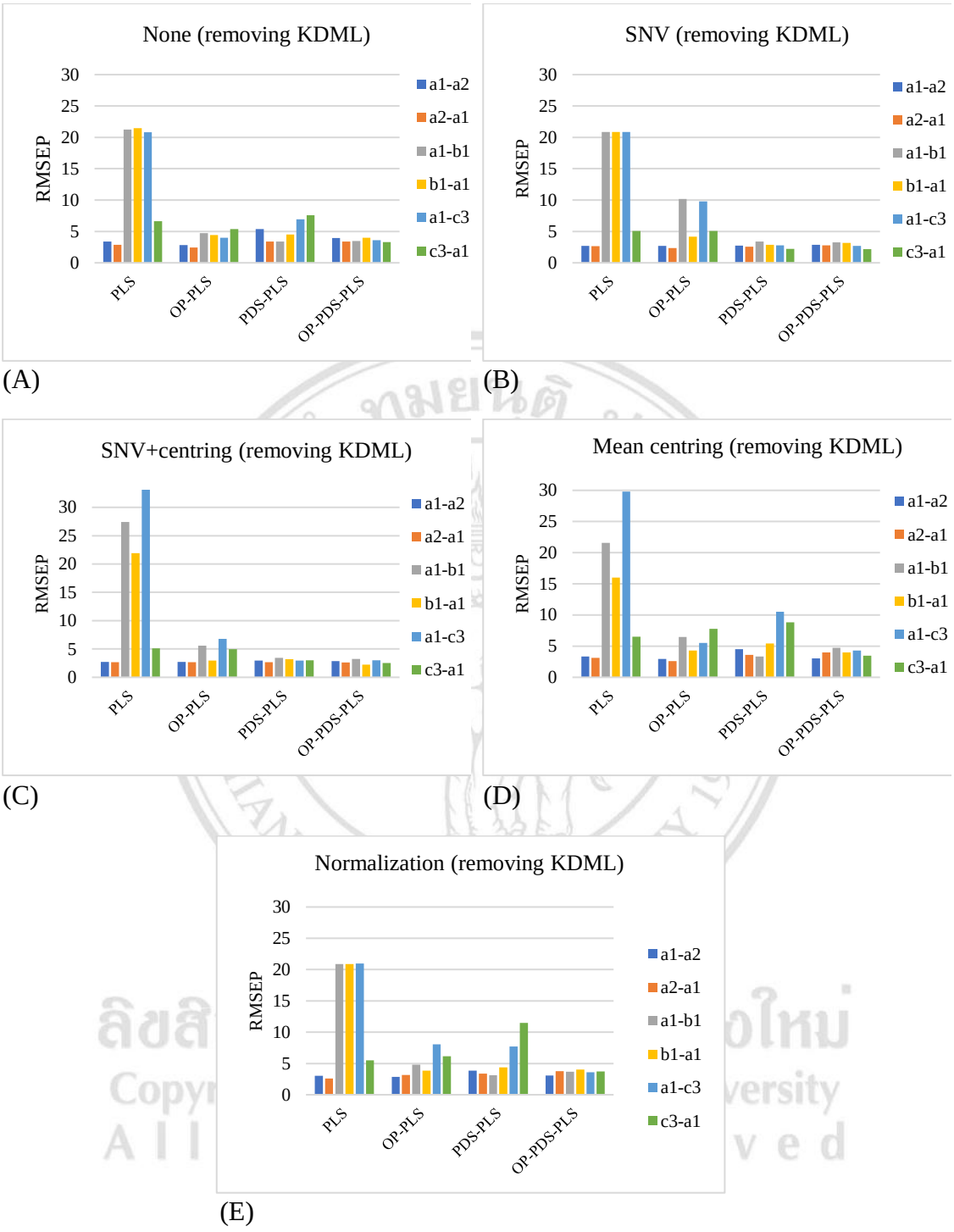


Figure 4.18 Comparison of four methods and six models removed KDML105 variation with different pretreatments. The predictive results represented in RMSEP of (A) none data preprocessing, (B) SNV, (C) SNV+ mean centring, (D) mean centring and (E) normalization

CHAPTER 5

Conclusions and suggestions

5.1 Conclusions

NIR is a potential technique that can be used to identify the adulteration in KDML105 rice grain. This technique offers a straightforward, non-destructive, rapid and cost-effective detection. It also allows non-complicated measurements without or less sample preparation. This research utilized some chemometrics to enhance the predictive performance as follows:

5.1.1 Data exploratory of rice samples

PCA could be successfully used to capture the variations in the pure KDML105, PT1 and WR rice. In addition, the difference in the mixture samples could be observed. Based on the PCA scores plot, the cluster of each dataset tended to separate into groups according to the types of the rice samples. In this research, SNV and mean centring data preprocessing were recommended since the variation was not too much relied on the first PC.

5.1.2 Untargeted classifications

To identify of the KDML105 rice has been adulterated, Q statistics could result in better classification rates than D statistics. This could be that the variation part of the KDML105 dataset could be well related to that of the mixture samples. Therefore, the summation of errors or non-systematic variation was high and could be captured by the Q statistics. When the decisions of both D and Q statistics were combined using SIMCA, the model could be effectively used to identify the adulterated samples at the accuracy and precision of 95.83% and 90.70%, respectively.

5.1.3 Quantification of adulteration level using OP-PDS-PLS

The quantification of adulteration in KDML105 rice was carried out using PLS with different modification including PLS, OP-PLS, PDS-PLS and OP-PDS-PLS. Three different data preprocessing including SNV, mean centring and normalization were also investigated. The samples used in this research consisted of (1) the samples having identical KDML105 with different WR, (2) the samples with different KDML105 mixed with an identical WR and (3) the samples having both different KDML105 and WR.

The results showed that using only PLS could not cope well with the datasets with different KDML105 and different WR. The OP-PLS could reduce the predictive errors by removing the non-related variation in the data such as the noise from different conditions and the effect of interferences. On the other hand, PDS-PLS could be used to standardize the model from different condition during experiment that the main problem in NIR measurement. OP-PDS-PLS could stabilize the predictive performance of the predication model especially when the different data preprocessing were used.

5.2 Suggestions

5.2.1 Normally, the multivariate data analysis attempts to investigate several variables or parameters to obtain the best result. However, in many cases, some parameters are not useful and cannot be related to the response and these useless parameters should be discarded from the model. Therefore, the raw NIR data should be identify the important parameters using variable selection such as variable importance on projection (VIP) and selectivity ratio (SR) in the future work.

5.2.2 The analysis of some other physical and chemical parameters of adulterated rice should be investigated to improve the prediction model.

5.3.3 Some other orthogonal projections could be investigated to improve the predictive performance in future task.

REFERENCES

[1] Ministry of Commerce. Annulment of Notification of Ministry of Commerce, Subject: Prescribing White Rice as a Standardised Commodity and Standards of White Rice B.E. 2555. Department of Foreign Trade. (2016). Available at: http://www.dft.go.th/th-th/DetailHotNews/ArticleId/8105/NOTIFICATION-OF-MINISTRY-OF-COMMERCE-Standard-Thailand-Rice. Accessed June 28, 2017.

[2] Vlachos, A. & Arvanitoyannis, I.S. (2008). A Review of rice authenticity/adulteration methods and results. *Food Science and Nutrition*, 48:553-598.

[3] Larrechi, M.S. & Callao, M.P. (2003). Strategy for introducing NIR spectroscopy and multivariate calibration techniques in industry. *Trends in Analytical Chemistry*, 22:634-640.

[4] Xu, L., Yan, S.M., Cai, C.B. & Yu, X.P. (2013). Untargeted detection of illegal adulterations in Chinese glutinous rice flour (GRF) by NIR spectroscopy and chemometrics: specificity of detection improved by reducing unnecessary variations. *Food Analytical Methods*, 6:1568-1575.

[5] Zhu, L., Sun, J., Wu, G., Wang, Y., Zhang, H., Wang, L., Qian, H. & Qi, X. (2018). Identification of rice varieties and determination of their geographical originin China using Raman spectroscopy. *Journal of Cereal Science*, 82:175-182.

[6] Golam, F., Yin, Y.H., Masitah, A., Afnierna, N., Majid, N.A., Khalid, N. & Osman, M. (2011). Analysis of aroma and yield components of aromatic rice in malaysian tropical environment. *Australian Journal of Crop Science*, 5:1318-1325.

[7] Timsorn, K., Lorjaroenphon, Y. & Wongchoosuk, C. (2017). Identification of adulteration in uncooked Jasmine rice by a portable low-cost artificial olfactory system. *Measurement*, 108:67-76.

[8] Bilge, G., Sezer, B., Eseller, K.E., Berberoglu, H., Topcu, A. & Boyaci, I.H. (2016). Determination of whey adulteration in milk powder by using laser induced breakdown spectroscopy. *Food Chemistry*, 212:183-188.

[9] Guler, A., Kocaokutgen, H., Garipoglu, A.V., Onder, H., Ekinci D. & Biyik, S. (2014). Detection of adulterated honey produced by honeybee (*Apis mellifera* L.) colonies fed with different levels of commercial industrial sugar (C_3 and C_4 Plants) syrups by the carbon isotope ratio analysis. *Food Chemistry*, 2014 155:155-160.

[10] Zhang, L., Li, P., Sun, X., Wang, X., Xu, B., Wang, X., Ma, F., Zhang, Q. & Ding, X. (2014). Classification and adulteration detection of vegetable oils based on fatty acid profiles. *Journal of Agricultural and Food Chemistry*, 2014 62:8745-8751.

[11] Vemireddy, L.R., Satyavathi, V.V., Siddiq, E.A. & Nagaraju, J. (2015). Review of methods for the detection and quantification of adulteration of rice: Basmati as a case study. *Journal of Food Science and Technology*, 52:3187-3202.

[12] Ganopoulos, I., Argiriou, A. & Tsaftaris, A. (2011). Adulteration in basmati rice detected quantitatively by combined use of microsatellite and fragrance typing with high resolution melting (HRM) analysis. *Food Chemistry*, 129:652-659.

[13] Pitiphunpong, S., Champangern, S. & Suwannaporn, P. (2011). The Jasmine rice (KDML 105 Variety) adulteration detection using physico-chemical properties. *Chiang Mai Journal of Science*, 38:105-115.

[14] Thind, G.K. & Sogi, D.S. (2005). Identification of coarse (IR-8), fine (PR-106) and superfine (Basmati-386) rice cultivars. *Food Chemistry*, 91:227–233.

[15] Cameron, D.K., Wang, Y. J. & Moldenhauer, K.A. (2008). Comparison of physical and chemical properties of medium-grain rice cultivars grown in California and Arkansas. *Journal of Food Science*, 73:72-78.

[16] Larrechi, M.S. & Callao, M.P. (2003). Strategy for introducing NIR spectroscopy and multivariate calibration techniques in industry. *Trends in Analytical Chemistry*, 22:634-640.

[17] Fischer, D., Sahre, K., Abdelrhim, M., Voit, B., Sadhu, V.B., Pionteck, J., Komber, H. & Hutschenreuter, J. (2016). Process monitoring of polymers by in-line ATR-IR, NIR and raman spectroscopy and ultrasonic measurements. *Comptes Rendus Chimie*, 9:1419-1424.

[18] Wu, J.G. & Shi, C.H. (2006). Calibration model optimization for rice cooking characteristics by near infrared reflectance spectroscopy (NIRS). *Food Chemistry*, 103:1054-1061.

[19] Brereton, R.G. (2003). Chemometrics: Data Analysis for the Laboratory and Chemical Plant. Wiley: Chichester, U.K.

[20] Krongchai, C., Funsueb, S., Jakmunee, J. & Kittiwachana, S. (2016). Application of multiple self organizing maps for classification of soil Samples in thailand according to their geographic origins. *Journal of Chemometrics*, 31:1-10.

[21] Brereton, R.G. (2011). One-class classification. *Journal of Chemometrics*, 25:225-286.

[22] Brereton, R.G. & Lloyd, G.R. (2010). Support vector machines for classification and regression. *Analyst*, 135:230-267.

[23] Kittiwachana, S., Ferreira, D.L.S., Lloyd, G.R., Fido, L.A., Thompson, D.R., Escott, R.E.A. & Brereton, R.G. (2010). One class classifiers for process monitoring illustrated by the application to on-line HPLC of a continuous process. *Journal of Chemometrics*, 24:96-110.

[24] Wold, S. & Sjöström, M. (1997). SIMCA: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109-130.

[25] Zábrodská, B. & Vorlová, L. (2015). Adulteration of honey and available methods for detection – a review. *Acta Veterinaria Brno*, 83:85-102.

[26] Shao, X., Bian, X., Liu, J., Zhang, M. & Cai, W. (2010). Multivariate calibration methods in near infrared spectroscopic analysis. *Analytical Methods*, 2:1662-1666.

[27] Pinto, R.C., Trygg, J. & Gottfries, J. (2012). Advantages of orthogonal inspection in chemometrics. *Journal of Chemometrics*, 26: 231–235.

[28] Hansen, P.W. (2001). Pre-processing method minimizing the need for reference analyses. *Journal of Chemometrics*,15: 123–131.

[29] Griffiths, M.L., Svozil, D., Worsfold, P. & Evans, E.H. (2006). The application of piecewise direct standardization with variable selection to the correction of drift in inductively coupled atomic emission spectrometry. *Journal of Analytical Atomic Spectrometry*, 21:1045-1052.

[30] Lianga, C, Yuan, H.F., Zhao, Z., Song, C.F. & Wang, J.J. (2016). A new multivariate calibration model transfer method of near-infrared spectral analysis. *Chemometrics and Intelligent Laboratory Systems*, 153:51-57.

[31] Thai Hom Mali rice. Wonder of Thai Hom Mali rice. Available at: http://www.thai-hommalirice.com/ewt_news.php?nid=1. Accessed July 4, 2018.

[32] Khonkaen rice seed center. RD15. Avilable at: http://kkn-rsc.ricethailand.go.th/index.php/e-library/varieties/434-rd15. Accessed May 10, 2018.

[33] Wangcharoen, W., Phanchaisri, C., Daengpok, W., Phuttawong, R., Hangsoongnern, T. & Phanchaisri, B. (2016). Consumer acceptance test and some related properties of selected KDML 105 rice mutants. *Journal of Food Science and Technology*, 53:3550-3556.

[34] Jalili, M. (2017). A review paper on melamine in milk and dairy products. *Journal of Dairy and Veterinary Sciences*, 4:1-3.

[35] Davies, A.M.C. (2005). An introduction to near infrared spectroscopy. NIR news, 7. Available at: http:// www.impublications.com/content/introduction-near-infrared-nir-spectroscopy. Accessed July 10, 2019.

[36] Ozaki, Y., (2012). Near-infrared spectroscopy-its versatility in analytical chemistry. *Analitical sciences*, 28:545-563.

[37] Osborne, B.G. (2006). Near-infrared spectroscopy in food analysis. *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Ltd.

[38] Pasquini, C. (2003). Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, 14:198-219.

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

[40] Kawata, S. 2002. New techniques in near-infrared spectroscopy. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.). *Near Infrared Spectroscopy: Principles, Instruments, Applications*. Wiley-VCH Verlag GmbH, Weinheim, pp. 75-84.

[41] Manley, M. & Baeten, V. (2018). Spectroscopic technique: near infrared (NIR) spectroscopy. In: Sun, D.W (Ed.). *Modern techniques for food authentication*. Elsevier, pp. 51-87.

[42] Osborne, B.G. (2000). Near-infrared spectroscopy in food analysis. In: Meyers, R.A. (Ed.). *Encyclopedia of Analytical Chemistry*. John Wiley & Sons Ltd, pp. 1-13.

[43] Cortés, V., Blasco, J., Aleixos, N., Cubero, S. & Talens, T. (2019). Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review. *Trends in Food Science & Technology*, 85:138-148.

[44] Penner, M.H. (1994). Basic principles of spectroscopy. In: Nielsen, S.S. (Ed.), Introduction to the Chemical Analysis of Foods. Jones and Bartlett Publishers, Boston, MA, pp. 317-324.

[45] Geladi, P., MacDougall, D. & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39:491-500.

[46] Osborne, B.G., Fearn, T. & Hindle, P.H. (1993). Practical NIR spectroscopy with application in food and beverage analysis. Addison-Wesley Longman Ltd: Harlow UK., Singapore.

[47] Blanco, M. & Villarroya, I. (2002). NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry*, 21, 240-250.

[48] Kumar, N., Bansal, A., Sarma, G.S. & Rawal, R.K. (2014). Chemometrics tools used in analytical chemistry: an overview. *Talanta*, 123:186-199.

[49] Roberts, J.J. & Cozzolino, D. (2016). An overview on the application of chemometrics in food science and technology—an approach to quantitative data analysis. *Food Analytical Methods*, 9:3258-3267.

[50] Biancolillo, A. & Marini, F. (2018). Chemometric methods for spectroscopy-based pharmaceutical analysis. *Frontiers in Chemistry*, 6:1-14.

[51] Theanjumpol, P., Wongzeewasakun, K., Muenmanee, N., Wongsaipun, S., Krongchai, C., Changrue, V., Boonyakiat, D. & Kittiwachana, S., (2019). Non-destructive identification and estimation of granulation in 'Sai Num Pung' tangerine fruit using near infrared spectroscopy and chemometrics. *Postharvest Biology and Technology*, 153:13-20.

[52] Xu, L., Shi, W., Cai, C.B., Zhong, W. & Tu, K. (2015). Rapid and nondestructive detection of multiple adulterants in kudzu starch by near infrared (NIR) spectroscopy and chemometrics. *LWT - Food Science and Technology*, 61:590-595.

[53] Lloyd, G.R., Ahmad, S., Wasim, M. & Brereton, R.G. (2009). Pattern recognition of inductively coupled plasma atomic emission spectroscopy of human scalp hair for discriminating between healthy and hepatitis C patients. *Analytica chimica acta*, 649:33-42.

[54] Gonzaga, F.B., Rocha, W.F.d.C. & Correa, D.N. (2015). Discrimination between authentic and false tax stamps from liquor bottles using laser-induced breakdown spectroscopy and chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 109: 24-30.

[55] Baranda, A.B., Etxebarria, N., Jimenez, R.M. & Alonso, R.M. (2005). Development of a liquid–liquid extraction procedure for five1,4-dihydropyridines calcium channel antagonists from human plasma using experimental design. *Talanta*, 67: 933-941.

[56] Westerlund E, Andersson R, Hämäläinen M et al. Principal component analysis - an efficient tool for selection of wheat samples with wide variation in properties. Journal of Cereal Science 1991; 14: 95-104

[57] Feng, Z. & Xu, T. (2011). Comparison of SOM and PCA-SOM in fault diagnosis of rround-testing bed. *Procedia Engineering*, 15:1271-1276.

[58] Ordoñez, J.C.B., Valdés, R.M.A. & Comendador, F.G. (2018). Energy efficient GNSS signal acquisition using singular balue decomposition (SVD). *Sensors*, 18: 1-27.

[59] Wold, S., Albano, C., Dunn, W.J., Esbensen, K., Hellberg, S., Johansson, E. & Sjöström, M. (1983). Pattern recognition: finding and using regularities in multivariate data food research, how to relate sets of measurements or observations to each other. *Food Research and Data Analysis*, 3:183-185.

[60] Wehrens, R., Putter, H. & Buydens, L.M.C. (2000). The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 54:35-52.

[61] Arlot, S. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40-79.

[62] Bartholomew, D.J. (2010). Principal components analysis. *The London School of Economics and Political Science*, London, U.K.

[63] Pizarro, C., González-Sáiz, J.M., Esteban-Díez, I., Rodríguez-Tecedor, S., Pérez-del-Notario, N. & Sáenz-González, C. (2009). Classification of archaeological sherds across the southeast united states based on variable selection from compositional fingerprints. *Analytica Chimica Acta*, 646:69-77.

[64] Brereton, R.G. (2011). One-class classifiers. *Journal of Chemometrics*, 25:225-246.

[65] Brereton, B.G. (2009). Chemometrics for pattern reccognition. Wiley:Chichester, U.K.

[66] Wold, S., Esbensen, K. & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:37-52.

[67] Kourti, T. & Macgregor, J.F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28:3-21.

[68] Nomikos, P. & Macgregor, J.F. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37:41-59.

[69] Wold, S. (1976). Pattern-recognition by means of disjoint principal components Models. *Pattern Recognition*, 3:127-139.

[70] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of Inequality of Variance in the one-way classification. *The Annals of Mathematical Statistics*, 25:290-302.

[71] Qin, S.J. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, 17:480-502.

[72] Pieszczek, L., Czarnik-Matusewicz, H. & Daszykowski, M. (2018). Identification of ground meat species using near-infrared spectroscopy and class modeling techniques - aspects of optimization and validation using a one-class classification model. *Meat Science*, 139:15-24.

[73] Messai, H., Farman, M., Sarraj-Laabidi, A., Hammami-Semmar, A. & Semmar, N. (2016). Chemometrics methods for specificity, authenticity and traceability analysis of olive oils: principles, classifications and applications. *Foods*, 5:77-113.

[74] Karunathilaka, S.R., Farris, F., Mossoba, M.M., Moore, J.C. & Yakes, B.J. (2017). Non-targeted detection of milk powder adulteration using Raman spectroscopy and chemometrics: melamine case study. *Journal Food Additives & Contaminants: Part A*, 34, 170-182.

[75] Hansen, P.W. (2001). Pre-processing method minimizing the need for reference analyses. *Journal of Chemometrics*, 15:123-131.

[76] Griffiths, M.L., Svozil, D., Worsfold, P. & Evans, E.H. (2006). The application of piecewise direct standardization with variable selection to the correction of drift in inductively coupled atomic emission spectrometry. *Journal of Analytical Atomic Spectrometry*, 21:1045-1052.

[77] Boulet, J.C. & Roger, J.C. (2012). Pretreatments by means of orthogonal projections. *Chemometrics and Intelligent Laboratory Systems*, 117:61-69.

[78] Wold, S., Antti, H., Lindgren, F. & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44:75-185.

[79] Andersson, C.A. (1999). Direct orthogonalization. *Chemometrics and Intelligent Laboratory Systems*, 47:51-63.

[80] Lianga, C., Yuan, H.F., Zhao, Z., Song, C.F. & Wang, J.J. (2016). A new multivariate calibration model transfer method of near-infrared spectral analysis. *Chemometrics and Intelligent Laboratory Systems*, 153:51-57.

[81] Wang, Z., Dean, T. & Kowalski, B.R. (1995). Additive background correction in multivariate instrument standardization. *Analytical Chemistry*, 67:2379-2385.

[82] Bouveresse, E. & Massart, D.L. (1996). Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 32:201-213.

[83] Funsueb, S., Krongchai, C., Mahatheeranont, S. & Kittiwachana. S. (2016). Prediction of 2-Acetyl-1-Pyrroline content in grains of Thai Jasmine rice based on panting condition, plant rrowth and yield component data using chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 156:203-210.

[84] Wongsaipun, S., Krongchai, C., Jakmunee, J. & Kittiwachana, S. (2018). Rice grain freshness measurement using rapid visco analyzer and chemometrics. *Food Analytical Methods*, 11:613–623.

[85] Geladi, P. & Kowalski, B. (1986). Partial least squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1-17.

[86] Farrés, M., Platikanov, S., Tsakovski, S. & Tauler, R. (2015). Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, 29:528-536.

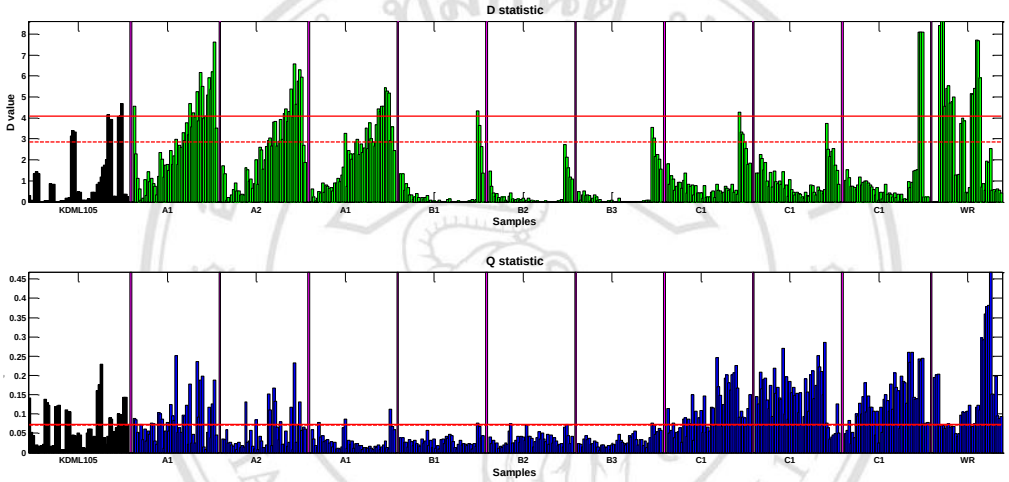
[87] Tran, T.N., Afanador, N.L., Buydens, L. & Blanchet, L. (2014). Interpretation of variable importance in Partial Least Squares with significance multivariate correlation (sMC). *Chemometrics and Intelligent Laboratory Systems*, 138:153-160.

APPENDIX I

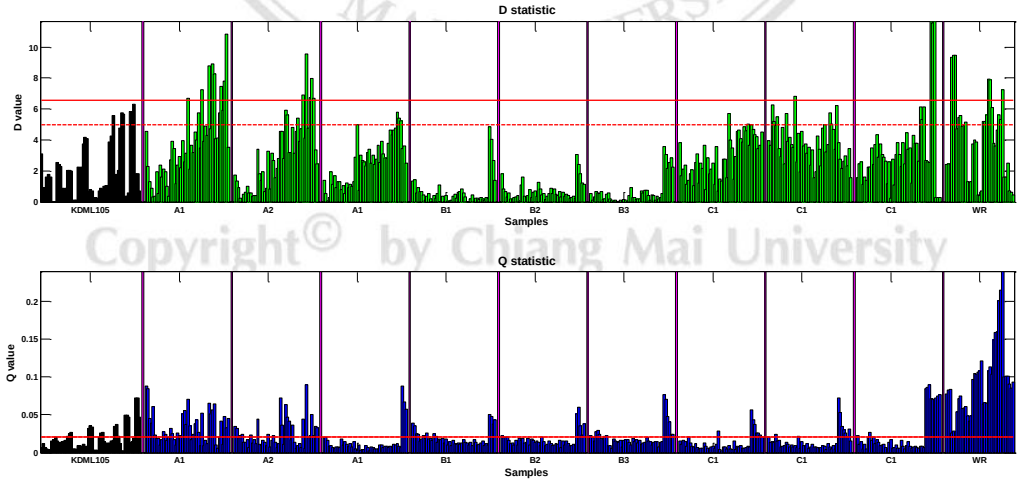
Supporting results

D and Q statictics of KDML105 in optimum number of PCs (1 – 8 PCs)

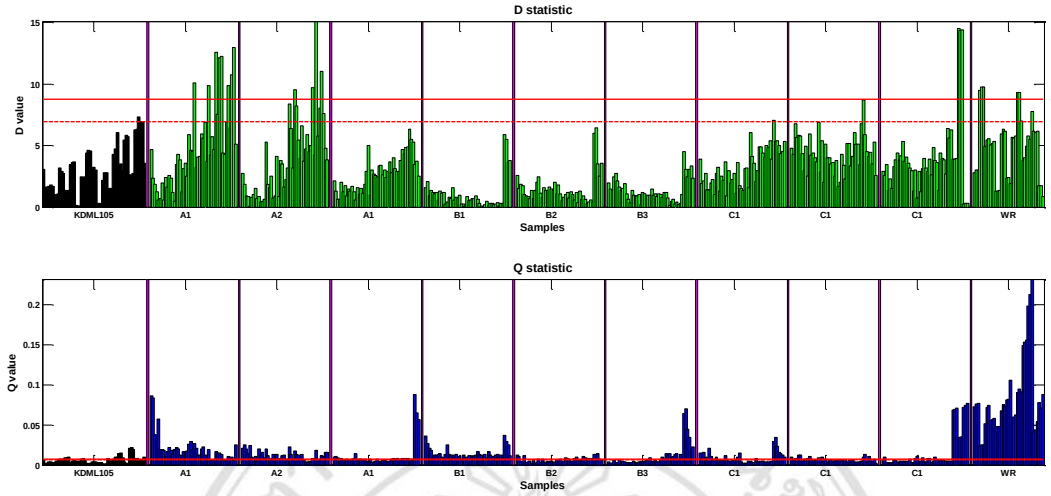
1 PC, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)



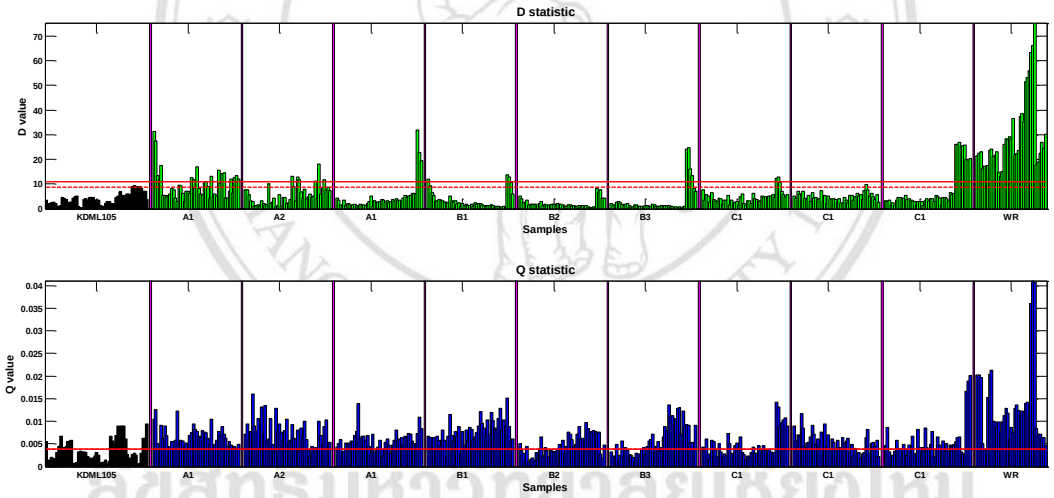
2 PCs, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)



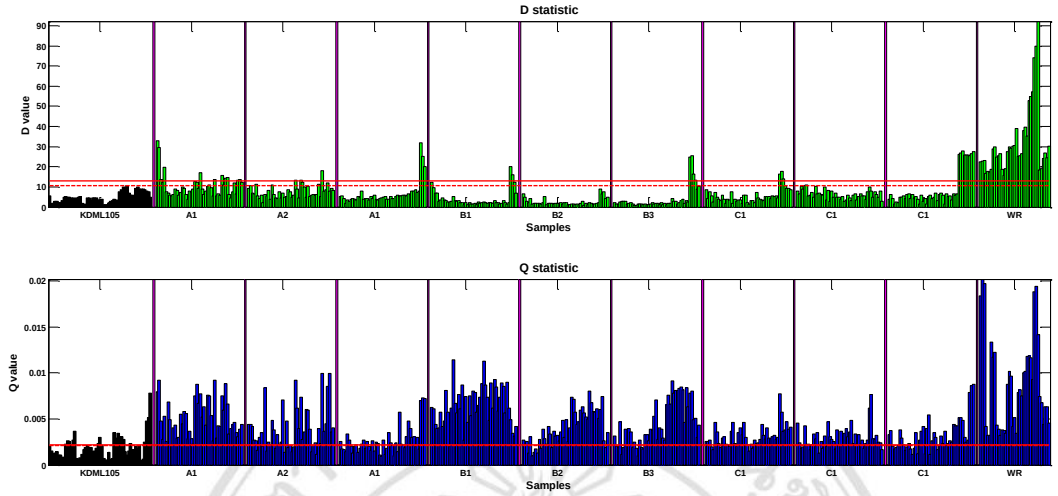
3 PC, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)



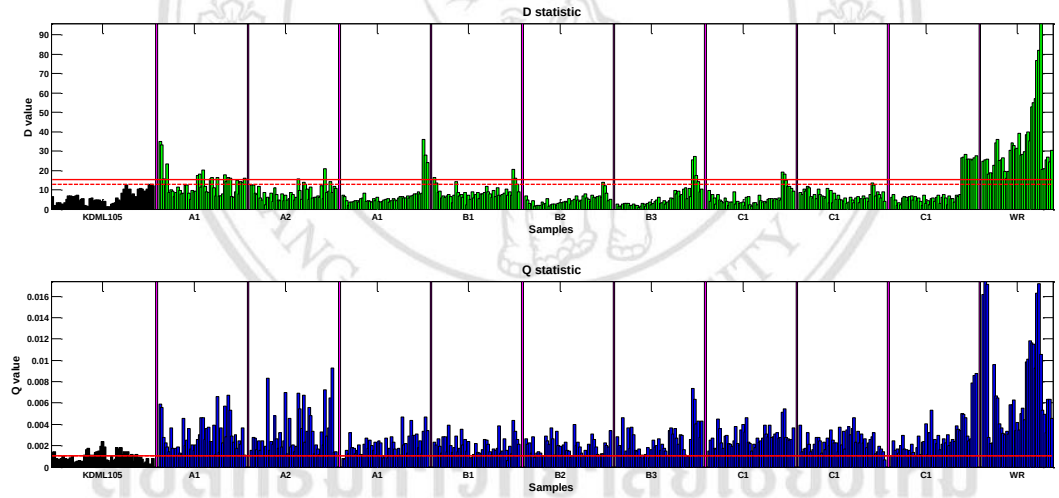
4 PCs, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)



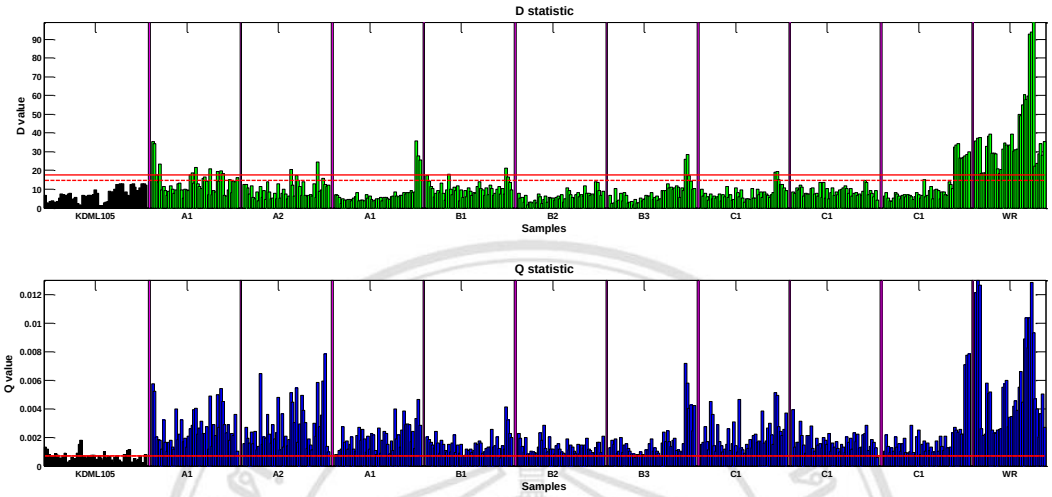
5 PC, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)



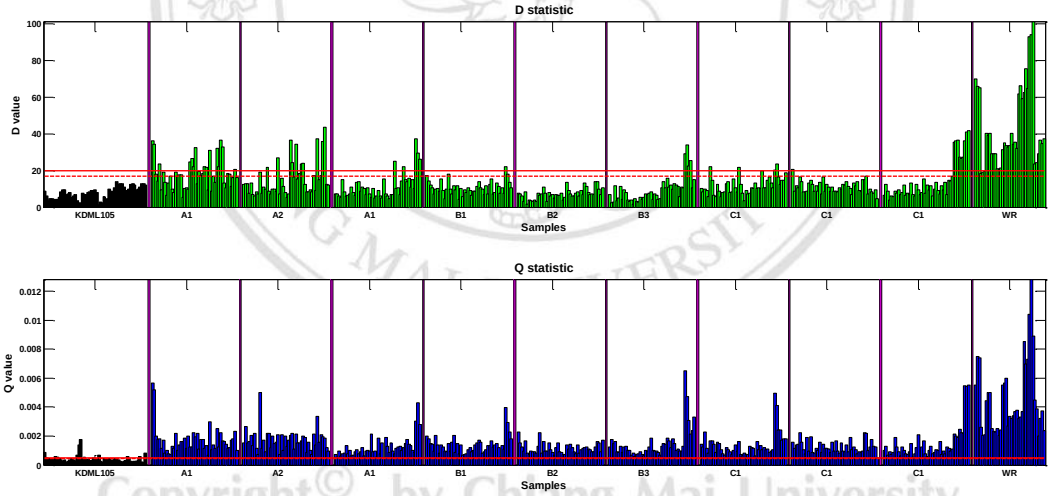
6 PCs, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)



7 PC, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)

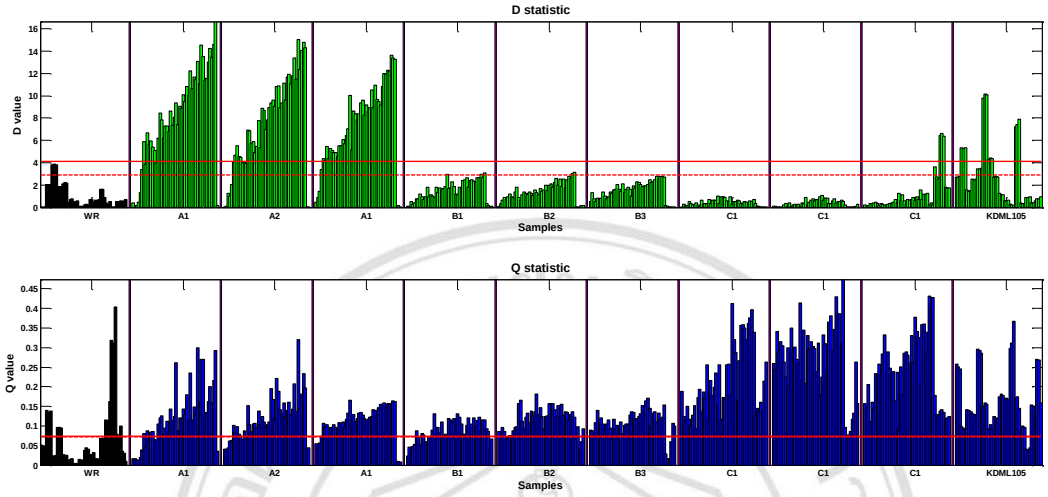


8 PCs, KDML105 model, SNV + mean centring at 90% (--) and 95% (-)

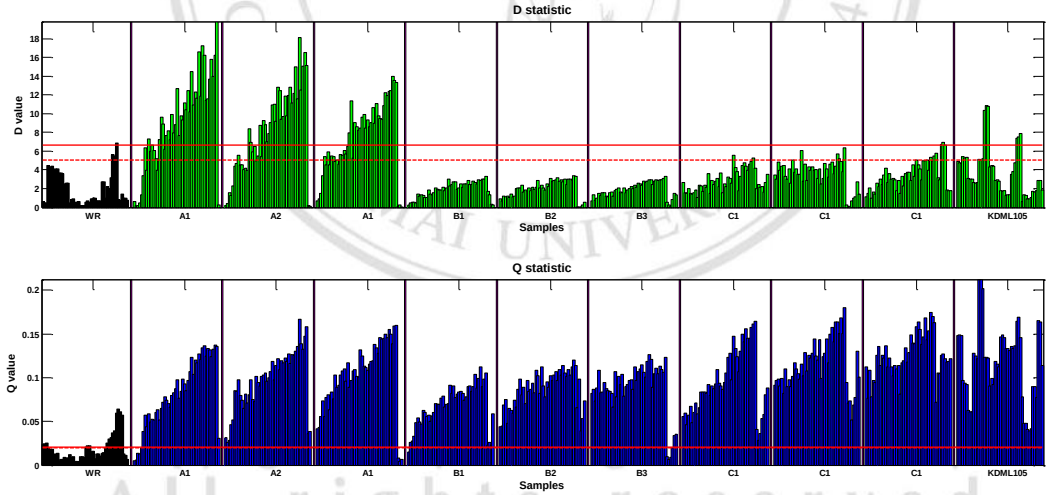


D and Q statictics of KDML105 in optimum number of PCs (1 – 8 PCs)

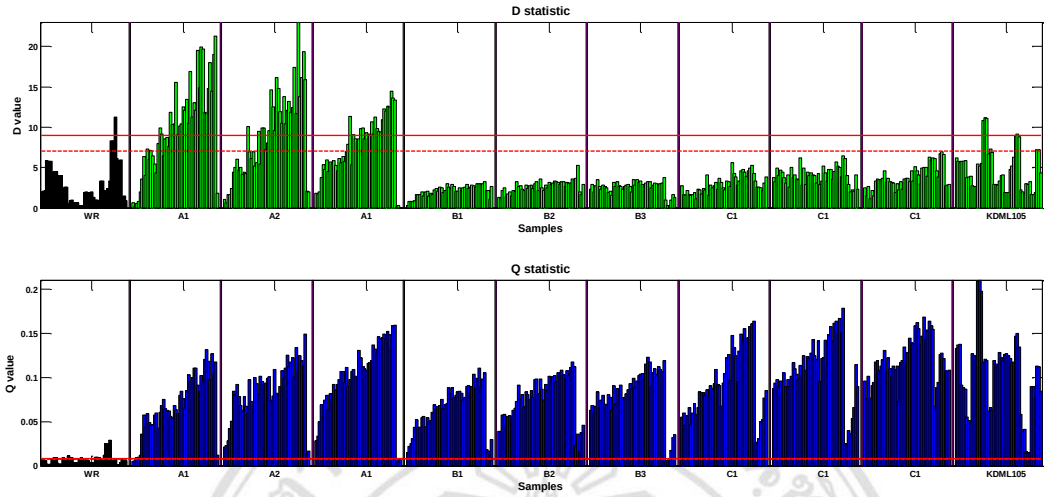
1 PC, WR model, SNV + mean centring at 90% (--) and 95% (-)



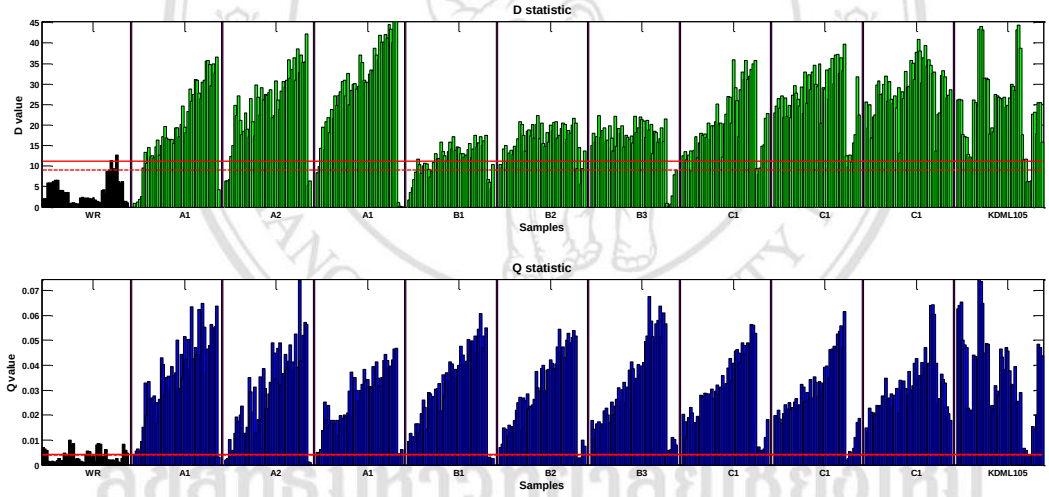
2 PCs, WR model, SNV + mean centring at 90% (--) and 95% (-)



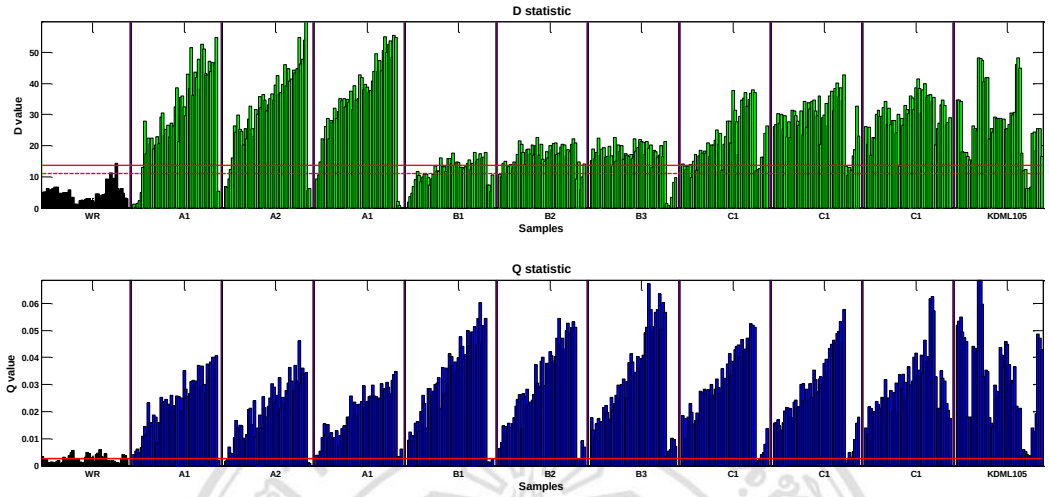
3 PCs, WR model, SNV + mean centring at 90% (--) and 95% (-)



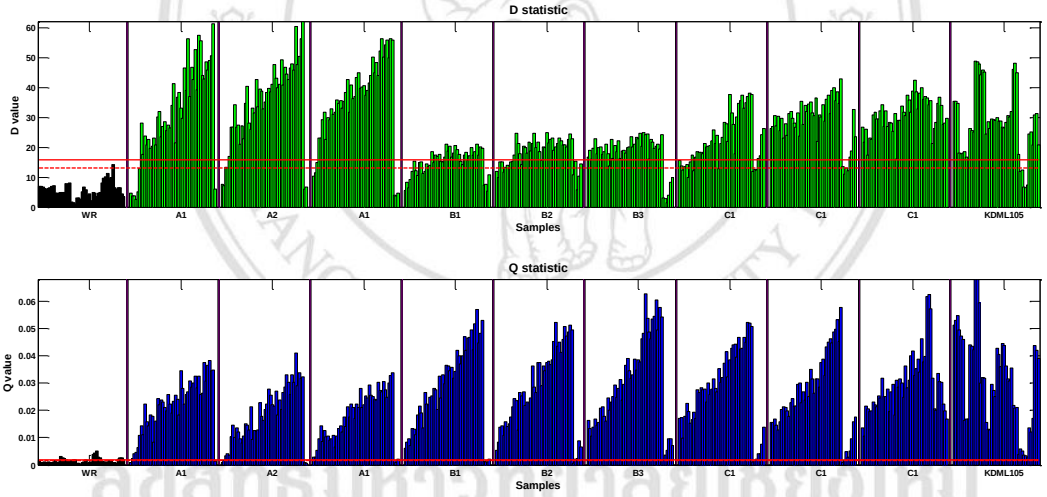
4 PCs, WR model, SNV + mean centring at 90% (--) and 95% (-)



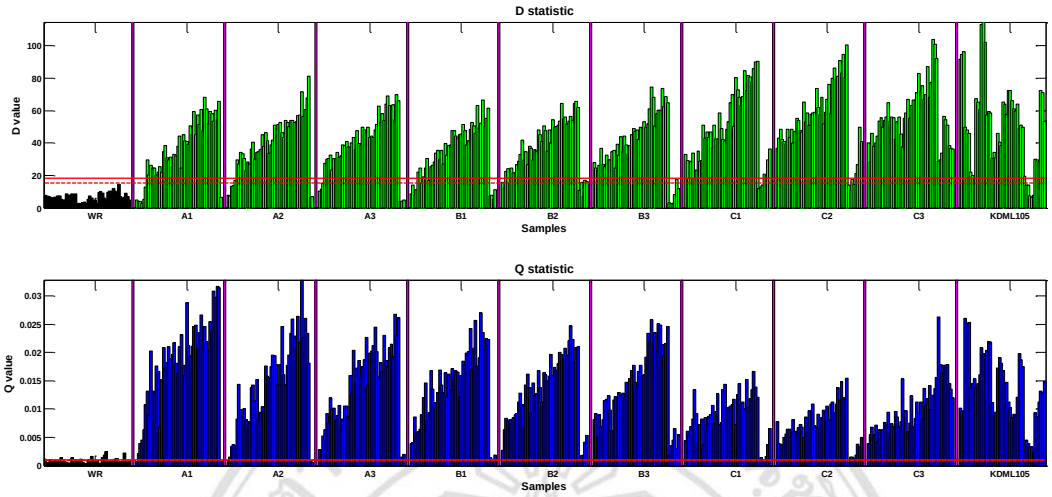
5 PCs, WR model, SNV + mean centring at 90% (--) and 95% (-)



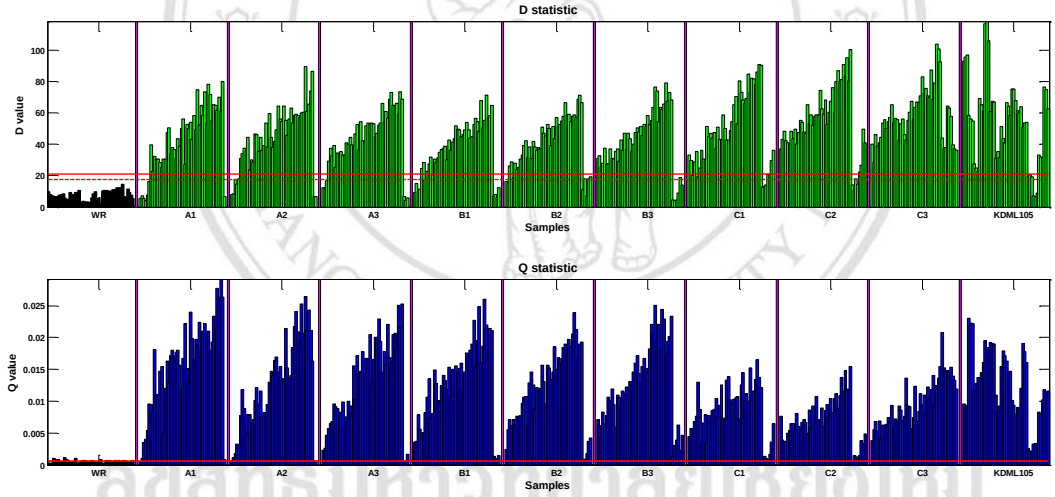
6 PCs, WR model, SNV + mean centring at 90% (--) and 95% (-)



7 PCs, WR model, SNV + mean centring at 90% (--) and 95% (-)

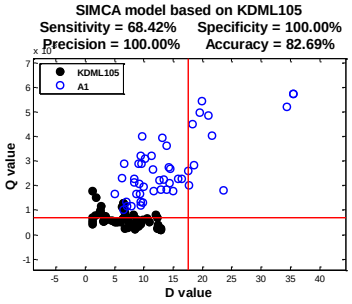


8 PCs, WR model, SNV + mean centring at 90% (--) and 95% (-)

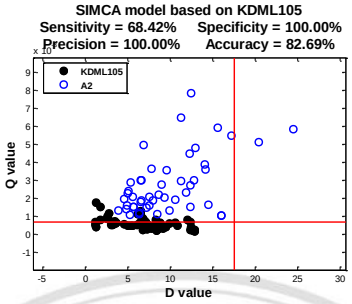


Classification of nine datasets using SIMCA based on KDML 105

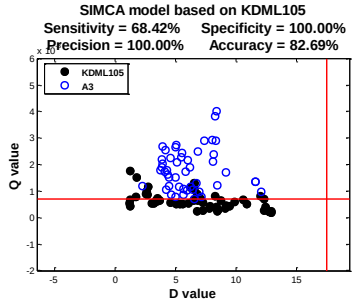
A1



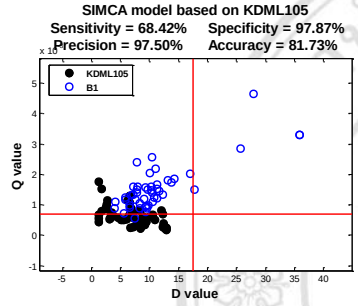
A2



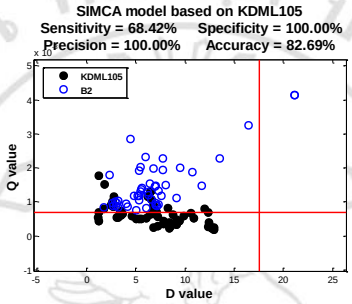
A3



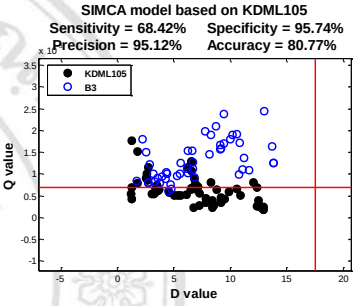
B1



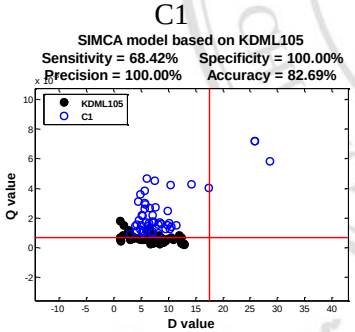
B2



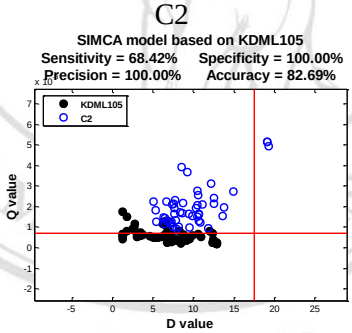
B3



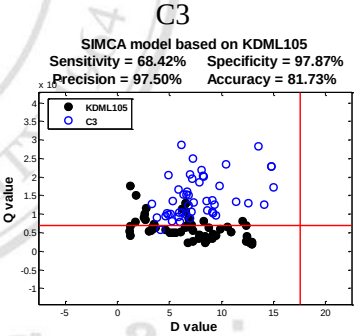
C1



C2

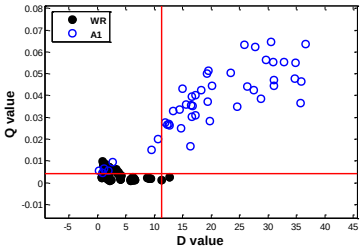


C3

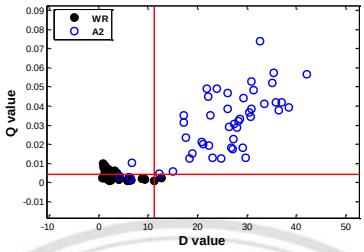


Classification of nine datasets using SIMCA based on WR

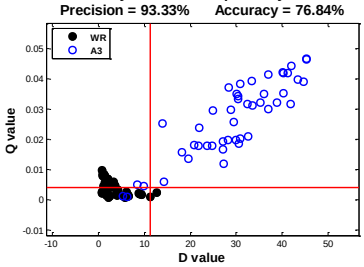
A1



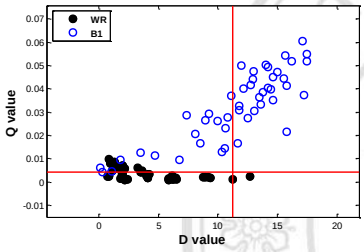
A2



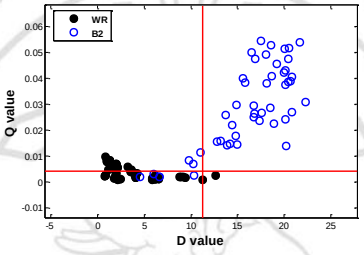
A3



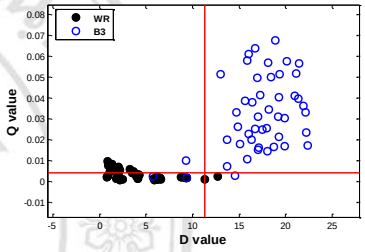
B1



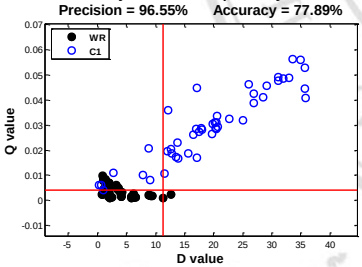
B2



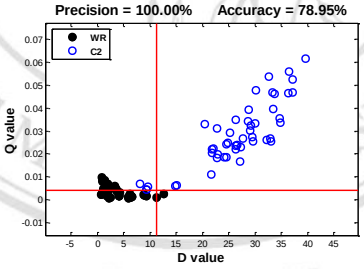
B3



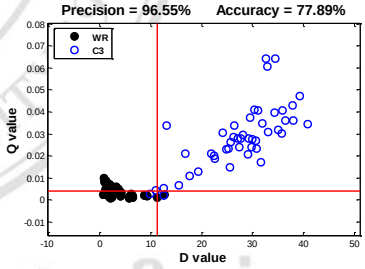
C1



C2



C3



Prediction of all datasets with SNV preprocessed

No.	Training set	Test set	PLS		OP-PLS		PDS-PLS		OP-PDS-PLS	
			Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP
1		A1	0.9871	2.37	0.9893	2.16	0.9871	2.37	0.9893	2.15
2		A2	0.9832	2.70	0.9852	2.54	0.9830	2.72	0.9813	2.85
3		A3	0.9845	2.60	0.9845	2.60	0.9796	2.98	0.9879	2.29
4		B1	-0.0015	20.85	0.8167	8.92	0.9734	3.40	0.9760	3.23
5	A1	B2	-0.0015	20.85	0.8808	7.19	0.9749	3.30	0.9792	3.00
6		B3	-0.0016	20.85	0.8895	6.92	0.9687	3.68	0.9691	3.66
7		C1	-0.0021	20.85	0.9128	6.15	0.9811	2.87	0.9780	3.09
8		C2	-0.0018	20.85	0.8557	7.91	0.9618	4.07	0.9648	3.91
9		C3	-0.0020	20.85	0.9181	5.96	0.9817	2.82	0.9831	2.71
1		A1	0.9836	2.67	0.9907	2.01	0.9845	2.59	0.9816	2.83
2		A2	0.9848	2.57	0.9866	2.41	0.9847	2.58	0.9866	2.42
3		A3	0.9807	2.89	0.9829	2.72	0.9810	2.87	0.9889	2.20
4		B1	-0.0015	20.85	0.8926	6.83	0.9818	2.81	0.9820	2.79
5	A2	B2	-0.0015	20.85	0.8901	6.91	0.9713	3.53	0.9690	3.67
6		B3	-0.0017	20.85	0.8762	7.33	0.9766	3.19	0.9772	3.14
7		C1	-0.0020	20.85	0.8205	8.83	0.9864	2.43	0.9863	2.44
8		C2	-0.0017	20.85	0.7940	9.45	0.9717	3.50	0.9731	3.41
9		C3	-0.0019	20.85	0.8631	7.71	0.9797	2.97	0.9778	3.11
1		A1	0.9793	3.00	0.9797	2.96	0.9816	2.83	0.9832	2.70
2		A2	0.9793	3.00	0.9825	2.76	0.9816	2.83	0.9749	3.30
3		A3	0.9797	2.97	0.9855	2.51	0.9797	2.96	0.9855	2.51
4		B1	-0.0015	20.85	0.4789	15.04	0.9758	3.24	0.9839	2.65
5	A3	B2	-0.0016	20.85	0.6372	12.55	0.9676	3.75	0.9735	3.39
6		B3	-0.0017	20.85	0.7945	9.44	0.9667	3.80	0.9692	3.65
7		C1	-0.0020	20.85	0.3305	17.05	0.9821	2.79	0.9814	2.84
8		C2	-0.0018	20.85	0.2831	17.64	0.9816	2.83	0.9820	2.80
9		C3	-0.0019	20.85	0.4964	14.78	0.9815	2.84	0.9810	2.87

Prediction of all datasets with mean centring preprocessed

No.	Training set	Test set	PLS		OP-PLS		PDS-PLS		OP-PDS-PLS	
			Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP
1		A1	-0.0029	20.86	0.9352	5.30	0.9813	2.85	0.9792	3.00
2		A2	-0.0028	20.86	0.9328	5.40	0.9759	3.23	0.9699	3.61
3		A3	-0.0027	20.86	0.9347	5.32	0.9790	3.02	0.9806	2.90
4		B1	0.9692	3.65	0.9767	3.18	0.9691	3.66	0.9768	3.18
5	B1	B2	0.9553	4.41	0.9630	4.01	0.9599	4.17	0.9647	3.91
6		B3	0.9699	3.61	0.9698	3.62	0.9718	3.50	0.9729	3.43
7		C1	0.9682	3.72	0.9682	3.72	0.9797	2.97	0.9802	2.93
8		C2	0.8979	6.66	0.9318	5.44	0.9665	3.81	0.9696	3.63
9		C3	0.9599	4.17	0.9650	3.90	0.9672	3.77	0.9730	3.42
1		A1	0.5065	14.64	0.9642	3.94	0.9866	2.42	0.9833	2.69
2		A2	0.2796	17.68	0.9496	4.68	0.9764	3.20	0.9721	3.48
3		A3	0.1645	19.04	0.9518	4.57	0.9779	3.10	0.9853	2.52
4		B1	0.9491	4.70	0.9713	3.53	0.9735	3.39	0.9863	2.44
5	B2	B2	0.9710	3.55	0.9772	3.15	0.9710	3.55	0.9772	3.15
6		B3	0.9753	3.27	0.9754	3.27	0.9740	3.36	0.9735	3.39
7		C1	0.6930	11.54	0.9383	5.17	0.9771	3.15	0.9765	3.19
8		C2	0.7817	9.73	0.9429	4.98	0.9668	3.80	0.9679	3.73
9		C3	0.7936	9.46	0.9832	2.70	0.9720	3.48	0.9730	3.42
1		A1	-0.0029	20.86	0.9480	4.75	0.9884	2.25	0.9887	2.22
2		A2	-0.0028	20.86	0.9562	4.36	0.9789	3.03	0.9790	3.02
3		A3	-0.0027	20.86	0.9608	4.13	0.9740	3.36	0.9756	3.25
4		B1	0.9409	5.06	0.9682	3.72	0.9847	2.58	0.9892	2.16
5	B3	B2	0.9667	3.80	0.9687	3.68	0.9663	3.82	0.9729	3.43
6		B3	0.9807	2.89	0.9807	2.90	0.9808	2.89	0.9807	2.90
7		C1	0.9159	6.04	0.9567	4.34	0.9788	3.03	0.9732	3.41
8		C2	0.9065	6.37	0.9295	5.53	0.9722	3.47	0.9736	3.39
9		C3	0.9373	5.21	0.9379	5.19	0.9820	2.80	0.9831	2.71

Prediction of all datasets with SNV + centring preprocessed

No.	Training set	Test set	PLS		OP-PLS		PDS-PLS		OP-PDS-PLS	
			Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP	Q^2	RMSEP
1		A1	0.8608	7.77	0.9732	3.41	0.9812	2.86	0.9790	3.02
2		A2	0.9115	6.20	0.9754	3.27	0.9822	2.78	0.9761	3.22
3		A3	0.9375	5.21	0.9398	5.11	0.9792	3.00	0.9843	2.61
4		B1	0.9405	5.08	0.9542	4.46	0.9740	3.36	0.9745	3.33
5	C1	B2	0.9281	5.59	0.9576	4.29	0.9601	4.16	0.9723	3.47
6		B3	0.9342	5.34	0.9372	5.22	0.9610	4.11	0.9639	3.96
7		C1	0.9870	2.37	0.9900	2.09	0.9871	2.37	0.9898	2.10
8		C2	0.8701	7.51	0.9571	4.32	0.9768	3.18	0.9720	3.48
9		C3	0.9727	3.44	0.9735	3.39	0.9762	3.22	0.9889	2.19
1		A1	0.7615	10.17	0.9544	4.45	0.9835	2.67	0.9811	2.87
2		A2	0.8682	7.56	0.9821	2.79	0.9793	3.00	0.9755	3.26
3		A3	0.8684	7.56	0.9724	3.46	0.9545	4.44	0.9803	2.93
4		B1	0.8103	9.07	0.9473	4.78	0.9597	4.18	0.9763	3.21
5	C2	B2	0.9302	5.51	0.9495	4.68	0.9579	4.28	0.9632	4.00
6		B3	0.9362	5.26	0.9485	4.73	0.9742	3.35	0.9670	3.79
7		C1	0.9451	4.88	0.9606	4.14	0.9650	3.90	0.9762	3.21
8		C2	0.9721	3.48	0.9721	3.48	0.9721	3.48	0.9721	3.48
9		C3	0.9610	4.12	0.9688	3.68	0.9699	3.61	0.9723	3.47
1		A1	0.9409	5.07	0.9484	4.73	0.9887	2.21	0.9881	2.27
2		A2	0.8785	7.26	0.9357	5.28	0.9771	3.15	0.9775	3.12
3		A3	0.9671	3.78	0.9696	3.63	0.9816	2.83	0.9855	2.51
4		B1	0.8758	7.34	0.9501	4.66	0.9764	3.20	0.9813	2.85
5	C3	B2	0.8520	8.01	0.9577	4.29	0.8520	8.01	0.9760	3.23
6		B3	0.8631	7.71	0.9634	3.99	0.9745	3.33	0.9711	3.54
7		C1	0.9529	4.52	0.9709	3.55	0.9756	3.26	0.9790	3.02
8		C2	0.9396	5.12	0.9712	3.54	0.9727	3.44	0.9723	3.47
9		C3	0.9818	2.81	0.9849	2.56	0.9817	2.82	0.9847	2.58

APPENDIX II

Supporting works

Journal publications

1. **Wongsaipun, S**., Krongchai, C., Jakmunee, J. & Kittiwachana, S. (2018). Rice grain freshness measurement using rapid visco analyzer and chemometrics. *Food Analytical Methods*, 11:613–623.

2. Theanjumpol, P., Wongzeewasakun, K., Muenmanee, N., **Wongsaipun, S**., Krongchai, C., Changrue, V., Boonyakiat, D. & Kittiwachana, S. (2019). Non-destructive identification and estimation of granulation in 'Sai Num Pung' tangerine fruit using near infrared spectroscopy and chemometrics. *Postharvest Biology and Technology*, 153:13-20.

3. Srithongkul, C., **Wongsaipun, S**., Krongchai, C., Santasup, C. & Kittiwachana, S. (2019). Investigation of mobility and bioavailability of arsenic in agricultural soil after treatment by various soil amendments using sequential extraction procedure and multivariate analysis. *Catena*, 181:1-11.

4. Thangsunan, P., **Wongsaipun, S**., Kittiwachana S. & Suree, N. (2019). Effective prediction model and determination of binding residues influential for inhibitors targeting HIV-1 integrase-LEDGF/p75 interface by employing solvent accessible surface area energy as key determinant. *Journal of Biomolecular Structure and Dynamics*, 27:1-14.

5. Krongchai, C., **Wongsaipun, S**., Funsueb, S., Theanjumpol, P., Jakmunee, J. & Kittiwachana, S. Comparison between linear and non-linear variable selection methods with applications to spectroscopic (UV-Vis/NIR) data. *Chiang Mai Journal of Science*. (Submitted)

Oral presentations

1. **Sakunna Wongsaipun**, Grissana Sudtasarn, Jaroon Jakmunee, Sila Kittiwachana*, "Application of Multiple Self-organizing Maps for Classification of Rice Grain Variety Based on Their Rapid Viscosity Analysis Profiles", The 13th Asian Conference on Analytical Sciences (ASIANALYSIS XIII), 8 – 11 December 2016, Chiang Mai, Thailand.

CURRICULUM VITAE

Author's Name	Miss Sakunna Wongsaipun
Date/Year of Birth	20 May 1992
Place of Birth	Phayao Province, Thailand
Education	2013, B.Sc., Chemistry, Chiang Mai University 2016, M.Sc., Chemistry, Chiang Mai University 2019, Ph.D., Chemistry, Chiang Mai University
Scholarship	Science Achievement Scholarship of Thailand (SAST) through the Commission on Higher Education, Ministry of Education