โครงการ
การเรียนการสอนเพื่อเสริมประสบการณ์

ชื่อโครงการ การตรวจวัดไมโครพลาสติกในดินด้วยเทคนิคเนียร์อินฟราเรดสเปกโทรสโกปี
และเคโมเมทริกซ์
Detection of microplastics in soil using near infrared spectroscopy
And chemometrics

ชื่อนิสิต นายณัฐภัทร สินสุวรรณาภรณ์   **เลขประจำตัว** 5833032723
ภาควิชา เคมี
ปีการศึกษา 2561

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

การตรวจวัดไมโครพลาสติกในดินด้วยเทคนิคเนียร์อินฟราเรดสเปกโทรสโกปี
และเคโมเมทริกซ์

Detection of microplastics in soil using near infrared spectroscopy
and chemometrics

โดย
นายณัฐภัทร สินสุวรรณาภรณ์

รายงานชิ้นนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต
ภาควิชาเคมี คณะวิทยาศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

Detection of microplastics in soil using near infrared

Spectroscopy and chemometrics
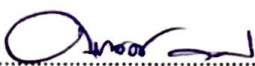
Mr. Nuttapat Sinsuvannaporn

In Partially fulfillment for the Degree of Bachelor of Science

Department of Chemistry, Faculty of Science

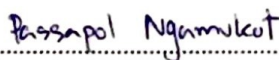Chulalongkorn University

Academic Year 2018

Project Title:	Detection of microplastics in soil using near infrared spectroscopy and chemometrics
Student Name:	Mr. Nuttapat Sinsuvannaporn
Field of Study:	Chemistry
Project Advisor:	Associate Professor Dr. Kanet Wongravee

Accepted by Department of Chemistry, Faculty of Science, Chulalongkorn University in Partial fulfillment of the Requirement for the Degree of Bachelor of Science

SENIOR PROJECT COMMITTEE

.... Chair Committee

(Associate Professor Viwat Vchirawongkwin, Dr. rer. nat.)

.... Project Advisor

(Associate Professor Kanet Wongravee, Ph.D.)

.... Committee

(Assistant Professor Passapol Ngamukot, Ph.D.)

This report was approved by head of Department of Chemistry

.... Head of Department of Chemistry

(Associate Professor Vudhichai Parasuk, Ph.D.)

Date......... Month May Year 2019

ชื่อโครงการ การตรวจวัดไมโครพลาสติกในดินด้วยเทคนิคเนียร์อินฟราเรดสเปกโทรสโกปีและเคโมเมทริกซ์

ชื่อนิสิตในโครงการ นายณัฐภัทร สินสุวรรณาภรณ์ **เลขประจำตัว** 5833032723

ชื่ออาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. คเณศ วงษ์ระวี

ภาควิชาเคมี คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีการศึกษา 2561

บทคัดย่อ

ไมโครพลาสติกในดินส่งผลกระทบในวงกว้างต่อสิ่งแวดล้อมอย่างมาก โดยเฉพาะสิ่งมีชีวิตที่อาศัยอยู่ใน ดิน ในการวิเคราะห์หาปริมาณไมโครพลาสติกด้วยวิธีดั้งเดิมเช่นเทคนิคโครมาโทรกราฟ ต้องใช้การสกัดแยกไม โครพลาสติกออกจากดินและมีความยุ่งยากในการวิเคราะห์ เนื่องจากต้องเตรียมตัวอย่าง ทำได้ยาก และต้องใช้ เวลานานในการวิเคราะห์ ในงานวิจัยนี้ผู้วิจัยได้เสนอการตรวจวัดไมโครพลาสติกในดินด้วยวิธีเนียร์อินฟราเรด สเปกโทรสโกปีซึ่งตรวจวัดในช่วงความยาวคลื่นตั้งแต่ 1100 - 2300 นาโนเมตร และวิเคราะห์ร่วมกับเทคนิค ทางเคโมเมทริกซ์ซึ่งเป็นวิธีการทางคณิตศาสตร์และสถิติ ข้อดีของเทคนิคเนียร์อินฟราเรดสเปกโทรสโกปี ได้แก่ ตรวจวัดได้ง่ายและรวดเร็ว ไม่ต้องเตรียมตัวอย่างก่อนการวิเคราะห์ ในการทดลองได้ใช้แก้วและจานพลาสติก ซึ่งทำมาจากพลาสติกชนิดพอลิเอทิลีนเทเรฟทาเลต (พีอีที) และพอลิสไตรีน ตามลำดับ นำมาบดและปั่นให้ ละเอียดด้วยเครื่องปั่นเพื่อทำให้พลาสติกมีขนาดเล็กลงจนได้ไมโครพลาสติกนำไมโครพลาสติกที่ได้มาวิเคราะห์ ขนาดและการกระจายด้วยกล้องจุลทรรศน์แบบใช้แสง ได้ขนาดของพีอีทีและพอลิสไตรีนคือ 44.45 ± 66.68 และ 103.39 ± 101.13 ไมครอนตามลำดับ ในการตรวจวัดด้วยเทคนิคเนียร์อินฟราเรดสเปกโทรสโกปีได้เตรียม ตัวอย่างไมโครพลาสติกในดินที่มีความเข้มข้นตั้งแต่ 1-10 ร้อยละโดยมวล และตรวจวัดโดยใช้โหมดสะท้อน กลับ ในการสร้างแบบจำลองสำหรับตรวจวัดไมโครพลาสติกในดินได้นำแบบจำลองกำลังสองน้อยที่สุดบางส่วน มาใช้ประกอบกับการเลือกช่วงความยาวคลื่นโดยใช้โปรแกรมแมตแล็บ พบว่าการตรวจวัดพีอีทีและพอลิสไตรีน ไมโครพลาสติกในดินได้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนจากการทำนายคือ 1.30 (R^2 = 0.8030) และ 1.19 (R^2 = 0.8253) ร้อยละโดยมวลตามลำดับ เมื่อทำการเลือกช่วงความยาวคลื่นพบว่าการตรวจวัดพีอี ทีและพอลิสไตรีนไมโครพลาสติกในดินได้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนจากการทำนายคือ 1.35 (R^2 = 0.7854, 47 ตัวแปร) และ 1.09 (R^2 = 0.8582, 56 ตัวแปร) ร้อยละโดยมวลตามลำดับ ในการตรวจวัด ไมโครพลาสติกที่ผสมกันทั้งสองชนิดในดินได้ทำการแปลงสเปกตรัมด้วยวิธีอีพีโอ พบว่าสามารถตรวจวัดไมโค รพลาสติกชนิดพีอีทีและพอลิสไตรีนได้ ซึ่งมีค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนจากการทำนายคือ 3.17 และ 3.26 ร้อยละโดยมวลตามลำดับ ดังนั้นสามารถตรวจวัดไมโครพลาสติกในดินได้ด้วยวิธีเนียร์ อินฟราเรดสเปกโทรสโกปีและเคโมเมทริกซ์

คำสำคัญ: ไมโครพลาสติก, เนียร์อินฟราเรดสเปกโทรสโกปี, เคโมเมทริกซ์, ดิน

Project Title	Detection of microplastics in soil using near infrared spectroscopy and chemometrics
Student Name	Mr. Nuttapat Sinsuvannaporn **Student ID** 5833032723
Advisor Name	Associate Professor Kanet Wongravee. Ph.D.

Department of Chemistry, Faculty of Science, Chulalongkorn University, Academic Year 2018

Abstract

Microplastic pollution in soil have a vital impact on organisms living in soil. To quantify amount of microplastics in soil, the preparation approach need to be performed in order to separate plastic from soil. Therefore, the conventional methods such as chromatography are complicated and time-consuming method to identify and quantify microplastics in soils. In the study, the powerful of near infrared spectroscopy (NIRS) ranging 1100 – 2300 nm combined with chemometrics tools are demonstrated as a technique for rapidly monitoring microplastics in soil as the preparation step is unrequired. Polyethylene terephthalate (PET) and polystyrene (PS) were grinded by blender to generate the artificially microplastics. Size of grinded microplastics of PET and PS are in the range of 44.45 ± 66.68 and 103.39 ± 101.13 µm, respectively. Sets of artificially polluted soil samples were prepared by mixing the microplastics in soil with various concentrations (1-10 %w/w). The reflectance mode of NIR were used to acquire the NIR spectra of soil contaminated with microplastic. The models for microplastic prediction were generated using machine learning algorithms, partial least squares regression (PLSR), with data pre-processing in MATLAB$^{®}$. The performance of models was evaluated showing that *RMSEP* for PET was 1.30 %w/w with R^2 = 0.8030 and *RMSEP* for PS was 1.19 %w/w with R^2 = 0.8253. After using variable selection method (shaving method), the *RMSEP* for PET was 1.35% w/w with R^2 = 0.7854 (47 variables) and *RMSEP* for PS was 1.09 %w/w with R^2 = 0.8582 (56 variables). The mixture of PET and PS microplastics in soil samples were quantified using external parameter orthogonalization (EPO) to extract only the reflectance spectra of target analyzes. The results presented *RMSEP* was 3.17 %w/w and 3.26 %w/w for PET and PS, respectively. It could be concluded that NIRS combined with chemometrics is potential technique for microplastics detection in soil.

Keywords: microplastics, near infrared spectroscopy, soil, chemometrics

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

1.1. Introduction

Every minute, tons of plastic have been produced continuously and there is enormous use of plastics. Microplastics are very small plastic particles that come from a variety of sources, including from a bulk plastic piece that gradually degrades into smaller pieces. Plastic is widely used and presented everywhere. Microplastics are increasingly considered as an environmental problem in recent years. People have been mainly focusing on environmental impacts from microplastics especially in the ocean and soil. In the ocean, effects of microplastics on marine organism is very concerned as a challenge problem to solve and to detect the microplastics particles. In contrast, microplastics in soil have been overlooked. Plastic wastes are landfilled in soil, therefore, microplastic pollution in soil or agricultural landscapes has potential consequences to plants and animals. Soils are essential components of landscapes in the world that can be polluted easily from waste or chemical substances. The potential consequences are biodiversity impacts and ecological effects. Although microplastics are presented in soils, the analytically available method has no standard method for microplastic detection including identification and quantification[1]. Many available protocols to qualify and quantify microplastics in soil have to extract microplastics from soil before analysis using chemical analysis instruments. Therefore, the conventional protocols are urgently needed to develop[2].

In last decade, various techniques for the microplastic detection have been published. Several standard techniques were used to determine amount and type of microplastics including Raman spectroscopy, Infrared (IR) spectroscopy and chromatography[2]. In 2018, Catarina F. Araujo. *et. al.*[2] reported the use of Raman spectroscopy to identify of microplastics. The technique can analyze of very small microplastic particles (< 20 micron). The authors have been reported that the technique drawbacks long measurement time of analysis. In 2018, Philipp M. Anger. *et. al.*[3] reported the use of Raman microspectroscopy technique to analyze microplastic particles. In 2018, Araujo C. F., *et. al.*[4] reported a review on occurrence of microplastics and detection including infrared spectroscopy, Raman spectroscopy, scanning electron microscope (SEM) coupled with energy dispersive X-ray spectroscopy. Although this

method is specific to individual microplastic signals in Raman spectrum, but they require single determination for each compound, which time consuming and expensive.

In many analytical tools, near infrared spectroscopy (NIRS) has been reviewed as a potential tool becoming popular technique to qualify and quantify a wide range of chemical analyses[6] in agriculture, food industry, pharmaceuticals and chemical process monitoring. The interesting benefits of this technique is mostly attributed to the rapid and easy detection, nondestructive analysis of bulk materials and unnecessary sample preparations. Near infrared (NIR) spectra are based on the rules of vibrational motions and observed in several bands including overtones and combination bands which differ from fundamental vibrational modes in mid infrared spectra. NIR spectroscopy is a vibrational spectroscopy which measures wavelengths from 800 nm to 2500 nm corresponding to overtone and combination bands of vibrational transitions in anharmonic oscillating potential model. The NIR absorptions are sensitive to C-H, O-H and N-H bonds which are mainly functional groups in chemical molecules. In contrast, NIR spectra containing overtone and combination bands are too complex to interpret using a conventional univariate calibration. The chemometrics or machine learning have to be employed to extract analytical signals and information from the NIR spectra. The development of chemometric models, involving mathematical and statistical method is required in order to interpret near infrared data. To solve NIRS analytical problems, the mathematical and statistical treatment are the major factor in order to develop a powerful predictive model. NIRS combined with chemometrics tools can be used in qualitative and quantitative analysis. Multivariate analysis or many machine learning algorithms are especially used for generating a complex model in NIRS to determine several targeted molecules.

Chemometrics, mathematical and statistical method for chemical data, is an important method in NIRS to interpret and extract relevant chemical data[6]. Multivariate analysis is mainly used for dimensionality reduction, data visualization, data classification, regression and statistical inference in NIR spectra. In regression problems, multivariate linear regression used for explaining the relationship between the predictor (X) and response (y). The predictor (X) consisting of independent variables are NIR spectra collected in matrix form and the response (y) is dependent variables which linearly related to the predictor (X). In the case of NIRS, the predictor (X) is the NIR spectra containing the reflection value of each wavelengths and the response (y) is the effects from the predictor such as chemical concentrations. The predictive models based on multivariate linear regression have two major parts including model training

with or without model selection and prediction to evaluate model validity and model performance. The training data set as a predictor (*X*) consists of NIR spectra in matrix form whose dimension are *M* rows (*M* observations) and N columns (*N* variables) and the response (*y*) consists of response values in vector form whose dimension are *M* rows corresponding to *M* observations and the number of columns is corresponding to the targeted responses usually have a one column. In linear regression model, the coefficient vector (*b*) were found by optimizing to minimize an error between predictor (*X*) of the training model and response (*y*) in L2-norm or high-dimensional Euclidean distance. The obtained coefficients were used to predict response called predicted response of the unknown predictor or test data sets.

In NIR spectroscopy, calibration curve built from a single selected peak from NIR spectra in univariate analysis fashion is obviously not sufficient. According to patterns of chemical signals in NIR spectra, chemometrics has been used to extract the pattern from all of wavelengths in NIR spectra for effective interpretation. Multivariate analysis can analyze the pattern from all of wavelengths in NIR spectra such as principal component analysis (PCA) and partial least squares (PLS) regression. Mostly, the PLS regression has been frequently used to generate the multivariate calibration model from NIR training data set. The PLS regression is mathematical model which is a relationship between partially latent variables from predictor (*X*) and response (*y*). The calibration model has been practically generated to quantify the predicted response for any system and using multivariate analysis in NIR spectroscopy is very importance to achieve the accuracy with high performance of the calibration model.
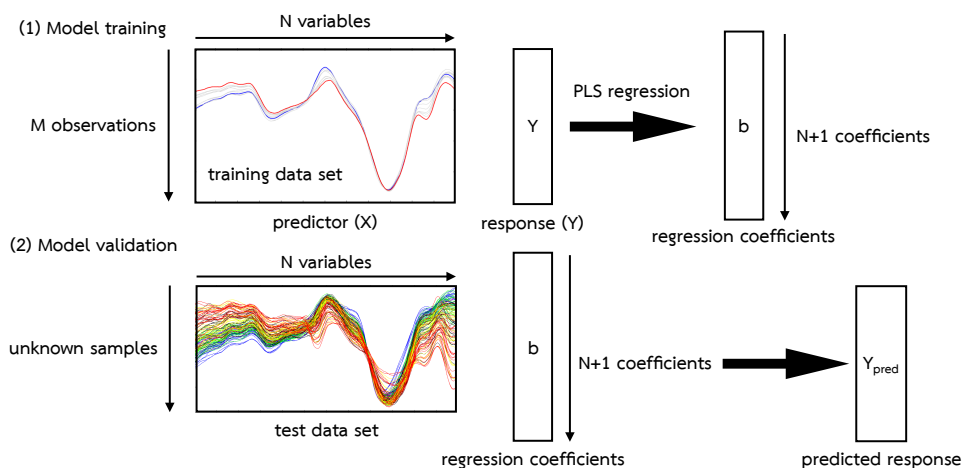
Figure 1.1 summary of partial least squares (PLS) model in calibration and prediction

There is only one article of using NIR spectroscopy for detection of microplastics. Recently, in 2019, Fabio Corradini *et. al.*[5] reported the use of near infrared (NIR) spectroscopy to quantify microplastic in soil. This research used multilinear regression baes on Bayesian approach. For the result, root mean squared deviation (*RMSD*) of 8, 18 and 10 g/kg for LDPE, PET and PVC, respectively. The root mean squared deviation (*RMSD*) in mixture treatment were reported of 8, 10, 15 g/kg for LDPE, PET and PVC, respectively.

In this work, the NIR spectroscopy was used and combined with chemometrics to propose calibration model using partial least squares (PLS) regression. Firstly, the predictor (*X*) were regressed on the response concentration vector to optimize number of PLS components. Secondly, the optimized number of PLS component was used to generate predictive model. Finally, the test data set was used to evaluate model performance and root mean squared error of prediction (*RMSEP*) was calculated.

1.2. Objective of this work

To develop and demonstrate a potential of predictive model based on partial least squares (PLS) regression to quantify polyethylene terephthalate (PET) and polystyrene (PS) in soil using near infrared spectroscopy and chemometrics

1.3. Scope of this work

This study involves the development of calibration model based on chemometrics to perform the predictive model to quantify polyethylene terephthalate (PET) and polystyrene (PS) in soil using near infrared spectroscopy. This study used data pre-processing techniques including smoothing, standard normal variate (SNV) and external parameter orthogonalization (EPO). The developed model, partial least squares (PLS) model with and without variable selection method, shaving method were used in this study.

1.4. Theoretical background
1.4.1. Near infrared spectroscopy

Near infrared (NIR) spectroscopy is vibrational spectroscopy between the infrared (IR) region and visible light that can be expressed in range of 800 – 2500 nm[6]. NIR spectroscopy is based on overtone and combination transitions of chemical bonds especially C-H, O-H and N-H. Combination bands are based on concurrently interaction of two or more vibrational

modes. Generally, the overtone bands are related to vibrational transitions that the vibrational quantum number is changed of two or more. According to the selection rules of quantum mechanics, NIR spectroscopy is based on anharmonic oscillator vibrational potential. Transition from ground ($\boldsymbol{V}$ = 0) to the first excited state, called fundamental vibration, which absorbs strongly light in infrared (IR) spectroscopy according selection rules. The transition from ground state to the second excited state with absorption of NIR light called first overtone. The transition from ground state to the third and fourth excited state with absorption of NIR light called second and third overtone, respectively (Figure 1.2.). The overtone and combination bands are characteristic bands of chemical compounds. As a result, it is difficult to interpret NIR spectrum without mathematical and statistical methods. Therefore, chemometrics combined with NIR spectroscopy for extracting relevant signals are very important.
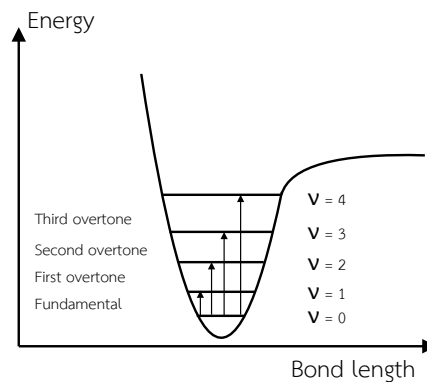
Figure 1.2. summary of vibrational transition in NIR spectroscopy

1.4.2. Chemometrics

1.4.2.1. Principal component analysis (PCA)

PCA is an unsupervised technique to reduce dimension of data and extract relevant features[6]. The concept of PCA is to find the orthonormal basis corresponding to the maximum variance extracted form data. The orthonormal basis is used to project the data onto the basis to obtain coefficients. The coefficients are a representative of the original data called scores. To perform PCA, the singular value decomposition was used to decomposition of data and find the loading (orthonormal basis). In summary, the PCA algorithm work as follows:

1.) Data matrix or predictor $\boldsymbol{X}$ whose dimension is M rows and N columns

2.) Orthogonal Decomposition of Covariance Matrix

$$C = \frac{1}{n-1}X^{T}X = WDW^{-1} = WDW^{T}$$

3.) Singular Value Decomposition (SVD) of data matrix X

$$X = UCW^T$$

$$X = \sigma_1 u_1 w_1^T + \sigma_2 u_2 w_2^T + \dots \sigma_r u_r w_r^T$$

For $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \dots \geq \sigma_r$ we select the w_1 and w_2 are orthonormal basis for orthogonal projection

4.) Orthogonal Projection of Data X

$$t_1 = Xw_1 \text{ and } t_2 = Xw_2 \text{ to obtain } T = XW \text{ and } X = TP^T + E$$

1.4.2.2. Partial least squares (PLS) model

In order to generate calibration model based on linear regression, partial least squares (PLS) is one of the most popularly used in spectroscopy[6]. Its purpose is to predict the output or response (y) related to predictor (X). The predictor (X) was projected onto coefficient subspace optimized from minimization of error term in L2-norm. In this regression, the training data set as a predictor (X) consists of NIR spectra in matrix form whose dimension are M rows (M observations) and N columns (N variables) and the response (y) consists of response values in vector form whose dimension are M rows corresponding to M observations and the number of columns is corresponding to the targeted responses usually have a one column. According to loading of the predictor (X) extracted from singular value decomposition (SVD) of X, the score matrix (T) is obtained from projecting predictor (X) onto loading. The following matrix equations are PLS algorithms and the PLS model is equation (1) by obtain the vector of regression coefficients (b).

$$X = TP^T + E \quad \text{and} \quad T = XW$$

$$Y = TQ^T + F$$

$$b = WQ^T$$

$$Y = Xb \qquad \qquad \dots (1)$$

1.4.2.3. Standard Normal Variate (SNV)

Standard normal variate (SNV) is a data pre-processing technique to normalize NIR spectra and remove unsystematic effects such as temperature and spectral baseline shift[6]. When NIR spectroscopy was performed, the obtained NIR spectra might be fluctuated from environmental effects and causing baseline shift of spectra. To normalize NIR spectra, SNV is

used as popular pre-processing method. This method used the average value (μ) and standard deviation (*SD*) of all of values in NIR spectra to subtract the interferences (equation 2).

$$X_{new} = \frac{x_i - \mu}{SD} \qquad \dots (2)$$

1.4.2.4. External parameter orthogonalization (EPO)

The external parameter orthogonalization was used as an algorithm to remove the effect of soil such as organic matters in soil[7]. The algorithm finds the regions in the spectra which are affected by target microplastic and projects the spectra data onto the orthogonal variation. The unsystematic and unwanted signals can be effectively removed. This analysis is related to principal component analysis (PCA). The subspace of spectra data was divided into three parts including chemical spectral responses, external parameter and residuals. In matrix form, the spectra or predictor (*X*) can be written as orthogonal decomposition of predictor[6].

$X = XP + XQ + R$

P is projection matrix corresponding to useful part of spectra

Q is projection matrix corresponding to not useful part of spectra

R is residual matrix

In summary, the EPO algorithm work as follows:

1.) Calculate difference spectra: $D = X_{predictor} - X_{plastic}$

2.) Perform singular value decomposition of *D*

3.) Define number of EPO components and corresponding *V* (eigenvector)

4.) Find $Q = VV^T$

5.) Calculate the projection matrix as $P = I - Q$

6.) The transformed spectra are calculated as $X_{new} = X_{predictor}P$

1.4.2.5. Variable selection (shaving method)

The shaving using selectivity ratio (*SR*) is a visualization tool for searching what are the important variables of a multivariate data set[8,9]. The ratio between the explained and the residual (unexplained) variance for each variable in the target projection vector (TP) defines the *SR* for the variables. This target projection utilizes both the predictive ability (regression vector) and the explanatory ability (spectral variance and covariance matrix) for the calculation

of the selectivity ratio. Given the PLS regression vector, **b**, target Projection is performed by projecting of the rows of **X** onto the normalized regression coefficients vector **b**. The score t_{TP} is proportional to the predicted values. The loadings, p_{TP}, are obtained by projecting the columns of **X** onto the score vectors, t_{TP}, which again is proportional to $y = Xb$

$$t_{tp} = Xb_{PLS}/\|b_{PLS}\|$$

$$p_{tp} = Xt_{tp}/(t_{tp}{}^T t_{tp})$$

The ratio of the explained variance $SS_{explained}$ (and of the residual variance for each variable $SS_{residual}$) in the sum of squares respectively, is used then to determine the variable importance using selectivity ratio (*SR*).

$$SS_{i,explained} = \|t_{tp,i}p_{tp,i}{}^T\|^2$$

$$SS_{i,residual} = \|e_{tp,i}\|^2$$

$$e_{tp} = X - t_{tp,i}p_{tp,i}{}^T$$

$$SR_i = SS_{i,explained}/SS_{i,residual}$$

1.4.2.6. Root Mean Squared Error (*RMSE*)

The root mean squared error (*RMSE*) is a measure of the differences between the values observed and values predicted by model. The *RMSE* is used for evaluating the model performance. Typically, the more complex model, the lower the bias, but the higher the variance reflected in mean squared error (*MSE*). The *RMSE* is the standard deviation of the error or residuals which can be calculated by the following equation.

$$RMSE = \sqrt{\sum_{i=1}^{N}(y_{predicted} - y_{observed})^2}$$

CHAPTER II

EXPERIMENTS

2. Materials and Method

2.1. Chemicals and Materials

Plastic glasses and bowls were purchased from local supermarkets (Gourmet market, Siam Paragon, Bangkok). Organic soils were purchased from HomePro (Home Product Center Public Company Limited), Ploenchit, Bangkok. The commercialized plastic glasses and bowls were collected from the supermarket. They were made from polyethylene terephthalate (PET) and polystyrene (PS), respectively. The plastics glassed and bowls were cut into smaller pieces and were later grinded for 30 minutes to obtain microplastics. The organic soil was heated in an oven at 60 $^{\circ}$C for 1 hour in order to remove excess water prior for further analysis. All glassware was cleaned up with detergent followed by DI water for several times.

2.2. Sample preparation

For preparation microplastic mixed in soil samples, each type of microplastics was prepared using % w/w to control the contamination. For example, to prepare 1 % w/w of microplastics in soil, the microplastics were weighed to approximately 1 g and then soil was added until the final weight of mixture is equal to 10 g, the microplastic concentrations of PET and PS in soil were 1 – 10 % w/w approximately for 10 contamination points (Table 2.1.). In mixture of microplastic in soil samples, the PET and PS microplastics were weighed and mixed with the total concentrations of to 10 % w/w as shown in Table 2.2. An external set of mixture of microplastics in soil samples was prepared with microplastic concentrations at 1, 3, 5, 7, 9 % w/w for each type of microplastics to be used as a test set for model validation.

Microplastic in soil samples (one type of microplastic)	Concentration of PET in soil (% w/w)	Concentration of PS in soil (% w/w)
1	0.93	1.00
2	1.96	2.03
3	2.96	3.09
4	3.92	4.09
5	5.04	5.11
6	5.82	6.14
7	6.59	7.19
8	7.43	7.86
9	8.78	8.67
10	9.76	10.05

Table 2.1. Summary of microplastic concentration of PET and PS in soil (1 – 10 %w/w)

Microplastic in soil samples (mixture of microplastic)	Concentration of PET in soil (% w/w)	Concentration of PS in soil (% w/w)
1	1.01	8.83
2	3.02	6.75
3	4.97	4.93
4	7.03	2.98
5	9.07	0.99

Table 2.2. Summary of microplastic concentration of mixture microplastic in soil samples for validation set

2.3. Spectral acquisition

NIR spectrometer with NIR256-2.5 detector, LS-1 tungsten halogen light source and fiber optic connector purchased from Ocean Optics was used to acquire NIR spectra of microplastics in soil samples. The window range of NIR spectra was set from 900 to 2500 nm with the resolution of 6.9 nm.
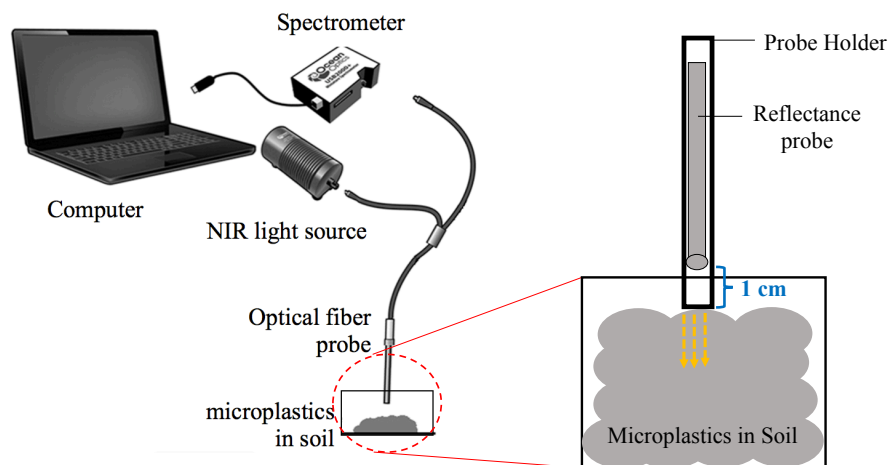
Figure 2.1. The setup scheme of NIR spectrometer

In this study, the spectral range were selected from 1100 – 2300 nm to avoid the fluctuation and unstable signals from environmental effects such as temperature. The number of total wavelengths was 176 wavelengths. The setup scheme of NIR instrument is shown graphically in Figure 2.1. The NIR spectra of the microplastic in soil samples were collected using reflectance mode with integration time of 100 milliseconds, 2 averaged scans with boxcar smoothing windows of 4. Each microplastic in soil samples was measured of five different areas and five replicated times for each area. Therefore, the number of acquired NIR spectra per one sample was 25 spectra. NIR spectra were preprocessed using spectral smoothing and standard normal variate (SNV) to remove multiplicative interferences of scattering and particle size effects. At each concentration of microplastic in soil samples, the five replicated times with different areas were averaged to obtain one representative NIR spectra (Figure 2.2.). The preprocessed NIR spectra was used for the further multivariate data analysis.
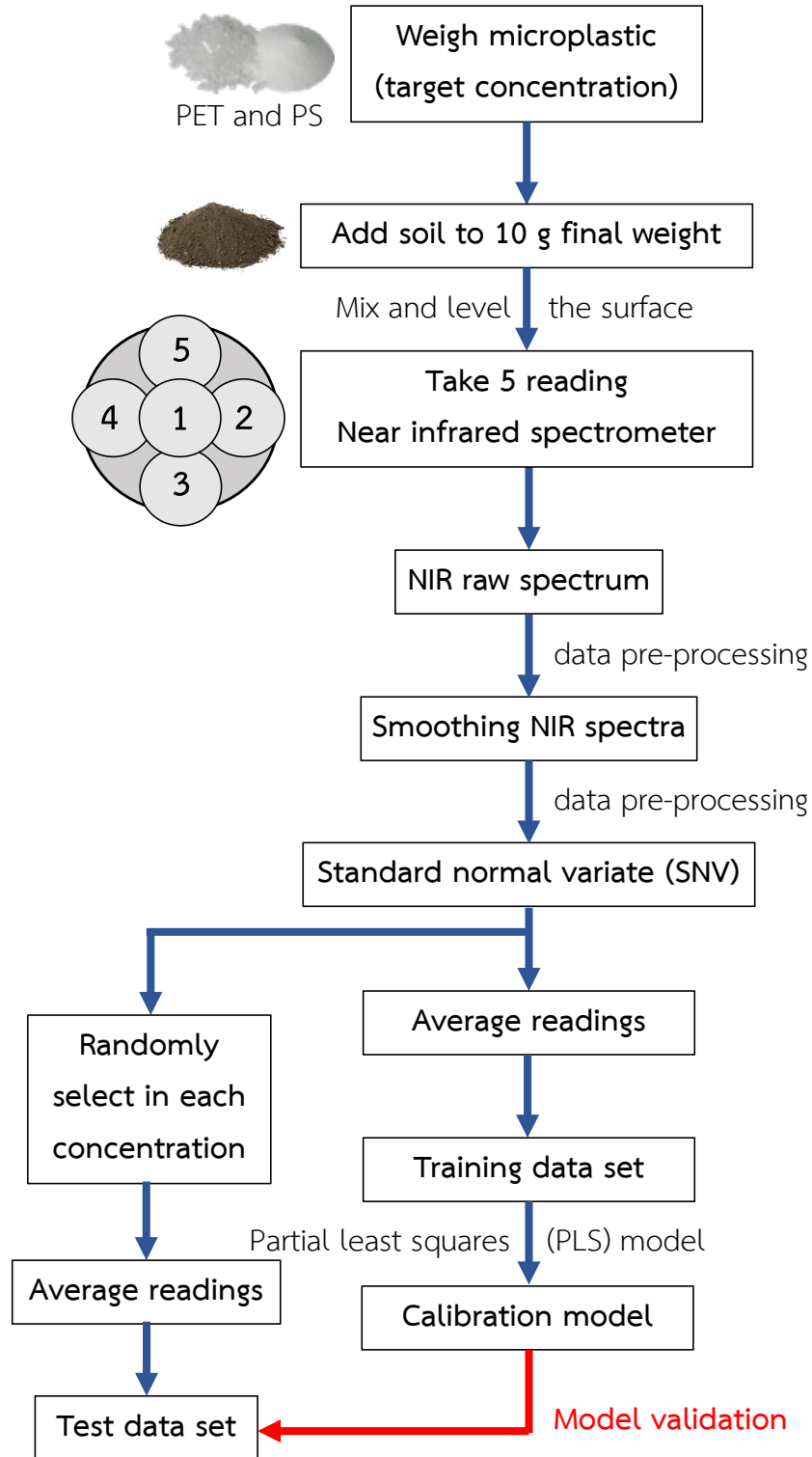
Figure 2.2. Sample preparation and analytical steps to obtain

training and test data set using data pre-processing

2.4. NIR spectra of training and test data set

In the analysis procedures, the total 250 NIR spectra were acquired. To eliminate the unsystematic variation, the NIR of five positions were averaged to a single NIR spectrum as a representative of replication. Therefore, the total of 50 averaged spectra was further used as a training data set for building the calibration model using partial least squares (PLS) to predict PET and PS microplastic concentration in soil. In addition, the total 125 NIR spectra of mixture of microplastics in soil samples were acquired. In the same way, to eliminate the unsystematic variation, the NIR of five positions were averaged to a single NIR spectrum as a representative of replication. Therefore, the total 25 averaged spectra was further used as a test data set for model validation.

2.5. Chemometrics

2.5.1. Regression model for prediction one type of microplastics in soil

The pre-processed NIR spectra were used as a training data set containing 50 spectra in the case of both PET and PS. The training data set was regressed on the concentration of microplastics using partial least squares (PLS) model and the root mean squared error of calibration (*RMSEC*) was calculated. In addition, the root mean squared error of cross validation (*RMSECV*) was calculated to select the optimized number of PLS components. The optimized PLS component was used to generate the PLS model to predict microplastic concentration. The test data set was used to evaluate the model performance by calculating root mean squared error of prediction (*RMSEP*) and R^2.

2.5.2. Regression model for prediction mixture of microplastics in soil

The pre-processed NIR spectra from the model for prediction of one type of microplastic in soil were used as a training data set for prediction mixture of microplastics in soil. Firstly, the training data set was processed by external parameter orthogonalization (EPO) before building PLS model. The EPO spectra for prediction of PET were calculated from the different matrix between the training data set (for PET prediction) and the PET spectrum in algebraic manipulation. In the same way, The EPO spectra for prediction of PS were calculated from the different matrix between the training data set (for PS prediction) and the PS spectrum in linear manipulation. Then, the EPO spectra used as a predictor was regressed on the concentration of microplastics using partial least squares (PLS) model and the root mean

squared error of calibration (*RMSEC*) was calculated. In addition, the root mean squared error of cross validation (*RMSECV*) was calculated to select the optimized number of PLS components. The optimized PLS component was used to generate the PLS model to predict microplastic concentration. The test data set was used to evaluate the model performance by calculating root mean squared error of prediction (*RMSEP*).

2.5.3. Variable selection method

According to the PLS model, to quantify microplastic concentration more precisely, a selection of signal region in NIR spectra from the target analytes might be necessary. Therefore, the wavelengths in NIR spectra related to the variations of PET and PS microplastic were identified and selected systematically. In this study, shaving method was used to select the informative regions. PLS model was performed on the training data set and the PLS coefficients were used to calculated selectivity ratio (*SR*) which are corresponded to the variable importance. Extended detailed following in this step:

Step 1: The NIR spectra of training data set was regressed on microplastic concentration and the number of PLS component was selected from optimization using k-fold cross validation. The PLS model was generated from the selected PLS components to obtain PLS coefficient vector. The test data set was used to evaluate the model performance by *RMSEP* and R^2

Step 2: The coefficient vector was used as a basis and the training data set was projected on to the basis to obtain coefficient of each sample. Then, the loading corresponding to the obtained coefficients was calculated.

Step 3: The coefficients and loading were used to calculate selectivity ratio (*SR*) of each variable corresponding to variable importance. The average selectivity ratio (*SR*) was calculated and used as criteria to determine selected wavelengths. The wavelengths whose selectivity ratio (*SR*) is higher than the average selectivity ratio (*SR*) were selected in spectral region.

Step 4: After selecting wavelengths in spectral region, the selected region considered as training data set was regressed on microplastic concentration and the number of PLS component was selected from optimization using k-fold cross validation again. The new PLS model was generated from the selected region NIR spectra to obtain PLS coefficient vector.

Step 5: The test data set was used to evaluate the model performance by *RMSEP* and R^2 using the PLS coefficients in Step 4. The model performances with and without variable selection were analyzed and discussed in advance.

CHAPTER III

RESULTS AND DISCUSSION

3.1. Microplastic particle size analysis

Microplastics were artificially made by grinding the bulk plastic pieces. The plastic glasses made from polyethylene terephthalate (PET) and the plastic bowls made from polystyrene (PS) were grinded by the blender. The bulk plastic pieces were grinded into smaller particles. In particle size analysis, an optical microscope (OM) was used to measure size of particles. The samples of microplastics were prepared by adding the microplastic particles into water forming a suspension. The suspension was dropped onto a glass slide and was analyzed by optical microscope. The condition used in this experiment was 20x objective lens and 10x eyepiece lens. The obtained pictures were snapped using dark field with the reference red 100 µm scale bar. The particle size of microplastics was analyzed using ImageJ and the particle sizes (> 300 particles) were collected in Excel to investigate size distribution. In size distribution, the model parameters were evaluated using maximum likelihood (ML) criteria to find mean and standard deviation of the particle sizes in probabilistic manipulation. The number of PET microplastic particles was 327 (N=327) and the number of PS microplastic particles was 312 (N=312). The results show that the size of grinded microplastics of PET and PS are in the range of 44.45 ± 66.68 µm (Figure 3.1. (a)) and 103.39 ± 101.13 µm (Figure 3.1. (b)), respectively. The PET microplastic particles were mostly smaller than the PS microplastic particles.
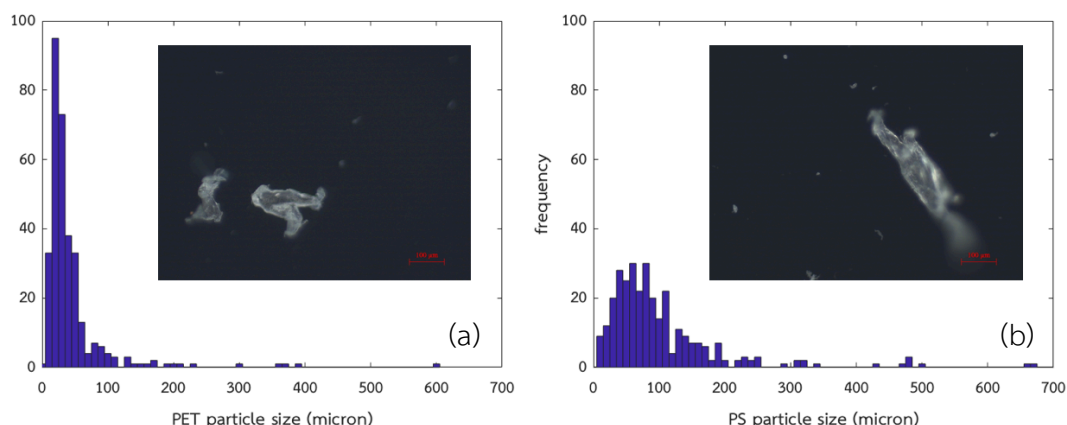
Figure 3.1. microplastic size distribution of (a) polyethylene terephthalate (PET) and (b) polystyrene (PS) measured by optical microscope (OM)

3.2. Near infrared spectrum of microplastics and soil

The near infrared (NIR) spectra of microplastics and soil in reflectance mode were measured using NIR reflectance probe. To visualize the characteristic of NIR spectrum consisting pattern in overtone and combination bands, NIR spectrum of pure soil (green line), pure PET microplastic (blue line) and pure PS microplastic (red line) were shown (Figure 3.2.). It can be seen that soil spectrum mainly differed from microplastic spectrum in the range of 1600 – 1700 nm and 2100 – 2200 nm. In PET microplastic spectrum, the dominant bands are 1400 – 1450 nm, 1650 – 1700 nm and 2100 – 2180 nm corresponding to (1) combination of C-H stretching and blending of -CH$_2$ (methylene) and –CH in aromatic ring, (2) first overtone of C-H stretching in aromatic ring and (3) another combination mode of C-H bond, respectively. In PS microplastic spectrum, the dominant bands are 1350 – 1450 nm, 1650 – 1750 nm and 2150 – 2200 nm corresponding to (1) combination of C-H stretching and blending of -CH$_2$ (methylene) and –CH in aromatic ring, (2) first overtone of C-H stretching in –CH$_2$ (methylene) and –CH of alkyl and aromatic ring and (3) another combination mode of C-H bond in –CH$_2$ (methylene) group, respectively. In soil spectrum, the dominantly distinct bands are 1400 – 1600 nm and 1900 – 2100 nm corresponding to first overtone of O-H and N-H stretching of water in soil and combination of O-H and N-H bonds due to organic matters in soil.
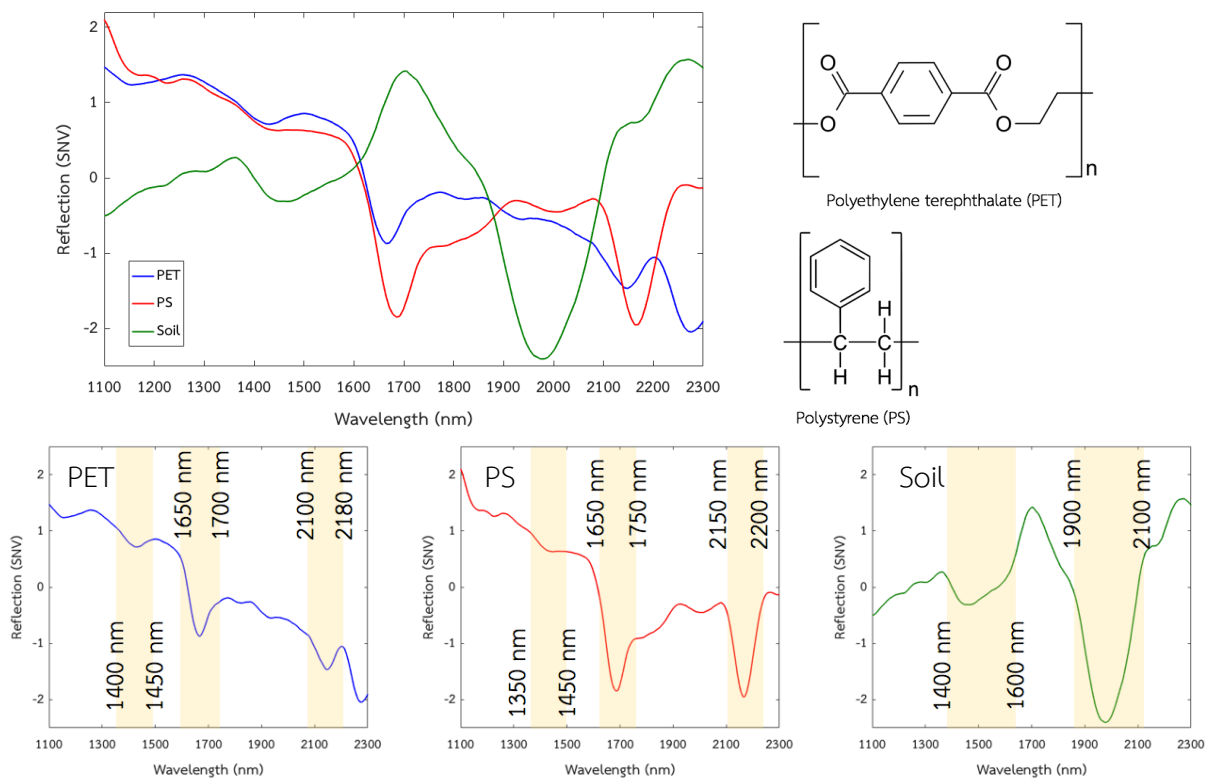
Figure 3.2. NIR spectrum of pure PET (blue), PS (red) microplastic and soil (green)

Although the NIR spectra of pure PET and PS are quite similar, the pattern and variation of each wavelength of these spectra are different. In multivariate analysis, principal component analysis (PCA) was employed to extract the variance of the spectral data. The NIR spectra of each type of microplastics in soil and pure soil were collected in matrix form. The data matrix was analyzed using PCA to extract the orthonormal basis corresponding to the mainly contributed components in spectral data with maximum variance. The scores were obtained by projecting the spectral data matrix onto the orthonormal basis which are first and second principal axes. The important role of PCA is dimensionality reduction of the data matrix by extracting the pattern of data. After applying PCA on the spectral data matrix, the scores were plotted (Figure 3.3.) and the result shows that the scores of microplastics in soil which are PET in soil (red dot) and PS in soil (blue dot) and pure soil (black dot) were clearly separated. According to the result, spectral signals of the microplastics can be extracted using PCA to qualify and quantify microplastics in soil using partial least squares (PLS) model.
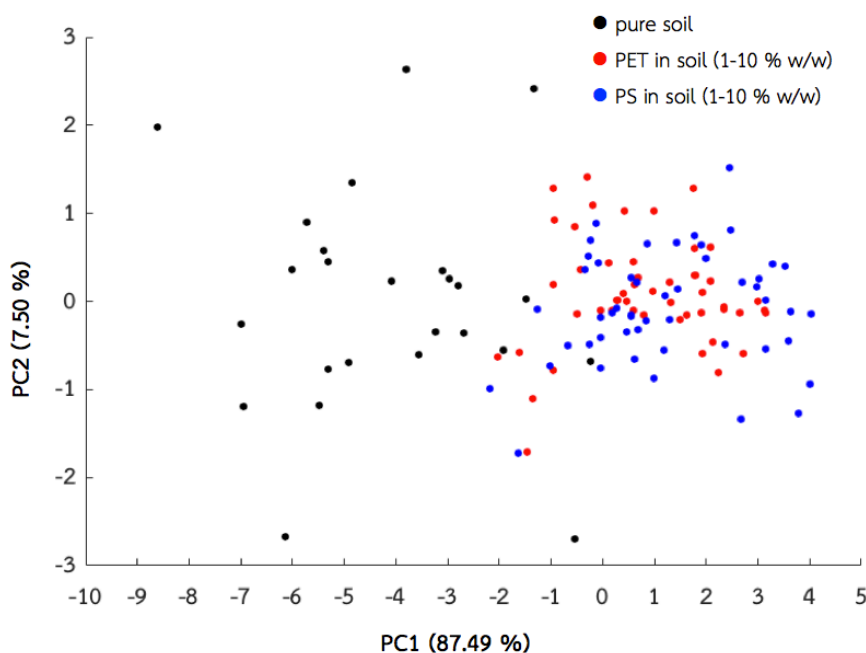
Figure 3.3. PCA scores plot of NIR spectral data, PET in soil (red dot), PS in soil (blue dot) and pure soil (black dot)

3.3. Variations of near infrared spectra of microplastics in soil

The microplastics in soil samples were measured using NIR spectrometer and the NIR spectra were collected at each microplastic concentration (1-10 % w/w). The concentration effects on NIR spectra were analyzed by visualization. To visualize the characteristic overtone and combination bands of analytes, the NIR spectra of microplastics in soil were averaged at

each concentration to obtain only one representative NIR spectra of each microplastic concentration in soil. The NIR spectra of PET in soil at each concentration (1-10 % w/w) were plotted (Figure 3.4. (a)) and can be clearly observed the variation of the NIR spectra with different concentrations. The characteristic overtone and combination bands of soil spectra were changed by adding different amount of microplastic in soil. Although the obtained NIR spectra were similar to soil spectra, the variation of the characteristic bands can be obviously observed by visualization. According to the NIR spectra of PET in soil at each concentration, there are three major bands with high variation were observed including 1100 – 1300 nm, 1450 – 1600 nm and 2150 – 2300 nm corresponding to (1) second overtone of –CH_2 (methylene) around 1140 – 1210 nm, (2) combination of –CH in aromatic around 1440 – 1450 nm and (3) combination of –CH_2 (stretching and bending), respectively. In addition, because of increasing in microplastic concentration, the amount of soil is reduced relatively. In the range of 1100 – 1300 nm and 1450 – 1600 nm, the reflection of the highest PET concentration (red line) was higher than the lowest PET concentration (blue line) due to the amount of soil and microplastics. The larger amount of soil, the more light is absorbed from soil which causes lower reflection of these bands. Although the PET microplastic concentrations were changed, the characteristic band of soil around 1900 – 2100 nm were not changed. Therefore, the NIR spectra are related to microplastic concentration only. In the range of 2150 – 2300 nm, the reflection of the highest PET concentration (red line) was lower than the lowest PET concentration (blue line) because of increasing in PET microplastic concentration. The range of 2150 – 2300 nm obviously corresponds to the combination mode of C-H bond in pattern of pure PET spectra. The variation of these bands are influenced by the amount of PET microplastic in soil. According to the results, it could be implied that these bands are strongly correlated to the concentration of PET microplastic in soil.

In the same way, The NIR spectra of PS in soil at each concentration (1-10 % w/w) were plotted (Figure 3.4. (b)) and can be similarly observed the variation of the NIR spectra with different concentrations. According to the NIR spectra of PS in soil at each concentration, there are three major bands with high variation were observed including 1100 – 1600 nm, 1650 – 1750 nm and 2150 – 2300 nm corresponding to (1) second overtone of –CH_2 (around 1140 – 1210 nm) and combination of –CH in aromatic (around 1440 – 1450 nm), (2) first overtone of –CH stretching (around 1755 – 1775 nm) and (3) combination of –CH_2 (around 2300 nm), respectively. In the range of 1100 – 1600 nm, the reflection of the highest PS

concentration (red line) was higher than the lowest PS concentration (blue line) due to the amount of soil and microplastics. Although The PS microplastic concentrations were changed, the characteristic band of soil around 1900 – 2100 nm were not changed. Similarly, the NIR spectra are related to microplastic concentration only. In the range of 1650 – 1750 nm and 2150 – 2300 nm, the reflection of the highest PS concentration (red line) was lower than the lowest PS concentration (blue line) because of increasing in PS microplastic concentration. These bands, 1650 – 1750 nm and 2150 – 2300 nm, obviously correspond to the first overtone of C-H stretching and combination of vibrational motion of C-H bond in $–CH_2$ (methylene) group of polystyrene (PS) which are similar to the pattern of pure PS spectra. The variation of these bands are influenced by the amount of PS microplastic in soil. According to the results, it could be implied that these bands are strongly correlated to the concentration of PS microplastic in soil. Therefore, in order to visualize the variability direction of NIR spectra along with overtone and combination bands pattern of microplastic in soil is required.
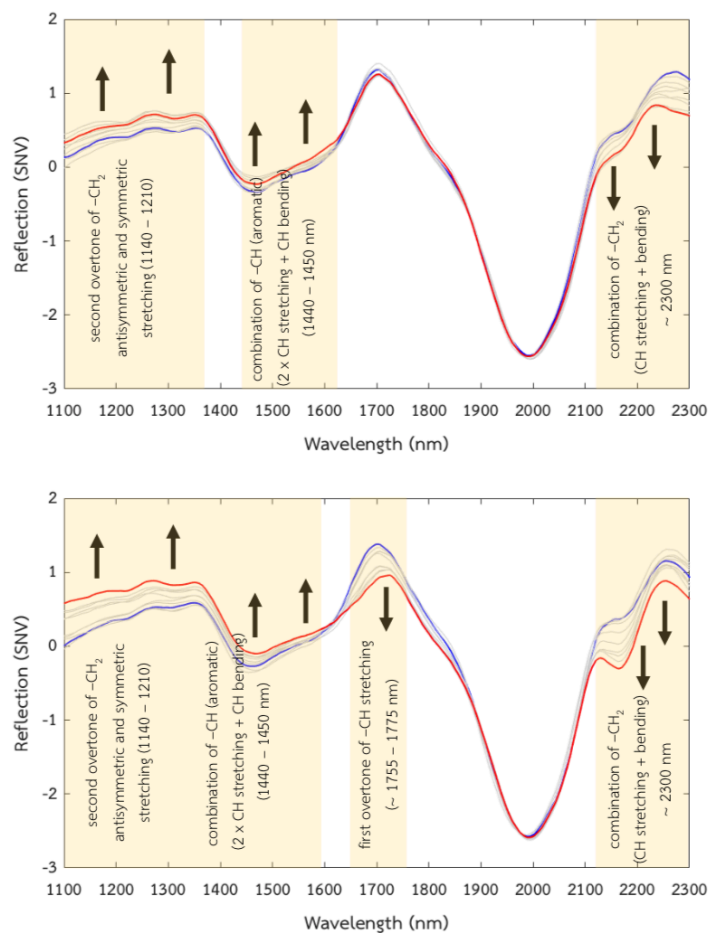

Figure 3.4. NIR spectrum of (a) PET in soil and (b) PS in soil with highest concentration (red line) and lowest concentration (blue line) and the arrows demonstrate the direction of variability of NIR spectrum

3.4. Calibration and prediction of microplastics in soil

Partial least squares regression (PLSR) model was employed to build the model for detection and determination of microplastics in soil using near infrared spectra as a training data set. Partial least squares (PLS) model is a linear regression model for multivariate analysis problem using linear least squares formulation to minimize prediction error. In calculation, near infrared spectra with all wavelengths were used as a variable to generate multivariate linear regression model. The data matrix collecting from near infrared spectra is an input data called predictor which strongly correlates to a response or microplastic concentration in this case. The data matrix consists of near infrared spectra whose row and column are sample and wavelengths respectively. In PLSR algorithm, the data matrix as an input data was regressed on response (concentration in this case) by optimization with loading of both data matrix and response. In near infrared spectroscopy, the data matrix has a number of columns or variables more than number of row or samples which caused the regression coefficients does not exist and has not a unique solution. Therefore, the partial least squares (PLS) model was frequently used to generate model in spectroscopy. In model building method, the training data set consists of total 50 near infrared spectra which are divided into 10 concentration points. Each concentration points have 5 spectra which are measured repeatedly. Therefore, the data matrix of training data set has 50 samples and 176 wavelengths.

In model selection method, the data matrix was fitted to response vector and a root mean squared error of calibration (*RMSEC*) for polyethylene terephthalate (PET) (Figure 3.5. (a)) and polystyrene (PS) (Figure 3.5. (b)) can be evaluated at each number of PLS components extracted from training data set. The *RMSEC* of each PLS components reduces while increasing in number of PLS components due to decreasing in model bias, in other words, the model is overfitting to training data set while increasing in number of PLS components. The optimized number of PLS components was selected at the optimum point of root mean squared error of cross validation (*RMSECV*) using resampling method, k-fold cross validation when k = 5, 10, 15 and 20 partitions. The optimized number of PLS components is 2 as it gave the lowest RMSE for most case in the model for predicting polyethylene terephthalate (PET) concentration (Figure 3.5. (c)) and optimized number of PLS components is 2 in the model for predicting polystyrene (PS) concentration (Figure 3.5. (d)) in soil samples. The test data set consisting of total 50 additional spectra for each type of microplastics was used to evaluated the PLSR model performance. The results show that the root mean squared error of prediction

(RMSEP) for polyethylene terephthalate (PET) was 1.30 % w/w with R^2 = 0.8030 (Figure 3.5. (e)) and the root mean squared error of prediction (*RMSEP*) for polystyrene (PS) was 1.19 % w/w with R^2 = 0.8253 (Figure 3.5. (f)). The PLSR model is a potential model for detection microplastics in soil using near infrared spectroscopy. Furthermore, the prediction of polystyrene (PS) has lower root mean squared error of prediction (*RMSEP*) and R^2 is greater than the prediction of polyethylene terephthalate (PET) because the NIR spectra of PS microplastic in soil have obviously variation in range of 2150 – 2300 nm corresponding to the pattern of pure PS spectrum. The PLSR model can be used to predicting microplastic concentration accurately in soil samples which consist of single microplastic component.
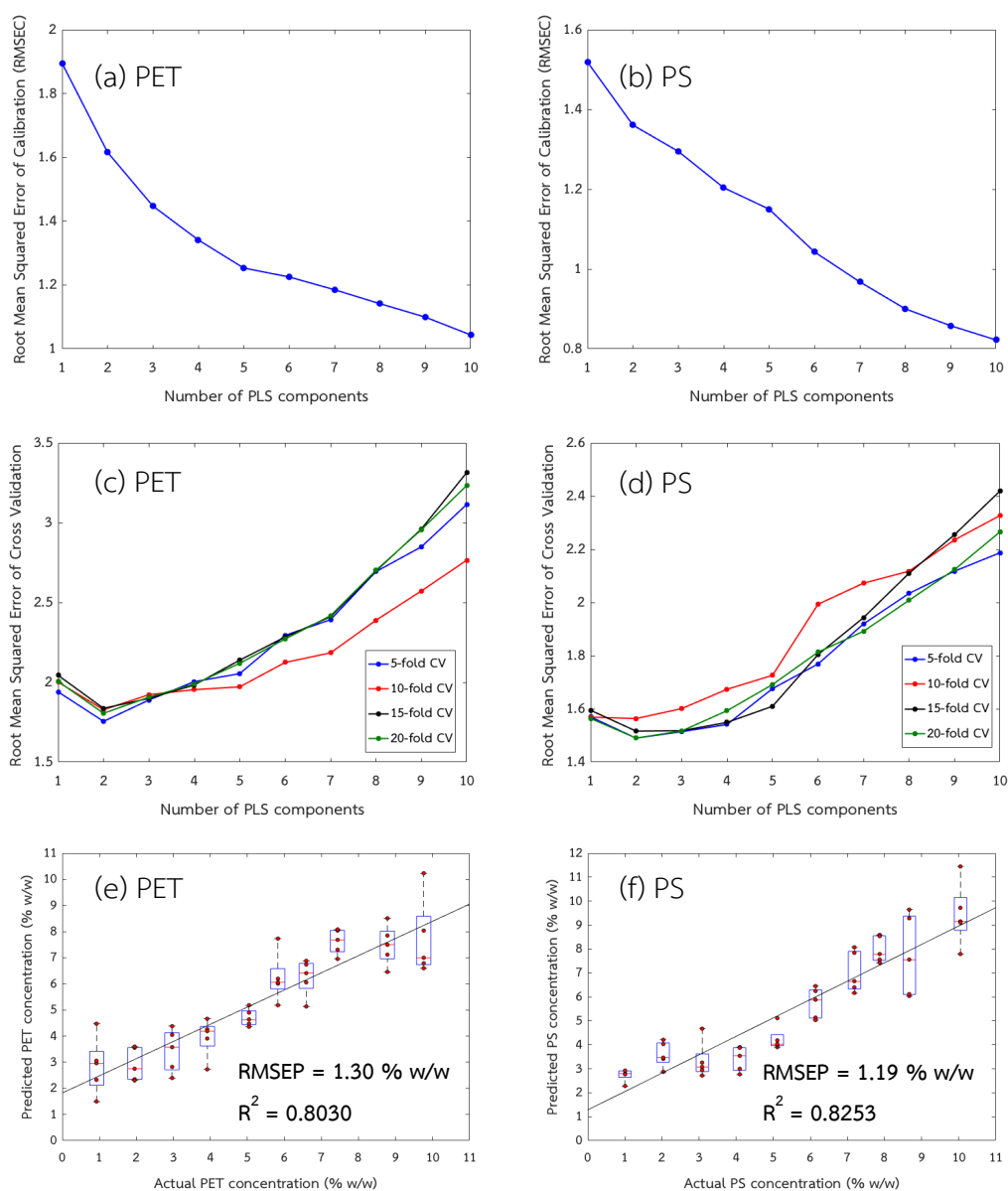

Figure 3.5. (a), (b) *RMSEC* plot (c), (d) *RMSECV* plot and (e), (f) correlation plot between actual and predicted concentration for PET and PS in soil, respectively

In general, partial least squares regression (PLSR) model in near infrared spectroscopy was employed to create multivariate linear regression with full spectral range 1100 – 2300 nm. However, some wavelengths or variables do not relevant to the underlying information of the targeted samples and these variables might be considered as some residuals or noise. The variable selection or feature selection method was used to increase performance of the model by systematically eliminate irrelevant variables using mathematical and statistical method. In spectroscopic aspect, the variable selection is a powerful approach to extract the informative signals which are correlated to chemical patterns. In this study, shaving method was used as a variable selection method. The parameter of each wavelengths called selectivity ratio (SR) was calculated using targeted projection to the PLS coefficients. The PLS coefficients were calculated from normal equation previously and the training data set was projected onto the PLS coefficients vector whose norm is one and the scores were obtained. The selectivity ratio (*SR*) was calculated by the ratio between the explained variance and the residual variance from covariance matrix extracted from training data set using the scores and loading. The variables with selectivity ratio (*SR*) higher than average of selectivity ratio (*SR*) were selected and the variables whose selectivity ratio (SR) is lower than average of selectivity ratio (*SR*) were eliminated from the model. The selectivity ratio (*SR*) for polyethylene terephthalate (PET) model (Figure 3.6. (a)) and the selectivity ratio (*SR*) for polystyrene (PS) model (Figure 3.6. (b)) was plotted with the dotted red line which represents the average of selectivity ratio (*SR*) to analyze the significant wavelengths for variable selection. The wavelengths for polyethylene terephthalate (PET) prediction model were selected in range 1433 – 1535 nm, 1650 – 1678 nm and 2127 – 2305 nm corresponding to (1) combination of C-H stretching and blending of -CH_2 (methylene) and –CH in aromatic ring, (2) first overtone of C-H stretching in aromatic ring and (3) another combination mode of C-H bond, respectively and whose selectivity ratio (SR) are higher than the average of selectivity ratio (SR) (Figure 3.6. (c)) and the wavelengths for polystyrene (PS) prediction model were selected in range 1439 – 1507 nm, 1630 – 1808 nm and 2184 – 2305 nm corresponding to (1) combination of C-H stretching and blending of -CH_2 (methylene) and –CH in aromatic ring, (2) first overtone of C-H stretching in –CH_2 (methylene) and –CH of alkyl and aromatic ring and (3) another combination mode of C-H bond in –CH_2 (methylene) group, respectively (Figure 3.6. (d)). The main wavelengths from shaving method are related to the variability of NIR spectra whose overtones and combination modes of the characteristic bands mentioned previously. The

selected variables of each model were used to fit the partial least squares regression model again and the number of PLS components was optimized again with the selected regions of predictor in training data set. The optimized number of PLS components was selected at the optimum point of root mean squared error of cross validation (RMSECV) using resampling method, k-fold cross validation when k = 5, 10, 15 and 20 partitions. The optimized number of PLS components is 2 as it gave the lowest RMSE for most case in the model for predicting polyethylene terephthalate (PET) concentration (Figure 3.5. (e)) and optimized number of PLS components is 3 in the model for predicting polystyrene (PS) concentration (Figure 3.5. (f)) in soil samples. When partial least squares regression (PLSR) was fitted again using selected variables as a predictor, the test data set was used to evaluated the PLSR model performance. The results show that the root mean squared error of prediction (RMSEP) for polyethylene terephthalate (PET) was 1.35 % w/w with R^2 = 0.7854 (Figure 3.5. (g)) and the root mean squared error of prediction (RMSEP) for polystyrene (PS) was 1.09 % w/w with R^2 = 0.8582 (Figure 3.5. (h)). The PLSR model combined with variable selection will reduce the number of variables which are fitted in the model. The number of selected variables for polyethylene terephthalate (PET) is 47 variables and the number of selected variables for polystyrene (PS) is 56 variables whereas the total variables before variable selection is 176 variables. The lower number of variables is preferred for model selection which excludes residual or noise from some uninformative variables. Although the root mean squared error of prediction (RMSEP) of predictive model for polyethylene terephthalate (PET) was slightly increased in comparison to the model with all variables but it is acceptable. On the other hand, the root mean squared error of prediction (*RMSEP*) of predictive model for polystyrene (PS) were obviously reduced with increasing in model performance. Therefore, the shaving method as a variable selection method can reduce number of variables in the model with maintaining a acceptable model performance or increasing in model performance.

It can be concluded that the promising values (Table 3.1.) for root mean squared error of prediction (*RMSEP*) were obtained to be 1.35 % w/w (R^2 = 0.7854) and 1.09 % w/w (R^2 = 0.8582) corresponding to the prediction of polyethylene terephthalate (PET) and polystyrene (PS), respectively. Therefore, it might be implied that the PLSR model can be used to predict amount of microplastic in soil.
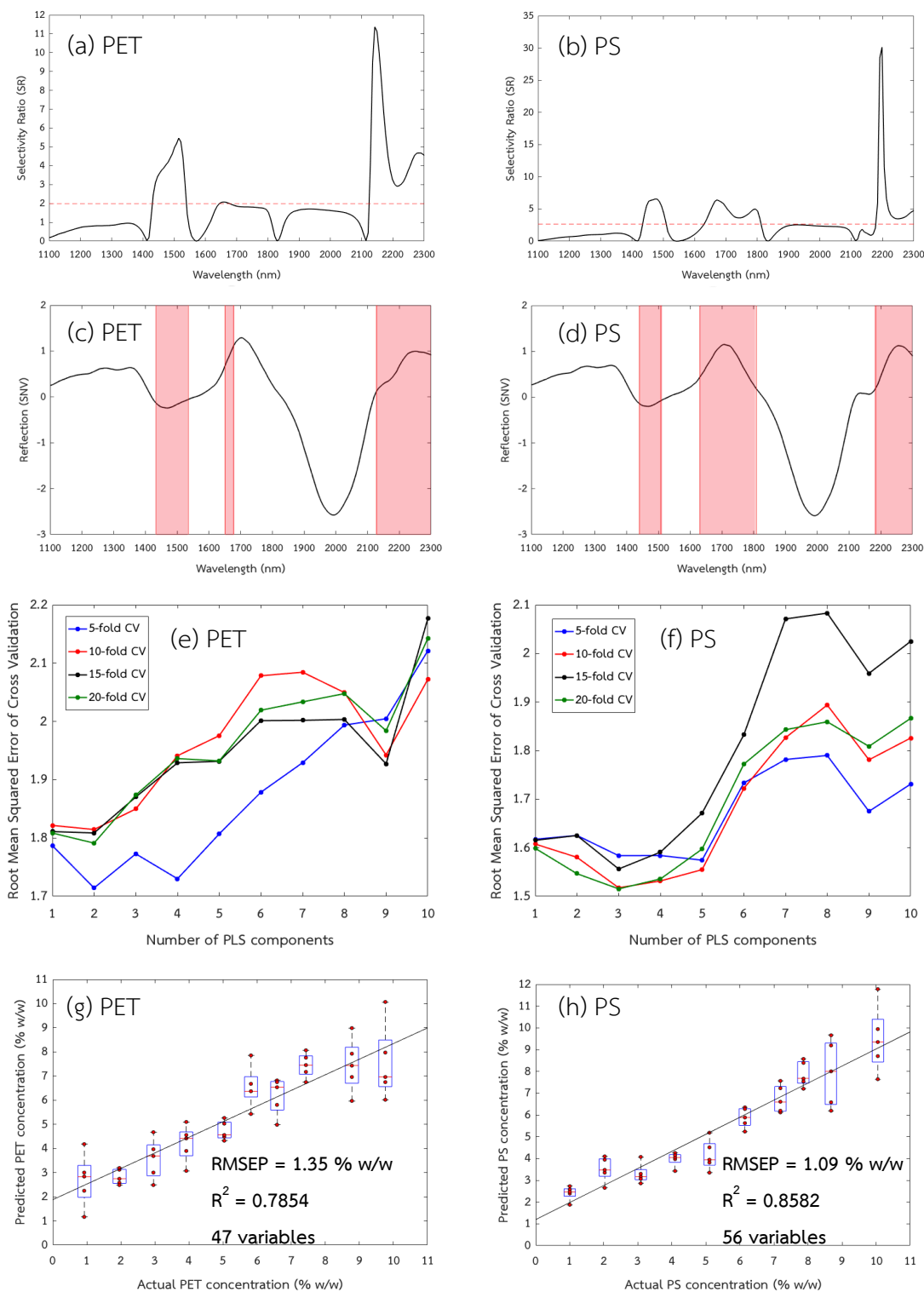
Figure 3.6. (a), (b) selectivity ratio plot with the average selectivity ratio in red line
(c), (d) selected wavelengths in NIR spectra (e), (f) *RMSECV* plot
and (g), (h) correlation plot between actual and predicted concentration
for PET and PS in soil, respectively

Generative model	Prediction of PET	Prediction of PS
Number of training samples	50 (smoothing + SNV)	50 (smoothing + SNV)
Number of test samples	50 (smoothing + SNV)	50 (smoothing + SNV)
Optimized number of PLS component (n_{comp})	2 (176 variables)	2 (176 variables)
RMSEP (test data set)	1.30 % w/w	1.19 % w/w
R^2	0.8030	0.8253
Selected region from shaving method	1433 – 1535 nm 1650 – 1678 nm and 2127 – 2305 nm (47 variables)	1439 – 1507 nm 1630 – 1808 nm and 2184 – 2305 nm (56 variables)
Optimized number of PLS component (n_{comp})	2 (47 variables)	3 (56 variables)
RMSEP (test data set)	1.35 % w/w	1.09 % w/w
R^2	0.7854	0.8582

Table 3.1. Summary of the model performance for detection of PET and PS microplastic in soil with and without variable selection

3.5. Calibration and Prediction of mixture of microplastics in soil

The near infrared spectra of microplastics which are composed of polyethylene terephthalate (PET) and polystyrene (PS) in soil are extracted for prediction the concentration of microplastics. The data pre-processing method called External Parameter Orthogonalization (EPO) was used to extract each spectrum or basis of each considered microplastic components. However, External Parameter Orthogonalization (EPO) was used when the signals of analyte are completely overlapped which cannot be determined directly. In this case, the signals of each microplastics are combined as mixture spectra. The data pre-processing method, External Parameter Orthogonalization (EPO), was used to extract of polyethylene terephthalate (PET) and polystyrene (PS) basis, respectively. Firstly, a different spectra between training set and microplastic average spectra was calculated as a matrix. The eigenvalue of the different spectra was calculated from covariance matrix of different spectra using eigenvalue decomposition or singular value decomposition (SVD). Secondly, the number

of eigenvector or EPO components was selected by the significant component with large eigenvalue. In this case, the number of EPO component for extracting polyethylene terephthalate (PET) was 1 corresponding to largest eigenvalue (Figure 3.7. (a)). Similarly, the number of EPO component for extracting polystyrene (PS) was 1 corresponding to the largest eigenvalue (Figure 4.7. (c)). The selected number of EPO components were 1 because the accumulative variance that calculated from eigenvalue was more than 99 percent of the total extracted variance. Finally, the basis of considered analyte was calculated in algebraic manipulation. The EPO spectra for prediction polyethylene terephthalate (PET) (Figure 3.7. (b)) was obtained from projection the training data set onto the subspace of the basis. The EPO spectra for prediction polyethylene terephthalate (PET) have pattern which is similar to polyethylene terephthalate (PET) spectrum at around 1400 – 1500 nm, 1650 – 1700 nm and at 2150 nm, but the EPO spectra also include the soil signal around 1900 - 1950 nm. The EPO spectra for prediction polystyrene (PS) (Figure 3.7. (d)) have pattern which is similar to polystyrene (PS) spectrum at around 1400 – 1500 nm, 1650 – 1700 nm and 2150 – 2200 nm, but the EPO spectra also include the soil signal around 1950 - 2050 nm. Both EPO spectra are different according to the training data set are projected onto the different signal's subspace. The EPO spectra was used for prediction of the microplastic concentrations of the mixture microplastic in soil samples consisting of both polyethylene terephthalate (PET) and polystyrene (PS) using the partial least squares regression with and without variable selection. The EPO spectra showed that EPO minimized the variability of soil spectra induced by other components. This adds flexibility to the implementation of transformed EPO spectra in predicting concentration of each type of microplstics in mixture microplastics in soil samples. EPO would be an important and essential part for the prediction microplastic in soil in NIR spectroscopy based soil sensing technology, because the chemical signal of organic matters in soil have to be exclueded for improving the prediction performance.
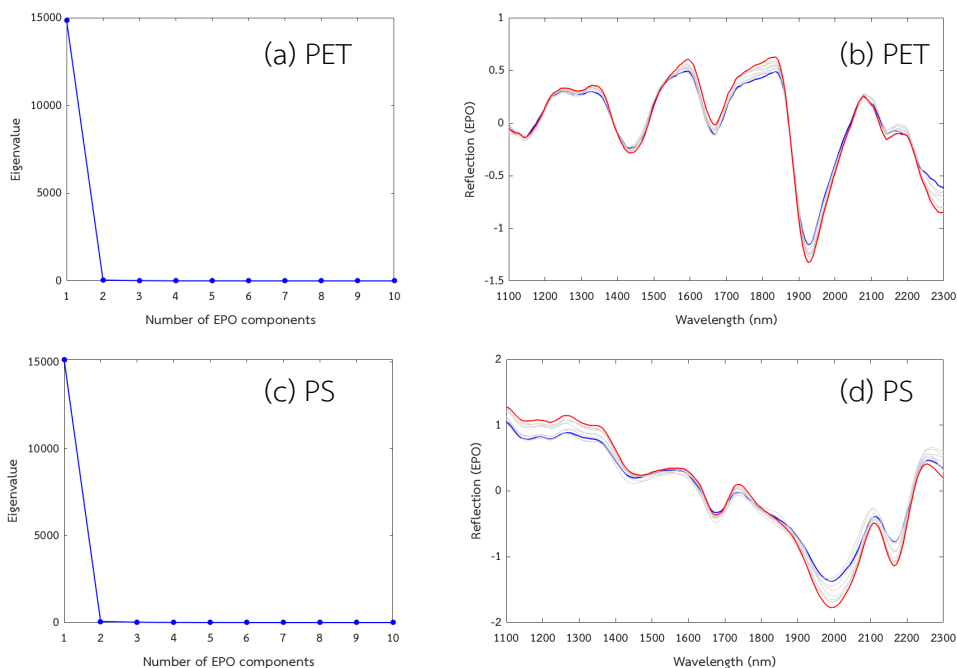
Figure 3.7. (a), (c) extracted EPO components corresponding to their eigenvalue of PET and PS, respectively. (b), (d) EPO spectra for prediction of PET and PS, respectively with the highest (red line) and the lowest (blue line) microplastic concentration

The EPO spectra was used as a training and test data set for prediction the microplastic concentrations of mixture microplastic in soil samples. The EPO spectra for prediction of polyethylene terephthalate (PET) and polystyrene (PS) has completely different from each others due to the signal's pattern of different type of microplastics. In model selection method, the EPO training data matrix was fitted to response vector and a root mean squared error of calibration (RMSEC) for polyethylene terephthalate (PET) (Figure 3.8 (a)) and polystyrene (PS) (Figure 3.8. (b)) can be evaluated at each number of PLS components extracted from EPO training data set. The RMSEC of each PLS components reduces while increasing in number of PLS components due to decreasing in model bias, in other words, the model is overfitting to training data set while increasing in number of PLS components. The optimized number of PLS components was selected at the optimum point of root mean squared error of cross validation (RMSECV) which is similar to the previous model selection using resampling method, k-fold cross validation when k = 5, 10, 15 and 20 partitions. The optimized number of PLS components is 1 in the model for predicting polyethylene terephthalate (PET) concentration (Figure 3.8. (c)) and proper number of PLS components is 3 in the model for predicting polystyrene (PS) concentration (Figure 3.8. (d)) of mixture

microplastic in soil samples. The EPO test data set consisting of total 25 spectra of mixture microplastics in soil samples was used to evaluated the PLSR model performance. The results show that the root mean squared error of prediction (RMSEP) for polyethylene terephthalate (PET) was 3.17 % w/w (Figure 3.8. (e)) and the root mean squared error of prediction (RMSEP) for polystyrene (PS) was 3.26 % w/w (Figure 3.8. (f)). The PLSR model can predict only high microplastic concentration around 5 – 10 % w/w approximately of each microplastic components but the prediction of low microplastic concentration around 1 – 3 % w/w approximately has an large error significantly. However, the samples which have high microplastic concentration can be predicted accurately especially polystyrene (PS) in mixture microplastic in soil samples.
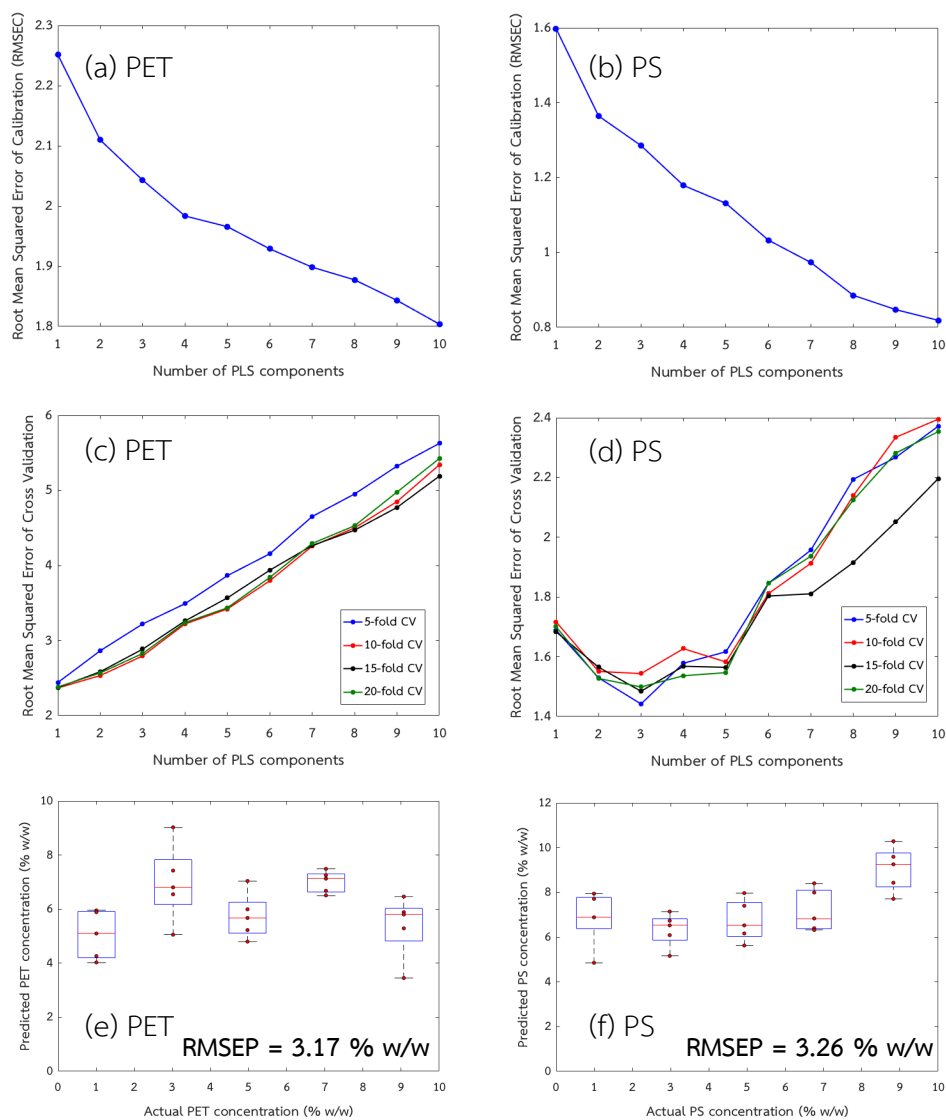

Figure 3.8. (a), (b) RMSEC plot (c), (d) RMSECV plot and (e), (f) correlation plot between actual and predicted concentration for PET and PS in soil, respectively using EPO

The prediction of microplastic concentration of mixture microplastic in soil sample is used the EPO spectra as a training data set for partial least squares regression (PLSR) model with full spectral range previously. The variable selection in partial least squares regression (PLSR) of EPO spectra is similar to the shaving method. The criteria parameter is variance of EPO reflection value of each wavelengths or variables which is comparable to the selectivity ratio (SR) in targeted projection. The variance of EPO reflection value of each wavelengths was calculated from the EPO training data set. The variance for polyethylene terephthalate (PET) (Figure 3.9. (a)) and for polystyrene (PS) (Figured 3.9. (b)) prediction is plotted against each wavelength and the dotted red line is the average of the variance of EPO reflection value. The variables whose variance is higher than average of variance were selected and the variables whose variance are lower than average of variance were eliminated from the model. The wavelengths for polyethylene terephthalate (PET) prediction model were selected in range 1100 – 1106 nm, 1582 – 1842 nm, 1897 – 1980 nm and 2233 – 2305 nm whose variance are higher than the average of variance (Figure 3.9. (c)) and the wavelengths for polystyrene (PS) prediction model were selected in range 1100 – 1242 nm, 1918 – 2106 nm, 2141 – 2191 nm and 2248 – 2305 nm whose variance are higher than the average of variance (Figure 3.9 (d)). The mainly selected wavelengths from variance criteria are related to the variability of NIR spectra whose overtone and combination modes of the characteristic bands that were mentioned previously. The selected variables of each model were used to fit the partial least squares regression model again and the number of PLS components was optimized again with the selected region of predictor in training data set. The optimized number of PLS components was selected at the optimum point of root mean squared error of cross validation (*RMSECV*) using resampling method, k-fold cross validation when k = 5, 10, 15 and 20 partitions. The optimized number of PLS components is 1 as it gave the lowest RMSE for most case in the model for predicting polyethylene terephthalate (PET) concentration (Figure 3.9. (e)) and optimized number of PLS components is 2 in the model for predicting polystyrene (PS) concentration (Figure 3.9. (f)) in soil samples. When partial least squares regression (PLSR) was fitted again using remained variables as a predictor, the test data set was used to evaluated the PLSR model performance. The results show that the root mean squared error of prediction (RMSEP) for polyethylene terephthalate (PET) was 3.00 % w/w (Figure 3.9. (g)) and the root mean squared error of prediction (*RMSEP*) for polystyrene (PS) was 3.53 % w/w (Figure 3.9. (h)). The PLSR model combined with variable selection will reduce the number of variables which

are fitted in the model. The number of selected variables for polyethylene terephthalate (PET) is 65 variables and the number of selected variables for polystyrene (PS) is 67 variables whereas the total variables before variable selection is 176 variables. The lower number of variables is preferred for model selection which is excluded residual or noise from elimination of some uninformative variables. The root mean squared error of prediction (*RMSEP*) of polyethylene terephthalate (PET) prediction was obviously lower than root mean squared error of prediction (RMSEP) of the previous model which is fitted with all spectral range but the root mean squared error of prediction (*RMSEP*) of polystyrene (PS) prediction was increased. The variance of EPO reflection value can help to select the range of informative variables. Although the root mean squared error of prediction (*RMSEP*) is lower after applying variable selection in case of detection of PET in soil, the prediction of low microplastic concentration around 1 - 3 % w/w approximately still has an large error significantly. However, the samples which have high microplastic concentration around 5 - 10 % w/w approximately can be predicted more accurately in polyethylene terephthalate (PET) and polystyrene (PS) in mixture microplastic in soil samples. According to the PCA scores plot (Figure 3.3.) of the NIR spectra, the resulted show that polyethylene terephthalate (PET) and polystyrene (PS) in soil samples cannot be extracted completely using principal component analysis (PCA). According to the extracted principal components and the data were regressed with the response using partial least squares (PLS) model, the microplastic component which is higher concentration than the others may be interference and caused to the large error for predicting the low concentration of mixture mixture microplastic in soil samples.
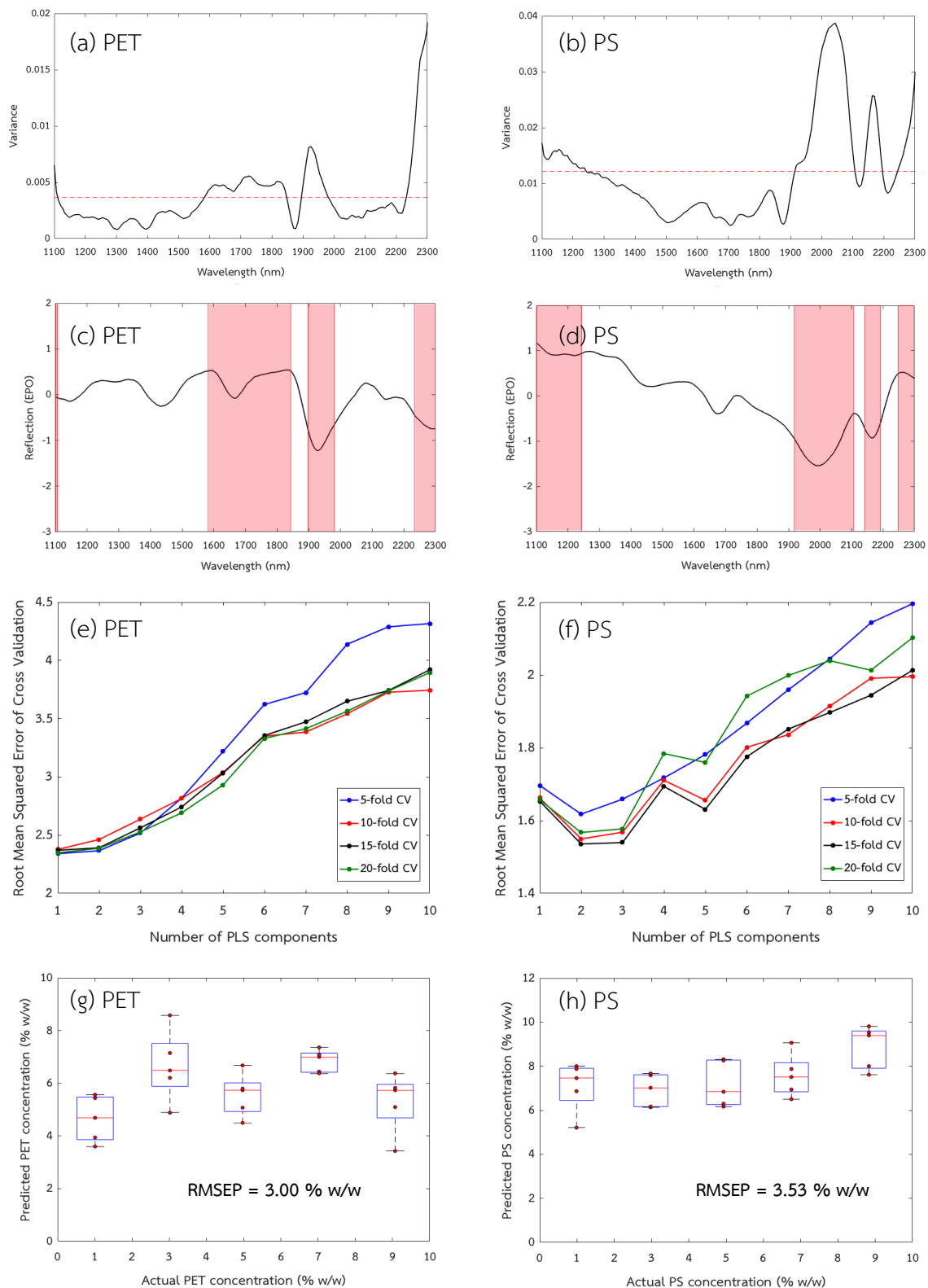
Figure 3.9. (a), (b) variance plot with the average variance in red line
(c), (d) selected wavelengths in EPO spectra (e), (f) RMSECV plot
and (g), (h) correlation plot between actual and predicted concentration
for PET and PS in mixture of microplastic in soil, respectively

It can be concluded that the promising values (Table 3.2.) for root mean squared error of prediction (RMSEP) were obtained to be 3.00 % w/w and 3.53 % w/w corresponding to the prediction of polyethylene terephthalate (PET) with 65 variables and polystyrene (PS) with 76 variables, respectively. Therefore, it might be concluded that the PLSR model combined with external parameter orthogonalization (EPO) as a pre-processing data treatment can be used to predict microplastic concentrations in soil especially in high concentration at 5 – 10 % w/w approximately.

Generative model	Prediction of PET	Prediction of PS
Number of training samples	50 (smoothing + SNV + EPO)	50 (smoothing + SNV + EPO)
Number of test samples	25 (smoothing + SNV + EPO)	25 (smoothing + SNV + EPO)
Optimized number of PLS component (n_{comp})	1 (176 variables)	3 (176 variables)
RMSEP (test data set)	3.17 % w/w	3.26 % w/w
Selected region from shaving method	1100 – 1106 nm 1582 – 1842 nm 1897 – 1980 nm and 2233 – 2305 nm (65 variables)	1100 – 1242 nm 1918 – 2106 nm 2141 – 2191 nm and 2248 – 2305 nm (67 variables)
Optimized number of PLS component (n_{comp})	1 (65 variables)	2 (67 variables)
RMSEP (test data set)	3.00 % w/w	3.53 % w/w

Table 3.2. Summary of the model performance for detection of PET and PS in mixture of microplastic in soil samples with and without variable selection using EPO spectra

CHAPTER IV

CONCLUSIONS

This study has demonstrated the potential of NIR spectroscopy combined with chemometrics to predict polyethylene terephthalate (PET) and polystyrene (PS) microplastic concentration in soil. The regression model, partial least squares (PLS), was proposed in order to construct the calibration model from training data set (1-10 % w/w) and use the model to predict microplastic concentration in soil. Three stages of methodology including data pre-processing, partial least squares (PLS) regression and variable selection were applied to NIR spectral data in order to obtain the optimized PLS model. The NIR spectra of microplastics in soil samples were collected in the range of 1100 – 2300 nm using reflectance mode, integration time of 100 milliseconds, 2 averaged scans with boxcar smoothing windows of 4. Standard normal variate (SNV) combined with smoothing is useful data pretreatment for raw spectra data in order to remove unsystematic signals.

According to characteristic overtone and combination bands of microplastics, PET spectrum shows dominated bands of 1400 – 1450 nm, 1650 – 1700 nm and 2100 – 2180 nm corresponding to (1) combination of C-H stretching and blending of -CH_2 (methylene) and –CH in aromatic ring, (2) first overtone of C-H stretching in aromatic ring and (3) another combination mode of C-H bond, respectively and PS spectrum shows dominated bands of 1350 – 1450 nm, 1650 – 1750 nm and 2150 – 2200 nm corresponding to (1) combination of C-H stretching and blending of -CH_2 (methylene) and –CH in aromatic ring, (2) first overtone of C-H stretching in –CH_2 (methylene) and –CH of alkyl and aromatic ring and (3) another combination mode of C-H bond in –CH_2 (methylene) group, respectively. According to the NIR spectra of PET in soil at each concentration, there are three major bands with high variation were observed including 1100 – 1300 nm, 1450 – 1600 nm and 2150 – 2300 nm corresponding to (1) second overtone of –CH_2 (methylene) around 1140 – 1210 nm, (2) combination of –CH in aromatic around 1440 – 1450 nm and (3) combination of –CH_2 (stretching and bending), respectively. The NIR spectra of PS in soil can be similarly observed the variation of the NIR spectra with different concentrations. According to the NIR spectra of PS in soil at each concentration, there are three major bands with high variation were observed including 1100 – 1600 nm, 1650 – 1750 nm and 2150 – 2300 nm corresponding to (1) second overtone of –CH_2 (around 1140 – 1210 nm) and combination of –CH in aromatic (around 1440 – 1450 nm), (2) first overtone of –CH

stretching (around 1755 – 1775 nm) and (3) combination of –CH$_2$ (around 2300 nm), respectively. There are mainly differences of characteristic bands of PET and PS in soil, it is seemly possible to extract each component of microplastic in soil.

In case of experimental part, it was used to quantify microplastic concentration in soil samples. The promising values for root mean squared error of prediction (*RMSEP*) were obtained to be 1.30 % w/w (without variable selection) and 1.09 % w/w (with variable selection) corresponding to PET and PS, respectively. These observations are in good agreement with high R^2. The prediction of microplastic concentration of mixture microplastics in soil samples is used the EPO spectra as a training data set for PLS model. It can be concluded that the promising values for root mean squared error of prediction (*RMSEP*) were obtained to be 3.00 % w/w (with variable selection) and 3.26 % w/w (without variable selection). Therefore, it might be implied that our PLS model approach can be used to predict microplastic concentration in soil.

REFERENCES

[1] Defu He., Yongming Luo., Shibo Lu., Mengting Liu., Yang Song., Lili Lei. Microplastics in soils: Analytical methods, pollution characteristics and ecological risks. *Trends in Analytical Chemistry*. **2018**. *109*. 163-172.

[2] Li, J.; Liu, H.; Paul Chen, J., Microplastics in freshwater systems: A review on occurrence, environmental effects, and methods for microplastics detection. *Water Research* **2018,** *137*, 362-374.

[3] Anger, P. M.; von der Esch, E.; Baumann, T.; Elsner, M.; Niessner, R.; Ivleva, N. P., Raman microspectroscopy as a tool for microplastic particle analysis. *TrAC Trends in Analytical Chemistry* **2018,** *109*, 214-226.

[4] Araujo, C. F.; Nolasco, M. M.; Ribeiro, A. M. P.; Ribeiro-Claro, P. J. A., Identification of microplastics using Raman spectroscopy: Latest developments and future prospects. *Water Research* **2018,** *142*, 426-440.

[5] Corradini, F.; Bartholomeus, H.; Huerta Lwanga, E.; Gertsen, H.; Geissen, V., Predicting soil microplastic concentration using vis-NIR spectroscopy. *Science of The Total Environment* **2019,** *650*, 922-932.

[6] H.W. Siesler, Y. Ozaki, S. Kawata, H.M.Heise. Near-Infrared Spectroscopy., Principles, Instruments, Applications., Wiley-VCH. Printed in the Federal Republication of Germany. **2002**

[7] Minasny, B.; McBratney, A. B.; Bellon-Maurel, V.; Roger, J. M.; Gobrecht, A.; Ferrand, L.; Joalland, S., Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma* **2011,** *167 - 168*, p. 118 - p. 124.

[8] Yong-Huan Yun., Hong-Dong Li., Bai-Chuan Deng., Dong-Sheng Cao. An overview of variable selection methos in multivariate analysis of near-infrared spectra. *Trends in Analytical Chemistry*. **2019**. *113*. 102-115.

[9] Mireia Farrés., Stefan Y. Platikanov., Stefan L. Tsakovski., Romà Tauler. Comparison of the Variable Importance in Prediction (VIP) and of the Selectivity Ratio (SR) variable selection methods in the analysis of three different data sets. *Journal of Chemometrics*. **2015**. 29. 528-536.

VITAE

Mr. Nuttapat Sinsuvannaporn was born on 27, Jauary 1997, in Bangkok, Thailand. He graduated high school from Suankularb Wittayalai School in 2015. After that, he continued his study in Department of Chemistry, Faculty of Science, Chulalongkorn Univeristy in 2015. Email: nuttapat.sin1997@gmail.com