Principle Component Analysis (PCA) - Classification of Arabica green bean coffee of North Sumatera Using FT–NIRS

S D Siregar, A Rindang and P C Ayu*

Department of Agricultural Engineering, Faculty of Agriculture, Universitas Sumatera Utara, Medan, Sumatera Utara, Indonesia.

E-mail: *pputricandra@usu.ac.id

Abstract. Coffee is one of the important export commodities in Indonesia. Indonesia is among the top four coffee producers in the world after Brazil, Vietnam and Columbia. A total of 301 samples of North Sumatera Arabica green coffee bean were obtained from different districts namely Dairi, Humbang Hasudutan and Mandailing Natal. In this research, Principal Component Analysis (PCA) was used as to classify the three North Sumatra Arabica coffee bean, and also used several pretreatment smoothing data: derivative 1 (D1), derivative 2 (D2), Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC) as spectra data correction methods. Result showed that the classification of the three North Sumatra Arabica coffee beans with PCA without pretreatment data was gaining 100% prediction accuracy.

1. Introduction

Coffee is one of the important export commodities in Indonesia. Indonesia is among the top four coffee producers in the world after Brazil, Vietnam and Columbia. Indonesia have many coffee industries, ranging from small industries to large industries and for domestic and foreign consumption. Indonesia also exports coffee to various countries in the world, such as the United States, Germany, Malaysia and other countries. The coffee used in this study was Arabica coffee beans from North Sumatera obtained from Mandailing Natal (Mandailing natal coffee), Humbang Hasudutan (Lintong coffee) and Dairi (Sidakalong coffee).

Principle Component Analysis (PCA) is a technique for constructing variables into new variables that are not overlapping and simple with the uniformity of data in accordance with the first component and a technique for grouping data into several different data groups. The classification of coffee in this study conducted by analysis of Principle Component Analysis (PCA) using Spectrometers Fourier Transform - Near Infrared Reflectance Spectroscopy (FT-NIRS). The advantage of Fourier transform (FT) is that radiation of all wavelengths is measured simultaneously by an interferometer, it’s called the Fourier Transform-Near Infrared Reflectance Spectroscopy (FT-IR) because it has a high sensitivity and speed [1][2]. NIRS with a wavelength of 1000-2500 nm use a non-destructive method that is to quickly and without damaging the coffee beans so that it is more efficiently applied to the specialty coffee industry in North Sumatra. Non-destructive methods are needed for the coffee industry because the process is fast with low costs and does not damage the chemical content in coffee.
2. Material and method

2.1 Materials and tools
This study was carried out in the Laboratory of Biosystem Engineering, Faculty of Agriculture, Universitas Sumatera Utara, Medan, Indonesia and the Laboratory of Food Processing and Agricultural Product Engineering (TPHP) Department of Mechanical and Biosystem Engineering, Faculty of Agricultural Technology, IPB University. The materials used in this study were coffee beans in the form of green bean originating from each region, namely Mandailing Natal Regency, Humbang Hasundutan and Dairi Regency. The tools used in this research were digital scales (0-200 grams, 0.001 gram accuracy), petri dishes, 50 mesh sieve, micro pipette, measuring cup, oven, desiccator, FT-NIR spectrometer (NIRFlex N-500, produced by BUCHI Labortechnik AG, Switzerland).

2.2 Research methods

2.2.1 Acquisition of coffee bean spectra. The acquisition of reflectance spectra was carried out on each sample of rice coffee beans. A total of 96 grams of coffee beans are placed in petri dishes evenly and tightly with 4 layers of piles [3]. Reflectance measurement were carried out by scanning each sample 3 times at three different points by adjusting the petri dish to rotate 360 degrees during the sample scan. The NIR wavelength range used was 1000-2500 nm with an interval of 2 nm.

2.2.2 Research data analysis. Spectra data obtained from measurements in the form of reflectant data were then transformed into absorbance spectra (log (1/R)) to obtain a linear correlation between NIR spectra values with chemical data. Data pretreatment method was used to reduce the noise caused by external influences. In this study the first and second derivative methods are used, multiple scatter correction (MSC) method, standard normal variate (SNV) and a combination of two pretreatment methods. Spectra data that has been processed by the pretreatment method was then analysed using the Principal Component Analysis (PCA) method.

3. Result and discussion

3.1 Spectra original North Sumatra Arabica coffee beans
Figure 1 shows the absorbance spectrum of NIRS scan results on North Sumatra Arabica coffee beans that have peaks and valleys at certain wavelengths as a sign of the presence of substances in coffee beans. The wavelength range used was 1000-2500 nm. The results of the acquisition of the NIRS spectra are then carried out the Principal Component Analysis (PCA).

![Figure 1. Absorbance spectra of North Sumatera coffee beans](image)

(a) Sidikalang coffee (b) Mandailing Natal coffee (c) Lintong coffee

3.2 Clustering analysis using PCA of spectra data pretreatment
Principle Component Analysis (PCA) is used in multivariable analysis to avoid multicolliniarity, as a modelling method that allows for exploratory data analysis and allows for sample classification by looking at differences and similarities between samples [4][5]. PCA aims to classify and produce new
variables in a smaller number than the original variables [6]. The principle of PCA is to group data into new data that are no longer correlated to determine differences between samples.

### 3.3 Data pretreatment

According to [7][8], data Pretreatment was done to improve the spectrum obtained. Pretreatment data for PCA analysis is using smoothing which functions to eliminate noise and is used in optimization of signal to noise rate, and is usually combined with other initial data processors. Derivative 1 and derivative 2 Savitzky Golay, because according to [9] Derivative 1 and derivative 2 Savitzky Golay were the most commonly used methods that function to remove the background and increase the resolution of the spectra and can clarify the peaks and valleys of absorbance spectra of NIRS data. *Multiplicative Scatter Correction* (MSC) is used to reduce the scattering and multiplicative effects of the NIR spectrum. Standard Normal Variate (SNV) to eliminate multiplicative interference from scatter effects on spectral data. The effect of SNV is on a vertical spectrum scale centered at zero and varies from -2 to 2 and almost similar to MSC. The difference is that SNV uses data from that spectrum while MSC uses the average spectrum from each set.

### 3.4 Outlier removal in the spectrum

The results obtained in PCA without pretreatment that was equal to 99.66 % with a total of 301 the number of data samples. In Figure 2 there was one outline of Sidikalang coffee types. According to [10] the outliers data was caused by the presence of different substances in the sample.

![Figure 2. Results of PCA analysis without pretreatment](image)

The results obtained in pretreatment smoothing of 99.66 %. There was one outlier data, namely from the type of sidikalang coffee in Figure 3. According to [10] the outliers data was caused by the presence of different substances in the sample.

![Figure 3. PCA + Smoothing analysis results](image)
Whereas in Figure 4 the results of PCA with pretreatment derivative 1 is equal to 98.00 % and there are six of data pen Cilan, three outliers from the data type of coffee Mandailing natal and three outliers from the data types Lintong coffee.

**Figure 4.** PCA + Derivative 1 analysis results

The results of PCA analysis with pretreatment Derivative 2 was 99.33 %. Where there were two outliers data of lintong coffee types as in Figure 5. According to [10] the outlier data was caused by the presence of different substances in the sample.

**Figure 5.** PCA + Derivative 2 analysis results

**Figure 6.** PCA + SNV analysis results
The results of PCA analysis in Figure 6 with SNV pretreatment of 99.66% where there was one data outliers of different types of coffee Mandailing Natal. According to [10] the outliers data was caused by the presence of different substances in the sample.

The results of PCA analysis with MSC pretreatment SNV was almost equal to that of 99.66% where there was one data outliers of different types of coffee Mandailing natal as in Figure 7. According to [10] the outliers data was caused by the presence of different substances in the sample.

![Figure 7. PCA + MSC analysis results](image)

3.5. Coffee Classification with PCA
Figure 8. Pretreatment spectra of absorbance by Smoothing, D1, D2, SNV and MSC

Clearly visible differences in clustering analysis in the three coffees with and without pretreatment. Figure 8 shows the PCA results of the three North Sumatra Arabica coffees with various pretreatments to see the different results given at each of the pretreatments. The analysis showed that the pretreatment spectra data with SNV and MSC can classify the three coffee well.

The results of PCA classification without pretreatment that was equal to 100% as in Figure 8 (original) can be seen the resulting PC1 score of 90% and PC2 score of 10%. For pretreatment smoothing in Figure 8 (smoothing) could also be classified well that was equal to 100%. While the results of the classification of derivative 1 pretreatment can be seen in Figure 8 (derivative 1) which was 91% but the classification results in pretreatment derivative 2 Figure 8 (derivative 2) were not good at 27% and PCA results in SNV and MSC pretreatment in Figure 8 (SNV, MSC) can be classified well that was equal to 93%.

Analysis of grouping by PCA on the three coffee can be distinguished. This means that the PCA was able to distinguish three types of North Sumatra Arabica coffee beans.

4. Conclusion
Results showed that the PCA classification without data pretreatment was able to classify 100% of the three North Sumatra Arabica coffee beans, namely Sidikalang coffee, Mandailing natal coffee and Lintong coffee and the results of PCA classification 100% for pretreatment smoothing, 91% for derivative, 93% for SNV and 93% for MSC can classify coffee well.

References
[1] Sablinskas V, Steiner G, Hof M 2005 Applications In: Gauglitz, G & Vo- Dinh, T (eds.) Handbook of spectroscopy (Germany: Wiley-VCH Verlag GmbH & Co. KGaA.)
[2] Larkin P 2011 Infrared and raman spectroscopy principles and spectral interpretation (Waltham: Elsevier Science)
[3] Budiastra I W, Sutrisno, Widayotomo S and Ayu P C 2018 Prediction of caffeine content in Java Preanger coffee beans by NIR spectroscopy using PLS and MLR method IOP Conf Series: Earth and Environmental Science 147 (2018) 012004 doi :10.1088/1755-1315/147/1/012004
[4] Cozzolino D, Cynkar W U, Shah N and Smith P 2011 Multivariate data analysis applied to spectroscopy: potential application to juice and fruit quality J Food Res Int 44 p 1888
[5] Kumar N, Bansal A, Sarma G S and Rawal R K 2014 Chemometrics tools used in analytical chemistry: an overview J Talanta 123 p 186
[6] Iriawan N, Astuti S P 2006 Mengolah data statistik dengan mudah menggunakan minitab 14 [Process statistical data easily using minitab 14] Yogyakarta: ANDI ISBN 979-763-111-7
[7] Cen H and He Y 2007 Theory and application of near infrared reflectance spectroscopy in determination of food quality J Trends in Food Sci & Technol 18 pp 72-83
[8] Munawar A A 2014 Multivariate analysis and artificial neural network approaches of near infrared spectroscopic data for non-destructive quality attributes prediction of mango Disertation Goettingen: Georg-August University

[9] Williams P and Karl N 1990 Near - Infrared technology in the agricultural and food industries (Minnesota: American Association of Cereal Chemists. Inc)

[10] Ramadhan S, Agus A M and Diswandi N 2016 Aplikasi NIRS dan principal component analysis (PCA) untuk mendeteksi daerah asal biji kopi Arabika (Coffea Arabica) [NIRS and principal component analysis (PCA) application for rapid classification of Arabica coffee beans (Coffea Arabica)] Jurnal ilmiah mahasiswa pertanian unsyiah 1 (1) pp 954-60

Acknowledgements
The research was funded by Universitas Sumatera Utara through the Non PNBP Fund Fiscal in 2019 based on the Implementation Contract of Universitas Sumatera Utara TALENTA Research Fiscal in 2019 between the Rector and the Chairman of Universitas Sumatera Utara Research Institute Number: 4167/UN5.1.R/PPM/2019 dated April 1st 2019.