Environmental origin classification of coffee beans using infrared spectroscopy

Yusmanizar*, A A Munawar
Department of Agricultural Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia.

*Email: yusmanizar@unsyiah.ac.id

Abstract. Coffee is one of tropical agricultural products cultivated in many counties and consumed by people worldwide. The main purpose of this study is to employ the infrared spectroscopy technique to rapidly classify the environmental origins of green coffee bean samples. To achieve this purpose, diffuse reflectance spectral data of coffee samples were collected and acquired in wavelength range of 1000 – 2500 nm. Classification models were established using principal component analysis (PCA) combined with linear discriminant analysis (LDA). The result showed that coffee bean sample can be classified based on their environmental origins with maximum total explained variance of the first two principal components is 97% (PC1 87% and PC2 10% respectively). Judging from the confusion matrix of the LDA, the classification accuracy reach 92%. It may conclude that infrared spectroscopy approach can be used to rapidly classify and sort coffee beans based on their geographical and environmental origins.

1. Introduction
Coffee is the main commodity of plantation commodities which is the mainstay of Indonesia's exports and plays an important role in the Indonesian economy. Based on 2019 BPS data, Indonesia's coffee production reached 742 thousand tons in 2019. Production is dominated by coffee from the island of Sumatra, namely Aceh (10%), North Sumatra (10%), South Sumatra (26%), Bengkulu (8%), Lampung (15%) and other provinces 31%. The total area of Indonesian coffee is 1,239.70 thousand ha, with the largest area being smallholder plantations at 1,215.50 thousand ha with a production of 731.60 thousand tons (98.6%), large state plantations covering an area of 14.50 thousand ha, producing 5.60 thousand tons (0.8%) and large private plantations with an area of 9.70 ha with a production of 4.40 thousand tons (0.6%) [1].

At this time drinking coffee has become a trend and lifestyle for millennial generation children. This can be seen from the number of franchised cafes and coffee shops that give their own nuances. The main quality of coffee that guarantees the current taste is also identified with geographical indications which are a growing trend. Geographical Indications (GI) of coffee indicate the area of origin of coffee products which due to natural and human factors or a combination of the two produce their own quality and taste of coffee. Coffee quality is determined based on the physical and chemical state of the coffee [2,3].
Determination of chemical content is usually carried out by laboratory tests accompanied by complex chemical analyzes that involve chemicals or other processes that sometimes take quite a long time. Chemical analysis can also cause environmental pollution, is expensive, ineffective, and damages the product so that it is not suitable to be applied in industries engaged in the production and distribution of coffee [4–6]. Alternative methods that are fast, environmentally friendly, effective and non-destructive are needed to predict coffee quality. One method currently being developed is the technology of diffuse reflectance infrared spectroscopy [7,8]. This study aims to test and apply NIR-based diffuse reflectance infrared spectroscopy technology in classifying and predicting the quality of Indonesian coffee, especially Sumatra, quickly, simultaneously and non-destructively.

2. Materials and methods

2.1. Study of the Electro-Optical Properties of Coffee

The materials used in this study were arabica and robusta rice coffee (dried coffee beans) with a water content of 12%. The equipment used is a NIR Spectrometer instrument type NIRFlex N-500 (Buchi Labortechnik AG, Switzerland), to produce a diffuse reflectance spectrum in the wavelength range of 1000–2500 nm [1]. One computer and Unscrambler® X software version 10.3 for spectrum analysis of data. The research method is quantitative with experiments conducted on 105 samples of Arabica and Robusta coffee originating from the island of Sumatra. Dried coffee beans to be analyzed are included in the sample holder. Before measuring coffee, the background measurement process is carried out first. The working principle of the Buchi NIRFlek N-500 instrument is shown in Figure 1.

![Figure 1. NIRFlek N-500 (Source: Operation Manual NIRFlek N-500 Buchi).](image-url)
2.2. Measurement of coffee qualities
The materials used in the chemical analysis were Arabica Gayo coffee beans, Mandailing, Koerintji, Lampung robusta coffee and Semendo. The equipment used in this research is a set of HPLC Water S-1500-Daries LS/AK 01-018 instruments. The columns used are LiChcoCART®250-4.6, Purosher®STAR RP-18e, Sorbent Lot No. HX109873, grinder, filter, micro pipette, test tube, measuring cup and a computer set with Ms. software Excel. The study used experimental methods to determine the chemical content of coffee, namely caffeine and chlorogenic acid in Sumatran Arabica and Robusta coffee destructively. Analysis of caffeine content was carried out using the AOAC method and chlorogenic acid was carried out using the Naegele method [9–11].

3. Results and discussion

3.1. Spectra properties of coffee samples
The results of the acquisition of the NIRS spectrum were averaged based on geographical indications of Gayo, Mandailing, Koerintji and Semendo robusta Arabica coffees, Lampung as shown in Figure 2. Arabica coffee spectrum is in the top position starting from Gayo coffee, Mandailing and Koerintji. The difference is clearly visible in the distance between the peaks of the Gayo, Mandailing and Koerintji coffee spectrum. In the robusta coffee spectrum, Semendo coffee has a spectrum peak below that of Lampung coffee. The spectrum of Arabica coffee has a higher peak than that of Robusta, this is due to the different types of varieties from the sample.

![Figure 2. Spectra features of Arabica coffee samples from different environmental origins.](image)

The results of the acquisition of the Sumatran coffee NIRS spectrum were then used in the Principal Component Analysis (PCA) analysis to see the relationship between the spectrum obtained and the properties of the material being measured [12,13]. The results of the analysis using the PCA+Hotelling T2 elliptical method are shown in Figure 3.
To find out which coffee samples were from which varieties and which areas were outside the ellipse, the data was grouped based on the area of origin of the coffee. The results of the plot of grouping coffee data based on the variety and area of origin with geographical indications are shown in Figure 4. The results of the PCA+Hottelling $\text{T}^2$ elliptical grouping are known that 4 coffee samples that are outside the ellipse of the arabica variety, while the robusta variety are all in the ellipse. The results of grouping 5 types of Sumatran coffee, samples that are outside the ellipse all come from the Aceh area (Gayo coffee). This is possible because the 4 samples of Aceh coffee have components (variables) that are much different from other samples, both in terms of chemical components, texture or other things. It can be concluded that Gayo coffee has different optical physico-optic characteristics compared to other regions, as can be seen from the results of the PCA+Hottelling $\text{T}^2$ elliptical analysis.

The results of the PCA+Hottelling $\text{T}^2$ elliptical model classification (Figure 5) show that the model is less than perfect in classifying coffee based on varieties and geographical indications. In the plot of the results of the classification of varieties, there is still mixing between robusta and arabica. In the coffee plot based on the GI, only Gayo and Mandailing arabica coffees appear to be separated, while Koerintji, Semendo and Lampung coffees are still in overlapping ellipses.
3.2. Classification of Arabica and Robusta Coffee
Classification of arabica coffee (Gayo, Mandailing, Koerintji) and robusta (Semendo, Lampung) was carried out using Linear Discriminant Analysis (LDA) and Support Vector Machine Classification (SVMC) models [14–16]. The LDA model used consists of linear, quadratic and mahalanobis models. The results of the LDA analysis are displayed in the form of a confusion matrix. The results of linear LDA analysis on the classification of 5 types of Sumatran coffee with geographical indications have an accuracy value of 87.62%. The results of the classification of the linear LDA model in the form of a confusion matrix are shown in Table 1. The results of the linear LDA analysis show that the model can classify the five types of Sumatran coffee well. In Arabica coffee, only Koerintji coffee is classified as robusta coffee (one sample is predicted to be Lampung robusta). In Gayo and Mandailing coffee, one Arabica coffee variety is still grouped, as well as Semendo and Lampung robusta coffee, only one sample of Lampung robusta coffee is predicted to be Koerintji Arabica coffee.

Table 1. Results of different LDA models classification (Linear, Quadratic, Mahalanobis).

| Origin of coffee | Actual | Gayo | Mandailing | Koerintji | Semendo | Lampung |
|------------------|--------|------|------------|-----------|---------|---------|
| Prediction       |        | 1    | 2          | 3         | 4       | 5       |
| **LDA Linear**   |        |      |            |           |         |         |
| Gayo             | 1      | 19   | -          | -         | -       | -       |
| Mandailing       | 2      | 1    | 13         | 3         | -       | -       |
| Koerintji        | 3      | -    | 3          | 16        | -       | 1       |
| Semendo          | 4      | -    | -          | -         | 24      | 3       |
| Lampung          | 5      | -    | -          | 1         | 1       | 20      |
| **LDA Quadratic**|        |      |            |           |         |         |
| Gayo             | 1      | 20   | -          | -         | -       | -       |
| Mandailing       | 2      | -    | 16         | -         | -       | -       |
| Koerintji        | 3      | -    | -          | 20        | -       | -       |
| Semendo          | 4      | -    | -          | -         | 25      | 1       |
| Lampung          | 5      | -    | -          | -         | -       | 23      |
| **LDA Mahalanobis**|      |      |            |           |         |         |
| Gayo             | 1      | 20   | -          | -         | -       | -       |
| Mandailing       | 2      | -    | 13         | -         | -       | -       |
| Koerintji        | 3      | -    | 3          | 20        | -       | -       |
| Semendo          | 4      | -    | -          | -         | 25      | 1       |
| Lampung          | 5      | -    | -          | -         | -       | 23      |
The SVM classification method is a method that seeks to find the best hyperplane that functions as a class separator in the input space [17–19]. The SVMC classification model produced correct predictions for Gayo and Mandailing arabica coffee, while for Koerintji coffee there were 4 samples that did not match the predictions. The results of the SVMC model on Robusta coffee there are still 2 samples in Semendo coffee and one sample in Lampung coffee which have not been predicted correctly (Table 2).

| Origin of coffee | Actual | Gayo | Mandailing | Koerintji | Semendo | Lampung |
|-----------------|--------|------|------------|-----------|---------|---------|
| Prediction      |        | 1    | 2          | 3         | 4       | 5       |
| Gayo            |        | 1    | 20         | 0         | 0       | 0       |
| Mandailing      |        | 2    | 0          | 16        | 1       | 0       |
| Koerintji       |        | 3    | 0          | 0         | 16      | 0       |
| Semendo         |        | 4    | 0          | 0         | 0       | 23      |
| Lampung         |        | 5    | 0          | 0         | 3       | 2       |

The results of the five calibration models used show that the linear LDA model has an accuracy value of 87.62%, quadratic 99.05%, mahalanobis 96.19% and SVMC with an accuracy value of 99.33% on training, 84.76% validation. From the results of the accuracy values obtained, it can be seen that SVMC has the highest accuracy value but based on the confusion matrix, Koerinji, Semondo and Lampung coffee are not classified perfectly. Therefore, in this study the LDA Quadratik model is the recommended model for classifying Sumatran coffee samples based on geographical indications.

4. Conclusion

The main purpose of this study is to employ the infrared spectroscopy technique to rapidly classify the environmental origins of green coffee bean samples. The result showed that coffee bean sample can be classified based on their environmental origins with maximum total explained variance of the first two principal components is 97% (PC1 87% and PC2 10% respectively). Judging from the confusion matrix of the LDA, the classification accuracy reach 92%. It may conclude that infrared spectroscopy approach can be used to rapidly classify and sort coffee beans based on their geographical and environmental origins.

References

[1] Yusmanizar, Setiasih I S, Nurjanah S, Muhaeimin M, Nurhadi B, Rosniawaty S and Munawar A A 2019 Fast and Non-Destructive Prediction of Moisture Content and Chologenic Acid of Intact Coffee Beans Using Near Infrared Reflectance Spectroscopy IOP Conference Series: Materials Science and Engineering vol 506 (Institute of Physics Publishing)

[2] Giraudo A, Grassi S, Savorani F, Gavoci G, Casiraghi E and Geobaldo F 2019 Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis Food Control 99 137–45

[3] Monakhova Y B, Ruge W, Kuballa T, Ilse M, Winkelmann O, Diehl B, Thomas F and Lachenmeier D W 2015 Rapid approach to identify the presence of Arabica and Robusta species in coffee using 1H NMR spectroscopy Food Chem. 182 178–84

[4] Bizzani M, Flores D W M, Colnago L A and Ferreira M D 2017 Non-invasive spectroscopic methods to estimate orange firmness, peel thickness, and total pectin content Microchem. J. 133 168–74

[5] Sezer B, Apaydin H, Bilge G and Boyaci I H 2018 Coffee arabica adulteration: Detection of wheat, corn and chickpea Food Chem. 264 142–8
[6] Assis C, Pereira H V, Amador V S, Augusti R, de Oliveira L S and Sena M M 2019 Combining mid infrared spectroscopy and paper spray mass spectrometry in a data fusion model to predict the composition of coffee blends Food Chem. 281 71–7

[7] Huang H, Yu H, Xu H and Ying Y 2008 Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review J. Food Eng. 87 303–13

[8] Correia R M, Tosato F, Domingos E, Rodrigues R R T, Aquino L F M, Filgueiras P R, Lacerda V and Romão W 2018 Portable near infrared spectroscopy applied to quality control of Brazilian coffee Talanta 176 59–68

[9] Scholz M B S, Pagiatto N F, Kitzberger C S G, Pereira L F P, Davrieux F, Charmetant P and Leroy T 2014 Validation of near-infrared spectroscopy for the quantification of cafestol and kahweol in green coffee Food Res. Int. 61 176–82

[10] Ribeiro J S, Ferreira M M C and Salva T J G 2011 Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy Talanta 83 1352–8

[11] Shan J, Suzuki T, Ogawa Y and Kondo N 2015 Coffee roasting degrees prediction in terms of weight loss with selected wavebands based on near-infrared spectroscopy Eng. Agric. Environ. Food 8 195–9

[12] Munawar A A, Syah H and Yusmanizar 2019 Fast and robust quality assessment of honeys using near infrared spectroscopy IOP Conference Series: Earth and Environmental Science vol 365 (Institute of Physics Publishing)

[13] Kusumiyati, Munawar A A and Suhandy D 2021 Fast, simultaneous and contactless assessment of intact mango fruit by means of near infrared spectroscopy AIMS Agric. Food 6 172–84

[14] Munawar A A, Hörsten D V., Mörllein D, Pawelzik E and Wegener J K 2013 Rapid and non-destructive prediction of mango sweetness and acidity using near infrared spectroscopy Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI) vol P-211 (Gesellschaft fur Informatik (GI)) pp 219–22

[15] Hayati R, Munawar A A and Fachruddin F 2020 Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango Data Br. 105571

[16] Kusumiyati, Hadiwijaya Y, Putri I E and Munawar A A 2021 Multi-product calibration model for soluble solids and water content quantification in Cucurbitaceae family, using visible/near-infrared spectroscopy Heliyon 7 e07677

[17] Pasquini C 2018 Near infrared spectroscopy: A mature analytical technique with new perspectives – A review Anal. Chim. Acta 1026 8–36

[18] Arendse E, Fawole O A, Magwaza L S and Opara U L 2018 Non-destructive prediction of internal and external quality attributes of fruit with thick rind: A review J. Food Eng. 217 11–23

[19] Munawar A A, Hörsten D V, Mörllein D, Pawelzik E and Wegener J K 2013 Rapid and non-destructive prediction of mango sweetness and acidity using near infrared spectroscopy Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI) vol P-211