SINET: Ethiop. J. Sci., 45(1): 1–25, 2022 ISSN: 0379–2897 (PRINT)
 eISSN: 2520–7997

Date received: August 14, 2021; Date revised: March 05, 2022; Date accepted: March 28, 2022
DOI: *https://dx.doi.org/10.4314/sinet.v45i1.1*

Improved Principal Component Analysis and Linear Discriminant Analysis for the Determination of Origin of Coffee Beans using

Endale Deribe Jiru [1], Berhanu Guta Wordofa [1], Mesfin Redi-Abshiro [2*]

[1] Department of Mathematics, College of Natural Sciences, Addis Ababa University, Addis Ababa, Ethiopia
[2] Department of Chemistry, College of Natural Sciences, Addis Ababa University, Addis Ababa, Ethiopia.
E-mail: mesfin.redi@aau.edu.et

ABSTRACT: In this work an improved Principal Component Analysis (PCA) method is used for better determination of geographical origins of Ethiopian Green Coffee Beans. In the commercially available and widely employed PCA methods the dataset is commonly normalized using Z-score procedure, which reduces the influence of the spread of data (or dispersion degree differences) on principal components (PCs). In the improved method, a new normalization procedure is introduced with the aim to improve the spread (dispersion) of data points around the mean. The PCs computed from the improved procedure could significantly better reflect information of the original dataset. The dispersion degree information in the original dataset was retained relatively much by using the improved PCA than the Z-score-based PCA. The improved PCA was then used to identify the most discriminating variables corresponding to the coffee samples and, based on that, Linear Discrimination Analysis (LDA) model was developed to classify and predict samples. The recognition and prediction abilities of the improved PCA and LDA at regional level respectively were 95.7% and 94% (using Chlorogenic Acids (CGA s) content), 91% and 97% (using Fatty Acids (FA) content), 99% and 100% (and using the combined CGA and FA contents). Mehari *et al.* (2016, 2019) reported recognition and prediction of the PCA, they applied on the same dataset, at regional level were 91% and 90% (using CGA s content) and 95% and 92 % (using FAs content), respectively. The result reveals that the newly introduced method is superior and the best discriminations of coffee beans were achieved. The combined analysis of CGA and FA concentrations is a useful tool for the determination of origin of coffee beans, and we recommend that the concerned bodies should use it to address the characterization, classification and authentication of Ethiopian coffee beans according to their geographical origins.

Key words/Phrases: Chlorogenic acid and Fatty acid, Classification, Dimensionality reduction, Linear discriminant analysis, Principal component analysis

INTRODUCTION

Coffee, which is the second most important commodity in international trade, involves networked trade covering both developing and developed countries. The price of coffees in the international market, which depends on the quality of the coffee beans, has a direct correlation with the taste of the final consumed product. The coffee's originality and traceability have been seen as important factors and hence, the determination of quality and originality of coffee is necessary (Kurniawan *et al.,* 2019). Coffee plays a vital role in the Ethiopia's economy and become a major source of foreign exchange earnings. The great diversity in country's coffee 'gene pools', which can be associated with the existence of diverse agro-ecology, has endowed Ethiopia with very diverse and unique quality coffee characteristics that associated with different localities (Ethiopian Coffee Science Society, ECSS, 2019). There is a price difference for Ethiopian coffee products depending on the region of production and that is determined by their flavors. This can be seen as the cause for the adulteration of expensive varieties with cheaper coffee varieties and fraud regarding to the production areas (origins) within the country (Bewketu Mehari *et al.,* 2016). This calls for a reliable and effective means of identifying the geographical origins of coffee beans. On the other hand, the development of a method that helps to make the desired classification needs high

*Author to whom correspondence should be addressed.

dimensional (multivariate) massive dataset that reveals different attributes of the coffee beans depending on the locations of their origin.

Indeed, technological advancement and innovations have brought massive high dimensional data, called Big Data, which has been encouraged advancement of computational techniques considering the major issues and challenges of those massive dataset like volume, speed, and variety that mainly related to dimensions (Kpigigbue *et al,* 2019). Moreover, recent development and advancement in research work and technologies resulted in an exponential growth in dataset with respect to sample size as well as dimensions. Laboratory instruments become more and more complex and report hundreds or more measurements for a single experiment and thus, dealing with and retrieving information from such high dimensional datasets create challenges for the users and researchers to automatically extract useful information, pattern and knowledge from them (Ullah *et al*., 2017). However, much of the data in those high dimensional datasets are highly redundant and can efficiently reduced down to a much smaller number of variables without a significant loss of information using the mathematical methods known as dimensionality reduction (DR) techniques and so, developing and applying effective and appropriate DR techniques based on the given dataset are currently a hot-research topic (Holmes and Huber, 2019).

Although many dimensionality reduction (DR) techniques have been developed and implemented, they are easy to misuse, and their results are often misinterpreted in practice (Nguyen and Holmes, 2019). In addition, most of the existing DR and classification techniques lack in producing easily interpretable features, understandable patterns and interesting results for different research areas and applications (Geng and Hamilton, 2006). It is known that the underlying assumption for dr techniques is that keeping the most useful information of the original high-dimensional datasets in a low-dimensional transformed subspace. Hence, the main goal of DR techniques is to get accurate and easily understandable representation of the original dataset with the removal of statistically redundant information (Jolliffe and Cadima, 2016; Breger, *et al*., 2020).

Principal component analysis (PCA) and Linear Discriminant Analysis (LDA) are two popular methods for DR and data visualization (Bishop, 2006). PCA is a data analysis method, which uses an orthogonal transformation to convert a set of possibly correlated observations into a new set of linearly uncorrelated components called *Principal Components,* which are ordered so that the first few retain most of the variation present in the original variables (Jolliffe, 2002; Mishra *et al*., 2017, Walker, 2020). In terms of linear algebra, PCA involves basically the eigenvalue decomposition of the covariance matrix of a dataset. On the other hand, LDA finds a linear combination of observation vectors, which separate two or more categories of objects by finding a low dimensional subspace that keeps data points from different classes far apart and those from the same class as close as possible (Bishop, 2006). In general, PCA and LDA are two widely used methods for DR and classification in the areas of machine learning, pattern recognition, and applications of science and engineering (Bishop, 2006; Charu, 2014; Tharwat, 2016, Walker, 2020).

Recently, several authors have developed, applied, and reviewed different dimensionality reduction (DR) and classification techniques. Gupta *et al*. (2002) introduced DR techniques that improve the performance of classification algorithms. They revealed that techniques such as PCA, LDA, and kernel based PCA and lda were used to reduce the dimensionality of original dataset. Similarly, Arunasakthi and Kamatchipriya (2014) conducted review on linear and non-linear DR techniques, and stated that PCA and LDA were the fundamental techniques for DR as well as retrieving effective variables of high dimensional data points. Despite the development of several PCA -based DR models, there are few studies focused on the retention outcome of key information from the original used dataset in PCA (Hosseini and Kaneko, 2011). Shang and Wang (2014) proposed improved the classical PCA by normalizing the data matrix using the mean of each feature of the original dataset. They applied it for comprehensive assessment on thermal power generation units and retained original information in better way than classical PCA. However, they couldn't implement it to applications involving negative data points. Data normalization is a data preprocessing technique for DR to transform the original dataset into a

desired range, which improves the outliers and data quality, removes the inconsistency and ambiguity in the original datasets and improves the performance of the techniques and algorithms (Rathod and Momin, 2012).

In most dimensionality reduction (DR) techniques, the primary step is identifying whether the selected variables are interrelated to each other, since information overlap could make evaluation results biased (Shirali *et al*., 2016). Studies show that pca models have been commonly implemented using original datasets for dr (Coussement *et al*., 2016; Rajesh *et al*., 2018). These original datasets contain key information mainly in the two aspects: the distribution information (or spread of data) among all variables, which is reflected by the variance; and the significant or insignificant relationships between the variables of the given dataset, which is reflected by correlation coefficient matrix (Shang and Wang, 2014). However, most studies used the *Z-score* standardization for data normalization in PCA model development that makes the variances of all indicators equal to 1, which eliminates the information of dispersion degree contained in the given dataset (Hao *et al.,* 2013; Shang and Wang, 2014).

Therefore, in order to apply normalization on the original dataset having large differences in the measured scales, essential consideration should be taken for PCA model development to avoid the loss of key information. Taking into account the fact that information overlap between variables is eliminated by applying PCA (Jolliffe, 2002; Shirali *et al*., 2016), this study aimed to introduce PCA with a new normalization method as DR and LDA classification technique to identify the most discriminating variables for the determination of the geographical origin of the various Ethiopian coffee beans.

Authentication of coffee origin is highly demanded by international consumers as additional attribute of quality, and thus consumers being willing to pay attractive prices for coffee varieties from particular areas. In this context, effective and reliable identification methods to prevent fraudulent practices become necessary. Bewketu Mehari *et al.* (2016) used the phenolic profiles of the Ethiopian coffee beans to identify characteristic chlorogenic acids according to their region of origin. They applied PCA on Pareto scaled data matrix and identified the concentrations of 3-cqa and 4,5-dicqa as the characteristic markers for Northwest and East coffees, respectively. Moreover, they applied LDA model for the classification of coffee samples and achieved the recognition and prediction abilities of 91% and 90%, respectively, at regional level, and 89% and 86%, respectively, at sub-regional level. Similarly, Kurniawan *et al.* (2019) developed da of pcs and applied it to dataset consisting three kinds of Java Arabica coffee beans namely Arabica Java Preanger, Arabica Bondowoso and Arabica Malang to classify them based on their origin. They confirmed that the best result for discriminating those three kinds of coffee beans was obtained with pc1 versus pc2 that classified Arabica coffee beans accurately 100%.

` On the other hand, the fatty acids contents, which decrease with increasing altitude of the coffee plants, are one of the major components that determine the quality and origin of coffee plants (Girmay *Tsegay et al.*, 2020). The article published by Bewketu Mehari *et al*. (2019) proposed analytical method to verify the production region of the Ethiopian coffee beans based on their fatty acid compositions. They applied PCA on Pareto scaled dataset to visualize data trends and LDA to construct classification models. The study identified Oleic, Linoleic, Palmitic, Stearic and Arachidic acids as the most discriminating compounds among the production regions. They achieved the recognition and prediction abilities of 95% and 92%, respectively, at regional level and 95% and 73%, respectively, at sub-regional level.

The article by Núñez *et al.* (2020) proposed techniques for the characterization, classification, and authentication of coffee samples according to their country of production, variety, and roasting degree. They used 306 commercially available coffee samples and divided them into three groups of samples: changing on the production country, coffee variety, and roasting degree. They applied pca and partial least squares regression-discriminant analysis (pls-da) and showed good discrimination capabilities among the different coffee production regions and coffee varieties (Arabica vs. Robusta).They revealed that pls-da provided classification rates higher than 89.3% and 91.7% for calibration and prediction, respectively.

However, there is no study on the characterization and classification of Ethiopian coffee beans according to their geographical origins using the combined analysis of

Chlorogenic acid and Fatty acid contents. Therefore, the aim of this study was using improved pca and lda to identify the most important discriminant compounds based on the composition of Chlorogenic acids and Fatty acids in the Ethiopian green coffee beans.

The novelty of the proposed dimensionality reduction and classification techniques of this study lies in following aspects.

1) The proposed pca technique improves the dispersion degree of the original dataset and the outlier robustness of pca and thus, allowed the normalized dataset retain more key information as compared to any previous techniques.
2) The proposed method uses the individual and combined analyses of Chlorogenic acid and Fatty acid contents, and hence, improves the characterization and classification of Ethiopian coffee beans according to their geographical origins with a high prediction success rate (100%) for the analysis of regional coffees as well as sub-regional coffee types.

The rest part of this work is organized as: Section two discusses the research methods for dimension reduction and classification using improved pca and lda, Section three present results and discussion of dimensionality reduction and classification applied to the Ethiopian coffee beans dataset, and the final section provides conclusion based on the findings of the study.

RESEARCH METHODS

Dataset

The data is taken from the published articles of Bewketu Mehari *et al.* (2016, 2019). The first article presents the measurements of eight different Chlorogenic acids (cgas) in each of 100 samples of green coffee beans collected from different part of Ethiopia; and the second article presents the measurements of 11 different Fatty acids (fas) in each of those samples. The data is constructed by applying the Box-Muller method using the mean, standard deviation and summarized values given for each regional and sub-regional category on the first article for Chlorogenic acids and for Fatty acids on the second article. The Box–Muller transform was developed and employed as a more computationally efficient alternative to the statistical inverse transform sampling method (Kloeden and Plate, 1992; Martino *et al.,* 2012). The coffee samples were collected from the four sampling regions, the major coffee production areas across Ethiopia, such as East, Northwest, West and South categories. Accordingly, the dataset contained a total of 100 samples, 27 from East (15 *Harar-A and 12 Harar-B*), 6 from Northwest (*3 Benishangul and 3 Finoteselam*), 18 from West (3 *Jimma-A,* 3 *Jimma-B* 10 *Kaffa and* 2 *Wellega*) and 49 from South (10 *Sidama-SA,* 29 *Sidama-SB and* 10 *Yirgachefe*) based on the 8 selected CGAs and 11 FAs found in green coffee beans. Consequently, the dataset of the measurements of CGA and FA are organized as observation matrices of sizes 8×100 and 11×100, respectively, and these are combined into 19×100 observation matrix. A column in each of the observation matrices is the measurement of corresponding variables of individual sample point.

Methods of Data Analysis

In this work, the statistical package for social science (spss) and matlab software were used to analysis the data. Pearson`s linear correlation coefficients were used for the determination of the variables with highest impact on the pca components' extraction process. Tabachnick and Fidell (2007) said that if there are few correlations above 0.3, it is a waste of time carrying on with the analysis. However, clearly, we do not have that problem, and the correlation matrix showed good consistency of results. Moreover, prior to constructing pca model, the suitability of the dataset for basis of pca was assessed using Kaiser-Meyer-Olkin (kmo) Measure of Sampling Adequacy and Bartlett's test of Sphericity (Maat *et al.,* 2011). The sampling is adequate if the value of kmo test is greater than 0.5 (Kaiser, 1974; Field, 2000), and the Bartlett's Test of Sphericity must be significant at $p < 0.05$ (Hair *et al.,* 2010; Tabachnick and Fidell, 2007).

Accordingly, the variables those satisfied the above tests were considered, and then one-way analysis of variance (anova) was used to test for the presence of significant differences between the mean concentrations of the variables (i.e. Chlorogenic acids, Fatty acids, and both) in the coffee beans from different categories. Differences were considered significant when $p < 0.05$. Next, new data normalization was applied to transform the raw data into a standard form, which enable good comparability between variables and

simplifies the algorithm's process. Consequently, the normalized data were analyzed using PCA models at regional and sub-regional levels. The PCA's Loadings plots and Score plots were used to identify the variables and the corresponding coffee samples. These plots and the significant differences revealed by ANOVA were used to select the suitable discriminant markers for the corresponding coffee samples. Finally, LDA was applied to develop classification models at both regional and sub-regional levels that could be used to classify the samples and predict the geographical origin of coffee beans.

Principal Component Analysis (PCA) and its Algorithm

PCA can employ the orthogonal projection to convert a large dataset of possibly interrelated variables into a smaller set of linearly uncorrelated variables so that the first linear combination captures the largest variance; the second linear combination explains the second largest variance; and so on (Jolliffe, 2002; Shlens, 2014; Walker, 2020). These new computed linear combinations are called the principal components (pcs) of the PCA model. After the pcs are computed, data analysis, visualization and interpretation can then be performed using those pcs instead of the original dimensions (variables) of the dataset (Breger *et al*., 2020).

Principal Component Analysis techniques

To prepare for PCA, let $\mathbf{X}=[X_1, X_2, \cdots, X_n]$ be a $d \times n$ observation matrix (a dataset of *n* multivariate items each of which with *d* components). That is, the dataset is organized so as its j-th column $X_j \in \Re^d$ is the j-th observation vector (sample point) with d components, i.e., $X_j = (x_{1j}, x_{2j}, \cdots, x_{dj})^T \; j = 1,2, \cdots, n$. Thus, $\mathbf{X}$ has d rows of variables where the i-th row $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ represents a particular feature (attribute) that varies over the *n* samples, $i = 1, 2, \cdots, d$. We assume, without loss of generality, $\mathbf{X}$ is mean-centered, i.e., the arithmetic mean for each row is zero. This can be always obtained by subtracting the row mean from each entry of the row. Consequently, the covariance matrix of the n samples is the $d \times d$ matrix S given by

$$S = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T$$

Indeed, the i-th diagonal element of S, $S_{ii} = \frac{1}{n-1} x_i x_i^T$, is the variance of the i-th variable x_i and its ij-th entry $S_{ij} = \frac{1}{n-1} x_i x_j^T$ is the covariance of x_i and x_j. Since S is positive semi-definite all its eigenvalues are real and non-negative. Moreover, as S is symmetric it has d orthonormal eigenvectors.

The goal of PCA is to reduce the dimensionality of the dataset using a linear transformation. The interpretation of PCA is that it finds the major axis (direction) of variation in the dataset such that the first pc defines the direction in the dataset with the greatest variance; and the i-th PC defines the direction orthogonal to the first $i-1$ PCs that maximizes the variance of the variables uncorrelated to each of the $i-1$ variables. Therefore, mathematically, the goal of PCA is to find an orthogonal $d \times d$ matrix $\mathbf{P}$ that determines the change of variable

$$X = \mathbf{P}Y$$

with the property that the new variables (components of Y) $y_1, y_2, \ldots, y_d$ are uncorrelated and are arranged in order of decreasing variance. Here $\mathbf{P}$ is orthogonal matrix means $\mathbf{P}^{-1} = \mathbf{P}^T$, so that the columns of $\mathbf{P}$ are orthonormal vectors (pair-wise orthogonal unit vectors), and $Y = \mathbf{P}^T X$. The new variables can have the desired properties if we take the matrix $\mathbf{P} = [v_1, v_2, \cdots, v_d]$ whose columns are the orthonormal eigenvectors of *S* corresponding to its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$, i.e., $Sv_i = \lambda_i v_i$, $v_i^T v_i = 1$, $v_i^T v_j = 0$ for $i \neq j$, and the eigenvalues are arranged in decreasing order so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$.

Now to justify that the new variables obtained by this transformation matrix $\mathbf{P}$ have the desired properties, first note that $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$ has zero mean (because the sum of its columns is equal to $\mathbf{P}^T \sum_{j=1}^{n} X_j = \mathbf{P}^T \mathbf{0} = \mathbf{0}$ as $\mathbf{X}$ has zero mean). So,

(i) The covariance of y_i and y_j for $i \neq j$, is

$$\text{cov}(y_i, y_j) = \frac{1}{n-1} y_i y_j^T = \frac{1}{n-1} v_i^T \mathbf{X}(v_j^T \mathbf{X})^T = v_i^T S v_j = \lambda_j v_i^T v_j = 0$$

Hence, the new variables y_1, y_2, ..., y_d are uncorrelated.

(ii) The variance of y_i is

$$\text{var}(y_i) = \frac{1}{n-1} y_i y_i^T = \frac{1}{n-1} v_i^T \mathbf{X}(v_i^T \mathbf{X})^T = v_i^T S v_i = \lambda_i v_i^T v_i = \lambda_i$$

, for each i =1, .., d.

Consequently,

$$\text{var}(y_1) = \lambda_1 \geq \lambda_i = \text{var}(y_i), \text{ for all } i = 2, 3, .., d.$$ (i.e., y_1 has the largest variance)

$$\text{var}(y_2) = \lambda_2 \geq \lambda_i = \text{var}(y_i), \text{for all } i = 3, 4, ..., d.$$ (i.e., y_2 has the largest variance among variables uncorrelated to y_1); and so on.

Therefore, taking the i-th pc to be the i-th unit eigenvector of v_i of S corresponding to the eigenvalue, λ_i, where the eigenvalues are sorted in decreasing order, v_1 is the first pc representing the direction of the largest (maximum) variance and, in general, the ith pc is v_i which is orthogonal to all previous i–1 pcs and represents the direction of maximum variance of all remaining uncorrelated variables.

Improved** PCA **techniques:

Due to the different measured scales among the variables of coffee beans, standardization is applied to the dataset to enable good comparability between variables.

Z-score Standardization: The original data matrix $\mathbf{X}_{d\times n}$ can be transformed into a standardized matrix $\mathbf{Y}_{d\times n}$ with zero mean and unit variance as shown below.

$y_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_{xj}}$ where y_{ij} is the standardized value of x_{ij}, while $\bar{x}_j$ and σ_{xj} are the *mean* and *standard deviation* of x_j, respectively.

However, Z-score makes the variance of each variable equal to 1, which reduces the influence of the spread of data (or dispersion degree differences) on pcs (Shang and Wang, 2014). Thus, pcs computed from the normalized dataset could not fully reflect information of the original dataset (Hosseini and Kaneko, 2011; Cai *et al.*, 2016). Based on this fact, an improved normalization method is proposed, that aimed to improve the spread (dispersion) of data points around the mean, as shown below

$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\beta_{xj}}$ so that z_{ij} is the standardized value of x_{ij} and

$$\beta_{xj} = \sqrt{(\max(x_j))^2 - (\min(x_j))^2} > 0.$$

The *mean* of j^{th} variable (z_j) is:

$$\bar{z}_j = \sum_{j=1}^{n} \frac{z_{ij}}{n} = \frac{\sum_{j=1}^{n}(x_{ij} - \bar{x}_j)}{n\beta_{xj}} = 0 \quad (1)$$

The *standard deviation* of j^{th} variable (z_j) is:

$$\sigma_{zj} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{\beta_{xj}}\right)^2} = \frac{\sigma_{xj}}{\beta_{xj}} \quad (2)$$

The *correlation coefficient matrix is*:

$$\rho_{z_j z_k} = \frac{S_{zj\,zk}}{\sigma_{zj}\sigma_{zk}} = \frac{1}{n-1}\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)(z_{ik} - \bar{z}_k)/(\sigma_{zj}\sigma_{zk}),$$

$S_{z_j z_k}$ is covariance between z_j and z_k

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_{ij} - \bar{x}_j)}{\beta_{xj}}\frac{(x_{ik} - \bar{x}_k)}{\beta_{xk}}/\left(\frac{\sigma_{xj}\sigma_{xk}}{\beta_{xj}\beta_{xk}}\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)/(\sigma_{xj}\sigma_{xk})$$

$$= \rho_{x_j x_k}. \quad (3)$$

According to *Equation (3)*, the original dataset $\mathbf{X}_{d\times n}$ and normalized dataset $Z_{d\times n}$ have the same correlation coefficient matrix, indicating that the improved standardization method keeps correlation information of all the variables. Importantly, $\sigma_{z_j}=\sigma_{x_j}/\beta_{x_j}$ at *Equation* (2) shows dispersion degree differences of all variables are partly retained, and the classical PCA (based on Z-score) is to some extent improved.

Algorithm 1: Improved PCA algorithm

1. *Standardize the original dataset*:

- The original data matrix $\mathbf{X}_{d\times n}$ can be transformed into a standardized matrix $Z_{d\times n}$ as:

$z_{ij}=(x_{ij}-\bar{x}_j)/\beta_{x_j}$ where z_{ij} is the standardized value of x_{ij} and $\bar{x}_j$ is the mean of x_j and

$$\beta_{xj}=\sqrt{(\max(x_j))^2-(\min(x_j))^2}>0$$

2. *Calculate the covariance matrix* S:

$S=(S_{z_i z_j})_{d\times d}=\frac{1}{n-1}ZZ^T$, where z_i and z_j are the ith and jth row vectors of $Z_{d\times n}$, respectively, $S_{z_i z_j}$ is the covariance value between z_i and z_j.

3. *Compute eigenvalues and eigenvectors of* **S**, using $\det(S-\lambda I)=0$ or using any available tool such as MATLB..

- Eigenvalues are arranged in descending order: $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$.
- The corresponding eigenvectors are calculated using $Sv_j=\lambda v_j$ and organized as columns of $V=(v_1,v_2,\ldots,v_d)$,

which are called *Principal Components* (PCs).

4. *Determine the number of principal components (PCs)*:

- A total cumulative percentage variance ≥ 75% is used.

5. *Identify the variables belonging to those determined* pcs:

- The loading of each variable on each those determined pcs is computed by $\theta_{ij}=v_{ij}\sqrt{\lambda_j}$, where λ_j is the eigenvalue corresponding to jth pc v_j and v_{ij} is i^{h} value of v_j.

6. *Calculate PCA scores*, using the projection $\mathbf{F}=V^T Z$.

LDA Techniques and its Algorithm

The goal of the LDA technique is to project the original data matrix onto a lower dimensional space. To achieve this goal, three steps needed to be performed. The first step is to calculate the distance between the means of different classes, which is called *the between-class variance (S_B)* or *between-class matrix (S_W)*, and followed by computing the distance between the mean and the data points of each class, which is called *the within-class variance* or *within-class matrix*. Finally, we construct *the* LDA *lower dimensional space*, which simultaneously maximizes between-class and minimizes within-class variances.

Algorithm 2: LDA algorithm

1) Given a set of *n* samples $\{X_i\}_{i=1}^{n}$, each of which is represented as a column vector of length *d*, and LDA is applied on data matrix $\mathbf{X}_{d\times n}$

2) Compute the mean of each class μ_j as: $\mu_j=\frac{1}{n_j}\sum_{x_i\in w_j}x_i$

3) Compute the total mean of all data μ as:

$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{c} \frac{n_i}{n} \mu_i$, where c represents the total number of classes

4) Calculate between-class matrix S_B as follows:

$$S_B = \sum_{i=1}^{c} n_i (\mu_i - \mu)(\mu_i - \mu)^T$$

5) Compute within-class matrix $\mathbf{S_W}$, as follows:

$S_W = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T$ where x_{ij} represents the i[th] sample in the j[th] class.

6) From Equations at (4) and (5), the matrix $\boldsymbol{W}$ that maximizing Fisher's formula is calculated as: $W = S_W^{-1} S_B$.
   - The eigenvalues and eigenvectors of $\boldsymbol{W}$ are then computed using: $\det(W - \lambda I) = 0$ (or any available tool)

7) Sorting eigenvectors in descending order according to their corresponding eigenvalues, the first *k*-eigenvectors are then used as a lower dimensional space ($\boldsymbol{V_k}$).

8) Project all the original samples ($\mathbf{X}$) onto the lower dimensional space of lda using the projection: $Y = V_k^T X$.

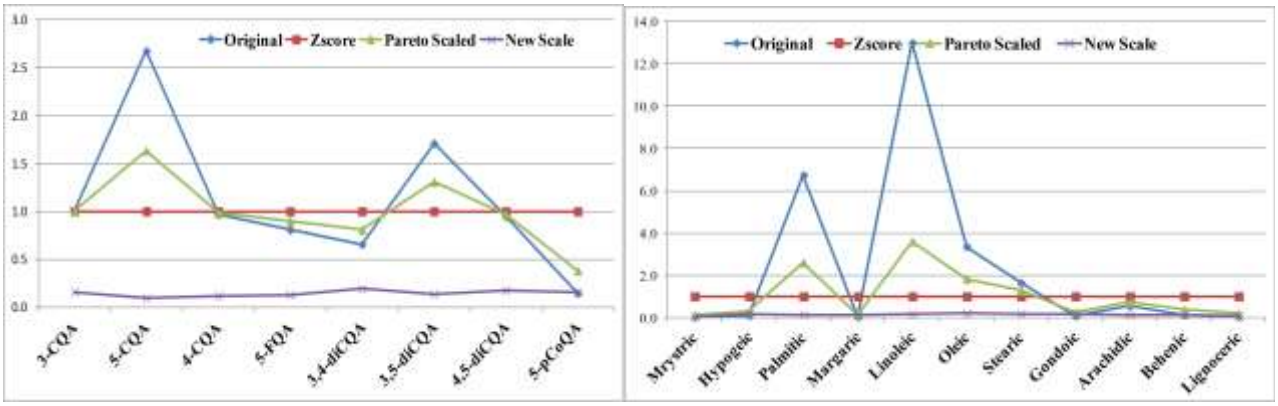
From the discussion in Section 2.3 and Section 2.4, one can notice that PCA involves basically the eigenvalue decomposition of the covariance matrix of a dataset. On the other hand, LDA finds a linear combination of observation vectors which separates two or more categories of objects by finding a low dimensional subspace that keeps data points from different classes far apart and those from the same class as close as possible.

RESULTS AND DISCUSSION

Analysis of Data Normalization

This study used improved PCA to identify the most discriminating variables (Chlorogenic acids and Fatty acids) of the Ethiopian green coffee beans according to their geographical origins. The result indicates that the improved standardization method keeps correlation information of all variables. On the other hand, the proposed method can make the standardized data retain more dispersion degree information of the original dataset compared to the PCA results with both Z-score and Pareto scaling, the most popular and widely used data normalization methods.

Figure 1 shows the distribution of standard deviations of CGAs and Fatty acids in four situations namely *from original dataset*, *Z-score standardization*, *Pareto scaling* and *improved scaling*. In Z-score and Pareto scaling, the original data points (x_{ij}) are normalized by applying $(x_{ij} - \bar{x}_j)/\sigma_{x_j}$ and $(x_{ij} - \bar{x}_j)/\sqrt{\sigma_{x_j}}$, respectively, where $\bar{x}_j$ and σ_{x_j} are the *mean* and *standard deviation* of x_j, respectively. The standard deviations of all the CGAs in four situations were 0.716, 0.000, 0.341, and 0.030, respectively, while the values were 3.916, 0.000, 1.117, and 0.039, respectively, for all Fatty acids. These confirmed that the dispersion degree information of the original dataset was retained relatively much by using the improved PCA.



(a) (b)

Figure 1. Standard deviation in four situations based on Chlorogenic acid (a) and Fatty acid (b) contents of green.

Suitability of dataset for basis of PCA

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's test of Sphericity were conducted on the dataset of the eight Chlorogenic acids and eleven Fatty acids contents of the Ethiopian green coffee beans. Accordingly, the KMO values were 0.788, 0.903 and 0.877 for Chlorogenic acids, Fatty acids, and both Chlorogenic acids and Fatty acids, respectively, exceeding the recommended minimum value of 0.5 (Field, 2000; Kaiser, 1974). It was also supported by Bartlett's test of Sphericity and found the results were significant at p-value less than 0.5. Moreover, with the exception of 5-pCoQA, the KMO value for each variable (Chlorogenic acid and Fatty acid) was above the recommended minimum value of 0.5. In addition, this result is supported by the test of one-way ANOVA ($p < 0.05$), which shows significant for all Fatty acids and Chlorogenic acids except 5-pCoQA at regional level, while significant for all at sub-regional level, indicating the mean content of the sample coffees for each variable differ significantly at both regional and sub-regional levels.

Table 1. Kaiser-Meyer-Olkin (KMO) and Bartlett's Test of Sphericity.

		Chlorogenic acids	Fatty acids	Chlorogenic acids and Fatty acids
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.788	0.903	0.877
Bartlett's Test of Sphericity	Approx. Chi-Square	240.84	863.40	1276.70
	Df	28	55	171
	Sig.	.000	.000	.000

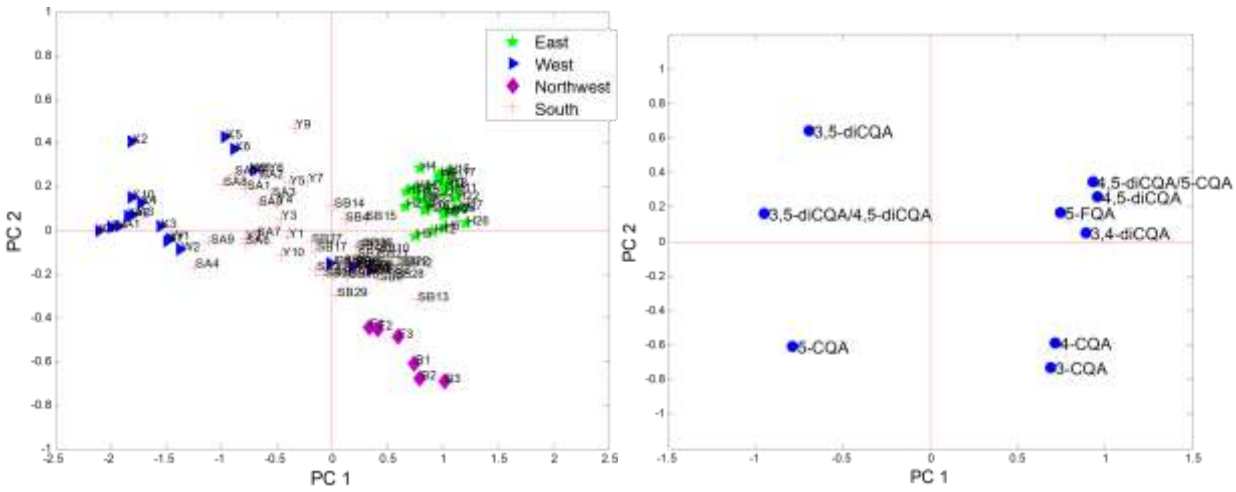
Principal Component Analysis (PCA) of the Coffee Samples

The normalized values of green coffee dataset were used directly with PCA to explore the presence of trends or patterns in the distribution of Chlorogenic acids and Fatty acids among the various regional (and sub-regional) green coffee beans. The PCA results are discussed in terms of scores and loadings. Scores are the transformed variable values, while loadings are the factors by which the original variables should be multiplied to obtain the scores. The distribution of the green coffee samples created by the scores of the first two principal components is displayed by the PCA Scores plot and the corresponding variables (Chlorogenic acids and Fatty acids) are displayed by the PCA Loadings plot.

PCA of Coffee Samples Exploration based on their Chlorogenic Acid contents

The data used for PCA at regional level consisted of 9 variables (corresponding to the selected 7 Chlorogenic acids and 2 concentration ratios) and 100 observations (corresponding to the number of

unique coffee samples). It was found that three components had chosen and the total variance of 83.1% is achieved from these three PCs. The first component explains 63.1% of the variance in the original dataset, while the second and third components explain 12.5% and 7.5% of the variances, respectively. From the scores plot, Figure 2(a), it is evident that the coffee samples tended to cluster according to their geographical origins. Coffee samples obtained from the eastern part of the country (Harar coffees) revealed marked differences in their Chlorogenic acid contents compared to coffee samples of other regions. Almost all of these coffees were separated from West, Northwest and South coffee samples by PC1. Similarly, coffee samples from Northwest region revealed marked differences in their Chlorogenic acid contents compared to coffee samples of other regions, and all of these coffees were separated from West, East and South coffees by PC2. From the score plot, some of the coffee samples originating from the West, particularly Jimma B, displayed similar Chlorogenic acid profiles to Sidama SB coffees from the South, whereas some of the coffee samples from the South, particularly Sidama SA, displayed similar Chlorogenic acid profiles to coffees from the West. On the other hand, among the green coffee beans from the South, Yirgachefe and Sidama SB differed from the other regions coffees with regard to their Chlorogenic acid profiles. These coffees typically grouped along the diagonal with coffees from Yirgachefe to the negative side of PC1 and Sidama SB to the positive side of PC1.



(a) (b)

Figure 2. PCA Scores plot (a) and Loadings plot (b) of the first two PCs at regional level based on the CGA contents of green coffee beans.

The PCA loadings plot for the first two principal components is presented in Figure 2(b). The plot displays how the individual Chlorogenic acids correlate with each other and contribute to the model. Chlorogenic acids that have high loadings (positive or negative) on each principal component have a strong impact on the model, whereas those with lower absolute values of loadings have a weaker influence.

Accordingly, using the first three principal components 4,5-diCQA; 3,5-diCQA/4,5-diCQA; 4,5-diCQA/5-CQA; 3,4-diCQA; 3-CQA; 3,5-diCQA and 5-FQA play the largest role in discriminating the green coffee beans from various regions. The first component is highly influenced by 4,5-diCQA; 3,5-diCQA/4,5-diCQA; 4,5-diCQA/5-CQA; 3,4-diCQA; and 5-FQA, whereas and 3-CQA and 3,5-diCQA contributed to separation by the second component. Accordingly, East coffees, which clustered on the positive side of PC1, were characterized mainly by their high concentrations of 4,5-diCQA, whereas Northwest coffees, which clustered on the negative side of PC2, were characterized mainly by high concentrations of 3-CQA. Moreover, with the exception of Jimma B, most coffees from West, which clustered on

negative side of PC1, were distinguished mainly by their higher 3,5-diCQA to 4,5-diCQA concentration ratios (i.e. high concentrations of 3,5-diCQA and low concentrations of 4,5-diCQA). On the other hand, most of the coffee samples from South region were characterized partly by high concentrations of 5-FQA and partly by 3,5-diCQA to 4,5-diCQA concentration ratios.

One way ANOVA test ($p < 0.05$) to assess the effect of region on the levels of the various Chlorogenic acids (CGAs) was carried out. Accordingly, the four regional coffee samples have been found to contain similar levels of total CGAs with no statistically significant difference among each other. However, the amounts of the various individual CGAs have been found to differ significantly among the regions. In line with this, Mehari, B. et al. (2016) found, with the exception of 5-pCoQA, similar levels of total Chlorogenic acids (CGAs) with no statistically significant difference in Ethiopian green coffee beans from different four studied regional categories. However, the study observed variations in the concentration of individual CGAs depending on the growing location of coffee beans.

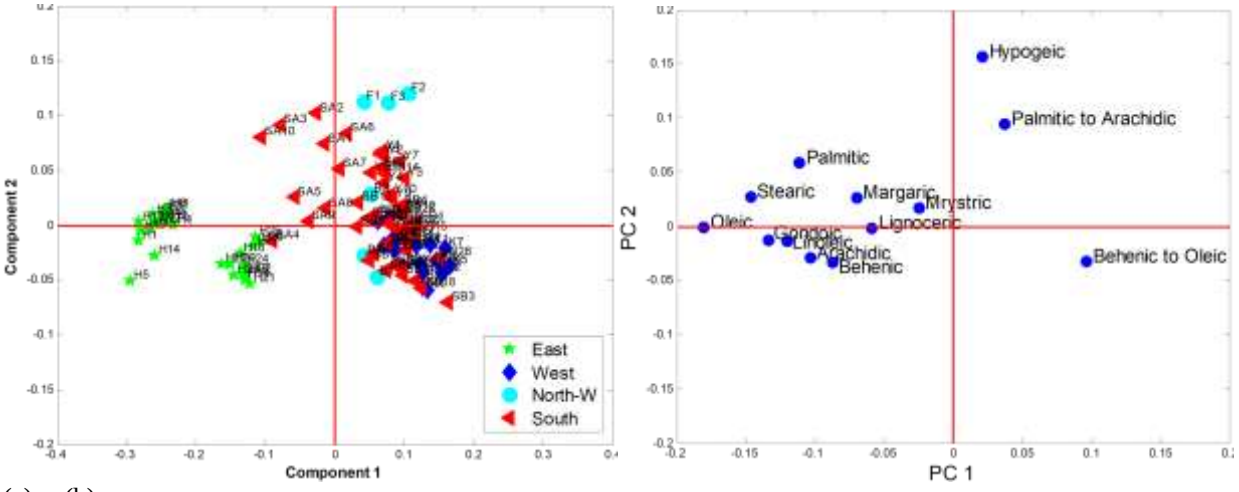
In addition to regional level, four components had chosen at sub-regional level and the total variance of 85.9% is achieved from these four PCs. The first component explains 52.3% of the variance in the original dataset, and the second, third and forth components explain 13.1%, 11.5% and 9.1% of the variances, respectively. From the score and loadings plots in PC1 versus PC2 at sub-regional levels, coffee samples from Hara were grouped on the negative side of PC1, and they were discriminated from the other sub-regional coffee types by 4,5-diCQA. These coffees contain significantly higher amount of 4,5-diCQA (average 4.9 mg/g) than coffee beans from the other regions (average 2.0–4.1 mg/g). Moreover, coffee samples from East contains higher amounts of 4,5-diCQA to 5- CQA concentration ratio (average 0.17) followed by coffees from Northwest (average 0.12) and Southern Ethiopia (average 0.11) while coffees from West contain the lowest level (average 0.08) of 4,5-di CQA to 5- CQA.The three coffee samples from West, i.e Kaffa, Jimma A and Wollega, and Sidama SA coffee samples from South were clustered on the positive side of PC1, and they were characterized by their higher 3,5-diCQA to 4,5-diCQA concentration ratios (i.e. higher content of 3,5-diCQA but smaller content of 4,5-diCQA). Moreover, results of one-way ANOVA ($p = 0.05$) indicated that the mean 3,5-diCQA /4,5-diCQA concentration ratio (3.7), calculated from those four sub-regional coffee types, differ significantly from the other sub-regional coffees, which were in the range of 1.2–2.2.

Similarly, PCA plots at the sub-regional level confirmed that coffees samples from Yirgachefe formed a separate cluster on the positive side of PC2, and they were differentiated from the other sub-regional coffees by their high 4,5-diCQA to 3,4-diCQA concentration ratios, while coffee samples from Jimma B, which clustered on positive side of PC2 and negative side of PC1, were differentiated from the other sub-regional coffees by their high 4,5-diCQA to 5-pCoQA concentration ratios. In addition, results of one-way ANOVA ($p = 0.05$) indicated that the mean 4,5-diCQA to 3,4-diCQA concentration ratio of Yirgachefe coffees (2.4) differs significantly from those of the other sub-regional coffees, which were in the range of 1.4–2.0, while the mean 4,5-di CQA to 5-pCoQA concentration ratio of Jimma B coffees (8.0) differs significantly from those of the other sub-regional coffees, which were in the range of 2.1–7.4. On the other hand, some coffee samples from Sidama SB show overlapping with Northwest and Jimma B coffees. However, applying the first four principal components, coffee samples from Benishangul and Finoteselam of Northwest region were distinguished other sub-regional coffee types by their higher contents of 3-CQA, while almost all Sidama SB coffees differed from the other sub-regional coffee types mainly because of their higher contents of 5-FQA. Moreover, results of one-way ANOVA ($p = 0.05$) also indicated that the mean 3-CQA content of Benishangul coffees (6.5 mg/g) and Finoteselam coffees (5.9 mg/g) differ significantly from the other sub-regional coffees, which were in the range of 2.5–4.2 mg/g.

PCA Coffee Samples Exploration based on their Fatty Acid contents

The data used for PCA consisted of 13 variables (corresponding to the selected 11 Fatty acids and 2 concentration ratios) and 100 observations (corresponding to the number of unique coffee samples) included. It was found that three components had chosen and the total variance of 82% is achieved from these three PCs. The first component explains 56% of the variance in the original dataset, while the second and third components explain 18% and 8% of the variances, respectively.

From the scores plot, Figure 3(a), it is evident that the coffee samples tended to cluster according to their geographical origins. Coffee samples obtained from the eastern part of the country (Harar coffees) revealed marked differences in their Fatty acid concentrations compared to other regions' coffee samples. All of these coffees were separated from West, Northwest and South coffee samples by the first component (PC1). Moreover, all coffee samples corresponding to Harar A and Harar B sub-regions from the East clearly distinguished from each other in their fatty contents. Some of the coffee samples originating from the South, particularly Sidama SB, show overlapping with some coffees from the West and Northwest. On the other hand, the green coffee beans originating from the South – i.e., Sidama SA and Yirgachefe – differed from the other regions coffees with regard to their Fatty acid contents. All coffee samples from Sidama SA were separated from East, West and Northwest coffee samples by PC1, while almost all of the coffees from Yirgachefe were grouped to the positive sides of PC1 and PC2. Similarly, Finoteselam coffee samples from Northwest region revealed marked differences in their Fatty acid compositions compared to other regions' coffee samples, and all of these coffees were separated from East, West and South coffee samples by PC2. Some of green coffee beans originating from the West, particularly Kaffa, differed from other regions coffees and almost all of these coffees were grouped to the far positive side of PC1.



(a) (b)

Figure 3. PCA Scores plot (a) and Loadings plot (b) of the first two components at regional level based on the Fatty acid contents of green coffee beans.

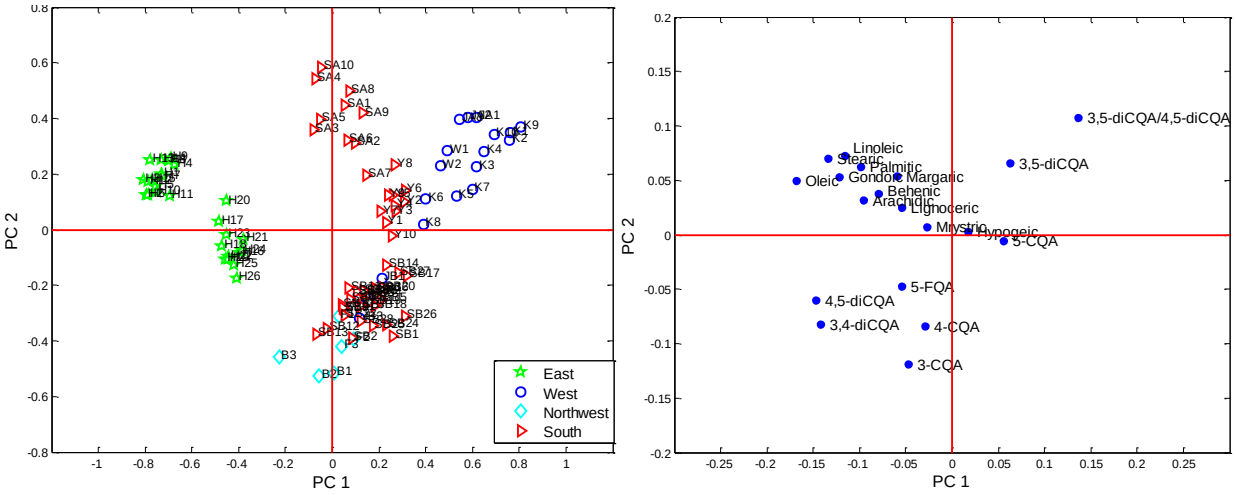
The PCA loadings plot for the first two principal components is presented in Figure 3(b). The plot displays how the individual Fatty acids correlate with each other and contribute to the model. Fatty acids that have high loadings (positive or negative) on each principal component have a strong impact on the model, whereas those with lower absolute values of loadings have a weaker influence. Accordingly, using the first two principal components Oleic, Gondoic, Stearic, Palmitic, Linoleic, Margaric, Hypogeic, Palmitic to Arachidic, and Behenic to Oleic acids play the largest role in discriminating the green coffee beans from various regions. The first component is highly influenced by Oleic acid, followed by Gondoic, Stearic, Palmitic, Linoleic and Behenic to Oleic concentration ratio, whereas Hypogeic and Palmitic to Arachidic concentration ratio contributed to separation by the second component.

Accordingly, East coffees (Harar A and Harar B), which clustered on the negative side of PC1, were characterized mainly by their high concentrations of Oleic acid, whereas Finoteselam coffees from Northwest, which clustered on the positive side of PC2, were characterized mainly by high concentrations of Hypogeic acid. On the other hand, among the green coffee beans originating from the South, Sidama SA and Yirgachefe coffees were characterized mainly by high concentrations of Margaric acid and Palmitic to Arachidic concentration ratio, respectively. Similarly, Kaffa coffee samples from West, which clustered on positive side of PC1, were distinguished mainly by their higher Behenic to Oleic concentration ratios (i.e. high concentrations of Behenic acid and low concentrations of Oleic acid).

One way ANOVA test ($p< 0.05$) to ascertain the effect of region and sub-region on the levels of the various Fatty acids was carried out. Accordingly, the four regional coffee samples have been found to contain similar levels of total Fatty acids with no statistically significant difference among each other. However, the amounts of the various individual Fatty acids have been found to differ significantly among the regions. In line with this, Mehari et al. (2019) found similar levels of total Fatty acids with no statistically significant difference in Ethiopian green coffee beans from different four studied regional categories.

PCA coffee samples exploration based on their Chlorogenic acid and Fatty Acid contents

PCA has been applied to the dataset consisted of 19 variables, corresponding to 8 Chlorogenic acids and 11 Fatty acids, and 100 observations, corresponding to the number of unique coffee samples. PCA performed on the normalized dataset of green coffee beans confirmed that coffee samples formed separate clusters in PC1 versus PC2 (64 % of total variability) according to their geographical origins. From the scores plot, Figure 4(a), coffee samples obtained from the eastern part of the country (Harar A and Harar B coffees) revealed marked differences in their Chlorogenic acid and Fatty acid contents compared to other regions' coffee samples. All of these coffees were separated from West, Northwest and South coffee samples by the first component. Moreover, all coffee samples corresponding to Harar A and Harar B sub-regions, from the East, clearly distinguished from each other in their Chlorogenic acid and Fatty acid contents. Similarly, with the exception of Jimma B coffees, all of the coffee samples from West region revealed marked differences in their Chlorogenic acid and Fatty acid contents compared to other regions' coffee samples, and these coffees were separated from Northwest, East and South coffee samples by the first component. Some of the coffee samples originating from the West, particularly Jima B, show overlapping with some coffees of Sidama SB from the South region. However, the coffee samples from the South differed from the other regions coffees with regard to their Chlorogenic acid and Fatty acid contents, and these coffee samples were separated from East and West by the first component, while they were separated from Northwest coffees by the second component. On the other hand, coffee samples from Northwest region revealed marked differences in their Chlorogenic acid and Fatty acid compositions compared to other regions' coffee samples, and almost all of these coffees were separated from East, West and South coffee samples by the second component.



(a) (b)

Figure 4. PCA Scores plot (a) and Loadings plot (b) of the first two components at regional level based on CGA and Fatty acid composition.

The loadings plot for the first two principal components displays how the individual Chlorogenic acids and Fatty acids correlate with each other and contribute to the PCA model. The variables (Chlorogenic acids and Fatty acids) that have high loadings (positive or negative) on each principal component have a strong impact on the model, whereas those with lower absolute values of loadings have a weaker influence. Accordingly, using the first two principal components Oleic acid, 4,5-diCQA, Gondoic acid, Arachidic acid, Stearic acid, Linoleic acid, 3,4-diCQA, 3,5-diCQA/4,5-diCQA, Palmitic acid, 3-CQA, and Margaric acid play the largest role in discriminating the green coffee beans from various regions. The first component is highly influenced by Oleic acid, followed by 4,5-diCQA, Gondoic acid, Arachidic acid, Stearic acid, Linoleic acid, 3,4-diCQA, 3,5-diCQA/4,5-diCQA, and Palmitic acid, whereas 3-CQ, 4-CQA and Margaric acid contributed to separation by the second component.

Accordingly, East coffees (Harar A and Harar B), which clustered on the negative side of PC1, were characterized mainly by their high contents of Oleic acid, whereas coffee samples from West, which clustered on the positive side of the first component, were characterized mainly by their high contents of 3,5-diCQA to 4,5-CQA concentration ratios. Similarly, coffee samples from Northwest, which clustered on negative side of the second component, were distinguished mainly by their high contents of 3-CQA. On the other hand, among the green coffee beans originating from the South, Sidama SA coffees were characterized mainly by high contents of Margaric acid.

One way ANOVA test ($p< 0.05$) to assess the effect of region and sub-region on the levels of the various Chlorogenic acids and Fatty acids was carried out. Accordingly, the four regional coffee samples have been found to contain similar levels of total Chlorogenic acids and Fatty acids with no statistically significant difference among each other. However, the amounts of the various individual Chlorogenic acids and Fatty acids have been found to differ significantly among the regions (and sub-regions).

Moreover, from the score and loadings plots in PC1 versus PC2 at sub-regional level, coffee samples from Hara A and Harar B were grouped on the negative side of PC1, and they were discriminated from the other sub-regional coffee types by Oleic acid contents. Next to Oleic acid, both these coffee sub-types were discriminated from the other coffee varieties by their high contents of 4,5-diCQA. The three coffee samples from West, i.e Kaffa, Jimma A and Wollega coffees were clustered on the positive side of PC1, and they were characterized by their higher 3,5-diCQA to 4,5-diCQA concentration ratios.

Similarly, PCA plots at the sub-regional level confirmed that coffees samples from Yirgachefe formed a separate cluster on the positive side of PC1 while coffee samples from Sidama SA clearly separated to the far negative side of PC2. Yirgachefe coffees were differentiated from the

other sub-regional coffees by their high 4,5-diCQA to 3,4-diCQA concentration ratios, while coffee samples from Sidama SA were differentiated from the other sub-regional coffees by their high Margaric acid contents. In the same way Finoteselam coffee sample, which clustered to the far positive side of PC2, were differentiated from the other sub-regional coffees by 3-CQA. On the other hand, some coffee samples from Sidama SB show overlapping with Benishangul and Jimma B coffee samples. However, applying the combination of the first four principal components, coffee samples from Benishangul and Jimma B were distinguished from other sub-regional coffee types by their higher contents of 3-CQA and 4,5-diCQA to 5-pCoQA concentration ratios, respectively, while Sidama SB coffees differed from the other sub-regional coffee types mainly because of their higher contents of 5-FQA.

Hence, the PCA results revealed that Fatty acid contents are suitable to clearly discriminate green coffee beans from the Eastern part (Harar-A and Harar-B) from other coffee varieties in Ethiopia, whereas coffees from Western and Northwest parts are clearly discriminated from other coffee varieties by their higher Chlorogenic acid contents than Fatty acid contents. On the other hand, green coffee beans from South are partly differentiated from other coffee types by Fatty acid contents, particularly Sidama SA, and partly by Chlorogenic acid contents, particularly Sidama SB and Yirgachefe green coffee beans.

LDA Coffee Samples Classification

Discriminant analysis at Regional level

Based on the concentrations of the identified Chlorogenic acids and Fatty acids, an attempt was made to construct classification model useful for the discrimination of the most important compounds, which helps the authentication of the geographical origin of the coffee beans. LDA, a technique that generates a set of discriminant functions, was applied to achieve this aim. These functions are based on linear combinations of the descriptor variables that provide the best discrimination among groups of coffee samples. The generated functions can then be applied to new samples that have measurements for the descriptor variables, but have unknown group membership, thus allowing the group membership to be predicted. All of the four groups of coffee samples – i.e., East, West, Northwest and South – were assigned equal prior probabilities. Finally, the reliability of the LDA model, at the regional level, was assessed in terms of its recognition and prediction abilities. For this, the entire sample set was divided into a training set and testing (an external validation) set. The testing set consisted of 30 (30%) randomly selected samples, while, the remaining 70 (70%) samples were used as training set to construct the LDA model at the regional level. Hence, 19 East, 13 West, 4 Northwest and 34 South samples of green coffee beans were included in training set and the remaining samples were included in testing set.

Discriminant analysis at Regional level based on Chlorogenic acids

All of the seven Chlorogenic acids and one concentration ratios were used simultaneously to construct the LDA model at the regional level. Three canonical discriminant functions were subsequently computed. The magnitude of Wilks' λ encompassing the three functions was 0.024, reflecting that the LDA model accounts for almost all of the variation in the dataset. Wilks' λ indicates the proportion of the total variance in the discriminant scores not explained by differences among the classified groups of samples. Smaller values of Wilks' λ indicate greater discriminatory ability of the computed functions. The first two discriminant functions together accounted for 90% of the total variance in the dataset.

The contribution of each Chlorogenic acid to the LDA model was assessed from the structure matrix and tests of equality of group means. The structure matrix (Table 2) indicates the correlation of each Chlorogenic acid with the discriminant functions. Accordingly, 4,5-diCQA is correlated most strongly to the first discriminant function, followed by 3,4-diCQA and 5-CQA, while the second function is correlated most strongly with 3,5-diCQA/4,5-diCQA, followed by 3-CQA and 4,5-diCQA. The test of equality of group means, which is a measure of a variable's potential before the discriminant model is created, is carried out with a one-way ANOVA for each variable, using the grouping variable as the factor. From the significance values in test of equality of group means, it can be concluded that (with p=0.05) each of the seven Chlorogenic acids and a concentration ratio contributed significantly to the model. Since Wilks' Lambda is also a measure of a variable's potential, 4,5-diCQA, followed by 3,5-diCQA/4,5-

diCQA; 3,4-diCQA and 3-CQA were indicated as the best discriminating variables to distinguish the four regional coffee samples.

Table 2. Structure Matrix based on CGAS.

	Function		
Chlorogenic Acids	**1**	**2**	**3**
5-CQA	**-.281**	.203	.180
3,5-diCQA/4,5-diCQA	-.273	**.817**	.201
3-CQA	.022	**-.733**	**.446**
4,5-diCQA	**.499**	**-.558**	-.016
4-CQA	.026	-.523	.245
3,5-diCQA	-.091	.494	-.023
3,4-diCQA	**.317**	-.446	.098
5-FQA	.185	-.214	**-.343**

The distribution of the bean samples on the plane formed by the discriminant function scores is indicated in Figure 5. The first function differentiates East coffees from the group formed by South, West and Northwest coffees. This function explains 74% of the variation in the dataset and is strongly associated and influenced to the positive side with 4,5-diCQA (Table 2). Hence, the grouping of East coffees to the far positive side of the first function can be attributed mainly to their higher content of 4,5-diCQA. This is also in agreement with the previous results of one-way ANOVA. Northwest coffees and most of the coffees from West cluster far to the negative and positive sides, respectively, of the second function. This function is correlated most strongly with 3,5-diCQA/4,5-diCQA, followed by 3-CQA. Hence, the grouping of Northwest coffees to the far negative side and most of West coffees to the far positive side of the second function can be attributed mainly to their higher content of 3-CQA and 3,5-diCQA/4,5-diCQA, respectively. On the other hand, coffee samples from South and few samples from West tend to show a similar Chlorogenic acid profile that is evident from their partial overlap on the PCA regional scores plot. However, they are significantly separated from one another by the combined effect of the first, second, and third discriminant functions. Accordingly, coffee samples from South region were separated by the third function, which is highly correlated and influenced to the negative side by 5-FQA and 3,5-diCQA.

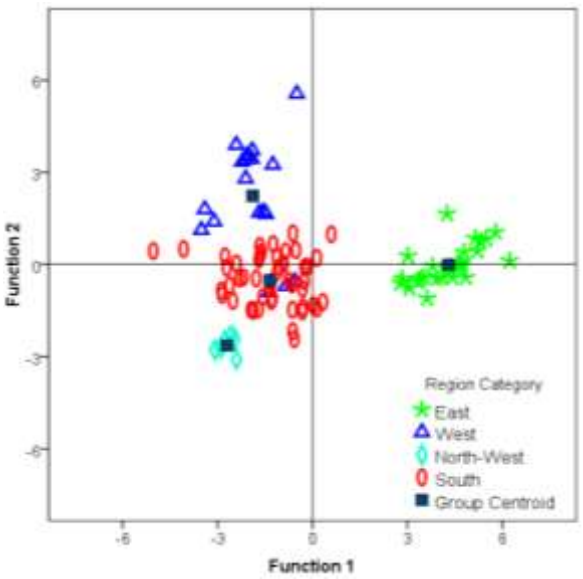


Figure 5: Scatter plot of the first two canonical discriminant function scores at regional level based on CGA contents

Analyzing the Chlorogenic acid content by LDA permitted successful classification of the coffee beans into the four regions studied. The computed three discriminant functions classified 96 of the 100 samples correctly, with three samples from West misclassified as South coffees, while one sample from South misclassified as West coffees. This overall proportion of correct classification obtained in this study (96%) is better than that achieved (92%) by Mehari, B. et al. (2016) for the classification of Ethiopian green coffee beans, based on Chlorogenic acid content, from the four regions studied, and (83%) correct classification obtained by Bertrand et al. (2008) as cited in (Mehari et al., 2016) following linear discriminant analysis of green Arabica coffee beans from three Colombian locations, based on their Chlorogenic acid contents.

The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions computed from the other samples in the dataset; i.e., each sample is treated as unknown and its class is determined based on the discriminant functions computed from the remainder of the samples. The percentage of cross-validated samples that were correctly classified provided an indication of the number of new samples, belonging to the groups of samples studied, that can be correctly classified by the LDA

model. Accordingly, 93% of cross-validated groups of samples were correctly classified.

In addition, calculating average of the values of five iterations, the recognition ability of the model, calculated as the percentage of the members of the training set that are correctly classified, was 95.7 %. Moreover, the prediction ability of the LDA model, calculated as the percentage of the members of the external validation set correctly classified by using the model developed in the training step, was 94 % at the regional level.

Discriminant analysis at Regional level based on Fatty acids

All of the 11 Fatty acids and 2 concentration ratios were used simultaneously to construct the LDA model. Three canonical discriminant functions were subsequently computed. The magnitude of Wilks' λ encompassing the three functions was 0.037, which indicates the proportion of the total variance in the discriminant scores not explained (i.e. 3.7%) by differences among the classified groups of samples. The first two discriminant functions together accounted for 91% of the total variance in the dataset.

The contribution of each fatty acid to the LDA model was assessed from the structure matrix and tests of equality of group means. The structure matrix (Table 3) indicates the correlation of each fatty acid with each discriminant function. Accordingly, Oleic acid is correlated most strongly to the first discriminant function, followed by Arachidic, Gondoic, and Stearic acids. The second function is correlated most strongly with Hypogeic acid, followed by Linoleic acid, while the third function is correlated most strongly with Stearic acid, followed by Margaric, Mrystic, Linoleic and Palmitic acids. The test of equality of group means, which is a measure of a variable's potential before the discriminant model is created, is carried out with a one-way ANOVA for each variable, using the grouping variable as the factor. From the significance values of test of equality of group means, it can be concluded that (with p=0.05) each fatty acid contributed significantly to the model. Since Wilks' λ is also a measure of a variable's potential, Oleic acid, followed by Gondoic, Arachidic, Stearic, and Linoleic acids were indicated as the best discriminating variables to distinguish the four regional coffee samples.

Table 3. Structure Matrix based on Fatty acids.

Fatty Acid	Function		
	1	2	3
Oleic Acid	.694*	.294	.263
Arachidic Acid	.492*	.005	.232
Gondoic Acid	.489*	-.076	.480
Behenic Acid	.367*	-.021	-.066
Lignoceric Acid	.347*	.092	.203
Palmitic/Arachidic	-.179*	.169	.117
Hypogeic Acid	-.141	.482*	.145
Stearic Acid	.300	.170	.682*
Margaric Acid	.437	.053	.628*
Mrystic Acid	.190	.027	.438*
Linoleic Acid	.089	.402	.417*
Palmitic Acid	.281	.192	.315*
Behenic/Oleic	.152	-.167	-.201*

*. *Largest absolute correlation between each variable and any discriminant function*

The distribution of the bean samples on the plane formed by the discriminant function scores is indicated in Figure 6. The first function differentiates East coffees from the group formed by South, West and Northwest coffees. This function explains 80.3% of the variation in the dataset and is strongly associated and influenced to the positive side with Oleic acid (Table 3). Hence, the grouping of East coffees to the far positive side of the first function can be attributed mainly to their higher content of Oleic acid. This is also in agreement with the previous results of one-way ANOVA. Almost all of the coffee samples from Northwest and most of the samples from West cluster to the positive and negative sides, respectively, of the second function. The second function is positively correlated with Hypogeic acid and negatively correlated with Lignoceric to Hypogeic acid concentration ratio. Hence, the grouping of Northwest and West coffees to the positive and negative sides, respectively, of the second function can be attributed mainly to their higher content of Hypogeic acid and Behenic acid to Oleic acid concentration ratios, respectively. On the other hand, some coffee samples from South and some samples from West tend to show a similar Fatty acid profile. However, they are significantly separated from one another by the combined effect of the first, second, and third

discriminant functions. Accordingly, coffee samples from South region were separated by the third function that is highly correlated and influenced to the positive side by Stearic acid, which is then by Margaric and Mrystic acids.

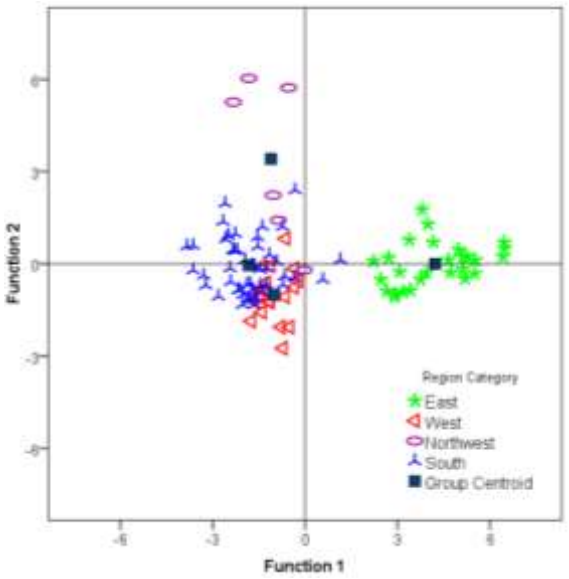


Figure 6. Scatter plot of the first two canonical discriminant function scores of green coffee beans at regional level based on Fatty acid composition.

Analyzing the Fatty acid content by LDA permitted successful classification of the coffee beans into the four regions studied. The computed three discriminant functions classified 94 of the 100 samples correctly, with five samples from South misclassified as West and Northwest coffees, while one sample from Northwest misclassified as south coffee. This overall proportion of correct classification obtained in this study (94%) is comparable with that achieved (96%) by Mehari et al. (2019) for the classification of Ethiopian green coffee beans, based on Fatty acid content, from the four regions studied. The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions computed from the other samples in the dataset. Accordingly, 86% of cross-validated groups of samples were correctly classified.

In addition, calculating average of the values from five iterations, the recognition ability of the model, calculated as the percentage of the members of the training set that are correctly classified, was 91%, and the prediction ability of the LDA model, calculated as the percentage of the members of the external validation set correctly classified by using the model developed in the training step, was 97% at the regional level.

Discriminant analysis at Regional level based on both Chlorogenic acids and Fatty acids

Here 8 Chlorogenic acids and all of the 11 Fatty acids, together 19 variables, were used simultaneously to construct the LDA model. Three canonical discriminant functions were computed and the magnitude of Wilks' λ encompassing the three functions was 0.006, reflecting that the LDA model accounts for almost all of the variation in the dataset. Wilks' λ indicates the proportion of the total variance in the discriminant scores not explained by differences among the classified groups of samples, and hence, this small value of Wilks' λ indicates greater discriminatory ability of the computed functions. The first two discriminant functions together accounted for 91% of the total variance in the dataset.

The contribution of each variable (Chlorogenic acid and Fatty acid) to the LDA model was assessed from the structure matrix, which indicates the correlation of each variable with each discriminant function. Accordingly, Oleic acid is correlated most strongly to the first discriminant function, followed by Gondoic acid, 4,5-diCQA, , Arachidic and Stearic acids. The second function is correlated most strongly with 3-CQA, followed by 4-CQA and 3,5-diCQA, while the third function has the largest absolute correlation with 5-FQA, followed by Margaric acid . The test of equality of group means, which is a measure of a variable's potential before the discriminant model is created, is carried out with a one-way ANOVA for each variable, using the grouping variable as the factor. From the significance values, it can be concluded that (with p=0.05) each Chlorogenic acid and each Fatty acid contributed significantly to the model. Since Wilks' λ is also a measure of a variable's potential, Oleic acid, followed by 4,5-diCQA, Gondoic acid, Arachidic acid, Stearic acid, 3,4-diCQA and 3-CQA were indicated as the best discriminating variables to distinguish the four regional coffee samples.

Table 4. Structure Matrix based on CGA and Fatty acid contents.

Chlorogenic acids and Fatty acids	Function 1	2	3
Oleic Acid	.496*	.104	.284
Gondoic Acid	.370*	-.014	-.060
4,5-diCQA	.369*	.302	-.048
Arachidic Acid	.355*	-.017	.118
Stearic Acid	.343*	.079	-.145
Lignoceric Acid	.253*	.038	.083
Behenic Acid	.250*	-.064	.231
Linoleic Acid	.243*	.106	-.052
Palmitic Acid	.214*	.112	.005
5-CQA	-.205*	-.089	.144
3-CQA	.033	.493*	.259
4-CQA	.032	.347*	.130
3,5-diCQA	-.077	-.308*	.030
Hypogeic Acid	-.090	.273*	-.036
3,4-diCQA	.236	.256*	.041
Mrystic Acid	.087	.253*	-.104
5-FQA	.138	.096	-.271*
Margaric Acid	.160	.064	-.164*

*. *Largest absolute correlation between each variable and any discriminant function*

The distribution of the bean samples on the plane formed by the discriminant function scores is indicated in Figure 7. The first function differentiates East coffees from the group formed by South, West and Northwest coffees. This function explains 73% of the variation in the dataset and is strongly associated and influenced to the positive side with Oleic acid (Table 4). Hence, the grouping of East coffees to the far positive side of the first function can be attributed mainly to their higher content of Oleic acid. This is also in agreement with the results of one-way ANOVA. Similarly, the second function differentiates Northwest coffees from the group formed by East, West and South coffee samples. This function is strongly correlated and influenced to the positive side with 3-CQA. Hence, the grouping of Northwest coffees to the positive side of the second function can be attributed mainly to their higher content of 3-CQA. Almost all of the coffees from West cluster to the negative sides of first and second functions, to some extent to the second function. The second discriminant function is strongly and negatively correlated with 3,5-diCQA. Hence, the grouping of West coffees to the negative sides of the second function can be attributed mainly to their higher content of 3,5-diCQA or 3,5-diCQA to 4,5-diCQA concentration ratios (i.e. higher content of 3,5-diCQA but lower content of 4,5-diCQA). On the other hand, the coffee samples from South are significantly separated from groups formed by West, Northwest and East coffees by the combined effect of the first and second discriminant functions. Accordingly, coffee samples from South region were separated from East coffees by the first function and from the West and Northwest coffees by the second function. Furthermore, using function 2 versus function 3, coffee samples from South were significantly separated from other regional coffees on the negative side of the third function, which has the largest absolute correlation with 5-FQA and Margaric acid.

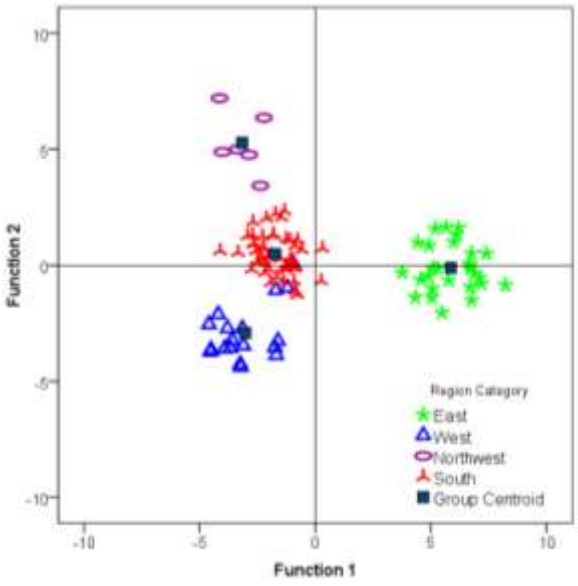


Figure 7: Scatter plot of the first two canonical discriminant function scores of green coffee beans at regional level based on CGA and Fatty acid composition

The result revealed that using both Chlorogenic acid and Fatty acid content of the green coffee beans by LDA permitted successful classification of the coffee beans into the four regional categories studied. The computed three discriminant functions classified 99 of the 100 samples correctly, with only one sample from West misclassified as South coffee. This overall proportion of correct classification of Ethiopian green coffee beans based on both Chlorogenic acid and Fatty acid contents (99%) is much better than that achieved (96%) for the classification green coffee beans based on Chlorogenic acid content and that (94%) for the classification coffee beans

based on Fatty acid content, from the four regions studied. The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions computed from the other samples in the dataset. Accordingly, 94% of cross-validated groups of coffee samples were correctly classified. In addition, calculating average of the values of five iterations, the recognition and prediction abilities of the model were 99% and 100%, respectively, for the classification of Ethiopian green coffee beans at the regional level based on the their Chlorogenic acid and Fatty acid contents.

Discriminant analysis at Sub-Regional level

Besides classifying the coffee beans into the four geographical regions studied, an attempt was also made to classify the sub-regional types of coffees based on their Chlorogenic acid content, Fatty acid content, and both Chlorogenic acid and Fatty acid. Since Harar, Jimma and Sidama were further subdivided into two subtypes, the coffee types totaled 11. According to Mehari et al. (2016), Sidama SA comprises Sidama A and Sidama B, whereas, Sidama SB comprises Sidama C and Sidama E coffees. Accordingly, Sidama SA category includes coffees originated from Arroressa, Chuko, Kercha, Nensebo, Uraga, Benssa, Berbere, Harena Buliki and Chire districts, while Sidama SB includes coffees from Segen Hizboch, Damot Sore, South Ari, Soddo Zuria, Konso, Arbaminch Zuria, Shone, Durame and Kedida Gamela districts. Hence, Sidama SB includes coffees originated from different districts of neighboring zones other than Sidama zone. In the same way, Jimma A category represents coffees from Limmu Kossa, Gomma and Gera districts of Jimma zone, whereas Jimma B category represents coffees from Bedelle and Dedessa districts of Illubabour zone (ECX, 2014; Mehari et al., 2016).

Finally, to assess the recognition and prediction abilities of the LDA model at sub-regional level, the entire sample set was divided into a Training set and Testing set. The testing set consisted of 30 randomly selected samples, while, the remaining 70 samples were used as training set to construct the LDA model. Due to their sample size, all of samples of Wollega (2), Jimma A (3), Jimma B (3), Benishangul (3) and Finoteselam (3) were used exclusively in the testing set.

A) Sub-Regional Classification Based on Chlorogenic acids

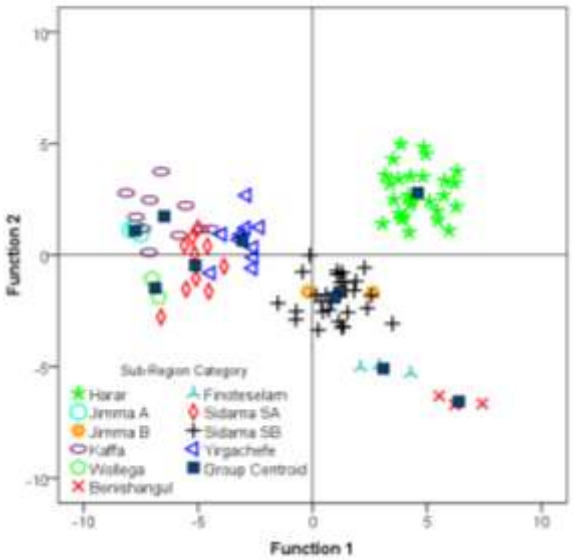


Figure 8. Scatter plot of the first two discriminant functions for green coffee beans by Sub-Region based on CGA contents

By considering all the sub-regional coffee types, a 97% correct classification was achieved. One sample of Kaffa coffee and two samples of Sidama SB coffee were incorrectly classified as Yirgachefe and Jimma B coffee types, respectively. The first and second canonical discriminant functions together explained 87% of the total variance in the dataset. The structure matrix shows that 4,5-diCQA and 3-CQA Chlorogenic acids and 3,5-diCQA to 4,5-diCQA concentration ratio were the major contributors to the first and second functions. This overall proportion of correct classification obtained in this study (97%) is better than that (89%) obtained by Mehari, B. et al. (2016), for the classification of the eight coffee varieties, i.e. Harar, Jimma (Jimma A and Jimma B in one class), Kaffa, Wollega, Sidama (Sidama SA and Sidama SB in one class), Yirgachefe, and Northwest coffees (Benishangul and Finoteselam in one class).

The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions derived from all samples other than that sample in the entire dataset. Accordingly, 91% of cross-validated groups of samples were correctly classified. In addition, the

recognition and prediction abilities of the LDA model at sub-regional level were 98% and 98.6%, respectively. Overall, the proportion of correct classification (97%) and cross-validation (91%) obtained for the classification of the various sub-regional types of coffees are comparable with that obtained at the regional level (96% and 93%, respectively). However, the recognition (98%) and prediction ability (98.6%) of the LDA model at sub-regional level were significantly higher than that obtained at the regional level (95.7% and 94%, respectively). Hence, Chlorogenic acids are a better indicator of the geographical sub-regions of origin of Ethiopian coffees than of the regional types. It is evident that discriminant models constructed from the Chlorogenic acid concentrations of green coffee beans are a useful tool for the authentication of Ethiopian coffees. A high prediction success rate (98.6%) was obtained relative to their sub-regions of origin than that prediction ability (94%), which was obtained with the four major regional coffee types.

B) Sub-Regional Classification Based on Fatty acids

By considering all the 11 sub-regional coffee types, a 96% correct classification was achieved. Four samples of Sidama SB coffee were incorrectly classified as Jimma A (1), Jimma B (2) and Kaffa (1) coffee types. The first and second canonical discriminant functions together explained 94% of the total variance in the dataset.

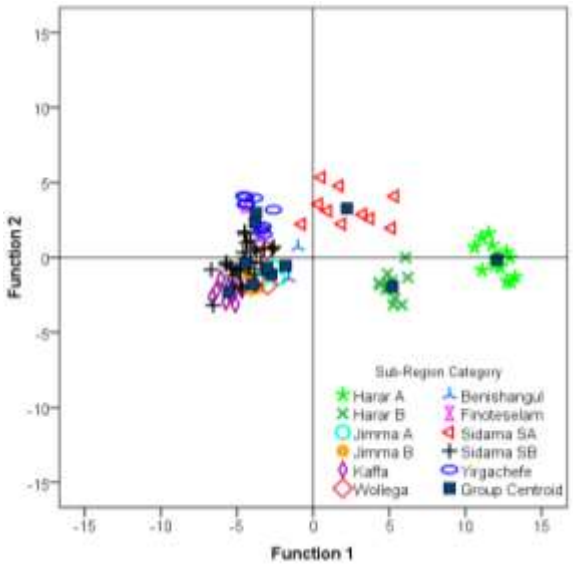


Figure 9: Scatter plot of the first two functions for Green Coffee beans by Sub-Region based on fatty acid contents

The structure matrix shows that Oleic, Stearic, Gondoic, Linoleic, Palmitic and Arachidic acids were the major contributors to the first function, and that was agreeable with the test of ANOVA. This overall proportion of correct classification obtained in this study (96%) is comparable with that (94%) obtained by Mehari et al. (2019), classification of Ethiopian green coffee beans based their geographical origin. The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions derived from all samples other than that sample in the entire dataset. Accordingly, 71% of cross-validated groups of samples were correctly classified.

The recognition and prediction abilities of the LDA model at sub-regional level based on Fatty acid contents were 96% and 96.7%, respectively. Overall, the proportion of correct classification and prediction ability of the model obtained for the classification of the various sub-regional coffee types based on their fatty acid contents were comparable with that obtained at the regional level. However, the recognition (96%) of the LDA model at sub-regional level was significantly higher than that obtained at the regional level (91%).

C) Sub-Regional Classification based on both Chlorogenic acids and Fatty acids

In this study, all the 11 sub-regional coffee types were considered and applying a linear discriminant analysis on dataset of Chlorogenic acids and Fatty acids, a 100% correct classification was achieved at sub-regional coffee types. The first and second canonical discriminant functions together explained 85% of the total variance in the dataset. The structure matrix shows that Oleic acid, followed by Stearic acid, 4,5-diCQA, 3,5-diCQA/4,5-diCQA, Gondoic acid, Linoleic acid, 3-CQA, Arachidic acid, and Palmitic acid were the major contributors to the first and second functions, and that was supported with the test of ANOVA. This overall proportion of correct classification obtained in this study (100%) is much better than that (97%) obtained in this study for classification of Ethiopian green coffee beans based on their Chlorogenic acid contents at sub-regional level and that (96%) obtained based on their Fatty acid content at sub-regional levels. The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions derived

from all samples other than that sample in the entire dataset. Accordingly, 94% of cross-validated groups of samples were correctly classified.

The recognition and prediction abilities of the LDA model based on both Chlorogenic acid and Fatty acid contents at sub-regional level were both 100%. Overall, the proportion of correct classification (100%) and cross-validation (94%) obtained for the classification of the various sub-regional coffee types were comparable with that obtained at the regional level (99% and 94%, respectively). Moreover, the highest recognition abilities of 99% and 100% at regional and sub-regional levels, respectively, and the highest prediction success rate (100%) for regional and sub-regional coffee samples classification were achieved in this study. Hence, undertaking the simultaneous analysis of Chlorogenic acids and Fatty acids is better for the discrimination of the geographical origins and discriminant models constructed from the Chlorogenic acid and Fatty acid concentrations of green coffee beans are a useful tool for the authentication of Ethiopian coffees.

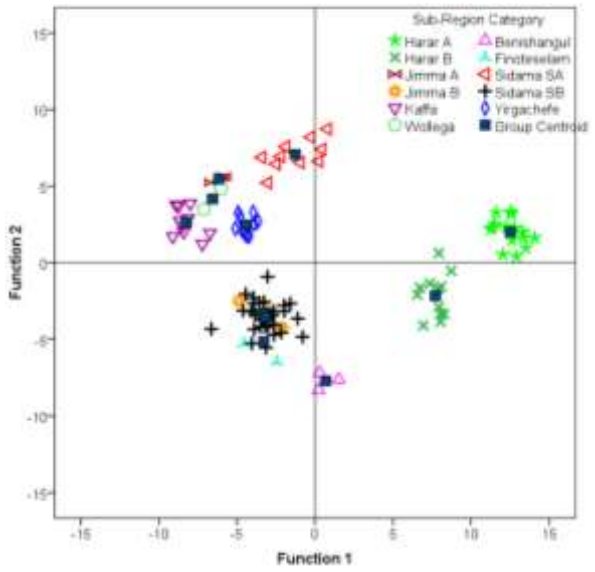


Figure 10. Scatter plot of the first two functions for Green Coffee beans by Sub-Region based on CGA and Fatty acid contents

To compare the results of our method with the previous ones, this study found that the recognition and prediction abilities of the improved PCA and LDA at regional level are: using CGA contents 95.7% and 94%, using FA contents 91% and 97%, and using the combined CGA and FA contents 99% and 100%, respectively. Similarly, the recognition and prediction abilities of the improved PCA and LDA at sub- regional level are: using CGAs 98% and 98.6%, using FAs 96% and 96.7%, and using the combined CGA and FA contents 100% and 100%, respectively. The previous researchers Bewketu Mehari et al. (2016, 2019) reported that the recognition and prediction of the PCA that they applied on the same dataset at regional level were: using CGAs 91% and 90%, using FAs 95% and 92%, respectively, while for sub-regional method these were: using CGA contents 89% and 86% , using FA contents 95% and 73%, respectively.

It is evident that our results which are obtained by using the improved PCA and LDA are superior than the results obtained by the previous researchers. In particular, the improved PCA and LDA methods that we have applied using the combination of CGA and FA contents achieved accuracy level of classification up to 99 to 100%. This higher level of accuracy is mainly due to the following two special features of our methods: (i) The proposed improved PCA method can make the standardized data retain more dispersion degree information of the original dataset compared to the usual PCA methods. (ii) The use of combination of the CGA and FA contents in the proposed method help to improve the accuracy levels of the classifications.

CONCLUSIONS

The study used improved PCA to identify the most important discriminant variables (Chlorogenic acids and Fatty acids), aiming to assist the concerned bodies to better understand the authentication of coffees based on their geographical origin. Exploratory analysis by improved PCA and LDA showed, in general, good discrimination capabilities among the different regional and sub-regional coffees varieties. Accordingly, 4, 5-diCQA; 3-CQA; and 3,5-diCQA to 4,5-diCQA concentration ratios were selected as suitable discriminant marker CGAs for green coffee beans originating from East, Northwest (Benishangul and Finoteselam) and West (Kaffa, Jimma A and Wollega), respectively, both at regional and sub-regional levels. Moreover, 4,5-diCQA to 3,4-diCQA; 4,5-diCQA to 5-pCoQA; and 5-CQA were found appropriate to differentiate green coffee beans from Yirgachefe, Jimma B and Sidama SB, respectively, at sub-regional level. Based on fatty acid contents, Oleic acid, followed by

Gondoic, Arachidic, Stearic, Linoleic, Margaric and Hypogeic acids were identified as the best discriminating fatty acids to distinguish the four regional coffee samples. Oleic acid has found to be suitable discriminant fatty acid for East (Hara A and Harar B) coffees. Moreover, Margaric acid and Palmitic to Arachidic concentration ratios were found to be suitable marker for Sidama SA and Yirgachefe coffee samples, respectively, at sub-regional level. The PCA results, from the dataset including both Chlorogenic acid and Fatty acid variables, revealed that fatty acid contents are suitable to clearly discriminate green coffee beans from the Eastern part (Harar) from other coffee varieties in Ethiopia, whereas coffees from Western and Northwest parts are clearly discriminated from other coffee varieties by their higher Chlorogenic acid contents than Fatty acid contents. On the other hand, green coffee beans from South are partly differentiated from other coffee types by Fatty acid contents, particularly Sidama SA, and partly by Chlorogenic acid contents, particularly Sidama SB and Yirgachefe green coffee beans. The results of LDA were in line with the PCA results, indicating that the LDA model was able to classify the coffee beans accurately based according to their regional as well as sub-regional geographical origin based on their Chlorogenic acid and Fatty acid contents. The recognition and prediction abilities of the LDA model were 95.7% and 94%, respectively, based on CGA contents; 91% and 97%, respectively, based on Fatty acid contents; and 99% and 100%, respectively, based on the combined analysis of CGA and Fatty acid contents, at the regional levels, while the values were 98% and 98.6%, respectively, based on CGA contents; 96% and 96.7%, respectively, based on Fatty acid contents; and both 100%, based on the combined analysis of CGA and Fatty acid contents, at the sub-regional levels. Finally, this study recommends that the government and concerned bodies should use the simultaneous analysis of Chlorogenic acid and Fatty acid contents to address the characterization, classification and authentication of Ethiopian coffee beans according to their geographical origins.

AKNOWLEGEMENT

Mr Endale Deribe Jiru expresses his gratirude to Addis Ababa University for partially sponsoring his study.

REFERENCES

1. Adnan U., Usman Q., Farhan H.K., and Saba B. (2017). Dimensionality Reduction Approaches and Evolving Challenges in High Dimensional Data. *Conference Paper*, October 2017. DOI: 10.1145/3109761.3158407
2. Arunasakthi K., and Kamatchipriya L. (2014), ‗A Review On Linear And Non-Linear Dimensionality Reduction Techniques, Machine Learning And Applications: *An Int. J. (Mlaij),* **1(1)**:.65-76.
3. Bertrand B, Villarreal D, Laffargue A, Posada H, Lashermes P, Dussert S (2008). Comparison of the effectiveness of Fatty acids, Fatty acids, and elements for the chemometric discrimination of coffee (Coffea arabica L.) varieties and growing origins. *J Agric Food Chem* **56:**2273–2280.
4. Bishop, C.M. (2006). *Pattern recognition and machine learning. Information science and statistics*. New York : Springer.
5. Breger, A., Orlando, J.I., and Harar, P. (2020). On Orthogonal Projections for Dimension Reduction and Applications in Augmented Target Loss Functions for Learning Problems. *J Math Imaging Vis* ***62,*** 376–394 (2020). https://doi.org/10.1007/s10851-019-00902-2
6. Cai, W., Dou, L.M., Si, G.Y., Cao, A.Y., He, J., and Liu, S. (2016). A principal component analysis/fuzzy comprehensive evaluation model for coal burst liability assessment. *Int. J. Rock Mech. Min. Sci.* 2016, **81**, 62–69.
7. Cattell, R. B. 1966. The screen test for the number of factors. Multivariate Behavioral Research 1:245-276.
8. Charu C.A. (2014). Data Classification, Algorithms and Applications. *Chapman and Hall/*CRC 2014, Pages 37–64 Print ISBN: 978-1-4665-8674-1 eBook ISBN: 978-1-4665-8675-8
9. Coussement, A.; Isaac, B.J.; Gicquel, O.; Parente, A. (2016). Assessment of different chemistry reduction methods based on principal component analysis: Comparison of the MG-PCA and score-PCA approaches. *Combust.* Flame 2016, 168, 83–97.
10. Doorsamy, W., and Cronje, W.A. (2015). A Method for Fault Detection on Synchronous Generators Using Modified Principal Component Analysis. *In Proceedings of the 2015 IEEE International Conference on Industrial Technology (ICIT)*, Seville, Spain, 17–19 March 2015; pp. 586–591.
11. Mendesil, E., Berecha, G., WeldeMichael, G., Belachew, K., and Kufa, T. ECSS (2019). (Eds.). Proceedings of the Ethiopian Coffee Science Society (ECSS): Enhancing Coffee science and Technology for Sustainable Development in

Ethiopia. *Inaugural Conference Held on 7 & 8 April 2017*, Jimma, Ethiopia, P 292.

12. Kurniawan, F., Budiastra, W. and Widyotomo, S.S. (2019). Classification of Arabica Java Coffee Beans Based on Their Origin using NIR Spectroscopy. *IOP Conf. Series: Earth and Environmental Science* 309 (2019) 012006.
13. Field, A. (2000). Discovering Statistics using SPSS for Windows. London-Thousand Oaks – New Delhi: Sage publications.
14. Geng, L., and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey, *ACM Computing Surveys (CSUR)*, **38(3)**: pp. 9.
15. Gupta, H., Agrawal, A. K., Pruthi, T., Shekhar, C., and Chellappa, R. (2002). An experimental evaluation of linear and kernel-based methods for face recognition, *Sixth IEEE Workshop on Computer Vision*, pp. 13-18.
16. Hair, J. F., Black, W.C., and Babin, B. J. (2010). *RE Anderson Multivariate data analysis: A global perspective*. New Jersey, Pearson Prentice Hall.
17. Hao, R.X., Li, S.M., Li, J.B., Zhang, Q.K., and Liu, F. (2013). Water Quality Assessment for Wastewater Reclamation Using Principal Component Analysis. *J. Environ. Inform*. 2013, **21**, 45–54.
18. Holmes S., and Huber W. (2019). *Modern Statistics for Modern Biology*. UK: Cambridge University Press.
19. Hosseini, H.M., and Kaneko, S. (2011). Dynamic sustainability assessment of countries at the macro level: A principal component analysis. Ecol. Indic. 2011, **11**, 811–823.
20. Jolliffe, I.T. (2002). *Principal Component Analysis*, 2nd Edition, Springer series in statistics 2002.
21. Jolliffe, I.T., and Cadima, J. 2016. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A* **374**: 20150202.
22. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas*., **20**:141-151.
23. Kaiser, H. F. (1974). An index of factor simplicity. Psychometrika **39**: 31–36.
24. Kloeden and Platen (1992). Numerical Solutions of Stochastic Differential Equations, pp. 11–12.
25. Kpigigbue N-A. (2019). Feature reduction and prediction for wine chemical component using principal component analysis (PCA) and linear discriminant analysis (LDA). *International IJCSMC*, Vol. 8, Issue. 12, December 2019, pg.34-45
26. Maat, S.M., Zakaria, E., Nordin, N.M., and Meerah, T.S.M. (2011). Confirmatory factor analysis of the mathematics teachers' teaching practices instrument. *World Applied Sciences Journal*, **12(11**): 2092-2096.
27. Martino, L., Luengo, D., and Míguez, J. (2012). "Efficient sampling from truncated bivariate Gaussians via Box-Muller transformation". *Electronics Letters*. 48 (24): 1533–1534. CiteSeerX 10.1.1.716.8683.
28. Bewketu M., Redi-Abshiro, M., Chandravanshi, B.S., Combrinck, S., Atlabachew M., and McCrindle, R. (2016). Profiling of phenolic compounds using UPLC–MS for determining the geographical origin of green coffee beans from Ethiopia. *J Food Compos Anal* **45:**16–25
29. Bewketu, M., Redi-Abshiro,M., Chandravanshi, B.S., Combrinck, S., McCrindle, R., and Atlabachew, M. (2019). GC-MS profiling of fatty acids in green coffee (Coffea arabica L.) beans and chemometric modeling for tracing geographical origins from Ethiopia. *J Sci Food Agric* **99**:3811–3823
30. Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., and Saikhom, R. (2017). Multivariate Statistical Data Analysis- Principal Component Analysis (PCA*). International Journal of Livestock Research*, **7(5):** 60-78.
31. Núñez, N., Collado, X., Martínez, C., Saurina, J. and Oscar, N. (2020). Authentication of the Origin, Variety and Roasting Degree of Coffee Samples by Non-Targeted HPLC-UV Fingerprinting and Chemometrics. Application to the Detection and Quantitation of Adulterated Coffee Samples. *Foods* 2020, 9, 378.
32. Nguyen L.H., and Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS Comput Biol* **15(6):** e1006907.
33. Pallant, J. (2013). SPSS Survival Manual. A step by step guide to data analysis using SPSS, 4th ed. Allen & Unwin.
34. Rajesh, S.; Jain, S.; Sharma, P. (2018). Inherent vulnerability assessment of rural households based on socio-economic indicators using categorical principal component analysis, *Uttarakhand. Ecol*. Indic. 2018, 85, 93–104.
35. Rathod, R. R., and Momin, B. F. (2012). Performance evaluation of Outlier Detection with normalized data set, International *Journal of Computer Science and Application*, ISSN: 0974-0767.
36. Shang, L., and Wang, S. (2014). Application of improved principal component analysis in comprehensive assessment on thermal power generation units. *Power Syst*. Technol. 2014, **38**, 1928–1933.
37. Shirali, G.A., Shekari, M., and Angali, K.A. (2016). Quantitative assessment of resilience safety culture using principal components analysis and numerical taxonomy: A case study in a

petrochemical plant. *J. Loss Prev. Process Ind*. **40**: 277–284.

38. Tabachnick, B.G. and Fidell, L.S. (2007). *Using Multivariate Statistics. 5th ed*. Boston, MA: Pearson Education. Inc.

39. Tharwat, A., Hassanien, E., and Elnaghi, A. (2016). Ba-based algorithm for parameter optimization of support vector machine, *Pattern Recognition Letters*.

40. Tsegay, G., Redi‐Abshiro, M., Chandravanshi, B.S., Ele, E., Mohammed, A.M., and Mamo, H. (2020). Effect of altitude of coffee plants on the composition of fatty acids of green coffee beans. *BMC Chemistry* (2020) **14**:36.