ANALYSIS OF EDIBLE OILS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND CHEMOMETRICS

By

MATTHEW TEMITOPE BAMIDELE

Bachelor of Science in Pure Chemistry

University of Ibadan

Ibadan, Oyo State, Nigeria

2008

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2019

ANALYSIS OF EDIBLE OILS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND CHEMOMETRICS

Thesis Approved:

Dr. Barry K. Lavine

Thesis Adviser

Dr. Ziad El Rassi

Dr. Nicholas Materer

Name: MATTHEW TEMITOPE BAMIDELE

Date of Degree: MAY, 2019

Title of Study: ANALYSIS OF EDIBLE OILS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND CHEMOMETRICS

Major Field: CHEMISTRY

Abstract:

The application of infrared spectroscopy and pattern recognition to the problem of discriminating edible oils by type is the subject of this research project. Previously published studies on this topic were generally limited to a small number of samples spanning five or six varieties of edible oils obtained from a single manufacturer over a limited production year range. In this study, infrared spectra obtained from 95 samples spanning 20 edible oils collected from supermarkets in the greater Newark, DE metropolitan area over a three year period were investigated using the three major types of pattern recognition methodology: mapping and display, clustering, and discriminant development. The edible oils could be partitioned into four distinct groups based on their degree of saturation and the ratio of polyunsaturated fatty acids to monounsaturated fatty acids. Edible oils assigned to one group could be readily differentiated from those assigned to other groups, whereas infrared spectra within the same group more closely resembled each other and therefore were difficult to accurately classify by type. These analyses highlight the need for adulteration studies that span multiple sources for each type of edible oil. The supplier to supplier variation for edible oils (and possibly the seasonal variation within a supplier) is greater than within supplier variation. This work also demonstrates that previous studies (which rely on only one source for each type of edible oil) provide an overly optimistic estimate of the ability to classify edible oils or to detect low levels of adulterants by infrared or Raman spectroscopy.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Edible oils are defined as a food substance other than a dairy product that is manufactured for human consumption from a fat or oil other than milk (Phillips, 2014). Edible oils are mainly composed of triglycerides (~ 95%), which are esters formed between glycerol and saturated, mono-unsaturated or poly-unsaturated fatty acids. The actual carboxylic acids depend upon the botanical origin of the edible oil (Gouilleux, Marchand, Charrier, Remaud, & Giraudeau, 2018). If free fatty acids are present in edible oils, this is due to cell damages in the vegetable tissue that occur during harvesting, storage, transport or initial processing. In addition, humidity and high temperature during storage of the edible oil can elevate the level of free fatty acids in these oils (Hammond, 2003). Triglycerides with poly-unsaturated fatty acids are prone to degradation by oxidative conditions, generating many degradation products including aldehydes, ketones and epoxides (Vaskova & Buckova, 2015).

Edible oils are an important component of the human diet as they add flavor and color to the foods that we consume. Edible oils also contain hydrophilic antioxidants (e.g., tocopherol, phenolic compounds, phytosterol), which aid in the metabolism of liposoluble vitamins (Gouilleux et al., 2018). The two essential fatty acids, omega-3 and omega-6, which cannot be synthesized by the human body are supplied by edible oils in the diet (Ng et al., 2018). In addition to the two

essential fatty acid, edible oils provide energy (Hou, Wang, Su, Wang, & Nie, 2019). Edible oils sold commercially include virgin olive oil, canola oil, corn oil, sesame oil, etc. In addition to being an important source of micronutrients and fat, cold-pressed edible oils which are neither refined nor processed have garnered attention as they have been shown to have promise in the prevention of premature aging, cancer, skin disorder, atherosclerosis and neurodegenerative diseases (Casoni, Simion, & Sârbu, 2019).

Edible oils are also used in oleochemical and pharmaceutical industries (Casoni et al., 2019). In the oleochemical industries, edible oils serve as raw materials for the production of compounds and biofuels. With the increasing cost of petroleum derived products, edible oils have found their way into these industries for the production of plastics, pharmaceuticals, inks, adhesives, and coatings. The advantage of using plant oils is that waste with lower or no toxicity is generated (Salimon, Salih, & Yousif, 2012). Greater demand in the use of edible oils for production of bioenergy and biofuel could create strong competition for edible oils being used as foods versus oleochemicals (Atadashi, Aroua, & Aziz, 2010).

Edible oils are important economically. In the Mediterranean region, extra virgin olive oil is considered to be one of the most valuable and profitable products. For this reason, the protected designation of origin (PDO) label has been used to designate extra virgin olive oils that originate in this region against potential counterfeit oils (Lastra-Mejías et al., 2019). Olive oil, an expensive edible oil that has health benefit, tops the list of the ten most adulterated foods ("Food Frauds: 10 Most Adulterated Foods", Monica Dybuncho, CBSN, January 22, 2013). As food supply chains are becoming increasingly complex with global challenges and risks emerging daily, the problems associated with adulteration of edible oils conducted for purposes of economic gain by food producers, processors, retailers, distributors and manufacturers is gaining increasing attention from governments and regulatory organizations (Jeffrey C. Moore, Spink, & Lipp, 2012). In a report commissioned by the Department of Homeland Security and funded by the National Center for

Food Protection and Defense, food fraud was defined as " a collective term that encompasses the deliberate substitution, addition, tampering, or misrepresentation of food, food ingredients, or food packaging, or false or misleading statements made about a product for economic gain" (Spink & Moyer, 2011). Similarly, the United State Pharmacopoeia (USP) Expert Panel on Food Ingredient Intentional Adulterants, has defined the intentional or economically motivated adulteration of food ingredients as "the fraudulent addition of nonauthentic substances or removal or replacement of authentic substances without the purchaser's knowledge for economic gain of the seller"(Jeffrey C. Moore et al., 2012).

Adulteration of food products is a well-known global issue with its associated economic and health burdens. It is unfortunate that (in this century) the techniques use by individuals or organizations to perpetrate this act is becoming more routine as improvements in food quality progress to the extent that some activities of these perpetrators are undetected (Lastra-Mejías et al., 2019). Individuals or organizations involved in food adulteration design fraudulent ingredients in a manner to evade existing quality assurance and quality control systems implemented by manufacturers.

The problem of food adulteration is also complicated by the fact that current food protection systems are not developed to detect an infinite number of potential adulterants in the food supply and the number of potential adulterants in the food supply is increasing each year (Spink, 2011; Spink & Moyer, 2011). It is unfortunate that consumers of edible oils are deprived of the benefits which they hope to derive t from the products which they purchase and in addition may experience health related problems as a result of the adulterants used. A practical example is the case of the Spanish toxic oil syndrome which was caused by counterfeit olive oil. The counterfeit olive oil was actually denaturalized rapeseed oil that contained aniline. The product was intended for industrial purposes, but it was labeled as olive oil for consumers. More than 20,000 consumers that purchased the olive oil for its sensory, nutritional and health benefits suffered

serious illness and even death as a result of this single act (Torrecilla, 2010). The situation has made it clear to consumers of edible oils that they are vulnerable to attack by these illicit activities.

A variety of method based on chromatographic and spectroscopic techniques are available for the authentication of edible oils. Infrared spectroscopy and high performance liquid chromatography are the most common analytical detection procedures for authentication of food products (Jeffrey C. Moore et al., 2012). Vibrational spectroscopic techniques differ from chromatographic techniques in their approach to authenticate edible oils. In contrast to chromatography which tends to isolate the components of the oil prior to analysis, vibrational spectroscopic methods consider the resultant sample spectrum as an unequivocal "holistic" fingerprint of a given oil sample. The analysis of the these fingerprint spectra by pattern recognition methods is required for detection of adulteration and successful authentication of edible oils (López-Díez, Bianchi, & Goodacre, 2003). Vibrational spectroscopy combined with multivariate analysis has the potential to discriminate between chemically similar edible oils (Baeten, Meurens, Morales, & Aparicio, 1996; Downey, McIntyre, & Davies, 2002). Detection of edible oil adulteration can be performed by direct analysis of specific component using infrared spectroscopy. However, there are indications that some adulterants cannot be detected if minor components are removed under extreme conditions of refining (Aparicio & Aparicio-Ruíz, 2000). Therefore, the pattern recognition approach has been employed for authentication of edible oils by vibrational spectroscopy.

Spectroscopic methods possess attributes that are desirable for analytical techniques used in food quality assessment. The attributes are speed, ease-of-use, minimal or no sample preparation and avoidance of sample destruction (Karoui, Downey, & Blecker, 2010). Infrared spectroscopy has received more attention because of the introduction of Fourier-transform infrared spectroscopy (FTIR). FTIR has three competitive advantages over dispersive infrared spectrometer that were used prior to 1982. The combination of Connes, Jacquinot's and Fellgett's advantages ensure that

FTIR has high wavelength precision, high optical throughput and high signal-to-noise compared to dispersive instrument. Furthermore, the development of infrared sampling technique such as attenuated total reflection (ATR) has probably the most useful development in FTIR during the past fifteen years. ATR is both simple and fast to use. To collect an infrared spectrum of an edible oil sample in ATR mode, the liquid is placed in contact with the surface of the crystal which is of high refractive index e.g., diamond.

For vibrational spectroscopy, the success of the analysis for detecting adulterants in the edible oil will depend on the data analysis methods used. Information in an infrared spectrum is captured by the positions, shapes and intensities of the bands. The band positions contain information about the chemical structure whereas information about the concentration of the compounds giving rise to specific bands are contained in the band intensities which are described by the Beer-Lambert law. The fact that edible oils contain hundreds of compounds explains the complex nature of their spectra which have overlapping bands. Hence, there is a need for the use of multivariate analysis to extract qualitative and quantitative information from the spectra (Karoui et al., 2010). Principal component analysis is a powerful data reduction tool that has been used to qualitatively and quantitatively analyze IR spectra to detect the presence of adulterants. Qualitatively, principal component analysis and related methods have been used to group known samples and to classify unknown samples with similar spectroscopic properties and to quantitatively determine the degree of adulteration of the samples (J. C. Moore, Lipp, & Griffiths, 2011).

The application of FTIR spectroscopy and pattern recognition methods to the problem of discriminating edible oils by variety has been investigated. Previously published studies on this subject were limited to approximately 30 samples spanning five or six varieties of edible oils obtained from a single manufacturer over a limited production year range (Jiménez-Carvelo, Osorio, Koidis, González-Casado, & Cuadros-Rodríguez, 2017; Lerma-García, Ramis-Ramos,

Herrero-Martínez, & Simó-Alfonso, 2010; Luna, da Silva, Ferré, & Boqué, 2013; Abdul Rohman & Che Man, 2011; A. Rohman & Man, 2010; Tay, Singh, Krishnan, & Gore, 2002; Vlachos et al., 2006; Wójcicki, Khmelinskii, Sikorski, & Sikorska, 2015; Zhang et al., 2012). In this study, FTIR spectra obtained from 95 samples spanning 20 varieties of edible oils collected from supermarkets in the greater Newark, DE metropolitan area over a three year period were examined using the three major types of pattern recognition methodology: mapping and display, clustering, and discriminant development. The 20 varieties of edible oils could be partitioned into four distinct groups based on their degree of saturation and the ratio of polyunsaturated fatty acids to monounsaturated fatty acids. Edible oils assigned to one group could be readily differentiated from those assigned to other groups, whereas FTIR spectra within the same group more closely resembled each other and therefore were difficult to classify by type. These analyses highlight the need for adulteration studies of edible oils that span multiple sources within each oil variety of edible oil. The supplier to supplier variation for edible oils (and possibly the seasonal variation within a supplier) is often greater than the within supplier variation. This work also demonstrates that previous studies (which rely on only one source for each type of edible oil) provide an overly optimistic estimate of the ability to classify edible oils or to detect low levels of adulterants by infrared or Raman spectroscopy.

References

Aparicio, R., & Aparicio-Ruíz, R. (2000). Authentication of vegetable oils by chromatographic techniques. *Journal of Chromatography A, 881*(1), 93-104. doi:https://doi.org/10.1016/S0021-9673(00)00355-1

Atadashi, I. M., Aroua, M. K., & Aziz, A. A. (2010). High quality biodiesel and its diesel engine application: A review. *Renewable and Sustainable Energy Reviews, 14*(7), 1999-2008. doi:https://doi.org/10.1016/j.rser.2010.03.020

Baeten, V., Meurens, M., Morales, M. T., & Aparicio, R. (1996). Detection of Virgin Olive Oil Adulteration by Fourier Transform Raman Spectroscopy. *Journal of Agricultural and Food Chemistry, 44*(8), 2225-2230. doi:10.1021/jf9600115

Casoni, D., Simion, I. M., & Sârbu, C. (2019). A comprehensive classification of edible oils according to their radical scavenging spectral profile evaluated by advanced chemometrics. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy, 213*, 204-209. doi:10.1016/j.saa.2019.01.065

Downey, G., McIntyre, P., & Davies, A. N. (2002). Detecting and quantifying sunflower oil adulteration in extra virgin olive oils from the Eastern Mediterranean by visible and near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry, 50*(20), 5520-5525. doi:10.1021/jf0257188

Gouilleux, B., Marchand, J., Charrier, B., Remaud, G. S., & Giraudeau, P. (2018). High-throughput authentication of edible oils with benchtop Ultrafast 2D NMR. *Food Chemistry, 244*, 153-158. doi:10.1016/j.foodchem.2017.10.016

Hammond, E. W. (2003). VEGETABLE OILS | Types and Properties. In B. Caballero (Ed.), *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (pp. 5899-5904). Oxford: Academic Press.

Hou, X., Wang, G., Su, G., Wang, X., & Nie, S. (2019). Rapid identification of edible oil species using supervised support vector machine based on low-field nuclear magnetic resonance relaxation features. *Food Chemistry, 280*, 139-145. doi:10.1016/j.foodchem.2018.12.031

Jiménez-Carvelo, A. M., Osorio, M. T., Koidis, A., González-Casado, A., & Cuadros-Rodríguez, L. (2017). Chemometric classification and quantification of olive oil in blends with any edible vegetable oils using FTIR-ATR and Raman spectroscopy. *LWT - Food Science and Technology, 86*, 174-184. doi:10.1016/j.lwt.2017.07.050

Karoui, R., Downey, G., & Blecker, C. (2010). Mid-infrared spectroscopy coupled with chemometrics: A tool for the analysis of intact food systems and the exploration of their molecular structure-quality relationships-A review. *Chemical Reviews, 110*(10), 6144-6168. doi:10.1021/cr100090k

Lastra-Mejías, M., Aroca-Santos, R., Izquierdo, M., Cancilla, J. C., Mena, M. L., & Torrecilla, J. S. (2019). Chaotic parameters from fluorescence spectra to resolve

fraudulent mixtures of fresh and expired protected designation of origin extra virgin olive oils. *Talanta, 195*, 1-7. doi:10.1016/j.talanta.2018.10.102

Lerma-García, M. J., Ramis-Ramos, G., Herrero-Martínez, J. M., & Simó-Alfonso, E. F. (2010). Authentication of extra virgin olive oils by Fourier-transform infrared spectroscopy. *Food Chemistry, 118*(1), 78-83. doi:https://doi.org/10.1016/j.foodchem.2009.04.092

López-Díez, E. C., Bianchi, G., & Goodacre, R. (2003). Rapid Quantitative Assessment of the Adulteration of Virgin Olive Oils with Hazelnut Oils Using Raman Spectroscopy and Chemometrics. *Journal of Agricultural and Food Chemistry, 51*(21), 6145-6150. doi:10.1021/jf034493d

Luna, A. S., da Silva, A. P., Ferré, J., & Boqué, R. (2013). Classification of edible oils and modeling of their physico-chemical properties by chemometric methods using mid-IR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 100*, 109-114. doi:https://doi.org/10.1016/j.saa.2012.06.034

Moore, J. C., Lipp, M., & Griffiths, J. C. (2011). Preventing the adulteration of food protein with better analytical methods: Avoiding the "next melamine". *INFORM - International News on Fats, Oils and Related Materials, 22*(6), 373-375.

Moore, J. C., Spink, J., & Lipp, M. (2012). Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *77*(4), R118-R126. doi:10.1111/j.1750-3841.2012.02657.x

Ng, T. T., Li, S., Ng, C. C. A., So, P. K., Wong, T. F., Li, Z. Y., . . . Yao, Z. P. (2018). Establishment of a spectral database for classification of edible oils using matrix-assisted laser desorption/ionization mass spectrometry. *Food Chemistry, 252*, 335-342. doi:10.1016/j.foodchem.2018.01.125

Phillips, T. (2014, October 29, 2018). Commonly Used Edible Oils and Their Benefits. Retrieved from https://www.thebalance.com/what-are-edible-oils-3973288

Rohman, A., & Che Man, Y. B. (2011). The use of Fourier transform mid infrared (FT-MIR) spectroscopy for detection and quantification of adulteration in virgin coconut oil. *Food Chemistry, 129*(2), 583-588. doi:https://doi.org/10.1016/j.foodchem.2011.04.070

Rohman, A., & Man, Y. B. C. (2010). Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. *Food Research International, 43*(3), 886-892. doi:https://doi.org/10.1016/j.foodres.2009.12.006

Salimon, J., Salih, N., & Yousif, E. (2012). Industrial development and applications of plant oils and their biobased oleochemicals. *Arabian Journal of Chemistry, 5*(2), 135-145. doi:https://doi.org/10.1016/j.arabjc.2010.08.007

Spink, J. (2011). The challenge of intellectual property enforcement for agriculture technology transfers, additives, raw materials, and finished goods against product fraud and counterfeiters. *Journal of Intellectual Property Rights, 16*(2), 183-193.

Spink, J., & Moyer, D. C. (2011). Defining the Public Health Threat of Food Fraud. *Journal of Food Science, 76*(9), R157-R163. doi:10.1111/j.1750-3841.2011.02417.x

Tay, A., Singh, R. K., Krishnan, S. S., & Gore, J. P. (2002). Authentication of Olive Oil Adulterated with Vegetable Oils Using Fourier Transform Infrared Spectroscopy.

LWT - Food Science and Technology, 35(1), 99-103. doi:https://doi.org/10.1006/fstl.2001.0864

Torrecilla, J. S. (2010). The olive : its processing and waste management.

Vaskova, H., & Buckova, M. (2015). Thermal Degradation of Vegetable Oils: Spectroscopic Measurement and Analysis. *Procedia Engineering, 100*, 630-635. doi:https://doi.org/10.1016/j.proeng.2015.01.414

Vlachos, N., Skopelitis, Y., Psaroudaki, M., Konstantinidou, V., Chatzilazarou, A., & Tegou, E. (2006). Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta, 573-574*, 459-465. doi:https://doi.org/10.1016/j.aca.2006.05.034

Wójcicki, K., Khmelinskii, I., Sikorski, M., & Sikorska, E. (2015). Near and mid infrared spectroscopy and multivariate data analysis in studies of oxidation of edible oils. *Food Chemistry, 187*, 416-423. doi:https://doi.org/10.1016/j.foodchem.2015.04.046

Zhang, Q., Liu, C., Sun, Z., Hu, X., Shen, Q., & Wu, J. (2012). Authentication of edible vegetable oils adulterated with used frying oil by Fourier Transform Infrared Spectroscopy. *Food Chemistry, 132*(3), 1607-1613. doi:https://doi.org/10.1016/j.foodchem.2011.11.129

CHAPTER II

MATERIALS AND METHODS

2.1. EDIBLE OIL SPECTRAL DATABASE

Ninety-five edible oil samples (see Table 1) were purchased over a three year period from supermarkets in the Newark, DE metropolitan area. Each sample represented a separate brand of edible oil or the same brand of edible oil manufactured in a different year. All edible oil samples used in this study were stored in glass containers with plastic caps at room temperature prior to infrared analysis. Some edible oils were associated with several brand names such as extra virgin olive oil (EVOO). For EVOO, the brand names were robust extra virgin olive oil, rich robust extra virgin olive, mild bidley extra virgin olive oil and extra virgin olive gentile oil. Extra light olive oil (ELOO) consisted of samples having the labels ELOO and extra light tasting olive oil, while olive oil was comprised of samples labeled as olive oil, classic olive and pure olive oil.

Adulterated edible oil samples were prepared to simulate adulteration of more expensive edible oils, for example, extra virgin olive oil or sesame oil through substitution or blending with less expensive oils (such as corn or canola oil), by mixing the appropriate volume of each edible oil to yield the adulterant mixture, e.g., 15% adulterated mixture of extra virgin olive oil by corn oil is prepared by mixing 850 μL of extra virgin olive oil and 150 μL of corn oil in a 15 ml sterile falcon tube using a Thermolyne MaxiMixPlus vortex mixer.

Table 1. Edible Oil Database

Edible Oil Type	Oil ID	Group ID	Number of Samples	Number of Spectra *Tset/Pset
EVOO	1	1	25	73/7
ELOO	2	1	7	20/4
Olive	3	1	8	23/3
Avocado	5	1	2	6/3
Peanut	6	1	4	12/3
Corn	7	3	8	26/19
Grapeseed	8	3	8	27/9
Safflower	9	1	2	6/3
Hazelnut	10	1	2	6/3
Canola	13	2	9	27/9
Canola-Vegetable	16	2	1	3/0
Vegetable	17	3	4	11/3
Canola-Sun-Soybean	18	2	1	2/7
Sunflower	19	1	1	3/0
Sweet Almond	23	1	2	6/0
Almond	27	1	4	12/3
Extra Virgin Sesame	28	2	3	9/6
Toasted Sesame	32	2	1	3/0
Walnut	33	4	2	6/4

Avocado-Olive-Flaxseed	34	1	1	3/0
Total	---	---	95	284/86 = 370

*Tset = training set. Pset = prediction set.

FTIR spectra were collected for each edible oil sample in triplicate over the spectral range 402 cm^{-1} to 3738 cm^{-1}. Above 3738 cm^{-1}, there were no absorption bands in any of the samples, and 400 cm^{-1} is the cut-off for the DTGS detector. Spectra were collected in attenuated total reflection (ATR) mode using a Thermo Scientific Nicolet iS50 FTIR spectrometer equipped with a deuterated triglycine sulfate (DTGS) detector. Apodization was performed using the Happ-Genzel function.

To collect an infrared (IR) spectrum, a drop of an edible oil sample was deposited on the crystal of the ATR module, and 64 scans were then performed at 4 cm^{-1} resolution for each sample. Background was collected before running each sample. Isopropanol and cotton swabs were used to clean the ATR crystal after each sample collection. The diamond crystal absorbance region (~2700 cm^{-1}- 1900 cm^{-1}) was zeroed out of each IR spectrum. An infrared absorption spectrum of an edible oil is shown in Figure 2.1.

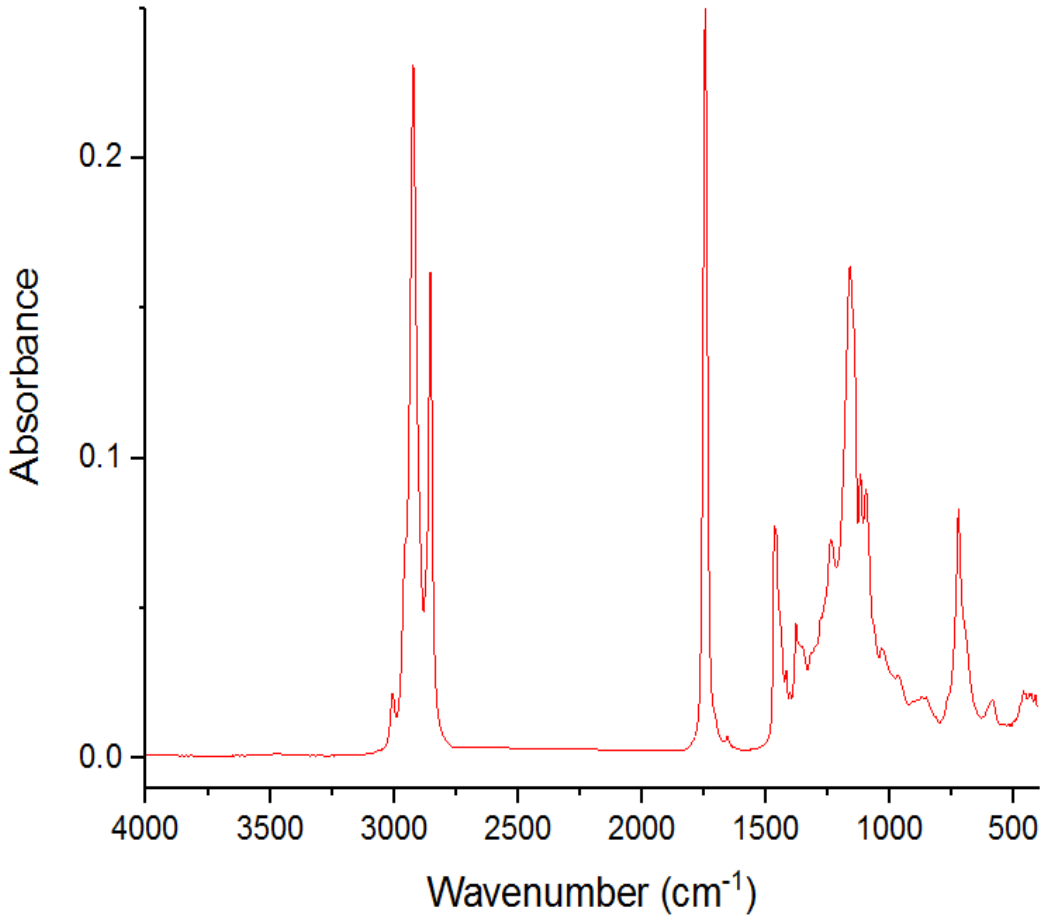

Figure 2.1. An infrared absorption spectrum of olive oil with the diamond region zeroed out.

2.2. PATTERN RECOGNITION ANALYSIS

Pattern recognition methods were originally developed to solve the class membership problem. In a typical pattern recognition study, samples are classified according to a specific property by using measurements that are indirectly related to the property of interest. An empirical relationship or classification rule is developed from a set of samples for which the property of interest and the measurements are known. The classification rule is then used to predict the property of samples that are not part of the original training set. The set of samples for which the property of interest and

measurements are known is called the training set, whereas the set of measurements that describe each sample in the data set is called a pattern. The determination of the property of interest by assigning a sample to its respective class is called recognition, hence the term pattern recognition.

For pattern recognition analysis, the IR spectrum of each sample is represented by a data vector $x = (x_1, x_2, x_3, \ldots\ldots x_j, \ldots.. x_n)$ where x_j is the absorbance at the j^{th} wavelength. Such a vector can be considered as a point in a high dimensional measurement space. The Euclidean distance between each pair of points is inversely related to their degree of similarity. Points representing spectra of samples from one class will cluster in a limited region of the measurement space separate from the other samples. Pattern recognition is a set of numerical methods for assessing the structure of the data space. The data structure is defined as the overall relation of each sample to every other sample in the data set.

In this section, the four major subdivisions of pattern recognition methodology are discussed: (1) data preprocessing, (2) mapping and display, (3) clustering, and (4) classification. The procedures that must be implemented in order to apply pattern recognition methods to a data set are also enumerated. Specific emphasis is placed on the application of these techniques to the problem of discrimination and authentication of edible oils by type.

2.2.1 DATA PREPROCESSING

The first step in any pattern recognition study is to convert the spectral data into computer compatible form. Normally, the spectral data are arranged in the form of a table or data matrix:

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & .... & x_{1N} \\ x_{21} & x_{22} & x_{23} & .... & x_{2N} \\ . & . & . & & . \\ . & . & . & & . \\ x_{M1} & x_{M2} & x_{M3} & & x_{MN} \end{bmatrix} \tag{2.1}$$

The rows of the matrix represent the spectra, and each column is the absorbance of the sample at a specific wavelength. It is essential that each column encodes the same information for all samples in the data set. If variable 5 is the absorbance at 3000 cm^{-1} for the FTIR spectrum of sample 1, it must also be the absorbance at 3000 cm^{-1} for the IR spectrum of samples 2, 3,…, M. Hence, alignment is crucial when spectra are translated into data vectors.

The next step is scaling. The objective is to enhance the signal to noise ratio of the data. In the application discussed in this thesis, two scaling techniques have been applied to the data: normalization and autoscaling. The procedures that should be used for a given data set, however, are highly dependent upon the nature of the problem under investigation.

Normalization involves setting the sum of the components of each data vector equal to some arbitrary constant. For mid-IR data, the data vectors are normalized to unit length. This is accomplished by dividing each vector by the square root of the sum of the squares of the components comprising the vector. Normalization compensates for variation in the spectra due to differences in the sample size or optical path length. However, normalization can introduce dependence between variables because of closure that can have an effect on the results of the investigation. Thus, one must take into account both of these factors when deciding whether or not to normalize data.

Autoscaling involves standardizing the variables such that each measurement has a mean of zero and a standard deviation of one, that is,

$$x_{i,new} = \frac{(x_{i,orig} - \bar{x}_{i,orig})}{s_{i,orig}} \tag{2.2}$$

where $\bar{x}_{i,orig}$ is the mean and $s_{i,orig}$ is the standard deviation of the original measurement variable. If autoscaling is not applied, the wavelengths with larger absorbance values will tend to dominate the analysis because their variance is generally larger. Autoscaling removes inadvertent weighing of the variables that otherwise would occur. Thus, each wavelength will have an equal weight in the analysis. Autoscaling affects the spread of the data by placing the data points inside a hypercube. However, it does not affect the relative distribution of the data points in the high dimensional measurement space.

2.2.2 MAPPING AND DISPLAY

Chemists often use graphical methods to study data. If there are only two or three measurements per sample, the data can be displayed as a graph or plot for direct viewing. By examining the plot, a chemist can search for similarities and dissimilarities among samples, find natural clusters, and even obtain information about the overall structure of the data set. If there are n measurements per sample (n>3), a two or three-dimensional representation of the measurement space is needed to visualize the relative position of the sample points in n-space. This representation must accurately reflect the structure of the data. A popular approach to this problem is using a mapping and display technique called principal component analysis (Du, 2019; Liu, Maljovec, Wang, Bremer, & Pascucci, 2017; Pořízka et al., 2018).

Principal component analysis is the most widely used multivariate analysis technique in science and engineering. It is a method for transforming the original measurement variables into new, uncorrelated variables called principal components. Each

principal component is a linear combination of the original measurement variables. Using this procedure, a set of orthogonal axes that represent the direction of greatest variance in the data is found. (Variance is defined as the degree to which the data are spread in the n-dimensional measurement space.) Usually, only two or three principal components are necessary to explain all of the information present in multivariate data. Hence, principal component analysis can be applied to multivariate data for dimensionality reduction, in order to identify outliers, display data structure, and classify samples.

Dimensionality reduction or data compression is possible with principal component analysis because of correlations between measurement variables. Consider Figure 2.2, which shows a plot of 15 samples in a two-dimensional space. The coordinate axes of this measurement space are defined by the variables x_1 and x_2. These two variables are correlated since fixing the value of x_1 limits the range of values possible for x_2. If x_1 and x_2 were uncorrelated, the enclosed rectangle shown in Figure 2.2 would be completely filled by data points. Because information can be defined as the scatter of points in a measurement space, correlations between measurement variables decrease the scatter or information content of the space. The data points are restricted to a small region of the measurement space due to correlations between the measurement variables. If the measurement variables are highly correlated, the data points may even reside in a subspace. This is shown in Figure 2.3. Each row of the data matrix is an object and each column is a measurement variable. Here x_3 is perfectly correlated with x_1 and x_2 since x_3 (third column) equals x_1 (first column) plus x_2 (second column). Hence, the seven data points lie in a plane (or 2-dimensional subspace) even though each point has three measurements associated with it because x_3 is a redundant variable.

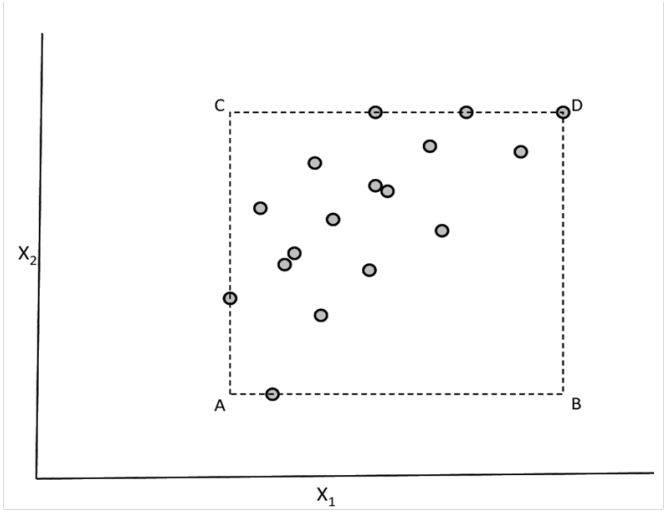

Figure 2.2. Seventeen hypothetical samples projected onto a 2-dimensional measurement space defined by the measurement variables X_1 and X_2. The vertices, A, B, C, and D, of the rectangle represent the smallest and largest values of X_1 and X_2. (Adapted from *NBS J. Res*., 1985, 190(6), 465-476)

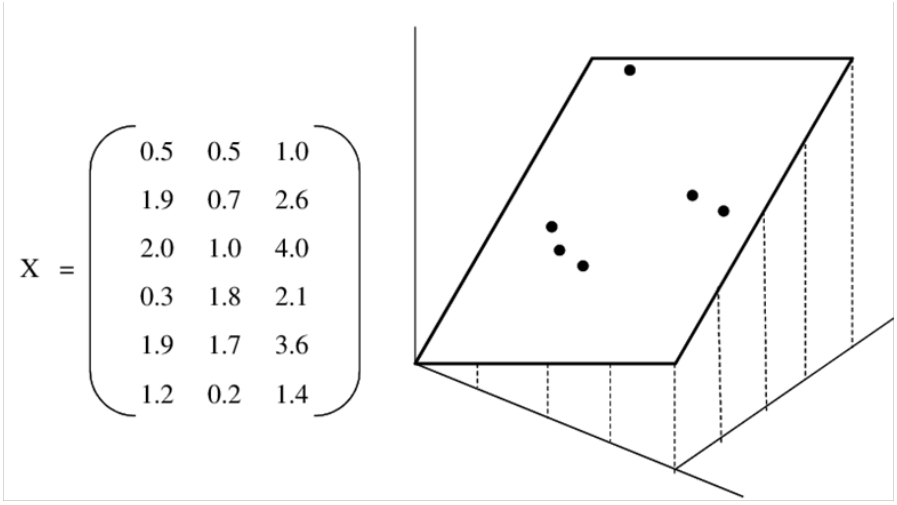

Figure 2.3. Six hypothetical samples projected onto a 3-dimensional measurement space. Because of strong correlations among the 3 measurement variables, the data points reside in a 2-dimensional subspace of the original measurement space. (Adapted from *Multivariate Pattern Recognition in Chemometrics*, Elsevier Science Publishers, Amsterdam, 1992)

Variables that contain redundant information are said to be collinear. High collinearity between variables is a strong indication that a new coordinate system can be found that will be better at conveying the information present in the data than one defined

by the original measurement variables. The new coordinate system for displaying the data is based on variance. (The scatter of the data points in the measurement space is a direct measure of the data's variance.) The principal components of the data define the variance-based axes of this new coordinate system. The first principal component is formed by determining the direction of largest variation in the original measurement space of the data and modeling it with a line fitted by linear least squares (see Figure 2.4) that passes through the center of the data. The second largest principal component lies in the direction of next largest variation. It passes through the center of the data and is orthogonal to the first principal component. The third largest principal component lies in the direction of next largest variation. It also passes through the center of the data; it is orthogonal to the first and second principal component, and so forth. Each principal component describes a different source of information because each defines a different direction of scatter or variance in the data. The orthogonality constraint imposed by the mathematics of principal component analysis also ensures that each variance-based axis will be independent.

A measure of the amount of information conveyed by each principal component is the variance of the data explained by it, which is expressed in terms of its eigenvalue. For this reason, principal components are arranged in order of decreasing eigenvalues. The most informative or largest principal component is the first and the least informative or smallest is the last. The amount of information contained in a principal component relative to the original measurement variables, i.e., the fraction of the total cumulative variance explained by the principal component, is equal to the eigenvalue of the principal component divided by the sum of all the eigenvalues. The maximum number of principal components that can be extracted from the data is the smaller of either the number of

samples or the number of variables in the data set, as this number defines the largest possible number of independent axes in the data.

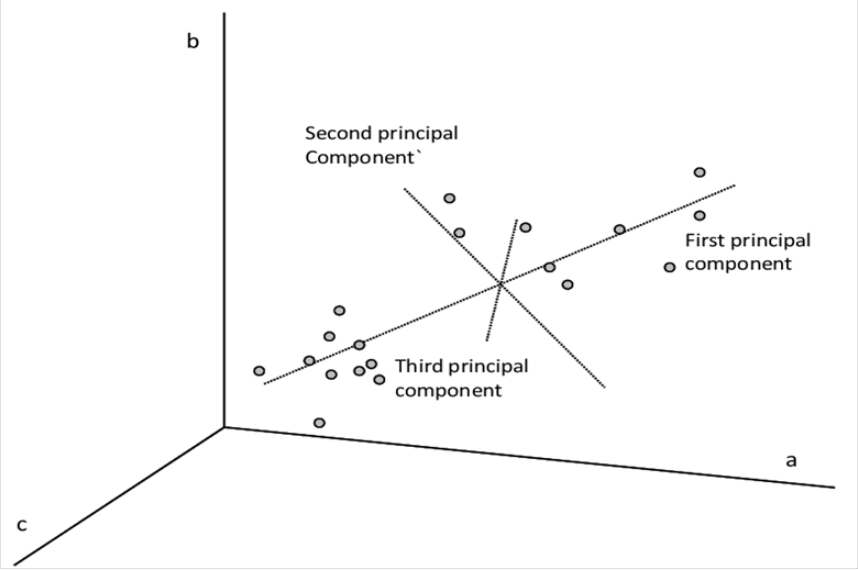

Figure 2.4. Principal component axes developed from the measurement variables a, b, and c. (Courtesy of Applied Spectroscopy, 1995, 49(12), 14A-30A)

If the data are collected with due care, one would expect that only the larger principal components would convey information about clustering since most of the information in the data should be about the effect of interest. However, the situation is not always as straightforward as implied. Each principal component describes some amount of signal and some amount of noise in the data because of accidental correlation between signal and noise. The larger principal components contain information primarily about signal, whereas the smaller principal components describe primarily noise. By discarding the smaller principal components, noise is discarded but so is a small amount of signal. However, the gain in signal to noise more than compensates for the biased representation of the data that occurs when plotting only its largest principal components.

Principal component analysis takes advantage of the fact that a large amount of data generated in scientific studies has a great deal of redundancy and therefore a great deal of collinearity. Because the measurement variables are correlated, 800-point spectra do not require 800 independent orthogonal axes to define the position of a sample point in the measurement space. Using principal component analysis, the original measurement variables that constitute a correlated axes system can be converted into an orthogonal axes system, which dramatically reduces the dimensionality of the data since only a few independent axes are needed to describe the data. Spectra of a set of samples often lie in a subspace of the original measurement space and a plot of the two or three largest principal components of the data can help one to visualize the relative position of the spectra in this subspace.

2.2.3 CLUSTERING

Exploratory data analysis techniques are often quite helpful in elucidating the complex nature of multivariate relationships. In the preceding section, the importance of using mapping and display techniques for understanding the structure of complex multivariate data sets was emphasized. In this section, some additional techniques will be discussed that also give insight into the structure of a data set. These methods attempt to find sample groupings or clusters within data using criteria developed from the data itself: hence the term cluster-analysis.

Cluster analysis is based on the principle that distances between pairs of points (i.e., samples) in the measurement space are inversely related to their degree of similarity. Although several different types of clustering algorithms exist, by far the most popular is hierarchical clustering [46-48], which is the focus here. The starting point for a hierarchical

clustering experiment is the similarity matrix, which is formed by first computing the distances between all pairs of points in the data set. Each distance is then converted into a similarity value

$$s_{ik} = 1 - \frac{d_{ik}}{d_{\max}} \tag{2.3}$$

where s_{ik} is the measure of similarity between samples i and k, d_{ik} is the Euclidean distance between samples i and k, and d_{max} is the distance between the two most dissimilar samples, which is also the largest distance in the data set. These similarity values, which vary from 0 to 1, are organized in the form of a table or square symmetric matrix. The similarity matrix is scanned for the largest value, which corresponds to the most similar point pair, and the two samples (comprising the point pair) are combined to form a new point located midway between the two original data points. After the rows and columns corresponding to the original two data points are removed, the symmetry matrix is then updated to include information about the similarity between the new point and every other point in the data set. The matrix is again scanned, the new nearest point pair is identified and combined to form a single point, the rows and columns of the two data points that were combined are removed, and the matrix is recomputed. This process is repeated until all points have been linked. The result of this procedure is a diagram called a dendogram, which is a visual representation of the relationships between samples in the data set (see Figure 4). Interpretation of the results is intuitive, which is the major reason for the popularity of these methods. For example, the dendogram shown in Figure 2.4 suggests that the data set can be divided into two groups (cluster 1 = samples 18, 15, 25, 12, 14, 13, 16, 17, 21, 22, 23, 20, 24; cluster 2 = samples 9, 11, 6, 10, 3, 19, 4, 1, 2, 5) or four groups (cluster 1 = samples 18, 15, 25, 12, 14, 13, 16, 17; cluster 2 = samples 21, 22, 23, 20, 24; cluster 3 = samples 9,

11, 6, 10; cluster 4 = samples 3, 19, 4, 1, 2, 5). Samples 7 and 8 are considered outliers by the dendogram.

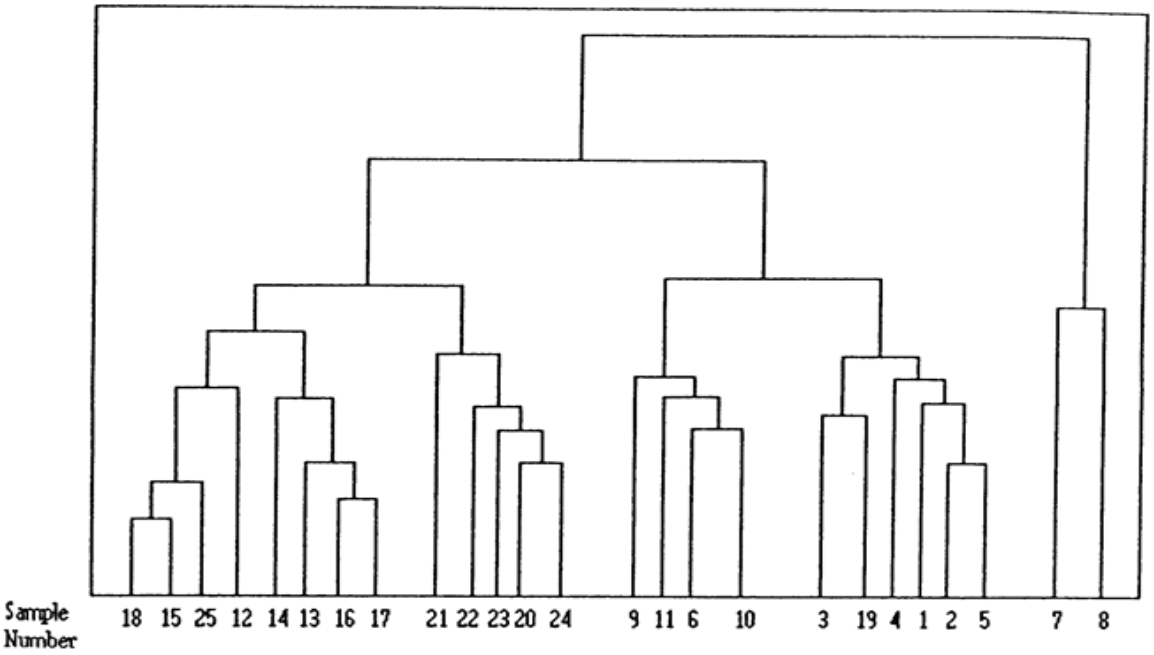

Figure 2.5. Dendogram of a data set that can be divided into two or four clusters. Samples 7 and 8 are outliers.

All clustering procedures yield the same results for data sets with well-separated clusters. However, the results differ when the clusters overlap because of space distorting effects. Single linkage favors the formation of large linear clusters instead of the usual elliptical or spherical clusters. As a result, poorly separated clusters are often chained together. Complete linkage, on the other hand, favors the formation of small spherical clusters. That is why it is a good idea to use at least two different clustering algorithms when studying a data set. If the dendograms are in agreement, then a strong case can be made for partitioning the data into distinct groups. If the cluster memberships differ, the data should be further investigated using principal component analysis or other mapping

and display techniques. As a general rule, it is recommended that hierarchical methods be used in tandem with principal component or self- organizing maps to detect clusters in multivariate data. Hierarchical methods are exploratory tools: the absolute validity of a dendogram is less important than insights and suggestions gained by the user about the data structure.

2.2.4 CLASSIFICATION

So far, only exploratory data analysis techniques, e.g., cluster and principal component analysis, have been discussed. These techniques attempt to analyze multivariate data without directly using the information about the class assignment of the samples to develop projections of the data. Although mapping and display techniques and cluster analysis are powerful methods for uncovering relationships in large multivariate data sets, they are often not sufficient for developing a classification rule. However, the overall goal of pattern recognition analysis is the development of a classification rule that can accurately predict the class membership of an unknown sample. In this section, classification methods will be discussed. The focus will on a genetic algorithm for pattern recognition and feature selection.

A block diagram of the pattern recognition GA is shown in Figure 2.6. The GA builds a population of binary strings, each of which represents a possible solution, i.e., a set of wavelengths. For a wavelength to be included in the subset, it is necessary for the corresponding bit in the string to be set at 1. If the bit is set to 0, the wavelength will not be included.

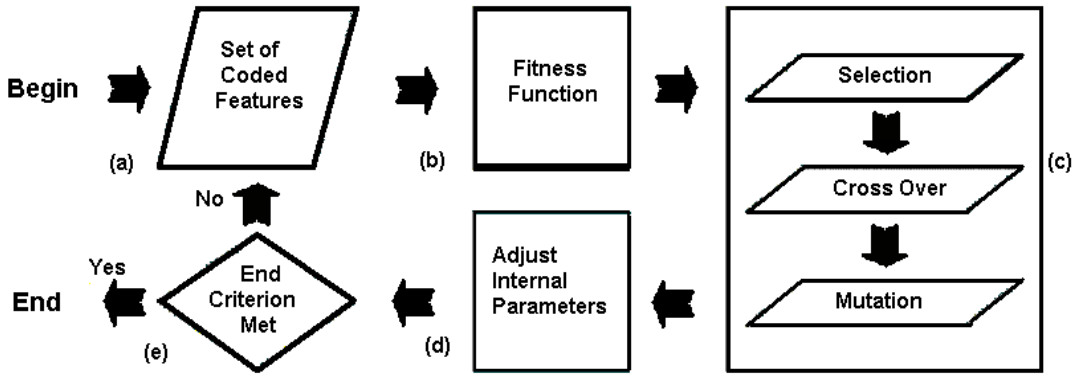

Figure 2.6. Block diagram of the pattern recognition GA.

During each generation, the strings are decoded yielding the actual parameter set, which is sent to a fitness function for evaluation. The string is assigned a value by the fitness function, which is a measure of the quality of the proposed solution. Solutions with high fitness values have a high probability of being selected for crossover. The power of the GA arises from crossover, which causes a structured yet randomized exchange of information between solutions, with the possibility that good solutions can generate even better ones. The new population of strings, which often yields better solutions to the problem, is evaluated using the fitness function. This entire process (evaluation, selection, crossover, mutation, and adjustment of internal parameters) is repeated until a specified number of generations or a feasible solution has been found.

The fitness function of the GA emulates human pattern recognition through machine learning to score the principal component plots and thereby identify a set of features that optimize the separation of the woods by sample type in a plot of the two largest principal components of the data. To facilitate the tracking and scoring of the principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see Equations 2.4 and 2.5, where CW(c) is the weight of class c

and c varies from 1 to the total number of classes in the data set, and $SW_c(s)$ is the weight of sample s in class c). The class weights sum to 100, whereas the sample weights in a class sum to a value equal to the corresponding class weight.

$$\mathrm{CW(c)} = 100\frac{\mathrm{CW(c)}}{\sum_{\mathrm{c}} \mathrm{CW(c)}} \tag{2.4}$$

$$\mathrm{SW}(s) = \mathrm{CW(c)}\frac{\mathrm{SW}(s)}{\sum_{\mathrm{s}\in\mathrm{c}} \mathrm{SW}\ (s)} \tag{2.5}$$

Each principal component plot generated for each feature subset is scored using the K-nearest neighbor (K-NN) classification algorithm (Pei, 1998). For a given data point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to largest. A poll is then taken of the point's K-nearest neighbors. For the most rigorous classification, K equals the number of samples in the class to which the point belongs. (K, which is assigned by the user, usually varies with the class.) The number of nearest neighbors with the same class label as the sample point in question, the so-called sample hit count (SHC), is computed ($0 \leq \mathrm{SHC(s)} \leq \mathrm{K_c}$), where $\mathrm{K_c}$ is the number of nearest neighbors to be calculated for each sample in class c. It then becomes a simple matter to score a principal component plot (see Equation 2.6 where F(d) is the fitness of the feature set being scored, SHC(s) is the number of nearest samples with the same class label as sample s, and SW(s) is the weight of sample s.)

$$F(d) = \sum_{\mathrm{c}} \sum_{\mathrm{s}\in\mathrm{c}} \frac{1}{\mathrm{K_c}} \times \mathrm{SHC(s)} \times \mathrm{SW(s)} \tag{2.6}$$

To better understand the scoring of the principal component plots, consider the hypothetical example of a data set with two classes having been initially assigned equal weights by the pattern recognition GA. Class 1 has 10 samples, and class 2 has 20 samples. Therefore, K_1 is 10 and K_2 is 20. At generation 0 (no children have yet been created), the samples in a given class will have the same weight. Therefore, each sample in class 1 has a sample weight of 5, whereas each sample in class 2 has a weight of 2.5. If sample 3, which is in class 1, has as its nearest neighbors 7 class one samples, then $SHC/K_1 = 0.7$ and $(SHC/K_1)*SW(3) = 0.7*5$ or 3.5. By summing $(SHC/K_c)*SW(s)$ for each sample point in the plot, the principal component map of the feature subset is scored.

The fitness function of the GA is able to focus on samples and classes that are difficult to classify by boosting their weights over successive generations. In order to boost, it is necessary to first compute the sample hit rate (SHR), which is the mean value of SHC/K_c over all feature subsets (ϕ) produced in a particular generation (see Equation 2.7). SHR, which is calculated over the entire population of solutions in a particular generation, provides consistent information about the difficulty in classifying a particular sample.

$$\mathrm{SHR(s)} = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC}_i(s)}{\mathrm{K}} \qquad (2.7)$$

Boosting is then performed in two steps. First, the class-hit rate (i.e., average sample hit rate for all samples in a class) is computed (see Equation 2.8 where $CHR_g(c)$ is the class hit rate for class c during generation g, AVG is the average, and $SHR_g(s)$ is the sample hit rate for each sample in class c during generation g). CHR like SHR provides consistent information about the difficulty in classifying a particular sample type. Classes and samples with low hit rates will be weighted more heavily, that is, they will have more influence in the fitness calculation, than classes or samples that score well.

Second, class and sample weights are adjusted during each generation using a perceptron (see Equations 2.9 and 2.10 where $CW_{g+1}(c)$ is the class weight for c during the current generation g+1, $CW_g(c)$ is the class weight for c during the previous generation g, P is the momentum, $CHR_g(c)$ is the class hit rate for c during generation g, $SW(s)_{g+1}$ is the sample weight for s during generation g+1, $SW(s)_g$ is the sample weight for s during generation g, and $SHR_g(s)$ is the sample hit rate for s during generation g.) The user must set the momentum, P.

During each generation, class and sample weights are updated using the class and sample hit-rates from the previous generation. After a certain number of generations, the class weights become fixed. Equation 2.9 is turned off, P is halved, and sample weights are renormalized using Equation 2. The GA then focuses on the troublesome samples; see Equation 2.10.

$$\mathrm{CHR}_g(c) = AVG(\mathrm{SHR}_g(s) : \forall_{s \in c}) \tag{2.8}$$

$$\mathrm{CW}_{g+1}(c) = \mathrm{CW}_g(c) + P(1 - \mathrm{CHR}_g(c)) \tag{2.9}$$

$$\mathrm{SW}_{g+1}(s) = \mathrm{SW}_g(s) + P(1 - \mathrm{SHR}_g(s)) \tag{2.10}$$

Boosting is crucial for the successful operation of the GA because it allows the values of the class and sample weights to change, thereby modifying the criteria for a good score. This can helps to minimize the problem of convergence to a local optimum. Hence, the fitness landscape of the GA changes as the population evolves towards a solution.

During each generation, the selection, crossover, and mutation operators are applied to the chromosomes to develop new and potentially better solutions. The selection operator used by the pattern recognition GA is implemented by ordering the population of

strings, i.e., the wavelength subsets, from best to worst fitness while simultaneously generating a copy of the same population and randomizing the order of the strings in this copy with respect to fitness. A fraction of the population is then selected as per the selection pressure, which is usually set at 0.5. The top half of the ordered population is mated with the top half of the random population, guaranteeing that the best 50% are selected for reproduction, while ensuring that every string in the randomized copy has a uniform chance of being selected due to the randomized selection criterion imposed on the strings in this population.

For each pair of strings selected for crossover, two new strings are generated using a variation of three-point crossover. As in the case of simple three-point crossover, the length of each new string or solution is the same as the dimensionality of the data. However, the crossover operator used by the pattern recognition GA is not compelled to preserve order among exchanged string fragments. This safeguards the loss of information or features in the population. As a result, it becomes less likely for the population variability to fall below a critical value due to the additional degree of freedom provided by the reordering. This variation of three-point crossover may also be useful in searching for good string arrangements. If the current population has bad ordering, where features with a high synergism are spaced apart at great distances, simple crossover would probably destroy these important allele packets. On the other hand, there is a chance to obtain good allele ordering, by using a crossover operator with a reordering algorithm embedded in it.

A mutation operator is applied to the new strings with the mutation probability usually set at 1%. This means that 1% of the new chromosomes (i.e., feature subsets) are selected at random for mutation. A chromosome marked for mutation has a single bit

flipped, which is selected at random. This allows the GA to explore other regions of the solution space. If the GA finds a better feature subset through mutation, the optimization will continue in a new direction.

The resulting population of strings, both parents and children, are sorted by their fitness (see Equation 2.6), and the top ϕ strings are retained for the next generation. The new population can be expected to perform better on average than its predecessor because the selection criterion used for reproduction exhibits bias for the higher-ranking strings. However, the reproduction operators also assure a significant degree of diversity in the population, since the crossover points and reordering of exchanged string fragments of each chromosome pair is selected at random.

References

Du, T. Y. (2019). Dimensionality Reduction Techniques for Visualizing Morphometric Data: Comparing Principal Component Analysis to Nonlinear Methods. *Evolutionary Biology, 46*(1), 106-121. doi:10.1007/s11692-018-9464-9

Liu, S., Maljovec, D., Wang, B., Bremer, P. T., & Pascucci, V. (2017). Visualizing High-Dimensional Data: Advances in the Past Decade. *IEEE Transactions on Visualization and Computer Graphics, 23*(3), 1249-1268. doi:10.1109/TVCG.2016.2640960

Pořízka, P., Klus, J., Képeš, E., Prochazka, D., Hahn, D. W., & Kaiser, J. (2018). On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review. *Spectrochimica Acta - Part B Atomic Spectroscopy, 148*, 65-82. doi:10.1016/j.sab.2018.05.030

Pei, M. (1998). Feature Extraction Using Genetic Algorithms. Proc. of International Symposium on Intelligent Data Engineering and Learning. https://ci.nii.ac.jp/naid/10010553087/en/

CHAPTER III

RESULTS AND DISCUSSION

Of the 370 IR spectra collected from 95 edible oil samples, 289 were assigned to the training set and 81 to the prediction set by random lot. Because of the large number of classes, a hierarchical classification scheme was employed to discriminate the FTIR spectra of the 20 varieties of edible oils in the training set. To implement this scheme, the average infrared spectrum of each variety of edible oil was computed. The 20 average spectra were then analyzed using both principal component analysis and hierarchical cluster analysis. Both the principal component plot and the dendogram of the average infrared spectra (see Figure 3.1) are in agreement and each indicates that dividing the 20 edible oils into four distinct groups is appropriate (see Table 3). For this reason, we chose to partition the edible oils into four groups (see Table 2.1 and Figure 3.2).

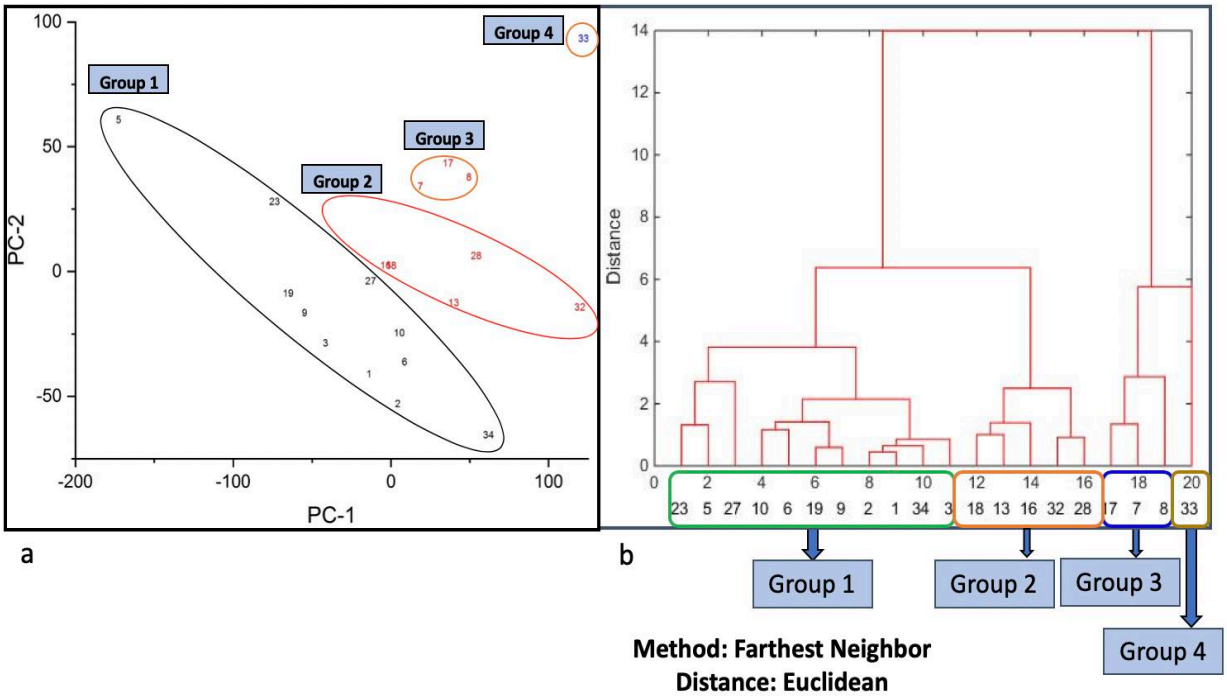

Figure 3.1. Principal component plot and dendogram (furthest linkage) of the average FTIR spectra of the 20 varieties of edible oils comprising the training set.

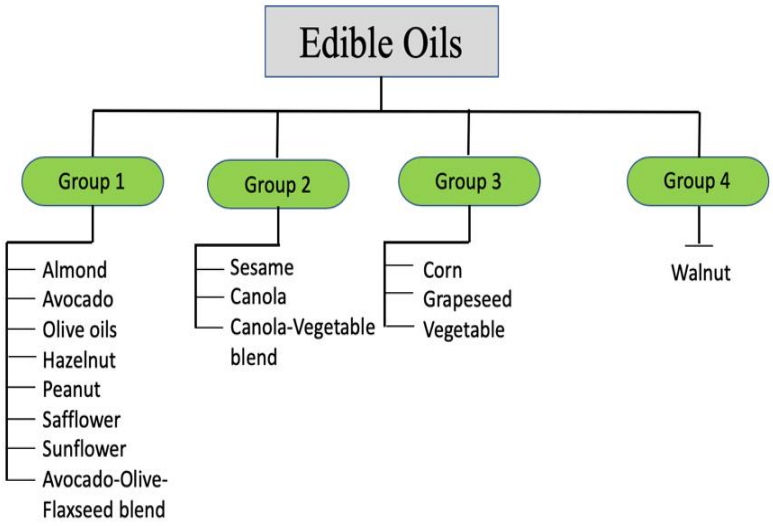

Figure 3.2. Partitioning of the IR spectra of the 20 edible oils into four groups

3.1 CLASSIFICATION OF EDIBLE OILS

The 289 infrared spectra of 6921 points each, which comprises the training set, were subject to principal component analysis. Figure 3.3 shows a plot of the two largest principal components of the 6921 features obtained from the 289 infrared spectra comprising the training set. Each spectrum is represented as a point in the principal component plot. The four groups of edible oils previously detected by cluster analysis are separated from each other in the plot. Since this projection is made without the use of information about the class assignment of each infrared spectrum, the resulting separation is, therefore, a strong indication of real differences in the infrared spectral profile of these four groups.

Feature selection was the next step as the deletion of uninformative features from the infrared spectral profiles often ensures that discriminatory information about edible oil group is the major source of variation in the data. Furthermore, feature selection can transform a difficult classification problem into a simple one. However, the feature selection method employed should be multivariate in nature to ensure that crucial features are not discarded. For these reasons, the pattern recognition GA was applied to the 289 training set spectra to identify wavelengths characteristic of each edible oil group by sampling key feature subsets, scoring their PC plots, and tracking those spectra and/or varieties of edible oils that were difficult to classify. After 200 generations, the pattern recognition GA identified 15 wavelengths whose PC plot (see Figure 3.4) displays four well separated clusters of infrared spectra on the basis of edible oil group. The first principal component appears to differentiate the edible oils by their degree of unsaturation as the edible oils comprising the first cluster, i.e., the Group 1 edible oils (extra virgin olive, extra light olive, pure olive, peanut, avocado, safflower, hazelnut, sunflower, sweet almond, almond, and avocado-olive-flaxseed) have a greater degree of unsaturation than those edible oils comprising Group 2 (canola, canola-vegetable, canola-sun-soybean, extra virgin sesame and toasted sesame), and Group 2 edible oils possess a greater degree of unsaturation than Group 3 edible oils (corn, grapeseed, and vegetable).

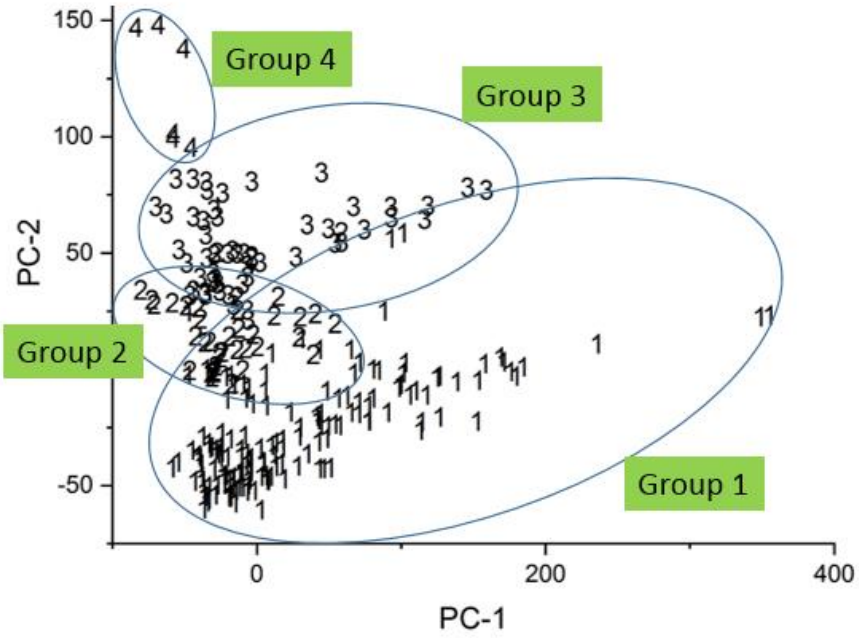

Figure 3.3. Plot of the two largest principal components of the 289 IR spectra and 6921 points comprising the training set. 1 = Group 1, 2 = Group 2, 3 = Group 3, and 4 = Group 4

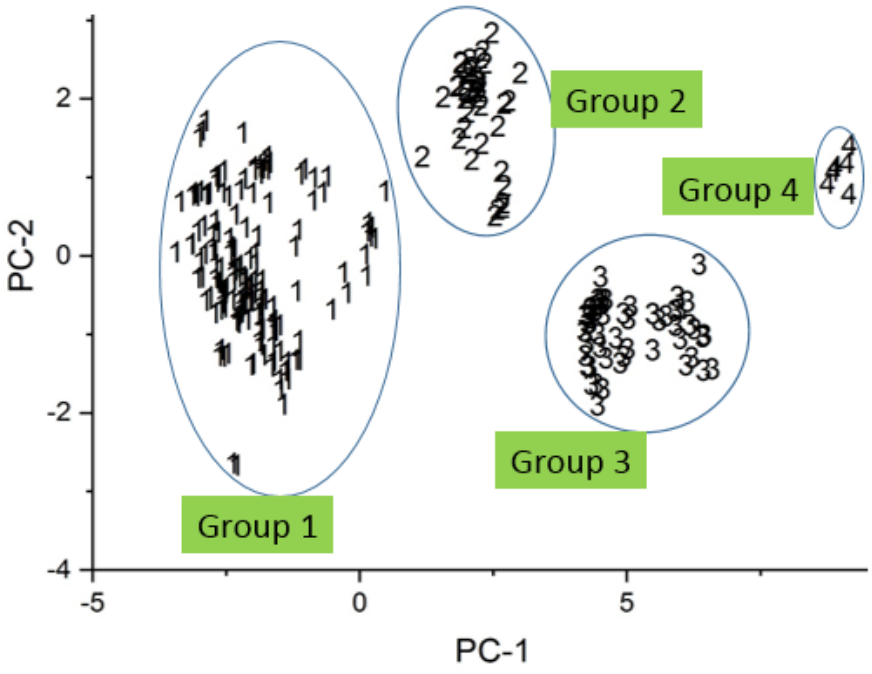

Figure 3.4. Plot of the two largest principal components of the 289 infrared spectra comprising the training set and the 15 spectral features identified by the pattern recognition GA. 1 = Group 1, 2 = Group 2, 3 = Group 3, and 4 = Group 4

A validation set of 81 infrared spectra was used to assess the predictive power of the principal component plot developed from the 289 infrared spectra comprising the training set and the 15 spectral features identified by the pattern recognition GA. The 81 infrared spectra were mapped onto the PC plot developed from the 15 spectral features. Figure 3.5 shows the validation set spectra projected onto the principal component plot. Each projected validation set spectrum lies in a region of the principal component plot with infrared spectra from the same edible oil group. This result suggests that the proposed hierarchical approach to the discrimination of the 20 varieties

of edible oils which involves dividing the oils into four groups is consistent with the structure of the data.

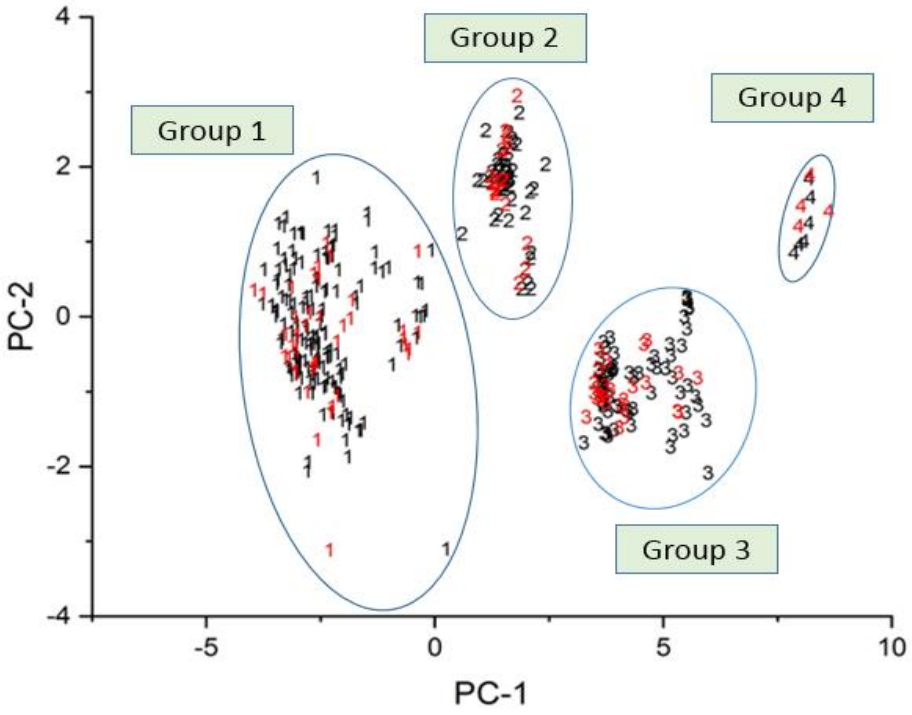

Figure 3.5. Projection of the 81 validation set spectra on the plot of the two largest principal components of the 289 infrared spectra comprising the training set and the 15 spectral features identified by the pattern recognition GA. Training set: 1 = Group 1, 2 = Group 2, 3 = Group 3, and 4 = Group 4.

Validation set: 1 = Group 1, 2 = Group 2, 3 = Group 3, and 4 = Group 4

The next step was to discriminate the edible oils by type within each group. The pattern recognition GA was applied to the 170 infrared spectra comprising the training set for Group 1 (see Table 2.1). The IR spectra of extra virgin olive oil, extra light olive oil, and pure olive oil could not be differentiated using the pattern recognition GA as the triglyceride fraction of these three oils is identical. For this reason, infrared spectra of extra virgin olive oil, extra light olive oil, and pure olive oil were merged into a single class. Figure 3.6 shows a plot of the two largest principal components of the 170 infrared spectra comprising the training set and the 12 spectral features identified by the pattern recognition GA. Each infrared spectrum is represented as a point in the principal component plot with the identity of the edible oil designated for each point. The 29 infrared spectra comprising the validation set for the Group 1 edible oils were projected onto the principal component plot of the 170 infrared spectra and the 12 spectral features identified by the pattern recognition GA (see Figure 3.7). All 29 infrared spectra comprising the validation set were correctly classified, i.e. each validation set sample was projected onto a region of the principal component plot containing infrared spectra of the same edible oil type.

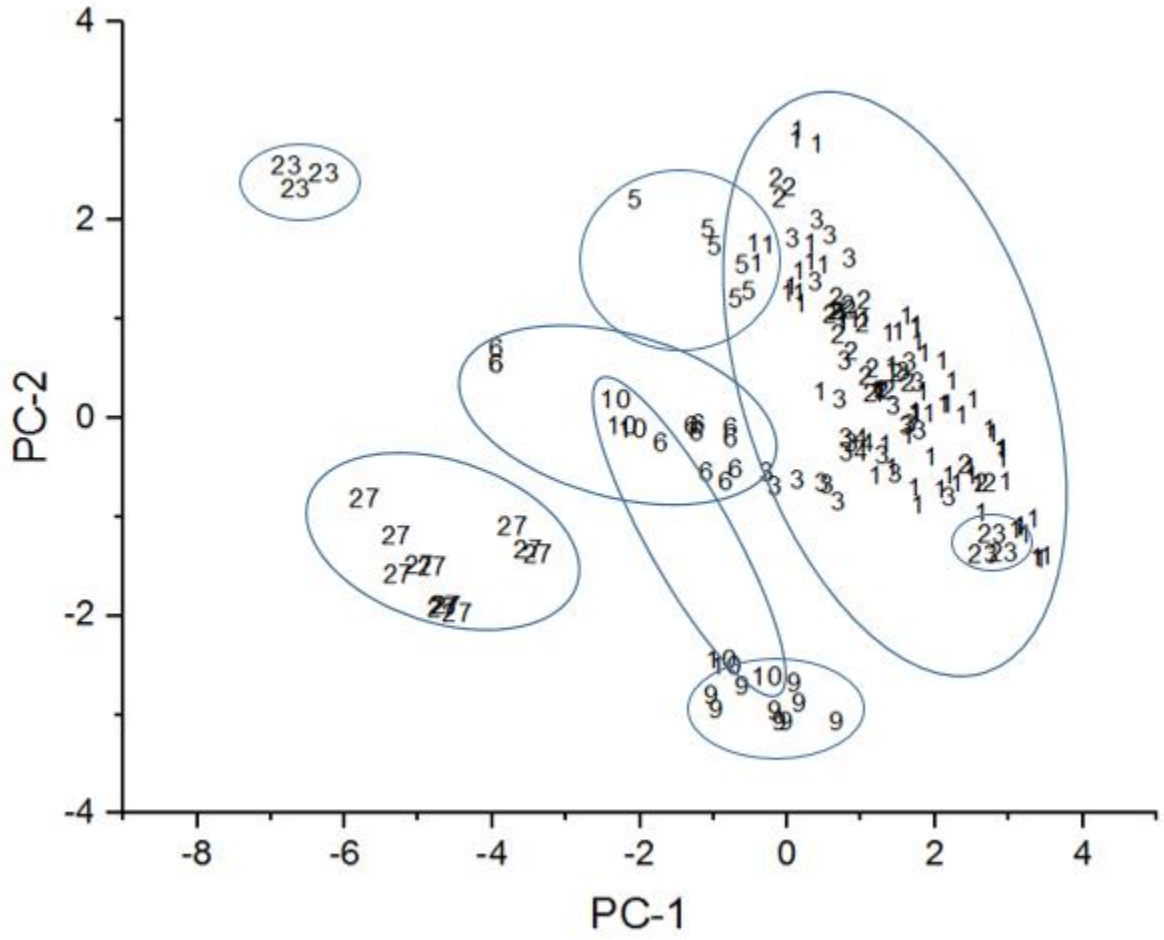

Figure 3.6. Plot of the two largest principal components of the 170 infrared spectra comprising the training set and the 12 spectral features identified by the pattern recognition GA. 1 = EVOO, 2 = ELOO, 3 = Olive Oil, 5 = Avocado, 6 = Peanut, 9 = Safflower, 10 = Hazelnut, 23 = Sweet Almond, and 27 = Almond.

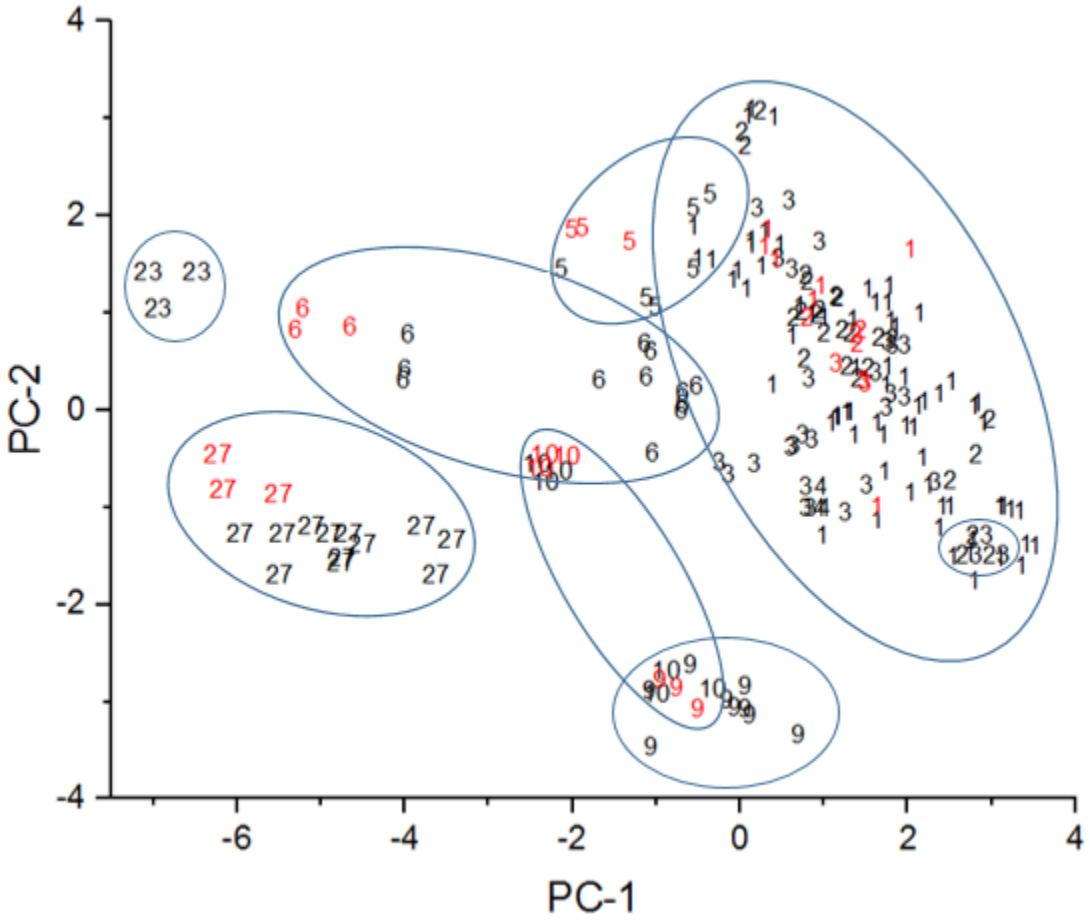

Figure 3.7. Projection of the 29 validation set spectra onto the plot of the two largest principal components of the 170 infrared spectra comprising the training set for Group 1 and the 12 spectral features identified by the pattern recognition GA. Training set: 1 = EVOO, 2 = ELOO, 3 = Olive Oil, 5 = Avocado, 6 = Peanut, 9 = Safflower, 10 = Hazelnut, 23 = Sweet Almond, and 27 = Almond. Validation set: 1 = EVOO, 2 = ELOO, 3 = Olive Oil, 5 = Avocado, 6 = Peanut, 9 = Safflower, 10 = Hazelnut, 23 = Sweet Almond, and 27 = Almond.

Figures 3.6 and 3.7 indicate that within-class variability is greater than class variability for sweet almond, peanut, and hazelnut. In previously published studies, workers have focused their

efforts on within supplier sample variation in presenting a best case scenario for classification. In this study, we considered sample variation within a supplier and between suppliers in the classification of edible oils, which would explain the results that we obtained for sweet almond, hazelnut, and peanut.

For the Group 2 edible oils (Canola, Canola-Sunflower-Soybean, Canola-Vegetable blend, Extra Virgin Sesame, and Toasted Sesame), the pattern recognition GA identified 18 spectral features. Figure 3.8 shows a plot of the two largest principal components of the 49 infrared spectra comprising the training set and the 18 spectral features identified by the pattern recognition GA. A validation set of 17 infrared spectra was used to assess the predictive power of the principal component plot developed from the 49 infrared spectra and 17 spectral features. The 17 infrared spectra were mapped onto the principal component plot developed from these 18 spectral features. Figure 3.9 shows the validation set spectra projected onto this principal component plot. Each projected validation set spectrum is located in a region of the principal component plot containing infrared spectra with the same class label. The results for the Group 3 edible oils are summarized in Figures 3.10 and 3.11.

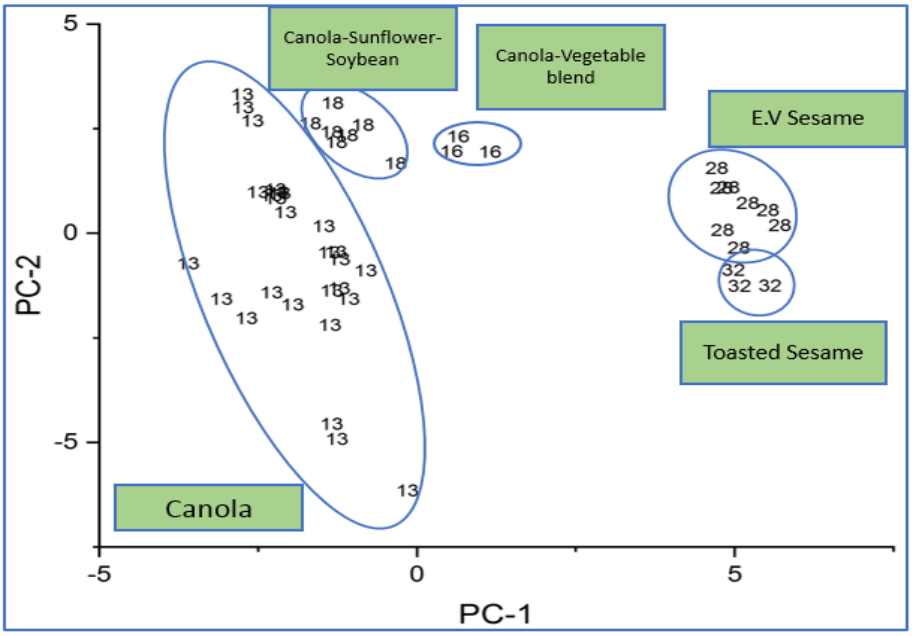

Figure 3.8. Plot of the two largest principal components of the 49 infrared spectra comprising the training set and the 17 spectral features identified by the pattern recognition GA. 13 = Canola, 16 = Canola-Vegetable blend, 18 = Canola-Sunflower-Soybean, 28 = Extra Virgin Sesame, and 32 = Toasted Sesame.

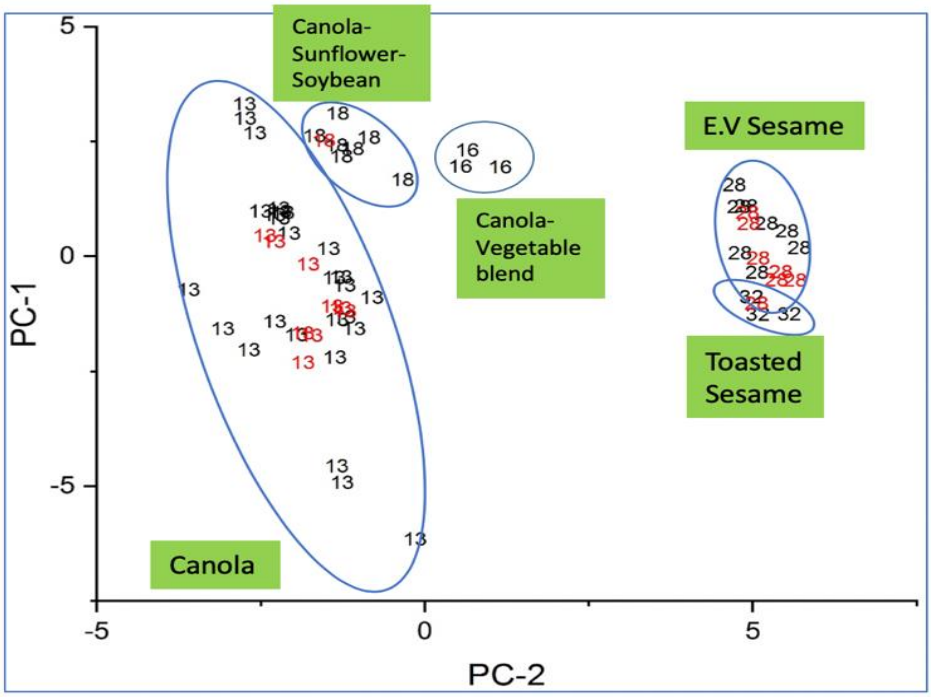

Figure 3.9. Projection of the 17 validation set spectra onto the plot of the two largest principal components of the 49 infrared spectra comprising the training set for Group 2 and the 17 spectral features identified by the pattern recognition GA. Training set: 13 = Canola, 16 = Canola-Vegetable blend, 18 = Canola-Sunflower-Soybean, 28 = Extra Virgin Sesame, and 32 = Toasted Sesame. Validation set: 13 = Canola, 16 = Canola-Vegetable blend, 18 = Canola-Sunflower-Soybean, 28 = Extra Virgin Sesame, and 32 = Toasted Sesame.

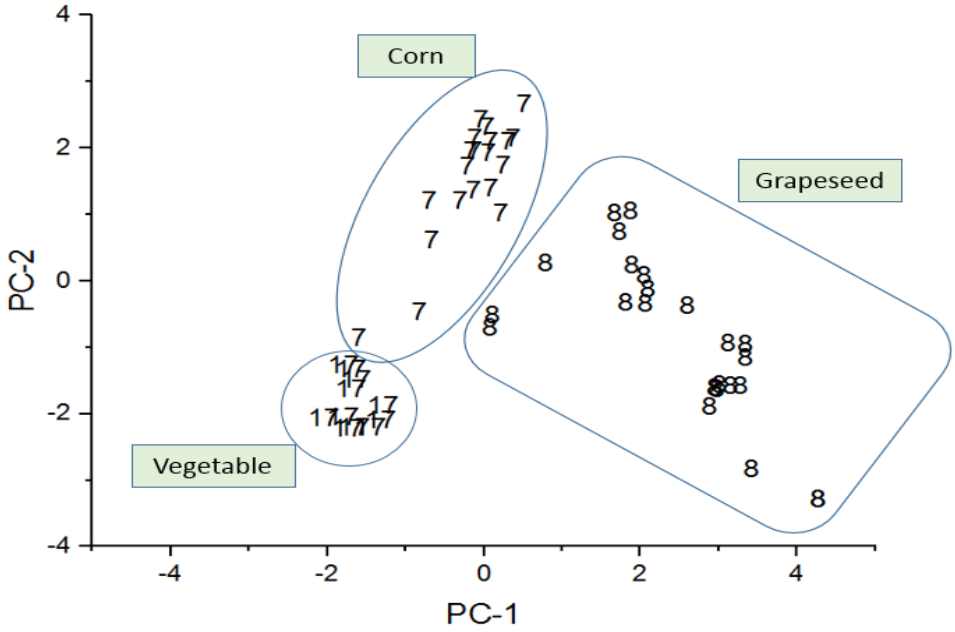

Figure 3.10. Plot of the two largest principal components of the 64 infrared spectra comprising the training set and the 10 spectral features identified by the pattern recognition GA. 7 = Corn Oil, 8 = Grapeseed Oil, and 17 = Vegetable Oil.

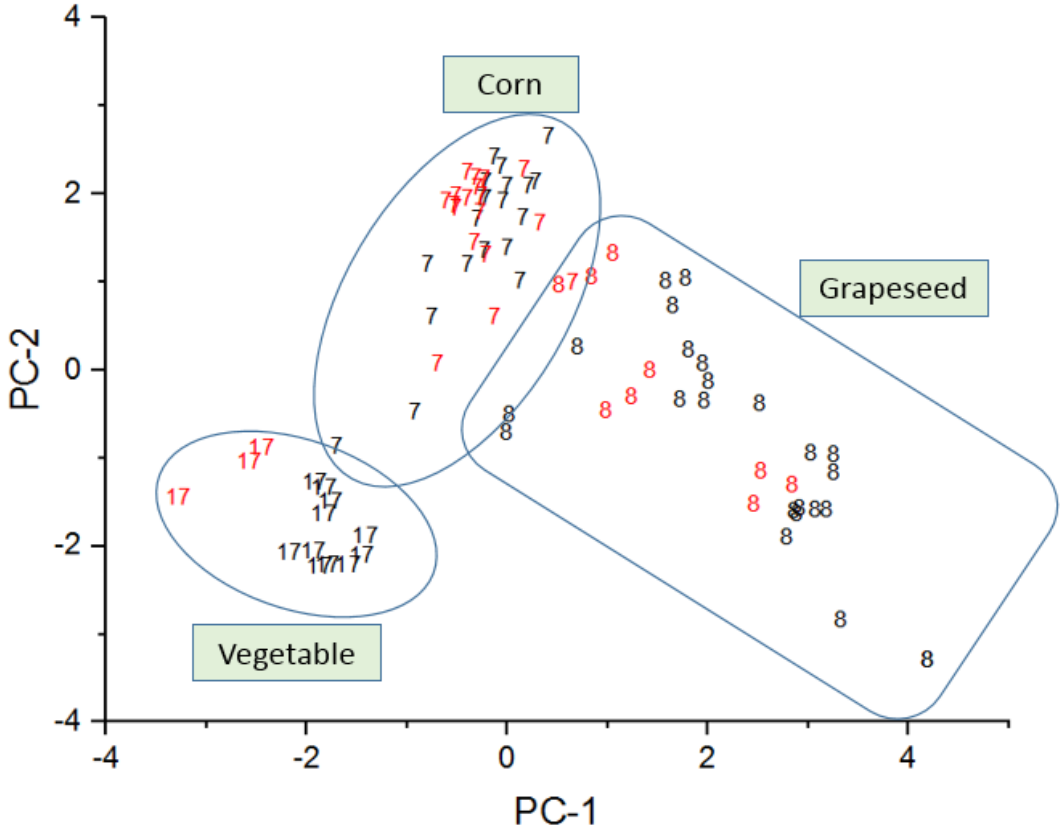

Figure 3.11. Projection of the 31 validation set spectra onto the plot of the two largest principal components of the 64 infrared spectra comprising the training set for Group 3 and the 10 spectral features identified by the pattern recognition GA. Training set: 7 = Corn Oil, 8 = Grapeseed Oil, and 17 = Vegetable Oil.

Validation set: 7 = Corn Oil, 8 = Grapeseed Oil, and 17 = Vegetable Oil.

3.2 ADULTERATION STUDIES

Adulteration studies were performed to simulate adulteration of olive oil through substitution with less expensive edible oils such as almond, hazelnut, corn and canola by mixing

the appropriate volume of each edible oil to yield the desired adulterant mixture. In the first study (see Figure 3.12), the pattern recognition GA was applied to a training set consisting of 11 almond oil spectra and the 56 EVOO spectra. The validation set consisted of 31 spectra of adulterated EVOO with the adulterant mixtures varying from 95% EVOO to 60% EVOO using almond oil as the adulterant. In the second study (see Figure 3.13), the training set consisted of 62 infrared spectra (56 spectra of EVOO and 6 spectra of hazelnut oil), and the validation set contained 31 spectra of EVOO and adulterated EVOO with the adulterant mixtures varying from 95% to 60% EVOO using hazelnut oil as the adulterant. Both almond oil and hazelnut oil as adulterants of EVOO could not be accurately detected by FTIR (see Figures 3.12 and 3.13). Both almond oil and hazelnut oil are in the same edible oil group as EVOO. This begs the questions as to the capability of FTIR to discriminate edible oils in the same group.

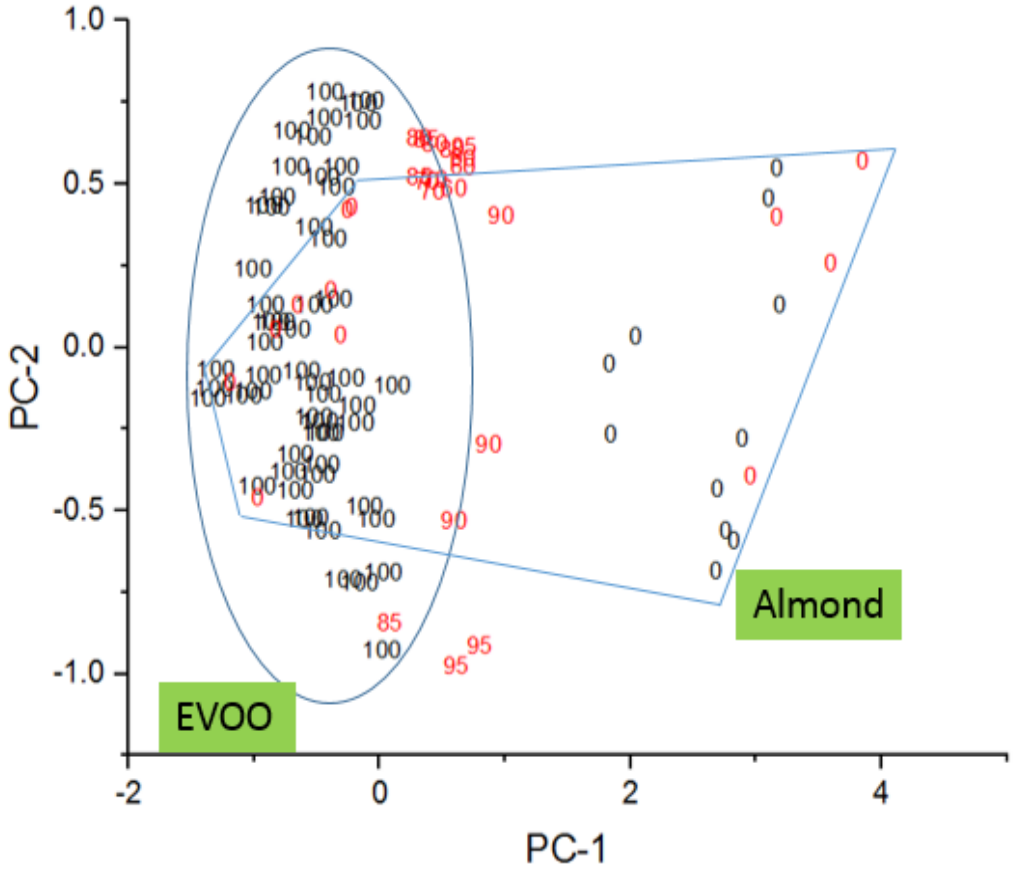

Figure 3.12. Projection of the validation set spectra onto the principal component plot of the 67 spectra (11 almond oil and 56 EVOO) comprising the training set and the two spectral features identified by the pattern recognition GA. Training set: 100 = EVOO and 0 = Almond. Validation set: 95 = 5% adulteration by almond, 90 = 10% adulteration by almond, 85 = 15% adulteration by almond, 80 = 20% adulteration by almond, 70 = 30% adulteration by almond, 60 = 40% adulteration by almond, and 0 = Almond.

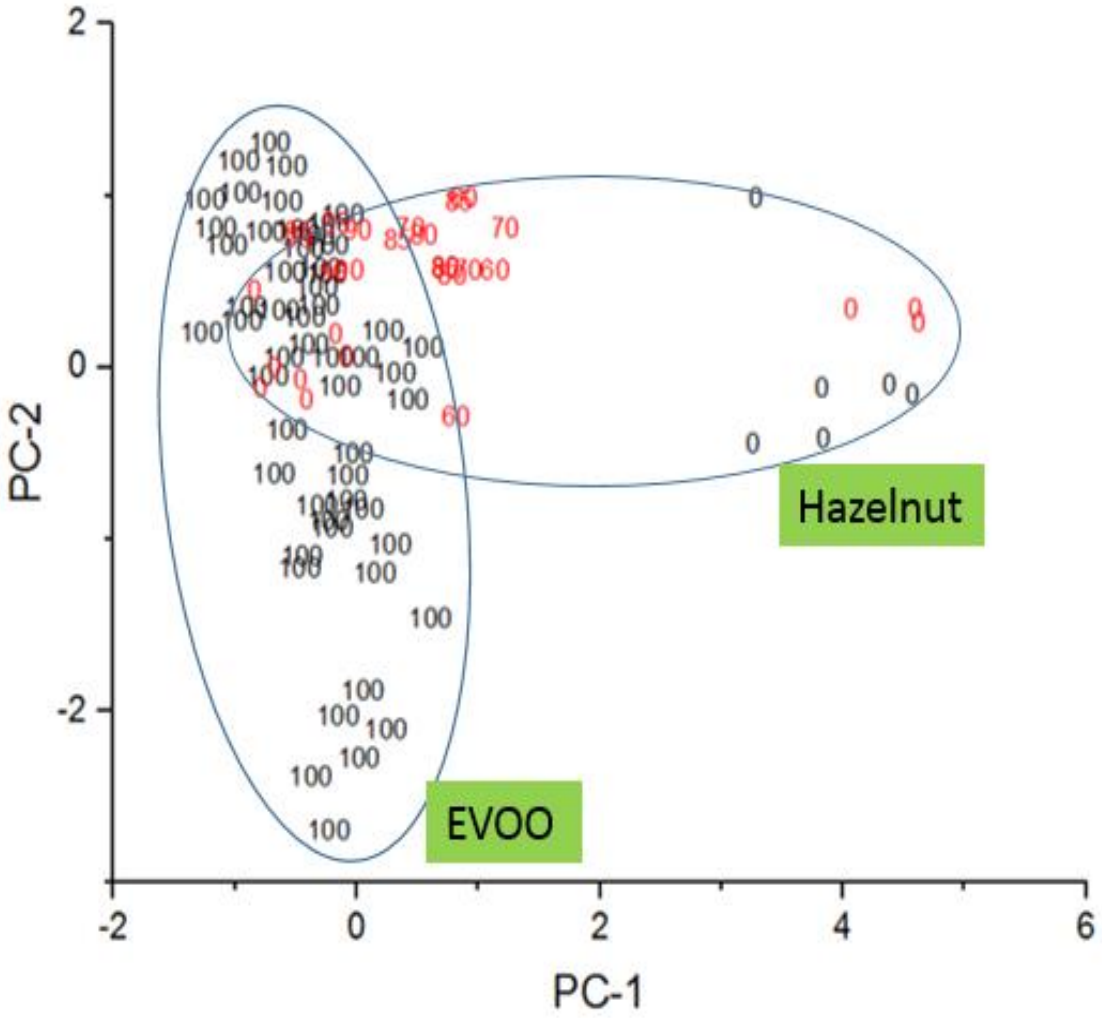

Figure 3.13. Projection of the validation set spectra onto the principal component plot of the 62 spectra (6 hazelnut oil and 56 EVOO) comprising the training set and the four spectral features identified by the pattern recognition GA. Training set: 100 = EVOO and 0 = Hazelnut. Validation set: 95 = 5% adulteration by hazelnut, 90 = 10% adulteration by hazelnut, 85 = 15% adulteration by hazelnut, 80 = 20% adulteration by hazelnut, 70 = 30% adulteration by hazelnut, 60 = 40% adulteration by hazelnut, and 0 = Hazelnut.

Adulteration studies were also performed on olive oil using adulterants from different edible oil groups. In one study, the pattern recognition GA was applied to a training set consisting of 9 ELOO and 24 corn oil spectra with the validation set consisting of 12 corn oil spectra, 3 ELOO spectra and 45 adulterant mixtures of ELOO and corn oil. Figure 3.14 summarizes the results of this study. In a second study, the training set consisted of 9 ELOO, 24 corn oil and 27 canola spectra and the validation set contained 3 canola, 3 ELOO, 12 corn oil, and 32 adulterant mixtures of ELOO, canola and corn oil. Both canola and corn oil are in a different edible oil group from ELOO. Furthermore, these two oils can be discriminated from ELOO using FTIR.

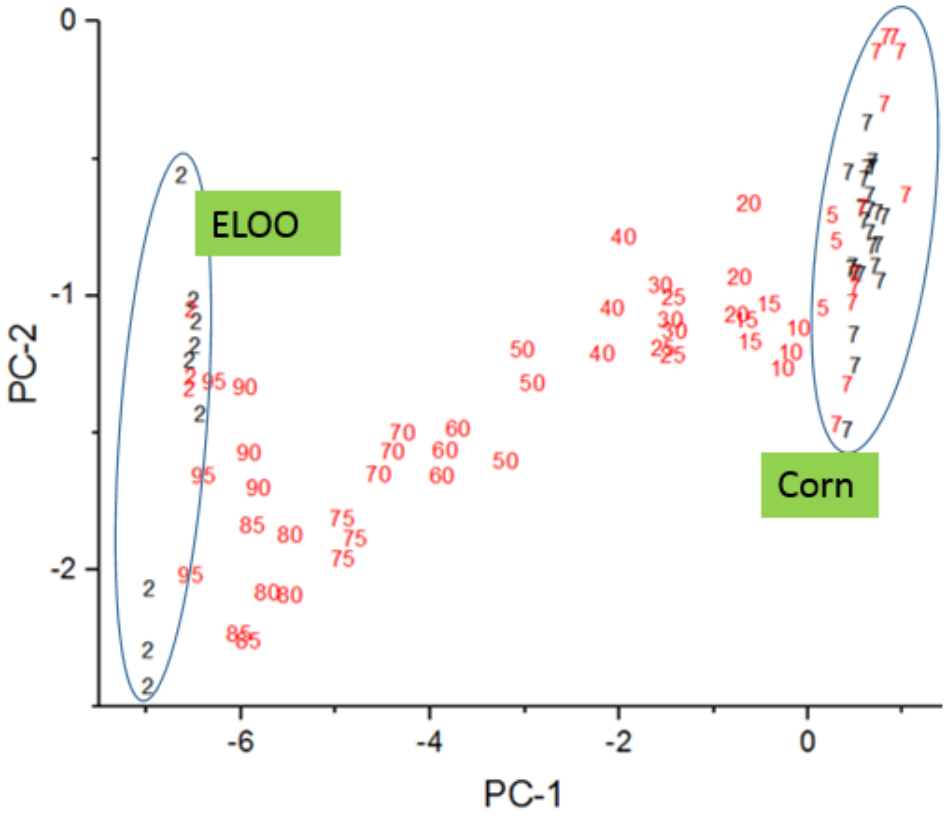

Figure 3.14. Projection of the validation set spectra onto the principal component plot of the 33 spectra (24 corn oil and 9 ELOO) comprising the training set and the 13 spectral features identified by the pattern recognition GA. Training set: 2 = ELOO and 7 = Corn. Validation set: 95 = 5% adulteration by corn, 90 = 10% adulteration by corn, 85 = 15% adulteration by corn, 80 = 20% adulteration by corn, 70 = 30% adulteration by corn, 60 = 40% adulteration by corn, 2 = ELOO and 7 = Corn.

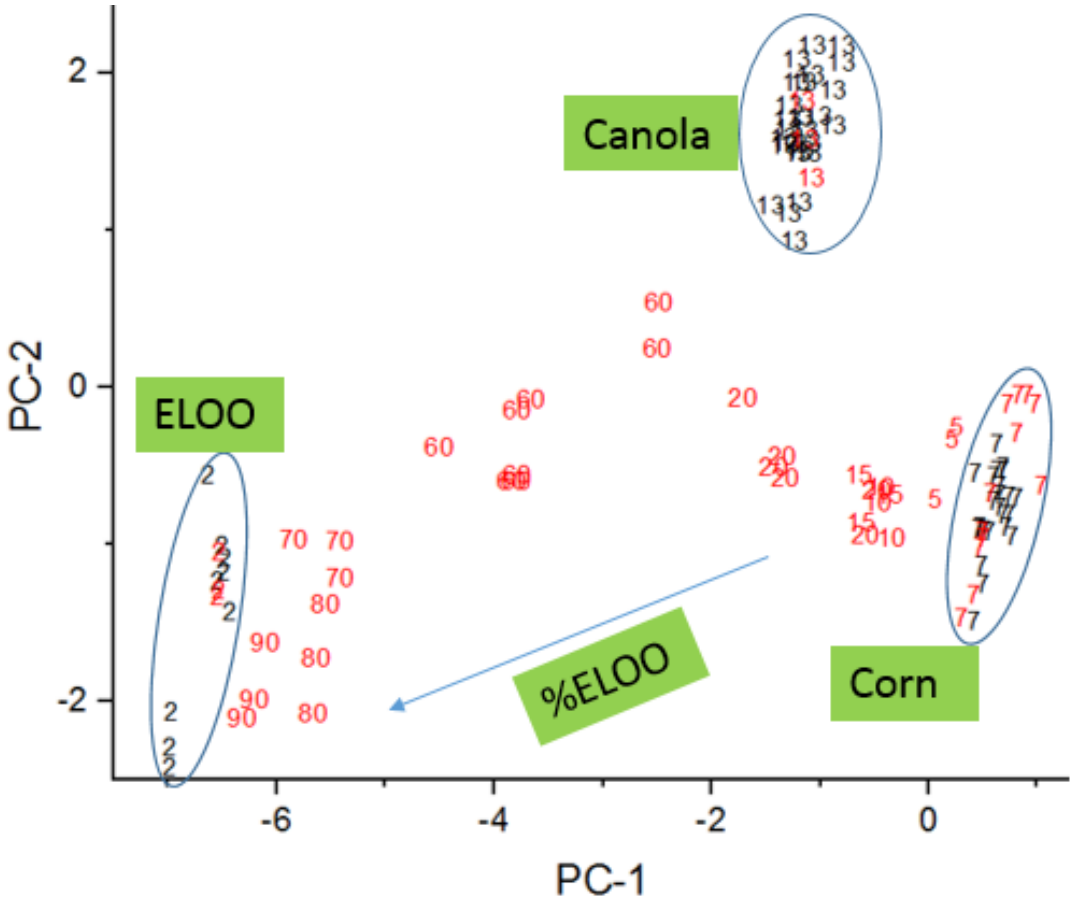

Figure 3.15. Projection of the validation set spectra onto the principal component plot of the 60 spectra (24 corn oil, 27 canola oil and 9 ELOO) comprising the training set and the 13 spectral features identified by the pattern recognition GA. Training set: 2 = ELOO, 7 = Corn and 13 = Canola. Validation set: 90 = 5% adulteration by corn and 5% adulteration by canola, 80 = 10% adulteration by corn and 10% adulteration by canola, 70 = 15% adulteration by corn and 15% adulteration by canola, 60 = 20% adulteration by corn and 20% adulteration by canola, 20 = 60% adulteration by corn and 20% adulteration by canola, 15 = 70% adulteration by corn and 15% adulteration by canola , 10 = 80% adulteration by corn and 10% adulteration by canola, 5 = 90% adulteration by corn and 5% adulteration by canola, 2 = ELOO, 7 = Corn and 13 = Can

CHAPTER IV

CONCLUSIONS

A hierarchical classification scheme has been developed to differentiate FTIR spectra of edible oils by type. The 20 edible oils investigated in this study were divided into 4 groups by cluster analysis. FTIR spectra of edible oils in the same group were more similar to each other than infrared spectra from different groups. Edible oils assigned to Group 1 could be readily differentiated from those assigned to Groups 2, 3, 4, and 5, whereas FTIR spectra of edible oils within the same group more closely resembled each other and therefore were more difficult to classify by type. For example, corn and canola oil (Groups 2 and 3) which previous workers have shown can be readily detected as adulterants in extra virgin olive oil (Group 1 oil) by FTIR spectroscopy, can be readily differentiated from extra light olive oil using the hierarchical based classification scheme described in this study.

Adulteration of extra virgin olive oil by less expensive grades of olive oil such as Grecian olive oil cannot be detected by FTIR spectroscopy because of the similarity of the triglyceride profile of extra virgin olive oil and that of other grades of olive oil. Furthermore, adulteration of extra virgin olive oil by almond or hazelnut oil (both of which are in the same group as olive oil) is a difficult analysis problem as shown by the

adulteration studies performed here. Although other workers have previously reported that PLS using all the wavelengths in the spectrum has the potential to detect hazelnut adulteration in extra virgin olive oil, the training set used was restricted to a single source of extra virgin olive oil and hazelnut oil. The reported PLS results may represent a best case scenario as the training set does not take into account within-supplier and between-supplier variability. Clearly, the hierarchical classification scheme formulated in this study affords the analyst the means to realistically assess the difficulty associated with detecting the presence of a particular adulterant in a specific variety of edible oil using FTIR spectroscopy.

These analyses also highlight the need for edible oil classification and adulteration studies that span multiple sources within each oil type. The supplier to supplier variation for edible oils (and possibly the seasonal variation within a supplier) may be greater than the within supplier variation. This work also indicates that previous studies (which rely on only one source for each type of edible oil) provide an overly optimistic estimate of the ability to spectroscopically classify edible oils or to detect low levels of adulterants.

REFERENCES

Aparicio, R., & Aparicio-Ruíz, R. (2000). Authentication of vegetable oils by chromatographic techniques. *Journal of Chromatography A, 881*(1), 93-104. doi:https://doi.org/10.1016/S0021-9673(00)00355-1

Atadashi, I. M., Aroua, M. K., & Aziz, A. A. (2010). High quality biodiesel and its diesel engine application: A review. *Renewable and Sustainable Energy Reviews, 14*(7), 1999-2008. doi:https://doi.org/10.1016/j.rser.2010.03.020

Baeten, V., Meurens, M., Morales, M. T., & Aparicio, R. (1996). Detection of Virgin Olive Oil Adulteration by Fourier Transform Raman Spectroscopy. *Journal of Agricultural and Food Chemistry, 44*(8), 2225-2230. doi:10.1021/jf9600115

Casoni, D., Simion, I. M., & Sârbu, C. (2019). A comprehensive classification of edible oils according to their radical scavenging spectral profile evaluated by advanced chemometrics. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy, 213*, 204-209. doi:10.1016/j.saa.2019.01.065

Downey, G., McIntyre, P., & Davies, A. N. (2002). Detecting and quantifying sunflower oil adulteration in extra virgin olive oils from the Eastern Mediterranean by visible and near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry, 50*(20), 5520-5525. doi:10.1021/jf0257188

Gouilleux, B., Marchand, J., Charrier, B., Remaud, G. S., & Giraudeau, P. (2018). High-throughput authentication of edible oils with benchtop Ultrafast 2D NMR. *Food Chemistry, 244*, 153-158. doi:10.1016/j.foodchem.2017.10.016

Hammond, E. W. (2003). VEGETABLE OILS | Types and Properties. In B. Caballero (Ed.), *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (pp. 5899-5904). Oxford: Academic Press.

Hou, X., Wang, G., Su, G., Wang, X., & Nie, S. (2019). Rapid identification of edible oil species using supervised support vector machine based on low-field nuclear magnetic resonance relaxation features. *Food Chemistry, 280*, 139-145. doi:10.1016/j.foodchem.2018.12.031

Jiménez-Carvelo, A. M., Osorio, M. T., Koidis, A., González-Casado, A., & Cuadros-Rodríguez, L. (2017). Chemometric classification and quantification of olive oil in blends with any edible vegetable oils using FTIR-ATR and Raman spectroscopy. *LWT - Food Science and Technology, 86*, 174-184. doi:10.1016/j.lwt.2017.07.050

Karoui, R., Downey, G., & Blecker, C. (2010). Mid-infrared spectroscopy coupled with chemometrics: A tool for the analysis of intact food systems and the exploration

of their molecular structure-quality relationships-A review. *Chemical Reviews, 110*(10), 6144-6168. doi:10.1021/cr100090k

Lastra-Mejías, M., Aroca-Santos, R., Izquierdo, M., Cancilla, J. C., Mena, M. L., & Torrecilla, J. S. (2019). Chaotic parameters from fluorescence spectra to resolve fraudulent mixtures of fresh and expired protected designation of origin extra virgin olive oils. *Talanta, 195*, 1-7. doi:10.1016/j.talanta.2018.10.102

Lerma-García, M. J., Ramis-Ramos, G., Herrero-Martínez, J. M., & Simó-Alfonso, E. F. (2010). Authentication of extra virgin olive oils by Fourier-transform infrared spectroscopy. *Food Chemistry, 118*(1), 78-83. doi:https://doi.org/10.1016/j.foodchem.2009.04.092

López-Díez, E. C., Bianchi, G., & Goodacre, R. (2003). Rapid Quantitative Assessment of the Adulteration of Virgin Olive Oils with Hazelnut Oils Using Raman Spectroscopy and Chemometrics. *Journal of Agricultural and Food Chemistry, 51*(21), 6145-6150. doi:10.1021/jf034493d

Luna, A. S., da Silva, A. P., Ferré, J., & Boqué, R. (2013). Classification of edible oils and modeling of their physico-chemical properties by chemometric methods using mid-IR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 100*, 109-114. doi:https://doi.org/10.1016/j.saa.2012.06.034

Moore, J. C., Lipp, M., & Griffiths, J. C. (2011). Preventing the adulteration of food protein with better analytical methods: Avoiding the "next melamine". *INFORM - International News on Fats, Oils and Related Materials, 22*(6), 373-375.

Moore, J. C., Spink, J., & Lipp, M. (2012). Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *77*(4), R118-R126. doi:10.1111/j.1750-3841.2012.02657.x

Ng, T. T., Li, S., Ng, C. C. A., So, P. K., Wong, T. F., Li, Z. Y., . . . Yao, Z. P. (2018). Establishment of a spectral database for classification of edible oils using matrix-assisted laser desorption/ionization mass spectrometry. *Food Chemistry, 252*, 335-342. doi:10.1016/j.foodchem.2018.01.125

Phillips, T. (2014, October 29, 2018). Commonly Used Edible Oils and Their Benefits. Retrieved from https://www.thebalance.com/what-are-edible-oils-3973288

Rohman, A., & Che Man, Y. B. (2011). The use of Fourier transform mid infrared (FT-MIR) spectroscopy for detection and quantification of adulteration in virgin coconut oil. *Food Chemistry, 129*(2), 583-588. doi:https://doi.org/10.1016/j.foodchem.2011.04.070

Rohman, A., & Man, Y. B. C. (2010). Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. *Food Research International, 43*(3), 886-892. doi:https://doi.org/10.1016/j.foodres.2009.12.006

Salimon, J., Salih, N., & Yousif, E. (2012). Industrial development and applications of plant oils and their biobased oleochemicals. *Arabian Journal of Chemistry, 5*(2), 135-145. doi:https://doi.org/10.1016/j.arabjc.2010.08.007

Spink, J. (2011). The challenge of intellectual property enforcement for agriculture technology transfers, additives, raw materials, and finished goods against product fraud and counterfeiters. *Journal of Intellectual Property Rights, 16*(2), 183-193.

Spink, J., & Moyer, D. C. (2011). Defining the Public Health Threat of Food Fraud. *Journal of Food Science, 76*(9), R157-R163. doi:10.1111/j.1750-3841.2011.02417.x

Tay, A., Singh, R. K., Krishnan, S. S., & Gore, J. P. (2002). Authentication of Olive Oil Adulterated with Vegetable Oils Using Fourier Transform Infrared Spectroscopy. *LWT - Food Science and Technology, 35*(1), 99-103. doi:https://doi.org/10.1006/fstl.2001.0864

Torrecilla, J. S. (2010). The olive : its processing and waste management.

Vaskova, H., & Buckova, M. (2015). Thermal Degradation of Vegetable Oils: Spectroscopic Measurement and Analysis. *Procedia Engineering, 100*, 630-635. doi:https://doi.org/10.1016/j.proeng.2015.01.414

Vlachos, N., Skopelitis, Y., Psaroudaki, M., Konstantinidou, V., Chatzilazarou, A., & Tegou, E. (2006). Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta, 573-574*, 459-465. doi:https://doi.org/10.1016/j.aca.2006.05.034

Wójcicki, K., Khmelinskii, I., Sikorski, M., & Sikorska, E. (2015). Near and mid infrared spectroscopy and multivariate data analysis in studies of oxidation of edible oils. *Food Chemistry, 187*, 416-423. doi:https://doi.org/10.1016/j.foodchem.2015.04.046

Zhang, Q., Liu, C., Sun, Z., Hu, X., Shen, Q., & Wu, J. (2012). Authentication of edible vegetable oils adulterated with used frying oil by Fourier Transform Infrared Spectroscopy. *Food Chemistry, 132*(3), 1607-1613. doi:https://doi.org/10.1016/j.foodchem.2011.11.129

Du, T. Y. (2019). Dimensionality Reduction Techniques for Visualizing Morphometric Data: Comparing Principal Component Analysis to Nonlinear Methods. *Evolutionary Biology, 46*(1), 106-121. doi:10.1007/s11692-018-9464-9

Liu, S., Maljovec, D., Wang, B., Bremer, P. T., & Pascucci, V. (2017). Visualizing High-Dimensional Data: Advances in the Past Decade. *IEEE Transactions on Visualization and Computer Graphics, 23*(3), 1249-1268. doi:10.1109/TVCG.2016.2640960

Pořízka, P., Klus, J., Képeš, E., Prochazka, D., Hahn, D. W., & Kaiser, J. (2018). On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review. *Spectrochimica Acta - Part B Atomic Spectroscopy, 148*, 65-82. doi:10.1016/j.sab.2018.05.030

Pei, M. (1998). Feature Extraction Using Genetic Algorithms. Proc. of International Symposium on Intelligent Data Engineering and Learning. https://ci.nii.ac.jp/naid/10010553087/en/

VITA

Matthew Temitope Bamidele

Candidate for the Degree of

Master of Science

Thesis: ANALYSIS OF EDIBLE OILS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND CHEMOMETRICS

Major Field: Chemistry

Biographical:

Education:

Completed the requirements for the Master of Science in Chemistry (Analytical Chemistry) at Oklahoma State University, Stillwater, Oklahoma in December, May, 2019.

Completed the requirements for the Master of Science in your Physical Chemistry at University of Ibadan, Ibadan, Oyo State Nigeria, 2012.

Completed the requirements for the Bachelor of Science in your Pure Chemistry at University of Ibadan, Ibadan, Oyo State Nigeria, 2008

Experience:

Teaching/Research Assistant, Department of Chemistry, Oklahoma State University (2015-2019)
Assistant Lecturer, Federal University Otuoke, Bayelsa State, Nigeria (2014-2015)
Assistant Lecturer, Wesley University of Science and Technology, Ondo, Ondo State, Nigeria (2013-2014)
Lecturer, School of Hygiene, Eleyele, Ibadan, Oyo State (2008-2013)

Professional Memberships:

Royal Society of Chemistry
American Chemical Society
Electrochemical Society
Phi Lambda Upsilon (PLU) Chemistry Society (Oklahoma State University Chapter)