VIBRATIONAL SPECTROSCOPY AND CHEMOMETRICS APPLIED TO THE FORENSIC ANALYSIS OF AUTOMOTIVE PAINTS AND EDIBLE OILS

By

FRANCIS KWOFIE

Bachelor of Science in Chemistry
University of Cape Coast
Cape Coast, Ghana
2012

Master of Science in Chemistry
East Tennessee State University
Johnson City, Tennessee
2015

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2019

VIBRATIONAL SPECTROSCOPY AND CHEMOMETRICS APPLIED TO THE FORENSIC ANALYSIS OF AUTOMOTIVE PAINTS AND EDIBLE OILS

Dissertation Approved:

Dr. Barry K. Lavine
Dissertation Adviser

Dr. Ziad El Rassi

Dr. Nicholas F. Materer

Dr. Sadagopan Krishnan

Dr. Albert T. Rosenberger

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to my research advisor, Dr. Barry Kenneth Lavine for his patience, understanding, guidance and motivation throughout my studies here at Oklahoma State University.

I would also like to thank my graduate advisory committee Dr. Ziad El Rassi, Dr. Nicholas Materer, Dr. Sadagopan Krishnan, and Dr. Albert T. Rosenberger for their guidance, advices and valuable comments.

I am also thankful to my lab members, Kaushalya, Isio, Matthew, Tom, George and Haoran for their friendship and help when I needed it. Special thanks goes to former lab mates Dr. Ayuba Fasasi, Dr. Collin White, Dr. Tao Ding, and Dr. Nuwan Don Perera for their helpful discussions and guidance.

Finally, I want to thank my family, especially my wife, Amy Nichole Kwofie and son Landen Alexander Kwofie for their unconditional love, tremendous support and encouragement. I thank my uncles George Kolog Gbinniya, Richard Owusu and my sister Joyce Owusuaa and late Margarete Owusuaa for their tremendous support for my career choice, without whom I would not be where I am today.

Name: FRANCIS KWOFIE

Date of Degree: JULY 2019

Title of Study: VIBRATIONAL SPECTROSCOPY AND CHEMOMETRICS APPLIED TO THE FORENSIC ANALYSIS OF AUTOMOTIVE PAINTS AND EDIBLE OILS

Major Field: CHEMISTRY

Abstract: Profiling of complex materials (e.g., automotive paint and cooking oil) with infrared and Raman spectroscopy is an active area of research with a large and growing literature. The object of profile analysis is to correlate a characteristic fingerprint pattern in a spectrum with the properties of the sample. Objective analysis of these profiles depends upon the use of multivariate curve statistical methods. In this regard, pattern recognition techniques have been found to be of enormous utility. In this dissertation, several projects were undertaken to demonstrate the advantages of chemical fingerprinting using spectroscopic techniques to solve problems in the areas of food chemistry and forensic science. In one study, Raman spectra of 15 varieties of edible oils obtained from 53 samples purchased over a 3 year period representing different production years and vendors were analyzed by pattern recognition methods using a hierarchical classification procedure. Supplier to supplier variability and seasonal variability within a supplier were the major sources of variation with the Raman spectral data. Edible oils assigned to one group could be readily differentiated from those assigned to other groups, whereas Raman spectra within the same group more closely resemble each other and therefore were more difficult to classify by type. In another study, IR microscopic imaging and a prototype pattern recognition library search system were applied to the forensic examination of automotive paint using a new methodology for cross sectioning paint samples and decatenating infrared spectral images. Successful methods developed in test experiments such as the studies described in this dissertation will become part of the routine analytical practices of chemists in the very near future.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Profiling of complex materials (e.g., automotive paint and cooking oil) with infrared and Raman spectroscopy continues to be an active area of research with a burgeoning literature. The object of profile analysis is to correlate a characteristic fingerprint pattern in a spectrum with the properties of a sample. Objective analysis of these profiles depends upon the use of multivariate statistical methods. In this regard, pattern recognition techniques have been found to be of enormous utility.

Pattern recognition methods are well suited for analyzing spectroscopic data because of the characteristics of the procedures. Methods are available that do not assume a mathematical model but rather seek relationships that provide definitions of similarity between groups of data. Pattern recognition methods are also able to deal with high dimensional data where more than three measurements are used to describe each sample. Finally, techniques are available for selecting important features from a large set of measurements. Thus, studies can be performed on systems where the exact relationships are not fully understood.

The research described in this dissertation is directed towards three specific goals: (1) development of a potential method to improve the quality of spectral library searches of automotive paints by applying infrared microscopy and alternating least squares to cross sectioned automotive paint samples, (2) coupling the FTIR imaging experiment with a prototype pattern recognition infrared library searching system previously developed to facilitate both the accuracy and speed of forensic automotive paint analysis, and (3) applying reliable variable selection methods to improve discrimination of edible oils by Raman spectroscopy. The significance of this research lies in the development of new methods to address problems of widespread interest in the areas of forensic science and food chemistry.

In the forensic examination of automotive paint, each layer of paint is analyzed individually by infrared spectroscopy. Forensic laboratories in the United States and Canada typically hand section each layer and present each separated layer to the spectrometer for analysis, which is time consuming. In addition, sampling too close to the boundary between adjacent layers can pose a problem as it produces an IR spectrum that is a mixture of the two layers. Not having a "pure" spectrum of each layer can prevent a meaningful comparison between each paint layer or in the situation of searching an automotive database will prevent the forensic paint examiner from developing an accurate hit list of potential suspects. These two problems have been addressed by collecting concatenated IR data from all paint layers in a single analysis by scanning across the cross sectioned layers of the paint sample using a FTIR imaging microscope. Decatenation of the IR data was achieved by multivariate curve resolution to obtain a pure IR spectrum of each automotive paint layer. Comparing the reconstructed IR spectrum of each layer

against the IR spectral library of the PDQ database demonstrated that it was possible to identify the correct line and model of the vehicle from these reconstructed spectra. This imaging approach to IR analysis of automotive paint, will not only save time and eliminate the need to analyze each layer separately, but also will ensure that the final spectrum of each layer is "pure" and not a mixture.

By coupling the proposed FTIR imaging experiment with the prototype pattern recognition infrared library searching system previously developed by the Lavine research group to search the paint data query (PDQ) automotive paint database, the forensic examination of automotive paint will be facilitated in terms of both accuracy and speed of the analysis. The prototype library searching system consists of two separate but interrelated components: search prefilters to cull the library spectra to a specific assembly plant or assembly plants and a cross correlation searching algorithm to identify spectra most similar to the unknown in the set of spectra identified by the search prefilters as potential matches. As the size of the library is culled for a specific match, the search prefilters will increase both the selectivity and accuracy of the search. Even in challenging trials where the paint samples evaluated were all the same make (General Motors) within a limited production year range, the respective assembly plant as well as the make and model of the vehicle could be identified from IR spectra of the clear coat and undercoat paint layers.

Manually coded text based searches performed using the current PDQ database tend to generate a large number of hits because the chemical information in the current database is described only in terms of generic chemical formulations. Furthermore, improper coding of the spectra and/or searching of the PDQ database may inadvertently include or exclude

certain motor vehicles from the hit list. Therefore, an added advantage of the proposed pattern recognition assisted approach to identify paint samples is an increase in accuracy because all IR spectra in the database are searched. In addition, the use of the proposed search prefilters gives far fewer hits in the database which translate into a significant time savings for the forensic scientist, ease of use, and fewer errors. Information derived from the proposed pattern recognition searches will also serve to quantify the general discrimination power of original automotive paint comparisons encountered in casework, and will further efforts to succinctly communicate the significance of the evidence to the courts.

Differentiation of edible oils by variety was investigated using Raman spectroscopy and pattern recognition methods. Raman spectra of 15 varieties of edible oils obtained from 53 samples purchased over a 3 year period representing different production years and vendors (possibly the same company but a different batch and from a different manufacturing plant) were collected at relatively short integration times to test the robustness of the Raman analysis to noise. By comparison, previously published studies on this subject have been limited to 20 samples obtained from a single brand within a limited production year range involving five or six edible represented by samples. The relatively large number of classes, samples, and spectra (i.e., replicates) in this study were necessary to build better statistical distributions of expected in-class variance to determine classification performance when developing discriminants from training sets and to have sufficient number of spectra to construct independent training and validation sets. Furthermore, the oils are a flexible platform from which we can collect data. The oil spectra have a low, but tunable net analyte signal to background (NAS/B). Edible oils are

essentially mixtures of triglycerides that differ in their relative composition of fatty acids (e.g., oleic, stearic, and linoleic). There are 5 major Raman features in our spectral window that vary in relative intensity and position among all samples. Experimental designs can be constructed with very similar oils (e.g., olive and sunflower) with relatively distinct spectra.

Supplier to supplier variability and seasonal variability within a supplier were major sources of variation within the Raman spectral data set as it is not only greater than variability within a supplier but was comparable in magnitude to the variability associated with edible oil type. The 15 varieties of edible oils could be partitioned into 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 Raman spectra within the same group more closely resembled each other and therefore were more difficult to classify by type.

This thesis is divided into seven chapters. The first chapter is the introduction which provides an overview of the research problems pursued in this dissertation. Chapters 2 and 3 provide the necessary background and theory and the research problems described in this dissertation are discussed in Chapters 4, 5 and 6. A summary of the results obtained in this dissertation research are outlined in Chapter 7.

CHAPTER II

LITTERATURE REVIEW OF AUTOMOTIVE PAINT AND EDIBLE OILS

2.1. The Composition of Paint

The primary function of paint is to protect and to improve the aesthetic nature of an object [1-3]. Paints or coatings can be liquids or powders that form adherent films on the surface of substrates. The origin of paints can be traced back to the Paleolithic era where a combination of sap extracted from plants and coloring agents obtained from berries and soil were used in cave paintings [4]. Paint is comprised of four components: binder(s), pigment(s), additive(s) and solvent [2, 4, 5].

2.1.1. Binder

The binder is a fluid or a polymeric constituent in which the pigment is suspended. It provides the necessary adhesion to ensure that both the pigments and additives are retained by the coating while ensuring that the paint will be attached to the object or substrate [2, 4, 6]. Upon curing, which can occur by evaporation, coagulation or polymerization [2, 4, 6], the binder serves as the foundation for the paint on the substrate. Films that are formed by evaporation often leave behind the binder, pigments and additives, which are known as lacquers. Lacquers can be easily re-dissolved upon addition of the

solvent. Increasing the molecular mass of the binder can improve the properties of the polymer film. Some common binders such as acrylic and polyvinyl acetate are synthetic binders whereas others such as casein, and cellulose are natural binders [7-9].

2.1.2. Pigments

Pigments are usually in the form of a powder and are responsible for providing color and as well as protecting the object against corrosion. Pigments can be classified as organic or inorganic. A large number of organic pigments exist [10]. Some organic pigments (natural or synthetic) are soluble in certain solvents. The advantages of organic pigments are richer colors and greater durability [10]. Inorganic pigments are less expensive, more resistant to ultraviolet light, more effective in protecting the substrate from corrosion and better heat stability. Extender pigments, which are a subgroup of inorganic pigments, do not contribute to color or corrosion resistance but enhance other coating properties such as flow, density, hardness and permeability. These properties make them attractive as they are able to reduce production costs [8, 9].

2.1.3. Additives

Additives are added to paints in small amounts to improve the performance characteristics of the finished coating [11]. Examples of additives include thickeners and surfactants, which reduce the surface tension of a liquid, or driers, which act as a catalyst for the natural process of oxidation to improve drying. There is a broad spectrum of additives that can affect and enhance the performance properties including sag resistance, de-foaming, gloss, viscosity, flexibility, ultraviolet, fire and microbial resistance [2, 5, 12].

However, some of the additives used present unique health issues. Because of these issues, there is on-going research to identify new additives so as to minimize the use of plasticizers [7, 9].

2.1.4. Solvents

Solvents play an important role in paint as it ensures that binders, pigments and additives are in a liquid solution facilitating easy application to the substrate [4, 12]. During the curing process, the solvent is usually lost after the application of heat. Volatile organic compounds are not used as solvents due to health and environmental reasons. For this reason, powder coatings have been developed which contain all other major constituents of paint with the exception of the solvent. Solvents with a high vapor pressure are classified as either fast or hot solvents. The coating properties is partially dependent on the rate of solvent evaporation [10].

2.1.5. The Composition and Structure of Automotive Paints

Modern automotive paint systems [13] are comprised of four distinct layers: clear coat, color coat (also known as base coat), surfacer-primer and electro-coat (which is referred to as the e-coat). With the exception of the clear coat, each paint layer contains fillers and pigments [14] and all the layers contain binders. Each automotive manufacturer tend to use a unique combinations of fillers and binders in each paint layer [14]. This unique combinations allow forensic scientists to determine the make and model of a vehicle within a limited production year range from an automotive paint chip recovered from a crime scene involving a vehicle fatality such as a hit-and-run. The original equipment manufacturer (OEM) automotive paint system is usually applied sequentially in a number of steps. Prior to depositing the paint system, all metallic components of the vehicle are

normally pre-treated *via* zinc electroplating (~ 1 μm), in order to prevent corrosion and inhibit rust [15]. The first paint layer applied to the vehicle is the e-coat, which is an approximately 20 μm thick epoxy based coating. It is electroplated onto the body of the vehicle: (1) to provide greater resistance against corrosion, (2) to protect the vehicle from stone chirping [16], (3) to hide any minor imperfections, and (4) to serve as an adhesive platform for the other layers. After the application of the e-coat, a primer surfacer, of approximately 30-40 μm alkyd based coating, is applied to hide any surface imperfections, and provide a uniform foundation that will be both amenable and more receptive to the application of the basecoat [15]. After the application of the primer surfacer, a roughly 15 μm thick pigment containing layer known as the basecoat is applied to achieve the desired color [16, 17]. The clear coat is the final coat applied in the automotive finishing process. The clear coat, which is the thickest of all the layers, is typically a 40 μm thick un-pigmented layer, consisting of UV absorbers and hindered amine light stabilizers. Their primary function is to protect the basecoat and underlying layers from UV degradation and weathering [16]. The clear coat also contributes important properties to the body of the vehicle such as hardness, and resistance to chemicals and solvents. A typical automotive paint system is shown in Figure 2.1.

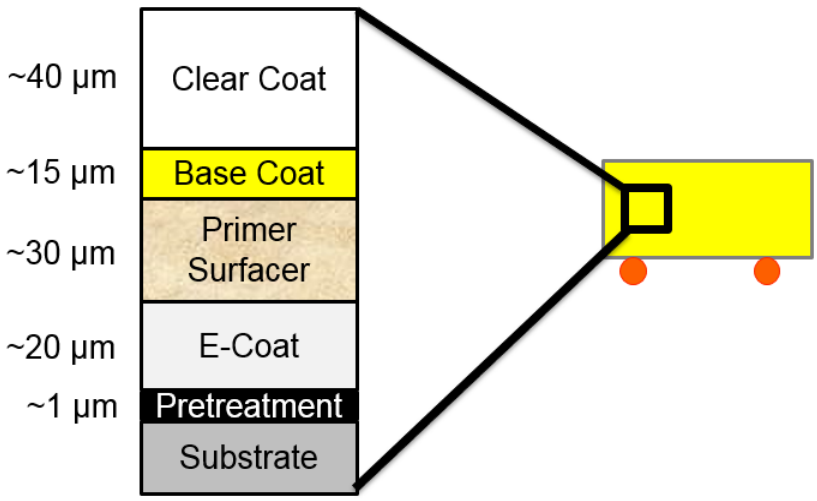

Figure 2.1. Typical automotive paint system comprising the clear coat, the basecoat, the primer-surfacer and the e-coat

2.1.6. Types of Automotive Paint Evidence

Automotive paint sample recovered from a crime scene can either be in the form of a paint chip or smear. Several factors such as the force of impact or collision and the nature of the victim's body surface determines the amount of paint evidence transferred [18]. If the force of collision between the vehicle and the victim is not great, it is most likely that only the top coat or the clear coat of the paint will be transferred. On the other hand, if the force of impact is great, then not only the clear coat but the other layers will also be transferred. Smears are usually generated by a glancing contact between the vehicle and the victim's body. Automotive paint chips, however, result from a greater or more forceful direct impact between the vehicle and the object in question leading to the deformation of the vehicle frame resulting in the generation of paint chips [18]. Paint chips usually contain all of the automotive paint layers, which makes it easier to determine the make and model of the vehicle. Furthermore, the individual layers comprising a paint chip can be hand sectioned whereas isolating the individual layers of a smear is problematic.

2.1.7. Automotive Paint Analysis

Chemical analysis of automotive paint samples such as paint chips or smears is typically done using Fourier transform infrared (FTIR) spectroscopy [19]. Most laboratories in North America are likely to hand-section each layer of the paint chip and present each separated layer to either an IR microscope fitted with an ATR accessory or collect transmission spectra. A diamond anvil cell is typically employed when a transmission spectra is to be collected. Other techniques employed in the analysis of automotive paints include but not limited to include microscopical examinations, colorimetric analysis usually known as microspectrophotometry (MSP), pyrolysis-gas chromatography/mass spectrometry (Py-GC/MS) and some elemental analysis techniques such as X-ray diffraction, X-ray fluorescence and scanning electron microscopy with energy dispersive X-ray spectroscopy [10].

2.1.8. Pyrolysis-Gas Chromatography/Mass Spectroscopy (Py-GC/MS)

Py-GC/MS is a very powerful and sensitive technique that has shown great potential in the forensic examination of polymer traces. Polymer binders may degrade during pyrolysis through a number of mechanisms such as monomer reversion, side group elimination or random scission. As a result, smaller compound which can easily be identified are formed [20]. It is also possible to identify separate peaks belonging to minor components of the paint in the pyrogram. Py-GC is primarily used for the comparative analysis of the organic components of the paint. The pyrolysis patterns of two or more samples are visually compared noting the absence or presence of peaks, their relative peak intensities and their retention times [20]. When Py-GC is coupled with MS, and subsequent library searching, the technique can be used to identify pyrolysis products [21-23]. Py-

GC/MS has been applied to the examination of samples of forensic interests including paints, fibers and adhesives [24, 25]. It has also been applied to the classification and identification of automotive paints [25-28]. It has been suggested by some authors that Py-GC/MS maybe the method of choice for the classification of chemically similar paints [26, 29]. Py-GC/MS however has a major drawback in that it is a destructive method and its applicability depends on the paint type and the amount of paint sample [20]. It provides better result when used to analyze individual paint layers [20].

2.1.9. Elemental Analysis of Paints

Elemental analysis of paints involve the use of bulk and trace instrumental techniques such as X-ray fluorescence (XRF), X-ray diffraction (XRD), scanning electron microscopy coupled with energy dispersive spectroscopy (SEM-EDS), laser ablation-inductively coupled plasma mass spectroscopy (LA-ICP-MS), and particle induced X-ray emission spectroscopy (PIXE) [10]. The primary use of the elemental analysis technique is for the characterization of the inorganic components of the paint sample. SEM-EDS is a non-destructive analytical method that provides information about particle size and distribution, and morphology. SEM-EDS is very sensitive to mid-range atomic weight elements [30]. Unlike SEM-EDS, XRF is more sensitive to higher atomic weight elements and hence useful in identifying extenders in the paint sample [30]. X-ray diffraction (XRD) provides information about the crystallinity of the material and is useful for studying or identifying the inorganic components in paint pigments [31, 32]. (LA-ICP-MS) also has the ability to detect elements presents in the paint matrix. Unlike SEM-EDS, it has the potential for trace elemental analysis [32]. The technique can be considered non-destructive as the sample does not require manipulation and only a small amount of

material for the analysis. LA-ICP-MS has been used for the detection of elements in lower concentrations in automotive paints [32]. A limitation is the lack of standards which presents the greatest challenge to the use of this technique [32].

2.2. Vegetable Oils

2.2.1. Introduction

Vegetable oils are used in industrial, pharmaceutical, nutritional and cosmetic products including products such as cooking oils, margarine, salad dressings, food coatings, paints, plasticizers, glycerol, synthetic fibers, hand creams, shower gels, detergents and many more [33, 34]. The term vegetable is given to any oil that originates from a plant source, which includes oils such as corn oil, sunflower oil, coconut oil, hazelnut oil, palm nut oil, olive oils and many more [33, 34]. Vegetable oils are known as edible oils which have been subjected to several processes to remove undesirable constituents [35]. In order for edible oils to be suitable for human consumption, most are subjected to refining processes such as neutralization, bleaching and deodorization. Some edible oils such as extra virgin olive oils can be consumed directly without refining [35]. Since the composition of edible oils depend on the type of oil, edible oils are typically characterized by their physical and chemical properties [36]. Vegetable oils are comprised of a complex mixture of which triacylglycerols form the major component while the minor components are polyphenols, aldehydes, sterols and a variety of volatile organic compounds [34, 37]. The major components of edible oils are of great importance due to their nutritional values. Polyphenols, vitamins and other anti-oxidants which makes up the minor components of edible oils are responsible for other health benefits that are associated with consuming vegetable oils such as their ant-oxidant properties [38, 39].

Antioxidants play an important role in human health, contributing to a decrease in the occurrence of diseases such as atherosclerosis and bowl syndromes [38-41]. According to the published literature, low-density lipoproteins which are rich in cholesterol and cholesteryl esters can be potentially harmful to human health and can cause diseases such as atherosclerosis [42]. Such diseases have been found to be a direct result of modified oxidative forms of low-density lipoproteins [42]. The presence of anti-oxidants such as polyphenols inhibit the extent of the oxidation of low-density lipoproteins. Edible oils or vegetable oils are an important part of the Mediterranean diet. It has been reported by Visioli *et al* that the occurrence of coronary heart disease and certain cancers are found to be lower in the Mediterranean regions [40]. Investigation into the biological activities of hydroxytyrosol, and luteolin among others found in certain vegetable oils such as olive oils, indicates that these compounds offer protective properties against the oxidation of low-density lipoproteins [42]. Oils with high phenol contents such as olive oils have been found to be beneficial to the human diet. It can therefore be concluded that not only are vegetable oils of great importance from an economic stance, but they also provide great health benefit to humans.

2.2.2. Constituents of Edible Oils

The constituents of edible oils can be divided into two major groups; saponifiable which constitute triacylglycerols, free fatty acid and phosphatides, and the unsaponifiable which consist of hydrocarbons, fatty alcohols and so on. The percent of the unsaponifiable fraction accounts, in general, 0.5-1.5 % of the oils [35].

2.2.2.1. Saponifiable Fraction

About 98.5-99.5 % of oils is made up of the saponifiable fraction. The major components of the saponifiable fraction are the triacylglycerols and free fatty acids, although other fatty acid derivatives such as mono- and diacylglycerols, phospholipids, waxes and sterol esters are also found [35].

2.2.2.1.1. Triacylglycerols

Triacylglycerols comprise about 98–99 % of the oils. They are made up of esters derived from the union of glycerol (1, 2, 3-propanetriol) and fatty acids [35]. In general, the fatty acids at the central position of the glycerol molecule are unsaturated, although saturated acids can be found at this position when the total concentration of saturated fatty acids in the oil is very high.

2.2.2.1.2. Mono and Diacylglycerols

Jointly with triacylglycerols, edible oils also contain partial glycerols such as mono- and diacylglycerols, comprising about 0.2 and 1.3 % of total fatty acids, respectively. Their presence in olive oil is an index of low quality [43]. For this reason, an oil quality marker is generally based on the relative amount of mono- and diacylglycerols [35].

2.2.2.1.3. Free Fatty Acids

The amount of free fatty acids in an edible oil depends on the degree of hydrolysis of triacylglycerols as their composition varies according to the botanical variety of the oil, or, in the case of olive oil, according to the genetic variety, climatic conditions, fruit maturity and geographical origin of olives [44-47]. Fatty acids differ from one another based on their degree of unsaturation, the number of carbons in the hydrocarbon chain as well as the relative positions of the double bonds in the hydrocarbon chain [42]. Fatty acids can be grouped into two groups based on the presence or absence of double bonds in the

hydrocarbon chain. Fatty acids with only C-C single bonds are referred to as saturated, while those with at least one double bond are known as unsaturated. Unsaturated fatty acids have the double bond either in a cis or trans configuration. Trans fatty acids are known to cause health issues and have been reported to be linked to diseases such as diabetes and heart attacks [48].

2.2.2.1.4. Phospholipids

Phospholipids are usually found in small amounts in freshly produced olive oils (40–135 mg/kg) [49]. The most important phospholipids in olive oils are phosphatidylcholine, phosphatidylethanolamine and phosphatidylinositol [45].

2.2.2.1.5. Waxes

Waxes are esters of fatty alcohols containing fatty acids. The main waxes detected in olive oils have a high and even carbon number, in particular, C_{36}– C_{46} esters. Their amount is low, not exceeding 35 mg/100 g [45].

2.2.2.2. Unsaponifiable Fraction

The unsaponifiable fraction of edible oils contains different compounds which are not chemically related to fatty acids and they include compounds such as hydrocarbons, aliphatic and fatty alcohols, vitamins, volatile compounds and aromatic hydrocarbons [35].

2.2.2.2.1. Hydrocarbons

Perhaps the most important hydrocarbon found in both extra virgin and refined olive oils is squalene [35]. Squalene comprises between 2,500 and 9,250 µg/g and has been found in olive oils in larger amounts compared to those found in other edible oils, which ranged from 16 to 370 µg/g. Other hydrocarbons also present in edible oils such as olive oils are C_{14}– C_{30} n-alkanes, some n-alkenes and terpene hydrocarbons such as a-farnesene.

The relative concentration of these hydrocarbons is ranges anywhere from 150–200 µg/g [50].

2.2.2.2.2. Aliphatic and Fatty Alcohols

Both aliphatic and fatty alcohols are minor components of edible oils and are also important constituents of edible oils, and, in the case of olive oil, they can be used to distinguish different olive oil types. Fatty alcohols can be linear (aliphatic) or nonlinear in structure [35]. Other alcohols, such as diterpene alcohols or acyclic diterpene alcohols are also found in olive oils. Aliphatic alcohols only have a linear structure. They are the precursors for the formation of waxes [35].

2.2.2.2.3. Vitamins

Vitamins play an important role in edible oils as they contribute to the stability of edible oils by protecting them from oxidation [51] thereby preventing lipid peroxidation in biological membranes [35]. Vitamins also provide antioxidants protection. Some of the vitamins presents in edible oils are Ts and T_3s [35]. While Ts is found in all edible oils, T_3s is found only in palm oil [52]. The relative concentrations of Ts and T_3s varies according to the type of oil [35].

2.2.2.2.4. Volatile and aromatic compounds

Volatile and aromatic compounds are primarily responsible for the aroma and flavor of most edible oils such as the olive oils [35]. There are more than one hundred components directly related to the aroma and flavor in edible oils, such as hydrocarbons, alcohols, aldehydes, esters, phenols, terpenes and derivatives of furan [53, 54]. Alcohols formed in the olive from polyunsaturated fatty acids and 6-carbon aldehydes are the most important components of olive oil aroma.

2.2.3. Analysis of Edible Oils

Over the years, a number of analytical techniques and methods have been developed for the identification and classification of edible oils [42]. Among these techniques are titrimetric techniques, which are commonly used for the determination of peroxide values in edible oils, chromatographic techniques such as high pressure liquid chromatography, gas chromatography, thin layer chromatography, mass spectrometry, and infrared, near infrared, Raman, nuclear magnetic resonance and ultraviolet-visible spectroscopy [42].

2.2.4. Adulteration of Edible Oils

Adulteration of edible oils involves the replacement or mixing of higher quality and higher cost edible oils such as extra-virgin olive oil, cocoa butter and milk fat with a lower-cost and lower quality edible oil such as corn oil, peanut oil, canola, sunflower, and soybean. Extra virgin olive oils are typically adulterated with corn or sunflower oil due to their similar composition and high degree of similarity between their IR and Raman spectra. Edible oils such as extra-virgin olive oils have high sensory qualities and great nutritional benefits and therefore they are the most likely targets for adulteration [55]. Olive oil is extracted by mechanical means from the first pressing of the olives and they do not undergo further processing [56] unlike lower grade oils which undergo several chemical treatments which usually results in the removal of most of their desirable nutritional constituents [56]. Adulteration of edible oils such as olive oils pose a serious problem for regulatory agencies such as the International Olive Council (IOC), edible oil suppliers and ultimately the consumer. When an edible oil such as olive oil is adulterated with peanut oil, for example, the nutritional value of the olive oil is reduced. This can pose

potential health issues due to the allergic reaction to peanut proteins [57]. The IOC exist as an international organization to provide quality standards for different grades of olive oils. Despite these legal standards, fraudulent activities in the olive oil industry continue [57]. Blended oils can be prepared for a specific purpose or product but it becomes an issue when the allowable mixture proportions are not followed or if the blend is marketed as genuine [58]. Turkey is one of the major producers of olive oil in the world market and rapeseed, sunflower and corn oils which have lower market prices are some of the most commonly found adulterants in olive oils [55].

References

1. Bayliss, D.A.a.K.A.C., *Paints and paint coatings, in Steelwork Corrosion Control.* Elsevier Science Publishers, Ltd.: Essex, 2002: p. 35-61.

2. Bentley, J., *Composition, manufacture and use of paint, in Forensic Examination of Glass and Paint: Analysis and Interpretation*, ed. B. Caddy. 2001, London: Taylor and Francis.

3. Ryland, S.G.a.E.M.S., *Analysis of paint evidence, in Forensic Chemistry Handbook*, ed. L. Kobilinsky. 2011, New Jersey: John Wiley & Sons, Inc.

4. Bender, L., *Chemistry/Trace/Paint and Coating: Overview, in Encyclopedia of Forensic Sciences*, ed. J.A.S.a.P.J. Saukko. 2013, London: Academic Press: . 245-249.

5. Wicks, Z.W.a.F.N.J., *Coatings, in Kirk-Othmer Encyclopedia of Chemical Technology*, ed. J.I.K.a.M. Howe-Grant. 2000, New Jersey: John Wiley & Sons, Inc.

6. Talbert, R., *Paint components, in Paint Technology Handbook*. 2007, Florida: Taylor and Francis Group.

7. Stoye, W.F.a.D., *Paints, coatings and solvents*. 2008: John Wiley & Sons.

8. Diejen, E.C.M.v., *Development of a fourier transform infrared database of paint binders.* Nederlands Forensisch Instituut, 2016.

9. Weldon, D.G., *Failure analysis of paints and coatings*. 2009: John Wiley & Sons.

10. Nordmann, V., *Spectroscopic and imaging methods in forensic investigations of modern paints and coatings.* 2016.

11. Turner, G.P.A., *Introduction to Paint Chemistry and Principles of Paint Technology*. 1988, London: Chapman and Hall Ltd.

12. Streitberger, H., et al, *Paints and coatings, 3. Paint Systems, in Ullmann's Encyclopedia of Industrial Chemistry*. 2014: John Wiley & Sons, Inc.

13. (Ed.), G.F., *Automotive Paints and Coatings*. 1995, New York, NY: VCH Publications.

14. White, C.G., *Variable selection to improve classification in structure-activity studies and spectroscopic analysis*. 2016, Oklahoma State University.

15. Bender, L., *Chemistry/Trace/Paint and Coating: Automotive Paint, in Encyclopedia of Forensic Sciences*. 2013, London: Academic Press.

16. Saferstein, R., *Criminalistics: An Introduction to Forensic Science*. 2001, New Jersey: Prentice Hall.

17. Houck, M.M.a.J.A.S., *Fundamentals of Forensic Science*. 2nd ed ed. 2010, Massachusetts: Academic Press.

18. Muehlethaler, C., L. Gueissaz, and G. Massonnet, *Chemistry/Trace/Paint and Coating: Forensic Paint Analysis, in Encyclopedia of Forensic Sciences*. 2013, London: Academic Press. 265-272.

19. S. Ryland, T.J., K. P. Kirkbride, *Current Trends in Forensic Paint Examination.* Forensic Sci. Rev, 2006. **18**: p. 97-117.

20. Janina Zi˛eba-Palus, G.Z., Jakub M. Milczarek, Paweł Ko´scielniak, *Pyrolysis-gas chromatography/mass spectrometry analysis as a useful tool in forensic examination of automotive paint traces.* Journal of Chromatography A, , 2007. **1179** p. 41–46.

21. Blackledge, R.D., Forensic Sci. Rev. 4, 1992. **2**.

22. D.T. Burns, K.P.D., Anal. Chim. Acta 539, 2005. **145**.

23. J.M. Milczarek, J.Z.e.-P., P. Ko´scielniak, , Problems Forensic Sci. LXI, 2005. **7**.
24. Challinor, J.M., Forensic Sci. Int. 21 1983. **269**.

25. J.M. Challinor, i.T.P.W.E., *Applied Pyrolysis Handbook*. 2004, Boca Raton, FL: CRC Press. 175.

26. A.R. Cassista, P.M.L.S., J., Can. Forensic Sci. 27 1994. **209**.

27. T.P. Wampler, G.A.B., W.J. Simonsick, J. Anal. Appl. Pyrol. 40/41, 1997. **79**.

28. Fukuda, K., Forensic Sci. Int. 29, 1985. **227**.

29. D.T. Burns, K.P.D., Anal. Chim. Acta 539, 2005. **157**.

30. Maric, M., *Chemical Characterisation and Classification of Forensic Trace Evidence*. 2014, Curtin University.

31. Henson, M.L.a.T.A.J., *Scanning electron microscopy and energy dispersive X-ray spectrometry (SEM/EDS) for the forensic examination of paints and coatings, in Forensic Examination of Glass and Paint Analysis and Interpretation*. 2001, London: Taylor & Francis.

32. Almirall, A.L.H.J.R., *Trace elemental analysis of automotive paints by laser ablation–inductively coupled plasma–mass spectrometry (LA–ICP–MS).* Anal Bioanal Chem, 2013. **376**: p. 1265–1271.

33. Gunstone, F.D., *Vegetable Oils in Food Technology; Composition, Properties and Uses*. 2000, USA and Canada: Blackwell Publishing, Ltd. 33.

34. Gunstone, F.D., *The Chemistry of Oils and Fats; Sources, Composition, Properties and Uses*. 2004, USA and Canada, : Blackwell Publishing, Ltd.

35. García, M.J.L., *Characterization and Authentication of Olive and Other Vegetable Oils.* pringer Theses, 2012.

36. JB, R., *Vegetable oils and fats. In: Rossell JB, Pritchard JLR (eds) Analysis of oilseeds, fats and fatty foods*. 1991, London: Elsevier Science. 261–327.

37. Gunstone, F.D., and Norris, F.A., *in "Lipids in Foods; Chemistry, Biochemistry and Technology".* Pergamon Press, 1983: p. 1-8, 29-32.

38. Manna, C., Galletti, P., Cucciola, V., Montedoro, G., and Zappia, V. , *Olive oil hydroxytyrosol protects human erythrocytes against oxidative damages.* J. Nutr. Biochem, 1999. **10**: p. 159-165.

39. Giovanni, C., Straface, E., Modesti, D., Coni, E., Cantafora, A., De Vincenzi, M.,Malorni, W., and Masella, R, *Biochemical and Molecular Action of Nutrients: Tyrosol, the Major Olive Oil Biophenol, Protects Against Oxidized-LDL-Induced Injury in Caco-2 Cells.* J. Nutr. , 1999. **129**: p. 1269-1277.

40. Visioli, F., and Galli, C. , *Olive Oil Phenols and Their Potential Effects on Human Health.* J. Agric. Food Chem, 1998. **46**: p. 4292-4296.

41. Lipworth, L., Martínez, M.E., Angell, J., Hsieh, C., and Trichopoulos, D. , *Review: Olive oil and Human Cancer: An Assessment of the Evidence.* Prev. Med., 1997. **26**: p. 181-190.

42. Retief, L., *Analysis of vegetable oils, seeds and beans by TGA and NMR spectroscopy*. 2011, University of Stellenbosch.

43. Mariani C, F.E., Riv Ital Sost Grasse, 1985. **62**: p. 3.

44. Aparicio R, A.V., Morales MT, Grasas Aceites 1994. **45**: p. 241.

45. D, B., *Olive oil In: Gunstone FD (ed) Vegetable oils in food technology*. CRC Press,. 2002, Oxford: Blackwell Publishing Ltd. 244–277.

46. D'Imperio M, D.G., Alfa M, Mannina L, Segre AL Food Chem 2007. **102**: p. 956.

47. Torres MM, M.D., Food Chem 2006. **96**: p. 507.

48. *The controversy over trans fatty acids: Effects early in life. Food and Chemical Toxicology*. Elsevier Ltd. Vol. 46 (12. 2008. 3571-3579.

49. Tiscornia E, F.N., Evangelisti F, Riv Ital Sost Grasse 1982. **59**: p. 519.

50. Lanzon A, A.T., Cert A, Gracian J, J Am Oil Chem Soc 1994. **71**: p. 285.

51. Blekas G, T.M., Bouskou D Food Chem 1995. **52**: p. 289.

52. Choo YM, Y.S., Ooi CK, Ma AN, Goh SH, Ong ASH J Am Oil Chem Soc 1996. **73**: p. 599.

53. Morales MT, T.M., *El In: Aparicio R, Harwood J (eds) Mundi-Prensa, papel de los compuestos volátiles y polifenoles en la calidad sensorial del aceite de oliva.* 2003, Madrid, Spain: Manual del aceite de oliva. Ed. 381–442.

54. Morales MT, A.R., J Am Oil Chem Soc, 1999. **76**: p. 295.

55. Gozde Gurdeniz, B.O., *Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data.* Food Chemistry, 2009. **116**: p. 519–525.

56. Yang, H., & Irudayaraj, J., *Comparison of near-infrared, Fourier transforminfrared, and Fourier transform-Raman methods for determining olive pomace oil adulteration in extra virgin olive oil.* Journal of the American Oil Chemists Society, 2001. **78**: p. 889–895.

57. Nick Vanstone, A.M., Perry Martos, and Suresh Neethirajan, *Detection of the adulteration of extra virgin olive oil by near-infrared spectroscopy and chemometric techniques.* Food Quality and Safety 2018. **2**: p. 189–198.

58. Ulberth, F., Buchgraber, M., *Authenticity of fats and oils.* European Journal of Lipid Science Technology, 2000. **102**: p. 687–694.

CHAPTER III

PATTERN RECOGNITION

3.1. Introduction

Many relationships in spectral data cannot be expressed in quantitative terms. These relationships are better expressed in terms of similarity and/or dissimilarity between diverse groups of spectra. The tasks which confront an analytical chemist when investigating these types of relationships are two-fold. First, can a useful structure based on distinct categories in the data be discerned? Second, can a sample as represented by its spectrum be classified into one of these categories for the prediction of a sample property? To develop a mathematic relation suitable for identifying and isolating classes within multivariate spectral data, analytical chemists have turned to pattern recognition which is a collection of mathematical, statistical, and numerical techniques based on formal logic and designed to solve the class membership problem [1-3].

Pattern recognition has its origin in the field of image and signal processing where techniques were developed to categorize samples on the basis of regularities in their observed data. In a typical pattern recognition study, samples are classified according to a specific property (which is often difficult to measure directly) using spectral 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 this property in samples that are not part of the original training set. The property in question may be the geographic origin of raw materials used to formulate a pharmaceutical tablet, and the measurements are the absorbances at specific wavelengths obtained directly from an infrared spectrum of the tablet [4]. The idea of an indirect relation between a spectrum and the property of a substance, first proposed by Hirschfeld and Martens [5-7], is plausible because the physical and chemical properties of many substances (such as pharmaceutical tablets) are governed by their chemical composition.

The set of samples for which the property of interest and measurements are known is called the training set. The set of measurements that describe each sample in the training set is called a pattern. The determination of the property of interest by assigning a sample to its respective category is called recognition – hence the term ‘‘pattern recognition’’ – because recognition is accomplished using the set of measurements that characterize each sample in the data set.

For pattern recognition analysis, each sample is represented as a data vector, x = (x1, x2, x3, ….. xj……xn) where component xj is a measurement, such as the absorbance at the jth wavelength. In other words, each sample can be considered as a point in an n-dimensional measurement space. The dimensionality of the measurement space corresponds to the number of measurements that are available for each sample. A basic assumption is that the distance between pairs of points in this measurement space is inversely related to the degree of similarity between the corresponding samples. Therefore, points representing samples from one class will cluster in a limited region of the

measurement space distant from the points corresponding to the other class. Pattern recognition is a set of methods for investigating data represented in this manner, in order to assess its overall structure, which is defined as the overall relationship of each sample to every other sample in the data set.

In this chapter, three major subdivisions of pattern recognition methods are discussed: (1) mapping and display, (2) clustering, and (3) feature selection. A summary of the techniques used in the studies described in this dissertation are included in this chapter. Special emphasis in the discussion of these techniques is placed on their problems in spectral pattern recognition.

3.2. Principal Component Analysis

Principal component analysis (PCA) is the best known of the unsupervised pattern recognition techniques and is the most widely used multivariate analysis method in science and engineering [8]. The overall goal of PCA is dimensionality reduction of a data set, while simultaneously retaining the relevant information present in the data. Dimensionality reduction or data compression is possible because chemical data sets are often redundant. That is, chemical data sets are not information rich. For this reason, PCA is often used as a mapping and display technique for exploratory data analysis. Dimensionality reduction is achieved by transforming the original measurements variables of the data matrix (i.e., the columns of the matrix) into principal components. Each principal component is a linear combination of the original measurement variables. Summarizing the information present in a spectral data set may require only two or three principal components.

Dimensionality reduction is possible using PCA because of correlations between the measurement variables. Consider a set of samples characterized by two measurement variables, X_1 and X_2 see Figure 3.1. X_1 appears to be correlated to X_2 because fixing the values of X_1 limit the range of values for X_2 in the space defined by these two variables. If these two measurement variables were uncorrelated, the entire enclosed rectangular area shown in Figure 3.1 would be populated by data points. Because information is defined as the scatter of points in a measurement space, it is evident that correlations between the measurement variables decrease the information content of this space. The data points, which are restricted to a small region of the measurement space due to correlations among the variables, could even reside in a subspace if the measurement variables are highly correlated.

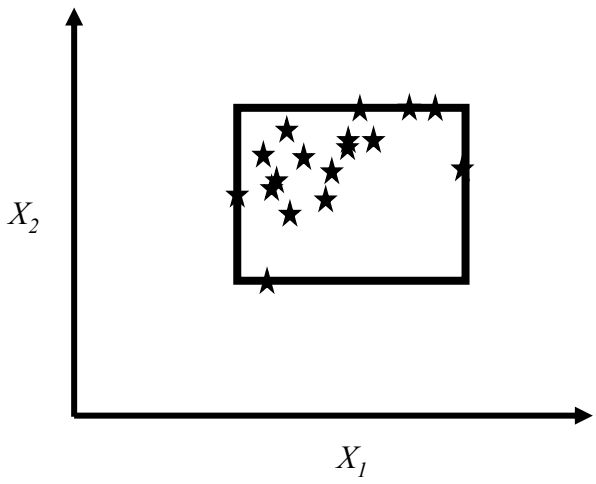

Figure 3.1. Seventeen hypothetical samples projected onto a 2-D space described by the measurements variables X_1 and X_2. A, B, C, and D defines the smallest and largest values of X_1 and X_2. (Adapted from NBD J. Res., 1985, 190(6), 465-476).

Collinearity between measurement variables is a strong indication that a set of basis vectors can be obtained that are better at conveying the information present in the data than

axes defined by the original measurements. This new basis set linked to variation can be used to develop a new coordinate system for displaying the data. The variance-based axes of this new coordinate system are defined by the principal components of the data. Determining the direction of largest variation in the original measurement (pattern) space and modeling it by a line fitted through the data points using linear least squares leads to the formation of the largest or first principal component of the data as shown in Figure 3.2. The second largest principal component of the data 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 passes through the center of the data and is orthogonal to the first and second largest principal components and so forth. Because each principal component is orthogonal to the other, different sources of information present in the data are captured by each principal component.

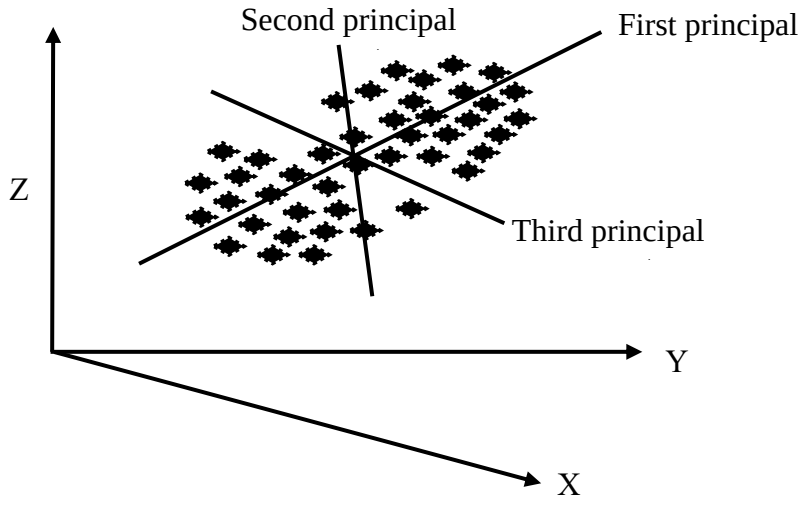

Figure 3.2. Graphical representation of principal component axes. The third principal component described only the noise in the data.

PCA involves performing an eigenvalue decomposition of the data matrix **X** (n x p) usually after mean centering or auto-scaling of the data has been performed using the singular value decomposition (SVD) algorithm [9]. Each principal component is associated with an eigenvalue (λ) which reveals the degree of variation in the data captured by the principal component. The first principal component has the largest eigenvalue followed by the second principal component and so forth. PCA decomposes **X** (n x p) into a score matrix **T** (n x f), loading matrix **P** (f x p), and residual matrix **E** (n x p), where n is the number of spectra in the data set, p is the number of points or features in each spectrum, and f is the number of principal components necessary to represent the spectral data. Usually, f is smaller than p due to correlations among the measurement variables. The decomposition of **X** is shown in Equation 3.1, where 1 is a column vector (n x 1) of ones and $\mathbf{x}_{mean}$ is a (1 x p) row vector representing mean of the data.

$$X = 1x_{mean} + TP' + E \tag{3.1}$$

The relationship between the original measurement variables (wavenumbers) and the principal components of the data are provided by the loading matrix and the coordinates of the samples in this principal components space are defined by the score matrix. The score matrix and the loading matrix describe the signal in the data, whereas the residual matrix represents the noise. By plotting the columns of the score matrix against each other, a representation of the distribution of the spectra in the p-dimensional multivariate space can be obtained. The number of principal components used to represent the spectral data is equal to the number of columns in the score matrix, which often is only two or three because of correlations among the measurements variables. By analyzing a data set using

PC (score) plots, it is possible to find relationships in the data, that is, to detect similarities and differences among groups of samples and to identify outliers present in the data.

3.3. Cluster Analysis

The objective of cluster analysis [10] is to uncover class structure in data. Cluster analysis is encountered in many fields, e.g., biology, geology, and geochemistry, under such names as unsupervised pattern recognition and numerical taxonomy. Clustering methods can be divided into three categories: hierarchical, object-functional, and graph theoretical. In this chapter, the focus is hierarchical clustering as this method is the most popular.

For cluster analysis, each sample is treated as a point in an n-dimensional measurement space. The coordinate axes of this space are defined by the measurements used to characterize the samples. Cluster analysis assesses the similarity between samples by measuring the distances between the points in the measurement space. Samples that are similar will lie close to one another, whereas dissimilar samples are distant from each other. For the studies described in this dissertation, the measurements used to characterize each sample are continuous variables, and the Euclidean distance is the distance metric used to assess similarity, because inter-point distances between samples are computed directly. However, there is a problem with using the Euclidean distance - the so-called scaling effect. It arises from inadvertent weighing of the variables in the analysis that can occur due to differences in the magnitude among the measurement variables. For example, consider a data set where each sample is described by two variables: the concentration of Na and the concentration of K as measured by an ion selective electrode. The concentration of Na varies from 50 to 500ppm, whereas the concentration of K in the same samples varies from

5 to 50ppm. A 10% change in the Na concentration will have a greater effect on Euclidean distance than a 10% change in K concentration. The influence of variable scaling on the Euclidean distance can be eliminated by auto-scaling the data, which involves standardizing the measurement variables using the standard deviation, so each variable has a mean of zero and a standard deviation of 1. Thus, a 10% change in K concentration has the same effect on the Euclidean distance as a 10% change in Na concentration when the data is auto-scaled. Clearly, autoscaling ensures that each measurement variable has an equal weight in the analysis. For cluster analysis, it is generally best to autoscale the data, since similarity is then determined by a majority vote of the measurement variables.

Hierarchical cluster analysis [10] 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. The starting point for a hierarchical clustering experiment is the similarity matrix. This matrix is formed by first computing the distances between all pairs of points in the data set. Each distance is converted into a similarity value using Equation 3.2.

$$S_{ik} = 1 - \frac{d_{ik}}{d_{max}} \tag{3.2}$$

where s_{ik} is the similarity between samples i and k which varies from 0 to 1, d_{ik} is the Euclidean distance between samples i and k, and d_{max} is the longest distance between two samples in the data set which corresponds to the two most dissimilar samples. The similarity values are organized in the form of a table or matrix which is then scanned to identify the most similar point pair (i.e., largest value). The two samples that comprise the point pair are combined to form a new point located midway between the two original points. Both the rows and columns corresponding to the old data points are removed from the matrix. The similarity matrix is then recomputed for the data set. In other words, the

matrix is updated to include information about the similarity between the new point and every other point in the data set. The new nearest point pair is identified, and combined to form a single point. This process is repeated until all points have been linked.

There are a variety of ways to compute the distances between data points and clusters in hierarchical clustering (see Figure 3.3). The nearest linkage method assesses similarity between a point and a cluster of points by measuring the distance to the closest point in the cluster. The farthest linkage method assesses similarity by measuring the distance to the point furthest away in the cluster. Mean linkage assesses the similarity by computing the distances between all point pairs where a member of each pair belongs to the cluster. The mean of these distances is used to compute the similarity between the data point and the cluster.

The results of a hierarchical clustering study are usually displayed as a dendogram, which is a tree shaped map of the inter-sample distances in the data set. The dendogram shows the merging of samples into clusters at various stages of the analysis and the similarities at which the clusters merge, with the clustering displayed hierarchically. Interpretation of the results is intuitive, which is the major reason for the popularity of these methods.

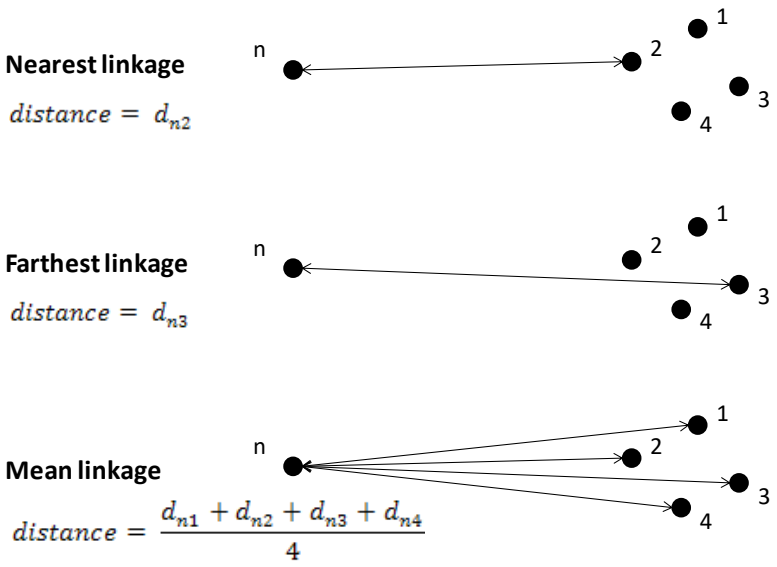

Figure 3.3. The distance between a data cluster and a point using (a) nearest linkage, (b) farthest linkage, and (c) mean linkage.

3.4. Genetic Algorithm for Pattern Recognition Analysis

Problems often arise when applying pattern recognition methods to chemical data. Classification success rates may vary with the pattern recognition method employed. Unfavorable classification results can be obtained despite a linearly separable training set. Automation of these techniques for the solution of a general class of pattern recognition methods is often difficult [11].

The basic premise underlying the pattern recognition methodology used in the studies discussed in this dissertation is that all classification methods work well when the problem is simple. By identifying the appropriate features, a "hard" problem can be reduced to a "simple" problem. Therefore, the goal is feature selection, in order to increase the signal to noise ratio of the data by discarding measurements that are not characteristic of the source profile of the classes in the dataset. To ensure identification of all relevant features,

it is best that a multivariate approach to feature selection is employed. The approach should also take into account the existence of redundancies in the data.

The approach to feature selection described in this chapter is based on a simple idea - identify a set of measurement variables that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these features is about differences between classes in the data set. Using this approach to feature selection, an eigenvector projection of the data is developed that discriminates classes in the dataset by maximizing the ratio of between- to within-group variance. This approach has a number of advantages. It avoids overly complicated solutions, which do not perform as well on the prediction set because of over-fitting. Although a principal component plot is not a sharp knife for discrimination, if we have a principal component plot that shows clustering, then our experience is that we will be able to predict robustly using this set of descriptors. Furthermore, the principal component plot displays the variability between large numbers of samples and show the major clustering trends present in the data. The user can visually identify the presence of confounding relationships in the data, thereby gaining insight into how a decision for classification is made.

In the studies described in this dissertation, the approach to feature selection described in the previous paragraph is implemented using a genetic algorithm. Genetic algorithms were developed by John Holland [12] to mimic the process of evolution. Genetic algorithms have advantages over conventional optimization search algorithms. They operate on the entire parameter set and simultaneously consider many points in the solution space unlike conventional methods that manipulate the parameters independently,

which can be a problem if an object function is overly sensitive to one parameter as the optimization function would tend to focus its effort on the troublesome parameter at the expense of the other parameters [11]. As genetic algorithms consider simultaneously many points in the search space, more of the response surface is probed reducing the chance of convergence to a local minimum since genetic algorithms utilize parallelism. A large number of potential solutions are considered simultaneously. Genetic algorithms require only information about the fitness of potential solutions. They make no assumptions about the topography of the solution surface and are not impacted by discontinuities or singularities that are disruptive to derivative and simplex optimization based methods [11]. By adjusting the parameters of the GA, it can be tailored to a particular application.

The genetic algorithm for feature selection used in the studies described in this dissertation (designated as the pattern recognition GA) [13-17] identifies a set of features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. The principal component plot of each feature subset, which is used by the fitness function of the pattern recognition GA acts as an embedded information filter. Sets of spectral features or wavelengths are selected based on their principal component plots, with a good principal component plot generated by features whose variance or information is primarily about differences between the classes or groups. This restricts the search to feature subsets of this type, thereby significantly reducing the size of the search space. In addition, the pattern recognition GA is able to focus on those classes and/or samples that are difficult to classify by boosting their weights over successive generations using a perceptron to adjust the values of the class and sample weights. Samples or classes that are consistently classified correctly are not as heavily weighted in

the analysis as those samples or classes that are difficult to classify. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection.

To track and score the PC plots generated by the pattern recognition GA in each generation, class and sample weights are computed (see Equations 3.3 and 3.4) where CW(c) is the weight of class c (with c varying from 1 to the total number of classes in the data set), and SWc(s) is the weight of samples in class c. Class weights sum to 100, and the sample weights for samples from a particular class sum to a value equal to its class weight.

$$\mathrm{CW(c)}=100\frac{\mathrm{CW(c)}}{\sum_{\mathrm{c}}\mathrm{CW\ (c)}} \tag{3.3}$$

$$SW(s) = CW(c)\frac{SW(s)}{\sum_{s\in c} SW(s)} \tag{3.4}$$

Each PC plot generated for each feature subset after extracting the features from its chromosome is scored using the K-nearest neighbor classification algorithm [18]. For each sample in the training set, Euclidean distances are computed between it and the other samples that are represented as points in the principal component (PC) plot. These distances are arranged from smallest to the largest. A poll is then taken of the sample's K_c nearest neighbors. For the most rigorous classification of the data, K_c is assigned a value corresponding to the number of samples in the class to which it is a member. The number of K_c nearest neighbors with the same class label as the data point (i.e., sample) in question, the so-called sample hit count, SHC(s), is computed ($0 < SHC(s) < Kc$) for sample. It is then a simple matter to score the PC plot (see Equation 3.5). First, the contribution to the

overall fitness score by each sample in class 1 is computed, with SHC for each sample comprising the class divided by K_c and multiplied by SW(s), and then summing up the product for the samples comprising the class to yield the contribution of this class to the overall fitness score. This same calculation is repeated for the other samples with the fitness score of each class summed to yield the overall fitness score, *F(d)*.

$$F(d) = \sum_{c} \sum_{s \in c} \frac{1}{K_c} \times SHC(s) \times SW(s) \tag{3.5}$$

The fitness function of the pattern recognition GA is able to focus on those samples and/or classes that are difficult to classify by boosting their weights over successive generations. To boost the sample and class weights, it is necessary to compute both the sample hit rate (SHR), which is the mean value of SHC/Kc over all feature subsets (which comprise the population of solutions) generated in a particular generation (see Equation 3.6), and the class-hit rate (CHR), which is the mean sample hit rate of all samples in a particular class (see Equation 3.6). ϕ in Equation 3.6 is the number of chromosomes in the population, and AVG in Equation 3.7 refers to the average or mean value.

$$SHR(s) = \frac{1}{\emptyset} \sum_{i=1}^{\emptyset} \frac{SHC_i(s)}{K_c} \tag{3.6}$$

$$CHR_g(c) = AVG\left(SHR_g(s): \forall_{s \in c}\right) \tag{3.7}$$

During each generation, class and sample weights are adjusted by a perceptron algorithm (see Equations 3.8 and 3.9) with the momentum, P, set by the user and with g + 1 being the current generation and g being the previous generation. Classes with a lower class hit rate are boosted more heavily than those classes that score well.

$$CW_{g+1}(s) = CW_g(s) + P(1 - CHR_g(s))) \tag{3.8}$$

$$SW_{g+1}(s) = SW_g(s) + P(1 - CHR_g(s))) \tag{3.9}$$

Boosting is crucial to the successful operation of the pattern recognition GA as it modifies the fitness landscape by adjusting the values of both the class and sample weights in each generation. This allows the pattern recognition GA to learn and assists it to obviate the problem of premature convergence to a local optimum. Thus, the fitness function of the pattern recognition GA is changing as the population evolves towards an optimal solution.

References

1. Gonzalez, J.T.T.a.R.C., *Pattern Recognition Principles*. 1974, MA: Addison Wesley Publishing Company, Reading.

2. Wold, B.R.K.a.S., *Pattern Recognition in Chemistry, in Classification, Pattern Recognition and Reduction of Dimensionality*. 1982, North Holland, Amsterdam.

3. B. R. Kowalski, a.C.F.B., *Pattern Recognition. A Powerful Approach for Interpreting Chemical Data.* J. Am. Chem. Soc., 1972. **94**: p. 5632-5640.

4. B. Rostaing, P.D., and Y. Roche, *Automation of Identification Process of Excipients and Pharmaceuticals. Fourier Transform Infrared Spectroscopy and Near-Infrared Diffuse Reflectance Spectroscopy Applied to Rapid in Process Control of Pharmaceutical Raw Material.* S. T. P. Pharma, 1988. **4(6)**: p. 509-515.

5. D. E. Honigs, G.M.H., and T. Hirschfeld, *A new Method for Obtaining Individual Component Spectra from those of Complex Mixtures.* Appl. Spec., 1984. **38**: p. 317-322.

6. Martens, H., *in Food Research and Data Analysis*. 1983, London: Applied Science Publisher. 5-38.

7. Martens, M.M.a.H., *Near Infrared Reflectance Determination of Sensory Quality of Peas.* Appl. Spec., 1986. **40**: p. 303-310.

8. I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, NY, 1986.

9. G. Golub, C.V.L., *Matrix Computations*. 1971, Baltimore: Johns Hopkins University Press.

10. Rousseeuw, L.K.a.P.J., *Finding Groups in Data – An Introduction to Cluster Analysis*. 1990, NY John Wiley & Sons.

11. White, C.G., *Variable selection to improve classification in structure-activity studies and spectroscopic analysis*. 2016, Oklahoma State University.

12. Holland, J.H., *Adaptation in Natural and Artificial Systems*. 6th ed. 2001, Cambridge, MA: MIT Press.

13. B.K. Lavine, C.G.W., T. Ding, M.M. Gaye, D.E. Clemmer, *Wavelet based classification of MALDI-IMS-MS spectra of serum N-Linked glycans from normal controls and patients diagnosed with Barrett's esophagus, high grade dysplasia, and esophageal adenocarcinoma.* Chemolab, 2018. **176** p. 74-81.

14. M. M. Gaye, T.D., H. Shion, A. Hussein, Y. HU, S. Zhou, Z. T. Hammoud, B. K. Lavine, Y. Mechref, J. C. Gelbler, D. E. Clemmer, *Delineation of disease phenotypes associated with esophageal adencarcinoma by MALDI-IMS-MS analysis of serum N-linked glycans.* Analyst, 2017. **142** p. 1525-1535.

15. A. Fasasi, N.M., R.-I. Stoian, C. White, M. Allen, Mark P. Sandercock, B. K. Lavine, *Pattern recognition assisted infrared library searching of automotive clear coat.* Appl. Spec., 2015. **69**: p. 84-94.

16. B. K. Lavine, A.F., N. Mirjankar, M. Sandercock, S. D. Brown, *Search prefilters for mid-IR spectra of clear coat automotive paint smears using stacked and linear classifiers.* J. Chemom., 2014. **28**: p. 385-39.

17. B. K. Lavine, C.W., C. Matthew Sundling, Curt Breneman, *Odor-structure relationship studies of tetralin and indan musks.* Chem. Senses, 2012. **37**: p. 723-736.

18. J. Tou, R.C.G., *Pattern Recognition Principles*. 1974, Reading, MA.: Addison-Wesley Publishing Company.

CHAPTER IV

TRANSMISSION INFRARED MICROSCOPY FOR THE FORENSIC EXAMINATION OF AUTOMOTIVE PAINT – SAMPLE PREPARATION

4.1. Introduction

Modern automotive paint consists of an e-coat, surfacer-primer and color coat layers protected by a thick clear coat layer [1]. All four paint layers contain binders, and the e-coat, surfacer-primer and color coat layers contain pigments and fillers. Automotive assembly plants use a unique combination of pigments and binders in each layer of paint which allows the forensic paint examiner to determine the make and model of a vehicle from an intact multilayered paint chip, which is often the only evidence recovered from a crime scene of a vehicle related fatality such as a hit-and-run or collision. Studies [2, 3] performed over 40 years ago by the Royal Canadian Mounted Police (RCMP) showed that vehicles could be differentiated by comparing the color, layer sequence and chemical composition of each layer in a manufacturer's automotive paint system.

The chemical composition of an automotive paint sample in forensic laboratories is typically determined using Fourier transform infrared (FTIR) spectroscopy [4]. Each layer is hand-sectioned by a scalpel, placed in a high pressure diamond anvil cell, and the corresponding infrared (IR) transmission spectrum of each layer is compared to paint from

a suspect's vehicle. As there are usually no witnesses at the crime scene, police are often unable to develop a suspect. In these situations, the IR spectrum of each layer of the paint chip can be matched to a particular make and model of a vehicle using an automotive paint database such as the paint data query (PDQ) database [3, 5-8]. However, the amount of time required to analyze a single automotive paint sample and perform a PDQ search can be lengthy. Furthermore, sampling too close to the boundary by using a scalpel to separate adjacent paint layers can produce an IR spectrum that is a mixture of two layers. Not having a "pure" spectrum of each layer prevents a meaningful comparison between each paint layer or, in the situation of searching an automotive paint database, will prevent the scientist from developing an accurate hit list of potential vehicles.

One way to decrease the time necessary for data collection (compared to the current method of hand sectioning and analyzing each layer separately by FTIR) is to collect IR data from all layers in a single analysis by scanning across a cross-sectioned paint sample using a FTIR imaging microscope equipped with an imaging detector. A complete scan can be performed in less than an hour. After the data has been collected, it can undergo decatenation using multivariate curve resolution [9, 10] to obtain the "pure" IR spectrum of each layer. This approach, not only eliminates the need to analyze each layer separately, but also will ensure that the final spectrum of each layer is "pure" and not a mixture. Minimizing the probability of collecting a mixed spectrum will result in a considerable time savings as well as objectively ensuring that only "pure" spectra from each layer are collected and used in subsequent searches of the PDQ database which will reduce the number of hits in a library search.

To collect IR spectra from a paint chip using an IR microscope, it is common practice to cast the chip in a block of epoxy. A microtome is used to cross section the paint chip to reveal the individual layers. Ideally, there should be no spectral interference from the epoxy. However epoxy adhesives when used for casting can prevent accurate library matching of both the original and refinished automotive paints by infiltrating the layers of the automotive paint [11].

Ideally, one would like to cross section each paint chip without using epoxy to cast the sample. This would make sample preparation faster and easier and more importantly, eliminate interfering peaks from the embedding media that otherwise could be present in the IR spectra of the OEM paint layers. In the study described in this chapter, it is demonstrate that automotive paint chips can be cross sectioned without casting the sample in epoxy. An IR image map of each paint chip was collected from the cross sectioned paint sample, and a transit (i.e., line) was passed through the image map of the sample with the IR spectra in contact with the transit extracted and collected to yield a line map of the data that was then analyzed by alternating least squares (ALS) to reconstruct the IR spectrum of each paint layer. Comparing each reconstructed IR spectrum against a library of IR spectra from the PDQ database, we show that high quality matches can be obtained, and the line and model of the vehicle from which the paint chip originated can be identified. This was not always the case when paint chips from the same automotive vehicles were cast in epoxy and then cross sectioned using a microtome.

4.2. Methodology

In the first study which involved paint chips cast in epoxy and cross sectioned (Data Set 1), three automotive paint samples (2001 General Motors Suburban, 2002 General

Motors Chevrolet Tahoe, and 2003 Toyota Highlander) were obtained from the RCMP. Each paint sample was removed from its metal substrate using a sharp scalpel, washed with methanol to remove dirt and particulate matter, embedded in a resin and then sectioned using a microtome (Reichert-Jung 2050) to generate a thin cross-section. Tuffleye® Finish blue light (Wet A Hook Technologies) and Slow-cure™ (Bob Smith Industries) thirty minute epoxy were the resins selected to prepare the embedded samples. The blue light epoxy block (which only consisted of resin) with the embedded sample was exposed to an intense blue light for approximately five minutes to achieve curing, whereas the Slow-cure™ epoxy resin block was prepared by mixing equal parts (by weight) of the resin and hardener. The uncured blue light resin and the thirty minute epoxy resin and hardener mixture was then poured into flat polyurethane embedding molds (BEEM®, Polysciences), and the automotive paint sample was then placed into the mold and oriented perpendicular to the bottom surface prior to polymerization of the epoxy. Paint samples in the Slow-cure™ thirty minute epoxy block were placed an oven at 60°C for ninety minutes to ensure total curing.

In the second study which involved four paint samples (2006 Buick Lacrosse, 2003 Nissan Murano, 2001 General Motors Suburban, and 2003 Toyota Highlander) cross sectioned without embedding media (Dataset 2). Each paint chip was placed between two rigid polyethylene plastic pieces and positioned in a microtome (Reichert-Jung 2050) to ensure that a thin cross section (approximately 4 to 5 μm thick) cut by the microtome contained a representative sampling of all four layers.

In both studies, each thin cross section was collected, deposited on a barium fluoride disk, and examined under a Leica light microscope for defects, which would

appear as dirt or cracks and crevices in an otherwise smooth surface. A portion of the barium fluoride disk without sample was run at 4cm^{-1} resolution to collect background before the image map of the sample was generated. For each cross sectioned paint sample, transmission IR image maps generated at 4cm^{-1} resolution using an iN10-MX microscope (Thermo-Nicolet, Madison, WI) equipped with a liquid nitrogen cooled mercury cadmium telluride (MCT) single imaging detector were collected. Both the aperture and step size of the single imaging MCT detector (50 micron x 375 micron) have adjustable values. For the analysis of each automotive paint sample, a 20 micron aperture and 5 micron step size yielded the best results when the microscope was operated in transmission mode.

A line map was extracted from the IR image map of each paint sample. The data for the line map was taken on as oblique transit as possible to include as many spectra of each layer and of the mixed interfacial region between layers. All layers were represented in the line map. The spectra comprising the line map were preprocessed. The region between 744cm^{-1} and 700cm^{-1} was deleted as data in this region was too noisy. The influence of CO_2 was suppressed by directly interpolating between 2280 cm^{-1} and 2400 cm^{-1}. Automatic baseline correction was applied to all IR spectra in the line map.

After preprocessing, spectra were extracted from the line map and examined for the presence of artifacts that may have been a direct result of the extraction procedure used. Spectra with aberrant peak intensities were discarded, and line maps with spectra that exhibited peak shifting were retaken. After verification of each line map, IR spectra constituting the investigated slice served as input for ALS.

4.3. ALS and Spectral Library Matching

Attempts to directly match IR spectra in each line map against the PDQ database were unsuccessful as the IR spectra in the line map were mixtures of the different layers of paint or were too noisy. For this reason, ALS was directly applied to the IR spectra comprising the line map. To ensure the success of ALS, the spectral data must have high signal to noise. For this reason, IR spectra that were noisy were deleted from the analysis. In order for ALS to perform well, each layer and the boundaries between the layers should be represented by as many spectra as possible, which was our motivation for using an oblique transit (line) to develop the line maps. Because an automotive paint fragment is a laminated structure, it is important to sample the layers in their order of presentation, which was the reason why ALS analysis was restricted to line maps.

ALS decomposes the data matrix, X, into three matrices (see Equation 4.1), where C is the concentration matrix, S is the spectral matrix, and E is the residual matrix. Equation 4.1 is solved iteratively in two constrained least squares steps (see Equations 4.2 and 4.3). To perform ALS, an initial estimate of C must be provided. Using this estimate of C, an estimate of S is computed. Using the estimate of S, C is computed. From the product of C and S, an estimate of the principal component analysis (PCA) reproduced data matrix, X_{PCA}, is calculated. This process is repeated until convergence has been achieved.

$$\mathbf{X} = \mathbf{CS} + \mathbf{E} \tag{4.1}$$

$$\mathbf{S}^T = (\mathbf{C}^T\mathbf{C})^{-1} (\mathbf{C}^T\mathbf{X}_{PCA}) \tag{4.2}$$

$$\mathbf{C} = (\mathbf{X}_{PCA}\mathbf{S}^T) (\mathbf{S}\mathbf{S}^T)^{-1} \tag{4.3}$$

In our study, the constraints used for ALS were nonnegative concentrations and nonnegative absorbances as the concentration of a particular layer and the absorbance at a particular wavelength should not be less than zero. Furthermore, the concentration profile of each layer was also constrained to be unimodal as a paint chip is a laminated structure. The use of the PCA reconstructed data matrix, instead of the original data matrix, stabilized the calculations and reduced the noise in the concentration and spectral matrices.

For each line map, three separate ALS models were computed to account for the rotational ambiguity associated with underdetermined system: a four component model, a six component model, and a fifteen component model. (In the case of the embedded paint sample, a five component model was substituted for the four component model.) All twenty-five components (twenty-six components in the case of an embedded paint sample) were used to find the pure spectra of the paint layers from the reconstructed IR spectra of the cross sectioned paint sample. Because the spectra of the clear coat, color coat, surfacer-primer, and e-coat layers are distinctive, the 25 IR spectra (or 26 IR spectra in the case of an embedded paint sample) could be readily divided into four groups (or five groups in the case of an embedded paint sample). The separate ALS models used to analyze the line maps allowed us to compensate for the rotational ambiguity associated with underdetermined systems of this type and improved the quality of the library matches obtained for each paint layer.

Our past experience with ALS has shown that initial estimates of either the concentration or spectral matrices are crucial for transforming these two matrices into physically meaningful solutions with ALS. For this reason, the varimax extended rotation (VER), previously developed by our research group [12, 13] (to identify the components

in an oil in water emulsion from Raman imaging data) was applied to each spectral line map to compute an initial estimate of the concentration matrix. VER utilizes a four-step procedure to determine the relative concentration and the IR spectrum of each paint layer in the cross sectioned sample. First, the spectra are preprocessed to identify the so-called extremum points (i.e., IR spectra where the proportion of a particular layer is maximized relative to all other layers). The IR spectra in the line map are normalized to constant row sum ensuring that each spectrum is weighted equally in the analysis while simultaneously reducing the number of degrees of freedom in the data by one). Next, each wavelength is range scaled. Range scaling allows for the extremum points in the data to be identified while recovering the lost degree of freedom). Range scaling also opens up the data recovering the dimension lost in the previous step. The final step is normalizing each spectrum to unit length. Normalization of the data to unit length accounts for changes in the optical path length ensuring that any variation in the data is due only to changes in the composition of the constituents. It also allows each spectrum to serve as a potential basis vector for a new coordinate system. Normalizing each row vector to unit length also reduces the dimensionality of the data by one. By using this preprocessing scheme, we have closed, opened, and closed the data again.

The second step involves principal component analysis [14], which reduces the dimensionality of the data while retaining the information present in the original data. In the third step, a Varimax rotation [15] is applied to the extracted principal components followed by a so-called extended rotation [16-18] that uses the extremum points to rotate the score and loading matrices towards a physically meaningful solution. After the concentration and spectral matrices have been rotated, they are transformed back to the

original measurement space in the fourth and final step of VER. The effects of normalization, range scaling, and row summing on the spectral data are removed. ALS is then applied to the estimates of the concentration matrix determined by VER to develop better estimates of both the concentration and spectral matrices. The spectral region used for decatenation of the line maps was 4000 cm^{-1} - 748 cm^{-1}.

Spectral library matching for each IR spectrum recovered by ALS was performed using OMNIC (Thermo Nicolet). All library searches were restricted to the spectral region between 1641 cm^{-1} and 860 cm^{-1} which in our previous studies [19] has been shown to contain information about the make, line and model of the vehicle. Outside of this region, the IR spectrum contains only the carbonyl band and C-H stretching bands, which are present in the spectra of all paint layers. Each IR spectral library searched was of the same manufacturer (e.g., General Motors) and within the same production year range (e.g., 2000-2006) as the automotive paint **s**ample from which the reconstructed IR spectra were obtained.

OMNIC library searchers were configured using correlation as the search type with Happ-Genzel apodization. The quality of each search was evaluated using the hit quality index (HQI). The top five hits of each search were reported as the identity of the unknown is expected to be captured in the top five hits. A library search where the correct sample or the correct line and model of the vehicle from which the sample was obtained is included in the top five hits was considered to be a successful match. The paint samples comprising the General Motors, Toyota, and Nissan libraries were similar in composition, making the library matching problem investigated in this study challenging.

4.4. Results and Discussion

4.4.1. Data Set 1

When a paint chip is embedded in blue light or thirty minute epoxy, the clear-coat layer and the e-coat layer of the chip are in direct contact with the embedding resin. The similarity of the IR spectrum of the embedding resin and the adjacent paint layer can prevent MCR from extracting an accurate IR spectrum for each of these layers. For blue light epoxy, the IR spectrum of the resin is similar to that of the clear coat layer (if the formulation of the clear coat layer is acrylic melamine styrene), whereas for the thirty minute epoxy, the IR spectrum of the resin is similar to the e-coat layer. This can adversely impact the results of MCR due to the mixing of the IR spectrum of the resin with that of the clear coat or e-coat layers. For this reason, three separate MCR models were computed for each line map: a four component model, a six component model, and a fifteen component model. All twenty-five components were used to find the pure spectra of the paint layers from the reconstructed IR spectra of the embedded paint sample. Some reconstructed IR spectra from the four component and six component MCR models differed from the IR spectra recovered from the fifteen component model. The rotational ambiguity associated with underdetermined system of this type was addressed when separate MCR models were used to analyze the line maps.

Because the spectra of the epoxy, clear coat, color coat, surfacer-primer and e-coat layers are distinctive for these three samples, the 25 reconstructed IR spectra were divided into five groups. Each group was library matched to the corresponding paint layer using IR spectra of General Motors or Toyota paint samples from the PDQ automotive paint database.

To assess the efficacy of this approach to forensic automotive paint analysis, three automotive paint samples were selected for this study. These three samples because of their small size are angled and are also representative of paint chips recovered from the clothing of hit-and-run victims. The innate difficulty associated with positioning each one for scanning makes them an excellent choice to evaluate the information content of the line maps. While the shape of each sample made it difficult to create a line map, spectral information about each layer was obtained using an oblique transit that fully bisected the chip. This type of cut maximized the number of spectra obtained for each line map, and allowed for the observed transitions between paint layers to occur over a large number of spectra, thereby providing more information about composition change as a function of position on the line map. Because the transit itself is angled off of the paint chip, the effects of any natural angling as a result of skewed sample positioning are obviated. The most informative transits for a given sample are those with start and endpoints that contained pure spectra of the epoxy used to embed the paint chip.

Figures 4.1 and 4.2 show the reconstructed IR spectra of the clear coat, surfacer-primer, and e-coat layers for UAZP00331 (2001 General Motors Suburban) from line maps generated using both blue light (59 IR spectra) and thirty minute epoxy (76 IR spectra). For each epoxy, the reconstructed IR spectrum is a good match for the IR spectrum of the same sample (see Figures 4.1 and 4.2) in the PDQ library. The reconstructed IR spectrum of each layer was also matched against IR spectra in the corresponding General Motors libraries for the clear coat, surfacer-primer, and e-coat layers with 628, 633, and 585 IR spectra respectively comprising each library. For the two undercoat layers (i.e., surfacer-primer and e-coat), the correct model was the first hit for both the blue light and thirty

minute epoxy paint samples (see Table 4.1). For the clear coat layer, the correct match was the first hit for the blue light epoxy sample, whereas the third hit was the correct match for the thirty minute epoxy paint sample (see Table 1). The hit quality index HQI) for each of these matches against the PDQ library was greater than 90%, which is indicative of a high quality match [20].

Table 4.1. Library search results for UAZP00331

Epoxy	**Layer**	**HQI Value of Match**	**Position in Search /(Top Five Hits)**
Thirty Minute	Clear Coat	97.10	3
Thirty Minute	Surfacer-Primer	95.10	1
Thirty Minute	E-Coat	97.70	1
Blue Light	Clear Coat	98.07	1
Blue Light	Surfacer-Primer	91.03	1
Blue Light	E-Coat	93.63	1

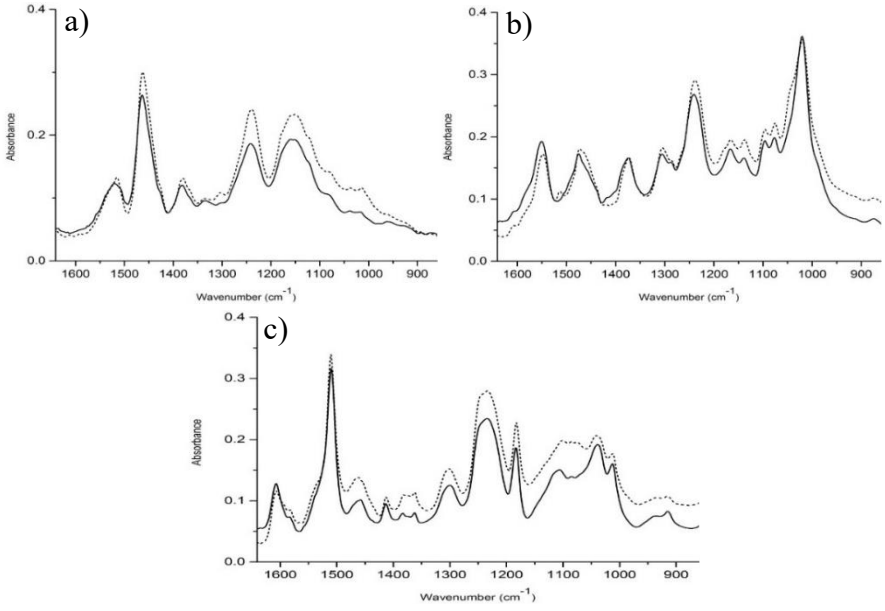

Figure 4.1. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00331) in the PDQ library (solid line) for the thirty minute epoxy. a) Clear Coat layer (OT2), b) Surfacer-primer layer (OU1), and c) e-coat layer (OU2).

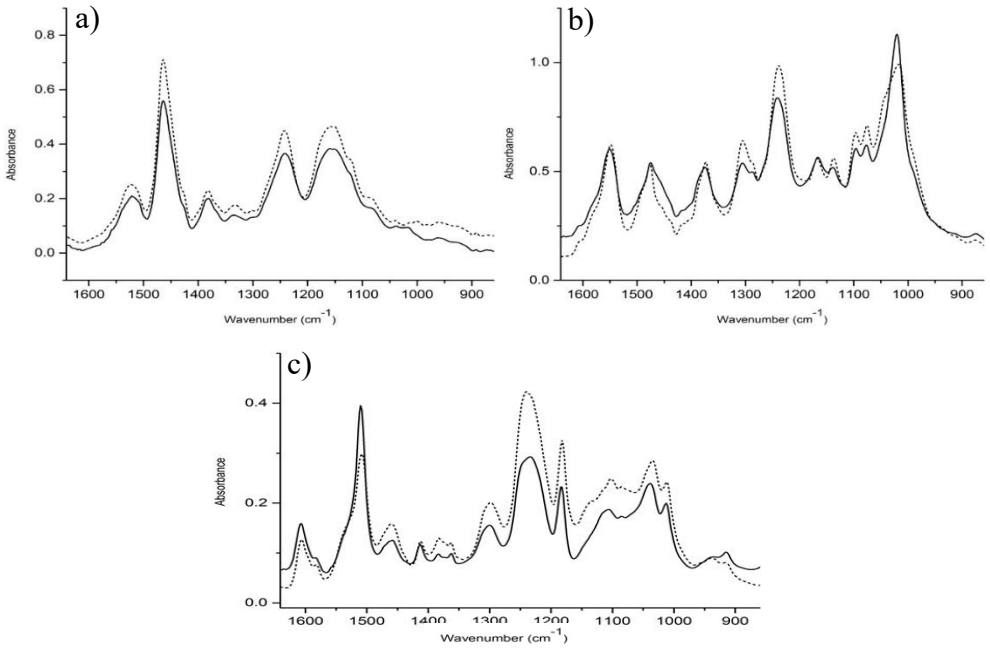

Figure 4.2. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00331) in the PDQ library (solid line) for the blue light epoxy. a) Clear Coat layer (OT2), b) Surfacer-primer layer (OU1), and c) e-coat layer (OU2).

The General Motors Suburban was the only model in the top five hits for each layer that was common in these six searches. Although the IR spectrum of the color coat layer can also be reconstructed from MCR for both the blue light and thirty minute epoxy samples, the large variation in the IR spectrum of the color coat layer due to the pigments present in the layer precluded its use in library matching against the General Motors spectral database. Previous studies performed in our laboratory have demonstrated that paint samples can be identified as to make and model of the vehicle using IR spectra of only the clear coat, surfacer-primer, and e-coat layers [21-24].

Figures 4.3 and 4.4 show the reconstructed IR spectra of the clear coat, surfacer-primer, and e-coat layers for UAZP00436 (2002 General Motors Chevrolet Tahoe) from line maps generated using both thirty minute epoxy (60 IR spectra) and blue light epoxy

(64 IR spectra). All reconstructed IR spectra, in the case of the thirty minute epoxy, were good matches for the same paint sample in the General Motors library (see Figure 4.3). As for the clear coat layer, the correct library match was the fifth hit, which corresponded to the actual paint sample in the General Motors library, whereas the correct library match for both the surfacer-primer and the e-coat layers was the third hit, which also corresponded to the actual paint sample in the library (see Table 4.2). Furthermore, Chevrolet Tahoe was the only model listed in the top five hits for each layer that was common in these three searches.

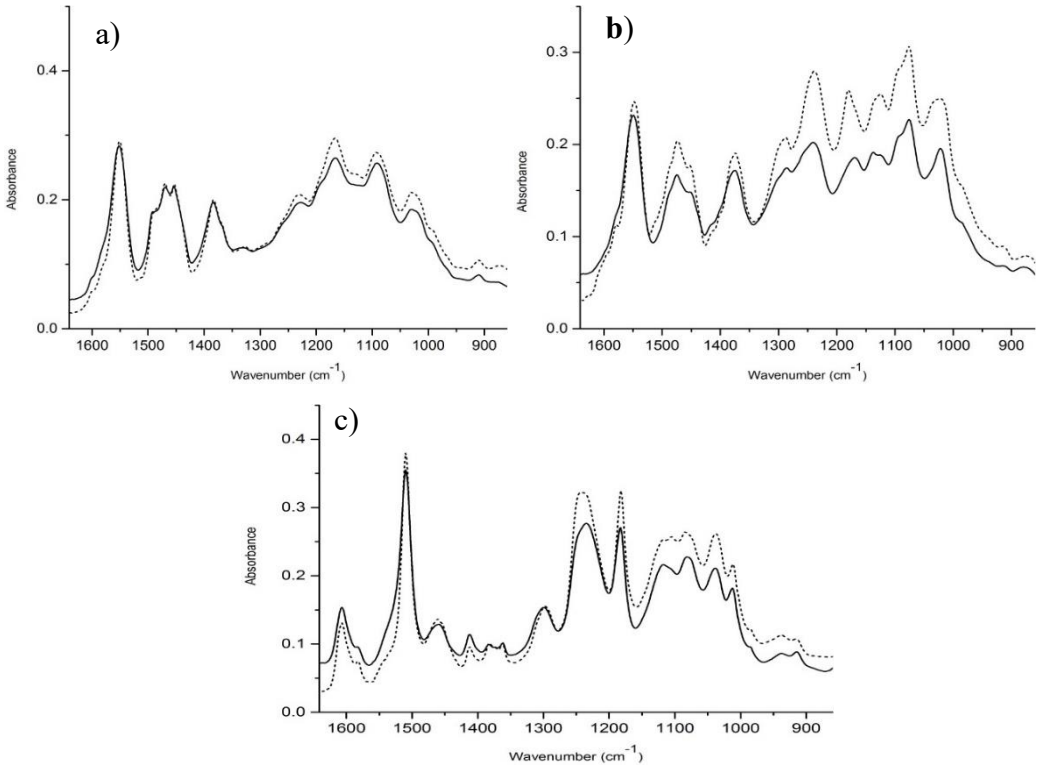

Figure 4.3. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00436) in the PDQ library (solid line) for the thirty minute epoxy. a) Clear Coat layer (OT2), b) Surfacer-primer layer (OU1), and c) e-coat layer (OU2). Each layer was a good match.

Table 4.2. Library search results for UAZP00436

Epoxy	Layer	HQI Value of Match	Position in Search (Top Five Hits)
Thirty Minute	Clear Coat	97.87	5
Thirty Minute	Surfacer-Primer	92.81	3
Thirty Minute	E-Coat	96.86	3
Blue Light	Clear Coat	97.43	2
Blue Light	Surfacer-Primer	94.01	1
Blue Light	E-Coat	76.94	No match

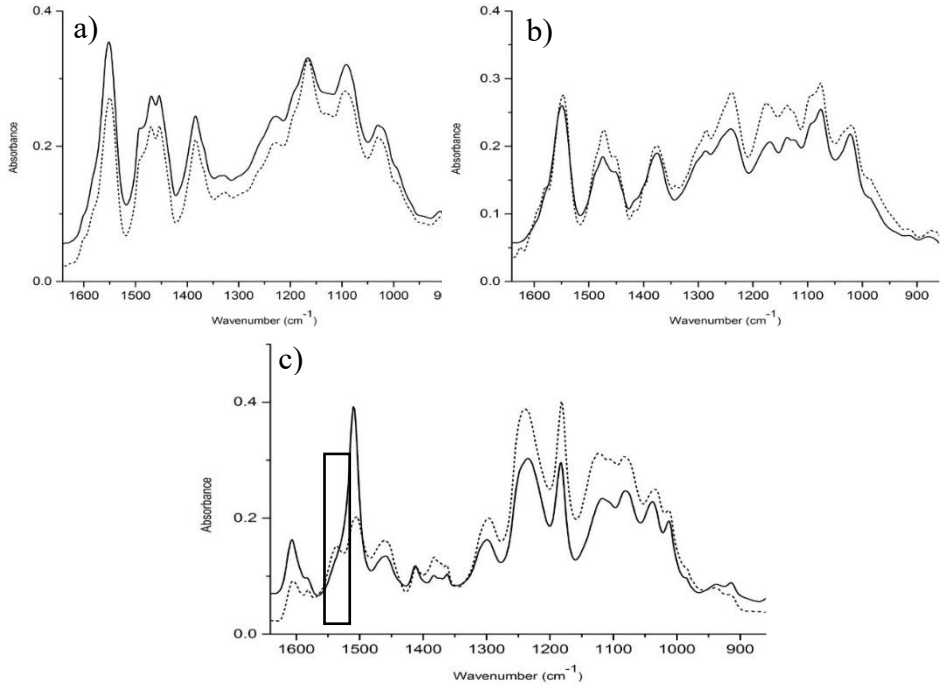

Figure 4.4. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00436) in the PDQ library (solid line) for the blue light epoxy. Although the clear coat and e-coat layers were a good match, there is substantial mixing of the blue light epoxy with the reconstructed IR spectra of OU2. a) Clear Coat layer (OT2), b) Surfacer-primer layer (OU1), and c) e-coat layer (OU2). 1550 cm^{-1} which is indicative of the blue light epoxy (see enclosed square in 4.4c) is absent in the IR spectrum of the e-coat layer (see solid line of Figs. 4.3c and 4.4c) but is present in the reconstructed e-coat layer IR spectrum.

In the case of the blue light epoxy (see Figure 4.4), the matches for both the clear coat and surfacer- primer layers were good. However, the blue light epoxy mixed with the

e-coat layer in the reconstructed IR spectrum of the e-coat layer. The correct library match for both the clear-coat and the surfacer-primer layers in the blue light embedded paint sample was the second and first hits respectively (see Table 4.2), which corresponded to the actual paint sample. However, the reconstructed IR spectrum of the e-coat layer could not be correctly matched in the library search of the PDQ database as reflected by its low HQI score (76.94%) for the top hit of the reconstructed IR spectrum of the e-coat layer (see Table 4.2).

Figures 4.5 and 4.6 show the reconstructed IR spectra of the clear coat, surfacer-primer, and e-coat layers for UAZP00484 (2003 Toyota Highlander) from line maps generated using both blue light and thirty minute epoxy. For both the blue light epoxy (45 IR spectra) and thirty minute epoxy (65 IR spectra), only two of the reconstructed IR spectra were good matches for the actual paint sample. For the thirty minute epoxy (see Figure 4.5), only the e-coat and the surfacer-primer layers were good matches, whereas the reconstructed clear-coat IR spectrum did not match the actual paint sample due to mixing of the clear coat IR spectra with that of the thirty minute epoxy (see Figure 4.7). The clear coat, surfacer-primer and e-coat reconstructed IR spectra were also matched against the IR spectra from the corresponding PDQ (Toyota) libraries (see Table 4.3) for the clear coat, surfacer-primer, and e-coat layers with 269, 308, and 298 IR spectra respectively comprising each library. For the two undercoat layers (e-coat and surfacer-primer), the actual sample was also the first hit (see Table 4.3) in the search. The reconstructed clear coat IR spectrum was a poor match for IR spectra in the Toyota library as the HQI value for the top hit in the search was only 83.99%, and the correct model was not in the top five hits.

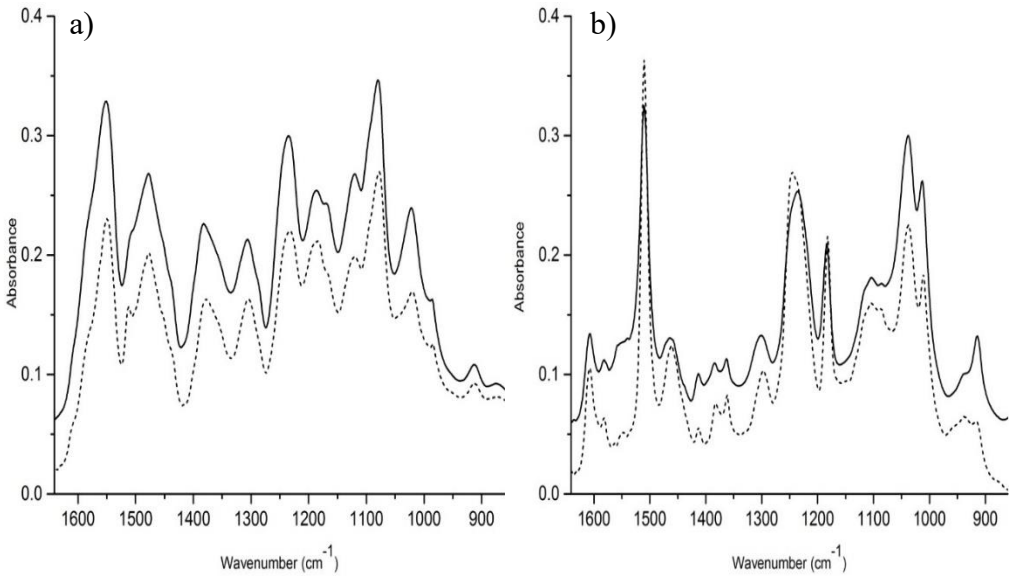

Figure 4.5. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00484) in the PDQ library (solid line) for the thirty minute epoxy. a) Surfacer-primer layer (OU1) and b) e-coat layer (OU2).

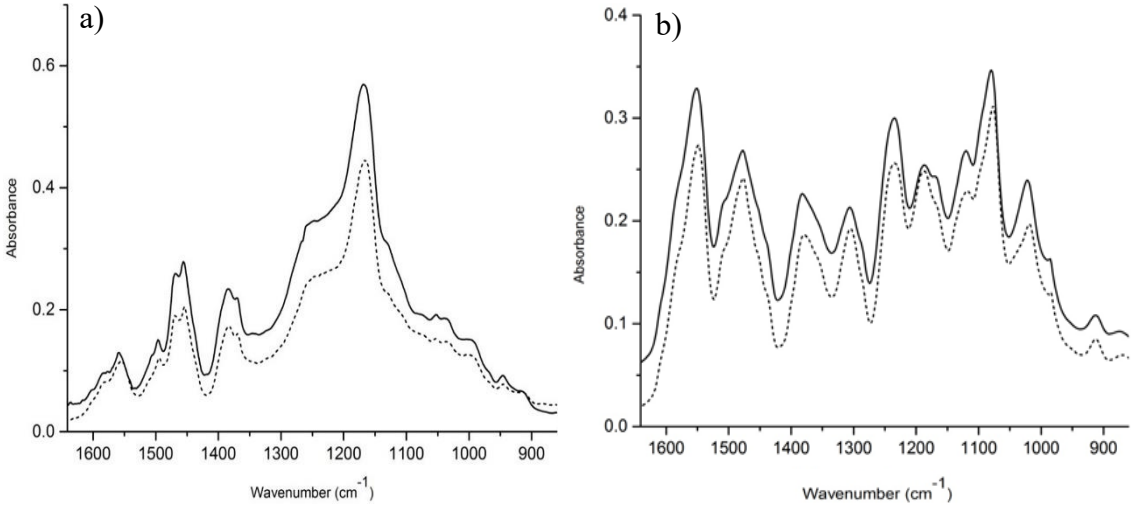

Figure 4.6. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00484) in the PDQ library (solid line) for the blue light epoxy. a) Clear coat layer (OT2) and b) Surfacer-primer layer (OU1).

Table 4.3. Library search results for UAZP00484

Epoxy	Layer	HQI Value of Match	Position in Search (Top Five Hits)
Thirty Minute	Clear Coat	83.99	No Match
Thirty Minute	Surfacer-Primer	95.39	1
Thirty Minute	E-Coat	96.01	1
Blue Light	Clear Coat	97.3	3
Blue Light	Surfacer-Primer	96.59	1
Blue Light	E-Coat	73.44	No Match

In the case of the blue light epoxy, only the clear coat and surfacer-primer layers were successfully reconstructed by MCR (see Figure 4.6). The reconstructed IR spectrum of the e-coat layer exhibited mixing with the blue light epoxy (see Figure 4.7). As a result, its IR spectrum did not yield a good match to the spectrum of the actual sample. When the reconstructed IR spectrum of the clear coat and surfacer-primer layers were matched against the corresponding Toyota libraries, the correct model was the third and first hit respectively (see Table 4.3). The reconstructed IR spectrum of the e-coat layer was a poor match for IR spectra in the Toyota library as its HQI value for its top hit in the search of the library was only 73.44% and the correct model was not in its top five hits. Clearly, the mixing of spectra of the epoxy with spectra of either the clear coat or e-coat layers can adversely impact the MCR results.

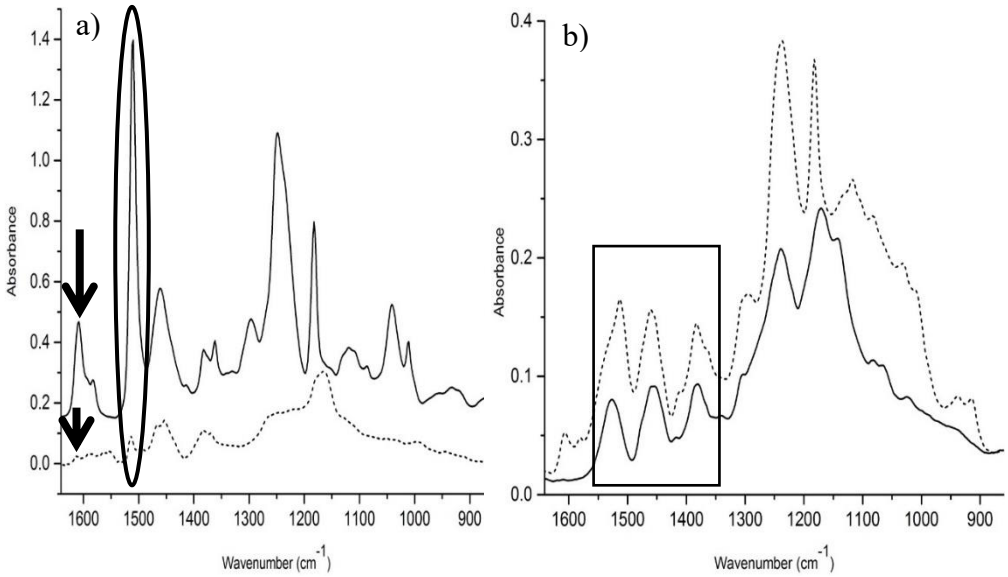

Figure 4.7. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the thirty minute and blue light epoxy (solid line). a) Mixing of the thirty minute epoxy spectrum with the reconstructed clear coat IR spectrum: 1510 cm^{-1} and 1609 cm^{-1} are absent in the IR spectrum of the clear coat layer (see solid line of Fig. 4.7a) for this sample. b) Mixing of the blue light epoxy spectrum with the reconstructed IR spectrum of the e-coat layer: the peaks present in the spectral region of 1350–1550 $cm^{-1}$are absent in the IR spectrum of the e-coat layer (see solid line of Fig. 4.7b) for this sample.

4.4.2 Data Set 2

If an automotive paint chip is cast in epoxy, the clear-coat and e-coat layers of the cross sectioned chip are in direct contact with the epoxy and infiltration of the epoxy into these layers can occur, which may prevent accurate reconstruction of their IR spectra by ALS. When the paint chip is not cast in epoxy prior to cross sectioning, the sample preparation is easier and extracting the IR spectrum of each paint layer by ALS is straightforward as there is no spectral interference from the epoxy. To demonstrate the advantages of this approach (i.e., not using epoxy to prepare the sample for cross sectioning), four automotive paint samples (each from a different vehicle) were selected

for infrared imaging and ALS analysis. All four paint chips were similar in size and shape to those cast in epoxy as they were angled because of their small size (3 mm in length). The starting point and endpoint for each transit in the IR image of the unembedded paint samples was the unoccupied region of the BaF_2 disk adjacent to the paint sample. For paint chips embedded in epoxy, the starting point and endpoint was the pure spectra of the epoxy used to cast the sample.

Four paint samples, each from the same vehicle as the unembedded cross sectioned paint samples, were also cast in an epoxy block for cross sectioning as part of this study to compare reconstructions of the IR spectra of the individual paint layers computed by ALS with and without the use of epoxy as an embedding medium. Figure 4.8 shows an image of a microtomed paint chip (UAZP00565) from a 2006 Buick Lacrosse in the presence and absence of epoxy on a BaF_2 disk. In Figure 4.8a, the cross sectioned paint chip without epoxy is displayed. All layers (clear coat, color coat, surfacer primer, and e-coat) are visible and the borders between the layers are well defined. By comparison, a large fraction of the same paint chip when cast in epoxy and cross sectioned (see Figure 4.8b) is barely visible. Most of the clear coat layer is buried under the epoxy, and the IR spectra collected in this region is likely to be more representative of the epoxy, not the clear coat layer of the paint sample.

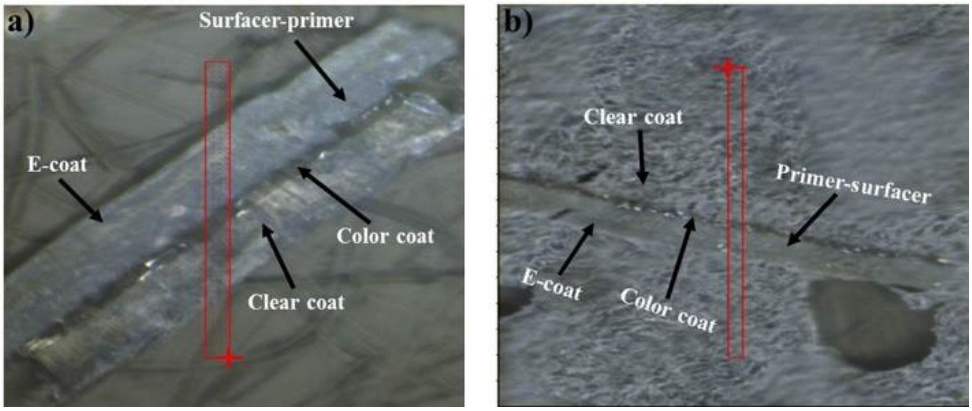

Figure 4.8. Image of a microtomed paint chip (UAZP00565) from a 2006 Buick Lacrosse in the presence and absence of epoxy on a BaF_2 disk. a) The cross sectioned paint chip without epoxy is displayed. All layers are visible and the borders between the layers are well defined. b) The same paint chip is cast in an epoxy block and cross sectioned. A large fraction of the paint chip is barely visible.

Figures 4.9 and 4.10 show reconstructed IR spectra of the clear coat, surfacer-primer, and e-coat layers from line maps of UAZP00565 samples generated with epoxy (60 spectra) and without epoxy (50 spectra). All reconstructed IR spectra for the paint chip that is not cast in epoxy were good matches for the same paint sample in the General Motors library (see Figure 4.9). As for the clear coat layer, the correct library match was also the first hit in the library search, which corresponded to the actual paint sample, whereas the correct match (i.e., the same assembly line and model of the vehicle) for the surfacer-primer and e-coat layers was the first and fifth hits respectively (see Table 4.4). When a UAZP00565 paint chip was cast in epoxy prior to cross sectioning (see Figure 4.10), the reconstructed IR spectra of both the e-coat and surfacer-primer layers were a poor match for the IR spectra of the e-coat and surfacer-primer layers of the UAZP00565 paint sample in the General Motors library. Furthermore, the reconstructed IR spectrum of the e-coat layer was not correctly matched (as to line and model of the vehicle) to any of the paint

samples in the hit-list (see Table 4.4) due to spectral interference from the epoxy. For both the surfacer-primer and clear coat layers, the reconstructed IR spectra were the fourth and third hits respectively in the library search (see Table 4.4). The hit quality index (HQI) value for each of these hits was greater than 90%. (A sample was judged to be correctly matched if the line and model of the vehicle corresponds to any of the samples in the top five hits using the HQI to rank the library IR spectra in the search).

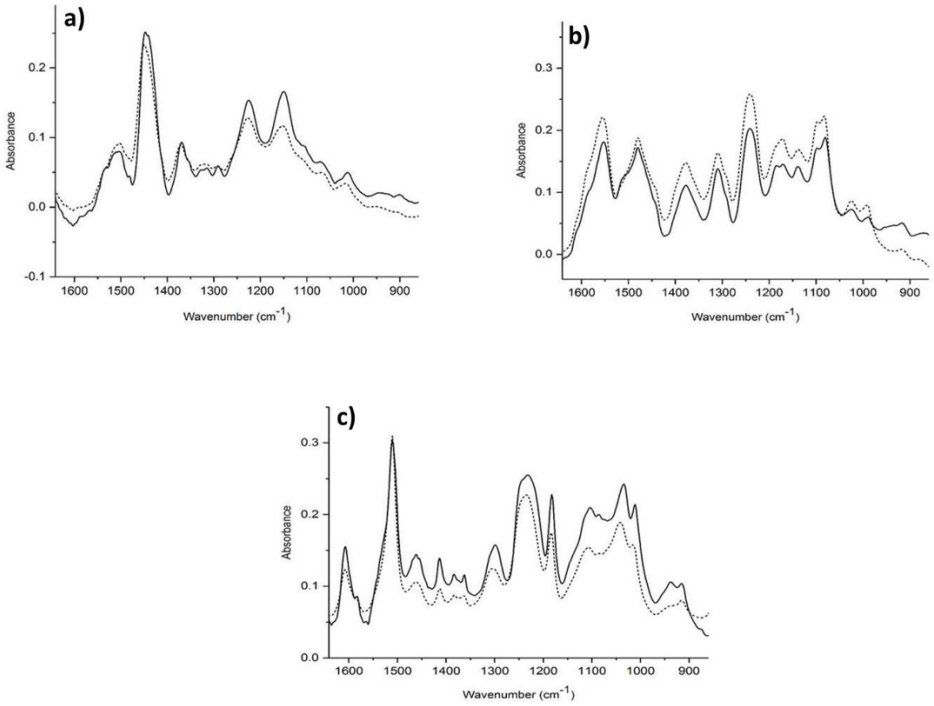

Figure 4.9. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00565 – 2006 Buick Lacrosse) in the General Motors spectral library for the paint sample not cast in epoxy. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

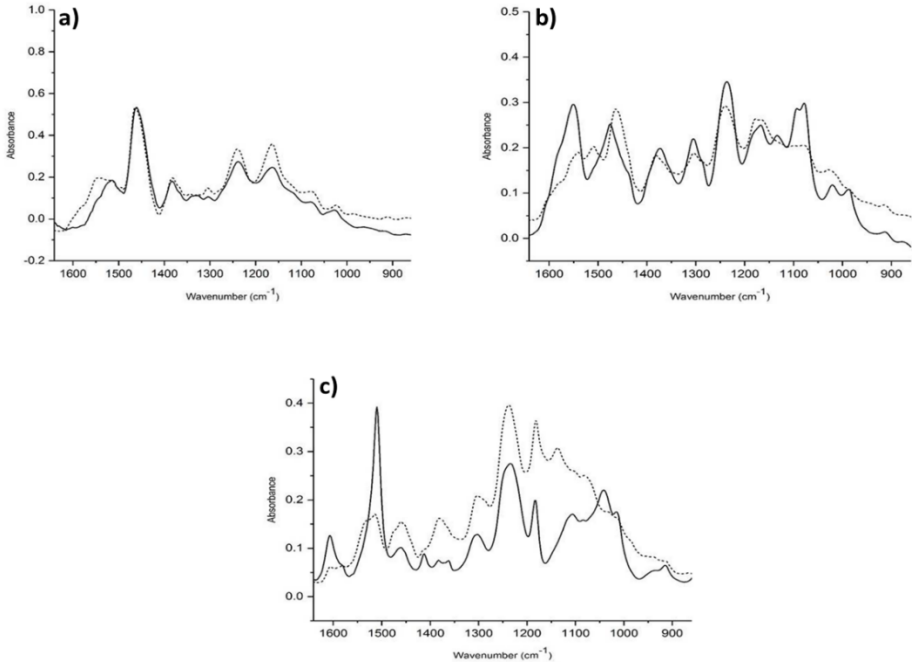

Figure 4.10. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00565 – 2006 Buick Lacrosse) in the General Motors spectral library for the paint sample cast in epoxy. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

Table 4.4. Library Search Results for UAZP00565

Medium	**Layer**	**HQI Value of Match**	**Position in Search /(Top Five Hits)**
None	Clear Coat[1,2]	97.07	1
None	Surfacer-Primer[1]	96.25	1
None	E-Coat[1]	94.31	5
Thirty Minute Epoxy	Clear Coat[1]	94.50	3
Thirty Minute Epoxy	Surfacer-Primer[1]	90.92	4
Thirty Minute Epoxy	E-Coat	68.53	No Match

[1]Correct model
[2]Actual paint sample

Figures 4.11 and 4.12 show reconstructed IR spectra of the clear coat, surfacer-primer, and e-coat layers from line maps of UAZP00731 (2003 Nissan Murano) generated with epoxy (35 spectra) and without epoxy (63 spectra). When the paint chip was not cast in epoxy prior to cross sectioning, all reconstructed IR spectra were again good matches

for the actual paint sample (see Figure 4.11), and the correct library match for each layer (see Table 4.5) was also the first hit, which corresponded to the actual paint sample. When the paint chip was cast in epoxy, all three reconstructed IR spectra (see Figure 4.12) were poor matches for the actual paint sample, and the reconstructed IR spectra of both the surfacer-primer and e-coat layers were not correctly matched to any of the paint samples in the hit-list (see Table 4.5) as the correct line and model was not in the top five hits. For the reconstructed clear coat layer, the correct match was the second hit (see Table 4.5), which gave the correct line and model of the vehicle from which the paint chip originated. However, the HQI value for this hit was less than 90%, which is indicative of the match not being of high quality. For both the clear coat and e-coat layers (see Figure 4.12), there were peaks in the reconstructed IR spectra due to the thirty minute epoxy.

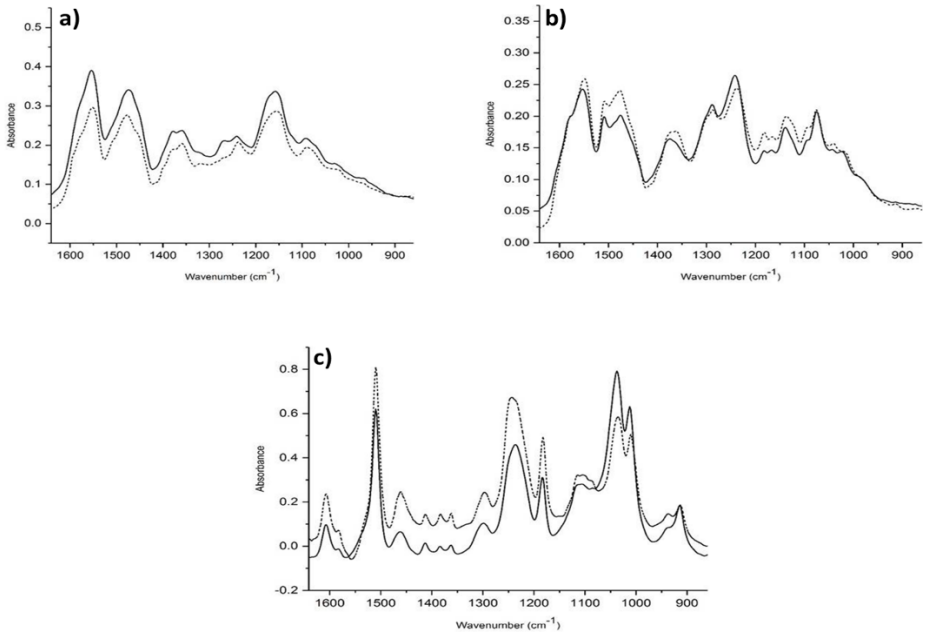

Figure 4.11. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00731 – 2003 Nissan Murano) in the Nissan spectral library for the paint sample not cast in epoxy. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

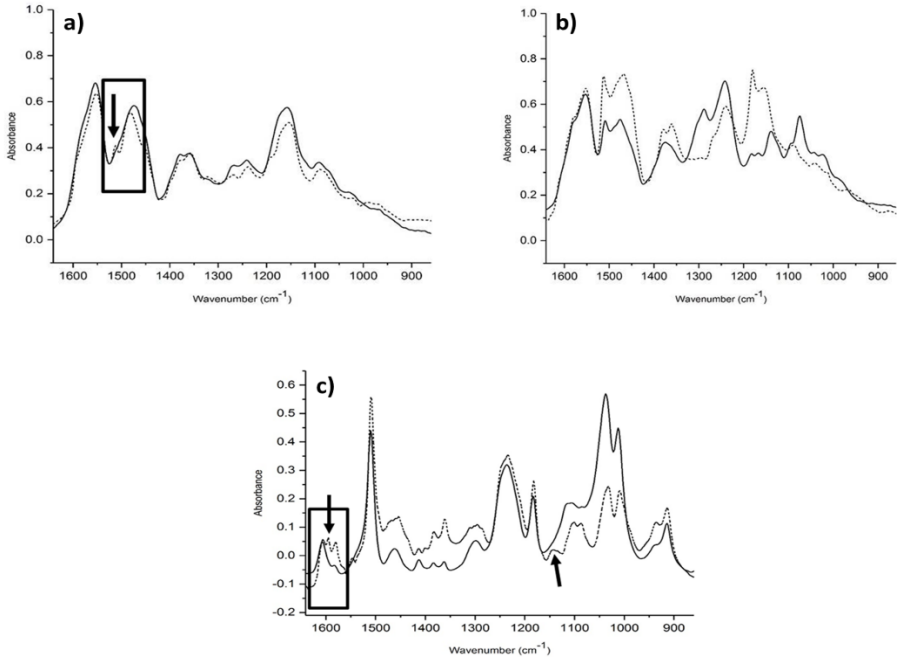

Figure 4.12. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00731 – 2003 Nissan Murano) in the Nissan spectral library for the paint sample cast in epoxy. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer. Peaks from the thirty minute epoxy are denoted by an arrow enclosed in a solid rectangle.

Table 4.5. Library Search Results for UAZP00731

Medium	**Layer**	**HQI Value of Match**	**Position in Search /(Top Five Hits)**
None	Clear Coat[1,2]	94.83	1
None	Surfacer-Primer[1,2]	95.49	1
None	E-Coat[1,2]	90.05	1
Thirty Minutes Epoxy Thirty Minute Epoxy	Clear Coat[1] Surfacer-Primer	89.36 62.45	2 No Match
Thirty Minute Epoxy	E-Coat	79.56	No Match

[1]Correct model
[2]Actual paint sample

Figures 4.13 and 4.14 show the reconstructed IR spectra of the clear coat, surfacer-primer, and e coat layers of UAZP00331 (2001 General Motors suburban) and UAZP00484 (2003 Toyota Highlander) generated from line maps (45 spectra for UAZPOO331 and 43

spectra for UAZPOO484) prepared from paint chips not cast in epoxy. All reconstructed IR spectra were good matches for the actual paint samples, and yielded good matches when searched against the General Motors and Toyota libraries. For UAZP00331 (see Table 4.6), the correct line and model was the first hit for both the clear coat and e-coat layers, whereas the second hit was the correct match for the surfacer-primer layer. Library search results for UAZP00484 (see Table 4.7) were also encouraging. For the surfacer primer and e-coat layers, the actual sample was also the first hit, whereas the correct line and mode was the top hit for the clear coat layer.

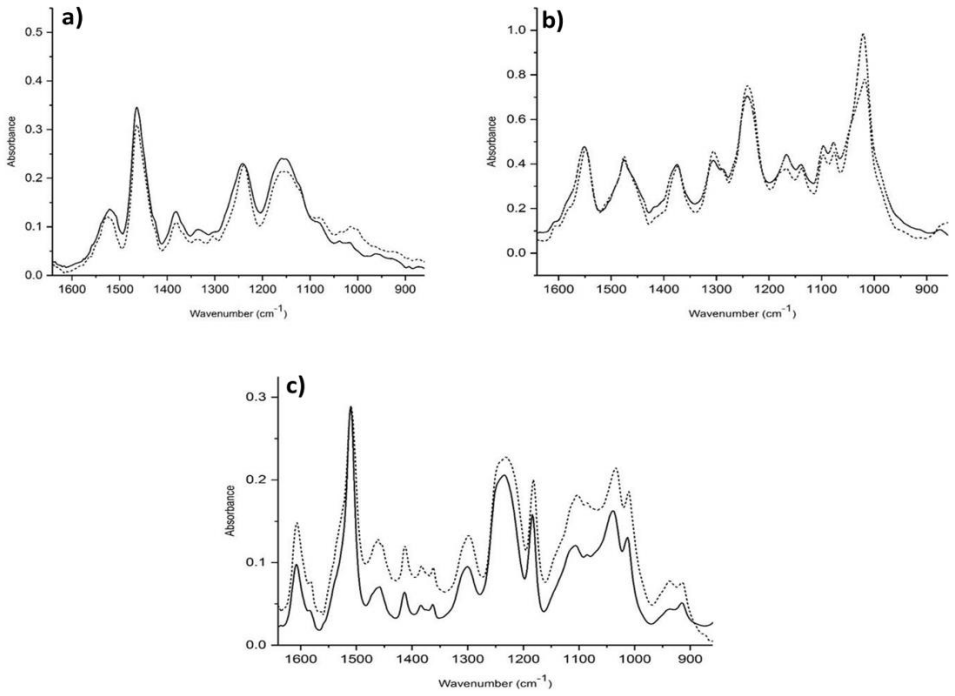

Figure 4.13. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00331 – 2001 General Motors Suburban) in the General Motors spectral library for the paint sample not cast in epoxy. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

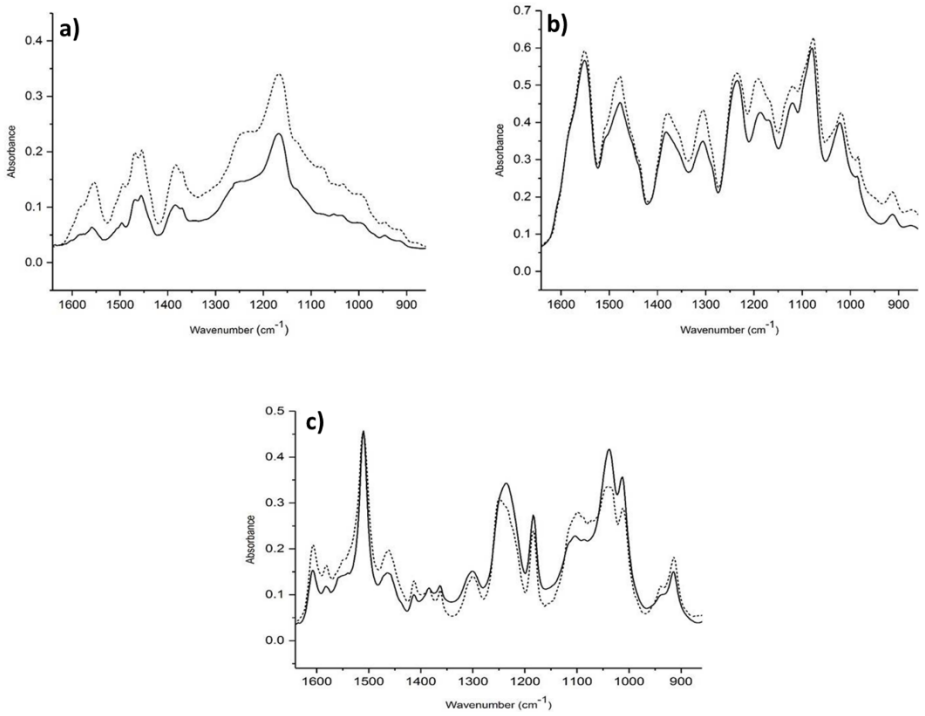

Figure 4.14. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00484 - 2003 Toyota Highlander) in the Toyota spectral library for the paint sample not cast in epoxy. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

Table 4.6. Library Search Results for UAZP00331

Medium	**Layer**	**HQI Value of Match**	**Position in Search /(Top Five Hits)**
None	Clear Coat[1]	97.70	1
None	Surfacer-Primer[1]	94.44	2
None	E-Coat[1]	96.58	1
Thirty Minute Epoxy	Clear Coat	97.10	3
Thirty Minute Epoxy	Surfacer-Primer	95.10	1
Thirty Minute Epoxy	E-Coat	97.70	1

[1]Correct model
[2]Actual paint sample

Table 4.7. Library Search Results for UAZP00484

Medium	Layer	HQI Value of Match	Position in Search /(Top Five Hits)
None	Clear Coat[1]	97.42	1
None	Surfacer-Primer[1,2]	90.97	1
None	E-Coat[1,2]	94.35	1
Thirty Minute Epoxy	Clear Coat	83.99	No match
Thirty Minute Epoxy	Surfacer-Primer[1,2]	95.39	1
Thirty Minute Epoxy	E-Coat[1,2]	96.01	1

[1]Correct model
[2]Actual paint sample

In a previously published study [20], reconstructed IR spectra of the clear coat, surfacer-primer, and e-coat layers for UAZP00331 and UAZP00484 were generated using line maps (76 IR spectra for UAZPOO331 and 65 IR spectra for UAZP00484) prepared from paint chips cast in epoxy prior to cross sectioning. For both UAZP00331 and UAZP00484, the correct match for the two undercoat layers (e-coat and surfacer-primer) was the first hit (see Tables 4.6 and 4.7). As for the clear coat layer, the third hit was the correct match for UAZP00331, whereas the reconstructed clear coat IR spectrum of UAZP00484 was a poor match for spectra in the Toyota library due to mixing of the thirty minute epoxy with the clear coat layer (see Tables 4.6 and 4.7). Clearly, the mixing of the IR spectra of the epoxy with spectra of the clear coat layer can adversely impact both the ALS and library search results.

The approach to automotive paint analysis described in this paper eliminates the need to analyze each layer of automotive paint separately and ensures that recovered IR spectra of the paint layers are not contaminated by epoxy or other media. This study, which is directly targeted to enhance current approaches to forensic automotive paint examination through decreased analyses times as compared to current practices and aid in evidential significance assessment, both at the investigative lead stage and at the courtroom

testimony stage, is a direct response to Recommendation 3 of the National Academies February 2009 Report [21].

4.5. Conclusion

IR spectra of the clear coat, surfacer-primer, and e-coat layers were collected in a single analysis from multi-layered automotive paint chips using a transmission FTIR imaging microscope. Decatenation of the spectral image as represented by a line map was achieved using VER/ALS to obtain a pure IR spectrum of each paint layer. The successful spectral reconstructions of each layer allowed us to quantify the discrimination power of the original automotive paint through library searching. The results of this study suggest that reconstructed IR spectra of the paint layers extracted from spectral line maps of IR images of cross sectioned paint samples using MCR can be searched against the PDQ database to yield good matches. The proposed IR imaging method is faster than hand sectioning and analyzing each layer separately by FTIR**.**

It has also been demonstrated that if the automotive paint chip is not embedded in an epoxy before cross sectioning, the IR spectra of the clear coat, surfacer-primer and e-coat layers can be successfully reconstructed and searched against a spectral library to yield a correct match. By comparison, this was not always the case when the paint chip was first cast in epoxy prior to cross sectioning with a microtome because the epoxy infiltrated specific layers of some automotive paint chips contaminating their IR spectra and preventing accurate library matching of the e-coat or surfacer-primer layers.

References

1. Dossel., H.S.a.K., *Automotive Paints and Coatings*. 2nd ed, ed ed. 2008: Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.

2. P.G. Rodgers, R.C., N.S. Cartwright, W.H. Clark, J.S. Deak, E.W.W. Norman, Can. Soc. Forens. Sci. J., 1976. **9**: p. 1–14.

3. N.S. Cartwright, P.G.R., Can. Soc. Forens. Sci. J, 1976. **9**: p. 145-154.

4. S. Ryland, T.J., K.P. Kirkbride, Forens. Sci. Rev, 2006. **18**: p. 97-117.

5. A. Beveridge, T.F., D. MacDougall, in: B. Caddy (ed.), *Forensic Examination of Glass and Paint: Analysis and Interpretation.* Taylor and Francis, NY, 2001: p. 227-233.

6. N.S. Cartwright, L.J.C., E.W.W. Norman, R. Cameron, W.H. Clark, D.A. MacDougal, *A Computerized System for the Identification of Suspect Vehicles Involved in Hit and Run Accidents.* Can. Soc. Forens. Sci. J. , 1982. **15** p. 105–115.

7. J.L. Buckle, D.A.M., R.R. Grant, *PDQ-Paint Data Queries: The History and Technology Behind the Development of the Royal Canadian Mounted Police Forensic Science Laboratory Services Automotive Paint Database.* Can. Soc. Forens. Sci. J. , 1997. **30** p. 199–212.

8. S.G. Ryland, E.M.S., in: L. Kobilinsky (ed.), Forensic Chemistry Handbook, John Wiley & Sons, NY 2012: p. 175.

9. S.C. Rutan, A.d.J., R. Tauler in: S. D. Brown, R. Tauler, B. Walczak (eds.), Comprehensive Chemometrics, Elsevier, Amsterdam, 2009. **Vol. 2**: p. 249–260.

10. A. de Juan, E.C., R. Tauler in: R. Meyers (ed.), Encyclopedia of Analytical Chemistry: Instrumentation and Applications, Wiley, New York, 2000. **vol. 11**: p. 9800–9837.

11. W-T Chang, T.-H.C., C-C Yu, and J-Y Kau, Forens. Sci. J. , 2002. **1** p. 55-60.

12. B.K. Lavine, J.P.R., E. Voigtman, Microchem. J. , 2002. **72** p. 163–178.

13. B.K. Lavine, C.E.D., J.P. Ritter, D. Westover, T. Hancewicz, Microchem. J., 2004. **76**: p. 173–180.

14. Jolliffe, I.T., *Principal Component Analysis*. 1986, NY Springer-Verlag.

15. Harman, H.H., *Modern Factor Analysis, 3rd ed.* University of Chicago Press, 1976.

16. Miesch, A.T., *Q-Mode Factor Analysis of Geological and Petrological Data Matrices with Constant Row Sums*. 1976, Washington: United States Government Printing Office.

17. A. T. Miesch, J.E.K., Comput. Geosci. , 1976. **1** p. 161-178.

18. W. E. Full, R.E., J. E. Klovan, Math. Geol., 1981. **13** p. 331-344.

19. A. Fasasi, N.M., R.-I. Stoian, C. White, M. Allen, Mark P. Sandercock, B. K. Lavine, *Pattern recognition assisted infrared library searching of automotive clear coat.* Appl. Spec., 2015. **69**: p. 84-94.

20. F. Kwofie, U.D.N.P., M. D. Allen, and B. K. Lavine, Talanta, 2018. **186** p. 662-669.

21. Academies, T.N., *Strengthening Forensic Science in the United States: A Path Forward*. 2009, Washington, DC: The National Academy Press.

CHAPTER V

TRANSMISSION INFRARED MICROSCOPY FOR THE FORENSIC EXAMINATION OF AUTOMOTIVE PAINT – PATTERN RECOGNITION ASSISTED INFRARED LIBRARY SEARCHING

5.1. Introduction

In the previous chapter, a new method to characterize automotive paint samples recovered from a crime scene involving a vehicle related fatality such as a hit-and-run was discussed. Infrared spectra data from all the layers of an automotive paint chip was collected in a single analysis by scanning across a cross-sectioned paint sample using an FTIR imaging microscope. After the data had been collected, it underwent decatenation using multivariate curve resolution to obtain the "pure" IR spectrum of each layer. Comparing the reconstructed IR spectrum of each layer against an IR spectral library from the PDQ database demonstrated that it was possible to identify the correct line and model of the vehicle using these reconstructed spectra. This imaging experiment not only saves time and eliminate the need to analyze each layer separately, but also ensures that the final spectrum of each layer is "pure" and not a mixture, which can occur when using a scalpel to separate the individual layers and sampling too close to the boundary between adjacent layers.

In this chapter, the coupling of the proposed FTIR imaging experiment with a prototype pattern recognition infrared library searching system previously developed by

Lavine and co-workers is discussed. The forensic examination of automotive paint is facilitated in terms of both the accuracy and speed of the analysis as a result of this coupling. The library searching system consists of two separate but interrelated components: search prefilters to cull the library spectra to a specific assembly plant or assembly plants and a cross correlation library searching algorithm to identify spectra most similar to the unknown in the set of spectra identified by the search prefilters as potential matches for the unknown. As the size of the library is culled for a specific match, the search prefilters increase both the selectivity and accuracy of the search.

5.2. Data Set

Thirty-two automotive paint samples from vehicles sold in North America between 2000 and 2006 were obtained from the Royal Canadian Mounted Police Forensic Services Laboratory. Each paint sample was from a metallic automobile component. Paint samples from plastic substrates were excluded as these components are often not painted in the same plant where the vehicle was assembled.

The colors of the paint samples were white, red, blue, silver and black. Thin cross sections of each paint chip in the range of 4-7μm for infrared microscopy were obtained using a microtome. The thirty-two paint samples obtained from cars and trucks spanned six different manufacturers: Chrysler, Ford, General Motors, Toyota, Honda, and Nissan. The PDQ identification number, the make, line and model as well as the vehicle type (car or truck) are listed in Table 5.1 for the transmission infrared microscopy data set.

Each paint sample was removed from the metal substrate using a shark knife, washed with methanol to remove dirt and particulate matter, and cast in thirty minute epoxy or placed between two rigid polyethylene plastic pieces to prepare thin sections which

involved exposing the edge of the paint sample. Each paint sample was positioned in the microtome to ensure that a thin cross section cut by the microtome contained all four paint layers. For automotive paint samples cast in blue light epoxy, mixing of the spectral features from the epoxy with the clear coat or e-coat layers occurred in 13 of these 32 paint samples. For these 13 samples, the entire procedure (including the sectioning of the embedded sample) was repeated several times with the same results. We believe this problem is linked to the compression of the cross sectioned paint sample by the epoxy, which caused a decrease in the thickness of each layer of the automotive paint. For an OEM automotive paint system, embedding a paint sample in an epoxy may be problematic for some samples when one or more layers are too thin.

Table 5.1. Paint Samples Analyzed by Transmission Infrared Microscopy

PDQ Number	Make	Line/Model	Type
UAZP00412	Chrysler	DOD/RAM	Truck
UAZP00421	Chrysler	JEE/JBT	Car
UAZP00451	Chrysler	CHR/CND	Car
UAZP00569	Chrysler	DOD/RAM	Truck
UAZP00600	Chrysler	DOD/NEO	Car
UAZP00401	Chrysler	DOD/DUR	Truck
UAZP00342	Ford	FOR/FOC	Car
UAZP00404	Ford	FOR/EPR	Car
UAZP00467	Ford	FOR/ECP	Car
UAZP00596	Ford	FOR/MUS	Car
UAZP00477	Ford	FOR/MUS	Car
UAZP00436	General Motors	CHE/CTA	Car
UAZP00271	General Motors	CHE/CTA	Car
UAZP00507	General Motors	CHE/TBZ	Car
UAZP00331	General Motors	CHE/SUB	Car
UAZP00499	General Motors	PON/BON	Car
UAZP00565	General Motors	BUI/LUC	Car
UAZP00729	Honda	Honda/CR-V	Car
UAZP00277	Honda	Honda/Odyssey	Car
CONT00726	Honda	Honda/Pilot	Car
CONT00736	Honda	Honda/Accord	Car
UAZP00730	Honda	Honda/Civic	Car

Table 5.1. Paint Samples Analyzed by Transmission Infrared Microscopy (Continue)

PDQ Number	Make	Line/Model	Type
UAZP00745	Nissan	Nissan/Titan	Car
UAZP00731	Nissan	Nissan/Murano	Car
UAZP00527	Nissan	Nissan/Altima	Car
UAZP00537	Nissan	Nissan/Pathfinder	Car
UAZP00381	Toyota	Toyota/Camry	Car
UAZP00313	Toyota	Toyota/Camry-Solara	Car
UAZP00733	Toyota	Toyota/Camry	Car
UAZP00561	Toyota	Toyota/Tacoma	Car
UAZP00440	Nissan	Nissan/Altima	Car

5.3. Results and Discussion

This section of the chapter is divided into three subsections. The first subsection is the reconstruction of the IR spectra of the individual paint layers of each sample from the spectral line maps using alternating least squares. The second and third subsections focus on the application of the prototype pattern recognition assisted infrared library search system using the reconstructed IR spectra of the automotive paint layers to identify the make and model of the vehicle from which the paint sample originated. All thirty-two of the original paint samples without epoxy were analyzed. Only twenty-seven of the thirty-two original paint samples were analyzed using epoxy resin because there was an insufficient amount of sample remaining after the analysis was completed in transmission and ATR mode with the unembedded paint samples.

5.3.1. Multivariate Curve Resolution

The thirty-two paints samples used in this study comprised a test bed to cross validate the multivariate curve resolution procedure discussed in the previous chapter. The presence or absence of epoxy did not change the method used to generate the sample line maps. Furthermore, the cross section was typically angled (whether in the presence or

absence of epoxy) as a result of the small size of the paint chip and the difficulty associated with the placement of the sample in the microtome. For this reason, the line map of each sample was generated from a cut taken along the diagonal of the paint chip to maximize the number of relevant spectra in the map. Earlier problems encountered in the generation of line maps with respect to these small samples were addressed through standardization of the method for taking the diagonal.

The cut sample to be scanned was positioned under a Thermo Nicolet iN 10 MX microscope equipped with a liquid nitrogen cooled MCT detector and the line map was generated along the diagonal of the sample. As in the previous experiment, all spectra within the line map had the region from 2280 cm^{-1} to 2400 cm^{-1} replaced with a blank line to negate the effects of CO_2 gas, and the spectral region below 748cm^{-1} was discarded because it was often too noisy. After this preprocessing was completed, spectra were extracted from the sample line map and examined for artifacts. Spectra with very large or small peak intensities were discarded from the line map as were spectra that exhibited signs of peak shifting. The paint samples generally did not exhibit peak shifting as was observed in our initial set of test samples analyzed using a Nicolet Magna-IR 550 Series II Spectrometer coupled to a Nic-Plan Analytical IR Microscope.

Using the entire spectral range (4000 cm^{-1} to 748 cm^{-1}) was more effective for decatenation by ALS than selecting a wavelength subset, e.g., the fingerprint region, as more information about how each layer changed as a function of sample position during scanning was obtained when the entire spectrum was subject to decatenation. Figures 5.1 through Figures 5.6 show the ALS spectral reconstructions of the clear coat, surfacer-primer, and e-coat layers from six automotive paint samples not encapsulated in epoxy.

Each paint sample is from a different manufacturer and is representative of the quality of the ALS reconstructions from that manufacturer. The clear coat, surfacer-primer and e-coat layers were successfully reconstructed for all six paint samples. All reconstructed IR spectra were good matches for the same paint sample in the PDQ library.

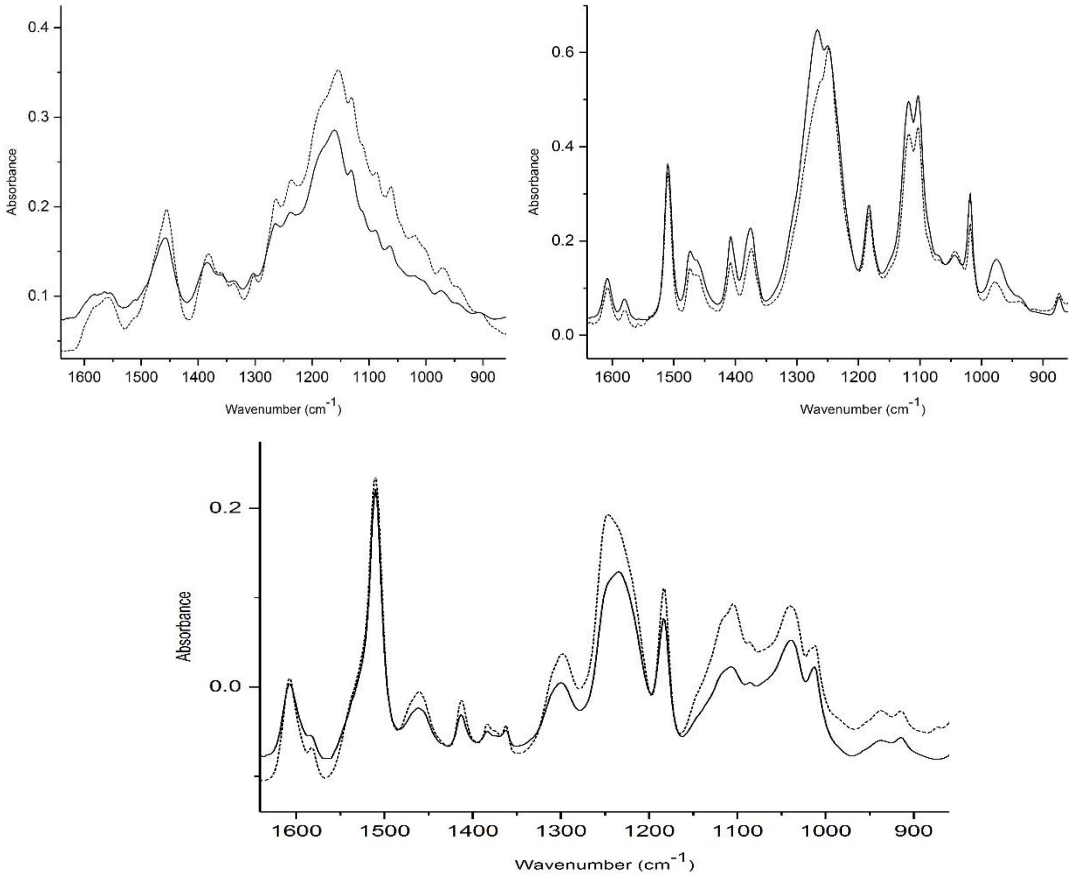

Figure 5.1. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00421-Chrysler Jeep) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

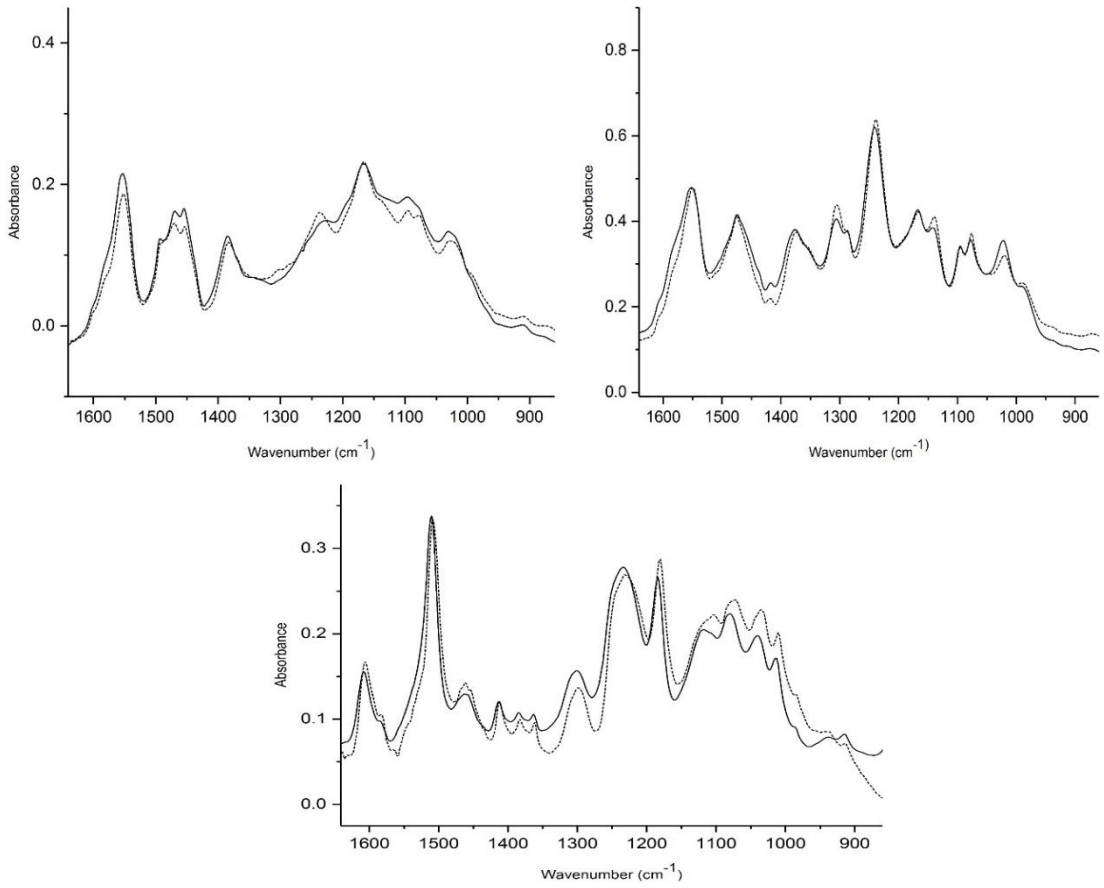

Figure 5.2. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00477-Ford Mustang) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

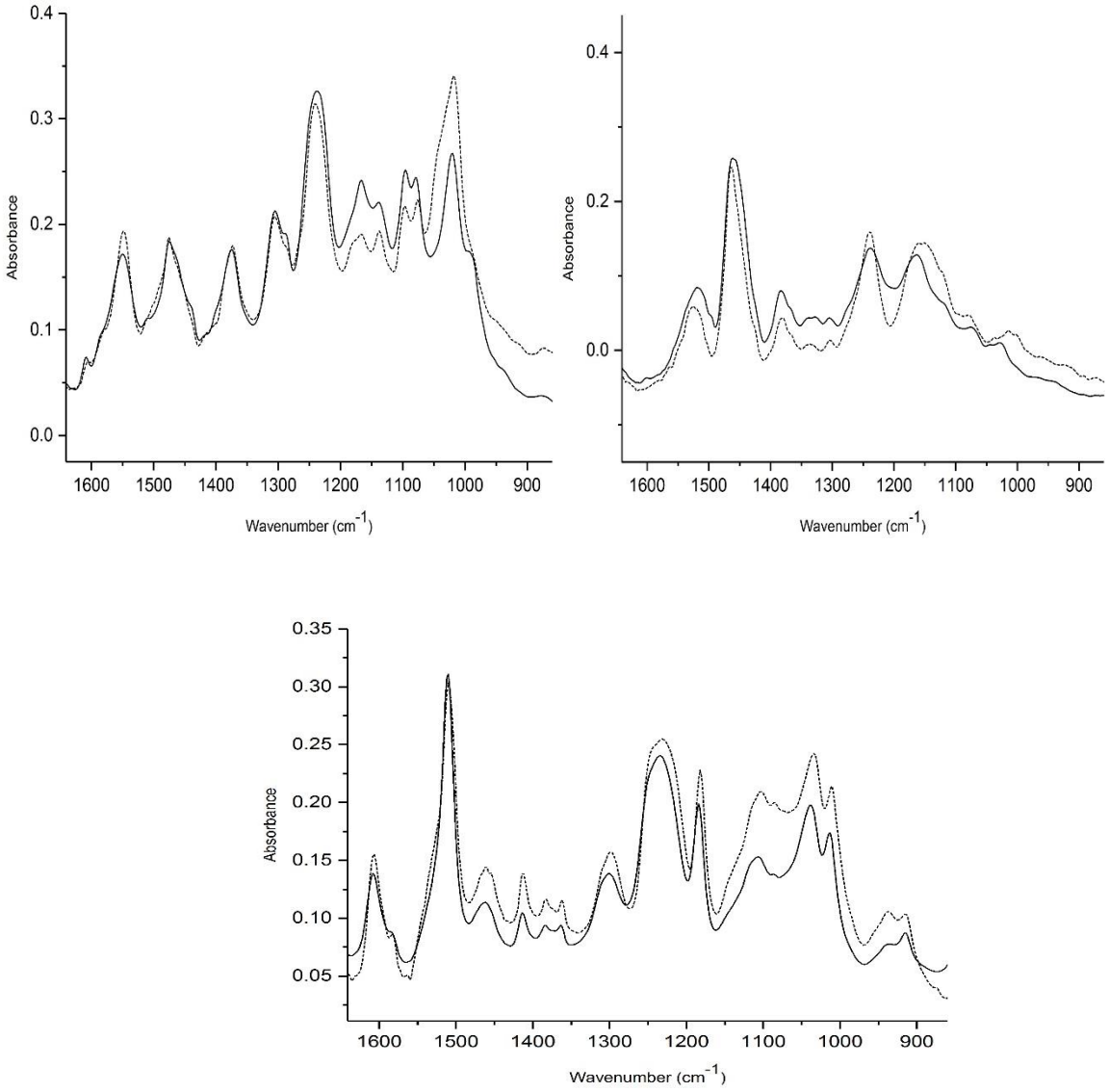

Figure 5.3. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00331-General Motors Chevrolet Suburban) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

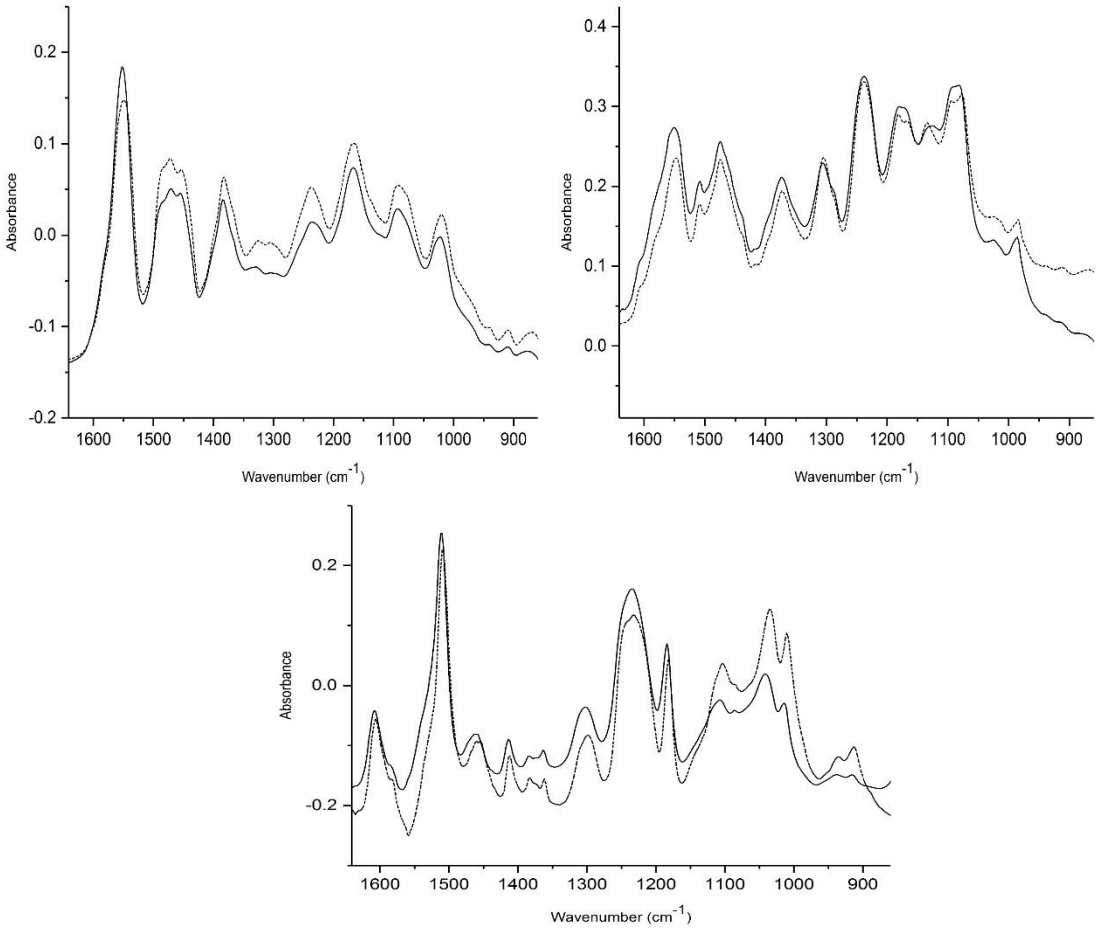

Figure 5.4. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (CONT00726-Honda Pilot) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

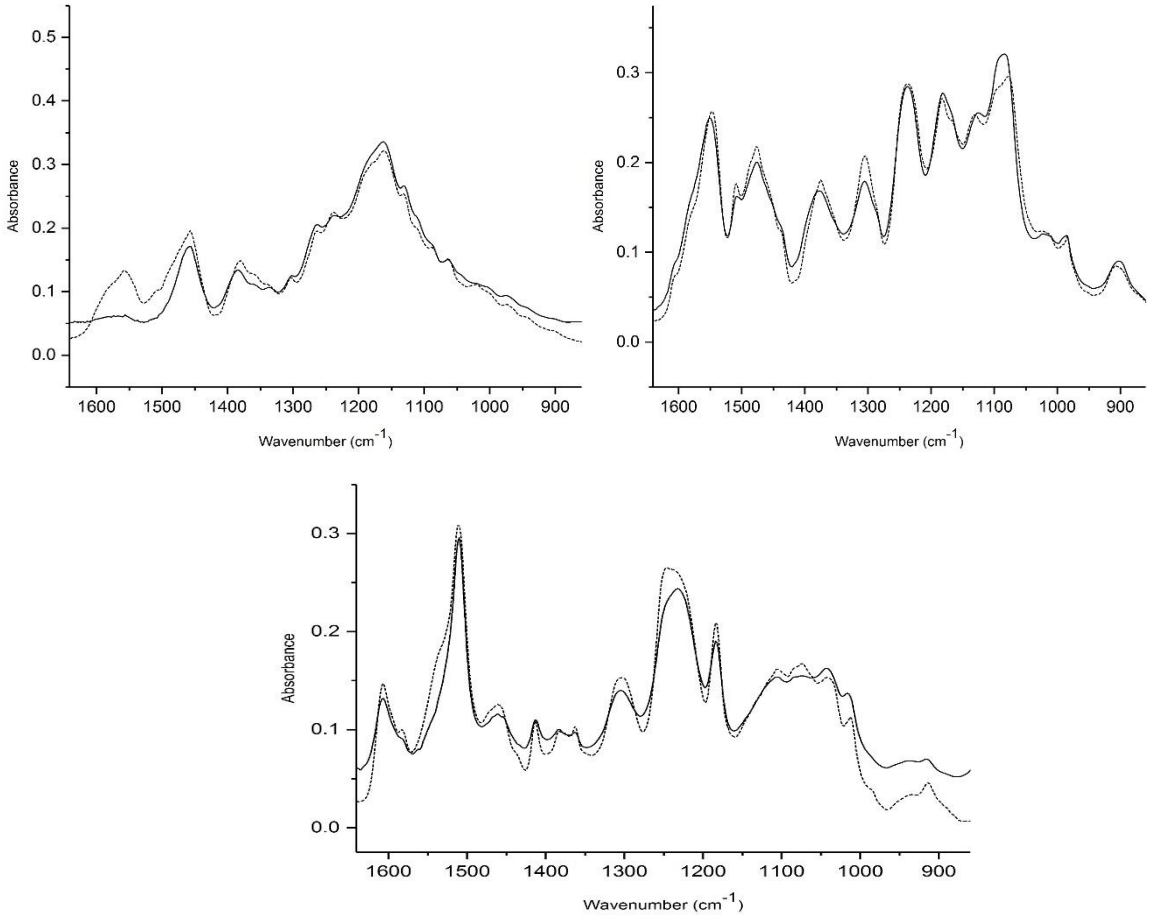

Figure 5.5. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00527-Nissan Altima) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

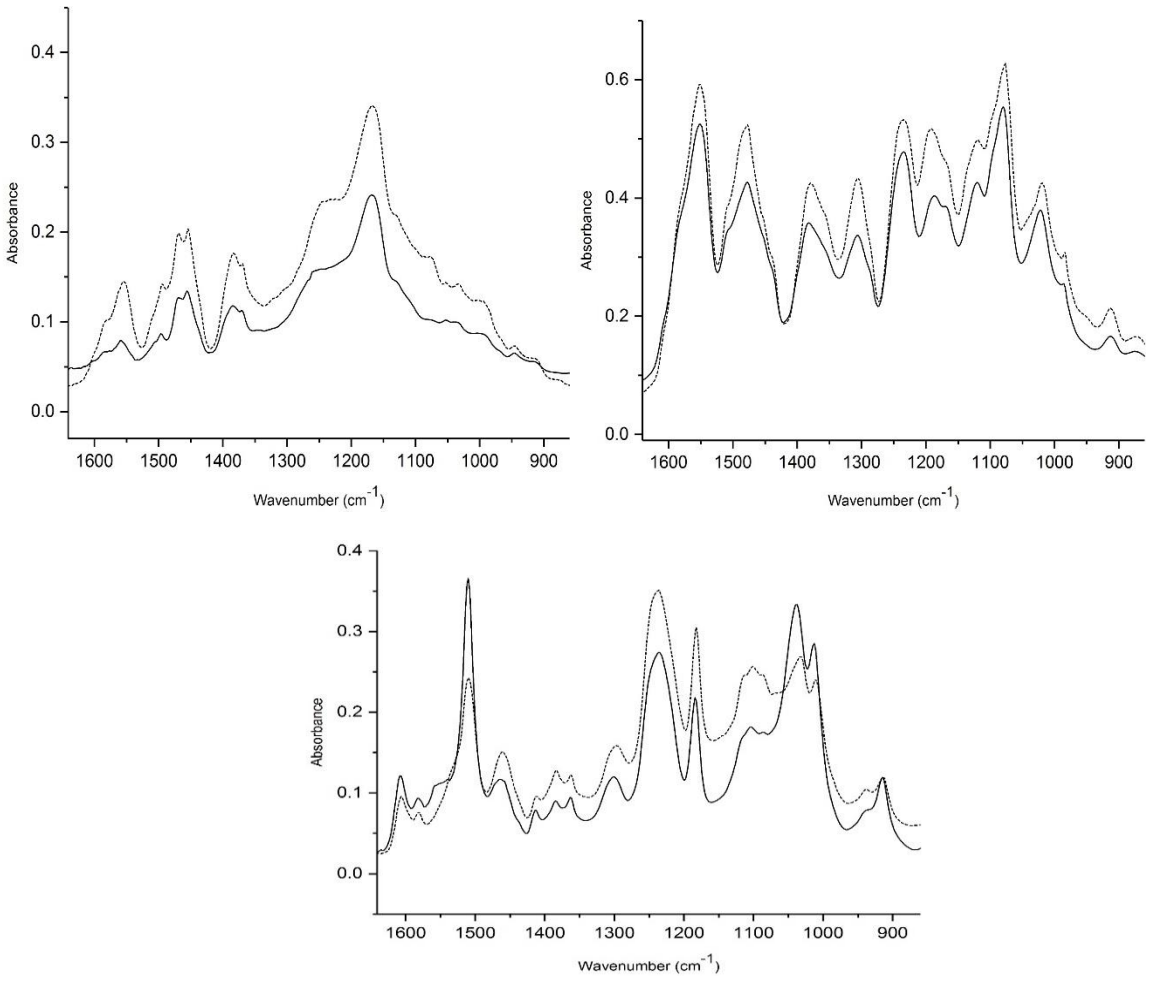

Figure 5.6. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00561-Toyota Tacoma) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

Figures 5.7 through Figures 5.12 show the ALS spectral reconstructions of the clear coat, surfacer-primer, and e-coat layers from six automotive paint samples encapsulated in epoxy. Each paint sample is from a different manufacturer and is representative of the quality of the ALS reconstructions from that manufacturer. The clear coat, surfacer-primer and e-coat layers were successfully reconstructed for all six paint samples and were good matches for the same paint sample in the PDQ library.

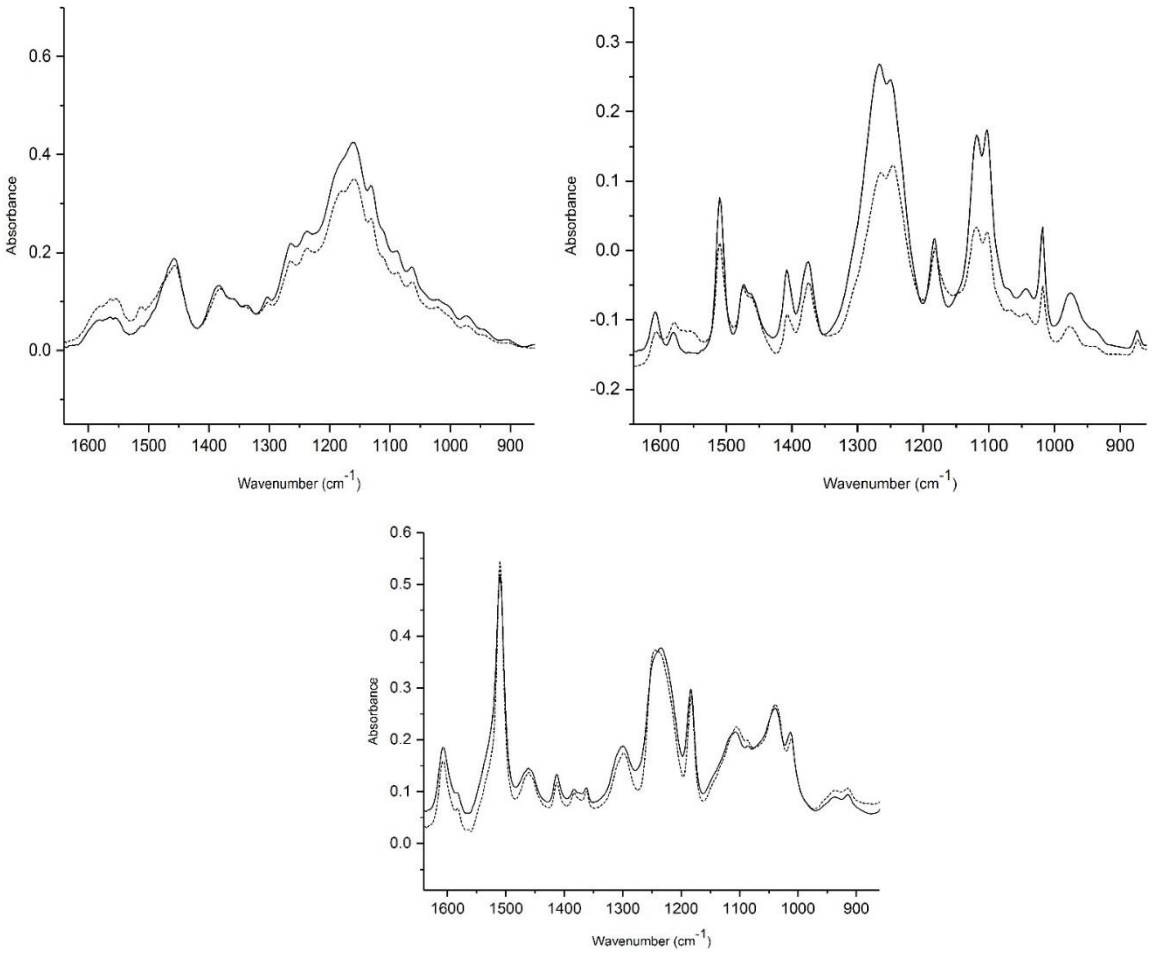

Figure 5.7. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00421-Chrysler Jeep) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

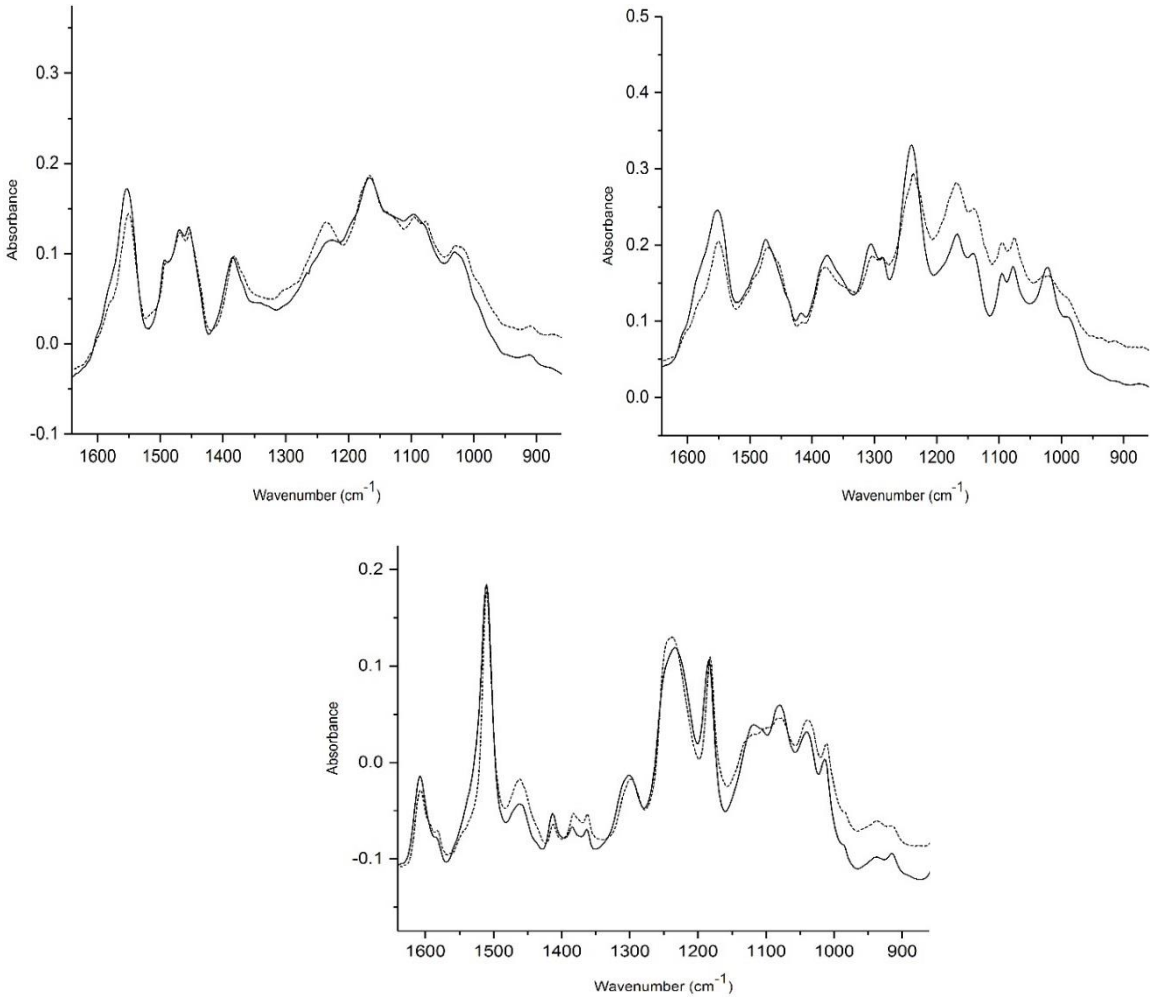

Figure 5.8. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00477-Ford Mustang) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

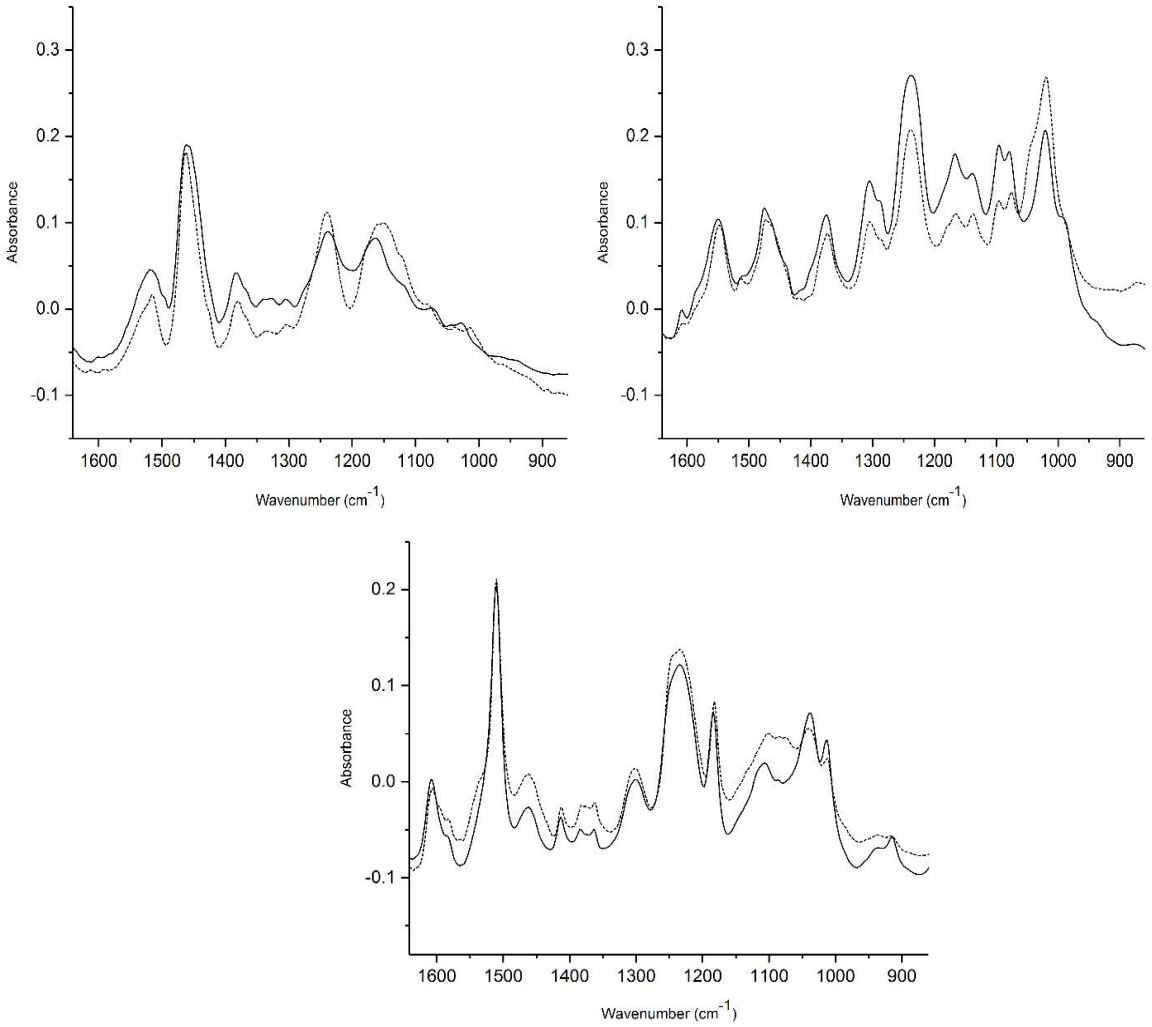

Figure 5.9. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00331-General Motors Chevrolet Suburban) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

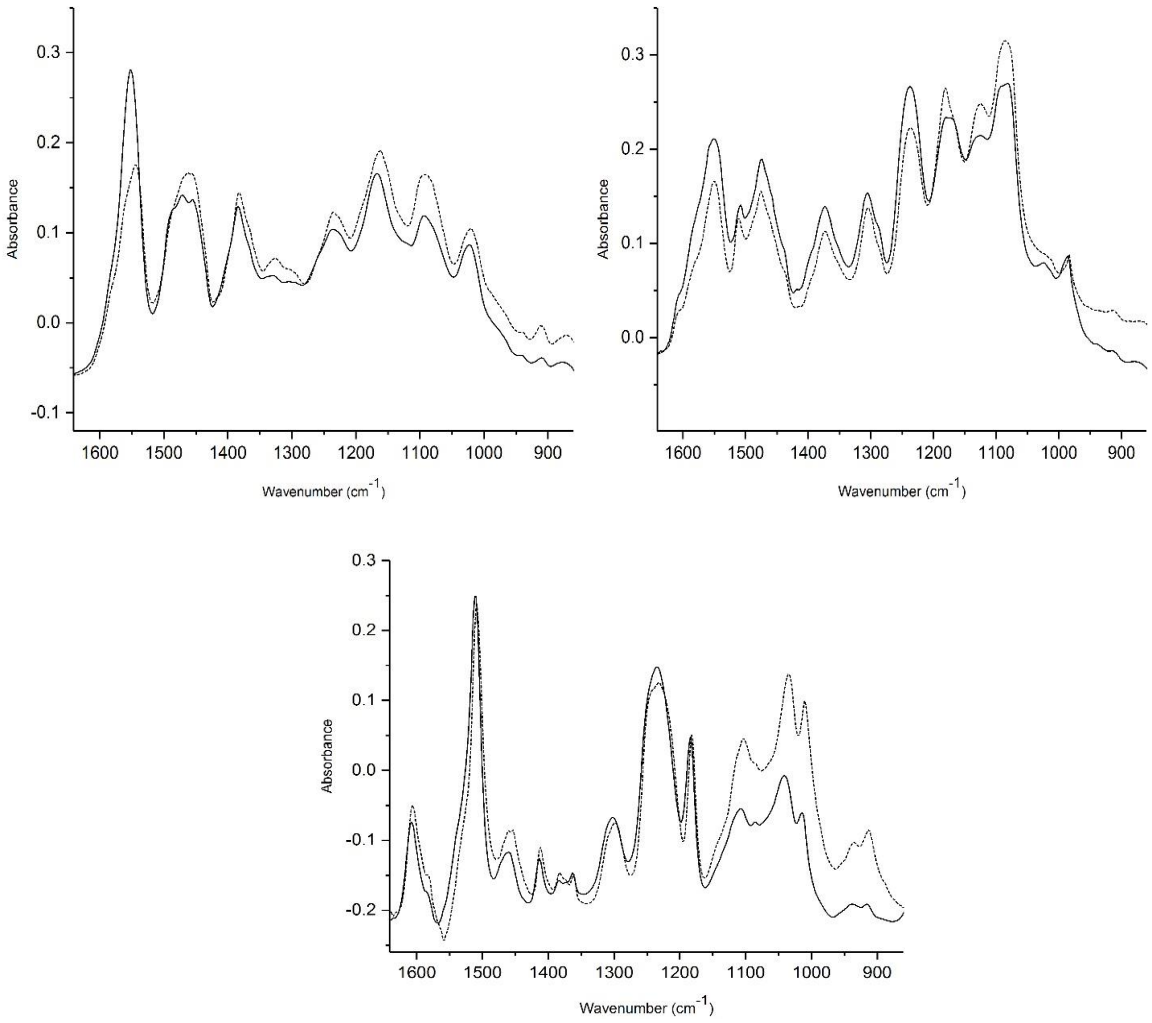

Figure 5.10. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (CONT00726-Honda Pilot) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

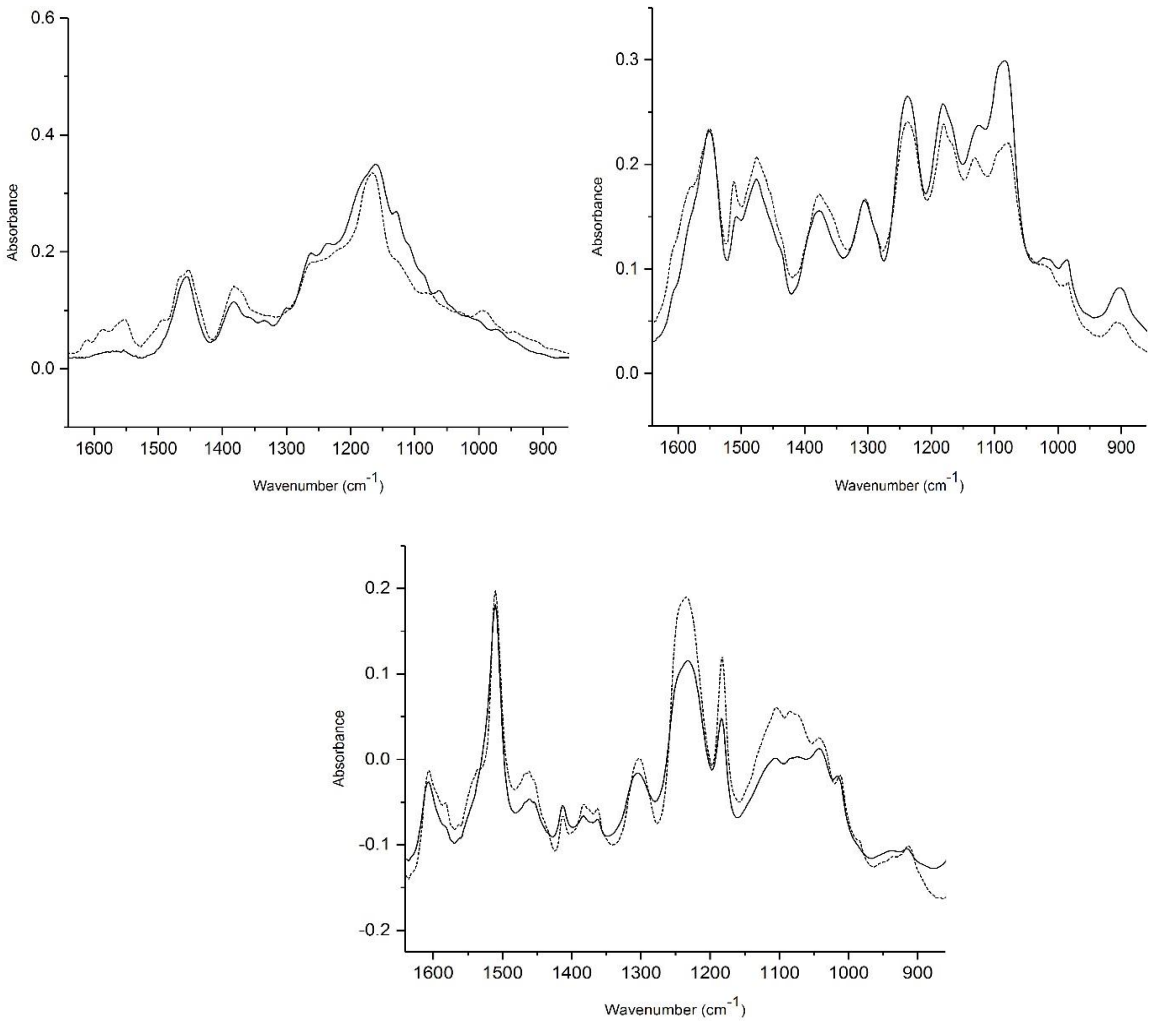

Figure 5.11. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00527-Nissan Altima) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

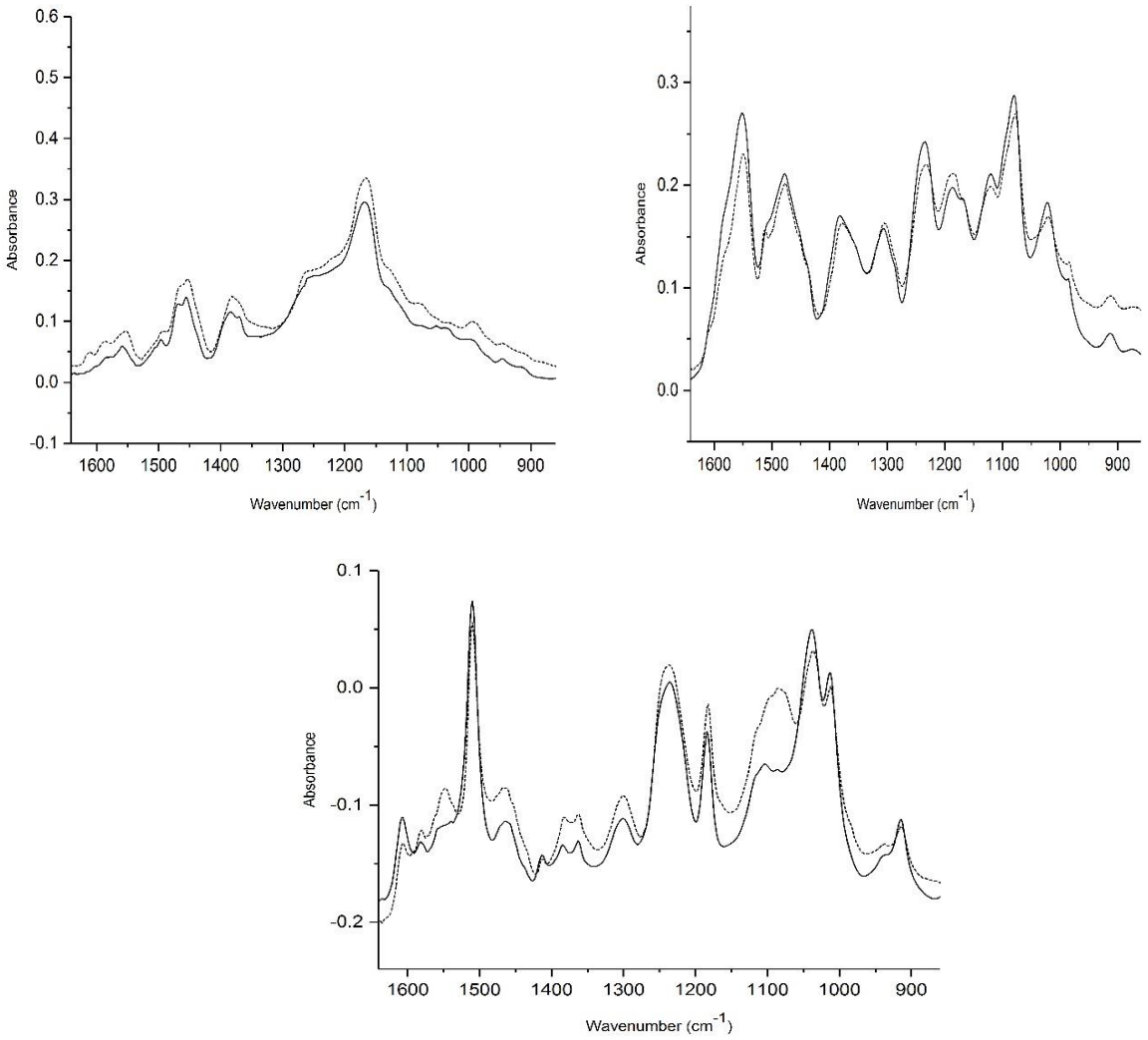

Figure 5.12. Comparison of the reconstructed IR spectrum (dashed line) to the IR spectrum of the actual paint sample (UAZP00561-Toyota Tacoma) in the PDQ spectral library. a) Clear coat layer, b) surfacer-primer layer, and c) e-coat layer.

ALS reconstructions of each paint layer for the thirty two automotive paint samples were evaluated using OMNIC library search routines that were configured as correlation for the search type with Happ-Genzel apodization. All library searches were restricted to the spectral region between 1641 cm^{-1} and 860 cm^{-1}. The quality of each search was assessed using the hit quality index (HQI). A library search was considered to be successful when the actual paint sample or the correct line and model of the vehicle from which the

paint sample originated was included in the top five hits of the search. Each library searched was the same manufacturer (e.g., General Motors) and the same production year range (e.g., 2000-2006) as the automotive paint **s**ample from which the reconstructed IR spectra were obtained. Table 5.2 summarizes the library search results for the paint samples without epoxy. For the clear coats, 29 of the 32 paint samples were correctly matched, whereas for the surfacer primer layer it was 31 out of 32 and 27 out of 32 for the e-coat layer. Table 5.3 summarizes the library search results for the paint samples with epoxy. For the clear coats, 23 of the 27 paint samples were correctly matched, whereas for the surfacer primer layer it was 22 out of 27 and 14 out of 27 for the e-coat layer.

An examination of the PDQ library spectra and the reconstructed IR spectra of the same paint sample for the IR spectra not correctly matched reveals large peak shifts (approximately $10cm^{-1}$) for some vibrational modes. When IR spectra from a high pressure diamond cell were compared to spectra obtained at ambient pressure from an IR microscope, frequency shifts for some modes were observed. Emmons et al., [1] attributed these observed frequency shifts to the removal of void spaces in the polymer (paint layer) which occurred during the compression of the paint sample by the diamond cell. For library searching algorithms based on correlation, these shifts will reduce the HQI value of a spectral match. To ensure accurate spectral library searching, it is necessary to address this problem. A brief summary of the solution to this problem is described in the next subsection.

Table 5.2. Library Search Results for the Unembedded Paint Samples

PDQ Number	Manufacturer	Clear Coat		Surfacer-Primer		E-Coat	
		HQI	% Match	HQI	% Match	HQI	% Match
UAZP00412	Chrysler	8	96.80	1	97.44	1	98.63
UAZP00421	Chrysler	1	95.49	1	96.28	3	98.58
UAZP00451	Chrysler	9	92.77	1	98.05	4	95.91
UAZP00569	Chrysler	1	94.56	1	98.65	4	96.47
UAZP00600	Chrysler	1	91.54	1	93.85	1	93.36
UAZP00401	Chrysler	1	97.81	1	98.61	1	97.30
UAZP00342	Ford	1	98.61	2	95.63	-	-
UAZP00404	Ford	1	98.18	2	95.67	3	93.83
UAZP00467	Ford	1	97.31	1	95.53	1	94.77
UAZP00596	Ford	1	98.80	1	92.39	1	98.11
UAZP00477	Ford	1	98.14	1	96.37	1	91.36
UAZP00436	General Motors	1	96.37	1	92.88	1	97.06
UAZP00271	General Motors	1	98.15	5	90.50	-	-
UAZP00507	General Motors	8	98.14	4	86.05	1	93.11
UAZP00331	General Motors	1	97.70	2	94.44	1	96.58
UAZP00499	General Motors	1	98.52	5	92.36	-	-
UAZP00565	General Motors	1	97.07	1	96.25	5	94.31
UAZP00729	Honda	1	96.52	9	91.93	1	97.06
UAZP00277	Honda	1	96.77	1	98.00	3	94.87
CONT00726	Honda	2	95.96	1	96.48	1	90.84
CONT00736	Honda	5	87.90	1	97.38	1	96.69
UAZP00730	Honda	2	95.28	1	97.88	2	94.38
UAZP00440	Nissan	1	98.10	1	95.19	1	89.12
UAZP00745	Nissan	1	96.38	1	85.73	1	97.47
UAZP00731	Nissan	1	94.83	1	95.49	1	90.05
UAZP00527	Nissan	1	96.93	1	95.68	1	97.07
UAZP00537	Nissan	2	91.07	3	94.77	-	-
UAZP00381	Toyota	2	92.78	5	96.55	3	95.96
UAZP00313	Toyota	1	94.74	1	92.19	-	-
UAZP00733	Toyota	1	96.26	2	98.08	4	94.96
UAZP00561	Toyota	1	97.51	1	96.85	2	92.30
UAZP00484	Toyota	1	97.42	1	90.97	1	94.35

Table 5.3. Library Search Results for the Embedded Paint Samples

PDQ Number	Manufacturer	Clear Coat		Surfacer-Primer		E-Coat	
		HQI	% Match	HQI	% Match	HQI	% Match
UAZP00412	Chrysler	1	97.13	4	98.09	1	97.63
UAZP00421	Chrysler	1	97.03	1	96.39	9	97.98
UAZP00451	Chrysler	5	96.03	1	96.77	7	97.90
UAZP00569	Chrysler	1	90.04	1	98.44	1	96.28
UAZP00600	Chrysler	1	98.05	3	98.74	10	97.54
UAZP00401	Chrysler	1	97.60	1	96.86	1	96.60
UAZP00342	Ford	1	96.47	1	93.95	8	97.36
UAZP00596	Ford	1	97.32	2	94.10	1	97.03
UAZP00436	General Motors	5	97.87	3	92.81	3	96.86
UAZP00507	General Motors	-	-	2	87.60	-	-
UAZP00331	General Motors	3	97.10	1	95.10	1	97.70
UAZP00565	General Motors	3	94.50	4	90.92	8	68.53
UAZP00277	Honda	1	97.46	1	97.49	3	98.15
CONT00736	Honda	-	-	1	97.11	1	94.00
UAZP00730	Honda	1	95.60	1	97.09	3	96.94
UAZP00440	Nissan	1	92.73	1	96.12	1	96.53
UAZP00731	Nissan	2	89.36	6	62.45	6	79.56
UAZP00527	Nissan	4	92.55	1	92.72	1	94.20
UAZP00537	Nissan	-	-	7	85.73	-	-
UAZP00381	Toyota	1	90.73	-	-	1	96.96
UAZP00733	Toyota	4	87.55	-	-	-	-
UAZP00484	Toyota	-	83.99	1	95.39	3	96.01
UAZP00385	Nissan	1	97.20	1	95.75	-	-
UAZP00404	Ford	1	92.14	2	85.93	9	95.14
UAZP00477	Ford	2	94.06	1	86.98	2	96.62
UAZP00729	Honda	2	85.46	-	-	10	92.75
UAZP00745	Nissan	2	90.73	2	89.27	-	-

5.3.2 Search Prefilters for Pattern Recognition Assisted Infrared Library Searching

The frequency shifts observed for some vibrational modes posed an even greater problem when using pattern recognition techniques. For almost all library searching algorithms, the Euclidean distance or the correlation coefficient between pairs of spectra is computed, whereas with pattern recognition methods, the metric that is computed minimizes within-source variability and maximize between-source variability (e.g., the assembly plant from which the vehicle originated). This more stringent requirement for spectral library matching necessitates higher quality data containing fewer artifacts. To solve the problem of frequency shifts encountered with some vibrational modes, IR transmission spectra from the PDQ database (generated using a high pressure diamond transmission cell) and the transmission IR microscope (generated at ambient pressure using a BaF_2 cell) were transformed to ATR spectra using an ATR simulation algorithm previously developed by Lavine [2, 3]. Using this correction algorithm, the large frequency shifts encountered for some vibrational modes in the polymer were diminished when applying this conversion to IR spectra in the PDQ library and to the paint samples analyzed by the infrared microscope. We attribute this success to removing the effect of the refractive index on the IR spectra. The correction algorithm uses a set of six equations to standardize the real and imaginary components of the refractive index for the IR spectra.

The reconstructed IR spectra of each paint sample were analyzed using a prototype pattern recognition library search engine [4-7] for multiple automotive paint layers (clear coat, surfacer primer, and e-coat layers) consisting of prefilters developed from the clear coat, surfacer-primer and e-coat layers for 1652 OEM paint systems spanning six manufacturers (General Motors, Ford, Chrysler, Honda, Nissan, and Toyota) within a

limited production year (2000-2006). A hierarchical classification scheme was utilized to identify the make and model of the vehicle from the reconstructed IR spectra. A search prefilter was developed to differentiate automotive paint samples by manufacturer. For each manufacturer, search prefilters were developed to identify the assembly plant (and hence the line and model of the vehicle) from FTIR spectra of OEM paint system using the clear coat, surfacer-primer and e-coat layers.

5.3.2.1. Methodology and Data Preprocessing for Search Prefilter Development

For pattern recognition analysis, each transmission spectrum was normalized to unit length. The discrete wavelet transform [8] using the 8sym6 mother wavelet (Symlet wavelet family, sixth smallest filter size, eighth level of decomposition) was applied to the fingerprint region (1641 cm^{-1} to 680 cm^{-1}) of each layer using the Matlab Wavelet toolbox 3.0.4 (The Mathworks Inc.). The Symlet 6 mother wavelet was chosen because the shape of its scaling function closely matched the shape of the bands comprising the IR spectra of the automotive paints. Three sets of wavelet coefficients were concatenated to form the sample pattern vectors used by the search prefilters as shown in Figure 5.13. Wavelet coefficients from the lower levels of decomposition were retained, resulting in 3426 wavelet coefficients per paint sample (i.e., 1142 coefficients each from the clear coat, surfacer–primer, and e-coat layers). Prior to pattern recognition analysis, the wavelet transformed spectra were autoscaled to ensure that each coefficient had a mean of zero and a standard deviation of one throughout all transformed IR spectra. Autoscaling removed any inadvertent weighing of the data that otherwise would occur due to differences in the magnitude among the wavelet coefficients comprising the spectral data.

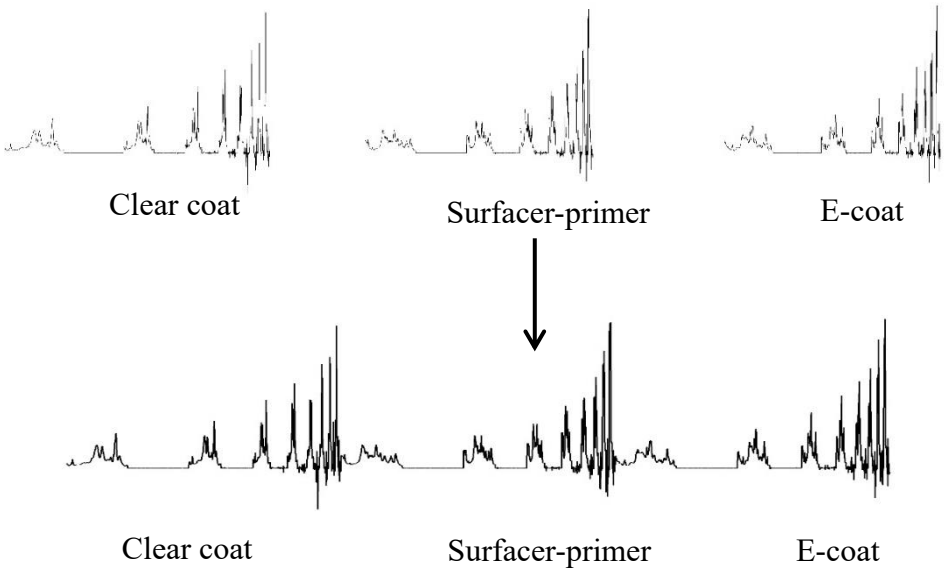

Figure 5.13. Clear coat, surfacer-primer, e-coat and fused wavelet preprocessed FT-IR data.

The automotive paint database used in this study consisted of 1652 paint samples from six manufacturers: General Motors (19 assembly plants), Chrysler (15 assembly plants), Ford (25 assembly plants), Honda (6 assembly plants), Nissan (6 assembly plants), and Toyota (5 assembly plants). These six manufacturers account for 80% of the vehicles purchased in North America.

Before the paint samples in this database were analyzed by the genetic algorithm for pattern recognition, they were investigated for outliers by examining principal component (PC) plots of each paint layer from each assembly plant and flagging samples that appeared discordant in the corresponding PC plots. These discordant observations were compared to the average IR spectrum of each layer for the given assembly plant. If the IR spectrum of the observation in question differed markedly from the average IR spectrum of the assembly plant in question, the sample was designated as an outlier and

removed from the analysis. Many of these discordant observations were paint samples obtained from replaceable automotive substrates that did not have the original manufacturer's paint system. For other discordant observations, the paint layers may have been mislabeled (e.g., primer surfacer labeled as the e-coat layer). The database of 1652 paint samples was divided into a training set of 1484 samples and a prediction set of 168 samples (see Table 5.4). Samples comprising the prediction set were selected by random lot.

Table 5.4. Automotive Paint Data

Manufacturer	Training	Prediction	Total
General Motors	408	44	452
Chrysler	350	40	390
Ford	345	39	384
Honda	126	16	142
Nissan	94	9	103
Toyota	161	20	181
Total	1484	168	1652

5.3.2.2. Manufacture Search Prefilter System

A hierarchical classification scheme was implemented to develop a search prefilter to identify the vehicle manufacturer of an intact paint chip from the IR spectra of the clear coat, surfacer–primer, and e-coat layers. The first step was to divide the automotive paint samples into two groups based on the chemical formulation of the clear coat layer. Modern automotive clear coats are either acrylic melamine styrene (singlet for the carbonyl band) or acrylic melamine styrene polyurethane (doublet for the carbonyl band). Paint samples whose clear coat layer exhibits a doublet for the carbonyl band were flagged and isolated

from the other paint samples. Nissan and Toyota were only represented by paint samples whose clear coat layer was acrylic melamine styrene (singlet for the carbonyl band).

A search prefilter was developed to classify IR spectra by vehicle manufacturer for paint samples with an acrylic melamine styrene polyurethane clear coat layer. Since the clear coat layer of the Nissan and Toyota paint samples was acrylic melamine styrene, this search prefilter was limited to four manufacturers: General Motors, Chrysler, Ford, and Honda. Table 5.5 summarizes the 209 training set samples analyzed by the pattern recognition GA in this phase of the study. Figure 5.14 shows a plot of the two largest PCs of the 3426 wavelet coefficients of the concatenated pattern vector (clear coat, surfacer–primer, and e-coat layers). Each sample is represented as a point in the PC plot of the wavelet transformed ATR spectral data. There is overlap between the four vehicle manufacturers in the PC plot.

Table 5.5. Acrylic Melamine Styrene Polyurethane

Manufacturer	**Training**	**Prediction**	**Total**
General Motors	104	16	120
Chrysler	65	6	71
Ford	23	3	26
Honda	17	2	19
Total	209	27	236

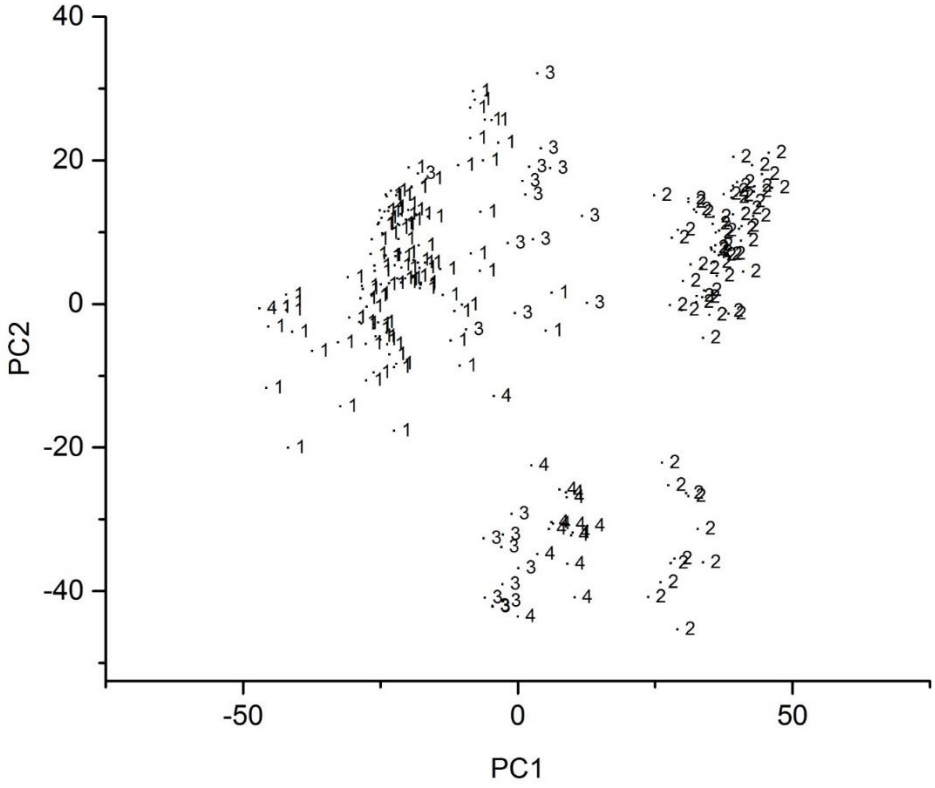

Figure 5.14. Principal component plot of the 3426 wavelet coefficients and the 209 concatenated IR spectra comprising the training set for those samples whose clear coats are defined by the formulation acrylic melamine styrene polyurethane. Each sample is represented as a point in the plot: 1 = GM, 2 = Chrysler, 3 = Ford, and 4 = Honda.

The pattern recognition GA identified wavelet coefficients characteristic of the manufacturer by sampling key feature subsets, scoring their PC plots, and tracking those samples or classes (i.e., automotive manufacturers) that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the pattern recognition GA identified 48 wavelet coefficients whose PC plot (Figure 5.15) shows clear delineation of the training set samples on the basis of automotive manufacturer. Projecting the 27 prediction set samples onto the PC plot developed from the 209 training set samples and the 48 wavelet coefficients identified by the pattern recognition GA showed that each projected prediction set sample was located

in a region of the PC plot (Figure 5.16) with samples from the same automotive manufacturer.

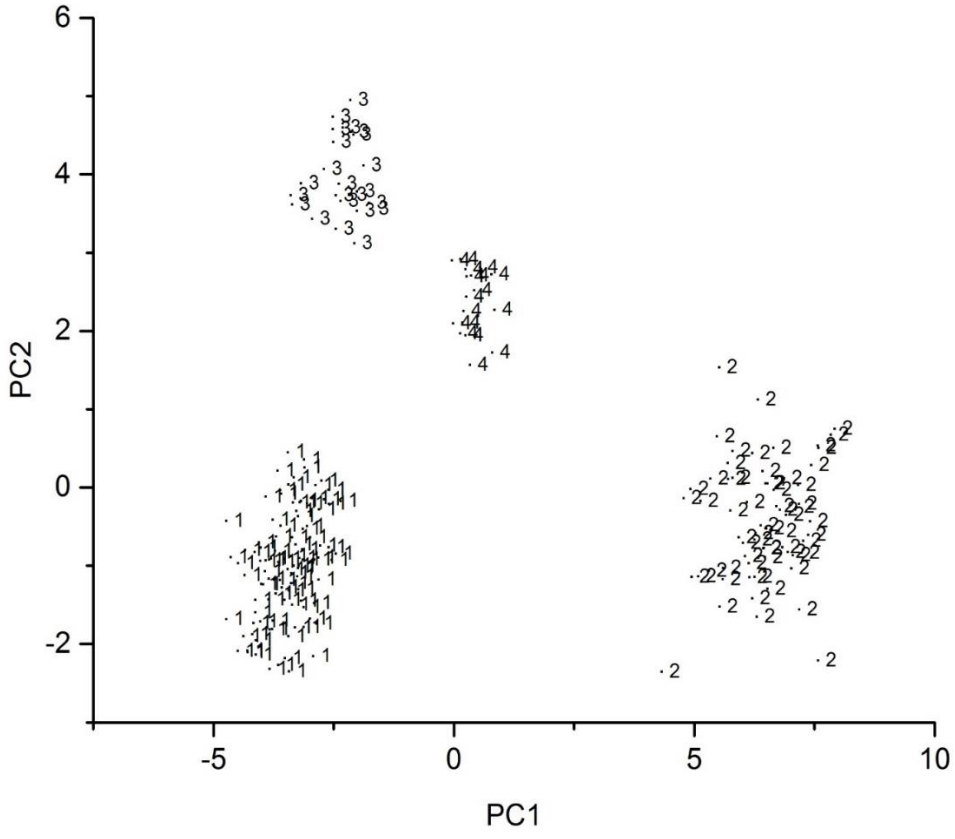

Figure 5.15. PC plot of the 48 wavelet coefficients identified by the pattern recognition GA and the 209 concatenated IR spectra comprising the training set. Training set: 1 = GM, 2 = Chrysler, 3 = Ford, 4 = Honda.

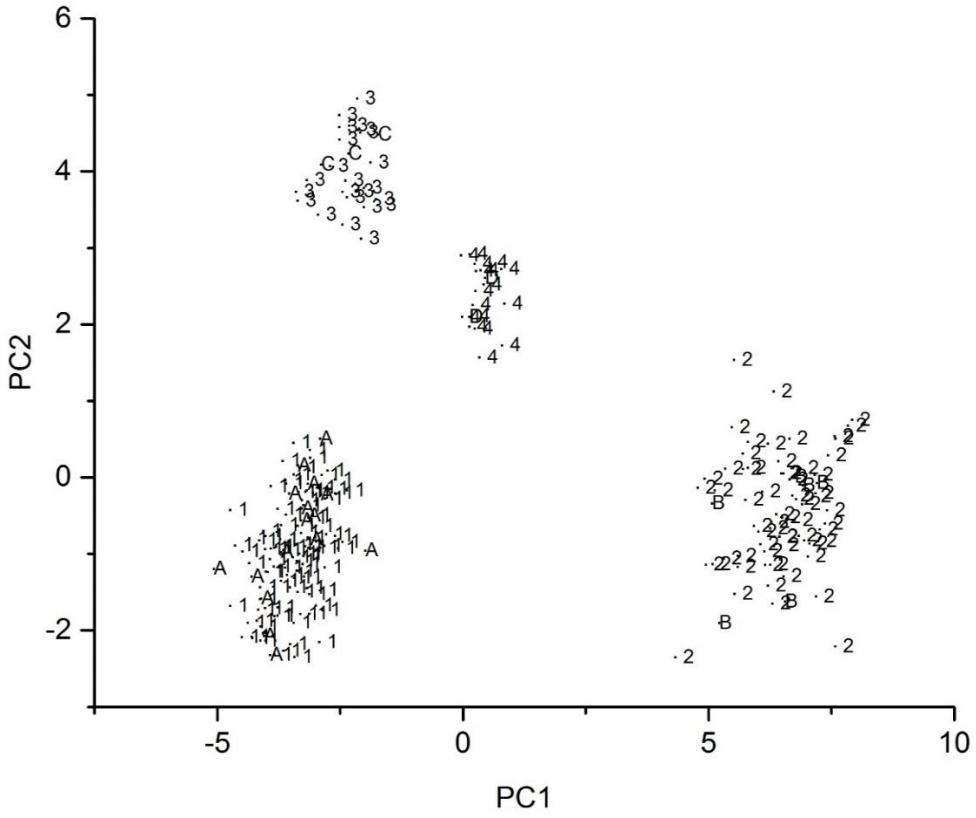

Figure 5.16. Projection of the 27 prediction set samples onto the PC plot of the 33 wavelet coefficients identified by the pattern recognition GA and the 209 concatenated IR spectra comprising the training set. Training set: 1 = GM, 2 = Chrysler, 3 = Ford, 4 = Honda. Prediction set: A = GM, B = Chrysler, C = Ford, and D = Honda.

The next step was to develop a classifier capable of discriminating IR spectra by manufacturer for paint samples that possessed an acrylic melamine styrene clear coat layer. All six manufacturers were represented by samples that exhibited a singlet for the carbonyl in their clear coat IR spectra. The 1416 paint samples in this phase of the study were divided into a training set of 1275 samples and a prediction set of 141 samples (see Table 5.6). Figure 5.17 shows a plot of the two largest PCs of the 1275 paint samples and the 3426 wavelet coefficients comprising the training set. A visual examination of the PC plot shows

two distinct clusters. One is for the Chryslers (6 assembly plants) and the other is for the other automotive manufacturers which includes some Chrysler assembly plants.

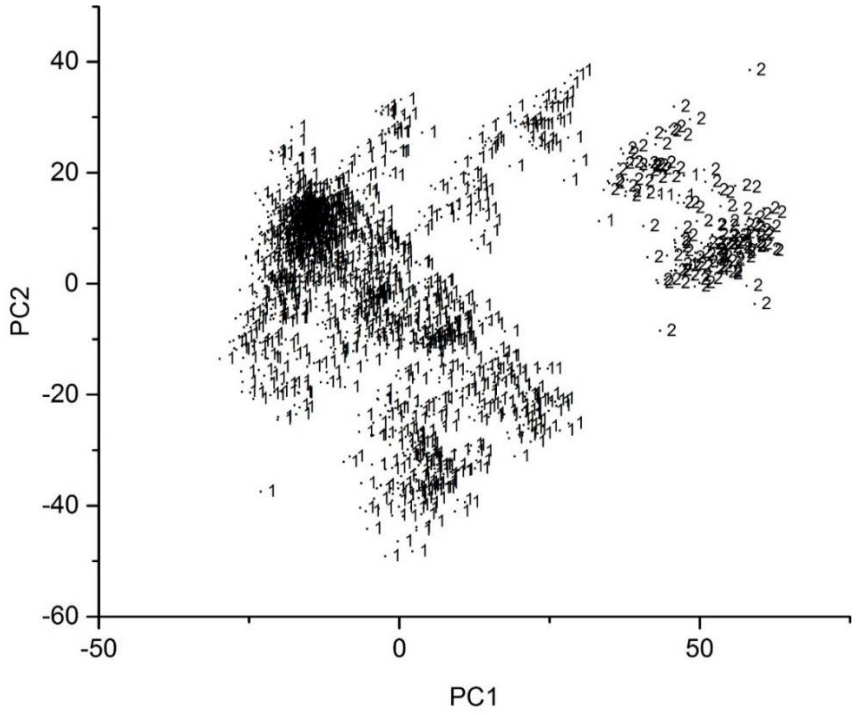

Figure 5.17. Principal component plot of the 3426 wavelet coefficients and the 1275 samples whose clear coats are formulated using acrylic melamine styrene. Each sample is represented as a point in the plot: 1 = General Motors, Chrysler, Honda, Nissan, and Toyota; 2 = Chrysler (3 plants).

Table 5.6. Acrylic Melamine Styrene

Manufacturer	**Training**	**Prediction**	**Total**
General Motors	304	28	332
Chrysler	285	34	319
Ford	322	36	358
Honda	109	114	123
Nissan	94	9	103
Toyota	161	20	181
Total	1275	141	1416

A classifier was developed using the pattern recognition GA for separating the Chryslers (6 assembly plants) from the other automotive manufacturers. Figure 5.18 shows a plot of the two largest PCs of the 1275 paint samples comprising the training set and the 19 wavelet coefficients identified by the pattern recognition GA for this two-way classification problem (Chrysler versus General Motors, Chrysler, Honda, Nissan, Ford and Toyota). The 6 assembly plants for Chrysler are well separated from the other vehicle manufacturers in the PC plot. The prediction set samples for this training set were then projected onto the PC plot of the 1275 paint samples and the 19 wavelet coefficients selected by the pattern recognition GA (see Figure 5.19). Each projected sample lies in a region of the PC plot with samples that are tagged with the same class label.

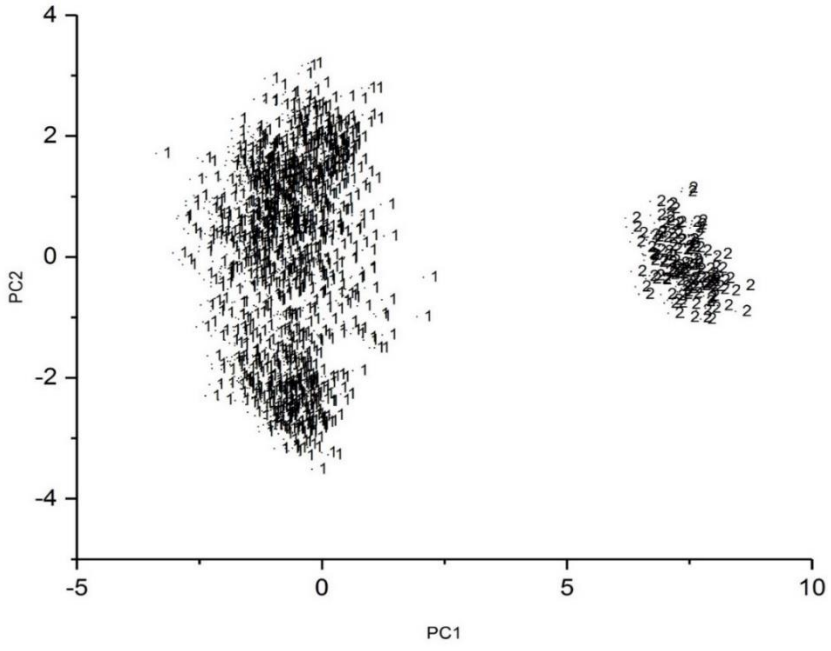

Figure 5.18. PC plot of the 1275 training set samples and the 19 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors, Chrysler, Honda, Nissan, and Toyota; 2 = Chrysler (3 plants).

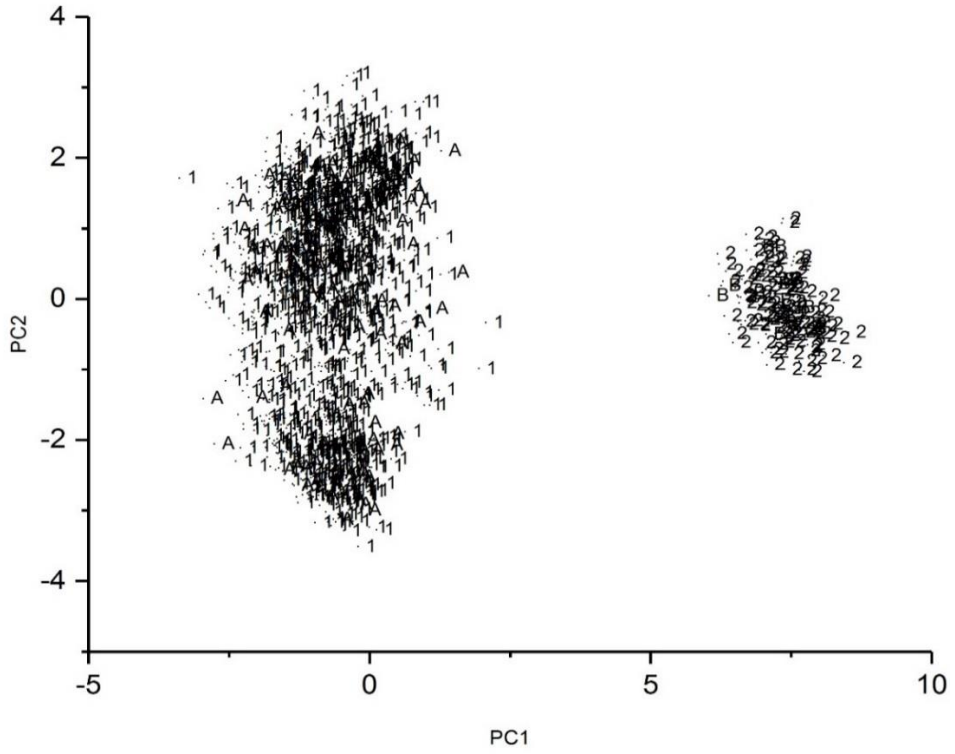

Figure 5.19. Projection of the prediction set samples onto the PC plot of the 1275 training set samples and the 19 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors, Chrysler, Honda, Nissan, and Toyota; 2 = Chrysler (3 plants). Prediction set: A = General Motors, Chrysler, Honda, Nissan, and Toyota; B = Chrysler.

A second discriminant was developed using the pattern GA for separating the 6 Chrysler and General Motors assembly plants from Toyota, Nissan, Honda, and the remaining Chrysler and General Motors automotive paint samples. Figure 5.20 shows a plot of the two largest PCs of the remaining 1135 training set samples (without the six Chrysler assembly plants) and the 45 wavelet coefficients identified by the pattern recognition GA for this classification problem: General Motors (4 assembly plants) and Chrysler (2 assembly plants) versus Honda, Nissan, Toyota, Ford and the remaining General Motors (11 assembly plants) and Chrysler (7 assembly plants) paint samples. The 6 General Motors and Chrysler assembly plants paint samples were well separated from

the others in the PC plot. Projection of the 127 prediction set samples onto the PC plot developed from the 1135 samples and the 45 wavelet coefficients identified by the pattern recognition GA is shown in Figure 5.21. All prediction set samples were located in a region of the PC map with samples that have the same class label.

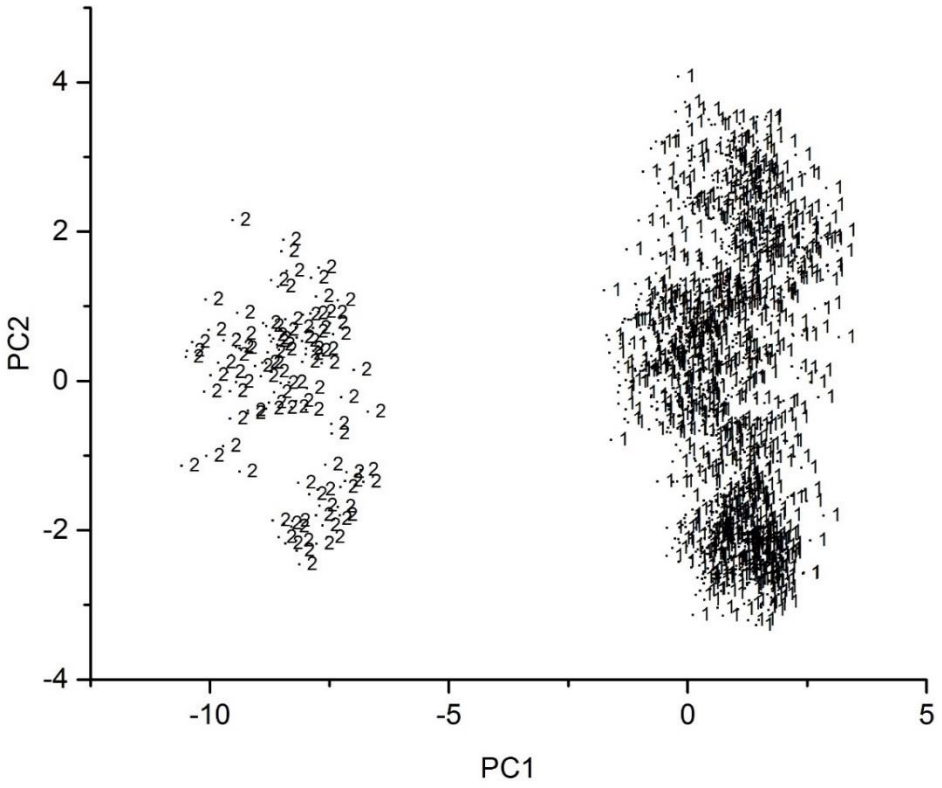

Figure 5.20. PC plot of the 1135 training set samples and the 45 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors, Chrysler, Honda, Nissan, Toyota and Ford; 2 = Chrysler (2 plants) and General Motors (4 plants).

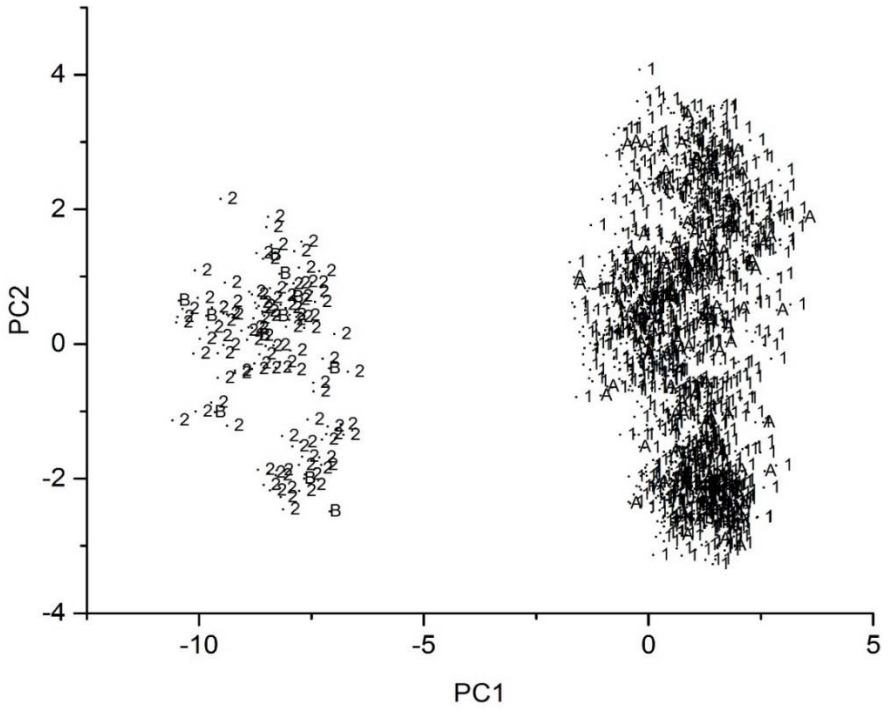

Figure 5.21. Projection of the 127 prediction set samples onto the PC plot of the 1135 training set samples and the 45 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors, Chrysler, Honda, Nissan, Toyota and Ford; 2 = Chrysler (2 plants) and General Motors (4 plants). Prediction set: A = All manufactures (General Motors, Honda, Nissan, Toyota, Ford and Chrysler); B = Chrysler (2 plants) and General Motors (4 plants).

The sample cluster in Figure 5.22 and Figure 5.23 corresponding to the four General Motors and two Chrysler assembly plant samples was subsequently divided into two categories according to vehicle manufacturer. Figure 5.22 shows a plot of the 103 paint samples and 12 wavelet coefficients identified by the pattern recognition GA for this two-class problem. Figure 5.23 shows the prediction set samples projected onto the PC plot shown in Figure 5.22. All prediction set samples were correctly classified.

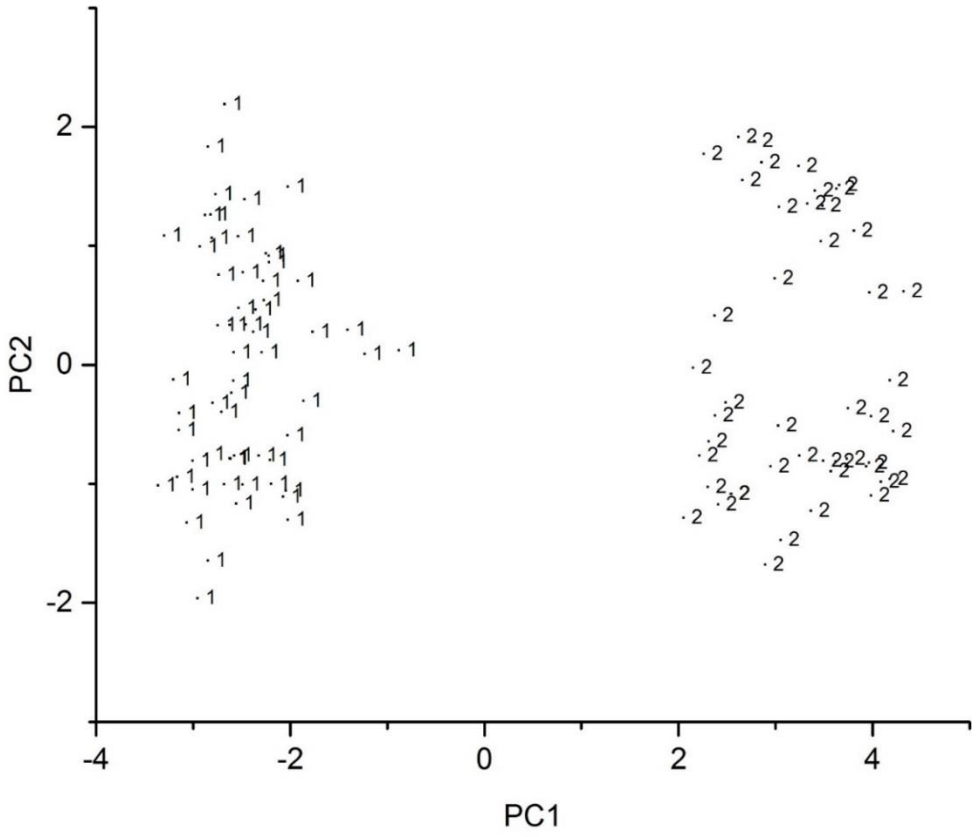

Figure 5.22. PC plot of the 103 training set samples and the 12 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors (4 plants); 2 = Chrysler (2 plants).

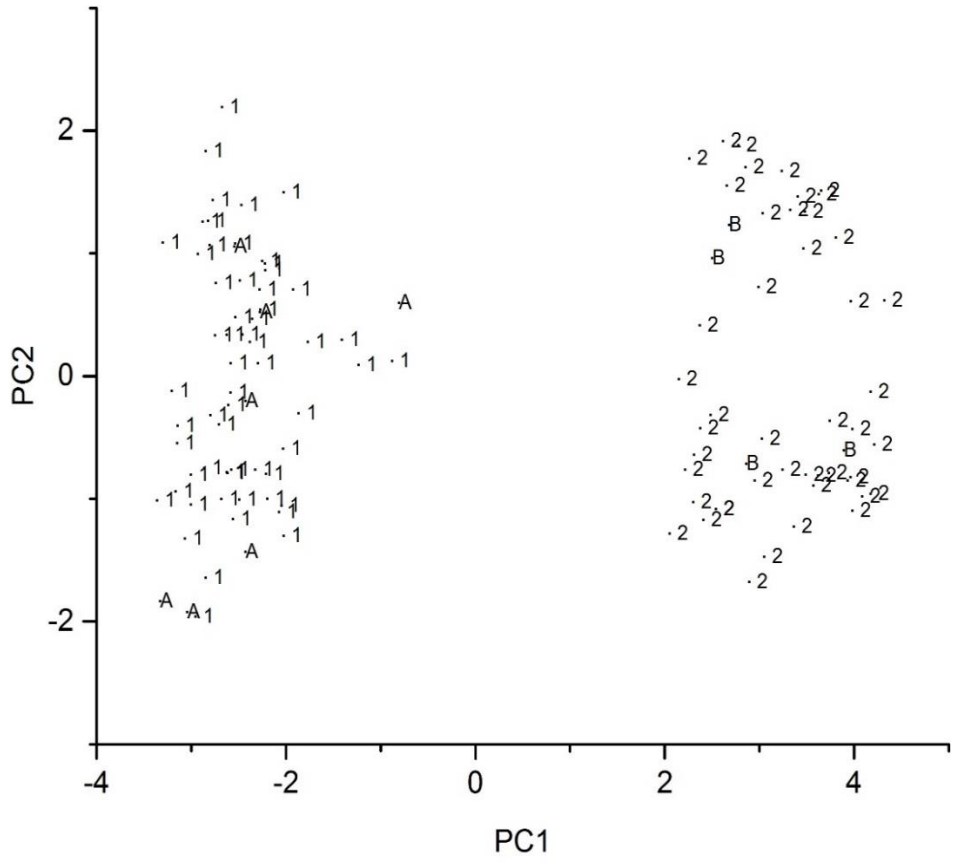

Figure 5.23. Projection of the prediction set samples onto the PC plot of the 103 training set samples and the 12 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors (4 plants); 2 = Chrysler (2 plants). Prediction set: A = General Motors (4 plants); B = Chrysler (2 plants).

Another classifier was developed using the pattern recognition GA to separate Chrysler (three assembly plants) from those of General Motors, Honda, Nissan, and Toyota as well as Chrysler (three other assembly plants). For this study, the 1050 training set samples (without the 2 Chrysler and the 4 General Motors assembly plants) were divided into two classes: three Chrysler assembly plants versus General Motors, Honda, Nissan, Ford, Toyota and the remaining three Chrysler assembly plants. The pattern recognition GA identified 44 wavelet coefficients that achieved separation for 3 Chrysler assembly plants from the General Motors, Honda, Nissan, Toyota and the remaining 3 Chrysler

assembly plants paint samples. Figure 5.24 shows a plot of the two largest PCs of the 1050 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA. The 3 Chrysler assembly plant paint samples are well separated from those of General Motors, Honda, Nissan, Toyota and the other 3 remaining Chrysler assembly plants. Figure 5.25 shows the projection of the 117 prediction set samples onto the plot of the two largest PCs of the 1050 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA. All prediction set samples were correctly assigned to their respective category.

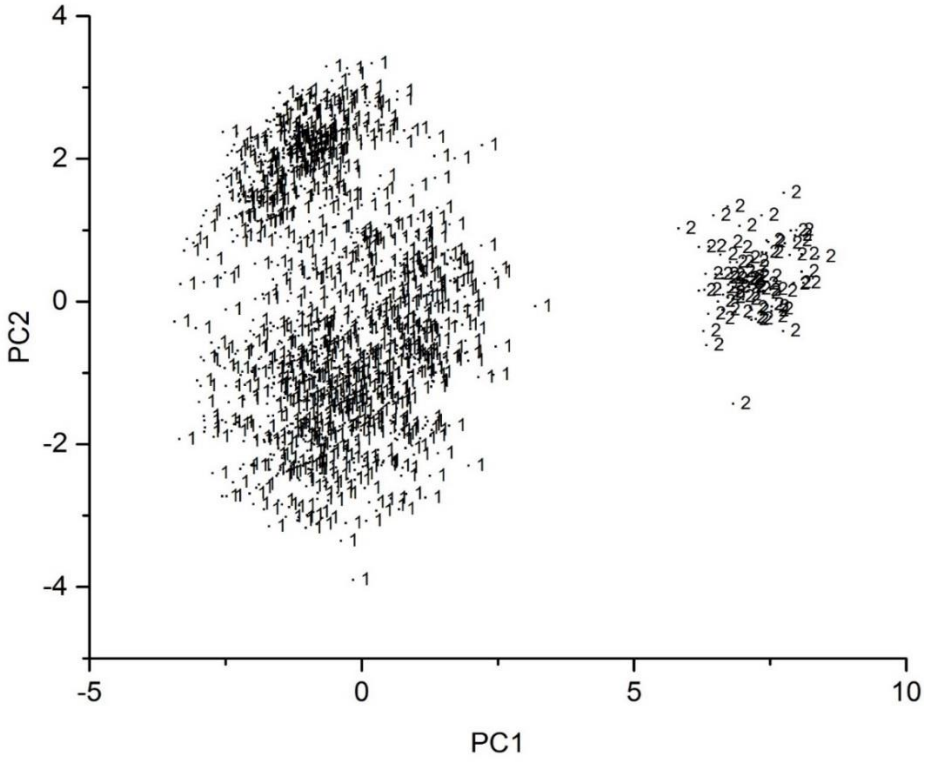

Figure 5.24. PC plot of the 1050 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors, Nissan, Toyota, Ford, Honda and Chrysler; 2 = Chrysler (3 plants).

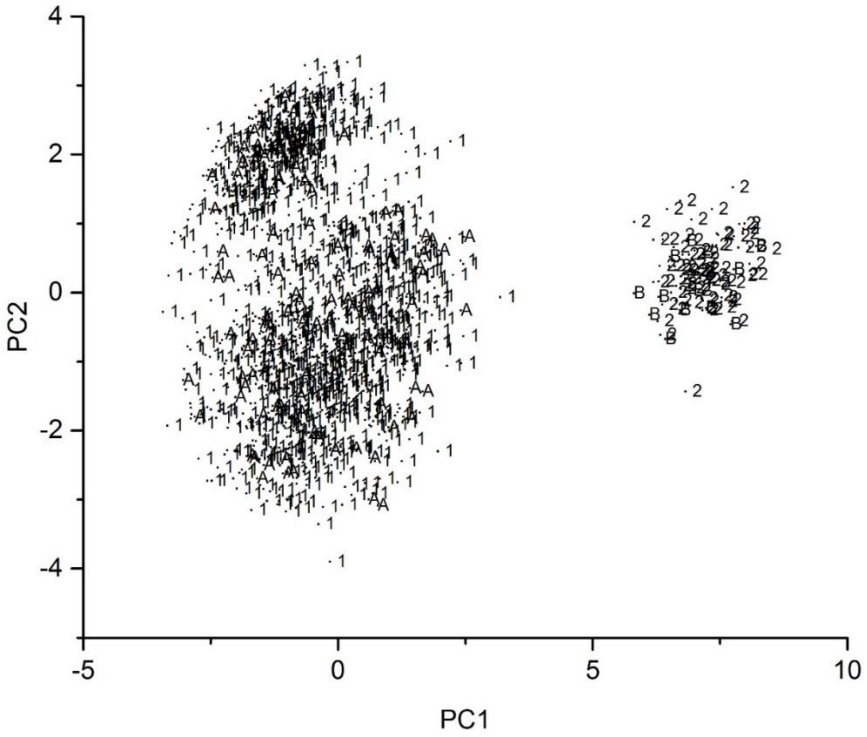

Figure 5.25. Projection of the 117 prediction set samples onto the PC plot of the 1050 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = General Motors, Nissan, Toyota, Ford, Honda and Chrysler; 2 = Chrysler (3 plants). Prediction set: A = General Motors, Nissan, Toyota, Ford, Honda and Chrysler; B = Chrysler (3 plants).

The remaining 1070 paint samples from the cluster representing General Motors, the remaining 3 Chrysler assembly plants as well as Honda, Nissan, Toyota and Ford were analyzed in a two-way classification study. Figure 5.26 shows a plot of the 966 training set samples comprising General Motors versus Honda, Toyota, Ford, Nissan and the remaining 3 assembly plants for Chrysler paint samples and the 22 wavelet coefficients identified by the pattern recognition GA for this two-way classification problem. Figure 5.27 shows a plot of the projection of the 104 prediction set samples onto the two largest PCs of the 966 training set samples and the 22 wavelet coefficients identified by the pattern recognition GA. General Motors paint samples are well separated from the paint samples from the other

manufacturers (Honda, Nissan, Ford, Toyota and the remaining 3 Chrysler assembly plants).

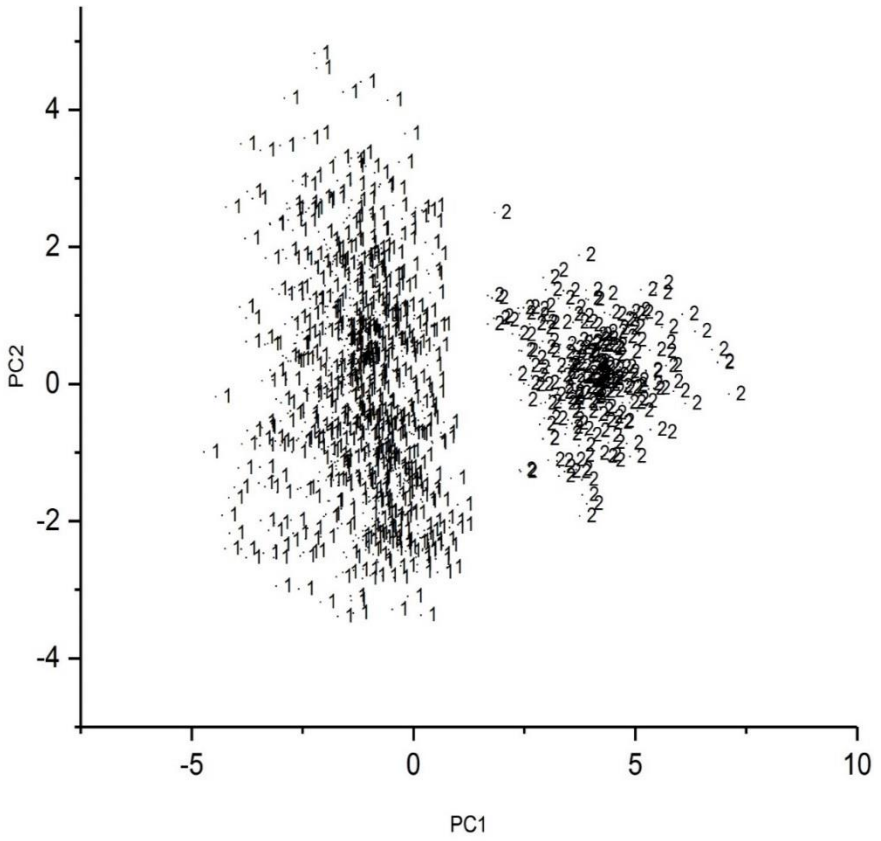

Figure 5.26. PC plot of the 966 training set samples and the 22 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Chrysler, Nissan, Toyota, Ford, and Honda; 2 = General Motors (all plants).

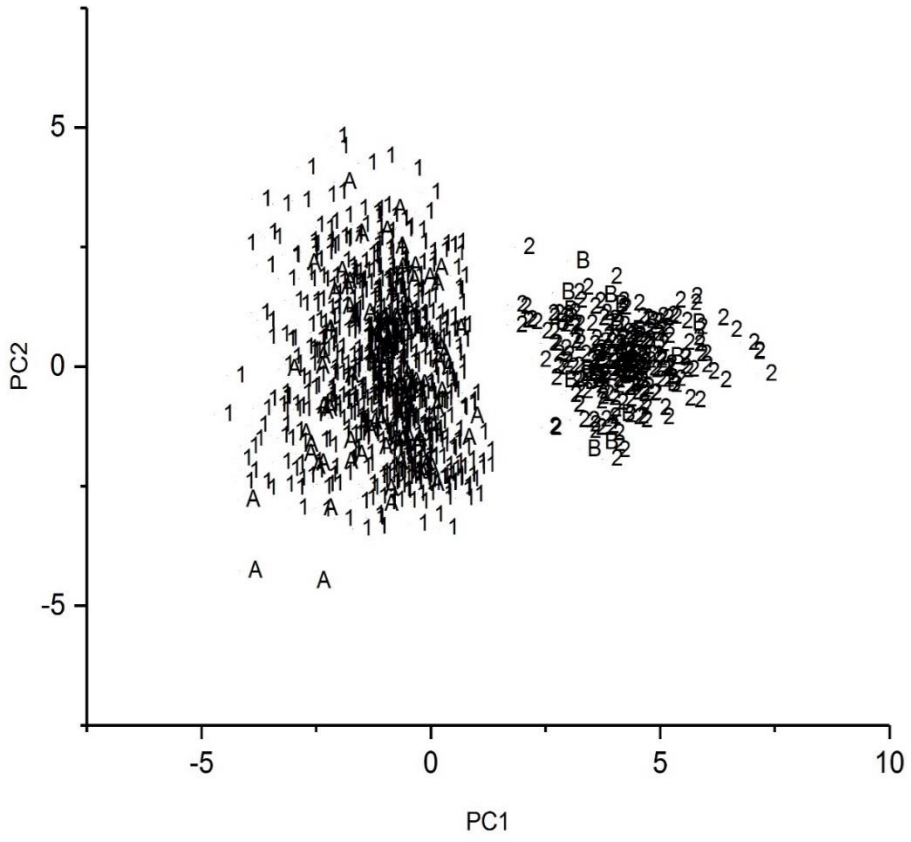

Figure 5.27. Projection of the 104 prediction set samples onto the PC plot of the 966 training set samples and the 22 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Chrysler, Nissan, Toyota, Ford, and Honda; 2 = General Motors (all plants). Prediction set: A = Chrysler, Nissan, Toyota, Ford, and Honda; B = General Motors (all plants).

The 799 paint samples from the cluster representing Honda, Nissan, Toyota, Ford and Chrysler were also analyzed in a two-way classification study. Figure 5.28 shows a plot of the 717 training set samples comprising Toyota, Honda, Nissan, Ford and Chrysler paint samples and the 28 wavelet coefficients identified by the pattern recognition GA for this two-way classification problem. Figure 5.29 shows a plot of the projection of the 82 prediction set samples onto the PC plot defined by the 717 Toyota, Honda, Nissan, Ford and the 3 Chrysler assembly plants paint samples and the 28 wavelet coefficients identified by the pattern recognition GA. The Toyota paint samples are well separated from Honda,

Nissan, Ford and Chrysler and all 82 prediction set samples are correctly classified by the PC plot. Each cluster in Figure 5.28 and Figure 5.29 was analyzed the pattern recognition GA. The 181 paint samples from Toyota could be differentiated from Honda, Nissan, Ford and Chrysler paint samples using the pattern recognition GA configured in the asymmetric classification mode. For this two-way classification study, the value of K was set at 161 for the Toyota assembly plant which represent the total number of Toyota paint samples (the target class) and 5 for the remaining paint samples (Honda, Nissan, Ford and Chrysler).

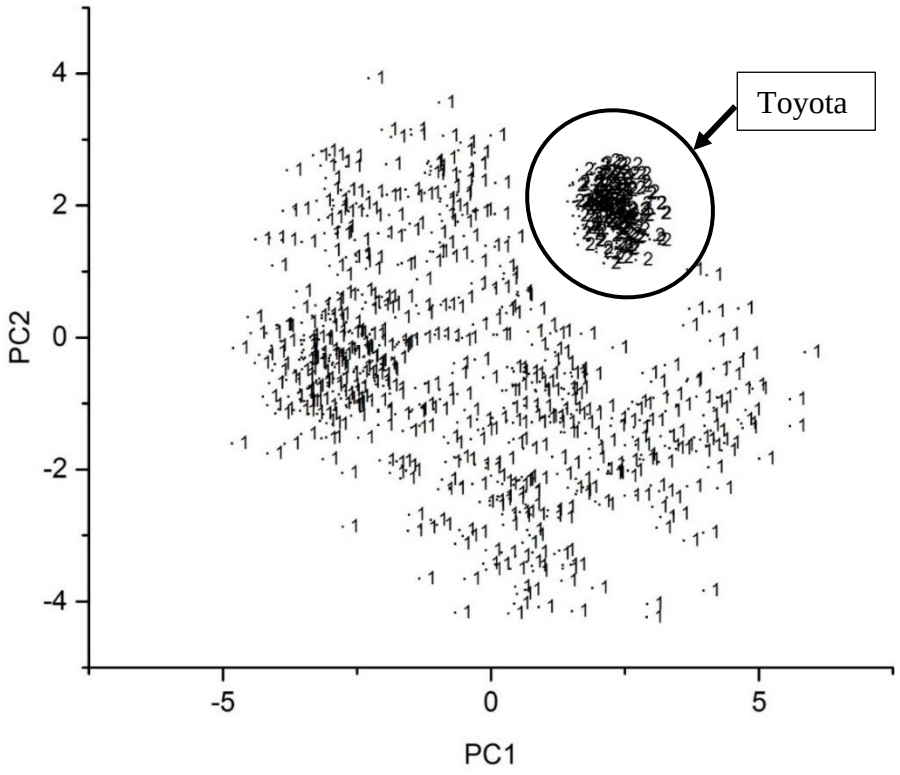

Figure 5.28. PC plot of the 717 training set samples and the 28 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Nissan, Ford, Honda and Chrysler; 2 = Toyota (all plants).

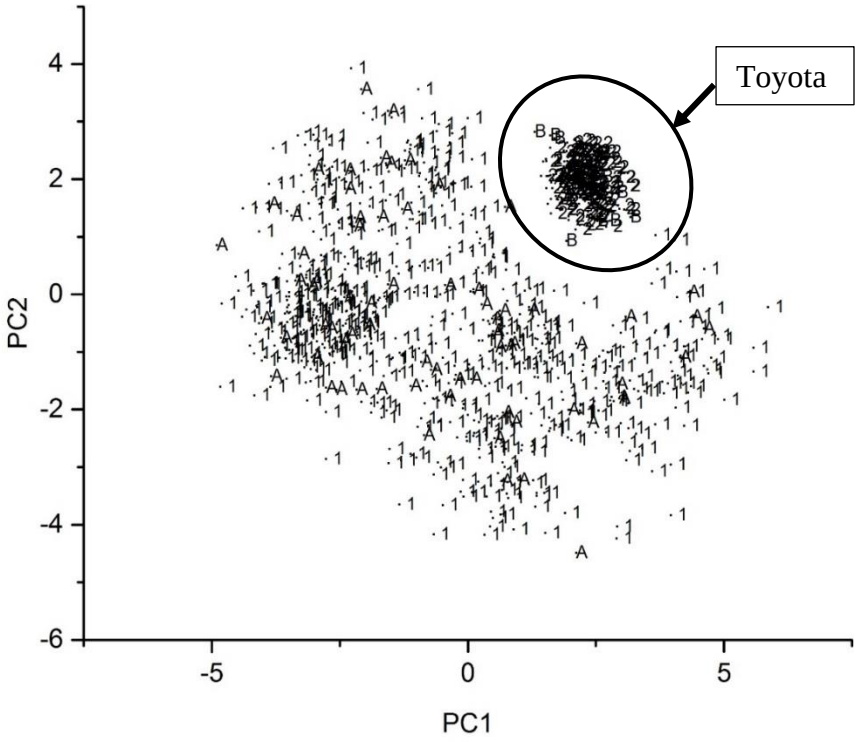

Figure 5.29. Projection of the 82 prediction set samples onto the PC plot of the 717 training set samples and the 28 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Nissan, Ford, Honda and Chrysler; 2 = Toyota (all plants). Prediction set: A = Nissan, Ford, Honda and Chrysler; B = Toyota (all plants).

Figure 5.30 shows a plot of the 618 comprising Chrysler (3 plants) and Ford assembly plants versus those of Nissan and Honda paint samples and the 36 wavelet coefficients identified by the pattern recognition GA. Figure 5.31 shows a plot of the projection of the 62 prediction set samples onto the PC plot defined by Chrysler (3 plants) and Ford assembly plants versus those of Nissan and Honda paint samples and the 36 wavelet coefficients identified by the pattern recognition GA. Both Honda and Nissan samples are well separated from those of Chrysler and Ford and all 62 prediction set samples were correctly assigned to their respective manufacturer.

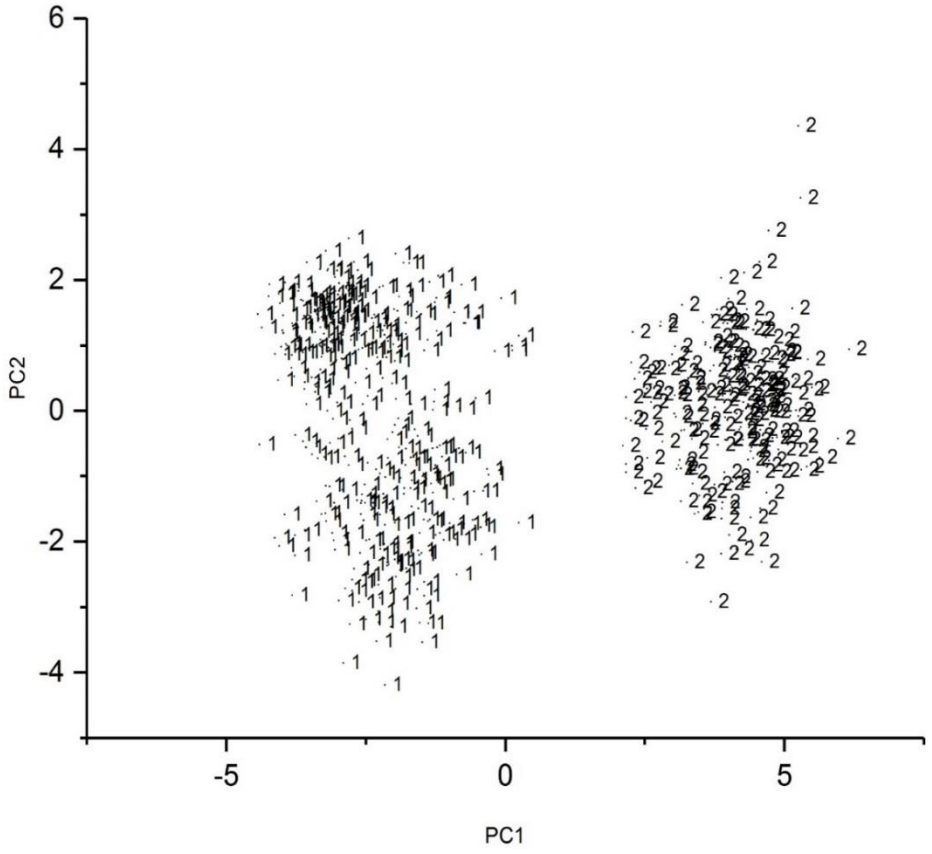

Figure 5.30. PC plot of the 618 training set samples and the 36 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Chrysler and Ford; 2 = Nissan and Honda

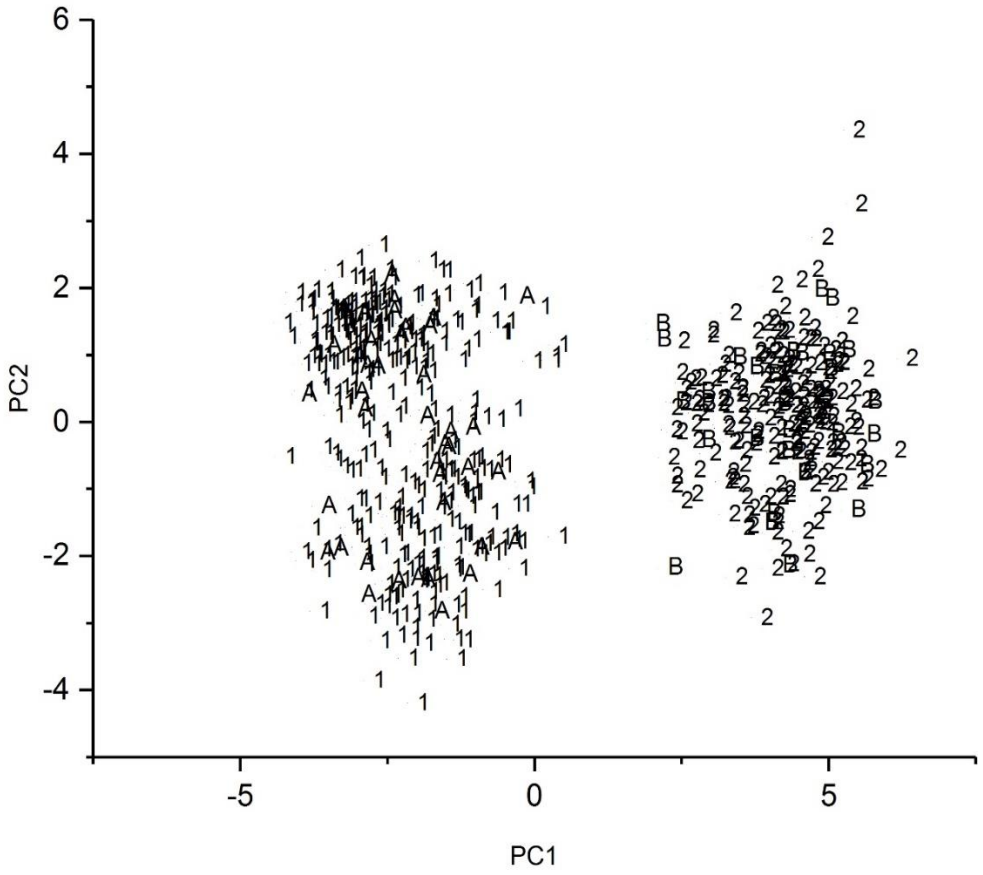

Figure 5.31. Projection of the 62 prediction set samples onto the PC plot of the 556 training set samples and the 36 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Chrysler and Ford; 2 = Nissan and Honda. Prediction set: A = Chrysler and Ford; B = Nissan and Honda.

The 123 Honda paint samples were differentiated from the 103 Nissan paint samples using the pattern recognition GA. Figure 5.32 shows a plot of the 203 Honda and Nissan training set samples and the 27 wavelet coefficients identified by the pattern recognition GA. Figure 5.33 shows the projection of the 23 prediction set samples associated with this training set onto the PC plot defined by the 226 Honda and Nissan paint samples and the 27 wavelet coefficients. The Honda paint samples form a compact cluster and all 23 prediction set samples were correctly classified by the PC plot.

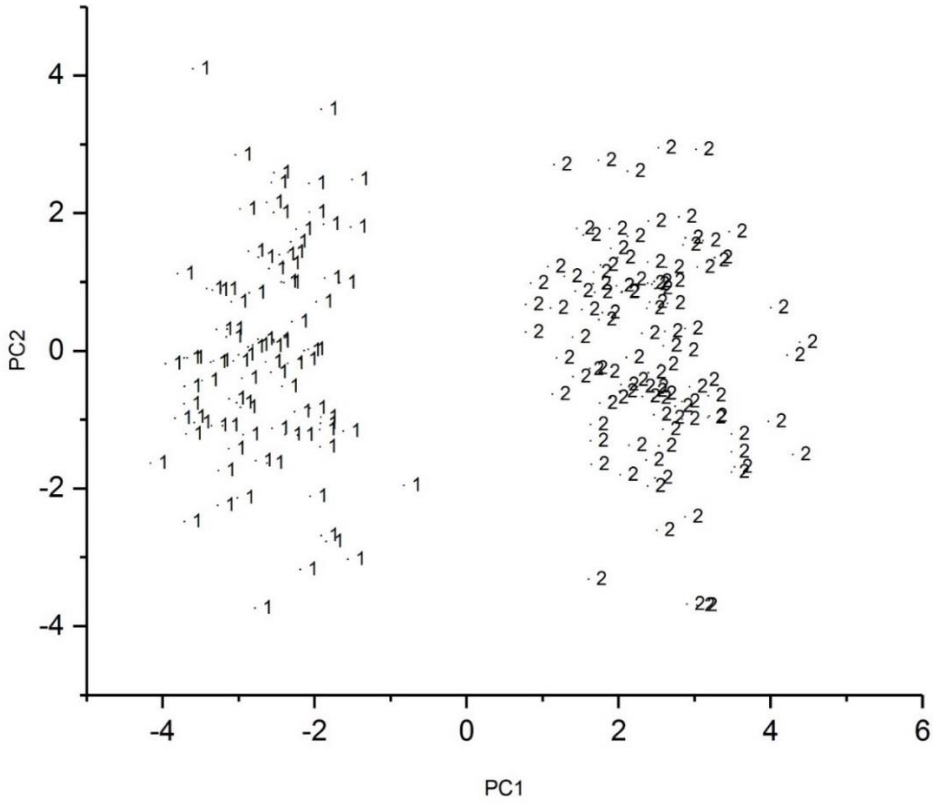

Figure 5.32. PC plot of the 203 training set samples and the 27 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Nissan; 2 = Honda.

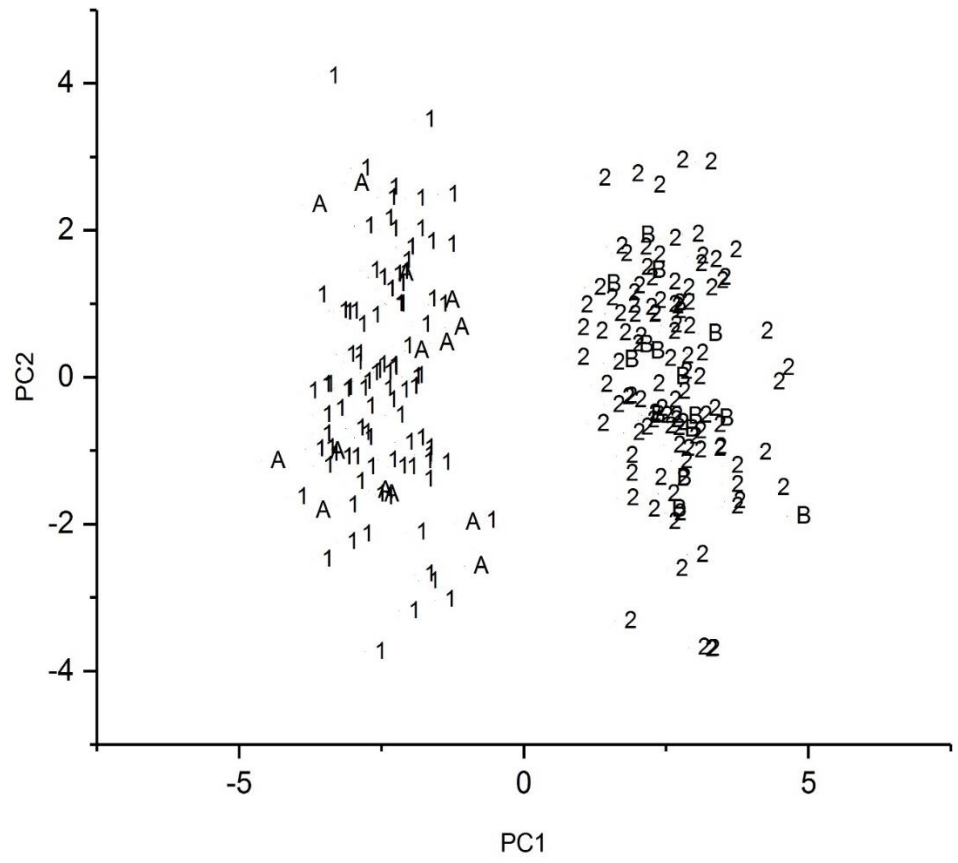

Figure 5.33. Projection of the 23 prediction set samples onto the PC plot of the 203 training set samples and the 27 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Nissan; 2 = Honda. Prediction set: A = Nissan; B = Honda.

Finally the 358 Ford paint samples were differentiated from the 34 Chrysler paint samples (3 assembly plants) using the pattern recognition GA. Figure 5.34 shows a plot of the 353 Ford and Chrysler training set samples and the 28 wavelet coefficients identified by the pattern recognition GA. Figure 5.35 shows the projection of the 39 prediction set samples associated with this training set onto the two largest PCs of the 353 Ford and Chrysler training set samples and the 28 wavelet coefficients. The Chrysler paint samples form a compact cluster and all 39 prediction set samples were correctly classified by the PC plot.

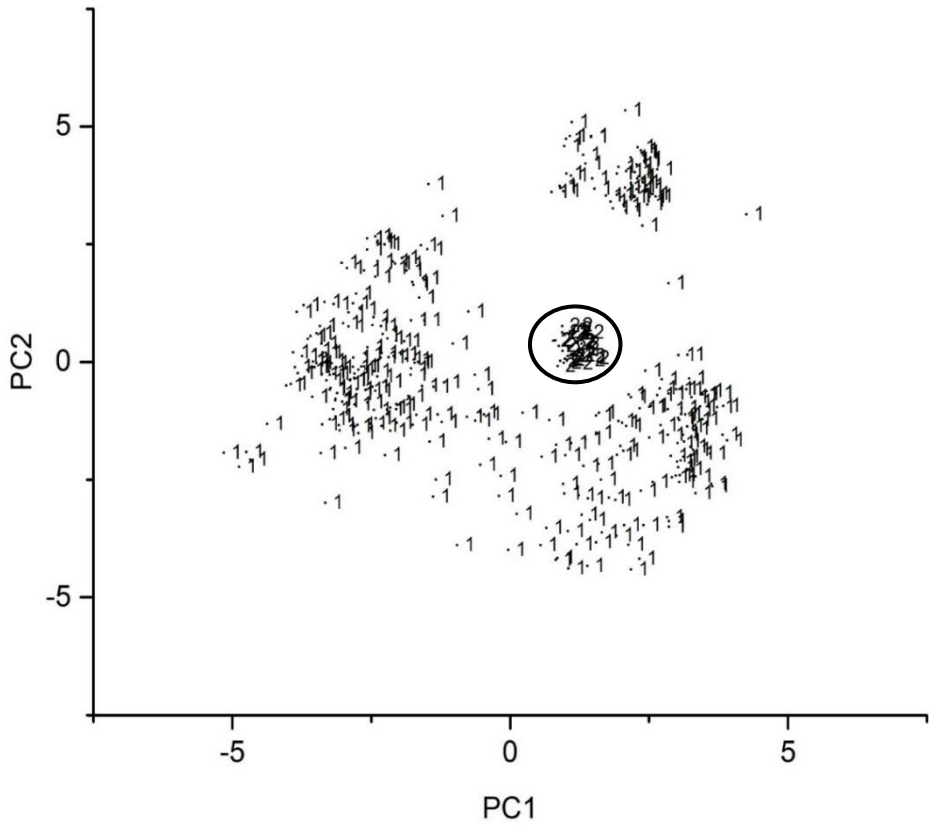

Figure 5.34. PC plot of the 353 training set samples and the 28 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Ford; 2 = Chrysler.

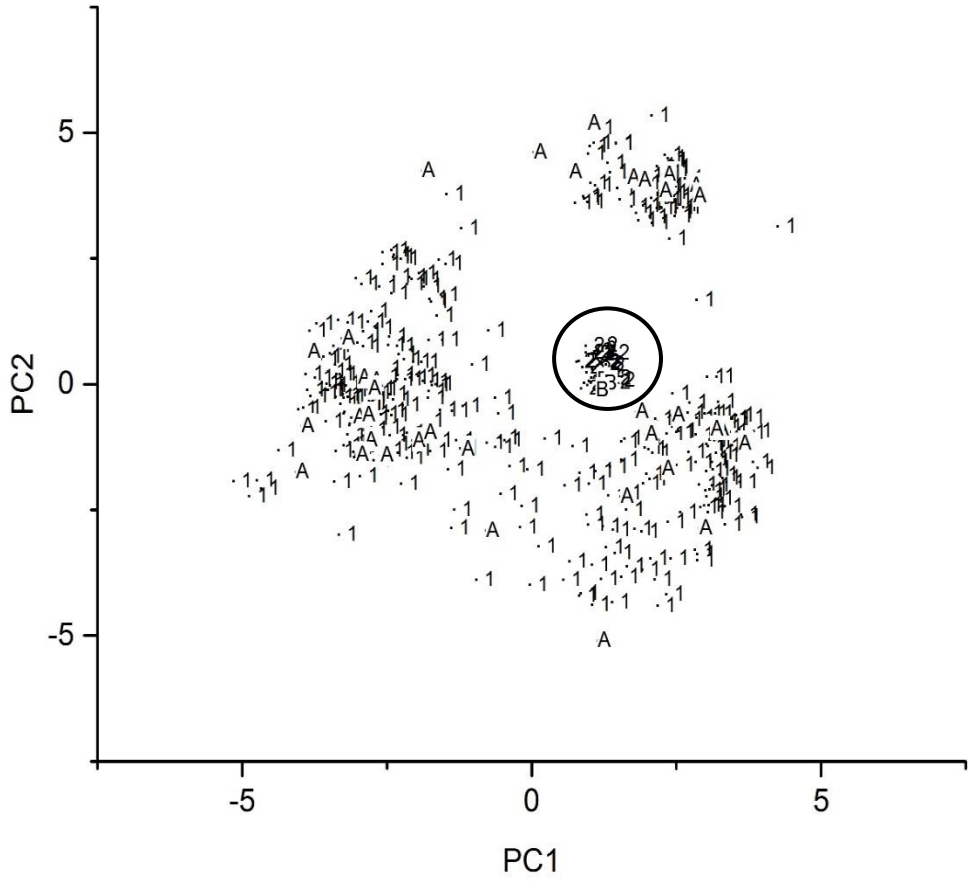

Figure 5.35. Projection of the 39 prediction set samples onto the PC plot of the 353 training set samples and the 28 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Ford; 2 = Chrysler. Prediction set: A = Ford; B = Chrysler.

Figure 5.36 provides an overview of the manufacturer search prefilter system developed from discriminants for paint samples whose clear coat layer is acrylic melamine styrene. A nine-tiered hierarchical classification scheme was developed by exploiting the linear separability of the sample classes comprising the training set. Classifier 1 separated the Chrysler (6 plants) from the other automotive manufacturers including some Chryslers. Classifier 2 separated Chrysler (2 plants) and General Motors (4 plants) from some General Motors, Chryslers, Honda, Nissan, Ford and Toyota. Classifier 3 separated the 2 Chrysler assembly plants from the 4 General Motors assembly plants. Classifier 4 was developed to

separate 3 Chrysler assembly plants from the remaining General Motors, Ford, Nissan, Toyota, Honda and the remaining 3 assembly plants for Chrysler. All the remaining paint samples for General Motors were separated from the remaining manufacturers using classifier 5. Classifier 6 effectively separated Toyota (all plants) from Honda, Nissan, Ford and the Chrysler (3 plants). Classifier 7 separated Chrysler and Ford from Honda and Nissan. Subsequently, Classifiers 8 and 9 were developed to effectively separate Honda from Nissan and Ford from the 3 remaining assembly plants for Chrysler respectively.

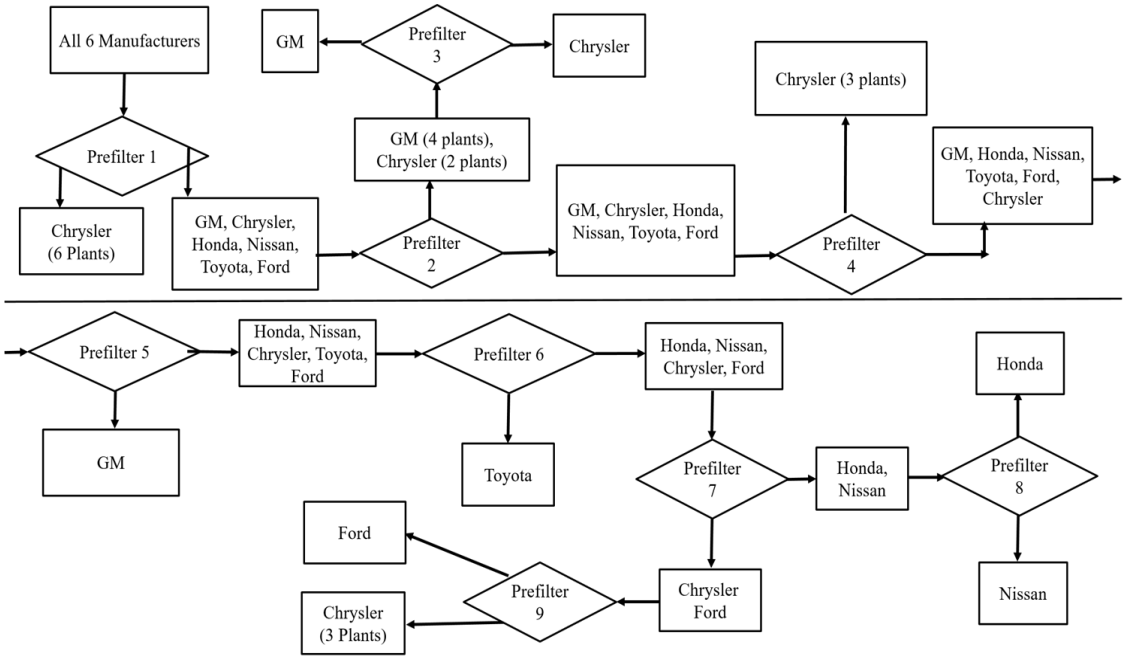

Figure 5.36. An overview of the manufacturer search prefilter system for paint samples whose clear coat layer is acrylic melamine styrene.

To demonstrate the operation of the manufacturer search prefilter system to identify the "make" of a vehicle from an unknown automotive paint sample whose clear coat layer is acrylic melamine styrene, infrared spectra of the clear coat, surfacer primer, and e-coat layers of a Chrysler (UAZP00421 – Jeep) were passed through the manufacturer search prefilter. Figure 5.37 depicts the steps implemented by the search prefilter system to identify the "make." In the first step (Prefilter 1), UAZP00421 was assigned to a collection of paint samples that spanned all six manufacturers (see Figures 5.18 and 5.38). In the second step (Prefilter 2), UAZP00421 was assigned to a smaller collection of paint samples that again spanned all six manufacturers (see Figures 5.20 and 5.38). In the third and final step (Prefilter 4), UAZP00421 was assigned to a collection of samples representing three Chrysler assembly plants (see Figure 5.24 and 5.40). Thus, the "make" of the vehicle is Chrysler.

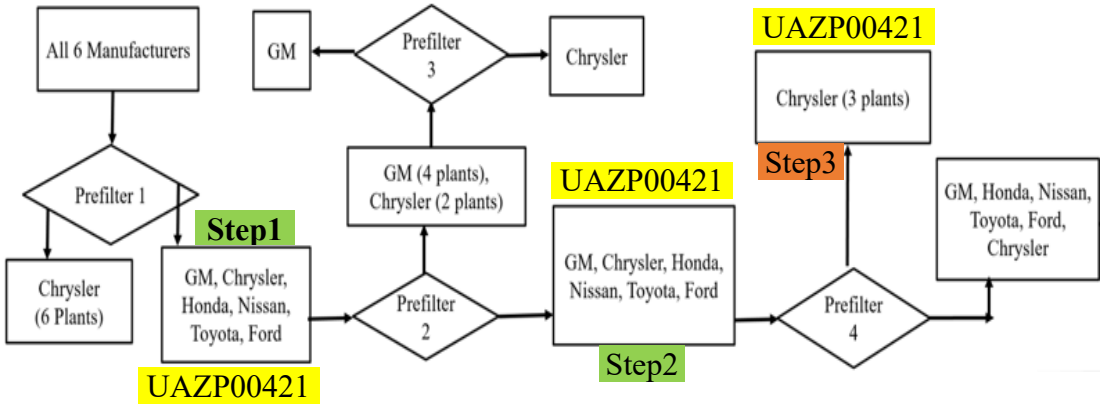

Figure 5.37. Flowchart explaining how the make for UAZP00421 was determined using the manufacturer search prefilter.

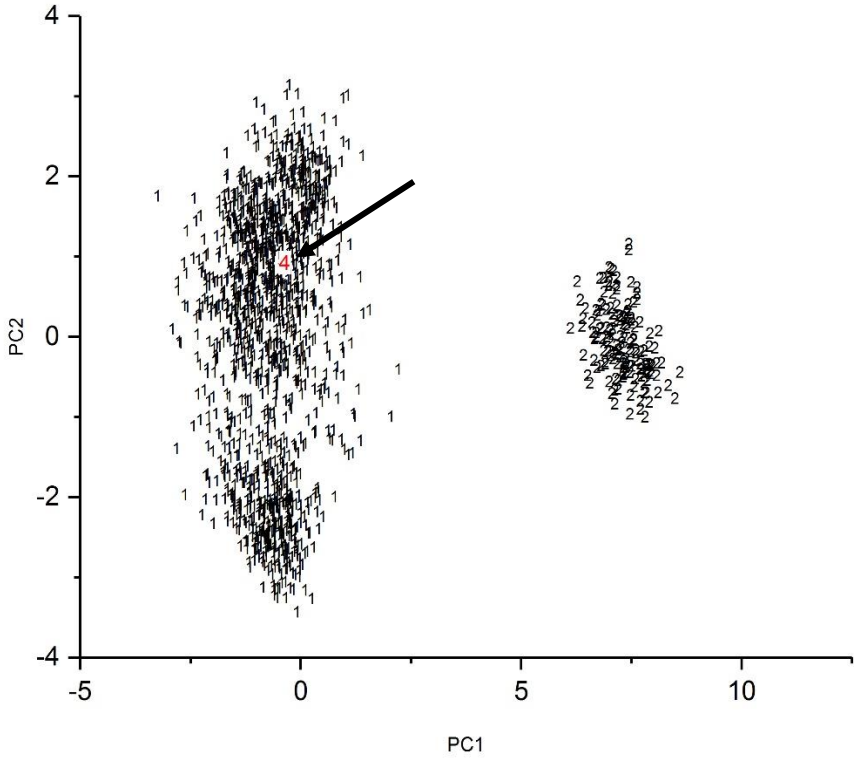

Figure 5.38. Assignment of UAZP00421 by Prefilter 1. 4 = UAZP00421.

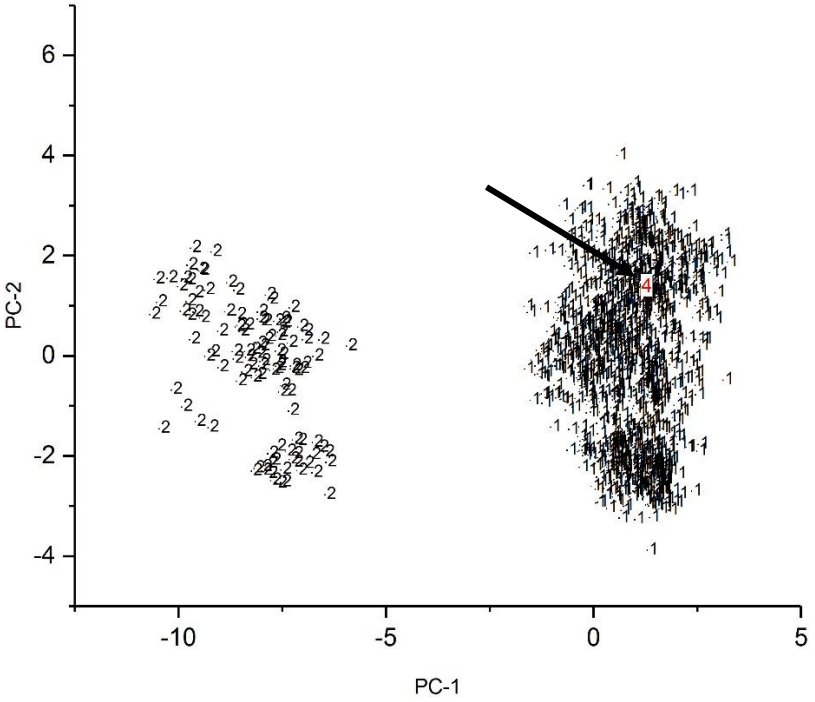

Figure 5.39. Assignment of UAZP00421 by Prefilter 2. 4 = UAZP00421.

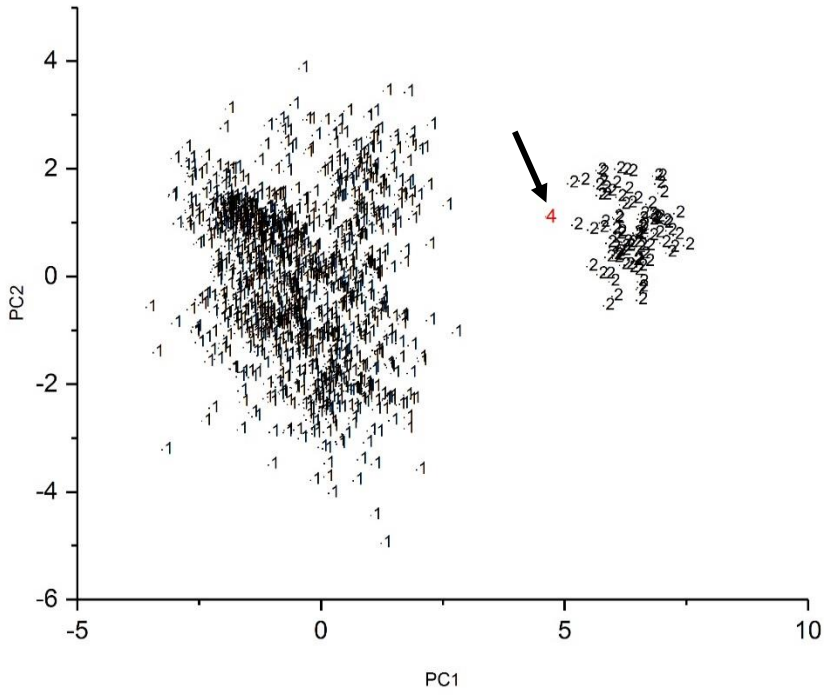

Figure 5.40. Assignment of UAZP00421 by Prefilter 4. 4 = UAZP00421.

To demonstrate the operation of the manufacturer search prefilter system to identify the "make" of a vehicle from an unknown automotive paint sample whose clear coat layer is acrylic melamine styrene polyurethane, a data vector consisting of the wavelet transformed infrared spectra of the clear coat, surfacer primer, and e-coat layers of a General Motors (UAZP00565 – Buick Lucerne) was projected onto a PC plot defined by the two largest principal components of the 48 wavelet coefficients identified by the pattern recognition GA and the 209 concatenated IR spectra comprising the training set (see Figures 5.15 and 5.41). Because the sample is projected into a region of the PC plot containing General Motors (GM) samples, the "make" of the vehicle is GM.

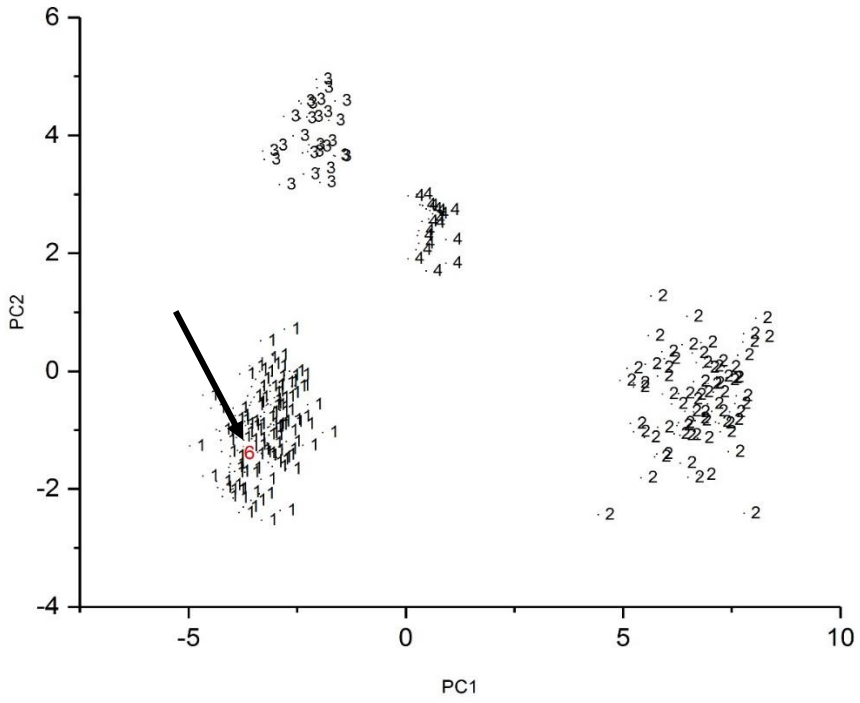

Figure 5.41. Projection of UAZP00565 onto the PC plot of the 33 wavelet coefficients identified by the pattern recognition GA and the 209 concatenated IR spectra comprising the training set for the manufacturer search prefilter developed for acrylic melamine styrene polyurethane. Training set: 1 = GM, 2 = Chrysler, 3 = Ford, 4 = Honda. 6 = UAZP00565.

Tables 5.7 and 5.8 summarize the results obtained from the manufacturer search prefilter system for identifying the "make" of the vehicle from the wavelet transformed spectra of the clear coat, surfacer-primer and e-coat layers of the automotive paint samples with and without epoxy. For the paint samples that were not cast in epoxy, the results are truly impressive. All 32 paint samples were correctly classified as to the "make" of the vehicle. For the paint samples cast in epoxy, the final results are not as impressive. Only twenty-two of twenty-seven paint samples were correctly classified as to the "make" of the vehicle. This can be attributed to the presence of epoxy in the clear coat and e-coat layers of these five misclassified paint samples.

Table 5.7 Unembedded Paint Samples

PDQ Number	Manufacturer	Search Prefilter Output
UAZP00412	Chrysler	Chrysler
UAZP00421	Chrysler	Chrysler
UAZP00451	Chrysler	Chrysler
UAZP00569	Chrysler	Chrysler
UAZP00600	Chrysler	Chrysler
UAZP00401	Chrysler	Chrysler
UAZP00342	Ford	Ford
UAZP00404	Ford	Ford
UAZP00467	Ford	Ford
UAZP00596	Ford	Ford
UAZP00477	Ford	Ford
UAZP00436	General Motors	General Motors
UAZP00271	General Motors	General Motors
UAZP00507	General Motors	General Motors
UAZP00331	General Motors	General Motors
UAZP00499	General Motors	General Motors
UAZP00565	General Motors	General Motors
UAZP00729	Honda	Honda
UAZP00277	Honda	Honda
CONT00726	Honda	Honda
CONT00736	Honda	Honda
UAZP00730	Honda	Honda
UAZP00440	Nissan	Nissan
UAZP00745	Nissan	Nissan
UAZP00731	Nissan	Nissan
UAZP00527	Nissan	Nissan
UAZP00537	Nissan	Nissan
UAZP00381	Toyota	Toyota
UAZP00313	Toyota	Toyota
UAZP00733	Toyota	Toyota
UAZP00561	Toyota	Toyota
UAZP00484	Toyota	Toyota

Table 5.8. Embedded Paint Samples

PDQ Number	Manufacturer	Search Prefilter Output
UAZP00412	Chrysler	Chrysler
UAZP00421	Chrysler	Chrysler
UAZP00451	Chrysler	Chrysler
UAZP00569	Chrysler	Chrysler
UAZP00600	Chrysler	Chrysler
UAZP00401	Chrysler	Chrysler
UAZP00342	Ford	Ford
UAZP00596	Ford	Ford
UAZP00436	General Motors	General Motors
UAZP00507	General Motors	General Motors
UAZP00331	General Motors	General Motors
UAZP00565	General Motors	General Motors
UAZP00277	Honda	Honda
CONT00736	Honda	Honda
UAZP00730	Honda	Honda
UAZP00440	Nissan	Nissan
UAZP00731	Nissan	Nissan
UAZP00527	Nissan	Nissan
UAZP00537	Nissan	Nissan
UAZP00381	Toyota	Toyota
UAZP00733	Toyota	Toyota
UAZP00484	Toyota	Toyota
UAZP00404	Ford	Not Correctly Classified
UAZP00477	Ford	Not Correctly Classified
UAZP00561	Toyota	Not Correctly Classified
UAZP00729	Honda	Not Correctly Classified
UAZP00745	Nissan	Not Correctly Classified

5.3.2.3. Assembly Plant Search Prefilters

After the "make" of the vehicle has been identified by the manufacturer search prefilter, the line and model of the vehicle was identified from IR spectra of OEM paint systems using a two-tiered process. First, the vehicle assembly plants of each manufacturer were divided into assembly plant groups by applying cluster analysis to the fingerprint region of the average IR spectrum of the clear coat layer which served as a prototypical data vector to represent the paint formulation used by each assembly plant. Second, each plant group was divided into its constituent assembly plants using the wavelet transformed IR spectra of the clear coat, surfacer-primer and e-coat layers to develop a discriminant. Further details on the formulation of these search prefilters to OEM paint systems can be found elsewhere [5-5 to 5-7].

To demonstrate the performance of the assembly plant search prefilters, a paint sample (UAZP00421 – Chrysler Jeep**)** correctly identified as Chrysler by the manufacturer search prefilter system for acrylic melamine styrene was assigned to one of the two Chrysler assembly plant groups (see Figure 5.42) and then one of the six assembly plants comprising the plant group (see Figure 5.43) by projecting the paint sample onto the PC plot defined by the two largest principal components of the 285 OEM paint systems and the 12 wavelet coefficients identified by the pattern recognition GA for plant group (see Figure 5.42) and the 155 OEM paint systems and 32 wavelet coefficients identified by the pattern recognition GA for assembly plant (see Figure 5.43). Using this hierarchical approach to classification, the assembly plant of the vehicle from which this paint sample (UAZP00421Chrysler Jeep) originated was correctly identified.

Tables 5.9 and 5.10 summarize the results obtained from the search prefilters for identifying the plant group and assembly plant of the vehicle from the wavelet transformed IR spectra of the clear coat, surfacer-primer and e-coat layers of the automotive paint samples with and without epoxy. If a paint sample is projected into a region of the PC plot with OEM paint systems from the same plant group or assembly plant as the sample, then the paint sample is correctly classified. Again, all 32 paint samples not cast in epoxy were correctly as classified and twenty-two of the twenty-seven paint samples cast in epoxy were correctly classified as to plant group and assembly plant.

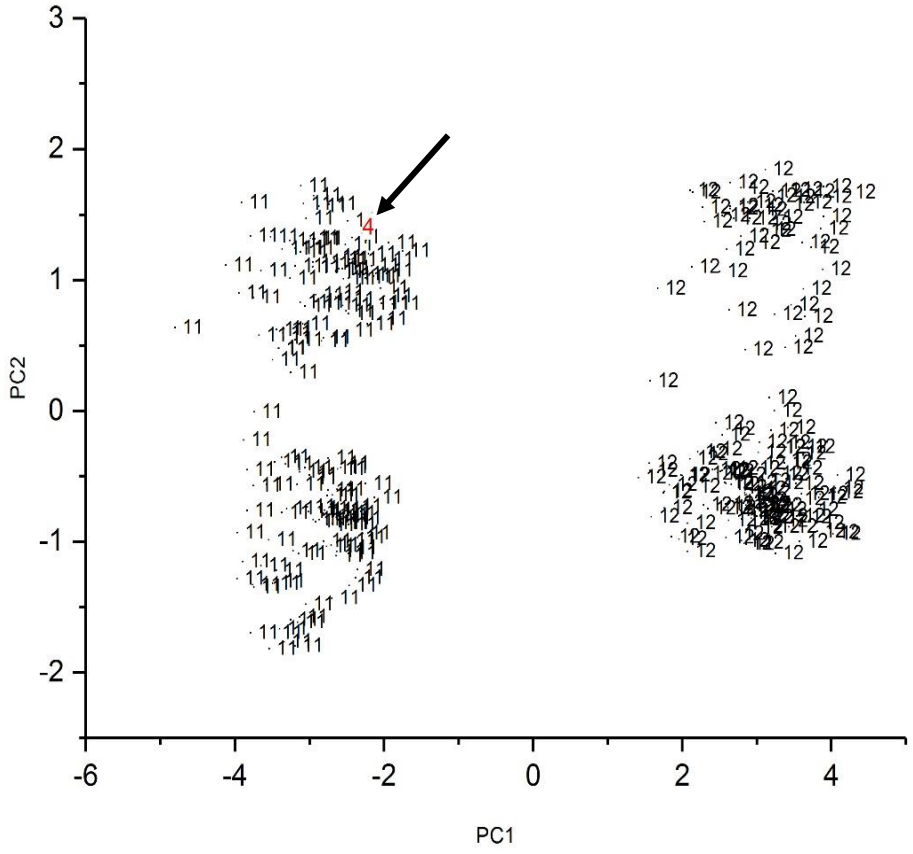

Figure 5.42. Assignment of UAZP00421 to Plant Group 11. 4 = UAZP00421.

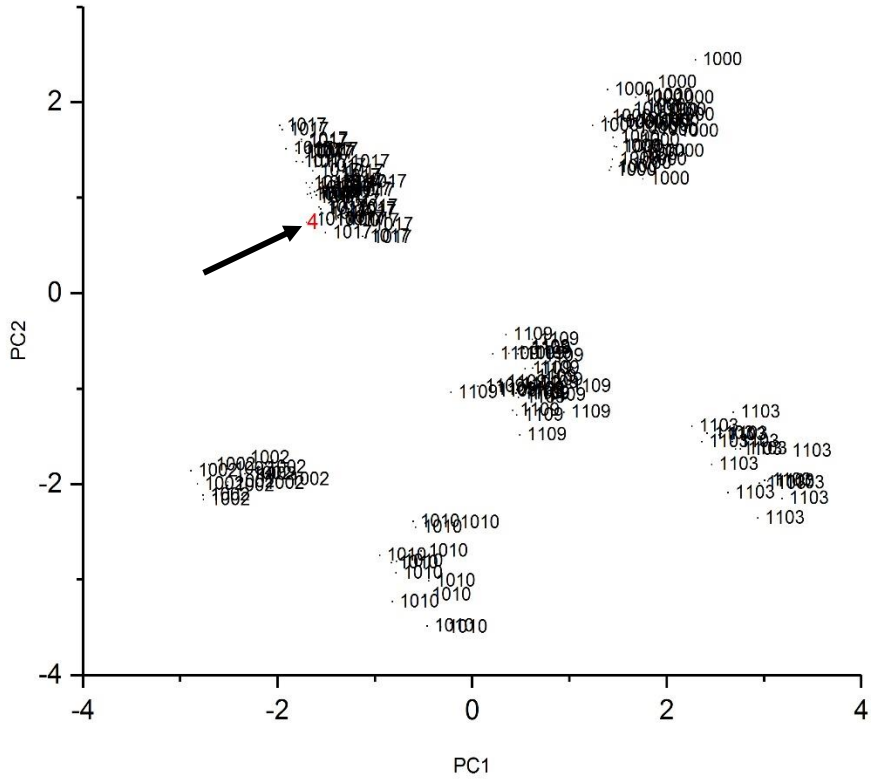

Figure 5.43. Assignment of UAZP00421 to Assembly Plant 1017 (Saltillo and Toluca). 4 = UAZP00421.

Tables 5.9. Unembedded Paint Samples

PDQ Number	Manufacturer	Plant Group	Assembly Plant	Search Prefilter Output
UAZP00412	Chrysler	11	1007	11/1007
UAZP00421	Chrysler	12	1009	12/1009
UAZP00451	Chrysler	12	1102	12/1102
UAZP00569	Chrysler	13	1109	13/1109
UAZP00600	Chrysler	11	1000	11/1000
UAZP00401	Chrysler	13	1006	13/1006
UAZP00342	Ford	22	2006	23/2006
UAZP00404	Ford	21	2010	21/2010
UAZP00467	Ford	21	2007	21/2007
UAZP00596	Ford	22	2005	23/2005
UAZP00477	Ford	22	2003	22/2003
UAZP00436	General Motors	1	1	1/1
UAZP00271	General Motors	4	12	4/12
UAZP00507	General Motors	1	20	1/20
UAZP00331	General Motors	5	26	5/26
UAZP00499	General Motors	2	21	2/21
UAZP00565	General Motors	2	10	2/10
UAZP00729	Honda	31	3007	31/3007
UAZP00277	Honda	31	3000	31/3000
CONT00726	Honda	31	3000	31/3000
CONT00736	Honda	31	3006	31/3006
UAZP00730	Honda	31	3002	31/3002
UAZP00440	Nissan	41	4006	41/4006
UAZP00745	Nissan	41	4001	41/4001
UAZP00731	Nissan	41	4017	41/2017
UAZP00527	Nissan	41	4017	41/4017
UAZP00537	Nissan	41	4001	41/4001
UAZP00381	Toyota	51	2347	51/2347
UAZP00313	Toyota	51	2347	51/2347
UAZP00733	Toyota	51	2347	51/2347
UAZP00561	Toyota	51	2347	51/2347
UAZP00484	Toyota	51	5005	51/5005

Tables 5.10. Embedded Paint Samples

PDQ Number	Manufacturer	Plant Group	Assembly Plant	Search Prefilter Output
UAZP00412	Chrysler	11	1007	11/1007
UAZP00421	Chrysler	12	1009	12/1009
UAZP00451	Chrysler	12	1102	12/1102
UAZP00569	Chrysler	13	1109	13/1109
UAZP00600	Chrysler	11	1000	11/1000
UAZP00401	Chrysler	13	1006	13/1006
UAZP00342	Ford	23	2006	23/206
UAZP00596	Ford	23	2005	23/2005
UAZP00436	General Motors	1	1	1/1
UAZP00507	General Motors	1	20	1/20
UAZP00331	General Motors	5	26	5/26
UAZP00565	General Motors	2	10	2/10
UAZP00277	Honda	31	3000	31/3000
CONT00736	Honda	31	3006	31/3006
UAZP00730	Honda	31	3002	31/3002
UAZP00440	Nissan	41	4006	41/4006
UAZP00731	Nissan	41	4017	41/4017
UAZP00527	Nissan	41	4017	41/4017
UAZP00537	Nissan	41	4001	41/4001
UAZP00381	Toyota	51	2347	51/2347
UAZP00733	Toyota	51	2347	51/2347
UAZP00484	Toyota	51	5005	51/5005

5.3.2.4. Forward and Reverse Library Searching

The prototype pattern recognition library search engine consists of search prefilters to reduce the size of the infrared spectral library to a specific assembly plant or plants and library search algorithms that utilize both forward and reverse searching to identify IR spectra most similar to the unknown in the truncated spectral library identified by the search prefilters. All library searches were restricted to the spectral region between 1641 cm^{-1} and 860 cm^{-1}. For the forward search, the spectra in the library that are most similar to the unknown are identified. The quality of each spectral match was evaluated using the OMNIC library search routines configured as correlation for the search type and Happ-

Genzel for apodization. The top five hits, i.e., the five library spectra with the largest hit quality index values are reported.

The reverse search was performed using a cross correlation library search algorithm [5-4 to 5-7]. The IR spectra were divided into three regions: 3675 to 2856 cm^{-1} (interval after absorption by the diamond cell), 1891 to 668 cm^{-1} (fingerprint interval and carbonyl band), and 1650 to 668 cm^{-1} (fingerprint interval). The overlap between the second and third spectral intervals enforces the relative scale of the peaks and captures the broader trends in the spectral data and effectively increases the importance of the fingerprint region in the spectral matching. It is better than using a disjoint set of intervals (e.g., 1650 cm^{-1} to 668 cm^{-1} with either single or double weighting, 1891 cm^{-1} to 1650 cm^{-1}, and 3675 cm^{-1} to 2856 cm^{-1}). Each region was normalized to unit length.

Within the intervals described, each comparison was made using a system of windows centered at the midpoint of the cross-correlated data interval. This midpoint corresponds to cross-correlation between the two signals with a zero time lag. From the midpoint, the windows expand in steps of 10 points or 100 points to include the entire cross-correlated spectrum. Because of the symmetry inherent to cross-correlation, the comparisons only need to be made from one side of the center burst. The Euclidian distance was used to evaluate the similarity index (see Equation 5.1) between the unknown and each library spectrum where s_{ij} is the similarity of the match, d_{ij} is the distance between the cross correlated (library versus sample) and autocorrelated (sample versus sample) spectrum and d_{max} is the largest distance in the set of cross correlated and autocorrelated spectra that were compared. The similarity metric in Equation 5.1 was used instead of the hit quality index, as it proved to be more informative.

$$s_{ij} = 1 - \frac{d_{ij}}{d_{\max}} \tag{5.1}$$

The backward search utilized autocorrelation and cross correlation to provide a probability index for the line and model of the unknown vehicle. For each window in each interval, the IR spectra in the library were ranked by their similarity index to the unknown, but only the label (i.e., the line and model of the automotive vehicle) of each of the top five hits in the window was preserved. After each window was processed, the number of hits for a specific line model and line was computed and divided by the number of comparisons that have been made by the algorithm. This generates a set of percentages that represent that likelihood of a particular line and/or model being a match for the unknown. Only those lines and models with a frequency of occurrence equal to or greater than 20% are included in the hit list.

While the forward search identified the library spectrum most similar to the unknown, the backward search provided insight into how well the library matched the unknown. For each unknown sample, the forward and backward searches were used in tandem to identify the corresponding vehicle information from the truncated PDQ library generated by the search prefilters. If there was agreement between the forward and reverse search results, the specific line and model of the vehicle common to both hit lists was always found to be the correct assignment. Samples assigned to the same line and model by both searches (forward and reverse) are well represented in the spectral library and also correlate well on an individual basis to a specific library sample. Further details about the cross correlation library search algorithm can be found elsewhere.

Tables 5.11 and 5.12 summarizes the results of the library searches for the unembedded and embedded paint samples for each layer using both forward and reverse searching. The forward search correctly matched all 32 unembedded paint samples. The correct line and model of the vehicle was present in the top five hits of the search for each paint sample. Although the reverse search did not perform as well as the forward search (18 of 32 were correctly identified as to line and model versus 32 out of 32), the reverse search has the advantage of providing insight into how well the truncated spectral library matched each unknown, rather than how well an individual sample matched spectra in the truncated spectral library. In all likelihood, the peak shifts encountered for some vibrational modes were not completely ameliorated using the ATR correction algorithm. As cross correlation is even more sensitive to changes in the band position than principal component analysis based methods, this discrepancy does not come as a surprise. For the embedded paint samples, the forward search correctly matched all of the samples successfully passed through the search prefilters. As for the reverse search, only 15 of the 22 paint samples were correctly identified as to line and model.

The prototype pattern recognition assisted infrared library search system applied to the reconstructed IR spectra of each paint layer categorized each unknown paint system by identifying successively smaller sets of spectra to which an unknown is assigned, thereby facilitating spectral library searching as the size of the library is culled to those spectra obtained from vehicles manufactured at the same assembly plant as that of the unknown. For the prototype pattern recognition library search system, the accuracy of the hit-list can be assessed as samples assigned to the same line and model by both the forward and reverse searches are always correctly matched. The infrared imaging experiment described in this

chapter when coupled to the prototype pattern recognition infrared library search system may be a potentially significant development that has the potential to enhance current approaches to forensic automotive paint examinations and aid in evidential significance assessment.

Tables 5.11. Unembedded Paint Samples

PDQ Number	Manufacturer	Forward Search Method			Reverse Search Method		
		OT2	OU1	OU2	OT2	OU1	OU2
UAZP00412	Chrysler						
UAZP00421	Chrysler						
UAZP00451	Chrysler						
UAZP00569	Chrysler						
UAZP00600	Chrysler						
UAZP00401	Chrysler						
UAZP00342	Ford						
UAZP00404	Ford						
UAZP00467	Ford						
UAZP00596	Ford						
UAZP00477	Ford						
UAZP00436	General Motors						
UAZP00271	General Motors						
UAZP00507	General Motors						
UAZP00331	General Motors						
UAZP00499	General Motors						
UAZP00565	General Motors						
UAZP00729	Honda						
UAZP00277	Honda						
CONT00726	Honda						
CONT00736	Honda						
UAZP00730	Honda						
UAZP00440	Nissan						
UAZP00745	Nissan						
UAZP00731	Nissan						
UAZP00527	Nissan						
UAZP00537	Nissan						
UAZP00381	Toyota						
UAZP00313	Toyota						
UAZP00733	Toyota						
UAZP00561	Toyota						
UAZP00484	Toyota						
Green = Correct, Red = Incorrect							

Tables 5.12. Embedded Paint Samples

PDQ Number	Manufacturer	Forward Search Method			Reverse Search Method		
		OT2	OU1	OU2	OT2	OU1	OU2
UAZP00412	Chrysler						
UAZP00421	Chrysler						
UAZP00451	Chrysler						
UAZP00569	Chrysler						
UAZP00600	Chrysler						
UAZP00401	Chrysler						
UAZP00342	Ford						
UAZP00596	Ford						
UAZP00436	General Motors						
UAZP00507	General Motors						
UAZP00331	General Motors						
UAZP00565	General Motors						
UAZP00277	Honda						
CONT00736	Honda						
UAZP00730	Honda						
UAZP00440	Nissan						
UAZP00731	Nissan						
UAZP00527	Nissan						
UAZP00537	Nissan						
UAZP00381	Toyota						
UAZP00733	Toyota						
UAZP00484	Toyota						
Green = Correct, Red = Incorrect							

References

1. E. D. Emmons, R.G.K., S. S. Duvvuri, J. S. Thompson, and A. M. Covington, *High Pressure Infrared Absorption Spectroscopy of Poly(Methyl Methacrylate).* J. Polymer Sci. B. Physics 2007. **45**: p. 358-367.

2. Barry K. Lavine, A.F., Nikhil Mirjankar, Koichi Nishikida, and Jay Campbell, *Simulation of Attenuated Total Reflection Infrared Absorbance Spectra – Applications to Forensic Analysis of Automotive Clearcoats.* Applied Spectroscopy, 2014. **68(5)**: p. 608-615.

3. Undugodage Don Nuwan Perera, K.N., and Barry K. Lavine, *Development of Infrared Library Search Prefilters for Automotive Clear Coats from Simulated ATR Spectra.* Applied Spectroscopy, 2018. **186**: p. 662-669.

4. Ayuba Fasasi, N.M., Razvan-Ionut Stoian, Collin White, Matthew Allen, Mark P. Sandercock and Barry K. Lavine, *Pattern Recognition Assisted Infrared Library Searching of Automotive Clear Coats.* Applied Spectroscopy, 2015. **69(1)** p. 84-94.

5. Barry K. Lavine, C.G.W., Matthew D. Allen, Ayuba Fasasi, and Andrew Weakley, *Evidential Significance of Automotive Paint Trace Evidence Using a Pattern Recognition Based Infrared Library Search Engine for the Paint Data Query Forensic Database.* Talanta, 2016. **159**: p. 317-329.

6. Barry K. Lavine, C.G.W., Matthew D. Allen, Ayuba Fasasi, and Andrew Weakley, *Forensic analysis of automotive paints using a pattern recognition assisted infrared library searching system: Ford (2000-2006).* Microchemical Journal, 2016. **129**: p. 173-183.

7. Barry K. Lavine, C.G.W., Matthew D. Allen, Andrew Weakley, *Pattern Recognition Assisted Infrared Library Searching of the Paint Data Query Database to Improve Investigative Lead Information from Automotive Paint Trace Evidence.* Applied Spectroscopy, 2017. **71(3)**: p. 480-495.

8. Walker, J.S., *A Primer on Wavelets and their Scientific Applications*. Vol. FL 33431. Boca Raton, : Chapman & Hall/CRC.

CHAPTER VI

ANALYSIS OF EDIBLE OILS USING RAMAN SPECTROSCOPY AND PATTERN RECOGNITION METHODS

6.1. Introduction

Edible oils are mixtures of triglycerides that differ in their relative composition of fatty acids (e.g. palmitic, oleic, steric and linoleic). Because of their high nutritional value, edible oils are an important component of the human diet [1]. They are a source of essential fatty acids and are a carrier of fat soluble vitamins [2, 3]. Edible oils are used in cooking and are also ingredients in many processed or precooked foods because of their sensory characteristics. Most edible oils sold commercially are derived from plants, e.g., olive, corn, canola, sunflower, peanut, and safflower oil, although some are derived from animals, e.g., tallow and lard.

Adulteration of more expensive edible oils, for example, extra virgin olive oil or sesame oil by substitution or by blending with less expensive oils (such as corn or canola oil) is a problem that is of concern to government and regulatory officials [4]. Analysis of edible oils for purposes of classification or authentication is usually carried out by gas chromatography/mass spectrometry (GC/MS) [5]. However, GC/MS analysis of edible oils can be laborious and time consuming. Approximately twenty years ago, analysis of edible

oils (e.g. olive, sesame, and canola) for authentication [6], adulteration [7], and classification [8] was demonstrated using Raman spectroscopy, which allowed spectra to be collected in a short time period without the need for sample preparation. More recently, multivariate classification and calibration methods have been applied to Raman and IR spectra of edible oils to improve classification success rates [9-15] and lower detection limits for adulterants [13, 16-18]. Generally, classification success rates of around 90% for edible oils as well as adulterant detection limits of around 10% have been reported in the literature. However, these studies, which used the 900 cm-1 to 1800 cm-1 region, were typically limited to 20 samples spanning five or six edible oils using PLS or linear discriminant analysis to perform a flat classification of the data. Furthermore, the edible oils investigated were represented by samples obtained from a single brand within a limited production year range. Many of these studies may have provided an overly optimistic estimate of the ability of Raman spectroscopy to classify edible oils by type or to detect low levels of adulterants present in these oils.

Raman spectra of edible oils usually contain weak narrow bands superimposed over a broad, high intensity fluorescence background (often referred to as baseline) that may distort the Raman bands of the components characteristic of the sample [19]. In these circumstances, baseline correction is necessary to ensure successful numerical processing of the Raman spectra by multivariate methods of data analysis such as principal component analysis (PCA) or partial least squares [20].

The application of Raman spectroscopy and pattern recognition methods to the problem of discriminating edible oils by type was investigated. In one study (designated as Data Set 1), 296 Raman spectra obtained from 53 samples spanning 15 varieties of edible

oils from different vendors (possibly the same company but a different batch and from a different manufacturing plant) were collected for 90 seconds at $2cm^{-1}$ resolution. The large number of classes (i.e., varieties of edible oils), samples, and spectra (i.e., replicates) were necessary to build better statistical distributions of expected in-class variance to determine classification performance when developing discriminants from training sets and to have sufficient number of spectra to construct independent training and validation sets. The Raman spectral data were then examined using the three major types of pattern recognition methodology: mapping and display, discriminant development and clustering. The 15 varieties of edible oils could be partitioned into five 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 Raman spectra within the same group more closely resembled each other and therefore would be more difficult to classify by type.

In another study (designated as Data Set 2), 215 Raman spectra of 15 edible oils/blends of edible oils were also collected at 2 cm^{-1} resolution. Using a genetic algorithm for pattern recognition, the discrimination of the edible oils by type was investigated. The 53 edible oil samples spanned multiple brands purchased over 3 years (representing different production years) for 9 of the 15 edible oils investigated. Supplier to supplier variation (and seasonal variation within a supplier) was a major source of variability within the Raman spectral data as it is not only greater than variability within a supplier but was comparable in magnitude to the variability associated with edible oil type. The novelty of these two studies arises from the incorporation of edible oils gathered systematically over three years, which introduces a heretofore unseen variability to the chemical compositions

of the edible oils that are being classified by type. This is the first time that many different edible oils and commercially available brands thereof have been classified simultaneously.

6.2. Experimental

Edible oils in the two studies discussed in this Chapter were purchased from supermarkets in the Newark, DE metropolitan area over 3 years, spanning 15 distinct varieties of edible oils (see Table 6.1 and Table 6.2). Each sample represented a different brand of edible oil or the same brand of oil but a different production year. In both studies, a sampling scheme was chosen to get as much variation as possible to simulate real world conditions, which was not the case in previously published work on this subject. We recognized this at the start of this investigation as this was our working hypothesis.

For the first study, an Ocean optics QE65000 Raman spectrograph (Dunedin, FL, USA) equipped with a Hamamatsu TE cooled CCD detector (Hamamatsu City, Shizuoka Pref., JP) and a Inphotonics (Norwood, MA, USA) fiber optic probe for sampling was used to collect 296 Raman spectra of the 15 edible oils/edible oil blends at 785nm. Each Raman spectrum (2000-50 cm^{-1}) was collected for 90 seconds integration at 2 cm^{-1} resolution and consisted of 1044 points. For the second study, 215 Raman spectra of 15 edible oils/blends of edible oils were collected at 785nm using a Kaiser Optical Systems Holospec *f*/1.8i spectrometer (Ann Arbor, MI) equipped with a liquid N_2 cooled detector (Princeton Instruments, Trenton, NJ) and a fiber optic probe (Kaiser Optical Systems, Ann Arbor, MI) for sampling. Each Raman spectrum (2000-50 cm^{-1}) was collected for 120 seconds through the side of 24mm glass vials.

Table 6.1. Edible Oil Data Set One

Oil Type	Oil ID	Number of Samples	Number of Spectra
Extra Virgin Olive Oil (EVOO)	1	12	63
Extra Light Olive Oil (ELOO)	2	4	31
Pure Olive Oil	3	6	75
Coconut Oil	4	1	4
Avocado Oil	5	1	14
Peanut Oil	6	3	10
Corn Oil	7	5	25
Grapeseed Oil	8	5	29
Safflower Oil	9	1	5
Hazelnut Oil	10	3	5
Flaxseed Oil	11	1	5
Canola Oil	13	7	11
Avocado/Flaxseed/Olive Oil	14	1	4
Sesame Oil	15	1	10
Vegetable Oil	17	2	5
Total	15 Oils	**53**	**296**

Table 6.2. Edible Oil Data Set Two

Oil Type	Oil ID	Number of Samples	Number of Spectra
Extra Virgin Olive Oil	1	12	44
Extra Light Olive Oil	2	4	13
Pure Olive Oil	3	6	24
Avocado	5	1	5
Peanut	6	3	13
Corn	7	5	24
Grapeseed	8	5	24
Safflower	9	1	3
Hazelnut	10	3	12
Canola	13	7	27
Sesame	15	1	4
Canol-Vegetable	16	1	8
Vegetable	17	2	5
Canola-Sun-Soybean	18	1	5
Sunflower	19	1	4
Total	15 Oils	53	215

6.3. Data Preprocessing and Pattern Recognition Analysis

For each Raman spectrum (see Figure 6.1a), the noise and background were stripped away from the signal by a three-step procedure performed using the PLS Toolbox 8.6 (Eigenvector Technology). First, a Whittaker filter [21] was applied to each spectrum. Next, the baseline corrected Raman spectrum was smoothed using a Savitzky-Golay linear filter with a 15 point window, and each spectrum was normalized to unit length. Figure 6.1 shows a representative Raman spectrum of corn oil that was baseline corrected, smoothed, and normalized to unit length. The bands at 2200cm^{-1} and 300cm^{-1} are an artifact of applying the Whittaker filter to the entire Raman spectrum (see Figure 6.1a). This artifact is due to the large changes in scattering intensity that occurs in these two regions. Furthermore, edible oils in the region 1000 cm^{-1} to 500 cm^{-1} do not have active Raman bands [22]. (If one looks at the signal to noise of all the Raman spectra for this region, it is very low.) Therefore, the baseline corrected, smoothed and normalized spectra (shown in Figure 6.1b) were truncated to 1772.6 cm^{-1} - 1127.6 cm^{-1} which corresponded to 361 points. Below this region, the spectra were too noisy, and above this region there did not appear to be any information present about edible oil type. Table 3.3 lists the wave-numbers of the five Raman active bands in this region characteristic of the major components (i.e., triglycerides) found in edible oils [22].

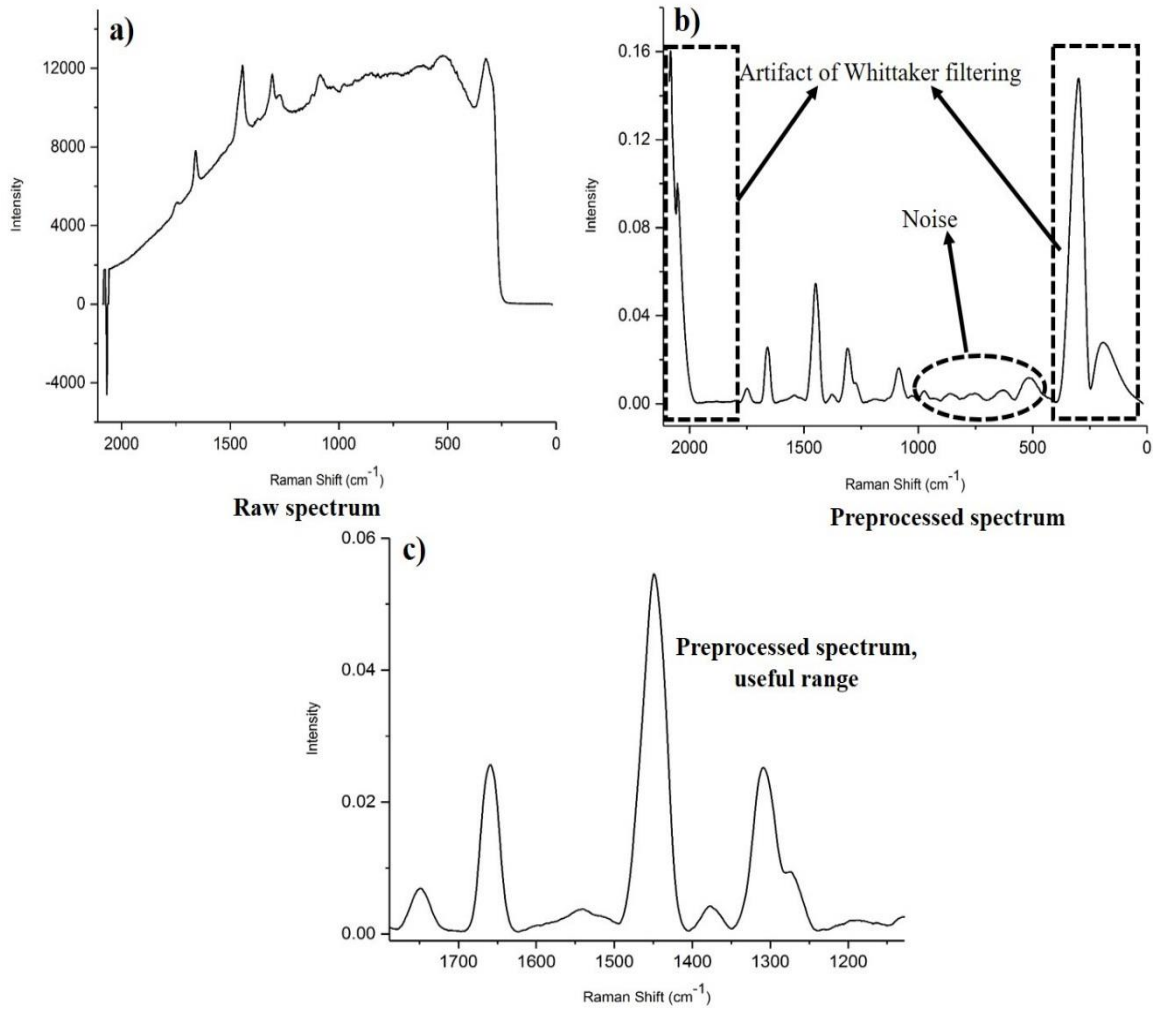

Figure 6.1. A representative Raman spectrum of corn oil: a) before baseline correction, smoothing, and normalization to unit length, b) after baseline correction, smoothing and normalization to unit length, and c) truncation of the uninformative regions to yield the spectral range (1772.6 cm^{-1} -1127.6 cm^{-1}) used for pattern recognition analysis.

Table 6.3. Raman Shift Assignments

Wavenumber	**Assignment**
1270 cm^{-1}	In plane =CH deformation in an unconjugated cis C=C
1305 cm^{-1}	In phase methylene twisting
1440 cm^{-1}	CH_2 scissoring deformation
1660 cm^{-1}	C=C stretching (cis)
1750 cm^{-1}	C=O stretching in an ester

6.4. Results and Discussion

6.4.1. Data Set 1

Outliers can obscure relationships present in data. For this reason, outlier analysis was performed on each class of edible oil in the training set using PCA prior to classification of the data. One corn oil spectrum (spectrum 296 in Figure 6.2) and one extra virgin olive oil spectrum (spectrum 149 in Figure 6.3) were judged to be outliers based on the PC plot of each edible oil and a comparison of the average Raman spectrum of corn oil and extra virgin olive oil to the Raman spectrum of the suspected outlier, see Figures 6.2 and 6.3. Therefore, these two Raman spectra were removed from the training set. In all likelihood, these two outliers were the result of small changes that occurred in the alignment of the fiber probe or the fiber optic connection to the monochromator of the Raman microscope. The spectra of the other replicates for each sample were very similar to the average Raman spectrum of corn or extra virgin olive oil.

Because of the large number of classes in the data set, a hierarchical classification scheme was employed to discriminate the Raman spectra of the 15 varieties of the edible oils comprising the training set. To implement this scheme, the average Raman spectrum of each variety of edible oil was computed. The 15 average Raman spectra were then analyzed using both PCA and cluster analysis [23]. For both PCA and hierarchical cluster analysis (Wards method), the spectral data were mean centered. Both the dendogram and the PC plot of the average Raman spectra (see Figure 6.4) were in good agreement, and each indicated that dividing the 15 classes of edible oils into five distinct groups is appropriate (see Table 6.4). A visual comparison of the average Raman spectra revealed that spectra in the same oil group were more similar to each other than spectra in different

oil groups. For this reason and because of the agreement in the results obtained from the PC score plot and the dendogram, we chose to partition the edible oils into five groups.

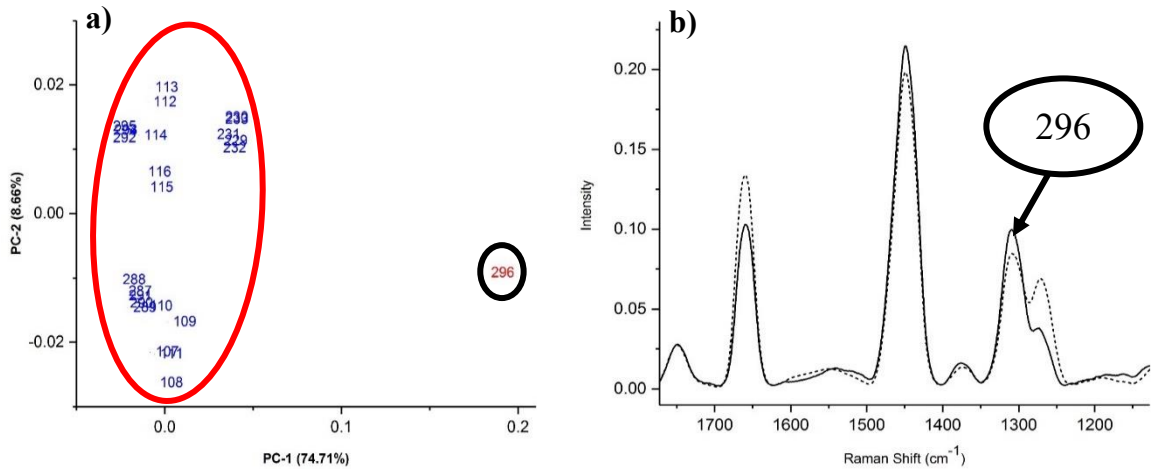

Figure 6.2. a) Plot of the two largest principal components of the Raman spectra of corn oil. One corn oil spectrum (spectrum id#296) appears as an outlier in the PC plot. b) Average Raman spectrum of corn oil (dashed line) and the Raman spectrum of the suspected outlier (solid line).

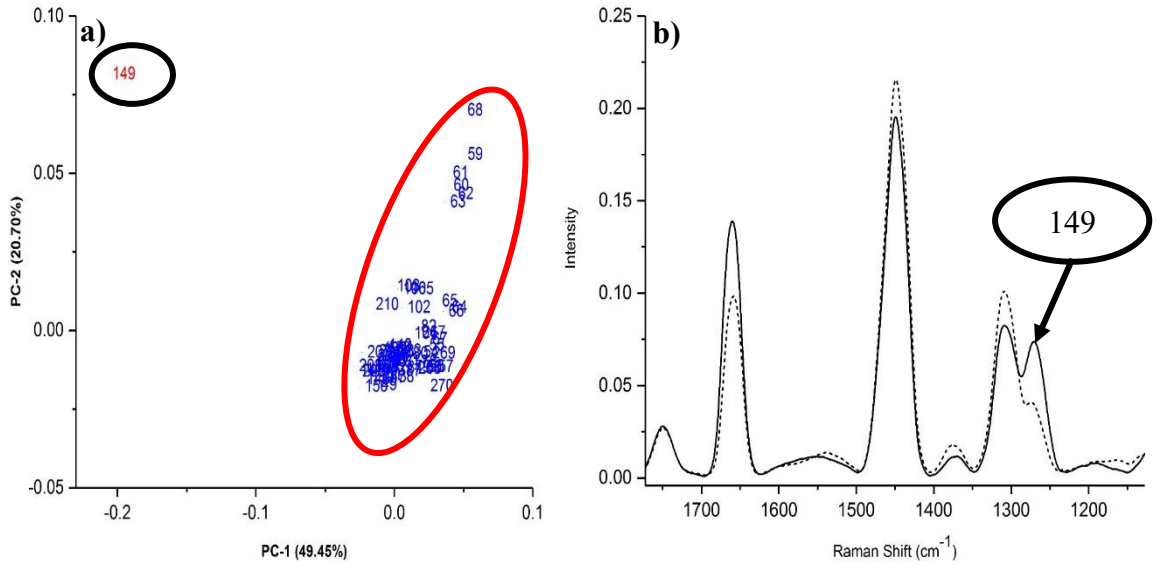

Figure 6.3. a) Plot of the two largest principal components of the Raman spectra of extra virgin olive oil. One extra virgin olive oil spectrum (spectrum id#149) appears as an outlier in the PC plot. b) Average Raman spectrum of extra virgin olive oil (dashed line) and the Raman spectrum of the suspected outlier (solid line).

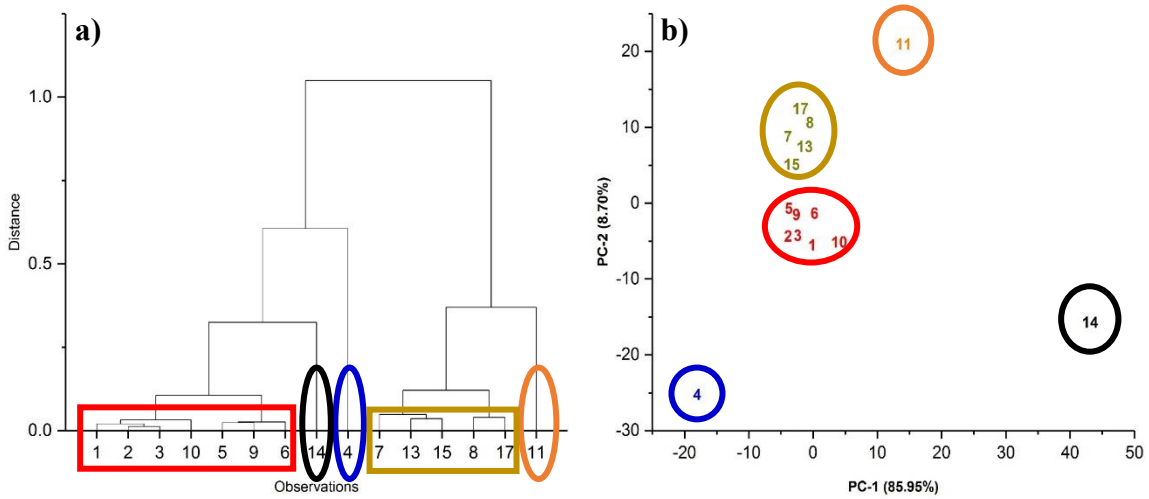

Figure 6.4. a) Dendogram (Wards method) and b) plot of the two largest principal components of the average Raman spectra of the 15 edible oils. The two plots are in agreement and each indicates that the 15 varieties of edible oils investigated in this study can be divided into five distinct groups. Group 1: 1 = extra virgin olive oil, 2 = extra light olive oil, 3 = pure olive oil, 6 = peanut oil, 9 = safflower oil, and 10 = hazelnut oil. Group 2: 7 = corn oil, 8 = grapeseed oil, 13 = canola oil, 15 = sesame oil, and 17 = vegetable oil. Group 3: 11 = flaxseed oil. Group 4: 4 = coconut oil. Group 5: avocado/flaxseed/olive oil.

Table 6.4. Edible Oil Group Assignments

Oil Type	Oil ID	Edible Oil Group
Extra Virgin Olive Oil	1	1
Extra Light Olive Oil	2	1
Pure Olive Oil	3	1
Coconut Oil	4	4
Avocado Oil	5	1
Peanut Oil	6	1
Corn Oil	7	2
Grapeseed Oil	8	2
Safflower Oil	9	1
Hazelnut Oil	10	1
Flaxseed Oil	11	3
Canola Oil	13	2
Avocado/Flaxseed/Olive Oil	14	5
Sesame Oil	15	2
Vegetable Oil	17	2

The set of data - 265 Raman spectra of 361 points each, which comprised the training set (see Table 6.5) – was subject to PCA. Figure 6.5 shows a plot of the two largest principal components of the 361 features obtained from the 265 Raman spectra comprising the mean-centered training set. Each spectrum is represented as a point in the PC plot. The five groups previously detected by cluster analysis are separated from each other in the PC plot. Since this projection is made without the use of information about the class assignment of each spectrum, the resulting separation is, therefore, a strong indication of real differences in the Raman spectral profile of these five edible oil groups.

A validation set of 29 Raman spectra (see Table 6.5) was employed to assess the predictive capability of the PC plot developed from the training set data. The 29 Raman spectra were directly mapped onto the PC plot defined by the 265 spectra and 361 spectral features comprising the training set. Figure 6.6 shows the validation set samples projected onto the PC plot of the training set data. Each projected validation set spectrum lies in a region of the plot with Raman spectra from the same edible oil group. This suggests that our approach taken for discriminating these 15 edible oils by first dividing them into five groups is supported by the data.

Table 6.5. Training and Validation Set for Edible Oil Group

Oil Type	Group ID	Oil ID	Training Set	Validation Set
Extra Virgin Olive Oil	1	1	54	8
Extra Light Olive Oil	1	2	29	2
Pure Olive Oil	1	3	68	7
Coconut Oil	4	4	4	0
Avocado Oil	1	5	12	2
Peanut Oil	1	6	10	0
Corn Oil	2	7	21	3
Grapeseed Oil	2	8	26	3
Safflower Oil	1	9	4	1
Hazelnut Oil	1	10	5	0
Flaxseed Oil	3	11	5	0
Canola Oil	2	13	11	0
Avocado/Flaxseed/Olive Oil	5	14	3	1
Sesame Oil	2	15	9	1
Vegetable Oil	2	17	4	1
Total	--------	--------	**265 spectra**	**29 spectra**

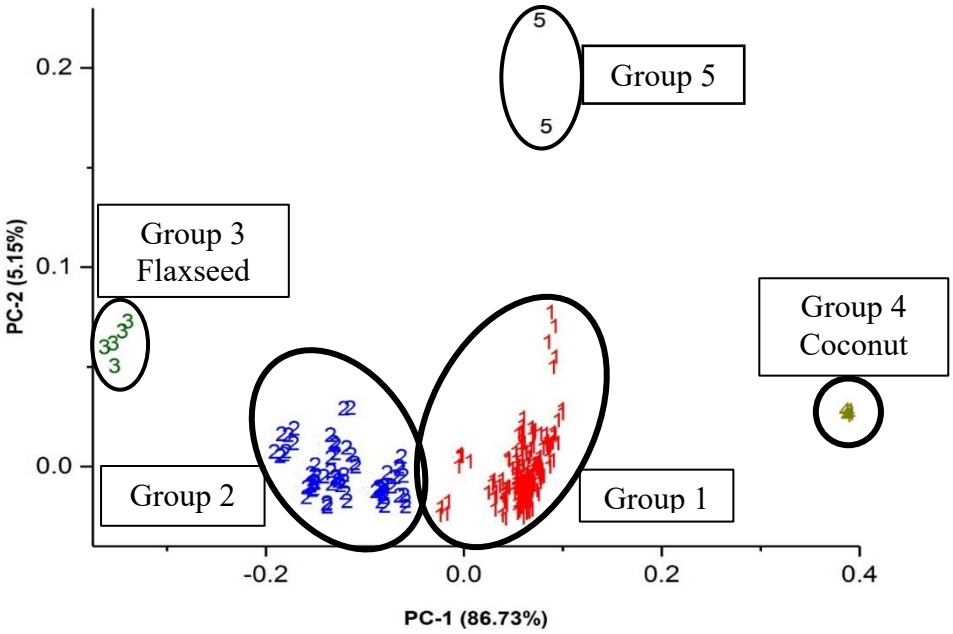

Figure 6.5. Plot of the two largest principal components of the 361 features obtained from the 265 Raman spectra comprising the mean-centered training set. 1 = Group 1, 2 = Group 2, 3 = Group 3, 4 = Group 4, and 5 = Group 5.

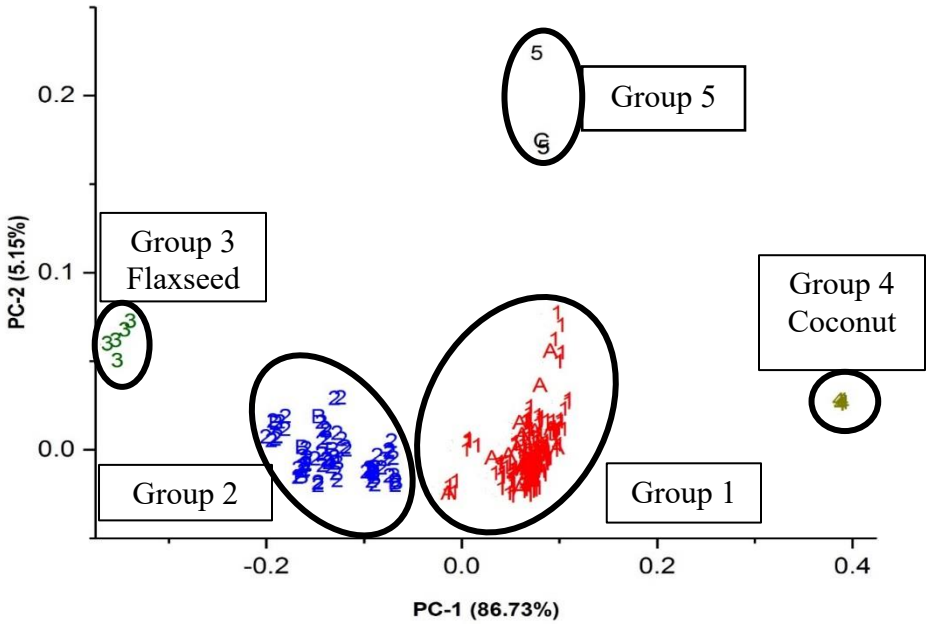

Figure 6.6. Projection of the 29 Raman spectra comprising the validation set onto the plot of the two largest principal components of the 361 features obtained from the 265 Raman spectra comprising the mean-centered training set. Training set: 1 = Group 1, 2 = Group 2, 3 = Group 3, 4 = Group 4, and 5 = Group 5. Validation set: A = Group 1, B = Group 2, and C = Group 5.

Three varieties of edible oils are well separated from the other twelve oils in the PC plot (see Figures 6.5 and 6.6): Coconut Oil (Group 4), Flaxseed Oil (Group 3) and the blend Avocado/Flaxseed/Olive (Group 5). The first principal component functionally distinguishes the oils by their degree of saturation. Of all the oils in this study (see Table 6.6), coconut oil has significantly higher percentage of saturated fats (~90% compared ~ 15% for the other oils) than the other oils while flax seed oil has one of the highest percentages of polyunsaturated fats (~70%). As for the remaining 12 edible oils, they were assigned to Groups 1 or 2. Group 1 edible oils (extra virgin olive, extra light olive, pure olive, peanut, avocado, safflower, and hazelnut oils) tend to have a greater degree of unsaturation than those from Group 2 (corn, vegetable, canola, sesame, and grapeseed oils).

This could manifest through a combination of a lower percentage of saturated fatty acids and a larger ratio of polyunsaturated fatty acids to monounsaturated fatty acids.

Table 6.6. Amounts of Saturated and Unsaturated Fats in Edible Oils

Edible Oil	Saturated	Monounsaturated	Polyunsaturated
Olive Oil	13%	74%	8%
Hazelnut	7%	80%	11%
Avocado Oil	10%	70%	20%
Peanut Oil	17%	49%	33%
Safflower Oil	9%	12%	74%
Grapeseed Oil	12%	17%	71%
Corn Oil	13%	24%	59%
Sesame Oil	14%	40%	46%
Canola Oil	7%	58%	32%
Vegetable Oil	15%	24%	61%
Flaxseed Oil	10%	19%	68%
Coconut Oil	87%	6%	2%

The next step was to investigate edible oils from Groups 1 and 2 individually using variable selection. The pattern recognition GA was applied to the 71 Raman spectra of Group 2 (see Table 6.7) to identify spectral features that were discriminatory as the spectral profiles of the edible oils comprising this Group were similar. For this study, the mutation rate of the GA was set at 0.4 and the number of chromosomes at 10,000. After 200 generations, the pattern recognition GA identified 23 spectral features that were correlated with edible oil type. These spectral features were identified by sampling key feature subsets, scoring their PC plots, and tracking those edible oils and/or Raman spectra that were difficult to classify. The boosting routine of the pattern recognition GA used this information to steer the population to an optimal solution. Figure 6.7 shows a plot of the two largest principal components of the 23 spectral features identified by the pattern recognition GA. The 8 validation set spectra assigned to Group 2 by the PC plot used to discriminate edible oils by their oil group (see Figure 6.7) were projected onto the PC plot

of the 71 Raman training set spectra from Group 2 and the 23 spectral features identified by the pattern recognition GA. All 8 Raman validation set spectra (see Table 6.7) were correctly classified, i.e., they were projected in a region of the PC plot (see Figure 6.8) with Raman spectra of the same type of edible oil.

Table 6.7. Training and Validation Set for Group 2 Edible Oils

Oil Type	Oil ID	Number of Samples	Training Set Spectra	Validation Set Spectra
Corn Oil	7	5	21	3
Grapeseed Oil	8	5	26	3
Canola Oil	13	7	11	0
Sesame Oil	15	1	9	1
Vegetable Oil	17	2	4	1
Total	-----	**20**	**71**	**8**

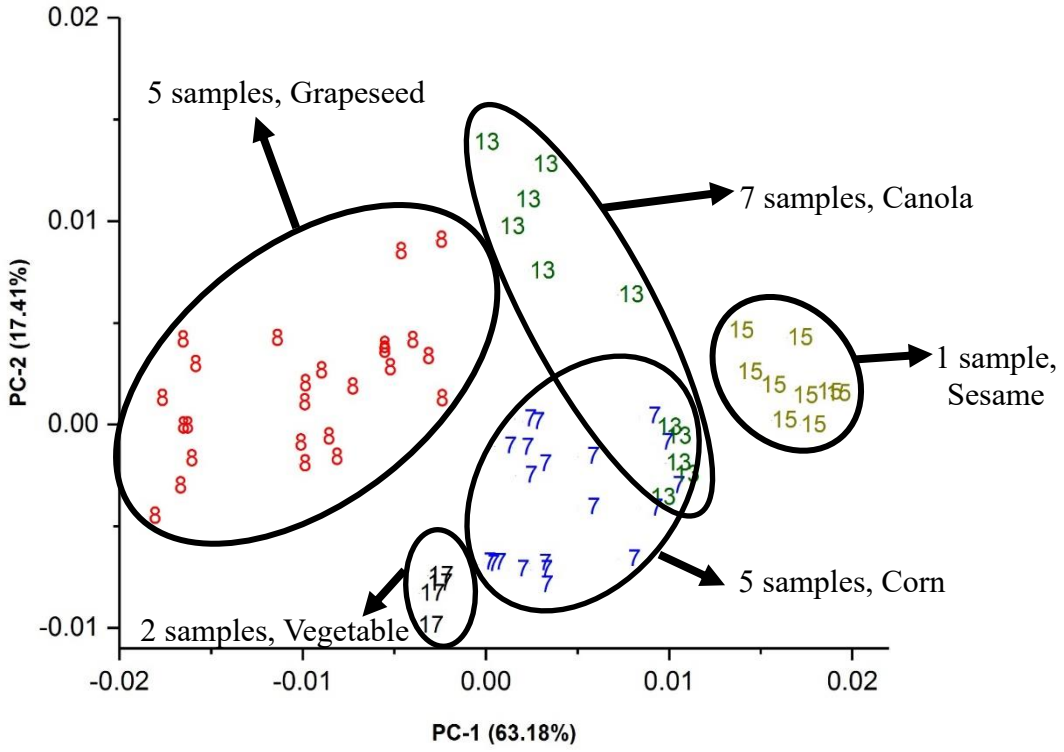

Figure 6.7. Plot of the two largest principal components of the 71 Raman spectra comprising the mean-centered training set for Group 2 and the 23 features identified by the pattern recognition GA. 7 = Corn oil, 8 = Grapeseed oil, 13 = Canola oil, 15 = Sesame oil, and 17 = Vegetable oil.

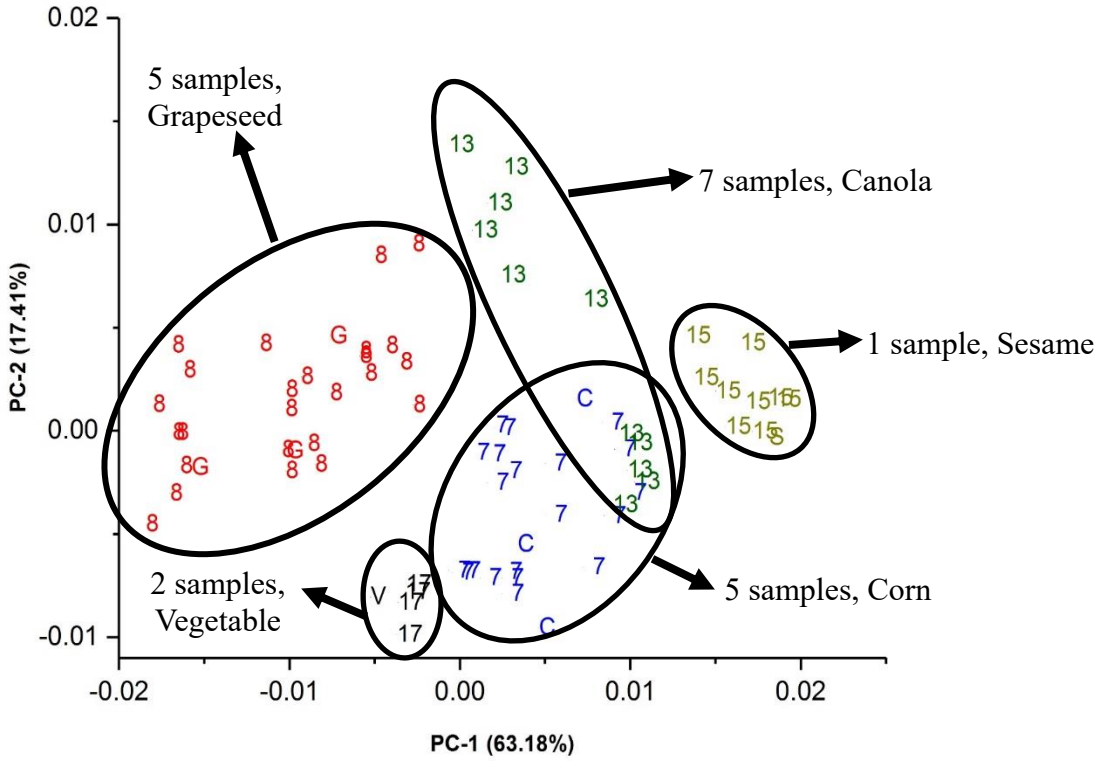

Figure 6.8. Validation set spectra projected onto the PC plot of the 71 Raman spectra comprising the training set for Group 2 and the 23 spectral features identified by the pattern recognition GA. Training set: 7 = Corn oil, 8 = Grapeseed oil, 13 = Canola oil, 15 = Sesame oil, and 17 = Vegetable oil. Validation set: C = corn oil, G = grapeseed oil, S = sesame oil, and V = vegetable oil.

Sesame, vegetable, and grapeseed oils were separated from each other and from the other edible oils in the plot (see Figures 6.7 and 6.8), whereas canola and corn oil overlap. In other studies on the application of vibration spectroscopy to discrimination of edible oils, many authors have reported that corn oil can be readily discriminated from canola oil [22]. In these studies, the authors only addressed within supplier sample variation in making their best case assessment for classification. In this study, both sample to sample variation within a supplier and between suppliers were considered for classification, which would explain the differences in our reported results for corn and canola oil from those reported in previously published studies.

The within class variability associated with each variety of edible oil in the PC plot is correlated to the number of samples representing the edible oil, not the number of Raman spectra comprising the class. Both vegetable and sesame oil yielded compact and well separated clusters in the PC plot as each oil is represented by only a single brand (i.e., only within supplier variation for sesame oil but seasonal variation within supplier variation for vegetable oil). Grapeseed, corn, and canola oils do not form compact clusters in the PC plot as each edible oil is represented by five or seven samples representing both within supplier including seasonal variation and between-supplier variation.

The 182 Raman spectra comprising the training set for Group 1 (see Table 6.8) were also investigated using the pattern recognition GA. Extra virgin olive oil, extra light olive oil, and pure olive oil could not be differentiated, which can probably be attributed to their triglyceride fraction being comparable [22]. For this reason, these three edible oils were merged into a single class. Variable selection was performed using the pattern recognition GA. After 200 generations, the pattern recognition GA identified 20 spectral features. Figure 6.9 shows a plot of the two largest principal components of the 174 Raman spectra comprising the training set and the 20 spectral features identified by the pattern recognition GA. Extra virgin olive oil, extra light olive oil, and pure olive oil are separated from the other edible oils in the PC plot (see Figure 6.9). Safflower, hazelnut, and avocado oil form compact and well defined clusters. Each contains spectra from only a single source (see Table 6.8). By comparison, peanut oil spectra in the plot are divided into two clusters. The three peanut oil samples (obtained from two different suppliers) capture both within-supplier and between-supplier variability. When taking into account this second source of

variability, it is plausible that peanut oil may not be able to be reliably differentiated from safflower oil using Raman spectroscopy.

Figure 6.10 shows the 20 validation set Raman spectra assigned to Group 1 by the PC plot used to discriminate edible oils by group (see Figure 6.9) projected onto the PC plot of the 182 Raman spectra comprising the training set for Group 1 and the 20 spectral features identified by the pattern recognition GA. The 17 extra virgin olive oil, extra light olive oil, and pure olive oil spectra (see Table 6.8) were assigned to the correct class in the PC plot. The Raman spectrum of safflower in the validation set was also correctly classified based on its projected location in the PC plot.

Table 6.8. Training and Validation Set for Group 1 Edible Oils

Oil Type	**Oil ID**	**Number of Samples**	**Training Set Spectra**	**Validation Set Spectra**
Extra Virgin Olive Oil	1	12	54	8
Extra Light Olive Oil	2	4	29	2
Pure Olive Oil	3	6	68	7
Avocado Oil	5	1	12	2
Peanut Oil	6	3	10	0
Safflower Oil	9	1	4	1
Hazelnut Oil	10	3	5	0
Total	-----	**30**	**182**	**20**

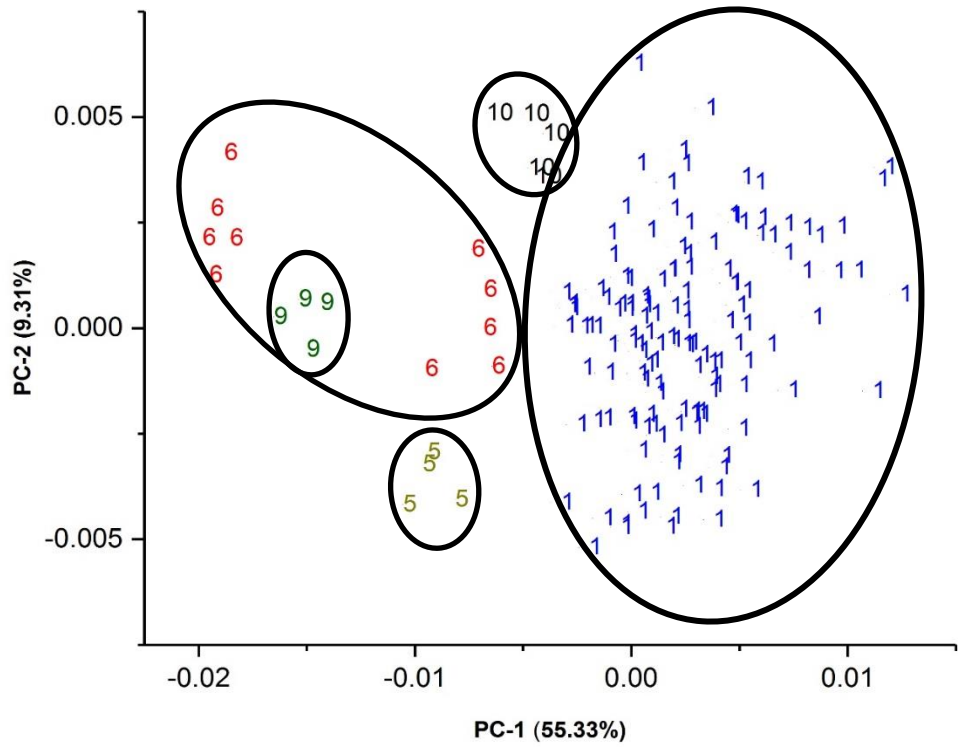

Figure 6.9. Plot of the two largest principal components of the 182 Raman spectra comprising the training set for Group 1 and the 20 spectral features identified by the pattern recognition GA. 1 = extra virgin olive oil, extra light olive oil and pure olive oil, 5 = avocado oil, 6 = peanut oil, 9 = safflower oil, 10 = hazelnut oil.

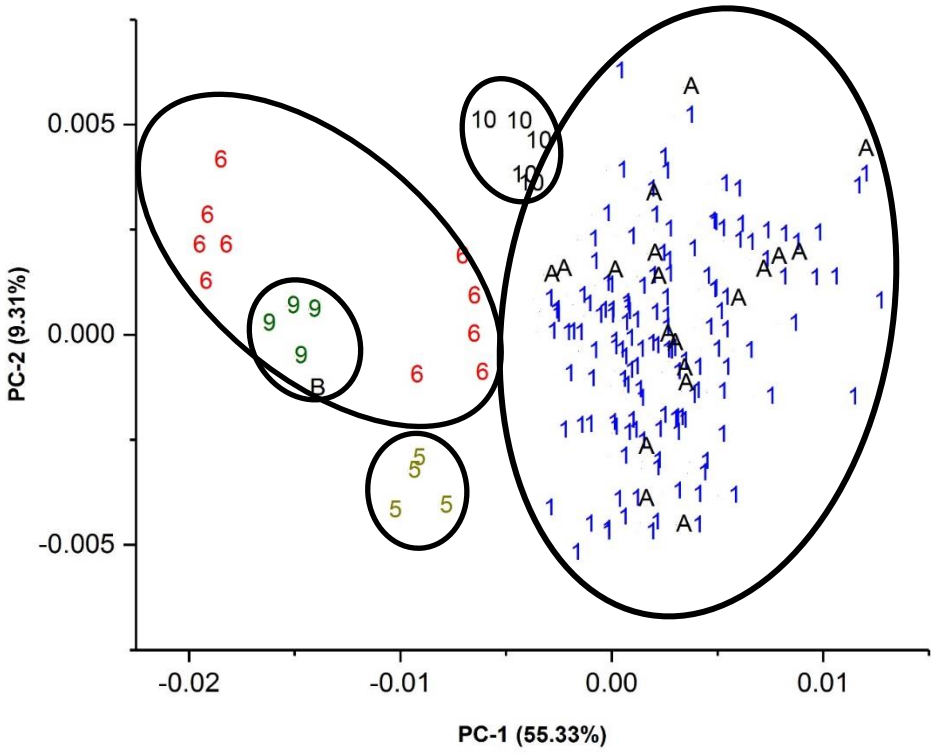

Figure 6.10. Validation set spectra projected onto the PC plot of the 182 Raman spectra comprising the training set for Group 1 and the 20 spectral features identified by the pattern recognition GA. Training set: 1 = extra virgin olive oil, extra light olive oil, and pure olive oil, 5 = avocado oil, 6 = peanut oil, 9 = safflower oil, 10 = hazelnut oil. Validation set: A = olive oils, B = safflower oil.

6.4.2. Data Set 2

A hierarchical classification scheme was used to discriminate the Raman spectra of the edible oils in the training set by edible oil type. To implement this scheme, the average spectrum of each variety or blend of edible oil in the training set was computed. The 15 average Raman spectra were then analyzed by principal component analysis (PCA) and hierarchical cluster analysis. For both PCA and hierarchical cluster analysis (i.e., Wards method), the spectral data were again mean centered. The PC plot and dendogram for the 15 average Raman spectra (see Figure 6.11), which are in agreement, show two clusters.

Therefore, the edible oils comprising the training set were divided into two groups of oils (see Table 6.9).

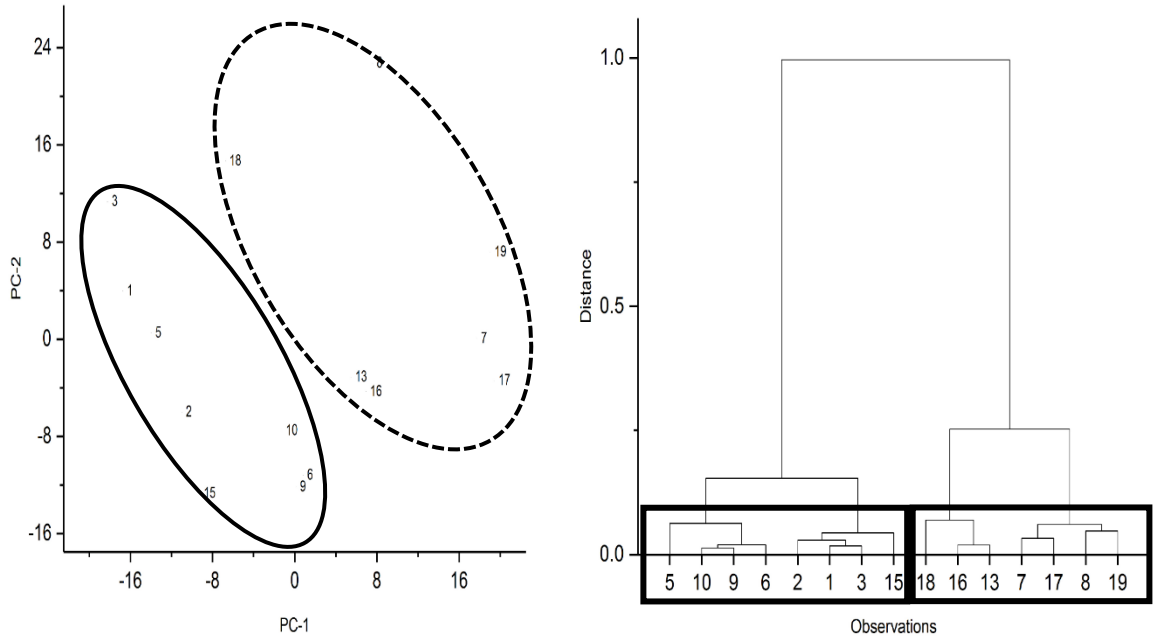

Figure 6.11. a) Plot of the two largest principal components and b) dendogram of the average Raman spectra of the 15 edible oils. Both the PC plot and the dendogram (Wards method) indicate that the edible oils can be divided into two oil groups.

Table 6.9. Group Assignments for Edible Oils

Oil Type	**Oil ID**	**Edible Oil Group**
Extra Virgin Olive Oil	1	1
Extra Light Olive Oil	2	1
Pure Olive Oil	3	1
Avocado	5	1
Peanut	6	1
Corn	7	2
Grapeseed	8	2
Safflower	9	1
Hazelnut	10	1
Canola	13	2
Sesame	15	1
Canola-Vegetable	16	2
Vegetable	17	2
Canola-Sun-Soybean	18	2
Sunflower	19	2

Figure 6.12 shows a plot of the two largest principal components of the 198 Raman spectra and 361 spectral features comprising the training set. Each spectrum is represented as a point in the PC plot. The two groups of edible oils previously detected by cluster analysis are separated in the PC plot.

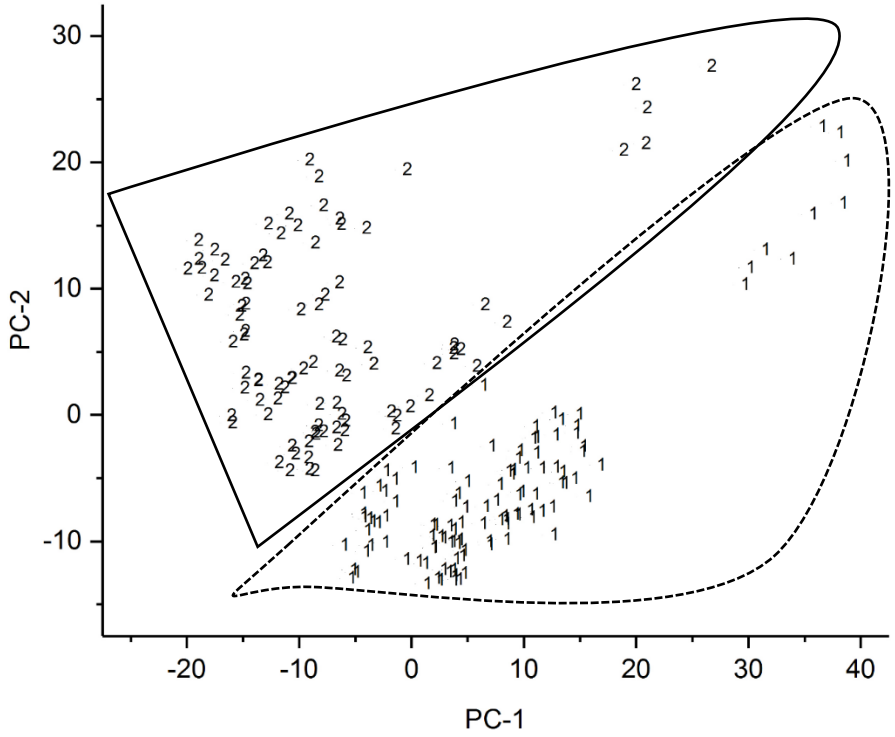

Figure 6.12. Plot of the two largest principal components of the 361 point Raman spectra comprising the training set. 1 = Group 1, and 2 = Group 2.

Variable selection was the next step as the deletion of uninformative features from the Raman spectral profiles ensures that discriminatory information about edible oil group is the major source of variation in the data. In addition, variable selection can transform a difficult classification problem into a simple one. However, the variable selection method employed should be multivariate in nature to ensure that crucial features will not be feature subsets, scoring their PC plots, and tracking those spectra and/or classes that were difficult

to classify. After 200 generations, the pattern recognition GA identified 11 wavelengths whose PC plot (see Figure 6.13) displayed two resolved and well separated clusters of spectra on the basis of the 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 (olive, peanut, avocado, safflower, hazelnut, and sesame) have a greater degree of unsaturation than those edible oils comprising Group 2 (corn, vegetable, canola, sunflower, canola-vegetable, canola-sun-soybean, and grapeseed). discarded. For these reasons, the pattern recognition GA was applied to the 198 training set spectra to identify discriminating wavelengths for edible oil group by sampling key.

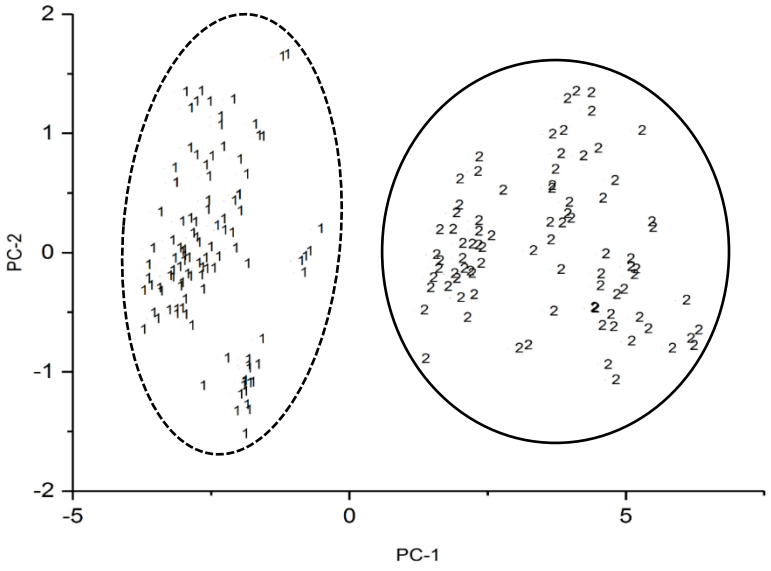

Figure 6.13. Plot of the two largest principal components of the 257 Raman spectra comprising the training set and the 11 spectral features identified by the pattern recognition GA. 1 = Group 1 and 2 = Group 2.

As for the 11 wavelengths selected by the pattern recognition GA (see Table 6.10), they correspond to two fundamental vibrational transitions: the =CH deformation and the C=C stretch-cis bands. Figure 6.14 shows the average Raman spectrum of the oils comprising each edible oil group. The 11 wavelengths selected by the pattern recognition GA for the training set spectra correspond to the bands in Figure 6.14 that are the most informative for discriminating these two groups based on a comparison of the average Raman spectrum of each group.

Table 6.10. Features Selected for Discrimination of Edible Oil Groups

Feature (Total of 361)	**Wavenumber (cm^{-1})**	**Assignment**
82	1284	=CH deformation
83	1286	=CH deformation
291	1657	C=C stretching (cis)
292	1658	C=C stretching (cis)
293	1660	C=C stretching (cis)
294	1662	C=C stretching (cis)
295	1663	C=C stretching (cis)
296	1665	C=C stretching (cis)
305	1680	C=C stretching (cis)
306	1682	C=C stretching (cis)
307	1684	C=C stretching (cis)

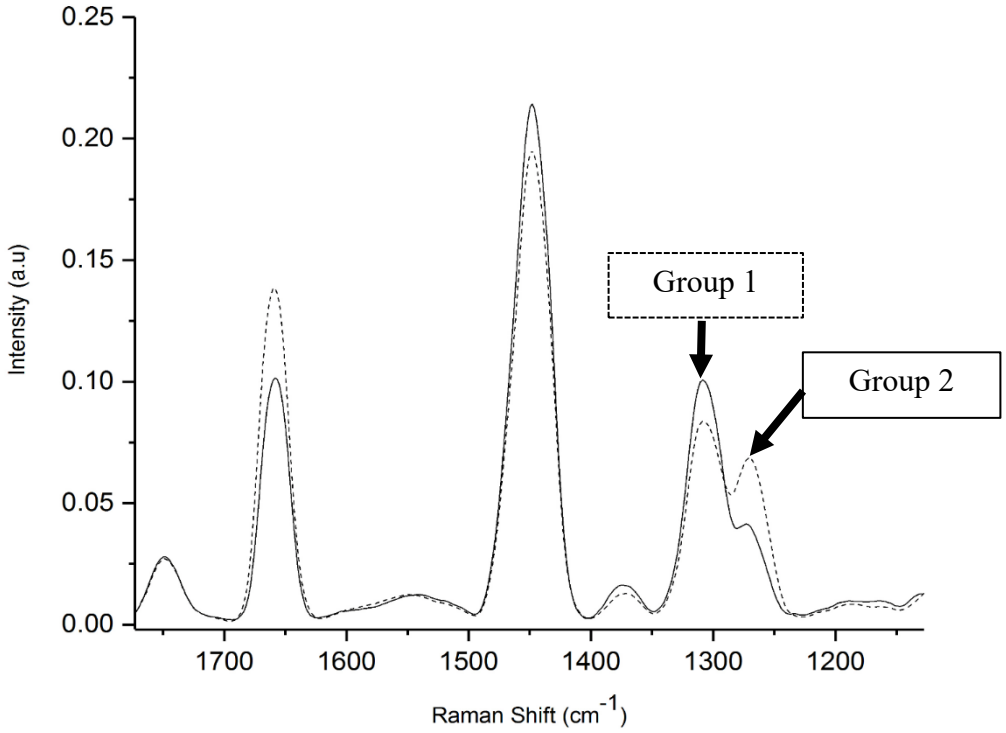

Figure 6.14. Average Raman spectra are shown for Group 1 (solid line) and Group 2 (dashed line). The 11 wavelengths selected by the pattern recognition GA correspond to the bands which are the most informative for discriminating these two groups based on a comparison of the average Raman spectrum computed for each oil group.

A validation set of 17 Raman spectra was used to assess the predictive power of the PC plot developed from the 198 Raman spectra comprising the training set and the 11 spectral features identified by the pattern recognition GA. The 17 Raman spectra were mapped onto the PC plot developed from these 11 spectral features. Figure 6.15 shows the validation set spectra projected onto this PC plot. Each projected validation set spectrum lies in a region of the PC plot with Raman spectra from the same edible oil group. This suggests that the hierarchical approach to discriminate the 15 varieties of edible oils by first dividing them into two groups is supported by the data.

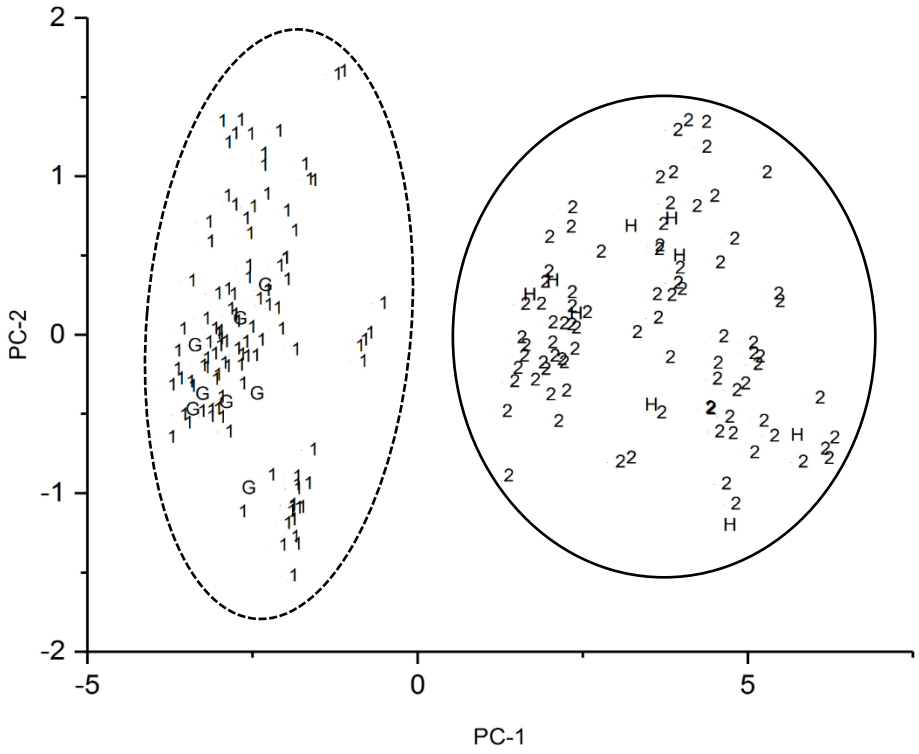

Figure 6.15. Projection of the 17 Raman spectra onto the PC plot developed from the 257 Raman spectra comprising the training set and the 11 spectral features identified by the pattern recognition GA. Training set: 1 = Group 1 and 2 = Group 2. Validation set: G = Group 1 and H = Group 2.

Variable selection was then performed on each edible oil group. The pattern recognition GA was applied to the 110 Raman spectra comprising the training set for Group 1 (see Table 6.11). The Raman spectra of EVOO, ELOO, and olive oil could not be differentiated using the pattern recognition GA as the triglyceride fraction of these three oils is identical. For this reason, the Raman spectra of EVOO, ELOO, and olive oil were combined into a single class. Furthermore, the pattern recognition GA was configured using edible oil type or sample identify as the object function against which variable selection was performed. Figure 6.16 shows a plot of the two largest principal components of the 110 Raman spectra comprising the training set and the 14 spectral features identified

by the pattern recognition GA using edible oil type as the object function (i.e., the Y-block of the classifier) against which variable selection was performed. (Sample identity refers to the spectra collected for a particular brand of edible oil purchased on a specific date). Figure 6.17 shows a PC plot of the 110 Raman spectra and 12 spectral features identified by the pattern recognition GA using sample identity as the object function. Each Raman spectrum is represented as a point in the two PC plots (see Figures 6.16 and 6.17) with the identity of the edible oil designated for each point in each plot. The 8 Raman spectra comprising the validation set for the Group 1 edible oils (see Table 6.11) were projected onto both PC plots. All 8 Raman spectra from the validation set were correctly classified in both plots, i.e. each spectrum lies in a region of the PC plot with Raman spectra of the same edible oil type.

Table 6.11. Training and Validation Set for Group 1 Edible Oils

Oil Type	**Oil ID**	**Number of Samples**	**Number of Training Set Spectra**	**Number of Prediction Set Spectra**
Extra Virgin Olive Oil	1	12	42	2
Extra Light Olive Oil	2	4	11	2
Pure Olive Oil	3	6	22	2
Avocado	5	1	5	0
Peanut	6	3	12	1
Safflower	9	1	3	0
Hazelnut	10	3	11	1
Sesame	15	1	4	0
Total		**31**	**110**	**8**

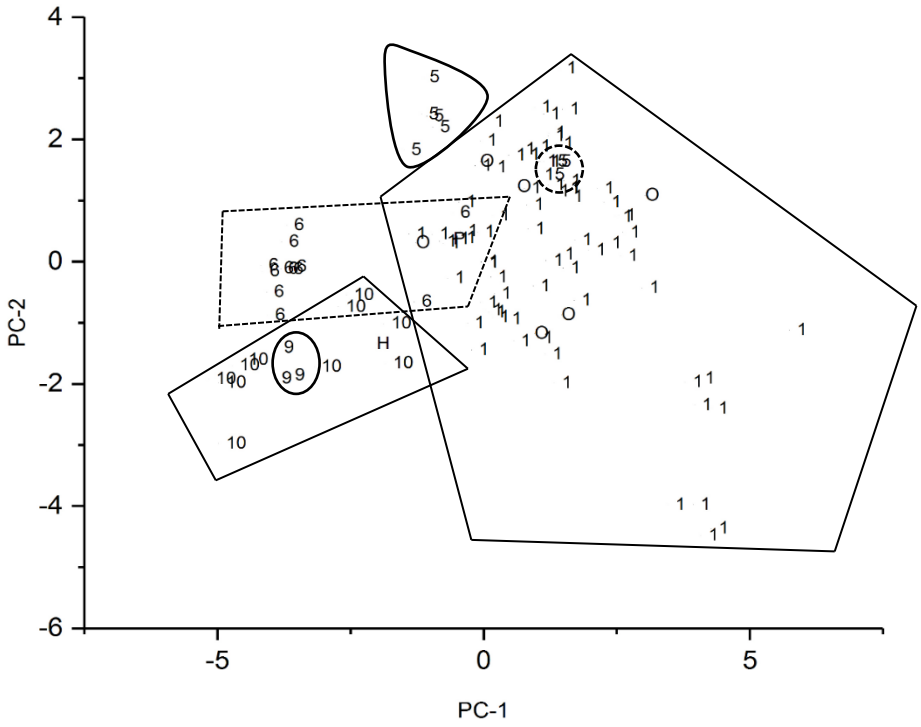

Figure 6.16. Plot of the two largest principal components of the 110 Raman spectra of the Group 1 edible oils comprising the training set and the 14 spectral features identified by the pattern recognition GA using edible oil type as the object function against which variable selection was performed by the pattern recognition GA. Validation set spectra for Group 1 are projected onto this PC plot. Training set: 1 = EVOO, ELOO, and olive oil, 5 = avocado oil, 6 = peanut oil, 9 = safflower oil, 10 = hazelnut oil, and 15 = sesame oil. Validation set: O = olive oils (EVOO, ELOO, and pure olive oil), H = hazelnut oil, P = peanut oil.

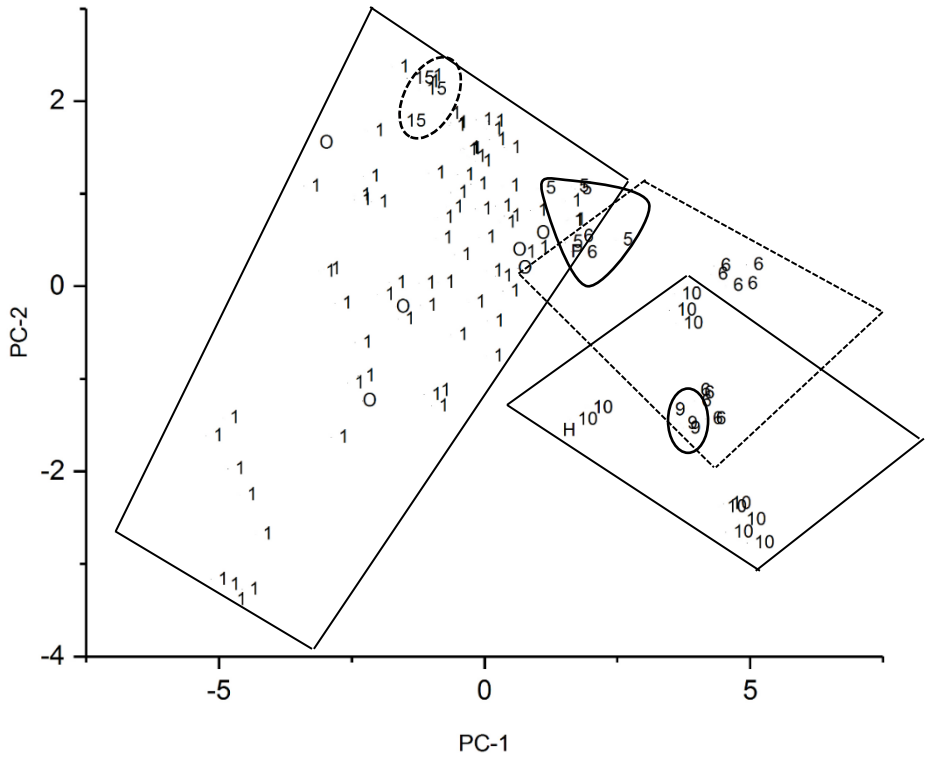

Figure 6.17. Plot of the two largest principal components of the 110 Raman spectra of the Group 1 edible oils comprising the training set and the 12 spectral features identified by the pattern recognition GA using sample identity as the object function against which variable selection was performed by the pattern recognition GA. Validation set spectra for Group 1 are projected onto this PC plot. Training set: 1 = EVOO, ELOO, and olive oil, 5 = avocado oil, 6 = peanut oil, 9 = safflower oil, 10 = hazelnut oil, and 15 = sesame oil. Validation set: O = olive oils (EVOO, ELOO, and pure olive oil), H = hazelnut oil, P = peanut oil.

When comparing these two PC plots, one observes the same classes being formed regardless of whether the object function used for variable selection is edible oil type or sample identity. The separation of the edible oils is better for some using edible oil type as the object function (e.g., avocado) and worse for others (e.g., hazelnut). Edible oils that have only samples from one source (e.g., avocado, sesame, and safflower oil) form tight clusters in the PC plot, whereas other edible oils that contain multiple sample sources (e.g., hazelnut, peanut, and the olive oils) are dispersed in the PC plot. Based upon an

examination of the PC plots shown in Figures 6.16 and 6.17, within-source variation (as represented by avocado, sesame and safflower oil) is small compared to between-source variation (as represented by hazelnut, peanut and the olive oils).

For the Group 2 edible oils (see Table 6.12), comparable results were obtained (see Figures 6.18 and 6.19). The same classes were formed, and the degree of separation between these classes in the two PC plots was the same regardless of the object function used by the pattern recognition GA (which was the variety of the edible oil or the identity of the sample). All Raman spectra comprising the validation set were correctly classified (see Figures 6.18 and 6.19). Within-source variation (as represented by the Raman spectra in the PC plot comprising canola-vegetable, canola-sun-soybean, and sunflower) is small compared to between-source variation (as represented by the Raman spectra in the PC plot comprising canola, corn, and grapeseed oil). In addition, between-source variation for the Group 2 edible oils was comparable to the variation associated with edible oil type for many Group 2 edible oils. For example, the between source variation for corn oil as represented by the average Raman spectra of samples 44 and 33 (see Figure 6.20) is comparable to the variation associated with edible oil type as represented by the average Raman spectra of canola and vegetable oil (see Figure 6.21).

Table 6.12. Training and Validation Set for Group 2 Edible Oils

Oil Type	Oil ID	Number of Samples	Number of Training Set Spectra	Number of Training Set Spectra
Corn	7	5	21	3
Grapeseed	8	5	21	3
Canola	13	7	25	2
Canola-Vegetable	16	1	8	0
Vegetable	17	2	5	0
Canola-Sun-Soybean	18	1	4	1
Sunflower	19	1	4	0
Total		**22**	**88**	**9**

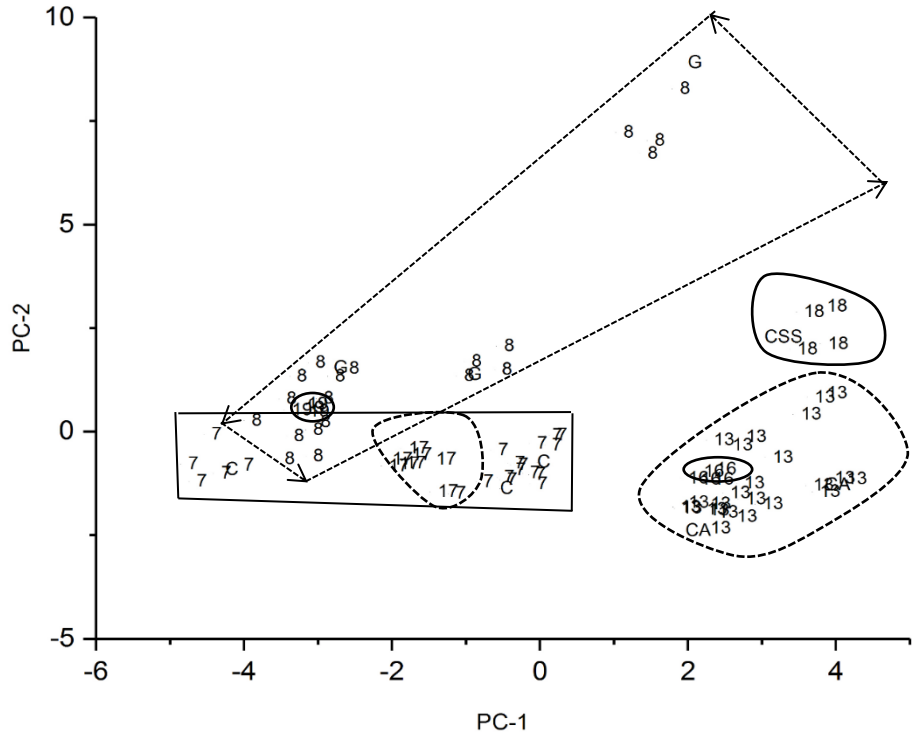

Figure 6.18. Plot of the two largest principal components of the 88 Raman spectra of the Group 2 edible oils comprising the training set and the 13 spectral features identified by the pattern recognition GA using edible oil type as the object function against which variable selection was performed by the pattern recognition GA. Validation set spectra for Group 2 are projected onto this PC plot. Training set: 7 = Corn oil, 8 = Grapeseed oil, 13 = Canola oil, 16 = Canola-Vegetable oil, 17 = Vegetable oil, 18 = Canola-Sun-Soybean oil, and 19 = Sunflower. Validation set: CSS = Canola-Sunflower-Soybean, CA = Canola, C = Corn, and G = Grapeseed.

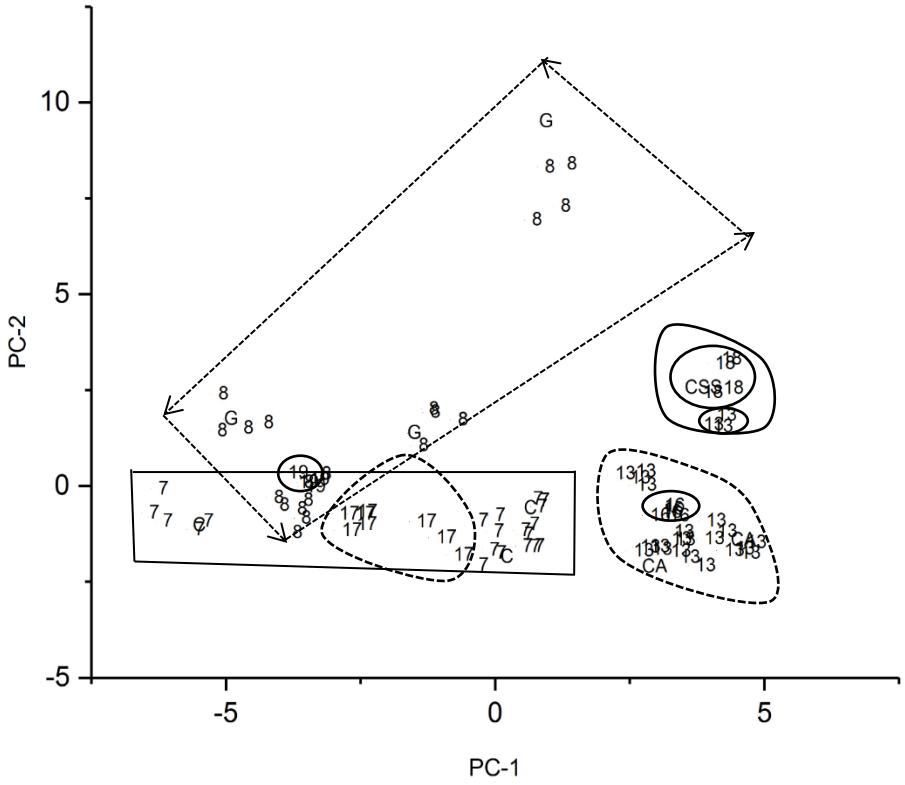

Figure 6.19. Plot of the two largest principal components of the 88 Raman spectra of the Group 2 edible oils comprising the training set and the 21 spectral features identified by the pattern recognition GA using sample identity as the object function against which variable selection was performed by the pattern recognition GA. Validation set spectra for Group 2 are projected onto this PC plot. Training set: 7 = Corn oil, 8 = Grapeseed oil, 13 = Canola oil, 16 = Canola-Vegetable oil, 17 = Vegetable oil, 18 = Canola-Sun-Soybean oil, and 19 = Sunflower. Validation set: CSS = Canola-Sunflower-Soybean, CA = Canola, C = Corn, and G = Grapeseed.

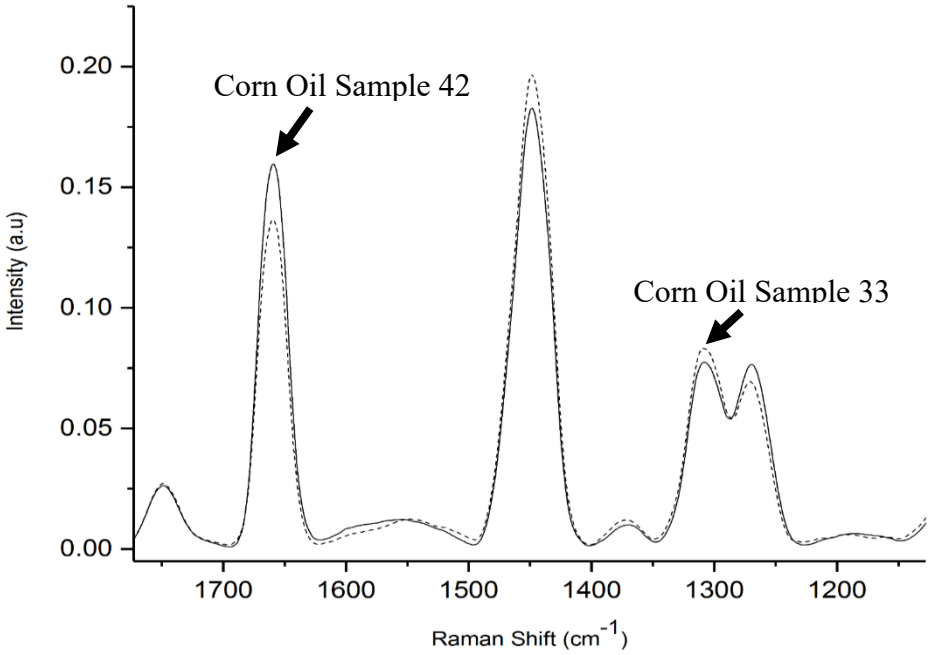

Figure 6.20. Average Raman spectrum of corn oil sample 44 (solid line) and the average Raman spectrum of corn oil sample 33 (dashed line). Sample 44 is comprised of the 5 Raman spectra that form a cluster adjacent to grapeseed oil in the PC plots shown in Figures 9 and 10 whereas the spectra comprising sample 33 are in the larger corn cluster.

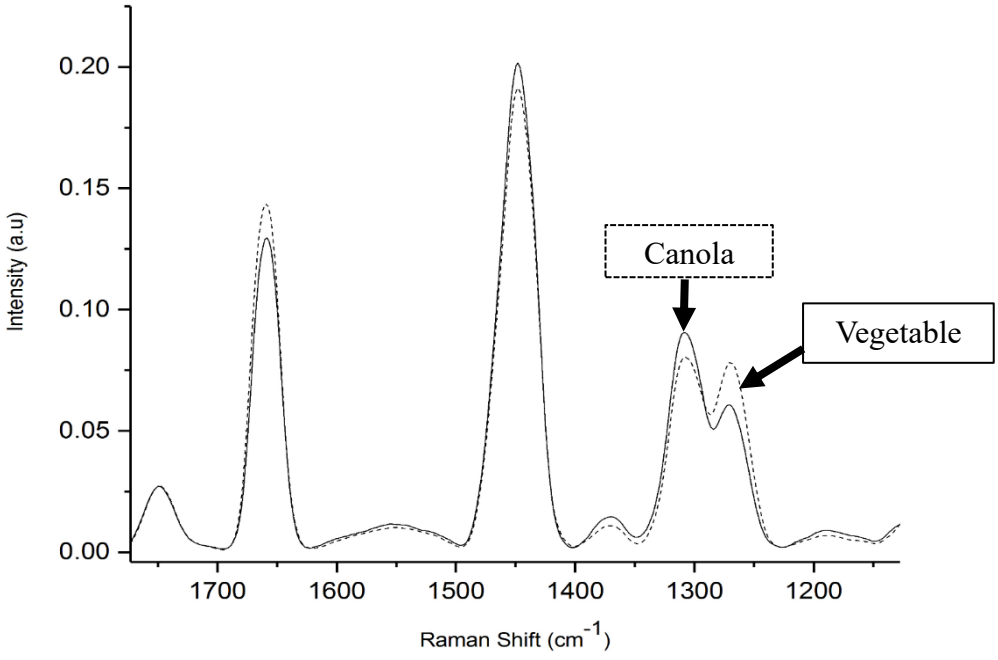

Figure 6.21. Average Raman spectrum of canola oil (solid line) and vegetable oil (dashed line).

As to the reason for why a successful classification of the Raman spectral data was obtained using sample identity as the GA's object function, either distinct separate subgroups exist within each oil type based on the sample identity or the classification of the Raman spectra of the edible oils in the training set is based on sample identity. Clearly, any classification of the triglyceride profiles of the edible oils as reflected by their Raman spectra can lead to spurious results and overfitting when one is attempting to discriminate edible oils by type within the same group.

6.5. Conclusion

This study reports on the use of Raman spectroscopic data to classify edible oils, using PCA, after the use of a genetic algorithm to perform wavelength selection. More importantly, the study's findings that large subsets of edible oils can be parsed using the PCA based methodology, and the validation sample subsets can be successfully classified would seem to indicate the validity of using a hierarchical classification scheme. The relative concentration among fatty acids and triglycerides varies among cultivar type, from season to season, and with the degree of ripeness of the fruit or seed at harvest [24]. Such variance propagates into the vibrational spectra of each oil as each fatty acid and triglyceride has a unique Raman spectral profile. To date, efforts to demonstrate the efficacy of vibrational spectroscopy to classify edible oils and to detect adulterants have presented 'best-case' scenarios with oils from a single batch. While edible oils from a particular batch are "nicely" clustered and can be differentiated from other classes of edible oils obtained from a single source, this study demonstrates that it is not possible to construct a single model that spans both seasonal and vendor variations for classification of edible oils and detection of adulterants in an edible oil.

References

1. P. Siri-Tarino, S.C., N. Bergeron, R. M. Krauss, *"Saturated Fats versus Polyunsaturated Fats versus Carbohydrates for Cardiovascular Disease Prevention and Treatment".* Ann. Rev. Nut. , 2015. **35**: p. 517-543.

2. J. Orsavova, L.M., J. V. Ambrozova, R., Vicha, J. Mlcek, *"Fatty Acids Composition of Vegetable Oils and its Contribution to Dietary Energy Intake and Dependence of Cardiovascular Mortality on Dietary Intake of Fatty Acids".* Int. J. Mol. Sci, 2015. **16(6)**: p. 12871–12890.

3. K. Chowdhury, L.A.B., S. Khan, A. Latif, *Studies on the Fatty Acid Composition of Edible Oil.* Bangladesh J. Sci. Ind. Res., 2007. **42(3)**: p. 311-316.

4. J. C. Moore, J.S., M. Lipp. , *"Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010".* J. Food Sci., 2012. **77(4)**: p. R118-R126.

5. T. Rezanka, P.M., *"Determination of Plant Triacylglycerols Using Capillary Gas Chromatography, High-Performance Liquid Chromatography and Mass Spectrometry".* J. Chromatog. , 1991. **542**: p. 145-159.

6. N. A. Marigheto, E.K.K., M. Defernez, R.H. Wilson, *"A Comparison of Mid-Infrared and Raman Spectroscopies for the Authentication of Edible Oils".* J. Amer. Oil Chem. Soc. , 1998. **75(8)**: p. 987-992.

7. V. Baeten, M.M., M.T. Morales, R. Aparicio., *"Detection of Virgin Olive Oil Adulteration by Fourier Transform Raman Spectroscopy".* J. Agricult. Food Chem. , 1996. **44(8)**: p. 2225-2230.

8. V. P. Baeten, M.T.H., M.T. Morales, R. Aparicio, *"Oil and fat Classification by FT-Raman Spectroscop".* J. Agricult. Food Chem. , 1998. **46(7)**: p. 2638-2646.

9. M. J. Lerma-Garcia, G.R.-R., J.M. Herrero-Martinez, E.F. Simo-Alfonso, *"Authentication of Extra Virgin Olive Oils by Fourier-Transform Infrared Spectroscopy".* Food Chem. , 2010. **118(1)**: p. 78-83.

10. O. O. Abbas Galtier, Y.L.D., C. Rebufa, J. Kister, J. Artaud, N. Dupuy, *"Comparison of PLS1-DA, PLS2-DA and SIMCA for Classification by Origin of Crude Petroleum Oils by MIR and Virgin Olive Oils by NIR for Different Spectral Regions".* Vibrat. Spec., 2011. **55(1)**: p. 132-140.

11. R. Korifi, Y.L.D., J. Molinet, J. Artaud, N. Dupuy, *Composition and Authentication of Virgin Olive Oil from French PDO Regions by Chemometric Treatment of Raman Spectra.* J. Raman Spectros. , 2011. **42(7)**: p. 1540-1547.

12. Y. Saucedo-Hernandez, M.J.L.-G., J.M. Herrero-Martinez, G. Ramis-Ramos, E. Jorge-Rodriguez, E.F. Simo-Afonso, *"Classification of Pumpkin Seed Oils According to Their Species and Genetic Variety by Attenuated Total Reflection Fourier-Transform Infrared Spectroscopy".* J. Agricul. Food Chem., 2011. **59(8)**: p. 4125-4129.

13. X. F. Zhang, M.Q.Z., X.H. Qi, F. Liu, C. Zhang, F. Yin, *Quantitative Detection of Adulterated Olive Oil by Raman Spectroscopy and Chemometrics".* J. Raman Spec. , 2011. **42(9)**: p. 1784-1788.

14. S. F. Li, X.R.Z., J.H. Zhang, G.Y. Li, D.L. Su, Y. Shan, *"Authentication of Pure Camellia Oil by Using Near Infrared Spectroscopy and Pattern Recognition Techniques".* J. Food Sci. , 2012. **77(4)**: p. C374-C380.

15. P. Samyn, D.V.N., G. Schoukens, L. Vonck, D. Stanssens, H. Van den Abbeele, *"Quality and Statistical Classification of Brazilian Vegetable Oils Using Mid-Infrared and Raman Spectroscopy".* Appl. Spec. . , 2012. **66(5)**: p. 552-565.

16. R. El-Abassy, M., P. Donfack, A. Materny *"Visible Raman Spectroscopy for the Discrimination of Olive Oils from Different Vegetable Oils and the Detection of Adulteration".* J. Raman Spec., 2009. **40(9)**: p. 1284-1289.

17. G. Gurdeniz, B.O., *"Detection of Adulteration of Extra-Virgin Olive Oil by Chemometric Analysis of Mid-Infrared Spectral Data".* Food Chem., 2009. **116(2)**: p. 519-525.

18. W. Dong, Y.Q.Z., B. Zhang, X.P. Wang, *Quantitative analysis of adulteration of extra virgin olive oil using Raman spectroscopy improved by Bayesian framework least squares support vector machines.* Anal. Meth. , 2012. **4** p. 2772-2777.

19. G. Schulze, A.J., M. M. L. Yu, A. Lim, R. F. B. Turner, M. W. Blades, *"Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation".* Appl. Spec., 2005. **59(5):**: p. 545-574.

20. M. A. Sharaf, D.L.I., B. R. Kowalski, *Chemometrics, .* 1986, NY John Wiley & Sons.

21. Eilers, P.H.C., *A Perfect Smoother.* Anal. Chem., 2003. . **75(14)**: p. 3631-3636.

22. Ruiz, C.J.-S.a.J.R., *"Use of Raman Spectroscopy for Analyzing Edible Vegetable Oils"* Appl. Spec. Rev. , 2016. **51(5)**: p. 417-430.

23. Lavine BK, D.C., Moores AJ, Griffiths PR. , *Raman spectroscopy and genetic algorithms for the classification of wood types.* Appl. Spec. , 2001. **55**: p. 960-966.

24. Beltran G, D.R.C., Sanchez S, Martinez L., *Influence of harvest date and crop on the fatty acid composition of virgin olive oils from Cv. Picual.* J. Agric. Food Chem., 2004. **52**: p. 3434-3440.

CHAPTER VII

CONCLUSION

The focus of this dissertation was forensic automotive paint analysis. Modern automotive paint systems consist of multiple layers of paint: a clear coat over a color coat, which in turn is over a surfacer-primer and e-coat layer. Since forensic laboratories in North America analyze each layer of paint individually by FTIR, time must be spent to hand-section each layer and then present each separated layer to the spectrometer for analysis. Sampling too close to the boundary between adjacent layers can produce an IR spectrum that is a mixture of two layers. In the situation of searching an automotive paint data base, not having a "pure spectrum" of each layer prevents a forensic paint examiner from developing an accurate hit list of potential suspects. One way to minimize the time necessary for data collection is to collect IR data from all layers in a single analysis by scanning across the cross-sectioned layers of the paint sample using a FTIR microscope equipped with an imaging detector. Once the data has been collected, it can then undergo decatenation using chemometrics to obtain a "pure" IR spectrum of each layer. This approach, not only eliminates the need to analyze each layer separately resulting in a considerable time savings, but can also ensure that the final spectrum of each layer is "pure" and not a mixture.

Thirty-two automotive paint samples from six manufacturers (General Motors, Chrysler, Ford, Toyota, Nissan, and Honda) within a limited production year range (2000-2006) were obtained from the Royal Canadian Mounted Police. Although several resins were investigated as embedding media in this project (including Tuffleye® Finish blue light – Wet A hook Technologies, Quick cure™ (Bob Smith Industries 5 min epoxy) and Embed-it™ Low viscosity epoxy kit (Polysciences®), it was the Slow-cure™ (Bob Smith Industries) thirty minute epoxy resin that was selected as the embedding medium. The thirty minute epoxy resin and hardener mixture were poured into flat polyurethane embedding molds (BEEM®, Polysciences), and the paint sample was placed into the mold and oriented perpendicular to the bottom surface prior to polymerization of the epoxy. Paint samples in the thirty minute epoxy block were then placed in an oven at 60°C for ninety minutes to ensure total curing. After hardening, the epoxy block was removed from the mold and positioned in the microtome to ensure that a thin cross section (approximately 4 to 5 µm thick) cut by the microtome contained all four paint layers.

Each thin cross section was collected, placed on a barium fluoride disk, and examined for defects, which would appear as dirt or cracks and crevices in an otherwise smooth surface when examined under a Leica light microscope. For embedded paint samples**,** a portion of the barium fluoride disk covered with cured epoxy without sample was run for background at $4cm^{-1}$ resolution before the image map of the embedded paint sample was obtained. Transmission IR image maps generated at $4cm^{-1}$ resolution using an iN10 MX microscope (Thermo-Nicolet, Madison, WI) equipped with a liquid nitrogen cooled mercury cadmium telluride (MCT) single imaging detector were collected for each cross sectioned automotive paint sample. For the analysis of the automotive paint samples,

a 20 micron aperture and 5 micron step size yielded the best results when the microscope was operated in transmission mode.

For multivariate curve resolution (MCR), a line map was extracted from the IR image of each cross sectioned paint sample. To obtain a line map, a transit (line) was passed through an IR image map of each paint sample. All spectra in contact with the transit were extracted, and the resulting collection of spectra was referred to as a line map. The data for the line map was taken on an oblique transit in order to include all paint layers and as many spectra of each layer and of the mixed interfacial region between the layers. Because the spatial resolution of the imaging microscope in transmission mode (for example) is 25 microns, the likelihood of capturing spectra characteristic of the boundary between two layers using the set of criteria for defining the oblique transit is high as the thickness of the undercoat (e-coat and surfacer primer) layers are approximately 10 μm and 20μm respectively and the clear coat layer is approximately 50μm thick.

For MCR analysis, it was necessary for the data to be free of both noise and experimental artifacts. For this reason, spectra were extracted from the line map and checked for artifacts that may have been a direct result of the extraction procedure used. The IR spectrum of each layer of the automotive paint was reconstructed from the line map using ALS. Our previous experience with ALS has shown that initial estimates of the concentration (score) or spectral (loading) matrices are crucial for rotating these two matrices towards a correct solution. For this reason**,** a varimax extended rotation developed by Lavine and used in two previous studies to resolve severely overlapped liquid chromatographic peaks or decatenate Raman images of oil in water emulsions was applied to the spectral line maps to compute the initial estimates of the concentration and spectral

matrices for ALS. The spectral region for decatenation of the IR spectra from each line map was $4000cm^{-1}$ - $748cm^{-1}$.

Library searching of IR spectra from the PDQ database was performed using search prefilters (i.e., discriminants) to identify the vehicle manufacturer and assembly plant of the vehicle from the reconstructed IR spectra of the clear coat, surfacer-primer and e-coat layers of the cross sectioned paint sample. To develop these search prefilters, the IR spectra were preprocessed using the discrete wavelet transform, which was applied to the fingerprint region of each layer to enhance subtle but significant features in the IR spectra. The Symlet mother wavelet (sixth smallest filter size, eighth level of decomposition) was chosen for preprocessing because the shape of its scaling function closely matched that of the shape of the bands comprising the IR spectra of the automotive paints. Three sets of wavelet coefficients, one for each layer, were concatenated (both approximation and detail coefficients) to form the sample pattern vectors used by the search prefilters. Wavelet coefficients characteristic of manufacturer or assembly plant were identified using a genetic algorithm for pattern recognition and feature selection. The wavelet transformed spectra were autoscaled to ensure that each coefficient has a mean of zero and a standard deviation of one throughout the entire set of transformed spectra. Search prefilters to identify automotive manufacturer were developed using 1652 OEM paint systems from General Motors, Chrysler, Ford, Honda, Nissan, and Toyota within a limited production year range (2000-2006). Search prefilters for assembly plant of a specific manufacturer were previously developed by members of our research group.

Initially, all cross sectioned paint samples were incorrectly classified by the library search prefilters for manufacturer and assembly plant when they were applied to the

reconstructed IR spectra obtained from the transmission line maps. Almost all peaks in the reconstructed IR transmission spectra were shifted compared to the corresponding PDQ transmission spectra of the same sample. The PDQ library consists of IR spectra collected by FTIR spectrometers, each equipped with a diamond cell. The diamond cell applies high pressure to the sample which causes shifts to occur in some IR bands due to a reduction in the free volume of the polymer. For some peaks in the fingerprint region, these shifts are large (at least $4cm^{-1}$), whereas for others they are smaller ($\sim 0.3cm^{-1}$).

To solve this problem, IR transmission spectra from the PDQ library used to develop the search prefilters were converted into ATR spectra using an ATR simulation algorithm previously developed by Lavine and coworkers. The ATR simulation algorithm, which was developed to convert transmission spectra from PDQ into ATR spectra, is able to compensate for most of these spectral shifts. The search prefilters were recomputed using the transformed IR spectral data, and the reconstructed IR spectra from the line maps were transformed to ATR spectra using the ATR simulation algorithm. The IR spectrum of each reconstructed paint layer was preprocessed in the same manner as the ATR spectra that comprised the training sets for the search prefilters. All thirty-two unembedded paint samples were then correctly classified as to the manufacturer, line, and model of the vehicle from which the paint sample originated.

These results are significant as there are clear advantages for cross sectioning paint samples without the use of epoxy in IR imaging. Sample preparation is faster and more straight-forward. Decatenation of the image data is also more straight-forward as spectral interference from the epoxy layer, which is a well-known problem among workers in IR and Raman microscopy, does not occur. Strict unimodality can be enforced in the

mathematics of the decatenation process implemented in this study using ALS since there are no epoxy layers to model. (The epoxy is in contact with both the clear coat and e-coat layer as they are the outer- and inner-most layers of the intact paint sample.) Third, library searches are also simplified as the number of components that need to be specified by the ALS algorithm are fewer and the accuracy of the reconstructed IR spectra for which the search is run is higher.

Only twenty-seven of the thirty-two original paint samples were analyzed using epoxy resin because there was an insufficient amount of sample remaining after analysis in transmission and ATR modes with unembedded paint samples. Of the twenty-seven embedded paint samples, twenty-two were correctly identified as to manufacturer and assembly plant using the search prefilters. Although improvements in baseline correction and restricting the MCR analysis to the fingerprint region improved the decatenation of the spectral line maps for these six samples, the problems encountered with these samples - the mixing of IR spectra of the epoxy with the clear coat or e-coat layers or the mixing of IR spectra of adjacent paint layers - remained. These problems appear to be linked to the compression of the cross sectioned paint sample by the epoxy, which causes a decrease in the thickness of each layer of the automotive paint. For an OEM automotive paint system, embedding a paint sample in an epoxy may be problematic for some paint systems when one or more layers are too thin.

Much of the research described in this final summary overview is directly targeted to enhance current approaches to forensic automotive paint analysis through decreased data collection times as compared to current practices and to aid in evidential significance assessment, both at the investigative lead stage and at the courtroom testimony stage.

Direct impact on over 75 local, state, and federal forensic laboratories that are currently using the PDQ database in the United States is anticipated. There may also be direct impact on international forensic laboratories using the database, including the Forensic Laboratory Services Division of the RCMP, the Centre of Forensic Sciences in Toronto, Canada, the ENFSI network of European forensic science institutes, the Australian Police Services, and the New Zealand Police Services. The research described in this dissertation is an international collaborative effort between the Lavine research group at Oklahoma State University and Mark Sandercock of the RCMP. The use of the prototype pattern recognition assisted infrared library search system previously developed by Lavine and Sandercock in tandem with FTIR imaging will ensure that fewer hits are generated in a PDQ library search. This can translate into a significant time savings for the forensic scientist. Furthermore, information derived from the search prefilters for vehicle manufacturer can serve to quantify the general discrimination power of original automotive paint comparisons and further efforts to succinctly communicate the significance of the evidence.

VITA

Francis Kwofie

Candidate for the Degree of

Doctor of Philosophy

Thesis: VIBRATIONAL SPECTROSCOPY AND CHEMOMETRICS APPLIED TO THE FORENSIC EXAMINATION OF AUTOMOTIVE PAINTS AND EDIBLE OILS

Major Field: Chemistry

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Chemistry major at Oklahoma State University, Stillwater, Oklahoma in July, 2019.

Completed the requirements for the Master of Science in Chemistry at East Tennessee State University, Johnson City, Tennessee in 2015.

Completed the requirements for the Bachelor of Science in Chemistry at University of Cape Coast, Cape Coast, Ghana in 2012.

Experience:

I served as a Graduate Research Associate and a Graduate Teaching Assistant at Oklahoma State University from January 2016 to July 2019.

I also served as a Graduate Teaching Assistant at East Tennessee State University from January 2014 to August 2015.

I served as a Teaching Assistant at University of Cape Coast from September 2012 to August 2013.

Leadership Positions:

President, Oklahoma State University Chapter of Phi Lambda Upsilon Honorary Chemical Society, 2018-2019.

Vice President, Oklahoma State University Chapter of Phi Lambda Upsilon Honorary Chemical Society, 2018-2019.