The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Document Title:	**Improving the PDQ Database to Enhance Investigative Lead Information from Automotive Paints**
Author(s):	**Barry K. Lavine, Collin White, Undugodage Perera, Koichi Nishikda, Matthew Allen**
Document No.:	**249893**
Date Received:	**May 2016**
Award Number:	**2012-DN-BX-K059**

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this federally funded grant report available electronically.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Cover Page

Report Title: Improving the PDQ Database to Enhance Investigative Lead Information from Automotive Paints

Award Number: 2012-DN-BX-K059

Authors: Barry K. Lavine, Collin White, Undugodage Perera, Koichi Nishikda, and Matthew Allen

Abstract:

New techniques for pattern recognition assisted infrared (IR) library searching of the Paint Data Query (PDQ) automotive database have been developed to determine the make and model of an automobile from an unknown paint sample recovered at a crime scene. Modern automotive paints have a thin color coat, which on a microscopic fragment, may be too thin to obtain accurate chemical information. Furthermore, the small size of the fragment also makes it difficult to accurately compare it with manufacturer's paint color standards. Since primer and clear coat paint layers are usually unique to the automotive assembly plant where these layers have been applied, combining chemical information obtained from the Fourier Transform infrared (FTIR) spectra of the two primer layers and from the clear coat layer should make it possible to rapidly and accurately identify the make and model of an automobile from its paint system alone. The application of IR search prefilters and a cross correlation library searching algorithm to the PDQ database, which is a major thrust of this research project, is crucial to extract investigative lead information from clear coat and primer layer paint smears. Information derived from these pattern recognition searches can serve to quantify the general discrimination power of original automotive paint comparisons encountered in casework and to succinctly communicate the significance of the evidence to the courts. To maintain relevancy of the newly designed pattern recognition techniques, the analysis of additional paint samples has been simultaneously undertaken to populate the database with samples from production years where there are insufficient data.

A correction algorithm to allow attenuated total reflection (ATR) spectra to be matched using the IR transmission spectra of the PDQ database has also been developed as part of this research project. ATR is a widely used sampling technique in IR spectroscopy because minimal sample preparation is required. As the penetration depth of the ATR analysis beam is shallow, the outer layers of a laminate or multi-layered paint sample can be preferentially analyzed with the entire sample intact. For this reason, forensic laboratories have taken advantage of ATR to collect IR spectra of automotive paint systems which may consist of three or more layers. However, the IR spectrum of a paint sample obtained by ATR will exhibit distortions, e.g., band broadening and lower relative intensities at higher wavenumbers, when compared to its transmission counterpart. This hinders library searching as most library spectra are measured in transmission mode. The correction algorithm to convert transmission spectra from the PDQ library into ATR spectra is able to address distortion issues such as the relative intensities and broadening of the bands, and introduction of wavelength shifts at lower frequencies, which prevent library searching of ATR spectra using archived IR transmission data.

Table of Contents

Abstract 1

Executive Summary 3

I. Introduction 5

II. Methodology 7

III. Results 15

Development of Search Prefilters for Manufacturer 15

Development of Search Prefilters for Chrysler 19

Development of Search Prefilters for General Motors 34

Development of Search Prefilters for Ford 50

Simulation of Attenuated Total Reflection Infrared Absorbance Spectra 68

IV. Conclusions 85

Discussion of Findings 85

Implications for Policy and Practice 86

Implications for Future Research 86

V. References 87

VI. Dissemination of Research Findings 89

Executive Summary

Automotive vehicles can be identified from paint fragments transferred onto another vehicle or onto the clothing of a pedestrian involved in a hit-and-run accident by comparing the color, layer sequence, and chemical composition of each individual layer of the paint [1, 2]. To make these comparisons possible, the Royal Canadian Mounted Police have developed a comprehensive database for forensic automotive paint analysis known as the paint data query (PDQ) database. PDQ is a database of the physical attributes, the chemical composition, and the infrared (IR) spectrum of each layer of the original manufacturer's automotive paint system. If the original automotive paint system is present in the recovered paint fragment, PDQ can assist in identifying both the make and model of the automotive vehicle within a limited production year range. Currently, PDQ contains 21,000 samples (street samples and factory panels) corresponding to over 84,000 individual paint layers representing the paint systems used in most domestic and foreign vehicles marketed in North America. Each year over 500 samples are painstakingly collected, analyzed and added to the PDQ database by the RCMP.

To use PDQ, the forensic chemist must translate the color of each automotive paint layer and its chemical composition based on the IR spectrum of each layer into specific text codes. The text based search and retrieval system of PDQ searches the database, comparing all records for make, model and year having a paint system similar to the coded information provided by the user. The final step in the process is to confirm these hits manually (as direct searching of IR spectra in PDQ does not exist) by comparing the IR spectrum of each unknown paint layer against the IR spectra in the database hit list. Topcoat color is compared to topcoat color charts to narrow down this hit-list to those manufacturers known to have used a similar topcoat color in those samples identified by the database search.

Searches performed in PDQ 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. For example, modern automotive clear coats in PDQ have one of only two possible formulations: acrylic melamine styrene or acrylic melamine styrene polyurethane. Furthermore, modern automotive paints contain a thin color coat, and on a microscopic fragment it may be too thin to obtain accurate chemical information. The small size of the fragment makes it difficult to accurately compare it with manufacturer's paint color standards. Another problem stems from forensic laboratories using attenuated total reflection (ATR) spectroscopy for infrared analysis of automotive paints. Although ATR is a widely used sampling technique in IR spectroscopy because minimal sample preparation is required, the IR spectrum of an automotive paint sample obtained by ATR exhibits distortions. Specifically, band broadening and lower relative intensities at higher wavenumbers occur when compared with its transmission counterpart. This hinders library searching as most library spectra are measured in transmission mode.

Fortunately, adhesion between layers in modern automotive paint systems is usually strong. Often, the primer layers are transferred during a collision when both the clear coat and color coat layers are also transferred. As the primer and clear coat layer are often unique to the assembly plant where these layers were applied, combining chemical information obtained from the Fourier Transform (FT) IR spectra of the two primer layers and from the clear coat layer makes it possible to rapidly and accurately identify the make and model of the automobile within a

limited production year range from the paint system alone. Applying data fusion techniques where data (e.g., spectra) from multiple sources (e.g., IR spectra of clear coat and primer paint layers) are combined and class membership information is extracted, search prefilters have been developed to determine the make and model of the vehicle from which an unknown paint sample was obtained. Even in challenging trials where the clear coat and undercoat layers evaluated were all the same manufacturer (e.g., General Motors, Chrysler, or Ford) within a limited production year range, the respective assembly plant of the vehicle was correctly identified using only the information from the FTIR spectra of the clear coat and undercoat paint layers. The development of a pattern recognition driven library search system, consisting of search prefilters to truncate the PDQ library to a specific assembly plant or assembly plants and a cross correlation library search algorithm to identify spectra that are most similar to the unknown in the set identified by the search prefilters (as well as characterizing the degree to which the truncated PDQ library fits the unknown paint sample), was crucial for extracting investigative lead information from clear coat and primer layers. Information derived from these pattern recognition searches also served to quantify the general discrimination power of original automotive paint material encountered in casework and to succinctly communicate the significance of the evidence to the courts.

A correction algorithm to allow ATR spectra to be searched using IR transmission spectra of PDQ has also been developed as part of this research project. Conversion of transmission spectra to ATR spectra were performed by taking advantage of a surface reflection phenomenon at the boundary between the sample and the internal reflection element of the spectrometer. In this procedure, the reflection of the incident beam from the internal reflection element is described at the boundary with the sample by Fresnel's equations. Since transmission spectra in the PDQ library do not exhibit sloping baselines or baseline offsets indicative of light scattering, the resulting ATR spectra obtained when applying the correction were of high quality. The proposed correction algorithm to convert transmission spectra from the PDQ library to ATR spectra is able to address distortion issues such as the relative intensities and broadening of the bands, and introduction of wavelength shifts at lower frequencies, which prevent library searching of ATR spectra using archived IR transmission data.

I. Introduction

In the forensic examination of automotive paint, each layer of paint is visually and chemically analyzed. The paint sample examined often consists of multiple and unique layers of paint. For architectural paint, forensic science is normally interested in comparing each layer from a crime scene, such as a door frame, to a suspect, such as a pry bar found in the suspect's possession. Likewise, for automotive paint, paint found on the clothing of a victim of a hit-and-run incident may be forensically compared to the paint from a suspect's vehicle. However, often there are no witnesses to a hit-and-run and police are unable to develop a suspect. In these situations the chemistry of the automotive paint layers recovered from the victim's clothing may be analyzed and, with the aid of an automotive paint database, the data can be correlated with a particular vehicle make and model within a limited production year range.

Modern automotive paint systems [1] consist of three or four layers: a clear coat over a color coat which in turn is over two undercoats. (White trucks often do not have a clear coat layer and so only have two primers and a color coat layer.) With the exception of the clear coat, each paint layer contains pigments and fillers (the colored component), and all layers contain binders (the matrix that holds the layer together). Automotive manufacturers tend to use unique combinations of fillers and binders in each layer of paint. It is this unique combination that allows forensic scientists to determine the possible manufacturer and model of a vehicle within a limited production year from an automotive paint chip recovered at the crime scene.

The chemical analysis of automotive paint samples in forensic laboratories is typically done using Fourier transform infrared (FTIR) spectroscopy [2]. Some laboratories, particularly in Europe, will embed the entire paint fragment, cross-section it, and then analyze each layer using an infrared (IR) microscope fitted with an attenuated total reflectance (ATR) accessory [3]. Other forensic laboratories, particularly in North America, are more likely to hand-section each layer and present each separated layer to either an IR microscope fitted with an ATR accessory, or collect transmittance spectra directly by placing the layer between diamond anvils.

Studies [4, 5] conducted over 35 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 individual layer in a paint system. To make the comparisons possible, a comprehensive database was developed as well as the means to search and retrieve information from it. Today, the Paint Data Query (PDQ) database contains over 21,000 samples (street samples and factory panels), that corresponds to over 84,000 individual paint layers, representing the paint systems used on most domestic and foreign vehicles marketed in North America. PDQ is a database of the physical attributes (i.e., color), the chemical composition and the IR spectrum of each layer of the original manufacturer's paint system. The PDQ concept is to narrow the list of possible vehicles to a number of suspects, not to identify a single vehicle [6, 7]. If the original paint layers are present in a recovered (i.e. unknown) paint chip, PDQ can assist in identifying the specific manufacturer and the production year of the automotive vehicle from which it came. The comparison of the IR spectrum of each paint layer in a paint system (clear coat, surface, and primer layers) to IR spectra in PDQ allows for the assembly plant at which the paint system was applied, and the production year within a limited production year range, to be identified. IR library searches based on the color coat layer are not performed because the paint chemistry may be color dependent and spectra from this layer are often of poor

quality as the IR signal is obscured by the metal and the pearlescent effect flakes mixed in the layer.

PDQ is comprised of data which contains the complete color, chemical composition, layer sequence and sourcing information on known paint systems, and search and retrieval software used to generate a hit list. To use PDQ, the forensic scientist must first translate the chemical formulation of the paint layer into specific text codes based on the IR spectrum, and then the scientist will enter the color, chemical composition, and layer sequence information derived from the examination and analysis of the unknown paint chip left at the scene of the crime. The software searches the database, comparing all records for make, model and year having a paint system similar to the coded information being searched. The final step in the process is to confirm the database hits by manually comparing the IR spectrum of each unknown paint layer against the spectra identified in the database hit list, which will often vary from 50 to 200 hits. Topcoat color is compared to topcoat color charts to narrow down the hit list so that only those manufacturers known to have used a similar topcoat color in the years indicated by the database search are reported.

A major problem with PDQ for modern automotive paint systems is its use of text to code the chemistry of each layer. Searches of the PDQ database require the user to code their FTIR spectrum according to the guidelines set out by the database, and to search these codes against the codes in the database. The coding used in PDQ is generic, and can lead to non-specific search criteria which results in a large number of spurious hits that a scientist must then work through and eliminate. This impairs the accuracy of a search. For example, the presence of styrene in the paint layer of a sample can be coded for that sample in the database but the amount could be small or large, a feature that could be easily distinguished by visual inspection of the spectrum but cannot be searched for using the text-based system of PDQ. Thus, initial PDQ searches for styrene will return a large number of hits that span multiple makes, models, and years.

Another problem is that modern automotive paint systems have a thin color coat which on a microscopic fragment may be too thin to obtain accurate chemical and topcoat color information. The small size of the fragment will make it difficult to accurately compare it with manufacturer's paint color standards. Most forensic laboratories rely on PPG or DuPont color refinish books for making color comparisons on paint chips recovered from crime scenes. The color represented in these books is intended for use by the refinish/auto body industry and are accurate on a macroscopic scale. While the color can be viewed microscopically, such as under a sterero-microscope, details such as effect flake size and distribution are not accurately reproduced and do alter the appearance of the color somewhat on a microscopic scale. The accuracy of such comparisons diminishes with the size of the paint chip recovered from the crime scene. In cases where the automotive paint sample is limited to the clear coat paint layer, the text based portion of PDQ cannot identify the automotive vehicle because modern clear coats in PDQ are coded as either acrylic melamine styrene or acrylic melamine styrene polyurethane.

A third problem is that forensic laboratories are using attenuated total reflection (ATR) as an infrared sampling technique with increasing frequency to collect FT-IR spectra of automotive paint samples because minimal sample preparation is required. As the penetration depth of the

ATR analysis beam is shallow, the outer layers of a laminate or multi-layered paint sample can be preferentially analyzed with the entire sample intact. For this reason, forensic laboratories have taken advantage of ATR to collect IR spectra of modern automotive paint systems which usually consists of four layers. However, an IR spectrum of a paint sample obtained by ATR exhibits distortions, e.g., band broadening and lower relative intensities at higher wavenumbers, when compared to its transmission counterpart. This can hinder library searching as most library spectra are measured in transmission mode.

II. Methodology

Pattern recognition assisted IR library searching techniques have been developed to search the spectral libraries of the PDQ database in an effort to differentiate between similar but nonidentical FTIR paint spectra and to correctly identify an unknown paint sample as to the manufacturer and model of the vehicle within a limited production year range. Paint samples are often recovered from hit-and-run accidents where damage to vehicles or injury or death to a pedestrian has occurred. Searches of the PDQ database using commercial software have met with only limited success. Because the PDQ automotive paint library is composed of a large number of similar spectra, commercial search algorithms have not proven to be sufficiently sensitive at distinguishing subtle but significant features in the data such as shoulders, unique shapes, and patterns, and minor peaks. All commercial library search algorithms involve some type of point-by-point numerical comparison between the IR spectrum of an unknown and each member of the library [8]. These algorithms lack interpretive ability because they treat the spectrum as a set of points rather than as a collection of specific bands. Furthermore, band shifting is not handled well and bands of low intensity, which may be highly informative, are often ignored.

Utilizing search prefilters, many of the problems encountered in library searching have been addressed. Most spectral comparisons performed during a search are of little use because the spectra in question are very dissimilar. A prefilter is a quick test to spot dissimilar spectra, thereby avoiding a complete spectral comparison. Prefilters used in this study allowed for more sophisticated but also for more time-consuming algorithms to be used for spectral comparisons since the library has been culled down for a specific match. The exceptionally high quality of the FTIR data in the PDQ database, and the comprehensiveness of this database, made it an excellent source of data for the development and subsequent validation of search prefilters.

To develop the search prefilters, chemical information from FTIR spectra of the two primer layers and the clear coat layer were combined and then subsequently analyzed using a genetic algorithm (GA) for features selection and pattern recognition. Spectral features in each FTIR spectrum characteristic of the assembly plant (and hence the manufacturer and model) of the vehicle were identified by the pattern recognition GA [9-22], which utilized both supervised and unsupervised learning to identify features that optimize the separation of the FTIR spectra by assembly plant 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 the features selected by the pattern recognition GA were about differences between the different classes (i.e., assembly plants) in the database. A principal component plot that shows separation of the data by class can only be generated using features whose variance or information is primarily about differences between these classes. This fitness criterion dramatically reduces the size of the search space since it limits the search to these types of feature subsets. In addition, the pattern

GA focused on those classes and/or samples that were difficult to classify as it trained by boosting the relative importance of classes and samples that consistently scored poorly. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The pattern recognition GA used in these studies integrated aspects of artificial intelligence and evolutionary computations to yield a "smart" one -pass procedure for wavelength selection and pattern classification.

This idea is demonstrated in Figure 1 which shows a plot of the two largest principal components of a data set prior to feature selection.

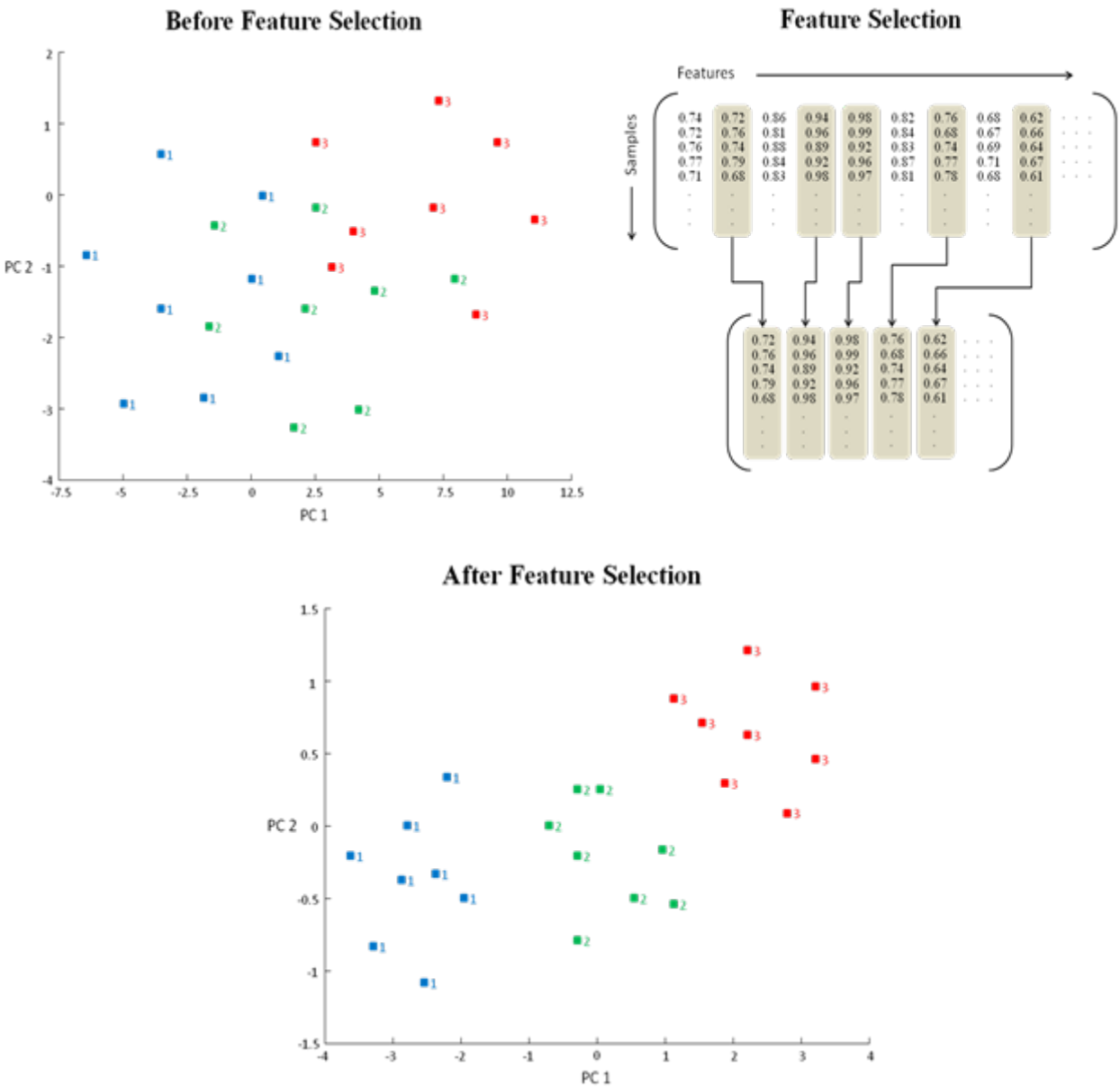

Figure 1. A plot of the two largest principal components developed from the 10 wavelengths in the data set does not show class separation. When principal components are developed from the wavelengths that contain information about the automobile model, clustering on the basis of model is evident.

The hypothetical data set consists of 30 IR spectra of clear coats (2000-2003) from Chryslers distributed between 3 assembly plants (1 = Newark/Durango, 2 = Toluca/PT Cruiser, and 3 = Toledo/Cherokee). Each paint sample in this example is characterized by 10 spectral features. However, only four of these features contain information about model type. When a principal component plot of the data is developed using only these four spectral features, clustering of spectra on the basis of the assembly plant (i.e., vehicle model) is evident.

To develop the search prefilters, FTIR spectra of the fingerprint region (1500 cm^{-1} to 600 cm^{-1}) of the clear coat and the two primer layers for each paint sample were combined into a single data vector. Since each FTIR spectrum in the PDQ database was collected with 4 cm^{-1} resolution, this region was characterized by 506 points in each FTIR spectrum. To combine the chemical information obtained from the clear coat and two primer layers, the first 506 elements of the data vector representing the paint sample will be the corresponding fingerprint region of the clear coat layer, the next 506 elements of the vector will be the first primer layer and the final 506 elements of the vector will be the second primer layer (see Figure 2). The pattern recognition GA will then identify the components of this data vector (i.e., specific features in each paint layer) that are correlated to the assembly plant of the vehicle from which the paint sample was obtained.

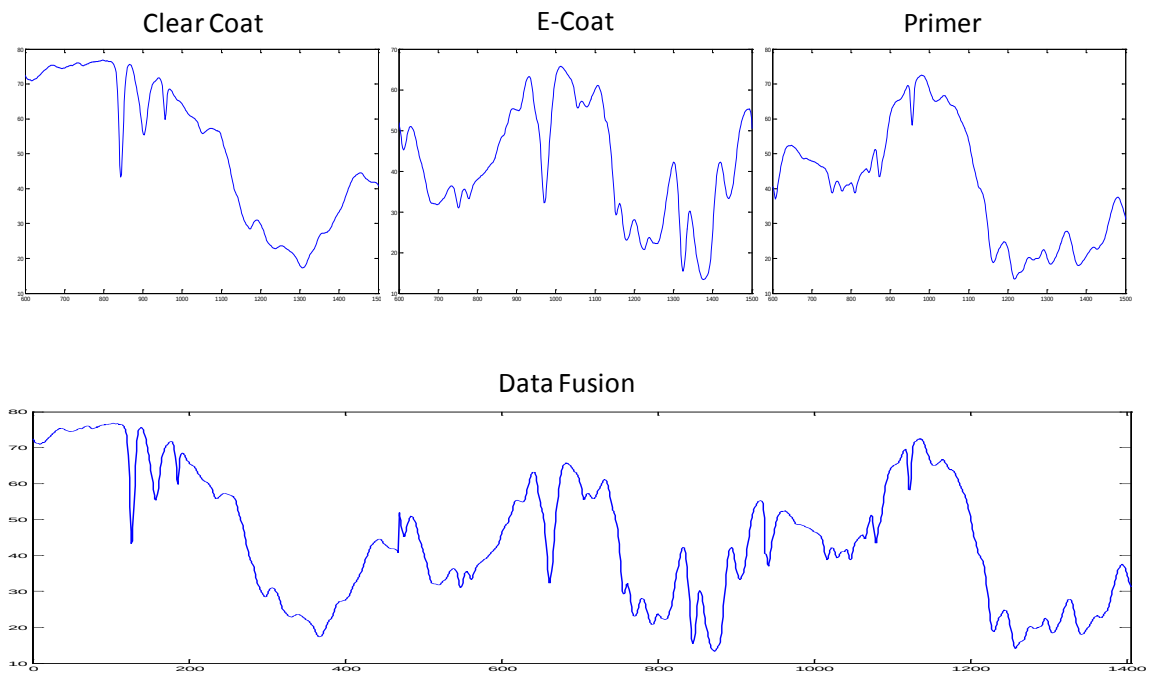

Figure 2. Clear coat, E-coat (first primer layer), primer (surfacer or second primer layer) and fused FT-IR spectrum

To develop the search prefilters, all fused IR spectra were preprocessed using wavelets [23-25] to enhance subtle but significant features in the data and to remove noise. Wavelets offer a different approach to removal of noise from multivariate data. Using wavelets, a new set of basis vectors that take advantage of the local characteristics of the data are developed that were better at conveying the information present in the data than axes defined by the original measurement variables (absorbance value at each wavelength in the IR spectrum). The mother wavelet selected to develop this new basis set is the one that best matches the attributes of the data. This

often circumvents the problem that occurs when an interfering source of variation in the data is correlated to information about the class membership of the samples, e.g., assembly plant, as a result of the design of the study or because of accidental correlations between signal and noise.

According to wavelet theory, a discrete signal such as a spectrum can be decomposed into approximation components and detail components. Deleting the approximation component with the lowest frequencies can result in the removal of baseline-like information from the data. If the scales representing signal are identified and retained and the scales representing background and noise are removed**,** an enhancement of signal to noise occurs with a reduction in the dimensionality of the data because of the elimination of the wavelet coefficients corresponding to noisy or uninformative spectral features. Classification of spectra is improved by selectively combining the scales.

Using wavelets, each fused spectrum is passed through two scaling filters: a high pass filter and a low pass filter. The low-pass filter will allow only the low frequency component of the signal to be measured as a set of wavelet coefficients which is called the "approximation". The high-pass filter will measure the high frequency coefficient set which is called the "detail". The detail coefficients usually correspond to the noisy part of the data. This process of decomposition is continued with different scales of the wavelet filter pair in a step-by-step manner to separate the noisy components from the signal until the necessary level of signal decomposition has been achieved. Figure 3 shows the first level of wavelet decomposition applied to an IR spectrum of an undercoat paint layer displayed in the transmittance mode.

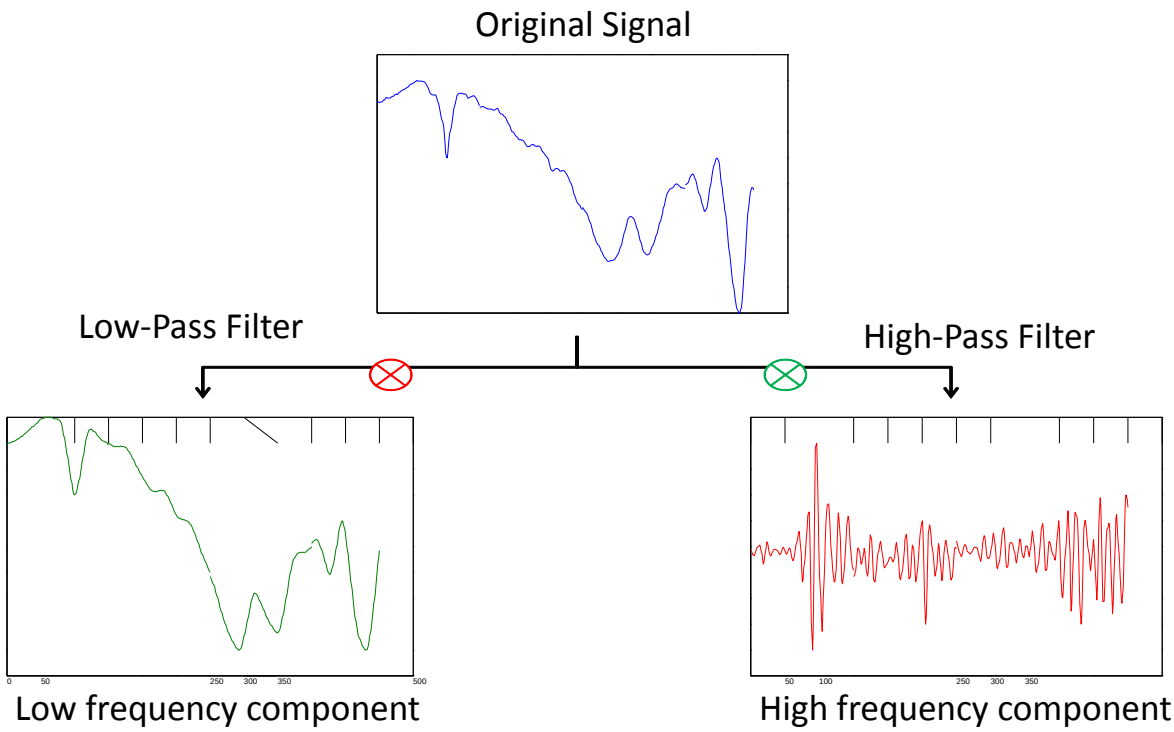

Figure 3. First level of wavelet decomposition applied to a clear coat paint spectrum displayed in the transmittance mode.

Wavelet coefficients from the fused IR spectra characteristic of the assembly plant of the vehicle were identified by the pattern recognition GA. The fitness function of the pattern recognition GA emulates human pattern recognition through machine learning to score the principal component plots and thereby identify a set of wavelet coefficients that optimize the separation of the automotive classes (i.e., assembly plants) in a plot of the two or three largest principal components of the data. To facilitate the tracking and scoring of the principal component plots, class weights (CWs) and sample weights (SWs), which are an integral part of the fitness function, are computed (see equations 1 and 2) where CW(c) is the weight of class c (with c varying from 1 to the total number of classes in the data set). $SW_c(s)$ is the weight of sample s in class c. The class weights sum to 100, and the sample weights for the objects comprising a particular class sum to a value equal to the class weight of the class in question.

$$\mathrm{CW(c)} = 100 \frac{\mathrm{CW(c)}}{\sum_{\mathrm{c}} \mathrm{CW(c)}} \qquad (1)$$

$$\mathrm{SW}(s) = \mathrm{CW(c)} \frac{\mathrm{SW}(s)}{\sum_{\mathrm{s} \in \mathrm{c}} \mathrm{SW}(s)} \qquad (2)$$

Scoring is performed on each principal component plot using the K-nearest neighbor (K-NN) classification algorithm. For a given sample point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to largest, and a poll is taken of the point's k nearest neighbors. For the most rigorous classification of the data, K equals the number of samples in the class to which the sample point belongs. The sample hit count (SHC), the number of like nearest neighbors, is $0 \leq \mathrm{SHC(s)} \leq K_c$. The fitness is computed using Equation 3.

To better understand how a principal component plot is scored, consider a data set with two classes, which have been assigned equal weights. Class 1 has 50 samples, and class 2 has 10 samples. At generation 0, the samples in a given class have the same weight. Thus, each sample in class 1 has a sample weight of 1, whereas each sample in class 2 has a weight of 5. Suppose a sample from class 2 has as its nearest neighbors 8 class one samples. Hence, SHC/K = 0.8, and (SHC/K)*SW = 0.8*5, which equals 4. By summing (SHC/K_c)*SW for each sample, each principal component plot can be scored - see Equation 3 where K_c is set by the user and for the most rigorous classification of the data K_c equals the number of samples in the class with the same class label and SHC (sample hit count) is the number of samples with the same class label as the sample in question. One advantage of using this procedure to score the principal component plots is that a class with a large number of samples does not dominate the analysis.

$$\sum_{\mathrm{c}} \sum_{\mathrm{s} \in \mathrm{c}} \frac{1}{K_c} \times \mathrm{SHC(s)} \times \mathrm{SW(s)} \qquad (3)$$

The fitness function of the pattern recognition GA is able to focus on those samples and/or assembly plants that are difficult to classify by boosting their sample and class weights over successive generations. In order to perform boosting, it is necessary to compute both the sample-hit rate (SHR), which is the mean value of SHC/K_c over all feature subsets formulated in a

particular generation (see equation 4), and the class-hit rate (CHR), which is the mean sample hit rate of all samples in a class (see equation 5). ϕ in equation 4 is the number of chromosomes in the population, and AVG in equation 5 refers to the average or mean value. During each generation, class and sample weights will be adjusted using a perceptron (see Equations 6 and 7) with the momentum, P, set by the user. (g + 1 refers to the current generation, whereas g is the previous generation.) Classes with a lower class hit rate are boosted more heavily than those classes that score well.

$$\mathrm{SHR(s)} = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC_i}(s)}{\mathrm{K_c}} \qquad (4)$$

$$\mathrm{CHR_g(c)} = AVG(\mathrm{SHR_g(s)} : \forall_{s\in c}) \qquad (5)$$

$$\mathrm{CW_{g+1}(s)} = \mathrm{CW_g(s)} + \mathrm{P(1 - CHR_g(s))} \qquad (6)$$

$$\mathrm{SW_{g+1}(s)} = \mathrm{SW_g(s)} + \mathrm{P(1 - SHR_g(s))} \qquad (7)$$

Boosting is crucial to ensure the successful operation of the pattern recognition GA because it modifies the fitness landscape by adjusting the values of the class and sample weights. This helps to minimize the problem of convergence to a local optimum. Hence, the fitness function of the pattern recognition GA is continually changing using information from previous generations as the population is evolving towards a solution.

Search prefilters (i.e. discriminants) have been developed from fused IR spectra of the fingerprint region of the clear coat, surfacer, and primer layers that extracted information from the fused IR spectrum of an unknown automotive paint sample to yield a response based on the assembly plant of the corresponding vehicle. Spectral features encoded in the wavelet coefficients identified by the pattern recognition GA have been used to develop the classifiers that serve as our search prefilters. In this project, we focused on the development of search prefilters to identify the assembly plant from fused IR spectra obtained from 25 General Motors (GM), 12 Chrysler, and 17 Ford car and truck assembly plants between the years 2000-2006. During this time period, GM, Chrysler, and Ford had the largest number of assembly plants in North America. If search prefilters can be developed that are able to discriminate automobiles assembled at one GM, Chrysler, or Ford plant from those assembled at another and can differentiate among different automobile manufacturers, we believe that this would be the best possible test of the proposed methodology to demonstrate the validity of this concept.

Search prefilters developed from fused spectra eliminated dissimilar spectra from the library search thereby providing the analyst with an opportunity to take advantage of more sophisticated but also more time-consuming search algorithms. Commercial infrared library search systems compare IR spectra by summing the squares of the difference between two spectra at every wave number. However, these algorithms do not perform well when differentiating between similar but nonidentical spectra as small peak shifts are not handled well and bands of low intensity,

which may be highly informative, are often ignored. For these reasons, a cross correlation function was used to provide the best match between an unknown and the spectra in the hit list generated by our search prefilters. The cross correlation function has been shown to correctly identify unknown spectra from similar but nonidentical spectra [26]. Although it is slower than conventional search algorithms, it is suitable as a post searching method to rank probable matches, which have been selected by a faster algorithm (e.g., search prefilters). Correlation based searches are insensitive to instrumental noise and very sensitive to changes in peak shape and in the relative peak position making them sensitive to structural differences.

Library matching was performed by cross correlating the unknown with each spectrum in the set of library spectra identified by the search prefilters and comparing each cross correlated spectra with the corresponding autocorrelated library spectra. Cross correlation is a measure of the similarity of two time varying functions. In signal processing, cross-correlation is a method used to estimate the correlation between two signals using a dot product after a time lag has been applied to one of the signals. The cross correlation function C_{ij} for the sampling interval Δt and relative displacement $n\Delta t$ between two signals s_i and s_j is estimated as shown in the following equation

$$C_{ij}(n\Delta t) = \frac{1}{T}\sum_{t=0}^{T} s_i(t)s_j(t) \qquad n = 0,1,2,....\frac{T}{\Delta t} \tag{8}$$

Autocorrelation is similar to cross correlation and is the signal being cross correlated with itself. Autocorrelation and cross correlation were performed by normalizing the spectra to unit length. Two different algorithms were used to perform cross correlation library searching. The first algorithm identifies the IR spectrum in the truncated library that is most similar to the unknown using three different modes of comparison:

1. Autocorrelated spectrum of unknown is compared to each cross-correlated unknown and library spectrum
2. Each autocorrelated library spectrum is compared to each cross-correlated unknown and library spectrum
3. Autocorrelated spectrum of unknown is compared to each autocorrelated library spectrum

Each comparison was made using a range of window sizes centered at the midpoint of the cross-correlated data interval (which corresponds to the cross correlation between two signals with zero lag) and increased in steps of 10 points or 100 points to include the entire cross correlated spectrum (see Figure 4). Because of the symmetry associated with cross correlation, the comparisons were made from only one side of the center burst. The Euclidian distance was used to evaluate the similarity index (see Equation 9) 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 and autocorrelated 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 9 was used instead of the hit quality index (HQI) used by commercial search algorithms, e.g., OMNIC, as it proved to be more informative for these spectra.

$$s_{ij} = 1 - \frac{d_{ij}}{d_{\max}} \tag{9}$$

Library spectra were arranged in descending order of similarity for each comparison and window size. The five most similar library spectra were then chosen from each comparison made for each window size, with sample identities preserved. After every window was analyzed, a histogram depicting the frequency of occurrence for the most similar spectra was generated. The frequency of occurrence for each spectra was then weighted by its average similarity index across every interval and every window. The 5 library spectra with the highest frequency of occurrence after weighting were selected as potential matches.

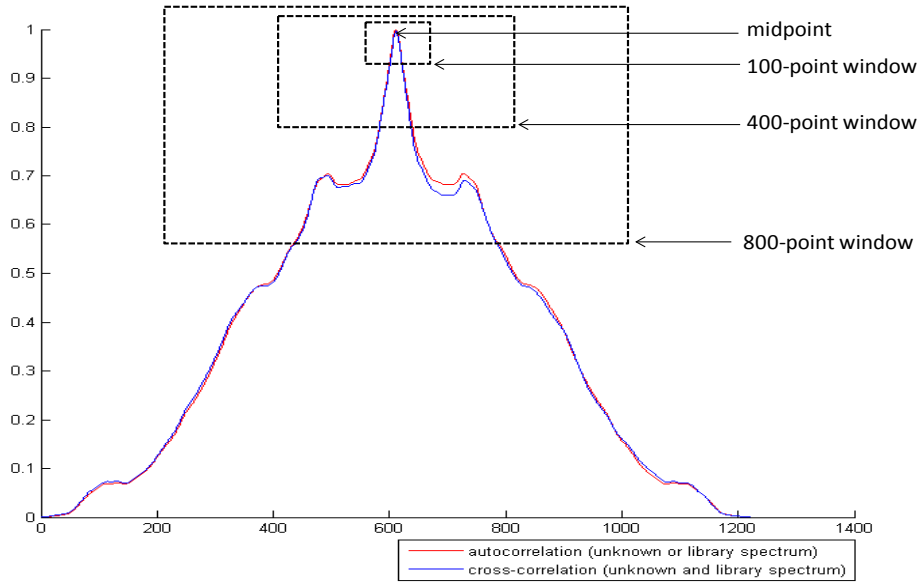

Figure 4. Comparison of the cross correlated unknown and PDQ library spectrum with the autocorrelated spectrum of the unknown or PDQ library spectrum.

All spectra were divided into three regions: 3675 to 2856 cm^{-1} (region after absorption by the diamond cell), 1891 to 668 cm^{-1} (fingerprint region and carbonyl band), and 1650 to 668 cm^{-1} (fingerprint region). The overlap between the second and third spectral regions, which effectively increases the importance of the fingerprint region in the spectral matching, proved superior to using a disjoint set of intervals (e.g., 1650 cm^{-1} to 668 cm^{-1} with double weighting, 1891 cm^{-1} to 1650 cm^{-1}, and 3675 cm^{-1} to 2856 cm^{-1}) as it enforces the relative scale of the peaks and captures the broader trends in the spectral data.

Spectral library matching using the cross correlations search algorithm was implemented using two schemes, both based upon transmittance spectra. The first scheme uses auto-cross correlation to identify spectra within the library that best matches the blind. At each window for each interval, the first method ranks all spectra in order of their similarity index value. The identity of the top five spectra (sample identification number, make, model, and line) is preserved for each window. After each window has been processed, the output of the algorithm is a set of five spectra with the most matches in the top five in each of the prior steps.

The second scheme uses auto-cross correlation to provide a probability index as to the model and line of the blind. At each window for each interval, the spectra in the library are ranked by their

similarity index with regard to the blind, but only the label (i.e., model and line of the automotive vehicle) of each of the top five spectra is preserved. After each window has been processed, the number of the hits for a specific model and line is divided by the sum of the number of comparisons performed. The output is a set of percentages that represents the likelihood that a particular model or line is a match to the blind. While the first scheme identifies the library spectrum most similar to the blind, the second scheme is assessing the similarity of the library spectra for the blind. In other words, the library is being matched to the blind which is the opposite of the first approach which is matching the blind to the library. The performance of the prototype pattern recognition driven library searching system (search prefilters and cross-correlation library search algorithm) for both schemes was compared to OMNIC, a commercial library searching algorithm used by Thermo Nicolet FTIR spectrometers.

III. RESULTS

Search prefilters were developed from 1182 automotive paint systems that spanned 3 automobile manufacturers (GM, Chrysler, and Ford) and 54 assembly plants (in North America) within a limited production year range (2000-2006). Because of the large number of classes (i.e., assembly plants) involved, a hierarchical classification scheme was employed. A search prefilter was developed to differentiate paint samples by automobile manufacturer. For each automobile manufacturer, search prefilters were developed to identify the assembly plant of the vehicle from the manufacturer's paint system conveyed by the sample. First, the assembly plants were divided into groups of plants based upon cluster analysis of the fingerprint region of the clear coat paint layer. Second, each plant group is divided into its constituent assembly plants using both the clear coat and the two undercoat paint layers. The search prefilter system is intended to categorize each unknown paint system by identifying successively smaller sets of vehicles to which the unknown is assigned. In the final step, library searching of an unknown is performed using IR spectra of vehicles assembled in the manufacturing plant identified by the search prefilters. A block diagram of the vehicle classification process used in the prototype pattern recognition driven library search system for the PDQ database is summarized in Figure 5.

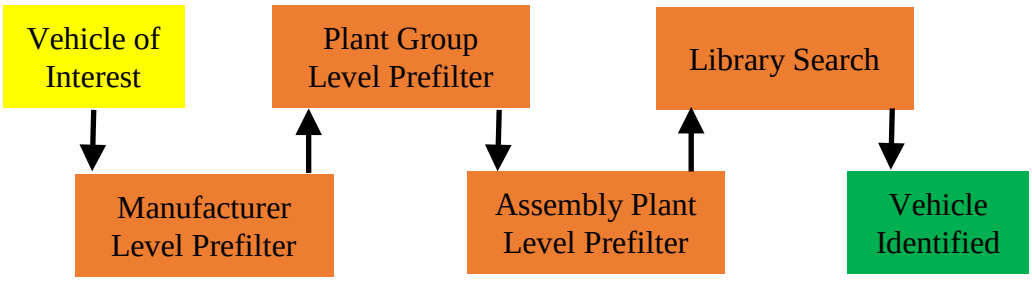

Figure 5. Block diagram of the vehicle classification process used in the prototype pattern recognition driven library search system for the PDQ database.

Development of Search Prefilters for Manufacturer

The initial focus of this study was to develop a search prefilter to classify the IR spectra of an unknown paint sample by manufacturer. To differentiate automotive paint samples by manufacturer, the 1306 paint systems investigated were divided into a training set of 1182 samples and a validation set of 124 samples. The validation set samples were chosen by random

lot. The training set of 1182 automotive paint systems was divided into 3 classes by automobile manufacturer (see Table 1).

Table 1. Training Set and Validation Set for Manufacturer Search Prefilter

Manufacturer	Training Set	Validation Set
GM (2000-2006)	429	44
Chrysler (2000-2006)	379	42
Ford (2000-2006)	374	38

All IR spectra were pre-processed for pattern recognition analysis as follows. After retaining only the fingerprint region in each layer, the spectra were smoothed using a Savitzky-Golay filter (fourth order polynomial, 17 point window). The smoothed IR spectra were vector normalized and then wavelet transformed using the Symlet 6 mother wavelet at the 8th level of decomposition (8Sym6). All wavelet coefficients for levels of decomposition less than the specified level were retained, such that the final result for each sample-paint layer combination is a row vector of wavelet coefficients, [A1 D1 A2 D2 … A8 D8], where A1 represents the set of first order approximation coefficients for the sample, D1 is the corresponding set of first order detail coefficients for the sample, A2 and D2 similarly represent the second order approximation and detail wavelet coefficients and so forth. Because the search prefilter utilizes both the clear coat and undercoat layers, the final step involves horizontally concatenating the wavelet coefficients from each layer into a single vector in the order of clear coat, surfacer, and primer.

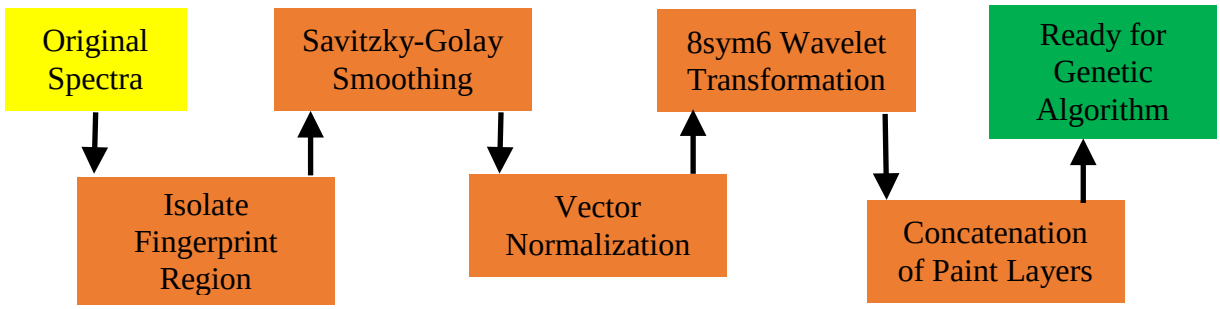

Figure 6. Block diagram of the spectral preprocessing procedure used to develop search prefilters.

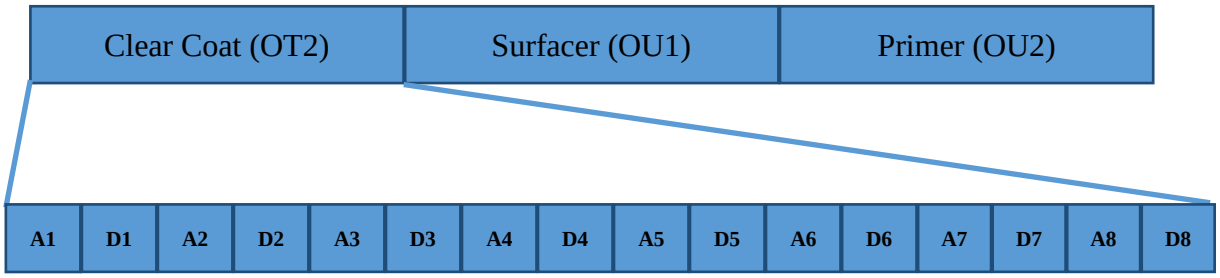

Figure 7. Diagram of the wavelet coefficient concatenation scheme used with the clear coat paint layer magnified to show how individual sets of approximation and detail coefficients are concatenated within each layer. Number of coefficients per block decreases with increasing level of decomposition (e.g., A1 or first order approximations and D1 or first order details contain an equal number of coefficients whereas A2 or second order approximations contains fewer coefficients than A1).

Figure 8 shows a PC plot of the two largest principal components of the 1182 training set samples and the 3450 wavelet coefficients comprising the training set data. Each paint sample is represented as a point in the PC map of the data (1 = GM, 2 = Chrysler, and 3 = Ford). The overlap of the wavelet transformed fused spectra of the fingerprint region for Chrysler and Ford is evident.

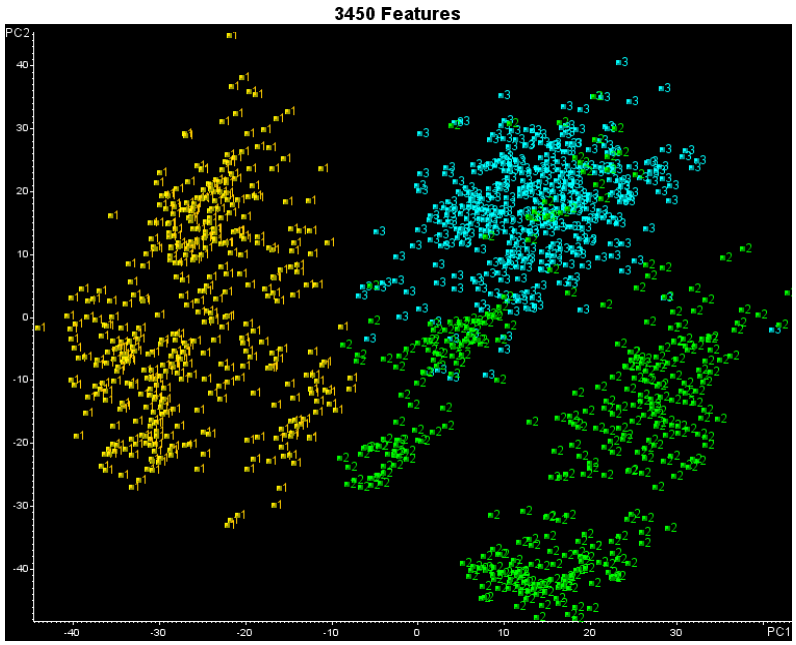

Figure 8. PC plot of the two largest principal components of the 1182 paint samples and the 3450 wavelet coefficients comprising the training set data (1 = GM, 2 = Chrysler, and 3 = Ford)

The next step was feature selection. A genetic algorithm for pattern recognition was used in the study to identify wavelet coefficients characteristic of automobile manufacturer. The pattern recognition GA identified wavelet coefficients by sampling key feature subsets, scoring their PC plots and tracking those paint samples or 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 39 wavelet coefficients (after deletion of 5 variables with low loadings) whose PC plot showed clustering of the data on the basis of automotive manufacturer (see Figure 9).

To assess the predictive ability of the 39 wavelet coefficients identified by the pattern recognition GA, a validation set of 124 paint samples was used. Figure 10 shows the validation set samples projected onto the PC plot of the data defined by the 1182 wavelet transformed fused IR spectra and the 39 wavelet coefficients identified by the pattern recognition GA. Each validation set sample lies in a region of the PC map with paint samples from the same automotive manufacturer. This result suggests that information about automotive manufacturer can be extracted from the wavelet transformed fused IR spectrum of an unknown paint sample.

1-NN [27] was also used to classify the 1182 wavelet transformed fused spectra in the training set. A classification rule developed from the 39 wavelet coefficients identified by the pattern recognition GA using 1-NN achieved a classification success rate of 100% for the training set.

To further test the predictive ability of this classifier, the validation set of 124 paint samples was employed. Again a classification success rate of 100% was achieved for the fused validation set spectra. The results from 1-NN are consistent with those results obtained using PCA.

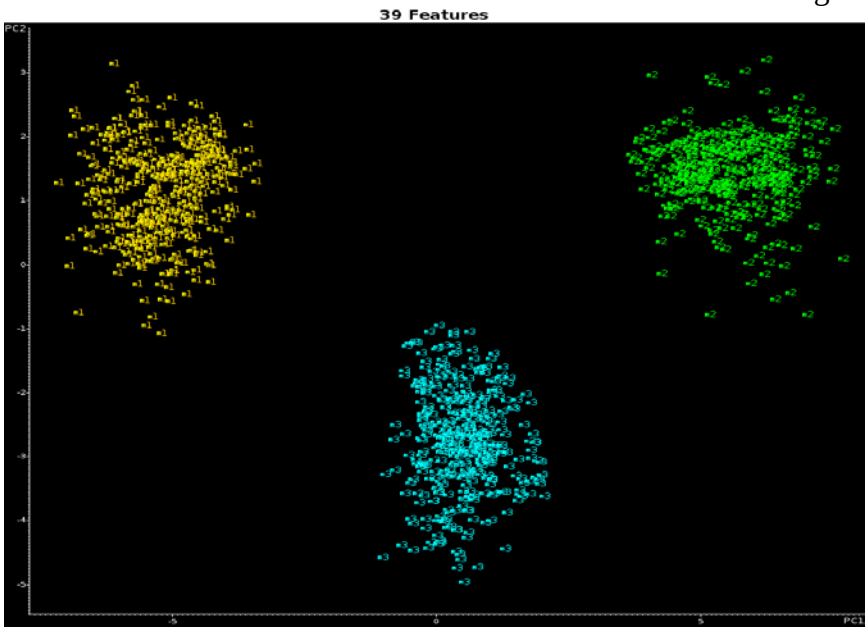

Figure 9. PC plot of the two largest principal components of the 1182 training set samples and the 39 wavelet coefficients identified by the pattern recognition GA (1 = GM, 2 = Chrysler, and 3 = Ford).

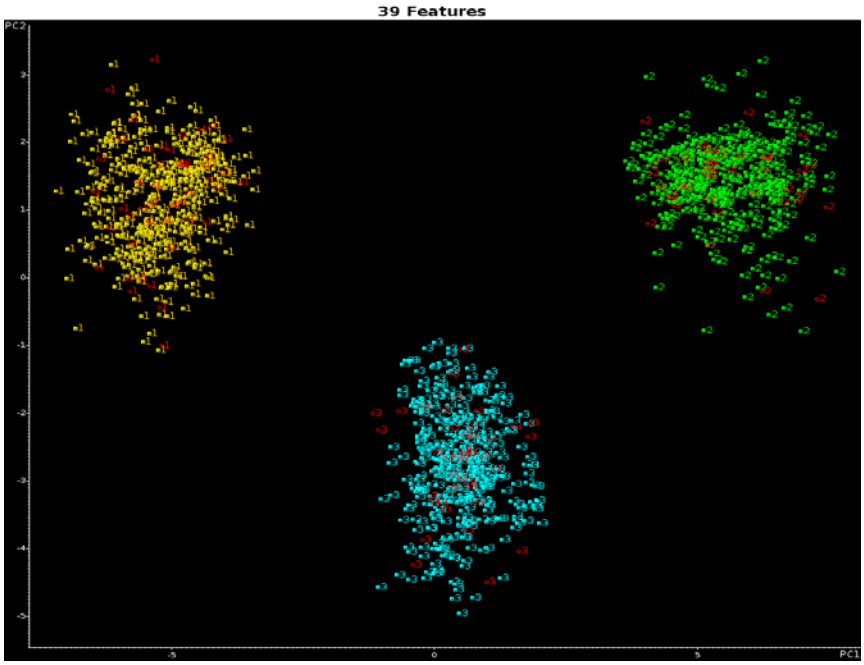

Figure 10. Validation set samples (in red) projected onto the PC plot of the data defined by the 1182 wavelet transformed fused IR spectra of the training set (green, yellow, cyan) and the 39 wavelet coefficients identified by the pattern recognition GA. (1 = GM, 2 = Chrysler, and 3 = Ford).

Development of Search Prefilters for Chrysler

The first step in the development of the Chrysler search prefilters was to differentiate Chrysler paint systems by plant group. To determine the composition of each plant group, Chrysler assembly plants (see Table 2) whose clear coat paint spectra exhibited a doublet for the carbonyl band (acrylic melamine styrene polyurethane) as opposed to a singlet (acrylic melamine styrene**)** were flagged. The two assembly plants (Jefferson North and Newark) whose clear coat paint spectra exhibited a doublet for the carbonyl were placed in Plant Group 13, whereas the assembly plants whose clear coat paint spectra exhibited a singlet for the carbonyl band were assigned to the other plant groups. Using only the clear coat paint spectra, each of the ten remaining assembly plants (see Table 2) was analyzed by principal component analysis [28] to assess its class structure. In four of the ten assembly plants (Bramalea/Brampton, Dodge Main, St. Louis, and Toledo), the PC plot of the clear coat paint spectra exhibited two distinct sample clusters (see Figures 11 through 14). For the Bramalea/Brampton assembly plant, clustering occurred on the basis of model: Dodge Charger and some Chrysler 300 lines versus Chrysler Concorde, Chrysler LHS, Dodge Intrepid, Dodge Magnum, and other Chrysler 300 lines, whereas for Dodge Main, clustering occurred on the basis of the production year of the vehicle: 2000-2002 versus 2003-2006. For the St. Louis assembly plant, clustering occurred on the basis of model and line: Dodge Caravan and Chrysler Town and Country versus Dodge Ram, whereas for Toledo, clustering was correlated to a specific vehicle: Jeep Liberty versus the other models and lines assembled at the plant. Because the average clear coat paint spectrum of each cluster was noticeably different when compared visually, the four assembly plants were further divided into subplants on the basis of the observed sample clustering.

Table 2. Chrysler Assembly Plants

PLANT	PID# (data label)	DIVIDED BETWEEN PLANT GROUPS	Plant GROUP
Belvidere (BEL)	1000	NO	11
Bloomington (BLO)	1001	NO	12
Bramalea/Brampton (BRA/BRP)	1002	YES	11, 12
Dodge Main (DOD)	1003	YES	11, 12
Jefferson North (JFN)	1004	NO	13
Newark (NEW)	1006	NO	13
Saltillo (SAL)	1007	NO	11
Sterling Heights (STH)	1008	NO	12
Saint Louis (STL)	1009	YES	11, 12
Toledo (TOL)	1010	YES	11, 12
Toluca (TOU)	1011	NO	11
Windsor (WIN)	1012	NO	12

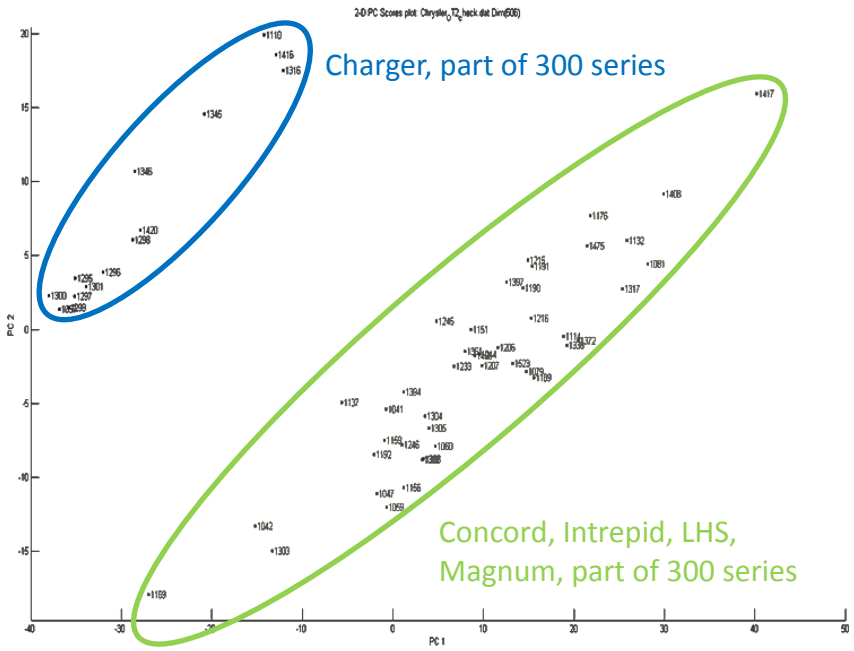

Figure 11. Plot of the two largest principal components of the clear coat paint spectra from the Bramalea/Brampton plant. Two distinct sample clusters are evident in the plot.

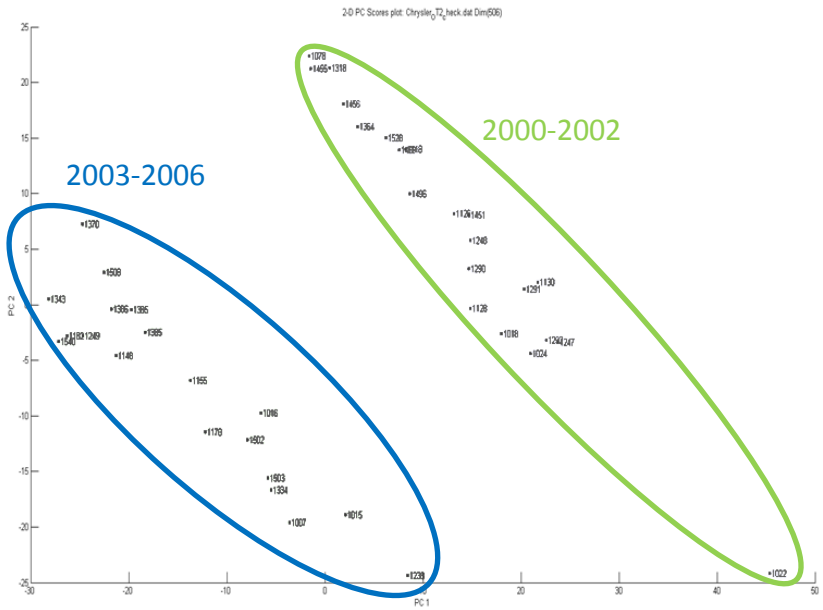

Figure 12. Plot of the two largest principal components of the clear coat paint spectra from the Dodge Main assembly plant. Two distinct sample clusters are evident in the plot.

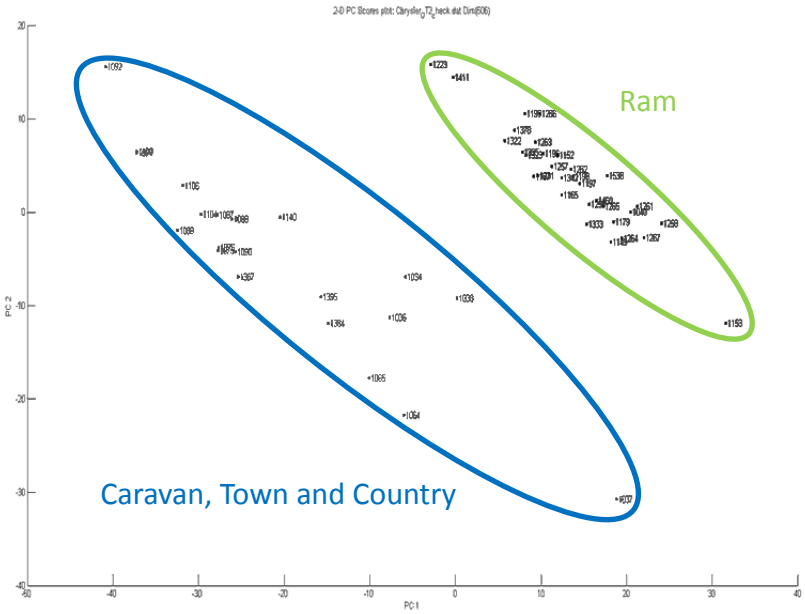

Figure 13. Plot of the two largest principal components of the clear coat paint spectra from the St Louis assembly plant. Two distinct sample clusters are evident in the plot.

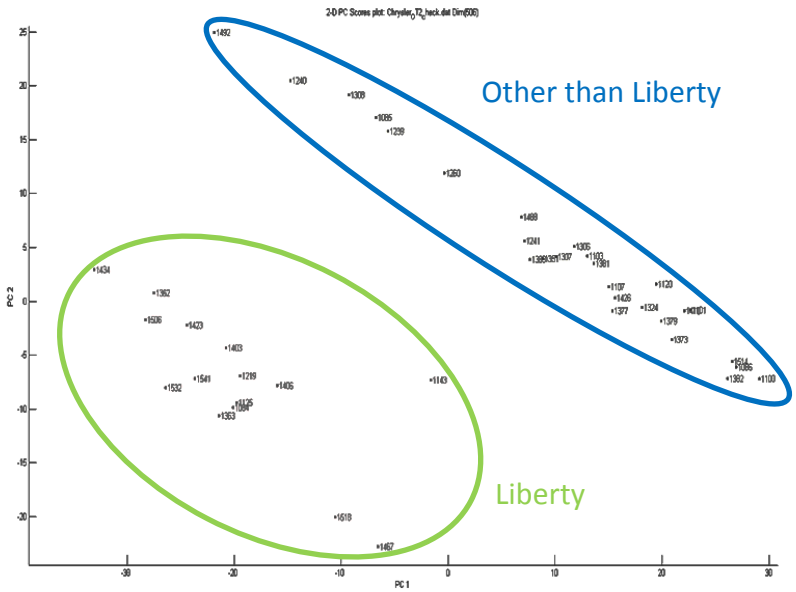

Figure 14. Plot of the two largest principal components of the clear coat paint spectra from the Toledo assembly plant. Two distinct sample clusters are evident in the plot.

To assign the remaining assembly plants and subplants to specific plant groups, the average IR spectrum (clear coat layer) of each assembly plant or subplant was computed. Principal component analysis and hierarchical clustering [29] were performed on the average spectra. Figures 15 and 16 summarize the results of the clustering study. Plant Group 11 consists of

Belvidere, Bramalea/Brampton (subplant), Dodge Main (subplant), Saltillo, St. Louis (subplant), Toledo (subplant), and Toluca assembly plants, whereas Plant Group 12 is comprised of Bloomington, Bramalea/Brampton (subplant), Dodge Main (subplant), Sterling Heights, St. Louis (subplant), Toledo (subplant), and Windsor assembly plants.

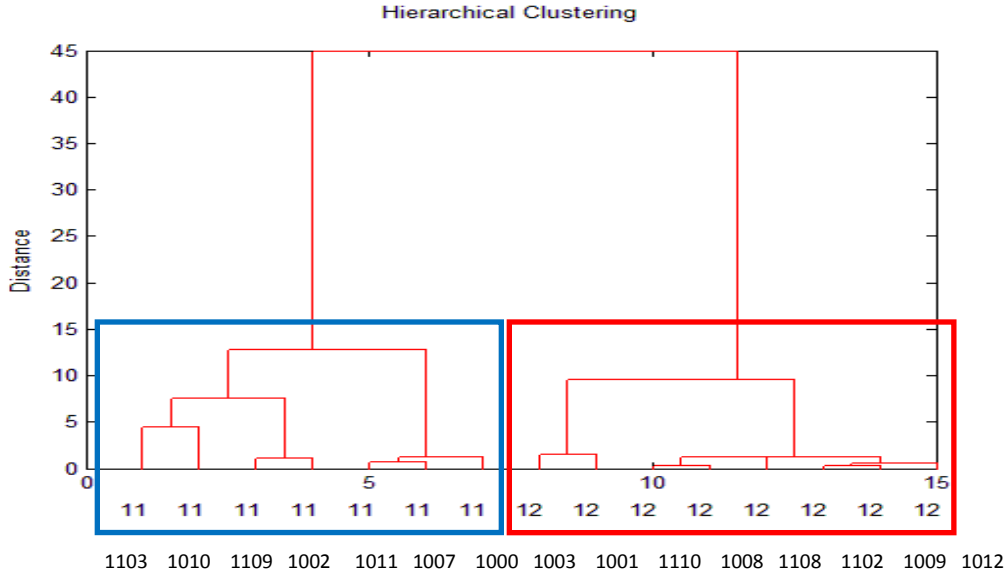

Figure 15. Hierarchical cluster analysis (Wards method) of the average IR spectrum (clear coats) of each assembly plant or subplant. 1000 = Belvidere, 1001 = Bloomington, 1002 = Bramalea/Brampton subplant, 1003 = Dodge Main subplant, 1007 = Saltillo, 1008 = Sterling Heights, 1009 = St. Louis subplant, 1010 = Toledo, 1011 = Toluca, 1012 = Windsor, 1102 = Bramalea/Brampton subplant, 1103 = Dodge Main subplant, 1109 = St. Louis subplant, and 1110 = Toledo subplant.

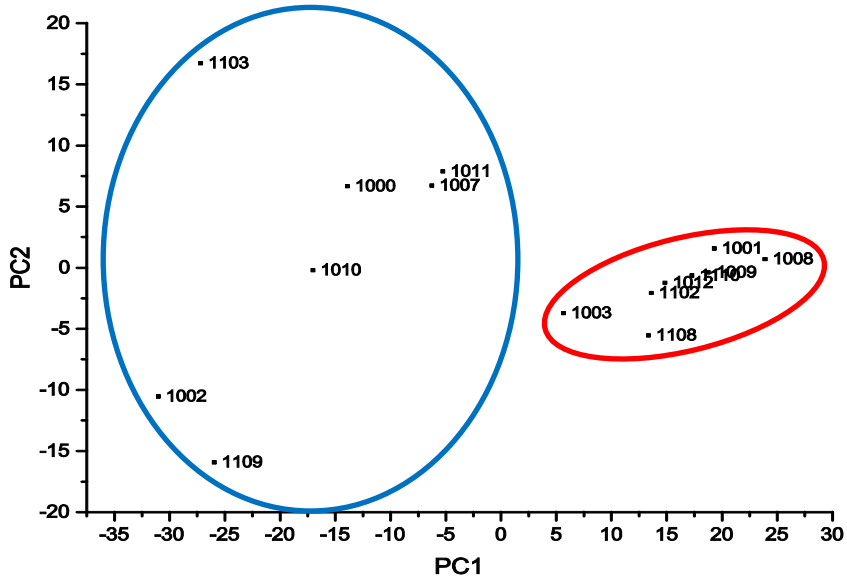

Figure 16. Principal component analysis of the average IR clear coat paint spectrum of each assembly plant or subplant. 1000 = Belvidere, 1001 = Bloomington, 1002 = Bramalea/Brampton subplant, 1003 = Dodge Main subplant, 1007 = Saltillo, 1008 = Sterling Heights, 1009 = St. Louis subplant, 1010 = Toledo, 1011 = Toluca, 1012 = Windsor, 1102 = Bramalea/Brampton subplant, 1103 = Dodge Main subplant, 1109 = St. Louis subplant, and 1110 = Toledo subplant.

Having ascertained the membership of each plant group for both the acrylic melamine styrene and acrylic melamine styrene polyurethane clear coats, the next step was discriminant development. Figure 17 shows a PC plot of the two largest principal components of the 379 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group (see Table 3). All IR spectra in the training set were vector normalized prior to the application of the wavelet transform (8Sym6), and all wavelet coefficients were autoscaled prior to principal component analysis. Each clear coat is represented as a point in the PC plot of the data. (11 = Plant Group 11 (acrylic melamine styrene), 12 = Plant Group 12 (acrylic melamine styrene), and 13 = Plant Group 13 (acrylic melamine styrene polyurethane)). Plant Group 12 is well separated from Plant Groups 11 and 13, whereas Plant Groups 11 and 13 overlap in the plot.

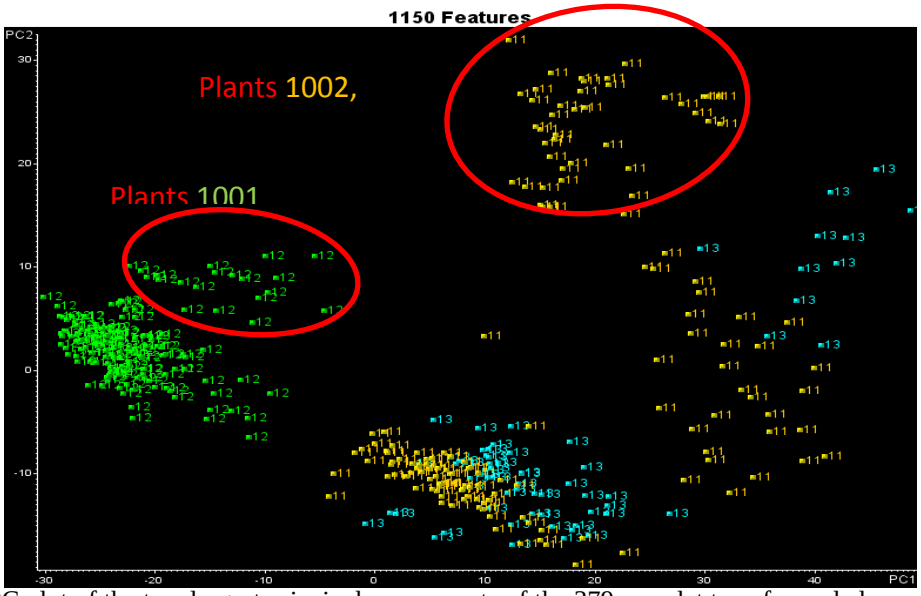

Figure 17. PC plot of the two largest principal components of the 379 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group. Each clear coat is represented as a point in the PC plot of the data. (11 = Plant Group 11, 12 = Plant Group 12, and 13 = Plant Group 13).

Table 3. Training Set and Validation Set for Chrysler Plant Groups

Group	Training	Validation
11	156	19
12	157	17
13	66	6

Feature selection was performed to identify wavelet coefficients characteristic of the profile of each plant group. The pattern recognition GA identified informative wavelet coefficients by sampling key feature subsets, scoring their PC plots, and tracking those plant groups/and or IR spectra 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 9 wavelet coefficients whose PC plot showed clustering (see Figure 18) of the IR clear coat paint spectra on the basis of plant group.

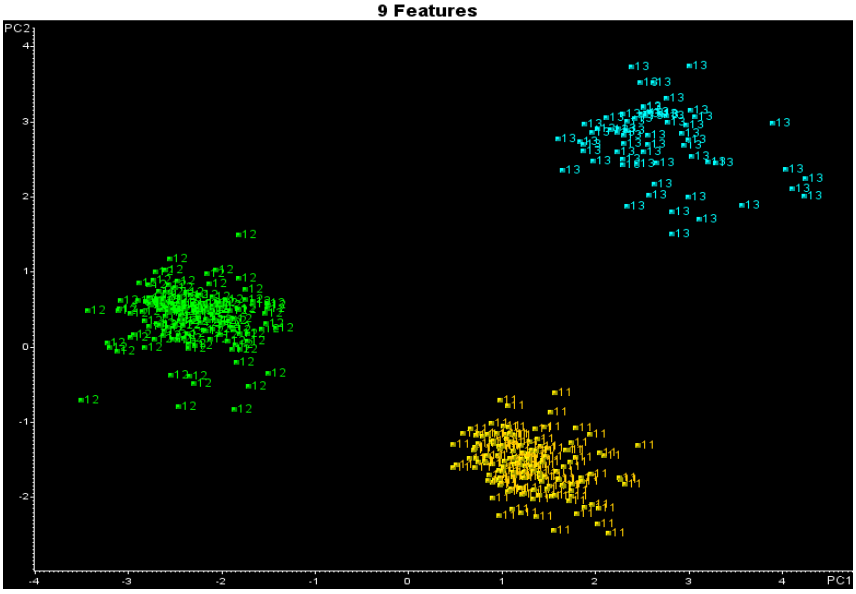

Figure 18. PC plot of the two largest principal components of the 379 training set samples and 9 wavelet coefficients identified by the pattern recognition GA (11 = Plant Group 11, 12 = Plant Group 12, 13 = Plant Group 13).

To assess the predictive ability of the 9 wavelet coefficients identified by the pattern recognition GA, a validation set of 42 IR spectra was used. IR spectra from the validation set were projected directly onto the PC plot developed from the 379 IR spectra and the training set and the 9 wavelet coefficients identified by the pattern recognition GA. Figure 19 shows the projection of the validation set samples onto the PC map of the training set data. All validation set samples are located in a region of the map with clear coats that have the same class label.

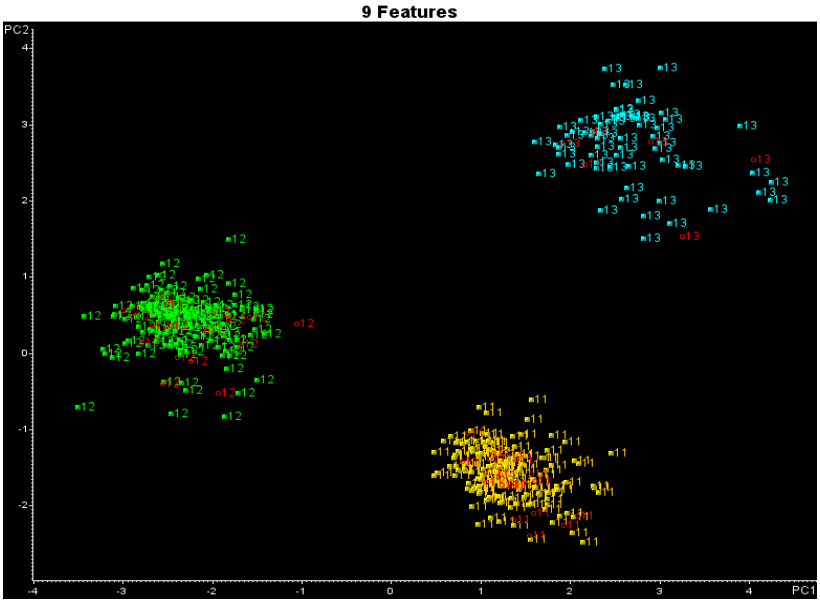

Figure 19. Validation set samples (in red) projected onto the PC plot of the data defined by the 379 wavelet transformed clear coat IR spectra of the training set (green, yellow, cyan) and the 9 wavelet coefficients identified by the pattern recognition GA. (11 = Plant Group 1, 12 = Plant Group 2, and 13 = Plant Group 13).

For each plant group, a search prefilter was developed to discriminate automotive paint samples by assembly plant using the fingerprint region of the clear coat and the two undercoat layers of the automotive paint samples. After retaining only the fingerprint region in each layer, all IR spectra were vector normalized and then wavelet transformed using the Symlet 6 mother wavelet at the 8th level of decomposition. Wavelet coefficients from each layer were horizontally concatenated into a single data vector in the order of clear coat, surfacer, and primer. The pattern recognition GA identified the components of this data vector (i.e., specific coefficients in each paint layer) correlated to the assembly plant of the vehicle from which the paint sample was obtained.

Table 4 lists the assembly plants or subplants comprising Plant Group 11, which consists of one assembly plant (Belvidere), four subplants (Bramalea/Brampton, Dodge Main, St. Louis, Toledo) and a plant subgroup consisting of two assembly plants (Saltillo and Toluca). Saltillo and Toluca were combined because the spectra of the clear coat, surfacer, and primer layers were superimposable for these two assembly plants.

Table 4. Assembly and Subplants Comprising Plant Group 11

Plant	Training	Validation
711 (Saltillo + Toluca plants)	47	10
1000 (Belvidere)	33	3
1002 (subplant of Bramalea/Brampton)	13	1
1010 (subplant of Toledo)	14	1
1103 (subplant of Dodge Main)	19	2
1109 (subplant of St Louis)	30	2

Figure 20 shows a plot of the two largest principal components of the 156 samples comprising Plant Group 11 and the 16 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant, subplant, and plant subgroup is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 11 onto the PC map showed that each projected validation set sample is located in a region of the PC plot with samples from the same assembly plant or subplant.

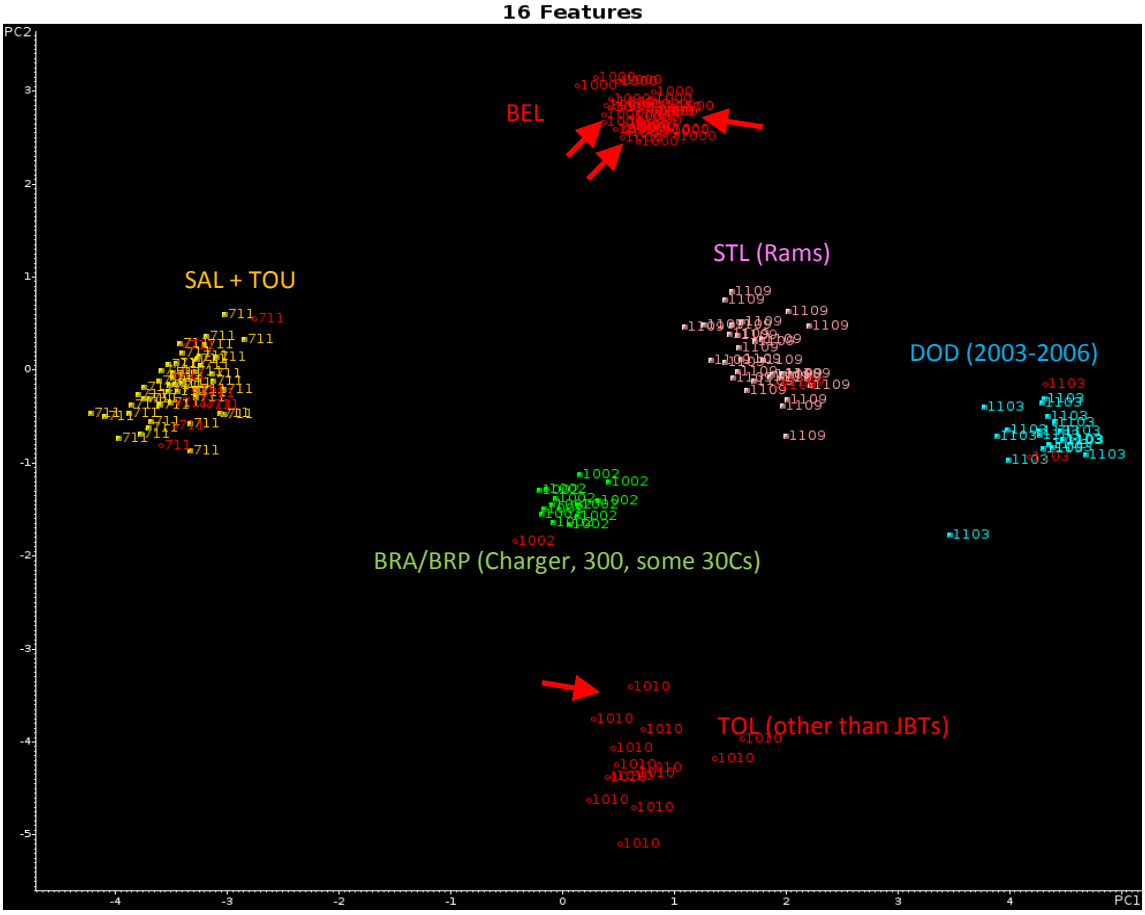

Figure 20. Validation set samples (in red or indicated by an arrow) projected onto the PC plot of the data defined by the 156 paint samples comprising Plant Group 11 (training set) and the 16 wavelet coefficients identified by the pattern recognition GA. (711 = plant subgroup containing Saltillo and Toluca, 1000 = Belvidere, 1002 = subplant of Bramalea/Brampton, 1010 = subplant of Toledo, 1103 = subplant of Dodge Main, 1109 = subplant of St Louis).

Table 5 lists the assembly plants or subplants comprising Plant Group 12, which consists of one assembly plant (Bloomington), three subplants (Bramalea/Brampton, Dodge Main, Sterling Heights), and two plant subgroups. During the course of the analysis, it was necessary to merge different plants and subplants into a single class referred to as a plant subgroup due to the similarity of their spectra. One plant subgroup was comprised of the Windsor assembly plant and the St. Louis subplant, whereas the other was comprised of two subplants (one from Sterling Heights and the other from Toledo). Although the principal component analysis plot of clear coat IR spectra from the Stirling Heights plant did not exhibit sample clustering, the corresponding principal component plot for the wavelet transformed concatenated IR spectral data (see Figure 21) showed clustering correlated to production year. Furthermore, the average spectra of the clear coat, surfacer, and primer layers for the 2002-2006 Stirling Heights vehicles were superimposable when compared to the average spectra of the clear coat, surfacer, and primer layers for the vehicles from the Toledo subplant assigned to this plant group. For this reason, this subplant was combined with Toledo to form the corresponding plant subgroup. Principal component analysis plots of fused IR spectral data from other assembly plants or subplants comprising Plant Group 12 did not exhibit sample clustering.

Table 5. Assembly and Subplants Comprising Plant Group 12

Plant	Training	Validation
912 (part of St. Louis + Windsor)	47	10
1001 (Bloomington)	33	3
1003 (part of Dodge Main)	13	1
1102 (part of Bramalea/Brampton)	14	1
1108 (part of Sterling Heights)	19	2
1810 (part of Sterling Heights + part of Toledo)	30	2

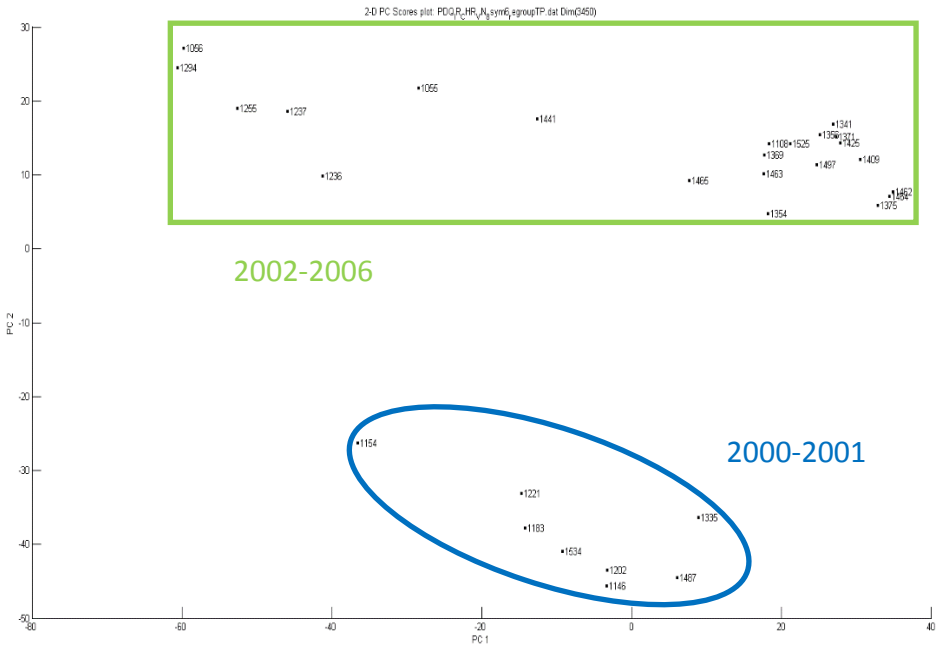

Figure 21. Plot of the two largest principal components of the wavelet transformed concatenated IR spectra from the Sterling Heights assembly plant. Two distinct clusters are evident in the plot.

Figure 22 shows a plot of the two largest principal components of the 157 samples comprising Plant Group 12 and the 22 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant, subplant, and plant subgroup forms a distinct and separated cluster in the PC plot of the training set data. Projecting the validation set samples assigned to Plant Group 12 onto the PC plot shows that each projected validation set sample is located in a region of the PC map with samples from the same assembly plant, subplant, or plant subgroup.

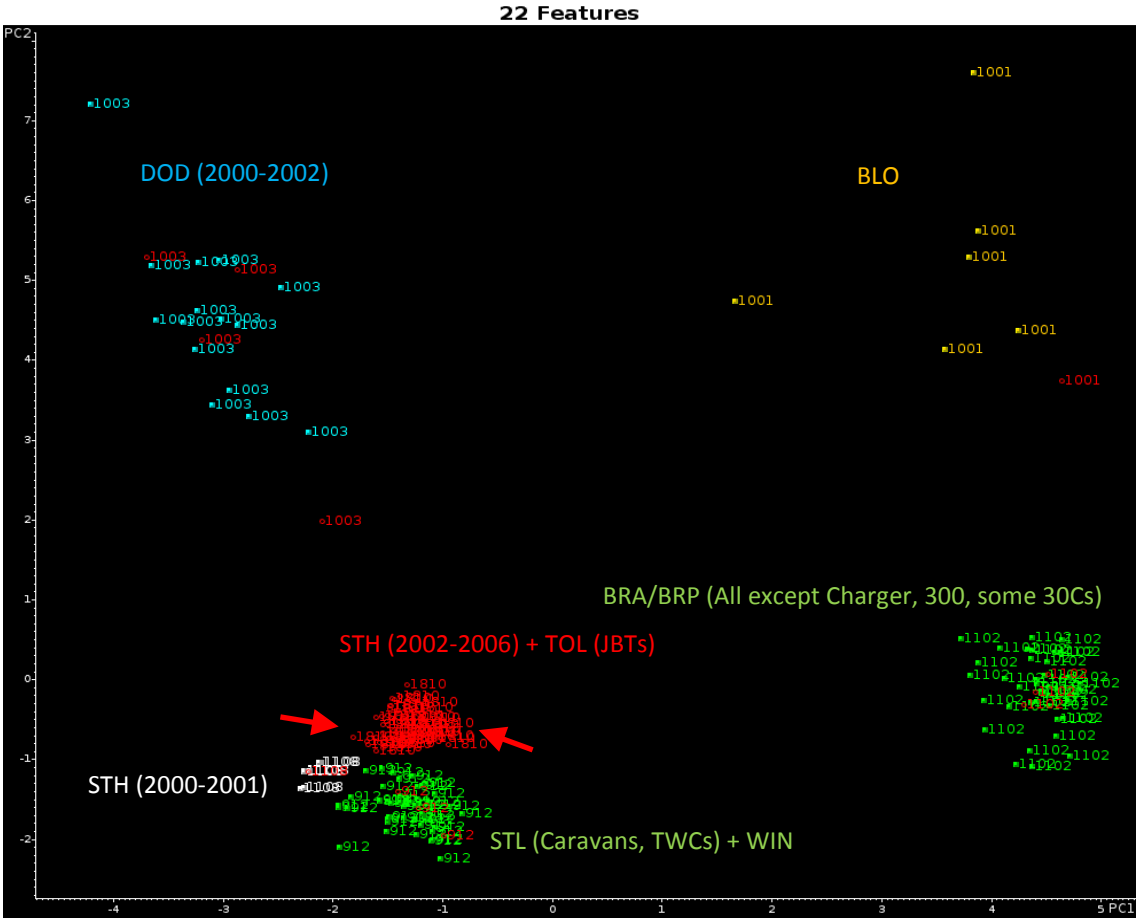

Figure 22. Validation set samples (in red or indicated by an arrow) projected onto the PC plot of the data defined by the 157 paint samples comprising Plant Group 12 (training set) and the 22 wavelet coefficients identified by the pattern recognition GA. (912 = plant subgroup containing Windsor and subplant of St. Louis, 1001 = Bloomington, 1003 = subplant of Dodge Main, 1102 = subplant of Bramalea/Brampton, 1108 = subplant of Sterling Heights, 1810 = subplant of Sterling Heights and subplant of St. Louis).

Plant Group 13 consisted of two assembly plants: Jefferson North and Newark. The pattern recognition GA was not able to identify a set of coefficients from the wavelet transformed concatenated IR spectra that could differentiate Jefferson North from Newark. To understand the reason for the lack of success, principal component analysis was performed on both the Newark and Jefferson North assembly plants. Clustering correlated to the production year of the vehicle was observed for Newark. For this reason, Newark was divided into two subplants. The Newark sample that was not a member of either cluster was tagged as an outlier and deleted from the analysis after comparing the spectra of each paint layer from this sample with the average spectra of each paint layer from each sample cluster identified in the PC plot of the wavelet transformed data. Table 6 describes the assembly plant and subplants comprising Plant Group 13.

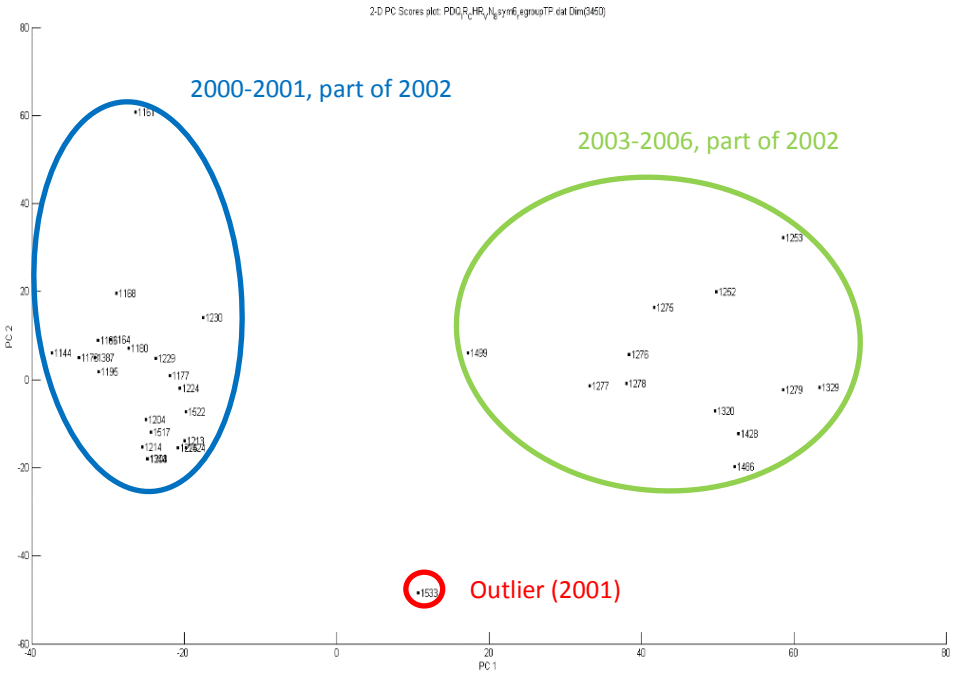

Figure 23. Plot of the two largest principal components of the wavelet transformed concatenated IR spectra from the Newark assembly plant. Two distinct sample clusters and one outlier are present in the PC plot.

Table 6. Assembly and Subplants Comprising Plant Group 13

Plant	Training	Validation
1004 (Jefferson North)	34	3
1006 (subplant of Newark)	20	2
1106 (subplant of Newark)	11	1

Figure 24 shows a plot of the two largest principal components of the 65 samples comprising Plant Group 13 and the 33 wavelet coefficients identified by the pattern recognition GA for the training set. The assembly plant and two subplants are well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 13 onto the PC plot showed that each projected validation set sample is located in a region of the PC plot with samples from the same assembly plant or subplant.

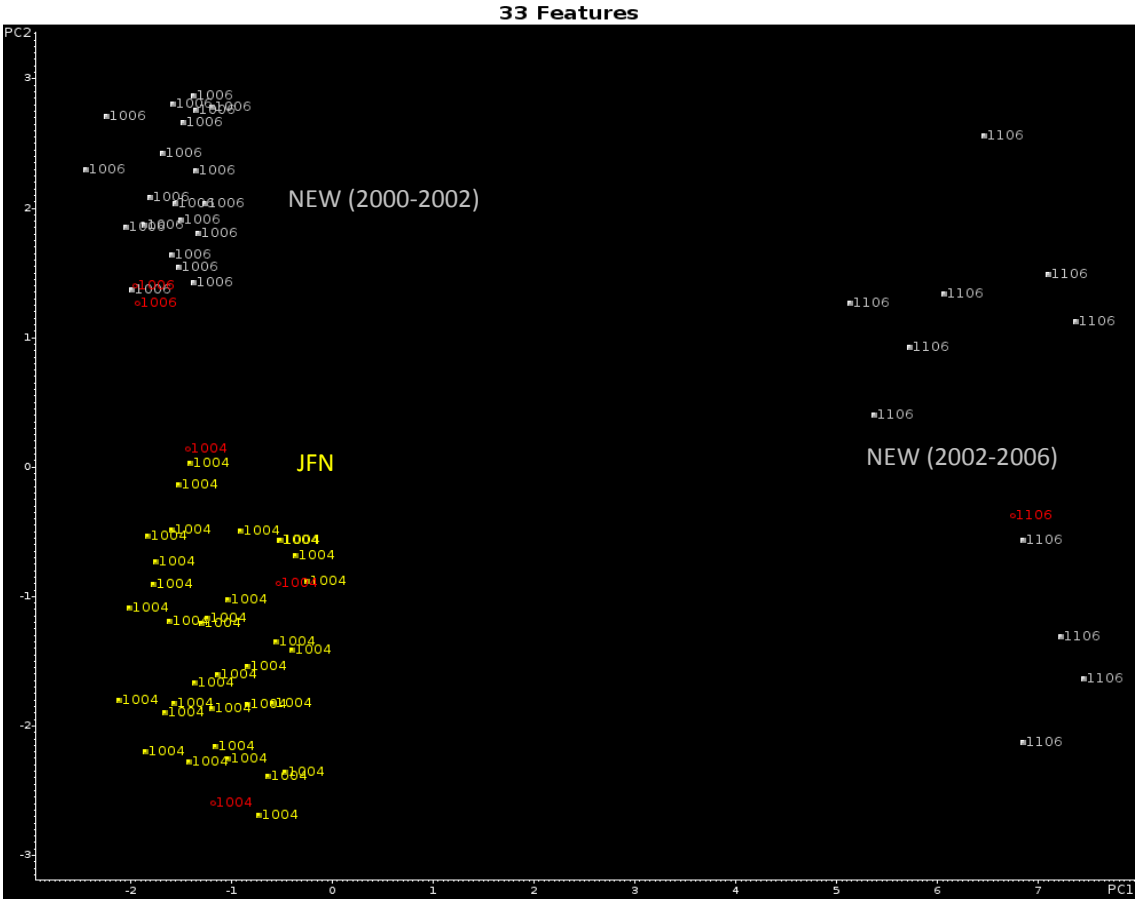

Figure 24. Validation set samples (in red) projected onto the PC plot of the data defined by the 65 paint samples comprising Plant Group 13 (training set) and the 33 wavelet coefficients identified by the pattern recognition GA. (1004 = Jefferson North, 1006 = Newark (2000-2002), and 1106 = Newark (2002-2006).

The PDQ spectral library consisted of the manufacturer's paint system for 1182 automotive Chrysler, General Motors, and Ford automotive vehicles within a limited production year range (2000-2006). To extract information about the model from the IR spectrum, library searching of each validation set sample was performed using the IR spectra of the automotive vehicles assembled in the assembly plant(s) identified by the search prefilters. During this phase of the study, each spectrum in the validation set was compared to library spectra using the regions from 3675 cm^{-1} to 2856 cm^{-1} and from 1891 cm^{-1} to 668 cm^{-1}. Within the region from 1891 cm^{-1} to 668 cm^{-1}, the spectral region from 1650 cm^{-1} to 668 cm^{-1}, corresponding to the fingerprint region, was doubled in relative weight to emphasize its importance. The region of the spectrum corresponding to absorption by the diamond transmission cell was omitted. Within each interval, up to 75 windows were employed from the center burst. The library search results for the validation set samples are summarized in Table 7 for 41 of the 42 samples. One validation set sample was deleted from this tabulation because the corresponding model and line of this paint sample was not present in the library. For Scheme 1, the remaining results were then weighted based on their average similarity index across all windows and spectral intervals. The 5 top matches were chosen from each comparison and a histogram depicting the frequency of occurrence for spectra most similar to the validation set sample was computed to find the closest

matching sample. Results from the cross correlation library searching method were compared with the top 5 matches identified by OMNIC (Thermo Nicolet), which accessed the entire PDQ library (1182 Chrysler, Plymouth, and Ford samples) for spectral matching. Library search results for both the cross correlation searching algorithm and OMNIC are summarized in Table 7. The prototype pattern recognition system outperformed OMNIC, which is a commercially available IR search algorithm considered by many workers in the field as the industry standard. To search all three paint layers simultaneously, the spectra (in absorbance mode) of the clear coat and undercoat paint layers for each sample were added to yield a single spectrum which was then converted to the transmittance mode for analysis by the cross correlation library search algorithm.

Table 7. Library Search Results for Chrysler

Layer	[1]OMNIC Searching	Prototype System (Scheme 1)	Prototype System (Scheme 2)
Clear Coat	28	40	38
Surfacer	31	39	35
Primer	26	37	31
[2]Fused (Addition)	14	38	37

[1]OMNIC searching was done using its built-in library searching routine. The diamond region, 2856 cm^{-1} to 1891 cm^{-1} was excluded from the search. The results of OMNIC were considered correct if there was a match within the top 5 hits found by OMNIC's search algorithm as ranked by the hit quality index.

[2]Fusion of the layers was achieved by taking spectra (in absorbance mode) of the clear coat and undercoat paint layers and directly adding them to yield a single spectrum which was then converted to the transmittance mode for analysis by the cross correlation library search algorithm.

All spectra incorrectly matched by Scheme 1 were also incorrectly matched by OMNIC. For Scheme 1, the sample corresponding to the only clear coat IR spectrum missed was also missed in the surfacer and primer layer searches. However, the additional sample missed in the surfacer layer search was correctly matched in the primer layer search. All paint samples incorrectly matched by Scheme 1 were also incorrectly matched by Scheme 2.

The top hit selected by Scheme 1 always yielded a reasonable match, usually superior to the match obtained between the validation set sample and the actual model for incorrectly matched paint samples (see Figures 25 through 27). This was true for all incorrectly matched paint samples. Although Scheme 2 did not perform as well as Scheme 1, it has the advantage of providing insight into how well the library matches the blinds, rather than how well an individual blind matches the library. This suggests that samples assigned the same model and line by both schemes are well represented in the library and correlate well on an individual basis to specific samples in the library. For these samples, one can have confidence in the quality of the match and high certitude regarding the accuracy of the match. A similar statement cannot be made regarding the accuracy and quality of individual matches by OMNIC as reflected by their hit quality index which are typically 98%-99% for the top 20 hits.

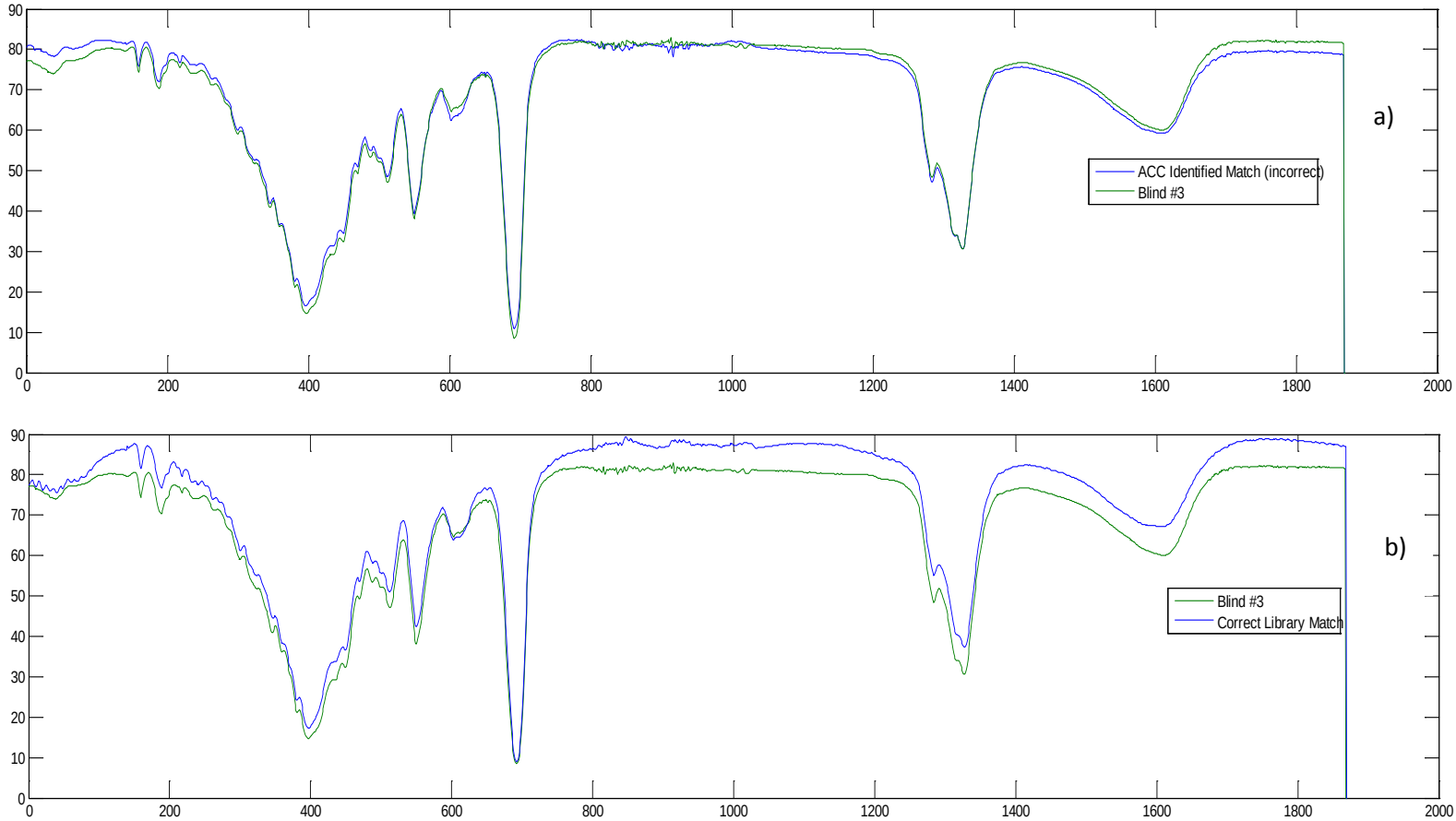

Figure 25: Chrysler Clear Coat spectral matches: a) top hit and b) correct match

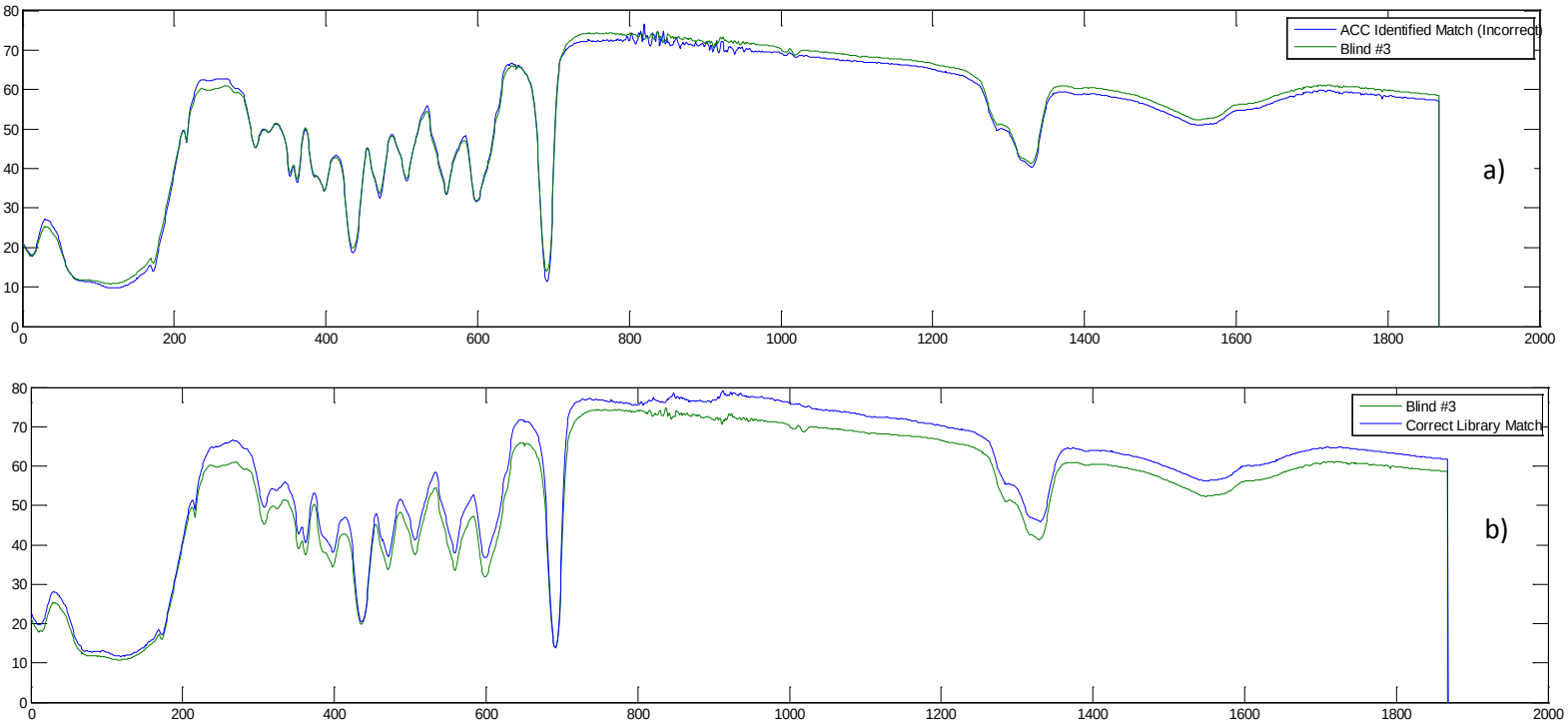

Figure 26: Chrysler Surfacer spectral matches: a) top hit and b) correct match

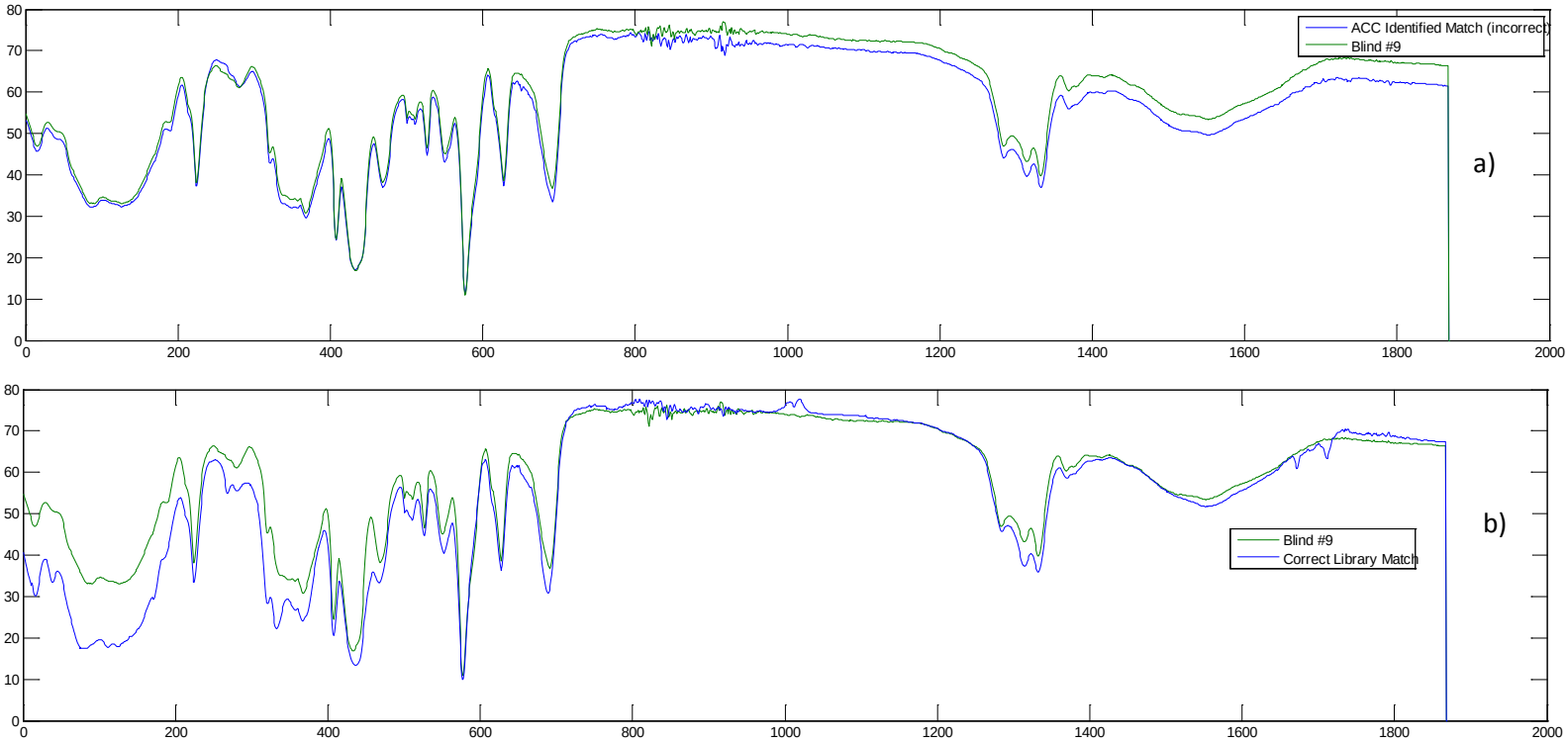

Figure 27: Chrysler Primer spectral matches: a) top hit and b) correct match

Development of Search Prefilters for General Motors

The first step in the development of the search prefilters for General Motors (GM) was to differentiate the paint systems by plant group. To determine the composition of each plant group, representative spectra were selected from each assembly plant. GM assembly plants (see Table 8) whose clear coat paint spectra exhibited a doublet for the carbonyl band (acrylic melamine styrene polyurethane) as opposed to a singlet (acrylic melamine styrene**)** were flagged. Average spectra of the clear coat layer from these nine assembly plants (Baltimore, Hamtramck, Orion, Ramos Arizpe, Silao, Spring Hill, Saint Therese, Wentzville, and Wilmington) were analyzed using cluster analysis. A separate cluster analysis was performed on the average spectra of the clear coat layer from the other sixteen assembly plants whose clear coat spectra corresponded to acrylic melamine styrene formulation (Arlington, Doraville, Fairfax, Flint, Fort Wayne, Fremont, Ingersoll, Janesville, Lansing, Linden, Lordstown, Moraine, Oklahoma City, Oshawa, Pontiac, and Shreveport).

Table 8. General Motors Assembly Plants

PLANT	PID# (Data Label)	DIVIDED BETWEEN PLANT GROUPS	Plant GROUP
Arlington (ARL)	1	NO	1
Baltimore (BAL)	2	NO	2
Doraville (DOR)	4	NO	1
Fairfax (FAI)	5	NO	1
Flint (FLI)	6	NO	3
Fort Wayne (FOR)	8	NO	1
Fremont (FRE)	9	NO	6
Hamtramck (HAM)	10	NO	2
Ingersoll (INE)	11	NO	3
Janesville (JAN)	12	NO	4
Lansing (LAN)	14	YES	1,5
Linden (LIN)	16	NO	3
Lordstown (LRD)	17	NO	6
Moraine (MOR)	18	NO	1
Oklahoma City (OKL)	20	YES	1,3
Orion (ORI)	21	NO	2
Oshawa (OSH)	22	YES	1,3,6
Pontiac (PON)	23	NO	1
Ramos Arizpe (RAM)	24	NO	5
Shreveport (SHR)	25	NO	3
Silao (SIL)	26	NO	5
Spring Hill (SPH)	27	NO	5
Saint Therese (THE)	28	NO	5
Wentzville (WEN)	29	NO	2
Wilmington (WIL)	30	NO	2

Prior to cluster analysis, principal component analysis was performed on each assembly plant to assess its class structure. PC plots of the clear coats from assembly plants corresponding to acrylic

melamine styrene polyurethane indicated that each plant was represented by a set of similar spectra. However, clustering was observed in the PC plots of three (Lansing, Oklahoma City, and Oshawa) of the seventeen assembly plants whose top coat layer corresponds to acrylic melamine styrene (see Figures 28 – 30). For Lansing, the two clusters in the PC plot were detected which were correlated with the carbonyl band being a singlet or a doublet. (This was missed in our initial analysis of the data when the assembly plants were divided into two groups based on whether the carbonyl band was a singlet or a doublet because the representative spectra selected were all singlets for the carbonyl.) For Oklahoma City, clustering was correlated to vehicle type (car or truck) with one paint sample not being a member of either cluster. This sample was tagged as an outlier after its IR spectrum was visually compared to the average clear coat paint spectrum of the two clusters detected in the principal component plot. In the case of the Oshawa assembly plant, three sample clusters were present in the principal component plot. Trucks formed one distinct cluster, whereas the automobiles were divided into two clusters on the basis of model and line. Because the average clear coat paint spectrum of each cluster in the PC plot was visually different, these three assembly plants were further divided into subplants. The Lansing subplant for the samples corresponding to the acrylic melamine styrene polyurethane formulation was transferred to the clustering study involving assembly plants whose clear coat layer was acrylic melamine styrene polyurethane.

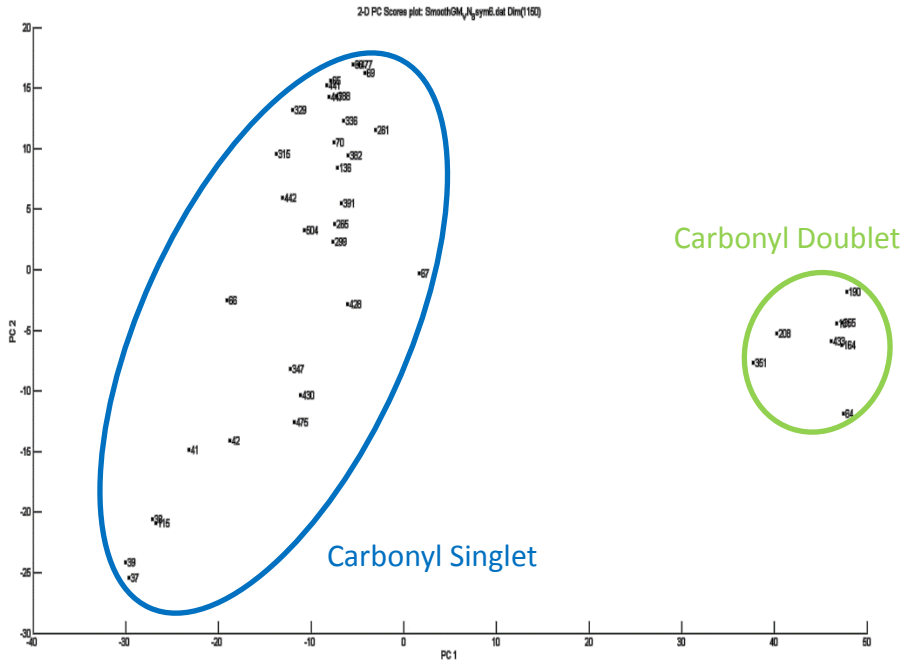

Figure 28. Plot of the two largest principal components of the wavelet transformed clear coat paint spectra from the Lansing plant. Two distinct sample clusters are evident in the plot.

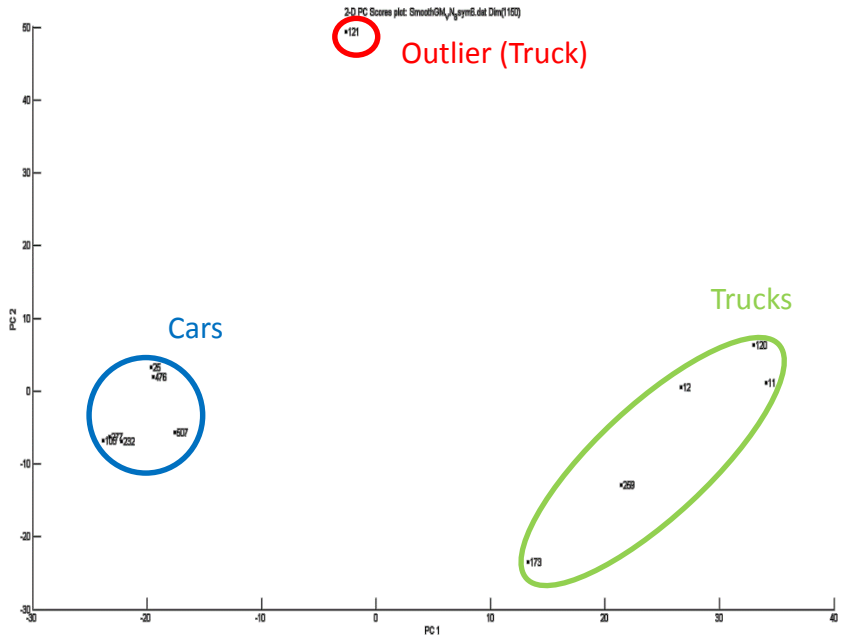

Figure 29. Plot of the two largest principal components of the wavelet transformed clear coat paint spectra from the Oklahoma City plant. Two distinct sample clusters and one outlier are evident in the plot.

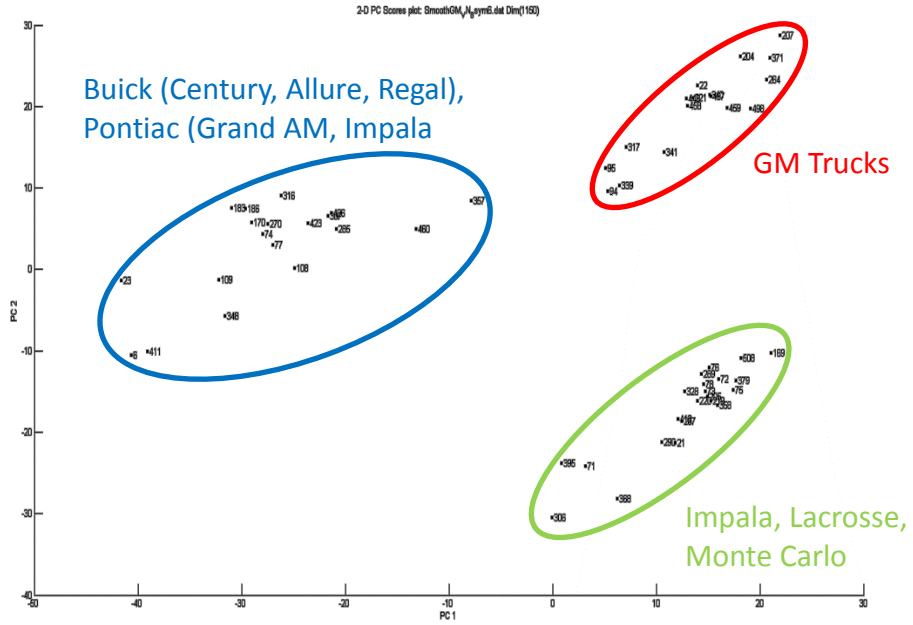

Figure 30. Plot of the two largest principal components of the wavelet transformed clear coat paint spectra from the Oshawa plant. Three distinct sample clusters are evident in the plot.

To assign the assembly plants and subplants to specific plant groups, the average IR spectrum of the clear coat layer of each assembly plant or subplant was computed. Principal component analysis and hierarchical clustering were performed on the average spectra. Figures 31 and 32 show the results of the principal component analysis and hierarchical cluster analysis for the nine assembly plants and one subplant (Baltimore, Hamtramck, Orion, Ramos Arizpe, Silao, Spring Hill, Saint Therese, Wentzville, Wilmington, and Lansing subplant) whose polymer formulation for the clear coat layer is acrylic melamine styrene polyurethane. From the results of the principal component analysis and cluster analysis, the nine assembly plants and one subplant were divided into two plant groups. Plant Group 2 consists of Baltimore, Hamtramck, Orion, Wentzville, and Wilmington assembly plants, whereas Plant Group 5 is comprised of the Lansing subplant and the Ramos Arizpe, Silao, Spring Hill, and Saint Therese assembly plants. Figures 33 and 34 show the results of principal component analysis and hierarchical clustering of the thirteen assembly plants and six subplants whose polymer formulation for the clear coat layer is acrylic melamine styrene. From the results of the principal component analysis and hierarchical clustering, the thirteen assembly plants and the six subplants were divided into four plant groups. Plant Group 1 consists of Arlington, Doraville, Fairfax, Fort Wayne, Lansing (subplant), Moraine, Oklahoma City (subplant), Oshawa (subplant) and Pontiac assembly plants, whereas Plant Group 3 is comprised of Flint, Linden, Oklahoma City (subplant), Oshawa (subplant) and Shreveport assembly plants. Plant Group 4 only contains the Janesville assembly plant, and Plant Group 6 consists of Fremont, Lordstown, and Oshawa (subplant).

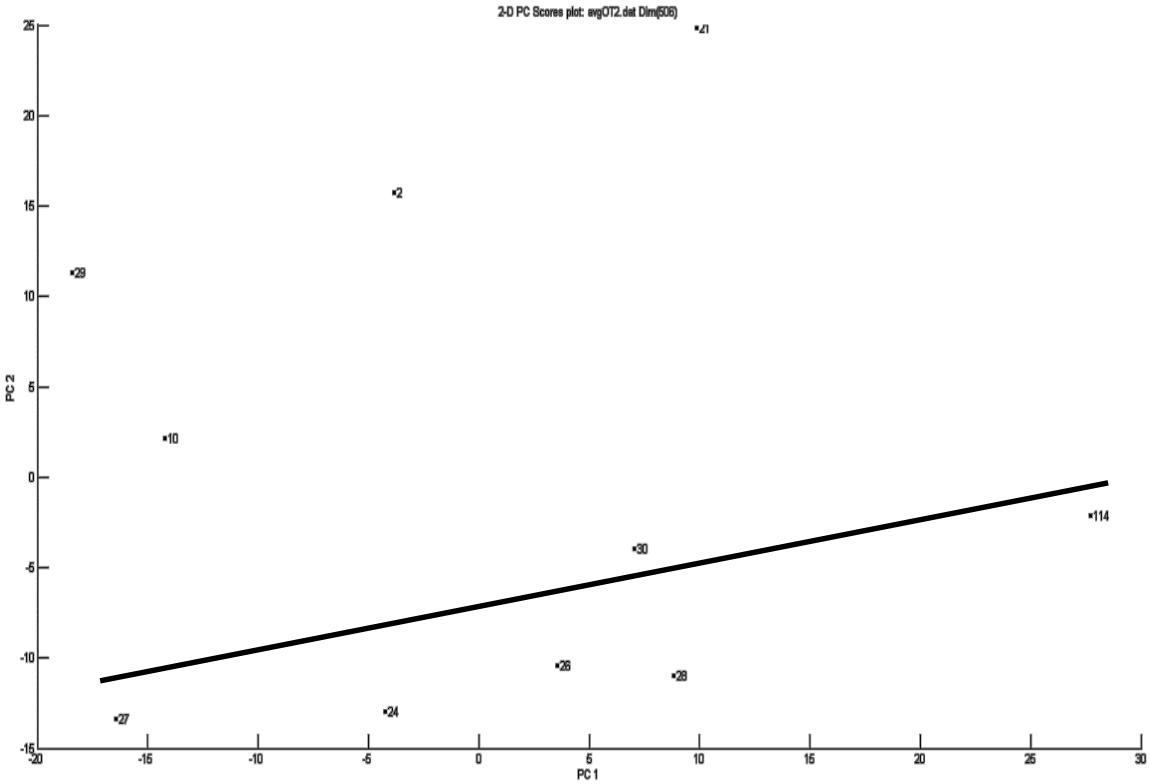

Figure 31. Principal component analysis of the average IR clear coat paint spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene polyurethane.

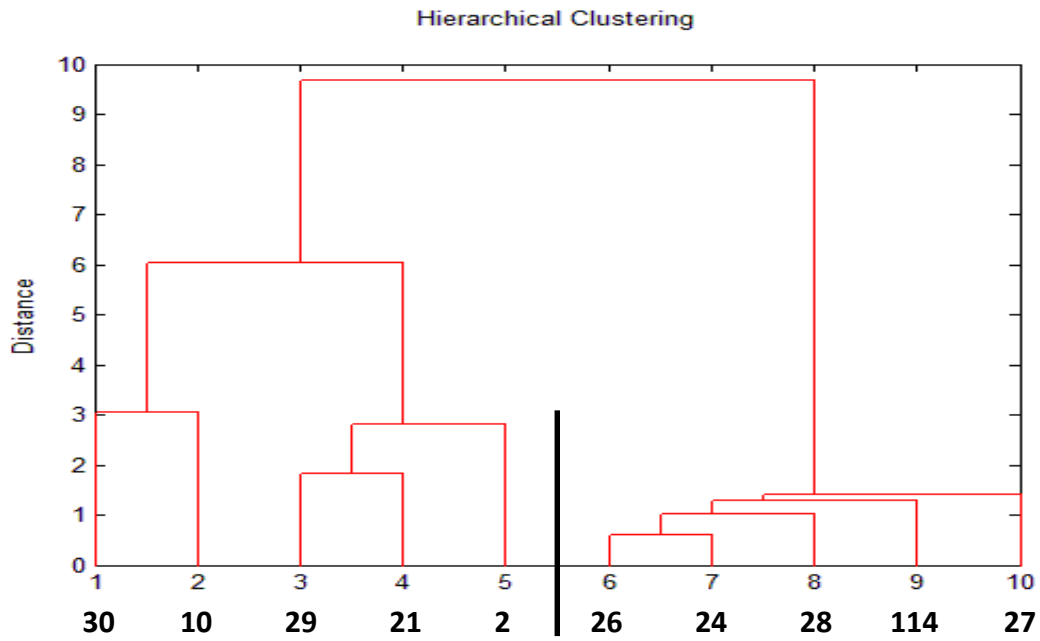

Figure 32. Hierarchical cluster analysis (Wards method) of the average IR spectrum (clear coats) of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene polyurethane.

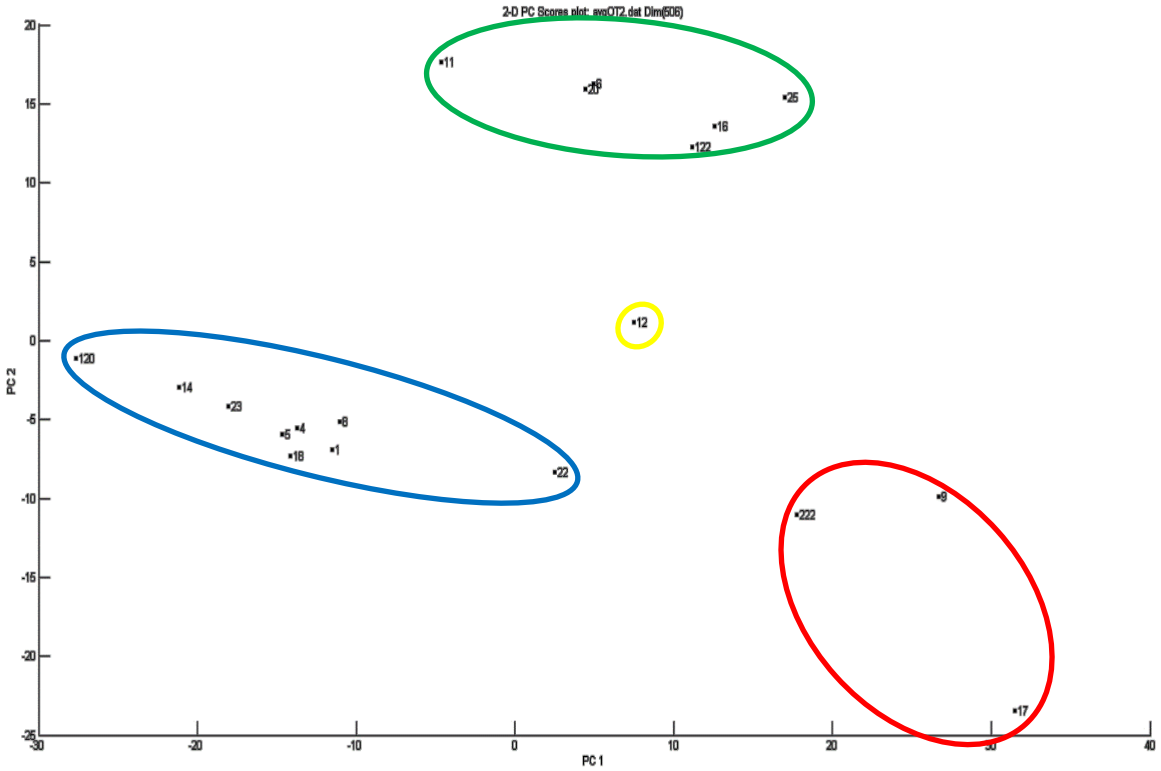

Figure 33. Principal component analysis of the average IR clear coat paint spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene.

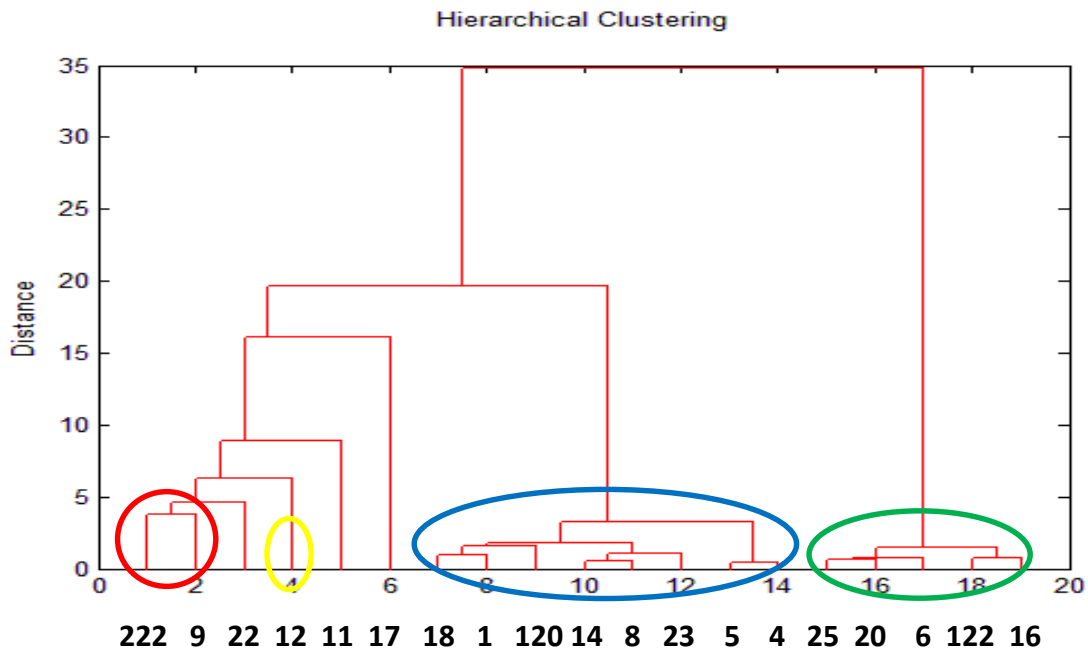

Figure 34. Hierarchical cluster analysis (Wards method) of the average IR spectrum (clear coats) of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene.

Having ascertained the membership of each plant group, the next step was classification. The training set and validation set used for Plant Group is listed in Table 9. During this phase of the study, 5 paint samples, which were flagged as discordant observations using the outlier routines of the pattern recognition genetic algorithm, were deleted from the analysis. Figure 35 shows a PC plot of the two largest principal components of the remaining 424 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group. All IR spectra in the training set were vector normalized and smoothed (Savitzky-Golay, 4th order, 17 point window) prior to the application of the wavelet transform (8Sym6), and all wavelet coefficients were autoscaled prior to principal component analysis. Each sample (clear coat) is represented as a point in the PC plot of the data.

Table 9. Training Set and Validation Set for GM Plant Groups

Group	Training	Validation
1	171	17
2	46	7
3	73	7
4	16	0
5	59	7
6	59	6

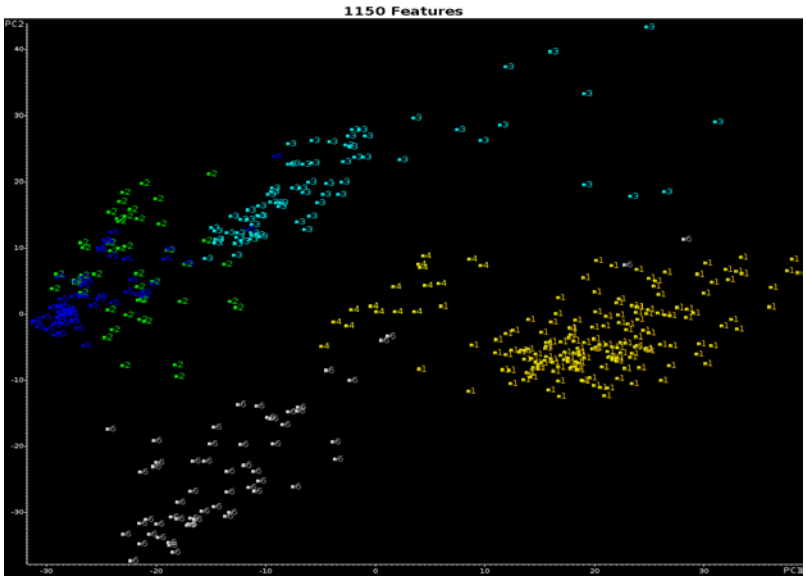

Figure 35. PC plot of the two largest principal components of the 424 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group. Each clear coat is represented as a point in the PC plot of the data. (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, 5 = Plant Group 5, 6 = Plant Group 6).

The next step was feature selection. The goal was to identify wavelet coefficients characteristic of the profile of each plant group. The pattern recognition GA identified informative wavelet coefficients by sampling key feature subsets, scoring their PC plots, and tracking those plant groups/and or IR spectra 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 26 wavelet coefficients whose PC plot showed clustering of the IR clear coat paint spectra on the basis of plant group (see Figure 36).

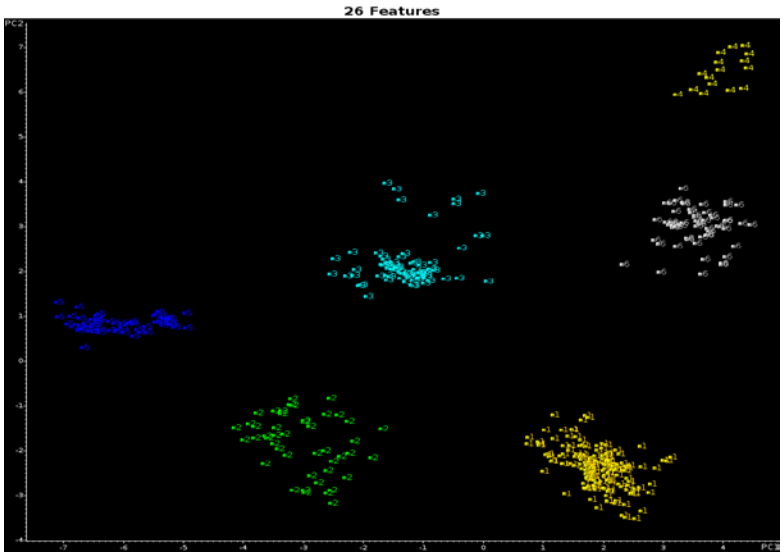

Figure 36. PC plot of the two largest principal components of the 424 training set samples and 26 wavelet coefficients identified by the pattern recognition GA (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, 5 = Plant Group 5, 6 = Plant Group 6).

To assess the predictive ability of the 26 wavelet coefficients identified by the pattern recognition GA, a validation set of 44 clear coat IR spectra was used. Clear coat IR spectra from the validation set were projected directly onto the PC plot developed from the 424 IR spectra of the training set and the 26 wavelet coefficients identified by the pattern recognition GA. Figure 37 shows the projection of the validation set samples onto the PC map of the training set data. All validation set samples are located in a region of the map with clear coats from the same Plant Group.

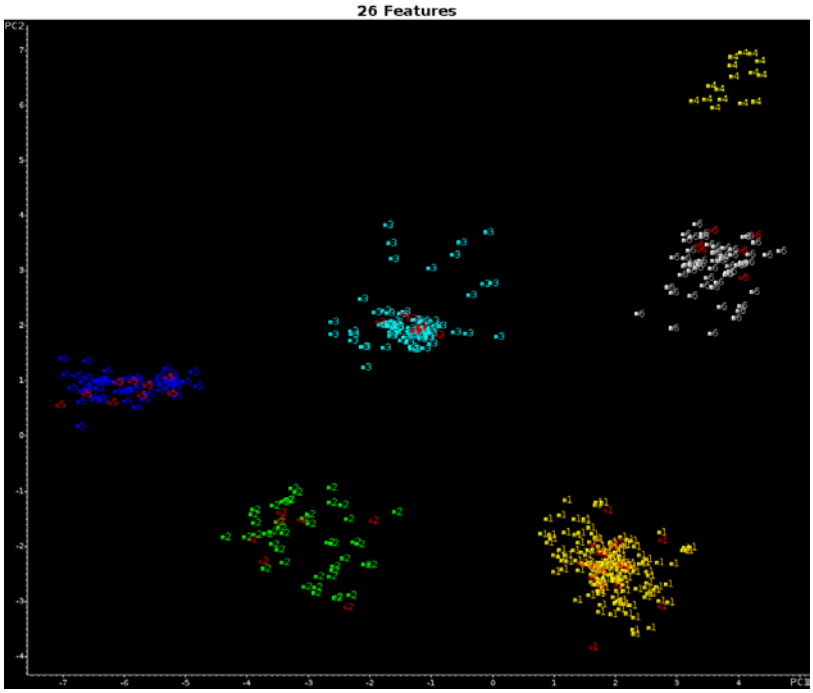

Figure 37. Validation set samples (in red) projected onto the PC plot of the data defined by the 424 wavelet transformed clear coat IR spectra of the training set and the 26 wavelet coefficients identified by the pattern recognition GA. (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, 5 = Plant Group 5, 6 = Plant Group 6).

For each plant group, a search prefilter was developed to discriminate automotive paint samples by assembly plant using the clear coat, surfacer, and primer layers. After retaining only the fingerprint region in each layer, all spectra were vector normalized and then wavelet transformed using the Symlet 6 mother wavelet at the 8th level of decomposition. Wavelet coefficients from each layer were horizontally concatenated into a single data vector in the order of clear coat, surfacer, and primer. The pattern recognition GA identified the components of this data vector (i.e., specific wavelet coefficients in each layer) correlated to the assembly plant of the vehicle from which the paint sample was obtained.

Table 10 lists the six assembly plants (Arlington, Doraville, Fairfax, Fort Wayne, Moraine, Pontiac) and three subplants (Lansing, Oklahoma City, Oshawa) comprising Plant Group 1. Doraville and the Lansing subplant were combined into a plant subgroup as were Fort Wayne and Pontiac because the spectra of their clear coat, surfacer, and primer layers were superimposable.

Table 10. Assembly and Subplants Comprising Plant Group 1

Plant	Training	Validation
1 (Arlington)	17	4
5 (Fairfax)	25	3
18 (Moraine)	27	2
20 (subplant of Oshawa)	18	1
122 (subplant of Oklahoma City)	4	1
414 (Doraville and subplant of Lansing)	52	5
823 (Fort Wayne and Pontiac)	28	1

Figure 38 shows a plot of the two largest principal components of the 171 samples comprising Plant Group 1 and the 55 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant, subplant, and plant subgroup is well separated from each other in the plot. When the validation set samples assigned to Plant Group 1 in the previous principal component plot (see Figure 37) were projected onto the PC plot shown in Figure 37, each projected validation set sample was located in a region of the map with paint samples from the same assembly plant.

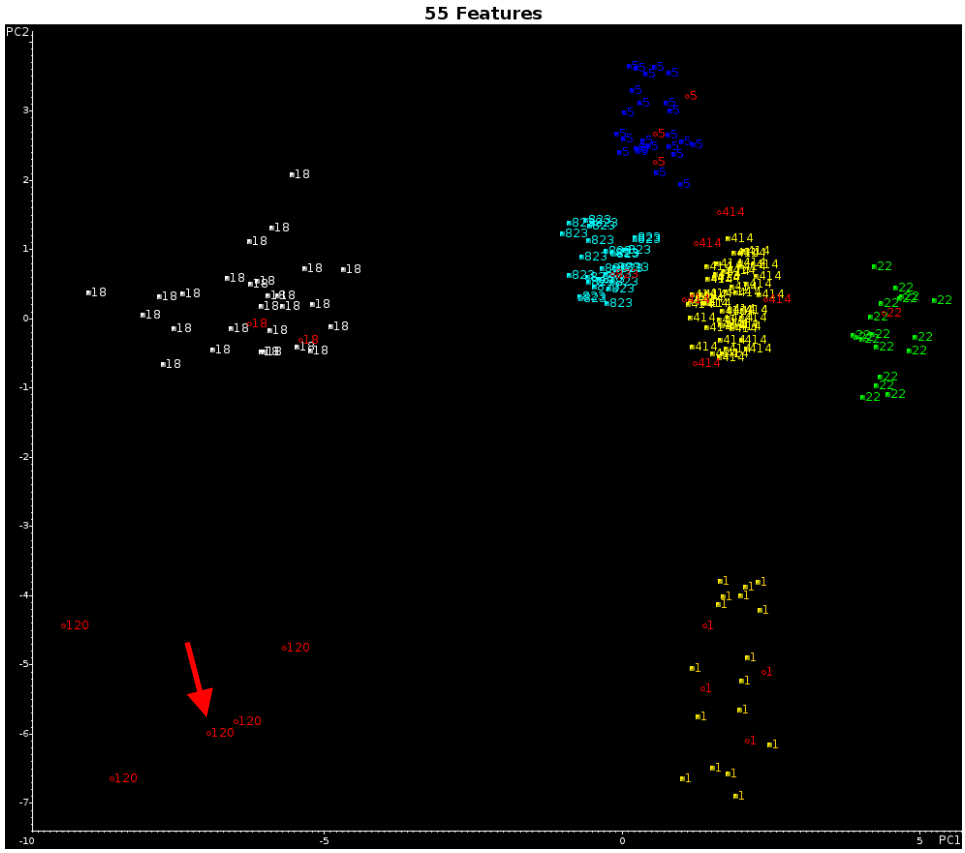

Figure 38. Validation set samples (in red or indicated by an arrow) projected onto the PC plot of the data defined by the 171 paint samples comprising Plant Group 1 (training set) and the 55 wavelet coefficients identified by the pattern recognition GA. 1 = Arlington, 5 = Fairfax, 18 = Moraine, 20 = subplant of Oshawa, 122 = subplant of Oklahoma City, 414 = Doraville and subplant of Lansing, and 823 = Fort Wayne and Pontiac.

Table 11 lists the five assembly plants (Baltimore, Hamtramck, Orion, Wentzville, Wilmington) comprising Plant Group 2. Figure 39 shows a plot of the two largest principal components of the 46 samples comprising Plant Group 2 and the 9 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 2 onto the PC map showed that each projected validation set sample is located in a region of the map with paint samples from the same assembly plant.

Table 11. Assembly and Subplants Comprising Plant Group 2

Plant	Training	Validation
2 (Baltimore)	8	0
10 (Hamtramck)	16	2
21 (Orion)	7	5
29 (Wentzville)	8	0
30 (Wilmington)	7	0

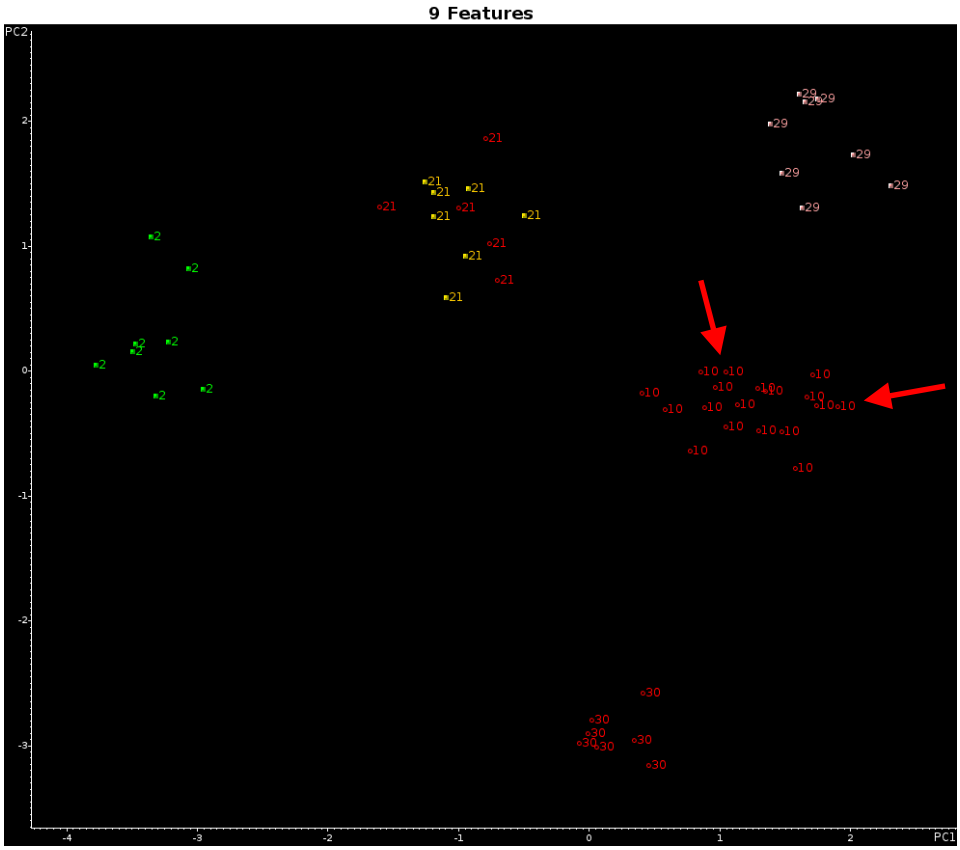

Figure 39. Validation set samples (in red or indicated by an arrow) projected onto the PC plot of the data defined by the 46 paint samples comprising Plant Group 2 (training set) and the 9 wavelet coefficients identified by the pattern recognition GA. 2 = Baltimore, 10 = Hamtramck, 21 = Orion, 29 = Wentzville, 30 = Wilmington.

Table 12 lists the four assembly plants (Flint, Ingersoll, Linden, Shreveport) and the two subplants (Oklahoma City, Oshawa) comprising Plant Group 3. Figure 40 shows a plot of the two largest principal components of the 72 samples comprising Plant Group 3 and the 43 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 3 by the plant group search prefilter onto the PC map showed that each projected validation set sample is located in a region of the map with paint samples from the same assembly plant.

Table 12. Assembly and Subplants Comprising Plant Group 3

Plant	Training	Validation
6 (Flint)	7	1
11 (Ingersoll)	10	0
16 (Linden)	11	3
20 (subplant of Oklahoma City)	7	0
25 (Shreveport)	18	1
122 (subplant of Oshawa)	20	2

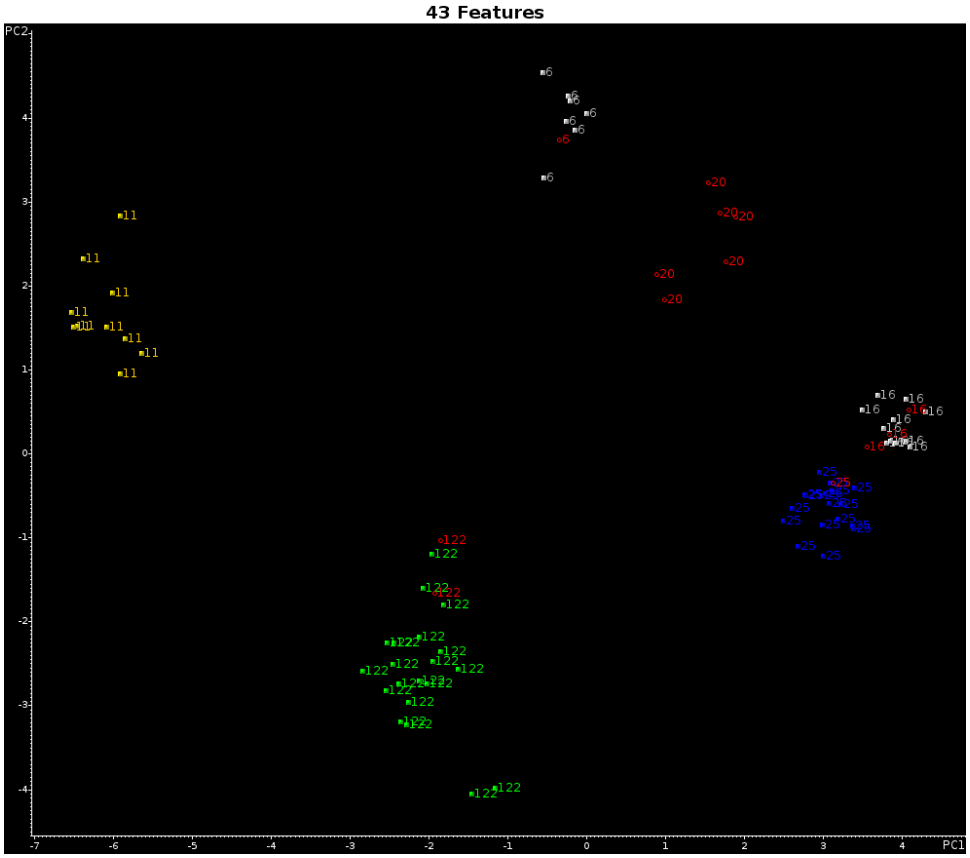

Figure 40. Validation set samples (in red or indicated by an arrow) projected onto the PC plot of the data defined by the 72 paint samples comprising Plant Group 3 (training set) and the 43 wavelet coefficients identified by the pattern recognition GA. 6 = Flint, 11 = Ingersoll, 16 = Linden, 20 = subplant of Oklahoma City, 25 = Shreveport, 122 = subplant of Oshawa.

Since Plant Group 4 consists solely of the Janesville assembly plant, no assembly plant prefilter is required. Table 13 lists the four assembly plants (Ramos Arizpe, Silao, Spring Hill, and St. Therese) and one subplant (Lansing) comprising Plant Group 5. Figure 41 shows a plot of the two largest principal components of the 59 paint samples comprising Plant Group 5 and the 30 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 5 onto the PC map showed that each projected validation set sample is located in a region of the map with paint samples from the same assembly plant.

Table 13. Assembly and Subplants Comprising Plant Group 5

Plant	Training	Validation
24 (Ramos Arizpe)	21	3
26 (Silao)	18	0
27 (Spring Hill)	8	1
28 (Saint Therese)	6	1
114 (subplant of Lansing)	6	2

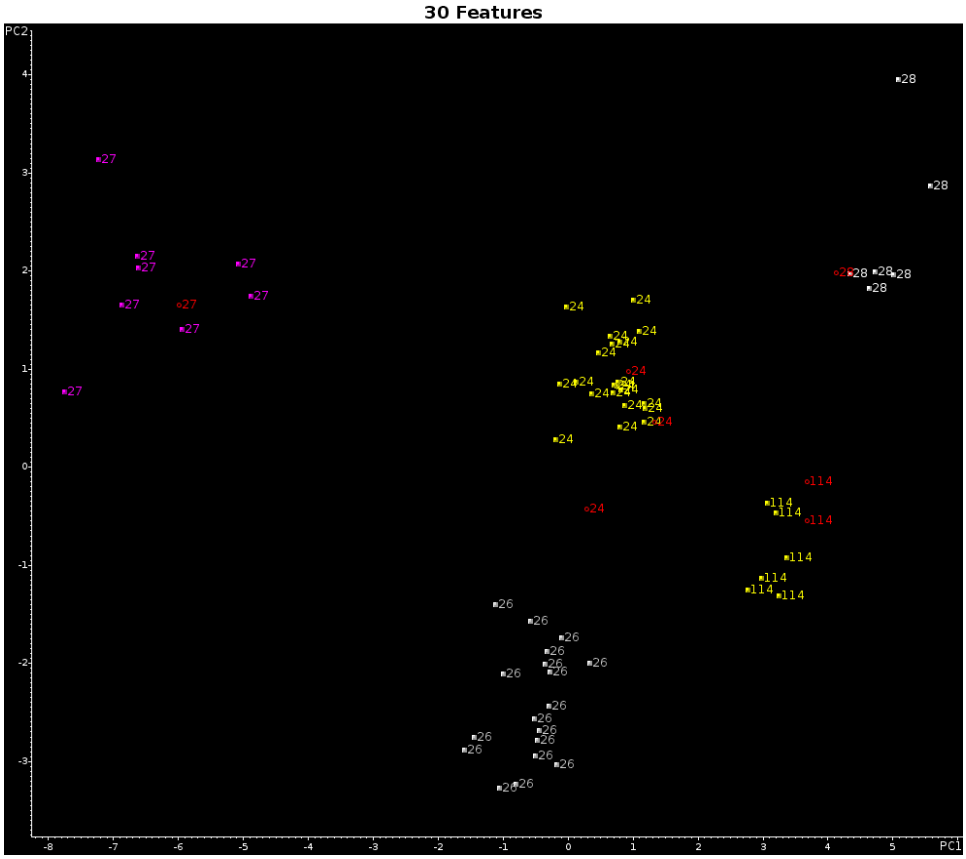

Figure 41. Validation set samples (in red or indicated by an arrow) projected onto the PC plot of the data defined by the 59 paint samples comprising Plant Group 5 (training set) and the 30 wavelet coefficients identified by the pattern recognition GA. 24 = Ramos Arizpe, 26 = Silao, 27 = Spring Hill, 28 = Saint Therese, 114 = subplant of Lansing.

Table 14 lists the two assembly plants (Fremont and Lordstown) and one subplant (Oshawa) comprising Plant Group 6. Figure 42 shows a plot of the two largest principal components of the 61 samples comprising Plant Group 6 and the 16 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant, is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 6 by the plant group search prefilter onto the PC map showed that each projected validation set sample is located in a region of the map with paint samples from the same assembly plant. The results of the GM study indicate that search prefilters developed from IR spectra of the clear coat and the two undercoat paint layers can characterize an unknown paint sample by the assembly plant of the vehicle from which the paint sample originated.

Table 14. Assembly and Subplants Comprising Plant Group 6

Plant	Training	Validation
9 (Fremont)	10	0
17 (Lordstown)	34	4
222 (subplant of Oshawa)	15	2

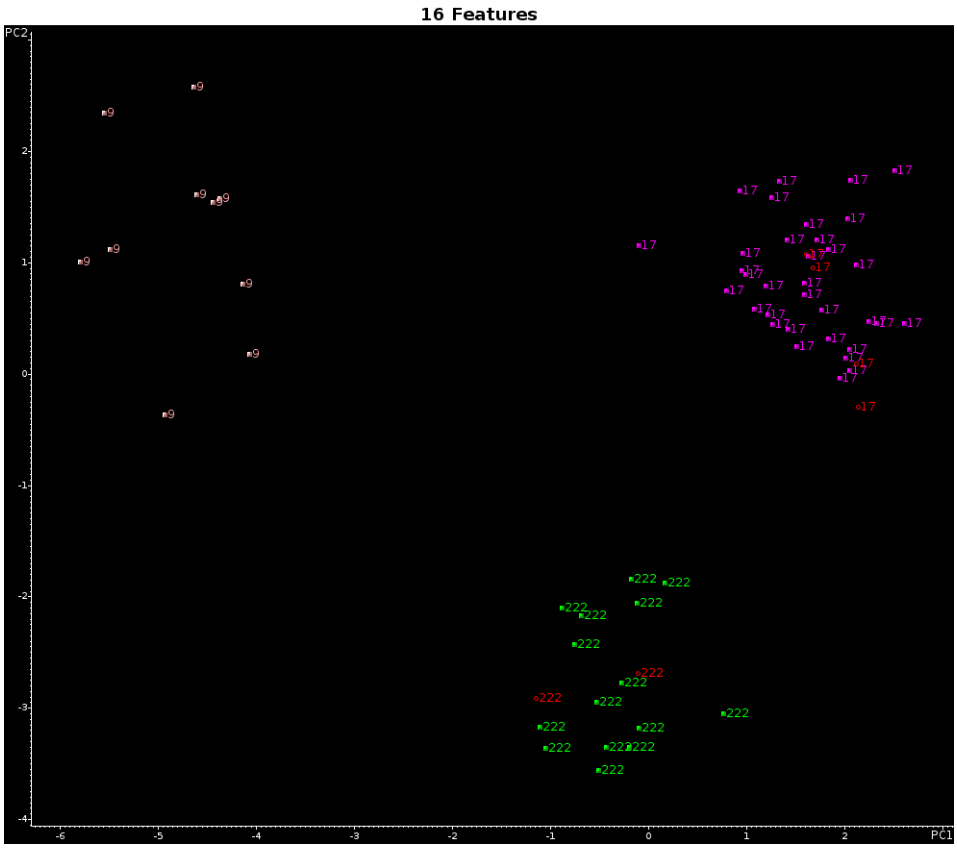

Figure 42. Validation set samples (in red or indicated by an arrow) projected onto the PC plot of the data defined by the 61 paint samples comprising Plant Group 6 (training set) and the 16 wavelet coefficients identified by the pattern recognition GA. 9 = Fremont, 17 = Lordstown, 222 = subplant of Oshawa.

Library searching of each validation set sample was performed using the IR spectra of the automotive vehicles assembled in the manufacturing plant identified by the General Motors search prefilters. Our spectral library consisted of the manufacturer's paint system for 1182 automotive Chrysler, General Motors, and Ford automotive vehicles within a limited production year range (2000-2006). To extract information about the model from the IR spectrum, library searching of each validation set sample was performed using the IR spectra of the automotive vehicles assembled in the assembly plant(s) identified by the search prefilters. The library search results for the validation set samples are summarized in Table 15 for the 44 validation set samples. Although several spectral intervals were investigated during this phase of the study (including the fingerprint region and the fingerprint region and the carbonyl band), the results tabulated here are for the region from 3675 cm^{-1} to 2856 cm^{-1} and from 1891 cm^{-1} to 668 cm^{-1}, where the relative weighting of the sub-region from 1650 cm^{-1} to 668 cm^{-1}, corresponding to the fingerprint region, was doubled relative to the other two spectral regions.

Within each interval, up to 75 windows were employed from the center burst. For Scheme 1, the remaining results were then weighted based on their average similarity index across all windows and spectral intervals. A histogram of the results was computed with the top 5 matches of each comparison used to find the closest matching sample. Results from the cross correlation library searching method were compared with the top 5 matches identified by OMNIC (Thermo Nicolet), which accessed the entire PDQ library for spectral matching. Search results for both schemes and OMNIC are summarized in Table 15. The prototype pattern recognition software system outperformed OMNIC (Thermo Nicolet) in all cases. There was one paint sample (UMNP00180) that was always misclassified by both OMNIC and the prototype pattern recognition software system. For UMNP00180, there were other models and lines that proved to be better matches. There were several samples continually missed by OMNIC. For example, both CONT00968 and UCAD00033 were usually missed by OMNIC in all the spectral regions investigated but in most cases were correctly classified by the prototype pattern recognition system. OMNIC misclassified the clear coat and the two undercoat layers of CONT01547 and UORS00235, whereas these two samples (for either the clear coat or undercoat layers) were correctly classified by the prototype pattern recognition system. For CONT01547, the correct model and line was a poor match, whereas for UORS00235, other models and lines proved to be better spectral matches.

All spectra incorrectly matched by Scheme 1 were also incorrectly matched by OMNIC. Two of the samples missed by Scheme 1 were missed in all three layers. The remaining samples missed were incorrectly matched in only some paint layers, dependent upon the quality of the individual spectra for the layer in question.

The top hit selected by Scheme 1 always yielded a reasonable match, usually superior to the match between the validation set sample and the actual model for incorrectly matched spectra (see Figures 43 through 45). This was true for all incorrectly matched paint samples. Although Scheme 2 did not perform as well as Scheme 1, it had the advantage of providing insight into how well the library matched the blinds, rather than how well an individual blind matched spectra in the library. This suggests that samples assigned the same model and line by both schemes are well represented in the library and correlate well on an individual basis to specific samples in the library. For these samples, one can have confidence in the quality of the match and high certitude regarding the accuracy of the match. A similar statement cannot be made regarding the accuracy and quality of

individual matches by OMNIC as reflected by their hit quality index which are typically 98%-99% for the top 20 hits.

Table 15. Library Search Results for General Motors

Layer	[1]OMNIC Searching	Prototype System (Scheme 1)	Prototype System (Scheme 2)
Clear coat	17	39	28
Surfacer	26	38	31
Primer	18	36	34
[2]Fused (Addition)	29	35	31

[1]OMNIC searching was done using its built-in library searching. The diamond region, 2856 cm^{-1} to 1891 cm^{-1} was excluded from the search. The results of OMNIC were considered correct if there was a match within the top 5 hits found by OMNIC's search algorithm as ranked by hit quality index.
[2]Fusion of the layers was achieved by taking spectra (in absorbance mode) of the clear coat and undercoat paint layers and directly adding them to yield a single spectrum which was then converted to the transmittance mode for analysis by the cross correlation library search algorithm.

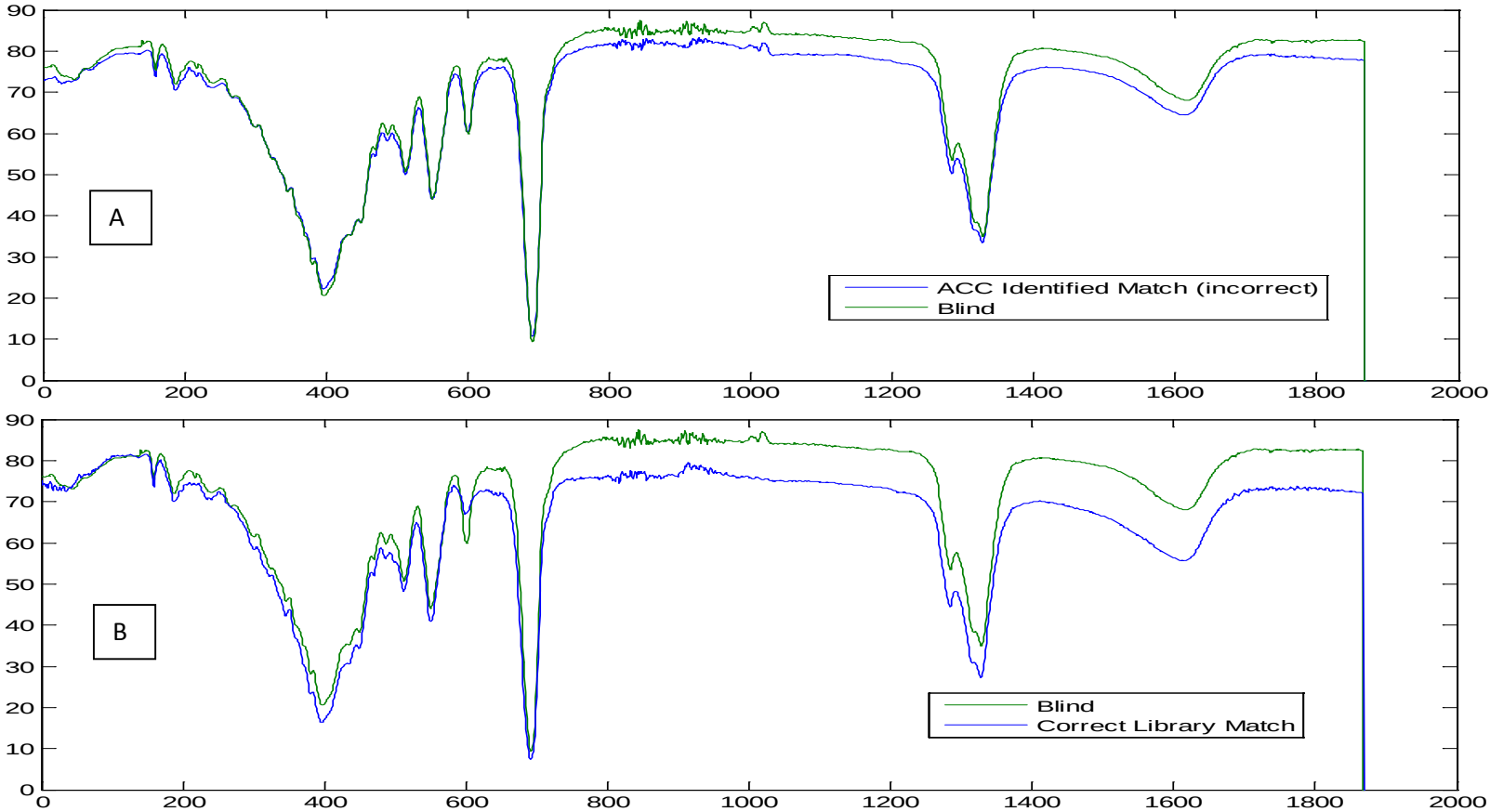

Figure 43: GM Clear Coat spectral matches: a) top hit and b) correct match

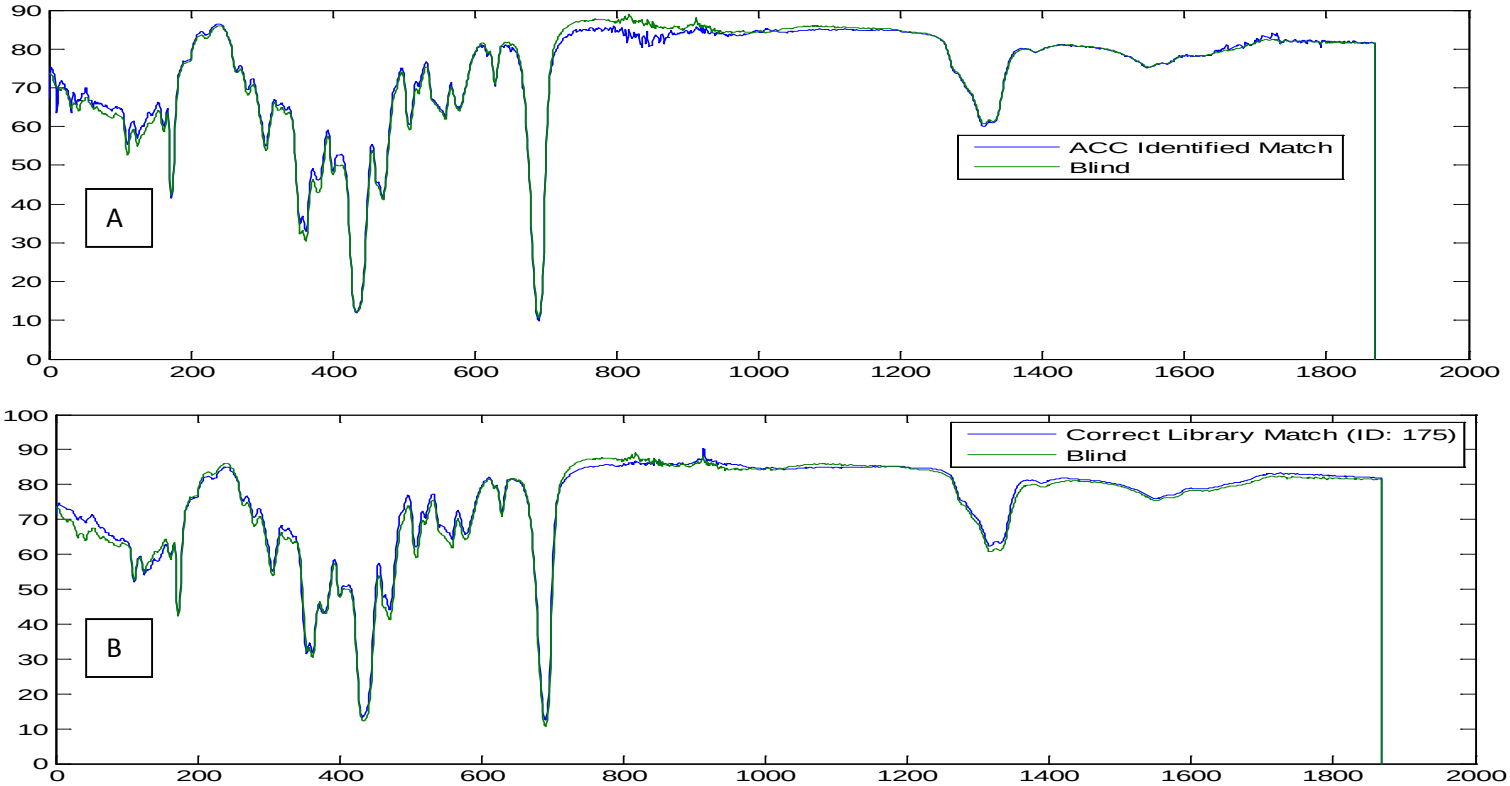

Figure 44: GM Surfacer spectral matches: a) top hit and b) correct match

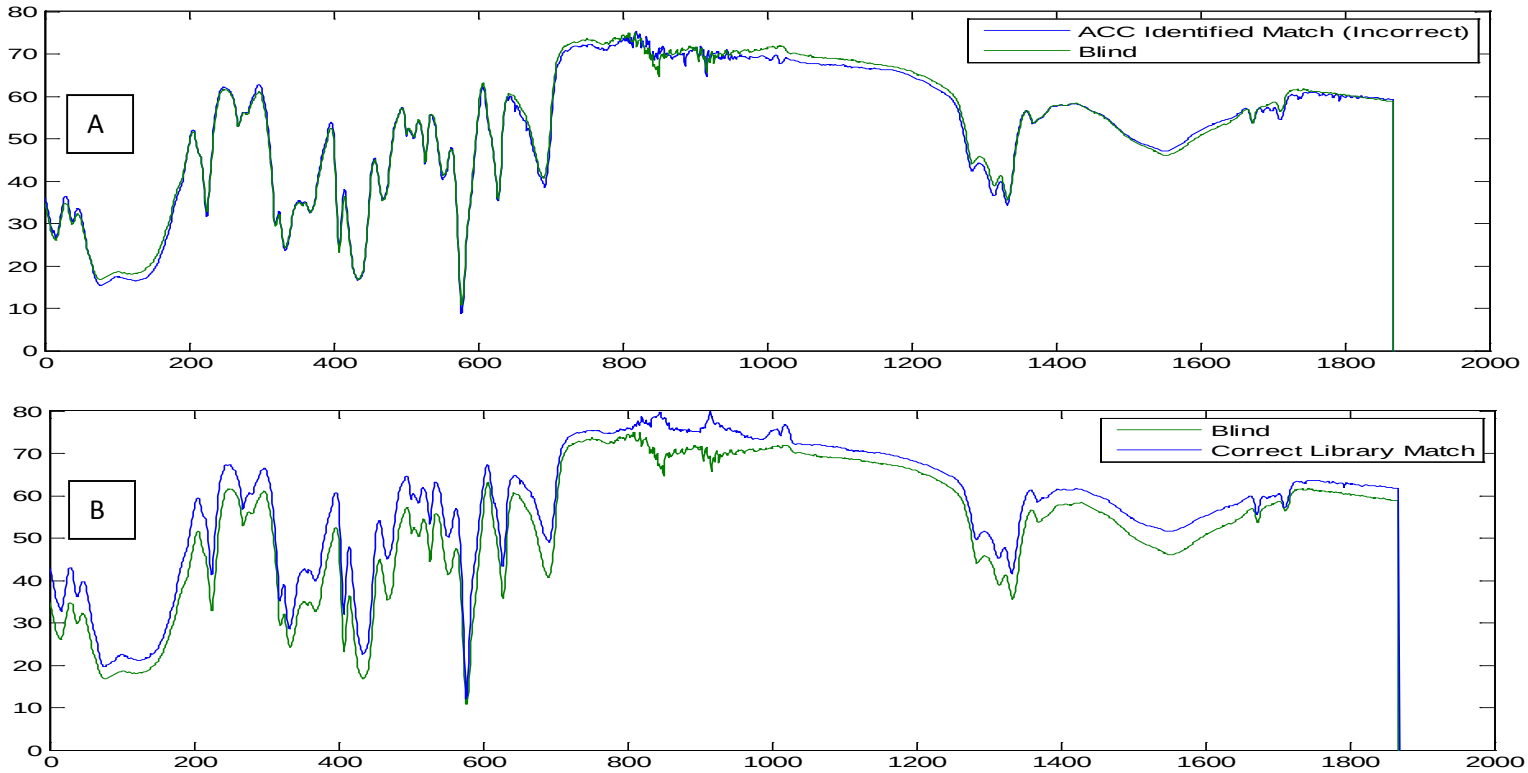

Figure 45: GM Primer spectral matches: a) top hit and b) correct match

Development of Search Prefilters for Ford

The first step in the development of the search prefilters for Ford was to differentiate the manufactured paint systems by plant group. To determine the composition of each plant group, representative spectra were selected from each assembly plant. The Ford assembly plants (see Table 16) whose clear coat paint spectra exhibited a doublet for the carbonyl band (acrylic melamine styrene polyurethane) as opposed to a singlet (acrylic melamine styrene**)** were flagged. The two assembly plants (St. Thomas-Talbotsville and Wixom) whose clear coat paint spectra exhibited a doublet for the carbonyl were assigned to Plant Group 24, whereas the other fifteen assembly plants whose clear coat spectra exhibited a singlet for the carbonyl were assigned to other plant groups.

Prior to cluster analysis, principal component analysis was performed on each Ford assembly plant to detect outliers and to assess class structure. Using only clear coat spectra, the Wixom plant was observed to contain an outlier which was attributed to the carbonyl band of the sample being a singlet, not a doublet. Distinct sample clusters were observed in the PC plots of six Ford assembly plants (Dearborn, Norfolk, Oakville, St. Thomas-Talbotsville, Twin Cities-St Paul, Wayne). For St. Thomas-Talbotsville and Dearborn (see Figures 46 and 47), the clustering was correlated to production year. However, the larger sample cluster (2000-2006) in the Saint Thomas-Talbotsville plant was represented by spectra that had a singlet for the carbonyl. For Norfolk, Oakville and Wayne (see Figures 48-50), clustering occurred on the basis of the model and the production year of the vehicle. For Twin Cities-Saint Paul (see Figure 51), clustering was correlated to production year. Since the average clear coat paint spectrum of each cluster was visually different, these six assembly plants were further divided into subplants. For St. Thomas-Talbotsville, the subplant corresponding to the acrylic melamine styrene formulation was removed from the Plant Group containing clear coats prepared using the acrylic melamine styrene polyurethane formulation.

Table 16. Ford Assembly Plant

PLANT	PID# (Data Label)	DIVIDED BETWEEN PLANT GROUPS	Plant GROUP
Atlanta (ATL)	2000	NO	21
Chicago (CHI)	2002	NO	21
Dearborn (DEA)	2003	YES	21,22
Edison (EDI)	2004	NO	21
Flat Rock (FLA)	2005	NO	22
Hermosillo (HER)	2006	NO	22
Kansas City (KAN)	2007	NO	23
Kentucky Truck (KTR)	2008	NO	21
Lorain (LOR)	2009	NO	22
Louisville (LOU)	2010	NO	23
Norfolk (NOR)	2011	YES	21, 22
Oakville (OAK)	2012	YES	21, 22
Saint Louis (STL)	2013	NO	22
Saint Thomas-Talbotsville (STT)	2014	YES	23, 24
Twin Cities-Saint Paul	2015	YES	21, 22
Wayne (WAY)	2016	YES	21, 22
Wixom (WIX)	2017	NO	24

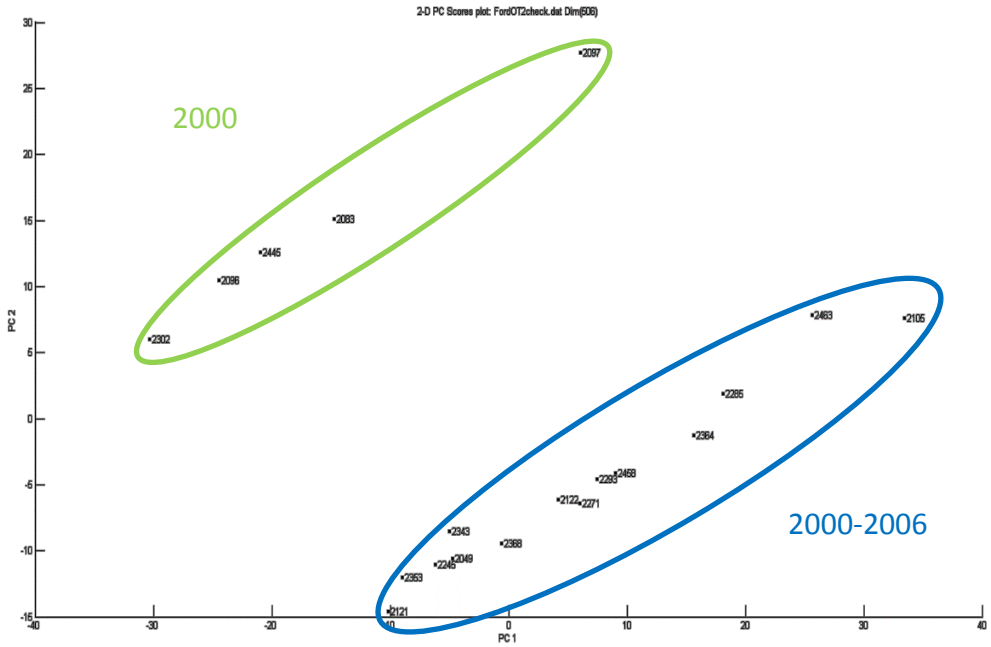

Figure 46. Plot of the two largest principal components of the wavelet transformed clear coat spectra from the Saint Thomas-Talbotsville plant. Two distinct sample clusters are evident in the plot.

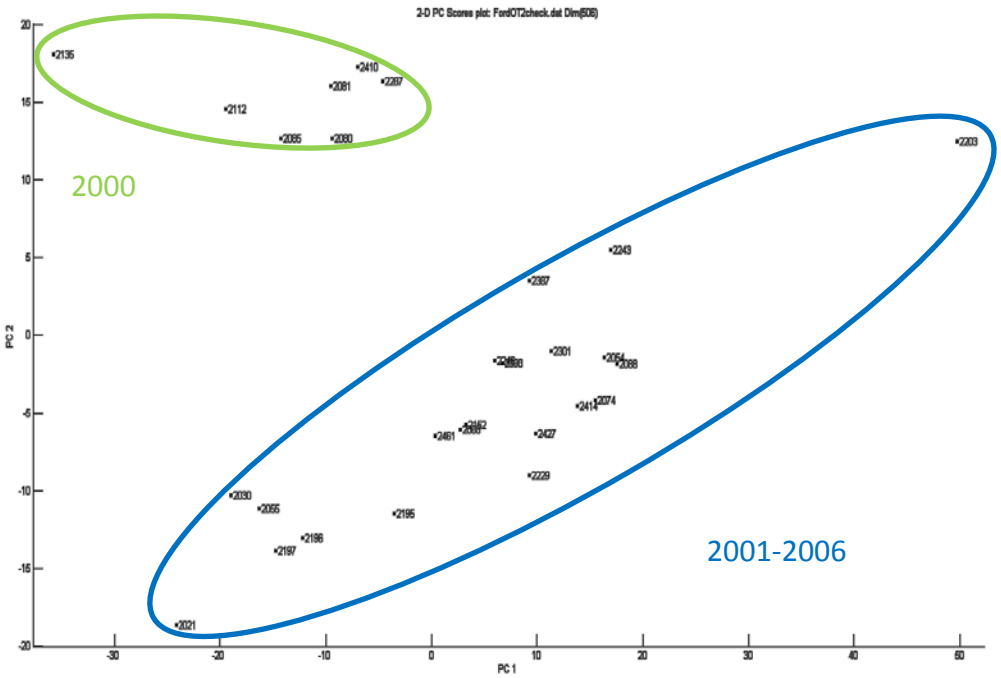

Figure 47. Plot of the two largest principal components of the wavelet transformed clear coat spectra from the Dearborn plant. Two distinct sample clusters are evident in the plot.

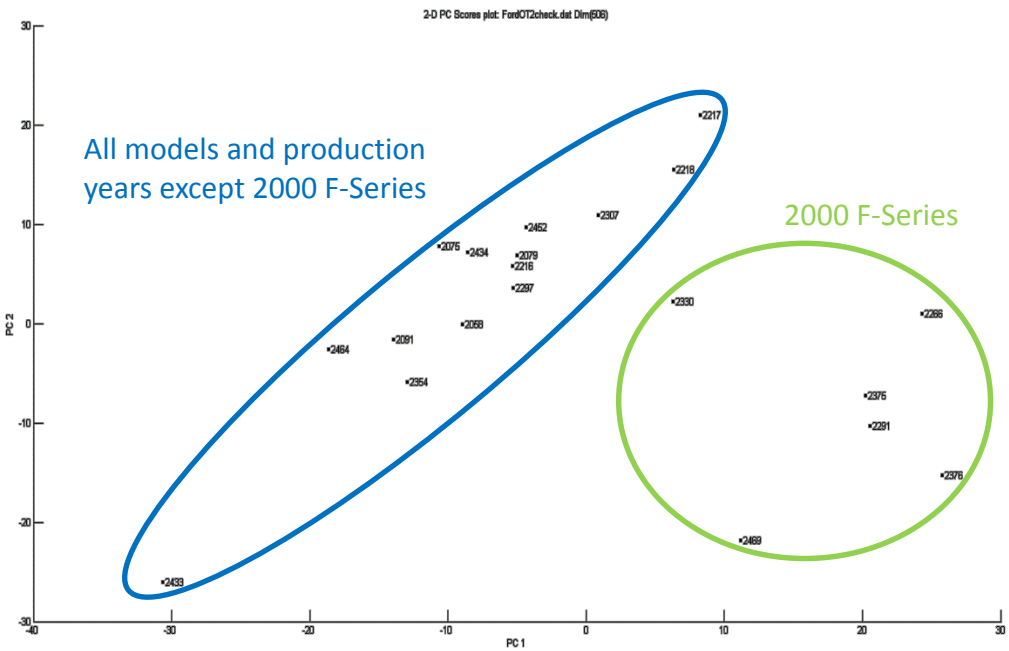

Figure 48. Plot of the two largest principal components of the wavelet transformed clear coat spectra from the Norfolk plant. Two distinct sample clusters are evident in the plot.

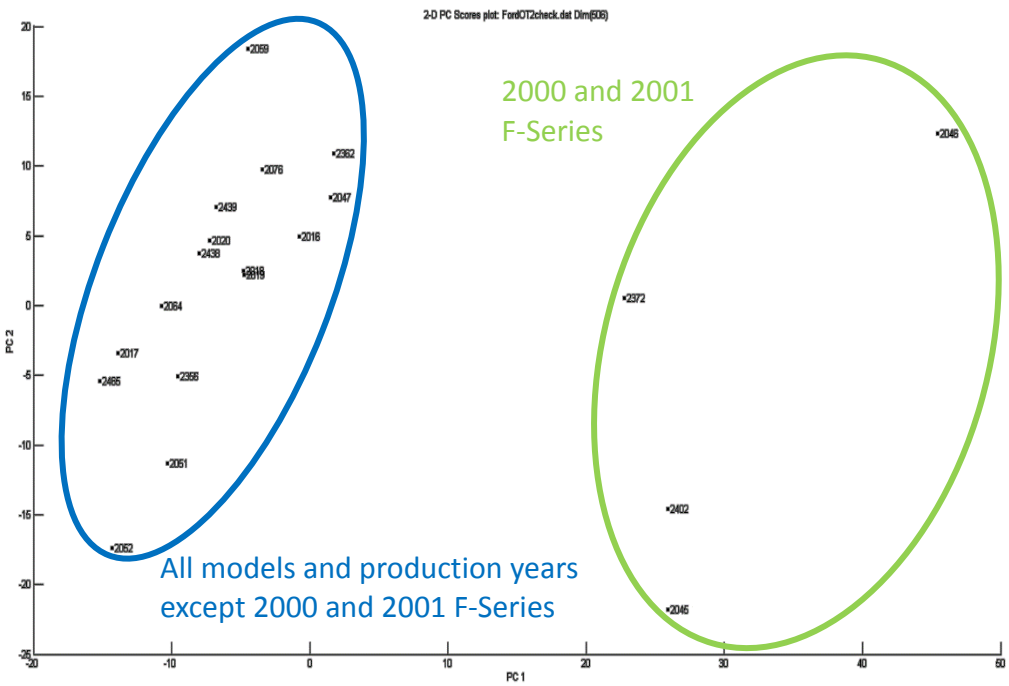

Figure 49. Plot of the two largest principal components of the wavelet transformed clear coat spectra from the Oakville plant. Two distinct sample clusters are evident in the plot.

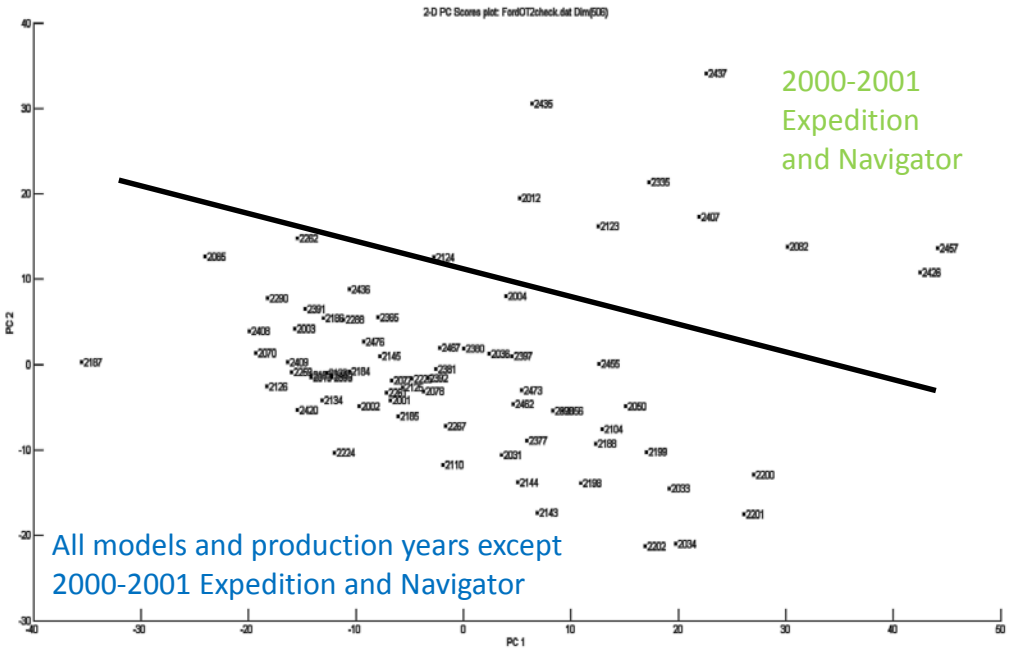

Figure 50. Plot of the two largest principal components of the wavelet transformed clear coat spectra from the Wayne plant. Two distinct sample clusters are evident in the plot.

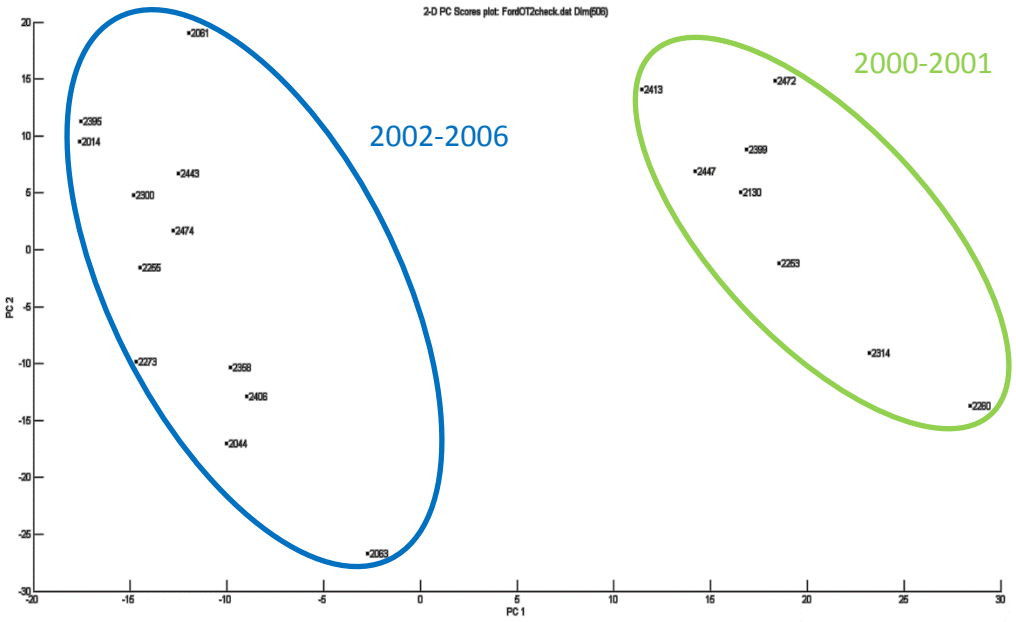

Figure 51. Plot of the two largest principal components of the wavelet transformed clear coat spectra from the Twin Cities-Saint Paul plant. Two distinct sample clusters are evident in the plot.

To assign the remaining eleven assembly plants (Atlanta, Chicago, Dearborn, Edison, Flat Rock, Hermosillo, Kansas City, Kentucky Truck, Lorain, Louisville, Saint Louis) and nine subplants (Norfolk/two subplants, Oakville/two subplants, St. Thomas-Talbotsville, Twin Cities – St. Paul/two subplants, Wayne/ two subplants) to specific plant groups, the average IR spectrum of the clear coat layer of each assembly plant or subplant was computed. Principal component analysis and hierarchical clustering were performed on the average spectra.

Figures 52 and 53 show the results of principal component analysis and hierarchical clustering for the twelve assembly plants and nine subplants whose polymer formulation for the clear coat layer was acrylic melamine styrene. From the results of the principal component analysis and hierarchical clustering, the eleven assembly plants and the nine subplants were divided into three plant groups. Plant Group 21 consists of Atlanta, Chicago, Dearborn (subplant), Edison, Kentucky Truck, Norfolk (subplant), Oakville (subplant), Twin Cities-St. Paul (subplant), and Wayne (subplant). Plant Group 22 consists of Dearborn (subplant), Flat Rock, Hermosillo, Lorain, Norfolk (subplant), Oakville (subplant), St. Louis, Twin Cities-St. Paul (subplant), and Wayne (subplant). Plant Group 23 consists of Kansas City, Louisville, St. Thomas-Talbotsville (subplant) and the lone Wixom paint sample whose clear coat spectrum has a singlet for the carbonyl band. (The Wixom sample was included in the cluster analysis for the purpose of assessing its similarity relative to other assembly plants or subplants as the possibility was raised about the sample being mislabeled.)

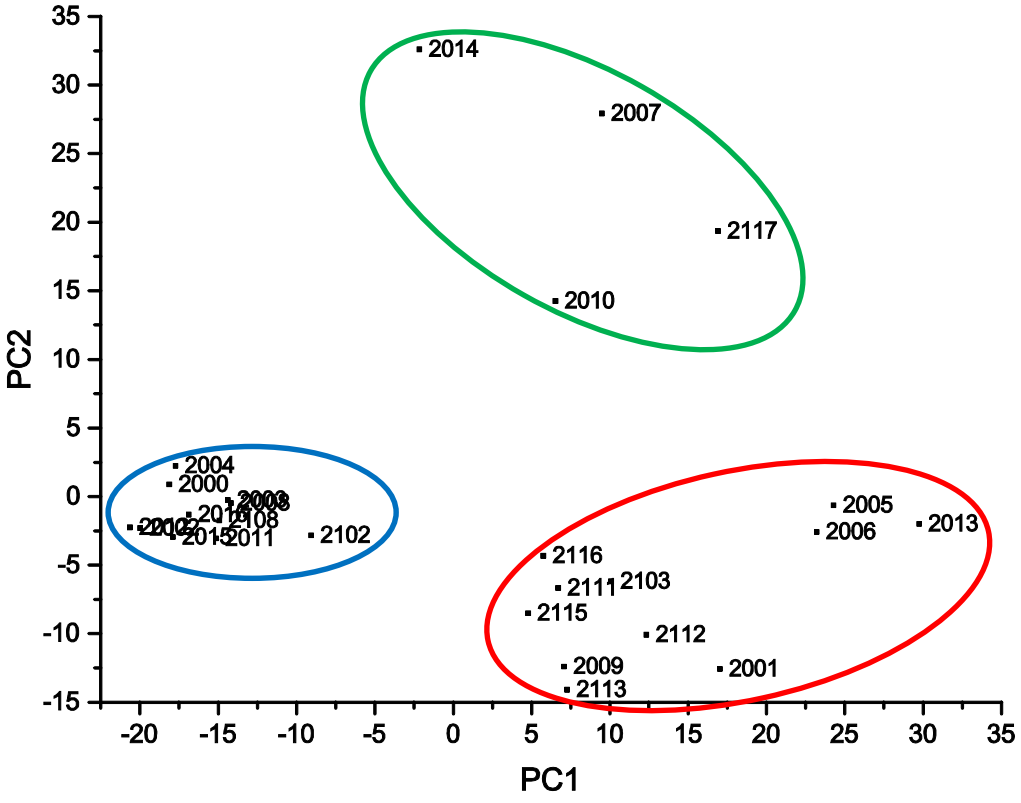

Figure 52. Principal component analysis of the average IR clear coat spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene.

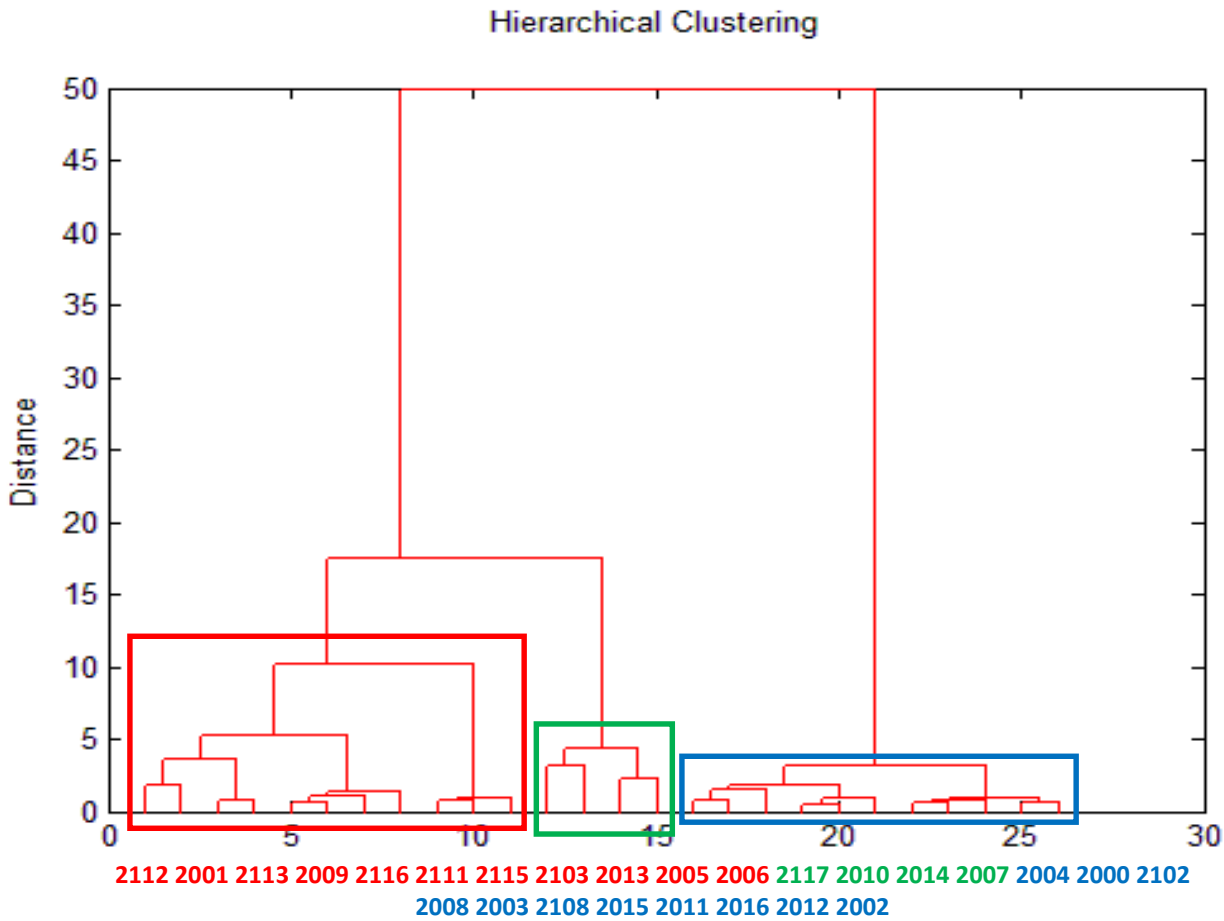

Figure 53. Hierarchical cluster analysis (Wards method) of the average IR spectrum (clear coats) of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene.

Having ascertained the membership of each plant group, the next step was classification. The training and validation sets used for Plant Group are listed in Table 17. Figure 54 shows a PC plot of the two largest principal components of the 375 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group. All IR spectra in the training set were vector normalized and smoothed (Savitzky-Golay, 4th order, 17 point window) prior to the application of the wavelet transform (8Sym6), and all wavelet coefficients were autoscaled prior to principal component analysis. Each sample (clear coat) is represented as a point in the PC plot of the data. The overlap between Plant Groups 22, 23, and 24 in the plot is evident.

Table 17. Training Set and Validation Set for Ford Plant Groups

Group	Training	Validation
21	180	15
22	86	12
23	85	8
24	23	3

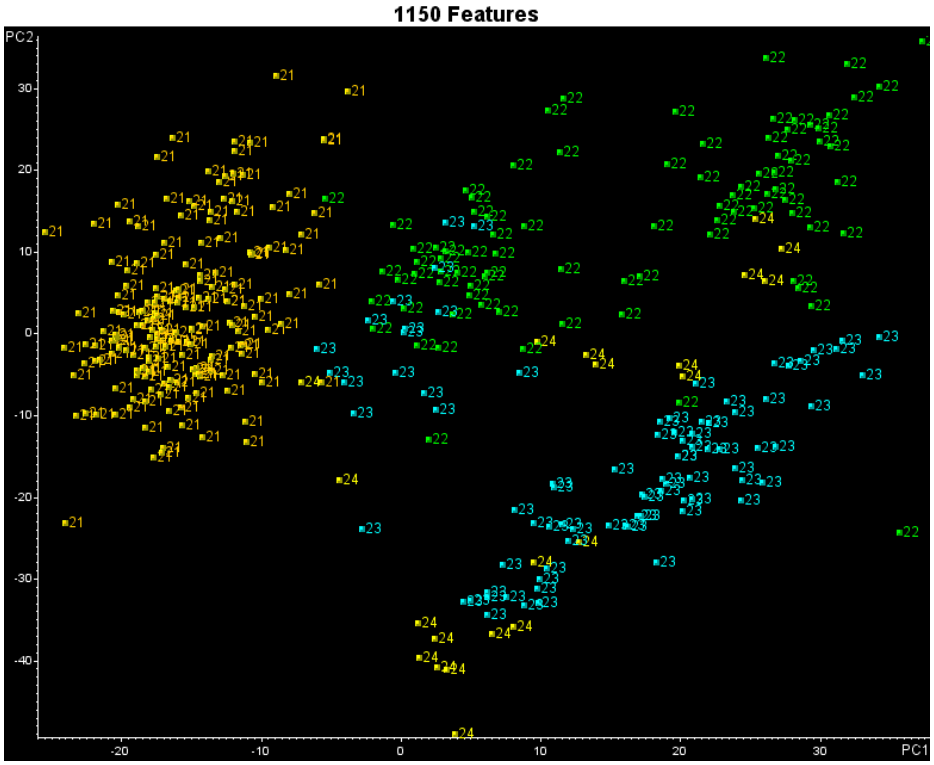

Figure 54. PC plot of the two largest principal components of the 374 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group. Each clear coat is represented as a point in the PC plot of the data. (21 = Plant Group 21, 22 = Plant Group 22, 23 = Plant Group 23, 24 = Plant Group 24).

The next step was feature selection. The goal was to identify wavelet coefficients characteristic of the profile of each plant group. The pattern recognition GA identified informative wavelet coefficients by sampling key feature subsets, scoring their PC plots, and tracking those plant groups/and or IR spectra 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 37 wavelet coefficients whose PC plot showed clustering of the IR clear coat paint spectra on the basis of plant group (see Figure 55).

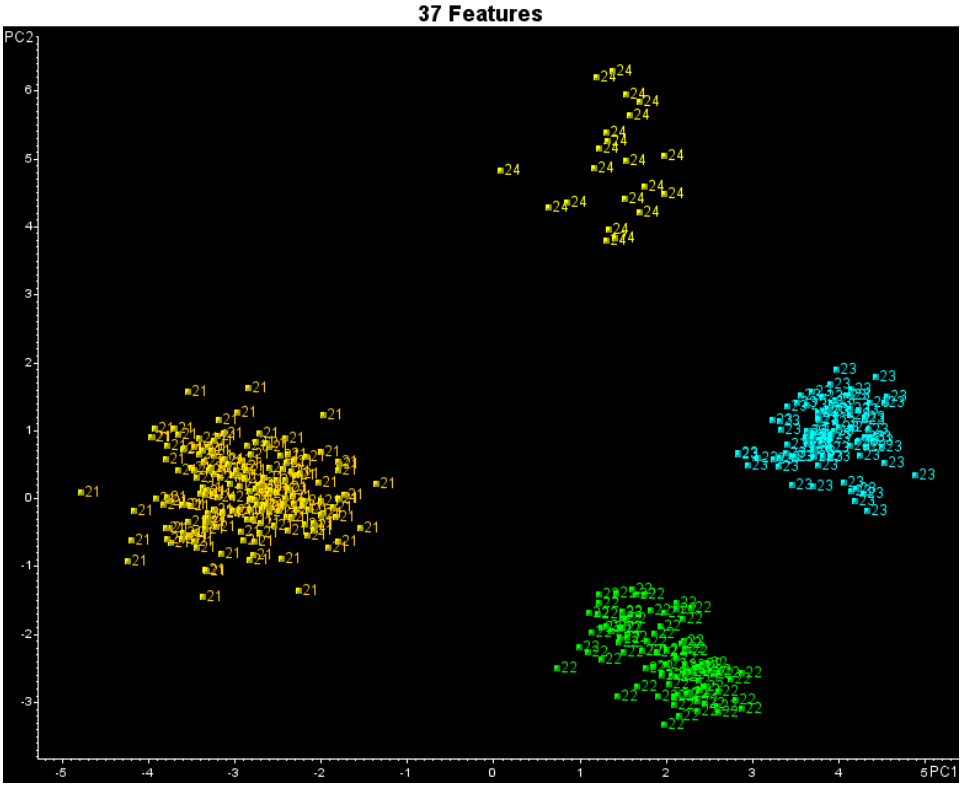

Figure 55. PC plot of the two largest principal components of the 374 training set samples and 37 wavelet coefficients identified by the pattern recognition GA (21 = Plant Group 21, 22 = Plant Group 22, 23 = Plant Group 23, 24 = Plant Group 24).

To assess the predictive ability of the 37 wavelet coefficients identified by the pattern recognition GA, a validation set of 38 clear coat IR spectra was used. Clear coat IR spectra from the validation set were projected directly onto the PC plot developed from the 374 IR spectra of the training set and the 37 wavelet coefficients identified by the pattern recognition GA. Figure 56 shows the projection of the validation set samples onto the PC map of the training set data. All validation set samples are located in a region of the map with clear coats from the same Plant Group.

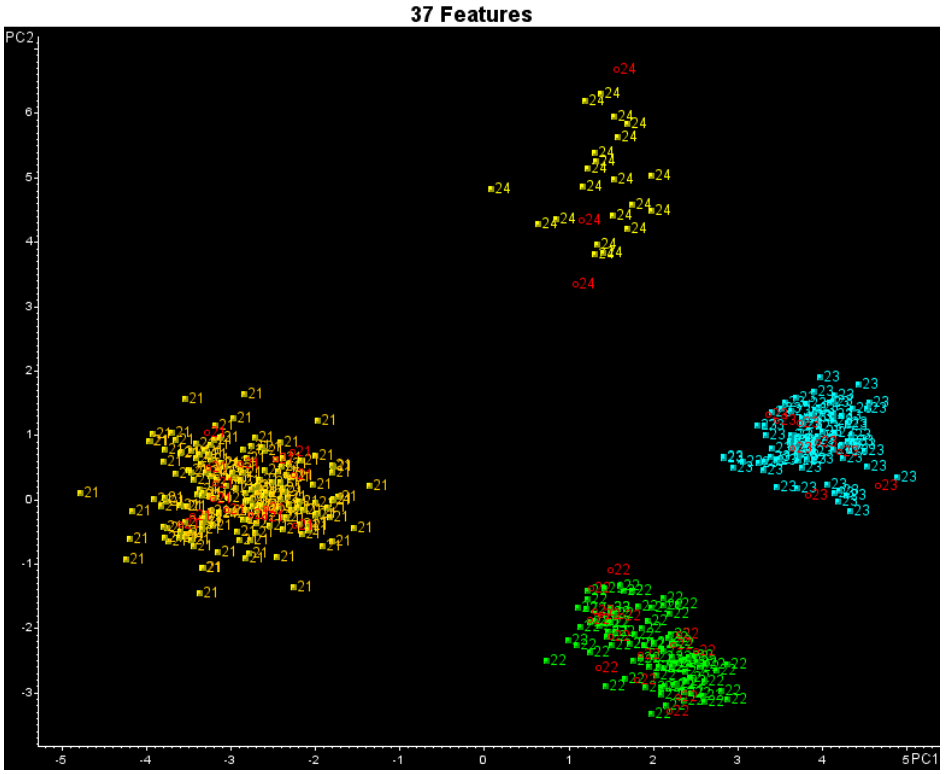

Figure 56. Validation set samples (in red) projected onto the PC plot of the data defined by the 374 wavelet transformed clear coat IR spectra of the training set and the 37 wavelet coefficients identified by the pattern recognition GA. (21 = Plant Group 21, 22 = Plant Group 22, 23 = Plant Group 23, 24 = Plant Group 24).

For each plant group, a search prefilter was developed to discriminate automotive paint samples by assembly plant using the clear coat, surfacer, and primer layers. After retaining only the fingerprint region in each layer, all spectra were vector normalized and then wavelet transformed using the Symlet 6 mother wavelet at the 8th level of decomposition. Wavelet coefficients from each layer were horizontally concatenated into a single data vector in the order of clear coat, surfacer, and primer. The pattern recognition GA identified the components of this data vector (i.e., specific wavelet coefficients in each layer) correlated to the assembly plant of the vehicle from which the paint sample was obtained.

Table 18 lists the subplant (Dearborn) and plant subgroup (Atlanta, Chicago, Edison, Kentucky Truck, Norfolk (subplant), Oakville (subplant), Twin Cities-St. Paul (subplant), Wayne (subplant)) comprising Plant Group 21. Figure 57 shows a plot of the two largest principal components of the 180 samples comprising Plant Group 21 and the 17 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. The validation set samples assigned to Plant Group 21 which are projected onto the PC map are shown in red. Only the Dearborn subplant could be reliably discriminated from the other assembly plants and subplants in this plant group. Because of the similarity of their spectra, the other assembly plants and subplants were combined into a single plant subgroup.

Table 18. Assembly and Subplants Comprising Plant Group 21

Plant	Training	Validation
2003 (subplant of Dearborn)	19	1
2481 (Atlanta, Chicago, Edison, Kentucky Truck, subplant of Norfolk, subplant of Oakville, subplant of Twin Cities-St. Paul, subplant of Wayne)	161	14

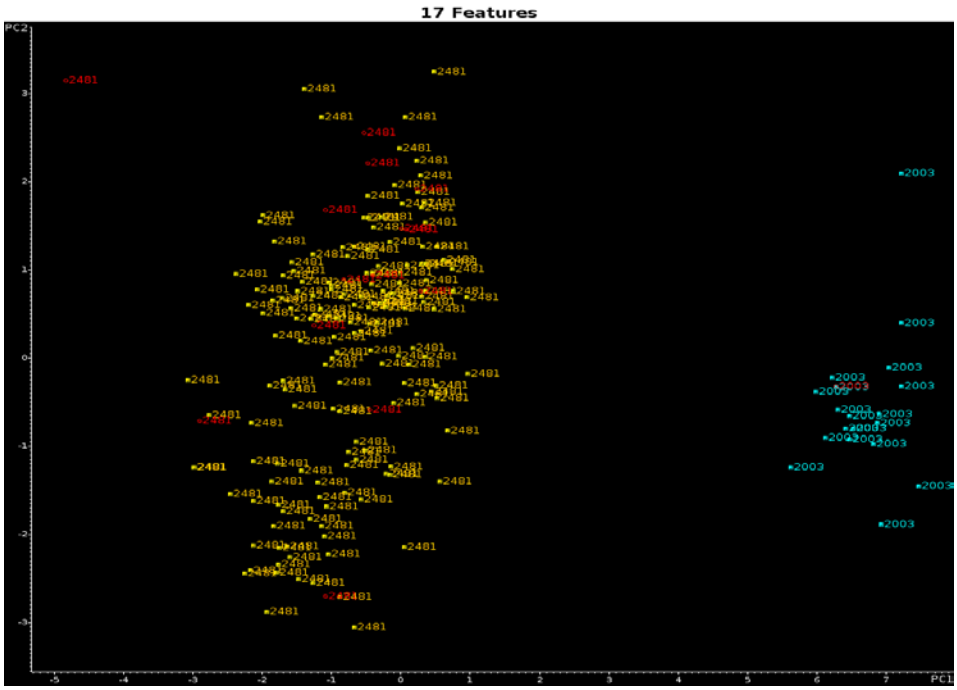

Figure 57. Validation set samples (in red) are projected onto the PC plot of the data defined by the 180 paint samples comprising Plant Group 21 (training set) and the 17 wavelet coefficients identified by the pattern recognition GA. 2003 = subplant of Dearborn, 2481 = Atlanta, Chicago, Edison, Kentucky Truck, Norfolk (subplant), Oakville (subplant), Twin Cities-St. Paul (subplant), Wayne (subplant).

Table 19 lists the three plant subgroups comprising Plant Group 22. Plant subgroup 1923 consists of Lorain, Oakville, and the St. Louis subplant, whereas Plant subgroup 3156 consists of the subplants of Dearborn, Norfolk, Twin Cities-St. Paul, and Wayne. Plant subgroup 5613 is comprised of the Flat Rock and Hermosillo plants, as well as the other St. Louis subplant. For St. Louis, it was necessary to divide the assembly plant into two subplants due to clustering observed in the surfacer layer (see Figure 58).

Figure 59 shows a plot of the two largest principal components of the 79 samples comprising Plant Group 22 and the 24 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. The validation set samples assigned to Plant Group 22, which are projected onto the PC map, are shown in red. Every validation set sample was projected onto a region of the map with paint samples from the same plant subgroup. During the course of feature selection, 6 training set samples were detected as outliers and were deleted from the analysis. Two were from Flat Rock, one was from Hermosillo, one was from Lorain, and two were from the Twin Cities-St. Paul subplant. These six samples

differed from their assembly plant or subplant in the surfacer and/or primer layers. A seventh sample (Avon Lake) was discarded from the analysis because the corresponding assembly plant was represented by only a single sample. Thus, the original training set of 86 wavelet preprocessed fused IR spectra was truncated to 79 spectra for assembly plant search prefilter development.

Table 19. Assembly and Subplants Comprising Plant Group 22

Plant	Training	Validation
1923 (Lorain, Oakville, subplant of St. Louis)	14	2
3156 (subplant of Dearborn, subplant of Norfolk, subplant of Twin Cities-St. Paul, subplant of Wayne)	25	5
5613 (Flat Rock, Hermosillo, subplant of St. Louis)	40	5

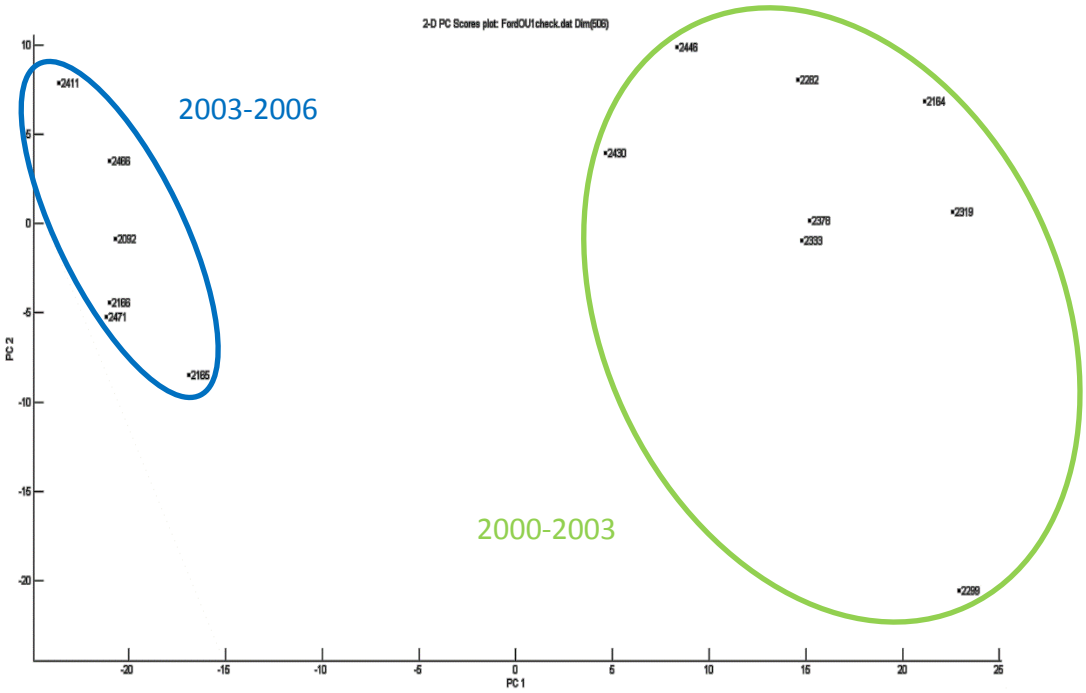

Figure 58. Plot of the two largest principal components of the surfacer spectra from the St. Louis assembly plant. Two distinct clusters are evident in the plot.

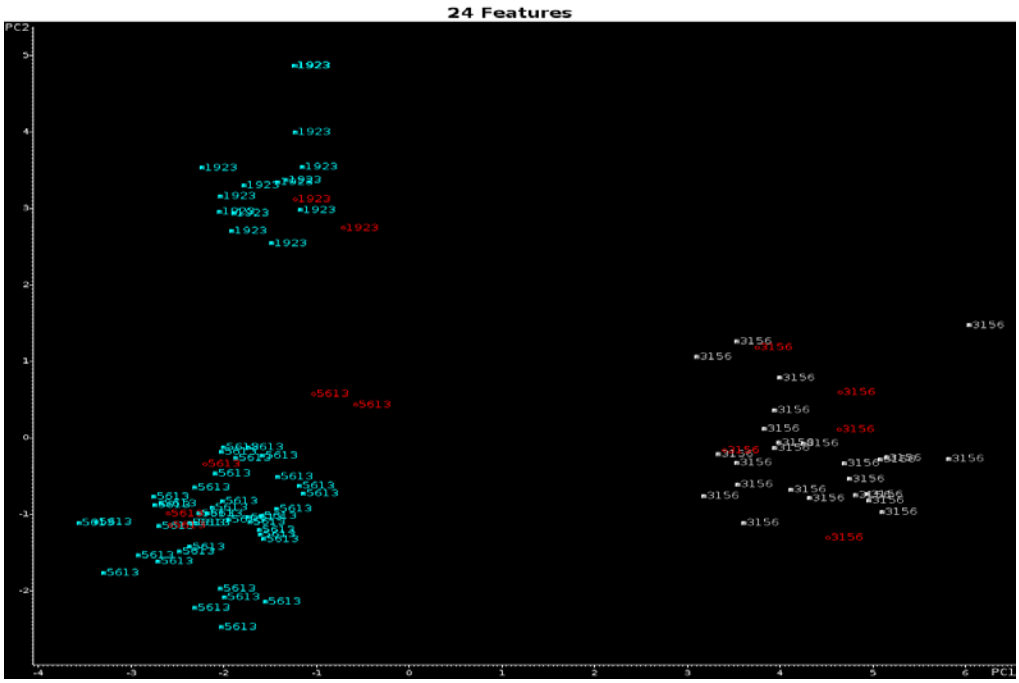

Figure 59. Validation set samples (in red) are projected onto the PC plot of the data defined by the 79 paint samples comprising Plant Group 22 (training set) and the 24 wavelet coefficients identified by the pattern recognition GA. 1923 = Lorain, Oakville, and St. Louis (subplant), 3156 = Dearborn (subplant), Norfolk (subplant), Twin Cities-St. Paul (subplant), and Wayne (subplant), 5613 = Flat Rock, Hermosillo, St. Louis (subplant).

Table 20 lists the two plants (Kansas City and Louisville) and the subplant (St. Thomas-Talbotsville) comprising Plant Group 23. Figure 60 shows a plot of the two largest principal components of the 80 samples comprising Plant Group 23 and the 51 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the PC plot. The validation set samples assigned to Plant Group 23 which are projected onto the PC map are shown in red. All validation set samples except for one from Louisville were projected onto a region of the map with paint samples that were from the same assembly plant or subplant. Five of the original 86 training set samples in Plant Group 23 were found to be outliers and discarded during the course of the pattern recognition analysis to develop an assembly plant and subplant search prefilter for this plant group. Four were from Kansas City, and the other was from Louisville. These five samples differed from their assembly plant or subplant in the surfacer and/or primer layers. A sixth sample (Wixom) was discarded because the corresponding subplant was represented by a single sample – the only carbonyl singlet among Wixom clear coat spectra.

Table 20. Assembly and Subplants Comprising Plant Group 23

Plant	Training	Validation
2007 (Kansas City)	39	4
2010 (Louisville)	28	3
2014 (subplant of St. Thomas-Talbotsville)	13	1

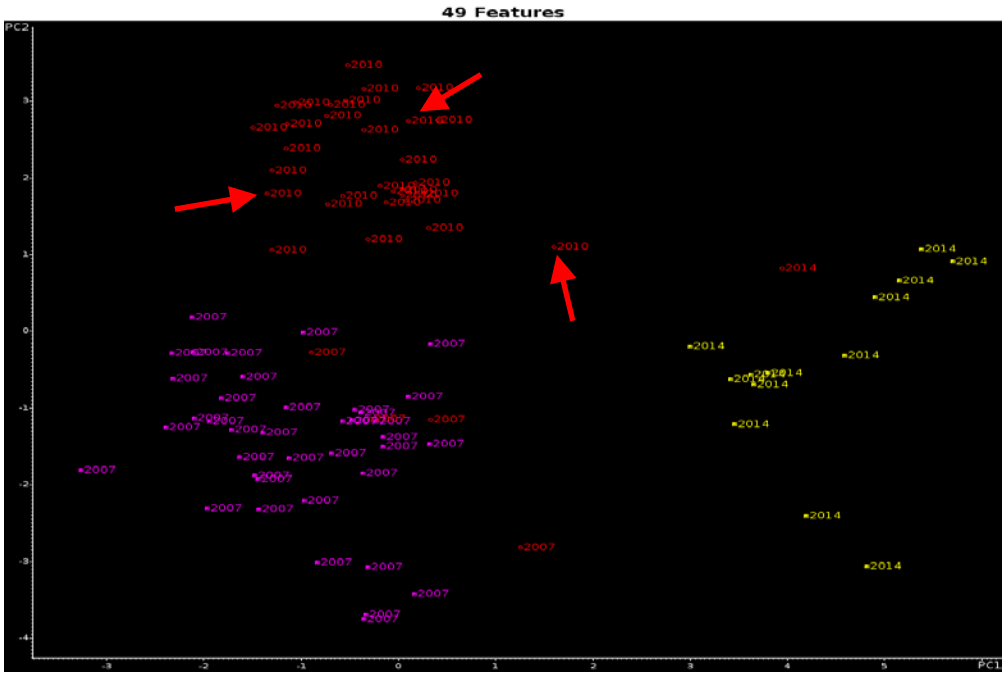

Figure 60. Validation set samples (in red or indicated by arrows) are projected onto the PC plot of the data defined by the 80 paint samples comprising Plant Group 23 (training set) and the 51 wavelet coefficients identified by the pattern recognition GA. 2007 = Kansas City, 2010 = Louisville, 2014 = St. Thomas-Talbotsville (subplant).

For Plant Group 24, the pattern recognition GA was not able to identify a set of coefficients from the wavelet transformed concatenated IR spectra that could differentiate Wixom from St. Thomas - Talbotsville. To understand the reason, principal component analysis was performed on both the Wixom and St. Thomas-Talbotsville assembly plants. Clustering correlated to the production year of the vehicle was observed for the Wixom assembly plant. For this reason, Wixom was divided into two subplants (see Figure 61). Table 21 lists the three subplants (one from St. Thomas-Talbotsville, two from Wixom) comprising Plant Group 24.

Table 21. Assembly and Subplants Comprising Plant Group 24

Plant	Training	Validation
2017 (subplant of Wixom)	9	1
2114 (subplant of St. Thomas-Talbotsville)	5	0
2217 (subplant of Wixom)	7	2

Figure 62 shows a plot of the two largest principal components of the 21 training set samples comprising Plant Group 24 and the 2 wavelet coefficients identified by the pattern recognition GA for the training set. All three subplants are well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 24 onto the PC plot showed that each projected validation set sample is located in a region of the PC map containing samples from the same subplant. As only two wavelet coefficients were necessary for the development of the search prefilter, it is logical to conclude that the underlying classification problem is simple.

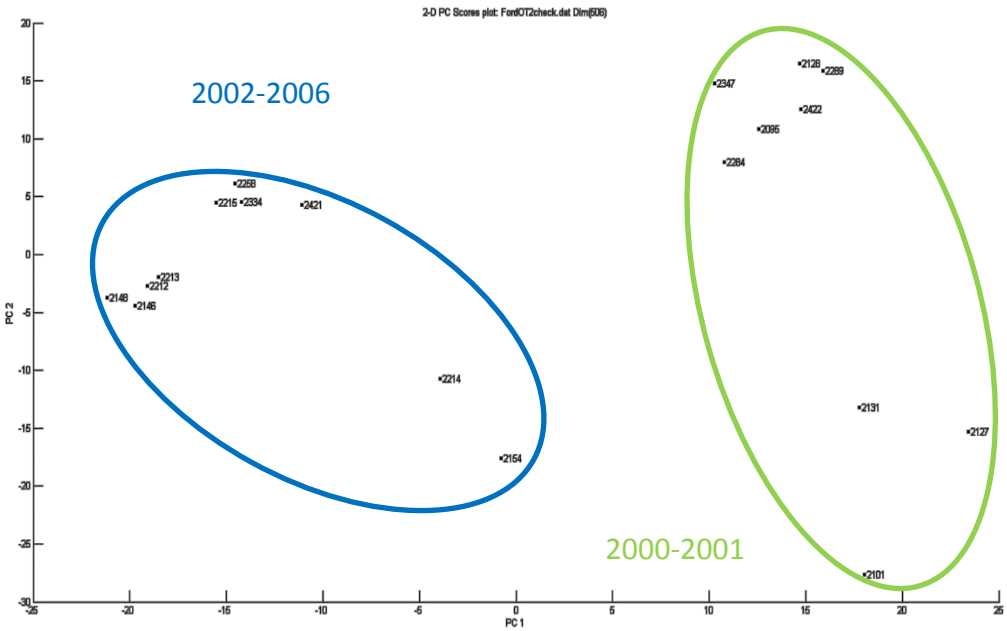

Figure 61. Plot of the two largest principal components of the surfacer spectra from the Wixom assembly plant. Two distinct clusters are evident in the plot.

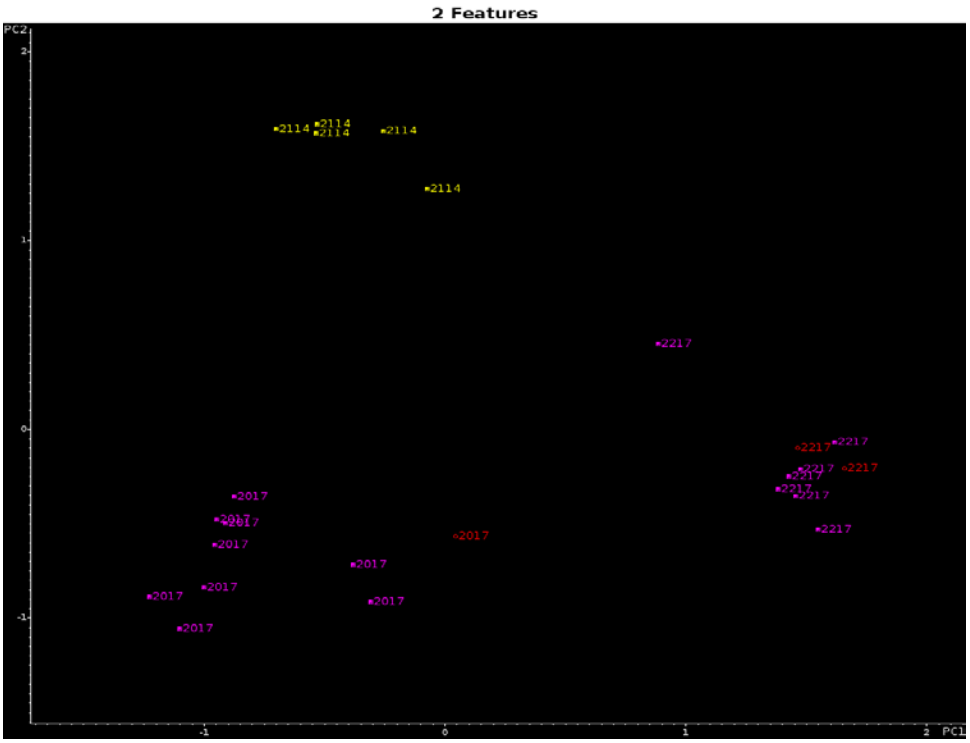

Figure 62. Validation set samples (in red or indicated by arrows) are projected onto the PC plot of the data defined by the 21 paint samples comprising Plant Group 24 (training set) and the 2 wavelet coefficients identified by the pattern recognition GA. 2017 = Wixom (subplant), 2114 = St. Thomas-Talbotsville (subplant), 2217 = Wixom (subplant).

The results of the Ford search prefilter development study suggest that Ford vehicles could pose challenges in the forensic examination of automotive paints as most assembly plants or subplants could not be differentiated. Ford automotive paint samples from the Atlanta, Chicago, Edison, and Kentucky Truck assembly plants and the Norfolk, Oakville, Twin Cities-St. Paul, and Wayne subplants produced automotive paint systems similar in both the top coat and under coat layers as reflected by their IR spectra. Furthermore, the Flat Rock and Hermosillo assembly plants and the St. Louis subplant produced automotive paint systems also similar in both top coat and undercoat layers. Only the Kansas City and Louisville assembly plants as well as the Dearborn, St. Thomas-Talbotsville, and Wixom subplants could be reliably differentiated by the search prefilters. This does not augur well for IR library searches involving Ford paint samples as there could be a high probability of obtaining a large number of spurious hits that a scientist must work through and eliminate thereby impairing the accuracy of the search. Nevertheless, the cross correlation library search algorithm was shown to be effective for these matches.

Library searching of each validation set sample was performed using the IR spectra of the automotive vehicles assembled in the manufacturing plant identified by the Ford search prefilters. The spectral library consisted of the manufacturer's paint system for 1182 automotive Chrysler, General Motors, and Ford automotive vehicles within a limited production year range (2000-2006). To extract information about the model from the IR spectrum, library searching of each validation set sample was performed using the IR spectra of the automotive vehicles assembled in the assembly plant(s) identified by the search prefilters. Three validation set samples were deleted from this analysis because the corresponding make and model of the corresponding paint sample was not present in the library. Although several spectral intervals were investigated during this phase of the study (including the fingerprint region and the fingerprint region and the carbonyl band), the tabulated results shown here are for the region from 3675 cm^{-1} to 2856 cm^{-1} and from 1891 cm^{-1} to 668 cm^{-1}. Additionally, the fingerprint region (1650 cm^{-1} to 668 cm^{-1}) was doubled in weight relative to the other two spectral intervals used.

The library search results for both schemes are summarized in Table 22 for 38 validation set samples. A histogram of the results with the top 5 matches for each comparison was computed to find the closest matching sample. For Scheme 1 this histogram was then weighted based on the average similarity index across all windows and intervals for each match candidate. The cross correlation library searching method was then compared with the top 5 matches identified by OMNIC, which used all 1182 spectra for library matching. Search results for both schemes and OMNIC are summarized in Table 22. The prototype pattern recognition software system outperformed OMNIC (Thermo Nicolet) in all cases and the results were comparable to General Motors whose search prefilters performed better than those of Ford.

There was one paint sample (UFLO00432) that was always misclassified by both the prototype pattern recognition system (both schemes) and OMNIC. For UFLO00432, there were other models and lines that were shown to be better matches. Two samples (UAZP00343 and UKYF00141) were always misclassified by OMNIC, but could be correctly classified by the prototype pattern recognition system. There were paint samples (UAZP00354, UAZP00474, and UOHL00133) that were usually missed by both OMNIC and the prototype pattern recognition system. For UAZP00354, the correct model was a poor match, whereas for UOHL00133, the

correct model was a better match but not an optimal match. For UAZP00474, the other models and lines were better matches.

Table 22. Library Search Results for Ford

Layer	[1]OMNIC Searching	Prototype System (Scheme 1)	Prototype System (Scheme 2)
Clear Coat	24	32	30
Surfacer	24	32	28
Primer	27	29	29
[2]Fused (Addition)	23	31	29

[1]OMNIC searching was done using its built-in library searching. The diamond region, 2856 cm^{-1} to 1891 cm^{-1} was excluded from the search. The results of OMNIC were considered correct if there was a match within the top 5 hits found by OMNIC's search algorithm as ranked by hit quality index.

[2]Fusion of the layers was achieved by taking spectra (in absorbance mode) of the clear coat and undercoat paint layers and directly adding them to yield a single spectrum which was then converted to the transmittance mode for analysis by the cross correlation library search algorithm.

All spectra incorrectly matched by Scheme 1 were also incorrectly matched by OMNIC. The sample corresponding to the sole clear coat IR spectrum missed was also missed in the surfacer and primer layer searches by the cross correlation library search algorithm. However, the additional paint sample missed in the surfacer layer search was correctly matched in the primer layer search. With one exception, all paint samples incorrectly matched by Scheme 1 were also incorrectly matched by Scheme 2. This exception was caused by two populations that were hit equally often. One population contained the correct model identifier, while the other had the correct line identifier. Because of this, the probabilistic system identified a fair chance of there being a correct match for the correct model and line in the top 5, when in reality is was observing the overlap between similar populations.

The top hit selected by Scheme 1 always yielded a reasonable match, usually superior to the match between the validation set sample and the actual model for incorrectly matched spectra (see Figures 63 through 65). This was true for all incorrectly matched samples. Although Scheme 2 did not perform as well as Scheme 1, it has the advantage of providing insight into how well the library matches the blinds, rather than how well an individual blind matches the library. This suggests that samples assigned the same model and line by both schemes are well represented in the library and correlate well on an individual basis to specific samples in the library. For these samples, one can have confidence in the quality of the match and high certitude regarding the accuracy of the match. A similar statement cannot be made regarding the accuracy and quality of individual matches by OMNIC as reflected by their hit quality index which are typically 98%-99% for the top 20 hits.

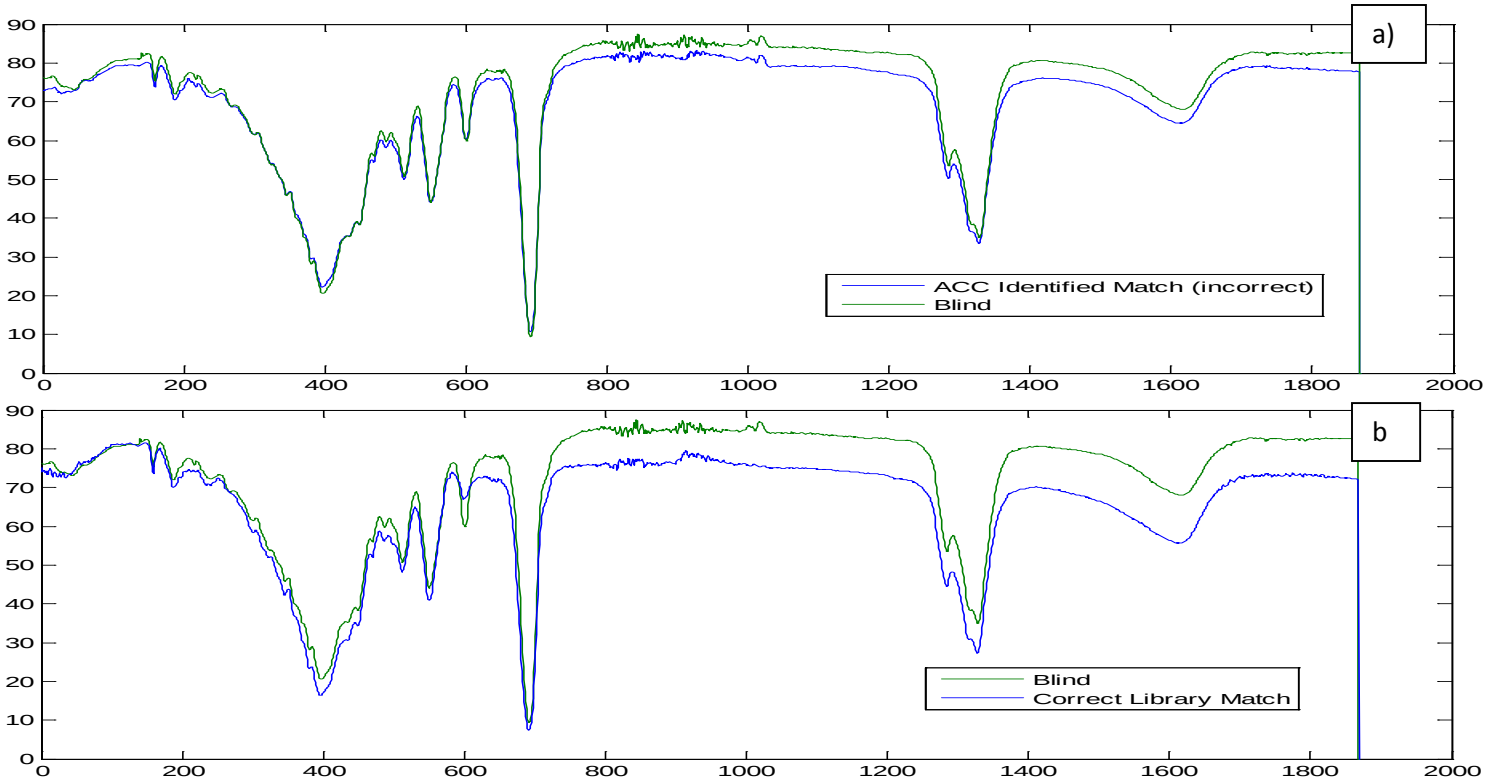

Figure 63. Ford GM clear coat spectral matches: a) top hit and b) correct match

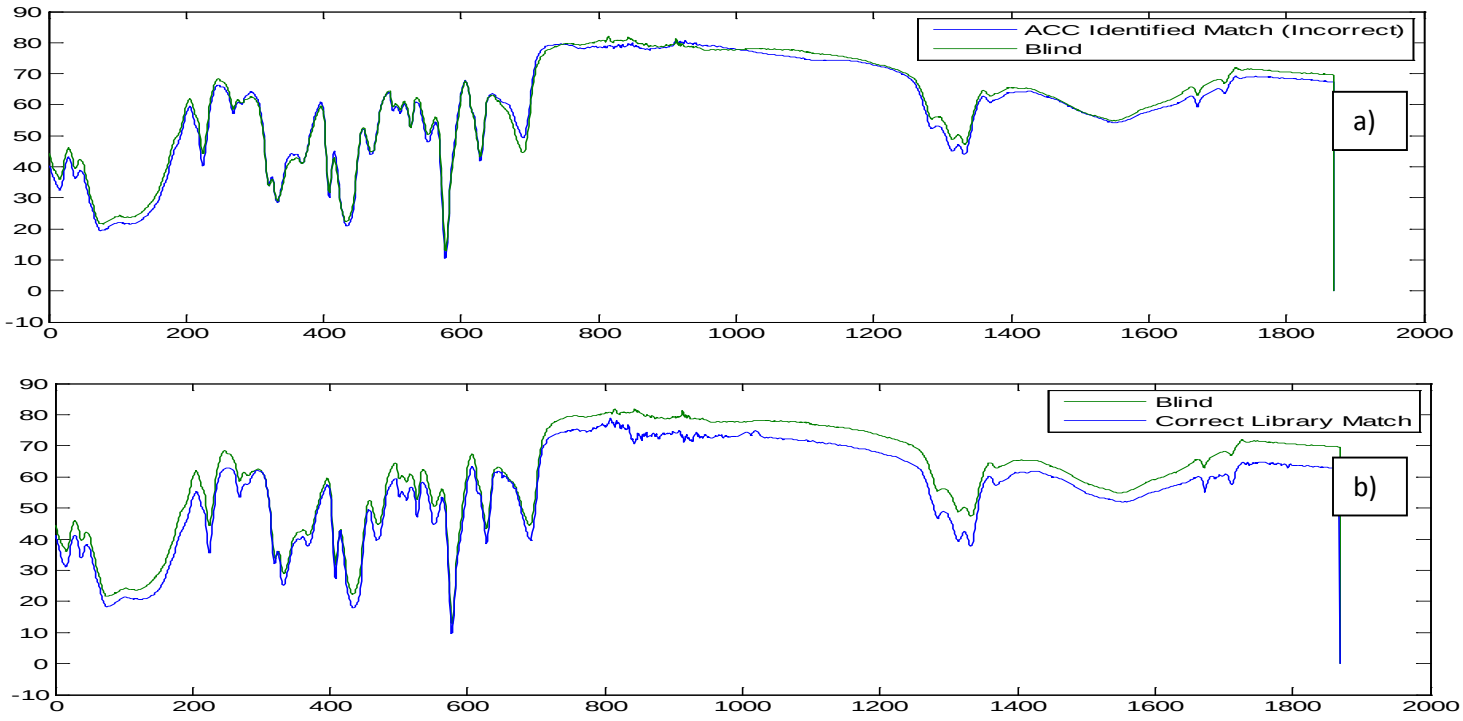

Figure 64. Ford surfacer spectral matches: a) top hit and b) correct match

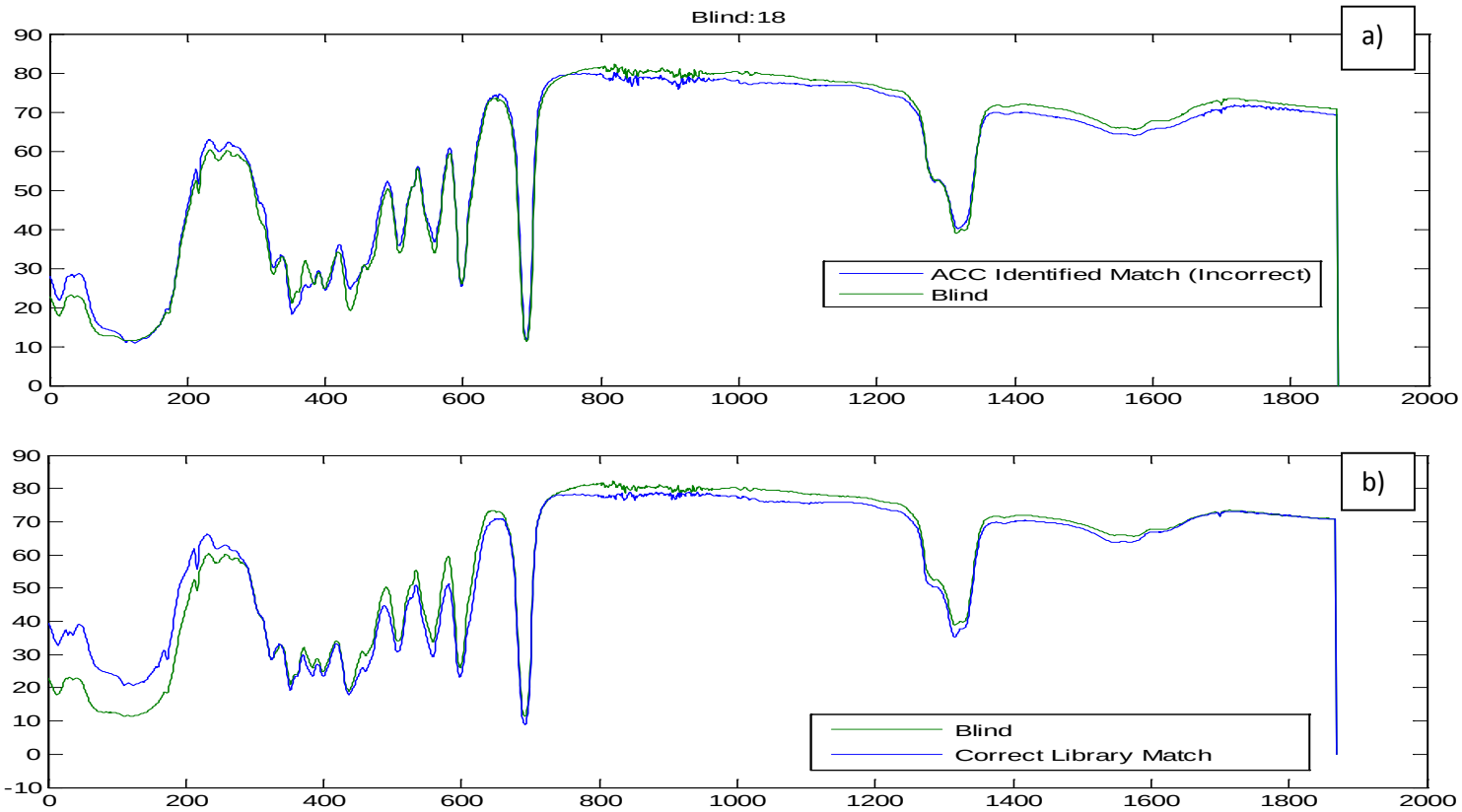

Figure 65: Ford Primer spectral matches: a) top hit and b) correct match

Simulation of Attenuated Total Reflection Infrared Absorbance Spectra

A correction algorithm to allow ATR spectra to be searched using IR transmission spectra of the paint data query (PDQ) automotive database is presented. The proposed correction algorithm to convert IR transmission spectra from the PDQ library to ATR spectra is able to address distortion issues such as the relative intensities and broadening of the bands, and introduction of wavelength shifts at lower frequencies, which prevent library searching of ATR spectra using archived IR transmission data. Conversion of an IR transmission spectrum to an ATR spectrum was performed by taking advantage of a surface reflection phenomenon at the boundary between the sample and the internal reflection element (IRE) of the spectrometer. In this procedure, the reflection of the incident beam from the IRE is described at the boundary with the sample by Fresnel's equations. In order to simulate an ATR spectrum the following data are required: optical constants, n- and k-indices of the sample, refractive index of IRE, sample thickness, incident angle of the beam and the number of internal reflections. By assigning values to these parameters, the reflectance of s- and p- polarized light can be calculated from Fresnel's equations. ATR spectra of p- and s-polarized incident light obtained are averaged to give the ATR spectrum with unpolarized light.

The optical constants of the paint sample were calculated from the transmission spectrum of the sample in absorbance units when the sample thickness is known. The process consists of two steps. First, the k-index is obtained using Equation 10.

$$k(\nu) = \frac{2.303A(\nu)\lambda}{4\pi d} \quad or\ k(\nu) = 2.303A(\nu)/4\pi\nu d \tag{10}$$

A(ν) is the transmission spectrum in absorbance units as a function of wavenumber, ν, λ (which equals 1/ν) is the wavelength of the incident beam, and d is the sample thickness, respectively. As two components, both real and imaginary, of a complex function such as the refractive index, $\tilde{n} = n + ik$, are mutually related through the Kramers-Kronig relationship, once the k-index has been obtained, the n-index is calculated by Kramers-Kronig integration of the k-index with the refractive index at high wavenumbers used as the anchor value, which is summarized in Equation 11.

$$n(\nu_a) = n(\infty) + \frac{2}{\pi} P \int_0^\infty \frac{\nu k(\nu)}{\nu^2 - \nu_a^2} d\nu \tag{11}$$

P denotes the principal value and n(∞) is defined as the index at high wavenumber, where there is no absorption. In our calculation, the double Fourier method was employed to perform calculations using Equation 11. Using Equation 11, the complex refractive index of the sample can be obtained.

If IR spectra that have been simulated using n(∞) values of 1.45, 1.50 or 1.55 are overlaid with the remaining parameter set unchanged, only small changes in the peak intensities of weak peaks which appear as shoulders of major (strong) peaks are observed. The peak positions of major peaks, for example the C=O stretching vibration, change only 0.7 cm^{-1} from 1.45 to 1.55. Hence, a spectral library search will not be influenced by this small discrepancy. Absolute values of the

peak intensities of major peaks such as C=O stretching vibration change approximately 20% from 1.45 to 1.55. Although this is not a small change, these changes can be easily compensated for by adjusting the film thickness of the transmission sample. Instead of an *ad hoc* selection of n(∞) and thickness for each sample, we have selected a self-consistent parameter set of 47° incident angle, n(∞) = 1.50, and sample thickness of 4.5 micrometers for these clear coat samples.

The set of n- and k- indices defines Fresnel's reflection coefficients for the p- and s-polarized beam (see Equation 12 and Equation 13).

$$r_s = \frac{n_0 \cos(\theta_0) - \tilde{n}_1 \cos(\theta_1)}{n_0 \cos(\theta_0) + \tilde{n}_1 \cos(\theta_1)} \tag{12}$$

$$r_p = \frac{\tilde{n}_1 \cos(\theta_0) - n_0 \cos(\theta_1)}{\tilde{n}_1 \cos(\theta_0) + n_0 \cos(\theta_1)} \tag{13}$$

Here, n_0 and $\tilde{n}_1$ are the refractive indices of the IRE and the sample. θ_0 and θ_1 are the angles of the beam in the IRE and in the sample respectively. Reflectance values for the ATR spectrum are given by Equation 14.

$$R = \frac{(R_p + R_s)}{2} = (\left|r_p\right|^2 + \left|r_s\right|^2)/2 \tag{14}$$

Using Equations 10 and 11, the k-index and n-index are calculated from the transmission spectrum. These values as well as the refractive index of the IRE are placed into Equations 12 and 13 to yield Fresnel's reflection coefficients. These coefficients are then used to compute the reflectance values of the corresponding ATR spectrum via Equation 14. Further details about Kramers-Kronig integration and the calculations of Fresnel's coefficients can be found elsewhere [30-32].

A program to convert an IR transmission spectrum to the corresponding ATR spectrum was written in Matlab (MathWorks, Nautick, MA U.S.A.) Since the double Fourier transform technique was used to implement Equation 11, the spectral data size was adjusted to 2^n points in excess of the original data length (analogous to zero filling) to the lower and higher frequency sides. In these calculations, the thickness of the paint sample is required. IR spectra in the PDQ library were collected using thin microtomed films of the surface materials of the individual paint layers compressed between diamond-windows at 40 psi with the film thickness estimated to be below 10 micrometers thick. To estimate the film thickness, 13 transmission spectra were selected from the PDQ library and ATR spectra of these same samples were measured using the built-in diamond single-reflection ATR accessory of the Nicolet iS50 FTIR spectrometer and compared with calculated ATR spectra for these samples using a sample thickness of 3 to 10 micrometers for every 1 micrometer increment. In all cases, the best match in both ordinate intensities and peak positions was observed when the sample thickness was assumed to be 4 to 5 micrometers. For this reason, a thickness of 4.5 micrometers for the sample film was used throughout our simulations.

In these simulation studies, ATR analysis was limited to the clear coat paint layer. Because the penetration depth of an ATR analysis is quite shallow and is considerably less than the thickness of an automotive clear coat, this layer can be preferentially analyzed directly on intact paint chips with little or no sample preparation. Use of an ATR objective on an infrared microscope is also the method of choice for the analysis of very thin automotive paint smears. Such smears may consist primarily of the clear coat paint layer transferred onto the impacted vehicle or other object such as a pedestrian. The Forensic Laboratory of the Royal Canadian Mounted Police who utilize high pressure diamond cells to acquire transmission IR spectra of automotive paint layers for the PDQ database have estimated that the thickness of the clear coat layer, once it is pressed between the two diamonds, as 5 micrometers.

For this program, it was also necessary to provide the incident angle of the ATR accessory in the equations used to compute the ATR spectrum. The incident angle of the iS50 ATR accessory was estimated using a simulation. As explained above, it is possible to simulate ATR spectrum when the optical constants of the sample (e.g., refractive index) are known. We have selected the optical constants of toluene published by Jones and coworkers [33]. The best match for the ATR spectrum of toluene measured on the iS50 built-in ATR accessory at 1 cm^{-1} spectral resolution was determined when the incident angle is changed from 43° to 50°. For medium strength (~0.53 Abs.) absorption such as the absorption band at 463 cm^{-1}, the best match in intensity was achieved using an incident angle of 47°, while for the strong band (~0.96 Abs.) at 725 cm^{-1}, 49° gave the best match in intensity. In this study, we have selected 47° for the effective incident angle of the built-in ATR accessory for our clear coat paint spectra as it is best for weak to medium strength absorptions. All transformations of transmission spectra to corresponding ATR spectra were performed with the following input data to the above-mentioned MATLAB codes: refractive index of diamond is 2.38, the refractive index at high wavenumber where there is no absorption of the sample n (∞), is 1.50, and the incident angle (relative to the normal) in the IRE is 47°. The value of 1.50 was searched from the list of refractive indices corresponding to the Sodium D-line. The clear coats used in this study are composed of acrylate polymer with modification by polyurethane and polystyrene. The refractive indices of these compounds are near 1.50, and this value was employed for n(∞) throughout our entire study.

Figure 66 shows a plot of the calculated versus measured ATR spectrum of a clear coat paint sample designated as UAZP00503 in the PDQ database. The carbonyl peak, which is a doublet at 1730 cm^{-1}, is accurately represented in the calculated ATR spectrum. Minor peaks and shoulders in the measured ATR spectrum are also accurately represented in the simulation. The noise in the measured ATR spectrum from 2400 cm^{-1} to 2000 cm^{-1} is caused by absorption bands of diamond in this region.

The incident angle used in this and in our other simulations is 47°. If the incident angle of the ATR accessory is 45°, the signal due to the p-polarized beam would be twice that of the s-polarized beam. As 47° incidence refers to a system that has more s-polarized beam than p-polarized beam, the signal of the absorption intensity is weaker than expected. This is evident when two spectra of stretched (oriented) polyethylene collected at two different orientations, each 90° apart, are compared (see Figure 67). The ratio of the intensity of the doublet that corresponds to the bending vibration is reversed in these two spectra. As these two spectra are different, one can only conclude that residual polarization exists with this spectrometer equipped with this ATR system. By

comparison, there is no difference in ATR spectra of unstretched polyethylene when it is oriented in these two perpendicular directions. For this reason, it is necessary to use a lower sample thickness in the calculations, which corresponds to a higher k value and therefore higher dispersion amplitude of the n index. The values specified in these simulations were self-consistent as they described the salient features of this system specified by equations 10 through 14.

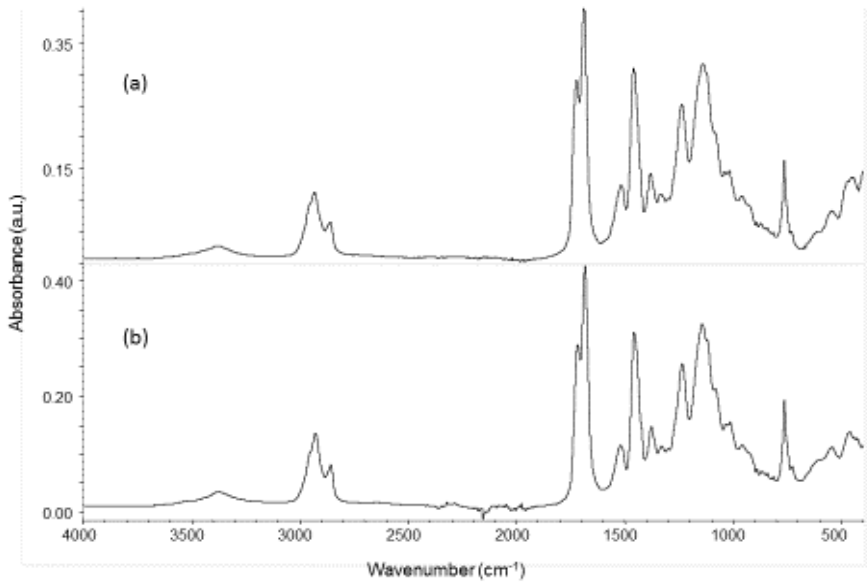

Figure 66. ATR spectrum of the (a) simulated and (b) measured ATR spectrum of the clear coat paint sample designated as UAZP00503 in the PDQ database. The ATR spectrum was measured at 4 cm^{-1} resolution.

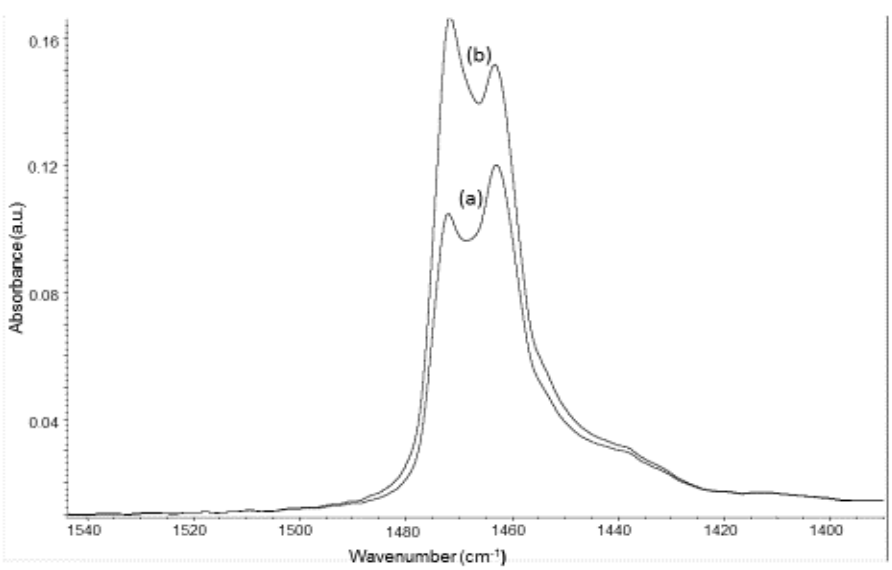

Figure 67. Spectra of stretched (oriented) polyethylene at two different orientations, each 90^0 apart: a) drawn axis of the sample is oriented parallel to the ridge of the diamond prism, b) drawn axis of the sample is oriented perpendicular to the ridge of the diamond prism.

One problem that impacted the quality of these simulations is the absence of any information about the angle of incidence from the manufacturer. If the incident angle is smaller, the absorption peak will be more intense but there will be more tailing towards lower frequency. If the incident angle is larger, the absorption peak will be less intense but it will be more symmetric. A second problem is the polarization of the entire optical system of the spectrometer used in this study including the interferometer and the ATR accessory. Our simulation assumes s- and p- polarization of 50:50. If the polarization deviates from 50:50, then the polarization must be measured. To make this measurement, it is necessary to insert a polarizer into the optical system, which is not an easy task. For this simulation study, we used 50:50 polarization in the calculations. In the future, we hope to refine this calculation by measuring the residual polarization of our spectrometer.

Although both the input incident angle and the sample refractive index at high wave-numbers must be provided by the user for these simulations, the same is true in the case of the ATR-to-Transmission correction method, as this routine also requires the input of the incident angle and the refractive index at high wavenumbers. For the ATR-to-Transmission correction method, data acquired with imperfect sample contact with the IRE and the resultant scattering of the evanescent wave at the boundary between the sample and the IRE can cause numerical overflow/underflow problems, because the offset from the baseline due to insufficient contact is interpreted as strong absorption, namely a large k-index in a wide wavenumber range, which in turn generates a large dispersion of the n-index in the same wavenumber range through Kramers-Kronig integrations. When such a problem occurs, the corresponding transmission spectrum may show, for examples, a straight line with a high absorbance value, or large ringing around strong peaks, or negative absorbance with opposite phase. Using the present method, the spectrum taken with insufficient sample contact with the IRE will at least show a spectrum if the intensity is weak or it will show a bent baseline. Even with such a poor quality spectrum, a library search is plausible using the present method.

During the course of this investigation, we discovered that it is not always possible to calculate an ATR spectrum of a sample from its corresponding transmission spectrum due to the presence of contaminants on the surface of some clear coat paint samples. ATR analysis is restricted to the surface of the sample and a region of a few micrometers below the surface of the sample. Transmission measurements, on the other hand, interrogate all regions of the sample along the path of the IR beam and are relatively insensitive to the presence of contaminants on the surface as these compounds would comprise a very small fraction of the sample. Figure 68 shows an ATR spectrum of the clear coat paint sample designated as ULH00050 in the PDQ database before and after washing with methanol. The peak between 1650 cm^{-1} and 1600 cm^{-1} corresponds to water, whereas the peak between 1050 cm^{-1} and 1000 cm^{-1} indicates the presence of silicates in the sample. These peaks are not present in the IR transmission spectrum of this sample. Washing the sample with methanol decreases the amount of water and silicates in the sample as reflected by the decrease in the intensities of these absorption bands. If water is trapped in the silicates, decreasing the amount of silicates in the sample by washing would also decrease the amount of water present in the clear coat paint sample.

Figure 69 shows the difference spectrum formed by subtracting the ATR spectrum of ULH00050 before and after sample washing. An examination of the difference spectrum reveals the presence of a water band at 1650 cm^{-1} and a silicate peak at 1000 cm^{-1}. Library matching of the difference

spectrum using OMNIC revealed the presence of bentonite (see Figure 70), which is a type of clay that contains alumino-silicates. Many paint samples in the PDQ library were obtained from salvage yards, and the presence of these contaminants could be due to prolonged exposure of the paint layers in the environment.

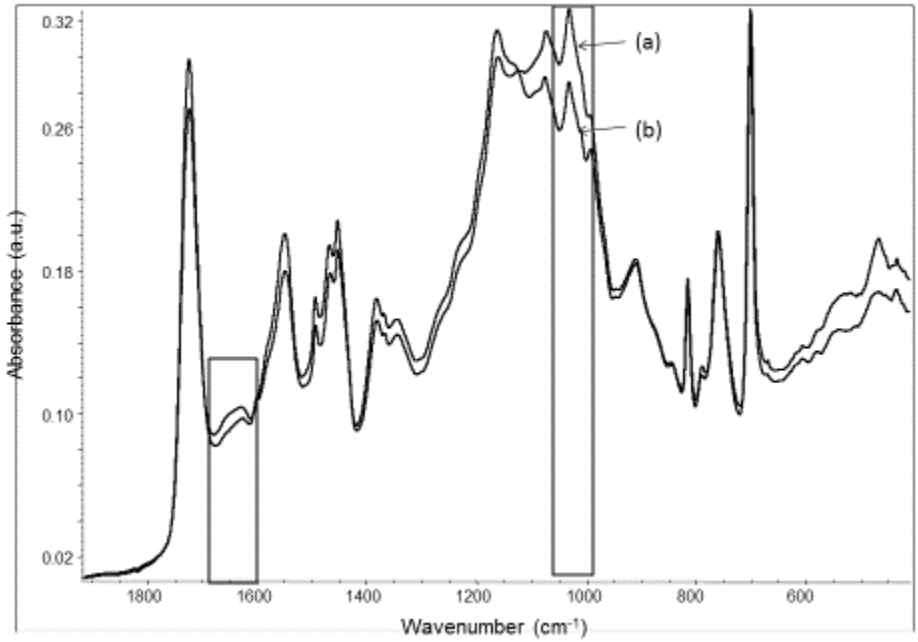

Figure 68. ATR spectra of ULH00050 (a) before and (b) after washing by methanol

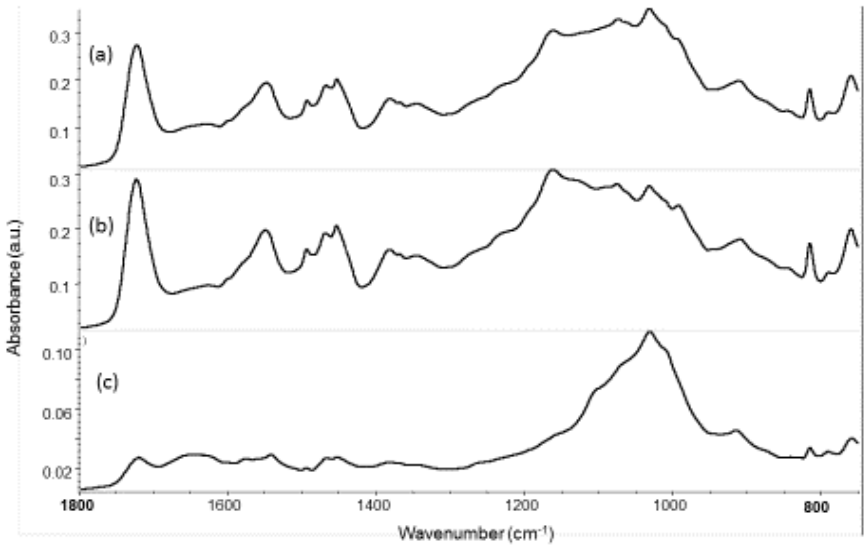

Figure 69. ATR spectra of ULH00050: a) before washing, b) after washing, and c) difference spectrum

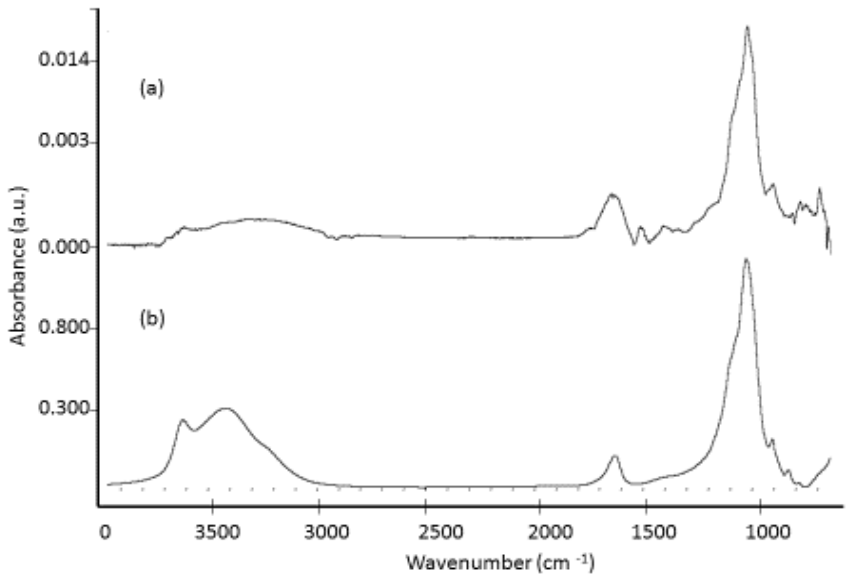

Figure 70. (a) Difference spectrum before and after washing and (b) library spectrum of bentonite selected by OMNIC as a match for the difference spectrum.

ATR spectra of ULH00050 were collected using a germanium IRE. However, these surface effects will also be present when a diamond IRE is used. Figure 71 shows the changes in the ATR spectrum of sample UOHL00050 after washing with methanol using a diamond IRE. Figure 72 shows the changes in the ATR spectrum of the same sample after washing with methanol using a germanium IRE. Although these surface contaminants are present in spectra collected using both diamond and germanium, they are more pronounced when germanium is used. This proves that the contaminant is near the surface as the penetration depth of the evanescent wave through germanium and the sample is less than it would be with diamond due to the higher refractive index of germanium.

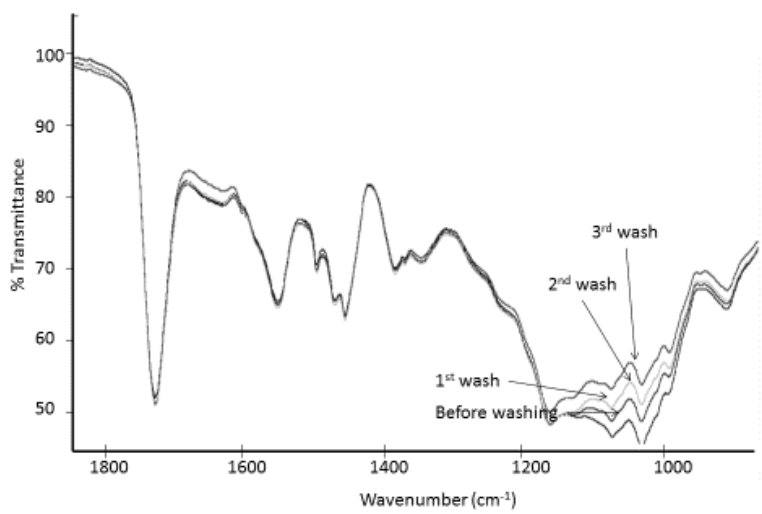

Figure 71. Effect of methanol washing on ATR spectra of UOHL00050 clear coat paint sample on diamond IRE

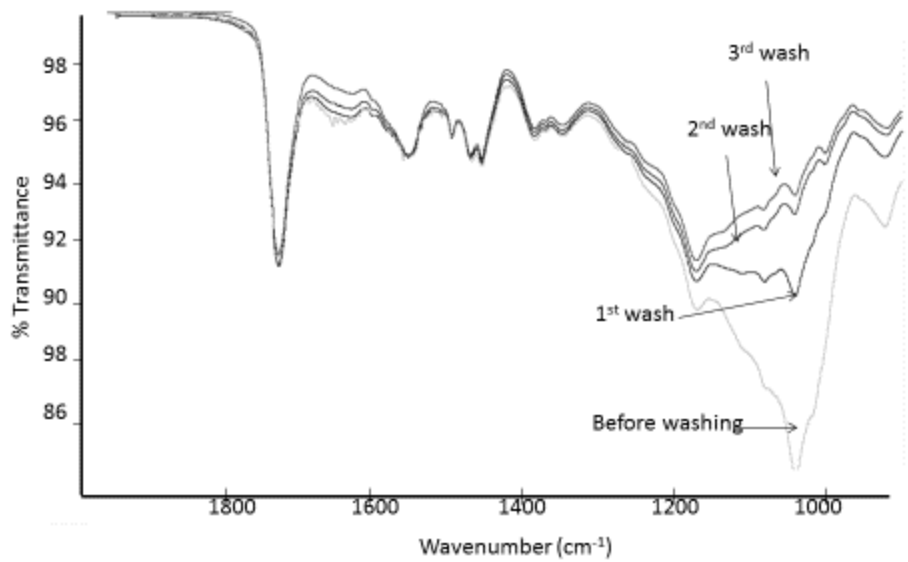

Figure 72. Effect of methanol washing on ATR spectra of UOHL00050 clear coat paint sample on germanium IRE

To further assess the efficacy of the ATR correction algorithm, search prefilters were developed from simulated ATR spectra for the purpose of identifying the assembly plant of a vehicle from an ATR spectrum of its clear coat paint smear. 456 transmission spectra from the PDQ library spanning 22 GM assembly plants (see Table 23) were transformed into ATR spectra using the ATR simulation algorithm. The spectral region used to formulate the search prefilters was the fingerprint region (1500 cm^{-1} to 600 cm^{-1}). In this study, both the simulated and experimental ATR spectra were preprocessed using the discrete wavelet transform, which increased the signal to noise of the data by concentrating the signal in specific wavelet coefficients. Using the pattern recognition GA, wavelet coefficients characteristic of the assembly plant of the vehicle were identified. In this study, the Symlet 6 mother wavelet at the 8th level of decomposition (8Sym6) was used to denoise and to resolve overlapping spectral responses in each ATR spectrum. Prior to pattern recognition analysis, the wavelet coefficients for each IR spectrum were organized as a data vector. Each wavelet coefficient was autoscaled to ensure that it had a mean of zero and a standard deviation of one. Auto-scaling removed inadvertent weighting of the wavelet coefficients that otherwise would occur due to differences in magnitude among the coefficients. This ensures that each wavelet coefficient had an equal weight in the pattern recognition analysis of the data.

Search prefilters were developed using the same hierarchical classification scheme from a previous study [reference] involving IR transmission data for GM assembly plants. First, an unknown was classified as to its plant group and then a search prefilter was used to identify a specific assembly plant or assembly plants within the plant group to which membership of the unknown was assigned. The assembly plants comprising each plant group are listed in Table 24. In the previous study, the IR spectra were initially divided into two categories based on the carbonyl band at 1709 cm^{-1}. In one category, the carbonyl band in each transmission spectrum was a singlet (Plant Groups

1, 3, and 5) which corresponds to acrylic melamine styrene and in the other category the carbonyl band was a doublet (Plant Groups 2 and 4) which corresponds to acrylic melamine styrene polyurethane. An examination of the expanded fingerprint region revealed five distinct spectral patterns with each pattern corresponding to a specific plant group.

Table 23. GM Plants used to develop the Search Prefilters

Plant ID	Plant	Model	Line
1	Arlington	CAD, CHE, GMC	SUB,YUK,ESD,CTA
3	Bowling Green	CAD,CHE	CVT,XLR
4	Doraville	PON	VTR,SIL,MTA,UPL,TAR
5	Fairfax	CHE,OLD,PON	GRA,MAL,ITR
6	Flint	CHE,GMC	SLV,SIE
8	Fort Wayne	CHE,GMC	SLV,SIE
9	Fremont	GMC	VIB,TAC,PVB,COA,GPR
10	Hamtramck	BUI,CAD,PON	BON,DEV,LUC,LES,SEV,ELD
11	Ingersoll	CHE, PON	EQU, MGM, TRA, TOR
12	Janesville	GMC	CTA,SUB,YUK
14	Lansing	PON	STS
16	Linden	CHE,GMC	BZR,JMY,S10
17	Lordstown	PON	SFR,CAV,COB,PST
18	Moraine	CHE,GMC,SAA	JMY,ENV,9S7,BZR,TBZ,SON
20	Oklahoma City	CHE, GMC	MAL,TBZ,ENV,EQU, XUV
21	Orion	PON,BUI	BON,PG6,LES,AUR, PKA
22	Oshawa	GMC,PON	ALL,REG
23	Pontiac	CHE,GMC	SLV,SIE,SIL
24	Ramos Arizpe	BUI,CHE,PON	CAV,SFR,RZV,AZT,HHR
25	Shreveport	CHE,GMC	S10,COL,SON
26	Silao	CHE,GMC,SAA	AVL,SUB,YXL
27	Spring Hill	STR	SSL,ION,SC1,SC2,SL1,VUE

Table 24. Assembly Plants Comprising Each Plant Group

Plant Group	Plant ID Number	Assembly Plant
1	1, 4, 5, 8, 14, 18, 23	Arlington, Doraville, Fairfax, Lansing, Moraine, Pontiac
2	3, 10, 21	Bowling Green, Hamtramck, Orion
3	6, 9, 11, 16, 17, 20, 22, 25	Flint, Fremont, Linden, Lordstown, Oklahoma City, Oshawa, Shreveport
4	12	Janesville
5	24, 26, 27	Ramos Arizpe, Silao, Spring Hill

The first step in this study was to apply principal component analysis (PCA) to the 456 wavelet transformed ATR spectra. Figure 73 shows a principal component (PC) plot of the 456 ATR spectra and the 1178 wavelet coefficients representing each simulated ATR spectrum. Each sample is represented as a point in the PC plot. The overlap of several plant groups in the PC plot of the data is evident.

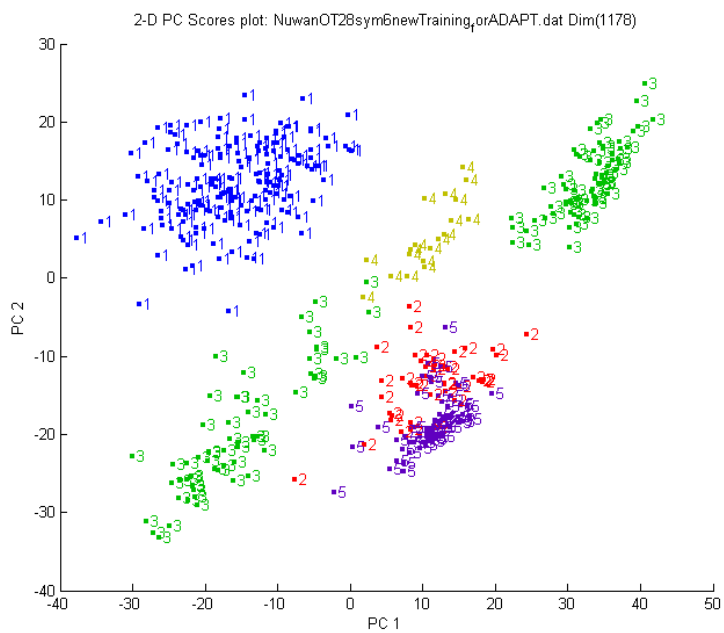

Figure 73. PC plot of the two largest principal components of the 456 ATR spectra and the 1178 wavelet coefficients comprising the training set. Each clear coat paint sample is represented as a point in the PC plot of the data (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

The next step was feature selection. The pattern recognition GA identified wavelet coefficients characteristic of the plant group by sampling key feature subsets, scoring their PC plots and tracking those plant groups and/or spectra that were most 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 27 wavelet coefficients whose PC plot (see Figure 74) showed clustering of the spectra on the basis of plant group.

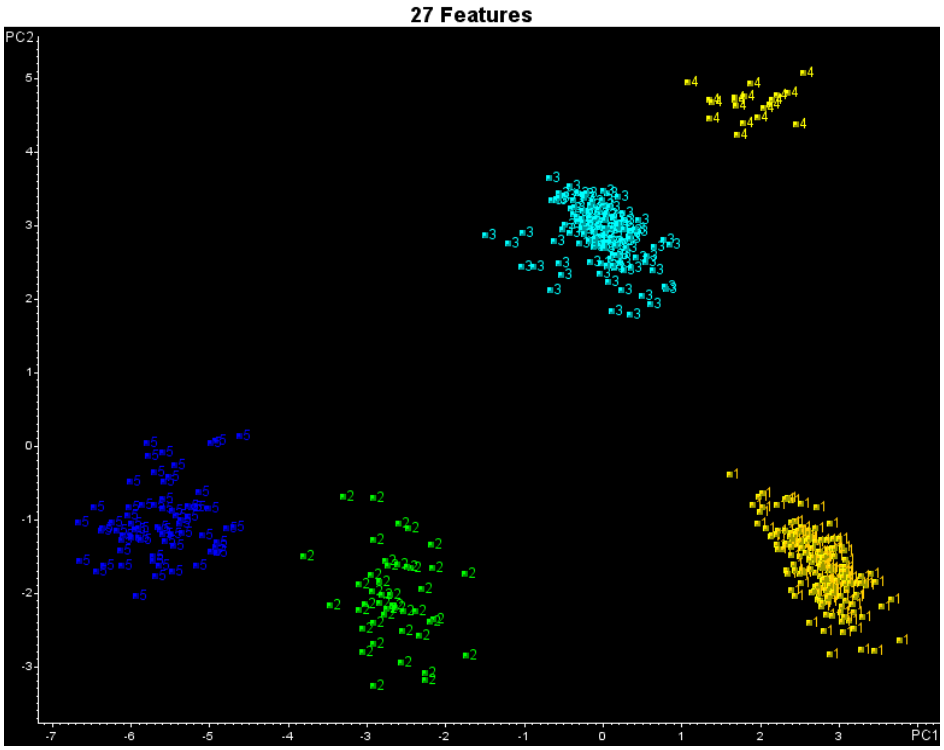

Figure 74. PC plot of the two largest principal components of the 456 ATR training set spectra and the 27 wavelet coefficients identified by the pattern recognition GA. Each clear coat paint sample is represented as a point in the PC plot of the data (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

To assess the predictive ability of these 27 wavelet coefficients, a validation set of 14 clear coat paint samples whose ATR spectra were obtained using a Nicolet iS50 FTIR spectrometer served as the validation set. ATR spectra from the validations set were projected onto the PC plot developed from the 456 simulated ATR spectra of the training set and the 27 wavelet coefficients identified by GA using the transductive learning routine of the pattern recognition GA [34]. Figure 75 shows the projection of the 14 validation set samples onto the PC plot of the training set data. All 14 clear coats were correctly classified.

The next step was to develop search prefilters to identify the spectra of the validation set samples by assembly plant. For each plant group, a search prefilter was developed to discriminate spectra by assembly plant within a plant group. Figure 76 shows a PC plot of the two largest principal components of 29 wavelet coefficients identified by the pattern recognition GA for Plant Group 1 (see Table 24). Each simulated spectrum is represented as a point in the PC plot of the data. Plant 18 (Moraine, OH) is well separated from the other assembly plants in the PC plot. The spectra from the other 6 assembly plants (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI) were similar, which prevented further discrimination by assembly

plant of these clear coats. Projecting the validation set samples assigned to Plant Group 1 onto this PC plot via transverse learning showed that each projected sample was located in a region of the map with paint samples of the same class label: either plant 18 or plants 1, 4, 5, 8, 14, and 23.

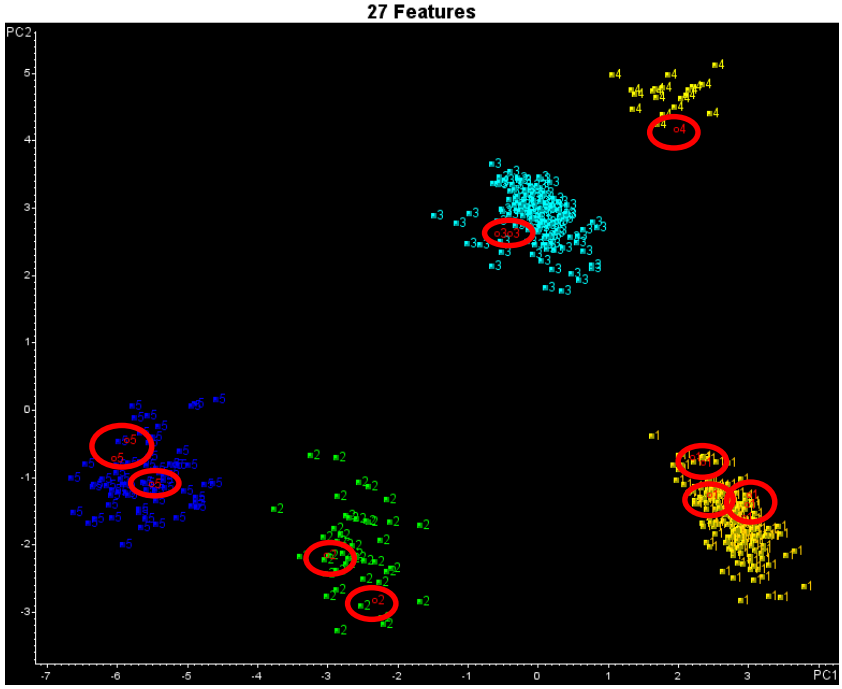

Figure 75. PC plot of the two largest principal components of the 456 ATR training set spectra and the 27 wavelet coefficients identified by the pattern recognition GA. Each clear coat paint sample is represented as a point in the PC plot of the data. Validation set samples are circled. (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

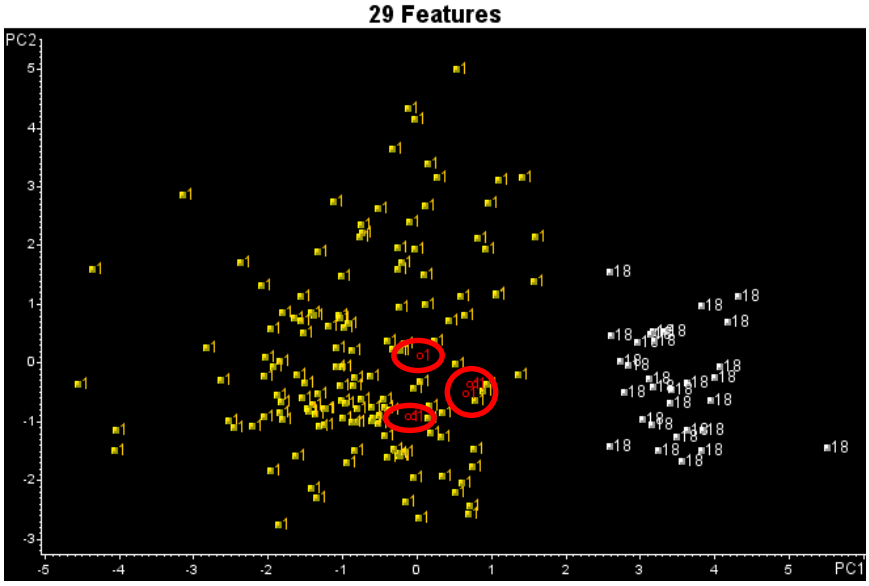

Figure 76. PC plot of the two largest principal components of the 180 Plant Group 1 simulated spectra and the 27 wavelet coefficients identified by the pattern recognition GA. Each simulated or experimental ATR spectrum is represented as a point in the PC plot of the data. Validation set samples are in red and circled. (1 = Arlington, Doraville, Fairfax, Fort Wayne, Lansing, Pontiac, and 18 = Moraine).

Figure 77 shows a plot of the two largest principal components of the 52 simulated ATR spectra and the 28 wavelet coefficients identified by the pattern recognition GA for the assembly plants comprising Plant Group 2. All 3 assembly plants (Bowling Green KY, Hamtramck MI, and Orion MI) are well separated from each other in the PC plot of the data. The two validation set samples assigned to Plant Group 2 also projected onto this plot lie in a region of the PC map that contains simulated spectra with the same class label.

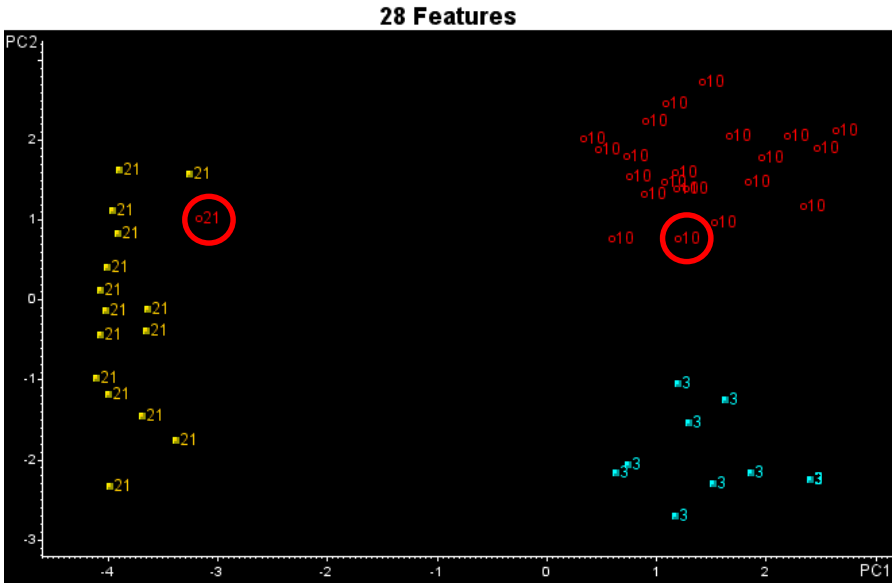

Figure 77. PC plot of the two largest principal components of the 52 Plant Group 2 simulated spectra and the 28 wavelet coefficients identified by the pattern recognition GA. Each simulated or experimental ATR spectrum is represented as a point in the PC plot of the data. Validation set samples are in red and circled. (3 = Bowling Green, 10 = Hamtramck, and 21 = Orion).

Figure 78 shows a plot of the two largest principal components of the 138 simulated ATR spectra from Plant Group 3 and the 33 wavelet coefficients identified by the pattern recognition GA. Clustering of the simulated spectra by assembly plant and model in the PC plot is evident for the spectra comprising the training set. Fremont California (Plant 9) and Lordstown Ohio (Plant 17) form distinct clusters in the PC plot as do the trucks from Oshawa (Plant 22). The two validation set samples assigned to Plant Group 3 are projected onto the PC plot in a region that contains simulated spectra from the same assembly plant.

Plant Group 4 contains only one assembly plant, and the three assembly plants from Plant Group 5 cannot be discriminated due to the similarity of their spectra. A summary of the results obtained for the 14 validation samples using the search prefilters developed from the simulated ATR spectra is shown in Table 3. For validation set samples 91001, 91013, 91027, and 92005, differences between the experimental ATR spectrum and the simulated ATR spectrum derived from the corresponding transmission spectrum of the same sample in the PDQ library (see Figures 79-82) can be attributed to weathering. The increase in the 1630 cm^{-1} band and the decrease and change in the spectral pattern for the 1085 cm^{-1} – 1030 cm^{-1} region can be attributed to the formation of primary amines and the loss of the C-O-C groups attached to the triazine ring of acrylic melamine

in acrylic melamine styrene [35-37]. By exposing a fresh surface of the clear coat automotive paint layer to the spectrometer (see Figure 83), interference due to weathering has been eliminated.

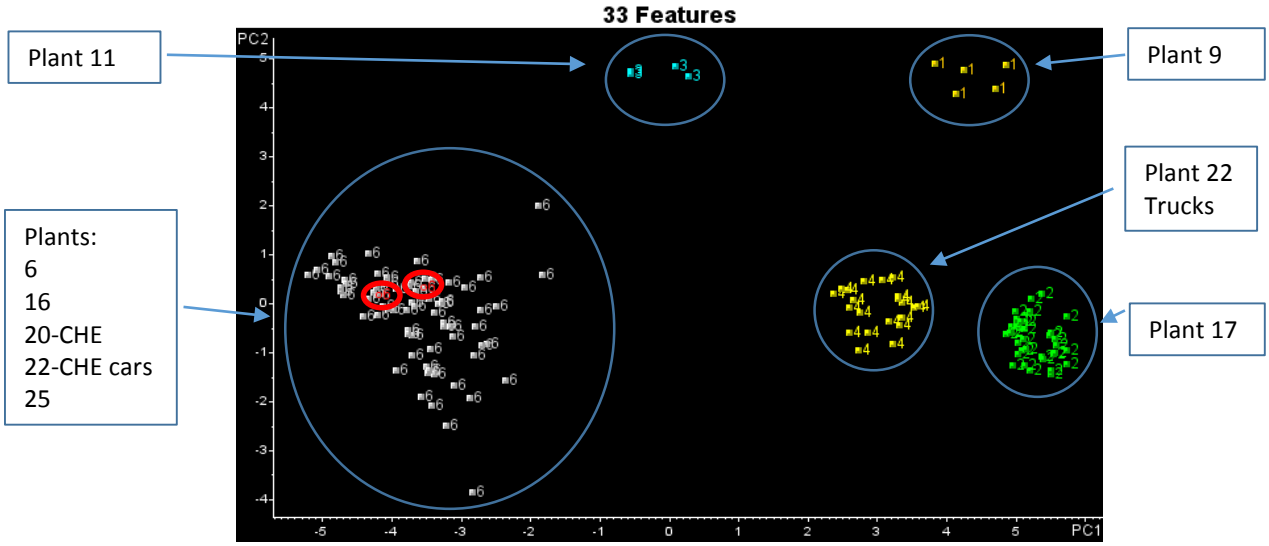

Figure 78. PC plot of the two largest principal components of 138 Plant Group 3 simulated ATR spectra and the 33 wavelet coefficients identified by the pattern recognition GA. Each simulated or experimental ATR spectrum is represented as a point in the PC plot of the data. Validation set samples are in red and circled. (1 = Fremont, 2 = Lordstown, 3 = Ingersoll, 4 = Oshawa, and 6 = Flint, Linden, Oklahoma City, Oshawa, Shreveport).

Table 25. Summary of Results for Validation Set Samples

Validation Sample	Assigned Plant Group	Assigned Plant(s)	ID of Validation Sample
91001	1	1,4,5,8,14, 22(Buick cars), 20(Trucks)	14
91011	5	24,26,27	26
91012	5	24,26,27	26
91013	1	1,4,5,8,14, 22(Buick cars), 20(Trucks)	14
91025	2	10	10
91027	1	1,4,5,8,14, 22(Buick cars), 20(Trucks)	14
91028	1	1,4,5,8,14, 22(Buick cars), 20(Trucks)	14
92001	1	1,4,5,8,14, 22(Buick cars), 20(Trucks)	1
92003	4	12	12
92004	3	6,16,20(CHE),22(CHE cars),25	6
92005	1	1,4,5,8,14, 22(Buick cars), 20(Trucks)	22
92006	5	24,26,27	24
92007	2	21	21
92008	3	6,16,20(CHE),22(CHE cars),25	25

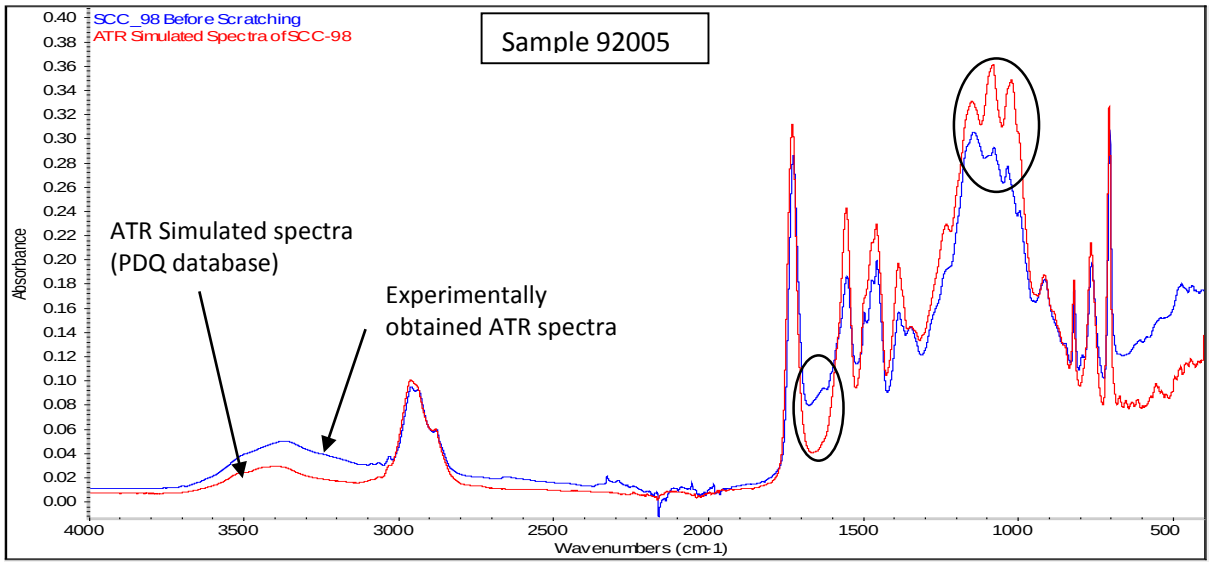

Figure 79. Simulated (red) and experimental (blue) ATR spectra for Sample 92005. For the experimental ATR spectrum, the increase in the 1630 cm^{-1} band and the decrease in the 1085 cm^{-1} and 1030 cm^{-1} bands are attributed to weathering of the clear coat layer.

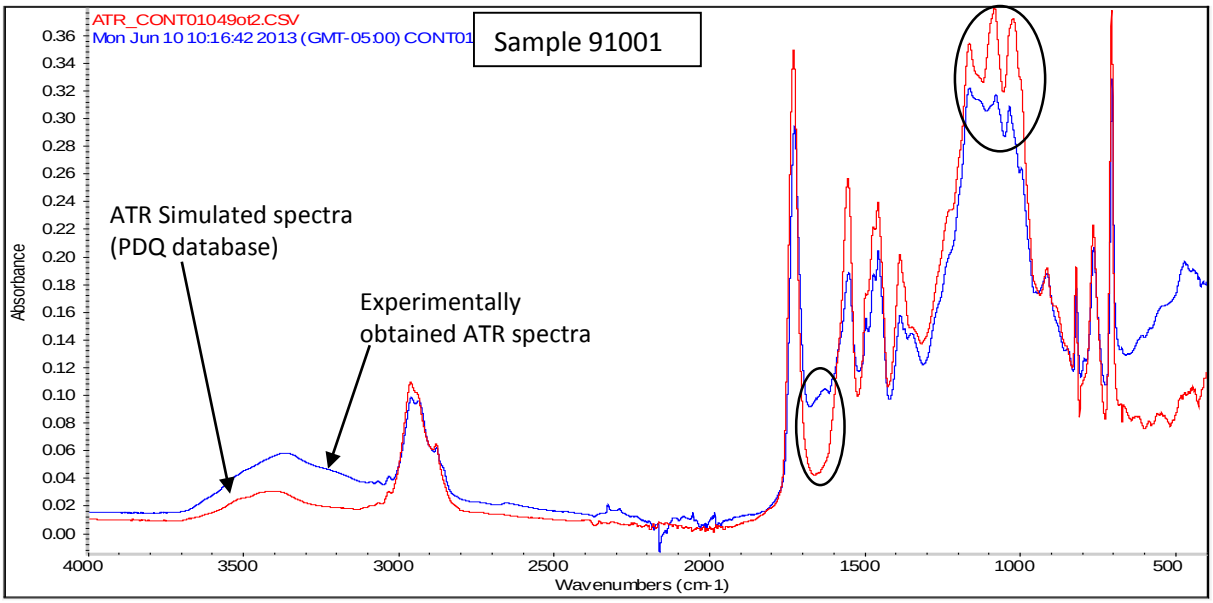

Figure 80. Simulated (red) and experimental (blue) ATR spectra for Sample 91001. For the experimental ATR spectrum, the increase in the 1630 cm^{-1} band and the decrease in the 1030 cm^{-1} and 1085 cm^{-1} bands are attributed to weathering of the clear coat layer.

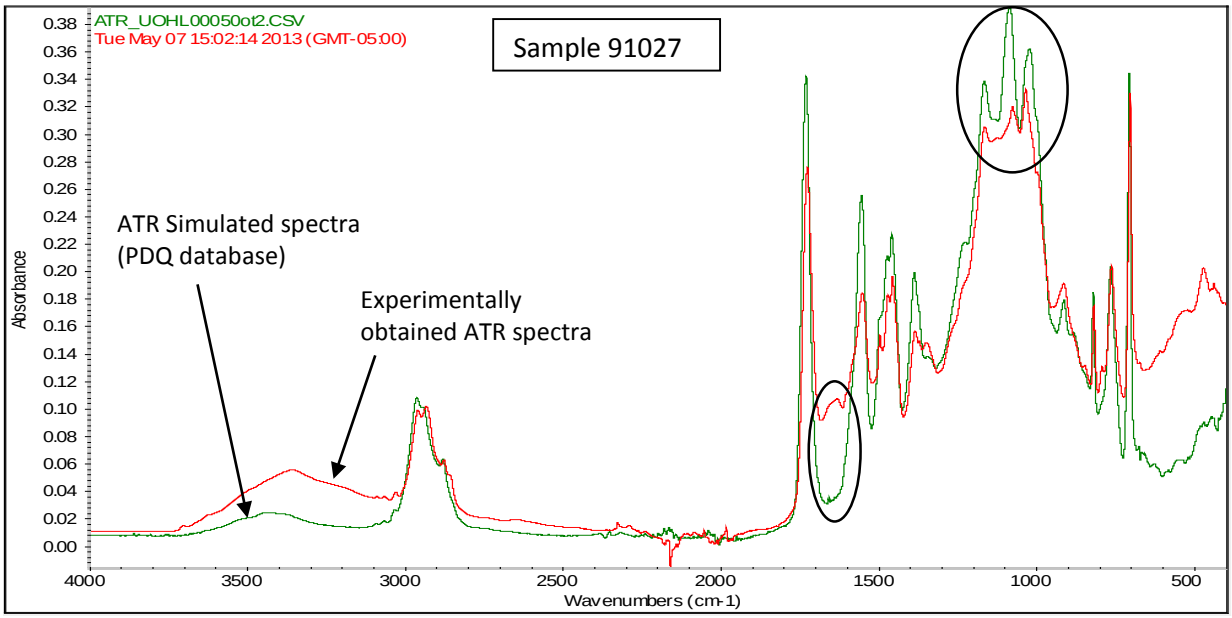

Figure 81. Simulated (green) and experimental (red) ATR spectra for Sample 91027. For the experimental ATR spectrum, the increase in the 1630 cm^{-1} band and the decrease in the 1030 cm^{-1} and 1085 cm^{-1} bands are attributed to weathering of the clear coat layer.

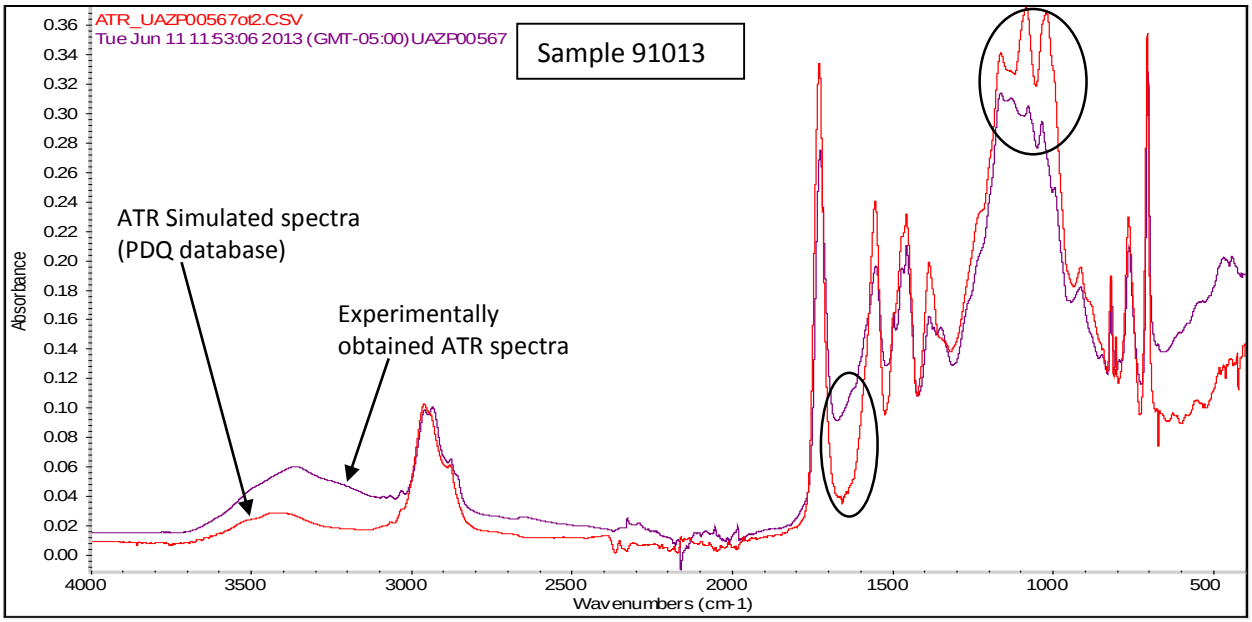

Figure 82. Simulated (red) and experimental (purple) ATR spectra for Sample 91013. For the experimental ATR spectrum, the increase in the 1630 cm^{-1} band and the decrease in the 1030 cm^{-1} and 1085 cm^{-1} bands are attributed to weathering of the clear coat layer.

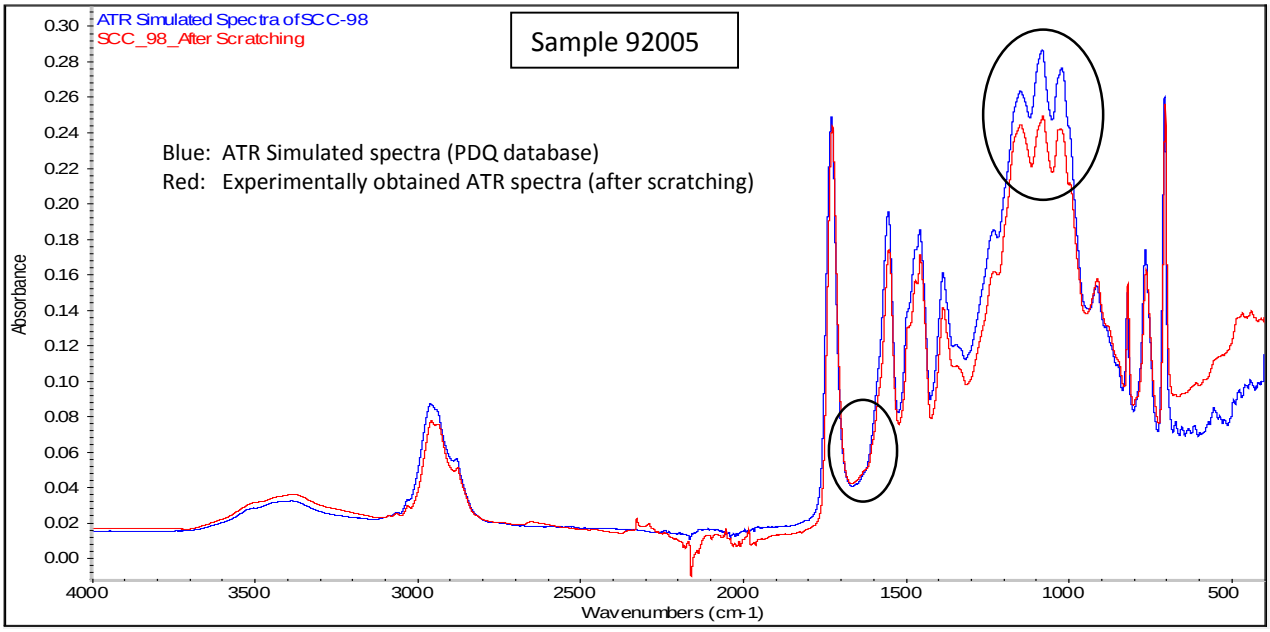

Figure 83. Simulated (red) and experimental (blue) ATR spectra for Sample 92005. For the experimental ATR spectrum, a fresh surface of the clear coat layer was exposed to the spectrometer to minimize the effects of weathering. The circled regions (1630 cm^{-1}, 1030 cm^{-1} and 1085 cm^{-1}) indicative of weathering are consistent with the simulated ATR spectrum.

Spectral libraries of archived transmission data of automotive paint samples can be transformed into ATR libraries using the correction algorithm. As ATR is a surface sensitive technique, the presence of contaminants at or near the surface of the clear coat paint layer, which is not a problem with transmission spectra, will pose a problem for ATR. Removal of these contaminants from automotive paint samples is crucial to quantify the general discrimination power of original automotive paint comparisons limited to clear coat paint smears when ATR spectroscopy is used to characterize automotive paint samples.

IV. Conclusions

Discussion of Findings: A search prefilter was developed to differentiate automotive paint samples by automobile manufacturer (Chrysler, General Motors, and Ford) using the clear coat and undercoat paint layers. For each automobile manufacturer, search prefilters were developed to identify the assembly plant of the vehicle from the manufacturer's paint system conveyed by the sample. First, the assembly plants were divided into groups of assembly plants based upon cluster analysis of the fingerprint region of the clear coat paint layer. Second, each plant group was divided into its assembly plants using both the clear coat and the two undercoat paint layers. The search prefilter system categorizes each unknown paint system by identifying successively smaller sets of vehicles to which the unknown is assigned. The search prefilters have the potential to facilitate spectral library searching as the size of the library is truncated to those spectra of paint samples obtained from assembly plants identified by the search prefilters.

The search prefilters developed for the GMC and Chrysler vehicles (2000-2006) can identify the assembly plant or subplant where the manufacturer's paint system was applied to the vehicle. This, in turn, allows for the model and line of the vehicle to be identified. As some models and lines are assembled in more than one assembly plant, identifying the specific plant that has assembled the vehicle reduces the size of the PDQ library to a smaller number of IR spectra than a search prefilter predicated on identifying the specific model and line of the vehicle. Although the results of the Ford search prefilters suggested that Ford vehicles could pose challenges in the forensic examination of automotive paints as most assembly plants or subplants could not be differentiated, the cross correlation library search algorithm was shown to be effective, yielding results for the matching of the validation set samples comparable to General Motors.

The cross correlation library searching algorithm in conjunction with the search prefilters was able to outperform OMNIC for all three U. S. manufacturers (Chrysler, General Motors, and Ford) and reduce the hit-list to five samples in each search. This is a potentially significant development as it can increase both the speed and accuracy of forensic automotive paint analysis.

A correction algorithm to allow attenuated total reflection (ATR) spectra to be matched using the IR transmission spectra of the PDQ database has also been developed as part of this research project. ATR is a widely used sampling technique in IR spectroscopy because minimal sample preparation is required. As the penetration depth of the ATR analysis beam is shallow, the outer layers of a laminate or multi-layered paint sample can be preferentially analyzed with the entire sample intact. For this reason, forensic laboratories have taken advantage of ATR to collect IR spectra of automotive paint systems which may consist of three or more layers. However, the IR spectrum of a paint sample obtained by ATR will exhibit distortions, e.g., band broadening and lower relative intensities at higher wavenumbers, when compared with its transmission counterpart. This hinders library searching as most library spectra are measured in transmission mode. The correction algorithm to convert transmission spectra from the PDQ library into ATR spectra is able to address distortion issues such as the relative intensities and broadening of the bands, and introduction of wavelength shifts at lower frequencies, which prevent library searching of ATR spectra using archived IR transmission data.

To maintain relevancy of the newly designed library search system, it is crucial to populate PDQ to assembly plants, makes, lines and production years including current years where there is insufficient data. Further validation of the search prefilters and further assessment of the cross correlation library search algorithms also needs to be undertaken to authenticate the performance of the proposed search prefilters and spectral library searching system. The studies discussed in this report have shown that evidential trace information can be obtained directly from clear coat and undercoat paint layers using search prefilters developed as part of a pattern recognition driven library matching search system.

Implications for Policy and Practice: The research project described in this report is directly targeted to enhance current approaches to data interpretation of forensic paint examinations and to aid in evidential significance assessment, both at the investigative lead stage and at the courtroom testimony stage. The search prefilters and cross correlation library searching algorithms for the PDQ database have the potential to enhance current approaches to data interpretation of forensic paint examinations and to aid in evidential significance assessment, both at the investigative lead level and at the courtroom testimony stage. There is also potential for direct impact with the 53 local, state, and federal forensic laboratories currently using the PDQ database in the United States as well as international forensic laboratories including the National Forensic Laboratory Services Division of the RCMP, the Centre of Forensic Sciences in Toronto, Canada, members of the ENFSI network of European forensic science institutes, and the Australian Police Services.

The R & D effort described in this report is an international collaborative effort between the Lavine research group at Oklahoma State University and Mark Sandercock of the RCMP. The advantages of using pattern recognition techniques to search the IR spectra of the PDQ database include extraction of investigative lead information from clear coat and undercoat paint layers and increased accuracy of searches as spectra from the entire database are searched. This is a significant improvement over the way searches are currently performed for automotive paints using the PDQ database. Information derived from the proposed pattern recognition searches will allow forensic scientists to quantify the general discrimination power of original automotive paint comparisons encountered in casework. Addressing these concerns is a direct response to Recommendation 3 of the National Academies' February 2009 report, "Strengthening Forensic Science in the United States: A Path Forward." It is anticipated that once these pattern recognition techniques have been developed, they may also be used to efficiently and accurately search other forensic spectral libraries, for example, illicit drug and pharmaceutical databases, textile fiber database and explosive databases.

Implications for Future Research: The proposed pattern recognition driven library searching system and the ATR correction algorithm have the potential to extract evidentiary lead information from automotive paint systems. In the forensic examination of automotive paint, each layer of paint is analyzed individually by FTIR. The more unique the paint layers are, the more information is contained in the sample, and the stronger are the forensic conclusions that can be drawn. Laboratories in North America 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 also be a problem as it produces an IR spectrum that is a mixture of two layers. Not having a "pure" spectrum of each layer will prevent 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 suspects. We have addressed these problems 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 an FTIR imaging microscope equipped with a linear array detector. Decatenation of the concatenated IR data can be achieved using multivariate curve resolution techniques to obtain a "pure" IR spectrum of each automotive paint layer. This approach, not only saves time and eliminates the need to analyze each layer separately, but also ensures that the final spectrum of each layer is "pure" and not a mixture. By integrating this imaging experiment including the use of multivariate curve resolution to improve spatial resolution with a prototype pattern recognition IR library searching system, the forensic examination of automotive paints can be facilitated in terms of both speed and accuracy.

V. References

1. G. Fettis (Editor), Automotive Paints and Coatings, VCH Publications, New York, 1995.
2. S. Ryland, T. Jegovich, K. P. Kirkbride, "Current Trends in Forensic Paint Examination," Forensic Sci. Rev. 18 (2006) 97-117.
3. K. Flynn, R. O'Leary, C. Lennard, C. Roux, B. J. Reedy, "Forensic Applications of Infrared Chemical Imaging: Multilayered Paint Chips," J. Forensic Sci. 50 (2005) 832-841.
4. P. G. Rogers, R. Cameron, N. S. Cartwright, W. H. Clark, J. S. Deak, "The Classification of Automotive Paint by Diamond Windows Infrared Spectrophotometry-Part I: Automotive Topcoats and Undercoats," E. W. W. Norman, Can. Soc. Forensic Sci., 9 (1976) 1-14.
5. P. G. Rogers, R. Cameron, N. S. Cartwright, W. H. Clark, J. S. Deak, E. W. W. Norman, "The Classification of Automotive Paint by Diamond Windows Infrared Spectrophotometry-Part II: Automotive Topcoats and Undercoats," Can. Soc. Forensic Sci., 9 (1976) 49-68.
6. N. S. Cartwright, L. J. Cartwright, 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. Forensic Sci. J., 15 (1982) 105-115.
7. J. L. Buckle, D. A. MacDougal, and 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. Forensic Sci. J. 30 (1997) 199-212.
8. S. R. Lowry, D. A. Huppler, and C. R. Anderson, Data Base Development and Search Algorithms for Automated Infrared Spectral Identification," J. Chem. Inf. Computer Sci., 25 (1985) 235-241.
9. B. K. Lavine and A. J. Moores, "Genetic Algorithms for Pattern Recognition Analysis and Fusion of Sensor Data," in Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, K. Siddiqui and D. Eastwood (Eds.), Proceedings of SPIES, 1999, pp. 103-112.
10. B. K. Lavine, A. Fasasi, N. Mirjankar, and M. Sandercock, "Search Prefilters to Assist in Library Searching of Infrared Spectra of Automotive Clear Coats," Talanta, 120 (2015) 182-190
11. B. K. Lavine, A. Fasasi, N. Mirjankar, and C. White, "Search Prefilters for Library Matching of Infrared Spectra in the PDQ Database using the Autocorrelation Transformation," Microchem. J., 113 (2014) 30–35.

12. B. K. Lavine, A. Fasasi, N. Mirjankar, M. Sandercock, and S. D. Brown, “Search Prefilters for Mid-IR Spectra of Clear Coat Automotive Paint Smears Using Stacked and Linear Classifiers,” J. Chem., 28 (2014) 385-394.
13. B. K. Lavine, N. Mirjankar, and S. Delwiche, “Classification of the Waxy Condition of Durum Wheat by Near Infrared Reflectance Spectroscopy using Wavelets and a Genetic Algorithm,” Microchem. J., 117 (2014) 178-182.
14. B. K. Lavine, K. Nuguru, N. Mirjankar, and J. Workman, “Pattern Recognition Assisted Infrared Library Searching,” Appl. Spec., 66 (2012) 917-925.
15. B. K. Lavine, K. Nuguru, N. Mirjankar, and J. Workman, “Development of Carboxylic Acid Search Prefilters for Spectral Library Matching,” Microchem. J., 103 (2012) 21-36
16. B. K. Lavine, N. Mirjankar, S. Ryland, and M. Sandercock, “Wavelets and Genetic Algorithms Applied to Search Prefilters for Spectral Library Matching in Forensics,” Talanta**,** 87 (2011) 46-52.
17. B. K. Lavine, D. Brzozowski, A .J. Moores, C. E. Davidson, and H.T. Mayfield, “Genetic Algorithm for Fuel Spill Identification,” Anal. Chim. Acta, 437 (2001) 233-246
18. G. A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F. K. Tadjimukhamedov, B. K. Lavine, and Nikhil Mirjankar, “Pattern Recognition Analysis of Differential Mobility Spectra with Classification by Chemical Family,” Anal. Chim. Acta, 579 (2006) 1-10
19. J. Karasinski, S. Andreescu, O. A. Sadik, B. Lavine, and M. N. Vora, “Multiarray Sensors with Pattern Recognition for the Detection, Classification, and Differentiation of Bacteria at Subspecies and Strain Levels,” Anal. Chem., 77 (2005) 7941-7949.
20. B. K. Lavine, C. E. Davidson, and W. T. Rayens, “Machine Learning Based Pattern Recognition Applied to Microarray Data,” Combinatorial Chem. High Through. Screening**, 7 (**2004) 115-131.
21. B. K. Lavine, C. E. Davidson, C. Breneman, and W. Katt, “Electronic Van der Waals Surface Property Descriptors and Genetic Algorithms for Developing Structure-Activity Correlations in Olfactory Databases,” J. Chem. Inf. Science, 43 (2003) 1890-1905.
22. B. K. Lavine, C. E. Davidson, A. J. Moores, and P. R. Griffiths, “Raman Spectroscopy and Genetic Algorithms for the Classification of Wood Types,” Appl. Spec., 55 (2001) 960-966.
23. B. B. Hubbard, The World According to Wavelets (Second Edition), A. K. Peters, Natick, MA 1998.
24. J. S. Walker, A Primer on Wavelets and Their Scientific Applications, Chapman & Hall, CRC Press, New York 1999.
25. F. Chau, Y. Liang, J. Fao, and X. Shao, Chemometrics – From Basics to Wavelet Transform, John Wiley & Sons, NY 2004.
26. L. A. Powell and G. M. Hieftje, “Computer Identification of Infrared Spectra by Correlation-Based File Searching,” Anal. Chim. Acta, 100 (1978) 313-327
27. M. James, Classification Algorithms, John Wiley & Sons, New York 1985.
28. J. E. Jackson, A User’s Guide to Principal Component Analysis, John Wiley & Sons, NY 1991.
29. D. L. Massart and L. Kaufman, The Interpretation of Analytical Chemical Data by the use of Cluster Analysis, John Wiley & Sons, NY 1983.
30. F. Woolen, “Dispersion Relations and Sum Rules,” in F. Woolen (Editor) Optical Properties of Solids, Academic Press, New York, 1972, pp. 173-185.

31. M. Born, W. Emil, A. B. Bhatia, P. C. Clemmow, D. Gabor, A. R. Stokes, A. M. Taylor, P. A. Wayman, and W. L. Wilcox, Principals of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light, Seventh Edition, Cambridge University Press, 2002
32. R. M. A. Azzam and N. M. Bashira, Ellipsometry and Polarized Light, Netherlands, Elsevier, 1977.
33. J. E. Bertie, R. N. Jones, Y. Apbelblat, C. D. Keefe, "Infrared Intensities of Liquids XIII: Accurate Optical Constants and Molar Absorption Coefficients Between 6500 and 435 cm^{-1} of Toluene at 25^0C from Spectra Recorded in Several Laboratories," Appl. Spectrosc., 48 (1994) 127-143.
34. B. K. Lavine, K. Nuguru, and N. Mirjankar, "One Stop Shopping - Feature Selection, Classification, and Prediction in a Single Step," J. Chem., 25 **(**2011) 116-129
35. T. Nguyen, J. Martin, E. Byrd, "Relating Laboratory and Outdoor Exposure of Coatings: IV. Mode and Mechanism for Hydrolytic Degradation of Acrylic-Melamine Coatings Exposed to Water Vapor in the Absence of UV Light," J. Coating Technol., 75 (2003) 37-45.
36. J. F. Larche, P. O. Bussiere, S. Therias, and J. L. Gardette, "Photooxidation of Polymers: Relating Material Properties to Chemical Changes," Poly. Degradation Stability, 97 (2012) 25-30.
37. J. F. Larche, P. O. Bussiere, P. Wong-Wah-Chung, and J. L. Gardette, "Chemical Structure Evolution of Acrylic-Melamine Thermoset upon Photo-Ageing," European Poly. J. 48 (2012) 172-180.

VI. Dissemination of Research Findings

The development of an ATR correction algorithm and a pattern recognition driven library searching system for automotive paints analysis and the demonstration of their efficacy against a large database of FTIR spectra is of significant interest to the wider scientific community. Therefore, publication of the initial results from this work will occur in more widely read journals such as Analytical Chemistry, Talanta, Analytica Chimica Acta, and Applied Spectroscopy. Additional publications, demonstrating the practical application of this prototype system to solve forensic casework will be published in one of the more widely read forensic science journals such as Journal of Forensic Sciences.

Oral and poster presentations focusing on the search prefilters have been made at several scientific meetings, e.g., American Academy of Forensic Sciences, Federation of Analytical Chemistry and Spectroscopy Societies, the Pittsburgh Conference, and the American Chemical Society. These presentations have generated interest among analytical chemists and forensic scientists as the pattern recognition driven library search system and the ATR correction algorithm have the potential to be used efficiently and accurately to search other forensic spectral libraries, for example, illicit drug and pharmaceutical databases, textile fiber database, explosive databases, architectural paint databases, plastic databases (motor vehicle parts), adhesive tape databases, caulks and sealant databases, and pigment databases (art forgeries).

We are currently working with the RCMP who have developed the PDQ database to engage the practitioner community and transition this research into practice. PDQ database users will be made aware of these research activities through the RCMP in their annual PDQ database updates. The PDQ Maintenance Team of the RCMP provides support and training on basic and advanced search

techniques for the database and they are often invited to the AAFS meetings to run a workshop. At this workshop, the maintenance team will be able to inform users about our research activities to improve the searching capabilities of the PDQ database.