| | |
|---|---|
| Document Title: | Improving Investigative Lead Information and Evidential Significance Assessment for Automotive Paint and the PDQ Database |
| Author(s): | Barry K. Lavine, Ayuba Fasasi, Nikhil Mirjankar, Mark Sandercock |
| Document No.: | 246642 |
| Date Received: | May 2014 |
| Award Number: | 2010-DN-BX-K217 |

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 Investigative Lead Information and Evidential Significance Assessment for Automotive Paint and the PDQ Database

**Award Number:** 2010-DN-BX-K217

**Authors:** Barry K. Lavine, Ayuba Fasasi, Nikhil Mirjankar, and Mark Sandercock

## Abstract

New pattern recognition techniques have been developed for searching infrared (IR) spectral libraries of the Paint Data Query (PDQ) automotive paint database to differentiate between similar but nonidentical Fourier transform infrared paint spectra, and to determine the assembly plant, model, and line of the vehicle from which an unknown paint sample originated. Currently, modern automotive paints use thinner undercoat and color coat layers protected by a thicker clear coat layer. As a result, a clear coat paint smear is sometimes the only layer of automotive paint left at the crime scene. In these cases, the text based portion of the PDQ database cannot identify the motor vehicle because of the reliance of the search on large variations in color and chemical formulation, which do not exist with clear coats. However, clear coat paint layers, like the undercoat and color coat paint layers, exhibit chemical features in their IR spectra indicative of the automobile manufacturing plant at which they were applied, so clear coat spectra may be used to identify the model, and line of a motor vehicle. An added advantage of using pattern recognition techniques to identify paint samples from their IR spectra will be an increase in accuracy because spectra from the entire database are searched. Information derived from these searches can serve to quantify the general discrimination power of original automotive paint comparisons encountered in casework, and will further efforts to succinctly communicate the significance of the evidence to the courts. 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.". To maintain relevancy of these newly designed pattern recognition techniques for spectral library searching, development of data resident in the database has been simultaneously undertaken to populate it for production years where there is insufficient data. The development of search prefilters and library searching algorithms for clear coat paint spectra in the PDQ database, which is the major goal of this project, is necessary to extract investigative lead information from clear coat paint smears.

# Table of Contents

# Executive Summary

Automotive vehicles can be identified from paint fragments left at the crime scene by comparing the color, layer sequence, and chemical composition of each individual layer of the paint. To make these comparisons possible, a comprehensive database has been developed as well as the means of searching and retrieving information from it by the RCMP. Today, the Paint Data Query (PDQ) database contains over 16,000 samples (street samples and factory panels), that corresponds to over 60,000 individual paint layers, representing the paint systems used in most domestic and foreign vehicles marketed in North America. The uniqueness of the PDQ database, and the support it has received from other forensic science laboratories around the world, has made PDQ a world-wide standard. Currently PDQ is the largest international automotive paint database in existence, and is being used by forensic scientists in Canada, USA, New Zealand, Singapore, Japan, and by the European Union Collection of Automotive Paints, populated by a number of EU countries, but largely maintained by the BKA for both sourcing and significance assessments.

PDQ is a database of the physical attributes, the chemical composition and the IR spectrum of each layer of the original manufacturer's paint system. If the original automotive paint layers are present in the recovered (i.e. unknown) paint chip, PDQ can assist in identifying the specific assembly plant, make and line of the motor vehicle. The PDQ concept is to narrow the list of possible vehicles to a reasonable number of suspects, not to identify a single vehicle. However, paint samples that do not contain the color coat or an undercoat layer pose a problem for PDQ because of the reliance of the text based search on the relatively large variation in both the color and chemical formulation in these layers.

The major problem encountered with the PDQ database is the use of text to code the chemistry of each paint layer. Searches of the PDQ database require the user to code their IR spectrum of the recovered paint sample according to the guidelines set out in 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. In other words, the accuracy of a search is impaired due to the conversion of spectral information to generic coded text. The text based system of PDQ also does not allow for the searching of clear coats because all modern clear coats applied to painted metal parts have only one of two possible formulations (i.e., they are coded as either acrylic melamine styrene, or acrylic melamine styrene polyurethane). There are no inorganic fillers or color with which to further discriminate a clear coat. As modern automotive paints use thinner undercoat and color coat layers protected by a thicker clear coat layer, a clear coat paint smear is sometimes the only layer of paint left at the crime scene. In these cases, the text based portion of the PDQ database, which relies on large variations in the chemical formulation for a successful search, cannot be used to identify the motor vehicle.

To assess the evidentiary information content of clear coat paint smear, pattern recognition techniques have been developed to search the infrared (IR) spectral libraries of the PDQ database to differentiate between similar but nonidentical IR paint spectra. At present, the capability to perform direct searching of IR spectra in PDQ does not exist, and spectral search algorithms commercially available cannot distinguish the subtle differences between clear coat paint spectra

from one vehicle model to the next. To tackle the problem of library searching in the PDQ database, a prototype library search system to identify the manufacturing plant, model, and line of an automobile from its clear coat paint spectrum has been developed. The system consists of **two separate but interrelated components**: **search prefilters** to cull the library spectra to a specific plant or plants and a **cross correlation searching algorithm** to identify spectra most similar to the unknown in the set identified by the search prefilters. As the size of the library is culled for a specific match, the use of the search prefilters increases both the selectivity and accuracy of the search.

Applying wavelets to denoise and deconvolve IR spectra of clear coats by decomposing each spectrum into wavelet coefficients which represent the sample's constituent frequency, a genetic algorithm (GA) for pattern recognition analysis has been used to identify wavelet coefficients characteristic of the manufacturing plant of the automobile from which the clear coat paint smear was obtained. Using the pattern recognition GA to identify wavelet coefficients characteristic of the manufacturing plant from which the clear coat paint sample was obtained, **search prefilters** have been developed to facilitate searching of the IR spectra from the PDQ data base. Even in challenging situations where the samples evaluated were all the same make (General Motors) with a limited production year range (2000-2006), the respective manufacturing plants were correctly identified. These search prefilters have been shown to be robust as spectra collected on two Thermo Nicolet instruments could be used to develop search prefilters that were able to identify the manufacturing plant from clear coat spectra obtained on two older BioRad instruments. To develop these search prefilters, it was necessary to solve the problem of classification transfer between instruments with different spectral line shapes. This was accomplished using convolution and deconvolution functions implemented in OMNIC.

The **library search algorithm cross correlates** an unknown with each IR spectrum in the set identified by the search prefilters. Each cross correlated spectrum is simultaneously compared to the autocorrelated spectrum of the unknown and of the library using windows that span different regions of the data from the midpoint. The autocorrelated unknown is also compared to the autocorrelated library spectrum as part of the matching. The top 5 hits in each search window are compiled and a histogram is computed that summarizes the frequency of occurrence that each library sample is selected. The 5 library samples with the highest frequency of occurrence are selected as potential hits. This approach is similar to one used in bootstrapping which attempts to minimize error in parameter estimation for multivariate calibration models.

Searches currently performed on the PDQ database tend to generate a large number of hits because the chemical information in the current database is described only in terms of generic information such as chemical formulations. An added advantage of using this pattern recognition approach to identify paint samples is an increase in the accuracy of the search because spectra from the entire database are searched. Improving the discrimination capability between spectra in the database using wavelets for spectral preprocessing and the pattern recognition GA to identify informative coefficients, permits the inter-comparison of all original equipment manufacturer (OEM) using the infrared spectra alone, thus making the comparison of all possible pairs in the database a realistic goal. By achieving this goal, the dependence on the text-based portion of the database is reduced resulting in improved ease of use and fewer errors.

By comparison, spectral library matching using commercial search algorithms has met with only limited success as automotive paint libraries are composed of a large number of similar spectra. Commercial IR search algorithms are not sufficiently sensitive at distinguishing subtle but significant features in these spectra, e.g., shoulders, unique peaks shapes and patterns, and minor peaks, which may be highly informative but are often ignored. This is due to the fact that all commercial search algorithms are limited to some form of numerical comparison (performed point by point) between the unknown and each spectrum in the library. Using the cross correlation search algorithm, these problems have been addressed.

# I. Introduction

Studies [1, 2] conducted over 30 years ago by the Royal Canadian Mounted Police (RCMP) have shown that automotive vehicles can be differentiated by comparing the color, layer sequence and chemical composition of each individual layer of automotive paint. To make the comparisons possible, a comprehensive database was developed as well as the means of searching and retrieving information from it. Today, the Paint Data Query (PDQ) database contains over 16,000 samples (street samples and factory panels), that corresponds to over 60,000 individual paint layers, representing the paint systems used on most domestic and foreign vehicles marketed in North America. Each year approximately 500 samples are painstakingly collected, analyzed and added to the PDQ database.

Attempts to keep PDQ current have been made through a collaborative effort between the RCMP, the United States Federal Bureau of Investigation (FBI), the German Bundeskriminalamt (the BKA) and the Japanese National Police Agency. Funding for expansion and maintenance of the database, including analysis of samples and entering the data into the database, was supplied by the National Institute of Science and Technology (NIST) Office of Law Enforcement Standards from 1995 to 1999 [3, 4]. The database is used by some seventy-five US local, state, and federal crime laboratories, including the FBI Laboratory.

Automotive paint systems [5] consist of three to five layers: a clear coat over a color coat which in turn is over one or more undercoats. With the exception of the clear coat, each paint layer contains pigments (the colored component) and fillers, and all layers contain binders (the glue 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 make and model of a vehicle within a limited production year range from a paint chip left at the scene of a crime.

The analytical method used to identify automotive paint relies on the selective absorption of infrared (IR) light by the components in the paint. For IR analysis, each paint layer is separated and placed between two diamond plates. Each component has a characteristic fingerprint which is captured in the IR spectrum. The comparison of the IR spectrum of each paint layer in a paint system (clear coat, color coat, and undercoats or primers) to spectra in the paint database allows for the manufacturing plant at which the paint system was applied and the model, line, and production year to be identified.

PDQ is a database of the physical attributes, the chemical composition and the IR spectrum of each layer of the original manufacturer's paint system. If the original paint layers are present in a recovered (i.e. unknown) paint chip, PDQ can assist in identifying the specific manufacturer and a limited production year range of the motor vehicle. The PDQ program is comprised of two components: 1) the data which contains the complete color, chemical composition, layer sequence and sourcing information on known paint systems; and 2) search and retrieval software used to generate a hit list. To use the program, 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 years 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 spectra of each unknown paint layer against the spectra identified in the database hit list. 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. The PDQ concept is to narrow the list of possible vehicles to a reasonable number of suspects, not to identify a single vehicle [6 - 8].

PDQ can also serve as a source of information to assist in assessing the significance of a match between an unknown and a known original paint system. One can access the database and determine how many other makes and models of vehicles in the database may have an original coating system like the one encountered in a current case. However, the number of like makes/models in the database cannot be relied upon to mimic what is on the road, particularly in a given region of North America. It can only be used to determine the number of like makes/models that are present in the database. PDQ can also provide a source of information to keep the forensic paint examiner abreast of current original automotive coating systems assuming current systems are being continuously added to PDQ as a rate consistent with these systems introduction to the marketplace.

Initially the database and search software was used exclusively within the RCMP forensic laboratory system to assist local investigators. In the early 1990's the software was removed from the RCMP mainframe and rewritten for stand-alone personal computers. The uniqueness of the PDQ database, and the support it has received from other forensic science laboratories around the world, has made PDQ the worldwide standard. Currently it is the largest international automotive paint database in existence and it is being used by forensic scientists in Canada, USA, Australia, New Zealand, Singapore, Japan, and many European countries within the EUCAP community.

The PDQ automotive paint database was originally designed as a general text-based search and retrieval system to eliminate the dependency on any one spectrometer or software package. This text-based coding of both physical and chemical characteristics serves as a potent pre-screen to a general infrared spectral search of materials that tend to be chemically very similar to one another. Incorporation of pattern recognition software into the database has the potential for more specific searches by relying less on subjective text-based characteristics. The database itself contains information on the complete topcoat (clear coat) and undercoat (primer) systems applied to most domestic and foreign vehicles marketed or imported into North America since the mid-1970s. Approximately thirty-three percent of the samples are factory panels received directly from automotive manufacturers and paint suppliers. The remaining sixty-seven percent are actual street samples collected to validate manufacturers' information.

PDQ is intended to be a source-based database, not a population-based database. Hence, information culled from querying the database will not result in frequency of occurrence population statistics for the number of vehicles on the road having a particular finish system. However, it will provide an indication of which vehicles manufactured over the past thirty years have similar paint layer systems.  This is valuable information in the quest for quantitative assessment of evidential significance.

The major problem encountered with the PDQ database is its use of text to code the chemistry of each layer.  Searches of the PDQ database require the user to code their IR spectrum according to the guidelines set out by the database, and to search these codes against the codes in the database.  Direct searching of IR spectra in the database does not exist, and commercial spectral search algorithms cannot distinguish subtle differences between spectra from one vehicle model to the next.  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.  In other words, the accuracy of a search is impaired due to the conversion of spectral information to generic coded text.  The text based system of PDQ also does not allow for the searching of clear coats because all modern clear coats applied to automotive components have only one of two possible formulations (i.e., they are coded as either acrylic melamine styrene, or acrylic melamine styrene polyurethane).  There are no inorganic fillers or color with which to further discriminate a clear coat sample.  Hence, paint samples that do not contain the color coat layer or at least one of the undercoat (primer) layers will pose a problem for PDQ because of the reliance of the text based search on the relatively large variations of color and chemical formulations in these layers. The inability to accurately search IR spectra in PDQ, and not to be able to search clear coat spectra at all, are both significant limitations to the current text-based PDQ database.

Research has been undertaken to develop new pattern recognition techniques for searching IR spectral libraries of the PDQ database. These new search techniques have been used to differentiate between similar but nonidentical IR paint spectra, and to determine the manufacturer, model, make and line of the vehicle from which an unknown paint sample originated. Currently, modern automotive paints use thinner undercoat and color coat layers protected by a thicker clear coat layer.  As a result, a clear coat paint smear is, all too often, the only layer of paint left at the crime scene.  In these cases, the text based portion of the PDQ database will not be able to identify the motor vehicle type.

Applying wavelets [9, 10] to denoise and deconvolve IR spectra of clear coats by decomposing each spectrum into wavelet coefficients which represent the sample's constituent frequency, a genetic algorithm (GA) for pattern recognition analysis [11, 12] has been used to identify wavelet coefficients characteristic of the manufacturing plant of the automobile from which the clear coat paint smear was obtained.  Even in challenging trials where the samples evaluated were all the same make (Chrysler) with a limited production year range, the respective manufacturing plants could be correctly identified [13, 14].

Based upon these previous studies [13, 14], clear coat paint layers, like the undercoat and color coat paint layers, exhibit chemical features in their IR spectra that are unique to the automobile manufacturing plant at which they were applied, so clear coat spectra have the potential to identify the make, model, and line of a motor vehicle.  The development of search prefilters and a cross correlation library searching algorithms for IR spectra in the PDQ database, which is the thrust of this project, is necessary in order to extract investigative lead information from clear coat paint smears.

Searches currently performed on the PDQ database tend to generate a large number of hits because the chemical information in the current database is described only in terms of generic information such as chemical formulations. An added advantage of using the pattern recognition approach to identify paint samples is an increase in the accuracy of the search because spectra from the entire database are searched. Improving the discrimination capability between spectra in the database, using wavelets for spectral preprocessing and the pattern recognition GA to identify informative coefficients, permits the inter-comparison of all original equipment material (OEM) using the IR spectra alone, thus making the comparison of all possible pairs in the database a realistic goal. By achieving this goal, the dependence on the text-based portion of the database is reduced resulting in improved ease of use and fewer errors. By coupling the proposed search prefilters with a library search algorithm that utilizes the cross correlation function, the accuracy of the library search is improved. Search prefilters cull the library to include only spectra from a specific manufacturing plant or a few manufacturing plants (which increases the selectivity of the search) for the final match which is performed using a cross correlation search algorithm. It is well known that a match between an unknown and a spectrum in the library increases as the size of the library increases even though the unknown is not present in the library. Since the size of the library is being dramatically reduced for a specific match when using these search prefilters, the likelihood of obtaining a match due to chance is diminished.

Spectral library matching using commercial IR search algorithms has met with only limited success as automotive paint libraries are composed of a large number of similar spectra. Commercial IR search algorithms are not sufficiently sensitive at distinguishing subtle but significant features in these spectra, e.g., shoulders, unique peaks shapes and patterns, and minor peaks, which may be highly informative but are often ignored [15]. This is due to the fact that all commercial search algorithms are limited to some form of numerical comparison (performed point by point) between the unknown and each spectrum in the library [16]. Using a cross correlation search algorithm, we have been able to address these problems.

The use of search prefilters generates fewer hits, with greater accuracy, translating into a significant time savings for the forensic scientist. Information derived from the proposed pattern recognition searches also serves to quantify the general discrimination power of original automotive paint comparisons encountered in casework, and will further efforts to succinctly communicate the significance of the evidence to the courts. Addressing these concerns is a direct response to Recommendation 3 of the National Academies' February 2009 report [17], "Strengthening Forensic Science in the United States: A Path Forward.". To maintain relevancy of the newly designed pattern recognition techniques, additional analysis of paint samples will be undertaken simultaneously with the development of data to populate PDQ to production years where there is insufficient data. It is anticipated that once these pattern recognition search techniques have been developed, they can also be used to efficiently and accurately search other forensic spectral libraries that utilize IR, for example, illicit drug and pharmaceutical databases, textile fiber databases and explosive databases.

## II. METHODOLOGY

A prototype IR library search system to identify the manufacturing plant, model, and line of an automobile from the IR spectrum of a clear coat has been developed as part of this research project. The system consists of two separate but interrelated components: search prefilters to cull the spectra in the PDQ library to a specific assembly plant and/or assembly plants and a cross correlation searching algorithm to identify IR spectra most similar to the unknown in the set identified by the search prefilters. As the size of the PDQ library is significantly reduced by the search prefilters, both the accuracy and the selectivity of the search will be increased. The cross correlation search algorithm is able to better delineate identical from very similar spectra in the PDQ database than commercial search algorithms, e.g., Know It All (BioRad) and OMNIC (Thermo-Nicolet), as these algorithms are not sufficiently sensitive at distinguishing subtle but significant features in the data such as minor peaks, shoulders, and peaks with unique shapes. Band of low intensity, which are often informative in clear coat spectra, are generally ignored by most commercial search algorithms.

**Search Prefilters.** Discriminants (i.e., search prefilters) have been developed to search the IR spectra in the PDQ database in an effort to differentiate between similar but nonidentical clear coat paint spectra and to correctly identify an unknown paint sample as to the manufacturer, model, make and line of the vehicle. Utilizing search prefilters, many problems previously encountered in IR library searching of the PDQ database have been addressed. Our motivation for using search prefilters is based on the fact that search prefilters will increase the selectivity of the search by eliminating library spectra from the search that are not from the same manufacturing plant or plant group. Because the size of the library is culled for a specific match when a search prefilter is used, the likelihood of obtaining a match due to chance is diminished. The exceptionally high quality of the IR data in the PDQ database, and the comprehensiveness of this database, makes it an excellent source of data for the development and subsequent validation of search prefilters of the type described in this report.

Information contained in any search prefilters should describe the relationship between the composition of the clear coat layer and the manufacturing plant, model, make, and line of the vehicle. However, every search function should have an appropriate degree of fuzziness. If the range of the prefilter is too narrow, there will be too many false negatives. On the other hand, there will be too many false positives if its range is too wide,

To develop search prefilters with the necessary degree of fuzziness, IR spectra in the PDQ database have been preprocessed using wavelets to enhance subtle but significant spectral 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 which are better at conveying the information present in the data than axes defined by the original measurement variables. The wavelet coefficients provide the coordinates of the samples in this new pattern space. The mother wavelet selected to develop the new basis set is the one that best matches the attributes of the data. This gets around 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., manufacturing 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 can be identified and retained and the scales representing background and noise are removed, an enhancement of signal to noise can occur with a reduction in the dimensionality of the data because of the elimination of the wavelet coefficients corresponding to noisy spectral features in the data. Classification of spectra can be improved by selectively combining scales, which will yield a discriminant with better performance than one trained on the original data or developed on an individual scale. Wavelets because of the nature of the basis vectors used to characterize the data are conducive to a variety of approaches for improving the quality of the input data for training.

Using wavelets, each IR library spectrum is passed through two scaling filters: a high pass filter and a low pass filter. The high-pass filter will allow only the high frequency component of the signal to be measured as a set of wavelet coefficients which is called the “approximation”. The low-pass filter will measure the low 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 1 shows the first level of wavelet decomposition applied to a clear coat paint spectrum displayed in the transmittance mode.

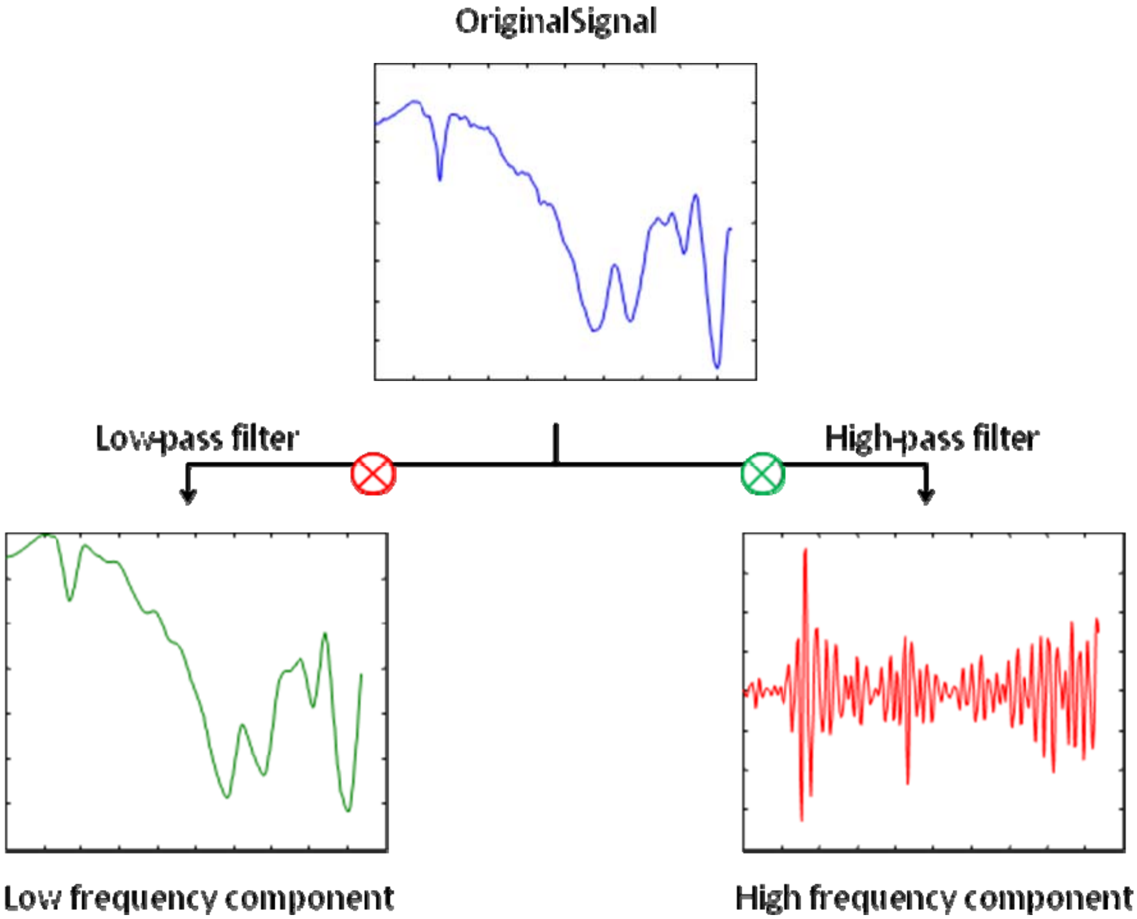


Figure 1. First level of wavelet decomposition applied to a clear coat paint spectrum displayed in the transmittance mode.

Wavelet coefficients characteristic of the model or manufacturer of the vehicle will be identified by a genetic algorithm (GA) for pattern recognition analysis [18-30] that utilizes both supervised and unsupervised learning to identify coefficients that optimize separation of the spectra by manufacturing 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 coefficients selected by the pattern recognition GA is about differences between the different classes (manufacturing plants) in the data. 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 the classes. This fitness criterion dramatically reduces the size of the search space. In addition, the pattern recognition GA focuses on those classes and/or samples that are difficult to classify as it trains by boosting the relative importance of classes and/or samples that are consistently misclassified. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one -pass procedure for feature selection and classification.

This idea is demonstrated in Figure 2, which shows a plot of the two largest principal components of a data set prior to feature selection. The data set consists of 30 IR spectra of clear coats (2000-2003) distributed between 3 classes (1 = Neon, 2 = PT Cruiser, and 3 = Chrysler 300). Each spectrum is characterized by 8 wavelet coefficients. However, only four of these coefficients contain information about the model type. When a principal component map of the data is developed using only these four coefficients, clustering of spectra on the basis of the model is evident. Chance classification will not be a serious problem because the bulk of the variance or information content of the feature subset selected is about the classification problem of interest. Furthermore, features that contain discriminatory information about a particular class membership problem are often correlated, which is why feature selection methods utilizing principal component analysis to display the information content of the data are preferred.

Implementation of the pattern recognition GA requires a population of candidate solutions and heuristics to manipulate them. The actual procedure involves several interrelated steps. First, an initial population of feature subsets is generated. During each generation, the feature subsets are sent to the fitness function for evaluation. Each feature subset is assigned a value by the fitness function, which is a measure of the quality of the proposed feature subset for the classification problem. Reproduction is then implemented. It involves three operators: selection, recombination, and mutation.

The selection and crossover operators are implemented by ordering the population of strings, i.e. potential solutions, from best to worst, while simultaneously generating a copy of the same population and randomizing the order of the strings in this copy with respect to their fitness. A fraction of the population is then selected as per the selection pressure which is set at 0.5. The top half of the ordered population is mated with strings from the top half of the random population, guaranteeing the best 50% are selected for reproduction, while every string in the randomized copy has a uniform chance of being selected. (This is due to the randomized selection criterion imposed on strings from this population.) If a purely biased selection criterion were used to select strings, only a small region of the search space would be explored. Within a few generations, the population would consist of only copies of the best strings in the initial population. Additional randomness or variability is achieved through the mutation operator, which deletes or adds a specific feature to the chromosome based on a certain probability. Finally, the boosting algorithm adjusts the internal parameters for the next iteration. The aforementioned procedure (evaluation, reproduction, and adjustment of internal parameters) is repeated until a specified number of generations have been executed or a feasible solution is found.

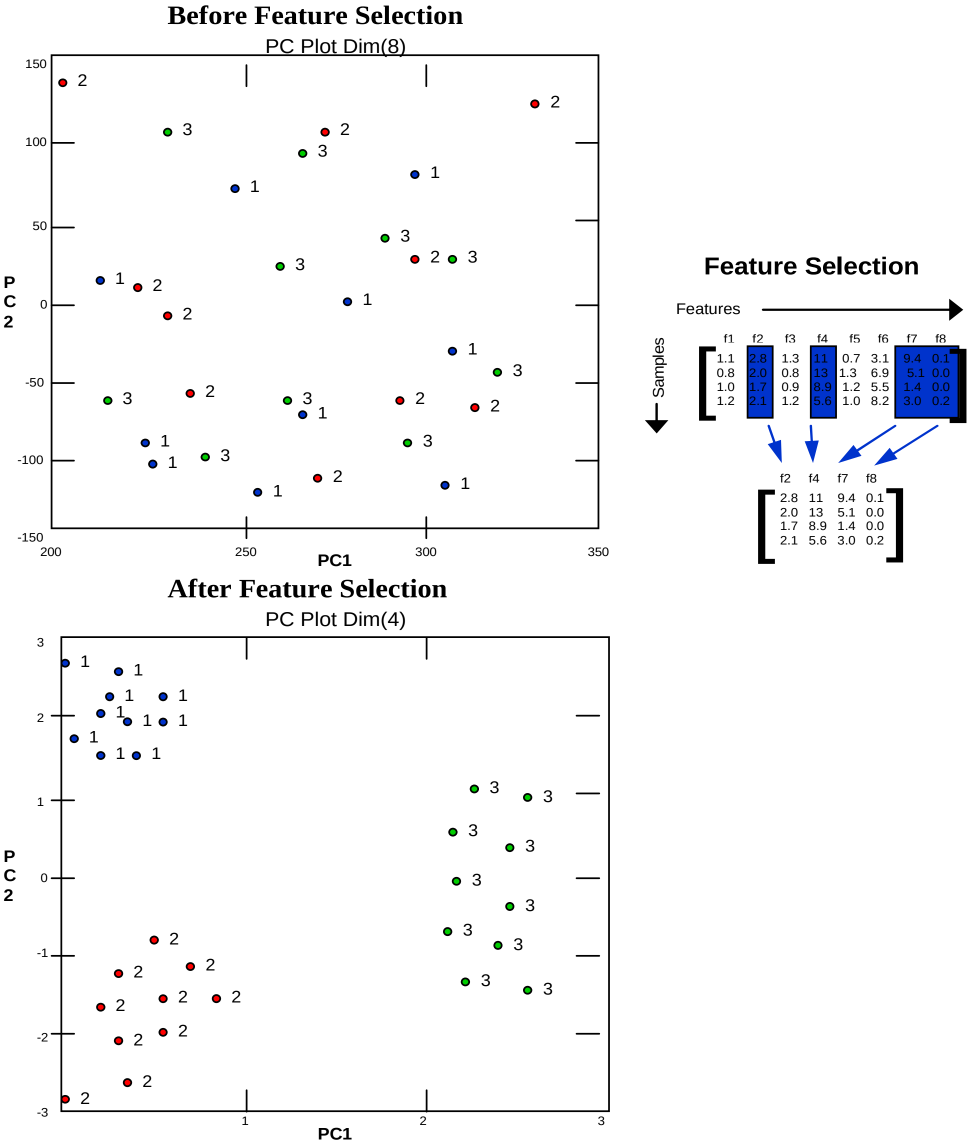


Figure 2. A plot of the two largest principal components developed from all of the wavelet coefficients in the data set does not show class separation. When principal components are developed from the coefficients that contain information about the class membership of the clear coat paint samples, clustering on the basis of the automobile model is evident.

A block diagram of the pattern recognition GA for feature selection that illustrates its operation is shown in Figure 3. The fitness functions used (PCKaNN for the studies described in this report), the reproduction operators, and the mechanism for adjusting the internal parameters of the GA (i.e., boosting) to guide the search in the appropriate direction through adjustment of the fitness function are important aspects of the pattern recognition GA that make it unique.

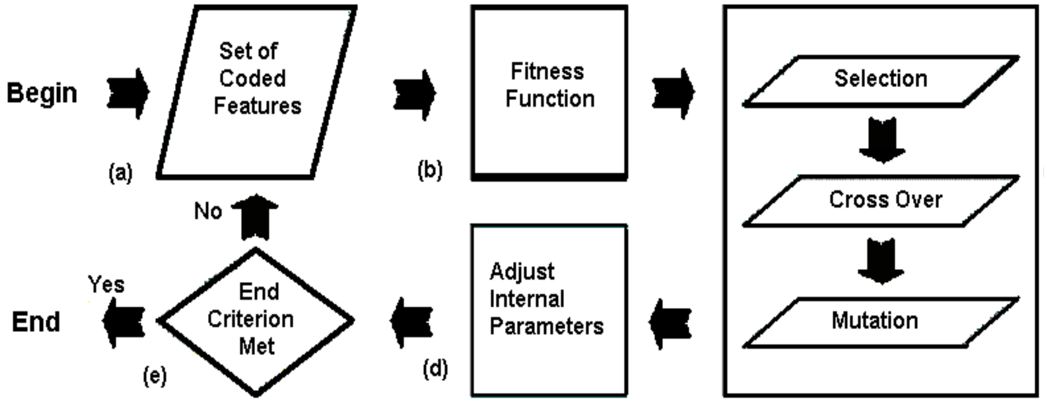


Figure 3. Block diagram of the pattern recognition GA for classification and feature selection.

The fitness function of the pattern recognition GA (which is called PCKaNN) emulates human pattern recognition through machine learning to score the principal component plots and thereby identify a set of features that optimize the separation of the classes 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 and sample weights, 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.

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). 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.

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

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

The fitness function of the pattern recognition GA is able to focus on those samples and classes that are difficult to classify by boosting their sample and class weights over successive generations. In order to perform boosting (designated as adjustment of internal parameters in Figure 3), 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). $\phi$ 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.

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

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

$$\mathrm{CHR}_g(c) = AVG(\mathrm{SHR}_g(s):\forall_{s\in c}) \qquad (5)$$

$$\mathrm{CW}_{g+1}(s) = \mathrm{CW}_g(s) + P(1\text{-}\mathrm{CHR}_g(s)) \qquad (6)$$

$$\mathrm{SW}_{g+1}(s) = \mathrm{SW}_g(s) + P(1\text{-}\mathrm{SHR}_g(s)) \qquad (7)$$

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., pattern classifiers) have been developed from the PDQ library spectral data that extract information from an IR spectrum of an unknown clear coat automotive paint sample to yield a response based on the manufacturing plant that produced the vehicle. Spectral features encoded in the wavelet coefficients identified by the pattern recognition GA have been used to develop these classifiers that serve as our search prefilters. In this study, we focused on the development of search prefilters to identify the model and manufacturing site from clear coat paint samples obtained from 24 General Motors (GM) car and truck manufacturing plants between the years 2000 and 2006. During this time period, GM had the largest number of vehicle plants in North America. If prefilters can be developed that are able to discriminate automobiles manufactured at one GM plant from those made at another, we believe this would be the best possible test of the proposed methodology to demonstrate the validity of this concept. Expanding these search prefilters to other vehicle manufacturers could then be undertaken in the future. For example, all of GM's car and truck lines producing cars between 2000 and 2006 could be compared against Ford, Chrysler, Honda, and Toyota. Currently, some GM vehicle models are not well represented over the 2000-2006 range in PDQ. However, additional samples have been analyzed and added to the database during this project. The addition of these paint samples to the database further strengthened the conclusions that we have drawn from this research.

IR spectra in the PDQ library for GM automobiles between the years 2000-2006 have been collected using four different spectrometers: BioRad 40A, BioRad 60A, and two Thermo-Nicolet 6700 FTIR spectrometers. Each spectrometer was nominally run at 4 cm$^{-1}$ resolution. Analogous to the situation found in multivariate calibration, a classification model (i.e., a search prefilter) developed from spectra measured on one instrument may not be valid when applied to spectra collected on a second instrument. Due to the large number and the variety of IR spectrometers sold, the ability to transfer multivariate classification models between IR spectrometers is crucial for the successful application of the search prefilters developed from IR data in the PDQ database. Therefore, the transfer of multivariate classification models between laboratory spectrometers has been investigated as part of this project. Algorithms to implement classification transfer can be divided into two groups. In the first group, a set of standards common to both the primary and secondary instruments is used to correct for unwanted instrumental variation. Of these approaches, the most popular are piecewise direct standardization [31] and orthogonal signal correction [32]. In the second group, signal preprocessing and data transformation techniques are used to correct for unwanted instrumental variation, e.g., finite impulse response filtering [33] and slope and bias correction [34], as transfer standards are unavailable.

Although there are discussions, reviews, and algorithm comparisons published on this subject, fundamental and first principle derivations are lacking. Often, practitioners are confronted with the situation of applying a potpourri of algorithms to their data to empirically determine what "works" best for their own application. They end up searching for models with fewer factors and smaller standard error of predictions (without confidence limits or tolerance verification) rather than applying a thorough understanding and rigor to seek a more fundamental solution to the problem of instrument transfer. In this study, classification transfer was accomplished by matching spectral line shapes of different instruments using convolution and deconvolution functions implemented with Nicolet's OMNIC software system. This allowed the spectra from one instrument appear to have been collected on a second instrument. The success of transforming spectral lines shapes between spectrometers will enable the new pattern recognition techniques developed for spectral library searching of the PDQ database to be successfully applied regardless of the IR spectrometer used to collect the data.

**Cross Correlation Library Searching Algorithm.** Library matching of IR spectra was performed by cross correlating the unknown with each spectrum in the set of library spectra identified by the search prefilters. 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 $\Delta t$ and relative displacement $n\Delta t$ between two signals $s_i$ and $s_j$ is estimated as shown in the following equation

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

Autocorrelation is similar to cross correlation with the signal cross correlated with itself. Cross correlation was performed by normalizing the sequence calculated such that the autocorrelated sequence calculated for all signals at zero lag is 1. Each unknown clear-coat spectrum was compared with clear-coat spectra from a plant or group of plants in the PDQ library identified by the search prefilters. Three different comparisons between the unknown and library spectra were made:

1. Autocorrelated unknown was compared with each cross-correlated unknown and library spectrum
2. Each autocorrelated library spectrum was compared with each cross-correlated unknown and library spectrum
3. Autocorrelated unknown was compared with each autocorrelated library spectrum

Each comparison was performed using a wide range of varying window sizes centered at the midpoint of the cross-correlated data. The window size was varied from the mid-point (which corresponds to the cross correlation between two signals with zero lag) to the entire data (see Figure 4).

The Euclidian distance was used to evaluate the similarity index (see Equation 8) 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 8 was used instead of the hit quality index (HQI) in OMNIC as it proved to be more informative for these spectra.

$$s_{ij} = 1 - \frac{d_{ij}}{d_{\max}} \qquad (9)$$

Library spectra were arranged in descending order of similarity for each comparison and window size. The five most similar library spectra were than chosen from each comparison for each window size. A histogram depicting the frequency of occurrence in these selected spectra was generated. The 5 library spectra with the highest frequency of occurrence were chosen as potential matches.

Library matching was performed in the spectral region 667-1844 cm$^{-1}$, which is primarily the fingerprint region. Spectral features from this region were found to be characteristic of the manufacturing plant. The cross-correlation library search routine was compared to the library matching algorithm used in Omnic under similar conditions and without the use of search pre-filters.

To further improve the performance of the library matching algorithm, the expanded fingerprint region investigated was further divided into multiple regions.  Each region was examined separately.  Library spectra were arranged in descending order of similarity for each region, and the ten most similar library spectra were chosen from each type of comparison and window size.  A histogram depicting the frequency of occurrence in the spectra selected was generated.  The five library spectra with the highest frequency of occurrence were chosen as potential matches.

Absorbance spectra, not transmittance spectra served as input for the cross-correlation algorithm.  Using absorbance, variations in the optical path length of the sample (e.g., the thickness of the clear coat paint smear on the diamond transmission cell) can be removed by normalizing each spectrum to unit length.  For transmittance data, variations in the optical path length are not linearly related to transmittance.  Thus, nonlinear variations in transmittance due to changes in the optical path between samples cannot be eliminated by judicious preprocessing of the data.  As the cross-correlation function is sensitive to these nonlinear variations, the quality of the match is diminished when transmittance data was substituted for absorbance data in the algorithm.

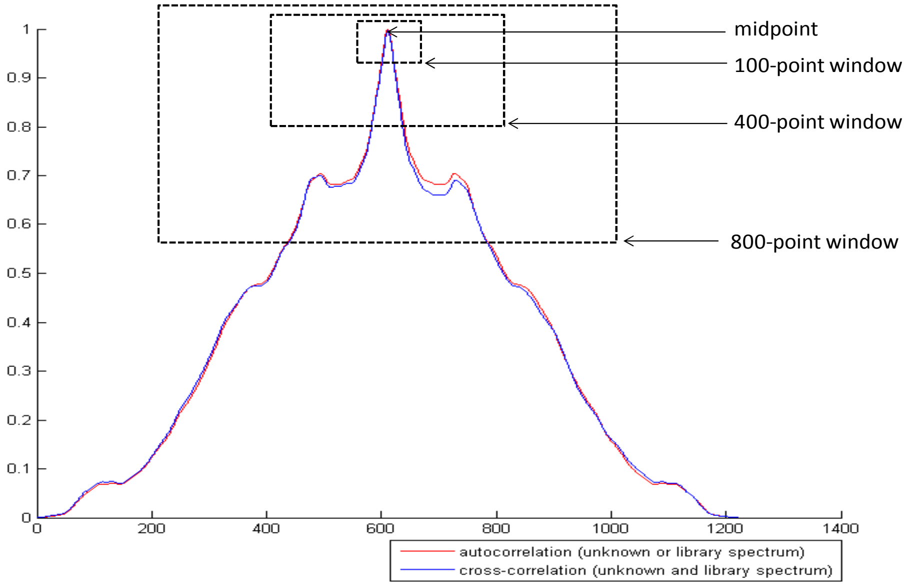


Figure 4.  Comparison of the cross correlated unknown and PDQ library spectrum with the autocorrelated spectrum of the unknown or PDQ library spectrum.

In a previous study, Hieftje and Powell [35] used cross correlation to perform library matching.  However, their approach to spectral matching was limited to comparing the midpoint of the cross correlation of the unknown and each library spectrum using the average value in a twenty-point window to normalize the midpoint.  Hieftje's approach did not prove successful with our data as the clear coat paint spectra from different plants are more similar than they are different.  By comparison, the cross correlation search algorithm that was developed as part of our study and which proved to be successful, took advantage of ideas from bootstrapping, which attempts to minimize error in multivariate classification and calibration models using a thorough and exhaustive comparison of the data, to compare similar spectra.

## III. RESULTS

A hierarchical classification scheme formulated from a visual inspection of the data was used to develop the search prefilters. The spectra were initially divided into two categories based on the carbonyl band at 1709 $cm^{-1}$. In one category, the carbonyl band in each spectrum is a singlet (Plant Groups 1, 3, and 4), and in the other category the carbonyl band is a doublet (Plant Groups 2 and 5). Spectra representative of these two categories are shown in Figures 5 and 6. An examination of the expanded fingerprint region (2000 – 400$cm^{-1}$) reveals five distinct spectral patterns with each pattern designated as a specific Plant Group.

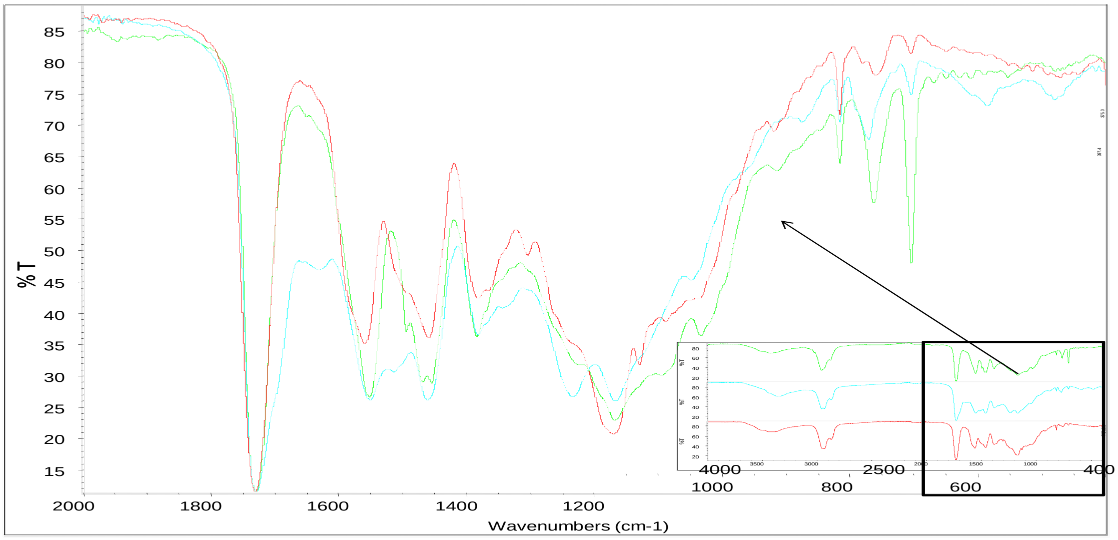


Figure 5. Average spectra of plant groups 1 (green), 3 (cyano), and 4 (red) overlayed with the region 2000 - 400 $cm^{-1}$ expanded

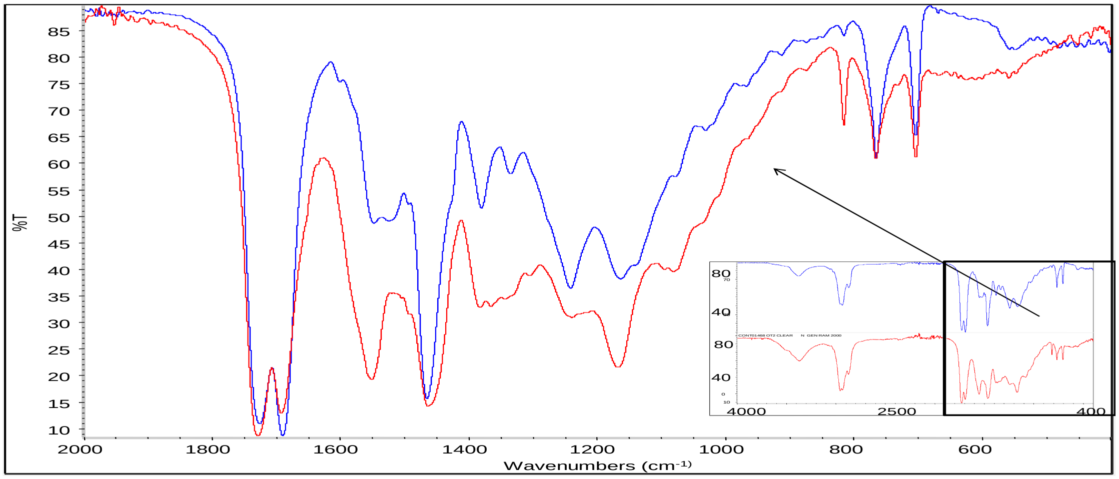
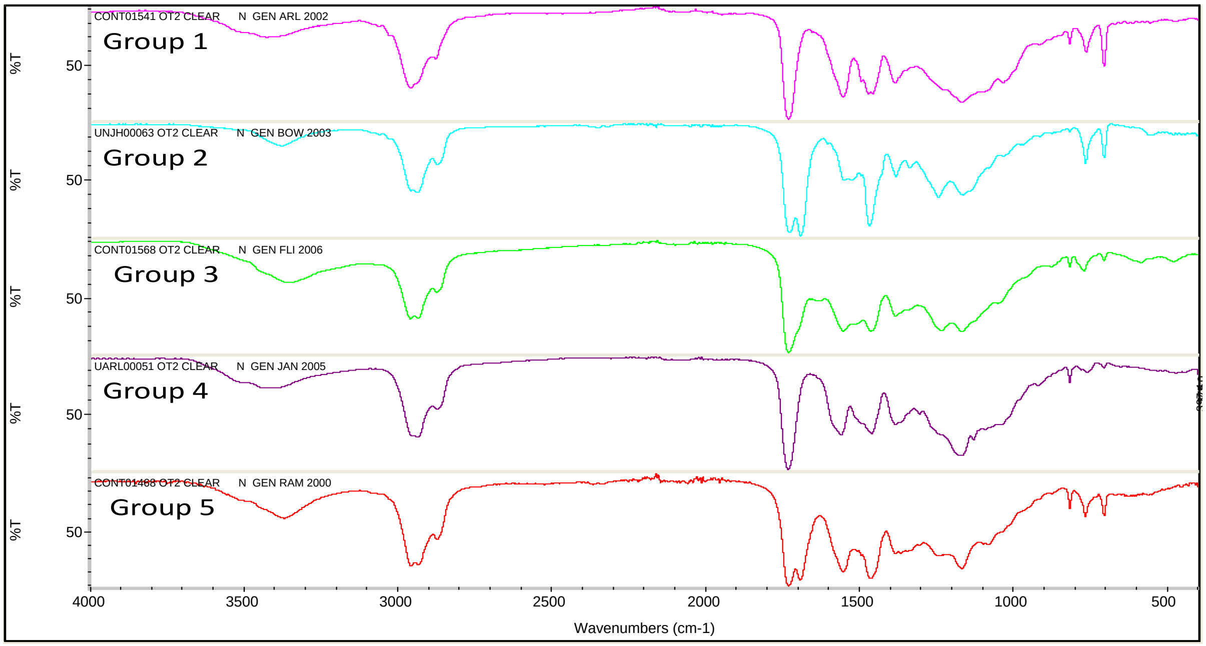


Figure 6. Average spectra of groups 2 (blue) and 5 (red) overlayed with the region 2000 - 400 $cm^{-1}$ expanded

IR spectra of clear coats from the 24 GM plants (2000-2006) were then assigned to one of the five plant groups, i.e., groups of manufacturing plants (see Figure 7). The spectral region used to differentiate the five plant groups includes both the fingerprint region and the carbonyl band of the clear coat. Initially, the carbonyl band was not excluded from the study because it contributed to discrimination of the spectra among the five plant groups. The spectral region from 4000cm$^{-1}$ to 2000 cm$^{-1}$ included the C-H stretch, which is common to all organic samples, and noise associated with the diamond transmission cell. As this spectral region would not be expected to contain information characteristic of the manufacturing plant of the paint sample, it was not used in the development of the search prefilters or in the cross correlation search algorithm for library matching.

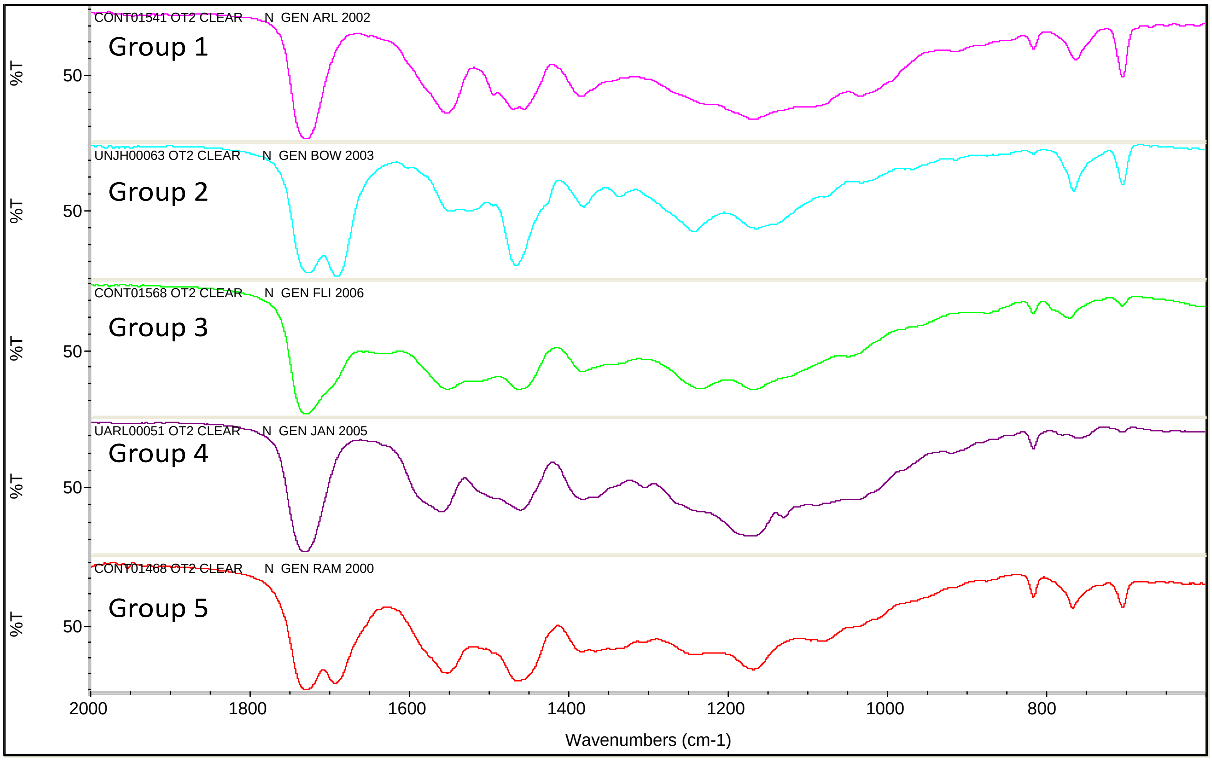


Spectral range: 4000cm$^{-1}$ to 400cm$^{-1}$

Figure 7. IR spectra of clear coat paint smears from the PDQ library for GM automobiles in the United States and Canada (2000-2006) could be assigned to one of five plant groups.

A search prefilter was developed to classify spectra into one of the five plant groups. Each plant group was further divided into individual manufacturing plants or into subgroups of manufacturing plants using a search prefilter to classify the individual spectra within each plant group. For the development of the search prefilter for plant groups, the spectral region of each clear coat paint sample was limited to the extended fingerprint region (2000cm$^{-1}$ to 600cm$^{-1}$), see Figure 8. Search prefilters for individual manufacturing plants were developed from the fingerprint region (1500 cm$^{-1}$ to 600cm$^{-1}$). We chose to limit ourselves to the fingerprint region to exclude the carbonyl band as it was not sufficiently discriminating for this level of classification.

For the development of the search prefilters, transmittance spectra, not absorbance spectra were used. The crucial issue for the development of the search prefilters was deconvolving overlapping spectral responses using wavelets, not removing noise associated with variations in the optical path length of each sample which was obviated by adjusting the thickness (amount) of the sample and the pressure applied by the diamond transmission cell to ensure that an absorbance of 1 was obtained for the carbonyl band in each sample. The carbonyl band was the most intense peak in all clear coat paint spectra.



Spectral range: 2000cm$^{-1}$ to 600cm$^{-1}$

Figure 8. The fingerprint region of the five plant-groups characteristic of GM clear coats (2000-2006)

**Development of Search Prefilters for IR Library Searching of Clear Coat Paint Smears**
Search prefilters developed from spectral data collected on two 6700 Thermo-Nicolet spectrometers were able to identify the respective manufacturing plant and the production line of the automotive vehicle from its clear coat paint smear using transmission spectra collected on a BioRad 40A or BioRad 60 FTIR spectrometer. All 4 spectrometers were equipped with DTGS detectors. An approach based on instrumental line functions was used to transfer the classification model between the Thermo-Nicolet and BioRad instruments. In this study, 209 IR spectra of clear coat paint smears that comprised the training set were collected using Thermo-Nicolet 6700 IR spectrometers, whereas the validation (test) set consisted of 242 IR spectra of clear coats obtained using two BioRad IR instruments.

The clear coat paint spectra used in this study were obtained from 21 General Motors (GM) industrial plants (2000-2006). This made the classification problem challenging as the samples evaluated were all from the same manufacturer (General Motors) with a limited production year range (2000-2006). Only clear coat paint spectra from metal substrates were used to develop the search prefilters. IR spectra of clear coats from bumpers and other plastic substrates were excluded as these components were often not painted in the same plant where the vehicle was assembled. Table 1 lists the GM manufacturing plants that were investigated in this study. A hierarchical classification scheme was used to develop the search prefilters, where the 21 GM manufacturing plants were divided into five major plant groups. The manufacturing plants comprising each plant group are listed in Table 2.

Using OMNIC, all spectra (both the training set and validation set) were aligned. The number of points collected in the wavelength range interrogated by the Thermo-Nicolet instrument varied from 1878 points to 1958 points whereas all spectra collected on the two BioRad instruments for the same wavelength range and resolution were represented by 1944 points. Band shifting was also observed in spectra collected on both the Thermo Nicolet and BioRad instruments. These problems were resolved using Nicolet's OMNIC software as an editor to process the BioRad spectra and the spectra from the Thermo Nicolet instrument using an appropriate estimate of the spectral line function of the two Thermo-Nicolet instruments. To further improve spectral alignment, we focused on the region from 1500 cm$^{-1}$ to 600 cm$^{-1}$. Each IR spectrum selected for processing was normalized to the helium neon laser frequency of 15798.0 cm$^{-1}$. This ensured proper spectral alignment along the x-axis for imported BioRad spectra to the Thermo-Nicolet instrument. For alignment along the y-axis (transmittance) of the spectra, we ensured that all spectra started from the same transmittance value. The quality of the diamond cell transmission spectra in the PDQ library (e.g., no sloping baseline or baseline offsets, and the value of the carbonyl absorbance peak in all library spectra being unity) proved pivotal in the successful alignment of these spectra along the y-axis.

Each clear coat paint sample was assumed to be between 3 and 4 micrograms and was run between diamond windows [36, 37]. Further details about the infrared and sampling conditions used to generate this data can be found elsewhere [38]. Each clear coat IR transmittance spectrum was normalized to unit length. The spectral region from 2000cm$^{-1}$ to 600cm$^{-1}$ used to develop the search prefilters for plant group was represented by 611 points. For pattern recognition analysis, each IR spectrum was initially represented as a data vector, $x = (x_1, x_2, x_3, \ldots x_j, \ldots\ldots x_{611})$ where $x_{611}$ is the transmittance of the clear coat paint sample for the 611th point.  All spectral features within this region were autoscaled to ensure that each measurement has a mean of zero and a standard deviation of one throughout all spectra. Autoscaling removed any inadvertent weighing of the data that otherwise would occur due to differences in the magnitude among the measurement variables comprising each IR spectrum.

**Table 1.  General Motors Plants Investigated in this Study**

| Plant ID | Plant | Make | Line |
|---|---|---|---|
| 1 | ARL | CAD, CHE, GMC | SUB,YUK,ESD,CTA |
| 3 | BOW | CAD,CHE | CVT,XLR |
| 4 | DOR | PON | VTR,SIL,MTA,UPL,TAR |
| 5 | FAI | CHE,OLD,PON | GRA,MAL,ITR |
| 6 | FLI | CHE,GMC | SLV,SIE |
| 8 | FOR | CHE,GMC | SLV,SIE |
| 9 | FRE | GM | VIB,TAC,PVB,COA,GPR |
| 10 | HAM | BUI,CAD,PON | BON,DEV,LUC,LES,SEV,ELD |
| 12 | JAN | GMC | CTA,SUB,YUK |
| 14 | LAN | PON | STS |
| 16 | LIN | CHE,GMC | BZR,JMY,S10 |
| 17 | LRD | PON | SFR,CAV,COB,PST |
| 18 | MOR | CHE,GMC,SAA | JMY,ENV,9S7,BZR,TBZ,SON |
| 20 | OKL | CHE, GMC | MAL,TBZ,ENV,EQU, XUV |
| 21 | ORI | PON,BUI | BON,PG6,LES,AUR,PKA |
| 22 | OSH | GMC, PON | ALL,REG |
| 23 | PON | CHE,GMC | SLV,SIE,SIL |
| 24 | RAM | BUI,CHE,PON | CAV,SFR,RZV,AZT,HHR |
| 25 | SHR | CHE,GMC | S10,COL,SON |
| 26 | SIL | CHE,GMC,SAA | AVL,SUB,YXL |
| 27 | SPH | STR | SSL,ION,SC1,SC2,SL1,VUE |

Table 2.  Manufacturing Plants Comprising Each Plant Group

| Plant Group | Plant ID Number | Manufacturing Plant |
|---|---|---|
| 1 | 1, 4, 5, 8, 14, 18, 23 | ARL, DOR, FAI, LAN, MOR, PON |
| 2 | 3, 10, 21 | BOW, HAM, IRI |
| 3 | 6, 9, 16, 17, 20, 22, 25 | FLI, FRE, LIN, LRD, OKL, OSH, SHR |
| 4 | 12 | JAN |
| 5 | 24, 26, 27 | RAM, SIL, SPH |

The initial focus of this study was to develop a search prefilter to classify the IR spectra by plant group. The training set of 209 IR spectra (Thermo Nicolet) was divided into 5 classes by plant group (see Table 3).  The first step in the study was to apply principal component analysis (PCA) to the normalized and autoscaled IR spectral data. PCA is a powerful method for uncovering hidden relationships in complex multivariate data sets. Using this procedure is analogous to finding a new coordinate system that is better at displaying the information in the data than axes defined by the original measurement variables.  The new coordinate system is linked to variation in the data.  The basis vectors of this new coordinate system are the principal components of the original data.  Each principal component is a linear combination of the original measurement variables.  Often, only the two or three largest principal components are necessary to explain most of the information present in a spectral data set due to the large number of interrelated measurement variables.

Figure 9 shows a principal component (PC) plot of the two largest principal components of the 209 IR spectra and the 611 features from each IR spectrum.  Each paint sample is represented as a point in the PC map 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 overlap of the clear coats from each plant group in the map of the data is evident.

The next step was feature selection.  A genetic algorithm for pattern recognition analysis was used in the study to identify spectral features characteristic of the profile of each plant group. The pattern recognition GA identified features by sampling key feature subsets, scoring their PC plots, and tracking those clear coat paint samples or plant groups that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution.  After 200 generations, the GA identified 12 spectral features (i.e., transmittance values at 12 specified wavelengths) whose PC plot showed clustering of the data on the basis of Plant Group ID (see Figure 10).

**Table 3. Training Set and Validation Set for Plant Group Search Prefilter**

| Group | Training Set Samples (Thermo-Nicolet) | Validation Set Samples (Bio-Rad) |
|---|---|---|
| 1 | 81 | 90 |
| 2 | 22 | 32 |
| 3 | 69 | 50 |
| 4 | 6 | 13 |
| 5 | 31 | 57 |
| Total | 209 | 242 |

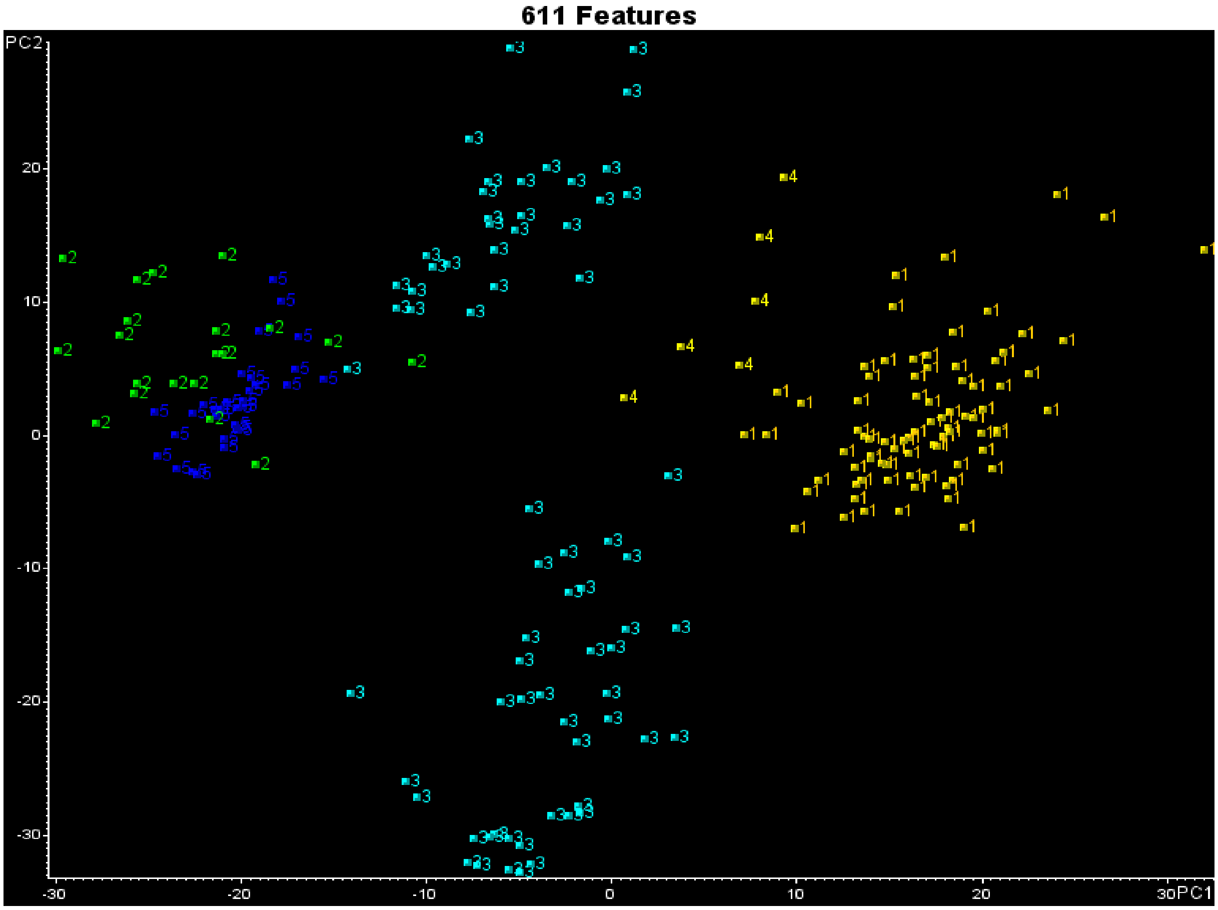


Figure 9. A PC plot of the two largest principal components of the 207 IR spectra and the 611 features of each spectrum is shown. Each paint sample is represented as a point in the PC plot of the data (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

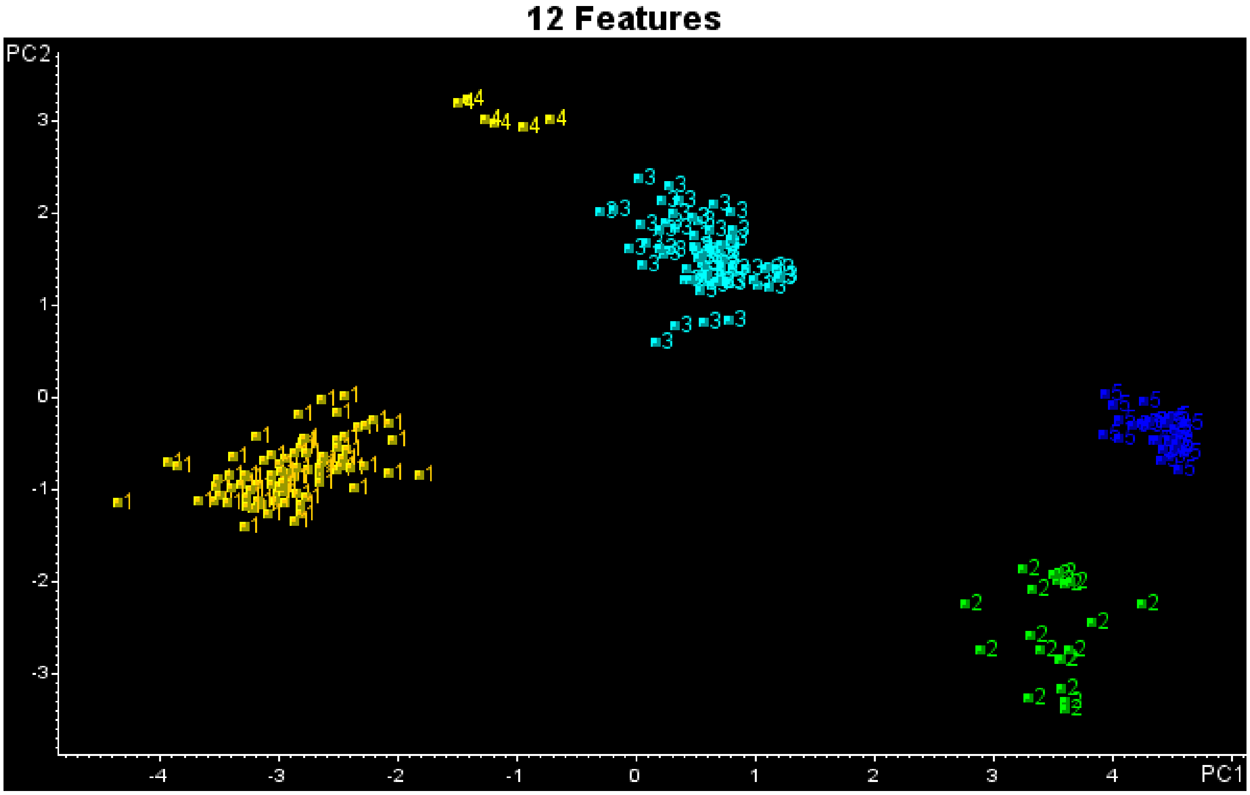


Figure 10. A PC plot of the two largest principal components of the 207 IR spectra and the 12 spectral features identified by the pattern recognition GA is shown. Each paint sample is represented as a point in the PC plot of the data (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

A validation set of 242 IR spectra of clear coats from two BioRad instruments was employed to assess the predictive ability of the 12 spectral features identified by the pattern recognition GA and the efficacy of the alignment procedure used to transfer the search prefilters for use by another instrument. Figure 11 shows the validation set samples projected onto the PC plot defined by the 209 IR spectra (Thermo-Nicolet) and the 12 spectral features identified by the pattern recognition GA. Each projected sample lies in a region of the map with paint samples from the same plant group. This result alone suggests that information about the manufacturing plant is contained in the IR spectrum of the clear coat paint smears.

Linear discriminant analysis (LDA) was also used to classify the 209 IR spectra in the training set. The training set data were divided into 5 classes on the basis of plant group. LDA was used to develop a classifier to separate the paint spectra by plant group. A discriminant developed from the 12 spectral features identified by the pattern recognition GA achieved a classification success rate of 100% for the training set. To further test the predictive ability of these 12 features and the discriminant associated with them, a validation set of 242 IR spectra of clear coat paint smears was employed. Again, a classification success rate of 100% was achieved for the IR spectra in the validation set. The results from the LDA study, which are summarized in Table 4, are consistent with the results obtained using PCA.

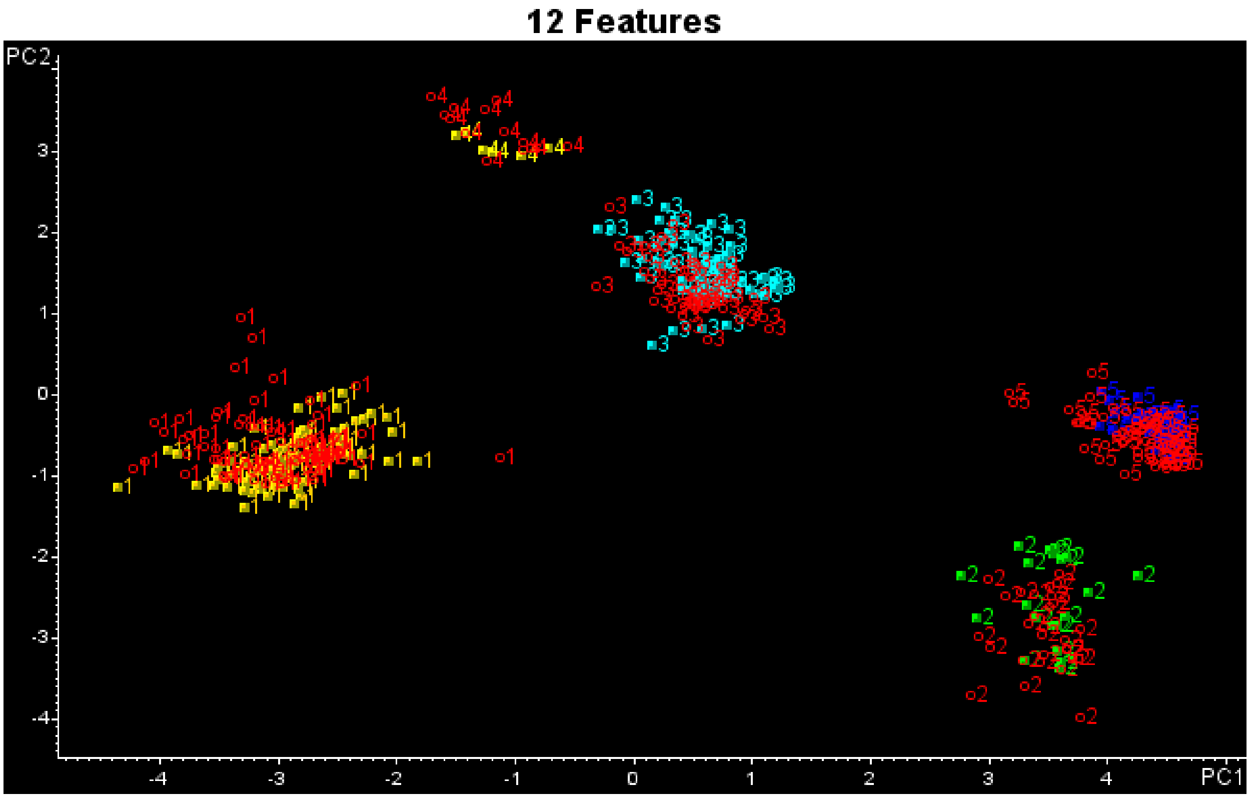


Figure 11. Validation set samples projected onto the PC plot of the Thermo-Nicolet training set samples defined by 12 spectral features identified by the pattern recognition GA.

**Table 4. LDA Analysis of 12 Spectral Features for Plant Group**

| Group | Samples | Misses | % Success | | Group | Samples | Misses | Success |
|---|---|---|---|---|---|---|---|---|
| 1 | 81 | 0 | 100 | | 1 | 90 | 0 | 100 |
| 2 | 22 | 0 | 100 | | 2 | 32 | 0 | 100 |
| 3 | 69 | 0 | 100 | | 3 | 50 | 0 | 100 |
| 4 | 6 | 0 | 100 | | 4 | 13 | 0 | 100 |
| 5 | 31 | 0 | 100 | | 5 | 57 | 0 | 100 |
| Total | 209 | 0 | 100 | | Total | 242 | 0 | 100 |

The next step in this study was to develop search prefilters to classify paint spectra by manufacturing plant of the paint sample. For each plant group, a search prefilter was developed to discriminate the spectra by manufacturing plant within a plant group. For this phase of the study, the spectral region used to formulate discriminants was from 1500cm$^{-1}$ to 600cm$^{-1}$. The carbonyl band, which was useful for discriminating the IR spectra by plant group, did not prove to be informative for discriminating spectra by manufacturing plant within a plant group due to the similarity of the shape and the intensity of the carbonyl band for manufacturing plants within a plant group. Because of the similarity of the IR spectra within a plant group, more powerful preprocessing methods were needed to extract information about manufacturing plant from the IR spectra of the clear coats. For this reason, wavelets were applied to each normalized spectrum using the MATLAB Wavelet toolbox 3.0.4 (MathWorks, Natick, MA). The wavelet packet transform [39] was implemented by iteratively passing each spectrum through pairs of wavelet filters at different levels of decomposition. Each wavelet filter is a scaled wavelet function that extracts high or low frequency signal components from the spectrum as wavelet coefficients. Each wavelet coefficient represents the similarity between a scaled wavelet and a section of the spectrum. The entire spectrum is passed through a scaled wavelet to generate a series of wavelet coefficients that comprise a wavelet packet. The high-pass wavelet filter is a low scaled wavelet (compressed wavelet with rapidly changing features) that extracts higher frequency signal components from the spectrum. The low-pass wavelet filter is a high scaled wavelet (stretched out smoother wavelet) that extracts lower frequency signal components from the spectrum. Each level of the filtering process involves a pair of high-pass and low-pass wavelet filters that break down the spectrum into a high frequency packet and a low frequency packet. Each wavelet packet in turn is broken down at the next level of decomposition using another pair of high-pass and low-pass wavelet filters. This process is continued until the required level of decomposition is achieved.

Wavelet coefficients from all nodes in the tree for each clear coat spectrum were organized as a data vector. For pattern recognition analysis, each wavelet coefficient was autoscaled such that it had a mean of zero and a standard deviation of one. In this phase of the study, the Symlet6 mother wavelet at the 8th level of decomposition, i.e., 8Sym6, was used to denoise and deconvolute each IR spectrum into 1080 wavelet coefficients. Criteria used to select this mother wavelet were based on its ability to extract information about the manufacturing plant from the data, which can then be exploited using the pattern recognition GA. There was a decrease in the ability of the pattern recognition GA to correctly classify the spectra when other wavelets such as the Daubachies 6 or 18 were used to denoise and deconvolute the data. This result can be explained by a well known empirical rule often applied to guide the selection of suitable wavelets for denoising data. If the signal contains sharp peaks or discontinuities, the use of the Haar or other compact wavelets would be indicated, whereas for signals that comprise broader peaks, a smoother wavelet such as the Symlet would be recommended. If the signal lies between these two extremes, such as mid-IR spectra, wavelets such as Symlet 4 through Symlet 8 would be expected to give good results.

Figure 12 shows a plot of the two largest principal components of the 19 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group (see Table 5). Each IR spectrum is represented as a point in the PC plot of the data. Plant 18 (Moraine OH) is well separated from the other manufacturing plants in the PC plot. The spectra from the other 6 manufacturing plants (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI) were super-imposable, which prevented further discrimination by manufacturing plant of these clear coats.

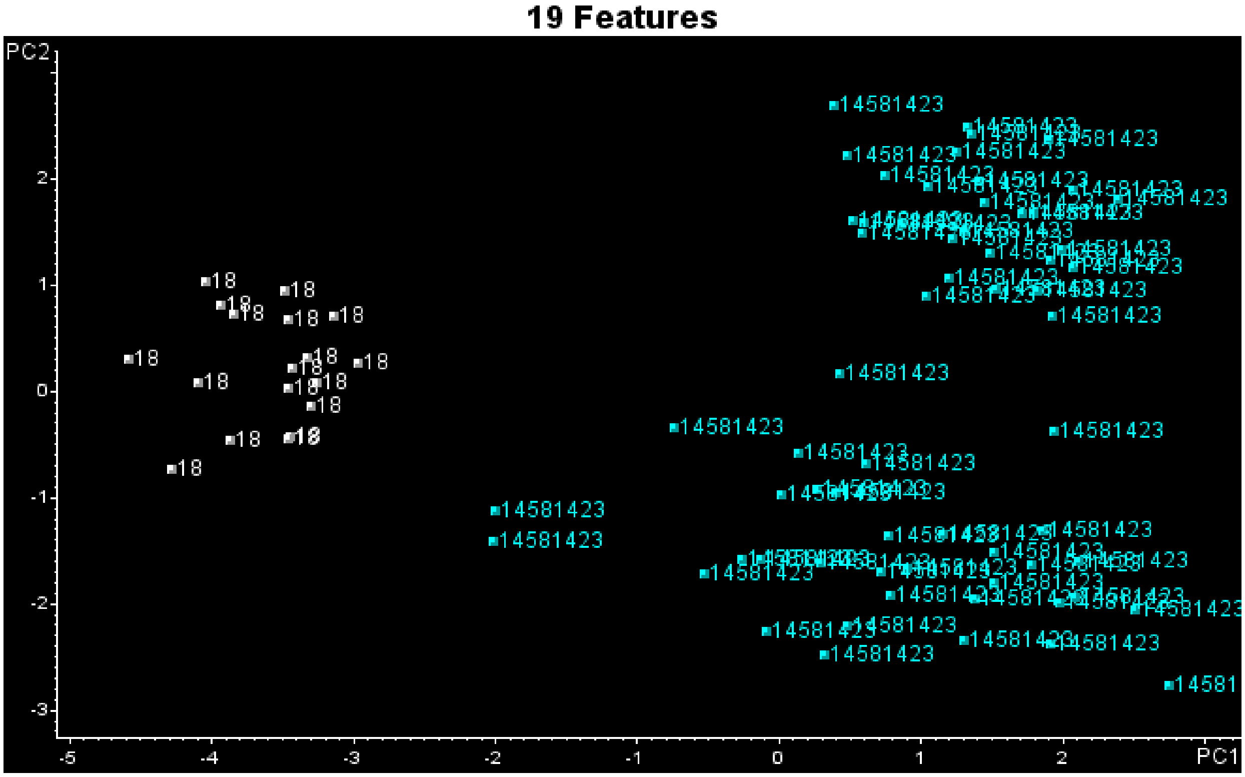


Figure 12. A plot of the two largest principal components of the 19 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group is shown. Each IR spectrum is represented as a point in the PC plot. Validations set samples are in red. 18 = Moraine OH, 14581423 = Arlington TX (1), Doraville GA (4), Fairfax KS (5), Fort Wayne IN (8), Lansing MI (14), Pontiac MI (23)

**Table 5. Training Set and Validation Set for Plant Group 1**

| Plants | Training set samples (Thermo-Nicolet) | Validation set samples (Bio-Rad) |
|---|---|---|
| 18 | 18 | 13 |
| 1,4,5,8,14,23 | 63 | 77 |
| Total | 81 | 90 |

A validation set of 90 IR spectra (see Table 5) was employed to assess the predictive ability of the 19 wavelet coefficients identified by the pattern recognition GA. We chose to map the 90 spectra directly onto the PC map defined by the 81 spectra of the training set and the 19 wavelet coefficients identified by the pattern recognition GA. Figure 13 shows the validations set samples projected onto the PC map developed from the training set data. Each projected sample lies in a region of the map with paint samples that have the same class label: either plant 18 or plants 1, 4, 5, 8, 14, and 23. Evidently, the pattern GA can identify wavelet coefficients characteristic of the manufacturing plant of a clear coat paint smear. This suggests that search prefilters developed from IR spectra of clear coats can be used to characterize paint smears by manufacturing plant or can identify a limited number of manufacturing plants associated with the clear coat paint layer.

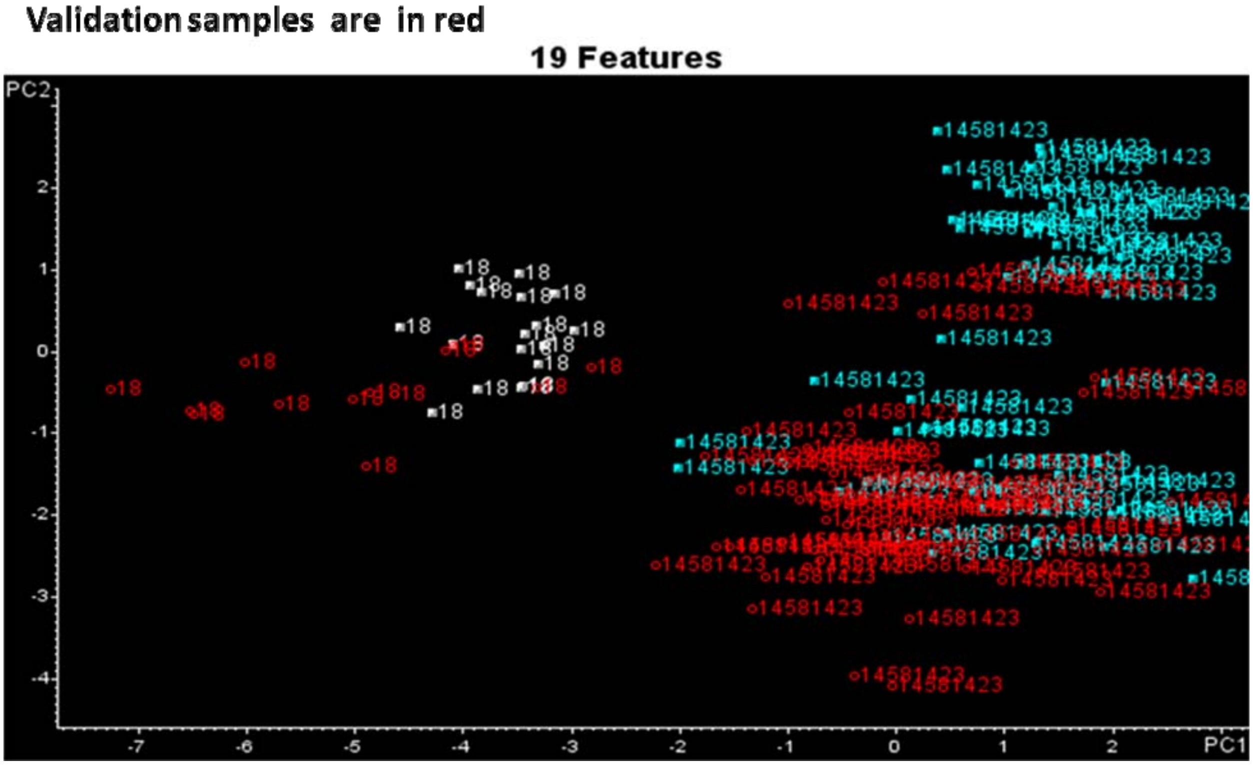


Figure 13. Projection of the validation set samples onto a plot of the two largest principal components of the 19 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group is shown. Each IR spectrum is represented as a point in the PC plot. Validation set samples are in red. 18 = Moraine OH, 14581423 = Arlington TX (1), Doraville GA (4), Fairfax KS (5), Fort Wayne IN (8), Lansing MI (14), Pontiac MI (23)

Figure 14 shows a plot of the two largest principal components of the 22 IR spectra of the training set and the 23 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the second plant group (see Table 6). Each IR spectrum is represented as a point in the plot. All 3 manufacturing plants (Bowling Green KY, Hamtramck MI, and Orion MI) are well separated from each other in the PC plot of the data. Figure 15 shows the validations set samples (see Table 6) projected onto the PC map developed from the 22 IR spectra of the training set and the 23 wavelet coefficients identified by the pattern recognition GA. Each projected sample lies in a region of the map with paint samples that have the same class label.

LDA was also used to classify the IR spectra in the training set. Table 7 summarizes the results of the LDA study for both the training set and validation set. Again, a classification success rate of 100% was achieved for the IR spectra in both the training and validation sets. For the IR spectra comprising the second plant group, it was necessary to truncate the last 15 points (610-600cm$^{-1}$) due to noise in the data. The two step procedure used to develop the search prefilters for manufacturing plant (which involved applying wavelets to decompose each spectrum into wavelet coefficients and using the pattern recognition GA to identify wavelet coefficients correlated with manufacturing plant) is well suited for the development of search prefilters to identify the source of a clear coat paint smear.

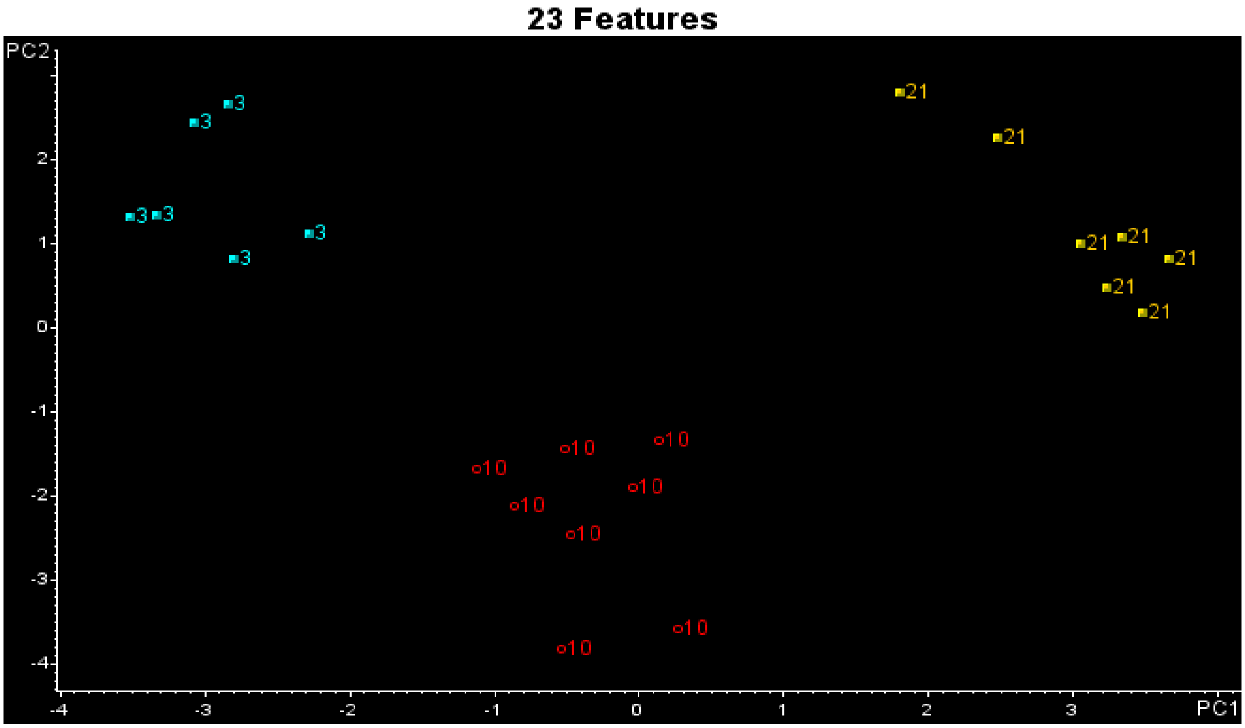


Figure 14. A plot of the two largest principal components of the 23 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the second plant group is shown. 3 = Bowling Green KY, 10 = Hamtramck MI, 21 = Orion MI

**Table 6. Training Set and Validation Set for Plant Group 2**

| Plants | Training Set Samples (Thermo-Nicolet) | Validation Set Samples (Bio-Rad) |
|---|---|---|
| 3 | 6 | 10 |
| 10 | 9 | 13 |
| 21 | 7 | 8 |
| Total | 22 | 31 |

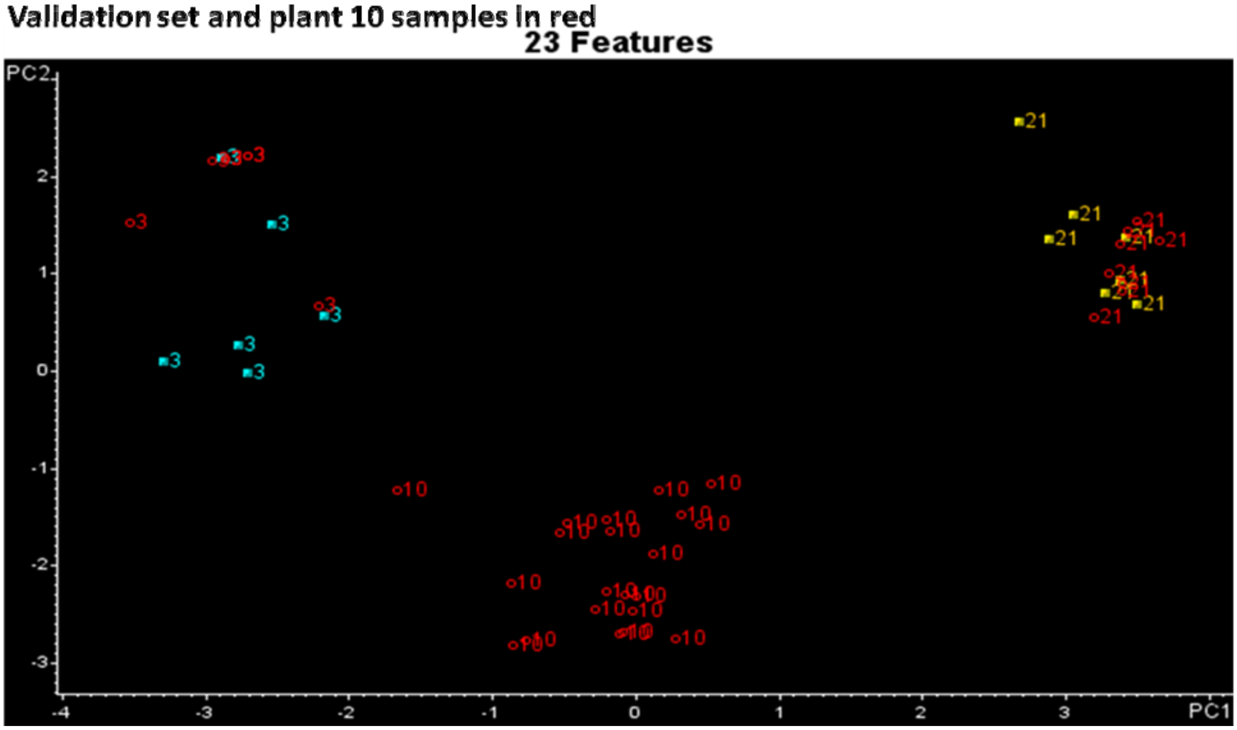


Figure 15. PC plot of the validations set samples (see Table 6) projected onto the PC map developed from the 21 IR spectra and 23 wavelet coefficients identified by the pattern recognition GA. 3 = Bowling Green KY, 10 = Hamtramck MI, 21 = Orion MI. Validation set samples and Plant 10 samples are in red.

**Table 7. LDA Results for Plant Group 2**

| Training - LDA | | | | Prediction - LDA | | | |
|---|---|---|---|---|---|---|---|
| Plant | Samples | Misses | % Success | Plant | Samples | Misses | % Success |
| 3 | 6 | 0 | 100 | 3 | 10 | 0 | 100 |
| 10 | 9 | 0 | 100 | 10 | 13 | 0 | 100 |
| 21 | 7 | 0 | 100 | 21 | 8 | 0 | 100 |
| Total | 22 | 0 | 100 | Total | 31 | 0 | 100 |

Figure 16 shows a plot of the two largest principal components of the 82 training set samples and the 9 wavelet coefficients identified by the pattern recognition GA for manufacturing plants comprising the third plant group (see Table 8). IR spectra from two manufacturing plants represented as 9 and 17 (Fremont CA and Lordstown OH) and trucks from Oshawa Ontario (Plant 22) cluster in distinct regions of the PC map of the data. GMC vehicles from Oklahoma

City (Plant 20) cluster in the same region of the map with Buicks from Oshawa Ontario (Plant 22). Vehicles from Linden NJ (Plant 16), trucks from Flint MI (Plant 6), trucks from Shreveport LA (Plant 25), Chevrolet cars from Oshawa Ontario (Plant 22), and Chevrolet trucks and cars from Oklahoma City (Plant 20) also form a distinct sample cluster. Two manufacturing plants, Oshawa Ontario (Plant 22), and Oklahoma City (Plant 20), have clear coat spectra that are distinct for specific models and lines. The net result is multiple clusters for automobiles or trucks from these two manufacturing plants.

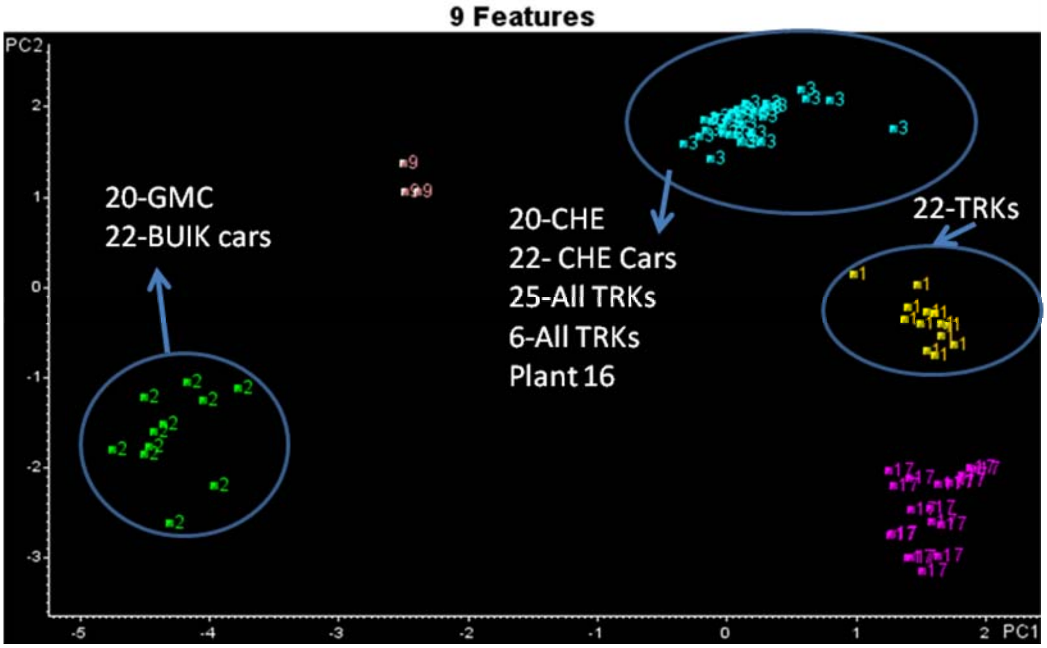


Figure 16. A plot of the two largest principal components of the 9 wavelet coefficients identified by the pattern recognition GA for manufacturing plants comprising the third plant group is shown. 2 = GMC (Oklahoma City) and Buick (Oshawa Ontario); 3 = Trucks (Flint MI), CHE and GMC (Linden NJ), Chevrolet (Oklahoma City), Chevrolet (Oshawa Ontario), and GMC (Shreveport LA); 9 = Fremont CA, 17 = Lordstown OH.

**Table 8. Training Set and Validation Set for Manufacturing Plants from Plant Group 3**

| Plants | Training Set Samples (Thermo-Nicolet) | Validation Set Samples (Bio-Rad) |
|---|---|---|
| 9 | 3 | 2 |
| 17 | 19 | 16 |
| 1 (Plant 22 trucks) | 13 | 9 |
| 2 (Plants 20-GMC, 22-Buick Cars) | 11 | 8 |
| 3 (Plants 6-all TRKS,16-all,20-CHE,22-CHE cars,25 all TRKS) | 36 | 37 |
| Total | 82 | 72 |

Figure 17 shows the validations set samples (see Table 8) projected onto the PC map developed from the 82 IR spectra of the training set and the 9 wavelet coefficients identified by the pattern recognition GA. Each projected paint sample lies in a region of the map with other paint samples that have the same class label.

In this study, paint samples from Plant Group 3, which were not part of the original study, were added to both the training set and validation set. These samples had been previously excluded from this study because they were identified as outliers due to excessive instrument noise. Water and carbon dioxide peaks were probably present in their spectra to varying amounts. These spectra were measured using one of the Bio-Rad instruments that had a peak in the region examined which was produced by the instrument probably due to a fault in a moving mirror. After rerunning these samples using the Thermo-Nicolet instrument, our data for Plant Group 3 was updated and the results were compared to a previous study which did not include these samples. As the results were the same, this supported our conclusion about the quality of this data and justified the decision to rerun these samples.

Because Plant Group 4 was a single plant (Janesville WI) and the IR spectra of all clear coats in Plant Group 5 were super-imposable, it was not necessary to develop additional search prefilters. Thus, a paint sample assigned to Plant Group 4 would be from the Janesville WI plant and a paint sample assigned to Plant Group 5 would be from the Ramos Arizpe (Mexico), Silao (Mexico), or Spring Hill Tennessee plants.

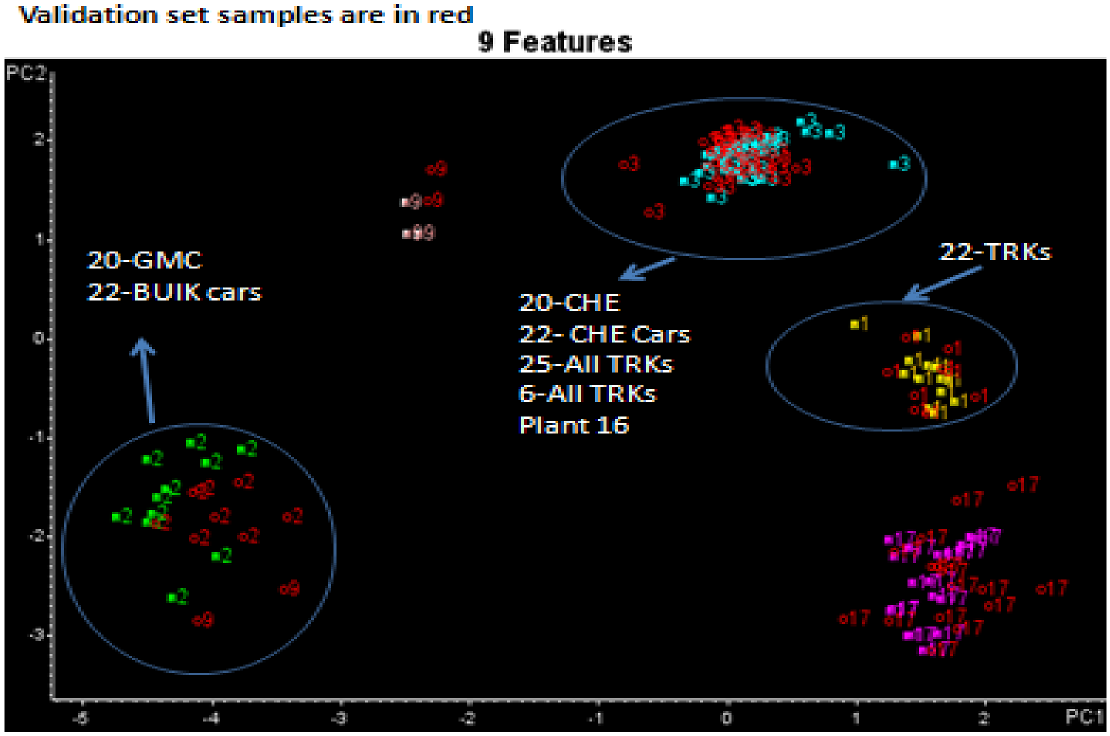


Figure 17. PC plot of the validation set samples (see Table 8) projected onto the PC map developed from the 82 IR spectra and 9 wavelet coefficients identified by the pattern recognition GA. 2 = GMC (Oklahoma City) and Buick (Oshawa Ontario); 3 = Trucks (Flint MI), CHE and GMC (Linden NJ), Chevrolet (Oklahoma City), Chevrolet (Oshawa Ontario), and GMC (Shreveport LA); 9 = Fremont CA, 17 = Lordstown OH. Validation set samples are in red.

In this study, a two step procedure for spectral pattern recognition is proposed. First, search prefilters are employed to divide the IR spectra of the clear coats into plant groups. A genetic algorithm for pattern recognition analysis is used to identify the discriminating wavelengths characteristic of each plant group. Second, search prefilters are developed for the IR spectra in each plant group to identify the specific manufacturing plant or the set of manufacturing plants that have similar IR spectra to the unknown. In this phase of the study, a wavelet packet tree is used to decompose each spectrum into wavelet coefficients that represent both high and low frequency components of the signal. A genetic algorithm for pattern recognition analysis is used to identify wavelet coefficients that contain information about the manufacturing plant of the paint samples. The proposed two step procedure is well suited for the development of search prefilters independent of the IR spectrometer used to generate the data. As these search prefilters are based on chemical information, they have the potential to facilitate spectral library searching in the PDQ database as the size of the library is culled to those spectra of paint samples obtained from manufacturing plants identified by the search prefilters.

**Pattern Recognition Assisted Infrared Library Searching of Clear Coat Paint Smears**

In this study, a novel approach to search IR spectral libraries of the PDQ database is presented. The first step in using this approach involves preprocessing the IR spectra using the discrete wavelet transform to enhance subtle but significant features in the spectral library data. In the second step, wavelet coefficients characteristic of the manufacturing plant (individual or plant group) are identified using a genetic algorithm for pattern recognition. Even in challenging trials where the samples evaluated were all the same model (General Motors) with a limited production year range (2000-2006), the manufacturing plant of an automobile from a paint chip could be correctly identified by search prefilters which were developed using spectral data from the PDQ library. Search prefilters developed as part of this study were successfully validated using a set of 10 blind samples.

For this study, 464 IR spectra of clear coats paint smears from metal substrates of automobiles and trucks assembled at 21 different General Motors (GM) assembly plants were collected using four different spectrometers: two Thermo-Nicolet 6700s FTIR spectrometers, one BioRad 40A and one BioRad 60 FTIR spectrometer. IR spectra of clear coats from bumpers and other plastic substrates were not used in the development of these search prefilters as plastic automotive components are often not subject to automotive paint at the same plant where the vehicle is assembled. Table 9 lists the 21 GM automotive plants used in this study. The GM manufacturing plants were divided into five major plant groups based on a visual analysis of the IR spectra of the clear coats.

Initial discrimination by plant group was done based on the characteristic absorption (stretching) band of carbonyl at around 1730 cm$^{-1}$ as Plant Group 2 and Plant Group 5 each showed a doublet peak, whereas Plant Group 1, Plant Group 3, and Plant Group 4 each showed a singlet peak within the aforementioned wavenumbers. Subsequent visual discrimination was made based on differences in spectral features resulting from differences in the vibrational modes in the finger print region of the IR spectra of the clear coat paint smears. The carbonyl band, which was useful for discriminating the IR spectra by plant group, did not prove to be informative for discriminating spectra by manufacturing plant within a plant group due to the similarity of the shape and intensity of the carbonyl band for each manufacturing plant within a plant group. For this reason, when developing search prefilters for both the plant group and the manufacturing plant, we chose to exclude the carbonyl band from the spectral range interrogated by our classifiers to promote simplicity as all spectra used by the search prefilters will contain the same number of points. In addition, we desired to maximize differences in the spectral features by minimizing the number of features that spectra had in common. Therefore, the spectral region used to formulate the discriminants was between 1500cm$^{-1}$ and 600cm$^{-1}$.

**Table 9. GM Plants used to Develop the Prefilters for Spectral Library Searching**

| Plant ID | Plant | Make | Line |
|---|---|---|---|
| 1 | ARL | CAD, CHE, GMC | SUB,YUK,ESD,CTA |
| 3 | BOW | CAD,CHE | CVT,XLR |
| 4 | DOR | PON | VTR,SIL,MTA,UPL,TAR |

| | | | |
|---|---|---|---|
| 5 | FAI | CHE,OLD,PON | GRA,MAL,ITR |
| 6 | FLI | CHE,GMC | SLV,SIE |
| 8 | FOR | CHE,GMC | SLV,SIE |
| 9 | FRE | GMC | VIB,TAC,PVB,COA,GPR |
| 10 | HAM | BUI,CAD,PON | BON,DEV,LUC,LES,SEV,ELD |
| 12 | JAN | GMC | CTA,SUB,YUK |
| 14 | LAN | PON | STS |
| 16 | LIN | CHE,GMC | BZR,JMY,S10 |
| 17 | LRD | PON | SFR,CAV,COB,PST |
| 18 | MOR | CHE,GMC,SAA | JMY,ENV,9S7,BZR,TBZ,SON |
| 20 | OKL | CHE, GMC | MAL,TBZ,ENV,EQU, XUV |
| 21 | ORI | PON,BUI | BON,PG6,LES,AUR, PKA |
| 22 | OSH | GMC,PON | ALL,REG |
| 23 | PON | CHE,GMC | SLV,SIE,SIL |
| 24 | RAM | BUI,CHE,PON | CAV,SFR,RZV,AZT,HHR |
| 25 | SHR | CHE,GMC | S10,COL,SON |
| 26 | SIL | CHE,GMC,SAA | AVL,SUB,YXL |
| 27 | SPH | STR | SSL,ION,SC1,SC2,SL1,VUE |

The manufacturing plants in each plant group are listed in Table 10. As Plant Group 4 consists of only a single manufacturing plant, the search prefilter developed for plant group will also serve to identify clear coat paint samples prepared in the Janesville WI plant. The 464 spectra that served as the training set for the development of the search prefilters for the PDQ database are summarized in Table 11. A hierarchical classification scheme was used to develop the search prefilters for the PDQ database. First, an unknown is classified as to its plant group and then a search prefilter is used to identify a specific plant or plants within the plant group to which membership of the unknown is assigned.

**Table 10. Manufacturing Plants Comprising Each Plant Group**

| Plant Group | Plant ID Number | Manufacturing Plant |
|---|---|---|
| 1 | 1, 4, 5, 8, 14, 18, 23 | ARL, DOR, FAI, LAN, MOR, PON |

| | | |
|---|---|---|
| 2 | 3, 10, 21 | BOW, HAM, IRI |
| 3 | 6, 9, 16, 17, 20, 22, 25 | FLI, FRE, LIN, LRD, OKL, OSH, SHR |
| 4 | 12 | JAN |
| 5 | 24, 26, 27 | RAM, SIL, SPH |

**Table 11. Training Set for Plant Group Search Prefilter**

| Group | Number of Training Set Samples |
|---|---|
| 1 | 164 |
| 2 | 54 |
| 3 | 141 |
| 4 | 21 |
| 5 | 84 |
| Total | 464 |

Data preprocessing was crucial in this study to ensure a successful analysis of the spectral data. IR spectra of the clear coats were not properly aligned along their x or y-axes as these spectra were collected on four different spectrometers. There were differences in alignment of the optical systems as these spectrometers were from different vendors and were manufactured in different years. Differentiating manufacturing plants within a plant group required finding small differences in IR spectra that are very similar.

The first step in this study was to align the spectra along the wavelength axis using OMNIC. Differences in the spectral resolution of the four instruments used (two Thermo-Nicolet 6700s FTIR spectrometers, one BioRad 40A and one BioRad 60 FTIR spectrometer) resulted in differences in the number of points in each spectrum for the same wavenumber range. The number of points collected per spectrum using the two Thermo-Nicolet instruments varied from 1878 points to 1958 points, whereas all spectra from the two BioRad instruments were represented by 1944 points. Band shifting was also observed in spectra collected on the Thermo Nicolet instrument. In order to remedy these two problems, each spectrum was normalized to the helium neon laser frequency of 15798.0. An instrumental line function representative of the Thermo Nicolet instruments and developed in OMNIC was applied to the Bio-Rad spectra to ensure that all measurements made by the Bio-Rad instrument were comparable to spectra collected using the two Thermo-Nicolet instruments. This ensured wavelength alignment along the x-axis for all clear coat spectra. After this preprocessing, each spectrum was comprised of 1869 points for the entire mid-IR range of 400 cm$^{-1}$ to 4000 cm$^{-1}$. The spectral region of 600 cm$^{-1}$ to 1500 cm$^{-1}$, which is comprised of 468 points, was subject to pattern recognition analysis.

Using the wavelet packet transform, the 464 spectra that served as the training set for the development of search prefilters for the PDQ database were passed through two scaling filters: a high pass filter and a low pass filter. This decomposition process, which utilizes wavelet coefficients that represent the high and low frequency components of the signal, was iterated using successive wavelet packets until the required level of signal decomposition was achieved. Applying the wavelet packet transform to the spectra is similar to using a magnifying lens to find minute details in spectra. Wavelets at different scales correspond to lenses of different magnifying power. The Wavelet packet tree separates signal from noise by isolating the signal in specific wavelet coefficients which were identified in this study using a genetic algorithm for pattern recognition. The pattern recognition GA identifies a set of wavelet coefficients that optimize the separation of the classes in a plot of the two or three largest principal component of data. Because principal components maximize variance, the bulk of the information encoded by these coefficients will be about differences between the classes in the training set. Chance classification will not be a serious problem since the bulk of the variance or information content of the feature set selected is about the classification problem of interest. The principal component plot functions as an embedded information filter for the genetic algorithm. Wavelet coefficients are selected based on their PC plots. A good PC plot can only be generated using coefficients whose variance or information content is primarily about class differences. Hence, PCA limits our search to these types of coefficients, thereby significantly reducing the size of the search space.

The first step in this study was to apply PCA to the wavelet transformed IR spectra. Using this procedure is analogous to finding a new coordinate system that is better at displaying the information content of the data than axes defined by the wavelet coefficients. This new coordinate system is linked to variation in the data. Often, only the two or three largest principal components are necessary to explain most of the information present in spectral data.

The Symlet6 mother wavelet at the 8th level of decomposition was used to deconvolve the spectra prior to PCA. Selection of this mother wavelet was based on its ability to extract information about the manufacturing plant from the data. The ability to correctly classify the data decreased when other mother wavelets such as the Daubachies 6 or Daubachies 18 were used to deconvolve the spectra. For spectra that contain sharp peaks, the use of the Haar or other compact wavelets is indicated, whereas for spectra containing broader peaks, a smoother wavelet such as the Symlet is recommended. For mid-IR spectra which lie between these extremes, wavelets such as Symlet 4 through Symlet 8 generally give good results.

Prior, the wavelet coefficients from all nodes in the tree generated 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. Autoscaling removes inadvertent weighting of the wavelet coefficients that otherwise would occur due to differences in magnitude among the coefficients. This ensures that each wavelet coefficient has an equal weight in the pattern recognition analysis of the data.

Figure 18 shows a PC plot of the two largest principal components of the 464 IR spectra and the 1080 wavelet coefficients representing each IR spectrum. Each paint sample is represented as a point in the PC map 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 overlap of the clear coats from each plant group in the map of the data is evident.

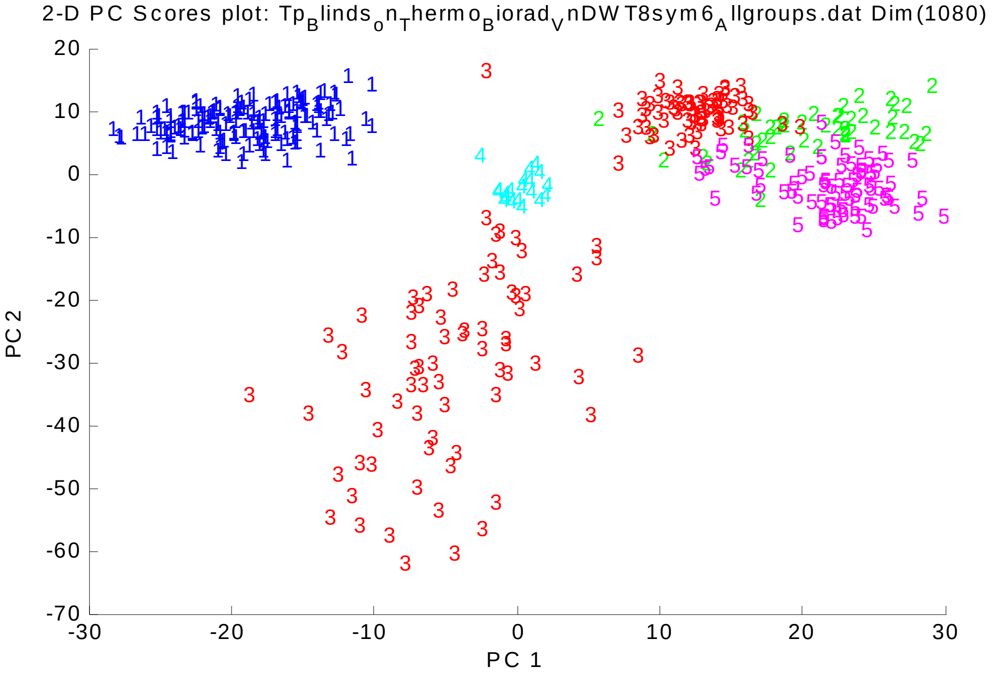


Figure 18. PC plot of the two largest principal components of the 464 IR spectra and the 1080 wavelet coefficients comprising the training set. Each paint sample is represented as a point in the PC plot of the data (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

The pattern recognition GA was used to identify wavelet coefficients characteristic of the plant group. Informative coefficients were identified by sampling key feature subsets, scoring their PC plots, and tracking those plant groups and/or IR 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 14 wavelet coefficients whose PC plot (see Figure 19) showed clustering of the IR spectra on the basis of plant group.

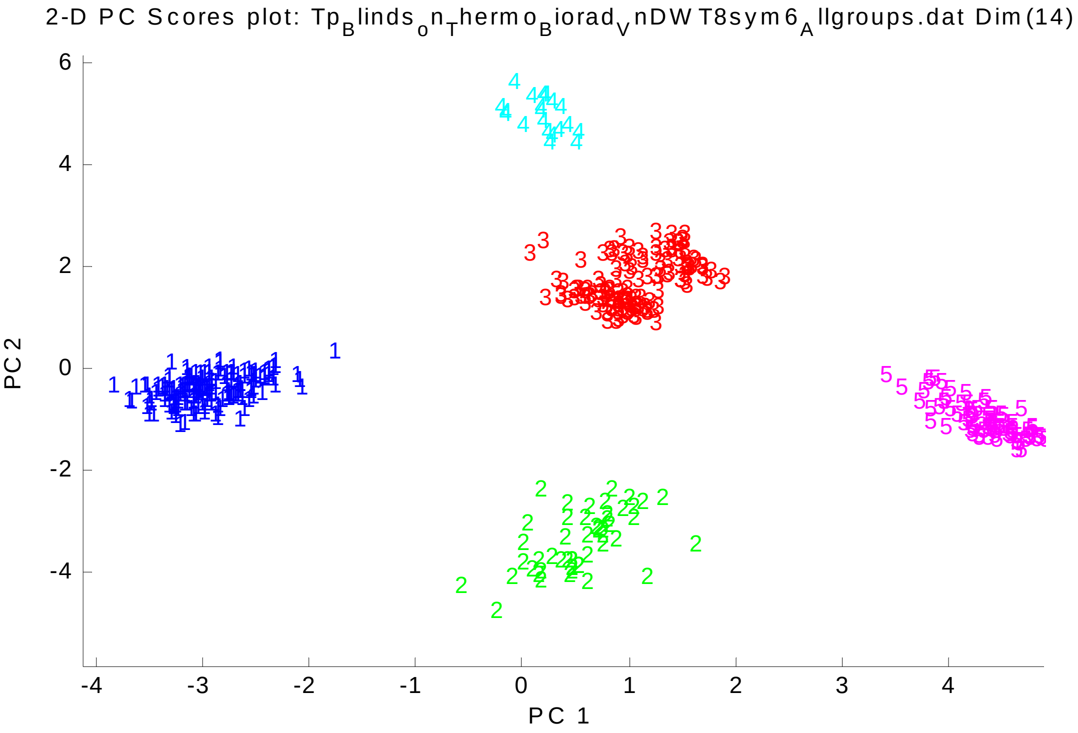


Figure 19. PC plot of the two largest principal components of the 464 IR spectra of the training set and the 14 wavelet coefficients identified by the pattern recognition GA. Each paint sample is represented as a point in the PC plot of the data (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

To assess the predictive ability of the 14 wavelet coefficients identified by the pattern recognition GA, a validation set of 10 blind samples was used. IR spectra from the validation set were projected directly onto the PC map developed from the 463 IR spectra of the training set and 14 wavelet coefficients identified by the pattern recognition GA. Figure 20 shows the projection of the blind (validation set) samples onto the PC map of the training set data. 9 of the 10 projected samples are located in a region of the map occupied by samples that have the same class label.

Figure 21 shows an IR spectrum of the misclassified blind sample (Sample #8) compared to an IR spectrum of a clear coat paint sample of the same make, model, line, year and manufacturing plant from the PDQ database. The dissimilarity between these two IR spectra is evident. The blind and library samples were measured on a Thermo-Nicolet 6700s FTIR spectrometer equipped with a deuterated triglycine sulfate detector with a resolution of 4 cm$^{-1}$. For the measured absorbance to be very nearly equal to the true absorbance, the interferograms were multiplied by the Norton-Beer medium apodization function before application of the Fourier transform.

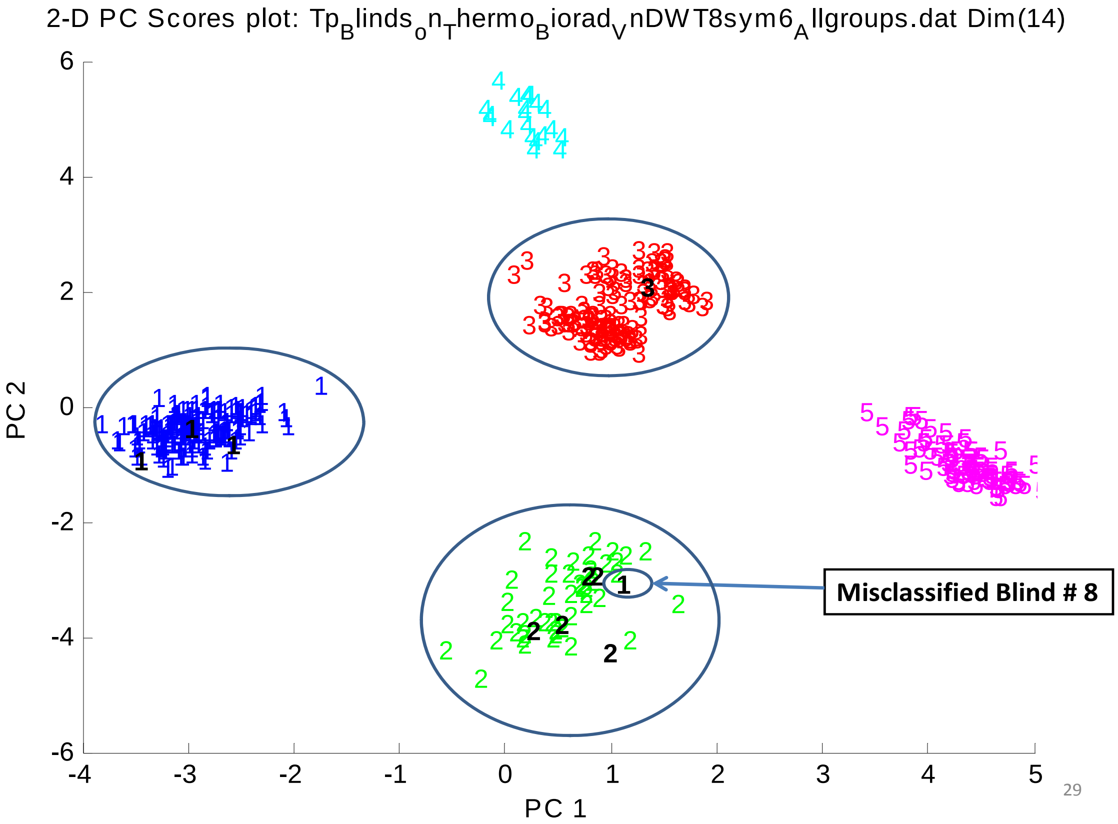


Figure 20. Projection of the blind (validation set) samples onto the PC map of the training set data. 9 of the 10 projected IR spectra are located in a region of the map occupied by spectra possessing the same class label.

Figure 22 shows the first hit from the PDQ database for blind sample #8 using the library search algorithm in OMNIC (Thermo Nicolet Corporation). Of all the modes available for the library search in OMNIC (correlation, absolute derivative, square derivative, absolute difference and square difference), correlation gave the most useful results in terms of the number of hits and the hit quality index. Consequently, the correlation mode was applied in the library search for all the blind samples using OMNIC. Using only the fingerprint region for the search gave the same results as using the entire mid-IR region after eliminating the noise present in the clear coat spectra caused by the diamond transmission cell. The hit obtained from the PDQ database is from a clear coat on a plastic bumper. In most cases, a clear coat is applied to plastic automotive components in the plant that manufactures the plastic component, not in the plant for assembling the vehicle. (For metal automotive components, the paint system for the component is applied in the same plant that assembles the vehicle.) Furthermore, bumpers and other plastic automotive components are replaceable and may not have the original automotive paint system used by the vehicle. For these reasons, the training set used in this study is limited to automotive paint from metal components.

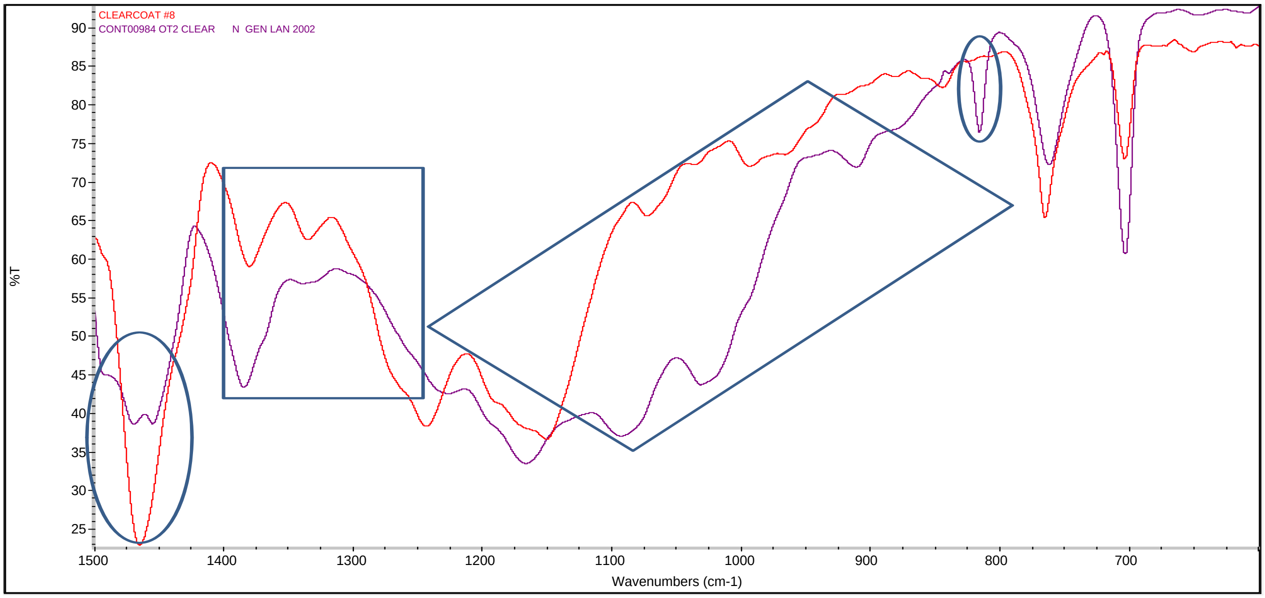


Figure 21. An IR spectrum of the misclassified blind sample (Sample #8) compared to an IR spectrum of a clear coat paint sample of the same make, model, line, year and manufacturing plant from the PDQ database.

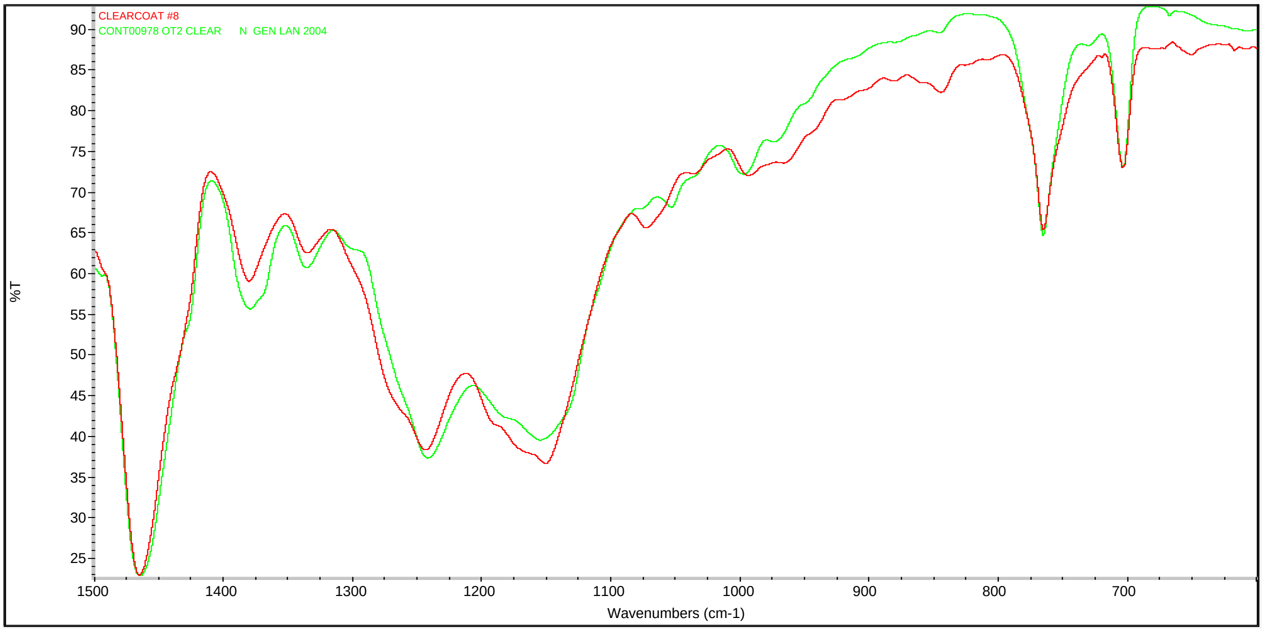


Figure 22. An IR spectrum of the misclassified blind sample and the first hit from the PDQ database using OMNIC

The next step in this study was to develop search prefilters to identify paint spectra of the blind clear coat paint samples by manufacturing plant. For each plant group, a search prefilter was developed to discriminate the spectra by manufacturing plant within a plant group. Figure 23 shows a plot of the two largest principal components of the 33 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group (see Table 10). Each IR spectrum is represented as a point in the PC plot of the data. Plant 18 (Moraine OH) is well separated from the other manufacturing plants in the PC plot. The spectra from the other 6 manufacturing plants (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI) were similar, which prevented further discrimination by manufacturing plant of these clear coats. However, the manufacturing plants comprising paint samples represented by the symbol 1 in the PC plot could be discriminated by year. Furthermore, projecting the blind samples assigned to Plant Group 1 onto the PC map showed that each projected sample lies in a region of the map with paint samples that have the same production year range and class label: either plant 18 or plants 1, 4, 5, 8, 14, and 23. This result indicates that search prefilters developed from IR spectra of clear coats can be used to characterize paint smears by manufacturing plant or by a limited number of plants and can specify a limited production year range for these samples.

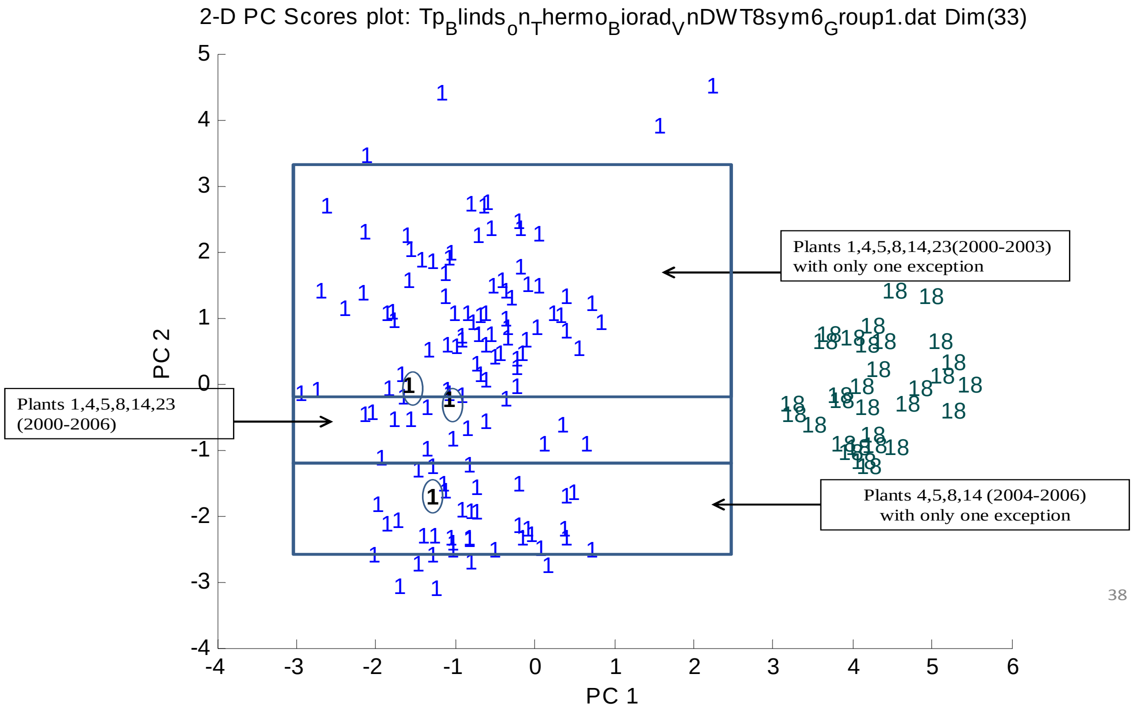


Figure 23. PC plot of the two largest principal components of the 164 IR spectra of the training set and the 33 wavelet coefficients identified by the pattern recognition GA. Each clear coat paint sample is represented as a point in the PC plot of the data (1 = ARL, DOR, FAI, FOR, LAN, PON, and 18 = MOR). Blind samples are in black and are circled

Figure 24 shows a plot of the two largest principal components of the 54 IR spectra and the 23 wavelet coefficients identified by the pattern recognition GA for clear coats from Plant Group 2 (see Table 10). Again, each IR spectrum is represented as a point in the PC plot. All 3 manufacturing plants (Bowling Green KY, Hamtramck MI, and Orion MI) are well separated from each other in the PC plot of the data.

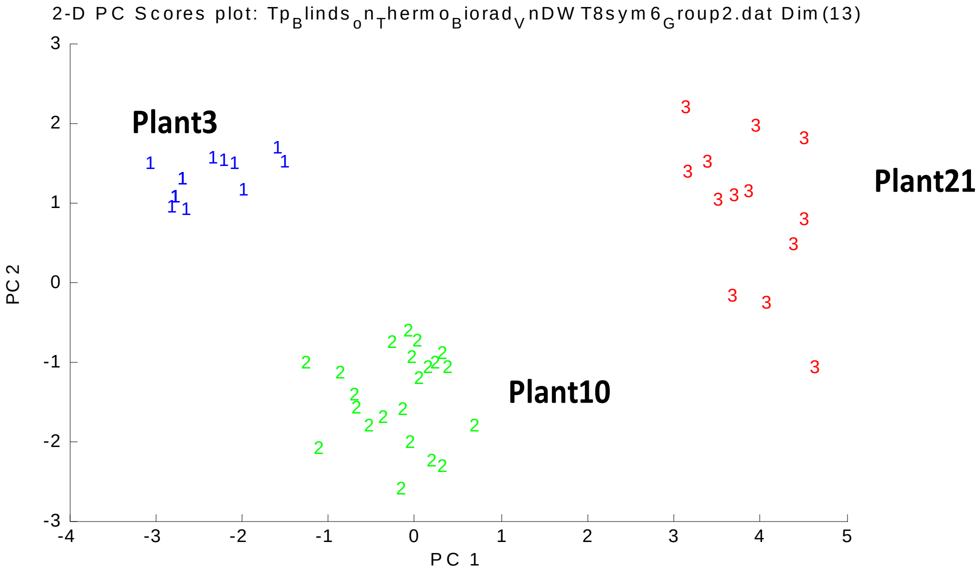


Figure 24. PC plot of the 54 IR spectra and 13 wavelet coefficients identified by the pattern recognition GA for spectra from Plant Group 2. 1 = Bowling Green KY, 2 = Hamtramck MI, 3 = Orion MI

Figure 25 shows the 5 blind (validation set) samples assigned to Plant Group 2 projected onto the PC plot of the second plant group. 4 of the 5 blind (projected) samples lie in a region of the map, which contain clear coat paint samples from the same manufacturing plant. Blind sample 4, which is from the Hamtramck Michigan Plant, was misclassified as an Orion Michigan Plant sample. There was only one IR spectrum of this model (which was a 2006 Buick Lucerne) in the PDQ database. Since the Lucerne appears to be a low volume vehicle, there is less chance of accurately characterizing it for casework.

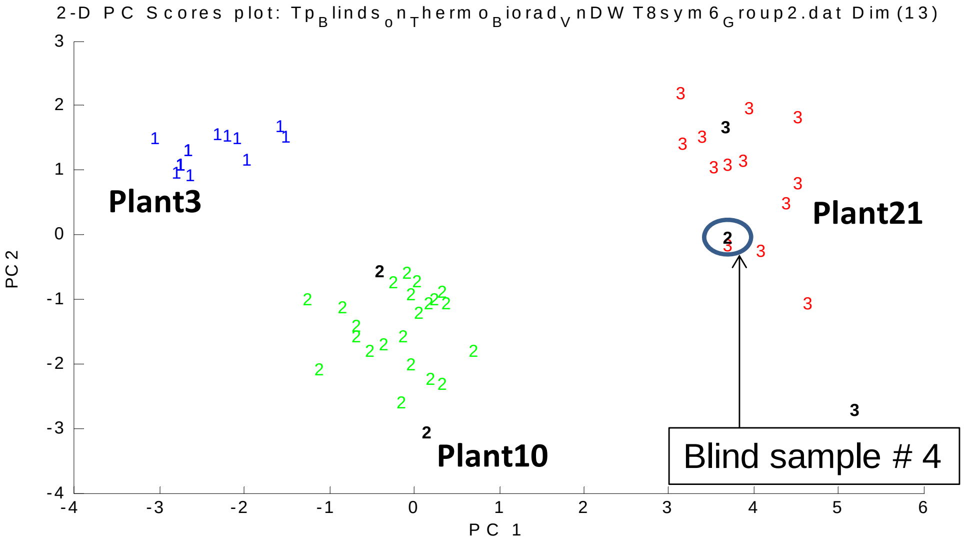


Figure 25. 4 blind (validation set) samples assigned to Plant Group 2 projected onto the PC plot of the second plant group. 1 = Bowling Green KY, 2 = Hamtramck MI, and 3 = Orion MI.

Figure 26 shows an IR spectrum of the misclassified Hamtramck Michigan Plant sample compared to the only IR spectrum of the same model, line, year, and assembly plant in the database. The peak present in the library spectrum (which is circled) but not present in the blind sample nullified the match. Figures 27 and 28 show the misclassified Hamtramck Michigan plant sample compared to the average IR spectrum from the Hamtramck MI plant and from the Orion Michigan plant. Although the match is better for Orion Michigan (the plant to which the blind sample has been assigned by the search prefilter), the similarity between the IR spectra for these two plants underscores the difficulties and challenges of matching clear coats to their respective manufacturing plant using search prefilters.

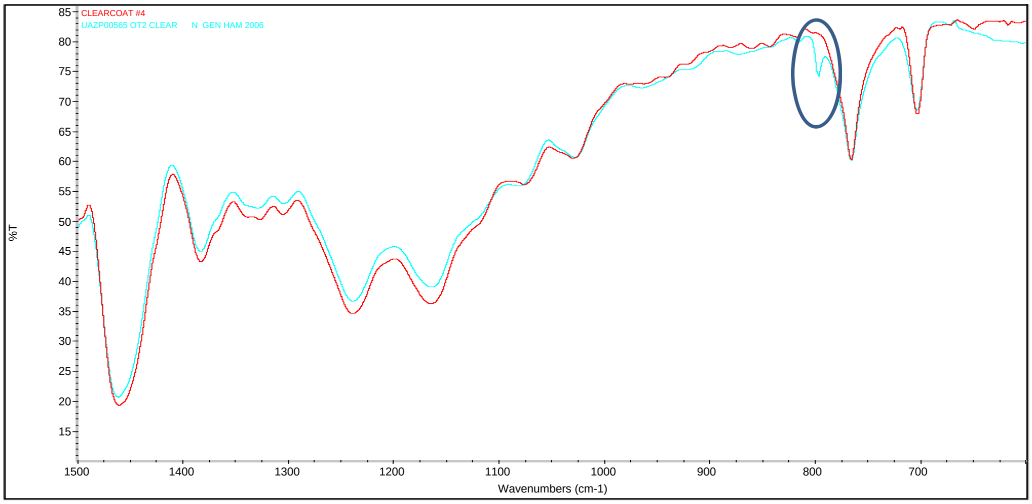


Figure 26. IR spectrum of the misclassified Hamtramck Michigan Plant sample (red) compared to the only IR spectrum of the same make, model, line, year and manufacturing plant in the database (cyano).

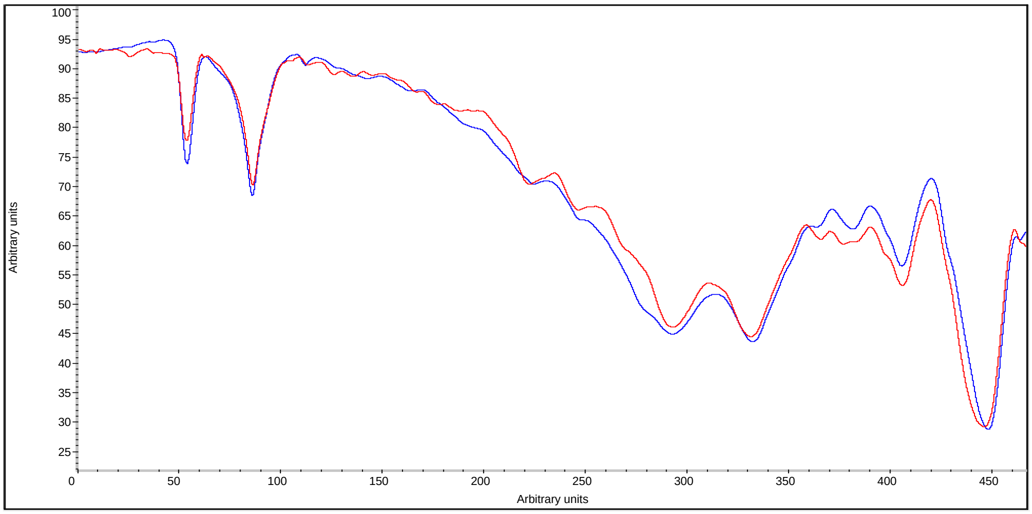


Figure 27. Misclassified Hamtramck Michigan plant sample (red) compared to the average IR spectrum (blue) from the Hamtramck MI plant

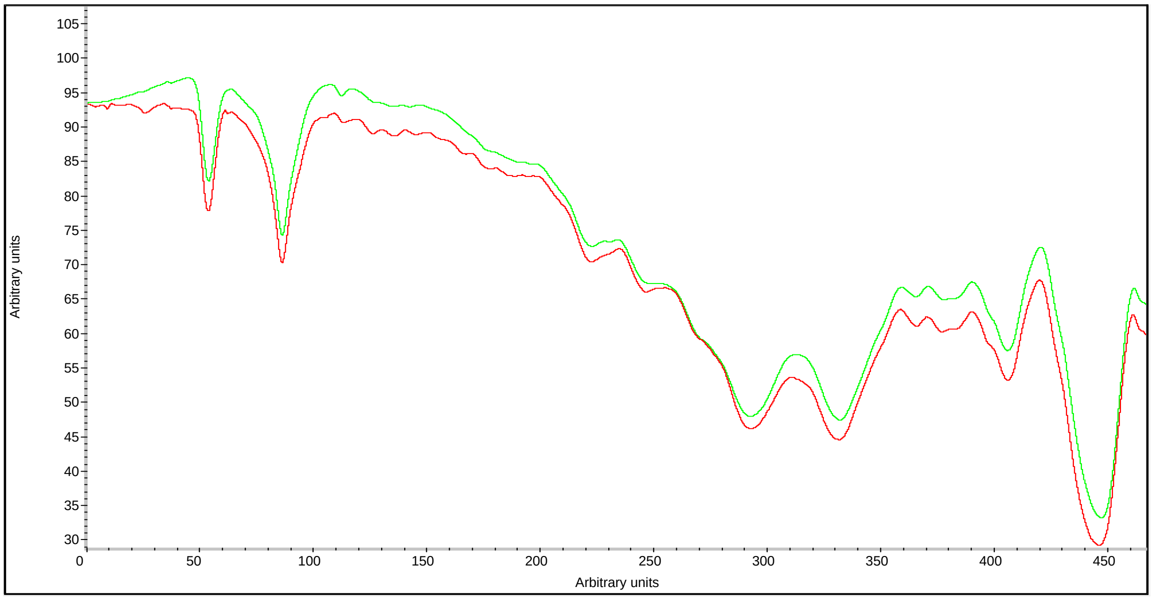


Figure 28. Misclassified Hamtramck Michigan plant sample (red) compared to the average IR spectrum (green) from the Orion Michigan plant

Figure 29 shows a plot of the two largest principal components of the 141 IR spectra and 13 wavelet coefficients identified by the pattern recognition GA for clear coats from Plant Group 3. Clustering of the spectra by manufacturing plant, model, and line in the PC plot is evident for the paint samples comprising this training set. Fremont California (Plant 9) and Lordstown Ohio (Plant 17) form distinct clusters in the PC plot of the data as do the trucks from Oshawa (Plant 22). The Buicks from Oshawa Ontario lie in another cluster with SUVs from the Oklahoma City Plant. Chevrolets from the Oklahoma City Plant and from Oshawa, GMC trucks from Shreveport Louisiana and the Flint Michigan Plant, and GMC and Chevrolet trucks from Linden, NJ form another cluster. The lone blind (validation set) sample assigned to Plant Group 3 is projected onto the PC plot in a region that contains clear coat paint samples from the same manufacturing plant.

A summary of the results obtained for the 10 blind samples is shown in Table 12. 8 of the 10 blind samples were correctly classified. One misclassified sample is probably a clear coat paint smear from a bumper and the other misclassified sample is from a manufacturing plant that is not well represented in the PDQ library. The search prefilters were developed from Bio-Rad and Thermo Nicolet IR spectra of paint samples aligned along both the x and y-axes using OMNIC software. This allowed for all spectra to be compared even if they were collected on different spectrometers.

.

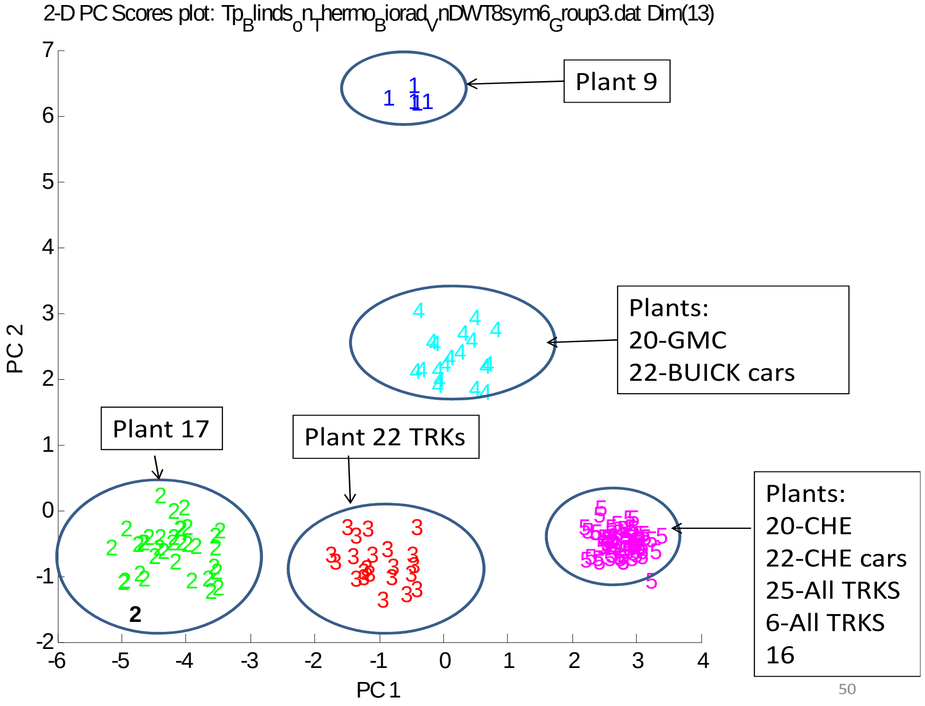


Figure 29. Plot of the two largest principal components of the 141 IR spectra and 13 wavelet coefficients identified by the pattern recognition GA for clear coats from Plant Group 3.

Table 12. Summary of the Results Obtained for the 10 Blind Samples

| Blind Sample | Assigned Plant Group | Assigned Plant(s) | ID of Blind Sample |
|---|---|---|---|
| 1 | 2 | 10 | 10 |
| 2 | 1 | 4, 5, 8, or 14 (2004-2006) | 14(2004) |
| 3 | 2 | 21 | 21 |
| 4 | 2 | 21 | 10 (wrongly classified at Plant level |
| 5 | 2 | 21 | 21 |
| 6 | 3 | 17 | 17 |
| 7 | 1 | 1,4,5,8,14 or 23 | 5 |
| 8 | 2 | NA | 14 (Wrongly classified at group level) |
| 9 | 2 | 10 | 10 |
| 10 | 1 | 1,4,5,8,14 or 23 | 14 |

**Search Prefilters Using Autocorrelated Infrared Absorbance Spectra**

The objective of this study is to investigate the feasibility of using the autocorrelation transformation in combination with pattern recognition methods to develop search prefilters for spectral library matching. The use of the autocorrelation transformation addresses many problems encountered with classification transfer as it eliminates translational differences between spectra along the wavelength axis. This transformation is also sensitive at distinguishing subtle but significant features in the data such as minor peaks, shoulders, and peaks with unique shapes. Dunn in previous studies [40-42] has demonstrated that peak shifts and other related spectral alignment problems can be obviated using the autocorrelation transformation. The autocorrelation transformation produces a histogram for each IR spectrum, which would be a more useful representation of spectra for pattern recognition involving data collected on different spectrometers. Dunn in his previous studies directly sampled points from these histograms using SIMCA pattern recognition to identify the informative regions. In our study, each histogram is first deconvolved to better capture signal by Coiflet wavelets with the informative wavelet coefficients identified using a genetic algorithm for pattern recognition.

In this study, 464 IR spectra of clear coats from metal substrates of vehicles (automobiles and trucks) assembled at 21 GM plants were collected using four spectrometers: two Thermo-Nicolet 6700s FTIR spectrometers, one BioRad 40A and one BioRad 60 FTIR spectrometer. IR spectra of clear coats from bumpers or from other plastic substrates were not used in the development of these search prefilters as plastic automotive components are often subjected to automotive paint at the plant that manufactures the component, not at the plant where the vehicle is assembled. Table 13 lists the 21 GM automotive plants used in this study. The 21 GM manufacturing plants were divided into five major plant groups based on visual analysis of the IR spectra of the clear coats.

One major challenge encountered when preprocessing the data was to develop a method that would obviate differences in the number points of the IR spectra of the clear coat samples as the spectra were generated using spectrometers from different vendors. Most likely, differences in the number of points, even with the same resolution, are caused by differences in alignment of the optical systems within the mid-IR range of the electromagnetic spectrum. To address this problem, OMNIC (Thermo-Nicolet) was set to yield 1869 points for a 4 cm$^{-1}$ within the mid-IR range (4000 cm$^{-1}$ to 400 cm$^{-1}$). This set of conditions was used to import all IR spectra collected using the Thermo-Nicolet and Bio-Rad spectrometers. This enabled us to assess the suitability of using autocorrelated IR spectra to develop search prefilters for the PDQ database.

A hierarchical approach to classification was taken because of the large number of manufacturing plants to be discriminated. In addition, the IR spectra of some manufacturing plants are quite distinctive whereas other manufacturing plants possessed similar IR signatures. As it was not possible to identify a set of features that could simultaneously differentiate all 21 plants, we chose to pursue a two-step approach that allowed us to first differentiate the manufacturing plants by plant group and then to apply the power of the pattern recognition GA to classify the IR spectra by manufacturing plant within each plant group.

Formulation of the plant groups was based on the characteristic absorption band of the carbonyl functional group at around 1709 cm$^{-1}$ as plant groups 2 and 5 each exhibited a doublet, whereas plant groups 1, 3, and 4 each showed a singlet at the aforementioned wavenumber. The doublet can probably be attributed to a polyurethane shoulder indicating a different chemical formulation than a clearcoat containing a carbonyl alone. Subsequent visual discrimination between plant groups was made based on differences in spectral features resulting from different vibrational modes in the fingerprint region of the IR spectra of these clear coats. The carbonyl band, which was useful for characterizing the IR spectra by plant group, was not informative for discriminating spectra by manufacturing plant within each plant group due to the similarity of the shape and intensity of the carbonyl band for each assembly plant within the same plant group. For this reason, we chose to exclude the carbonyl band from the spectral range interrogated by our search prefilters when developing classifiers both for plant group and for assembly plant. This simplified the data analysis problem as all spectra used by the search prefilters contained the same number of points. In addition, we wanted to maximize differences in the spectra by minimizing the number of features that all spectra had in common. Therefore, the spectral region used to formulate discriminants in this study was between 1500cm$^{-1}$ and 600cm$^{-1}$.

**Table 13. GM Plants used to Develop the Prefilters for Spectral Library Searching**

| Plant ID | Plant | Make | Line |
|---|---|---|---|
| 1 | ARL | CAD, CHE, GMC | SUB,YUK,ESD,CTA |
| 3 | BOW | CAD,CHE | CVT,XLR |
| 4 | DOR | PON | VTR,SIL,MTA,UPL,TAR |
| 5 | FAI | CHE,OLD,PON | GRA,MAL,ITR |
| 6 | FLI | CHE,GMC | SLV,SIE |
| 8 | FOR | CHE,GMC | SLV,SIE |
| 9 | FRE | GMC | VIB,TAC,PVB,COA,GPR |
| 10 | HAM | BUI,CAD,PON | BON,DEV,LUC,LES,SEV,ELD |
| 12 | JAN | GMC | CTA,SUB,YUK |
| 14 | LAN | PON | STS |
| 16 | LIN | CHE,GMC | BZR,JMY,S10 |
| 17 | LRD | PON | SFR,CAV,COB,PST |
| 18 | MOR | CHE,GMC,SAA | JMY,ENV,9S7,BZR,TBZ,SON |
| 20 | OKL | CHE,GMC | MAL,TBZ,ENV,EQU, XUV |
| 21 | ORI | PON,BUI | BON,PG6,LES,AUR, PKA |
| 22 | OSH | GMC,PON | ALL,REG |
| 23 | PON | CHE,GMC | SLV,SIE,SIL |
| 24 | RAM | BUI,CHE,PON | CAV,SFR,RZV,AZT,HHR |
| 25 | SHR | CHE,GMC | S10,COL,SON |
| 26 | SIL | CHE,GMC,SAA | AVL,SUB,YXL |
| 27 | SPH | STR | SSL,ION,SC1,SC2,SL1,VUE |

Assembly plants in each plant group are listed in Table 14. As Plant Group 4 consists of only a single manufacturing plant, the search prefilter developed for plant group also serves to identify clear coats from the Janesville WI plant. The 464 IR spectra that comprise the training set for the development of the search prefilters for the PDQ database are summarized in Table 15. The classification scheme used to implement these search prefilters for the PDQ database involves classifying an unknown by plant group and then identifying the specific plant or plants within the plant group to which the unknown sample is assigned.

**Table 14. Manufacturing Plants Comprising Each Plant Group**

| Plant Group | Plant ID Number | Manufacturing Plant |
|---|---|---|
| 1 | 1, 4, 5, 8, 14, 18, 23 | ARL, DOR, FAI, LAN, MOR, PON |
| 2 | 3, 10, 21 | BOW, HAM, IRI |
| 3 | 6, 9, 16, 17, 20, 22, 25 | FLI, FRE, LIN, LRD, OKL, OSH, SHR |
| 4 | 12 | JAN |
| 5 | 24, 26, 27 | RAM, SIL, SPH |

**Table 15. Training Set for Plant Group Search Prefilter**

| Group | Number of Training Set Samples |
|---|---|
| 1 | 164 |
| 2 | 54 |
| 3 | 141 |
| 4 | 21 |
| 5 | 84 |
| Total | 464 |

The Coiflet 4 mother wavelet at the 4th level of decomposition was used to deconvolve the autocorrelated spectra prior to PCA. Selection of this mother wavelet was based on its ability to extract information about the assembly plant from the autocorrelated data. For spectra that contain sharp peaks, the use of the Haar or other compact wavelets is indicated, whereas for spectra that contain broader peaks, a smoother wavelet such as the Symlet is recommended. For autocorrelated data, we selected the Coiflet mother wavelet because it is symmetric and is well suited to represent autocorrelated spectra, which is also symmetric. Our previous experience with wavelets indicates that matching the shape of the mother wavelet with the spectral bands to be deconvolved is crucial for the successful application of wavelets (see Figure 30). The fourth level of decomposition was used since the high frequency component extracted after the 4th level was smoother than the previous level. This indicated removal of signal and not noise.

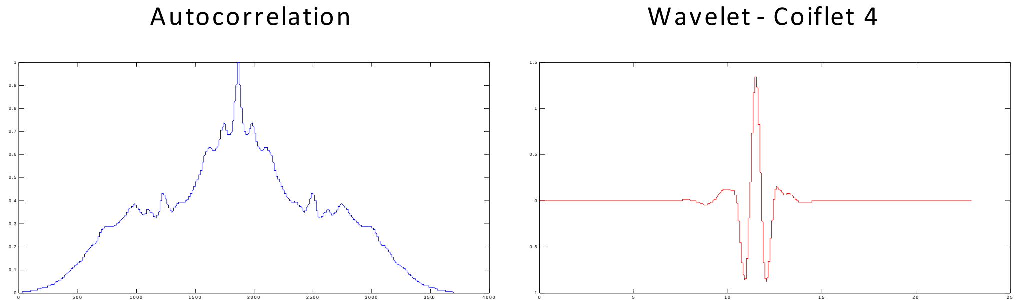


Figure 30. Autocorrelated IR spectrum of a clear coat and the Coiflet 4 mother wavelet function

Figure 31 shows a PC plot of the two largest principal components of the 464 IR spectra and the 1014 wavelet coefficients comprising the training set. Each paint sample is represented as a point in the PC plot (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5). The overlap of the clear coats by plant group is evident in the PC plot of the data.

The pattern recognition GA identified wavelet coefficients characteristic of plant group by sampling key feature subsets, scoring their PC plots, and tracking those plant groups and/or autocorrelated IR 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 28 wavelet coefficients whose PC plot (see Figure 32) showed clustering of the autocorrelated spectra on the basis of plant group.

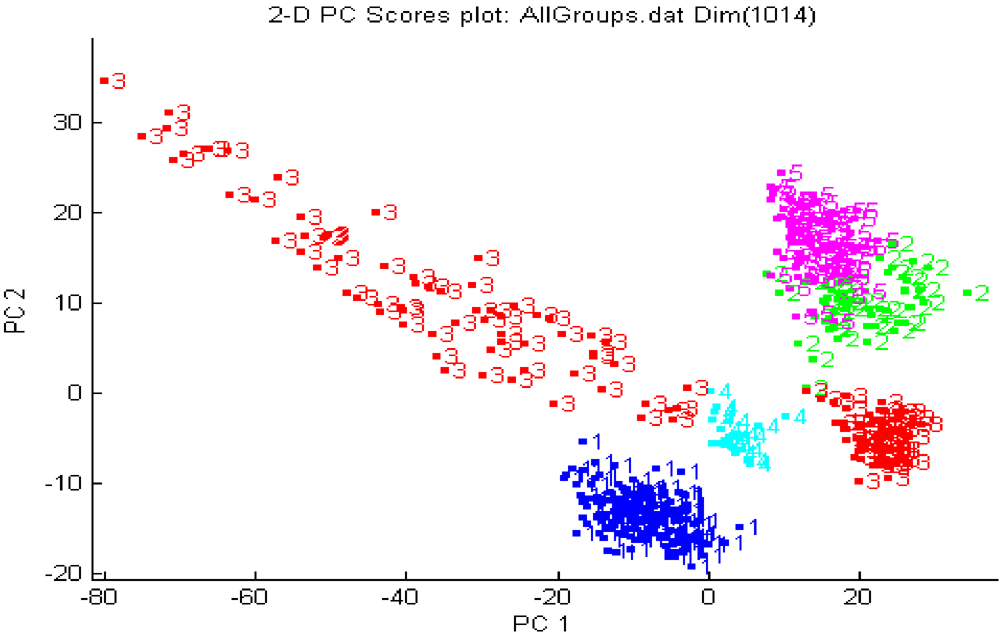


Figure 31. PC plot of the two largest principal components of the 464 IR spectra and 1014 wavelet coefficients comprising the training set. Each paint sample is represented as a point in the PC plot. 1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5.

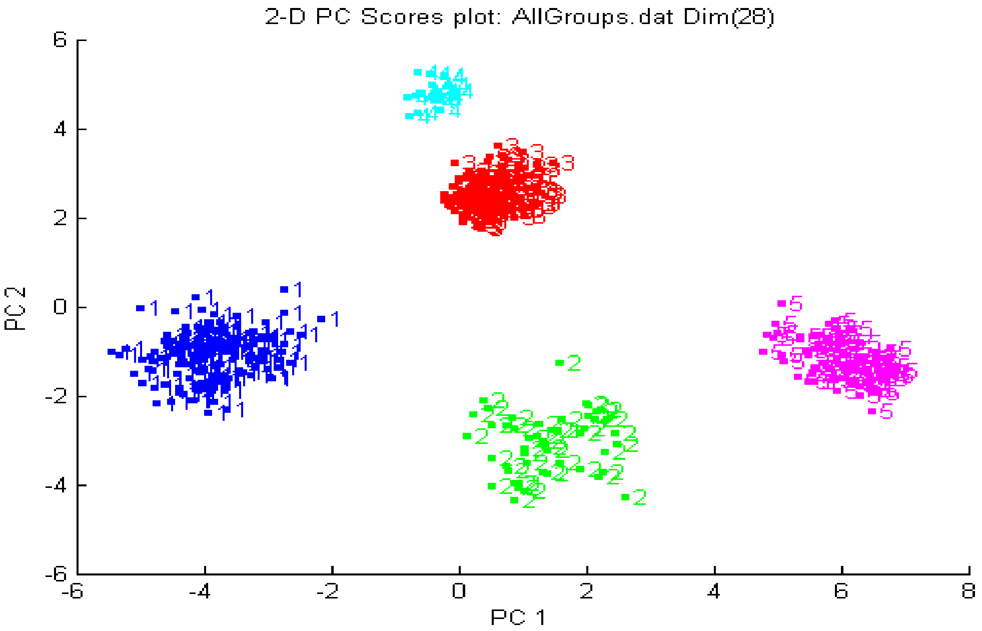


Figure 32. PC plot of the two largest principal components of the 464 IR spectra and 28 wavelet coefficients comprising the training set. Each paint sample is represented as a point in the PC plot. 1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5.

To assess the predictive ability of the 464 IR spectra and the 28 wavelet coefficients identified by the pattern recognition GA, a set of 10 blind (validation set) samples was used. Spectra from the blind samples were projected directly onto the PC map developed from the 463 autocorrelated spectra and 40 wavelet coefficients identified by the pattern recognition GA. Figure 33 shows the projection of the blind samples onto the PC map of the training set data. 9 of the 10 projected blind samples are located in a region of the map occupied by spectra possessing the same class label. The misclassified blind sample was a clear coat obtained from a bumper made of plastic. Often, the clear coat is applied to plastic automotive components in the plant that manufactures the component, not in the plant that assembles the vehicle. Furthermore, bumpers and other plastic automotive components are replaceable and may not contain the original automotive paint system used by the vehicle.

The next step in this study was to develop search prefilters to identify the blind samples by manufacturing plant. For the first three plant groups, we developed a search prefilter to discriminate the autocorrelated spectra of the clear coats by manufacturing plant within each plant group. Figure 34 shows a plot of the two largest principal components of the 28 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group (see Table 14). Each autocorrelated spectrum is represented as a point in the PC plot of the data. Plant 18 (Moraine OH) is well separated from the other 6 manufacturing plants in the PC plot. As for these other six manufacturing plants, the spectra (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI) are very similar, which prevented further discrimination by manufacturing plant for these clear coats

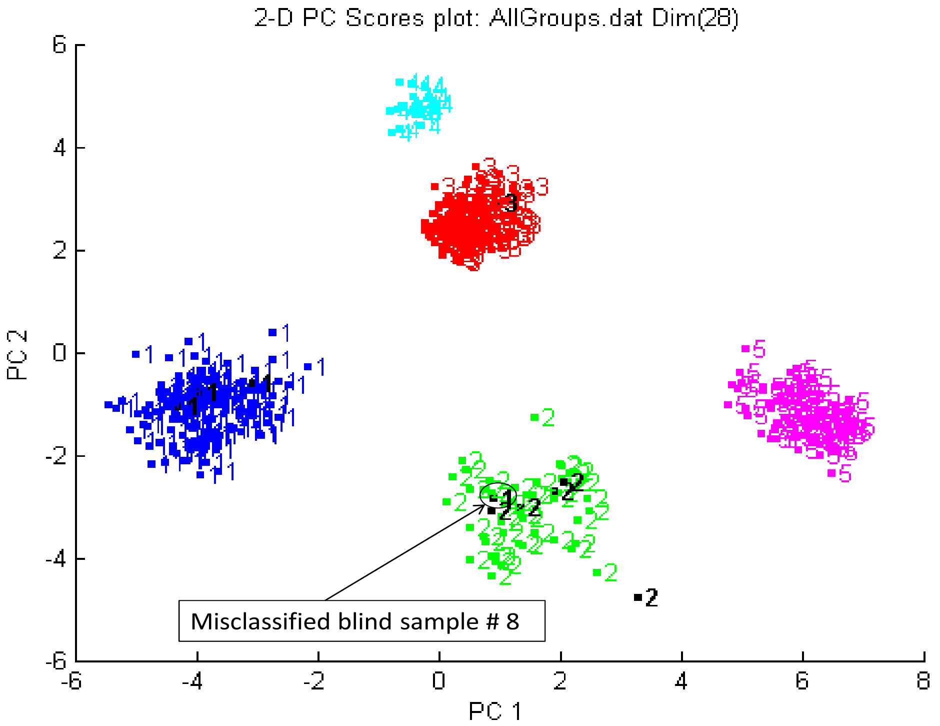


Figure 33. Projection of the blind samples (which are in black) onto the PC map of the training set data. Each paint sample is represented as a point in the PC plot. 1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5.

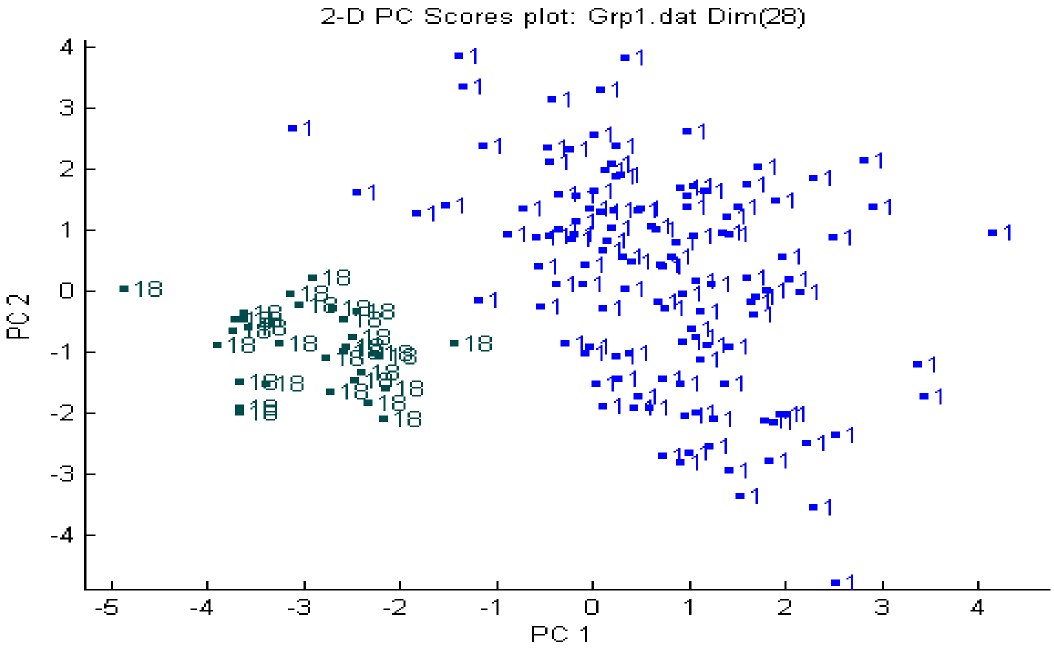


Figure 34. Plot of the two largest principal components of the 28 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group (see Table 14). Each autocorrelated spectrum is represented as a point in the PC plot of the data. 1 = Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI and 18 = Moraine OH

Projection of the blind samples assigned to Plant Group 1 onto the PC plot of the data defined by the 28 wavelet coefficients identified by the pattern recognition GA show that each blind sample is located in a region of the map with paint samples that have the class label: either Plant 18 or Plants 1, 4, 5, 8, 14, and 23 (see Table 13). This result indicates that search prefilters developed from autocorrelated spectra of clear coats can be used to characterize paint smears by manufacturing plant or can specify a limited number of manufacturing plants that would be potential manufacturing plants for assembly of the automobile.

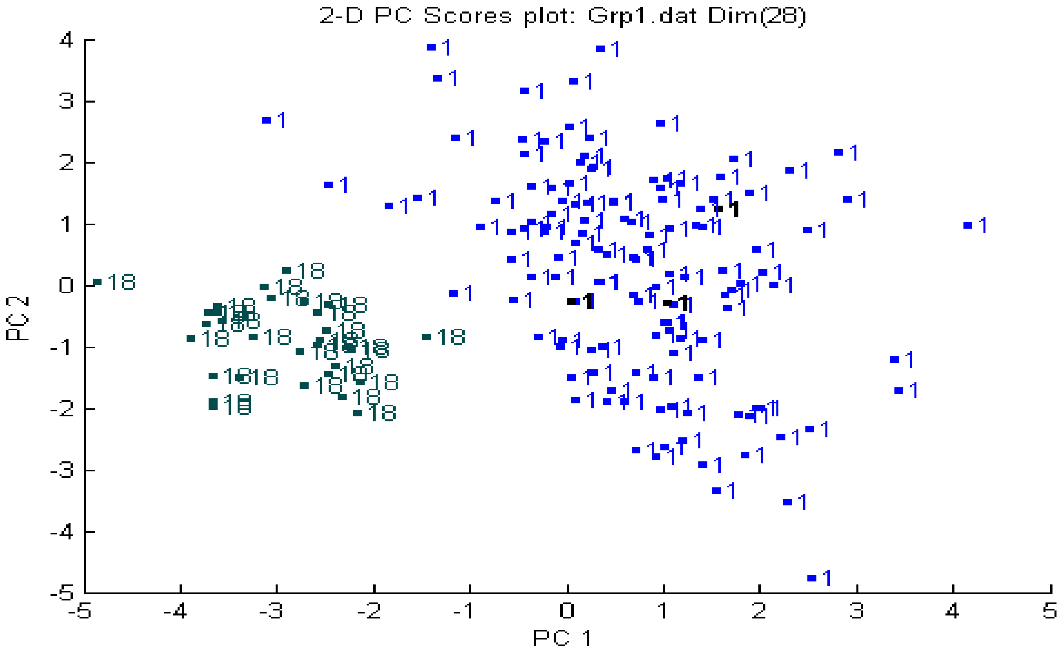


Figure 35. Projection of the blind samples (which are in black) onto the PC map of the data show that each projected sample is located in a region of the map with paint samples that have the class label. 1 = Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI and 18 = Moraine OH

Figure 36 shows a plot of the two largest principal components of the 54 autocorrelated spectra and the 18 wavelet coefficients identified by the pattern recognition GA for clear coats from Plant Group 2 (see Table 14). Again, each autocorrelated spectrum is represented as a point in the PC plot. The Bowling Green KY Plant is well separated from the Hamtramck MI and Orion MI plants in the PC plot of the data. However, the Hamtramck MI plant overlaps with the Orion plant in the plot due to the similarity of the IR spectra from these two plants.

The 5 blind samples assigned to Plant Group 2 were used to assess the predictive ability of the 28 wavelet coefficients identified by the pattern recognition GA. We chose to map the 5 IR spectra directly onto the PC map defined by the 54 IR spectra and 28 wavelet coefficients. Figure 37 shows the 5 blind (validations set) samples assigned to Plant Group 2 projected onto the PC plot of the second plant group. All 5 blind samples lie in a region of the map, which contain clear coat paint samples from the same manufacturing plant.

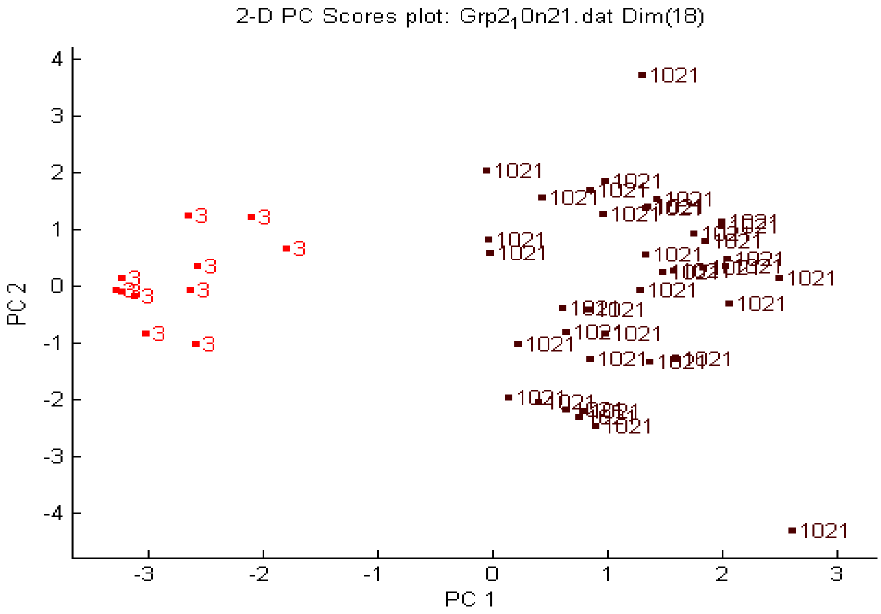


Figure 36. Plot of the two largest principal components of the 54 autocorrelated spectra and the 18 wavelet coefficients identified by the pattern recognition GA for clear coats from Plant Group 2. Each autocorrelated spectrum is represented as a point in the PC plot of the data. 3 = Bowling Green KY, 10 = Hamtramck MI, and 21 = Orion MI.

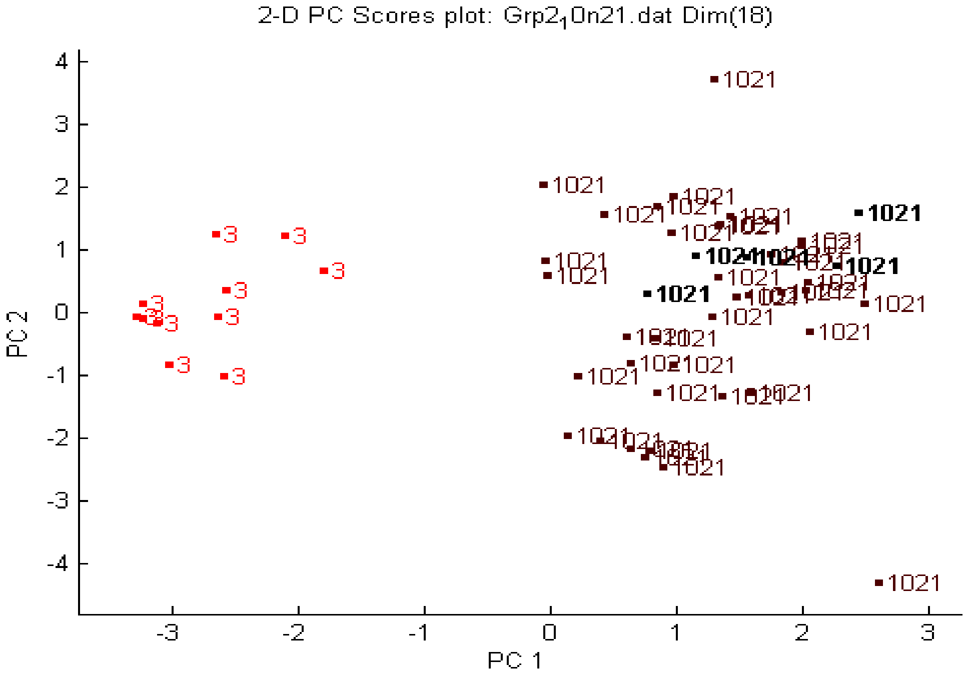


Figure 37. Projection of the blind samples (which are in black) onto the PC map of the data show that each projected sample is located in a region of the map with paint samples with the class label. 3 = Bowling Green KY, 10 = Hamtramck MI, and 21 = Orion MI.

Figure 38 shows a plot of the two largest principal components of the 141 autocorrelated spectra and 17 wavelet coefficients identified by the pattern recognition GA for clear coats from Plant Group 3. Clustering of the autocorrelated spectra by manufacturing plant, automobile model, and line is evident from an examination of the PC plot. Trucks from Oshawa (Plant 22) form a distinct cluster in the PC plot of the data as do trucks and automobiles from Fremont California (Plant 9) and Lordstown Ohio (Plant 17). SUVs from Oklahoma City (Plant 20) lie in another cluster with Buicks from Oshawa Ontario. Chevrolets from Oklahoma City (Plant 20) and from Oshawa (Plant 22), GMC trucks from Shreveport Louisiana (Plant 25) and Flint Michigan (Plant 6), and GMC and Chevrolet trucks from Linden, NJ (Plant 16) all lie in another cluster. The lone blind (validation set) sample assigned to Plant Group 3 is projected onto the PC plot in a region containing clear coat paint samples from the same manufacturing plant (Lordstown).

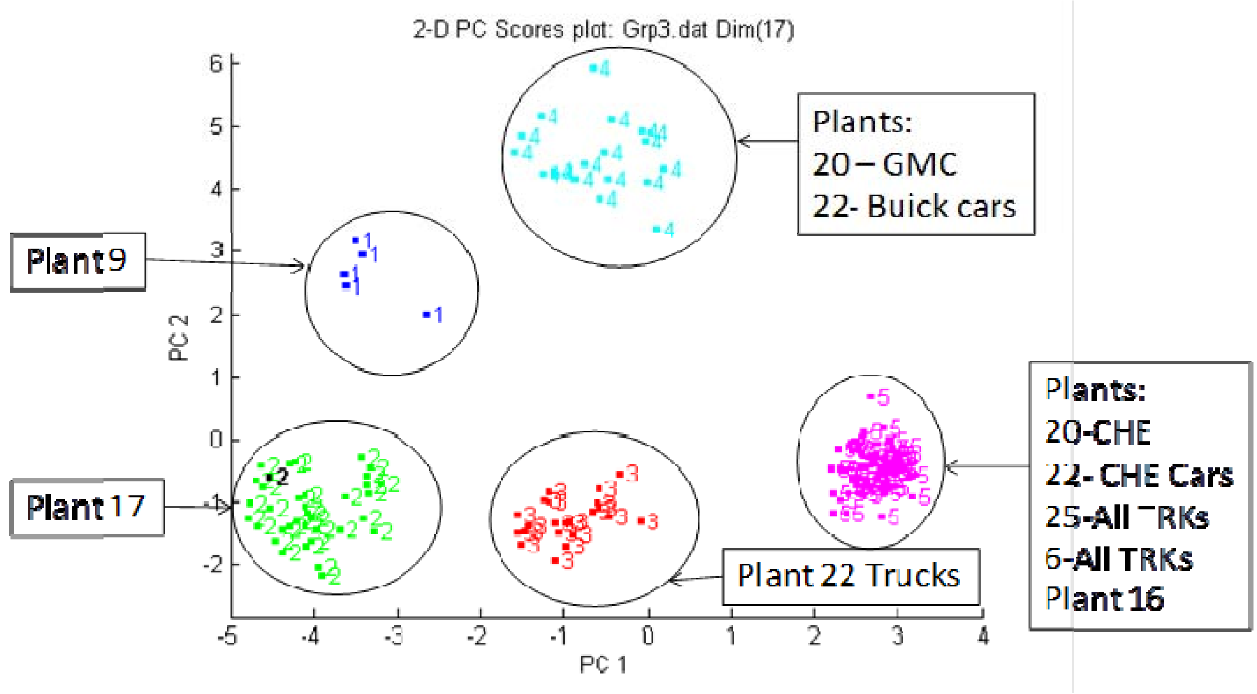


Figure 38. Plot of the two largest principal components of the 141 autocorrelated spectra and 17 wavelet coefficients identified by the pattern recognition GA for clear coats from Plant Group 3. Each autocorrelated spectrum is represented as a point in the PC plot of the data. The lone blind sample (in black) is projected onto the PC map of the data. 1 = FRE, 2 = LRD, 3 = OSH, 4 = OKL and OSH, and 5 = OKL, OSH, SHR, FLI, and LIN.

A summary of the results obtained for the 10 blind samples is shown in Table 16. 9 of the 10 blind samples were correctly classified. The lone misclassified blind sample is a clear coat paint smear from a bumper. The search prefilters developed from Bio-Rad and Thermo Nicolet IR spectra that were autocorrelated to address alignment issues along both the x and y-axes allowed for all spectra to be compared even if they were collected on different spectrometers.

Table 16. Summary of the Results Obtained for the 10 Blind Samples

| Blind | Assigned Plant Goup | As Identified by Auto Corr | Actual Identity |
|---|---|---|---|
| 1 | 2 | 10, 21 | HAM(Plant 10) |
| 2 | 1 | 1,4,5,8,14 or 23 | LAN (Plant 14) |
| 3 | 2 | 10,21 | ORI (Plant 21) |
| 4 | 2 | 10,21 | HAM(Plant 10) |
| 5 | 2 | 10,21 | ORI (Plant 21) |
| 6 | 3 | 17 | LRD (Plant 17) |
| 7 | 1 | 1,4,5,8,14 or 23 | FAI (Plant 5) |
| 8 | 2 | NA | LAN (Plant 14)<br>Wrongly classified at Group level |
| 9 | 2 | 10,21 | HAM(Plant 10) |
| 10 | 1 | 1,4,5,8,14 or 23 | LAN(Plant 14) |

**Search Prefilters for IR Spectra of Clear Coats using Stacked and Linear Classifiers**

The approach taken in the previous studies outlined and discussed in this section of the report utilizes a genetic algorithm for feature selection and pattern recognition to select individual wavelengths or wavelet coefficients for the development of search prefilters. An alternate approach to variable selection first proposed by Kalivas [43] and investigated in this study attempts to improve upon classification results by selecting spectral regions rather than searching for distinct wavelengths in the IR spectrum. The informative spectral regions in the clear coat paint spectra were identified using a recently developed classification technique [44] called stacked PLS discriminant analysis (SPLSDA), to create classification models for search prefilters. Stacking, a concept first proposed by Brieman [45], is similar to ensemble modeling [46], a concept that has appeared in many fields.

In the closely related technique, stacked partial least squares regression (SPLSR), small intervals of the data matrix comprising the X-block are each regressed on the Y block values separately [47]. The simple regression models are then combined, giving a simpler and often better regression model than a global model that utilizes all regions of the spectrum. For classification, a discriminant analysis-based classifier is used on each of the small intervals to classify the samples. One reason for using SPLSDA over other classification techniques is the inherent dimension reduction obtained for each PLS model - most are simple, with few latent variables needed to describe the class-related information in the data.

In a stacked PLS model, the calibration spectra comprising the training set are first partitioned into a set of n disjoint wavelength regions of equal width. All spectra (training set and validation set) are partitioned in the same way. For the calibration set, n interval PLS models are developed between the target property vector (manufacturing plant) and each of the n intervals, and a set of PLS interval regression vectors are obtained. These interval PLS models are then combined using a set of weighting values determined by a cross validation procedure to form a stacked model where each model has a specific weight defined by the reciprocal of the cross validated error rate of the PLS model developed on the $k^{th}$ interval normalized to the sum of the reciprocal of the cross validated error for all of the models. Direct application of the individual PLS regression models to a validation set partitioned as above gives the value for the class membership of the validation set samples using the previously established weights.

There is a resampling step iterated about this entire process to minimize the chance of overly optimistic results which is crucial for the success of this method. This step involves separating the data into two parts, one for the cross-validation step and one for the prediction step. Each of the intervals in the first half of the data must be cross-validated to determine the root-mean-square-error of cross-validation (RMSECV). The RMSECV (see Equation 10 where $y_i$ is the sample class membership value, $x_k$ is the $k^{th}$ spectrum interval and $b_k$ is the regression vector for the kth spectrum interval calculated for the PLS model) is used in the formulation of weights for each interval in the final, stacked, model. For the $k^{th}$ interval with $s_k$ as the reciprocal of the RMSECV, the weight $w_k$ is calculated as shown below (see Equation 11 where $s_k^2$ is the reciprocal of the cross validated error rate for the $k^{th}$ PLSDA model). The summation normalizes all of the weights to a unit sum. If the RMSECV is zero for an individual interval, an appropriate weight is used instead ($s_k^2$ = 10). The purpose of calculating the weights in this way is to ensure that a high weight is assigned to any interval where samples are assigned values close to their class membership (coded Y) values.

$$RMSECV_k = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - x_k b_k\right)^2} \tag{10}$$

$$w_k = \frac{s_k^2}{\sum_{k=1}^{n} s_k^2} \tag{11}$$

The calculated weight matrix is then used to effectively scale each interval's regression coefficients, which are then all summed together to obtain a single regression coefficient matrix. Ideally, the regression coefficient matrix creates predicted Y (class membership) values such that the target class' Y values are well separated from all others. The threshold Y value to use is one that leads to the minimum overlap of the target class from all others. The predicted y (class membership) values are calculated as seen in the equation below (Equation 12), where X is the second half of the data and β are the regression coefficients for the k<sup>th</sup> PLS model.

$$\hat{y}_u = \sum_{k=1}^{n} w_k X_{k,u} \beta_{k,SPLS} \tag{12}$$

Once the establishment of a classification model using SPLSDA has been completed, the classification of clear coat paint samples can begin. Another set of paint samples, collected using a FTIR spectrometer from a different manufacturer will use the previously calculated regression coefficient matrix to find the predicted class membership values. The same set of discriminants used in the final classification in SPLSDA is also used to classify the paint samples in the validation set.

The discriminants in SPLSDA were optimized using the technique of cross-validation. A subset of the training (calibration) set data that is withheld from the construction of the discriminant is used for testing. The p samples for testing are removed randomly with the calibration done on the remaining (m – p) samples. Repeated discriminant development and test cycles are averaged over all samples and evaluated as a function of the number of latent variables. Cross validation may involve leaving out one sample per test cycle (full cross validation) or leaving out every third or fourth sample in each test cycle (segmented cross validation).

In this study, Venetian blind cross-validation [48] was employed with a 0.5 hold-out fraction and 10 repeats (for most spectra) to optimize the number of latent variables in each (calibration set) spectrum for each PLS model developed. However, some classes had fewer samples and smaller hold out fractions. Venetian blind cross validation was used as it is computationally inexpensive and is reliable when there are many samples. Venetian blind cross validation was applied to interval sizes ranging from 2 to 25 sub-regions of the spectral response to optimize both the number of latent variables for each interval and the interval size at once. (In this study, the number of latent variables was limited to 20 for each interval. However, some classes had fewer samples and required fewer latent variables. The stacking process guards against any overfit or underfit of the intervals by PLS as these would predict poorly in the cross validation and receive low weights in the stacked model.

A 2-way cross-validation was performed to identify the spectral windows and the number of latent variables in each spectral window that yielded the minimum prediction error. The best spectral windows were selected to give the minimal cross-validation error in stacking. For stacked classification, all spectra were preprocessed using Savitzky-Golay first derivative smoothing with a default window size of 15 wavelengths followed by mean-centering inside the cross-validation. The spectra used in this study were obtained from 18 General Motors (GM) assembly plants (2000-2006). Only the clear coat paint layer from metallic parts was used to develop the search prefilters. Clear coats from plastic substrates (e.g., bumpers) were excluded as these substrates are often not painted in the same plant where the vehicle is assembled. Table 17 designates the 18 GM manufacturing plants investigated in this study.

**Table 17. GM Assembly Plants used to Develop Stacked Classifiers as Search Prefilters**

| Plant ID | Plant | Make | Line |
|---|---|---|---|
| 1 | ARL | CAD, CHE, GMC | SUB,YUK,ESD,CTA |
| 3 | BOW | CAD,CHE | CVT,XLR |
| 4 | DOR | PON | VTR,SIL,MTA,UPL,TAR |
| 5 | FAI | CHE,OLD,PON | GRA,MAL,ITR |
| 8 | FOR | CHE,GMC | SLV,SIE |
| 10 | HAM | BUI,CAD,PON | BON,DEV,LUC,LES,SEV,ELD |
| 12 | JAN | GMC | CTA,SUB,YUK |
| 14 | LAN | PON | STS |
| 16 | LIN | CHE, GMC | BZR,JMY,S10 |
| 17 | LRD | PON | SFR,CAV,COB,PST |
| 18 | MOR | CHE,GMC,SAA | JMY,ENV,9S7,BZR,TBZ,SON |
| 21 | ORI | PON,BUI | BON,PG6,LES,AUR, PKA |
| 22 | OSH | GMC,PON | ALL,REG |
| 23 | PON | CHE,GMC | SLV,SIE,SIL |
| 24 | RAM | BUI,CHE,PON | CAV,SFR,RZV,AZT,HHR |
| 25 | SHR | CHE,GMC | S10,COL,SON |
| 26 | SIL | CHE,GMC,SAA | AVL,SUB,YXL |
| 27 | SPH | STR | SSL,ION,SC1,SC2,SL1,VUE |

A hierarchical classification scheme was used to develop the search prefilters. The 18 GM assembly plants were divided into five major plant groups. Assembly plants comprising each plant group are listed in Table 18. An unknown is classified as to its plant group using a search prefilter and then a second search prefilter is used to identify the specific plant or plants within the plant group to which membership of the unknown is assigned. The clear coat paint spectra that served as the training set used in the development of both the Plant Group and Plant search prefilters for the PDQ database are summarized in Table 19.

**Table 18. Manufacturing Plants Comprising Each Plant Group**

| Plant Group | Plant ID Number | Manufacturing Plant |
|---|---|---|
| 1 | 1, 4, 5, 8, 14, 18, 23 | ARL, DOR, FAI, FOR, LAN, MOR, PON |
| 2 | 3, 10, 21 | BOW, HAM, IRI |
| 3 | 16, 17, 22, 25 | LIN, LRD, OSH, SHR |
| 4 | 12 | JAN |
| 5 | 24, 26, 27 | RAM, SIL, SPH |

**Table 19. Training Set and Validation Set for Plant Group Search Prefilter**

| Plant Group | Training Set Samples (Thermo-Nicolet) | Validation Set Samples (Bio-Rad) |
|---|---|---|
| 1 | 78 | 80 |
| 2 | 20 | 31 |
| 3 | 69 | 51 |
| 4 | 6 | 13 |
| 5 | 21 | 43 |
| Total | 194 | 221 |

Spectra from both the BioRad and Thermo-Nicolet instruments were aligned using OMNIC. The number of points collected in the wavelength range by the Thermo-Nicolet instrument varied from 1878 points to 1958 points whereas the spectra collected on the two BioRad instruments for the same wavelength range and resolution were represented by 1944 points. Peak shifting was also observed in spectra collected on the Thermo Nicolet instrument. Both problems were resolved using Nicolet's OMNIC software as an editor to process the BioRad spectra and the spectra from the Thermo Nicolet instrument using an appropriate estimate of the spectral line function of the two Thermo-Nicolet instruments. For alignment along the y-axis (transmittance) of the spectra, we ensured that all spectra started from the same transmittance value. The quality of the diamond cell transmission spectra in the PDQ library (e.g., no sloping baseline or baseline offsets, and the absorbance of the carbonyl absorbance peak in all library spectra being unity) proved pivotal for successful alignment of these spectra along the y-axis.

The first step in this study was to develop a search prefilter to classify the clear coats by plant group. A five-way classification study was undertaken for the Thermo Nicolet IR spectra. Each IR spectrum in the training set was partitioned into n adjacent spectral intervals of equal length where the number of intervals was varied from 2 to 25. For the training set, n PLS discriminant models were developed between the class membership of the samples and each of the n spectral intervals. The performance of each PLS model was evaluated with the contribution of each PLS model to the overall discriminant weighted according to the cross validated error rate of the model. Figure 39 summarizes the cross validated error rates for the 194 Thermo Nicolet spectra as a function of the number of spectral intervals and the number of latent variables for each PLS model. The cross validated error rate for each plant group was 0% when the number of latent variables for each interval was greater than 3. Double cross validation identified the specific wavelength windows in each spectrum used for stacking. The number of spectral intervals used for stacking varied from 2 to 4. By comparison, the pattern recognition GA was also able to correctly classify every Thermo Nicolet IR spectrum in the training set (see Figure 40).

Figure 39 also summarizes the error rate for the 221 (BioRad) IR spectra in the validation set. An error rate of 0% for each plant group can be achieved when the number of latent variables used to model the wavelength windows comprising each stacked model is 5. The principal component plot developed from the 11 wavelengths identified by the pattern recognition GA was also able to correctly classify every sample in the validation set (see Figure 41). When linear discriminant analysis was used to develop a classifier for these same 11 wavelengths, 100% correct classification was again achieved for both the training set and validation set (see Table 20).

The next step in this study was to develop search prefilters to classify the IR spectra by assembly plant for Plant Groups 1, 2, and 3. Plant Group 4 contains only a single assembly plant and the IR spectra from the three assembly plants comprising Plant Group 5 cannot be differentiated as their spectra are superimposable. In this phase of the study, Plant Groups 1, 2, and 3 were first investigated using the pattern recognition GA to search for significant structure in the data using PCKaNN to score the features. We selected the pattern recognition GA as our benchmark because the pattern recognition GA is not limited to using mathematics for modeling per se but is also discovery as the GA can serve as a data microscope to sort, probe and to look for hidden relationships in the data and to uncover the class structure of the data.

Because IR spectra within a plant group are more similar than IR spectra from different plant groups, more powerful preprocessing methods were judged to be necessary to extract information about assembly plant from the IR spectra of the clear coats. The Symlet6 mother wavelet at the 8th level of decomposition, i.e., 8Sym6, was implemented using the discrete wavelet transform and applied to each vector normalized IR spectrum to denoise and deconvolute the IR data into wavelet coefficients. The criterion used to select this mother wavelet is based solely on the ability of 8Sym6 to extract information about assembly plant from the spectra. A decrease in the ability of the GA to correctly classify the IR spectra was observed when other mother wavelets were used to denoise and deconvolute the spectra.

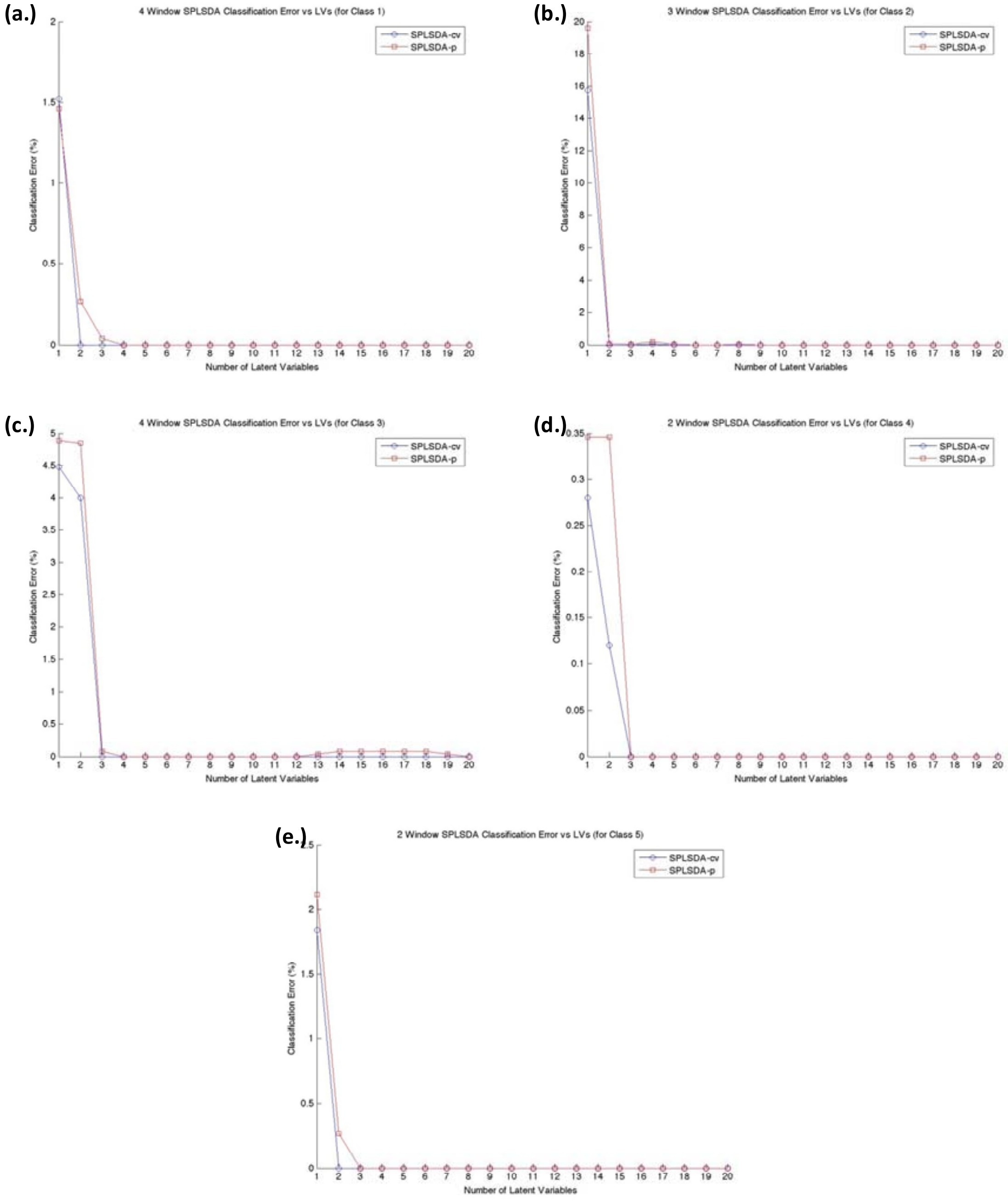


Figure 39. Cross validated error rates for the 194 Thermo Nicolet spectra (black) and 221 BioRad spectra (red) as a function of the number of spectral intervals and the number of latent variables for each PLS model: a) Plant Group 1, b) Plant Group 2, c) Plant Group 3, d) Plant Group 4, and e) Plant Group 5.

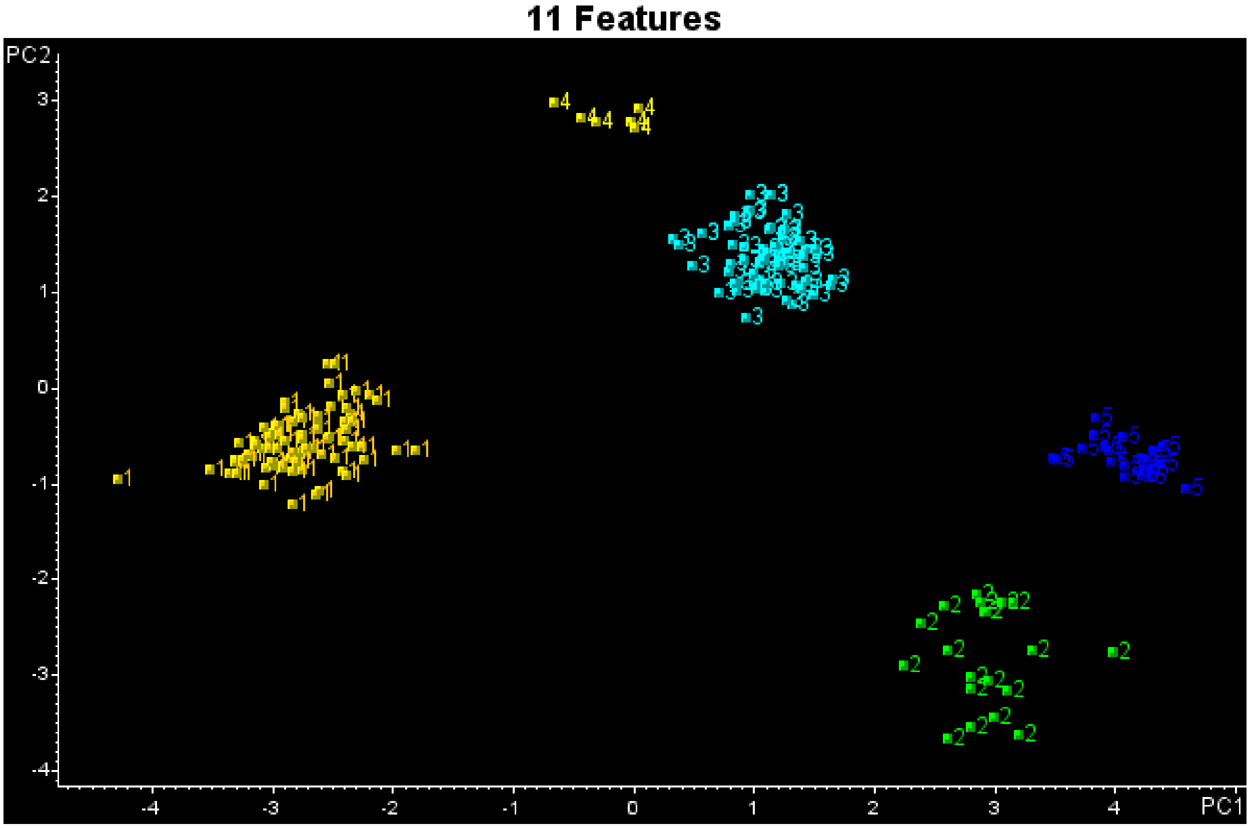


Figure 40. Plot of the two largest principal components of the 11 wavelengths identified by the pattern recognition GA. Each paint sample is represented as a point in the PC plot (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

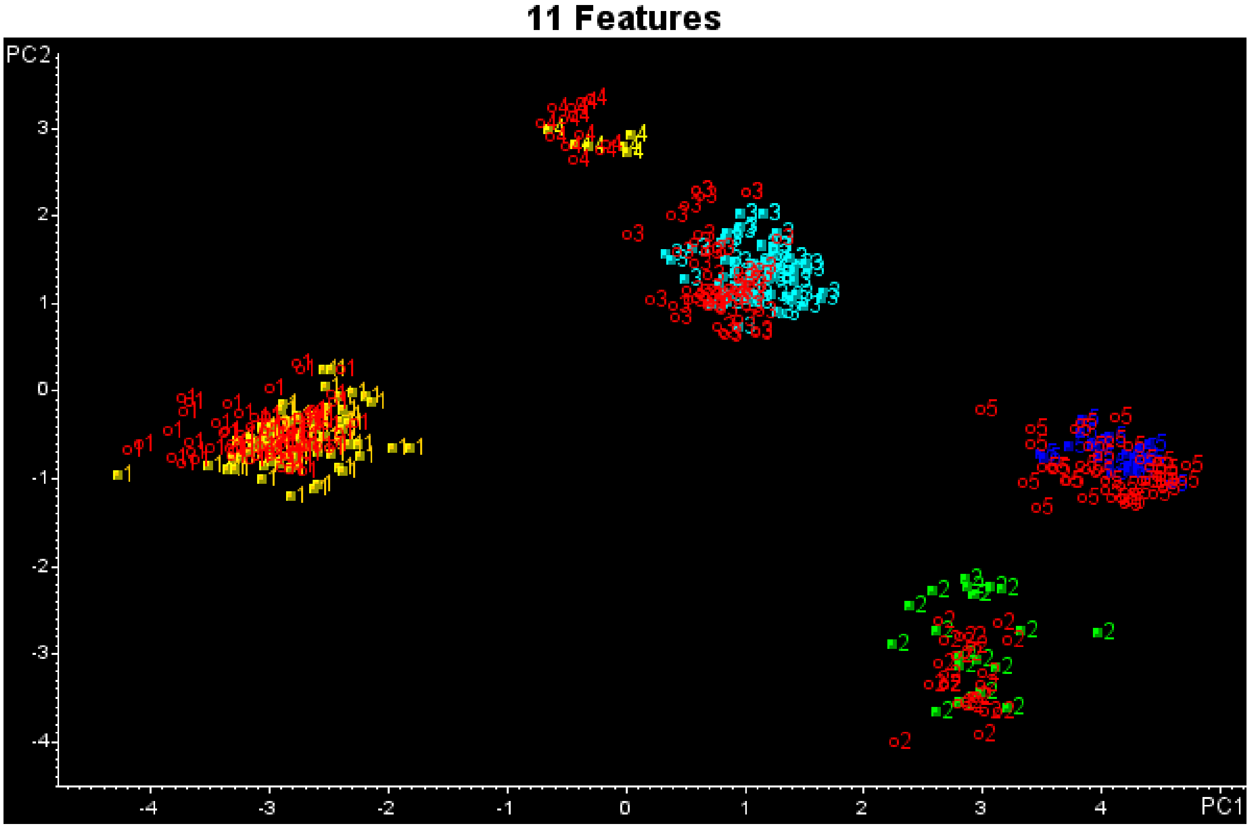


Figure 41. Validation set samples (red) projected onto the PC plot of the Thermo-Nicolet training set samples characterized by 11 wavelengths identified by the pattern recognition GA. (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

**Table 20. LDA Analysis for Plant Group Using 11 Spectral Features**

| Group | Samples | Misses | Success | | Group | Samples | Misses | Success |
|---|---|---|---|---|---|---|---|---|
| 1 | 78 | 0 | 100 | | 1 | 80 | 0 | 100 |
| 2 | 20 | 0 | 100 | | 2 | 31 | 0 | 100 |
| 3 | 69 | 0 | 100 | | 3 | 51 | 0 | 100 |
| 4 | 6 | 0 | 100 | | 4 | 13 | 0 | 100 |
| 5 | 21 | 0 | 100 | | 5 | 43 | 0 | 100 |
| Total | 194 | 0 | 100 | | Total | 221 | 0 | 100 |

Figure 42 shows a plot of the two largest principal components of the 26 wavelet coefficients identified by the pattern recognition GA for the assembly plants comprising the first plant group (see Table 21). Each IR 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. IR spectra from the other 6 manufacturing plants (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI) are superimposable, which prevented discrimination by assembly plant for these clear coats. Although the pattern recognition GA was parameterized to search for wavelet coefficients to separate all 7 assembly plants, the class structure of the data detected by the GA when performing feature selection indicated that only a single assembly plant can be identified among the 7 assembly plants that constitute this plant group. This result was verified when individual spectra from the other six plants was examined visually and overlayed for comparison.

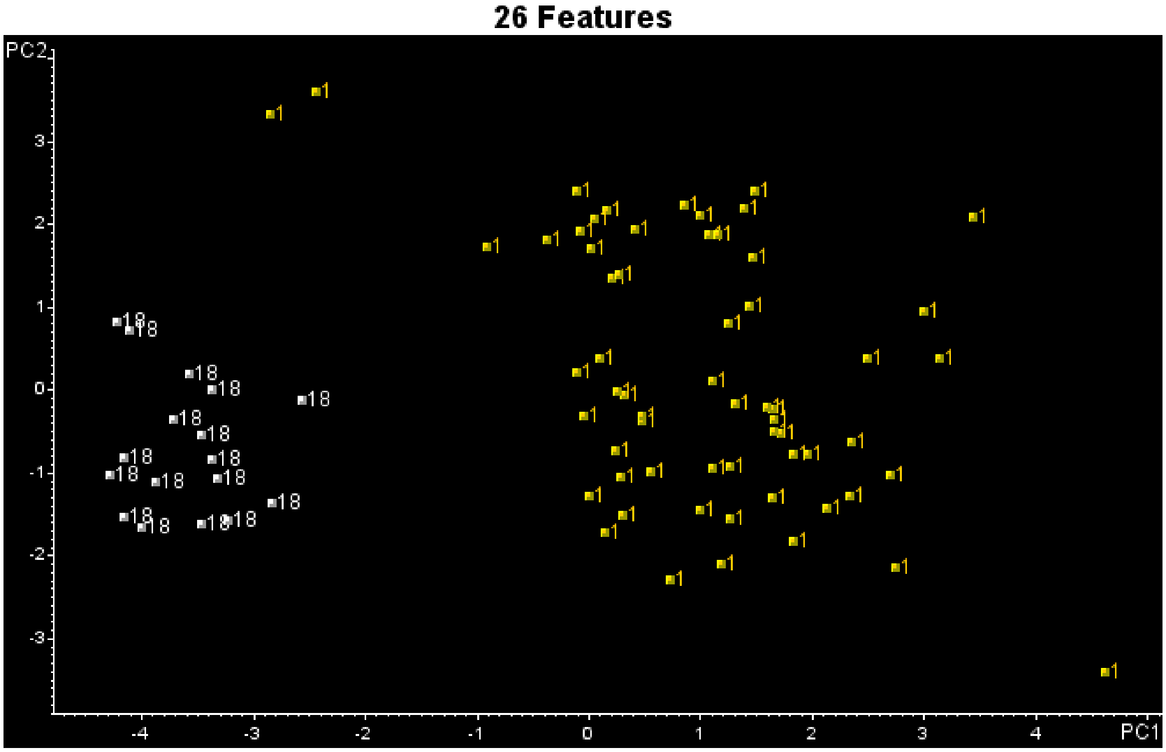


Figure 42. PC plot of the of the 26 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group is shown. Each IR spectrum is represented as a point in the PC plot. 1 = Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, Pontiac MI, and 18 = Moraine OH

A validation set of 80 IR spectra (see Table 21) was employed to assess the predictive ability of the 26 wavelet coefficients identified by the pattern recognition GA. We chose to map the 80 spectra directly onto the PC map defined by the 78 IR spectra of the training set and the 26 wavelet coefficients identified by the pattern recognition GA. Figure 43 shows the validation set samples projected onto the PC map developed from the training set data. Each projected sample is in a region of the map with paint samples that have the same class label: either plant 18 or plants 1, 4, 5, 8, 14, and 23.

**Table 21. Training Set and Validation Set for Plant Group 1**

| Plants | Training set samples (Thermo-Nicolet) | Validation set samples (Bio-Rad) |
|---|---|---|
| 18 | 17 | 13 |
| 1,4,5,8,14,23 | 61 | 67 |
| Total | 78 | 80 |

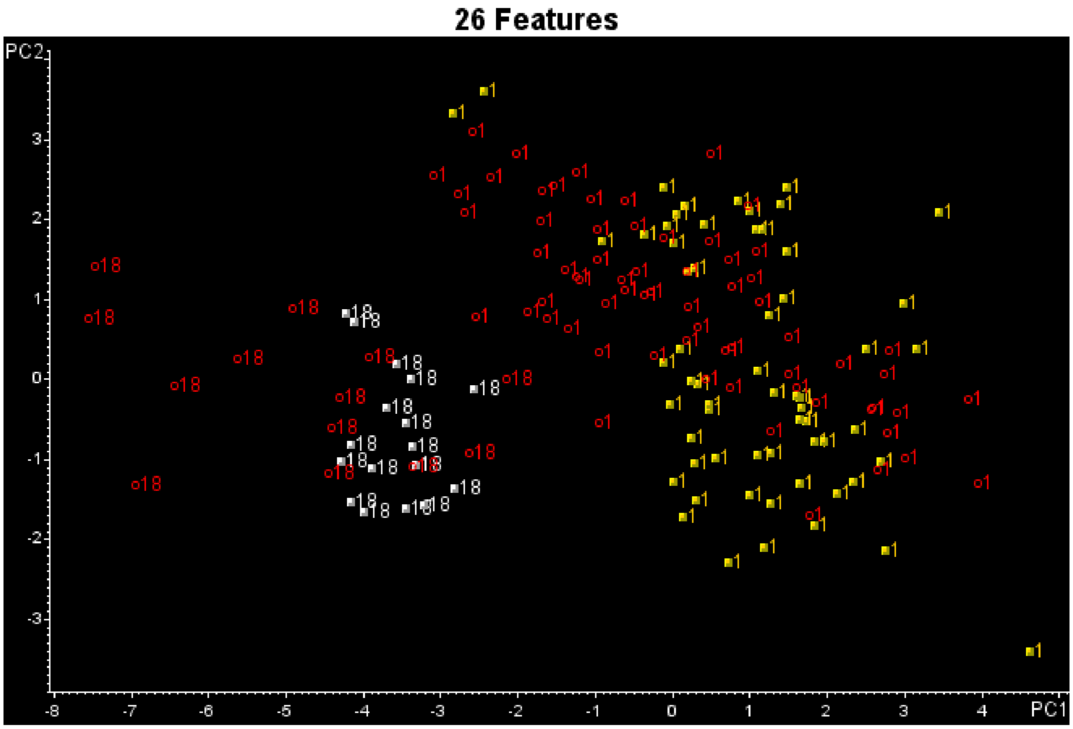


Figure 43. A plot of the two largest principal components of the 26 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group is shown. Each IR spectrum is represented as a point in the PC plot. Validations set samples are in red. 1 = Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, Pontiac MI, and 18 = Moraine OH

Figure 44 shows a plot of the two largest principal components of the 20 IR spectra of the training set and the 26 wavelet coefficients identified by the pattern recognition GA for assembly plants comprising Plant Group 2 (see Table 22). Each IR spectrum is represented as a point in the plot. All 3 manufacturing plants (Bowling Green KY, Hamtramck MI, and Orion MI) form distinct and well separated clusters in the PC plot. Only one training set sample (Hamtramck MI) is misclassified.

A validation set of 31 IR spectra (see Table 22) was employed to assess the predictive ability of the 26 wavelet coefficients identified by the pattern recognition GA. The 31 IR spectra were projected onto the PC plot defined by the 20 IR spectra of the training set and 24 wavelet coefficients identified by the pattern recognition GA (see Figure 45). All validation set samples (except for a Hamtramck MI clear coat) are in a region of the map with paint samples that have the same class label.

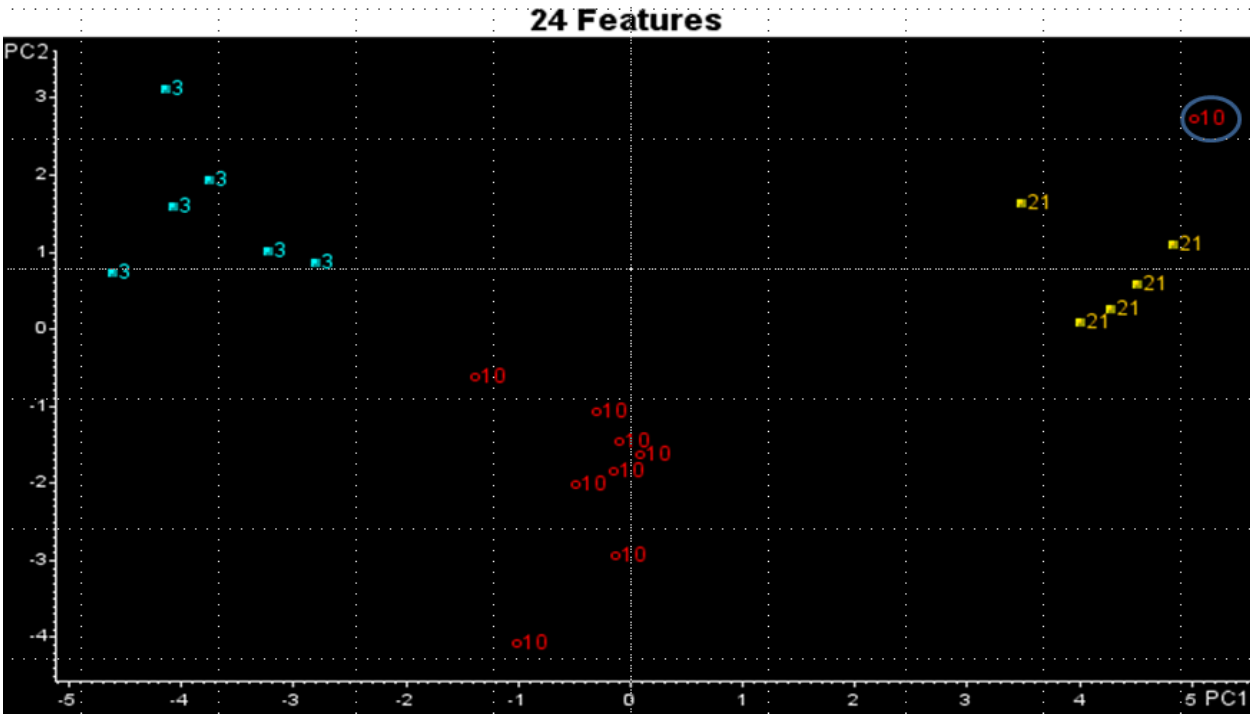


Figure 44. A plot of the two largest principal components developed from the 20 IR spectra of the training set and the 24 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the second plant group is shown. 3 = Bowling Green KY, 10 = Hamtramck MI, and 21 = Orion MI

**Table 22. Training Set and Validation Set for Plant Group 2**

| Plants | Training Set Samples (Thermo-Nicolet) | Validation Set Samples (Bio-Rad) |
|---|---|---|
| 3 | 6 | 10 |
| 10 | 9 | 13 |
| 21 | 5 | 8 |
| Total | 20 | 31 |

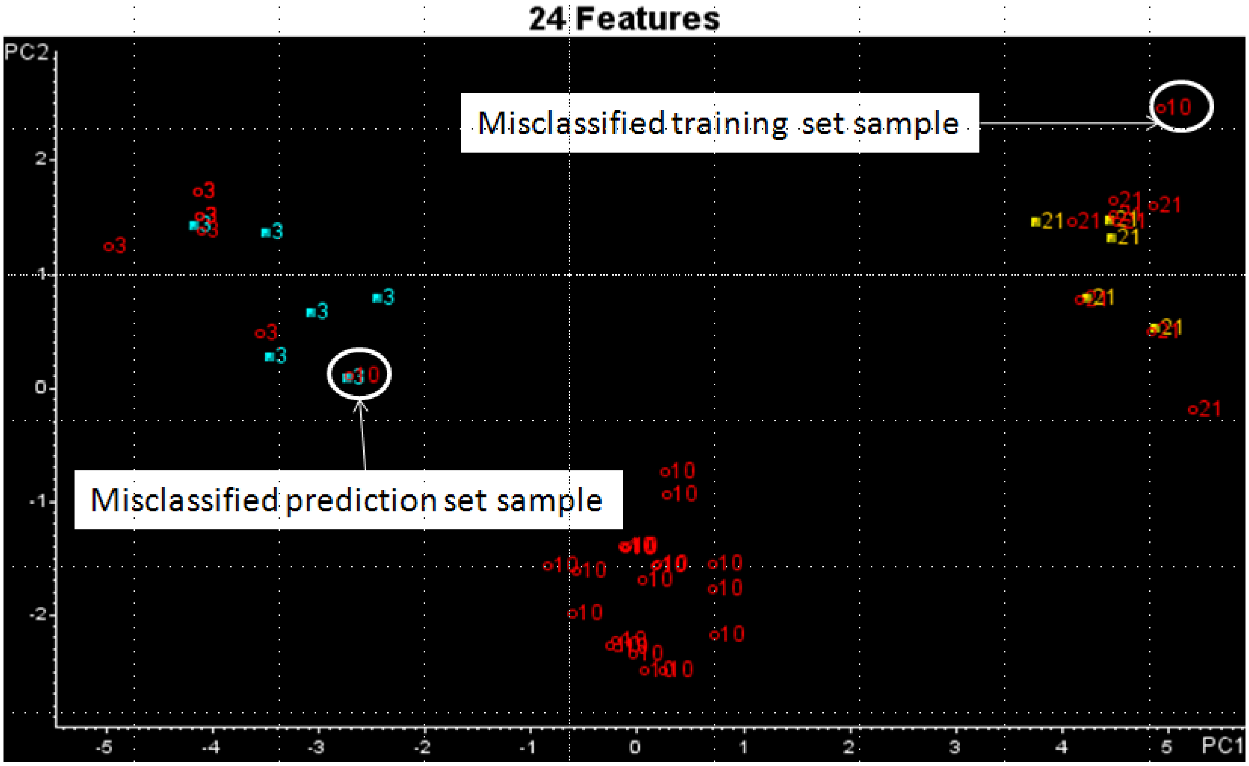


Figure 45. PC plot of the validations set spectra (see Table 21) projected onto the PC map developed from the 20 IR spectra and 24 wavelet coefficients identified by the pattern recognition GA. 3 = Bowling Green KY, 10 = Hamtramck MI, 21 = Orion MI. Validation set samples and Plant 10 samples are in red.

Figure 46 shows a plot of the two largest principal components of the 69 training set samples and the 5 wavelet coefficients identified by the pattern recognition GA for manufacturing plants comprising the third plant group (see Table 23). Clear coats from Oshawa Ontario (Plant 22) are divided into 3 distinct sample groups: Chevrolets, Buicks, and GMC trucks. The Buicks form a separate cluster in the PC plot as do the GMC trucks. Automobiles from Lordstown OH (Plant 17) also cluster in a distinct region of the PC map of the data. There is a fourth cluster consisting of Chevrolets from Oshawa Ontario (Plant 22), automobiles from Linden NJ (Plant 16), and trucks from Shreveport LA (Plant 25).

Figure 47 shows the validations set samples (see Table 23) projected onto the PC plot developed from the 54 IR spectra of the training set and the 5 wavelet coefficients identified by the pattern recognition GA. Each validation set sample lies in a region of the map with other paint samples that have the same class label. For Plant Group 3, wavelet transformed IR spectra of clear coats were differentiated by assembly plant and also by model and line for a given assembly plant.

The significance of the GA runs are two-fold: (1) search prefilters can be developed to extract information from clear coats independent of the instrument used to generate the data, and (2) the manufacturing plant responsible for the clear coat paint layer can be narrowed down to a single plant or a few plants. Wavelets played a pivotal role in uncovering information about assembly plant from the IR spectra through deconvolution of overlapping spectral responses.

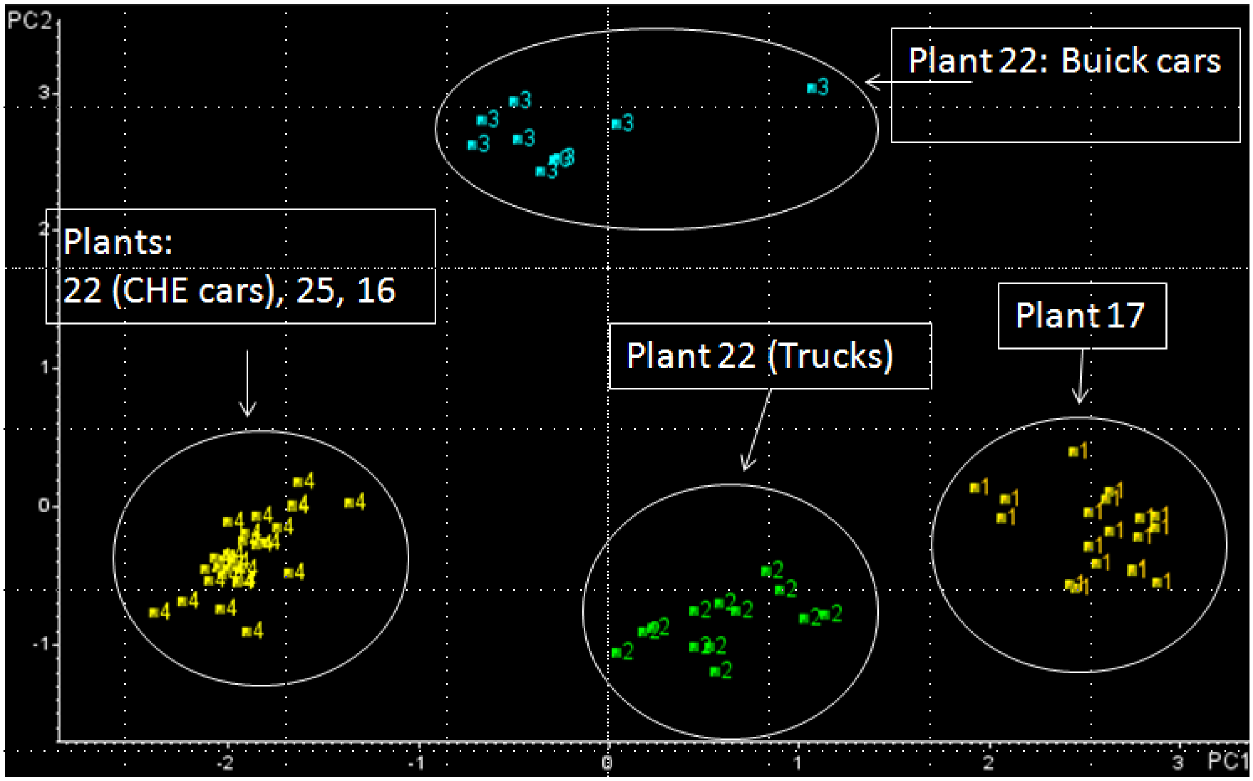


Figure 46. Plot of the two largest principal components of the 5 wavelet coefficients identified by the pattern recognition GA for manufacturing plants comprising the third plant group is shown. 1 = Plant 17, 2 = Plant 22 (GMC trucks), 3 = Plant 22 (Buick automobiles), and 4 = Plants 16, 22 (Chevrolet automobiles), and Plant 25 (GMC trucks)

**Table 23. Training Set and Validation Set for Manufacturing Plants from Plant Group 3**

| Plants | Training Set Samples (Thermo-Nicolet) | Validation Set Samples (Bio-Rad) |
|---|---|---|
| 17 | 19 | 16 |
| 22 (Trucks) | 13 | 8 |
| 22 (Buick Cars) | 9 | 5 |
| 16, 22 (CHE cars), 25 (all TRKS) | 28 | 25 |
| Total | 69 | 54 |

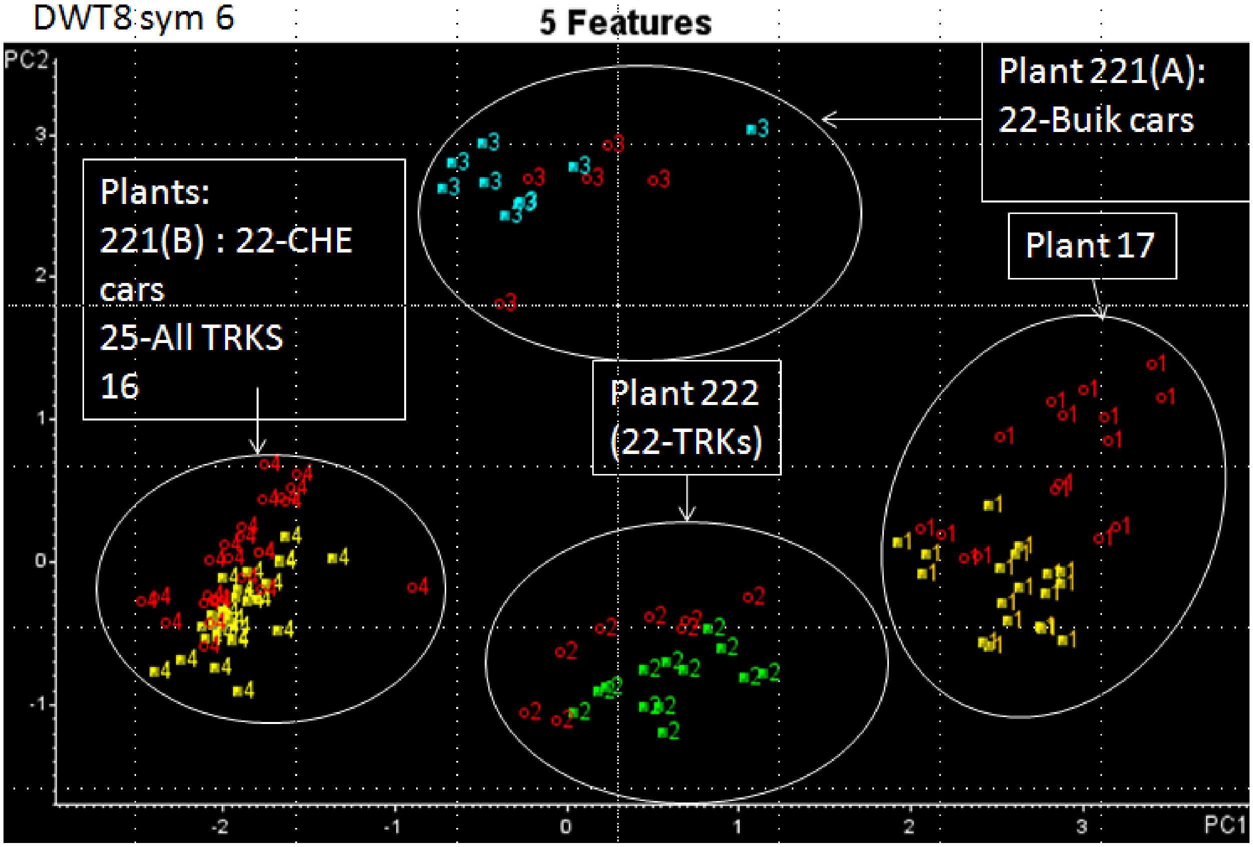


Figure 47. PC plot of the validation set samples (see Table 22) projected onto the PC map developed from the 69 IR spectra of the training set and 5 wavelet coefficients for assembly plants comprising the third plant group. 1 = Plant 17, 2 = Plant 22 (GMC trucks), 3 = Plant 22 (Buick automobiles), and 4 = Plants 16, 22 (Chevrolet automobiles), and Plant 25 (GMC trucks). Samples from the validation set are in red.

Search prefilters were also developed for the IR spectra in each of the three plant groups investigated by the pattern recognition GA using the stacked PLS discriminants. For Plant Group 1, the development of the search prefilter for assembly plant focused on the binary classification problem: Plant 18 (Moraine OH) versus the other six assembly plants (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI). Figure 48 summarizes the error rates for both the training and validation sets using the stacked PLS models. The cross validated error rate for the training set was 0% when 9 spectral intervals identified by double cross validation and 6 latent variables for each of the intervals were used for stacking. These same conditions yielded an error of 18% for the validation set. By comparison, the classification success rates obtained by the pattern recognition GA for Plant 18 are 100% for both the training set and validation set. The superior performance of the PC plot, which is a simpler classification model for the data, can be attributed to two factors: (1) the use of wavelets to preprocess the IR spectra, and (2) the approach taken by the pattern recognition GA to identify specific features in the data. The approach taken by the stacked classifiers attempts to identify informative spectra regions, which in turn limits the preprocessing of the IR spectra to the first or second derivative.

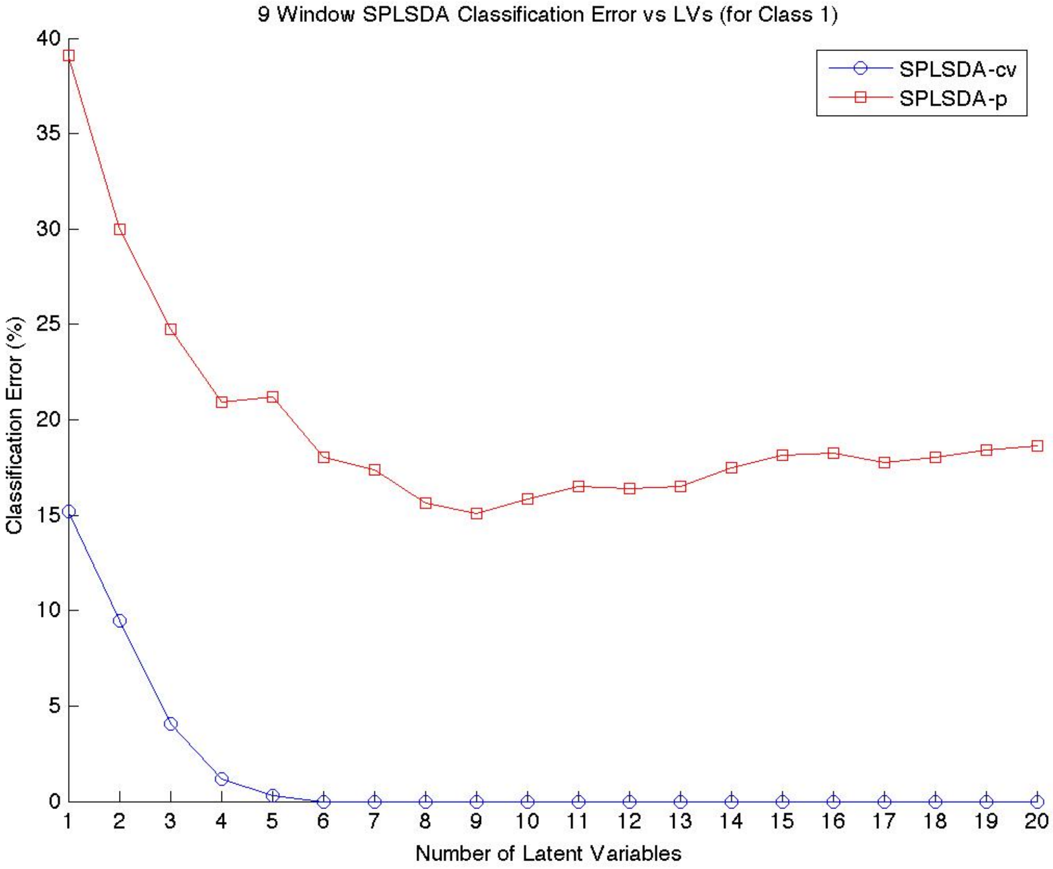


Figure 48. Cross validated error rate (blue) and prediction error (red) for the Moraine OH plant as a function of the number of latent variables for the 9 wavelength windows identified by double cross validation for the stacked PLS classifier.

For Plant Group 2, the development of a search prefilter for assembly plant required the solution of a 3-way classification problem (Bowling Green KY/3, Hamtramck MI/10, and Orion MI/21). Training set and validation set results for the stacked PLS classifier developed from the IR spectra of Plant Group 3 are summarized in Figure 49. The error rate for each assembly plant in the training set is 0% when 4 latent variables are used to model the wavelength windows used for stacking. For the validation set, the average error rate is 38% (Bowling Green/Plant 3), 2% (Hamtramck MI/Plant 10), and 16% (Orion MI/Plant 21). By comparison, the pattern recognition GA identified 24 features that correctly classified all but one sample in both the training set and validation set. The lone training set sample and lone validation set sample incorrectly classified by the GA were from the Hamtramck MI assembly plant.

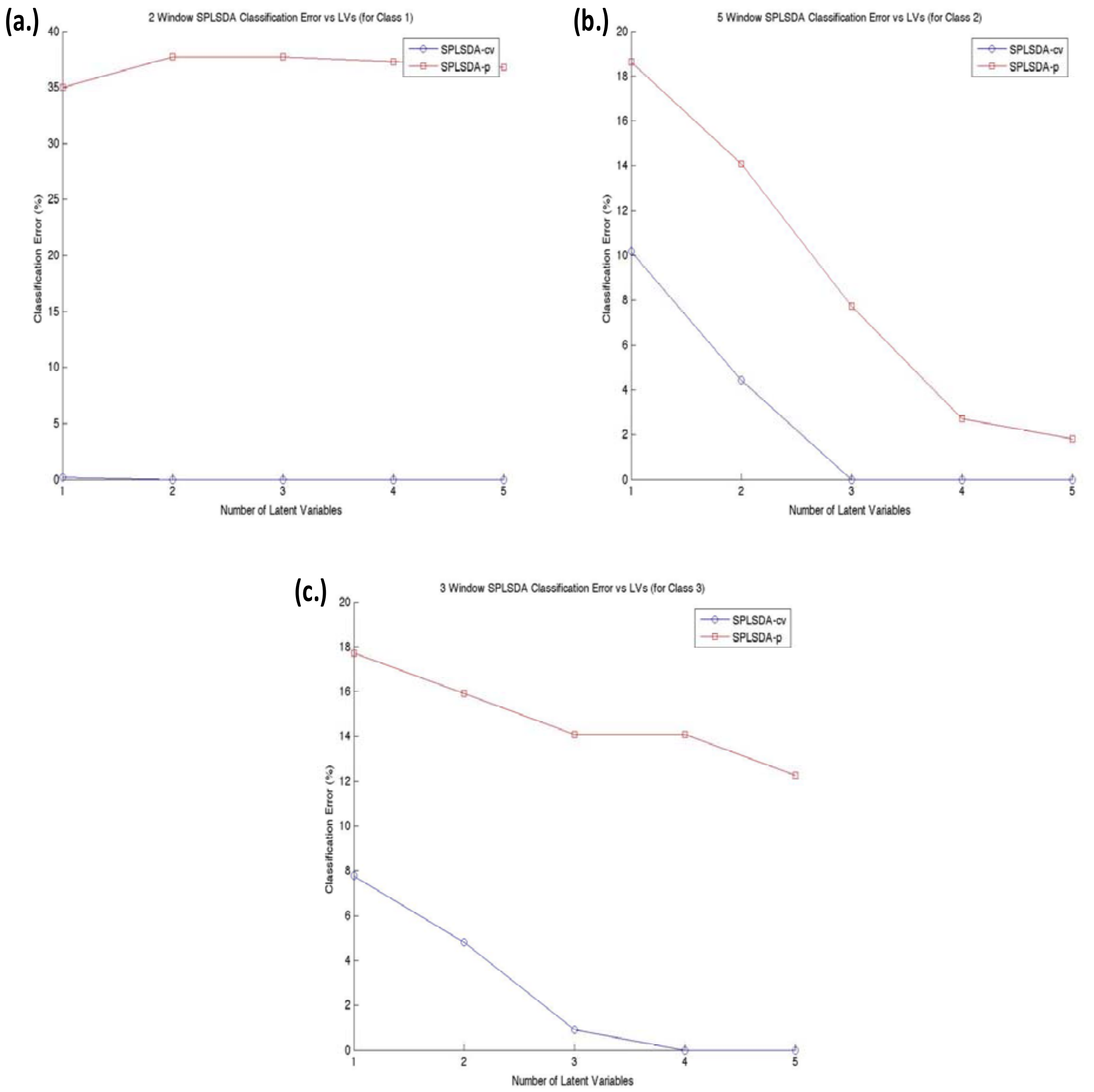


Figure 49. Cross validated error rate (blue) for the training set and prediction error rates (red) for the validation set as a function of the number of latent variables for each of the wavelength windows used in stacking: a) Bowling Green, b) Hamtramck MI, and c) Orion MI.

Figures 50 shows the cross validated (i.e., training set) and prediction (i.e., validation set) error rates for each assembly plant or plant subgroup comprising Plant Group 3 which was detected by the pattern recognition GA. Search prefilters developed for assembly plant by the stacked PLS classifiers achieved classification success rates of 100% for all assembly plants or plant subgroups comprising Plant Group 3. However, error rates for the validation set were 20% (Lordstown/Plant 17), 25% (Buicks from Oshawa Ontario/Plant 22), 25% (GMC trucks from Oshawa Ontario/Plant 22), and 30% (automobiles from Linden NJ/Plant 16, Chevrolets from

Oshawa Ontario/Plant 22, and trucks from Shreveport LA/Plant 25. By comparison, the pattern recognition GA achieved 100% correct classifications for both the training set and validation set.

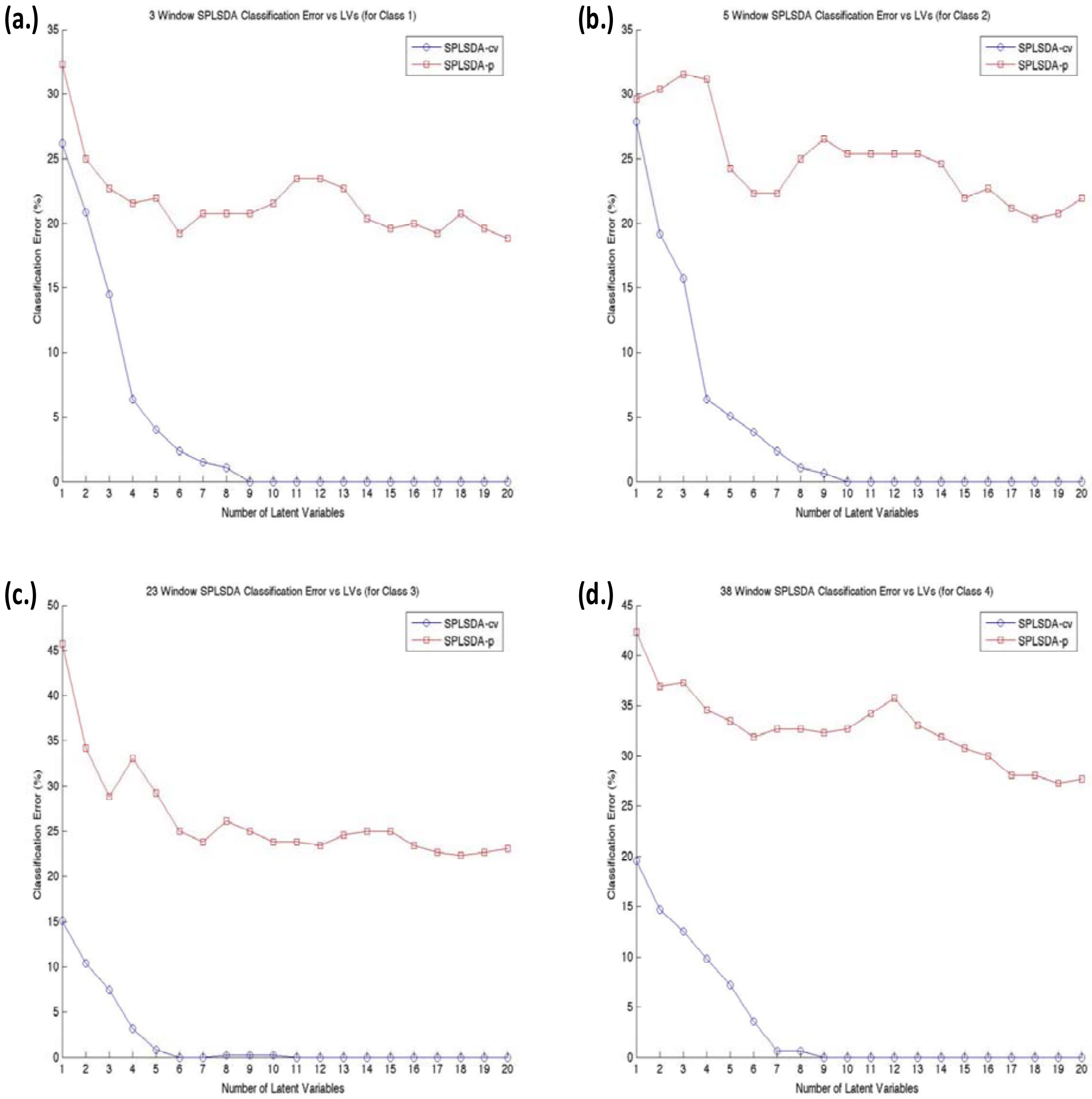


Figure 50. Error rates for the training set (blue) and prediction error rate (red) for the validation set as a function of the number of latent variables for the wavelength windows used in stacking: a) Lordstown/Plant 17, b) Buicks from Oshawa Ontario/Plant 22, c) GMC trucks from Oshawa Ontario/Plant 22, and d) automobiles from Linden NJ/Plant 16, Chevrolets from Oshawa Ontario/Plant 22, and trucks from Shreveport LA/Plant 25

The results obtained in this study suggest that identifying specific wavelengths in the data is superior to identifying informative wavelength regions when applying pattern recognition methods to IR spectra from the PDQ database when differentiating paint samples by assembly plant. The proposed two step procedure for development of search prefilters, which involves the application of wavelets to decompose each IR spectrum into wavelet coefficients that represent both the high and low frequency components of the signal and the use of a genetic algorithm for pattern recognition analysis to identify wavelet coefficients that contain information about the assembly plant of the paint samples, is superior to the use of stacked PLS discriminants in the wavelength domain where the reciprocal of the cross validated error rate for each PLS classifier is used as the weighting value in the stacked models. Search prefilters developed using specific wavelengths or wavelet coefficients outperformed search prefilters that utilized specific wavelength windows. The similarity of the IR spectra within a plant group and the noise present in the spectra may be obscuring information present. Clear coat paint spectra from the PDQ database may not be well suited for stacking as there are few spectral intervals which can reliably distinguish the different sample groups (i.e., assembly plants) in the data. Furthermore, the information contained in the IR spectra about assembly plant may not be highly compartmentalized in an interval which also works against stacking. The large difference in the error rate between the validation set and training set for the stacked models could be indicative of overfitting by PLS.

### Search Prefilters and Cross Correlation Library Searching Algorithm for Identification of Mid Infrared Spectra of Clear Coat Paint Smears

The objective of this study is to assess the evidentiary information content of clear coat paint smears using pattern recognition techniques to search the IR spectral library of the PDQ database to differentiate between similar but nonidentical IR spectra. The PDQ database uses a text based search system which relies on the large variation in color and chemical formulation of the color coat and undercoat layers of automotive paint to identify unknown paint samples. Currently, modern automotive paints are using thinner undercoat and color coat layers protected by a thicker clear coat. As a result, only a clear coat paint smear is often the only layer of paint left at the crime scene. In these cases, PDQ cannot identify the make, line, and model of the motor vehicle as the clear coat paint layer does not contain color pigments. Furthermore, the chemical composition of clear coats is limited to only one of two possible formulations: acrylic melamine styrene or acrylic melamine styrene polyurethane. Pattern recognition methods applied directly to the IR spectra of clear coats may hold the potential for more specific searches of the PDQ database.

At present, the capability to perform direct searching of IR spectra in the PDQ database does not exist, and commercially available spectral library search algorithms often cannot distinguish subtle differences between clear coat paint spectra from one model or line to the next. To tackle the problem of IR library searching in the PDQ database, a prototype library search system has been developed. The prototype system consists of two distinct but interrelated components: **search prefilters** to cull the library spectra in the PDQ database to a specific assembly plant or assembly plants and a **cross correlation searching algorithm** to identify IR spectra that are most similar to the unknown in the subset of spectra identified by the search prefilters.

**Search prefilters** utilizing pattern recognition techniques have been applied to IR spectral databases with only limited success [49-51]. The reasons can be attributed to the nature of the modeling problem, which is often complex as structure-spectrum relationships cannot always be solved using a single spectral band indicative of the functional group present in a compound or the property of a material. The most significant wavelengths used to develop functional group search prefilters for an IR spectral library often bear no relationship to the characteristic frequencies of the specific functional group. Some spectral features are included in the search prefilter for a negative classification of potential interfering compounds rather than for a positive identification of the functional group of the compound or the property of the material [52].

The **cross correlation** function used by the **cross correlation searching algorithm** is able to differentiate between similar but nonidentical IR spectra and identify unknown spectra [35]. Other advantages of correlation based searches include sensitivity to minor peaks and to small shoulders on peaks, and insensitivity to instrumental noise. By comparison, most commercial infrared library search algorithms compare IR spectra by summing the squares of the difference between spectra at each wavenumber. Minor peaks and shoulders on peaks, which can be crucial for identifying an unknown paint sample, are generally ignored when the Euclidean distance is used as a similarity metric to match spectra, while instrumental noise can often obscure matches when using the Euclidean distance to compare IR spectra.

The specific goals of this study are two-fold: (1) Development of search prefilters to identify the manufacturing plant of an automobile from a wavelet transformed IR spectrum of a clear coat paint smear, and (2) Development of a library search algorithm based on the cross correlation function to identify potential matches in the library after culling library spectra using search prefilters to increase the selectivity and accuracy of the search. Using the wavelet packet transform, clear coat paint spectra are passed through two scaling filters: a high pass filter and a low pass filter. This decomposition process yields wavelet coefficients that represent both high and low frequency components of the signal. The data are then iterated using successive wavelet packets until the required level of signal decomposition has been achieved.

The wavelet packet transform is analogous to using a magnifying lens to find minute detail in IR spectra. Wavelets at various scales are analogous to lenses of different magnifying power. Converting an IR spectrum to the wavelet domain will deconvolute it. Wavelets increase the signal to noise of the data by concentrating the signal in specific wavelet coefficients. Wavelet coefficients characteristic of the vehicle assembly plant can be readily identified using the pattern recognition GA.

A major challenge in this study is that IR spectra of the clear coats were not properly aligned along their x or y-axes as these spectra were collected by four different IR spectrometers (BioRad 40A, BioRad 60A, and two Thermo-Nicolet 6700 FTIR spectrometers). There are differences in the alignment of the optical systems as the spectrometers are from different vendors and were manufactured in different years. Careful and judicious preprocessing of the data is necessary for the development of the search prefilters and for the implementation of library matching using either the cross correlation search algorithm or commercial library searching algorithms. In this study, spectral line shapes between instruments are matched by convolution and deconvolution functions developed using Nicolet's OMNIC software system. An instrumental line function representative of the Thermo Nicolet instruments and developed in OMNIC was applied to the Bio-Rad spectra to ensure that all measurements made by the Bio-Rad instrument were comparable to spectra collected on the two Thermo-Nicolet instruments. This ensured wavelength alignment along the x-axis for all clear coat spectra in the PDQ library.

For alignment along the y-axis (transmittance) of the spectra, the editor in OMNIC was used to ensure that all IR spectra started from the same transmittance value. The thickness of the sample and the pressure applied by the transmission diamond cell in collecting the spectra were such that an absorbance of unity was obtained for the carbonyl stretching band in all paint spectra in the PDQ library. Spectral alignment along the y-axis using the carbonyl stretching band was straight forward as there are no sloping baselines or baseline offsets in the diamond cell transmission spectra.

In this study a genetic algorithm for pattern recognition analysis was used to identify wavelet coefficients that optimized the separation of the classes (automobile assembly plants) in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these feature sets will be about differences between classes in the data set. Chance classification is not a serious problem as the bulk of the variance or information content of the wavelet coefficients selected is about the classification problem of interest. Wavelet coefficients that contain discriminatory information about a particular classification problem may be correlated, which is why feature selection methods that utilize PCA can be advantageous.

The PC plot in the fitness function of the pattern recognition GA functions as an embedded information filter. Wavelet coefficients are selected based on their PC plots. A good PC plot can only be generated using coefficients whose variance or information is primarily about differences between the classes. Hence, PCA limits the search to these types of features, significantly reducing the size of the search space.

For this study, search prefilters were developed using 478 IR spectra of clear coats paint smears from automobiles and trucks assembled at 24 different General Motors (GM) industrial plants. The paint samples were obtained from metallic automotive substrates. The automotive assembly plants were divided into five major plant groups based on visual analysis of the IR spectra. Table 24 lists the 24 GM automotive plants used in this study.

**Table 24. GM Plants used to Develop the Prefilters for Spectral Library Searching**

| Plant ID | Plant | Make | Line |
|---|---|---|---|
| 1 | ARL | CAD, CHE, GMC | SUB,YUK,ESD,CTA |
| 3 | BOW | CAD,CHE | CVT,XLR |
| 4 | DOR | PON | VTR,SIL,MTA,UPL,TAR |
| 5 | FAI | CHE,OLD,PON | GRA,MAL,ITR |
| 6 | FLI | CHE,GMC | SLV,SIE |
| 8 | FOR | CHE,GMC | SLV,SIE |
| 9 | FRE | GMC | VIB,TAC,PVB,COA,GPR |
| 10 | HAM | BUI,CAD,PON | BON,DEV,LUC,LES,SEV,ELD |
| 11 | INE | CHE, PON | EQU, GEM, TKR, TOR |
| 12 | JAN | GMC | CTA,SUB,YUK |
| 14 | LAN | PON | STS |
| 16 | LIN | CHE,GMC | BZR,JMY,S10 |
| 17 | LRD | PON | SFR,CAV,COB,PST |
| 18 | MOR | CHE,GMC,SAA | JMY,ENV,9S7,BZR,TBZ,SON |
| 20 | OKL | CHE, GMC | MAL,TBZ,ENV,EQU, XUV |
| 21 | ORI | PON,BUI | BON,PG6,LES,AUR, PKA |
| 22 | OSH | GMC,PON | ALL,REG |

| | | | |
|---|---|---|---|
| 23 | PON | CHE,GMC | SLV,SIE,SIL |
| 24 | RAM | BUI,CHE,PON | CAV,SFR,RZV,AZT,HHR |
| 25 | SHR | CHE,GMC | S10,COL,SON |
| 26 | SIL | CHE,GMC,SAA | AVL,SUB,YXL |
| 27 | SPH | STR | SSL,ION,SC1,SC2,SL1,VUE |
| 28 | THE | CHE, PON | CMR, FBD |
| 30 | WLM | PON, STR | SOL, LS1, LS2, LSN, LW1, LWT |

A hierarchical classification scheme was used to develop the search prefilters for the clear coat IR spectra of the PDQ database. First, an unknown clear coat paint sample is classified by plant group and then a second search prefilter is used to identify the assembly plant or assembly plants within the plant group to which the unknown is assigned. Assembly plants for each plant group are listed in Table 25. The 478 IR spectra that comprise the training set for the development of the search prefilters are summarized in Table 26.

**Table 25. Manufacturing Plants Comprising Each Plant Group**

| Plant Group | Plant ID Number | Manufacturing Plant |
|---|---|---|
| 1 | 1, 4, 5, 8, 14, 18, 23 | ARL, DOR, FAI, LAN, MOR, PON |
| 2 | 3, 10, 21, 30 | BOW, HAM, IRI, WLM |
| 3 | 6, 9, 11, 16, 17, 20, 22, 25 | FLI, FRE, INE, LIN, LRD, OKL, OSH, SHR |
| 4 | 12 | JAN |
| 5 | 24, 26, 27, 28 | RAM, SIL, SPH, THE |

**Table 26. Training Set for Plant Group Search Prefilter**

| Group | Number of Training Set Samples |
|---|---|
| 1 | 164 |
| 2 | 60 |
| 3 | 146 |
| 4 | 20 |
| 5 | 88 |
| Total | 478 |

The first step in this study was to apply PCA to the wavelet transformed IR spectral data. Using this procedure is analogous to finding a new coordinate system better at displaying the information content of the data than axes defined by the original measurement variables. This new coordinate system is linked to variation in the data. Often, only two or three principal components are necessary to explain most of the information present in spectral data due to the large number of interrelated measurements.

The Symlet6 mother wavelet at the 8th level of decomposition was used to deconvolve the IR spectra prior to PCA. Figure 51 shows the plot of the two largest principal components of the 478 IR spectra and 1080 wavelet coefficients of the training set. Each paint sample is represented as a point in the PC plot. The overlap of the clear coats by plant group is evident from an examination of the PC plot of the data.

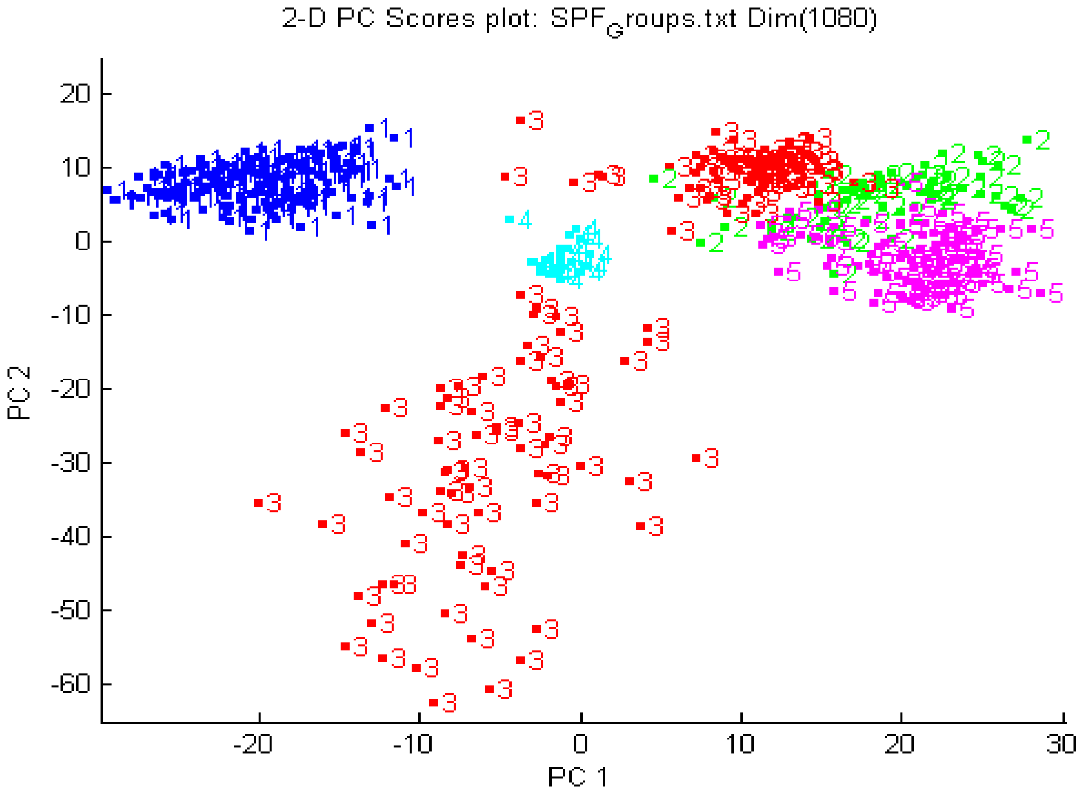


Figure 51. Plot of the two largest principal components of the 478 IR spectra and 1080 wavelet coefficients of the training set. Each paint sample is represented as a point in the PC plot: 1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5.

The pattern recognition GA identified wavelet coefficients characteristic of plant group by sampling key feature subsets, scoring their PC plots, and tracking those plant groups and/or wavelet transformed 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 20 wavelet coefficients whose PC plot (see Figure 52) showed clustering of the IR spectra on the basis of plant group.

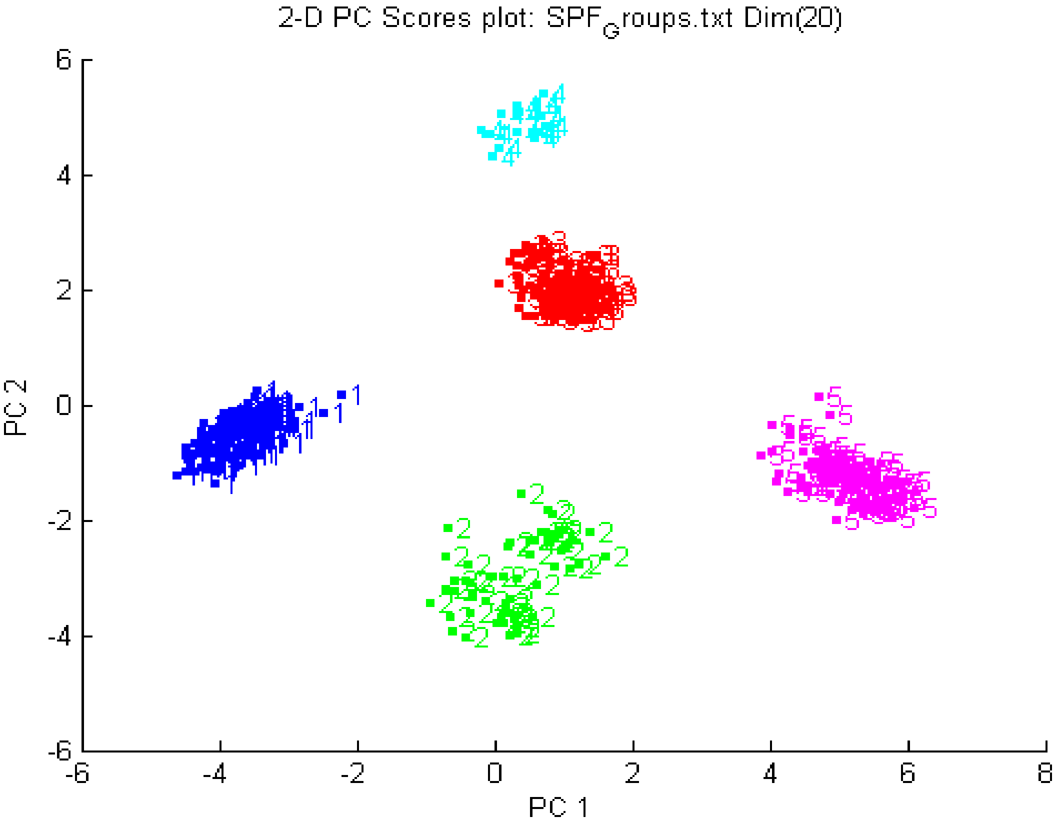


Figure 52. Plot of the two largest principal components of the 478 IR spectra and 20 wavelet coefficients identified by the pattern recognition GA. Each paint sample is represented as a point in the PC plot: 1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5.

To assess the predictive ability of the 478 IR spectra and the 20 wavelet coefficients identified by the pattern recognition GA, 22 blind (validation set) samples were used. Spectra from the blind samples were projected directly onto the PC map developed from the 478 wavelet transformed spectra and the 20 wavelet coefficients identified by the pattern recognition GA. Figure 53 shows the projection of the blind samples onto the PC map of the training set data. 21 of the 22 projected blind samples are located in a region of the PC map occupied by IR spectra possessing the same class label. The misclassified blind sample is a clear coat (blind sample 8) obtained from a plastic bumper. Often, the clear coat layer is applied to plastic automotive components at the plant where the component is manufactured, not at the plant where the vehicle is assembled. For this reason, all paint samples used to develop search prefilters in this study have been limited to metal automotive substrates.

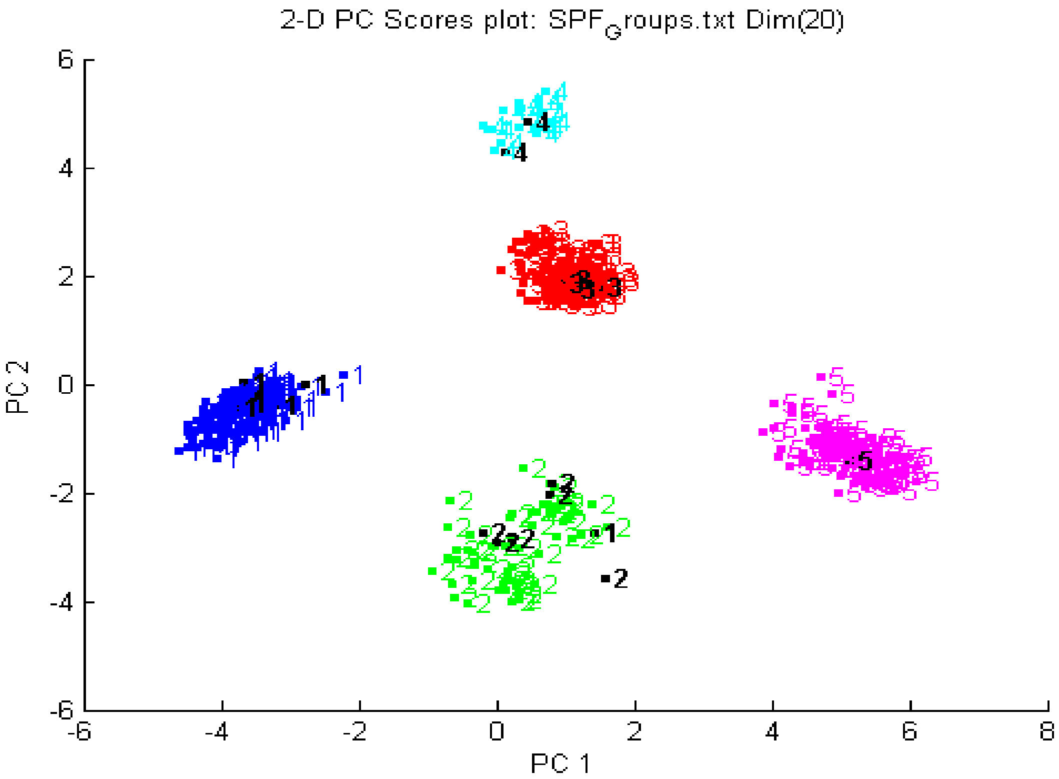


Figure 53. Projection of 22 blind samples (in black) onto the PC map developed from the training set. Each paint sample is represented as a point in the PC plot (1 = Plant Group 1, 2 = Plant Group 2, 3 = Plant Group 3, 4 = Plant Group 4, and 5 = Plant Group 5).

The next step in this study is to develop search prefilters to identify the 22 blind samples by assembly plant. For the first three plant groups, a search prefilter was developed for each plant group to discriminate the wavelet transformed IR spectra of the clear coats by assembly plant. (Plant Group 4 consists of a single plant and the IR spectra from the 4 assembly plants comprising Plant Group 5 cannot be differentiated as their spectra are too similar.) Figure 54 shows a plot of the two largest principal components of the 164 IR spectra from Plant Group 1 (see Table 22) and the 33 wavelet coefficients identified by the pattern recognition GA which contain information about the assembly plant. Each IR spectrum is represented as a point in this PC plot. Plant 18 (Moraine OH) is well separated from the other 6 manufacturing plants. IR spectra from the other 6 assembly plants (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI) are superimposable, which prevented further discrimination by assembly plant for these clear coats.

The predictive ability of the 33 wavelet coefficients identified by the pattern recognition GA was assessed by projecting the 8 blind samples assigned to Plant Group 1 (using the search prefilter developed for Plant Group) directly onto the PC map of the 164 IR spectra comprising the training set. Figure 55 shows the projection of the 8 blind samples onto the PC map of the data. Every blind sample is in a region of the map occupied by IR spectra possessing the same class label.

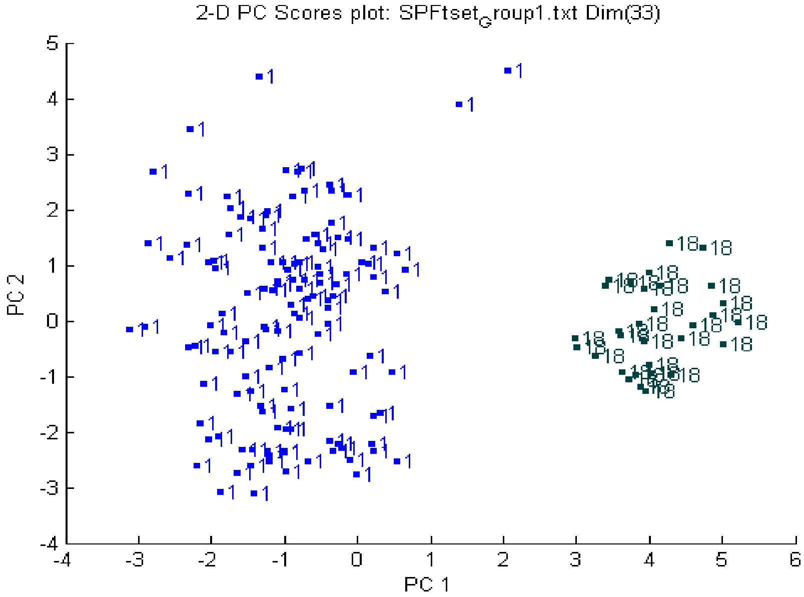


Figure 54. Plot of the two largest principal components of the 164 IR spectra from Plant Group 1 and the 33 wavelet coefficients identified by the pattern recognition GA for assembly plant. Each paint sample is represented as a point in the PC plot: 1 = Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, Pontiac MI, and 18 = Moraine OH

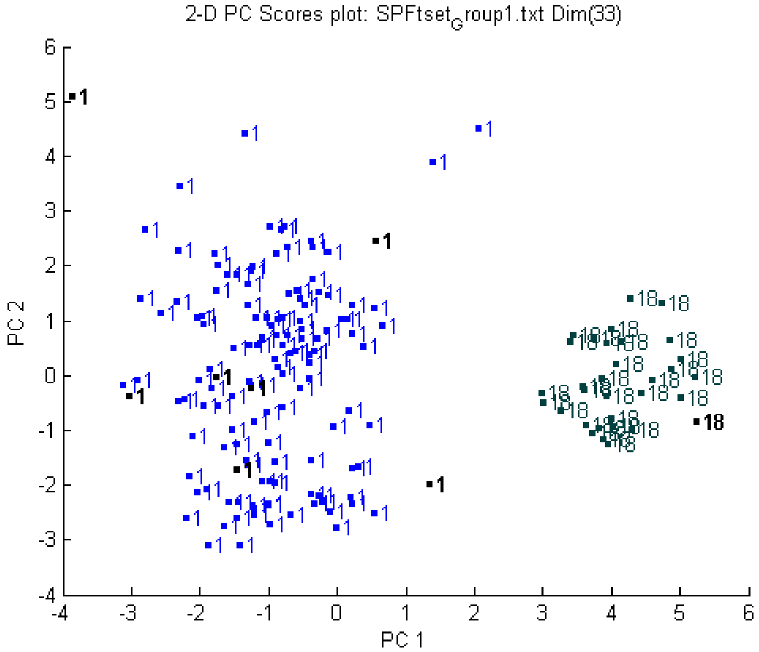


Figure 55. Projection of the 8 blind samples (black) onto the PC map of the 164 IR spectra from Plant Group 1 and the 33 wavelet coefficients identified by the pattern recognition. Each paint sample is represented as a point in the PC plot: 1 = Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, Pontiac MI, and 18 = Moraine OH

Figure 56 shows a PC plot of the 60 IR spectra and the 33 wavelet coefficients identified by the pattern recognition GA for the clear coats from Plant Group 2. Again, each IR spectrum is represented as a point in the PC plot. The Bowling Green KY and Wilmington DE Plants are well separated from each other and from the Hamtramck MI and Orion MI plants in the PC plot of the data. However, the paint samples from Hamtramck and Orion MI plants overlap in the plot due to the similarity of their IR spectra.

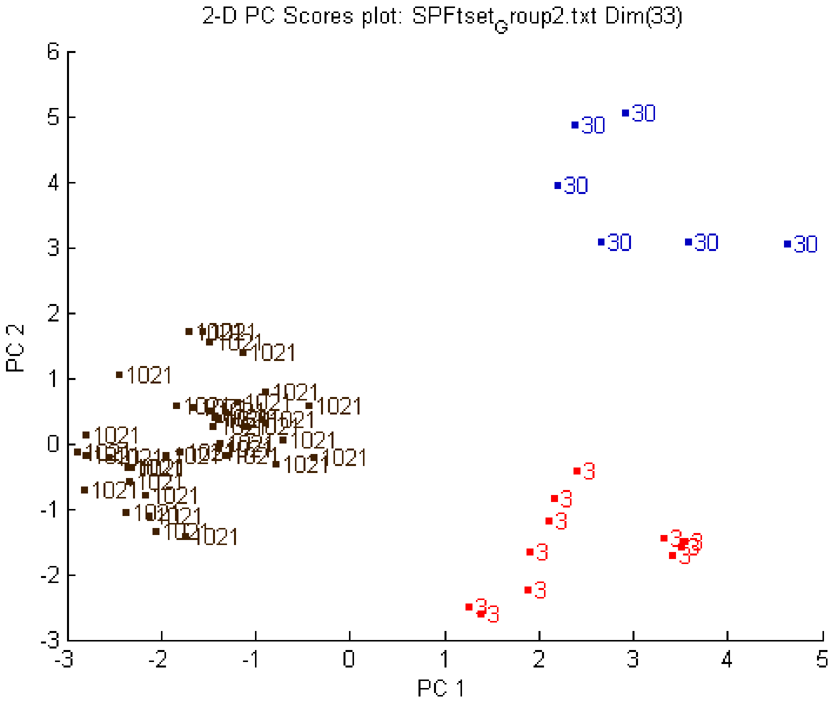


Figure 56. Plot of the two largest principal components of the 60 wavelet transformed IR spectra and the 33 wavelet coefficients identified by the pattern recognition GA for the paint samples from Plant Group 2. Each spectrum is represented as a point in the PC plot of the data. 3 = Bowling Green KY, 10 = Hamtramck MI, 21 = Orion MI, and 30 = Wilmington DE.

6 of the 7 blind samples assigned to Plant Group 2 were used to assess the predictive ability of the 33 wavelet coefficients identified by the pattern recognition GA. (The misclassified blind sample from a plastic automotive component was excluded.) We chose to map the 6 IR spectra directly onto the PC map defined by the 60 transformed IR spectra and the 33 wavelet coefficients identified by the pattern recognition GA. Figure 57 shows the 6 blind (validations set) samples projected onto the PC plot of the assembly plants which define Plant Group 2. All 6 blind samples are in a region of the map that contains paint samples from the same automotive assembly plant.

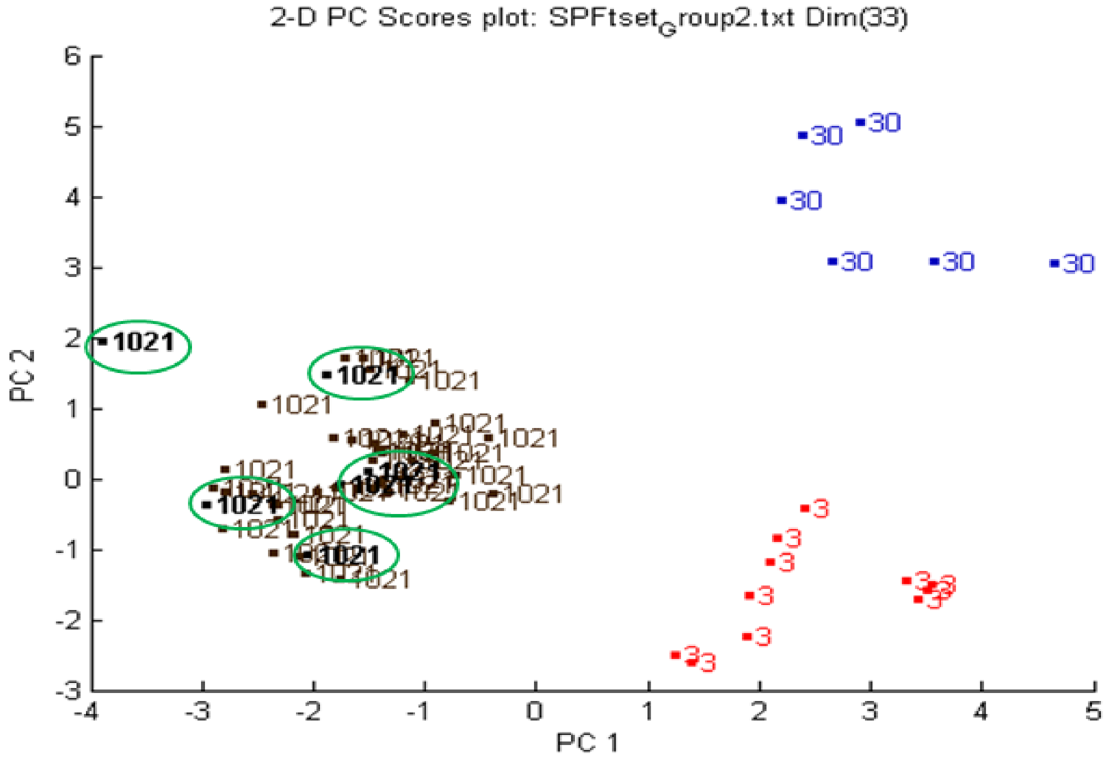


Figure 57. Projection of the 6 blind samples (in black and circled) onto the PC plot of 60 wavelet transformed spectra and the 33 wavelet coefficients identified by the pattern recognition GA**.** Each projected paint sample is located in a region of the map with paint samples that have the class label. 3 = Bowling Green KY, 10 = Hamtramck MI, 21 = Orion MI, and 30 = Wilmington DE.

Figure 58 shows a PC plot of the 146 IR spectra and the 15 wavelet coefficients identified by the pattern recognition GA to differentiate clear coats by assembly plant for Plant Group 3. Clustering of the paint spectra by assembly plant and automotive model is evident from an examination of this PC plot. Automotive vehicles from Plants 9 (Freemont CA), 11 (Ingersoll ON) and 17 (Lordstown OH) and trucks from Plant 22 (Oshawa ON) form distinct clusters in the plot. Trucks from Plant 20 (Oklahoma City) and Buicks from Plant 22 (Oshawa Ontario) form another cluster. A plant subgroup consisting of Chevrolet automobiles (Plant 20 - Oklahoma City and Plant 22 – Oshawa Ontario), GMC trucks (Plant 25 -Shreveport Louisiana and Plant 6 - Flint Michigan), and Chevrolet trucks (Plant 16 from Linden NJ) lie in another cluster.

The 4 blind samples assigned to Plant Group 3 were used to assess the predictive ability of the 15 wavelet coefficients identified by the pattern recognition GA. The 4 blind samples were directly mapped onto the PC plot in Figure 58. In Figure 59, the results of this mapping are summarized. All 4 blind samples lie in a region of the map with paint samples that are from the same assembly plant.

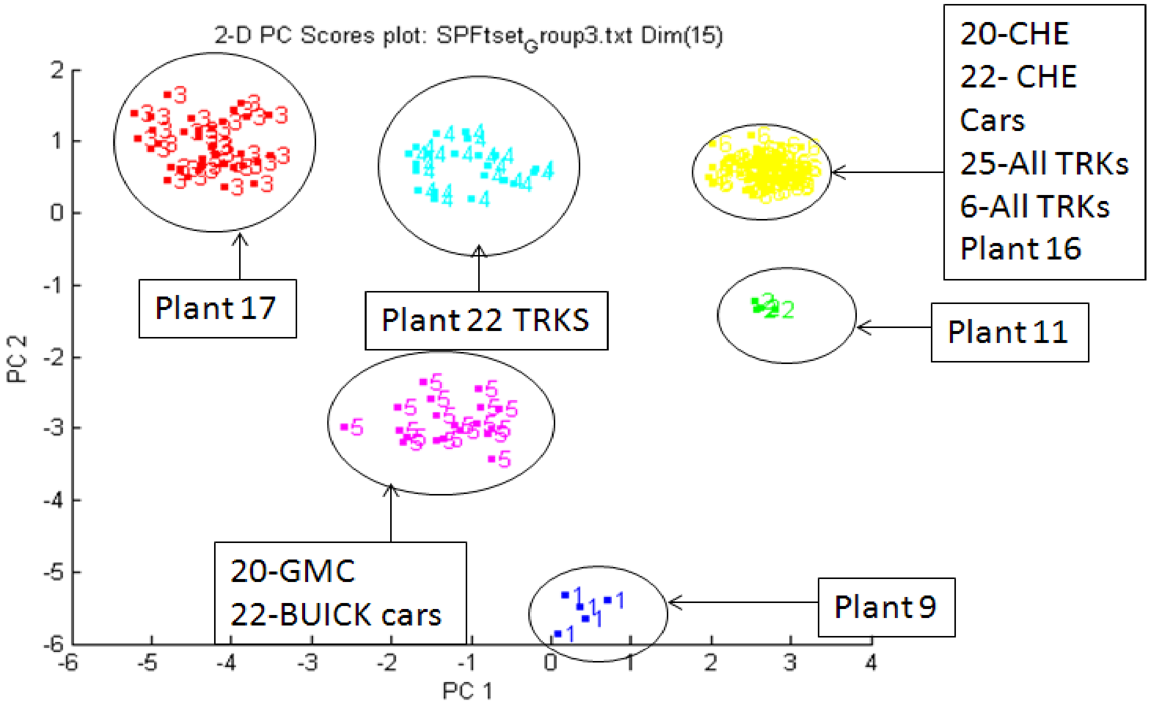


Fig 58. Plot of the two largest principal components of the 146 transformed IR spectra and 15 wavelet coefficients identified by the pattern recognition GA. 1 = Fremont CA, 2 = Ingersoll ON, 3 = Lordstown OH, 4 = Oshawa ON, 5 = Oklahoma City and Oshawa ON, and 6 = Oklahoma City, Oshawa ON, Shreveport LA, Flint MI, Linden NJ.

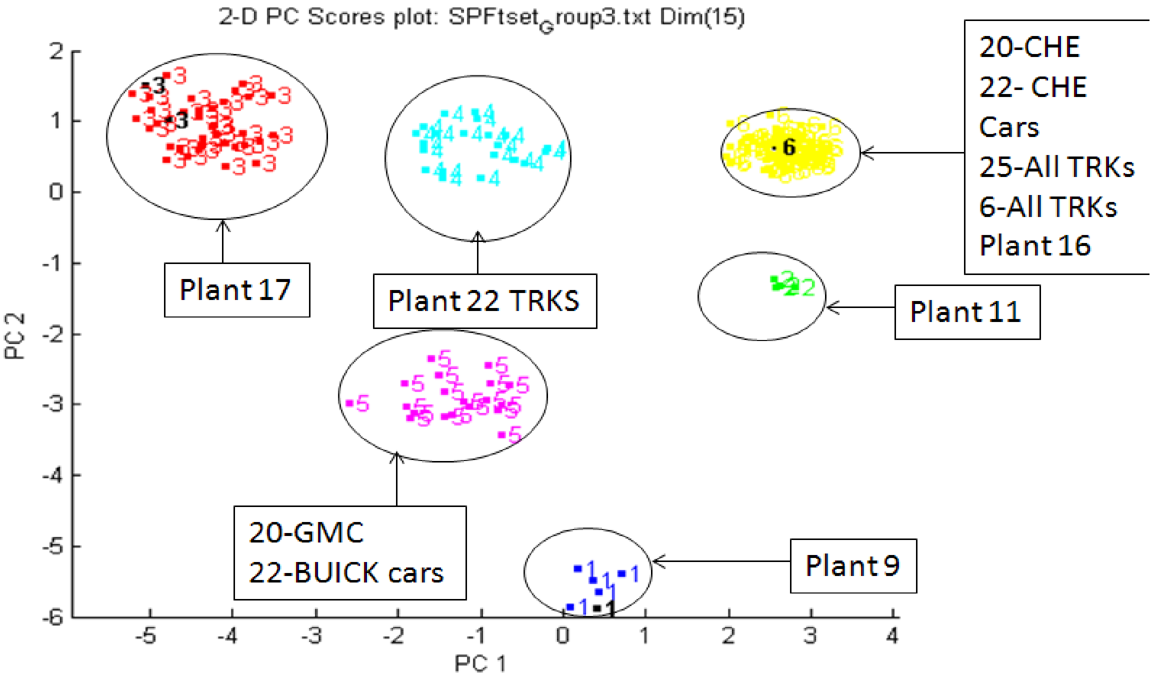


Fig 59. Projection of the 4 blind samples (in black) assigned to Plant Group 3 onto the PC plot of the 146 transformed IR spectra and 15 wavelet coefficients identified by the pattern recognition GA. 1 = Fremont CA, 2 = Ingersoll ON, 3 = Lordstown OH, 4 = Oshawa ON, 5 = Oklahoma City and Oshawa ON, and 6 = Oklahoma City, Oshawa ON, Shreveport LA, Flint MI, and Linden NJ.

A summary of the search prefilter results for the blinds is given in Table 27. 21 of the 22 blind samples were correctly classified by plant group and by assembly plant using the hierarchical classification scheme employed. The misclassified blind sample is a clear coat paint smear from a plastic bumper. The search prefilters were developed using IR spectra collected on four different instruments from two different manufacturers that were aligned along both the wavelength and absorbance axes using OMNIC software. This approach allowed for a direct comparison of spectra even if they are collected on different spectrometers manufactured several years apart.

**Table 27. Summary of Search Prefilter Results for 22 Blinds**

| **Blind Sample** | **Assigned Plant Group** | **Assigned Plant(s)** | **ID of Blind Sample** |
|---|---|---|---|
| 1 | 2 | 10, 21 | 10 |
| 2 | 1 | 1, 4, 5, 8, 14, 23 | 14 |
| 3 | 2 | 10, 21 | 21 |
| 4 | 2 | 10, 21 | 10 |
| 5 | 2 | 10, 21 | 21 |
| 6 | 3 | 17 | 17 |
| 7 | 1 | 1,4,5,8,14,23 | 5 |
| 8 | 2 | Not Applicable | 14 (incorrectly classified) |
| 9 | 2 | 10, 21 | 10 |
| 10 | 1 | 1,4,5,8,14,23 | 14 |
| 11 | 1 | 1,4,5,8,14,23 | 14 |
| 12 | 1 | 1,4,5,8,14,23 | 1 |
| 13 | 4 | 12 | 12 |
| 14 | 1 | 1,4,5,8,14,23 | 14 |
| 15 | 1 | 18 | 18 |
| 16 | 4 | 12 | 12 |
| 17 | 2 | 10,21 | 10 |
| 18 | 3 | 9 | 9 |
| 19 | 5 | 24,26,27, 28 | 26 |
| 20 | 3 | 17 | 17 |
| 21 | 3 | 6, 16, 20, 22, 25 | 6 |
| 22 | 5 | 24, 26, 27, 28 | 24 |

To extract information about the model of the automobile from the blinds, the cross correlation searching algorithm was used to identify the IR spectra most similar to the unknown in the subset of IR spectra identified by the search prefilters. During this phase of the study, the IR spectra were vector normalized and compared to library spectra using the region from 1843 to 667 $cm^{-1}$. This spectral region contains the carbonyl band, which was excluded from the spectral region interrogated by the search prefilters. Two different types of comparisons were undertaken for spectral library matching using the cross correlation algorithm: (1) each autocorrelated blind was compared with the cross correlated blind and library spectrum, and (2) each autocorrelated library spectrum was compared with the cross correlated blind and library spectrum. Comparisons involved different size windows across the midpoint. A histogram of the results was made with the top 5 matches of each comparison used to find the closest matching sample. The cross correlation library searching method was then compared with the top 5 matches identified by OMNIC, a commercially available IR search algorithm considered by many spectroscopists in the field as the industry standard. Library search results for both the cross correlation searching algorithm and OMNIC are summarized in Table 25. The cross correlation searching algorithm outperformed OMNIC. Spectral library matching using OMNIC yielded very similar results whether the entire PDQ library was accessed by OMNIC or whether the search by OMNIC was restricted to a subset of IR spectra identified by the search prefilters.

**Table 28. Results from Library Search of PDQ Database**

| **Library Searching Method** | **Number of Correct Matches In Top 5** | **Total Number of Blinds** |
|---|---|---|
| Cross Correlation | 20 | 21 (plastic sample discarded) |
| OMNIC | 12 | 21 (plastic sample discarded) |

The IR spectrum of the blind sample that was not correctly identified by the cross correlation searching algorithm was compared to the IR spectrum of a clear coat of the same model and line and from the same assembly plant as the blind (see Figure 60a). When these two IR spectra are overlaid and compared to the overlay between the top hit identified by the cross correlation searching algorithm and the blind (see Figure 60b), it is evident that a better match is achieved using the cross correlation searching algorithm. This result strongly suggests that our cross correlation searching algorithm is able to identify the IR spectrum in the prescreened PDQ library most similar to the unknown.

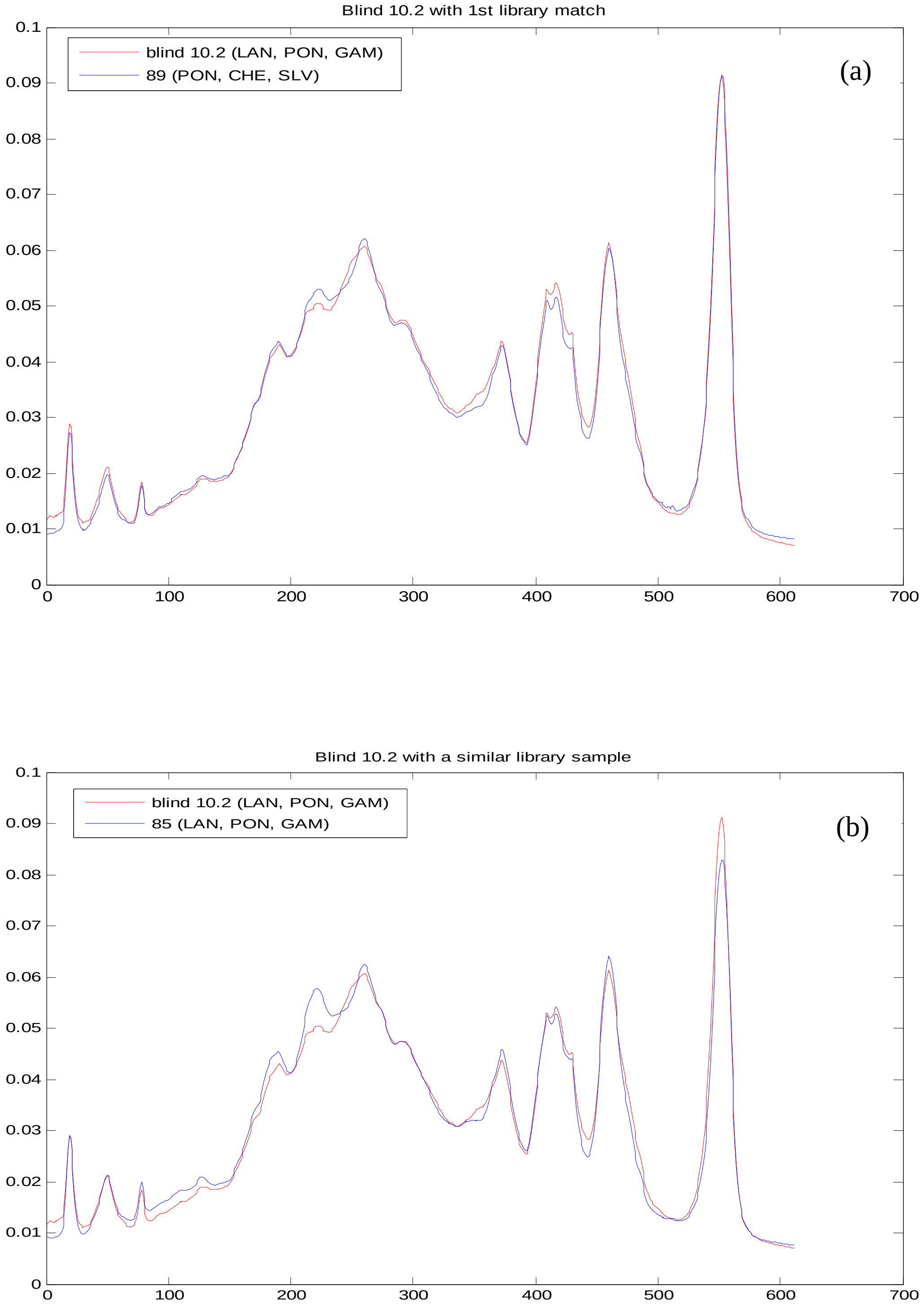


Figure 60. IR spectrum of blind sample (red) overlaid with (a) top hit from cross correlation search algorithm (blue) and b) with PDQ library spectrum (blue) of the same model and line and from the same assembly plant.

A second validation set was provided by the RCMP to further assess the performance of the search prefilters and the cross correlation library search algorithm. This second set consisted of GM vehicles from 2007-2013, not 2000-2006 which is the production year range of GM vehicles in the PDQ database used to identify these validation set samples. Of the 171 samples comprising this second validation set, 39 did not fit the search prefilters developed from IR spectra in the PDQ database as the formulation of the clear coats used for these specific lines had changed in the most recent production years. For 43 validation set samples, the corresponding lines did not exist in the PDQ database, and for 14 samples, the IR spectra of the specific lines from the PDQ database corresponding to these samples did not resembled these clear coats. Thus, 75 IR spectra of clear coats from the second validation set were available for matching as their lines are present and their IR spectra are comparable to those in the PDQ library. In this study, the IR spectra were divided into three regions (see Figure 61) with each region given equal weight in the analysis. The cross-correlation library search routine was compared to the library matching routine used in OMNIC under similar conditions without the use of search pre-filters. Library matching using OMNIC was configured for correlation as search type with Happ-Genzel apodization as these produced the best results. 75 unknown clear-coat IR spectra were investigated using both OMNIC and the cross correlation searching algorithm. 55 of the 75 IR spectra were correctly matched by the cross-correlation searching algorithm whereas 44 of the 75 IR spectra were correctly matched by OMNIC.

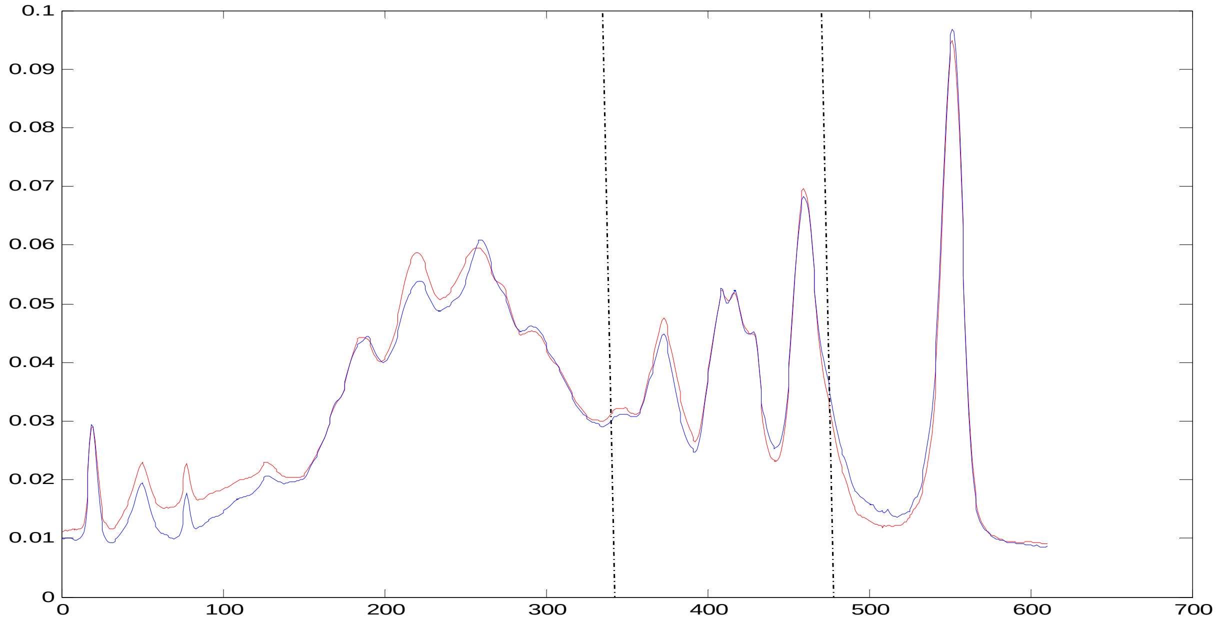


Figure 61. Each clear coat IR spectrum was divided into three regions for spectral library matching. The three regions are 1-335 (667 – 1350 cm$^{-1}$), 336-470 (1351 – 1575 cm$^{-1}$), and 471-611 (1576 – 1843 cm$^{-1}$).

The clear-coat paint spectra incorrectly matched by the cross-correlation search algorithm were also incorrectly matched by OMNIC. The top hit selected by the cross-correlation searching algorithm always yielded a reasonable match, often superior to the match between the validation set sample and the actual line (see Figure 62). When compared to IR spectra of the same line from 2004-2006 (see Figure 63), several clear coat paint spectra from 2007-2010 showed reversals in the carbonyl intensities of the doublet. When these same IR spectra were reanalyzed using the cross correlation searching algorithm without the carbonyl band, they were correctly matched to the appropriate automotive line. In all likelihood, a few assembly plants had a formula change (i.e., the relative amount of polyurethane was changed in their formulation) between 2006 and 2007, which would explain reversals in their carbonyl bands. The larger fraction of IR spectra from this validation set that were incorrectly matched compared to the previous 22 blind sample set can probably be attributed to small changes in the formula of these clear coat paint samples from 20007-2013. Although the search prefilters were insensitive too these changes, the cross correlation searching algorithm was affected due to its ability to recognize small spectral differences among a set of similar spectra. This problem can be readily addressed by ensuring that PDQ is populated to current production years.

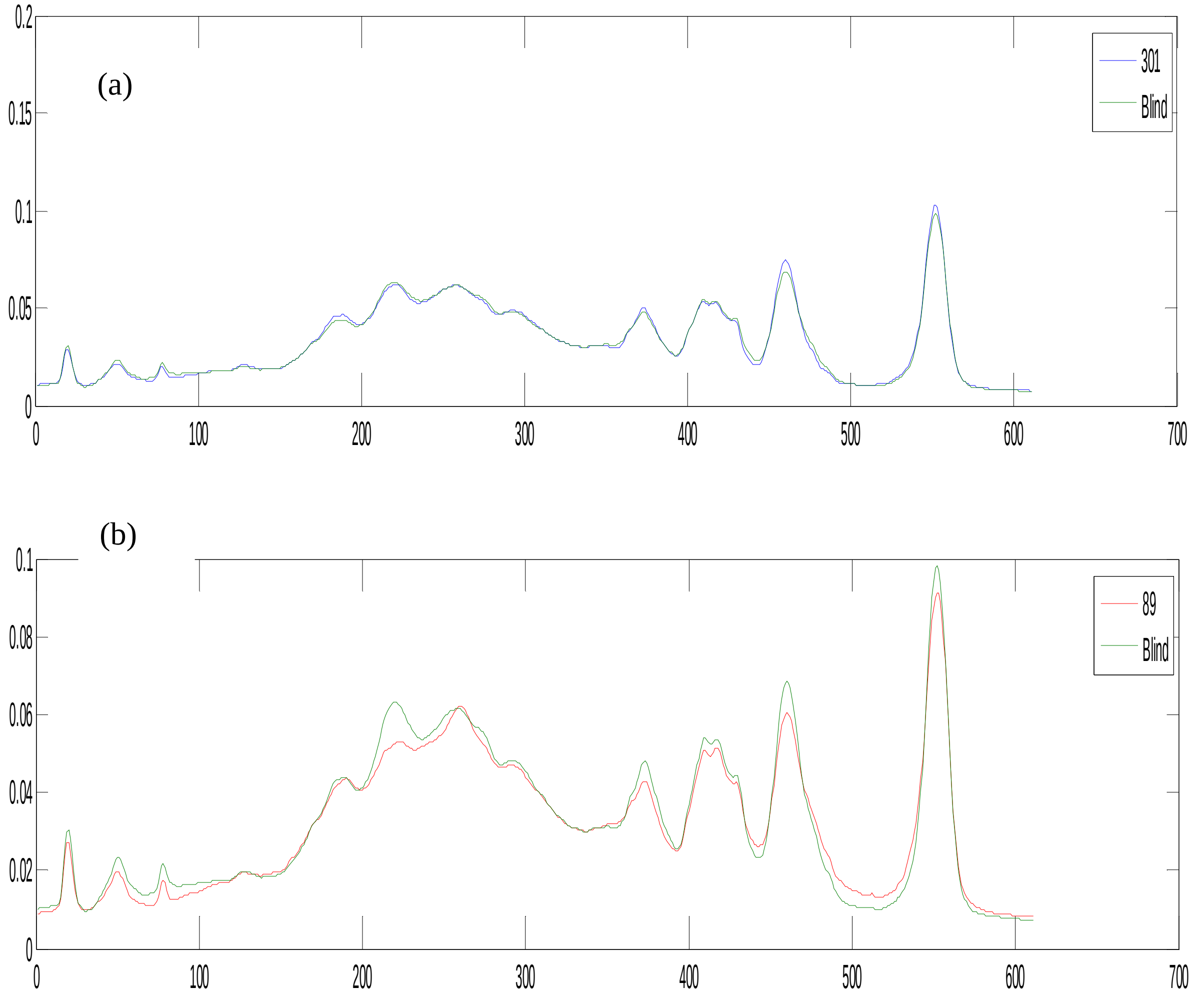
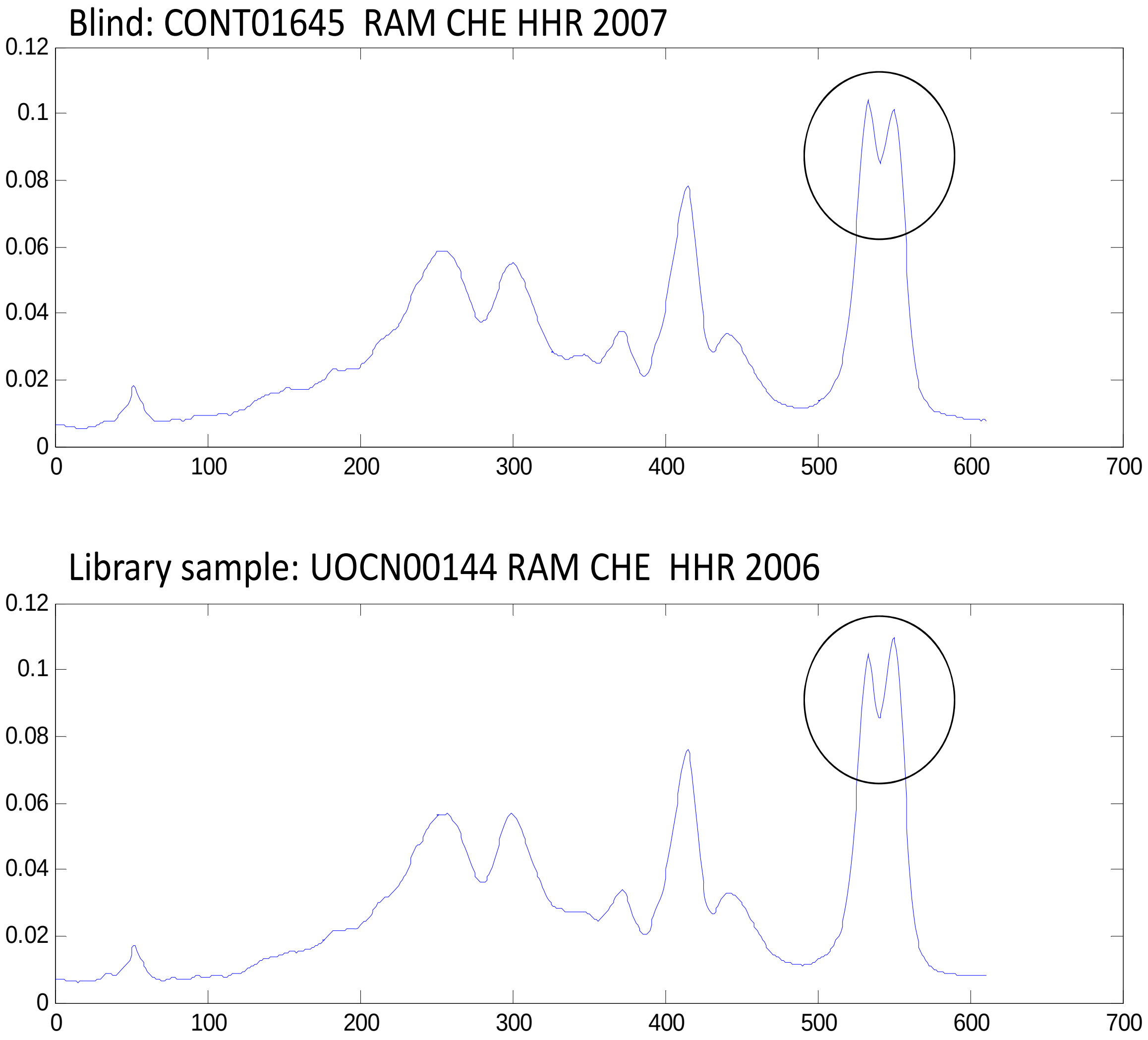


Figure 62. Comparison of validation set (blind) sample with a) first hit, and b) actual line



Blind: CONT01645 RAM CHE HHR 2007

Library sample: UOCN00144 RAM CHE HHR 2006

Figure 63. IR spectra of a Chevrolet HHR from the Ramos Arizpe assembly plant in a) 2007 and b) 2006. Reversal in the carbonyl intensities of the doublet (see circled peaks) can be attributed to a change in the amount of polyurethane in the formulation used to prepare the clear coat paint layer.

## IV. Conclusions

**Discussion of Findings:** A two step procedure for search prefilters has been developed to identify the manufacturing plant of an automobile from its clear coat paint smear. First, search prefilters are employed to divide the IR spectra of the clear coats into distinct plant groups. A genetic algorithm for pattern recognition analysis is used to identify discriminating spectral features characteristic of each plant group. Second, search prefilters are developed for the IR spectra in each plant group to identify the specific manufacturing plant or a subset of manufacturing plants that have similar IR spectra to the unknown. 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 vehicles (2000-2006) can identify the plant or a group of plants where the vehicle was assembled by pattern recognition analysis of the IR spectrum of a clear coat paint smear. 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 plant, identifying the manufacturing 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.

Transfer of search prefilters between FTIR spectrometers can be achieved using the appropriate estimate of the spectral line function for the master instrument. This will ensure that all IR spectra appear to originate from the same instrument. Using OMNIC, it has been shown that alignment of spectra along the wavelength and absorbance axes can be achieved. This is crucial as it permits spectra from different instruments to be used in the development of classifiers for search prefilters.

The autocorrelation transformation can be applied to IR spectra of clear coats to develop search prefilters. Advantages of this preprocessing include the elimination of translational differences between spectra along the wavelength axis and sensitivity at distinguishing subtle but significant features in the data such as minor peaks, shoulders, and peaks with unique shapes, which can play an important role in identifying the model or line of a vehicle from a paint sample. For the clear coat paint spectra investigated, the autocorrelation transformed IR spectra also proved advantageous for the development of search prefilters

Stacked classifiers which utilize ensemble modeling for the identification of wavelength windows within the IR spectra indicative of the manufacturing plant of the paint sample have been investigated and compared to the pattern recognition GA for feature selection. The pattern recognition GA outperformed the stacked classifiers demonstrating that searching for unique spectral features as opposed to informative wavelength windows is a better strategy for feature selection with mid-IR data.

The cross correlation library searching algorithm is sensitive to minor peaks and to shoulders, which can be highly informative in automotive paint layer searches whereas commercial search algorithms often ignore these spectral features due to the nature of the numerical comparison made between each library spectrum and the unknown.

Our studies have also shown that memory effects from previous samples can confound library matching of the clear coats. For this reason, we recommend the following set of procedures when acquiring IR data of clear coats. First, collect a background after running a spectrum and store the interferogram of both the sample and the background. This is crucial when dealing with so-called sticky samples. It would be best not to use a stored background when performing multiple runs because of sample memory effects which occur when under cleaning the substrate or when over cleaning the substrate. Substrates used for both the blind and the library sample should be scrupulously clean. Apodization functions used should be the same for both the library sample and for the unknown. Finally, one should routinely take a single beam spectrum of only the substrate and look for changes over time indicative of either contamination of the substrate or changes in the optical system of the instrument.

The pattern recognition based library searching system utilizes well defined criteria to extract investigative lead information from raw IR data. The techniques that have been found to be the most useful for the development of search prefilters do not attempt to fit the IR data to an exact functional form. Rather, relationships are developed which provide definitions of similarity between diverse groups of IR spectra.

If the limitations of the methodology utilized in the proposed library searching system are not fully understood, the danger of misinterpretation and misuse of spectroscopic data can be significant. For this reason, experimental artifacts must be controlled or otherwise taken into account. Each manufacturing plant, model, and line must be well represented in the database. To maintain relevancy of the newly designed library search system, it is crucial to populate PDQ to manufacturing 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 search algorithm also needs to be undertaken to authenticate the performance of the proposed spectral library searching system. The studies discussed in this report have shown that evidential trace information can be obtained directly from clear coat paint smears using the proposed spectral 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 prototype pattern recognition library search system for the PDQ database has 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. The potential for direct impact exists on over 75 local, state, and federal forensic laboratories that are currently using the PDQ database in the United States. There may also be direct impact on international forensic laboratories using the database, including the Forensic Laboratory Services Division of the RCMP, the Centre of Forensic Sciences in Toronto, Canada, the ENFSI network of European forensic science institutes, the Australian Police Services, and the New Zealand Police Services.

The 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 paint smears 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 clear coats 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 based library searching system has the potential to extract evidentiary lead information from clear coat paint smears. Due to the limited number of formulations used to prepare clear coats, the use of additional information in the search would improve both the selectivity and accuracy of searches undertaken using the proposed library search system. Modern automotive paints have a thin color coat which on a microscopic fragment may be too thin to obtain accurate chemical information. The small size of the fragment will also make it difficult to accurately compare it with manufacturer's paint color standards. Fortunately, adhesion between paint layers is usually very strong and one or both primer layers can be transferred during a collision if the clear coat and color coat layers are also transferred. As the primer layers and clear coat layer are often characteristic of the automotive manufacturing plant where these layers are applied, combining chemical information obtained from the Fourier Transform (FT) IR spectra of the two primer layers and from the clear coat layer should make it possible to rapidly and accurately identify the model, line and the assembly plant of an automobile. 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 can be developed to differentiate between similar but nonidentical FT-IR paint spectra, and to determine the make and model of the vehicle from which an unknown paint sample originated.

## V. References

1. 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. Forens. Sci. J**. 1997, 30, 199-212.

2. N. S. Cartwright, and P. G. Rodgers, "A Proposed Data Base for the Identification of Automotive Paint," **Can. Soc. Forens. Sci. J.** 1976, 9(4), 145-154.

3. G. A. Bishea, J. L. Buckle, and S. G. Ryland, "International Forensic Automotive Paint Database," Proceedings – Investigation and Forensic Science Technologies; International Society of Optical Engineering (SPIE), Vol. 3576, February, 1999, p. 73.

4. A. Beveridge, T. Fung, D. MacDougall, "Use of infrared spectroscopy for the

Characterization of Paint Fragments," In: Forensic Examination of Glass and Paint: Analysis and Interpretation, B. Caddy, ed., Taylor and Francis, NY, NY, pp. 220- 233, 2001.

5. G. Fettis (Editor), Automotive Paints and Coatings, VCH Publications, New York, 1995.

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. Forens. Sci. J.,** 1982, 15(3/4), 105-115.

7. A. Hobbs, “Sifting Through the Layers: The Application of Forensic Databases to Tape and Paint Analyses,” Trace Evidence Symposium, August 13-16, Clearwater Beach, FL, 2007.

8. B. B. Christy, “The use of the PDQ (Paint Data Query) Database along with other resources to provide vehicle information for hit and run fatalities within Virginia,” Trace Evidence Symposium, August 13-16, Clearwater Beach, FL, 2007.

9. F. Chau, Y. Liang, J. Gao, and X. Shao, Chemometrics – From Basics to Wavelet Transform, John Wiley & Sons, NY 2004.

10. J. S. Walker, Primer on Wavelets and Their Scientific Applications, Chapman & Hall/CRC, Boca Raton, FL 1999.

11. J. Karasinski, S. Andreescu, O. A Sadik, B. Lavine, and M. N. Vora, “Multi-array Sensors with Pattern Recognition for the Detection, Classification and Differentiation of Bacteria at Subspecies and Strain Levels,” **Anal. Chem**., 2005, 77(24), 7941-7949.

12. B. K. Lavine, C. E. Davidson, and W. T. Rayens, “Machine Learning Based Pattern Recognition Applied to Microarray Data,” **Combinatorial Chemistry & High Throughput Screening,** 2004, 7, 115-131.

13. Barry K. Lavine and Nikhil Mirjankar, “Wavelets and Genetic Algorithms Applied to Spectral Pattern Recognition in Forensics,” Federation of Analytical Chemistry & Spectroscopy Societies,” Memphis, TN, October 16, 2007

14. Barry K. Lavine, Nikhil Mirjankar, Scott Ryland, and Mark Sandercock, “Wavelets and Genetic Algorithms Applied to Search Prefilters for Spectral Library Matching in Forensics,” **Talanta,** 2011, 87, 46-52.

15. J. C. W. Bink and H. A. Van’t Klooster, “Classification of Organic Compounds by Infrared Spectroscopy with Pattern Recognition and Information Theory,” **Anal. Chim. Acta**, 1983, 150, 53-59.

16. 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**., 1985, 25, 235-241.

17. “Strengthening Forensic Science in the United States: A Path Forward,” National Research Council of the National Academies Press, Washington, DC, February 2009.

18. B. K. Lavine, A. J. Moores, and L. K. Helfend, "A Genetic Algorithm for Pattern Recognition Analysis of Pyrolysis Gas Chromatographic Data," **J. Anal. Appl. Pyrolysis**, 1999, 50, 47-62

19. B. K. Lavine, A. J. Moores, H. T. Mayfield, and A. Faruque, “Genetic Algorithms Applied to Pattern Recognition Analysis of High Speed Gas Chromatograms of Aviation Turbine Fuels Using an Integrated Jet-A/JP-8 Data Base,” **Microchemical Journal**, 1999, 61, 69-78

20. B. K. Lavine, J. Ritter, A. J. Moores, M. Wilson, A. Faruque, and H. T. Mayfield, “Source Identification of Underground Fuel Spills by Solid Phase Micro-extraction/High-Resolution Gas Chromatography/Genetic Algorithms,” **Anal. Chem**., 2000, 72(2), 423-431

21. B. K. Lavine, D. Brzozowski, J. Ritter, A. J. Moores[*1], and H. T. Mayfield, “Fuel Spill Identification by Selective Fractionation Prior to Gas Chromatography I. Water Soluble Components,” **J. Chromat. Sci.,** 2001, 39(12), 501-506

22. B. K. Lavine, D. Brzozowski, A .J. Moores, C. E. Davidson, and H.T. Mayfield, “Genetic Algorithm for Fuel Spill Identification,” **Anal. Chim. Acta**, 2001, 437(2), 233-246

23. B. K. Lavine, C. E. Davidson, A. J. Moores, and P. R. Griffiths, “Raman Spectroscopy and Genetic Algorithms for the Classification of Wood Types,” **Applied Spectroscopy**, 2001, 55(8), 960-966.

24. B. K. Lavine, A. Vesanen, D. M. Brzozowski, and H. T. Mayfield, “Authentication of Fuel Standards using Gas Chromatography/Pattern Recognition Techniques,” **Anal Letters**, 2001, 34(2), 281- 294

25. B. K. Lavine, C. E. Davidson, and A. J. Moores, “Innovative Genetic Algorithms for Chemoinformatics, “**Chemometrics & Intelligent Laboratory Instrumentation**, 2002, 60(1), 161-171.

26. B. K. Lavine, C. E. Davidson, and A. J. Moores, “Genetic Algorithms for Spectral Pattern Recognition,” **Vibrational Spectroscopy**, 2002, 28(1), 83-95.

27. 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,** 2003, 43, 1890-1905.

28. B. K. Lavine, C. E. Davidson, and D. J. Westover, "Spectral Pattern Recognition Using Self Organizing Maps," **J. Chem. Inf. Comp. Science,** 2004, 44(3), 1056-1064

29. G. A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F. K. Tadjimukhamedov, Barry K. Lavine, and Nikhil Mirjankar, "Pattern Recognition Analysis of Differential Mobility Spectra with Classification by Chemical Family," **Anal. Chim. Acta**, 2006, 579(1), 1-10

30. J. Karasinski, L. White, Y. Zhang, E. Wang, S. Andreescu, O. A Sadik, B. Lavine, and M. N. Vora, "Detection and Identification of Bacteria Using Antibiotic Susceptibility and a Multi-array Electrochemical Sensor with Pattern Recognition," **Biosensors & Bioelectronics,** 2007, 22(11), 2643-2649

31. Y.-D. Wang, D. J. Veltkamp, B. R. Kowalski, "Multivariate Instrument Standardization," **Anal. Chem**., 1991, 63, 2750-2756.

32. S. Wold, H. Antti, F. Lindgren, and J. Ohman, "Orthogonal Signal Correction of Near Infrared Spectra," **Chemom. Intell. Lab. Syst**., 1998, 44, 175-185.

33. T. B. Blank, S. T. Sum, and S. D. Brown, "Transfer of Near-Infrared Multivariate Calibrations without Standards," **Anal. Chem**., 1996, 68, 2987-2995.

34. A. J. Myles, T. A. Zimmerman, and S. D. Brown, "Transfer of Multivariate Classification Models between Laboratory and Process Near-Infrared Spectrometers for the Discrimination of Green Arabica and Robusta Coffee Beans," **Appl. Spec**., 2006, 60, 1198-1203.

35. L. A. Powell and G. M. Hieftje, Computer Identification of Infrared Spectra by Correlation-Based File Searching," **Anal. Chim. Acta**, 1978, 100, 313-327

36. P. G. Rodgers, R. Cameron, N. S. Cartwright, W. H. Clark, J. S. Deak, and E. W. W. Norman, "The Classification of Automotive Paint by Diamond Windows Infrared Spectrophotometry-Part I: Automotive Topcoats and Undercoats," **Can. Soc. Forensic. Sci. J**., 1976, 9(1), 1-14.

37. P. G. Rodgers, R. Cameron, N. S. Cartwright, W. H. Clark, J. S. Deak, and E. W. W. Norman, "The Classification of Automotive Paint by Diamond Windows Infrared Spectrophotometry-Part II: Automotive Topcoats and Undercoats," **Can. Soc. Forensic. Sci. J.,** 1976, 9(2), 49-68.

38. P. G. Rodgers, R. Cameron, N. S. Cartwright, W. H. Clark, J. S. Deak, J. S., and E. W. W. Norman, "The Classification of Automotive Paint by Diamond Windows Infrared Spectrophotometry-Part III: Automotive Topcoats and Undercoats," **Can. Soc. Forensic. Sci. J.,** 1976, 9(3), 103-111.

39. Foo-tim Chan, Yi-zeng Liang, Jumbin Gao, and Xue-guang Shao, Chemometrics – From Basics to Wavelets, Wiley Interscience, Volume 164, New York, 2004

40. M. Sarker, W. Graham Glen, L. Ym, and W. J. Dunn III, “Comparison of SIMCA Pattern Recognition and Library Search Identification of Hazardous Compounds from Mass Spectra,” **Anal. Chim. Acta**, 1992, 257, 229-238.

41. W. J. Dunn III, S. L. Emery, and W. G. Glen, “Preprocessing Variable Selection, and Classification Rules in the Application of SIMCA Pattern Recognition to Mass Spectral Data,” **Environ. Sci. Technol**., 1989, 23, 1499-1505.

42. W. J. Dunn III, M. G. Koehler, S. L. Emery, and D. R. Scott, “Application of Pattern Recognition to Mass Spectral Data of Toxic Organic Compounds in Ambient Air,” **Chem. Intell. Lab. Systems**, 1987, 1, 321-334.

43. J. M. Brenchley, U. Horchner, and J. H. Kalivas, “Wavelength Selection Characterization for NIR Spectra,” **Applied Spectroscopy**, 1997, 51(5), 689-699.

44. Steven D. Brown and Wangdong Ni, “Wavelet OSC and Stacked Methods for Classification, “FACSS, Louisville, KY, October 21, 2009.

45. L. Brieman, “Stacked Regression,” **Machine Learning**, 1996, 24, 49-64.

46. A. J. Myles and S. D. Brown, “Decision Pathway Modeling,” **J. Chemometrics**, 2004, 18, 286-293.

47. W. Ni., S. D. Brown, and R. Man, “Data Fusion in Multivariate Calibration Transfer,” **Anal. Chem**., 2010, 661, 133-142.

48. John Chalmers (Ed.), Spectroscopy in Process Analysis, Sheffield Academic Press, Mansion House, 19 Kingfield Road, Sheffield S11 9AS, England, p. 351.

49. C. P. Wang and T. L. Isenhour, Infrared Library Search on Principal Component-Analyzed Fourier Transformed Absorption Spectra,” **App. Spec**., 1987, 41(2), 185-194.

50. R. J. Anderegg, and D. J. Pyo, “Selective Reduction of Infrared Data,” **Anal. Chem.,** 1987, 59, 1914-1917.

51. G. W. Small, “Automated Spectral Interpretation,” **Anal. Chem.,** 1987, 59(7), 535A-546A.

52. L. Domokos, I. Frank, G. Matolcsy, and G. Jalsovszky, Pattern Recognition Applied to Vapor Phase Infrared Spectra,” **Anal. Chim. Acta**, 1983, 154, 181-189.

## VI. Dissemination of Research Findings

The development of a prototype IR library search system to identify the manufacturing plant, model, and line of an automotive vehicle from a clear coat paint smear, and the demonstration of its 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 with clear coats 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 and the prototype searching system have been made at several scientific meetings, e.g., American Academy of Forensic Sciences, Federation of Analytical Chemistry and Spectroscopy Societies, and the American Chemical Society. These presentations have generated interest among analytical chemists and forensic scientists as the pattern recognition and library search techniques developed for the proposed library search system 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.