VARIABLE SELECTION TO IMPROVE

CLASSIFICATION IN STRUCTURE-ACTIVITY

STUDIES AND SPECTROSCOPIC ANALYSIS

By

COLLIN GARETH WHITE

Bachelor of Science in Chemistry
Oklahoma State University
Stillwater, Oklahoma
2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2016

VARIABLE SELECTION TO IMPROVE

CLASSIFICATION IN STRUCTURE-ACTIVITY

STUDIES AND SPECTROSCOPIC ANALYSIS

Dissertation Approved:

Barry K. Lavine

Dissertation Adviser

Ziad El Rassi

John Gelder

Nicholas F. Materer

R. Russell Rhinehart

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Barry K. Lavine, for his support, guidance and patience. I would also like to thank my graduate advisory committee: Dr. Ziad El Rassi, Dr. Nicholas Materer, Dr. John Gelder, and Dr. R. Russell Rhinehart. I am also thankful to present and former members of the Lavine research group: Nuwan Perera, Tao Ding, Kaushalya S. Dahal, Francis Kwofie, Nikhil S. Mirjankar, Sandhya R. Pampati, and Razvan I. Stoian. Special thanks go to Dr. Ayuba Fasasi and Matthew D. Allen for their helpful discussions and guidance. Additionally, I would like to thank Dr. Douglas R. Heisterkamp from the computer science department and former engineering student Solomon Gebreyohannes for their assistance.

Finally, I wish to thank my parents, Karl A. White and Barbara L. White, and my brother, Ethan M. White for their support and encouragement.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: COLLIN GARETH WHITE

Date of Degree: MAY 2016

Title of Study: VARIABLE SELECTION TO IMPROVE CLASSIFICATION IN STRUCTURE-ACTIVITY STUDIES AND SPECTROSCOPIC ANALYSIS

Major Field: CHEMISTRY

Abstract: A genetic algorithm for variable selection to improve classifications is explored and validated on a wide range of data. In one study, 147 tetralin and indan musks and nonmusks compiled from the literature for the purpose of investigating the relationship between molecular structure and musk odor quality were correctly classified by 45 molecular descriptors identified by the pattern recognition GA which revealed an asymmetric data structure. A 3-layer feed-forward neural network trained by back propagation was used to develop a discriminant that correctly classified all of the compounds in the training set as musk and nonmusk. The neural network was successfully validated using an external prediction set of 37 compounds. In another study, 172 tetralin-, indan- and isochroman-like compounds were combed from the published literature to investigate the relationship between chemical structure and musk odor quality. The 20 molecular structural descriptors selected by the pattern recognition GA yielded a discriminant that was successfully validated using an external validation set consisting of 19 compounds. In a third study, the development of a prototype pattern recognition library search system for the infrared spectral libraries of the paint data query database to improve the discrimination capability and permit quantification of discriminant power for automotive paint comparisons involving the original equipment manufacturer is described. The system consists of two separate but interrelated components: search prefilters to cull the library spectra to a specific assembly plant and a cross correlation library search algorithm that utilizes both forward and backward searching to identify the year, line and model of the unknown in the spectral set identified by the search prefilters. The genetic algorithm was able to identify spectral variables from the clear coat, surfacer-primer and e-coat layers of the original manufacturer's automotive paint that were characteristic of the assembly plant of the vehicle.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Humans have an ability to recognize and classify patterns. However, they are poor at processing and are also error prone (and slow) at performing computations involving repetitive tasks. This limits the amount of information that can be gained as a result of using the innate pattern recognition skills possessed by humans. For this reason, computers have been developed. Computers are viewed (in the context of this dissertation) as a tool to aid in the performance of human pattern recognition. However, data analysis should be viewed as an interface between man and computers, and this dissertation relates the connection between the two, although rarely has this concept been explicitly addressed.

The first step in presenting this work is the introduction of pattern recognition and its implementation using a genetic algorithm which is the topic of Chapter 2. Proposed as a method for variable selection, the strengths of the genetic algorithm for pattern recognition match the difficulties inherent in feature selection. Developed using a visual framework, this method allows the user to play an interactive role in the data analysis and the interpretation of the results.

There are also a number of advantages associated with using a pattern recognition methodology based on feature selection. For techniques such as discriminant analysis, variable selection is important as signal is averaged with noise over a large number of variables with a loss of discernible signal amplitude when noisy features are present in the data. With neural networks, the presence of irrelevant measurement variables can cause the network to focus its attention on the idiosyncrasies (i.e., noise) of individual training set samples due to the net's ability to approximate a variety of complex functions in higher dimensional space, thereby causing it to lose sight of the broader picture, which is essential for generalizing any relationship beyond the training set. Variable selection is also necessary because of the sheer enormity of many classification problems. Variable selection improves the reliability of a classifier because noisy variables increase the chances of false classification and decrease classification success-rates on new data. It is important to identify and delete features from the data set that contain information about experimental artifacts or other systematic variations in the data not related to legitimate chemical differences between the classes represented in the study. Variable selection can also lead to an understanding of the essential features that play an important role in governing the behavior of the process under investigation. Measurements that are informative and those measurements that are uninformative can be identified. For all of these reasons, variable selection should be the principal focus in any pattern recognition study.

Chapter 3 describes the development of a prototype pattern recognition library search system for the infrared spectral libraries of the paint data query database to improve the discrimination capability and permit quantification of discrimination power for automotive paint comparisons involving original equipment material. 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 library 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 truncated 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, surfacer-primer and e-coats by decomposing each spectrum into wavelet coefficients which represent the sample's constituent frequency, the 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 automotive paint sample was obtained. Using the pattern recognition GA to identify wavelet coefficients characteristic of the assembly plant from which the paint sample was obtained, search prefilters were developed to facilitate searching of IR spectra from the PDQ data base. Even in challenging situations where the samples evaluated were all the same make (General Motors, Chrysler, or Ford) within a limited production year range (2000-2006), the respective assembly plant could be correctly identified.

The best match between each unknown and library spectra in the hit list generated by the search prefilters was identified using a cross-correlation library search algorithm that performed both forward and backward searching. In the forward search, spectra were divided into intervals. The top five hits identified in each search window were compiled and a histogram was computed that summarized the frequency of occurrence for each library sample. IR spectra most similar to the unknown were flagged. The backward search computed the frequency and occurrence of each line and model without regard to the identity of the individual spectra. Only those lines and models with a frequency of occurrence greater than or equal to 20% were included in the final hit list. If there was agreement between the forward and backward search results, the specific line and model that was common to both hit lists was always the correct assignment. Samples assigned to the same line and model by both searches were always well represented in the library and correlate well on an individual basis to specific library samples. For these samples, one can have confidence in the accuracy of the match.

Chapter 4 presents the results from two structure activity relationship studies involving polycyclic musks. In one study, 147 tetralin and indan-like compounds were compiled from the literature for the purpose of investigating the relationship between molecular structure and musk odor quality. Each compound was represented by 1344 molecular descriptors. A genetic algorithm for pattern recognition analysis was used to identify a subset of these molecular descriptors that could differentiate musks from nonmusks in a plot of the two largest principal components of the data. A principal component plot of the 110 compounds in the training set using 45 molecular descriptors identified by the pattern recognition GA revealed an asymmetric data structure. Tetralin and indan musks were found to occupy a small, but well defined region of the principal component (descriptor) space, with the nonmusks randomly distributed in the principal component plot. A 3-layer feed-forward neural network trained by back propagation was used to develop a discriminant that correctly classified all of the compounds in the training set as musk or nonmusk. The neural network was successfully validated using an external prediction of 37 compounds.

In another study, 191 tetralin-, indan-, and isochroman-like compounds were combed from the published literature to investigate the relationship between chemical structure and musk odor quality. Each compound in this database was represented by 1368 molecular descriptors. A genetic algorithm for pattern recognition analysis was used to identify a subset of the 1368 molecular descriptors that could differentiate musks from nonmusks in a plot of the two largest principal components of the data. The 20 molecular structural descriptors selected by the pattern recognition GA contained information about the shape, electronic properties, and intermolecular interactions of these compounds. Due to the risk of model interpretation when performing variable selection, model cross validation was performed using an external validation set of 19 compounds. Discriminants (in the form of a principal component plot of the 20 molecular descriptors and the 172 compounds comprising the training set) were successfully validated using these 19 compounds.

Chapter 5 presents concluding remarks and identifies areas of future research. The potential impact of the prototype pattern recognition library search system on forensic automotive paint analysis is discussed. The methodology used for structure-activity correlations described in Chapter 4 offer prospects for the intelligent design of musk odorants.

CHAPTER II

METHODOLOGY

2.1. INTRODUCTION

Often, relationships in chemical data cannot be expressed in quantitative terms. These relationships are better expressed in terms of similarity and dissimilarity among diverse groups of data. The task confronting a scientist when investigating these types of relationships is two-fold. First, can the data be divided into distinct groups for the prediction of some property? Second, can the features necessary for differentiating these groups be identified? Pattern recognition [2-1] is a name given to a collection of methods well suited for tackling both of these tasks since these techniques can display variability between large numbers of samples and show the major clustering trends present in a large data set.

Pattern recognition refers to a set of methods originally developed to solve the class membership problem. In a pattern recognition study, samples are classified according to a specific property using measurements indirectly related to that property. An empirical relationship or classification rule is developed from a set of samples for which the property of interest and the measurements are known. The classification rule is then used to predict this property in samples that are not part of the original training set. The property in question may be the odor of a molecule and the measurements may be molecular descriptors that convey information about the shape, size, and electronic properties of the molecules comprising the dataset.

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

The basic premise underlying the pattern recognition methodology used in the studies described in this dissertation is that all classification methods work well when the problem is simple. By identifying the appropriate features, a "hard" problem can be reduced to a "simple" problem. Thus, the goal is feature selection, in order to increase the signal to noise of the data by discarding measurements that are not characteristic of the source profile of the classes in the dataset. To ensure identification of all relevant features, it is best that a multivariate approach to feature selection is employed. The approach should also take into account the existence of redundancies in the data.

In this chapter, a genetic algorithm (GA) for pattern recognition analysis and feature selection of multivariate chemical data is described. The pattern recognition GA identifies a set of features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Since principal components maximize variance, the bulk of the information encoded by the selected features is about differences between the classes in the data set. The principal component plot used by the fitness function acts as an embedded information filter. Sets of features are selected based on their principal component plots, with a good principal component plot generated by features whose variance or information is primarily about differences between the classes. This limits the search to these types of feature subsets, thereby significantly reducing the size of the search space. In addition, the GA can focus on those classes and/or samples that are difficult to classify by boosting their weights over successive generations using a perceptron to learn class and sample weights. Samples that consistently classify correctly are not as heavily weighted in the analysis as samples that are difficult to classify. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection.

2.2. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis [2-2] is the oldest and best known of the techniques used in multivariate data analysis. The overall goal of principal component analysis is to reduce the dimensionality of multivariate data without a loss of significant information. In principal component analysis, the original measurement variables are transformed into a new set of variables called principal components. Each principal component is a linear combination of the original measurement variables. Often, only two or three principal components are necessary to explain all of the information present in multivariate data. By plotting the data in a coordinate system defined by the two or three largest principal components, it is possible to identify key relationships, that is, find similarities and differences among samples represented as chromatograms or spectra in a data set.

Dimensionality reduction is possible with principal component analysis because chemical datasets are often redundant. In other words, chemical datasets are measurement rich, but they are not information rich. This is best understood by considering a dataset consisting 17 samples characterized by two measurements, X_1 and X_2. Figure 2.1 shows a plot of this data in a 2-dimensional coordinate system defined by the variables X_1 and X_2 which serve as basis vectors for the measurement space of these samples. X_1 and X_2 are correlated since fixing the value of X_1 limits the range of values possible for X_2. The enclosed rectangle in Figure 2.1 would be fully populated if these two measurement variables are uncorrelated. Since information is often defined as the scatter of points in a measurement space, it is evident that correlations among measurement variables decrease the information content of this space. The data points are restricted to a small region of this vector space due to correlations among the measurement variables.

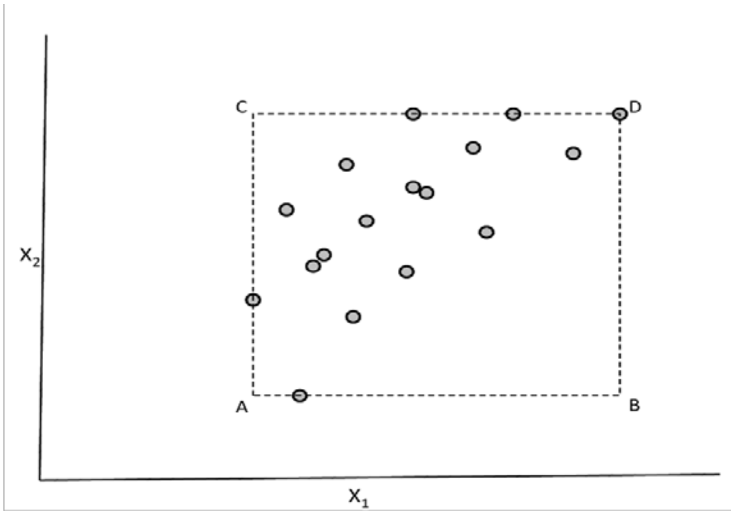


Figure 2.1. Seventeen hypothetical samples projected onto a 2-dimensional measurement space defined by the measurement variables X_1 and X_2. The vertices, A, B, C, and D, of the rectangle represent the smallest and largest values of X_1 and X_2. (Adapted from *NBS J. Res*., 1985, 190(6), 465-476)

Redundancy in data is due to collinearity (i.e., correlations) among the measurement variables. For measurement variables that are highly correlated, the data points will lie in a subspace. In Figure 2.2, X_3 is perfectly correlated with X_1 and X_2 since X_1 plus X_2 equals X_3. These six samples lie in a plane even though each data point has three measurements associated with it.

High collinearity between variables is a strong indication that a new set of basis vectors exist that are better at conveying the information present in the data than axes defined by the original measurement variables. This new basis set linked to the variance of the data can be used to develop a new coordinate system for displaying the data. The principal components of the data, which are a linear combination of the original measurement variables, define the variance-based axes of this new coordinate system.

$$X = \begin{pmatrix} 0.5 & 0.5 & 1.0 \\ 1.9 & 0.7 & 2.6 \\ 2.0 & 1.0 & 3.0 \\ 0.3 & 1.8 & 2.1 \\ 1.9 & 1.7 & 3.6 \\ 1.2 & 0.2 & 1.4 \end{pmatrix}$$

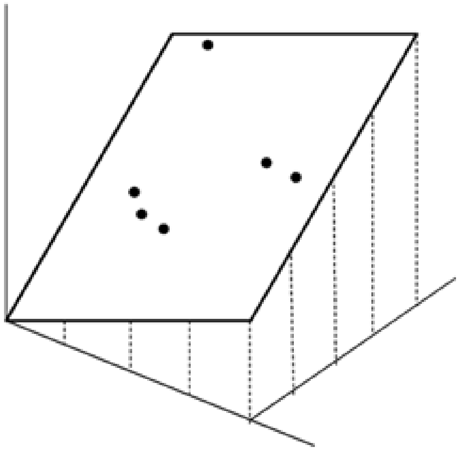


Figure 2.2. Six hypothetical samples projected onto a 3-dimensional measurement space. Because of strong correlations among the 3 measurement variables, the data points reside in a 2-dimensional subspace of the original measurement space. (Adapted from *Multivariate Pattern Recognition in Chemometrics*, Elsevier Science Publishers, Amsterdam, 1992)

Variables that are highly correlated will have a great deal of redundancy and are said to be collinear. High collinearity between a set of measurements variables is a strong indication that a new set of basis vectors can be found that are better at conveying the information content present in data than axes defined by the original measurement variables. This new basis set linked to variation in the data can be used to develop a new coordinate system for displaying the data. The principal components of the data define the variance-based axes of this new coordinate system.

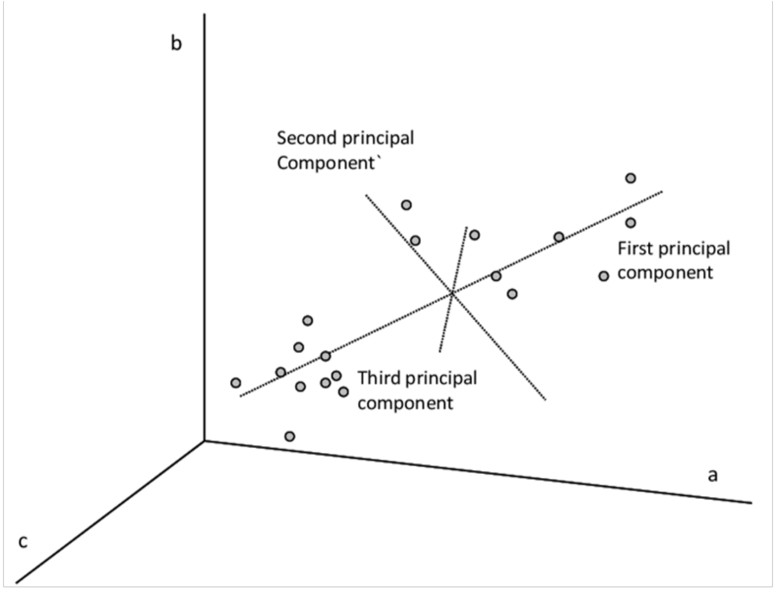


Figure 2.3. Principal component axes developed from the measurement variables a, b, and c. (Courtesy of Applied Spectroscopy, 1995, 49(12), 14A-30A)

The largest principal component is formed by determining the direction of largest variation or scatter in the original measurement space and fitting it with a line using linear least squares that runs through the center of the data (see Figure 2.3). The second largest principal component, which also passes through the center of the data and is orthogonal to the first principal component, lies in the direction of next largest variation. The third largest principal component lies in the direction of the next largest variation, passes through the center of the data and is orthogonal to the first and second principal components, and so forth. Each principal component describes a different source of information because each describes a different direction of variation or scatter in the data. The scatter of the data points in the measurement space which is linked to information is also a direct measure of the data's variance. (In other words, variance is synonymous with information.)

Furthermore, the orthogonality constraint imposed by the mathematics of principal component analysis ensures that each variance-based axis is independent.

The information content of each principal component can be assessed by measuring the percentage of variance explained by it, which is calculated from its eigenvalue. The first principal component, which has the largest eigenvalue, is more informative than the second principal component, which has the second largest eigenvalue and so forth. The percentage of variance explained is often calculated on a cumulative basis to assess the number of principal components required to explain a certain fraction of the variance. The number of principal components computed is the smaller of the number of samples or measurements that comprise the dataset. If the training set consists of 350 samples, with 300 features per sample, the number of principal components that can be computed cannot exceed 300 and could even be less because of collinearities among the measurement variables.

2.3. GENETIC ALGORITHM FOR PATTERN RECOGNITION ANALYSIS

Genetic algorithms were developed by John Holland [2-3]. They simulate the process of evolution to solve optimization problems. A search of the solution space is performed by utilizing knowledge contained in a population of solutions while simultaneously employing operators simulating reproduction and mutation to generate new and potentially better solutions. Each generation will be at least as good, and often better than the previous generation in terms of its most fit member (i.e., the feature subset within the generation yielding the highest classification score) and the average fitness score of the population for that generation.

There are several advantages for using genetic algorithms to solving ill-defined optimization problems such as the variable selection problem in pattern recognition. While conventional optimization techniques will manipulate parameters independently, genetic algorithms operate on the entire parameter set and simultaneously consider many points in the solution space. Genetic algorithms require only information about the fitness of the potential solutions, i.e., individuals feature sets. They make no assumptions about the topography of the solution surface, and are not disturbed by discontinuities or singularities that would be disruptive to derivative and simplex optimization based methods. By adjusting the parameters of the genetic algorithm, it can be tailored to individual classifications problems such as an asymmetric classification which can be accommodated by changing the parameters in the pattern recognition GA assigned to the non-structured class.

The initial population of solutions is usually generated randomly, but the option of seeding the initial population with known solutions is available. Once the initial population is generated, the genetic algorithm proceeds through a well-defined sequence of steps which occurs at each generation to form the new population of solutions. This sequence of steps is described in Figure 2.4.

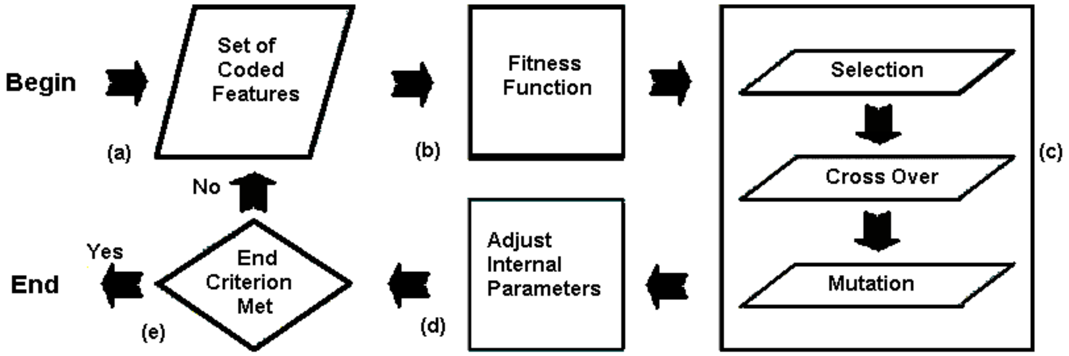


Figure 2.4. Block diagram of the pattern recognition GA.

The fitness function evaluates each variable subset in the population and assign a score to each solution. These scores, which summarize the quality of the proposed solution, are used to select the chromosomes (i.e., variable subsets) for reproduction. The process of reproduction is implemented using three operators: selection, recombination, and mutation. In selection, the fitness is used to select variable subsets for recombination. Solutions with a high fitness have a higher probability of being selected. These solutions then undergo a structured yet randomized exchange of information using an operator called crossover with the expectation that good solutions will generate even better ones (i.e., recombination). In the studies described in this dissertation, two point crossover was used. The data vector corresponding to each potential solution for a pair of solutions is broken up at two different points with the fragments swapped to yield a new solution. Additional randomness or variability is achieved by the mutation operator, which flips the state of single bits based on certain probabilities. This allows the genetic algorithm to explore other regions of the solution space. If the genetic algorithm identifies a better solution in this region, then potential solutions from this point can invade the population, with the optimization continuing in a new direction. The boosting algorithm adjusts the genetic algorithm's internal parameters for the next iteration. The aforementioned procedure (fitness evaluation, reproduction, and adjustment of internal parameters) is repeated until a specified number of generations have been executed or a feasible solution is found.

The selection and crossover operators were implemented by ordering the population of 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 solution in the random population has a uniform chance of being selected. If a purely biased selection criterion were used to select solutions for crossover, only a small region of the search space would be searched. Within a few generations, the population would consist of only copies of the best solutions in the initial population.

2.3.1 PCKaNN Fitness Function. The original fitness function of the pattern recognition GA, known as PCKaNN, scores the principal component plots and thereby identifies a set of features that optimize the separation of the classes in a plot of the two or three largest principal components of the data [2-4 – 2-10]. Because principal component analysis is used to determine the information present in a given subset of features, it is precisely this variation in principal components (different coordinate system for each feature subset) that allows meaningful comparisons to be made between sets of features.

To track and score the principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see Equations 2.1 and 2.2) 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. Class weights sum to 100, and the sample weights for the samples comprising each class sum to a value equal to the class weight for the particular class.

$$CW(c) = 100\frac{CW(c)}{\sum_{i=1}^{c} CW(c)} \tag{2.1}$$

$$SW_c(s) = CW(c)\frac{SW_c(s)}{\sum_{s \in C} SW_c(s)} \tag{2.2}$$

Each principal component plot for each feature subset in the population is scored using the Euclidean distance. For a given data point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to the largest. A poll is taken of the point's K_c nearest neighbors. For the most rigorous classification of the data, K_c equals the number of samples in the class to which the sample point belongs. (K_c is usually assigned a different value for each class.) For each sample in the principal component plot, the number of K_c nearest neighbors with the same class label as the sample point in question, the so-called sample hit count, SHC(s), is computed ($0 \leq \text{SHC(s)} \leq K_c$). It is then a simple matter to score the principal component plot (see Equation 2.3). First, the contribution to the overall fitness by each sample in class 1 is computed, with the scores of the samples comprising the class summed to yield the contribution by this class to the overall fitness. This same calculation is repeated for the other with the scores from each class summed to yield the overall fitness, *F (d).*

$$F(d) = \sum_{c} \sum_{sec} \frac{1}{K_c} * SHC(s) * SW(s) \tag{2.3}$$

To understand the scoring of principal component plots by PCKaNN, consider a dataset consisting of two classes. One class has 50 samples and the other has 25 samples. Because class weights are required to sum to 100, the class weights at generation 0 are 50.

For uniformly distributed sample weights which is the case at generation 0, one class contains samples whose weights are assigned a value of 1 and the other has samples whose weights are assigned a value of 2. Suppose a sample in class 1 has, as its nearest neighbors, 40 samples from class 1 in a principal component plot representing a particular feature subset. For this sample, $\mathrm{SHC(s)/K_c}$ = 40/50 when K for the class is set at 50, and the contribution of the sample to the fitness function for the particular feature subset is 0.8 * 1 or 0.8. Multiplying $\mathrm{SHC/K_c}$ by SW(s) for each sample and summing up the corresponding product for the 75 samples in the dataset yields the score of the principal component plot and the value of the fitness function for this particular feature subset.

To steer the population towards an optimal solution, internal parameters (i.e., sample and class weights) are adjusted using a process known as boosting [2-11 – 2-17]. In order to boost, the following two parameters must be computed: the sample-hit rate (SHR), which is the average value of $\mathrm{SHC/K_c}$ over all feature subsets produced in a particular generation (see Equation 2.4) and the class-hit rate (CHR), which is the mean sample hit rate of all samples in the class (see Equation 2.5). ϕ in Equation 2.4 is the number of feature subsets in the population, and AVG in Equation 2.5 is the average or mean value.

$$\mathrm{SHR(s)} = \frac{1}{\phi}\sum_{\mathrm{i=1}}^{\phi}\frac{\mathrm{SHC_i}(s)}{\mathrm{K_c}} \qquad (2.4)$$

$$\mathrm{CHR_g(c)} = AVG(\mathrm{SHR_g(s)} : \forall_{\mathrm{s \in c}}) \qquad (2.5)$$

Using a perceptron (see Equations 2.6 and 2.7), sample and class weights are adjusted with the momentum term, 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 and samples with a lower sample hit rate are boosted more heavily than those classes and samples that score well.

The momentum of the perceptron is initially set to 0.5. When the average change in the class weights fall below a specified threshold value, the class weights will become fixed. The perceptron for the class weights will be deactivated and the sample weights for each class will become uniformly distributed according to the class weight. The perceptron for the sample weights remains operational with the momentum decreased to 0.25. This value has been chosen in part because it facilitates learning by the genetic algorithm but does not cause a particular sample to dominate the calculation, which would result in the other samples not contributing to the scoring by the fitness function.

$$CW_{g+1}(c) \ = \ CW_g(c) \ + \ P\left(1 - CHR_g(c)\right) \tag{2.6}$$

$$SW_{g+1}(s) \ = \ SW_g(s) \ + \ P\left(1 - SHR_g(s)\right) \tag{2.7}$$

Boosting is important for PCKaNN and the pattern recognition GA because it modifies the fitness landscape by adjusting the values of the class and sample weights using information from the population to guide these changes. This helps to minimize the problem of convergence to a local optimum and enable the GA to identify the global optimum. During each generation, class and sample weights are updated using the class and sample hit-rates from the previous generation. Evaluation, reproduction, and boosting of potential solutions are repeated until a specified number of generations are executed or a feasible solution is found.

2.3.2. Modifications of PCKaNN. Modifications to PCKaNN have been made to improve the performance of the pattern recognition GA and generalize the algorithm. Three different modifications and/or changes to PCKaNN have been undertaken: (1) incorporation of the Hopkins statistic and modified Hopkins statistic into PCKaNN for transductive learning, (2) incorporation of the root mean square error of calibration (RMSEC) into PCKaNN for multivariate calibration, and (3) variable selection based on pinch ratio clustering which projects the data onto a nonlinear subspace.

2.3.2.1 Transductive Learning. The Hopkins statistic [2-18] and the modified Hopkins statistic [2-19] search for features that increase sample clustering whereas PCKaNN identifies features that optimize class separation. The Hopkins statistic is given by Equation 2.8.

$$H = \frac{\sum U}{\sum U + \sum W} \tag{2.8}$$

For the fitness function of the genetic algorithm, U are the distances between randomly selected locations and their nearest neighbors in the principal component plot and W refers to the distances between randomly selected data points and their nearest neighbors in the same principal component plot. The number of random locations and data points is an adjustable parameter, which is typically set to 10% of the number of data points in the training set. Because the value of the Hopkins statistic varies from 0.5 (no clustering) to 1.0 (perfect clustering), it was found necessary to scale its value using a sigmoid transfer function. (For uniformly distributed random data, the sum of the U terms and the sum of the W terms will be equal, so H is 0.5. In the case of a feature subset yielding very tight and well separated clusters, the sum of the W terms will be very small compared to the sum of the U terms with the limiting value of H approaching unity.)

For datasets that contain more features than samples, it is well known that even multivariate normal distributions with no outliers will have subsets of variables that produce eigenvector projections containing points that appear as outliers in a principal component plot. The Hopkins statistic will generate high values for such projections and thereby be distracted from other types of data structures. For this reason, it was necessary to robustify the Hopkins statistic using an influence function for principal components.

The influence function for principal components is given in Equation 2.9, where t_{ij} is the influence of the *i*th sample on the *j*th principal component, y_{ij} is the score of the *i*th sample on the *j*th principal component, λ_j is the eigenvalue of the *j*th principal component, and *n* is the number of samples. Influence values are normalized across all samples to sum to 1 (Equation 2.10).

$$t_{ij} = \frac{y_{ij}^2}{(n-1)\lambda_j} \tag{2.9}$$

$$\sum_{i=1}^{n} t_{ij} = 1 \tag{2.10}$$

The influence function identifies observations with high leverage (i.e., outliers) and deweights their contribution to the Hopkins statistic. The robustification procedure for the Hopkins statistic is defined in Equation 2.11 where max(influence$_i$) is the influence value (which is a fraction) for the data point having the greatest influence on the eigenvalue of the ith principle component.

$$H_{adj} = H - H\sum_{i=1}^{PC} \max(influence_i) \tag{2.11}$$

The Hopkins statistic is then combined with PCKaNN through the use of a user specified weighting factor r, see Equation 2.12, where P is the score from PCKaNN, H_{adj} is the contribution to the overall fitness score from the Hopkins statistic, and F is the overall fitness score. The range of r is from 0 to 1. Setting r to 0 is the same as running PCKaNN, while setting r to 1 yields a fitness function based solely on the Hopkins statistic.

$$F = (1 - r)P + rH_{adj} \tag{2.12}$$

By simply tuning the relative contributions of the Hopkins statistic and PCKaNN, the structure of a data set can be explored. For example, new classes can be discovered by simply changing the relative contribution of the Hopkins statistic and the original fitness function to the overall fitness score. For training sets with small amounts of labeled data and large amounts of unlabeled data, this approach is guaranteed to perform better [2-20] than a learning model developed from a set of features using only the labeled data points for the training set since information in the unlabeled data is used by the fitness function to guide feature selection which also prevent overfitting. This approach to feature selection is semi-supervised learning as it incorporates aspects of both supervised and unsupervised learning to develop a new paradigm for multivariate data analysis where classification, clustering, variable selection, and prediction are combined in a single step enabling a more careful analysis of the data.

Unlabeled samples in the training set can be tracked using the boosting routines of the pattern recognition GA through modification of the Hopkins statistic (see Equations 2.13 - 2.15). The score of the modified Hopkins statistic, MH, is given by Equation 2.13. Sample j is an unlabeled sample whereas sample i is the training set sample closest to the unlabeled sample in the principal component plot, u is the number of unlabeled samples, USW is the weight of the unlabeled sample, and d is the distance between the unlabeled sample and the labeled sample closest to it in the principal component plot. Each unlabeled sample is initially assigned a sample weight of $100/u$. The distance between sample j and the labeled sample in the training set, which is its nearest neighbor, is computed for each feature subset in the entire population from which $avgD(j)$ for sample j is calculated using Equation 2.14, where φ is the number of potential solutions in the population. The weights for the unlabeled samples are adjusted using Equation 2.15 with $\mathrm{USW}_{\mathrm{g+1}}$(j) denoting the weight for sample j in the current generation g+1, $\mathrm{USW}_{\mathrm{g}}$ (j) denotes the weight for sample j in the previous generation g and P is the momentum term of the perceptron. Boosting further minimizes the probability of the genetic algorithm converging to a local optimum for a training set which includes unlabeled samples.

$$MH = \sum_{j=1}^{u} \frac{1}{1 + d_{ij}} USW_j \qquad (2.13)$$

$$avgD(j) = \frac{1}{\varphi} \sum_{i=1}^{\varphi} \frac{1}{1 + d_{ij}} \qquad (2.14)$$

$$USW_{g+1}(j) \; = \; USW_g(j) \; + \; P\left(1 - avgD_g(j)\right) \qquad (2.15)$$

2.3.2.2 Multivariate Calibration. The PCKaNN fitness function can be applied to the problem of variable selection in multivariate calibration. Variables that have a principal component plot where the position of each sample data point directly corresponds to its y-block value, are identified using a variation of Equation 2.3. Each principal component plot generated for each feature subset in the population is scored using Euclidean distances. The dimensionality of the principal component subspace scored by the fitness function is defined by the user. For a given data point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to largest. A poll is then taken of the sample's K-nearest neighbors where K is a user defined parameter with smaller values of K corresponding to more stringent scoring conditions. The fraction of samples within K corresponding to data points that should lie near this sample based on the value of the dependent variable is computed. This fraction designated as SHC(s)/K_S in Equation 2.3 is used to score each principal component plot.

The RMSEC, which is a measure of the difference between the actual (observed) y-block value and the value fitted (calculated) by the model for the N sample in the training set (see Equation 2.16), can be integrated into PCKaNN using an approach similar to one used for the Hopkins statistic or modified Hopkins statistic.

$$RMSEC = \sqrt{\sum_{i=1}^{N}\left(Y_{actual,i} - Y_{calculated,i}\right)^2} \qquad (2.16)$$

Prior to its integration, the RMSEC value for each feature subset in the population is scaled using Equation 2.17 where RMSEC$_{\text{allF}}$ is the root mean square error of calibration for the N samples in the training set using all spectral features, RMSEC$_{\text{subset}}$ is the root mean square error of calibration for the N samples in the training set using a subset of features identified by the genetic algorithm, and I is the scaled RMSEC value used by the fitness function. Integration of the scaled RMSEC value with PCKaNN is accomplished using Equation 2.18 where r is a user defined value to control the weighting between the PCKaNN and RMSEC scores.

$$I = 100 * \left(\frac{RMSEC_{allF} - RMSEC_{subset}}{RMSEC_{allF}}\right) \tag{2.17}$$

$$F = (1 - r)PCKaNN + rRMSEC \tag{2.18}$$

2.3.2.3. Pinch Ratio Clustering. PCKaNN works well for linearly separable datasets. For more complex data analysis problems, nonlinear discriminants such as Pinch Ratio Clustering (PRC) may be the preferred method. PRC [2-21] utilizes Topologically Intrinsic Lexicographic Ordering (TILO), based on knot theory [2-22]. In TILO, an ordered list of samples is given to the algorithm, along with a set of edge weights, each associated with a pair of samples. The edge weights are calculated using information about distances between samples in the vector space defined by the feature subset (i.e., potential solution) to be evaluated. This information can serve as the basis for penalties, with the infraction being that two samples defining the edge are in different PRC clusters. The objective of TILO is to reorder the list with the goal of moving sample pairs with high edge weights together, thereby providing PRC with an optimized sample order which can be divided into distinct clusters based on the class membership values of the samples. It does this by considering samples for cyclic shifting within the order.

For each position in the ordered list, boundary values are calculated as the sum of all edge weights that connect a sample either at the position in question or before it with a sample after that position. These boundary values are sorted into descending order to form an array known as the width of the current ordering. Widths are compared on the basis of the first element for which they are different (e.g., if width A and width B have their two highest values equal to each other, but width A's third highest value is less than that of width B, width A is less than width B). If a potential cyclic shift is found to result in an ordering with a smaller width, the cyclic shift is carried out. This continues until the TILO algorithm cannot find any additional swaps which reduce edge widths within the ordered list.

Once TILO finds a strongly irreducible ordered list, this list is returned to PRC, which will check the boundary array associated with this ordered list, seeking local minima. A local minimum in the boundary value array is an indication that a location associated with a local minimum can serve as a good split location to divide the ordered list into clusters. Information about candidate split locations is then passed to a priority queue, with priority determined by the pinch ratio, given by Q_i in equation 2.19, where B is the boundary value at the position indicated by its subscript, i is the position being evaluated, p is the earliest position of interest, and q is the latest position of interest. Splits are then performed, with each resulting cluster being reevaluated after each split, until the desired number of clusters has been obtained, or no further split locations can be found as only locations with pinch ratios less than 1 can become split locations.

$$Q_i = \frac{B_i}{\min\left(\max_{p \le j \le i}\left(B_j\right), \max_{i < k \le q}\left(B_k\right)\right)} \tag{2.19}$$

The pinch score is actually determined from two additional components. The first is the class membership score, calculated by determining if the members of the same cluster have the same class membership values. A penalty is assessed every time a sample is found to lie in the same cluster as a sample from a different class. The second is a score based on the pinch ratios of the splits. This pinch ratio score is given by S_b in Equation 2.20, in which N is the number of splits performed. (The splits are arranged in the order in which they were performed). In Equation 2.20, S_b is a weighted average (scaled to 100): each split's pinch ratio is given half the influence of the pinch ratio of the previous split. S_b can only reach 100 if all of the pinch ratios are zero, which indicates an optimal solution.

$$S_b = 100 * \left(\frac{\sum_{i=1}^{N} \left(\frac{1 - Q_{i-1}}{2^{i-1}} \right)}{\sum_{i=1}^{N} 2^{1-i}} \right) \tag{2.20}$$

The two scores are considered in tandem, with the class membership score serving as a sieve. If a perfect class membership score is achieved, the pinch ratio score is given full influence. When the class membership score is less than 100, the pinch ratio score will not be calculated. The class membership score is issued a penalty via division by 2^{N+1}, where N is the number of classes. For the simplest classification problem, that of a binary classification, the class membership score is divided by 8. This penalty ensures that only feature subsets achieving a perfect class membership score will have the highest score among potential solutions for each generation.

2.4. SOFTWARE DESIGN AND IMPLEMENTATION

The pattern recognition GA is written in Java. Originally, the system was run under Windows. Currently, it is run under LINUX. The conversion to LINUX involved the substitution of the parallel processing method from TCP to MPJ as TCP is incompatible with LINUX. This involved a significant overhaul of the original code, including the removal of the original daemon system as MPJ has a daemon system of its own.

The change from TCP to MPJ offered advantages other than allowing the new 40-thread server to run the pattern recognition GA. The TCP based code required the computer on which it was running to “self-connect”, using a specific IP address contained in an input file. Some computers with multiple internet ports had to have an internet line run between two of these port. Loss of an internet connection to a computer running the previous version of the GA would often times result in an inability to start new daemons because the IP address of the computer had changed. This would require the user to run the “ipconfig” command to look up the old IP address, and change the input file so that the address contained therein matched the computer's new address. The LINUX version with MPJ based parallel processing bypassed these issues. Additionally, LINUX is becoming increasingly popular for scientific and computational applications, so the added advantage of compatibility represents an important development for the Java GA.

2.4.1. Use of the Java GA. In order to use the Java GA, the raw data must first be preprocessed into the format required by the Java GA. The raw data must be arranged into a matrix with sample and class labels attached. The sample data vectors typically occupy rows, and the measurement variables (features) typically occupy columns. The sample and class labels are concatenated onto the left side of the data matrix, along with any other header information that the user wishes to include. When validation set samples or blind samples are used, the class label column must be duplicated, and in the copy, the validation set samples are labeled as "NaN" (not a number). Any mathematical transformations (e.g., vector normalization) that the user wishes to apply are performed off-line using MATLAB or other programs.

Two input files are required to start a project using the pattern recognition GA. The first is the aforementioned data file, which is given to the pattern recognition GA in the .dat file format. The second is a simple text file which tells the GA whether the samples are represented by row vectors or column vectors (this is controlled by an integer, typically set to 1 to indicate row vectors but with the value of 2 also being valid and indicating column vectors), how many header columns are present, and a label for each header column. This file is provided with the .dsc file extension.

When the Java GA is started, a graphical user interface (GUI) is launched which manages the processes of managing a new project, starting a new run within a project, and recalling a plot resulting from a previous run. In any of these cases, the first step is to load the required files, which is managed by the first two panels of the pattern recognition GA. If a new run is started, an additional three or four panels (depending on whether a calibration or classification will be performed) prompt the user for the required input parameters. On the last of these panels is the “Start Run” button which, when pressed, closes the GUI and initiates the execution of the run. When the run is complete, another GUI pane appears and displays the plot corresponding to the best solution in the final generation. The best solution from earlier generations can be readily accessed by the user in the run files created by the pattern recognition GA.

2.4.2. Algorithm Implementation. The Java GA represents chromosomes as arrays of positive integers, each integer corresponding to a specific feature (they are numbered in the same order that they appear in the data matrix). Input parameters are arranged into an object within the Java framework, with the appropriate parameter called when needed. To start the GA, the user must execute its main file called PR. This file contains the first GUI panel, and will then call another file for the loading of the input files, depending on whether a calibration or classification run is chosen. One pair of files handles the assignment of parameters for a new run, and another handles the execution of parallel calibration and parallel classification. In both cases, the fitness function options are implemented as methods within another pair of classes, which make calls to PCA objects common to both calibration and classification. Other tasks, such as plotting and boosting are handled by other objects in Java.

Parallel processing involves the use of daemons - individual threads tasked with a portion of the work. In the original version of the pattern recognition GA, a daemon registry had to be started, and then each daemon started separately before starting a run. In the LINUX version, a certain number of daemons is specified as a parameter in the command that starts the pattern recognition GA. The number of daemons to be started should never exceed the number of threads available in the processor of the computer (this number is 40 for the LINUX server machine used for the studies described in this dissertation). Up to 40 daemons, more daemons generally means more efficiency, and faster execution. Attempting to use too many daemons will result in loss of efficiency (two or more daemons will compete for the same computer resources) with slower execution as the outcome.

The daemons are assigned numbers as they are allocated for a task. Because this is a Java application, these numbers start at 0 instead of 1, and will run to N-1, where N is the number of daemons that have been allocated. The daemon whose number is zero is known as the root. The root daemon is responsible not only for its share of the work, but for issuing work to the other daemons, receiving the appropriate response messages from them when their tasks are completed, and once all daemons have replied and the root has finished its own share of the work, the root must perform further steps involved in the transition between generations. The result is a slightly poorer execution speed than what would be achieved were it possible to parallelize all steps, since there would be a short period of time at the beginning and end of each generation in which only the root is performing work.

REFERENCES

2-1. J. T. Tou, R. C. Gonzalez, Pattern Recognition Principles. Addison Wesley Publishing, Reading, MA 1974.

2-2. I. T. Jolliffe, Principal Component Analysis. Springer Verlab, NY 1986.

2-3. J. H. Holland, “Adaptation in Natural and Artificial Systems”, 6th edition, MIT Press, Cambridge, MA, 2001.

2-4. B. K. Lavine, C. E. Davidson, A.J. Moores, “Innovative genetic algorithms for chemometrics”, Chem. Int. Lab. Sys., 2002, 60, 161-171.

2-5. B. K. Lavine, D. Brzozowski, A. J. Moores, C. E. Davidson, H. T. Mayfield, “Genetic algorithm for fuel spill identification”, Anal. Chim. Acta, 2001, 437, 233-246.

2-6. B. K. Lavine, C. E. Davidson, A. J. Moores, P. R. Griffiths, “Raman spectroscopy and genetic algorithms for the classification of wood types”, Appl. Spectrosc., 2001, 55, 960–966.

2-7. B. K. Lavine, C. E. Davidson, A. J. Moores, “Innovative Genetic Algorithms for Chemoinformatics”, Chem. Int. Lab. Sys., 2002, 1, 161–171.

2-8. B. K. Lavine, C. E. Davidson, A. J. Moores, “Genetic Algorithms for Spectral Pattern Recognition”, Vib. Spectrosc., 2002, 28, 83–95.

2-9. B. K. Lavine, C. E. Davidson, R. K. Vander Meer, S. Lahav, V. Soroker, A. Hefetz, “Genetic Algorithms for Deciphering the Complex Chemosensory Code of Social Insects”, Chem. Int. Lab. Sys., 2003, 66, 51–62.

2-10. B. K. Lavine, C. E. Davidson, C. Breneman, W. Katt, “Electronic Van der Waals Surface Property Descriptors and Genetic Algorithms for Developing Structure-Activity Correlations in Olfactory Databases”, J. Chem. Inf. Sci., 2003, 43, 1890-1905.

2-11. B. K. Lavine, C. E. Davidson, C. Breneman, W. Katt, “Genetic algorithms for clustering and classification of olfactory stimulants”, In: J. Bajorath (Ed.), Chemoinformatics: Methods and Protocols. Humana Press, Totowa, NJ, 2004, 399-426.

2-12. Y. Freund, R. E. Schapire, “A decision theoretic-generalization of online learning and an application to boosting”, Journal of Computer and System Sciences, 1997, 55, 119-139.

2-13. Y. Freund, R. E. Schapire, “Experiments with a new boosting algorithm”, Proceedings of Machine Learning: Thirteenth International Conference, 1996, 148-156.

2-14. R. E. Schapire, “The boosting approach to machine learning: An overview”, In Nonlinear Estimation and Classification, Springer, 2003.

2-15. R. E. Schapire, M. Rochery, M. Rahim, N. Gupta, “Boosting with prior knowledge for call classification”, IEEE Transactions on Speech and Audio Processing, 2005, 13.

2-16. C. Rudin, I. Daubechies, R. E. Schapire, “On the dynamics of boosting”**,** In Advances in Neural Information Processing Systems 16, 2004.

2-17. A. J. Moores, “The learning genetic algorithm: a hybrid learning system”, Master’s thesis, Clarkson University, Potsdam, NY, 2000.

2-18. B. Hopkins, “A New Method of Determining the Type of Distribution of Plant Individuals,” Ann. Bot., 1954, 18, 213-216.

2-19. B. K. Lavine, K. Nuguru, and N. Mirjankar, “One Stop Shopping – Feature Selection, Classification, and Prediction n a Single Step,” J. Chemomet., 2011, 25, 116-129.

2-20. V. Vapnik, “Statistical Learning Theory”. John Wiley & Sons: New York, 1998.

2-21. D. Heisterkamp, J. Johnson, “Pinch Ratio Clustering from a Topologically Intrinsic Lexicographic Ordering”, 2013 SIAM International Conference on Data Mining (SDM13), Austin, Texas, 560-568, 2013.

2-22. D. Gabai, “Foilations and the Topology of 3-Manifolds. III,” J. Differ. Geom., 1987, 26(3), 479-536.

CHAPTER III

IMPROVING THE PDQ DATABASE TO ENHANCE INVESTIGATIVE LEAD INFORMATION FROM AUTOMOTIVE PAINTS

3.1. INTRODUCTION

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

Modern automotive paint systems [3-1] consist of four layers: a clear coat over a color coat which in turn is over two undercoats, which are the surfacer-primer and the e-coat. (White trucks often do not have a clear coat layer and so only have two primers and a color coat layer.) With the exception of the clear coat, each paint layer contains pigments and fillers (the colored component), and all layers contain binders (the matrix that holds the layer together). Automotive manufacturers tend to use unique combinations of fillers and binders in each layer of paint. It is this unique combination that allows forensic scientists to determine the manufacturer and line of a vehicle within a limited production year range from an automotive paint chip recovered at the crime scene.

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

Studies [3-4, 3-5] conducted over 40 years ago by the Royal Canadian Mounted Police (RCMP) showed that vehicles could be differentiated by comparing the color, layer sequence and chemical composition of each individual layer in a paint system. To make the comparisons possible, a comprehensive database was developed as well as the means to search and retrieve information from it. Today, the Paint Data Query (PDQ) database contains over 24,000 samples (street samples and factory panels), that corresponds to over 96,000 individual paint layers, representing the paint systems used on most domestic and foreign vehicles marketed in North America. PDQ is a database of the physical attributes (i.e., color), the chemical composition and the IR spectrum of each layer of the original manufacturer's paint system. The PDQ concept is to narrow the list of possible vehicles to a number of suspects, not to identify a single vehicle [3-6, 3-7]. If the original paint layers are present in a recovered (i.e. unknown) paint chip, PDQ can assist in identifying the specific manufacturer and the production year of the automotive vehicle from which it came. The comparison of the IR spectrum of each paint layer in a paint system (clear coat, surface-primer, and e-coat layers) to IR spectra in PDQ allows for the assembly plant at which the paint system was applied, and the production year within a limited range, to be identified. IR library searches based on the color coat layer are not performed because the paint chemistry may be color dependent and spectra from this layer are often of poor quality as the IR signal is obscured by the metal and the pearlescent effect flakes mixed in the layer.

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

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

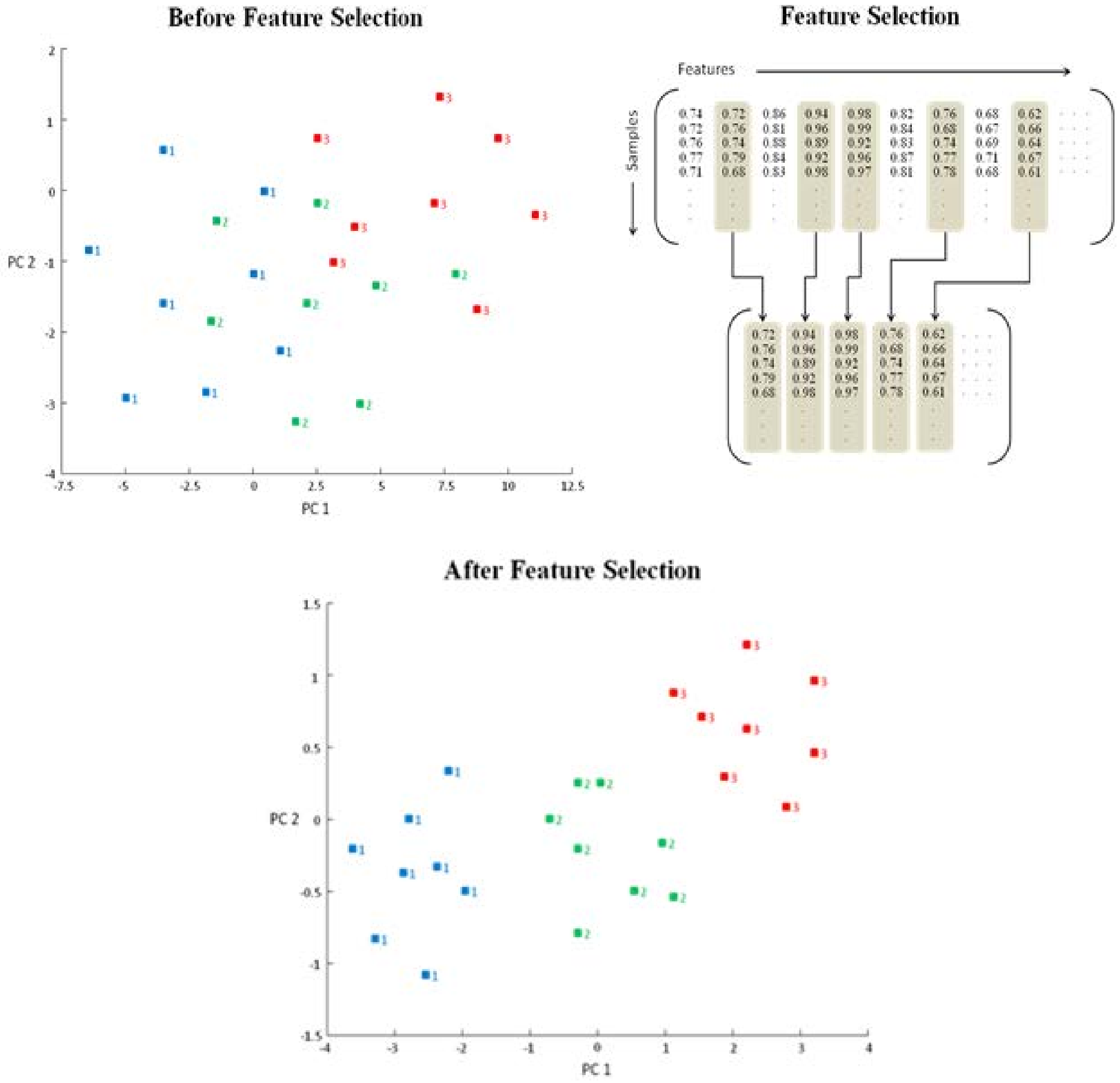
Another problem is that modern automotive paint systems have a thin color coat which on a microscopic fragment may be too thin to obtain accurate chemical and topcoat color information. The small size of the fragment will make it difficult to accurately compare it with manufacturer's paint color standards.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

3.2. METHODOLOGY

Pattern recognition assisted IR library searching techniques have been developed to search the spectral libraries of the PDQ database in an effort to differentiate between similar but non-identical FTIR paint spectra and to correctly identify an unknown paint sample as to the manufacturer and model of the vehicle within a limited production year range. Paint samples are often recovered from hit-and-run accidents where damage to vehicles or injury or death to a pedestrian has occurred. Searches of the PDQ database using commercial software have met with only limited success. Because the PDQ automotive paint library is composed of a large number of similar spectra, commercial search algorithms have not proven to be sufficiently sensitive at distinguishing subtle but significant features in the data such as shoulders, unique shapes, and patterns, and minor peaks. All commercial library search algorithms involve some type of point-by-point numerical comparison between the IR spectrum of an unknown and each member of the library [3-8]. These algorithms lack interpretive ability because they treat the spectrum as a set of points rather than as a collection of specific bands. Furthermore, band shifting is not handled well and bands of low intensity, which may be highly informative, are often ignored.

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

3.2.1. Search Prefilters. To develop the search prefilters, chemical information from FTIR spectra of the two primer layers and the clear coat layer were combined and then subsequently analyzed using a genetic algorithm (GA) for features selection and pattern recognition. Spectral features in each FTIR spectrum characteristic of the assembly plant (and hence the manufacturer and model) of the vehicle were identified by the pattern recognition GA [3-9 – 3-22], which utilized both supervised and unsupervised learning to identify features that optimize the separation of the FTIR spectra by assembly plant in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by the features selected by the pattern recognition GA were about differences between the different classes (i.e., assembly plants) in the database. A principal component plot that shows separation of the data by class can only be generated using features whose variance or information is primarily about differences between these classes. This fitness criterion dramatically reduces the size of the search space since it limits the search to these types of feature subsets. In addition, the pattern GA focused on those classes and/or samples that were difficult to classify as it trained by boosting the relative importance of classes and samples that consistently scored poorly. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The pattern recognition GA used in these studies integrated aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for wavelength selection and pattern classification. This idea is demonstrated in Figure 3.1 which shows a plot of the two largest principal components of a data set prior to feature selection.



0.74	0.72	0.86	0.94	0.98	0.82	0.76	0.68	0.62
0.72	0.76	0.81	0.96	0.99	0.84	0.68	0.67	0.66
0.76	0.74	0.88	0.89	0.92	0.83	0.74	0.69	0.64
0.77	0.79	0.84	0.92	0.96	0.87	0.77	0.71	0.67
0.71	0.68	0.83	0.98	0.97	0.81	0.78	0.68	0.61

0.72	0.94	0.98	0.76	0.62
0.76	0.96	0.99	0.68	0.66
0.74	0.89	0.92	0.74	0.64
0.79	0.92	0.96	0.77	0.67
0.68	0.98	0.97	0.78	0.61

Figure 3.1. A plot of the two largest principal components developed from the 10 wavelengths in the data set does not show class separation. When principal components are developed from the wavelengths that contain information about the automobile model, clustering on the basis of model is evident.

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

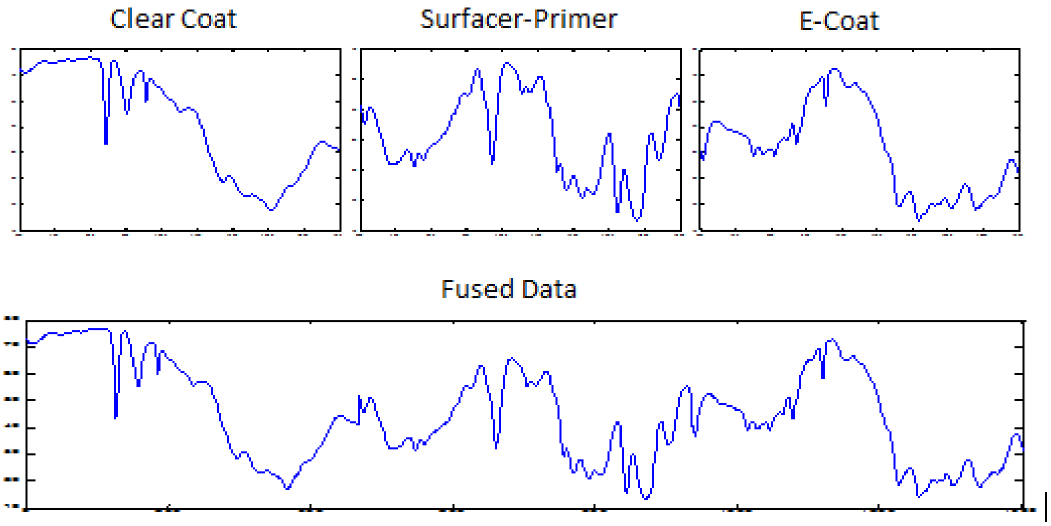


Figure 3.2. Clear coat, surfacer-primer, e-coat, and fused FT-IR spectrum

To develop the search prefilters, FTIR spectra of the fingerprint region (1640 cm^{-1} to 667 cm^{-1}) of the clear coat and the two primer layers for each paint sample were combined into a single data vector. Since each FTIR spectrum in the PDQ database was collected with 4 cm^{-1} resolution, this region was characterized by 506 points in each FTIR spectrum. To combine the chemical information obtained from the clear coat and two primer layers, the first 506 elements of the data vector representing the paint sample will be the corresponding fingerprint region of the clear coat layer, the next 506 elements of the vector will be the first primer layer and the final 506 elements of the vector will be the second primer layer (see Figure 3.2). The pattern recognition GA will then identify the components of this data vector (i.e., specific features in each paint layer) that are correlated to the assembly plant of the vehicle from which the paint sample was obtained.

To develop the search prefilters, all fused IR spectra were preprocessed using wavelets [3-23 – 3-25] to enhance subtle but significant features in the data and to remove noise. Wavelets offer a different approach to removal of noise from multivariate data. Using wavelets, a new set of basis vectors that take advantage of the local characteristics of the data are developed that were better at conveying the information present in the data than axes defined by the original measurement variables (absorbance value at each wavelength in the IR spectrum). The mother wavelet selected to develop this new basis set is the one that best matches the attributes of the data. This often circumvents the problem that occurs when an interfering source of variation in the data is correlated to information about the class membership of the samples, e.g., assembly plant, as a result of the design of the study or because of accidental correlations between signal and noise.

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

Using wavelets, each fused spectrum is passed through two scaling filters: a high pass filter and a low pass filter. The low-pass filter will allow only the low frequency component of the signal to be measured as a set of wavelet coefficients which is called the "approximation". The high-pass filter will measure the high frequency coefficient set which is called the "detail". The detail coefficients usually correspond to the noisy part of the data. This process of decomposition is continued with different scales of the wavelet filter pair in a step-by-step manner to separate the noisy components from the signal until the necessary level of signal decomposition has been achieved. Figure 3.3 shows the first level of wavelet decomposition applied to an IR spectrum of an undercoat paint layer displayed in the transmittance mode.

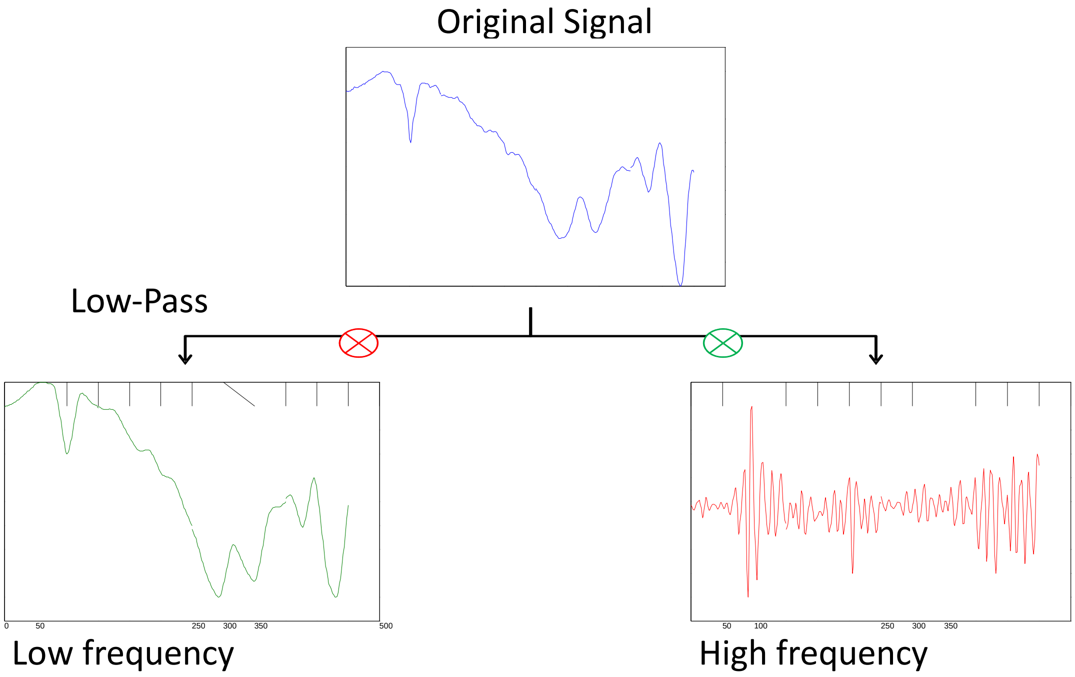


Figure 3.3. First level of wavelet decomposition applied to a clear coat paint spectrum displayed in the transmittance mode.

Wavelet coefficients from the fused IR spectra characteristic of the assembly plant of the vehicle were identified by the pattern recognition GA. The fitness function of the pattern recognition GA emulates human pattern recognition through machine learning to score the principal component plots and thereby identify a set of wavelet coefficients that optimize the separation of the automotive classes (i.e., assembly plants) in a plot of the two or three largest principal components of the data.

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

Search prefilters developed from fused spectra eliminated dissimilar spectra from the library search thereby providing the analyst with an opportunity to take advantage of more sophisticated but also more time-consuming search algorithms. Commercial infrared library search systems compare IR spectra by summing the squares of the difference between two spectra at every wave number. However, these algorithms do not perform well when differentiating between similar but non-identical spectra as small peak shifts are not handled well and bands of low intensity, which may be highly informative, are often ignored. For these reasons, a cross correlation function was used to provide the best match between an unknown and the spectra in the hit list generated by our search prefilters. The cross correlation function has been shown to correctly identify unknown spectra from similar but non-identical spectra [3-26]. Although it is slower than conventional search algorithms, it is suitable as a post searching method to rank probable matches, which have been selected by a faster algorithm (e.g., search prefilters). Correlation based searches are insensitive to instrumental noise and very sensitive to changes in peak shape and in the relative peak position making them sensitive to structural differences.

3.2.2. Cross-Correlation Library Search System. Library matching was performed by cross-correlating the unknown with each spectrum in the set of library spectra identified by the search prefilters and comparing each cross-correlated spectra with the corresponding autocorrelated library spectra. Cross-correlation, which is a measure of the similarity of two time varying functions, estimates the correlation between two signals using a dot product after a time lag has been applied to one of the signals. The cross-correlation function C_{ij} for the sampling interval Δt and relative displacement nΔt between signals s_i and s_j is estimated as shown in Equation 3.1.

$$C_{ij}(n\Delta t) \ = \frac{1}{T}\sum_{t=0}^{T} s_i(t) * \ s_j(t) \tag{3.1}$$

Autocorrelation is similar to cross-correlation, and is the signal being cross-correlated with itself. Each IR spectrum is normalized to unit length prior to the application of autocorrelation and cross correlation. Two versions of the cross correlation algorithm, each based upon transmittance spectra, were used in this study. The forward search divides the spectra are divides the spectra into three intervals: 3675 to 2856 cm^{-1} (interval after absorption by the diamond cell), 1891 to 668 cm^{-1} (fingerprint interval and carbonyl band), and 1650 to 668 cm^{-1} (fingerprint interval). The overlap between the second and third spectral intervals enforces the relative scale of the peaks and captures the broader trends in the spectral data as shown in our previous studies [3-27]. This effectively increases the importance of the fingerprint region in the spectral matching, and is superior to a disjoint set of intervals (e.g., 1650 cm^{-1} to 668 cm^{-1} with either single or double weighting, 1891 cm^{-1} to 1650 cm^{-1}, and 3675 cm^{-1} to 2856 cm^{-1}). Each region is normalized to unit length.

IR spectra in the truncated library most similar to the unknown are identified using three different modes of comparison: (1) autocorrelated spectrum of the unknown is compared to each cross-correlated unknown and library spectrum, (2) each autocorrelated library spectrum is compared to each cross-correlated unknown and library spectrum, and (3) the autocorrelated spectrum of the unknown is compared to each autocorrelated library spectrum.

Each comparison was made using a range of window sizes centered at the midpoint of the cross-correlated data interval (which corresponds to the cross correlation between two signals with zero lag) and increased in steps of 10 points or 100 points to include the entire cross correlated spectrum (see Figure 3.4). Because of the symmetry associated with cross correlation, the comparisons were made from only one side of the center burst. The Euclidian distance was used to evaluate the similarity index (see Equation 3.2) 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 3.2 was used instead of the hit quality index [3-28] used by commercial search algorithms, e.g., OMNIC, as it proved to be more informative for these spectra.

$$s_{ij} = 1 - \frac{d_{ij}}{d_{max}} \tag{3.2}$$

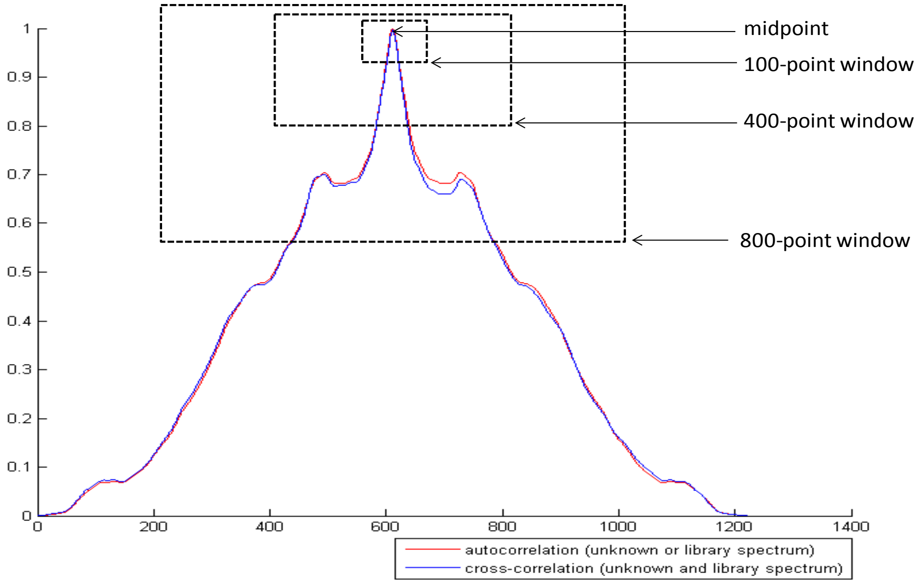


Figure 3.4. Comparison of the cross correlated unknown and PDQ library spectrum with the autocorrelated spectrum of the unknown or PDQ library spectrum.

At each window for each interval, the spectra were ranked in order of their similarity. Library spectra were arranged in descending order of similarity for each comparison and window size. The five most similar library spectra were then chosen from each comparison made for each window size, with sample identities preserved (sample identification number, make, line, and model). After every window was analyzed, a histogram depicting the frequency of occurrence for the most similar spectra was generated. The frequency of occurrence for each sample within this histogram was then weighted by its average similarity value across each comparison made at each window and interval. From the weighted histogram, the 5 library spectra with the highest weighted value were selected. Thus, the final result consists of the five library spectra with the highest frequency of occurrence after weighting.

The backward search utilizes autocorrelation and cross correlation to provide a probability index for the line and model of the unknown vehicle. For each window in each interval, the IR spectra in the library were ranked by their similarity (see Equation 2) with regard to the unknown, but only the label (i.e., the line and model of the automotive vehicle) of each of the top five spectra was preserved. After each window was processed, the number of hits for a specific line and model was computed and divided by the number of comparisons made by the algorithm. A set of percentages is generated that represents the likelihood of a particular line and/or model of being a match for the unknown. Only those lines and models with a frequency of occurrence equal to or greater than 20% are included in the hit list.

While the forward search identifies the library spectrum that is most similar to the unknown, the backward search provides insight into how well the library matches the unknown. For each unknown sample, the forward and backward searches were used in tandem to identify the corresponding vehicle information from the truncated PDQ library spectra generated by the search prefilters.

3.3. RESULTS AND DISCUSSION

Search prefilters were developed from 1179 automotive paint systems that spanned 3 automobile manufacturers (GM, Chrysler, and Ford) and 54 assembly plants (in North

America) within a limited production year range (2000-2006). Because of the large number of classes (i.e., assembly plants) involved, a hierarchical classification scheme was employed. A search prefilter was developed to differentiate paint samples by automobile manufacturer. For each automobile manufacturer, search prefilters were developed to identify the assembly plant of the vehicle. First, the assembly plants were divided into groups of plants based upon cluster analysis of the fingerprint region of the clear coat layer. Second, each plant group was divided into its constituent assembly plants using the clear coat, surfacer-primer, and e-coat layers. The search prefilters are intended to categorize each unknown manufacturer's paint system by identifying successively smaller sets of automotive vehicles to which an unknown is assigned. In the final step, library searching of an unknown is performed using IR spectra of vehicles manufactured in the assembly plant identified by the search prefilters. A block diagram of the vehicle classification process used in the prototype pattern recognition assisted library search system for the PDQ database is summarized in Figure 3.5.

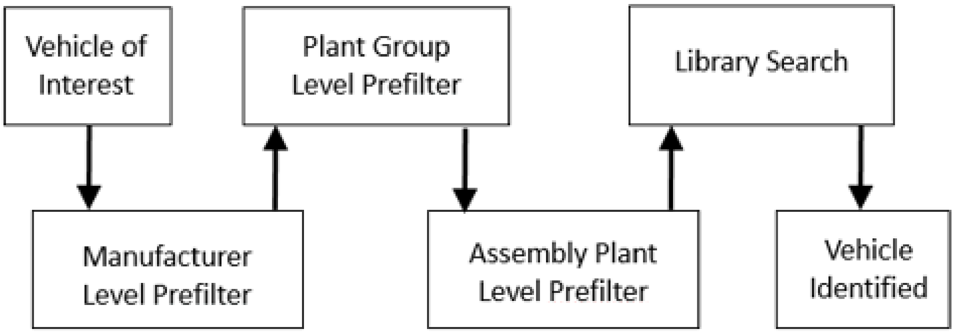

Vehicle of Interest
Manufacturer Level Prefilter
Plant Group Level Prefilter
Assembly Plant Level Prefilter
Library Search
Vehicle Identified

Figure 3.5. Block diagram of the vehicle classification process used in the prototype pattern recognition driven library search system for the PDQ database.

3.3.1. Manufacturer Search Prefilter. The initial focus of this study was to develop a search prefilter to classify IR spectra by manufacturer. To differentiate the automotive paint systems by manufacturer, the 1303 paint systems investigated were divided into a training set of 1179 samples and a validation set of 124 samples. The validation set samples were chosen by random lot. The training set of 1179 automotive paint systems was divided into 3 classes by automobile manufacturer (see Table 3.1).

Table 3.1. Training Set and Validation Set for Manufacturer Search Prefilter

Manufacturer	**Training Set**	**Validation Set**
GM (2000-2006)	429	44
Chrysler (2000-2006)	379	42
Ford (2000-2006)	371	38

All IR spectra were preprocessed for pattern recognition analysis using the scheme described in Figure 3.6. After retaining only the fingerprint region in each layer, the spectra were smoothed using a Savitzky-Golay filter (fourth order polynomial, 17 point window). The smoothed IR spectra were vector normalized and then wavelet transformed using the Symlet 6 mother wavelet at the 8th level of decomposition (8Sym6). All wavelet coefficients for the levels of decomposition less than the specified level were retained, such that the final result for each sample-paint layer combination is a row vector of wavelet coefficients, [A1 D1 A2 D2 … A8 D8], where A1 represents the set of first order approximation coefficients for the sample, D1 is the corresponding set of first order detail coefficients for the sample, A2 and D2 similarly represent the second order approximation and detail wavelet coefficients and so forth. Because the search prefilter utilizes both the clear coat, surfacer primer, and e-coat layers, the final step involves horizontally concatenating the wavelet coefficients from each layer into a single vector in the order of clear coat, surfacer-primer, and e-coat (see Figure 3.7).

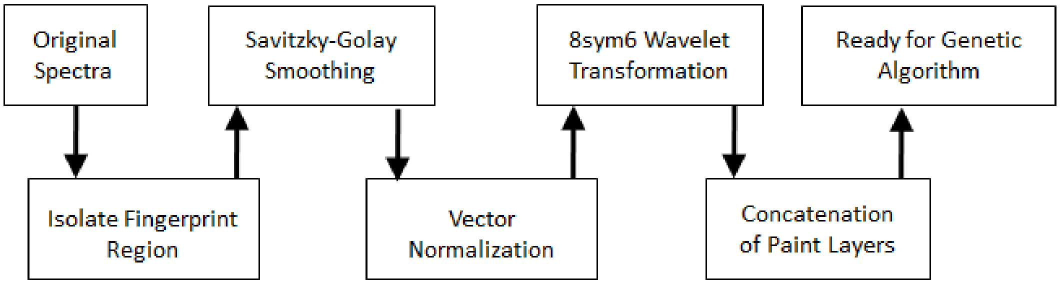

Original Spectra
Isolate Fingerprint Region
Savitzky-Golay Smoothing
Vector Normalization
8sym6 Wavelet Transformation
Concatenation of Paint Layers
Ready for Genetic Algorithm

Figure 3.6. Block diagram of the spectral preprocessing procedure used to develop search prefilters.

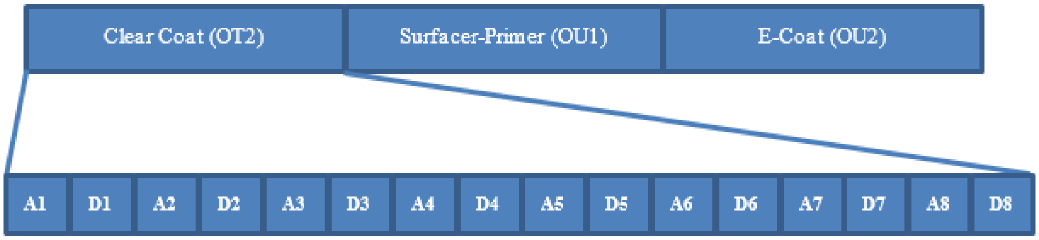


Figure 3.7. Diagram of the wavelet coefficient concatenation scheme used with the clear coat paint layer magnified to show how individual sets of approximation and detail coefficients are concatenated within each layer. Number of coefficients per block decreases with increasing level of decomposition (e.g., A1 or first order approximations and D1 or first order details contain an equal number of coefficients whereas A2 or second order approximations contains fewer coefficients than A1).

Figure 3.8 shows a PC plot of the two largest principal components of the 1179 training set samples and the 3450 wavelet coefficients comprising the training set data. Each paint sample is represented as a point in the principal component (PC) plot of the data (1 = GM, 2 = Chrysler, and 3 = Ford). The overlap of the wavelet transformed fused IR spectra of the fingerprint region for Chrysler and Ford is evident.

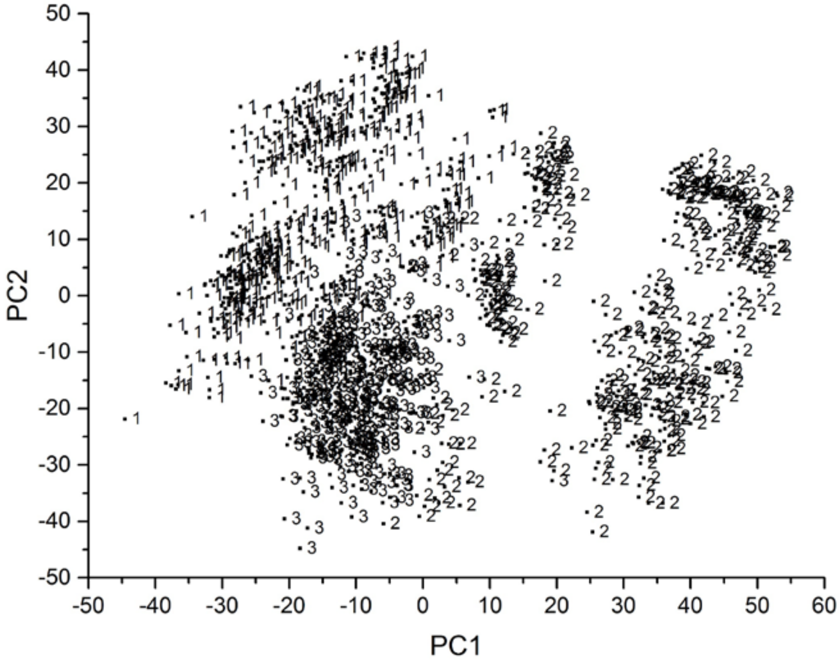


Figure 3.8. PC plot of the two largest principal components of the 1182 paint samples and the 3450 wavelet coefficients comprising the training set data (1 = GM, 2 = Chrysler, and 3 = Ford).

The next step was feature selection. A genetic algorithm for feature selection and pattern recognition analysis was used in this study to identify wavelet coefficients characteristic of automotive manufacturer. The pattern recognition GA identified wavelet coefficients by sampling key feature subsets, scoring their PC plots and tracking those paint samples or automotive manufacturers that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the pattern recognition GA identified 39 wavelet coefficients whose PC plot showed clustering of the fused IR spectra on the basis of automotive manufacturer (see Figure 3.9).

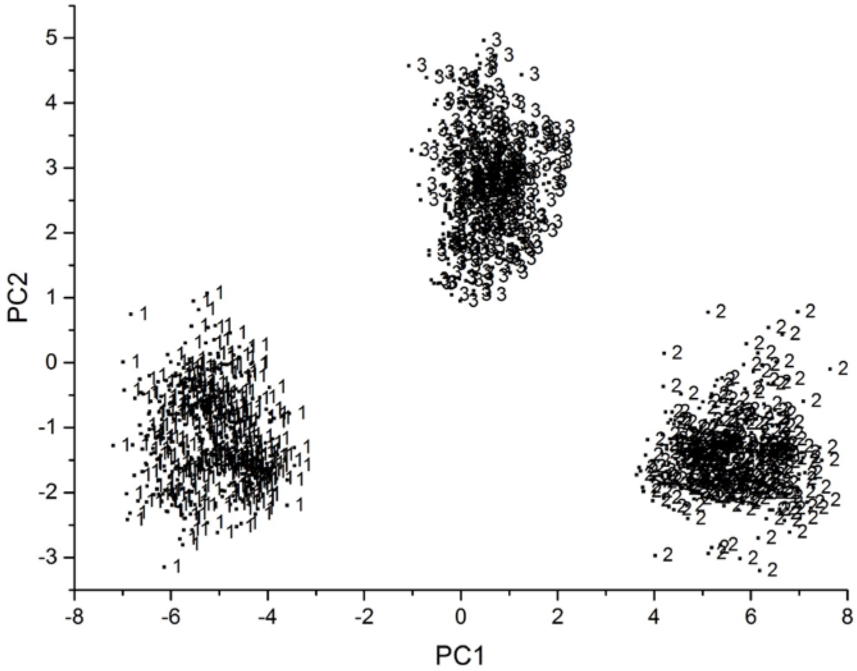


Figure 3.9. PC plot of the two largest principal components of the 1182 training set samples and the 39 wavelet coefficients identified by the pattern recognition GA (1 = GM, 2 = Chrysler, and 3 = Ford).

To assess the predictive ability of the 39 wavelet coefficients identified by the pattern recognition GA, a validation set of 124 paint samples was employed. In Figure 3.10, the validation set samples are projected onto the PC plot of the data defined by the 1179 wavelet transformed fused IR spectra and the 39 wavelet coefficients identified by the pattern recognition GA. Each validation set sample lies in a region of the PC plot with paint systems from the same automotive manufacturer. This result suggests that information about automotive manufacturer can be extracted from the wavelet transformed fused IR spectrum of an unknown paint sample.

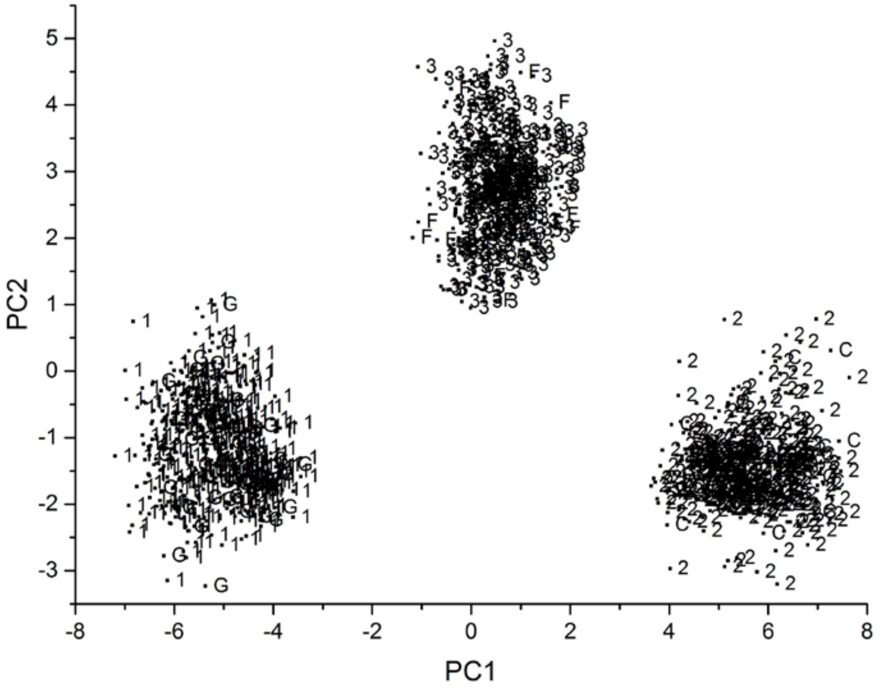


Figure 3.10. Validation set samples (in red) projected onto the PC plot of the data defined by the 1179 wavelet transformed fused IR spectra of the training set (green, yellow, cyan) and the 39 wavelet coefficients identified by the pattern recognition GA. (1 = GM, 2 = Chrysler, and 3 = Ford).

3.3.2. General Motors. The first step in the development of search prefilters for General Motors was to differentiate the General Motors paint systems by plant group. Our focus in this phase of the study was component classification rather than just IR spectral comparisons. Not only will this improve the quality of the spectral comparisons, it will also strengthen the credibility of the testimony by the forensic paint examiner during the presentation of evidence in legal proceedings.

To determine the composition of each plant group, representative spectra were selected from each assembly plant. General Motors assembly plants (see Table 3.2) whose clear coat IR spectra exhibited a doublet for the carbonyl band (indicative of the polymer acrylic melamine styrene polyurethane) as opposed to a singlet (indicative of the polymer acrylic melamine styrene**)** were flagged. Average IR spectra of the clear coat layer for each of the nine assembly plants (Baltimore, Hamtramck, Orion, Ramos Arizpe, Silao, Spring Hill, Saint Therese, Wentzville, and Wilmington) whose carbonyl exhibited a doublet were analyzed using cluster analysis. A separate cluster analysis was also performed on the average IR spectra of the clear coat layer for each of the other sixteen assembly plants (Arlington, Doraville, Fairfax, Flint, Fort Wayne, Fremont, Ingersoll, Janesville, Lansing, Linden, Lordstown, Moraine, Oklahoma City, Oshawa, Pontiac, and Shreveport) whose clear coat IR spectra exhibited a singlet for the carbonyl band.

Table 3.2. Distribution of GM Assembly Plants into Plant Groups

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

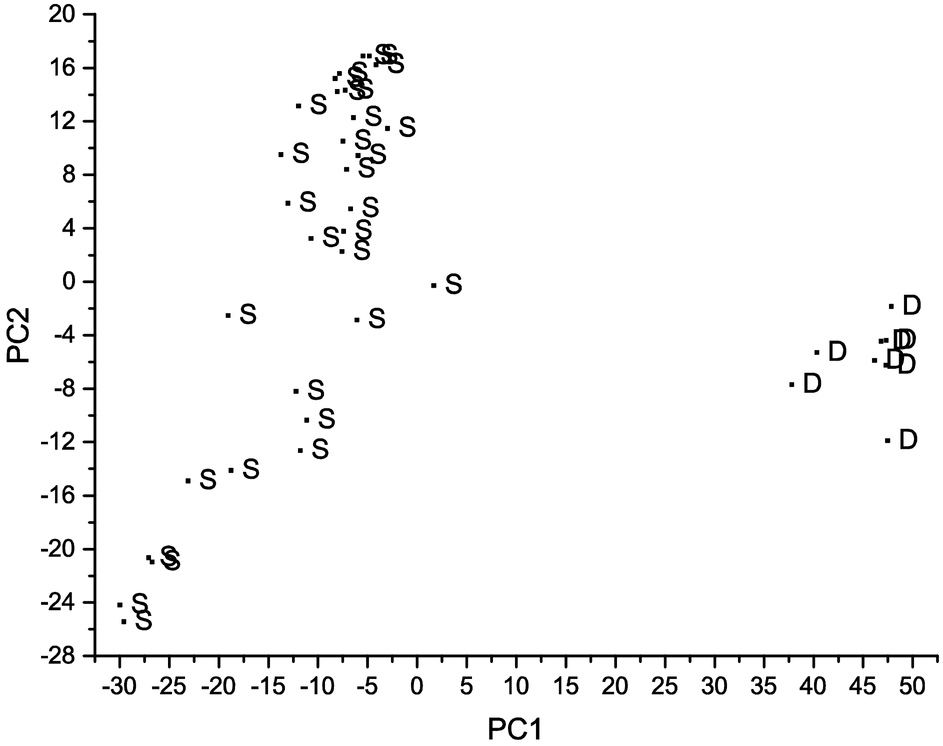


Figure 3.11. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Lansing plant. Two distinct sample clusters are evident in the plot. S = singlet (carbonyl) and D = doublet (carbonyl).

Prior to cluster analysis, principal component analysis [3-29] was performed on each assembly plant to assess its class structure. PC plots of the clear coats from the assembly plants corresponding to acrylic melamine styrene polyurethane indicated that each plant was represented by a set of similar spectra. However, clustering was observed in PC plots of three of the sixteen assembly plants (Lansing, Oklahoma City and Oshawa) whose clear coat layer corresponds to acrylic melamine styrene. For Lansing (Figure 3.11), the two clusters in the PC plot were correlated to the carbonyl band being a singlet or a doublet. (This was missed in our initial analysis of the data when the assembly plants were divided into two groups based on whether the carbonyl band was a singlet or a doublet because the representative spectra initially selected for this assembly plant were only singlets for the carbonyl.) For Oklahoma City (Figure 3.12), clustering was correlated to vehicle type (car or truck) with one sample not assigned to either sample cluster. This unique sample was tagged as an outlier after its IR spectrum was visually compared to the average clear coat IR spectrum of each of the two clusters detected in the PC plot. In the case of the Oshawa (Figure 3.13) assembly plant, three clusters were detected in the PC plot. Trucks formed a distinct cluster, whereas the automobiles were divided into two clusters on the basis of their line and model.

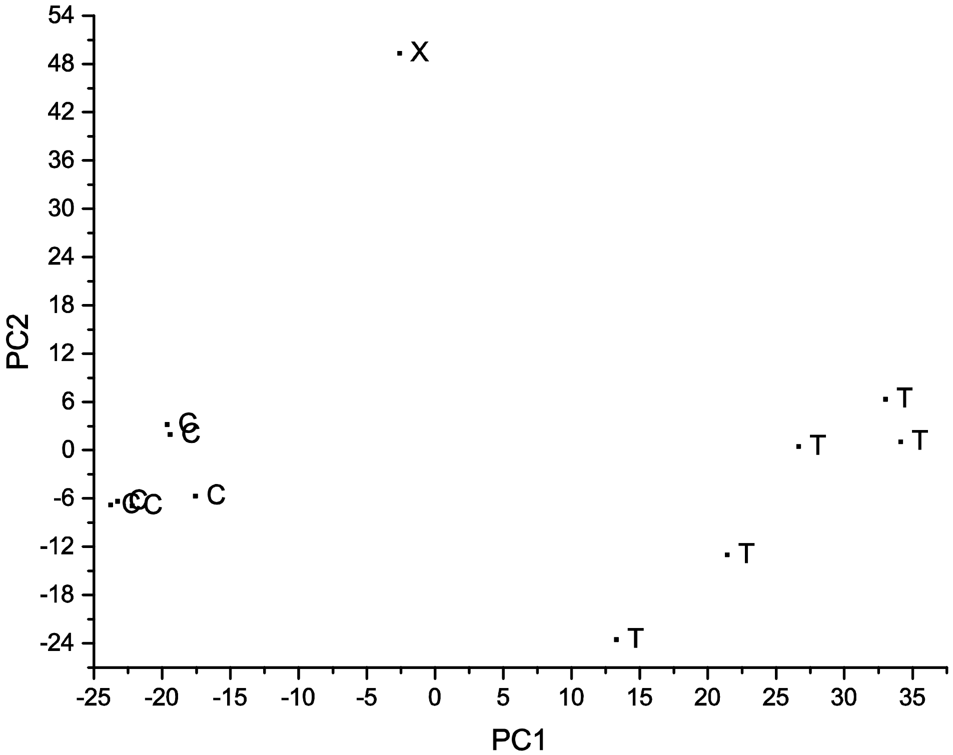


Figure 3.12. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Oklahoma City plant. Two distinct sample clusters and one outlier are evident in the plot. C = car, T = truck, and X = outlier.

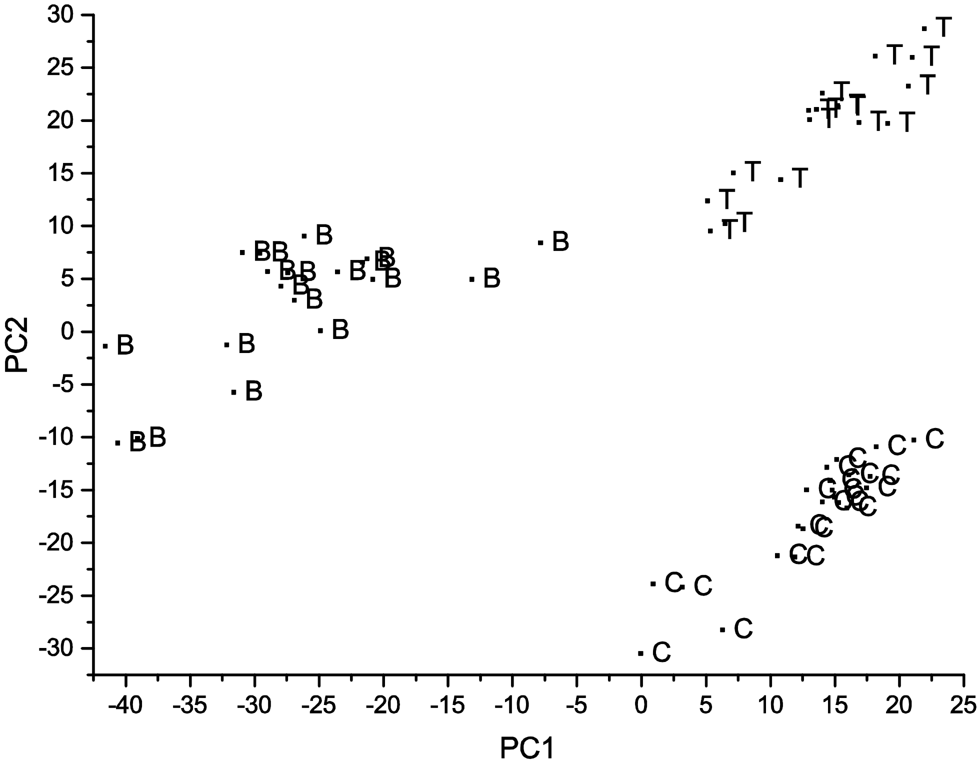


Figure 3.13. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Oshawa plant. Three distinct sample clusters are evident in the plot. B = Buick, T = truck, and C = cars other than Buick.

Because the average clear coat IR spectrum of each detected cluster in the PC plots was visually different, the three assembly plants were divided into subplants. Both the Lansing and Oklahoma City assembly plants were each divided into two subplants, and Oshawa was divided into three subplants. The Lansing subplant corresponding to the samples whose IR spectra exhibited a doublet for the carbonyl band was transferred to the cluster analysis study involving assembly plants whose clear coat layer is acrylic melamine styrene polyurethane.

To assign either assembly plants or subplants to specific plant groups, the average IR spectrum of the clear coat layer of each assembly plant or subplant was computed. Principal component analysis and hierarchical clustering [3-30] were performed on the average IR spectra. As the mathematics underlying these two methods are quite different, a strong case can be made for partitioning the data into distinct groups if both the PC plot and dendogram are in agreement.

The results of the principal component analysis (Figure 3.14) were compared to the hierarchical cluster analysis (Figure 3.15) for the nine assembly plants and one subplant (Baltimore, Hamtramck, Orion, Ramos Arizpe, Silao, Spring Hill, Saint Therese, Wentzville, Wilmington, and Lansing subplant) whose polymer formulation for the clear coat layer is acrylic melamine styrene polyurethane. From these results, the nine assembly plants and one subplant were divided into two plant groups. Plant Group 2 consists of Baltimore, Hamtramck, Wentzville, Orion, and Wilmington assembly plants, whereas Plant Group 5 consists of the Lansing subplant and the Ramos Arizpe, Silao, Spring Hill, and Saint Therese assembly plants.

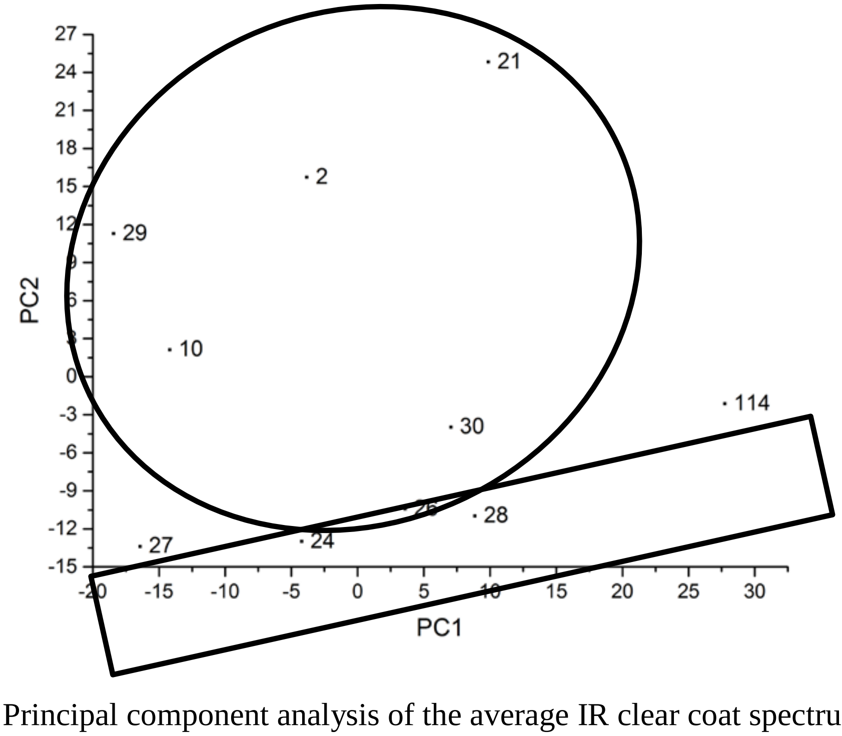


Figure 3.14. Principal component analysis of the average IR clear coat spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene polyurethane. The average spectra are divided into two distinct clusters (Plant Groups 2 and 5).

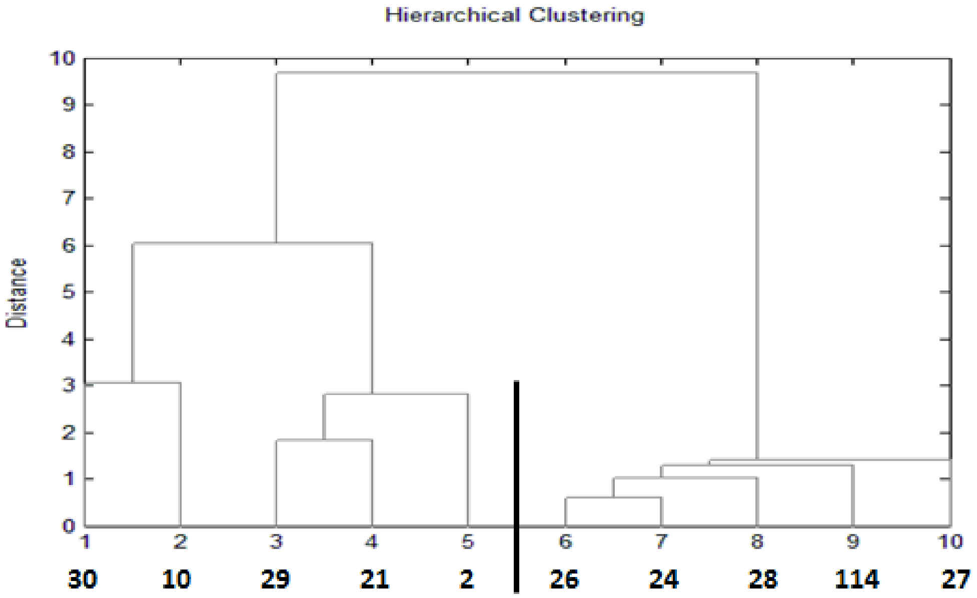


Figure 3.15. Hierarchical cluster analysis (Wards method) of the average IR clear coat spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene polyurethane. The average spectra are divided into two distinct clusters (Plant Groups 2 and 5).

The results of principal component analysis (Figure 3.16) and hierarchical clustering (Figure 3.17) of the thirteen assembly plants and six subplants whose polymer formulation for the clear coat layer is acrylic melamine styrene were compared. From these results, the thirteen assembly plants and six subplants were divided into three plant groups. Plant Group 1 consists of Arlington, Doraville, Fairfax, Fort Wayne, Lansing (subplant), Moraine, Oklahoma City (subplant), Oshawa (subplant) and Pontiac, whereas Plant Group 3 is comprised of Flint, Linden, Oklahoma City (subplant), Oshawa (subplant) and Shreveport, and Plant Group 4 consists of Janesville, Fremont, Lordstown, and Oshawa (subplant). Although the Oshwa subplant (in Plant Group 1) and the Ingersoll assembly plant (in Plant Group 3) were assigned to Plant Group 4 by the dendogram, the results from the pattern recognition GA supported the configuration shown in the PC plot.

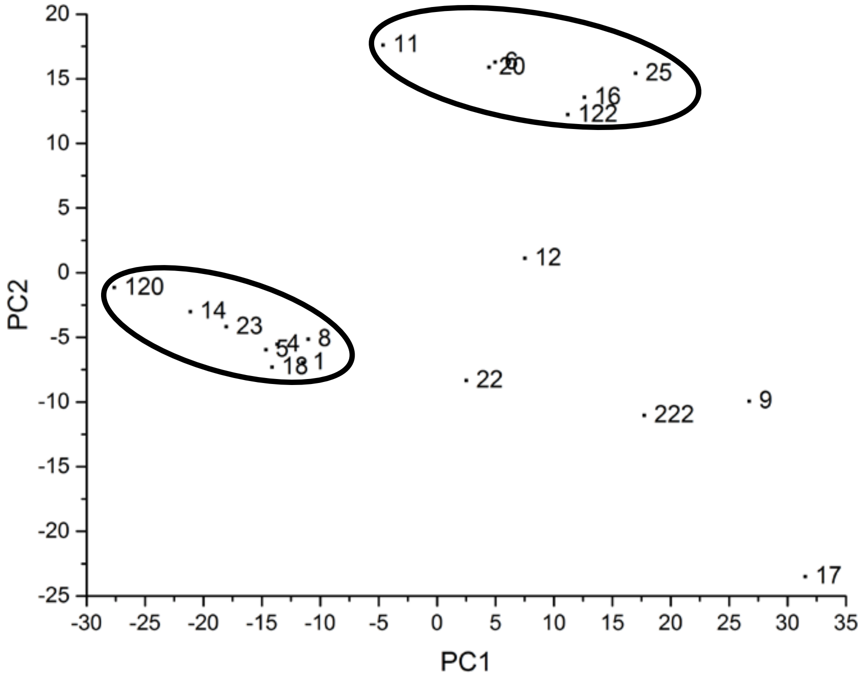


Figure 3.16. Principal component analysis of the average IR clear coat spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene. The average spectra are divided into three distinct clusters (Plant Groups 1, 3 and 4).

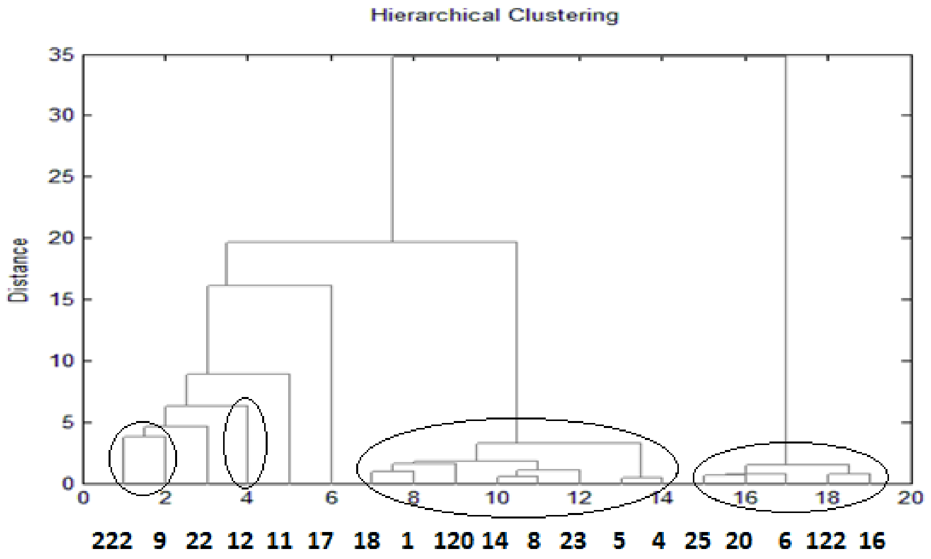


Figure 3.17. Hierarchical cluster analysis (Wards method) of the average IR clear coat spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene. The average spectra are divided into three distinct clusters (Plant Groups 1, 3 and 4).

Having ascertained the membership of each plant group, the next step was classification. The training and validation sets for Plant Group are summarized in Table 3.3. In this phase of the study, five samples were flagged as discordant observations using the outlier routines of the pattern recognition GA and were deleted from the analysis. The training set (424 wavelet transformed clear coat IR spectra and 1150 wavelet coefficients) was divided into five categories on the basis of Plant Group. A PC plot of the 424 training set samples and 1150 wavelet coefficients is shown in Figure 3.18.

Table 3.3. Training Set and Validation Set for General Motors Plant Groups

Plant Group	Training	Validation
1	171	17
2	46	7
3	73	7
4	75	6
5	59	7

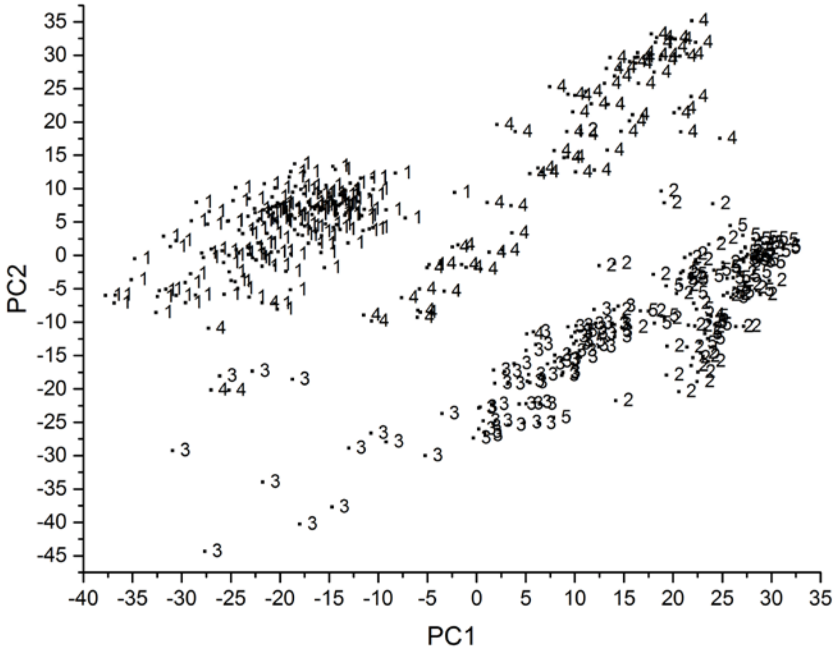


Figure 3.18. PC plot of the 424 training set samples and the 1150 wavelet coefficients of the clear coat layer.

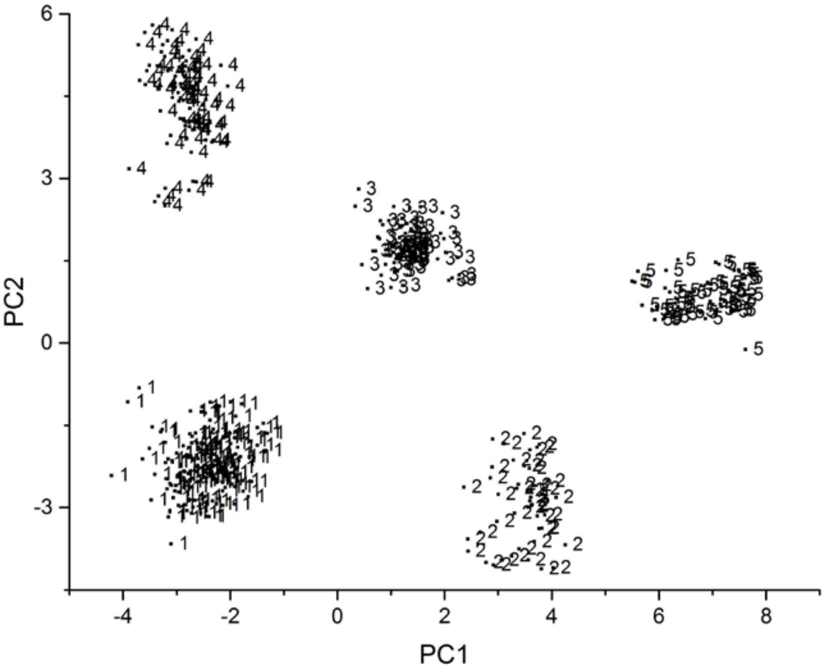


Figure 3.19. PC plot of the 424 training set samples and the 33 wavelet coefficients identified by the pattern recognition GA.

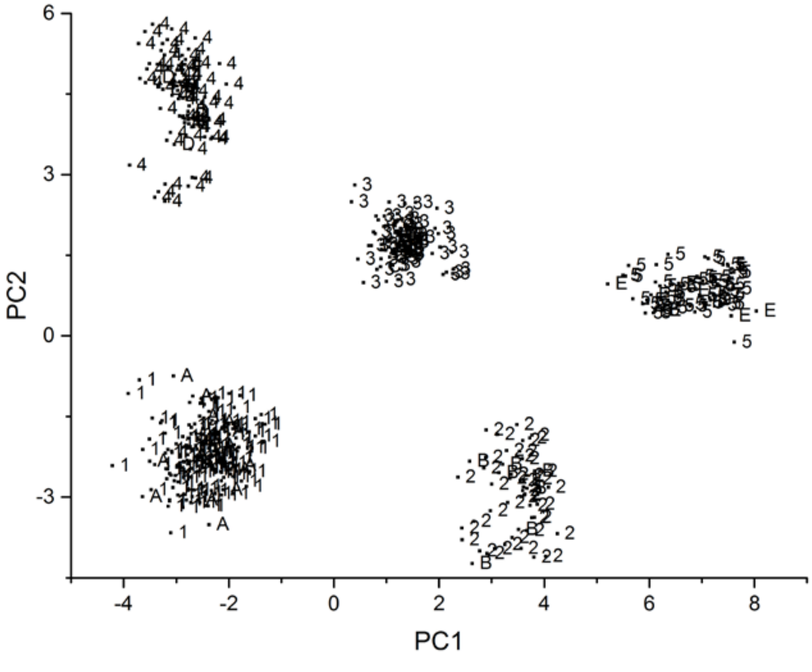


Figure 3.20. Projection of the 44 validation set samples onto the PC plot of the 424 training set samples and the 33 wavelet coefficients identified by the pattern recognition GA. Validation set: A = Plant Group 1, B = Plant Group 2, C = Plant Group 3, D = Plant Group 4, E = Plant Group 5.

The pattern recognition GA identified wavelet coefficients characteristic of plant group by sampling key feature subsets, scoring their PC plots and tracking those samples or groups 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 33 wavelet coefficients whose PC plot (see Figure 3.19) showed clustering of the fused IR spectra on the basis of their plant group. The 44 validation set samples were then projected onto the PC plot (see Figure 3.20) define by the 424 training set samples and the 33 wavelet coefficients identified by the pattern recognition GA. Each validation set sample lies in a region of the PC plot with paint systems from the same automotive manufacturer.

Table 3.4. Assembly Plants and Subplants Comprising Plant Group 1

Assembly Plant	Training Set	Validation Set
1 (Arlington)	17	4
2 (Fairfax)	25	3
3 (Moraine)	27	2
4 (subplant of Oshawa)	18	1
5 (subplant of Oklahoma City)	4	1
6 (Doraville and subplant of Lansing)	52	5
7 (Fort Wayne and Pontiac)	28	1

For each plant group, a search prefilter was developed to discriminate the samples by assembly plant or subplant using the clear coat, surfacer-primer and e-coat layers. The pattern recognition GA identified specific wavelet coefficients in each layer correlated to the assembly plant of the vehicle from which the paint sample was obtained. Table 3.4 lists the six assembly plants (Arlington, Doraville, Fairfax, Fort Wayne, Moraine, Pontiac) and three subplants (Lansing, Oklahoma City, Oshawa) comprising Plant Group 1. Doraville and the Lansing subplants were combined into a plant subgroup as were the Fort Wayne and Pontiac assembly plants because the average spectra of their clear coat, surfacer-primer and e-coat layers were superimposable.

The pattern recognition GA identified wavelet coefficients characteristic of assembly plant by sampling key feature subsets, scoring their PC plots and tracking those samples and/or classes 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 55 wavelet coefficients whose PC plot shows clustering on the basis of assembly plant (see Figure 3.21).

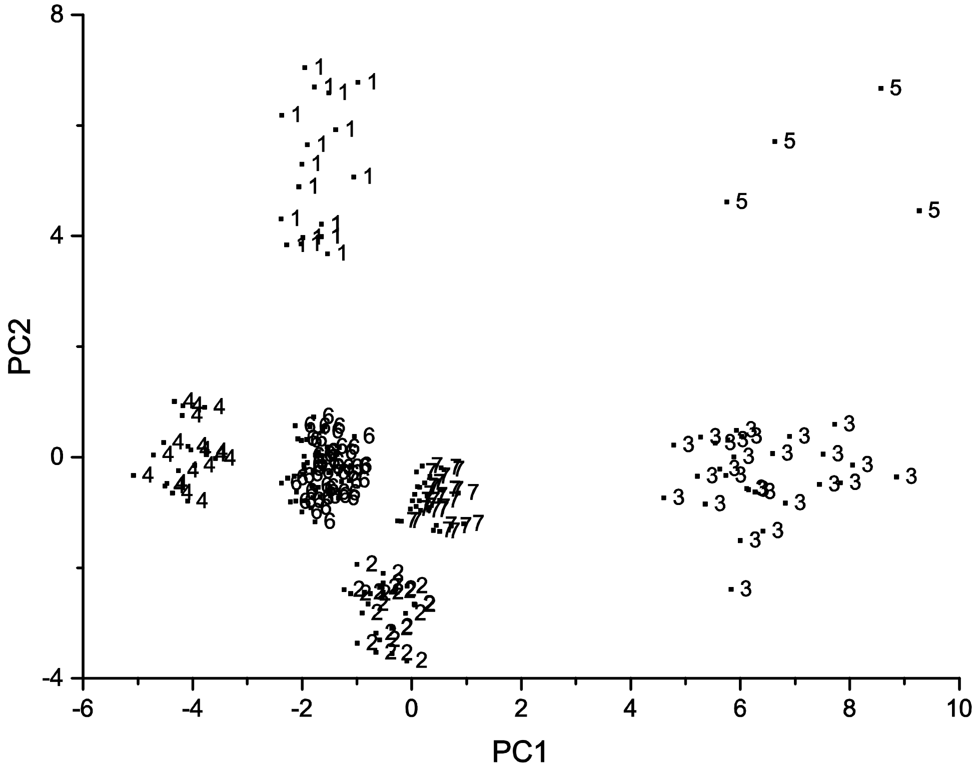


Figure 3.21. Plot of the two largest principal components of the 171 samples comprising Plant Group 1 and the 55 wavelet coefficients identified by the pattern recognition GA for the training set. (1 = Arlington, 2 = Fairfax, 3 = Moraine, 4 = subplant of Oshawa, 5 = subplant of Oklahoma City, 6 = Doraville and subplant of Lansing, and 7 = Fort Wayne and Pontiac).

When the validation set samples assigned to Plant Group 1 (see Table 4) were projected onto the PC plot of the 171 training set samples and 55 wavelet coefficients identified by the pattern recognition GA, each projected validation set sample was located in a region of the map with paint samples from the same assembly plant or subplant (see Figure 3.22).

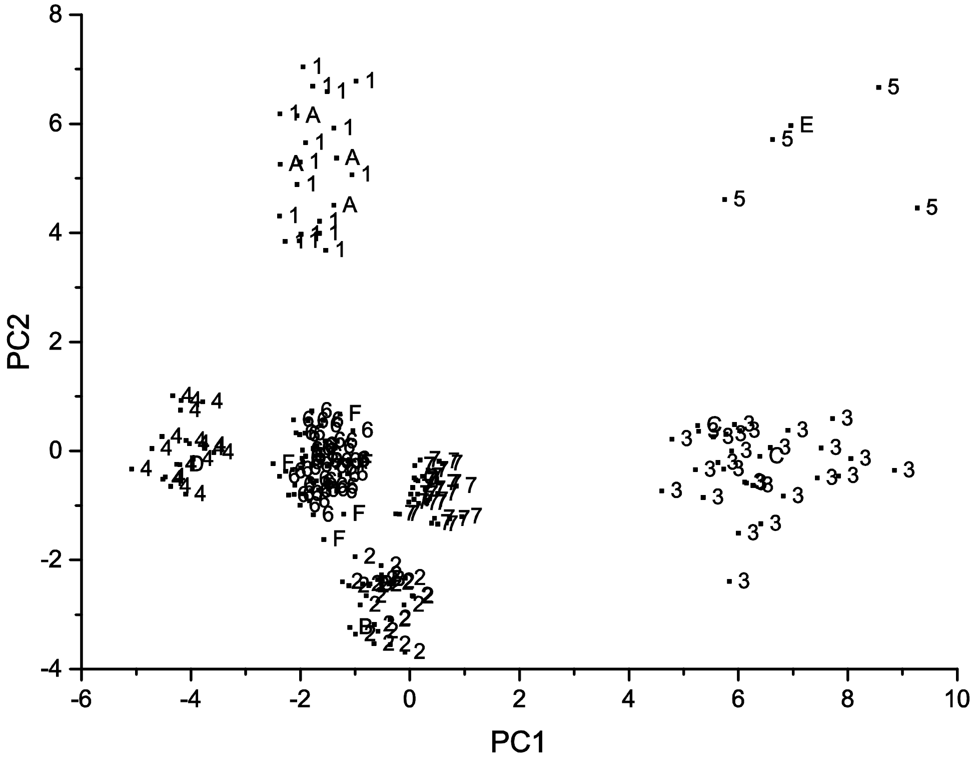


Figure 3.22. Projection of the validation set samples assigned to Plant Group 1 onto the PC plot defined by the 171 samples comprising the training set and the 55 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Arlington, 2 = Fairfax, 3 = Moraine, 4 = subplant of Oshawa, 5 = subplant of Oklahoma City, 6 = Doraville and subplant of Lansing, and 7 = Fort Wayne and Pontiac. Validation set: A = Arlington, B = Fairfax, C = Moraine, D = subplant of Oshawa, E = subplant of Oklahoma City, F = Doraville and subplant of Lansing, and G = Fort Wayne and Pontiac assembly plants.

Table 3.5 lists the five assembly plants (Baltimore, Hamtramck, Orion, Wentzville, Wilmington) comprising Plant Group 2. Figure 3.23 shows a plot of the two largest principal components of the 46 samples comprising Plant Group 2 and the 9 wavelet coefficients identified by the pattern recognition GA. Again, all assembly plants are well separated from each other in the PC plot. Projecting the validation set samples assigned to Plant Group 2 (see Table 3.5) onto the PC plot shows that each projected validation set sample lies in a region of the PC map with samples from the same assembly plant (see Figure 3.24).

Table 3.5. Assembly Plants Comprising Plant Group 2

Assembly Plant	Training	Validation
1 (Baltimore)	8	0
2 (Hamtramck)	16	2
3 (Orion)	7	5
4 (Wentzville)	8	0
5 (Wilmington)	7	0

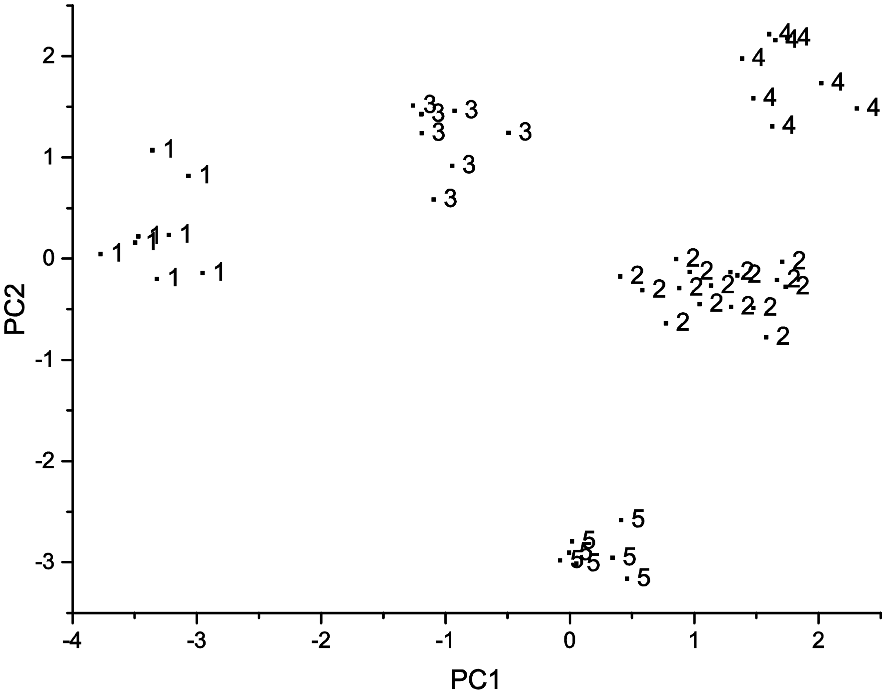


Figure 3.23. Plot of the two largest principal components of the 46 samples comprising Plant Group 2 and the 9 wavelet coefficients identified by the pattern recognition GA for the training set. 1 = Baltimore, 2 = Hamtramck, 3 = Orion, 4 = Wentzville, and 5 = Wilmington assembly plants.

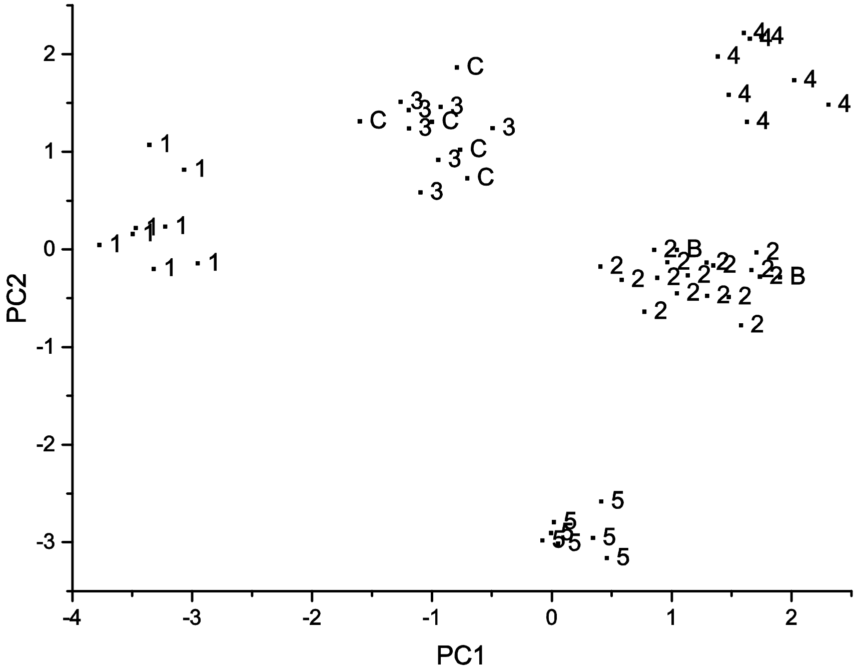


Figure 3.24. Projection of validation set samples assigned to Plant Group 2 onto the PC plot defined by the 46 samples of the training set and the 9 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Baltimore, 2 = Hamtramck, 3 = Orion, 4 = Wentzville, and 5 = Wilmington. Validation set: B = Hamtramck, C = Orion assembly plants.

Table 3.6 lists the four assembly plants (Flint, Ingersoll, Linden, Shreveport) and two subplants (Oklahoma City, Oshawa) comprising Plant Group 3. Figure 3.25 shows a plot of the two largest principal components of the 72 samples comprising Plant Group 3 and the 43 wavelet coefficients identified by the pattern recognition GA. Again, every assembly plant is well separated in the PC plot. Projecting the validation set samples assigned to Plant Group 3 by the plant group search prefilter (see Table 3.6) onto the PC map shows that each projected validation set sample is located in a region of the map with paint samples from the same assembly plant (see Figure 3.26).

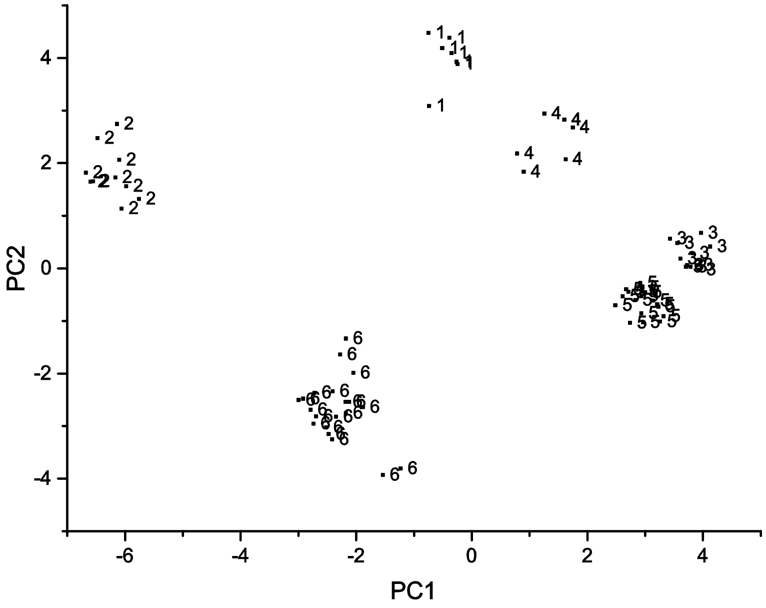


Figure 3.25. Plot of the two largest principal components of the 72 samples comprising Plant Group 3 and the 43 wavelet coefficients identified by the pattern recognition GA for the training set.

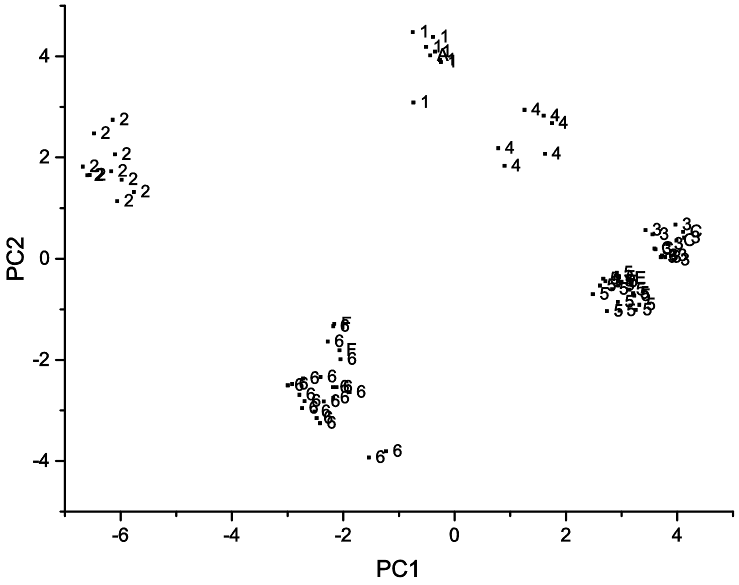


Figure 3.26. Projection of validation set samples assigned to Plant Group 3 (see Table 5) onto the PC plot defined by the 72 samples of the training set and the 43 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Flint, 2 = Ingersoll, 3 = Linden, 4 = subplant of Oklahoma City, 5 = Shreveport, 6 = subplant of Oshawa. Validation Set: A = Flint, C = Linden, E = Shreveport, and F = subplant of Oshawa.

Table 3.6. Assembly Plants and Subplants Comprising Plant Group 3

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

Table 3.7 lists the three assembly plants (Fremont, Janesville, Lordstown) and one subplant (Oshawa) comprising Plant Group 4. Figure 3.27 shows a plot of the two largest principal components of the 59 samples comprising Plant Group 4 and the 20 wavelet coefficients identified by the pattern recognition GA. Every assembly plant is well separated in the PC plot. Projecting the validation set samples assigned to Plant Group 4 by the plant group search prefilter (see Table 3.7) onto the PC map showed that each projected validation set sample is located in a region of the map with samples from the same assembly plant (see Figure 3.28).

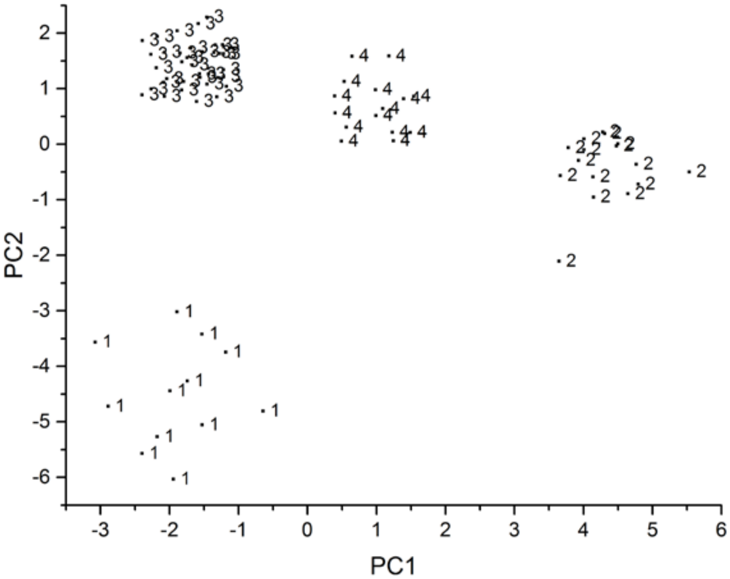


Figure 3.27. Plot of the two largest principal components of the 59 samples comprising Plant Group 4 and the 20 wavelet coefficients identified by the pattern recognition GA for the training set.

Table 3.7. Assembly Plants and Subplant Comprising Plant Group 4

Assembly Plant	Training	Validation
1 (Fremont)	10	0
2 (Janesville)	16	0
3 (Lordstown)	34	4
4 (subplant of Oshawa)	15	2

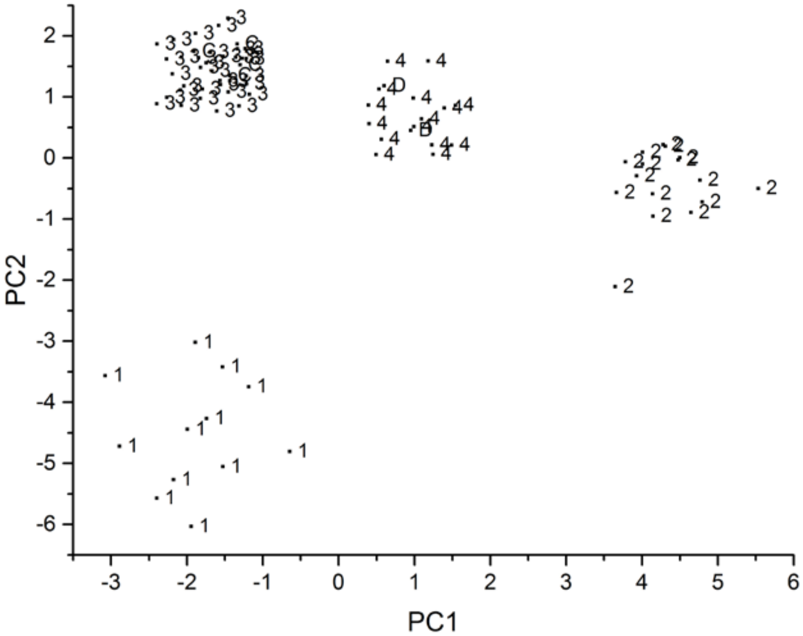


Figure 3.28. Projection of validation set samples assigned to Plant Group 4 onto the PC plot defined by the 59 samples of the training set and the 20 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Fremont, 2 = Janesville, 3 = Lordstown, and 4 = subplant of Oshawa. Validation set: C = Lordstown, and D = subplant of Oshawa.

Table 3.8 lists the four assembly plants (Ramos Arizpe, Silao, Spring Hill, and St. Therese) and one subplant (Lansing) comprising Plant Group 5. Figure 3.29 shows a plot of the two largest principal components of the 59 paint samples comprising Plant Group 5 and the 30 wavelet coefficients identified by the pattern recognition GA. Every assembly plant is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 5 (see Table 3.8) onto the PC map shows that each projected validation set sample is located in a region of the map with samples from the same assembly plant (see Figure 3.30). The results from the General Motors study show that search prefilters developed from IR spectra of the clear coat and the two undercoat layers can characterize an unknown paint sample by assembly plant or the subplant of the vehicle from which the paint sample originated.

Table 3.8. Assembly Plants and Subplant Comprising Plant Group 5

Assembly Plant	Training	Validation
1 (Ramos Arizpe)	21	3
2 (Silao)	18	0
3 (Spring Hill)	8	1
4 (Saint Therese)	6	1
5 (subplant of Lansing)	6	2

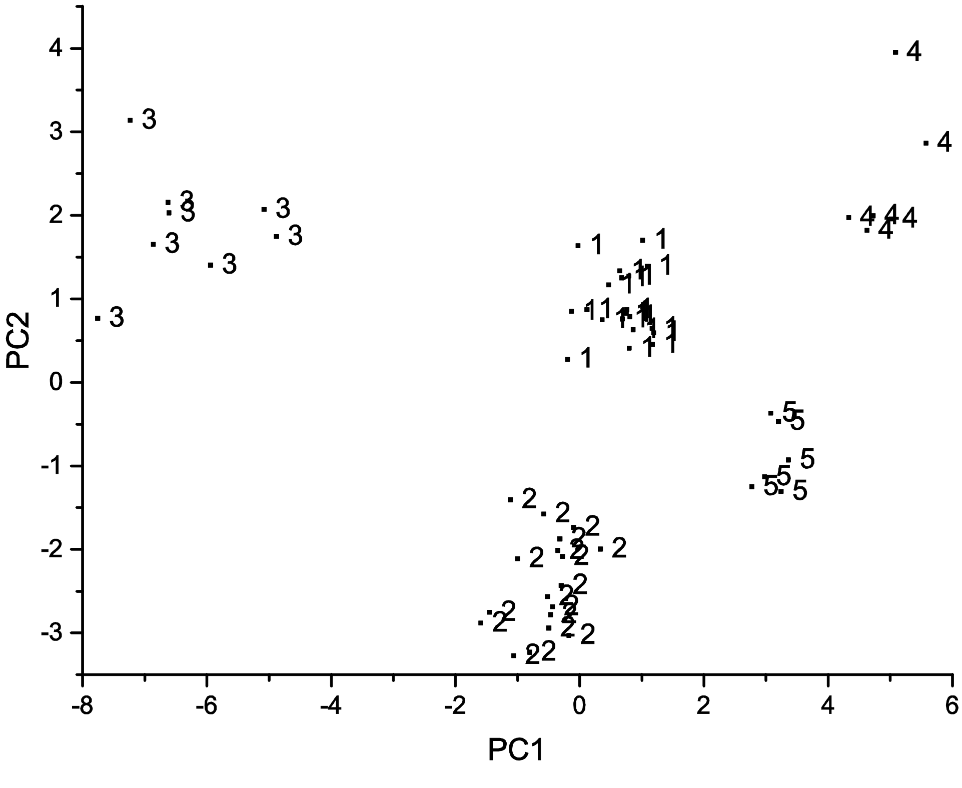


Figure 3.29. Plot of the two largest principal components of the 59 samples comprising Plant Group 5 and the 30 wavelet coefficients identified by the pattern recognition GA for the training set.

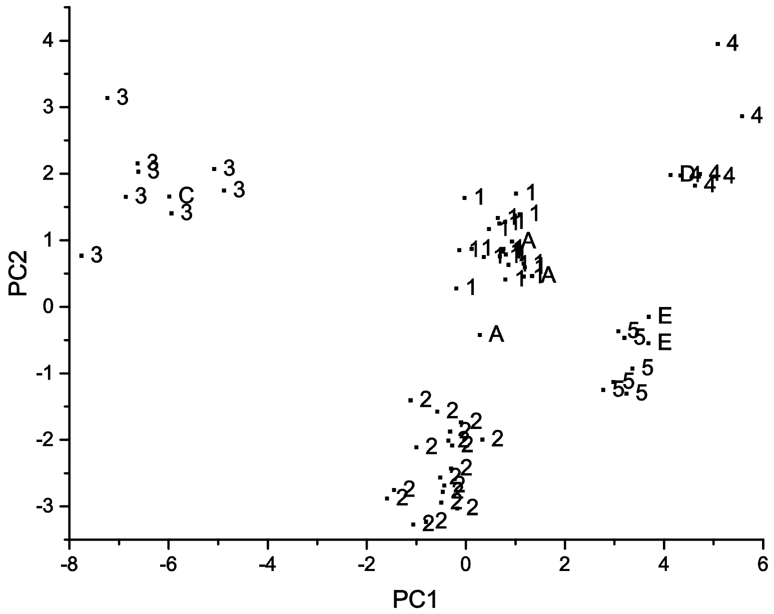
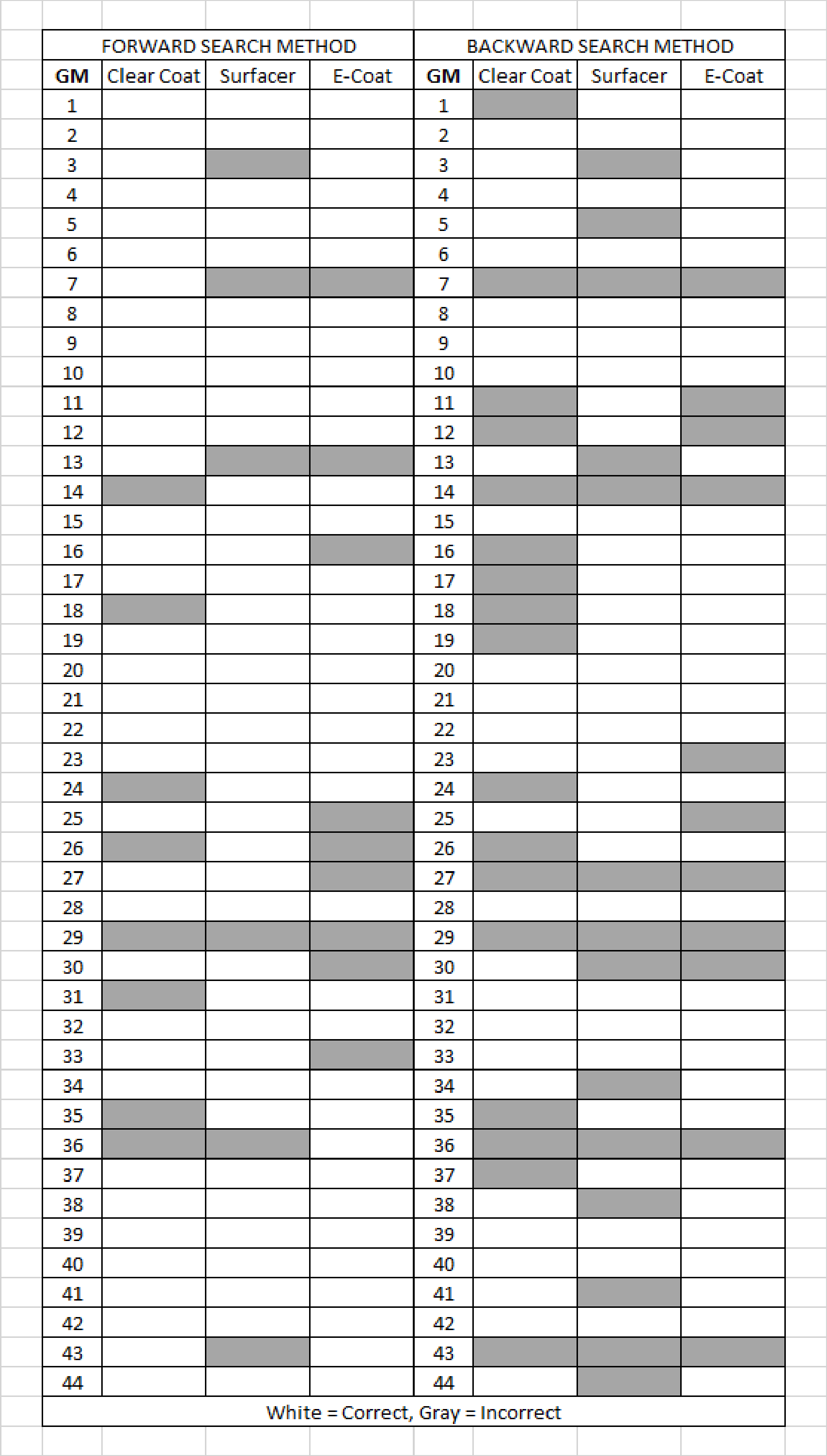


Figure 3.30. Projection of validation set samples assigned to Plant Group 5 (see Table 5) onto the PC plot defined by the 59 samples of the training set and the 30 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Ramos Arizpe, 2 = Silao, 3 = Spring Hill, 4 = Saint Therese, 5 = subplant of Lansing. Validation set: A = Ramos Arizpe, C = Spring Hill, D = Saint Therese, and E = subplant of Lansing.

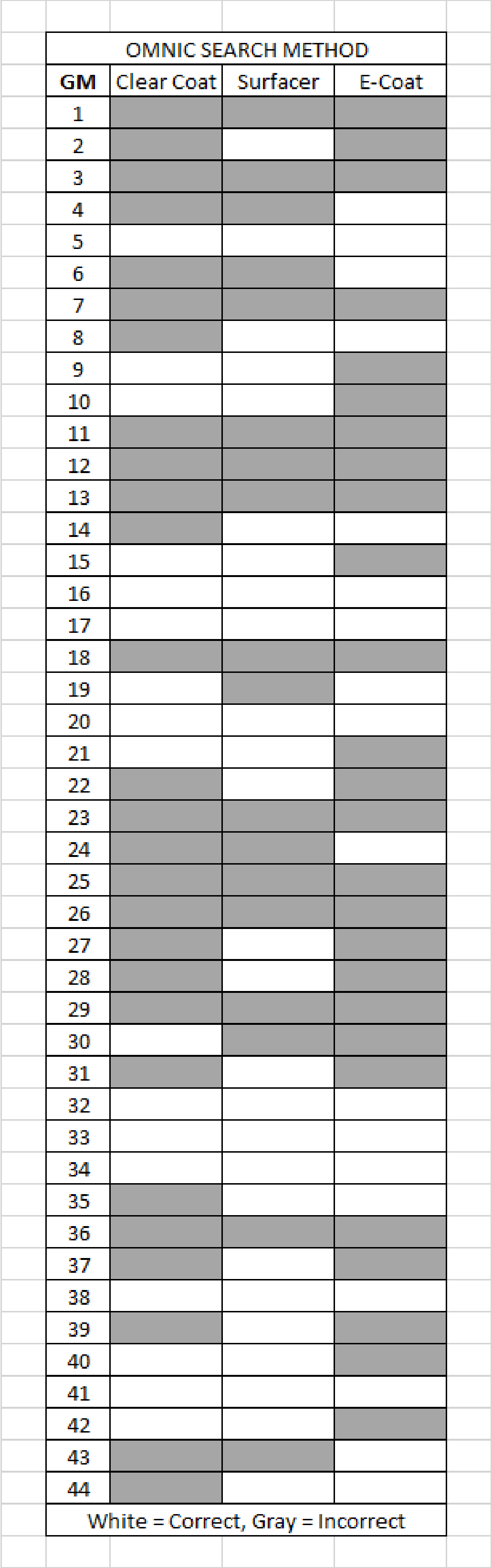
To extract information about the line and model of the vehicle, the two cross correlation library search algorithms were employed in tandem to identify library spectra most similar to the unknown and to assess the degree to which the sample is represented in the truncated IR library identified by the search prefilters. Library search results are summarized in Table 3.9 for the 44 validation set samples using the prototype pattern recognition search system and in Table 3.10 for OMNIC. Both cross correlation search algorithms outperformed OMNIC. 27 of the 44 samples were correctly matched in all 3 layers by the first cross correlation algorithm (which identifies the IR spectrum in the library most similar to the unknown), whereas only 17 samples were correctly matched in all 3 layers by the second cross correlation algorithm (which computes a probability index as to the line and model of the unknown vehicle). Only 9 samples were correctly matched in all 3 layers by OMNIC.

Table 3.9. Prototype Cross Correlation Library Search Results for GM



FORWARD SEARCH METHOD				BACKWARD SEARCH METHOD			
GM	Clear Coat	Surfacer	E-Coat	**GM**	Clear Coat	Surfacer	E-Coat
1				1	▨		
2				2			
3		▨		3		▨	
4				4			
5				5		▨	
6				6			
7		▨	▨	7	▨	▨	▨
8				8			
9				9			
10				10			
11				11	▨		▨
12				12	▨		▨
13		▨	▨	13		▨	
14	▨			14	▨	▨	▨
15				15			
16			▨	16	▨		
17				17	▨		
18	▨			18	▨		
19				19	▨		
20				20			
21				21			
22				22			
23				23			▨
24	▨			24	▨		
25			▨	25			▨
26	▨		▨	26	▨		
27			▨	27	▨	▨	▨
28				28			
29	▨	▨	▨	29	▨	▨	▨
30			▨	30		▨	▨
31	▨			31			
32				32			
33			▨	33			
34				34		▨	
35	▨			35	▨		
36	▨	▨		36	▨	▨	▨
37				37	▨		
38				38		▨	
39				39			
40				40			
41				41		▨	
42				42			
43		▨		43	▨	▨	▨
44				44		▨	
White = Correct, Gray = Incorrect							

Table 3.10. OMNIC Library Search Results for GM



	OMNIC SEARCH METHOD		
GM	Clear Coat	Surfacer	E-Coat
1	■	■	■
2	■	□	■
3	■	■	■
4	■	■	□
5	□	□	□
6	■	■	□
7	■	■	■
8	■	□	□
9	□	□	■
10	□	□	■
11	■	■	■
12	■	■	■
13	■	■	■
14	■	□	□
15	□	□	■
16	□	□	□
17	□	□	□
18	■	■	■
19	□	■	□
20	□	□	□
21	□	□	■
22	■	□	■
23	■	■	■
24	■	■	□
25	■	■	■
26	■	■	■
27	■	□	■
28	■	□	■
29	■	■	■
30	□	■	■
31	■	□	■
32	□	□	□
33	□	□	□
34	□	□	□
35	■	□	□
36	■	■	■
37	■	□	■
38	□	□	□
39	■	□	■
40	□	□	■
41	□	□	□
42	□	□	■
43	■	■	□
44	■	□	□
White = Correct, Gray = Incorrect			

The first cross correlation library search algorithm always yielded a reasonable match, which was always superior to the match between the validation set sample and the actual model even for incorrectly matched samples and layers. Although the second cross correlation library search algorithm did not perform as well as the first algorithm, it has the advantage of providing insight into how well the truncated library matches each validation set sample, rather than how well an individual validation set sample matches spectra in the truncated library. For the clear coat layer, 36 samples were correctly matched by the forward search, 27 samples by the backward search, and 26 samples were assigned to the same line and model by both searches. For the surfacer-primer, 38 samples were correctly matched by the forward search, 30 by the backward search, and 30 samples were assigned to the same line and model by both searches. For the e-coat, 35 samples were correctly matched by the forward search, 33 by the backward search, and 29 samples were assigned to the same model and line by both searches. If a specific line and model was common to both hit lists (forward and backward cross-correlation searches), the assignment for the corresponding validation set sample was always correct. Validation set samples assigned to the same line and model by the two cross correlation search algorithms were always well represented in the library and correlated well on an individual basis to specific samples present in the library. For these validation set samples, one can have confidence in the accuracy of the match as both the forward and backward matches produced the same result.

Library searches performed by OMNIC yielded hit quality index values for the five top matches that typically exceeded 99% for correctly and also for incorrectly matched samples. (A sample was judged to be incorrectly matched if the model and line of the vehicle did not correspond to any of the samples in the hit-list.) The gap between the top hit and the twentieth hit was typically less than 2%, and the gap between the top hit and the hundredth hit was typically less than 3%, which suggests that a close visual inspection of the results from these searches was necessary. For the clear coat layer, 17 samples were correctly matched, whereas the number of samples correctly matched for the surfacer-primer and the e-coat layer by OMNIC were 26 and 18 respectively.

Neither the actual hit quality index value nor the size of the gap was a reliable indicator of the uniqueness of the matches in these OMNIC searches. Clearly, a value of the hit quality index by itself was not indicative of the quality of a spectral match for the clear coat, surfacer-primer or e-coat layers as incorrectly matched samples often had higher hit quality index values than correctly matched samples. For the forensic paint examiner, assessing the accuracy of a library search for an unknown paint sample using a commercial library search algorithm such as OMNIC is often problematic due to the similarity of the IR spectra of the clear coat, surfacer-primer, and e-coat paint layers.

3.3.3. Chrysler. The first step in the development of the Chrysler search prefilters was to differentiate the Chrysler paint systems by plant group. To determine the composition of each plant group, representative spectra were selected for each assembly plant. Chrysler assembly plants (see Table 3.11) whose clear coat IR spectra exhibited a doublet for the carbonyl band (indicative of acrylic melamine styrene polyurethane) as opposed to a singlet (indicative of acrylic melamine styrene**)** were flagged. The two assembly plants (Jefferson North and Newark) whose clear coat IR spectra exhibited a doublet for the carbonyl band were placed in Plant Group 13, whereas the other 10 assembly plants whose clear coat IR spectra exhibited a singlet for the carbonyl band were assigned to other plant groups.

Table 3.11. Distribution of Chrysler Assembly Plants into Plant Groups

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

Using only clear coat IR spectra, each of the 10 remaining assembly plants (see Table 3.11) was analyzed by principal component analysis to assess the class structure. In four of the 10 assembly plants (Bramalea/Brampton, Dodge Main, St. Louis, and Toledo), the corresponding principal component plot of the clear coat layer exhibited two distinct sample clusters. For the Bramalea/Brampton assembly plant (Figure 3.31), clustering occurred on the basis of model: Dodge Charger and some Chrysler 300 lines versus Chrysler Concorde, Chrysler LHS, Dodge Intrepid, Dodge Magnum, and other Chrysler 300 lines, whereas for Dodge Main (Figure 3.32), clustering occurred on the basis of the production year of the vehicle: 2000-2002 versus 2003-2006. For the St. Louis assembly plant (Figure 3.33), clustering occurred on the basis of model and line: Dodge Caravan and Chrysler Town and Country versus Dodge Ram, whereas for Toledo (Figure 3.34), clustering was correlated to a specific vehicle: Jeep Liberty versus the other models and lines assembled at the plant. Because the average clear coat paint spectrum of each cluster was noticeably different when compared visually, the four assembly plants were further divided into subplants on the basis of the observed sample clustering.

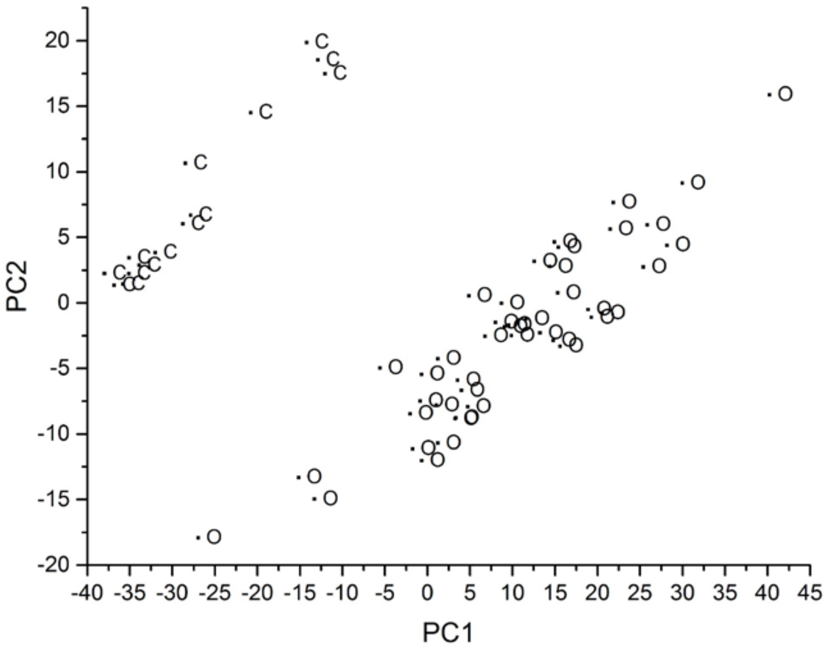


Figure 3.31. Plot of the two largest principal components of the clear coat spectra from the Bramalea/Brampton assembly plant. Two distinct clusters are present in the plot. C = Charger, some 300 series, O = rest of 300 series, other lines and models.

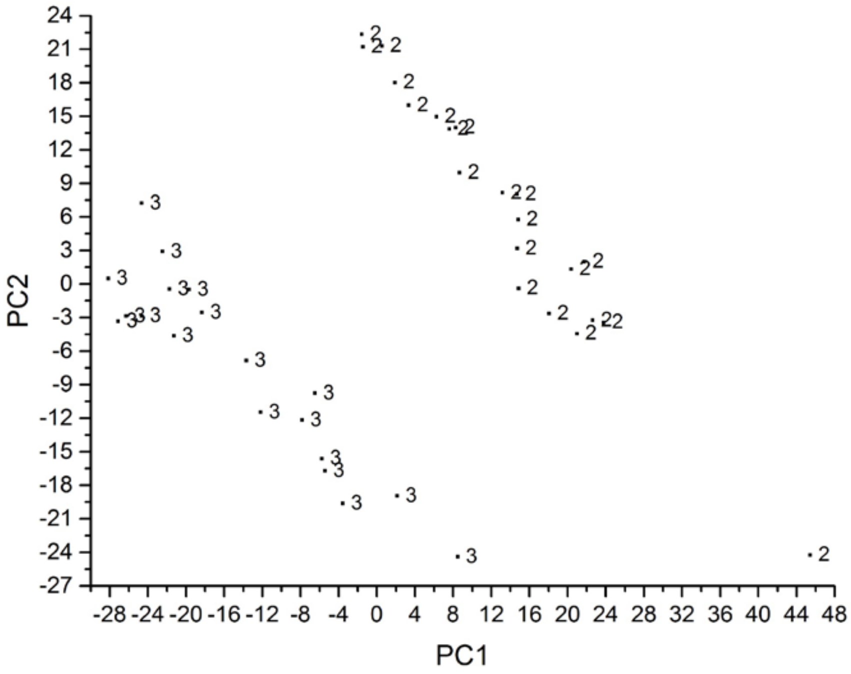


Figure 3.32. Plot of the two largest principal components of the clear coat spectra from the Dodge Main assembly plant. Two distinct clusters are present in the plot. 2 = 2000-2002, 3 = 2003-2006.

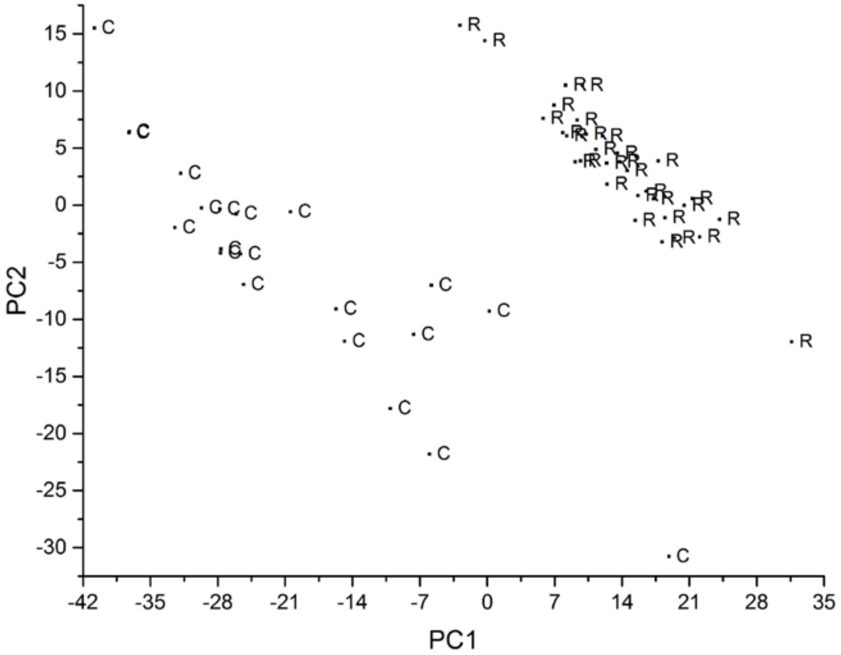


Figure 3.33. Plot of the two largest principal components of the clear coat spectra from the St. Louis assembly plant. Two distinct clusters are present in the plot. C = Caravan, Town and Country, R = Ram.

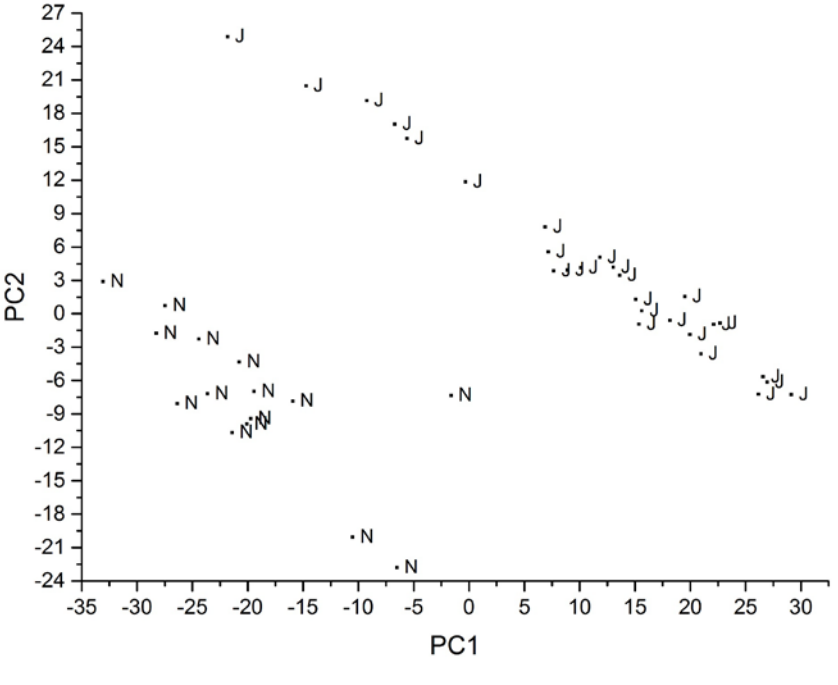


Figure 3.34. Plot of the two largest principal components of clear coat spectra from the Toledo assembly plant. Two distinct clusters are present in the plot. J = Liberty, N = other lines and models.

To assign the remaining 6 assembly plants and the 8 subplants to specific plant groups, the average IR spectrum of the clear coat layer of each assembly plant or subplant was computed. Principal component analysis (Figure 3.35) and hierarchical clustering (Figure 3.36) were performed on these average spectra. Plant Group 11 consists of Belvidere, Bramalea/Brampton (subplant), Dodge Main (subplant), Saltillo, St. Louis (subplant), Toledo (subplant), and Toluca assembly plants, whereas Plant Group 12 is comprised of Bloomington, Bramalea/Brampton (subplant), Dodge Main (subplant), Sterling Heights, St. Louis (subplant), Toledo (subplant), and the Windsor assembly plants.

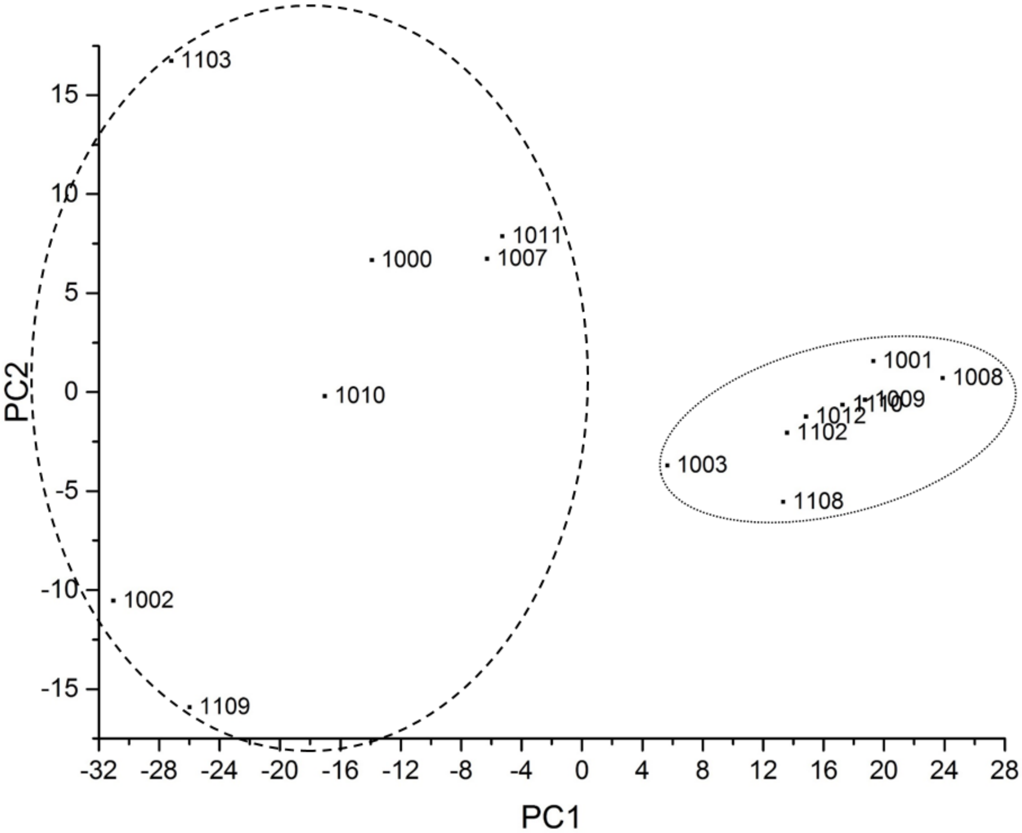


Figure 3.35. Principal component analysis of the average IR spectrum (clear coats) of each assembly plant or subplant. 1000 = Belvidere, 1001 = Bloomington, 1002 = Bramalea/Brampton subplant, 1003 = Dodge Main subplant, 1007 = Saltillo, 1008 = Sterling Heights, 1009 = St. Louis subplant, 1010 = Toledo, 1011 = Toluca, 1012 = Windsor, 1102 = Bramalea/Brampton subplant, 1103 = Dodge Main subplant, 1109 = St. Louis subplant, and 1110 = Toledo subplant.

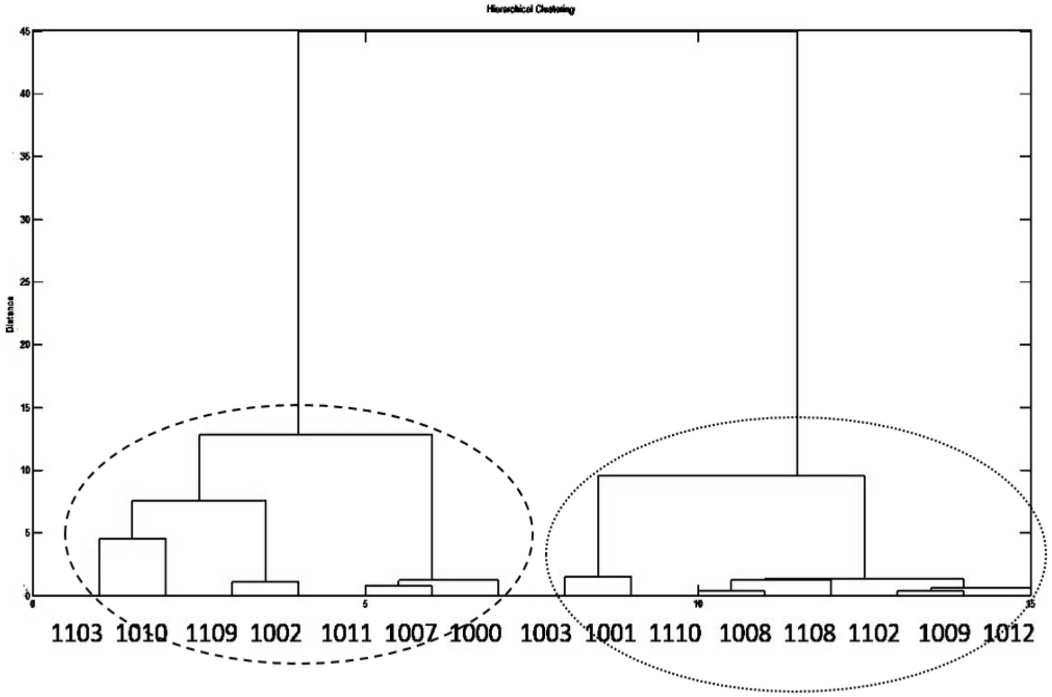


Figure 3.36. Hierarchical cluster analysis (Wards method) of the average IR spectrum (clear coats) of each assembly plant or subplant. 1000 = Belvidere, 1001 = Bloomington, 1002 = Bramalea/Brampton subplant, 1003 = Dodge Main subplant, 1007 = Saltillo, 1008 = Sterling Heights, 1009 = St. Louis subplant, 1010 = Toledo, 1011 = Toluca, 1012 = Windsor, 1102 = Bramalea/Brampton subplant, 1103 = Dodge Main subplant, 1109 = St. Louis subplant, and 1110 = Toledo subplant.

Having ascertained the membership of each plant group, the next step was classification. The training set and validation set for the Chrysler plant groups are summarized in Table 3.12. The training set (379 wavelet transformed clear coat IR spectra and 1150 wavelet coefficients) was divided into 3 classes. A PC plot of the 379 training set samples and 1150 wavelet coefficients is shown in Figure 3.37.

Table 3.12. Training Set and Validation Set for Chrysler Plant Groups

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

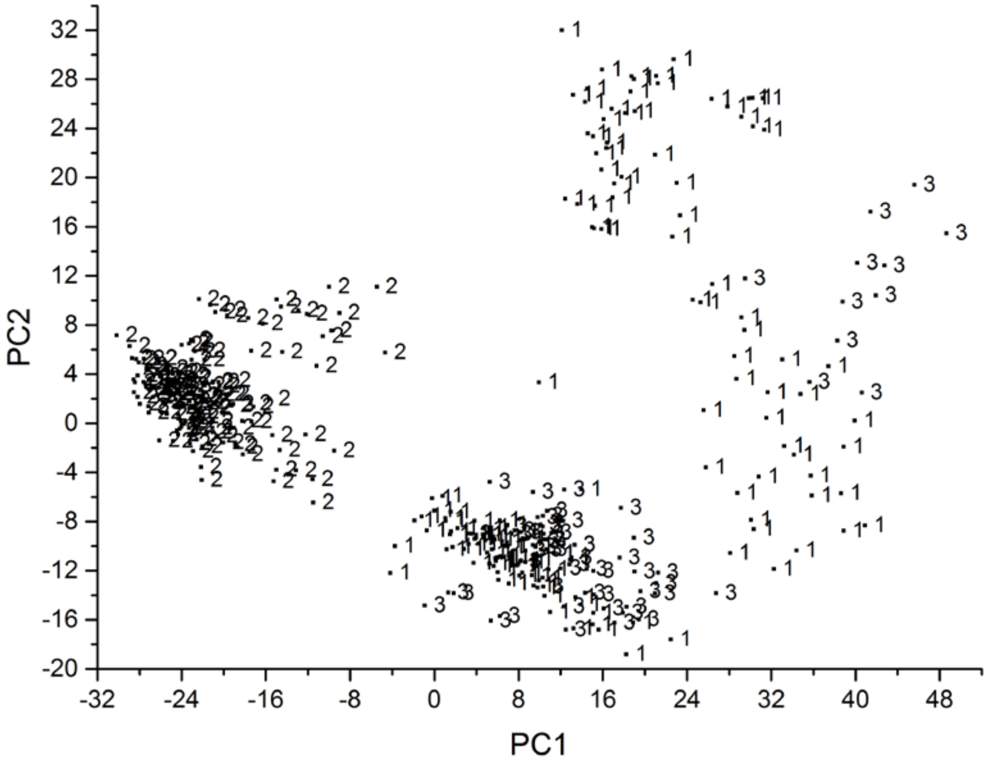


Figure 3.37. PC plot of the 379 training set samples and the 1150 wavelet coefficients of the clear coat layer.

The pattern recognition GA identified wavelet coefficients characteristic of plant group by sampling key feature subsets, scoring their PC plots and tracking those samples or groups that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the pattern recognition GA identified 9 wavelet coefficients whose PC plot (see Figure 3.38) showed clustering of the fused IR spectra on the basis of their plant group. The 42 validation set samples were then projected onto the PC plot (see Figure 3.39) define by the 379 training set samples and the 9 wavelet coefficients identified by the pattern recognition GA. Each validation set sample lies in a region of the PC plot with paint systems from the same automotive manufacturer.

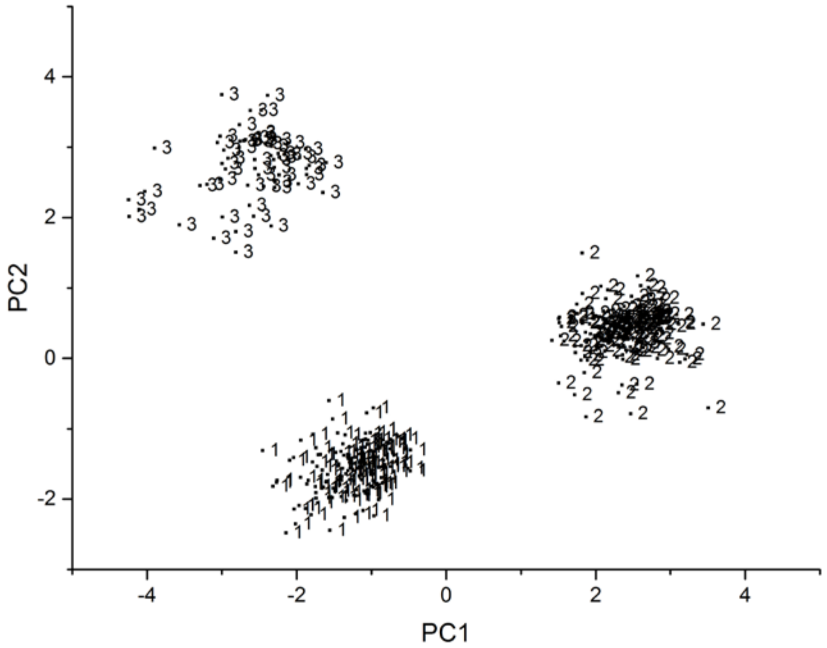


Figure 3.38. PC plot of the 379 training set samples and the 9 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Plant Group 11, 2 = Plant Group 12, 3 = Plant Group 13.

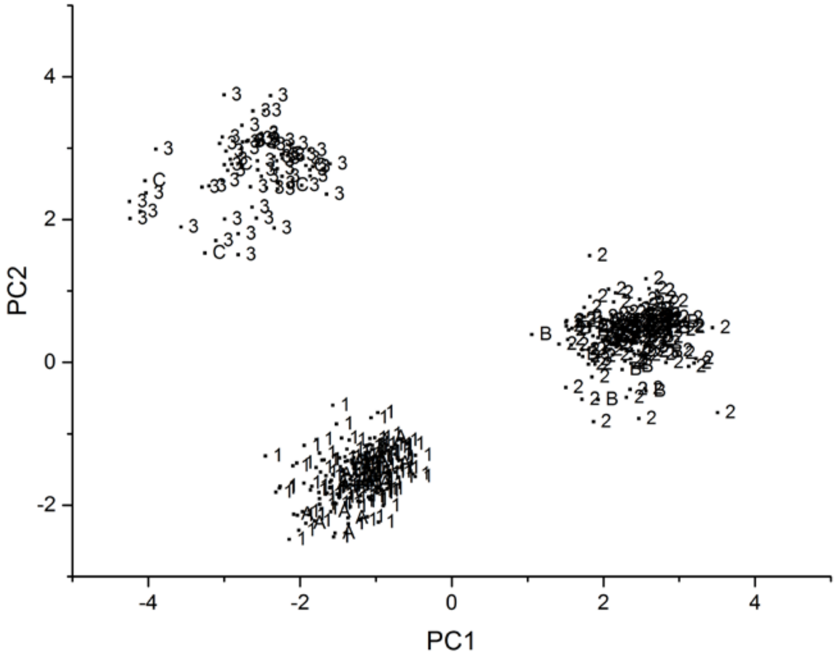


Figure 3.39. Projection of the 42 validation set samples onto the PC plot of the 379 training set samples and the 9 wavelet coefficients identified by the pattern recognition GA. Validation set: A = Plant Group 11, B = Plant Group 12, C = Plant Group 13.

For each plant group, a search prefilter was developed to discriminate automotive paint samples by assembly plant using the fingerprint region of the clear coat, surfacer-primer and e-coat layers. The pattern recognition GA identified specific coefficients in each paint layer correlated to the assembly plant of the vehicle from which the paint sample was obtained by sampling key feature subsets, scoring their PC plots and tracking those samples and/or classes (i.e., assembly plants or subplants) that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution.

Figure 3.40 shows a plot of the two largest principal components of the 156 samples comprising Plant Group 11 (see Table 3.13) and the 16 wavelet coefficients identified by the pattern recognition GA. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant, subplant, and plant subgroup is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 11 onto the PC plot showed that each projected validation set sample is located in a region of the PC plot with samples from the same assembly plant or subplant.

Table 3.13. Assembly Plants and Subplants Comprising Plant Group 11

Plant	Training	Validation
Saltillo, Toluca	47	10
Belvidere	33	3
subplant of Bramalea/Brampton	13	1
subplant of Toledo	14	1
subplant of Dodge Main	19	2
subplant of St. Louis	30	2

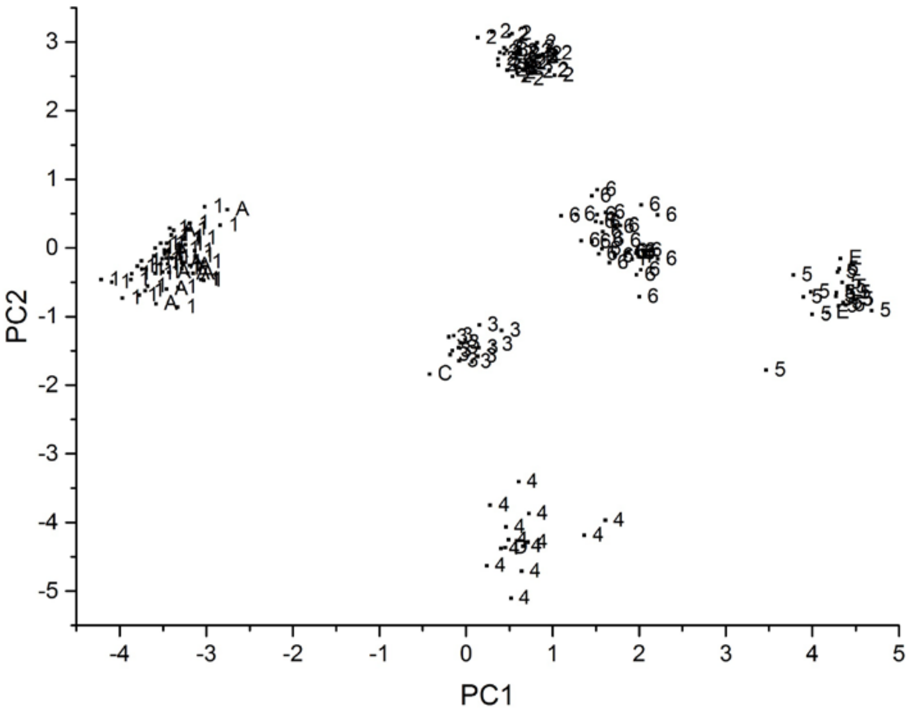


Figure 3.40. Validation set samples projected onto the PC plot of the data defined by the 156 paint samples of the training set for Plant Group 11 and the 16 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = plant subgroup containing Saltillo and Toluca, 2 = Belvidere, 3 = subplant of Bramalea/Brampton, 4 = subplant of Toledo, 5 = subplant of Dodge Main, 6 = subplant of St. Louis. Validation set: A = plant subgroup containing Saltillo and Toluca, B = Belvidere, C = subplant of Bramalea/Brampton, D = subplant of Toledo, E = subplant of Dodge Main, F = subplant of St. Louis.

Table 3.14 lists the assembly plants or subplants comprising Plant Group 12, which consists of one assembly plant (Bloomington), three subplants (Bramalea/Brampton, Dodge Main, Sterling Heights), and two plant subgroups. During the course of this analysis, it was necessary to merge different plants and subplants into a single class which is referred to as a plant subgroup due to the similarity of their spectra. One plant subgroup was comprised of the Windsor assembly plant and the St. Louis subplant, whereas the other was comprised of two subplants (one from Sterling Heights and the other from Toledo). Although the principal component analysis plot of clear coat IR spectra from the Sterling

Heights plant did not exhibit sample clustering, the corresponding principal component plot for the wavelet transformed concatenated (clear coat, surfacer-primer and e-coat layers) IR spectral data (see Figure 3.41) showed clustering which was correlated to production year. The average spectra of the clear coat, surfacer-primer and e-coat layers for the 2002-2006 Sterling Heights vehicles were superimposable when compared to the average spectra of the clear coat, surfacer-primer and e-coat layers for the vehicles from the Toledo subplant that was assigned to this plant group. For this reason, this subplant was combined with Toledo to form a plant subgroup. Principal component analysis plots of the fused IR spectra (clear coat, surfacer-primer and e-coat layers) from the other assembly plants or subplants comprising Plant Group 12 did not exhibit clustering as was the case with Sterling Heights.

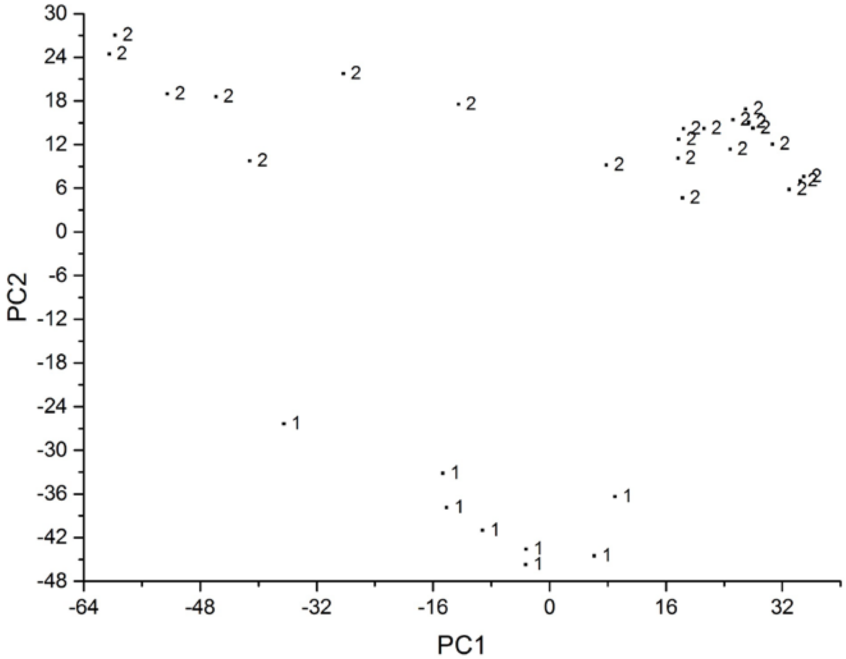


Figure 3.41. Plot of the two largest principal components of the wavelet transformed concatenated IR spectra (clear coat, surfacer-primer, and e-coat layers) from the Sterling Heights assembly plant. Two distinct clusters are present in the plot. 1 = 2000-2001, 2 = 2002-2006.

Table 3.14. Assembly Plant and Subplants Comprising Plant Group 12

Plant	Training	Validation
subplant of St. Louis, Windsor	47	10
Bloomington	33	3
subplant of Dodge Main	13	1
subplant of Bramalea/Brampton	14	1
subplant of Sterling Heights	19	2
subplant of Sterling Heights, subplant of Toledo	30	2

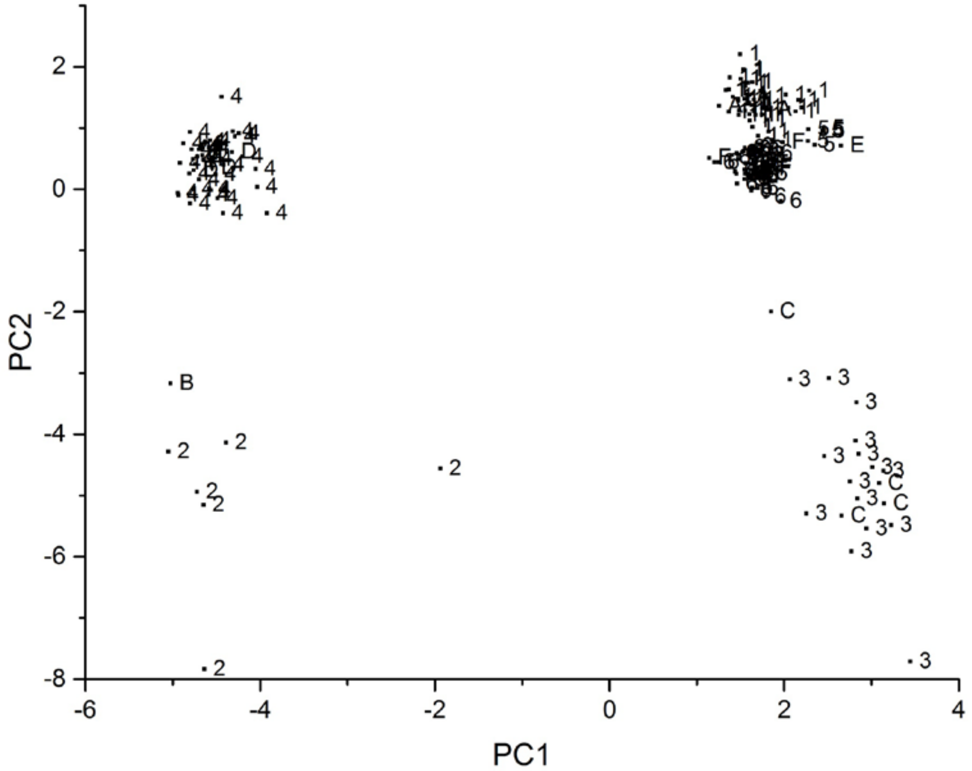


Figure 3.42. Validation set samples projected onto the PC plot of the data defined by the 157 paint samples of the training set comprising Plant Group 12 and the 22 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = plant subgroup containing Windsor and subplant of St. Louis, 2 = Bloomington, 3 = subplant of Dodge Main, 4 = subplant of Bramalea/Brampton, 5 = subplant of Sterling Heights, 6 = subplant of Sterling Heights and subplant of St. Louis. Validation set: A = plant subgroup containing Windsor and subplant of St. Louis, B = Bloomington, C = subplant of Dodge Main, D = subplant of Bramalea/Brampton, E = subplant of Sterling Heights, F = subplant of Sterling Heights and subplant of St. Louis.

Figure 3.42 shows a plot of the two largest principal components of the 157 samples comprising Plant Group 12 and the 22 wavelet coefficients identified by the pattern recognition GA. Each fused IR spectrum is represented as a point in the PC plot. Every assembly plant, subplant, and plant subgroup forms a distinct and separated cluster in the PC plot. Projecting the validation set samples assigned to Plant Group 12 onto the PC plot shows that each projected validation set sample is located in a region of the PC plot with samples from the same assembly plant, subplant, or plant subgroup.

Plant Group 13 consisted of two assembly plants: Jefferson North and Newark. The pattern recognition GA was not able to identify a set of coefficients from the wavelet transformed concatenated IR spectra that could differentiate Jefferson North from Newark. To understand the reason for this lack of success, principal component analysis was performed on both the Newark and Jefferson North assembly plants. Clustering correlated to the production year of the vehicle was observed for Newark (see Figure 3.43) from the concatenated IR data (clear coat, surfacer-primer and e-coat layers). For this reason, the Newark assembly plant was divided into two subplants. The Newark sample that was not a member of either cluster was tagged as an outlier and deleted from the analysis after comparing the spectra of each paint layer from this sample with the average spectra of each paint layer from each sample cluster identified in the PC plot of the wavelet transformed data. Table 3.15 describes the assembly plant and subplants comprising Plant Group 13.

Table 3.15. Assembly Plant and Subplants Comprising Plant Group 13

Plant	Training	Validation
Jefferson North	34	3
subplant of Newark	20	2
subplant of Newark	11	1

Figure 3.44 shows a plot of the two largest principal components of the 65 samples comprising Plant Group 13 and the 33 wavelet coefficients identified by the pattern recognition GA for the training set. The assembly plant and two subplants are well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 13 onto the PC plot showed that each projected validation set sample is located in a region of the PC plot with samples from the same assembly plant or subplant.

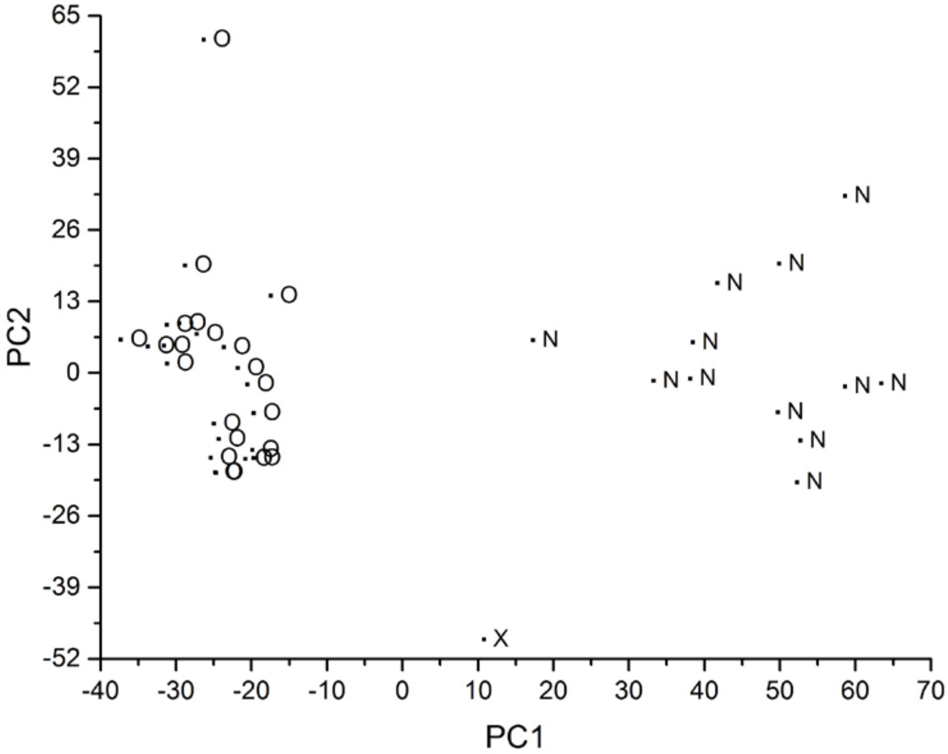


Figure 3.43. Plot of the two largest principal components of the wavelet transformed concatenated IR spectra (clear coat, surfacer-primer and e-coat layers) from the Newark assembly plant. Two distinct sample clusters and one outlier are present in the PC plot. O = 2000-2002, N = 2002-2006, and X = sample outlier.

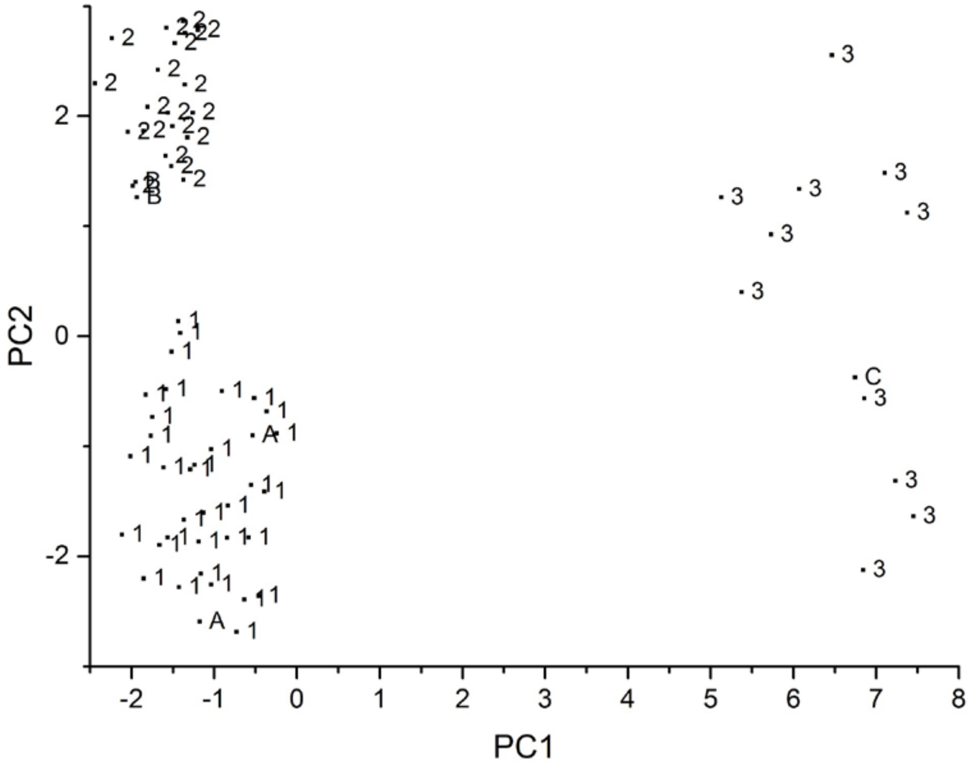
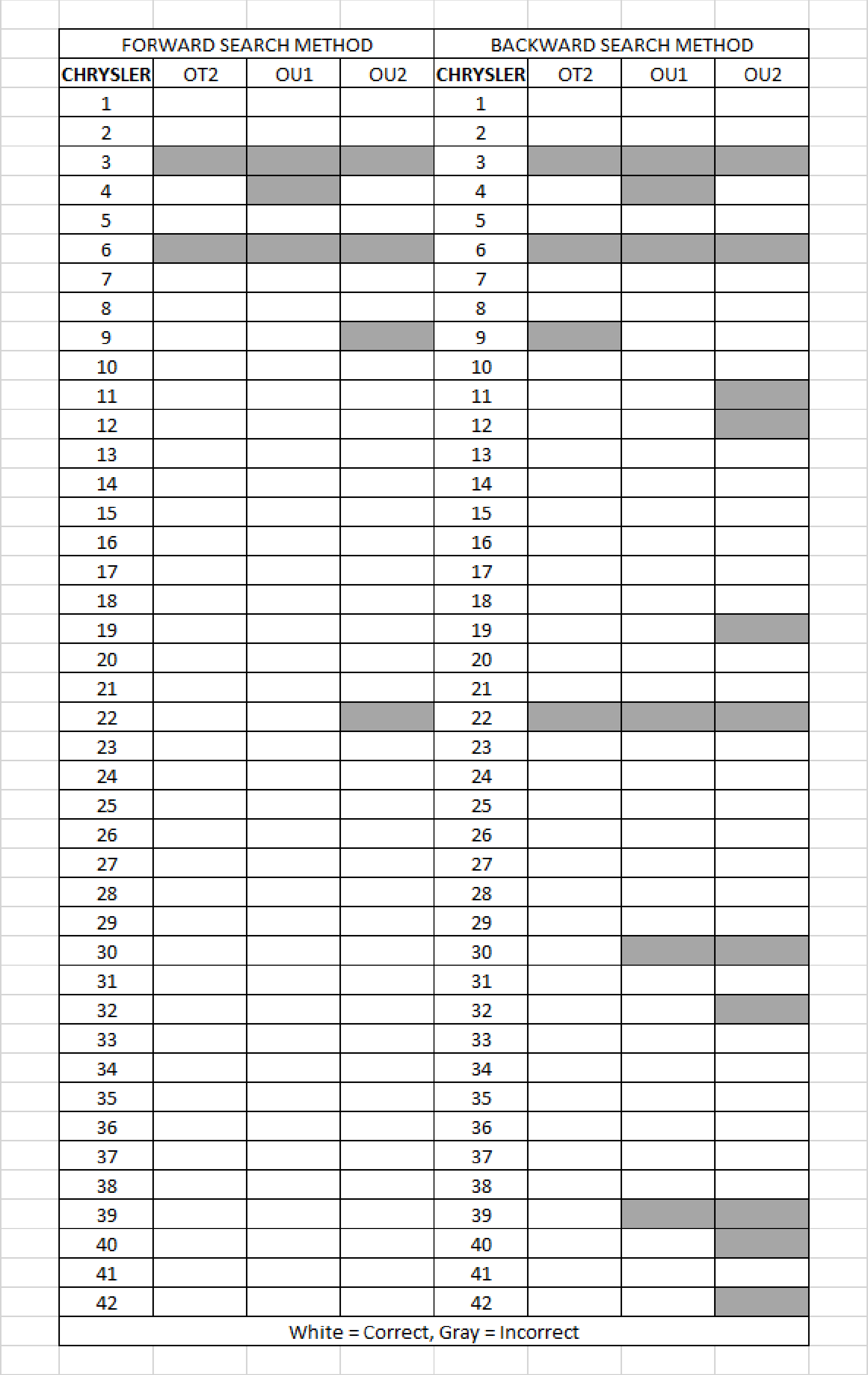


Figure 3.44. Validation set samples projected onto the PC plot of the data defined by the 65 paint samples of the training set comprising Plant Group 13 and the 33 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Jefferson North, 2 = Newark (2000-2002), and 3 = Newark (2002-2006). Validation set: A = Jefferson North, B = Newark (2000-2002), and C = Newark (2002-2006).

To extract information about the line and model of the vehicle, both a forward and backward search was performed using the cross correlation library search algorithm. Library search results are summarized in Table 3.16 for the 42 validation set samples using the prototype pattern recognition search system. Table 3.17 lists the results for the 42 validation set samples using OMNIC.

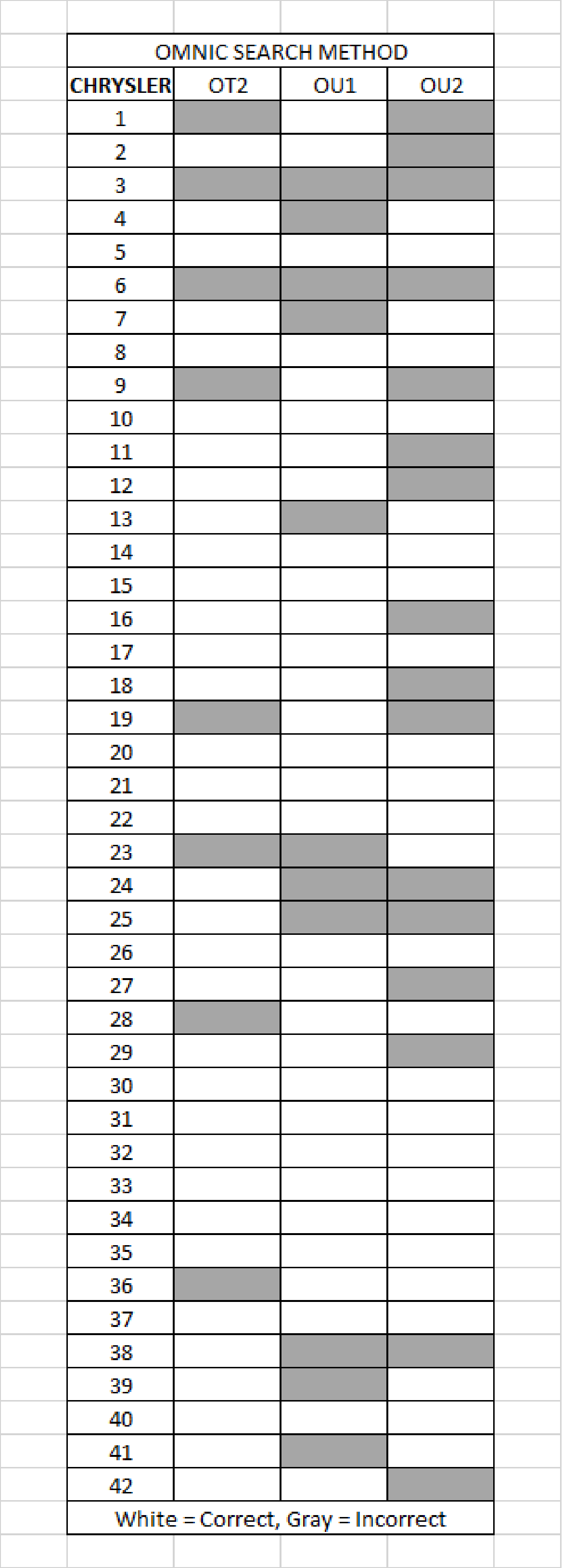
Both the forward and backward searches outperformed OMNIC. 37 of the 42 samples were correctly matched in all 3 layers by the forward search (which identifies spectra in the truncated PDQ spectral library most similar to the unknown), whereas only 29 samples were correctly matched in all 3 layers by the backward search (which computes a probability index as to the line and model of the unknown vehicle). Only 17 samples were correctly matched in all 3 layers by OMNIC. Of the 42 validation set samples, one was not represented in the 1183 spectral library as to the line and model of the vehicle.

Table 3.16. Prototype Cross-Correlation Library Search Results for Chrysler



FORWARD SEARCH METHOD				BACKWARD SEARCH METHOD			
CHRYSLER	OT2	OU1	OU2	**CHRYSLER**	OT2	OU1	OU2
1				1			
2				2			
3	▨	▨	▨	3	▨	▨	▨
4		▨		4		▨	
5				5			
6	▨	▨	▨	6	▨	▨	▨
7				7			
8				8			
9			▨	9	▨		
10				10			
11				11			▨
12				12			▨
13				13			
14				14			
15				15			
16				16			
17				17			
18				18			
19				19			▨
20				20			
21				21			
22			▨	22	▨	▨	▨
23				23			
24				24			
25				25			
26				26			
27				27			
28				28			
29				29			
30				30		▨	▨
31				31			
32				32			▨
33				33			
34				34			
35				35			
36				36			
37				37			
38				38			
39				39		▨	▨
40				40			▨
41				41			
42				42			▨
White = Correct, Gray = Incorrect							

Table 3.17. OMNIC Search Results for Chrysler



OMNIC SEARCH METHOD			
CHRYSLER	OT2	OU1	OU2
1	▨		▨
2			▨
3	▨	▨	▨
4		▨	
5			
6	▨	▨	▨
7		▨	
8			
9	▨		▨
10			
11			▨
12			▨
13		▨	
14			
15			
16			▨
17			
18			▨
19	▨		▨
20			
21			
22			
23	▨	▨	
24		▨	▨
25		▨	▨
26			
27			▨
28	▨		
29			▨
30			
31			
32			
33			
34			
35			
36	▨		
37			
38		▨	▨
39		▨	
40			
41		▨	
42			▨
White = Correct, Gray = Incorrect			

The forward search always yielded a reasonable spectral match, which was always superior to the match between the validation set sample and the actual line and model even for incorrectly matched samples. Although the backward search did not perform as well as the forward search, the backward search has the advantage of providing insight into how well the truncated library matched each validation set sample, rather than how well an individual validation set sample matched spectra in the truncated spectral library. For the clear coat layer, 40 samples were correctly matched by the forward search and 38 samples by the backward search with 38 samples assigned to the same line and model by both searches. For the surfacer-primer layer, 39 samples were correctly matched by the forward search and 36 by the backward search with 36 samples assigned to the same line and model by both searches. For the e-coat layer, 38 samples were correctly matched by the forward search and 31 by the backward search with 30 samples assigned to the same line and model by both searches. If a specific line and model was common to both hit lists (forward and backward cross-correlation searches), the assignment for the corresponding validation set sample was always correct. Validation set samples assigned to the same line and model by the forward and backward searches were always well represented in the library and correlated well on an individual basis to specific samples present in the library. For these validation set samples, one can have confidence in the accuracy of the match.

Library searches performed by OMNIC yielded hit quality index values for the five top matches that always exceeded 99% for correctly and incorrectly matched samples. (A sample was judged to be incorrectly matched if the line and model of the vehicle did not correspond to any of the samples in the hit-list.) The gap between the top hit and the thirtieth hit was typically less than 2%, and the gap between the top hit and the hundredth hit was typically less than 3%. This strongly suggests that a close visual inspection of the results from OMNIC searches will be necessary. For the clear coat layer, only 34 samples were correctly matched, whereas the number of samples correctly matched for the surfacer-primer and the e-coat layer by OMNIC were 31 and 26 respectively.

The hit quality index value was not a reliable indicator of the uniqueness of spectral matches in OMNIC searches for the clear coat, surfacer-primer or e-coat layers as incorrectly matched samples often had higher hit quality index values than correctly matched samples. For the forensic paint examiner, assessing the accuracy of a library search for an unknown automotive paint sample using a commercial library search algorithm such as OMNIC will be problematic.

3.3.4. Ford. The first step in the development of the search prefilters for Ford was to differentiate the manufacturer's paint systems by plant group. To determine the composition of each plant group, representative IR spectra of the clear coat layer were selected from each assembly plant. The Ford assembly plants (see Table 3.18) whose clear coat spectra exhibited a doublet for the carbonyl (which is indicative of the polymer acrylic melamine styrene polyurethane) as opposed to a singlet (acrylic melamine styrene**)** were flagged. These two assembly plants (St. Thomas-Talbotsville and Wixom) were assigned to Plant Group 24, whereas the other fifteen assembly plants whose clear coat spectra exhibited a singlet for the carbonyl (acrylic melamine styrene) were assigned to three other plant groups by cluster analysis.

Table 3.18. Distribution of Ford Assembly Plants into Plant Groups

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

Prior to cluster analysis, principal component analysis was performed on each Ford assembly plant to detect outliers and to assess class structure. For the assembly plants comprising Plant Group 24, we observed an outlier in the Wixom plant which we attributed to the carbonyl band of this sample being a singlet (i.e., acrylic melamine styrene), not a doublet (acrylic melamine styrene polyurethane). The formulation of the clear coat for this sample is different from the other Wixom assembly plant samples and this sample was therefore deleted from our pattern recognition analysis. For St. Thomas-Talbotsville (see Figure 3.45), clustering was correlated to production year. However, the larger sample cluster (2000-2006) in the principal component plot of the Saint Thomas-Talbotsville assembly plant consisted of spectra that had a singlet for the carbonyl (i.e., acrylic melamine styrene, not acrylic melamine styrene polyurethane). For St. Thomas-Talbotsville, the subplant corresponding to the acrylic melamine styrene formulation was removed from Plant Group 24 which we had restricted to clear coats prepared from acrylic melamine styrene polyurethane.

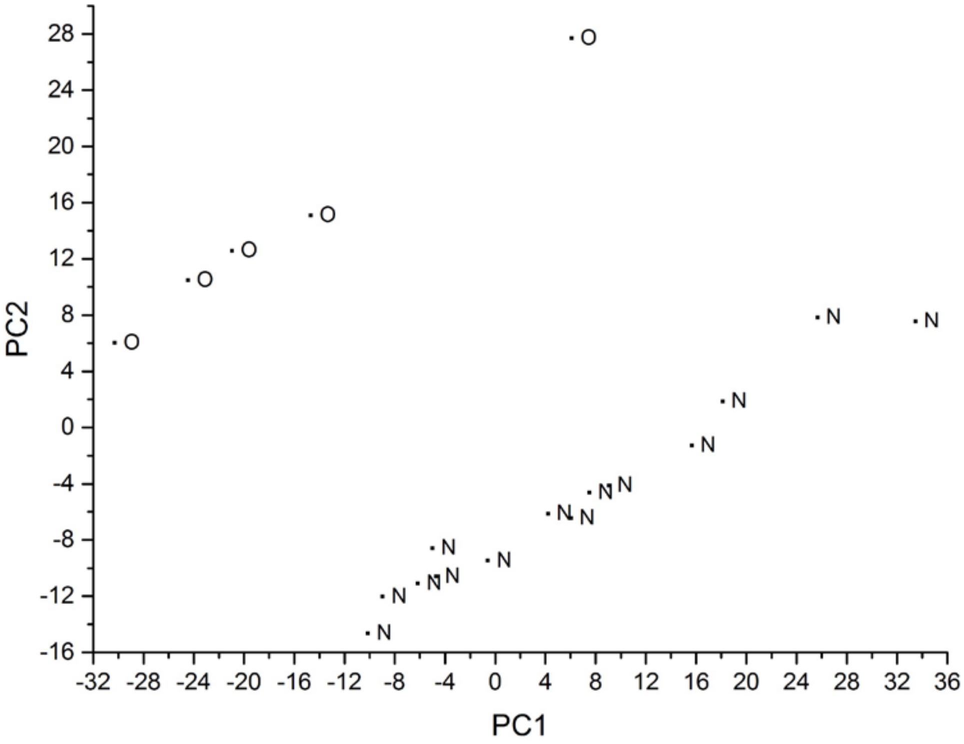


Figure 3.45. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Saint Thomas-Talbotsville assembly plant. Two distinct clusters are present in the plot. O = 2000, N = 2000-2006.

Clustering was also observed in the PC plots of six other Ford assembly plants (Dearborn, Louisville, Norfolk, Oakville, Twin Cities-St Paul, and Wayne). For Dearborn (Figure 3.46), Louisville (Figure 3.47), and Twin Cities-Saint Paul (Figure 3.48), clustering occurred on the basis of production year. For Norfolk (Figure 3.49), Oakville (Figure 3.50) and Wayne (Figure 3.51), clustering was correlated to the line and model of the vehicle.

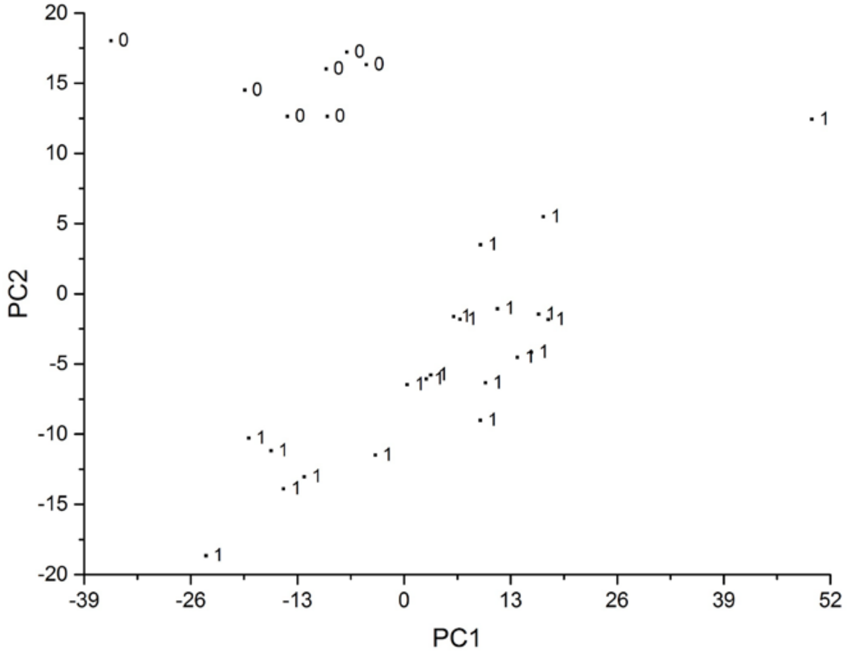


Figure 3.46. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Dearborn assembly plant. Two distinct clusters are present in the plot. 0 = 2000, 1 = 2001-2006.

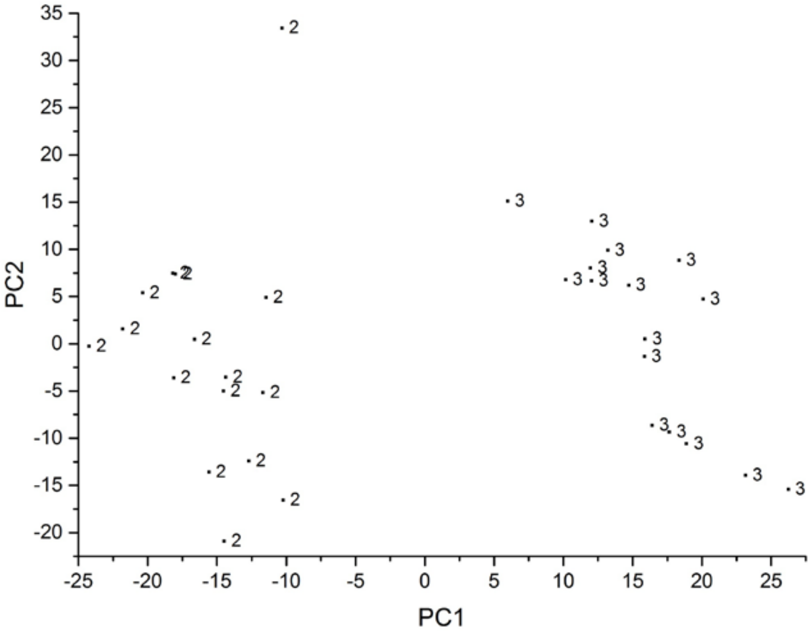


Figure 3.47. Plot of the two largest principal components of the clear coat IR spectra from the Louisville assembly plant. Two distinct clusters are present in the plot. 2= 2000-2002, 3 = 2003-2006.

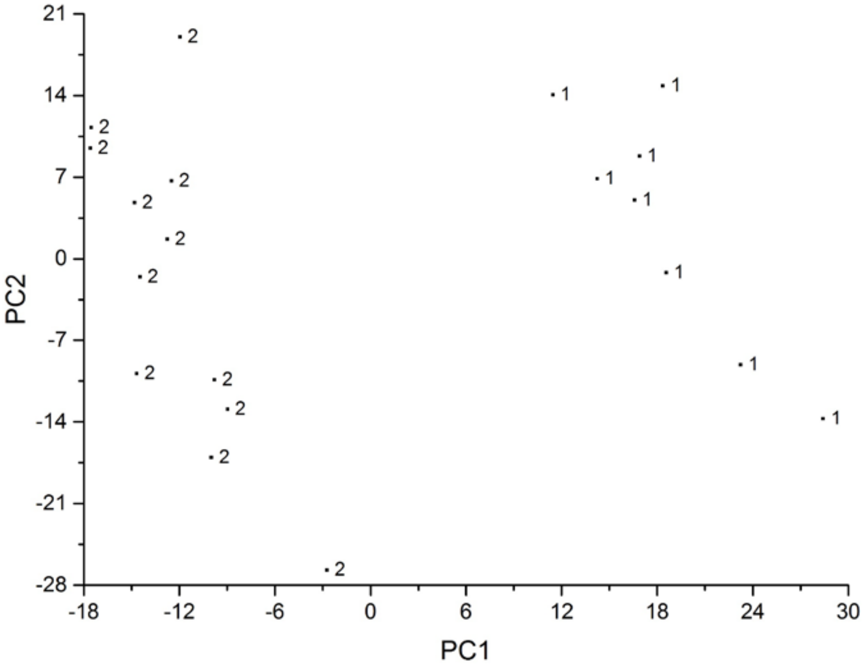


Figure 3.48. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Twin Cities-Saint Paul assembly plant. Two distinct clusters are present in the plot. 1 = 2000-2001, 2 = 2002-2006.

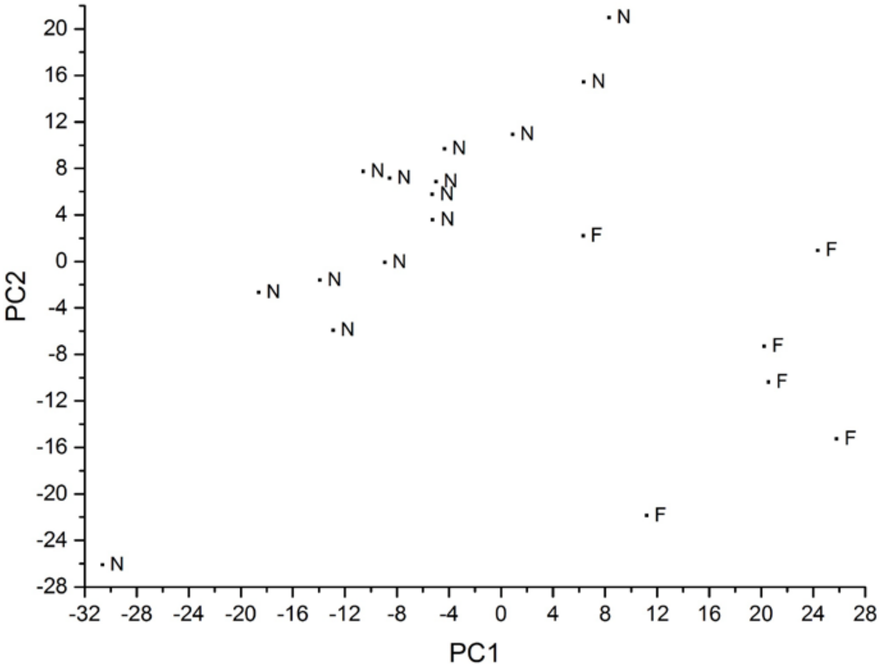


Figure 3.49. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Norfolk assembly plant. Two distinct clusters are present in the plot. N = all except 2000 F-series, F = 2000 F-series.

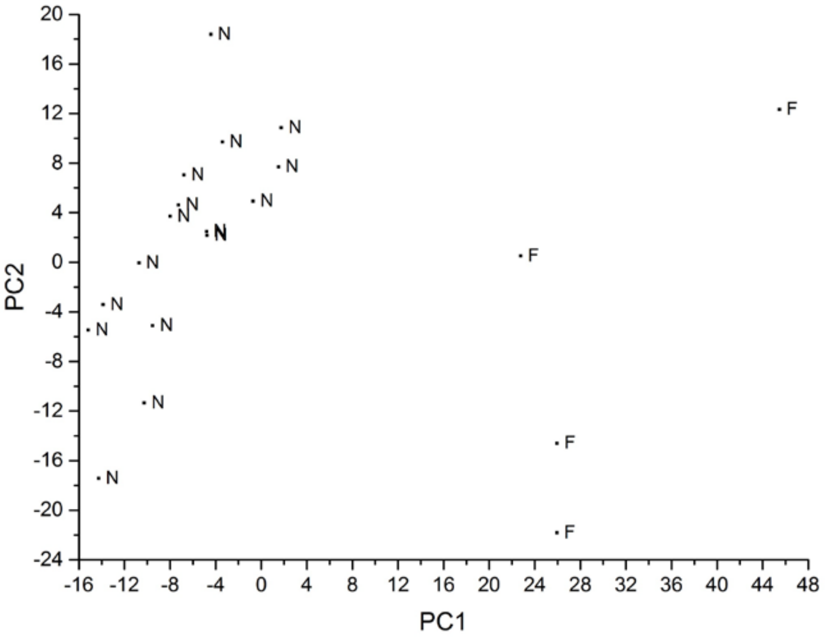


Figure 3.50. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Oakville assembly plant. Two distinct clusters are present in the plot. N = all except 2000-2001 F-series, F = 2000-2001 F-series.

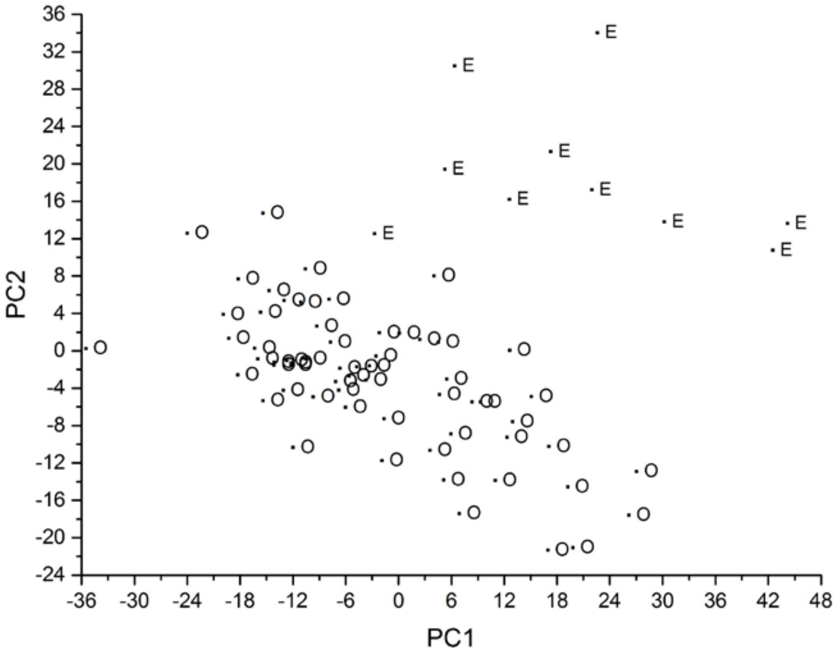


Figure 3.51. Plot of the two largest principal components of the wavelet transformed clear coat IR spectra from the Wayne assembly plant. Two distinct clusters are present in the plot. E = 2000-2001 Expedition and Navigator, O = all except 2000-2001 Expedition and Navigator.

To assign the remaining nine assembly plants (Atlanta, Chicago, Edison, Flat Rock, Hermosillo, Kansas City, Kentucky Truck, Lorain, Saint Louis) and thirteen subplants (Dearborn /two subplants, Louisville/two subplants, Norfolk/two subplants, Oakville/two subplants, St. Thomas-Talbotsville, Twin Cities – St. Paul/two subplants, Wayne/ two subplants) to specific plant groups, the average IR spectrum of the clear coat layer of each assembly plant or subplant was computed. Principal component analysis (Figure 3.52) and hierarchical clustering (Figure 3.53) were performed on the average IR spectra of each assembly plant or subplant.

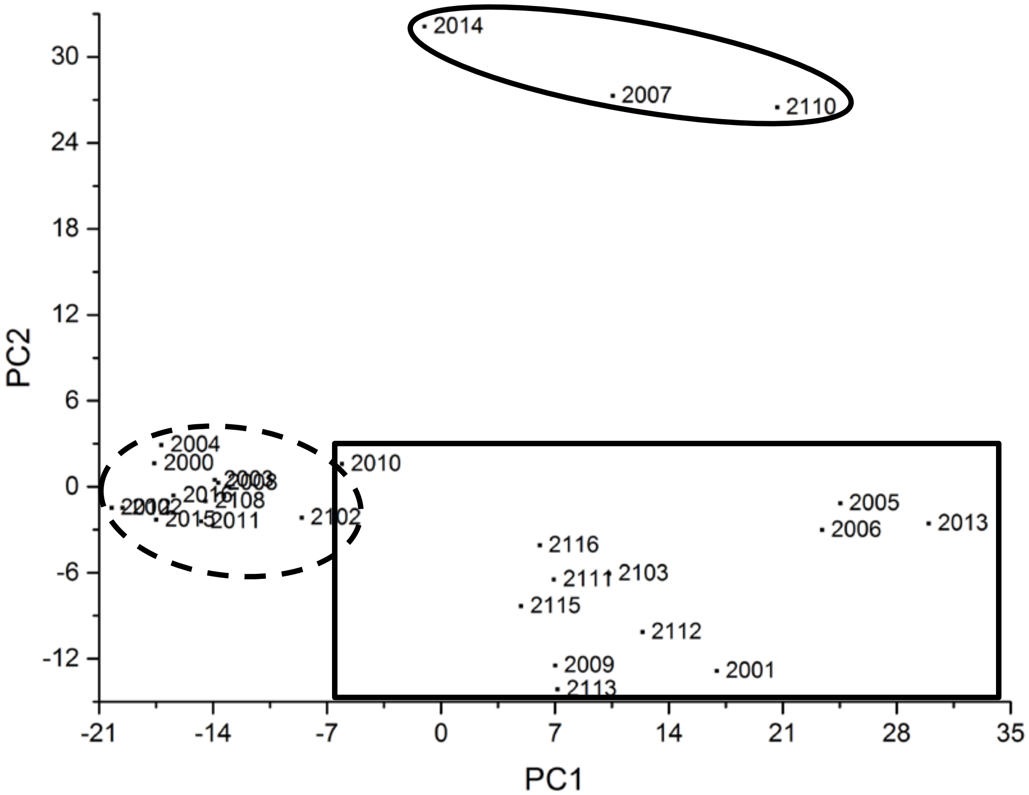


Figure 3.52. Principal component analysis of the average IR clear coat spectrum of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene.

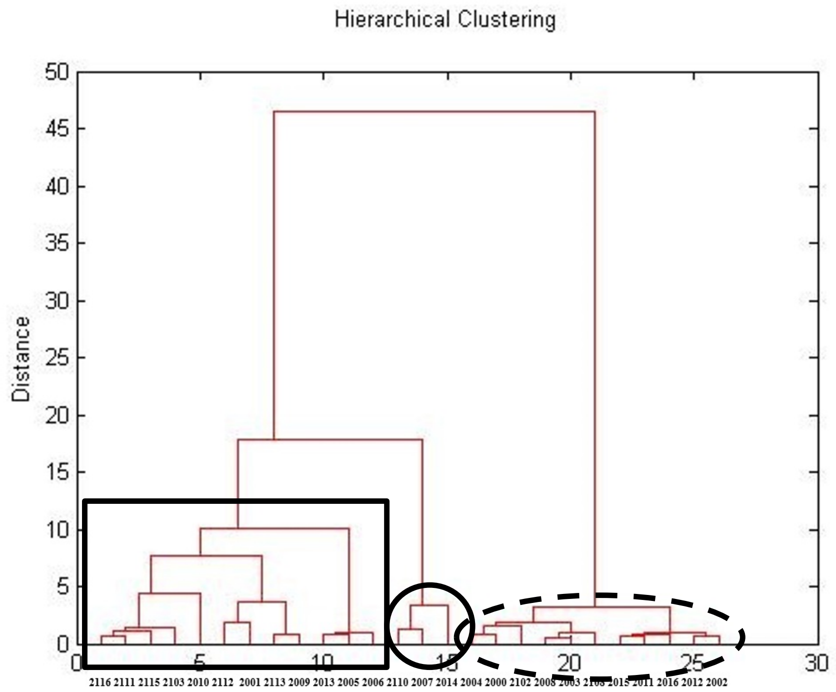


Figure 3.53. Hierarchical cluster analysis (Wards method) of the average IR spectrum (clear coats) of each assembly plant or subplant whose polymer formulation for the clear coat layer is acrylic melamine styrene.

From the results of the principal component analysis and hierarchical clustering, the nine assembly plants and the thirteen subplants were divided into three plant groups. Plant Group 21 consists of Atlanta, Chicago, Dearborn (subplant), Edison, Kentucky Truck, Norfolk (subplant), Oakville (subplant), Twin Cities-St. Paul (subplant), and Wayne (subplant). Plant Group 22 consists of Dearborn (subplant), Flat Rock, Hermosillo, Lorain, Norfolk (subplant), Oakville (subplant), St. Louis, Twin Cities-St. Paul (subplant), and Wayne (subplant). Plant Group 23 consists of Kansas City, Louisville (both subplants), and the St. Thomas-Talbotsville (subplant). For one of the two Louisville subplants (Plant ID 2010), principal component analysis and hierarchical clustering disagreed as to the plant group assignment for this subplant. The pattern recognition GA for the training set yielded consistent results when this subplant was assigned to Plant Group 23.

Having determined the membership of each plant group, developing a classifier to differentiate the plant groups was the next step. The number of samples in the training set and validation set for Plant Group are listed in Table 3.19. Figure 3.54 shows a PC plot of the two largest principal components of the 371 wavelet transformed clear coat IR spectra and 1150 wavelet coefficients comprising the training set for Plant Group. Each sample is represented as a point in the PC plot. The overlap between Plant Groups 22, 23, and 24 is evident.

Table 3.19. Training Set and Validation Set for Ford Plant Groups

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

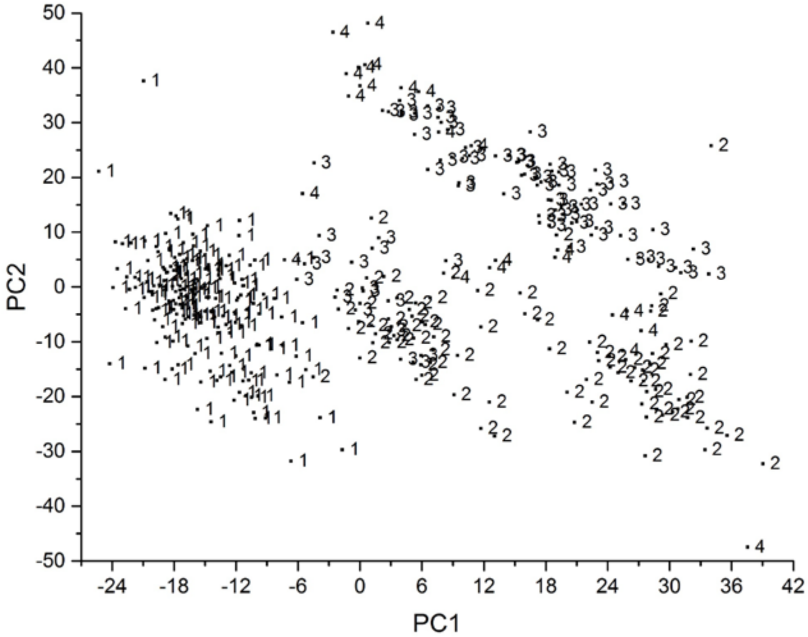


Figure 3.54. PC plot of the two largest principal components of the 371 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group. Each clear coat is represented as a point in the PC plot of the data. (1 = Plant Group 21, 2 = Plant Group 22, 3 = Plant Group 23, 4 = Plant Group 24).

The next step was feature selection. The goal was to identify wavelet coefficients characteristic of the profile of each plant group. The pattern recognition GA identified informative wavelet coefficients by sampling key feature subsets, scoring their PC plots, and tracking those plant groups/and or IR spectra that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the pattern recognition GA identified 37 wavelet coefficients whose PC plot showed clustering of the IR clear coat paint spectra on the basis of plant group (see Figure 3.55).

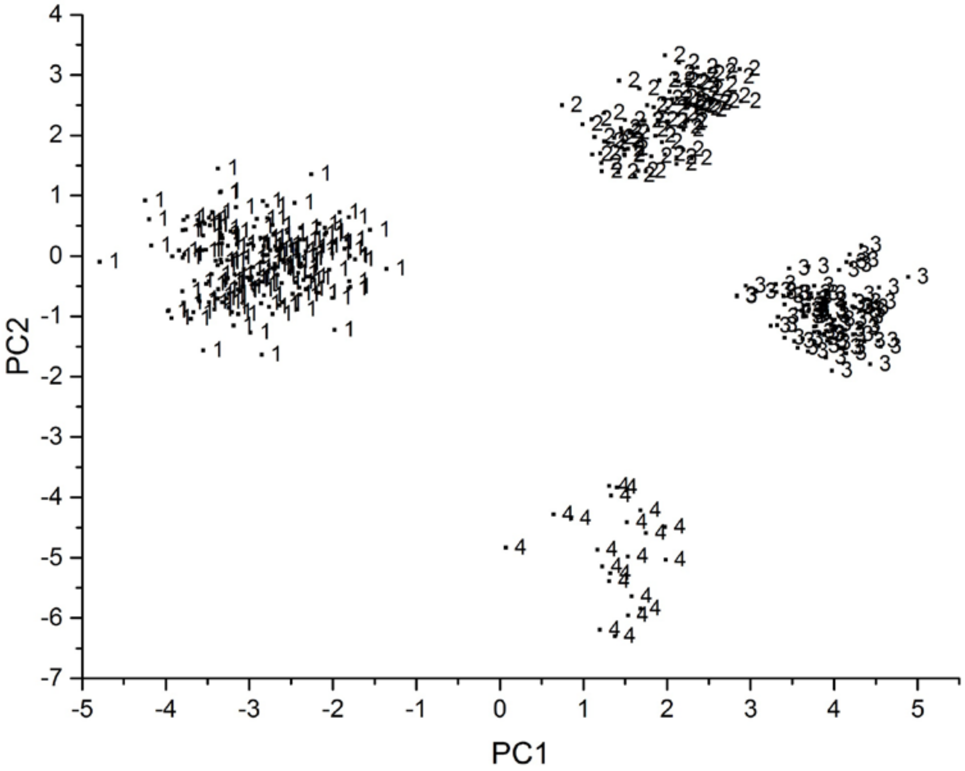


Figure 3.55. PC plot of the two largest principal components of the 371 training set samples and 37 wavelet coefficients identified by the pattern recognition GA (1 = Plant Group 21, 2 = Plant Group 22, 3 = Plant Group 23, 4 = Plant Group 24).

The predictive ability of the 37 wavelet coefficients identified by the pattern recognition GA was assessed using a validation set of 37 clear coat IR spectra identified as Fords by the manufacturer search prefilter. Clear coat IR spectra from the validation set were projected directly onto the principal component plot developed from the 371 IR spectra of the training set and the 37 wavelet coefficients identified by the pattern recognition GA. Figure 3.56 shows the projection of the validation set samples onto the principal component plot of the training set data. All validation set samples were correctly classified, i.e., they were located in a region of the principal component plot with samples from the same Plant Group.

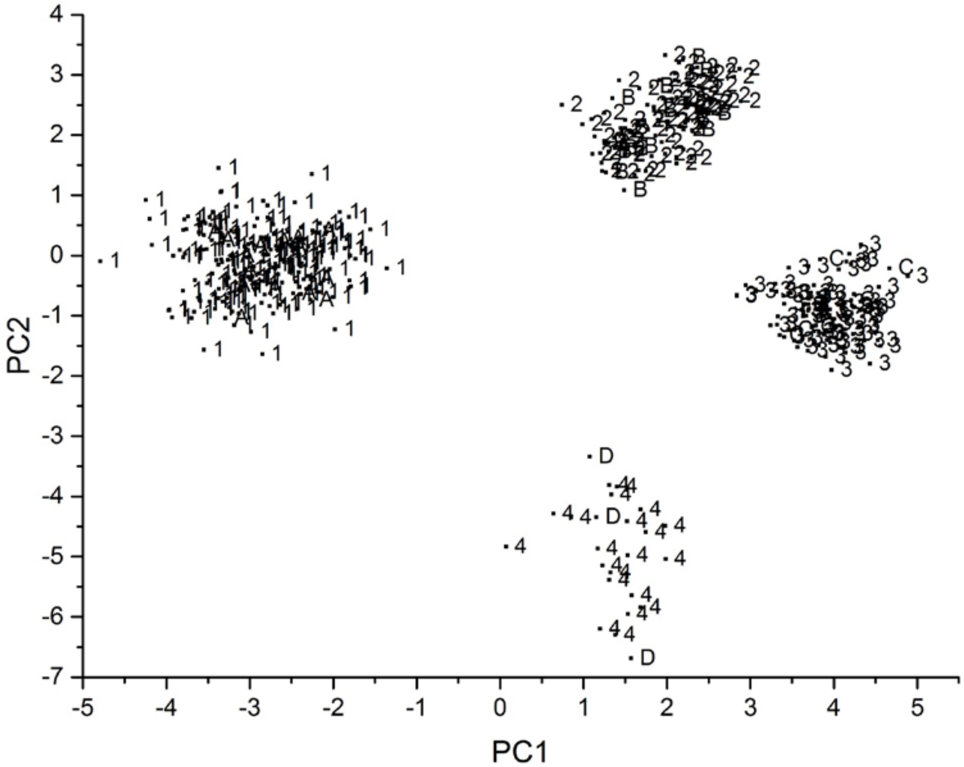


Figure 3.56. Validation set samples projected onto the PC plot of the data defined by the 371 wavelet transformed clear coat IR spectra of the training set and the 37 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Plant Group 21, 2 = Plant Group 22, 23 = Plant Group 23, 24 = Plant Group 24. Validation set: A = Plant Group 21, B = Plant Group 22, C = Plant Group 23, D = Plant Group 24.

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

Table 3.20 lists the assembly plants or subplants comprising Plant Group 21. Figure 3.57 shows a plot of the 180 samples comprising Plant Group 21 and the 17 wavelet coefficients identified by the pattern recognition GA for the training set. Each fused IR spectrum is represented as a point in the principal component plot. The validation set samples assigned to Plant Group 21 by the Plant Group search prefilter were projected onto this PC plot (see Figure 3.57). Only the Dearborn subplant could be differentiated from the other assembly plants and subplants for this plant group.

Table 3.20. Assembly Plants and Subplants Comprising Plant Group 21

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

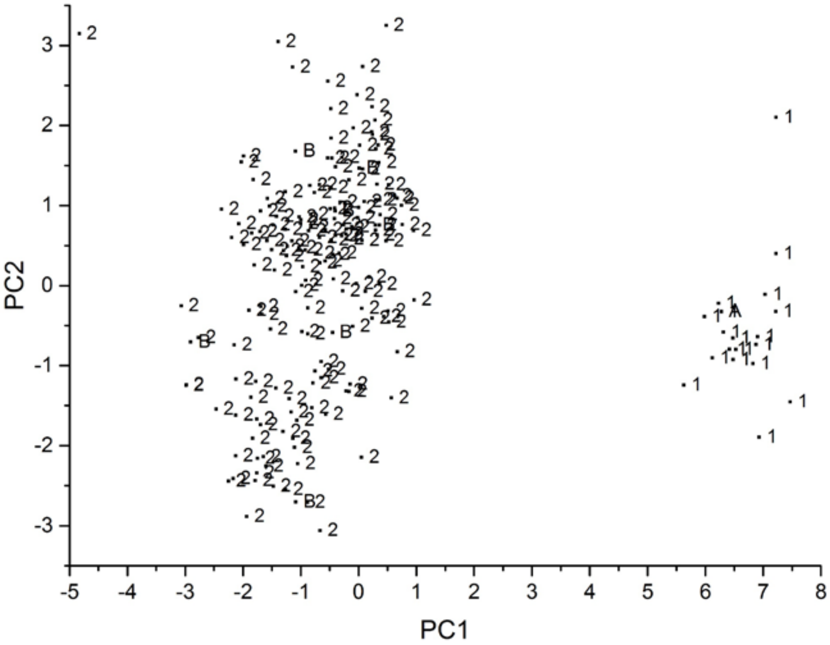


Figure 3.57. Validation set samples are projected onto the PC plot of the data defined by the 180 paint samples comprising Plant Group 21 (training set) and the 17 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = subplant of Dearborn, 2 = Atlanta, Chicago, Edison, Kentucky Truck, Norfolk (subplant), Oakville (subplant), Twin Cities-St. Paul (subplant), Wayne (subplant). Validation set: 1 = subplant of Dearborn, 2 = Atlanta, Chicago, Edison, Kentucky Truck, Norfolk (subplant), Oakville (subplant), Twin Cities-St. Paul (subplant), Wayne (subplant).

Table 3.21 lists the assembly plants or subplants comprising Plant Group 22. The pattern recognition GA divided the four assembly plants and six subplants into three groups designated as plant subgroups. During the course of variable selection, 6 training set samples were detected as outliers and subsequently were deleted from the analysis. Two were from Flat Rock, one was from Hermosillo, one was from Lorain, and two were from the Twin Cities-St. Paul subplant. These six samples differed from the other samples in their respective assembly plant or subplant in the spectra of the surfacer-primer and/or e-coat layers. Thus, the original training set of 85 wavelet preprocessed fused IR spectra was truncated to 79 spectra for discriminant development. It was also observed that IR spectra from the St. Louis assembly plant should be divided into two subplants on the basis of production year because of the surfacer-primer layer (see Figure 3.58).

Table 3.21. Assembly Plants and Subplants Comprising Plant Group 22

Plant Group	Training Set	Validation Set
Lorain, Oakville, subplant of St. Louis	14	2
subplant of Dearborn, subplant of Norfolk, subplant of Twin Cities-St. Paul, and subplant of Wayne	25	5
Flat Rock, Hermosillo, subplant of St. Louis	40	5

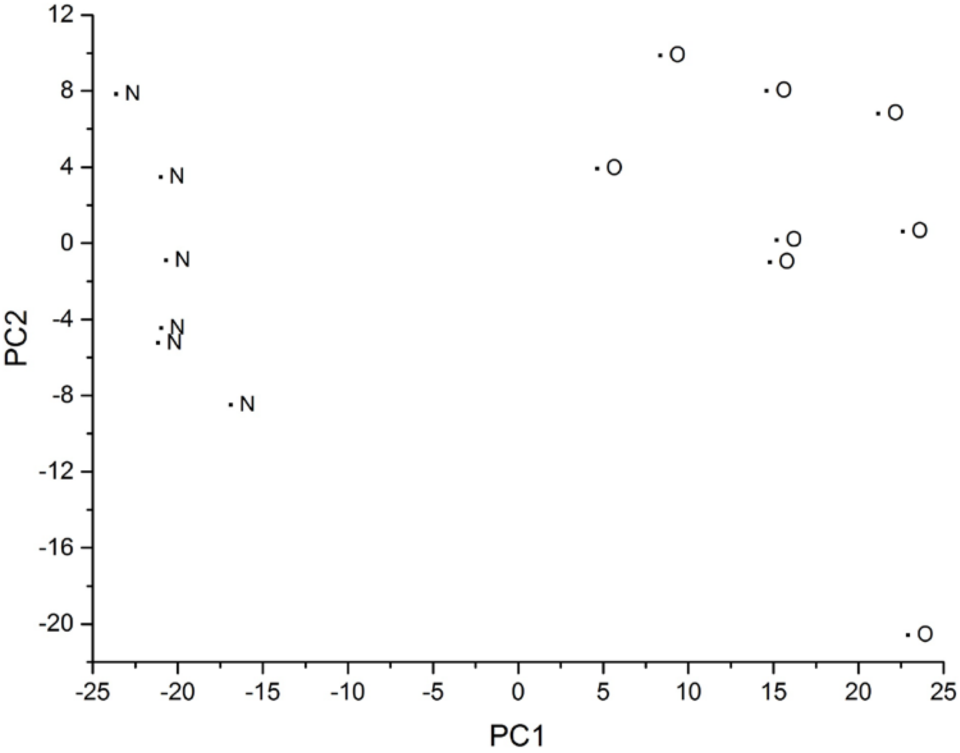


Figure 3.58. Plot of the two largest principal components of the surfacer-primer spectra from the St. Louis assembly plant. Two distinct clusters are evident in the plot. O = 2000-2002, N = 2003-2006.

Figure 3.59 shows a plot of the two largest principal components of the 79 samples comprising Plant Group 22 and the 24 wavelet coefficients identified by the pattern recognition GA. The validation set samples assigned to Plant Group 22 by the Plant Group search prefilter were projected onto this PC plot (see Figure 3.59) in regions occupied by samples from the same plant subgroup.

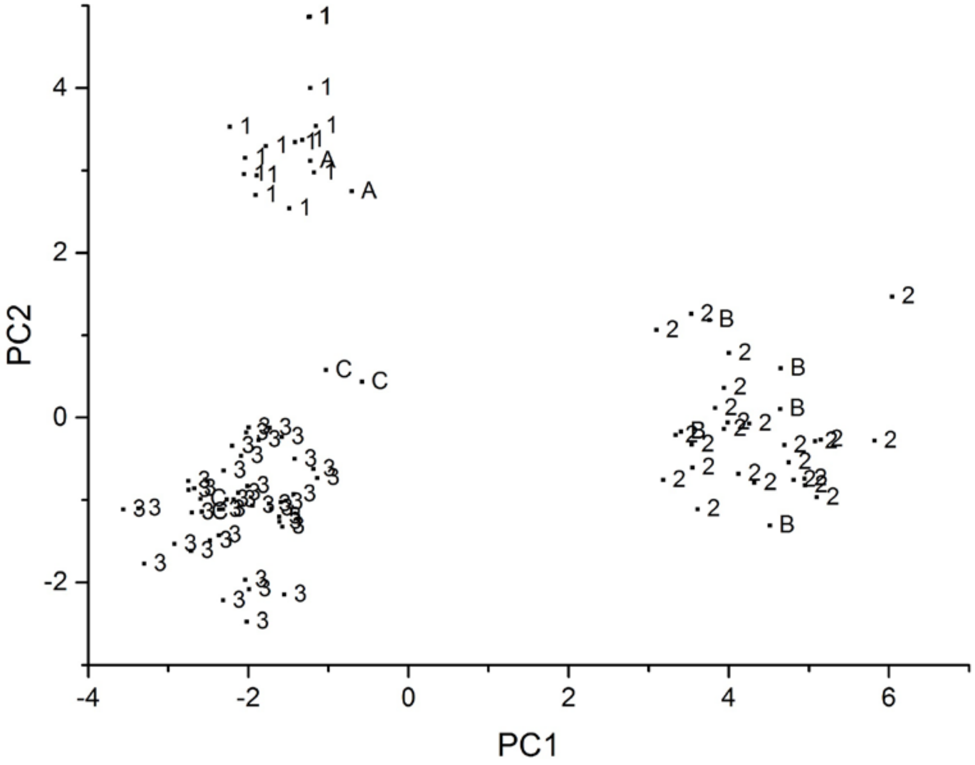


Figure 3.59. Validation set samples are projected onto the PC plot of the data defined by the 79 paint samples comprising Plant Group 22 (training set) and the 24 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Lorain, Oakville, and St. Louis (subplant), 2 = Dearborn (subplant), Norfolk (subplant), Twin Cities-St. Paul (subplant), and Wayne (subplant), 3 = Flat Rock, Hermosillo, St. Louis (subplant). Validation set: A = Lorain, Oakville, and St. Louis (subplant), B = Dearborn (subplant), Norfolk (subplant), Twin Cities-St. Paul (subplant), and Wayne (subplant), C = Flat Rock, Hermosillo, St. Louis (subplant).

Table 3.22 lists the assembly plants or subplants comprising Plant Group 23. Figure 3.60 shows a plot of the two largest principal components of the 80 samples of Plant Group 23 and the 38 wavelet coefficients identified by the pattern recognition GA. Each fused IR spectrum is represented as a point in the PC plot. During the course of the pattern recognition analysis, it was necessary to combine one of the two Louisville subplants with the Kansas City assembly plant due to the similarity of their spectra to form a plant subgroup. The validation set samples assigned to Plant Group 23 were projected onto the PC plot of the 80 samples and 38 wavelet coefficients. All validation set samples were located in a region of the plot with samples from the same assembly plant, subplant or plant subgroup (see Figure 3.60). Five of the original 85 training set samples in Plant Group 23 were found to be outliers and were discarded during the course of the pattern recognition analysis. Four were from Kansas City, and the other was from Louisville. These five samples differed from their assembly plant or subplant in the surfacer-primer and/or e-coat layers.

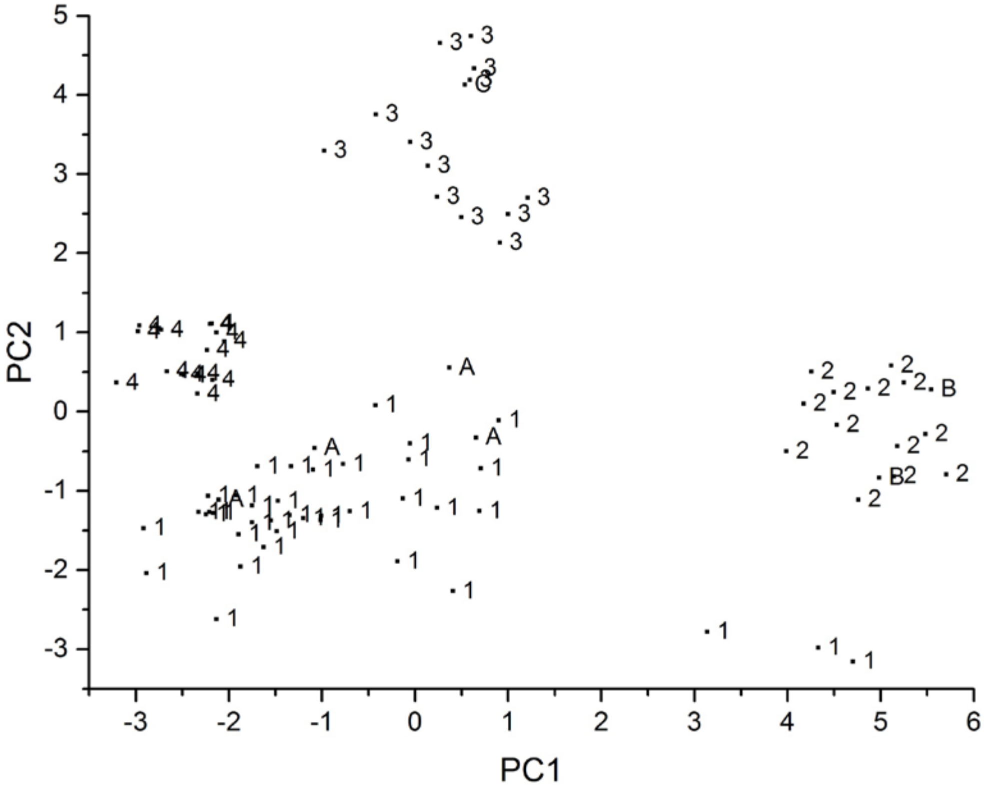


Figure 3.60. Validation set samples are projected onto the PC plot of the data defined by the 80 paint samples comprising Plant Group 23 (training set) and the 38 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Kansas City, Louisville (subplant), 2 = Louisville (subplant), 3 = St. Thomas-Talbotsville (subplant). Validation set: 1 = Kansas City, Louisville (subplant), 2 = Louisville (subplant), 3 = St. Thomas-Talbotsville (subplant).

Table 3.22. Assembly Plants and Subplants Comprising Plant Group 23

Plant Group	Training Set	Validation Set
Kansas City, subplant of Louisville	54	4
subplant of Louisville	13	2
subplant of St. Thomas-Talbotsville	13	1

Table 3.23 lists the subplants comprising Plant Group 24. For Plant Group 24, the fused spectra are divided among three subplants. Previously, St. Thomas-Talbotsville was divided into two subplants on the basis of the carbonyl of the clear coat layer (acrylic melamine styrene polyurethane versus acrylic melamine styrene) with the subplant corresponding to acrylic melamine styrene transferred to Plant Group 23. When developing an assembly plant search prefilter for Plant Group 24, we came to the conclusion that spectra from the Wixom assembly plant should be divided into two subplants on the basis of production year because of the surfacer-primer layer (see Figure 3.61).

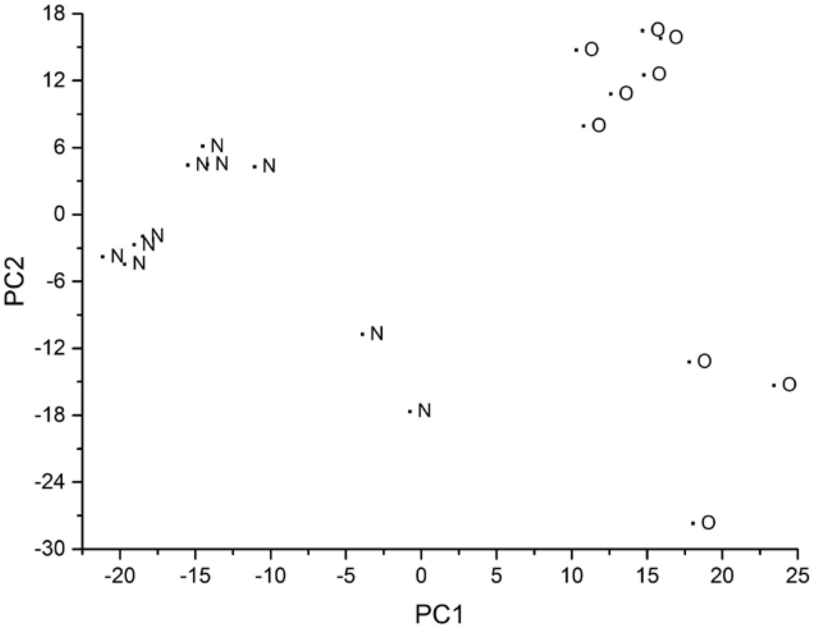


Figure 3.61. Plot of the two largest principal components of the surfacer-primer spectra from the Wixom assembly plant. Two distinct clusters are evident in the plot. O = 2000-2001, N = 2002-2006.

Table 3.23. Assembly Plants and Subplants Comprising Plant Group 24

Plant Group	Training Set	Validation Set
subplant of Wixom	9	1
subplant of St. Thomas-Talbotsville	5	0
subplant of Wixom	7	2

Figure 3.62 shows a plot of the two largest principal components of the 21 samples of Plant Group 24 and the 2 wavelet coefficients identified by the pattern recognition GA. All three subplants were well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 24 onto the PC plot (see Figure 3.62) showed that each projected validation set sample was located in a region of the plot containing samples from the same subplant. As only two wavelet coefficients were necessary for the development of this search prefilter, the classification problem underlying the development of the search prefilter is simple.

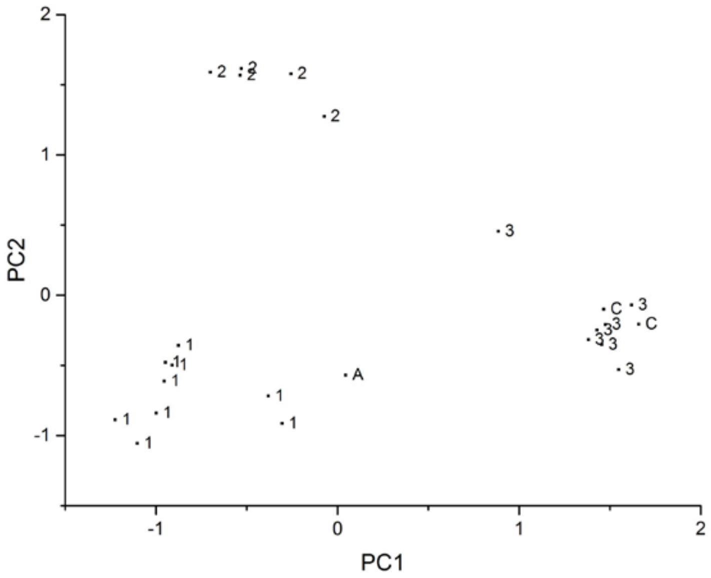
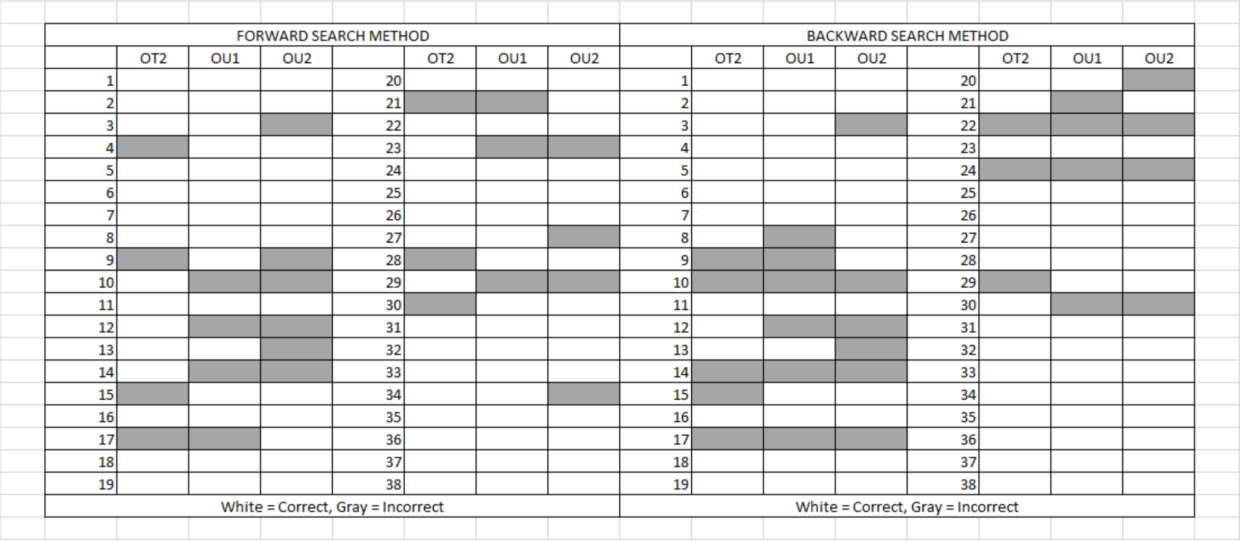


Figure 3.62. Validation set samples are projected onto the PC plot of the data defined by the 21 paint samples comprising Plant Group 24 (training set) and the 2 wavelet coefficients identified by the pattern recognition GA. Training set: 1 = Wixom (subplant), 2 = St. Thomas-Talbotsville (subplant), 3 = Wixom (subplant). Validation set: A = Wixom (subplant), B = St. Thomas-Talbotsville (subplant), C = Wixom (subplant).

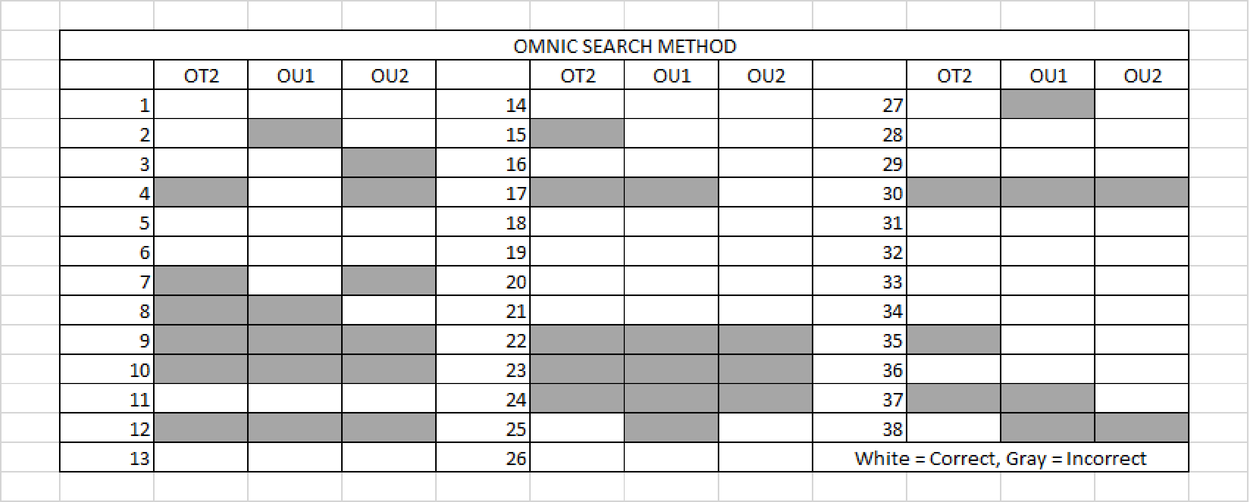
To extract information about the line and model of the vehicle, the cross correlation library search algorithms was applied to the truncated PDQ library spectra using both modes (i.e., forward and backward searching). Library search results are summarized for the 38 validation set samples in Table 3.24 for the prototype pattern recognition library search system and in Table 3.25 for OMNIC. Both the forward and backward search of the cross correlation library search algorithm outperformed OMNIC. For the clear coat layer, 31 samples were correctly matched by the forward search, 30 samples by the backward search, and 28 samples were assigned to the same line and model by both searches. For the surfacer-primer, 31 samples were correctly matched by the forward search, 28 by the backward search, and 28 samples were assigned to the same line and model by both searches. For the e-coat, 28 samples were correctly matched by the forward search, 28 by the backward search, and 25 samples were assigned to the same line and model by both searches. When a specific line and model was common to both hit lists (i.e., forward and backward cross-correlation library searches), the assignment by the algorithm for the corresponding validation set sample was always correct. Validation set samples assigned to the same line and model by the forward and backward searches were always well represented in the library and correlated well on an individual basis to specific samples present in the truncated spectral libraries. For these validation set samples, one can have confidence in the accuracy of the match as both the forward and backward searches produced the same result.

Table 3.24. Prototype Cross-Correlation Library Search Results for Ford



FORWARD SEARCH METHOD								BACKWARD SEARCH METHOD							
	OT2	OU1	OU2		OT2	OU1	OU2		OT2	OU1	OU2		OT2	OU1	OU2
1				20				1				20			▨
2				21	▨	▨		2				21		▨	
3			▨	22				3			▨	22	▨	▨	▨
4	▨			23		▨	▨	4				23			
5				24				5				24	▨	▨	▨
6				25				6				25			
7				26				7				26			
8				27			▨	8		▨		27			
9	▨		▨	28	▨			9	▨	▨		28			
10		▨	▨	29		▨	▨	10	▨	▨	▨	29	▨		
11				30	▨			11				30		▨	▨
12		▨	▨	31				12		▨	▨	31			
13			▨	32				13			▨	32			
14		▨	▨	33				14	▨	▨	▨	33			
15	▨			34			▨	15	▨			34			
16				35				16				35			
17	▨	▨		36				17	▨	▨	▨	36			
18				37				18				37			
19				38				19				38			
White = Correct, Gray = Incorrect								White = Correct, Gray = Incorrect							

Table 3.25. OMNIC Library Search Results for Ford



OMNIC SEARCH METHOD											
	OT2	OU1	OU2		OT2	OU1	OU2		OT2	OU1	OU2
1				14				27		▨	
2		▨		15	▨			28			
3			▨	16				29			
4	▨		▨	17	▨	▨		30	▨	▨	▨
5				18				31			
6				19				32			
7	▨		▨	20				33			
8	▨	▨		21				34			
9	▨	▨	▨	22	▨	▨	▨	35	▨		
10	▨	▨	▨	23	▨	▨	▨	36			
11				24	▨	▨	▨	37	▨	▨	
12	▨	▨	▨	25		▨		38		▨	▨
13				26				White = Correct, Gray = Incorrect			

Library searches performed by OMNIC yielded hit quality index values for the five top matches that typically exceeded 99% for correctly and incorrectly matched samples. (A sample was considered incorrectly matched if the line and model of the vehicle did not correspond to any sample in the hit-list.) For these searches, the gap between the top hit and the twentieth hit was less than 2%, and the gap between the top hit and the hundredth hit was less than 3%. For the clear coat layer, 24 samples were correctly matched, whereas the number of samples correctly matched for the surfacer-primer and the e-coat layers by OMNIC were 24 and 27 respectively.

Neither the value of the hit quality index nor the size of the gap could serve as a reliable indicator of the uniqueness of the matches in these OMNIC searches. A value of the hit quality index by itself was not indicative of the quality of a spectral match for the clear coat, surfacer-primer or e-coat layers as incorrectly matched samples often had higher values than correctly matched samples. For the forensic paint examiner, assessing the accuracy of a library search for an unknown paint sample using a commercial library search algorithm such as OMNIC is problematic using these figures of merit.

3.4. CONCLUSION

Search prefilters were developed to differentiate automotive paint samples by manufacturer (Chrysler, General Motors, and Ford) using the clear coat, surfacer-primer, and e-coat layers. For each automotive manufacturer, search prefilters were developed to identify the assembly plant of the vehicle from the manufacturer's paint system. First, the assembly plants were divided into groups of assembly plants based upon cluster analysis of the fingerprint region of the clear coat layer. Second, each plant group was divided into its respective assembly plants using the clear coat, surfacer-primer and e-coat layers. The search prefilter system categorized each unknown paint sample by identifying successively smaller sets of vehicles to which an unknown paint sample was assigned. The search prefilters have the potential to facilitate spectral library searching as the size of the library is truncated to those spectra of automotive paint samples obtained from the same assembly plant as that of the unknown.

The General Motors and Chrysler search prefilters were able to identify the assembly plant or subplant of the vehicle. This, in turn, allows for the line and model of the vehicle to be identified. As some lines and models are produced by more than one assembly plant, identifying the specific assembly plant that reduces the size of the PDQ library to a smaller number of IR spectra than a search prefilter based on identifying the specific model and line of the vehicle.

The cross correlation library searching algorithm in conjunction with the search prefilters outperformed OMNIC and reduced the hit-list to five samples in each search. Furthermore, the accuracy of the hit-list could be assessed as samples assigned to the same line and model by the forward and backward search were always correctly matched. This is a potentially significant development that can enhance current approaches to data interpretation of forensic automotive paint examinations and aid in evidential significance assessment, both at the investigative lead stage and at the courtroom testimony stage.

REFERENCES

3-1. G. Fettis (Ed.), “Automotive Paints and Coatings”. VCH Publications, New York, NY, 1995.

3-2. S. Ryland, T. Jegovich, K. P. Kirkbride, “Current Trends in Forensic Paint Examination”, Forensic Sci. Rev., 2006, 18, 97-117.

3-3. K. Flynn, R. O’Leary, C. Lennard, C. Roux, B. J. Reedy, “Forensic Applications of Infrared Chemical Imaging: Multilayered Paint Chips”, J. Forensic Sci., 2005, 50, 832-841.

3-4. P. G. Rogers, R. Cameron, N. S. Cartwright, W. H. Clark, J. S. Deak, “The Classification of Automotive Paint by Diamond Windows Infrared Spectrophotometry-Part I: Automotive Topcoats and Undercoats”, E. W. W. Norman, Can. Soc. Forensic Sci., 1976, 9, 1-14.

3-5. P. G. Rogers, R. Cameron, N. S. Cartwright, W. H. Clark, J. S. Deak, E. W. W. Norman, “The Classification of Automotive Paint by Diamond Windows Infrared Spectrophotometry-Part II: Automotive Topcoats and Undercoats”, Can. Soc. Forensic Sci., 1976, 9, 49-68.

3-6. N. S. Cartwright, L. J. Cartwright, E. W. W. Norman, R. Cameron, W. H. Clark, D. A. MacDougal, “A Computerized System for the Identification of Suspect Vehicles Involved in Hit and Run Accidents”, Can. Soc. Forensic Sci. J., 1982, 15, 105-115.

3-7. J. L. Buckle, D. A. MacDougal, R. R. Grant, “PDQ-Paint Data Queries: The History and Technology Behind the Development of the Royal Canadian Mounted Police Forensic Science Laboratory Services Automotive Paint Database”, Can. Soc. Forensic Sci. J., 1997, 30, 199-212.

3-8. S. R. Lowry, D. A. Huppler, C. R. Anderson, “Data Base Development and Search Algorithms for Automated Infrared Spectral Identification”, J. Chem. Inf. Computer Sci., 1985, 25, 235-241.

3-9. B. K. Lavine, A. J. Moores, “Genetic Algorithms for Pattern Recognition Analysis and Fusion of Sensor Data”, In: K. Siddiqui and D. Eastwood (Eds.), Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, Proceedings of SPIES, 1999, 103-112.

3-10. B. K. Lavine, A. Fasasi, N. Mirjankar, M. Sandercock, "Search Prefilters to Assist in Library Searching of Infrared Spectra of Automotive Clear Coats", Talanta, 2015, 120, 182-190.

3-11. B. K. Lavine, A. Fasasi, N. Mirjankar, C. G. White, "Search Prefilters for Library Matching of Infrared Spectra in the PDQ Database using the Autocorrelation Transformation," Microchem. J., 2014, 113, 30–35.

3-12. B. K. Lavine, A. Fasasi, N. Mirjankar, M. Sandercock, S. D. Brown, "Search Prefilters for Mid-IR Spectra of Clear Coat Automotive Paint Smears Using Stacked and Linear Classifiers", J. Chem., 2014, 28, 385-394.

3-13. B. K. Lavine, N. Mirjankar, S. Delwiche, "Classification of the Waxy Condition of Durum Wheat by Near Infrared Reflectance Spectroscopy using Wavelets and a Genetic Algorithm", Microchem. J., 2014, 117, 178-182.

3-14. B. K. Lavine, K. Nuguru, N. Mirjankar, J. Workman, "Pattern Recognition Assisted Infrared Library Searching", Appl. Spec., 2012, 66, 917-925.

3-15. B. K. Lavine, K. Nuguru, N. Mirjankar, J. Workman, "Development of Carboxylic Acid Search Prefilters for Spectral Library Matching", Microchem. J., 2012, 103, 21-36.

3-16. B. K. Lavine, N. Mirjankar, S. Ryland, M. Sandercock, "Wavelets and Genetic Algorithms Applied to Search Prefilters for Spectral Library Matching in Forensics", Talanta, 2011, 87, 46-52.

3-17. B. K. Lavine, D. Brzozowski, A. J. Moores, C. E. Davidson, H. T. Mayfield, "Genetic Algorithm for Fuel Spill Identification", Anal. Chim. Acta, 2001, 437, 233-246.

3-18. G. A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F. K. Tadjimukhamedov, B. K. Lavine, N. Mirjankar, "Pattern Recognition Analysis of Differential Mobility Spectra with Classification by Chemical Family," Anal. Chim. Acta, 2006, 579, 1-10.

3-19. J. Karasinski, S. Andreescu, O. A. Sadik, B. K. Lavine, M. N. Vora, "Multiarray Sensors with Pattern Recognition for the Detection, Classification, and Differentiation of Bacteria at Subspecies and Strain Levels", Anal. Chem., 2005, 77, 7941-7949.

3-20. B. K. Lavine, C. E. Davidson, W. T. Rayens, "Machine Learning Based Pattern Recognition Applied to Microarray Data", Combinatorial Chem. High Through. Screening, 2004, 7, 115-131.

3-21. B. K. Lavine, C. E. Davidson, C. Breneman, 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.

3-22. B. K. Lavine, C. E. Davidson, A. J. Moores, P. R. Griffiths, “Raman Spectroscopy and Genetic Algorithms for the Classification of Wood Types”, Appl. Spec., 2001, 55, 960-966.

3-23. B. B. Hubbard, The World According to Wavelets (Second Edition). A. K. Peters, Natick, MA, 1998.

3-24. J. S. Walker, A Primer on Wavelets and Their Scientific Applications. Chapman & Hall, CRC Press, New York, NY, 1999.

3-25. F. Chau, Y. Liang, J. Fao, X. Shao, Chemometrics – From Basics to Wavelet Transform. John Wiley & Sons, New York, NY, 2004.

3-26. L. A. Powell, G. M. Hieftje, “Computer Identification of Infrared Spectra by Correlation-Based File Searching”, Anal. Chim. Acta, 1978, 100, 313-327.

3-27. B. K. Lavine, C. G. White, M. D. Allen, A. Fasasi, “Improving Investigative Lead Information in the Forensic Examination of Automotive Paints”, In: B. K. Lavine, K. Booksh, S. Brown (Eds). 40 Years of Chemometrics – From Bruce Kowalski to the Future. ACS Symposium Series, 2015, 1199, 195-218.

3-28. M. Boruta, “FT-IR Search Algorithm – Assessing the Quality of a Match”, Spectroscopy, 2012, 27, 1-6.

3-29. J. E. Jackson, A User’s Guide to Principal Component Analysis. John Wiley & Sons, New York, NY, 1991.

3-30. D. L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis. John Wiley & Sons, New York, NY, 1983.

CHAPTER IV

ODOR-STRUCTURE RELATIONSHIP STUDIES OF POLYCYCLIC MUSKS

4.1 INTRODUCTION

Compounds exhibiting musk odor have been the subject of intensive study for many years [4-1, 4-2] due to their characteristic odor and fixative properties. Musks are the source of the primary odor in cologne, and are a secondary odor note in other fragrance products (e.g., perfumes where floral odor is the primary odor note). Musks are also interesting from a structural point of view since they span a variety of structural manifolds thereby making musk odor prediction challenging [4-3]. A plethora of information about musks is available in the scientific literature.

The first compound that was given the designation of musk occurred naturally as a dark red to black-brown grainy secretion produced by the glands of the male Asian musk deer (Moschus moschiferus). It is from this macrocyclic compound that commercial musk products were initially derived [4-4]. In 1891, Baur produced the first synthetic musk, one of many nitroaromatic compounds exhibiting an odor similar to that of natural musk. Currently, polycyclic musks are the largest and most important group of synthetic musks as they include some of the most powerful musks known, e.g., Phantolid, Tonalid and galoxolide [4-5]. The commercial success of indan and tetralin musks has been attributed to their chemical stability as well as their similarity in odor to naturally occurring musks.

Odor-structure relationship (OSR) studies, which have been the subject of several reviews [4-6 – 4-9] have played an important role in the development of new synthetic musks. The goal of any OSR study is to find the relationship between the biological activity of interest (e.g., musk odor) and the chemical structure of the compound. However, this goal cannot be achieved by directly comparing the chemical structure and the odor quality of the compound. Instead, an indirect approach must be utilized, which involves generating a set of molecular descriptors that quantify information about the size, shape, and electronic properties of the molecule for a set of compounds and then correlating the olfactory properties of these compounds to a subset of the molecular descriptors using statistical or pattern recognition methods.

OSR analysis using pattern recognition methods provides an approach for analyzing structure-activity relationships of musks. At the heart of this approach is the identification of a set of molecular descriptors that can discriminate musks from nonmusks that have a similar chemical structure. Several studies published in the literature have demonstrated the potential efficacy of this approach. In 1986, Narvaez [4-10] identified a set of 14 molecular descriptors that could discriminate musk odorants from nonmusks in a training set of 148 bicyclo- and tricyclo-benzenoid compounds, and 15 of the 16 compounds in the validation set were correctly classified by a discriminant developed from the 14 topological, fragment based, and geometrical descriptors. The 14 molecular descriptors selected for discriminant development contained information about the number of rings, the number of quaternary centers, distances between polar heteroatoms and quaternary centers and the nearest methyl group, and the degree of branching of substituents attached to the nonaromatic ring.

CASE methodology was applied by Klopman and Ptscelintsev [4-11] to a set of 152 non nitroaromatic musk and nonmusk compounds. 23 descriptors were identified of which 9 were structural fragments responsible for musk odor and 7 were fragments encountered exclusively in the nonmusks. The 23 descriptors correctly classified 18 of the 20 compounds in the validation set. 3-layer neural networks were employed by Charastrette [4-12] and Cherqaoui [4-13] to classify indan and tetralin musks and their nonmusk analogues using linear free energy relationship parameters as descriptors for the aromatic substituents. These two studies employed either fragment based descriptors or some variation of Hansch analysis which limited these studies to a set of homologous compounds.

Most recent OSR studies of musks [4-14 – 4-17] included the use of a genetic algorithm for variable selection to identify informative descriptors from a library of Breneman Transferable Atom Equivalent (TAE) descriptors [4-18 – 4-20]. From these descriptors, a molecule was mapped to a large set of spatially-resolved property descriptors correlating with key intermolecular interaction modes. Breneman's descriptors included spatially resolved shape/property hybrid electron density based property encoded surface translation (PEST) descriptors [4-21] which can achieve strong correlations to biological responses. These molecular descriptors are not subject to the limitations of topological, geometric, and 2D fragment based descriptors or limited to homologous compound sets, which is the case for Hansch analysis [4-22]. The OSR of nitroaromatic musks, which was previously not well understood, because of the complex substitution pattern and the varied polyfunctional character of the nitro group was successfully modeled using 6 TAE based descriptors. These 6 descriptors were found to contain information about molecular interactions which may be important in olfaction.

In this chapter, two OSR studies of polycyclic musks and nonmusks which are similar in structure to the musks are presented. The first study involves indan and tetralin musks, whereas the second study is a progression of the first study and includes isochroman musks.

4.2. EXPERIMENTAL

4.2.1. Musk Compounds. An olfaction database was compiled from literature reports [4-23 – 4-32] of chemical structure and odor quality for three structural classes of musks: indans, tetralins, and isochromans. Indans typically contain a carbonyl group attached directly to a benzene ring which shares one of its bonds with a cyclopentane ring. Tetralins are similar to indans, but a cyclohexane ring is substituted for a cyclopentane ring. Isochromans, which are similar to the indans and tetralins, contain two nonaromatic rings. However, one carbon in one of the non-aromatic rings is substituted with an oxygen. Examples of these three polycyclic classes of compounds are displayed in Figure 4.1.

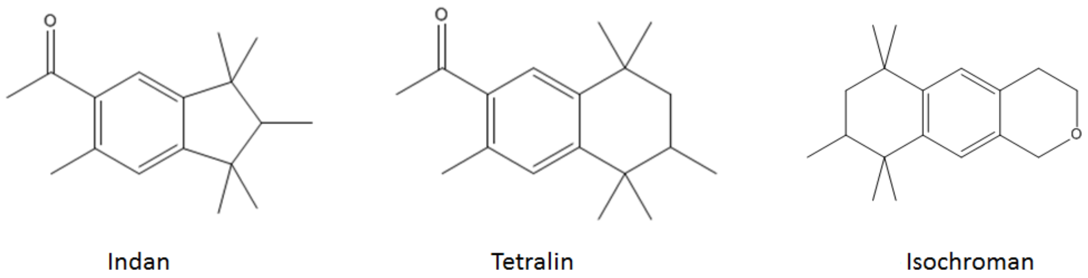


Figure 4.1. Examples of indan, tetralin, and isochroman compounds

Nonmusks that were similar in chemical structure to the musks were also included in the database. This not only contributed to the additional challenge of separating very similar structures according to odor quality but also would increase our understanding of how small structural changes affect odor quality.

Two pattern recognition studies were undertaken to explore structure-activity relationships of musks. The first study focused on 147 indans and tetralin musks and nonmusks. 70 were musks and 77 were nonmusks. The musks are of strong, medium or weak odor intensity and the nonmusks are odorless or have an odor other than musk. Of the 147 compounds, 110 comprised a training set and 37 were assigned to the validation set. Table 4.1 lists the training set compounds and Table 4.2 lists the validation set compounds used in the first study. The second study focused on 191 indans, tetralins and isochroman musks and nonmusks. 99 were musks, and 92 were nonmusks. The training set consisted of 172 compounds and the validation set consisted of 19 compounds. Table 4.3 lists the training set compounds and Table 4.4 lists the validation set compounds used in the second study.

Many of the compounds listed in Tables 4.1 - 4.4 possess atoms that can serve as stereocenters. However, the olfactory data gleaned from the literature for these compounds does not specify the chirality of the molecule. Furthermore, the molecular descriptor generation routines used in these studies cannot distinguish between enantiomers. Although some information has been lost about the relationship between structure and olfactory quality, the molecular attributes, which characterize musk odor modality, are captured by the topological and shape-aware molecular descriptors used in these two musk studies.

Table 4.1. List of training set compounds in the indan and tetralin study

Compound Name	Odor Quality
6-(1,1-dimethylethyl)-4-ethyl-2,3-dihydro-1,1-dimethyl-1H-indene	OLES
6-acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
7-acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
8-acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
5-acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
7-acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
8-acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
5-acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
6-acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
5-acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
6-acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
7-acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
1-(2,3-dihydro-1,1-dimethyl-1H-inden-4-yl)-ethanone	OLES
1-(2,3-dihydro-3,3-dimethyl-1H-inden-4-yl)-ethanone	OLES
1-[3-(1,1-dimethylethyl)-5,6,7,8-tetrahydro-1-naphthalenyl]-ethanone	OLES
1-(2,3-dihydro-1,1-dimethyl-1H-inden-5-yl)-ethanone	OLES
1-(2,3-dihydro-3,3-dimethyl-1H-inden-5-yl)-ethanone	OLES
1-(2,3-dihydro-1,2,3,3,6-pentamethyl-1H-inden-5-yl)-ethanone	OLES
5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	OLES
1-(5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-5-naphthalenyl)-ethanone	OLES
5,6,7,8-tetrahydro-4-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	OLES
3-acetyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarbonitrile	OLES
2-methyl-1-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-1-propanone	OLES
1-(5,8-diethyl-5,6,7,8-tetrahydro-3,5,8-trimethyl-2-naphthalenyl)-ethanone	OLES
1-(3-ethyl-5,6,7,8-tetrahydro-8,8-dimethyl-2-naphthalenyl)-ethanone	OLES
1-[2,3-dihydro-1,3,3,6-tetramethyl-1-(2-methylpropyl)-1H-inden-5-yl]-ethanone	OLES
1-[5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenyl]-propanone	OLES
1-(5,8-dihydro-5,5,7,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
1-(5,8-dihydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
1-(5,8-dihydro-5,5,6,8,8-pentamethyl-1-naphthalenyl)-ethanone	OLES
1-[5,6,7,8-tetrahydro-3,8,8-trimethyl-5-(1-methylethyl)-2-naphthalenyl)-ethanone	OLES
1-[5,6,7,8-tetrahydro-3,5,5-trimethyl-8-(1-methylethyl)-2-naphthalenyl)-ethanone	OLES
4-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-3-buten-2-one	OLES
1,2,3,4-tetrahydro-5-methoxy-1,1,4,7,7-pentaethylnaphthalene	OLES
5,6,7,8-tetrahydro-1-methoxy-4,5,5,8,8-pentaethylnaphthalene	OLES
5,6,7,8-tetrahydro-1,3-dimethoxy-5,5,8,8-tetraethylnaphthalene	OLES
3-amino-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxylic acid methyl ester	OLES
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenol	OLES
5,6,7,8-tetrahydro-2-methoxy-5,5,8,8-tetramethylnaphthalene	OLES
2-ethoxy-5,6,7,8-tetrahydro-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-tetrahydro-3-methylbutoxy-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenol	OLES
5,6,7,8-tetrahydro-2-methoxy-3,5,5,8,8-pentamethylnaphthalene	OLES
5,6,7,8-tetrahydro-2,3-dimethoxy-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-tetrahydro-2-methoxymethyl-3,5,5,8,8-pentamethylnaphthalene	OLES
3-ethyl-5,6,7,8-tetrahydro-2-methoxy-5,5,8,8-tetramethylnaphthalene	OLES
2-methoxy-1-(5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	OLES
3-acetyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxylic acid methyl ester	NONM
1-[5,6,7,8-tetrahydro-3-(methoxymethyl)-5,5,8,8-tetramethyl-2-naphthalenyl]-ethanone	NONM
1-(5,6,7,8-tetrahydro-3-hydroxy -5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid	NONM
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid diethyl ester	NONM
1-(3,5,8-triethyl-5,6,7,8-tetrahydro-5,8-dimethyl-2-naphthalenyl)-ethanone	NONM
1-{5,8-diethyl-5,6,7,8-tetrahydro-5,8-dimethyl-3-(1-methylethyl)-2-naphthalenyl]-ethanone	NONM
1-(5,6,7,8-tetrahydro-3,8,8-trimethyl-2-naphthalenyl)-ethanone	NONM
1,3,3-trimethyl-1-propyl-1H-indene	NONM
1-(2,3-dihydro-1,1,2,3,3-pentamethyl-1H-inden-5-yl)-1-propanol	MMUS

1-[2,3-dihydro-1,1,3,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MMUS
1-[3-(1,1-dimethylethyl)-2,3-dihydro-1,1,2,6-tetramethyl-1H-inden-5-yl]-ethanone	MSTR
1-[2,3-dihydro-1,1,2,2,6-pentamethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[6-ethyl-2,3-dihydro-1,1,2,2-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[6-ethyl-3-dihydro-1,1,2,2-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-1-propanone	MUSK
1-ethyl-2,3-dihydro-1,3,3,6-tetramethyl-1H-indene-5-carboxyaldehyde	MSTR
2,3-dihydro-1,1,3,3,6-pentamethyl-1H-indene-5-carboxyaldehyde	MSTR
1-(2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-inden-5-yl)-ethanone (“Phantolid”)	MSTR
1-(2,3-dihydro-1,1,2,3,6-pentamethyl-1H-inden-5-yl)-ethanone	MWEA
1-(2-ethyl-2,3-dihydro-1,1,3,3,6-pentamethyl-1H-inden-5-yl)-ethanone	MMUS
1-(6-ethyl-2,3-dihydro-1,1,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MMUS
1-(6-ethyl-2,3-dihydro-1,1,2,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MMUS
1-(2,3-dihydro-1,1,3,3,5,6-hexamethyl-1H-inden-4-yl)-ethanone	MMUS
1-(2,3-dihydro-1,1,2,3,3,-pentamethyl-1H-inden-5-yl)-ethanone	MMUS
1-[2,3-dihydro-1,1,2,3,3-pentamethyl-6-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-dihydro-1,1,3,3-pentamethyl-6-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-dihydro-1,1,2,6-tetramethyl-3-(trimethylsilyl)-1H-inden-5-yl]-ethanone	MSTR
2,3-dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-indene-5-carboxaldehyde	MSTR
1-[2-ethyl-2,3-dihydro-1,1,6-trimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MSTR
1-[2,3-dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MSTR
2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-indene-5-carbonitrile	MSTR
3-ethyl-5,6,7,8-tetrahydro-1-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	MMED
5,6,7,8-tetrahydro-1-methoxy-3,3,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
5,6,7,8-tetrahydro-1-hydroxy-3,3,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxaldehyde	MMED
5’,6’,7’,8’-tetrahydro-5’,5’,8’,8’-tetramethyl-3’-(methylthio)-2’-acetonaphthone	MUSK
5’,6’,7’,8’-tetrahydro-5’,5’,6’,8’,8’-pentamethyl-3’-(methylthio)-2’-acetonaphthone	MUSK
5’,6’,7’,8’-tetrahydro-5’,5’,6’,7’,8’,8’-pentamethyl-3’-(methylthio)-2’-acetonaphthone	MUSK
3’-(ethylthio)-5’,6’,7’,8’-tetrahydro-5’,5’,8’,8’-tetramethyl-2’-acetonaphthone	MUSK
1-(3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-propanone	MWEA
1-(3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone (“Versalide”)	MSTR
1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-ethanone (“Tonalid”)	MSTR
1-(5,6,7,8-tetrahydro-2-naphthalenyl)-ethanone	MSTR
1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-carbonitrile	MMED
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	MSTR
1-(5,8-dihydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MMUS
1-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenyl)-ethanone	MWEA
1-(5,6,7,8-tetrahydro-3,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MMED
1-(5-ethyl-5,6,7,8-tetrahydro-3,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MWEA
1-(3-ethyl-5,6,7,8-tetrahydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	MSTR
1-(5,6,7,8-tetrahydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	MMUS
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenecarbonitrile	MMUS
3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-pentamethyl-2-naphthalenecarbonitrile	MMUS
1-(1-ethyl-2,3-dihydro-1,3,3,5,6-pentamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(3-ethyl-2,3-dihydro-1,1,3,5,6-pentamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(1-ethyl-2,3-dihydro-1,3,3-trimethyl-1H-inden-4-yl)-ethanone	MWEA
1-(3-ethyl-2,3-dihydro-1,1,3,6-tetramethyl-1H-inden-5-yl)-ethanone	MWEA
1-(2,3-dihydro-1,1,2,3,3,5,6-heptamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(2,3-dihydro-1,1,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MWEA
1-[3-(1,1-dimethylethyl)-5,6,7,8-tetrahydro-5,5-dimethyl-1-naphthalenyl]-ethanone	MWEA
1-(5,8-dihydro-3,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	MWEA

OLES = Odorless, NONM = Nonmusk, MSTR = Strong Musk, MMUS = Musk of medium or strong odor intensity, MMED = Medium Musk, MUSK = Musk of Unspecified Odor Intensity, MWEA = Weak Musk

Table 4.2. List of validation set compounds in the indan and tetralin study

Compound Name	Odor Quality
8-acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
1-(2,3-dihydro-1,3,3,6-tetramethyl-1-propyl-1H-inden-5-yl)-ethanone	OLES
1-(5,8-dihydro-5,5,7,8,8-pentamethyl-1-naphthalenyl)-ethanone	OLES
1-(5,6,7,8-tetrahydro-1-methoxy-3,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-(3-methylbutoxymethyl)-naphthalene	OLES
2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-indene-5-carboxaldehyde	MSTR
1-(2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-inden-5-yl)-1-propanone	MSTR
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenecarboxaldehyde	MMUS
1-(5-ethyl-5,6,7,8-tetrahydro-3,5,8-trimethyl-2-naphthalenyl)-ethanone	MWEA
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid dimethyl ester	NONM
1-(3-ethyl-2,3-dihydro-1,1,3-trimethyl-1H-inden-4-yl)-ethanone	MWEA
5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenylcarboxaldehyde	MSTR
(5,6,7,8-tetrahydro-1,3,5,5,8,8-hexamethyl-2-naphthalenylcarboxaldehyde	MSTR
1-(5,6,7,8-tetrahydro-1,3,5,5,8,8-hexamethyl-2-naphthalenyl)-ethanone	OLES
1-(5,6,7,8-tetrahydro-4,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	MWEA
1-(5,6,7,8-tetrahydro-4,5,5,8,8-pentamethyl)-2-naphthalenecarboxaldehyde	MWEA
1-(5,6,7,8-tetrahydro-1,4,5,5,8,8-hexamethyl-2-naphthalenyl)-carboxaldehyde	MWEA
1-(5,6,7,8-tetrahydro-1,3,4,5,5,8,8-heptamethyl)-2-naphthalenecarboxaldehyde	MWEA
1-(5,6,7,8-tetrahydro-1,3,5,5,6,8,8-heptamethyl)-2-naphthalenecarboxaldehyde	MSTR
1-(5,6,7,8-tetrahydro-1,3,5,5,7,8,8-heptamethyl)-2-naphthalenecarboxaldehyde	MSTR
1-(5,6,7,8-tetrahydro-1,3,5,5,6,8,8-heptamethyl-2-naphthalenyl)-ethanone	OLES
Trans-1-(5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl)-ethanone	MSTR
Trans-1-(5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl)-carboxaldehyde	MSTR
Cis-1-(5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl)-carboxaldehyde	MSTR
1-(3-chloro-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(3-bromo-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(3-iodo-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-[2,3-dihydro-1,1,6-trimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-[6-ethyl-2,3-dihydro-1,1-dimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-[2,6-diethyl-2,3-dihydro-1,1-dimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-(1,2,3,4,5,6,7,8-octahydro-2,3,8,8-tetramethyl-2-naphthalenyl)-ethanone	AMBER
1-(1,2,3,4,5,6,7,8-octahydro-3,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(1,2,3,4,6,7,8,8a-octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(1,2,3,4,5,6,7,8-octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(1,2,3,5,6,7,8,8a-octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
2,3-dihydro-β,1,1,2,3,3-hexamethyl-1H-indene-5-ethanol	WOODY
1-[5,6,7,8-tetrahydro-5,5-dimethyl-3-(1-methylethyl)-2-naphthalenyl)-ethanone	WOODY

OLES = Odorless, NONM = Nonmusk, MSTR = Strong Musk, MMUS = Musk of medium or strong odor intensity, MMED = Medium Musk, MUSK = Musk of Unspecified Odor Intensity, MWEA = Weak Musk, AMBER = Amber odor, WOODY = Woody odor

Table 4.3. List of training set compounds in the indan, tetralin and isochroman study

Compound Name	Odor Quality
6-(1,1-dimethylethyl)-4-ethyl-2,3-dihydro-1,1-dimethyl-1H-indene	OLES
7-acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
8-acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
5-acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
7-acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
8-acetyl-3,4-dihydro-1,1,4,4,6-pentamethyl-2(1H)-naphthalenone	OLES
5-acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
6-acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES

5-acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
7-acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
8-acetyl-3,4-dihydro-1,1,4,4,7-pentamethyl-2(1H)-naphthalenone	OLES
1-(2,3-dihydro-1,1-dimethyl-1H-inden-4-yl)-ethanone	OLES
1-(2,3-dihydro-3,3-dimethyl-1H-inden-4-yl)-ethanone	OLES
1-(2,3-dihydro-1,3,3,6-tetramethyl-1-propyl-1H-inden-5-yl)-ethanone	OLES
1-(5,8-dihydro-5,5,7,8,8-pentamethyl-1-naphthalenyl)-ethanone	OLES
1-[3-(1,1-dimethylethyl)-5,6,7,8-tetrahydro-1-naphthalenyl]-ethanone	OLES
1-(2,3-dihydro-1,1-dimethyl-1H-inden-5-yl)-ethanone	OLES
1-(2,3-dihydro-3,3-dimethyl-1H-inden-5-yl)-ethanone	OLES
1-(2,3-dihydro-1,2,3,3,6-pentamethyl-1H-inden-5-yl)-ethanone	OLES
5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	OLES
1-(5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-5-naphthalenyl)-ethanone	OLES
1-(5,6,7,8-tetrahydro-1-methoxy-3,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
5,6,7,8-tetrahydro-4-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	OLES
3-acetyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarbonitrile	OLES
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-(3-methylbutoxymethyl)-naphthalene	OLES
1-(5,8-diethyl-5,6,7,8-tetrahydro-3,5,8-trimethyl-2-naphthalenyl)-ethanone	OLES
1-(3-ethyl-5,6,7,8-tetrahydro-8,8-dimethyl-2-naphthalenyl)-ethanone	OLES
1-[2,3-dihydro-1,3,3,6-tetramethyl-1-(2-methylpropyl)-1H-inden-5-yl]-ethanone	OLES
1-[5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenyl]-propanone	OLES
1-(5,8-dihydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
1-(5,8-dihydro-5,5,6,8,8-pentamethyl-1-naphthalenyl)-ethanone	OLES
1-[5,6,7,8-tetrahydro-3,8,8-trimethyl-5-(1-methylethyl)-2-naphthalenyl)-ethanone	OLES
1-[5,6,7,8-tetrahydro-3,5,5-trimethyl-8-(1-methylethyl)-2-naphthalenyl)-ethanone	OLES
4-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-3-buten-2-one	OLES
1,2,3,4-tetrahydro-5-methoxy-1,1,4,7,7-pentaethylnaphthalene	OLES
5,6,7,8-tetrahydro-1-methoxy-4,5,5,8,8-pentaethylnaphthalene	OLES
5,6,7,8-tetrahydro-1,3-dimethoxy-5,5,8,8-tetraethylnaphthalene	OLES
3-amino-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxylic acid methyl ester	OLES
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenol	OLES
5,6,7,8-tetrahydro-2-methoxy-5,5,8,8-tetramethylnaphthalene	OLES
2-ethoxy-5,6,7,8-tetrahydro-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-tetrahydro-3-methylbutoxy-5,5,8,8-tetramethylnaphthalene	OLES
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenol	OLES
5,6,7,8-tetrahydro-2-methoxy-3,5,5,8,8-pentamethylnaphthalene	OLES
5,6,7,8-tetrahydro-2,3-dimethoxy-5,5,8,8-tetramethylnaphthalene	OLES
3-ethyl-5,6,7,8-tetrahydro-2-methoxy-5,5,8,8-tetramethylnaphthalene	OLES
2-methoxy-1-(5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	OLES
3-acetyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxylic acid methyl ester	NONM
1-[5,6,7,8-tetrahydro-3-(methoxymethyl)-5,5,8,8-tetramethyl-2-naphthalenyl]-ethanone	NONM
1-(5,6,7,8-tetrahydro-3-hydroxy -5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid diethyl ester	NONM
1-(3,5,8-triethyl-5,6,7,8-tetrahydro-5,8-dimethyl-2-naphthalenyl)-ethanone	NONM
1-{5,8-diethyl-5,6,7,8-tetrahydro-5,8-dimethyl-3-(1-methylethyl)-2-naphthalenyl]-ethanone	NONM
1-(5,6,7,8-tetrahydro-3,8,8-trimethyl-2-naphthalenyl)-ethanone	NONM
1-(3-fluoro-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(3-chloro-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(3-bromo-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(3-iodo-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-[2,3-dihydro-1,1,6-trimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-[6-ethyl-2,3-dihydro-1,1-dimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM
1-[2,6-diethyl-2,3-dihydro-1,1-dimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	NONM

1-(2,3-dihydro-1,1,2,3,3-pentamethyl-1H-inden-5-yl)-1-propanol	MMUS
1-[2,3-dihydro-1,1,3,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MMUS
1-[3-(1,1-dimethylethyl)-2,3-dihydro-1,1,2,6-tetramethyl-1H-inden-5-yl]-ethanone	MSTR
1-(2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-inden-5-yl)-ethanone	MSTR
1-[2,3-dihydro-1,1,2,2,6-pentamethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[6-ethyl-3-dihydro-1,1,2,2-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-1-propanone	MUSK
1-ethyl-2,3-dihydro-1,3,3,6-tetramethyl-1H-indene-5-carboxyaldehyde	MSTR
2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-indene-5-carboxaldehyde	MSTR
2,3-dihydro-1,1,3,3,6-pentamethyl-1H-indene-5-carboxyaldehyde	MSTR
5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenylcarboxaldehyde	MSTR*
Trans-5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl-carboxaldehyde	MSTR*
Cis-5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl-carboxaldehyde	MSTR*
1-(2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-inden-5-yl)-ethanone (“Phantolid”)	MSTR
1-(2,3-dihydro-1,1,2,3,6-pentamethyl-1H-inden-5-yl)-ethanone	MWEA
1-(2-ethyl-2,3-dihydro-1,1,3,3,6-pentamethyl-1H-inden-5-yl)-ethanone	MMUS
1-(6-ethyl-2,3-dihydro-1,1,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MMUS
1-(6-ethyl-2,3-dihydro-1,1,2,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MMUS
1-(2,3-dihydro-1,1,3,3,5,6-hexamethyl-1H-inden-4-yl)-ethanone	MMUS
1-(2,3-dihydro-1,1,2,3,3,-pentamethyl-1H-inden-5-yl)-ethanone	MMUS
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenecarboxaldehyde	MMUS
1-[2,3-dihydro-1,1,2,3,3-pentamethyl-6-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-dihydro-1,1,3,3-pentamethyl-6-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
1-[2,3-dihydro-1,1,2,6-tetramethyl-3-(trimethylsilyl)-1H-inden-5-yl]-ethanone	MSTR
2,3-dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-indene-5-carboxaldehyde	MSTR
1-[2-ethyl-2,3-dihydro-1,1,6-trimethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MSTR
1-[2,3-dihydro-1,1,2,6-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MSTR
(5,6,7,8-tetrahydro-1,3,5,5,8,8-hexamethyl-2-naphthalenylcarboxaldehyde	MSTR*
2,3-dihydro-1,1,2,3,3,6-hexamethyl-1H-indene-5-carbonitrile	MSTR
3-ethyl-5,6,7,8-tetrahydro-1-methoxy-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	MMED
5,6,7,8-tetrahydro-1-methoxy-3,3,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
5,6,7,8-tetrahydro-1-hydroxy-3,3,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxaldehyde	MMED
5’,6’,7’,8’-tetrahydro-5’,5’,8’,8’-tetramethyl-3’-(methylthio)-2’-acetonaphthone	MUSK
5’,6’,7’,8’-tetrahydro-5’,5’,6’,8’,8’-pentamethyl-3’-(methylthio)-2’-acetonaphthone	MUSK
5’,6’,7’,8’-tetrahydro-5’,5’,6’,7’,8’,8’-pentamethyl-3’-(methylthio)-2’-acetonaphthone	MUSK
3’-(ethylthio)-5’,6’,7’,8’-tetrahydro-5’,5’,8’,8’-tetramethyl-2’-acetonaphthone	MUSK
1-(3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-propanone	MWEA
1-(3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone (“Versalide”)	MSTR
1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-ethanone (“Tonalid”)	MSTR
1-(5,6,7,8-tetrahydro-2-naphthalenyl)-ethanone	MSTR
1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-carbonitrile	MMED
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenecarboxaldehyde	MSTR
3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenecarboxaldehyde	MSTR
5,6,7,8-tetrahydro-1,3,5,5,7,8,8-heptamethyl-2-naphthalenecarboxaldehyde	MSTR*
Trans-1-(5,6,7,8-tetrahydro-3,5,5,6,7,8,8-heptamethyl-2-naphthalenyl)-ethanone	MSTR*
1-(5,8-dihydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MMUS
1-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-3-(1-methylethyl)-2-naphthalenyl)-ethanone	MWEA
1-(5-ethyl-5,6,7,8-tetrahydro-3,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MWEA
1-(5,6,7,8-tetrahydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	MMUS
5,6,7,8-tetrahydro-3,5,5,8,8-pentamethyl-2-naphthalenecarbonitrile	MMUS
3-ethyl-5,6,7,8-tetrahydro-5,5,8,8-pentamethyl-2-naphthalenecarbonitrile	MMUS
6-ethyl-2,6,8,8-tetramethyl-cyclopenta[g]-2-benzopyran	MMUS

1,3,4,6,7,8-hexahydro-6-(1-methylethyl)-4,7,8,8-tetramethyl-cyclopenta[g]-2-benzopyran	MMUS
1,3,4,6,7,8-hexahydro-8-(1-methylethyl)-4,6,6,7-tetramethyl-cyclopenta[g]-2-benzopyran	MMUS
3,4-dihydro-4-methyl-5,7-bis-(1-methylethyl)-1H-2-benzopyran	MMUS
2,3-dihydro-β-(1-methylethyl)-1H-Indene-5-ethanol	MMUS
4-ethyl-1,3,4,6,7,8-hexahydro-6,6,7,8,8-pentamethyl-cyclopenta[g]-2-benzopyran	MMXT
4,6-diethyl-1,3,4,6,7,8-hexahydro-6,8,8-trimethyl-cyclopenta[g]-2-benzopyran	MMXT
3,4,6,7,8,9-hexahydro-6,6,8,9,9-pentamethyl-1H-naphtho[2,3-C]pyran	MSTR
3,4,6,7,8,9-hexahydro-6,6,7,9,9-pentamethyl-1H-naphtho[2,3-C]pyran	MSTR
1,2,3,4,6,7,8,9-octahydro-6,6,9,9-tetramethyl-benz[g]isoquinoline	MWEA
1,2,3,4,6,7,8,9-octahydro-2,6,6,9,9-pentamethyl-benz[g]isoquinoline	MWEA
1-(3-ethyl-2,3-dihydro-1,1,3-trimethyl-1H-inden-4-yl)-ethanone	MWEA
1-(5-ethyl-5,6,7,8-tetrahydro-3,5,8-trimethyl-2-naphthalenyl)-ethanone	MWEA
1-(1-ethyl-2,3-dihydro-1,3,3,5,6-pentamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(3-ethyl-2,3-dihydro-1,1,3,5,6-pentamethyl-1H-inden-4-yl)-ethanone	MWEA
1-(1-ethyl-2,3-dihydro-1,3,3-trimethyl-1H-inden-4-yl)-ethanone	MWEA
1-(3-ethyl-2,3-dihydro-1,1,3,6-tetramethyl-1H-inden-5-yl)-ethanone	MWEA
1-(2,3-dihydro-1,1,3,3-tetramethyl-1H-inden-5-yl)-ethanone	MWEA
1-[3-(1,1-dimethylethyl)-5,6,7,8-tetrahydro-5,5-dimethyl-1-naphthalenyl]-ethanone	MWEA
1-(5,6,7,8-tetrahydro-4,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	MWEA*
1-(5,6,7,8-tetrahydro-4,5,5,8,8-pentamethyl)-2-naphthalenecarboxaldehyde	MWEA*
1-(5,6,7,8-tetrahydro-1,3,4,5,5,8,8-heptamethyl)-2-naphthalenecarboxaldehyde	MWEA*
1-(5,6,7,8-tetrahydro-1,4,5,5,8,8-hexamethyl-2-naphthalenyl)-carboxaldehyde	MWEA*
1-(5,8-dihydro-3,5,5,8,8-pentamethyl-2-naphthalenyl)-ethanone	MWEA
1,2,3,4,6,7,8,9-octahydro-4,6,6,9,9-pentamethyl-benz[g]isoquinoline	MMED
6,9-diethyl-3,4,6,7,8,9-hexahydro-4,6,9-trimethyl-1H-naphtho [2,3-C]pyran	MWEA
1,3,4,6,7,8-hexahydro-4,6,6,7,8,8-hexamethyl-cyclopenta[G]-2-benzopyran	MSTR
3,4,6,7,8,9-hexahydro-6,6,9,9-tetramethyl-1H-naphtho[2,3-C]-pyran	MSTR
4-ethyl-3,4,6,7,8,9-hexahydro-6,6,9,9-tetramethyl-1H-naphtho[2,3-C]pyran	MSTR
5,6,7,8-tetrahydro-1,3,5,5,6,8,8-heptamethyl-2-naphthalenecarboxaldehyde	MSTR*
3,4,6,7,8,9-hexahydro-6,6,9,9,10-pentamethyl-1H-naphtho[2,3-C]pyran	MMED
3,4,6,7,8,9-hexahydro-6,6,7,8,9,9-hexamethyl-1H-naphtho[2,3-C]pyran	MMED
1,3,4,6,7,8-hexahydro-6,6,8,8-tetramethyl-cyclopenta[G]-2-benzopyran	MWEA
1,3,4,6,7,8-hexahydro-4,6,6,8,8-pentamethyl-cyclopenta[G]-2-benzopyran	MWEA
1,3,4,6,7,8-hexahydro-6,6,7,8,8-pentamethyl-cyclopenta[G]-2-benzopyran	MWEA
3,4,6,7,8,9-hexahydro-4,6,6,7,9,9-hexamethyl-1H-naphtho[2,3-C]-pyran	MMED
1,3,4,6,7,8-hexahydro-4,4,8,8-tetrahydro-cyclopenta[G]-2-benzopyran	OLES
6,8-diethyl-1,3,4,7-tetrahydro-6,8-dimethyl-cyclopenta[G]-2-benzopyran	OLES
Dodecahydro-6,6,9,9-tetrahydro-1H-naphtho[2,3-C]pyran	OLES
3,4,6,7,8,9-hexahydro-1,6,6,9,9-pentamethyl-1H-naphtho[2,3-C]pyran	OLES
3,4,6,7,8,9-hexahydro-3,6,6,9,9-pentamethyl-1H-naphtho[2,3-C]pyran	OLES
1,3,5,6,7,8-hexahydro-5,5,8,8-tetramethyl-naphtho[2,3-C]thiophene	OLES
1,3,5,6,7,8-hexahydro-5,5,8,8-tetramethyl-naphtho[2,3-C]thiophene-2-oxide	OLES
6,7,8,9-tetrahydro-6,6,9,9-tetramethyl-benzo[G]-phthalazine	OLES
2,3,7,8-tetraahydro-4,4,6,6,8,8-hexamethyl-cyclopenta[G]-1-benzopyran	OLES
6,9-diethyl-3,4,7,8-tetrahydro-6,9-dimethyl-1H-naphtho[2,3-C]pyran	OLES
3,4,6,7,8,9-hexahydro-6,6,9,9-tetramethyl-2H-naphtho[2,3-B]pyran	NONM
2,3,4,7-tetrahydro-4,6,6,7,8,8-hexamethyl-cyclopenta[G]-1-benzopyran	NONM
1-(1,2,3,4,5,6,7,8-octahydro-2,3,8,8-tetramethyl-2-naphthalenyl)-ethanone	AMBER
1-(1,2,3,4,5,6,7,8-octahydro-3,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(1,2,3,5,6,7,8,8a-octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
2,3-dihydro-β,1,1,2,3,3-hexamethyl-1H-indene-5-ethanol	WOODY
1-(1,2,3,4,5,6,7,8-octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-[5,6,7,8-tetrahydro-5,5-dimethyl-3-(1-methylethyl)-2-naphthalenyl)-ethanone	WOODY

1-(1,2,3,4,6,7,8,8a-octahydro-4,8,8-trimethyl-2-naphthalenyl)-ethanone	WOODY
1-(3-ethyl-5,6,7,8-tetrahydro-5,5,6,8,8-pentamethyl-2-naphthalenyl)-ethanone	MSTR
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid dimethyl ester	NONM
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedimethanol	MWEA
1-(1-ethyl-2,3-dihydro-1,6-dimethyl-1H-inden-5-yl)-ethanone	MWEA
1-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	NONM
1-(2,3-dihydro-1,1,2,3,3-pentamethyl-1H-inden-5-yl)-1-propanone	MMUS
1-(1,2,6,7,8,8a-hexahydro-3,6,6,8a-tetramethyl-4-acenaphthalenyl)-ethanone	MSTR*
1,2,6,7,8,8a-hexahydro-3,6,6,8a-tetramethyl-4-acenaphthalenecarboxaldehyde	MSTR*
(4,5,5a,6,7,8-hexahydro-1,4,4,6-tetramethyl-2-acenaphthalenyl)-carboxaldehyde	MWEA*
1-(2,3,6,7,8,9-hexahydro-1,1-dimethyl-1H-benz[e]inden-4-yl)-ethanone	MSTR

OLES = Odorless, NONM = Nonmusk, MSTR = Strong Musk, MMUS = Musk of medium or strong odor intensity, MMED = Medium Musk, MUSK = Musk of Unspecified Odor Intensity, MWEA = Weak Musk, MMXT = Mixture of Musks, AMBER = Amber odor, WOODY = Woody odor.

Table 4.4. List of validation set compounds in the indan, tetralin, and isochroman study

Compound Name	Odor Quality
1-(5,6,7,8-tetrahydro-1,3,5,5,8,8-hexamethyl-2-naphthalenyl)-ethanone	OLES*
1-(5,6,7,8-tetrahydro-1,3,5,5,6,8,8-heptamethyl-2-naphthalenyl)-ethanone	OLES
2-methyl-1-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-1-propanone	OLES
1,3,3-trimethyl-1-propyl-1H-indene	NONM
5,6,7,8-tetrahydro-2-methoxymethyl-3,5,5,8,8-pentamethylnaphthalene	OLES
5,6,7,8-tetrahydro-1,3-dimethoxy-5,5,8,8-tetramethylnaphthalene	OLES
1-(5,8-dihydro-5,5,7,8,8-pentamethyl-2-naphthalenyl)-ethanone	OLES
1,3,5,6,7,8-hexahydro-5,5,8,8-tetramethyl-naphtho[2,3-C]furan	MSTR
5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2,3-naphthalenedicarboxylic acid	NONM
6-acetyl-3,4-dihydro-1,1,4,4,8-pentamethyl-2(1H)-naphthalenone	OLES
1-[6-ethyl-2,3-dihydro-1,1,2,2-tetramethyl-3-(1-methylethyl)-1H-inden-5-yl]-ethanone	MUSK
6-acetyl-3,4-dihydro-1,1,4,4,5-pentamethyl-2(1H)-naphthalenone	OLES
3,4-dihydro-4-methyl-6,8-bis-(1-methylethyl)-1H-2-benzopyran	MSTR
2,3-dihydro-1,1,3,3,6-pentamethyl-1H-Indene-5-carboxaldehyde	MSTR
3,4,6,7,8,9-hexahydro-4,4,6,6,9,9-hexamethyl-1H-naphtho[2,3-C]pyran	MMED
3,4,6,7,8,9-hexahydro-4,6,6,9,9-pentamethyl-1H-naphtho[2,3-C]pyran	MSTR
1-(5,6,7,8-tetrahydro-3-methoxy-5,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	OLES
1-(5,6,7,8-tetrahydro-3,5,8,8-tetramethyl-2-naphthalenyl)-ethanone	MMED
1-(2,3-dihydro-1,1,2,3,3,5,6-heptamethyl-1H-inden-4-yl)-ethanone	MWEA

OLES = Odorless, NONM = Nonmusk, MSTR = Strong Musk, MMED = Medium Musk, MUSK = Musk of Unspecified Odor Intensity, MWEA = Weak Musk, AMBER = Amber odor, WOODY = Woody odor

4.2.2. Molecular Descriptor Generation. The chemical structure of each compound was translated into a mole file using ChemDraw (Cambridge Soft). The resulting mole files or topological tables served as input for the modeling program Quanta (Molecular Simulations), which utilizes the CHARMM force field. Quanta produced 3D coordinates for each compound. Two descriptor generation routines were run to obtain the molecular descriptors used in these two OSR studies.

Traditional molecular property descriptors (e.g., specific atom counts or partial positive surface area) were computed for each compound using CODESSA (CompuDrug International, Sedona, AZ). In order for a molecular descriptor to be used in these studies, it must be computable for every compounds in the dataset. However, some compound-descriptor combinations were uncalculatable and the corresponding CODESSA descriptors were excluded from the study. As a result, some CODESSA descriptors were unique to a particular study whereas other CODESSA descriptors were common to both studies. 400 CODESSA descriptors were computed for the indan and tetralin study, while 161 were computed for the study involving the isochromans.

Breneman's transferrable atom equivalent (TAE) descriptors were generated using a sequence of steps managed through a Java GUI and starting with the use of Jaguar (Schrodinger). Within this sequence, RECON methodology was employed to compute TAE's molecular surface property reconstructions, and the Property Encoded Surface Translator (PEST) algorithm which generated both hybrid shape/property descriptors and wavelet coefficient descriptors (WCD). The final output after utilization of the Java GUI was a set of 1207 descriptors, which have been shown to have advantages [4-17] over traditional molecular derived descriptors.

Between the final two steps in TAE descriptor calculation, there is a global limit calculation which affects the reported values in the final step. The significance is that values obtained for TAE descriptors are dependent on the dataset itself: adding or removing a molecule from the dataset will cause the descriptor values to change. Beyond this point, no additional compounds can be added to the dataset without regenerating the TAE descriptors. Failing to regenerate the TAE descriptors when a sample is to be added will result in an artificial separation because the added samples will have been run under different global limits, resulting in drastically different descriptor values.

4.3. INDAN AND TETRLIN STUDY

The first OSR study focused on indan and tetralin musks and their corresponding nonmusks. Two examples of strong musks are phantolid (a well-known indan) and tonalid (similar to phantolid, but a tetralin). The nonmusks are either odorless or exhibit an odor other than musk. Many of the nonmusks in were similar in their structure to the musks. Even small changes in structure, such as the addition or removal of a methyl group or a small change in the connectivity of a molecule would result in a loss of musk odor quality. Some examples of the effects of these structural changes on indan and tetralin musks are shown in Figure 4.2.

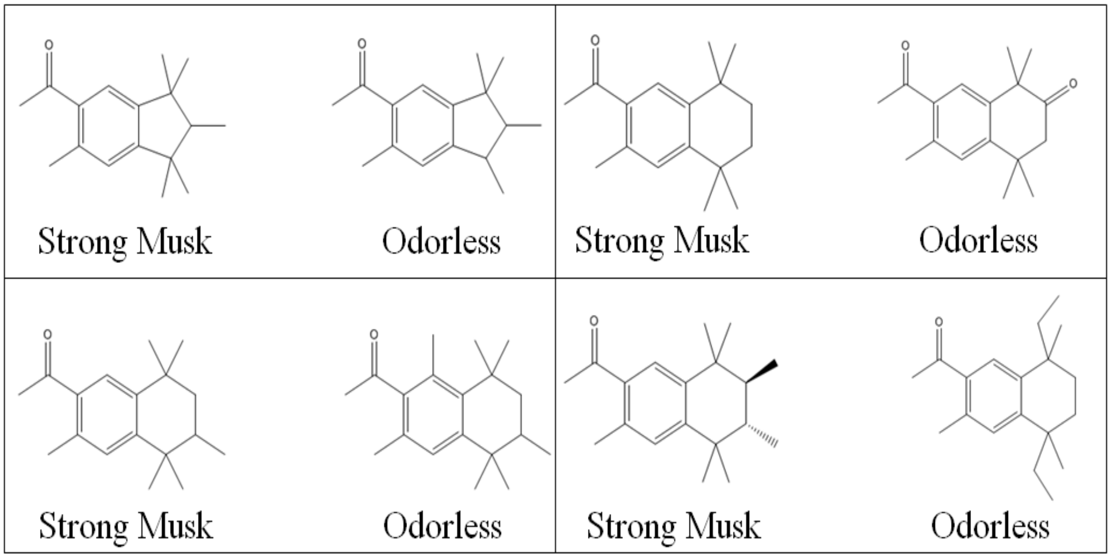


Figure 4.2. Examples of small differences between strong musks and nonmusks.

4.3.1. Pattern Recognition Results. After removal of invariant descriptors, each compound in the dataset was represented by a set of 1344 molecular descriptors (970 TAE / PEST, 374 CODESSA). Figure 4.3 shows a principal component (PC) plot of the 110 training set samples and the 1344 molecular descriptors. Each compound is represented as a point in the plot (1 = nonmusk and 2 = musk). This PC plot is an "all features" plot (i.e., before any variable selection is performed). The two classes (musk and nonmusk) overlap which is not surprising since the nonmusks were chosen to be similar in chemical structure to the musks.

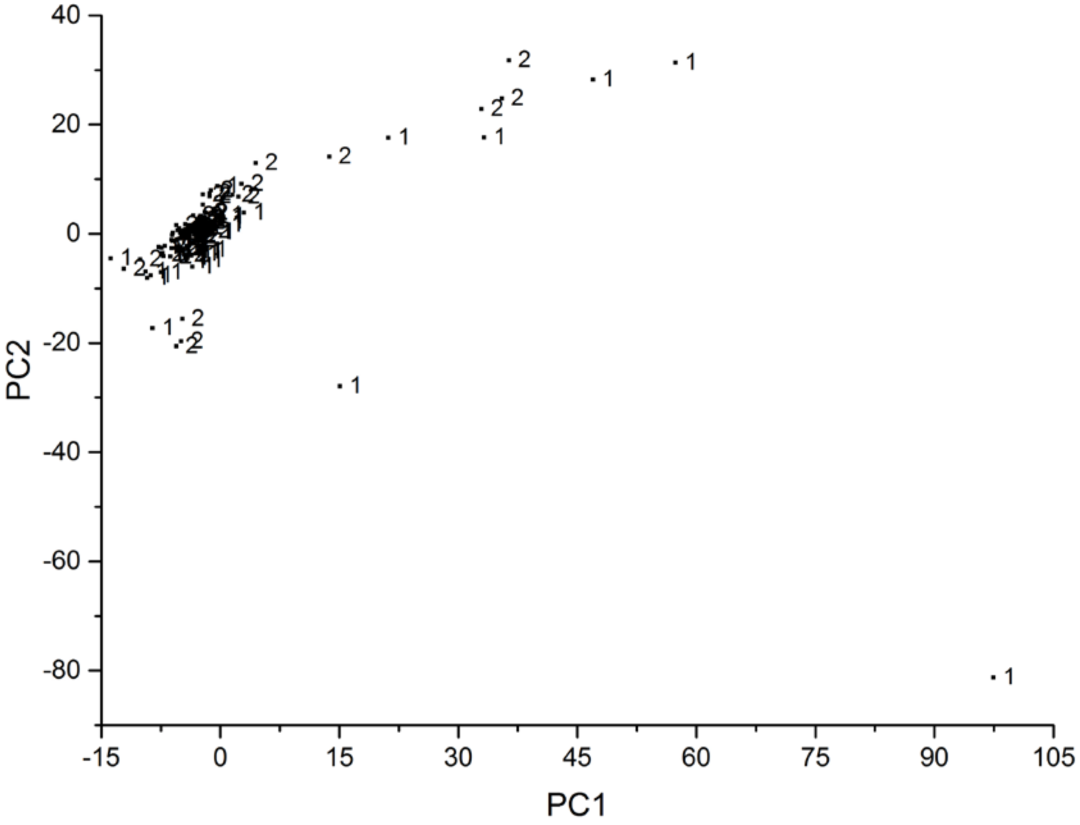


Figure 4.3. PC plot of the indan and tetralin training set compounds using all 1344 descriptors (1 = nonmusk, 2 = musk).

The pattern recognition GA was used to identify molecular descriptors from which a discriminating relationship could be found for the tetralin and indan musks. For this study, sixteen runs were performed with each run performed using different initial populations and mutation rates. A histogram depicting the descriptors most frequently selected by the pattern recognition GA during each generation was constructed and the 100 most frequently selected descriptors were extracted and subjected to further analysis by the pattern recognition GA. The pattern recognition GA then identified the most informative of these 100 descriptors by sampling key feature subsets, scoring their PC plots and tracking those classes and/or compounds that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, a set of 45 descriptors were identified whose PC plot (see Figure 4.4) showed clustering on the basis of odor.

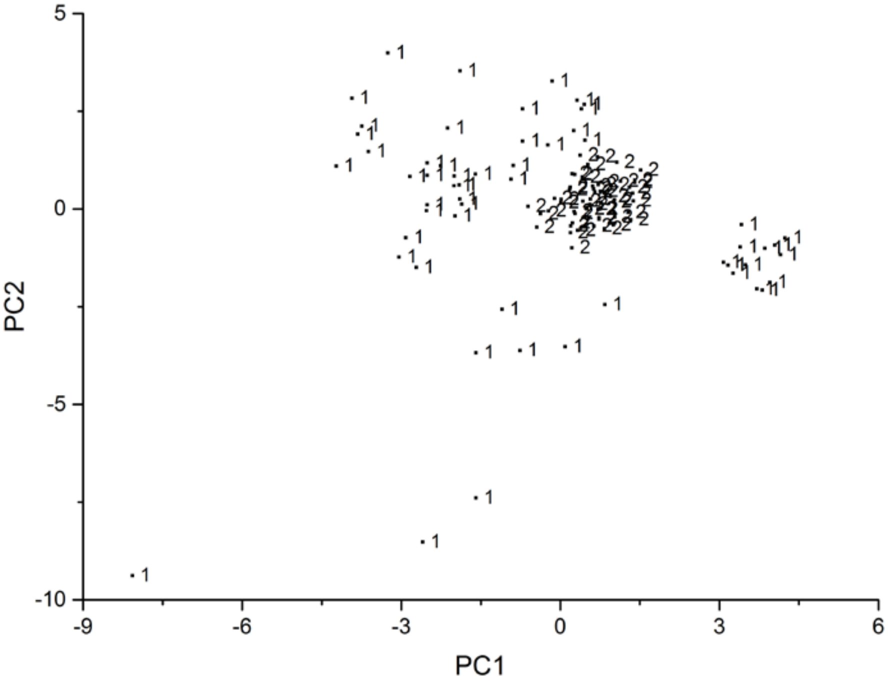


Figure 4.4. PC plot of the indan and tetralin training set with 45 selected features. (1 = nonmusk, 2 = musk).

The data structure of the tetralin and indan musk classification problem is asymmetric [4-33 – 4-35]. The tetralin and indan musks occupy a small and well defined region of the descriptor space, whereas the nonmusks are randomly distributed in this space. The musks form a class of compounds whose odor is a well behaved function of the 45 molecular descriptors identified by the pattern recognition GA, whereas the change in the structure of the nonmusks cannot be modeled using these descriptors because the compounds are inactive for a variety of reasons.

A validation set of 37 compounds (see Table 2) was used to assess the predictive ability of the 45 molecular descriptors identified by the pattern recognition GA. The 37 musks and nonmusks were projected on the PC plot defined by the 110 compounds of the training set and 45 descriptors identified by the pattern recognition GA. 34 of 37 validation set compounds were correctly classified, as shown in Figure 4.5.

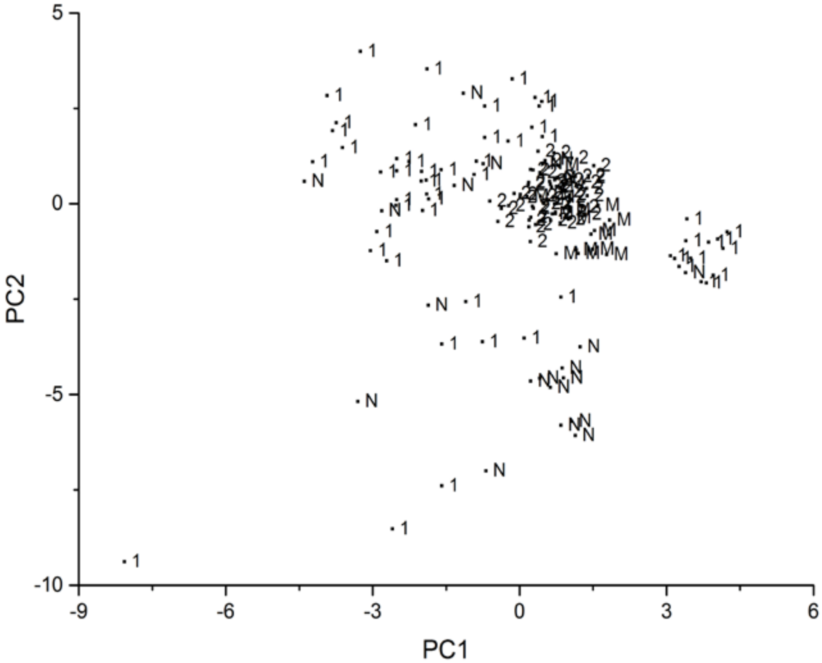


Figure 4.5. Projection of the validation set onto the PC plot of the indan and tetralin training set with 45 selected features. (1 = nonmusk training, 2 = musk training, N = nonmusk validation, M = musk validation).

The three misclassified compounds (see Figure 4.6) are all nonmusks. These compounds differ structurally in the lipophilic region, i.e., cyclohexane ring, from tetralin and indan musks. It is believed that odorant-receptor interactions involving this portion of the molecular are more selective than interactions involving the polar functional group of the molecule [4-26]. The TAE, PEST and CODESSA derived descriptors were not able to adequately delineate changes in shape resulting from alkyl substitution in this region of the molecule.

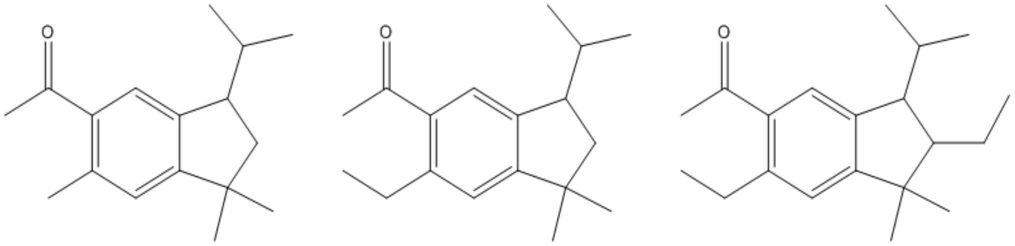


Figure 4.6. Three compounds misclassified by the GA.

The 45 descriptors identified by the pattern recognition GA were given to a back propagation neural network with topography of 45:2:1. A performance goal of 0.001 was set for the output of the neural network with a momentum constant of 0.99, learning rate of 0.01, ratio to decrease learning rate of 0.7, and ratio to increase learning rate of 1.5. The back propagation neural network correctly classified the training set, as well as 34 of the 37 validation set samples. The same three samples misclassified by the principal component plot were also the same three samples missed by the neural network.

4.3.2. Interpretation of Selected Descriptors. In order to understand the meaning of the separation achieved in the indan and tetralin musk study, it is necessary to examine the identities of the descriptors and the information they convey. Of the 45 descriptors, 41 were TAE descriptors and 4 were CODESSA descriptors. Their identities are given in Table 4.5.

Table 4.5. Identities of the 45 Selected Descriptors for the Indan and Tetralin Study

BNP.MIN	BNP.H0	BNP.W16
BNP.B36	BNP.B51	PIP.W3
PIP.W15	PIP.W17	PIP.W31
PIP.W3	FUK.W0	FUK.W4
FUK.W11	FUK.W20	FUK.W28
FUK.B56	EP.H3	EP.W5
EP.W14	EP.W22	EP.W31
EP.B02	G.W10	G.W29
K.W6	K.W21	K.B34
K.B46	G.B12	DKN.STDN
DKN.W27	DGN.AVGN	DGN.W27
DGN.B26	DRN.W0	DRN.W4
LAPL.W10	LAPL.W18	LAPL.W31
ANGLE.B06	ANGLE.B36	WPSA-3
HOMO-LUMO GAP	TOTAL HYBRID.	RCNS

The TAE descriptors are given by a set of letters identity the descriptor type, followed by a dot and a set of letters and number identifying the specific descriptor. The four CODESSA descriptors (WPSA-3, HOMO-LUMO GAP, TOTAL HYBRID, RCNS) provide information about electronic structure and surface charge distribution, which supplements the information provided by the TAE descriptors. The first 5 TAE descriptors are bare nuclear potential (BNP) descriptors. The name bare nuclear potential is a reflection of the fact that this quantity is being mapped onto the electron density isosurface. BNP.MIN, BNP.H0 and BNP.W16 capture polar and hydrogen bonding interactions. BNP.B36 and BNP.B51 capture information about shape and polarity. The “3” indicates that rays of average size are important, “6” indicates that relatively high BNP values are important, “5” indicates that long rays are important, and “1” indicates that relatively low BNP values are important.

The next 4 TAE descriptors are Politzer Ionization Potential (PIP) descriptors. These describe several aspects of the surface electrostatic potential distribution of the compounds in the data set, and have been shown to be correlated with a number of intermolecular binding modes, not the least of which is induced-dipole interactions. PIP.W3, PIP.W15, PIP.W17, and PIP.W31 are wavelet representations of the local ionization potential of the molecule. PIP.W3 describes the low value range of this property whereas the other three descriptors describe the high value range. These four PIP descriptors are known to convey information about hydrogen bonding and acidity.

Six of the TAE descriptors are Fukui radical reactivity indices (FUK). Like the PIP descriptors, the Fukui descriptors involve a perturbation expression which is meant to describe the spatial distribution of radical reactivity. However, for the Fukui descriptors, there is a selectable denominator term which places the reactivity index on a cationic, radical, or anionic scale. Five of the six Fukui descriptors are wavelet representations with FUK.W0 and FUK.W4 describing the low value range of this property and FUK.W11, FUK.W20, and FUK.W28 describing the high value range. FUK.B56 is a shape descriptor with near median values of ray length and property. The next six descriptors pertain to electrostatic potential (EP). EP.H3 is a histogram descriptor and EP.W5, EP.W14, EP.W22, and EP.W31 are scale coefficient wavelet descriptors. These five descriptors are correlated to the solvation energy of the molecule with EP.W14, EP.W22, and EP.W31 in the positive part of the EP range. EP.B02 is a shape property descriptor denoting short ray length distances and negative electrostatic potentials.

The G and K descriptors refer to G and K kinetic energy reconstructions. G.W10, G.W29, K.W6, and K.W21, which are wavelet descriptors derived from the G and K kinetic energy reconstructions normal to and away from the surface of the molecule, describe hydrogen bonding interactions. K.B34, K.B46 and G.B12 are shape descriptors derived from the ray traces of the G and K kinetic energy reconstructions normal to and away from the surface of the molecule. These descriptors capture more of the interior volume, i.e., local shape, as opposed to conformational information. The DKN and DGN descriptors are also related to G and K kinetic energy reconstructions. DKN.STDN and DKN.W27 which describe the rate of change in the K kinetic energy density normal to and away from the surface of the molecule are correlated to both hydrophobicity and polarizability. DGN.AVGN and DGN.W27, which characterize the rate of change in the G kinetic energy density normal to and away from the surface of the molecule describes weak bonding interactions. DGN.B26, which is a PEST descriptor, suggests that molecular shape is important. DRN descriptors are related to the DKN and DGN descriptors, with the wavelet descriptors DRN.W0 and DRN.W4 representing the rate of fall off of the electron density.

The last two groups of TAE descriptors identified by the GA are LAPL and ANGLE descriptors. LAPL.W10, LAPL.W18, and LAPL.W31 are also wavelet descriptors. They are derived from the second derivative of the electronic energy distribution and are important in characterizing donor/acceptor relationships. ANGLE.B06 and ANGLE.B36 are pure shape descriptors, which favor planar, disk shaped molecules.

4.4. INDAN, TETRALIN AND ISOCHROMAN STUDY

The second study is similar to the first study but includes isochroman musks. Furthermore, the nonmusks were chosen to be as similar in structure to the musks as possible (see Figure 4.7). Isochroman musks differ in odor quality from indan and tetralin musks. A trained perfumer can differentiate isochromans from indans and tetralins. For this reason, the isochroman musks were treated as a separate class of musks different from the indan and tetralin musks.

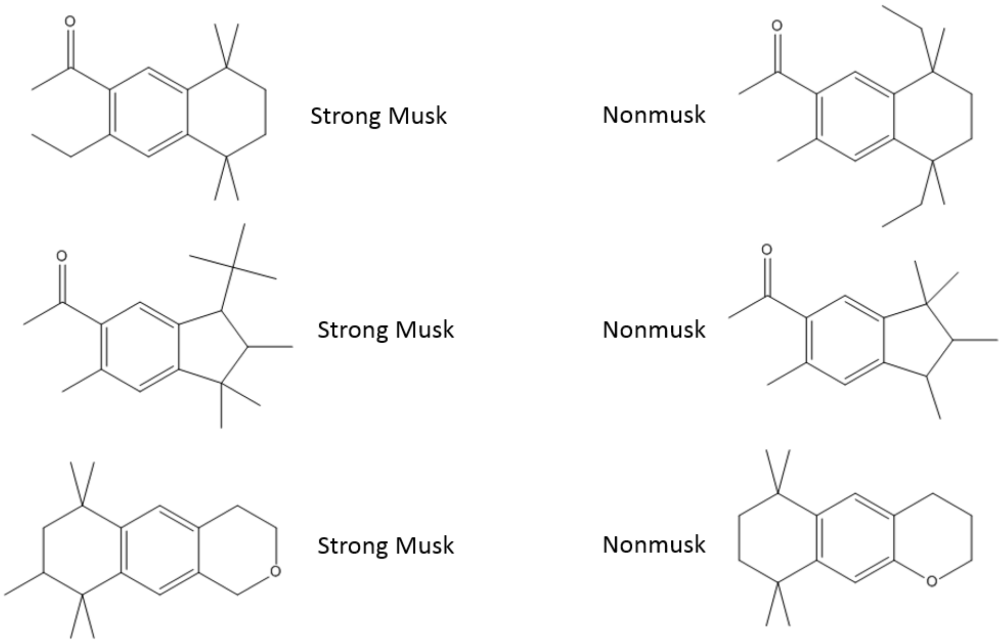


Figure 4.7. More examples small differences between strong musks and nonmusks.

4.4.1. Pattern Recognition Results. Each compound in this dataset was represented as by 1369 molecular descriptors. Figure 4.8 shows a PC plot of the 172 training set samples and 1369 molecular (TAE and CODESSA) descriptors. Each compound in the training set is represented by a point in the PC plot. The 1's are nonmusks, the 2's are indan and tetralin musks, and the 3's are isochroman musks. The overlap of the musks and nonmusks in the PC plot of the 1369 molecular descriptors is not surprising in view of the similarity of the chemical structures of the musks and the nonmusk compounds.

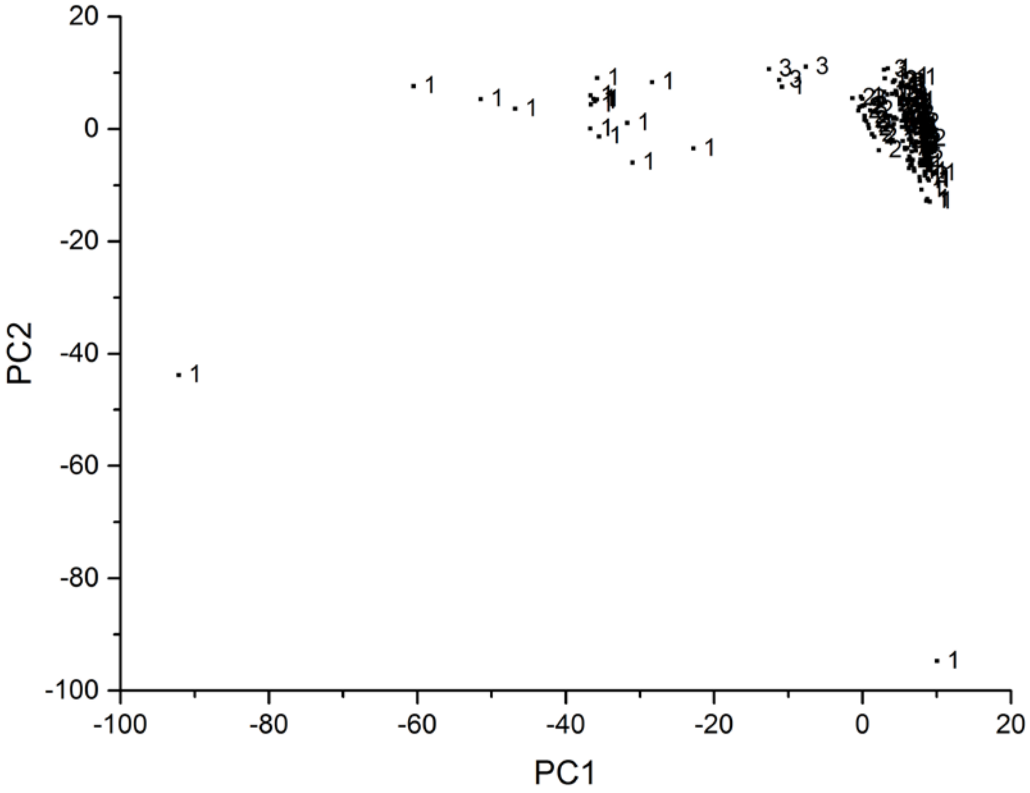


Figure 4.8. PC plot of the isochroman included training set using all 1369 descriptors. (1 = nonmusk, 2 = indan or tetralin musk, 3 = isochroman musk)

The pattern recognition GA was used to identify molecular descriptors correlated to musk odor quality for the tetralin, indan, and isochroman musks. Key descriptors were identified by sampling key feature subsets and scoring their PC plots, while simultaneously tracking those classes and compounds that were difficult to classify. The pattern recognition GA identified 21 molecular descriptors whose PC plot (see Figure 4.9) showed clustering of the compounds in the training set (see Table 4.3) on the basis of odor. For this set of GA runs, it was necessary to configure the pattern recognition GA in the asymmetric classification mode. The tetralin and indan musks represented as 2's in the PC plot of the descriptor space spanned by the 21 descriptors occupy a well-defined region in the plot and the isochroman musks (represented as 3's) also occupy a well-defined region in the PC plot. However, the nonmusks (represented as 1's) are randomly distributed in the PC plot. The two distinct clusters of compounds in the PC plot suggests that isochroman musks possess an OSR that is different from those of tetralin and indan musks.

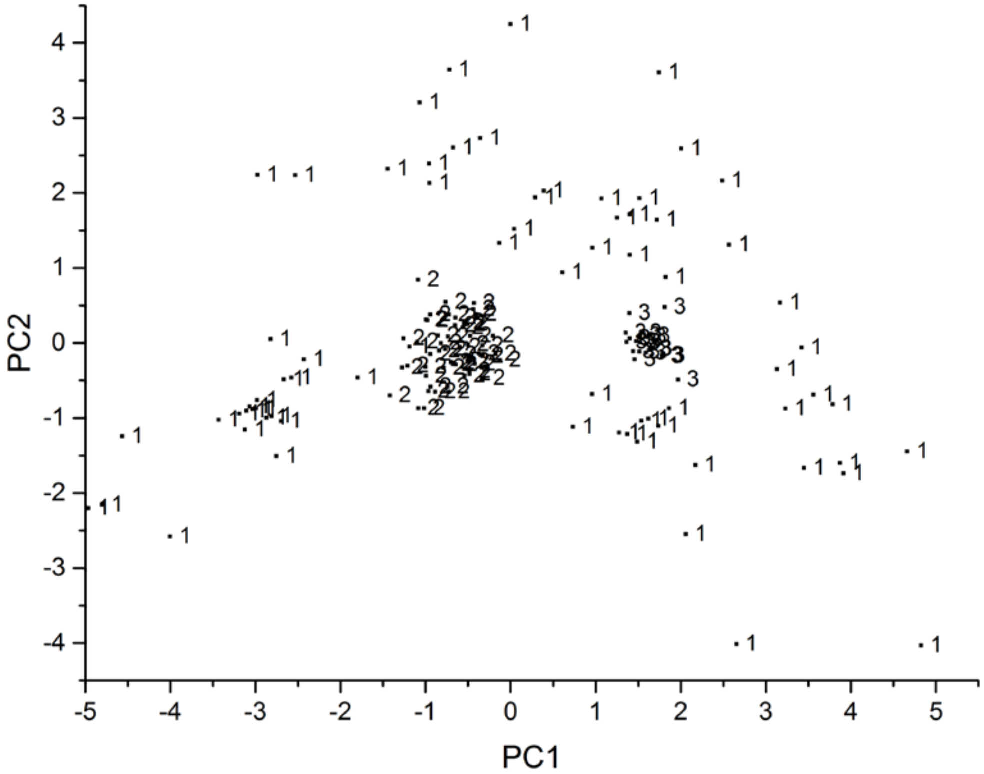


Figure 4.9. PC plot of the isochroman included training set and 21 selected descriptors. (1 = nonmusk, 2 = indan or tetralin musk, 3 = isochroman musk)

The predictive ability of the 21 molecular descriptors identified by the pattern recognition GA was assessed using a validation set consisting of 19 compounds (see Table 4.4). 11 of these compounds were nonmusks, 2 were indan musks, 2 were tetralin musks and 4 were isochroman musks. Of the 11 nonmusks, 9 were odorless and 2 had an odor other than musk. The 19 musks and nonmusks were projected onto the PC plot defined by the 172 compounds of the training set and the 21 molecular descriptors identified by the pattern recognition GA. Figure 4.10 shows the projection of the 19 validation set compounds onto the PC plot developed from the training set samples and the 21 descriptors identified by the pattern recognition GA. All the validation set samples are correctly classified, i.e., they lie in a region of the map with compounds that have the same class label. For this mapping, the validation set samples are designated as N (nonmusk) or M (musk). Tetralin and indan musks in the validation set lie in a region of the map with other tetralin and indan musks and isochroman musks from the validation set lie in a region of the map with other isochroman musks.

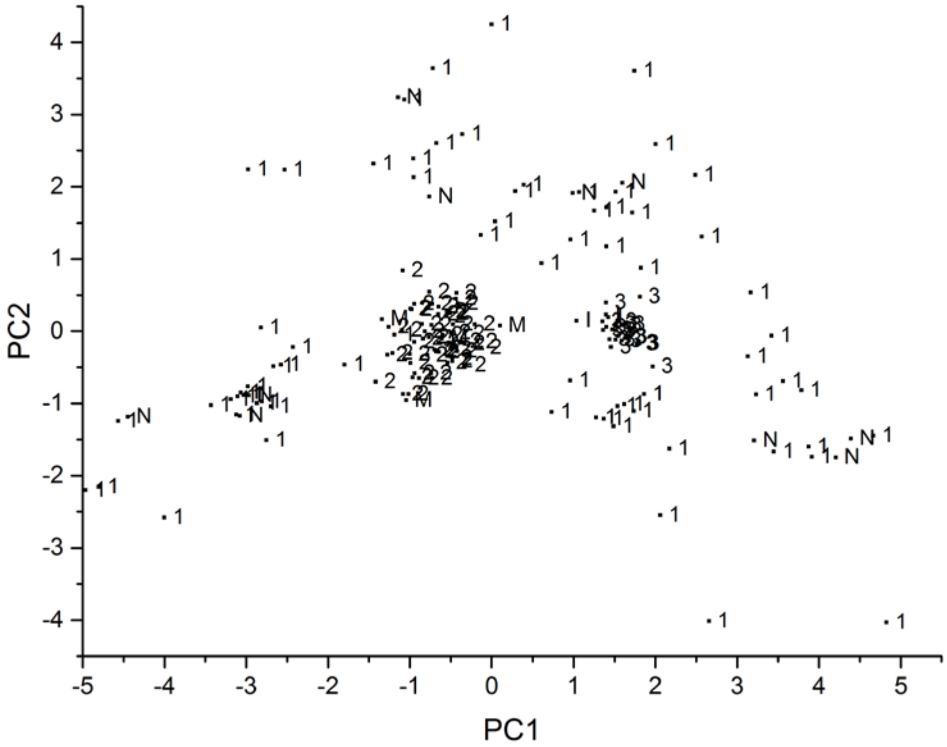


Figure 4.10. Projection of the validation set onto the PC plot from Figure 4.9. (1 = nonmusk training, 2 = indan or tetralin musk training, 3 = isochroman musk training, N = nonmusk validation, M = indan or tetralin musk validation, I = isochroman musk validation)

4.4.2. Interpretation of Selected Descriptors. Of the 21 descriptors identified by the pattern recognition GA for discrimination of the indan, tetralin, and isochroman musks, 17 were TAE descriptors and 4 were CODESSA descriptors. These 21 descriptors are listed in Table 4.6.

Table 4.6. Identities of the 21 Selected Descriptors for the Isochroman Included Study

BNP.B37	PIP.STD	PIP.STDN
PIP.W6	PIP.B20	FUK.B55
EP.W30	G.B17	K.B01
DKN.STDP	DKN.W16	DGN.W17
DGN.W23	DGN.B46	DRN.B27
DRN.B77	ANGLE.B65	# S ATOMS
MOLECULAR WEIGHT	MOMENT PRODUCT AB	YZ SHADOW

Of the 17 TAE descriptors, 9 contained information about shape. BNP.B37 captured information about the polarity and molecular shape of the compound. The 3 indicates that rays of average size are important and the 7 indicates that relatively high BNP values are important. The other 8 shape aware TAE descriptors are PIP.B20, FUK.B55, G.B17, KB01, DGN.B46, DRN.B27, DRN.B77, and ANGLE.B65. PIP.B20 and FUK.B55 are shape aware descriptors developed from local ionization potential and Fukui radical reactivity indices. G.B17 and KB01 are shape descriptors derived from the ray traces of the K and G kinetic energy reconstructions normal to and away from the surface of the molecule. DGN.B46, DRN.B27 and DRN.B77 are molecular shape descriptors that capture more of the interior volume, i.e., local shape, as opposed to conformational information. ANGLE.B64 is a pure molecular shape descriptor.

PIP.STD, PIP.STDN, AND PIP.W6 convey information about the local average ionization potential of the molecule and EP.W30 is a scale coefficient wavelet descriptor that is correlated to the solvation energy of the molecule with EP.W30 in the positive part of the EP range. DKN.STDP and DKN.W16 which describe the rate of change in the K kinetic energy density normal to and away from the surface of the molecule are correlated to both hydrophobicity and polarizability. DGN.W17 and DGN.W23 characterize the rate of change in the K kinetic energy density normal to and away from the surface of the molecule and are also correlated to both hydrophobicity and polarizability. Of the 4 CODESSA descriptors, Moment Product AB and YZ shadow are shape descriptors that favor planar disk shaped molecules. The other two CODESSA descriptors are fragment based descriptors conveying information about the number of sulfur atoms and the molecular weight of the molecules comprising the training set.

4.5. CONCLUSION

The SAR of tetralin, indan, and isochroman musks has been reported in the literature. Geometric considerations have been shown to be important for these musks. A particular skeletal arrangement, the position of the polar functional group relative to the bulky substituents in the molecule, has been shown to be correlated with the musk odor modality. Another important variable is the steric environment of the polar functional group. The presence of methyl groups near the polar heteroatom has been shown to diminish the musk odor modality presumably due to steric hindrance. Although a polar functional group must be present for indans, tetralins and isochromans in order to elicit musk odor, the introduction of a second functional group can diminish the musk odor modality unless the two functional groups are not distant from each other which would allow some form of cooperation between them. Information about centers of branching in the molecule, for example, the degree of branching of substituents attached to the nonaromatic ring, is another important variable for identifying musk odorants. Indans appear to require extensive substitution in the 5 member ring for this class of compounds to exhibit musk odor. The indans selected for this study exhibit these trends, and the PC plot of the descriptors identified by the pattern recognition GA has effectively captured this information through a graphically oriented structure-activity correlation.

The molecular descriptors identified by the pattern recognition GA contain information about the shape, and electronic surface properties of the compounds. No descriptor by itself could correctly classify more than 65% of the compounds in the training set for either study, and no two descriptors had a pair-wise correlation greater than 0.85. Therefore, one can conclude that musk odor quality does not correlate with any single molecular property. The musk odor quality of tetralin and indan musks, for example, was correlated to 45 molecular descriptors. The need for such a large pool of descriptors to differentiate indan and tetralin musks from nonmusks can be probably be attributed in some measure to the nature of the descriptors used. A more likely explanation is that musk odor is a subtle function of several molecular properties.

The main processes associated with olfaction are volatility, mucous transport, receptor binding and receptor activation. An examination of the molecular descriptors identified by the pattern recognition GA suggests that volatility and mucous transport were not important factors in this study as the molecular descriptors correlated to volatility (e.g., molecular size and lower order molecular connectivity indices) and transport (e.g., Log P and dipole moment) were not selected by the pattern recognition GA to differentiate musks from nonmusks. This result, which is in direct conflicts with the theory that diffusion rate through the mucous membrane is a major determinant of odor intensity, can probably be explained by the fact that nonmusks selected for this study were similar in structure to that of the musks. Receptor binding and activation appear to be the factors for discrimination of musks from nonmusks in these two studies. Although the descriptors selected by the pattern recognition GA may be more indicative of structural features that are lacking in the nonmusks than features involved in the mechanism governing odor quality of the indan, tetralin and isochroman musks, the asymmetric data structure encountered for the compounds in the training set compounds would suggest the opposite.

REFERENCES

4-1. J. E. Amoore, “Molecular Basis of Odor”. Charles C. Thomas, Springfield, IL, 1970.

4-2. C. Jennings-White, “Human Primary Odors”, Perfum. Flavor, 1985, 9, 46-58.

4-3. B. K. Lavine, C. G. White, N. Mirjankar, C. M. Sundling, C. Breneman, “Odor-Structure Relationship Studies of Tetralin and Indan Musks”, Chemical Senses, 2012, 37, 723-736.

4-4. C. Diete, “Manufacture of Perfumery”. Henry Carey & Baird Company, Philadelphia, PA, 1982.

4-5. B. K. Lavine, C. G. White, “Odor-Structure Relationship Studies of Indan, Tetralin, and Isochroman Musks”, ACS Symposium Series, 2015, 1191, 333-359.

4-6. K. J. Rossiter, “Structure-Odor Relationships”, Chem. Rev., 1996, 96, 3201-3240.

4-7. P. Kraft, J. A. Bajfrowicz, C. Denis, G. Frater, “Odds and Trends: Recent Developments in the Chemistry of Odorants”, Angewandte Chemie Int. Ed., 2000, 39, 2980-3010.

4-8. P. Kraft, “Brain Aided Musk Design”, Chem. Biodiver, 2004, 1, 1957-1974.

4-9. G. Frater, J. A. Bajgrowicz, P. Kraft, “Fragrance Chemistry”, Tetrahedron, 1998, 54, 7633-7703.

4-10. J. N. Narvaez, B. K. Lavine, P. C. Jurs. “Structure-Activity Studies of Musk Odorants Using Pattern Recognition: Bicyclo and Tricyclo-Benzenoids”, Chemical Senses, 1986, 11, 145-156.

4-11. G. Klopman, D. Ptscelintsev, “Application of the Computer Automated Structure Evaluation Methodology to a QSAR Study of Chemoreception. Aromatic Musky Odorants.” J. Agric. Food Chem., 1992, 40, 2244-2251.

4-12. M. Chastrette, D. Zakarya, J. F. Peyraud, “Structure-Musk Odor Relationships for Tetralins and Indans using Neural Networks (on the contribution of descriptors for classification)”, Eur. J. Med. Chem., 1994, 29, 343-348.

4-13. D. Cherqaoui, M. Esseffar, Villemin, J. M. Cense, M. Chastrette, D. Zakarya, “Structure-Musk Odor Relationship of Tetralin and Indan Compounds Using Neural Networks”, New Journal of Chemistry, 1998, 22, 839-843.

4-14. B. K. Lavine, C. E. Davidson, A. J. Moores, “Innovative Genetic Algorithms for Chemoinformatics”, Chem. Intell. Lab. Instrumen., 2002, 60, 161-171.

4-15. B. K. Lavine, C. E. Davidson, C. Breneman, W. Katt, “Electronic van der Waals Surface Property Descriptors and Genetic Algorithms for Developing Structure-Activity Correlations in Olfactory Databases”, J. Chem. Inf. Comput. Sci., 2003, 43, 1890-1905.

4-16. B. K. Lavine, C. E. Davidson, C. Breneman, and W. Katt, “Genetic algorithms for clustering and classification of olfactory stimulants”, in J. Bajorath (Ed.), Chemoinformatics: Methods and Protocols. Humana Press, Totowa, NJ, 2004, 399-426.

4-17. B. K. Lavine, K. Nuguru, N. Mirjankar, “One Stop Shopping - Feature Selection, Classification, and Prediction in a Single Step”, J. Chem., 2011, 25, 116-129.

4-18. C. Breneman, T. R. Thompson, M. Rhem, M. Dung, “Electron Density Modeling of Large Systems Using the Transferable Atom Equivalent Method”, Comput. Chem., 1995, 19, 161-172.

4-19. M. Song, C. Breneman, J. Bi, N. Sukumar, K. P. Bennett, S. Cramer, N. Tugcu, “Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression”, J. Chem. Inf. Comput. Sci., 2002, 42, 6, 1347-1357.

4-20. C. E. Whitehead, C. Breneman, N. Sukumar, M. D. Ryan, “Transferable Atom Equivalent Multi-centered Expansion Method”, J. Comput. Chem., 2003, 24, 512-529.

4-21. C. Breneman, C. M. Sundling, N. Sukumar, L. Shen, W. Katt, M. J. Embrechts, “New developments in PEST Shape/Property Hybrid Descriptors”, J. Comp. Aid. Molec. Design., 2003, 17, 231-240.

4-22. C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology. American Chemical Society, Washington, DC, 1995.

4-23. M. G. J. Beets, “Structure-Response Relationships in Chemoreception”, in C. J. Cavalitto (Ed.), Structure Activity Relationships, Pergamum Press, New York, NY, 1973.

4-24. M. G. J. Beets, “Odor, Taste, and Molecular Structure”, in: I. Morton, D. N. Rhodes (Eds.), The Contribution of Chemistry to Food Supplies, Butterworth and Co., London, England, 1974, 99-152.

4-25. M. G. J. Beets, “Structure Activity Relationships in Human Chemoreception”, Applied Science Publishers, London, England, 1978.

4-26. G. Ohloff, B. Winter, C. Fehr, “Chemical Classification and Structure-Odor Relationships”, in: P. M. Muller, D. Lamparsky (Ed.), Perfumes: Art, Science, and Technology, Elsevier Applied Science, Amsterdam, Netherlands, 1991, 310-325.

4-27. E. T. Theimer (Ed.), Fragrance Chemistry, The Science of the Sense of Smell. Academic Press, New York, NY, 1982.

4-28. I. B. Bersuker, A. S. Dimoglo, M. Y. Gorbachov, P. F. Vlad, M. Pesaro, “Origin of Musk Fragrance Activity: The Electron-Topological Approach”, New J. Chem, 1991, 15, 307-320.

4-29. T. F. Wood, Chemistry of the Aromatic Musks. Givaudanian, Clifton, NJ, 1970, 1-37.

4-30. C. Fehr, J. Galindo, R. Haubrichs, R. Perret. “New Aromatic Musk Odorants: Design and Synthesis”, Hel. Chim. Acta, 1989, 72, 1537-1553.

4-31. D. H. Pybus C. S. Sell (Eds.), The Chemistry of Fragrances. Royal Society of Chemistry. Cambridge, England, 1999.

4-32. R. Tenahsi, “Odor and Molecular Structure”, in G. Ohloff, A. F. Thomas (Eds.), Gustation and Olfaction. Academic Press, London, England, 1971.

4-33. W. J. Dunn, S. Wold, “Structure-Activity Analyzed by Pattern Recognition: The Asymmetric Case”, J. Med. Chem., 1980, 23, 595-599.

4-34. W. J. Dunn, S. Wold, “SIMCA Pattern Recognition and Classification”, in H. Van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design. VCH, New York, NY, 1995, 187.

4-35. V. S. Rose, J. Wood, H. J. H. MacFie, “Generalized Single Class Discrimination (GSCD). A New Method for the Analysis of Embedded Structure-Activity Relationships.” QSAR, 1992, 11, 492-504.

CHAPTER V

CONCLUSION

5.1. FORENSIC AUTOMOTIVE PAINT ANALYSIS

The prototype pattern recognition library search engine (search prefilters and cross correlation library searching algorithm developed for the PDQ database and described in this dissertation) is 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. There is also potential of this search engine to have direct impact with 53 local, state, and federal forensic laboratories currently using the PDQ database in the United States as well as international forensic laboratories including the National Forensic Laboratory Services Division of the RCMP, the Centre of Forensic Sciences in Toronto, Canada, members of the ENFSI network of European forensic science institutes, and the Australian Police Services.

The advantages of using pattern recognition techniques to search the IR spectra of the PDQ database include extraction of investigative lead information from clear coat, surfacer-primer, and e-coat layers and increased accuracy of searches as spectra from the entire database are searched. This is a significant improvement over the way searches are currently performed for automotive paints using the PDQ database. Information derived from the proposed pattern recognition searches will allow forensic scientists to quantify the general discrimination power of original automotive paint comparisons encountered in casework. Addressing these concerns is a direct response to Recommendation 3 of the National Academies' February 2009 report, "Strengthening Forensic Science in the United States: A Path Forward."

In the forensic examination of automotive paint, each layer of paint is analyzed individually by FTIR. The more unique the paint layers are, the more information is contained in the sample, and the stronger are the forensic conclusions that can be drawn. Laboratories in North America hand-section each layer and present each separated layer to the spectrometer for analysis which is time consuming. In addition, sampling too close to the boundary between adjacent layers can also be a problem as it produces an IR spectrum that is a mixture of two layers. Not having a "pure" spectrum of each layer will prevent a meaningful comparison between each paint layer or, in the situation of searching an automotive paint database, will prevent the scientist from developing an accurate hit list of potential suspects.

The Lavine research group is currently addressing these problems by collecting concatenated IR data from all paint layers in a single analysis by scanning across the cross-sectioned layers of the paint sample using an FTIR imaging microscope equipped with a linear array detector. Decatenation of the concatenated IR data can be achieved using multivariate curve resolution techniques to obtain a "pure" IR spectrum of each automotive paint layer. This approach, not only saves time and eliminates the need to analyze each layer separately, but also ensures that the final spectrum of each layer is "pure" and not a mixture. By integrating this imaging experiment including the use of multivariate curve resolution to improve spatial resolution with a prototype pattern recognition IR library searching system, the forensic examination of automotive paints can be facilitated in terms of both speed and accuracy.

5.2. STRUCTURE ACTIVITY CORRELATIONS OF MUSKS

Utilizing large existing olfactory databases available through the open scientific literature, a new structure/activity correlation methodology to facilitate the intelligent design of new odorants with specialized properties has been developed. In the first step, each molecule in the database is characterized by an appropriate set of molecular descriptors. To accomplish this task, an enhanced version of Breneman's Transferable Atom Equivalent descriptor methodology was used to create a large set of electron density derived shape/property hybrid, wavelet coefficient and TAE histogram descriptors. These molecular property descriptors have been chosen to represent the problem because they have been shown to contain pertinent shape and electronic properties of the molecule and correlate with key modes of intermolecular interactions.

Traditional QSAR methodologies, which employ fragment based descriptors, have been shown to be effective for QSAR development within homologous sets of molecules but are less effective when applied to datasets containing a great deal of structural variation. In contrast to previous attempts at structure-activity relationships, our use of shape-aware electron density based molecular property descriptors has removed many of the limitations brought about by the use of descriptors based on substructure fragments, molecular surface properties, or other whole molecule descriptors. Another reason for the mixed success of past QSAR efforts can be traced to the nature of the underlying modeling problem, which is often quite complex. To meet these challenges, a genetic algorithm for pattern recognition analysis has been developed that selects descriptors which create class separation in a plot of the two largest principal components of the data while simultaneously searching for features that increase clustering of the data. The efficacy of this methodology was successfully validated in the two studies described in this dissertation.

VITA

Collin Gareth White

Candidate for the Degree of

Doctor of Philosophy

Thesis: VARIABLE SELECTION TO IMPROVE CLASSIFICATION IN STRUCTURE-ACTIVITY STUDIES AND SPECTROSCOPIC ANALYSIS

Major Field: Chemistry

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy/Education in your major at Oklahoma State University, Stillwater, Oklahoma in May, 2016.

Completed the requirements for the Bachelor of Science in Chemistry at Oklahoma State University, Stillwater, Oklahoma in May, 2010.

Experience:

Teaching/Research Assistant, Department of Chemistry, Oklahoma State University (August 2010-May 2016).

Professional Memberships:

American Chemistry Society, Phi Lambda Upsilon Honorary Chemical Society

Leadership Positions:

President, OSU Chapter of Phi Lambda Upsilon Honorary Chemical Society, 2015-2016

Vice President, OSU Chapter of Phi Lambda Upsilon Honorary Chemical Society, 2014-2015

Awards:

First Place, Oral Presentations, Automation Block, Oklahoma State University Research Day 2016.