PATTERN RECOGNITION STUDIES OF COMPLEX

SPECTROSCOPIC DATA SETS

By

TAO DING

Bachelor of Science in Pharmacy
Shenyang Pharmaceutical University, China
Shenyang, Liaoning
1992

Master of Science in Chemistry
Oklahoma State University
Stillwater, OK
2009

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
December, 2016

PATTERN RECOGNITION STUDIES OF COMPLEX

SPECTROSCOPIC DATA SETS

Dissertation Approved:

Dr. Barry K. Lavine

Dissertation Adviser

Dr. John Gelder

Dr. Ziad El Rassi

Dr. Sadagopan Krishnan

Dr. Donghua Zhou

ACKNOWLEDGEMENTS

I would like to express my appreciation to my dissertation advisor, Dr. Barry K. Lavine, for his guidance and assistance while a student in his research group. I wish to also extend my appreciation to the other members of the dissertation committee: Dr. John Gelder, Dr. Ziad El Rassi, Dr. Sadagopan Krishnan and Dr. Donghua Zhou. Thanks to the Lavine research group, in particular, Nikhil Mirjankar, Collin White, Matthew D. Allen, and Nuwan Perera for their patience and assistance. Finally, I wish to thank my family for their unconditional and continuous love and support throughout my life.

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

Name: Tao Ding

Date of Degree: DECEMBER 2016

Title of Study: PATTERN RECOGNITION STUDIES OF COMPLEX SPECTROSCOPIC DATA SETS

Major Field: Chemistry

Abstract:

Profiling of complex samples using spectroscopic techniques continues to be an active area of research with a large and burgeoning literature. The overall goal of profile analysis is to correlate a characteristic fingerprint pattern in a spectrum with the properties of a sample or in biomedical studies with the presence or absence of disease in a patient or animal from which the sample was taken. Fingerprinting experiments of this type often yield profiles containing hundreds of constituents. Multivariate statistical and pattern recognition techniques can be effective methods for the analysis of such complex data. However, the classification of complex samples on the basis of their spectroscopic profiles is complicated by several factors: (1) confounding of the desired group information by experimental variables or other systematic variations in the data, and (2) the presence of noisy data and irrelevant variables that unnecessarily enlarge the data space and the complexity of the classification model developed from the data, an effect that tends to increase both the error rate and to reduce robustness of data-driven predictions. Several interesting projects involving these effects and methods for dealing with them are highlighted in this dissertation. In one study, the identification of N-linked glycan biomarkers in serum samples measured by MALDI-IMS-MS and analyzed by pattern recognition techniques to screen a population at risk for esophageal adenocarcinoma (EAC) is discussed. In another study, search prefilters were developed as part of a prototype pattern recognition library search system to facilitate searching of infrared spectra in the Paint Data Query database and to improve discrimination capability for automotive paint comparisons involving the original equipment manufacturer. A genetic algorithm for variable selection to improve classifications was used in both of these studies. The approach taken by the genetic algorithm for pattern recognition relies heavily on graphics for the presentation of results.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Profiling of complex samples using spectroscopic techniques is an active area of research with a large and burgeoning literature. The object of profile analysis is to correlate a characteristic fingerprint pattern in a spectrum with the properties of a sample (e.g., the make and model of an automobile from a paint sample recovered at a crime scene) or in biomedical studies with the presence or absence of disease in a patient or animal from which the sample was taken. Fingerprinting experiments of this type often yield profiles containing hundreds of constituents. Objective analysis of these profiles depends upon the use of multivariate statistical methods. However, only a few studies that focus on the development of methods to handle such data have been undertaken.

Pattern recognition methods are well suited for analyzing profile data because of the attributes of the procedures. First, methods are available which assume no mathematical model but rather seek relationships which provide definitions of similarity between groups of data. Pattern recognition techniques are also able to handle high dimensional data where more than three measurements are used to describe each sample. Finally, techniques are available to select important features from a large set of measurements. This allows studies to be undertaken on systems where the exact relationships are not well understood.

Pattern recognition techniques can be effective methods for the analysis of complex profile data. However, the classification of complex samples on the basis of their spectroscopic profiles is complicated by two factors: (1) confounding of the desired group information by experimental variables or other systematic variations in the data, and (2) the presence of noisy data and irrelevant variables that unnecessarily enlarge the data space and the complexity of the classification model developed from the data, an effect that tends to increase both the error rate and to reduce robustness of data-driven predictions. Two projects involving these effects and methods for dealing with them are discussed in this dissertation.

In the first project, N-linked glycans, extracted from patient sera and healthy control individuals, were analyzed by matrix-assisted laser desorption ionization (MALDI) in combination with ion mobility spectrometry (IMS), time of flight mass spectrometry (MS) and pattern recognition methods. MALDI-IMS-MS data were collected in duplicate for 58 serum samples obtained from individuals diagnosed with Barrett's esophagus (BE, 14 patients), high-grade dysplasia (HGD, 7 patients), esophageal adenocarcinoma (EAC, 20 patients) and disease-free controls (NC, 17 individuals). A combined mobility distribution of 9 N-linked glycans was established for 90 MALDI-IMS-MS spectra (45 training set samples) and analyzed using a genetic algorithm for feature selection and classification. Two models for phenotype delineation were subsequently developed and as a result, the four phenotypes (BE, HGD, EAC and NC) were unequivocally differentiated. Next, these two models were tested against 26 blinds (13 serum samples). The model based on the original concatenated ion distribution data allowed for the correct phenotype prediction of

20 blinds and the model based on the wavelet transformed ion distribution profiles correctly predicted 23 of the 26 blinds. Although applied to a limited number of blind samples, this methodology appears promising as a means of discovering molecules from serum that may have capabilities as markers of esophageal disease.

In the second study, a prototype infrared (IR) search engine has been developed to search the Paint Data Query (PDQ) database, the largest automotive paint database in the world, to identify the line and model of an automotive vehicle using the IR spectra of the clear coat, surfacer-primer, and e-coat layers in an effort to improve discrimination capability and permit quantification of discrimination power for automotive paint comparisons involving the original equipment manufacturer (OEM). Multi-layered automotive paint fragments, which are one of the most complex materials encountered in the forensic science laboratory, provide crucial links in criminal investigations and prosecutions. To determine the origin of these paint fragments, forensic automotive paint examiners have turned to PDQ, which allows the forensic examiner to compare the layer sequence and color, texture and composition of the sample to OEM paint systems. However, modern automotive paints have a thin color coat and this layer on a microscopic fragment is often too thin to obtain accurate chemical and topcoat color information. As part of this study, search prefilters have been developed for the IR spectral libraries of the PDQ database in an effort to improve discrimination capability and permit quantification of discrimination power for OEM automotive paint comparisons. The similarity of IR spectra of the corresponding layers of various records for original finishes in the PDQ database often results in poor discrimination using commercial library search algorithms. A pattern recognition approach employing search prefilters has been employed to significantly improve the discrimination of IR spectra in the PDQ database and thus improve the accuracy of a search. This improvement permits inter-comparison of OEM automotive paint layer systems using IR spectra alone. Such information can serve to quantify the discrimination power of the original automotive paint encountered in casework and further efforts to succinctly communicate trace evidence to the courts.

The two studies described in this dissertation share several common attributes. First, the data sets from these two studies were underdetermined, i.e., there were more features than samples. Second, the data sets investigated were redundant, i.e., the measurement variables were highly correlated.   Third, variable selection was crucial for the successful development of the pattern classifier. The fundamental problem investigated in this dissertation, variable selection involving collinear and underdetermined data sets to improve modeling, is currently a problem of great interest in process monitoring, quality assurance and quality-by-design applications. The anticipated broader impact of the dissertation research will be to reduce the cost-of-ownership of classification models. This need is expected to increase as more processes are monitored by spectroscopic methods in efforts to improve control or to monitor quality more closely.

CHAPTER II

PATTERN RECOGNITION METHODOLOGY

2. 1. OVERVIEW OF PATTERN RECOGNITION

Pattern recognition is a collection of methods to categorize samples on the basis of regularities in observed data. Pattern recognition methods were originally developed to solve the class membership problem (e.g., differentiating between different disease states using constituents measured from a serum sample). In a typical pattern recognition study, samples are categorized according to a specific property using measurements indirectly related to the property of interest. An empirical classification rule is developed using a collection of objects for which the specific property of interest is known, i.e., a training set. This relationship or classification rule is then used to predict this property in objects that are not part of the original training set. A pattern is a set of measurements that characterize a test sample, whereas recognition is the process of assigning a sample to its respective class.

In a pattern recognition study, each object or sample is represented as a point in a high dimensional measurement space. The number of dimensions of the space corresponds to the number of measurements that are available for each object. A basic assumption is that distances between pairs of points are inversely related to their degree of similarity. Points representing objects from one class will cluster in a limited region of this space distant from the points corresponding to the other class. Pattern recognition is a collection of methods to investigate data represented in this manner for the purpose of assessing the general structure of the data space. The structure of a data space is defined as the overall relation of each object to every other object in the data set.

To apply pattern recognition techniques to a data set, there are a series of operations that must be performed. A summary of these operations and the pattern recognition techniques used in the two studies described in this dissertation are covered in the remaining sections of this chapter. Specific emphasis is placed on the application of techniques to problems in profile analysis.

2.1.1. Data Representation

The first step in a pattern recognition study is to convert the raw data into a string of scalar measurements comprising a pattern vector: $X = (x_1, x_2, x_3, \ldots\ldots\ldots x_N)$. Each component of the pattern vector represents a physically measureable quantity. For an infrared spectrum from an automotive paint sample, each component of the pattern vector is the absorbance or transmittance at a specified wavelength. The pattern vectors, in turn, are arranged in the form of a data matrix. Each row of the data matrix represents an observation or sample and the columns represent the values for each measurement.

$$\begin{bmatrix} X_{11}, & X_{12}, & X_{13}, & \ldots\ldots X_{1N} \\ X_{21}, & X_{22}, & X_{23}, & \ldots\ldots X_{2N} \\ X_{31}, & X_{32}, & X_{33}, & \ldots\ldots X_{3N} \\ . & . & . \ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots & . \\ X_{M1}, & X_{M2}, & X_{M3} & \ldots\ldots X_{MN} \end{bmatrix} \tag{2.1}$$

It is crucial that features encode the same information for all samples in the data matrix. If the second measurement in the data matrix, for example, is the transmittance for a specific wavelength that corresponds to the carbonyl in sample one, it must also be the transmittance for the carbonyl in samples 2, samples 3, ….. M. Hence, alignment is crucial when spectra (either infrared or mass spectra) are translated into data vectors. Peak matching can be a challenging problem in the case of two-dimensional mass spectrometry or infrared spectroscopy involving spectra from different instrument manufacturers.

2.1.2 Data Preprocessing

The next step is preprocessing. The preprocessing procedures used for a given data set will depend upon the nature of the problem and the attributes of the data. Preprocessing is crucial for a successful analysis of a data set using pattern recognition techniques. This aspect of pattern recognition has not been adequately investigated. In the two studies discussed in this dissertation, the preprocessing procedures used include scaling (e.g., normalization and auto-scaling) and transformations (e.g., wavelets).

Normalization involves setting the sum of the squares of the components of each data vector to the same arbitrary constant. In other words, all data vectors will have the same length. This operation is typically performed with absorption data in infrared spectroscopy. In mass spectrometry, each component of the data vector is normalized by taking the ratio of each variable to the measurement variable with the largest intensity. Scaling techniques, such as normalization, focus the pattern recognition analysis on questions about variations in relative composition of samples rather than absolute concentration measures.

Auto-scaling involves adjusting the measurements such that each has a mean of zero and variance of unity (see Equation 2.2). This scaling technique removes any inadvertent weighting of the variables that would otherwise arise due to differences in magnitude among the measurements comprising each pattern vector. After auto-scaling, all of the measurements will have equal weight and therefore an equal effect in the analysis.

$$x_{i,new} = \frac{x_{i,orig} - m_i}{s_{i,orig}} \qquad (2.2)$$

$m_{i,orig}$ = mean of the original measurement

$s_{i,orig}$ = standard deviation of the original measurement

The wavelet transform resolves overlapping spectral responses while simultaneously reducing the noise. Mother wavelets used in analytical chemistry to interpret data include the Haar, Daubechies, Symlet and Coilet. The criterion for selection of the mother wavelet for a specific data set is based on comparing the shape of the mother wavelet to that of the bands comprising the spectra. Both the Symlet and Daubechies mother wavelets are effective for preprocessing infrared and mass spectral data. Symlets are more symmetrical than Duabeshies and Symlets were selected for the two studies described in this dissertation.

In the two studies described in this dissertation, Symlets were computed using the discrete wavelet transform [2-1]. For applying any wavelet to multivariate data, it is necessary to specify the level of decomposition and the filter size. For example, 8Sym6 refers to the 8th level of decomposition with a filter size of 6 for decomposition of spectral data into its constituent frequencies using the Symlet mother wavelet. The filtering process used by the discrete wavelet transform is summarized in Figure 2.1. Each sample spectrum is decomposed into wavelet coefficients representing both the high and low frequency components of the signal. The high-pass filter will only allow for the high frequency component of the signal (known as the detail coefficients) to pass, while the low frequency component of the signal can only be transmitted through the low-pass filter (known as the approximation coefficients), see Figure 2.2. Wavelet coefficients from all nodes in the tree (see Figure 2.1) from each sample spectrum are organized into a vector for further data analysis [2-2].

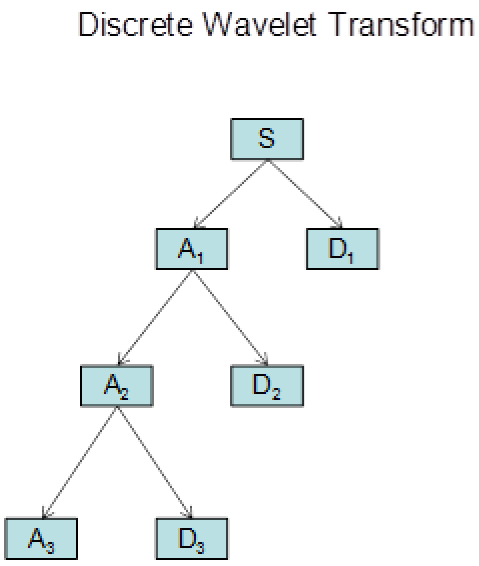


Figure 2.1. The diagram of discrete wavelet transform of original signals S to give approximations A_n and details D_n (n = decomposition level)

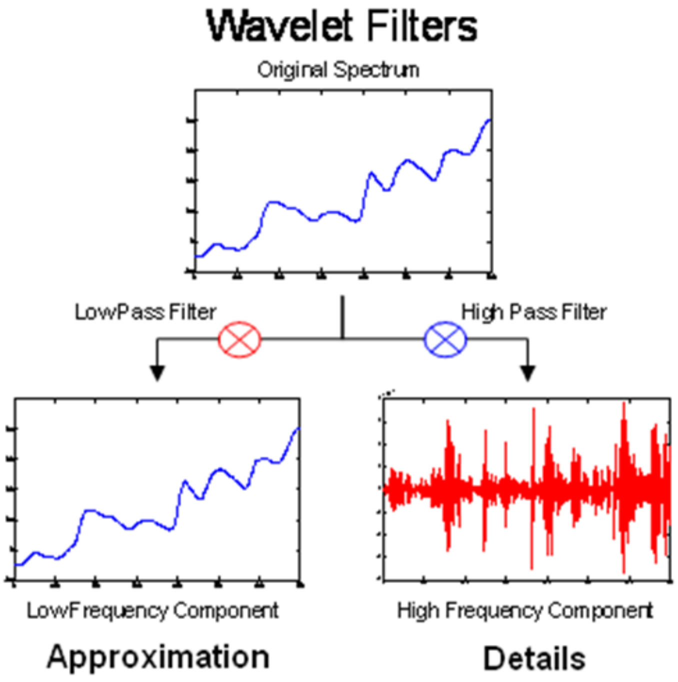


Figure 2.2. Decomposition of a sample spectrum using wavelet filters

2.2. PRINCIPAL COMPONENT ANALYSIS

In the two studies highlighted in this dissertation, principal component analysis [2-3 – 2-5] played a central role in classification, variable selection, and prediction. Principal component analysis is the most widely used multivariate analysis method in science and engineering. The goal of principal component analysis is to reduce the dimensionality of a data set while preserving the information present in the original data. This reduction is achieved by transforming the original measurement variables into new variables called principal components. Each principal component can be expressed as 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 data sets in which there are a large number of interrelated measurement variables.

Dimensionality reduction occurs using principal component analysis because of correlations between the measurement variables. Consider a set of samples characterized by two measurements, x_1 and x_2. Figure 2.3 shows a plot of these samples in a 2-dimensional measurement space. The coordinate axes (or basis vectors) of this measurement space are the variables x_1 and x_2. There appears to be a relationship between these two measurement variables. This relationship suggests that x_1 and x_2 are correlated, since fixing the value of x_1 limits the range of values possible for x_2.

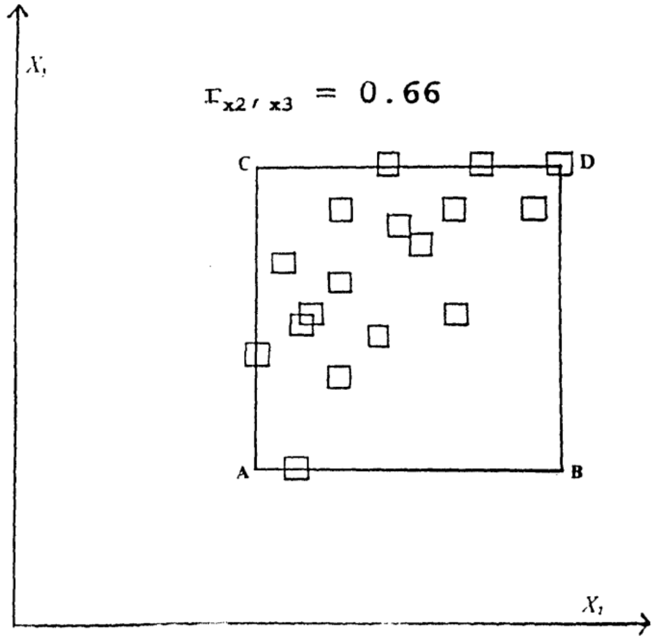


Figure 2.3. Plot of 16 samples in a measurement space defined by the variables x_1 and x_2 which are correlated. (Adapted from *NBS J. Res.*, 1985, 190(6), 465-476)

If the two variables, x_1 and x_2, were uncorrelated, the enclosed rectangle in Figure 2.3 would be fully populated by data points. Because information is defined as the scatter of points in a measurement space, it is evident that correlations between the variables decrease the information content of the measurement space. The data points are restricted to a small region of the measurement space due to correlations between the variables and may even reside in a subspace when the measurement variables are highly correlated (see Figure 2.4). Variables that are highly correlated or have a great deal of redundancy are called collinear.

$$\mathbf{X} = \begin{bmatrix} 0.5 & 0.5 & 1.0 \\ 1.9 & 0.7 & 2.6 \\ 2.0 & 2.0 & 4.0 \\ 0.3 & 1.8 & 2.1 \\ 1.9 & 1.7 & 3.6 \\ 1.2 & 0.2 & 1.4 \\ 1.9 & 0.9 & 2.8 \end{bmatrix}$$

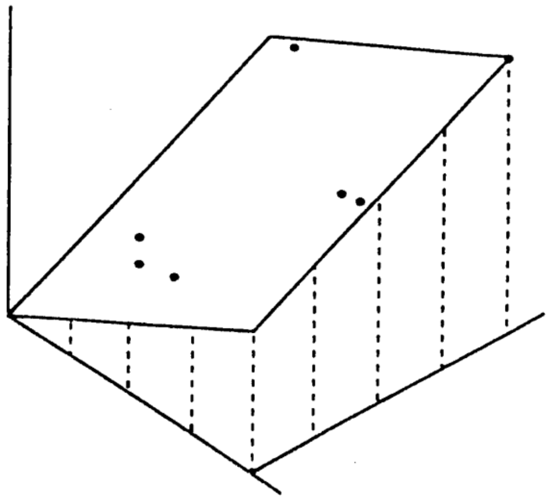


Figure 2.4. The three measurement variables used to characterize the six samples are highly correlated as the addition of the first two columns (variables) of the data matrix yields the third column (third measurement variable). (Adapted from *Multivariate Pattern Recognition in Chemometrics*, Elsevier Science Publishers, Amsterdam, 1992.)

High collinearity between variables - as measured by their correlation or covariance - is a strong indication that a new set of basis vectors can be found that is better at conveying the information content present in the data than axes defined by the original measurement variables. The 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. The largest principal component is formed by determining the direction of largest variation in the original measurement space and modeling it using a line fitted by linear least squares (see Figure 2.5). The second largest principal component lies in the direction of next largest variation: it passes through the center of the data and is orthogonal to the largest principal component. The third largest principal component lies in the direction of next largest variation: it passes through the center of the data and is orthogonal to the first and second largest principal components, and so forth. The number of principal components that can be extracted from the data is the smaller of either the number of samples or number of measurements in the data set, as this number defines the largest number of independent variables in our data.

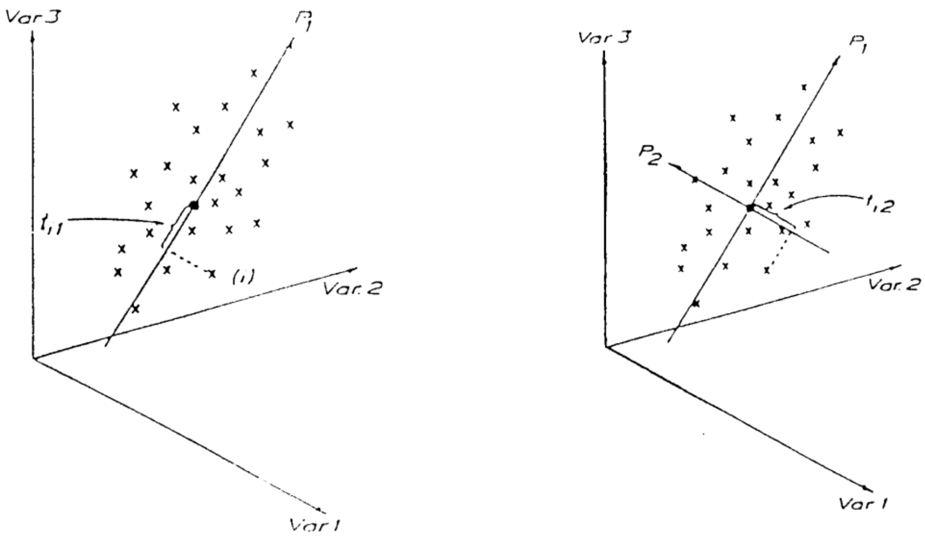


Figure 2.5. Development of a new set of basis vectors (i.e.,principal components) from the original measurement variables. (Adapted from Chemometrics: Mathematics and Statistics in Chemistry, NATO ASI Series, D. Reidel Publishing Co., 1983.

One measure of the amount of information conveyed by each principal component is its variance. For this reason, the principal components are usually arranged in order of decreasing variance: the most informative principal component is first, and the least informative is the last. Hence, one would expect that only the first few principal components should convey information about the signal, if the data are collected with due care, since most of the information in the data should be about the effect which we seek to study. The situation, however, is not always so straightforward. Each principal component describes some amount of signal and some amount of noise in the data because of accidental correlation between signal and noise. The larger principal components primarily describe signal, whereas the smaller principal components essentially describe the noise. When smaller principal components have been deleted, noise has been discarded from the data, but so has a small amount of signal. However, the gain in signal to noise more than compensates for the biased representation of the signal that results from discarding principal components that contain a small amount of signal but a large amount of noise. This approach to describing a data set in terms of important and unimportant variation is known as soft modeling in latent variables.

Principal component analysis takes advantage of the fact that a large amount of data is generated in a pattern recognition study. The data have a great deal of redundancy and therefore a great deal of collinearity. Because the measurement variables are correlated, 100-point spectra do not necessarily require 100 independent axes to define the position of the sample points. By employing principal component analysis, the original measurement variables, which constitute a correlated axis system, can be converted into a system which removes correlation by forcing the new axes to be independent and orthogonal, a requirement that greatly simplifies the data because the correlations present in the spectral data usually allows us to use far fewer axes to represent the sample points. Spectra for a set of automotive paints may reside in a subspace of the original 100-dimensional measurement space for the spectra, and a plot of the two or three largest principal components of the data can help us to visualize the relative position of the samples in this subspace.

2.3 GENETIC ALGORITHM FOR VARIABLE SELECTION

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

A potential solution to these problems is variable selection [2-6]. Irrelevant features can introduce so much noise that a good classification of the data cannot be obtained. When these irrelevant features are removed, a clear and well-separated class structure in the data can be found. The deletion of irrelevant variables is, therefore, a major goal of any pattern recognition study since noisy variables increase the chances of false classification and decrease the classification success-rates obtained with new data. Feature selection is also necessary because of the sheer size of many classification problems, e.g., DNA array data, which consists of thousands of descriptors per observation but only 50 or 100 observations distributed equally between two classes.

The approach to feature selection used in the two studies described in this dissertation is based on a simple idea - identify a set of measurement variables that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these features will be about differences between classes in the data set. This idea is demonstrated in Figure 2.6, which shows a plot of the two largest principal components of a data set prior to feature selection. The data set consists of 30 samples distributed between 3 classes (good, better, and best). Each sample is characterized by 10 measurements. However, only 4 of these measurements contain information about the classification problem. When a principal component map of the data is developed using only these 4 measurements, sample clustering on the basis of class is evident.

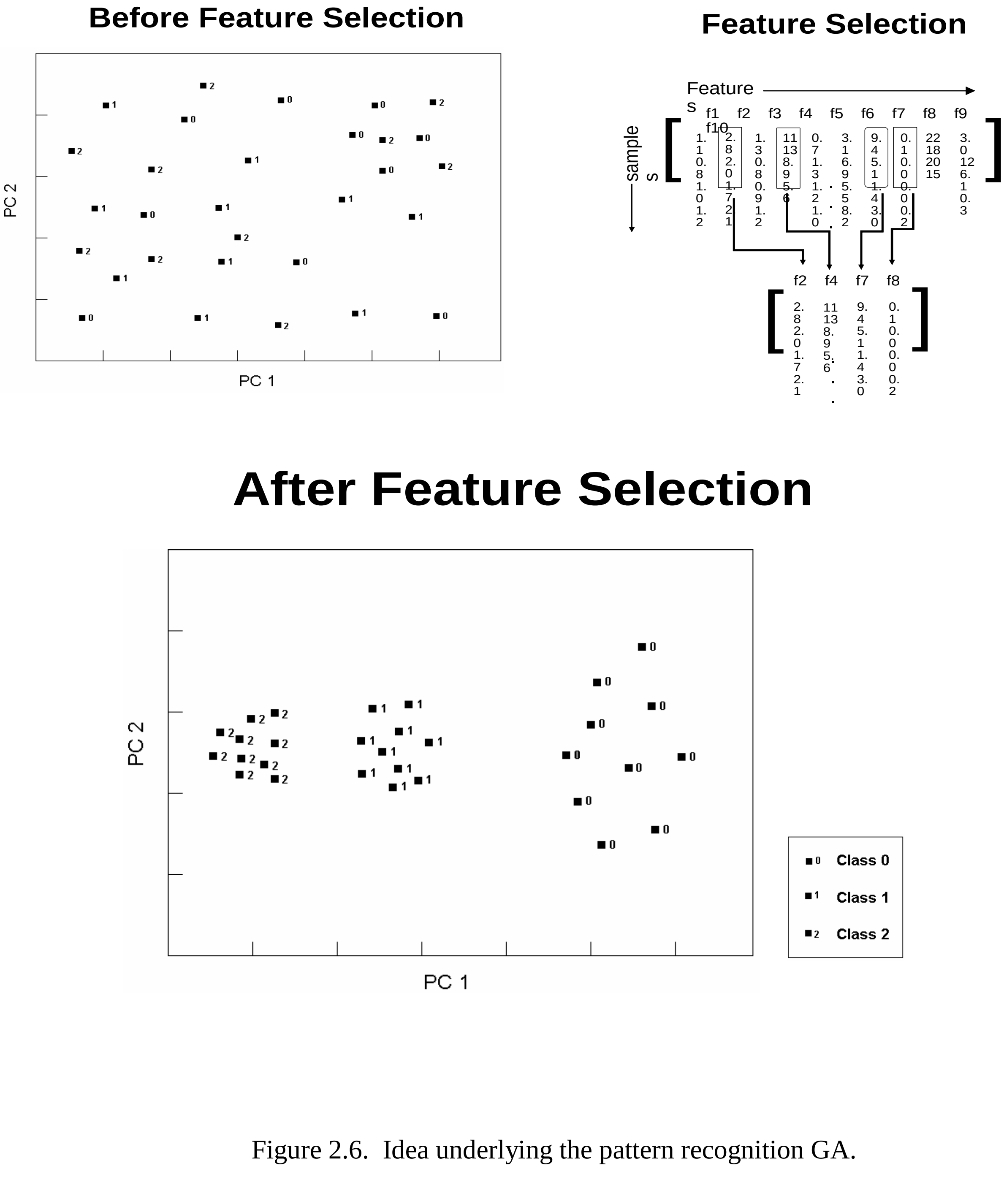


Figure 2.6. Idea underlying the pattern recognition GA.

Using this approach to feature selection, an eigenvector projection of the data is developed that discriminates classes in a data set by maximizing the ratio of between- to within-group variance (which is the same criterion used in canonical variate analysis to develop projections of the data for classification). This approach to feature selection has a number of advantages. It avoids overly complicated solutions that do not perform as well on the prediction set because of over-fitting, which is a serious problem with most wrapper methods [2-7]. Although a principal component plot is not a sharp knife for discrimination, if we have a principal component plot that shows clustering, then our experience is that we will be able to predict robustly using this set of descriptors. For redundant features, noise reduction and better class separation can be achieved if principal component analysis is used to characterize the information content of the redundant measurement variables. Furthermore, a principal component plot displays variability between large numbers of samples and shows the major clustering trends present in the data; the user can visually identify the presence of confounding relationships in the data, thereby gaining insight into how the decision is made for a classification. Although filters, which select variables by ranking them using either the Fisher ratio or the variance weight [2-8] are preferred by many workers because of their computational and statistical scalability, the variables selected by filters are usually not optimal for a given predictor because they score features individually and independent of each other and as such cannot determine which feature combinations give the best classification results.

To identify these features, a genetic algorithm [2-9], which exploits knowledge contained in a population of solutions (i.e., feature subsets) to generate new and better feature subsets while simultaneously using random choice as a tool to guide a highly exploitive search of the data, has been employed in the two studies described in this dissertation. A block diagram of the pattern recognition GA used in the two studies described in this dissertation is shown in Figure 2.7. Selected feature subsets are sent to a fitness function for evaluation. The fitness function assigns a score to each feature subset, which is a measure of the quality of the feature subset for the classification problem. The score is used to select feature subsets for recombination. Feature subsets with a higher fitness score have a higher probability of being selected. Selected chromosomes undergo a structured yet randomized exchange of information, with the expectation that good solutions (i.e., feature subsets) will generate even better ones through recombination. To ensure that all features are represented in the population at any given time, a mutation operator is used to fine tune the diversity of the population. A feature subset marked for mutation has a single random bit flipped, which allows for the pattern recognition GA to explore other regions of the search space. If the GA finds a better solution, the optimization will then continue in a new direction. A boosting algorithm adjusts the internal parameters of the pattern recognition GA for the next iteration (generation).

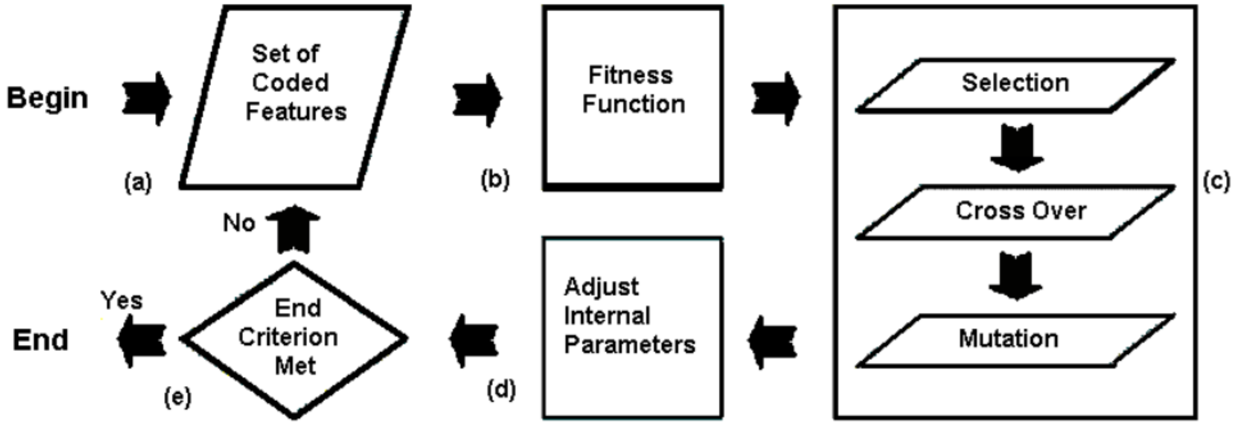


Figure 2.7. Block diagram of the genetic algorithm for pattern recognition analysis

During each generation, class and sample weights are computed as shown by Equations 3 and 4 respectively where CW(c) is the weight of class c, and SW(s) is the weight of sample s in class c. The sum of the sample weights for the spectra assigned to a particular class is equal to the class weight, and the sum of all class weights in the data set is equal to 100.

$$CW(c) = 100 \frac{CW\ (c)}{\sum_c CW(c)} \qquad (2.3)$$

$$SW(s) = CW(c) \frac{SW\ (s)}{\sum_{s\epsilon c} SW(s)} \qquad (2.4)$$

For a given data point, Euclidean distances are computed between it and every other point in the PC plot. These distances are arranged in ascending order. A survey is taken of the point's K_c-nearest neighbors. (K_c is provided by the user, and for the most rigorous classification of the data, K_c equals the number of samples in the class to which the point belongs.) The number of K_c-nearest neighbors with the same class label as the sample point in question, known as the sample hit count (SHC), is computed (0 < SHC(s) < Kc). It is then a simple matter to score a principal component plot (see Equation 2.5).

$$F(d) = \sum_c \sum_{s\epsilon c} \frac{1}{Kc} \times SHC(s) \times SW(s) \quad (2.5)$$

To better understand the scoring of the PC plots, consider a data set comprised of two classes, with each assigned equal class weights. One class has 20 samples, and the other has 40 samples. At generation 0, all classes will have the same class weight and all samples in a given class have the same sample weight. Thus, each sample in class 1 (20 samples) has a sample weight of 2.5, whereas each sample in class 2 (40 samples) has a weight of 1.25. Suppose a sample from class 1 has 9 samples from class 1 as its nearest neighbors. For this sample, SHC/K = 0.45, and (SHC/K)*SW = 0.45*2.5, which equals 1.125. By summing (SHC/K_c)*SW for all samples, each PC plot is scored. A PC plot with a higher score indicates greater separation among the classes in the variable subset from which the plot was generated.

PCKaNN is able to focus on those samples (i.e., spectra) and classes (e.g.., disease state of a patient) that are difficult to classify by boosting their sample and class weights over successive generations. In order to boost, it is necessary to calculate both the sample-hit rate (SHR), SHR is the mean value of SHC/K_c over all feature subsets produced in a particular generation (see Equation 2.6), and the class-hit rate (CHR), which is the mean sample hit rate of all samples in a class (see Equation 2.7). The variable Ø, in Equation 2.6, is the number of chromosomes in the population, whereas ∀ and AVG in equation 2.7 refer to all samples in the class and the average or mean value. During each generation, class and sample weights are adjusted using a perceptron (see Equations 2.8 and 2.9) with the momentum, P, set by a user. (g + 1 is the current generation, whereas g is the previous generation.) Classes with a lower class hit rate are boosted more heavily than classes that score well.

$$SHR(s) = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{SHCi(s)}{Kc} \qquad (2.6)$$

$$CHRg(c) = AVG(SHRg(s)\colon \forall s\epsilon c) \qquad (2.7)$$

$$CWg + 1(s) = CWg(s) + P(1 - CHRg(s)) \quad (2.8)$$

$$SWg + 1(s) = SWg(s) + P(1 - SHRg(s)) \quad (2.9)$$

Boosting is important for the successful operation of the pattern recognition GA using PCKaNN as its fitness function since it modifies the fitness landscape by adjusting the values of the class and sample weights which are an integral part of the fitness function. This mitigates the problem of convergence to a local optimum because the fitness function of the pattern recognition GA is changing as the population evolves towards a solution. Boosting minimizes the potential problem of a deceptive fitness landscape [2-10].

2.4 HIERARCHICAL CLUSTER ANALYSIS

The goal of cluster analysis is to determine the structural characteristics of a data set by organizing the data into subgroups or clusters. These methods are based on the following principle: the distance between pairs of points (samples) is inversely related to their degree of similarity. Although several different types of clustering algorithms exist, by far, the most popular is hierarchical clustering [2-11]. This particular algorithm works by measuring the distances between all pairs of points in the data set, identifying the nearest pair, combining them into a new point which is located midway between the two original points, and recalculating the distances from this new point to every other point in the data set. One then finds the new nearest pair, combines them and so on. This process is continued until all the points have been linked. The result of this procedure is a diagram called a dendogram, which is a visual representation of the sample groupings. Dendograms can be analyzed for clustering using a variety of criteria.

REFERENCES

2-1 Hayeck TJ, Kong C, Spechler S.J, Gazelle G S, Hur C, The prevalence of Barrett's esophagus in the US: estimates from a simulation model confirmed by SEER data, Dis Esophagus. 2010, 23(6): 451-457.

2-2 Walker, J. S., A Primer on Wavelets and their Scientific Applications, Chapman & Hall/CRC, Boca Raton, FL, 1999.

2-3 Chan, F-t., Liang, Y-z, Gao, J., and Shao, X-guang, Chemometrics – From Basics to Wavelet Transform, Wiley-Interscience, NY 2004.

2-4 Jackson, J. E. A User's Guide to Principal Component Analysis, Wiley Interscience, NY 1991.

2-5 Jolliffe, I. T., Principal Component Analysis, Springer Verlag, NY, 1986.

2-6 Sharaf, M. A., Illman, D. L., and Kowalski, B. R., Chemometrics, John Wiley & Sons, NY 1986.

2-7 Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. A., Feature Extraction – Foundations and Applications, Springer, Netherlands, 2006.

2-8 Guyon, I., Elisseeff, A., An Introduction to Variable and Feature Selection, J. Mach. Learn. Res., 3 (2003) 1157-1182.

2-9 Kowalski, B. R., and Bender, C. F., Pattern Recognition. Powerful Approaches for Interpreting Chemical Data, J. Am. Chem. Soc., 1972, 94, 5632-5639.

2-10 Holland, J. N., Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA. 1975.

2-11 Goldberg, D. E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley, Reading, MA 1989.

2-12 Massart, D.L, and Kaufman, L, The Interpretation of Analytical Chemical Data by the use of Cluster Analysis, Wiley, New York, 1983

CHAPTER III

DISCOVERY OF ESOPHAGEAL ADENOCARCINOMA USING MALDI-IMS-MS DATA OF SERUM N-LINKED GLYCANS

3.1 INTRODUCTION

During the past two decades, the incidence of esophageal adenocarcinoma (EAC) has increased in many countries including the United States at a rate exceeding that of any other form of cancer [3-1]. The increase in the rate of esophageal cancer can probably be attributed to gastro esophageal reflux disease (GERD) and Barrett's esophagus (BE). Approximately 30 million adults in the U.S suffer from GERD. Acid reflux causes the structure and form of the epithelial lining of the esophagus to change in order to protect the esophagus from stomach acid. Barret's esophagus (BE) occurs when stratified epithelial cells of the esophagus convert to columnar epithelial cells. It is generally accepted that EAC, which represents 60% to 90% of all esophageal cancers [3-2], develops from a premalignant lesion of the esophagus referred to as BE. Patients afflicted with BE are 30 to 125 times at greater risk for EAC than the general population, and 0.5% to 1% of patients with BE are expected to develop EAC each year [3-3]. The prevailing view is that patients with BE who eventually succumb to EAC do so by a gradual progression at the cellular level from a normal squamous cell to a metaplastic columnar cell (which is synonymous with BE). The columnar cell, in turn, may undergo a dysplastic transformation resulting in high grade dysplasia (HGD) that ultimately can result in a malignant cell. Each of these pathological states is clearly identifiable histologically, and one or more of them may be observed in the same esophagus. The precise underlying molecular mechanism by which this progression occurs has yet to be elucidated.

Less than 20% of patients with esophageal cancer survive beyond 3 years [3-4, 3-5]. Several factors contribute to this low survival rate. The most important factor is that a majority of patients demonstrate an advanced state of the disease at diagnosis. By comparison, patients diagnosed with early stage esophageal cancer have a better prognosis. Currently, there is no noninvasive test to screen patients who are at risk for the development of esophageal cancer [3-6, 3-7]. The identification of biomarkers indicative of EAC has the potential to result in an early diagnosis and to improve the outcome of patients diagnosed with EAC. Biomarkers indicative of the malignant transformation can lead to the prevention of EAC by identifying patients who are in one of the premalignant states where intervention can be carried out successfully. For example, BE can be treated with anti-reflux surgeries or drugs, such as proton pump inhibitors, while radio frequency ablation is a safe and effective treatment for HGD [3-8].

Aberrant glycosylation has been implicated in several types of cancers. The plasma membranes of cells are coated with glycol-conjugates, such as glycoproteins, glycolipids, and proteoglycans, which play an important role in normal molecular functions. N-glycans, which are the glycans attached to proteins through asparagine residues, are involved in various biological functions, such as signal transduction, cell adhesion, cell motility and proliferation. It is accepted that alterations in glycans are associated with cancer [3-9]. Moreover, they are amenable to population-screening strategies which involve low cost testing. Serum N-glycans may hold the potential to be used as esophageal cancer biomarkers to diagnose BE, HGD and EAC or for evaluating the therapeutic efficacy of cancer treatments. Identifying sensitive and specific N-glycans using mass spectrometry to monitor the progression of EAC has attracted great interest and has been the subject of several studies.

Hammoud and colleagues [3-10], using MALDI-TOF-MS, analyzed permethylated N-glycans from normal controls, BE, HGD, and EAC. The intensities of 98 glycans, 26 of which correspond to known glycan structures, were found to be characteristic of the disease state of the subjects investigated. The mean relative intensities of fucosylated N-glycans exhibited larger significant changes than sialylated ones for the four EAC phenotypes investigated. This study suggested that MS-based serum glycomic profiling had potential for future EAC prediction and diagnosis. The significance of this research lies in its highly reproducible permethylated glycan MS data. However, this study was not validated using blind samples.

In another study, Mann and coworkers [3-11] used AAL and LTA lectin affinity chromatography to explore fucosylated glycans. They determined that the ratio of the relative intensities of fucosylated glycans from serum could differentiate disease free from HGD and EAC. However, only a few samples were used in this study and the results were not validated using an external prediction set (blind samples).

Microchip electrophoresis with laser-induced fluorescence detection [3-12] was also used to analyze altered glycans released by the surface protein from esophageal cancerous cells. Serum N-glycan samples were obtained from disease-free individuals (NC) and from patients with BE, HGD, and EAC. Multivariate analysis of the microchip electrophoregrams showed that four disease phenotypes could be differentiated based on the intense peaks from both the native (50 most intense peaks) and the desialylated N-glycan samples (75 most intense peaks). However, microchip electrophoresis could not elucidate the structure of N-glycans associated with EAC nor provide information about glycan isomers.

Utilizing serum N-glycans as biomarkers for EAC disease progression will require the characterization of the chemical structure of the corresponding glycans including their isomers. Monosaccharide composition, monosaccharide linkage position and anomeric configuration can lead to a large number of potential glycan isomers. Even when MS/MS has been applied to the structural elucidation of serum glycans, identifying the chemical structure of the various glycan isomers has been proven to be problematic. This situation is further confounded when the glycans contain multi units of monosaccharides with different branching patterns [3-13]. Ion mobility spectrometry/tandem mass spectrometry (IMS-MS) has been developed to address analysis problems of this type. When compared to MALDI-TOF-MS, MALDI-IMS-MS was reported to reduce chemical noise, increase sequence coverage and allow high through put separation of a complex mixture [3-14]. In the study described in this chapter, IMS-MS combined with pattern recognition methods was shown to have advantages over MALDI-TOF-MS for biomarker discovery. The integrated drift time intensities of 9 N-glycan ions played a major role in differentiating

EAC, BE, HGD and NC, whereas MALDI-TOF analysis of a similar set of serum samples could distinguish EAC from normals but did not provide phenotype delineation of the 4 disease states (Normal, BE, HGD, and EAC) [3-15]. Based on the results of the study discussed in this chapter, it is likely that IMS-MS combined with multivariate data analysis can probably be extended to include the problem of esophageal disease phenotype delineation.

3.2. ION MOBILITY/TIME-OF- FLIGHT MASS SPECTROMETRY INSTRUMENTATION

Ion mobility/time-of-flight mass spectrometry (IMS-TOF-MS) was selected to characterize aberrant glycosylation related to the disease stage of the subject because of the attributes of the methodology. IMS-TOF-MS couples matrix-assisted laser desorption/ionization (MALDI) to an ion mobility tandem time-of-flight mass spectrometer. After a sample is injected into an IMS-TOF-MS, the ions are initially separated by differences in their mobility and then isolated by their mass-to-charge ratio in a time-of-flight mass analyzer. Individual ion mobility distributions and m/z ratios are measured independently. Because the flight times of ions through the mass spectrometer are shorter than their residence times in the drift tube, an entire mass spectrum can be obtained for ions of a specific mobility. The resulting spectrum of the N-glycan serum sample is three-dimensional as it contains information about drift time, mass-to-charge, and ion abundance. A schematic diagram of the instrument used in this study is shown in Figure 3.1 [3-15].

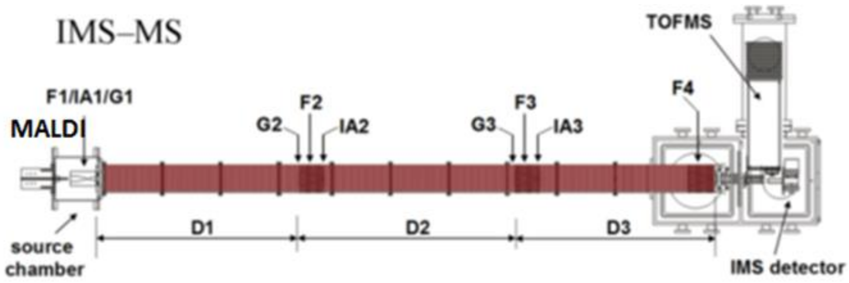


Figure 3.1. Schematic diagram of the ion mobility /time-of flight instrument.

The spectrometer was constructed in the laboratory of David Clemmer at Indiana University. In this study, the drift tube D1 and D2 were directly connected to the time of fly mass spectrometer. G1 is an ion-gate which prevents neutral species from entering the drift tube from the MALDI source, F1 is the desolvation chamber housing and hourglass ion funnel. Applying a gating voltage to the G2 ion gate will select drift (mobility) distributions. Mobility-selected ions accumulate in the ion funnel F2, which is used to focus the diffuse ion cloud to the center axis of the drift interface region. IA2 is an ion activation region to fragment ions. Drift time, flight time and collision cross sections are defined as t_D, t_F and Ω [3-16 – 3-21].

$$t_D = \frac{L}{E_D K} \tag{3.1}$$

$$t_F = l\left(\frac{m}{2ZE_{TOF}}\right)^{1/2} \tag{3.2}$$

$$\Omega = \frac{(18\pi)^{1/2}}{16} \frac{ze}{(k_B T)^{1/2}} \left[\frac{1}{m_l} + \frac{1}{m_B} \right]^{1/2} \frac{t_D E}{L} \frac{760}{P} \frac{T}{273.2} \frac{1}{N} \qquad (3.3)$$

where K is the mobility of the ions; L is the length of the drift field; E_D is the applied drift field; I is the length of the field-free region; m/m_A is the ion mass; z is the ion charge state; E_{TOF} is the kinetic energy of the ions; N is the Boltzmann constant; m_B is the buffer gas mass; P is the pressure of the drift tube; and T is the temperature of drift tube.

3.3 METHOD AND MATERIALS

A method was developed to identify the different stages of EAC using the associated serum N-glycans. The method included glycan extraction, purification, mass spectral identification and pattern recognition analysis. Blood serum samples were obtained from 116 normals or patients diagnosed at different stage of esophageal cancer (see Table 3.1).

Table 3.1 Composition of the IMS-MS Esophageal Adenocarcinoma Dataset

Sample Type	Number of Sample Spectra
Normal Controls (NC)	28
Barrett's Esophagus (BE)	20
High-grade Dysplasia (HGD)	10
Esophageal Adenocarcinoma (EAC)	32
Total training samples	90
Blinds	26

3.3.1 Experimental Materials

Materials used to analyze the serum samples were peptide-N-glycosidase F (PNGase F, EC 3.5.1.52; Sigma), ammonium bicarbonate (≥99.0%, Sigma), sodium hydroxide beads (97%, Sigma), methyl iodide (99%, Sigma), dithiothreitol (DTT, ≥98%, Sigma) and iodoacetamide (Sigma, St. Louis, MO), chloroform (99.8%, Aldrich), trifluoroacetic acid

(TFA, 99%, Aldrich), dimethyl sulfoxide (DMSO, 99.9%, J. T. Baker), micro-spin columns and C18 Sep-Pak cartridges(J. T. Baker, Phillipsburg, NJ), Harvard Apparatus (Holliston, MA and Water, Milford, MA), and β-N-acetylglucosaminidase (Endo-M,TCI, Portland, OR).

3.3.2 Sample Preparation

To obtain N-glycan from human blood serum, 10µL of human serum plasma was mixed with 200µL of 100mM ammonium bicarbonate buffer solution, with 5µL of 10mM DTT then added. The solution was incubated at 56°C for 45 min. After cooling to room temperature, 200µL of 55mM iodoacetamide prepared in 100mM ammonium bicarbonate buffer solution was added to the mixture. The sample was placed in the dark for 30 minutes. 100mM phosphate buffer was used to adjust sample solution pH to 7.5. The N-glycans were truncated from the tryptic digest using 5-mU aliquots of PNGase F and Endo-M which was added to the mixture and incubated overnight (18-22 h) at 37 °C.

C18 Sep-Pak cartridges were used to preconcentrate the glycans. The cartridges were preconditioned with ethanol and deionized water. The eluting N-glycan solution was further purified using a home-packed activated carbon microspin column, which was preconditioned with acetonitrile and equilibrated with 0.1% TFA aqueous solution. An aliquot of the diluted sample was injected into the activated carbon microspin column, which was first washed with 0.1% TFA aqueous solution. The glycans were eluted from the microspin column using 50% acetonitrile-0.1% TFA aqueous solution. Each samples was dried under vacuum and permethylated using a previously published procedure [3-22].

3.3.3 MALDI-IMS-MS Measurement

For MALDI-IMS-TOF-MS analysis, each glycan enriched sample (prepared using the procedure described above) was dissolved in 2µL methanol /water solution (1:1, v: v) and mixed with 2µL of the MALDI matrix (2, 5-dihydroxybenzoic acid) prepared at 10 mg.mL^{-1} in methanol/ water (1:1, v: v) using 2mM sodium acetate. Duplicate spotting of each serum sample was performed (2 µL each, one immediately following the other) on two 96-wells MALDI plates (Plate 1 and Plate 2), with dextran spotted after every ten samples as a control. Data were collected using a Synapt G2-S travelling wave ion mobility mass spectrometer (TWIMS) operated in positive mode. The Nd/Yag laser (355 nm) for MALDI was fired 1000 times/sec with an energy of 450au in a reverse-spiral pattern. The ion mobility cell was set with 40 V as the peak height voltage and 350 m/s as the T-wave velocity. A MassPREPTM calibration mix containing polyethylene glycol (Waters Corporation, Milford, MA) was used as an external standard.

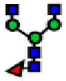
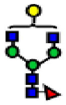
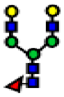
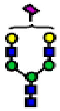
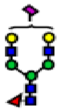
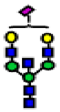
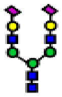
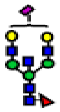
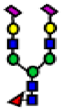
The mass to charge ratio of the time of flight mass analyzer was set from 1000 to 5000 m/z. For each sample, an entire mass spectrum was collected in three minutes and all data were collected over a 24 hour window. The N-linked glycans were extracted using Driftscope software (Waters Corporation, Manchester, UK) from a diagonal selection across the drift bin (m/z) two-dimensional spectrum (2D-plot). A single N-linked glycan ion $[M+Na]^+$ was acquired using box selection described in previous studies [3-14, 3-23, 3-24]. For each sample, the data included in a diagonal selection across mobility distributions of the selected N-linked glycans and their relative mass over charge intensity was obtained from each MALDI-IMS-MS spectrum.

3.3.4 Data Set Preparation for Pattern Recognition Analysis

The chemical structures of the 9 glycans used in this study are shown in Table 3.2. F represents fucose (red triangles), H represents hexose (mannose is green circle and galactose is yellow circle), N represents N-acetyl glucosamine (blue square) and S represents sialic acid (purple diamond). 6 out of the 9 ions depicted in Table 3.2 are fucosylated species.

Each glycan ion mobility distribution was extracted from the mass spectal image (see Figure 3.2) of a serum sample by Driftscope software (Waters Corporation, Manchester, UK) from a diagonal selection across the drift bin (m/z) two-dimensional spectrum (2D-plot). Ion intensities were represented by a color code in which blue represents the lowest intensity and red represents the highest intensity. Figure 3.3 shows the ion mobility distribution of three N-linked glycan ions extracted from the mass spectral image of four serum sample: NC, BE, HGD and EAC. If an ion distribution corresponded to a single gas phase ion, then the distribution would be Gaussian. For the ion distributions shown in Figure 3.3, the glycans may exist as distinct conformers in the gas phase or their mobility ion profiles denote the presence of structural isomers. A visual examination of the mobility distributions for these three glycans from a single individual in each phenotype group suggest differences in peak shapes, peak intensities and peak intensity ratios across disease phenotypes. Because of the number of samples (see Table 3.1) and features in the data set, a more systematic analysis of the data using pattern recognition methods is necessary to establish a correlation between N-linked glycan mobility distributions and disease phenotypes.

Table 3.2. Nine N-linked Glycans used in this Study

glycan composition[a]	*m/z*[b] $[M+Na]^+$	structure
$F_1H_3N_4$	1835.9	
$F_1H_4N_4$	2040.0	
$F_1H_5N_4$	2244.1	
$S_1H_5N_4$	2431.2	
$S_1F_1H_5N_4$	2605.3	
$S_1H_5N_5$	2676.3	
$S_2H_5N_4$	2792.4	
$S_1F_1H_5N_5$	2850.4	
$S_2F_1H_5N_4$	2966.5	

[a] F represents fucose (red triangle), H represents hexose (mannose green circle, galactose yellow circle), N represents N-acetyl glucosamine (blue square) and S represents sialic acid (purple diamond).
[b] Permethylated glycans with free reducing end.

The ion mobility distribution for each glycan from a serum sample was translated into a data vector where each element corresponds to the ion intensity at a specified drift time for a fixed m/z value. All 9 ion mobility distributions for a sample were then concatenated into a single data vector. In other words, mobility distributions of the 9 N-linked glycans were sequentially spliced together in a single mobility distribution across an arbitrary drift bin axis. Each sample data vector consisted of 1791 ion intensity values normalized to the largest peak intensity in the vector. The data set of 116 ion mobility distribution spectra, which was divided into a training set of 90 spectra and a validation set of 26 blinds (see Table 3.1), was analyzed using pattern recognition methods.

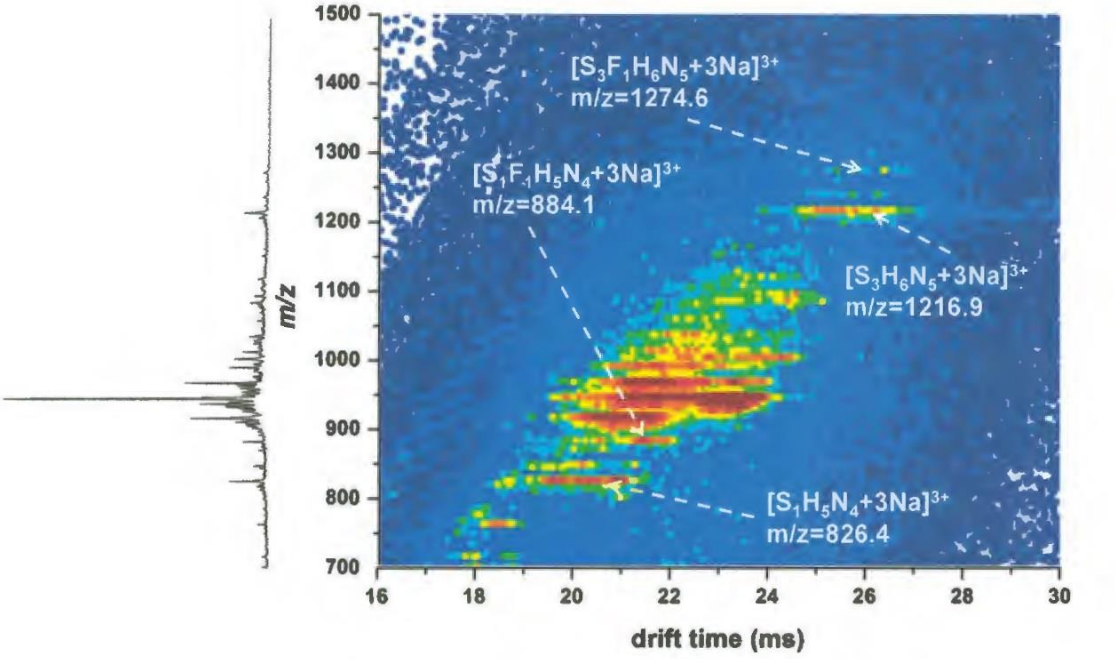


Figure 3.2. A mass spectral image of a serum sample enriched in N-linked glycans from a Normal patient. Ion intensity is shown as a function of drift time and m/z values. A color code is used to represent ion intensity with blue representing the lowest intensity and red representing the highest intensity.

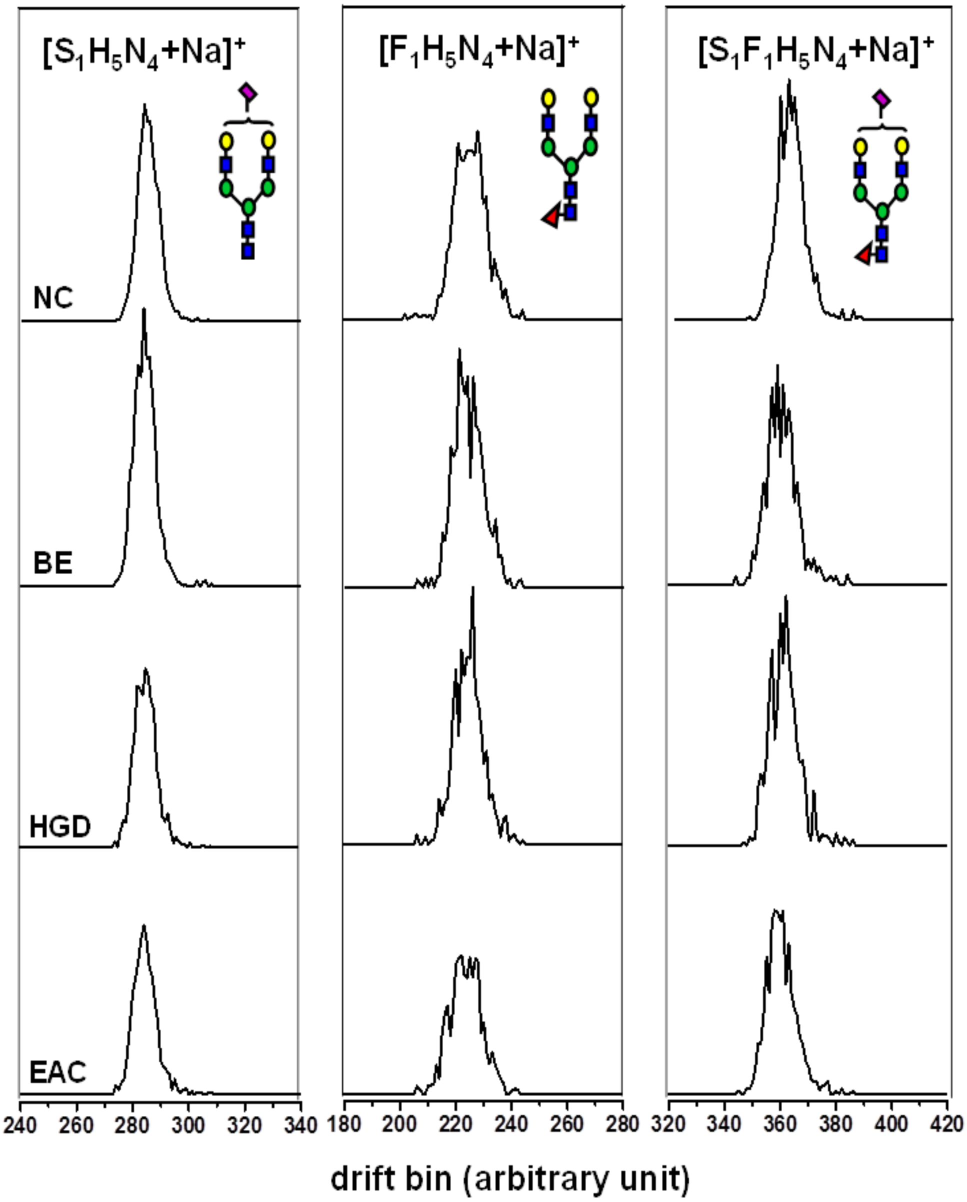


Figure 3.3. Ion mobility distribution of N-linked glycan ions $[S_1H_5N_4+Na]^+$, $[F_1H_5N_4+Na]^+$ and $[S_1F_1H_5N_4+Na]^+$. Esophageal adenocarcinoma (EAC), high grade dysplasia (HGD), Barrett's esophagus (BE) and normal control (NC) phenotypes are represented by a single individual. Glycan structures are shown as insets: F represents fucose (red triangle), H represents hexose (mannose green circle, galactose yellow circle), N represents N-acetylglucosamine (blue square) and S represents sialic acid (purple diamond).

The ion mobility distribution for each glycan from a serum sample was translated into a data vector where each element corresponds to the ion intensity at a specified drift time for a fixed m/z value. All 9 ion mobility distributions for a sample were then concatenated into a single data vector. In other words, mobility distributions of the 9 N-linked glycans were sequentially spliced together in a single mobility distribution across an arbitrary drift bin axis. Each sample data vector consisted of 1791 ion intensity values normalized to the largest peak intensity in the vector. The data set of 116 ion mobility distribution spectra was divided into a training set of 90 spectra and a validation set of 26 blinds (see Table 3.1).

3.4 RESULTS AND DISCUSSION

For each training set sample (MALDI plates 1 and 2), the corresponding 9 glycan composite mobility distribution contained 1791 drift bins. Because many of the drift bins were zero (before and after each individual glycan) or have similar intensities, only the ion intensities from 404 drift bins were considered for pattern recognition analysis. Figure 3.4 shows a plot of the two largest principal components of these 404 mass spectral features. Each training set sample is represented as a point in the PC plot. EAC samples are partially resolved from the other three phenotypes but more noticeably, three outliers are present (two NC and one HGD) in this plot. A visual examination of the composite mobility distribution data reveals that these three outliers are represented by profiles that are markedly different from the other mobility distribution profiles in the training set. For this reason, these three outliers were removed and principal component analysis was again performed on the truncated training set.

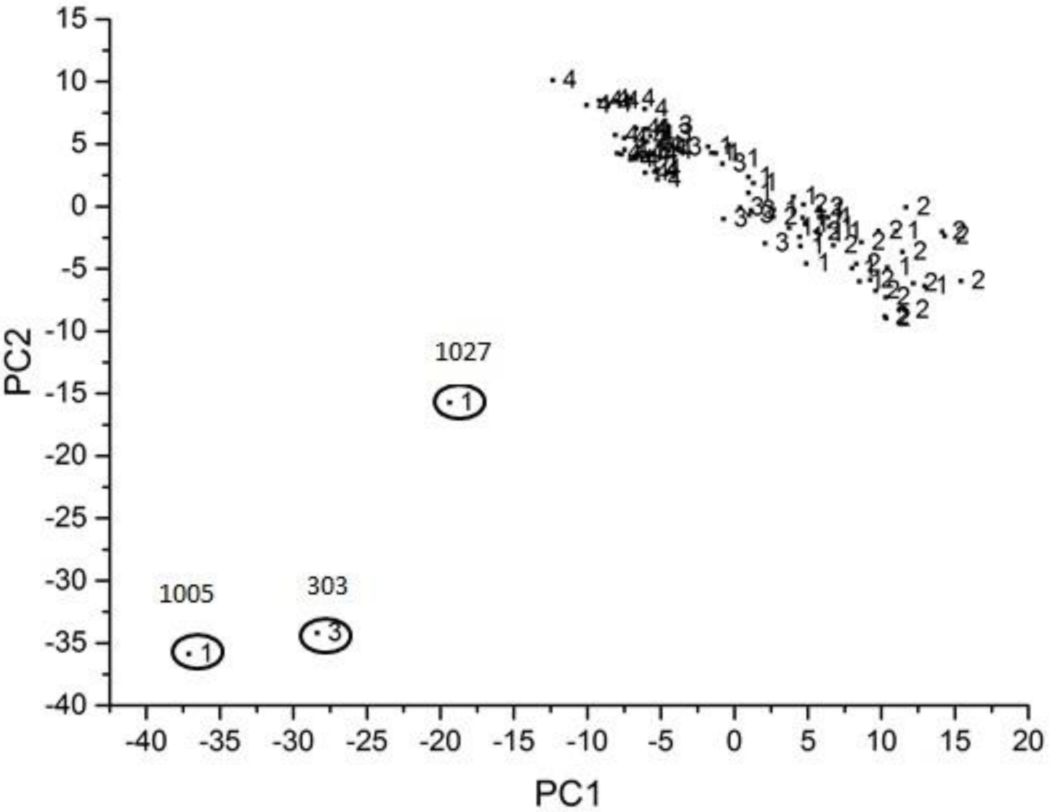


Figure 3.4. Plot of the two largest principal components of the 90 mobility distribution profiles and the 404 mass spectral features of the training set. 1 = Normal control (NC, 28 spectra), 2 = Barrett's esophagus (BE, 20 spectra), 3 = high grade dysplasia (HGD, 10 spectra) and 4 = esophageal adenocarcinoma (EAC, 32 spectra).

A plot of the two largest principal components of the truncated training set is shown in Figure 3.5. The training set can be divided into two groups (see solid line along the second principal component axis and parallel to the first principal component axis delineating the separation of the mobility distribution profiles in Figure 3.5). All samples above the solid line are from the first MALDI spot for each sample within MALDI plate 1 (duplicates were spotted back to back on the MALDI plate, the first spot labeled a, and the second spot labeled b). Examining the origin of each sample, it was observed that all samples from the second spot on plate 1 and both spots on plate 2 lie below the solid line in Figure 3.4. There were no variations in instrumental parameters within one MALDI plate or between the two plates as the instrument used was continually tuned using an external standard. The observed clustering in the PC plot is probably due to the quality of the sample spotting technique, which improved during the course of the experiment.

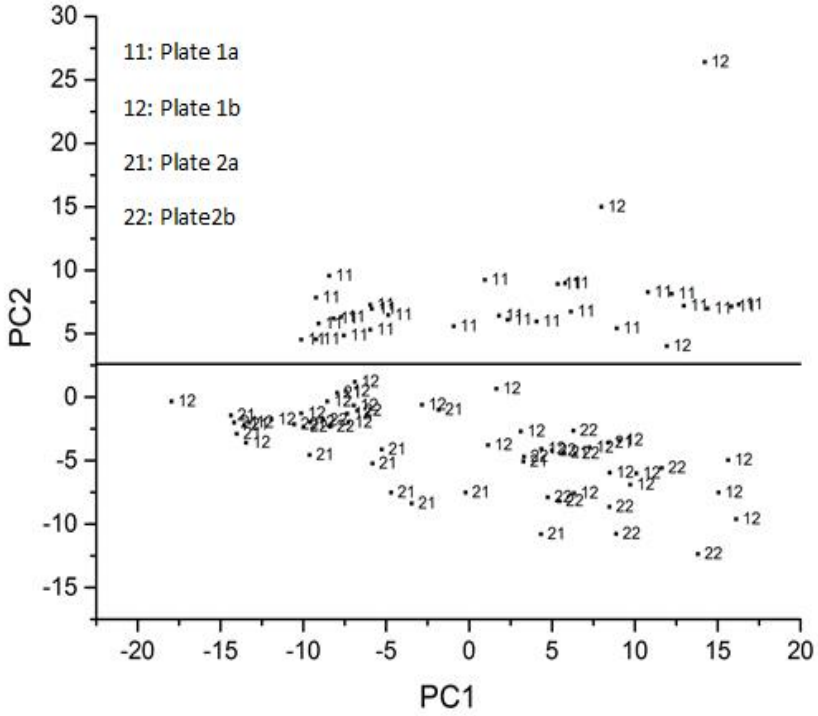


Figure 3.5. A plot of the two largest principal components of the truncated training set (with the three outliers removed from the data set) is shown. Almost all samples above the solid line are from the first MALDI spot for each sample whereas the samples from the second spot on plate 1 and both spots on plate 2 lie below the solid line.

Following this line of investigation, the first set of ion mobility distribution profiles from plate 1 (denoted as plate 1a samples) and the second set of distribution profiles from plate 1 combined with both sets of spectra from plate 2 (denoted as plate 1b, plate 2a and 2b samples) were analyzed separately using principal component analysis (see Figures 3.6 and 3.7) and the pattern recognition GA, which identified drift bin intensities characteristic of disease phenotype by sampling key feature subsets, scoring their principal component plots and tracking those samples and/or phenotypes that were difficult to classify. Separation of the different phenotype groups is observed in both principal component plots after feature selection. (see Figures 3.8 and 3.9). Although sample spotting is a major source of variation in the data, our hypothesis is that differences between phenotypes represent a larger source of variation. Because of this, data from all plates were analyzed in a single training set by the pattern recognition GA. This phase of the analysis (described below), which allowed for an evaluation of the training set samples with respect to a possible bias introduced by sample preparation, is valuable as it helps us to ensure that observed differences are due to phenotype rather than to the experimental conditions used to generate the data.

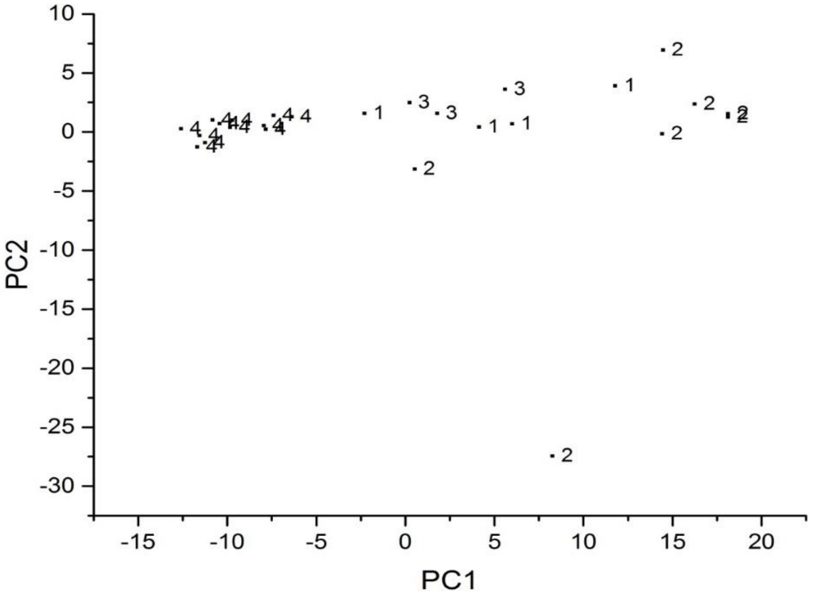


Figure 3.6. Plot of the two largest principal components of the plate 1a ion distribution profiles and 404 mass spectral features from the training set. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

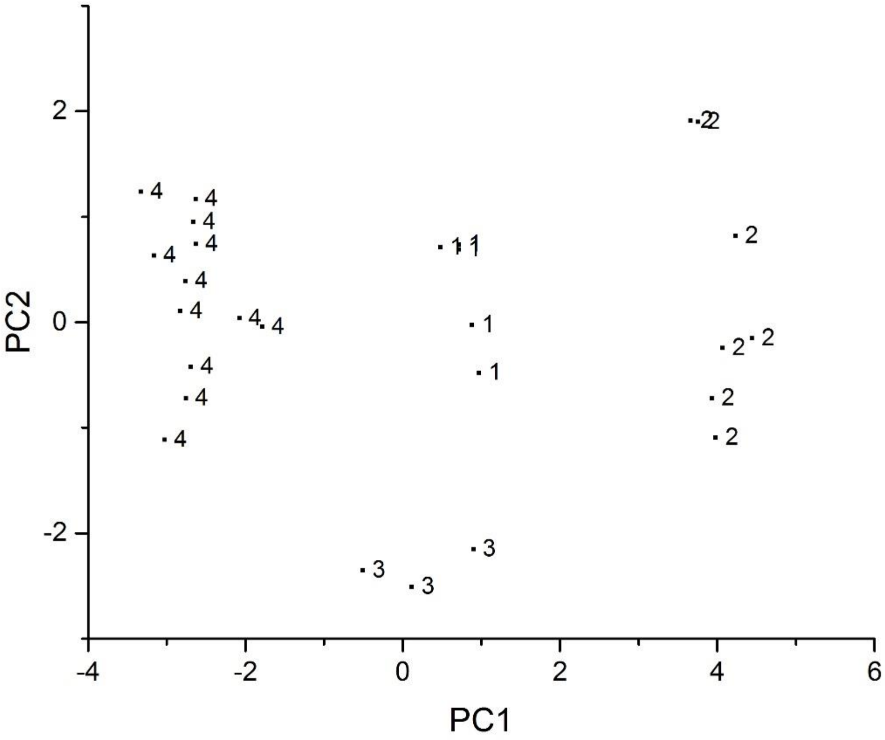


Figure 3.7. Plot of the two largest principal components of the plate 1a ion distribution profiles from the training set and the 12 mass spectral features identified by the pattern recognition GA. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

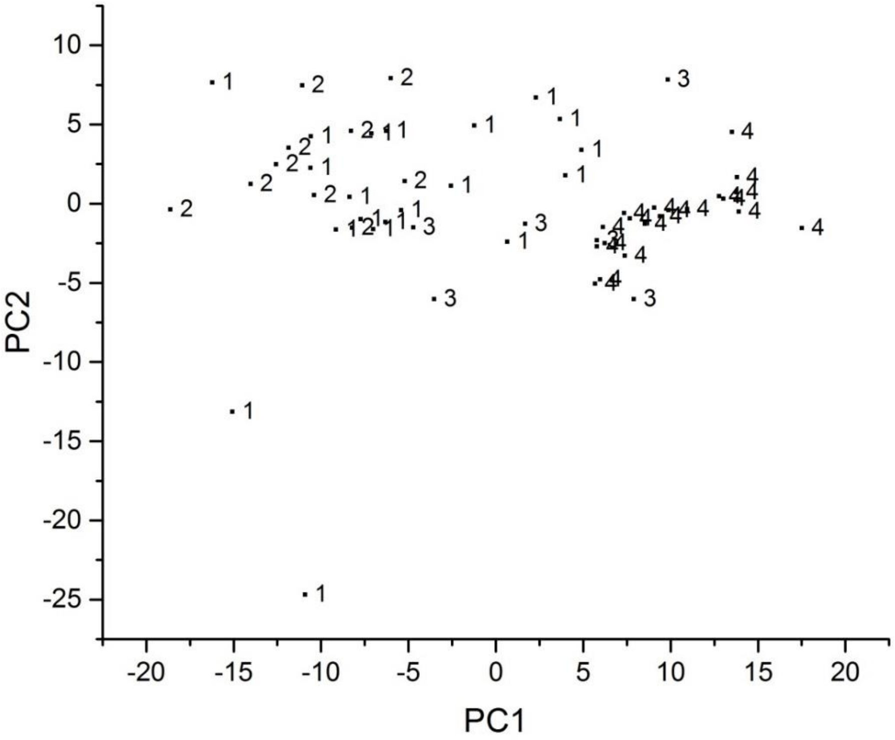


Figure 3.8. Plot of the two largest principal components of the plate 1b, 2a, and 2b ion distribution profiles and the 404 mass spectral features from the training set. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

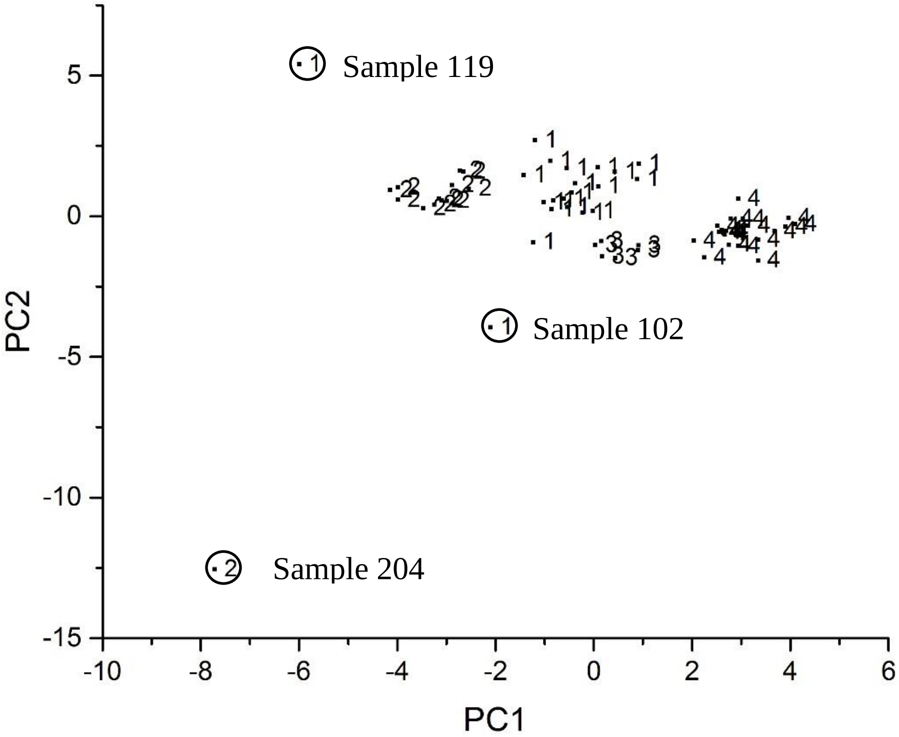


Figure 3.9. Plot of the two largest principal components of the plate 1b, 2a, and 2b ion distribution profiles from the training set and the 26 mass spectral features identified by the pattern recognition GA. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

For experiments of the type that we are considering, there will be relationships among the set of conditions used to generate the data and the patterns that result. One must realize this in advance when approaching the task of analyzing such data. The problem is utilizing information characteristic of the pathological alteration characteristic of the various disease states (BE, HGD, and EAC) of the patients without being inundated by the large amount of quantitative data due to variations in the experimental conditions contained in the complex MALDI-IMS-MS spectral images. The study design used here serves as a test to determine whether information characteristic of the disease state of the subject can be extracted from two-dimensional mass spectral data.

The truncated training set of 87 ion distribution profiles and 404 drift bin features was analyzed by the pattern recognition GA to identify informative features (correlated to disease phenotype) in the MALDI-IMS-MS dataset by sampling key feature subsets (chromosomes) and scoring their PC plots. After 200 generations, the boosting routine of the pattern recognition GA steered the population to an optimal solution.

The capability of the set of features identified by the pattern recognition GA to delineate between NC, BE, HGD and EAC phenotypes was assessed using principal component analysis. Figure 3.9 shows a plot of the two largest principal components of the 24 features identified by the pattern recognition GA. Remarkably, the four phenotypes are delineated. This is an improvement from our previous study where only NC and EAC phenotypes were differentiated [3-14].

During the course of this analysis, the pattern recognition GA identified 5 additional samples that were discordant. 3 of these 5 samples were previously identified as problematic with respect to classification of the different phenotype groups in the principal component plot of Plates 1a, 2a, and 2b after feature selection (see Figures 3.10). These 5 samples were deleted from the analysis. The four phenotypes are unequivocally distinguished in the principal component plot and in addition, all clusters are tight (only one sample lies outside of a cluster). The observation of tight phenotype clusters suggests high prediction power for these features. This was the case when examining the projection of the 26 blind samples onto the principal component plot comprising these 24 features (see Figure 3.11). All 26 blind samples fall within a given phenotype. Glycans contributing to phenotype differentiation were identified by examining the position along the arbitrary drift bin axis of the features selected by the pattern recognition GA. Most of the selected features were localized on the mobility distribution of five glycans: $S_1F_1H_5N_4$, $F_1H_4N_4$, $F_1H_5N_4$, $S_1H_5N_{4,}$ and $S_2H_5N_4$.

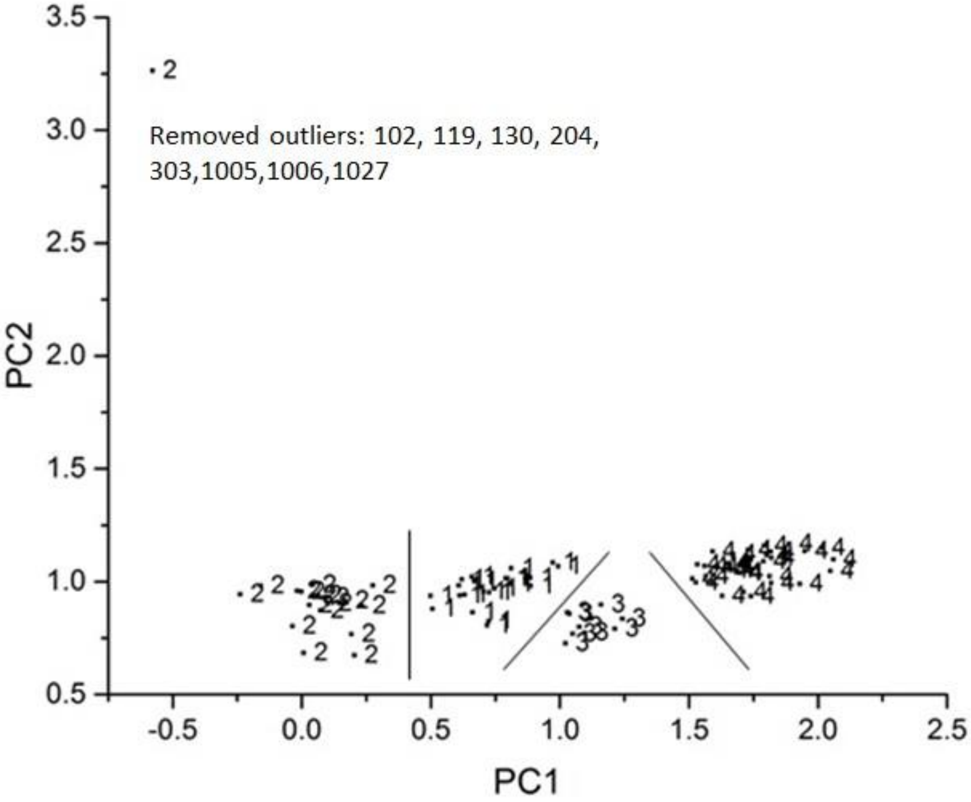


Figure 3.10. Plot of the two largest principal components of the 82 training set samples (8 samples were deleted because they were outliers) and the 24 features identified by the pattern recognition GA. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

The predictions for the 26 blinds are summarized in Table 3.3. Among the 26 blinds, 20 are correctly predicted. The number of false positives was 2, and the number of false negatives was 4. Within the false negative predictions, two samples were predicted as NC instead of BE and two as NC instead of HGD; and within the false positive predictions, one EAC and one HGD prediction were incorrectly made. The 24 features from which the discriminant (principal component plot) was developed yielded 80% sensitivity and 66% specificity. Although a larger blind sample set would be necessary for a true clinical evaluation of the sensitivity and specificity of a discriminant developed from these 24 features, this methodology appears promising for disease phenotype delineation.

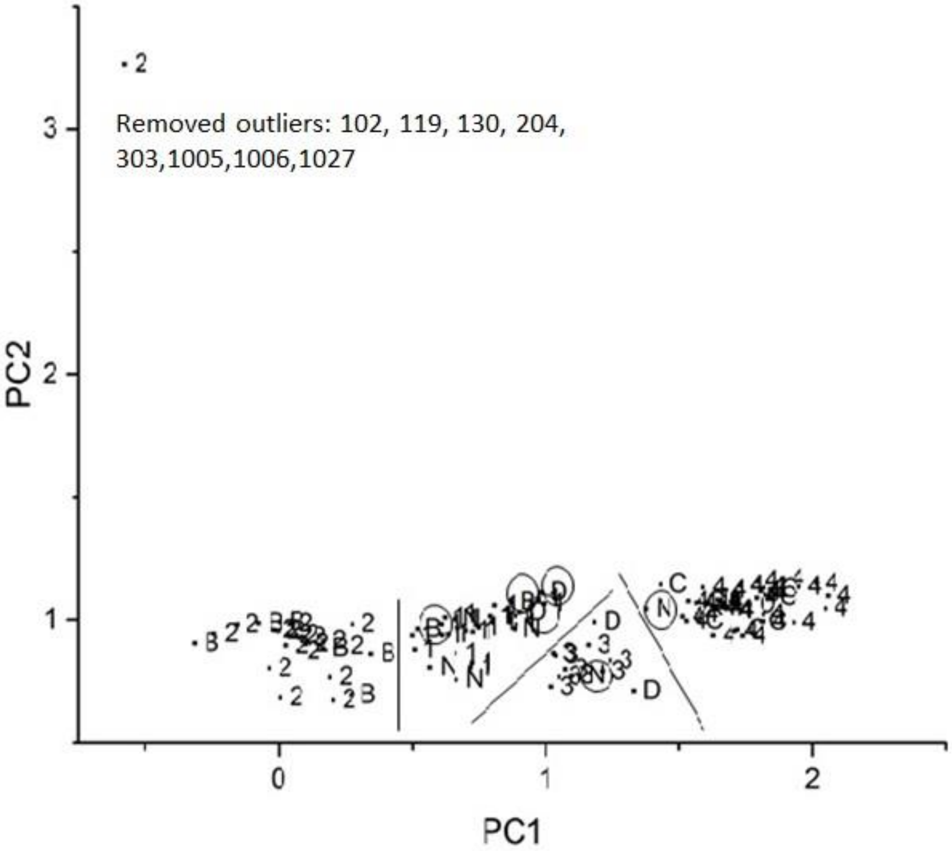


Figure 3.11. Blind samples projected onto the principal component plot defined by the 82 training set samples and the 24 features identified by the pattern recognition GA. For the training set, 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC. For the blinds, N = Normal Controls, B = Barrett's Esophagus, H = High-grade dysplasia, C = Esophageal adenocarcinoma. Circled blinds are incorrectly predicted by the principal component map of the data.

Table 3.3. Phenotype Prediction Results

Blind Sample	Phenotype	Prediction
U_1	BE	BE
U_2	BE	NC
U_3	BE	BE
U_4	EAC	EAC
U_5	EAC	EAC
U_6	EAC	EAC
U_7	HGD	HGD
U_8	HGD	HGD
U_9	NC	HGD
U_10	NC	NC
U_11	NC	EAC
U_12	EAC	EAC
U_13	BE	BE
U_14	BE	BE
U_15	EAC	EAC
U_16	HGD	NC
U_17	NC	NC
U_18	EAC	EAC
U_19	HGD	NC
U_20	BE	BE
U_21	BE	NC
U_22	EAC	EAC
U_23	NC	NC
U_24	BE	BE
U_25	EAC	EAC
U_26	NC	NC

Of the 24 features selected by the pattern recognition GA, 7 were found to be significant for overexpression or under-expression of a specific protein or protein fragment for EAC or an intermediate stage of EAC using a one-way unstacked ANOVA implemented via Minitab 13.1. Table 3.4 summarizes the results for these 7 features identified as significant at the $p < .002$ level. Although these 7 features were found to be significant using a univariate means test to identify mass features correlated to overexpression or under-expression, successful classification of this data required all 24 features.

Table 3.4. ANOVA Results for the 24 GA Selected Features

Glycan (BIN)	Significance ($P < 0.002$)
$F_1H_3N_4$ (93)	None
$F_1H_3N_4$ (112)	None
$F_1H_4N_4$ (306)	Underexpressed for EAC
$F_1H_4N_4$ (307)	Underexpressed for EAC
$F_1H_4N_4$ (317)	Underexpressed for EAC
$F_1H_5N_4$ (508)	None
$F_1H_5N_4$ (517)	Underexpressed for EAC
$F_1H_5N_4$ (518)	Underexpressed for EAC
$F_1H_5N_4$ (533)	Underexpressed for EAC
$F_1H_5N_4$ (536)	None
$S_1H_5N_4$ (732)	Underexpressed for EAC
$S_1H_5N_4$ (758)	None
$S_1H_5N_4$ (764)	None
$S_1F_1H_5N_4$ (916)	None
$S_1F_1H_5N_4$ (929)	None
$S_1F_1H_5N_4$ (953)	None
$S_1H_5N_5$ (1128)	None
$S_1H_5N_5$ (1136)	None
$S_2H_5N_4$ (1333)	None
$S_2H_5N_4$ (1372)	None
$S_2H_5N_4$ (1376)	None
$S_1F_1H_5N_5$ (1542)	None
$S_1F_1H_5N_5$ (1572)	Overexpressed for HGD
$S_2F_1H_5N_5$ (1740)	None

The ion mobility distributions of the three glycans shown in Figure 3.3 suggest the presence of structural isomers for these glycans as the ion distribution profile for a single compound should be Gaussian. For this reason, the discrete wavelet transform was applied to the concatenated ion mobility distribution profiles of the 9 glycans. Wavelets can resolve overlapping spectral responses while simultaneously increasing signal to noise by separating the signal from noise in distinct wavelet coefficients. The motivation for applying the discrete wavelet transform to the concatenated glycan ion mobility distribution profiles is shown in Figure 3.12.

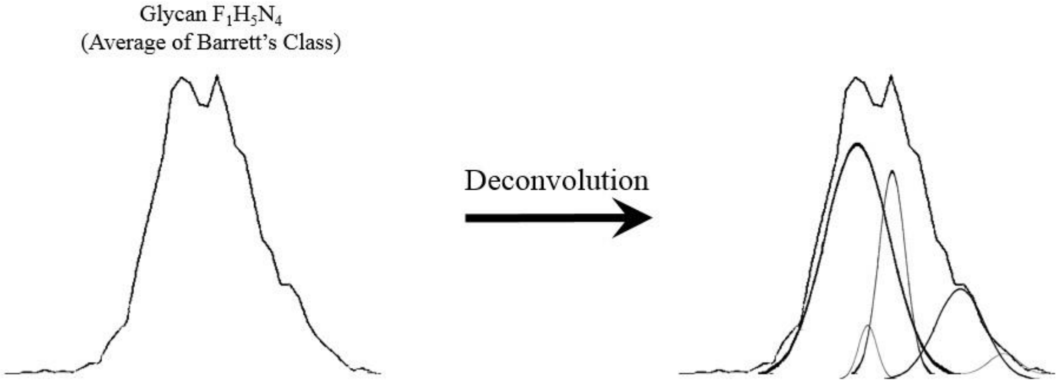


Figure 3.12. Concept underlying the motivation for applying the discrete wavelet transform to the concatenated ion mobility distribution profiles

Figure 3.13 shows a plot of the two largest principal components of the 90 training set samples and the 2696 wavelet coefficients of the ion distribution profiles using the Symlet 6 mother wavelet at the 8th level of decomposition (8Sym6). Each wavelet preprocessed distribution profile is represented as a point in the principal component plot. Three outliers (two NC samples, 1005 and 1027 and one HGD sample, 303) that are present in this plot are also present in the principal component plot of the full training set shown in Figure 3.4 for the original drift bin data. The wavelet transformed ion distribution profiles of these three outliers were again markedly different from the other mobility ion distribution profiles in the training set. For this reason, these three outliers were again removed, and principal component analysis was performed on the truncated training set.

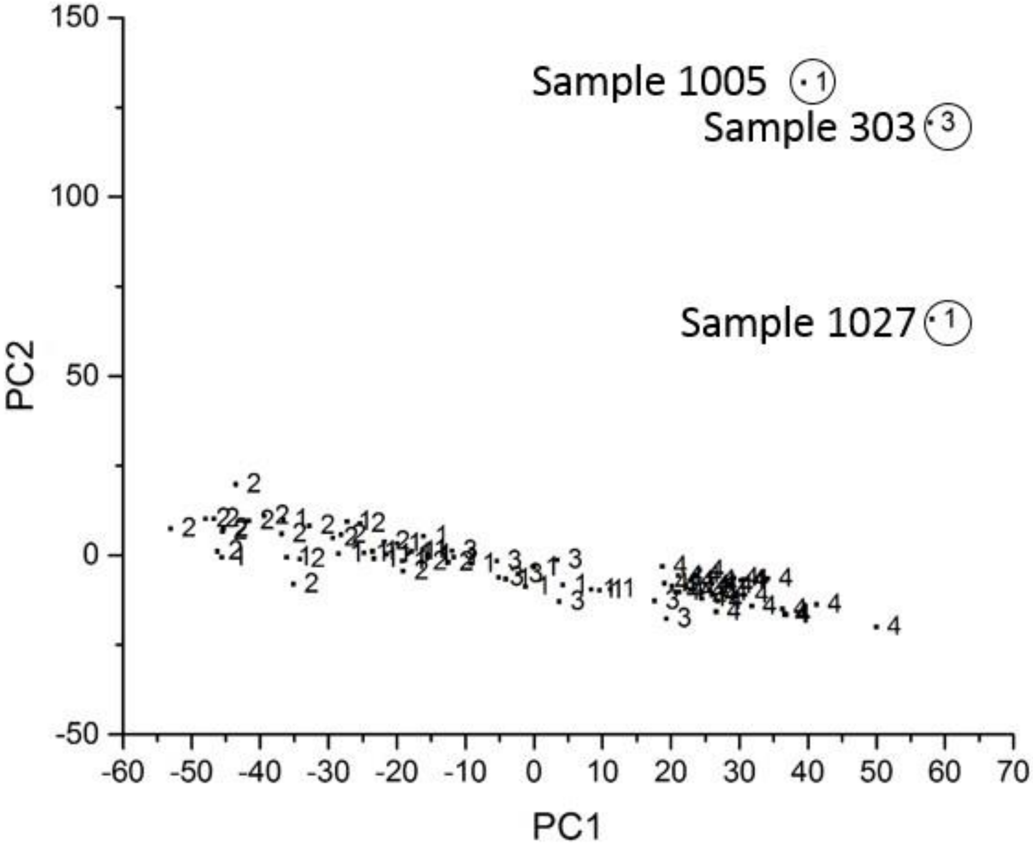


Figure 3.13. Plot of the two largest principal components of the 90 wavelet transformed mobility distribution ion profiles and the 2696 wavelet coefficients. 1 = Normal control (NC, 28 spectra), 2 = Barrett's esophagus (BE, 20 spectra), 3 = high grade dysplasia (HGD, 10 spectra) and 4 = esophageal adenocarcinoma (EAC, 32 spectra).

Figure 3.14 shows a plot of the two largest principal components of the wavelet transformed data (2696 wavelet coefficients) for the truncated training set. Again, the training set can be divided into two groups (see solid line along the second principal component axis and parallel to the first principal component axis delineating the separation of the mobility distribution profiles in Figure 3.14). All samples above the solid line are from the first MALDI spot for each sample within MALDI plate 1 (duplicates were spotted back to back on the MALDI plate, the first spot labeled a, and the second spot labeled b). Examining the origin of each sample, it was observed that all samples from the second spot on plate 1 and both spots on plate 2 lie below the solid line in Figure 3.14. Similar results were obtained for the original ion distribution profile data (see Figure 3.5). Again, the observed clustering in the principal component plot is probably due to the quality of the sample spotting technique, which improved during the course of the experiment.

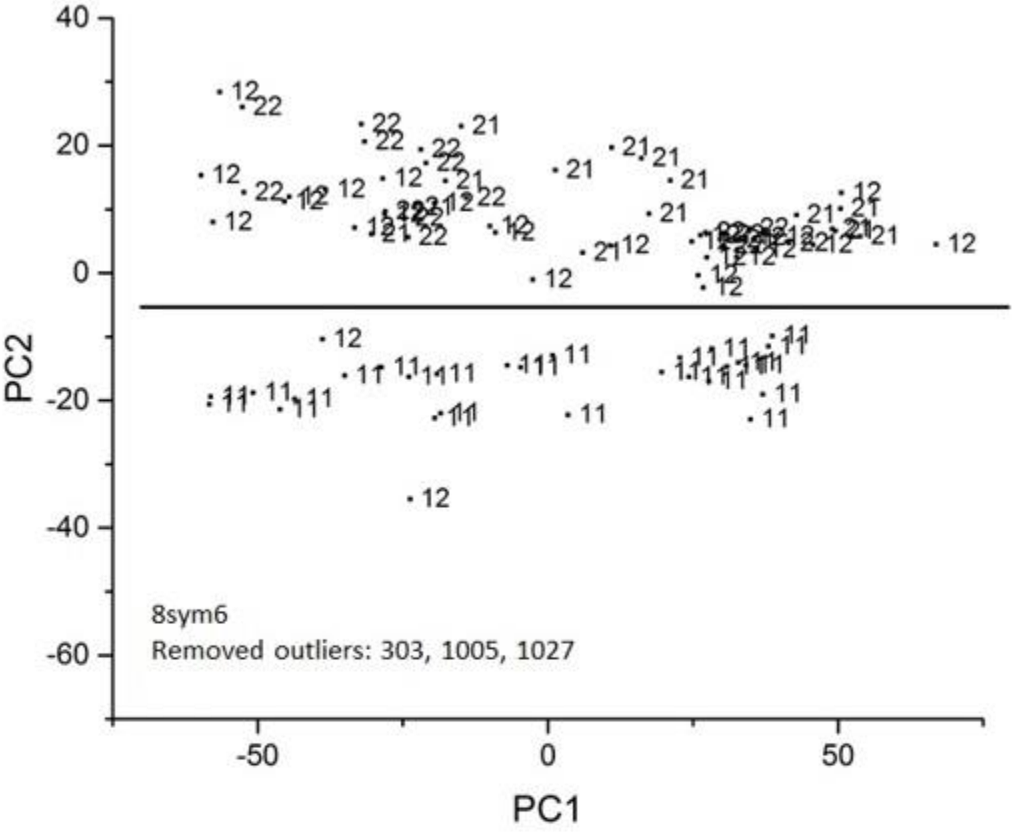


Figure 3.14. A plot of the two largest principal components of the truncated training set of 87 spectra (with the three outliers removed) and 2696 wavelet coefficients. Almost all of the samples above the solid line are from the first MALDI spot whereas the samples from the second spot on plate 1 and both spots on plate 2 lie below the solid line.

The set of ion mobility distribution profiles from plate 1 (denoted as plate 1a samples) and the set of ion mobility distribution profiles also from plate 1 combined with the two sets of spectra from plate 2 (denoted as plate 1b, plate 2a and 2b samples) were analyzed individually using principal component analysis. The pattern recognition GA, which identified wavelet coefficients characteristic of disease phenotype by sampling key feature subsets, scoring their principal component plots and tracking those samples and/or phenotypes that were difficult to classify, was applied to each set of ion distribution profiles. Separation of the different phenotype groups is observed in both principal component plots after feature selection. (see Figures 3.15 and 3.16).

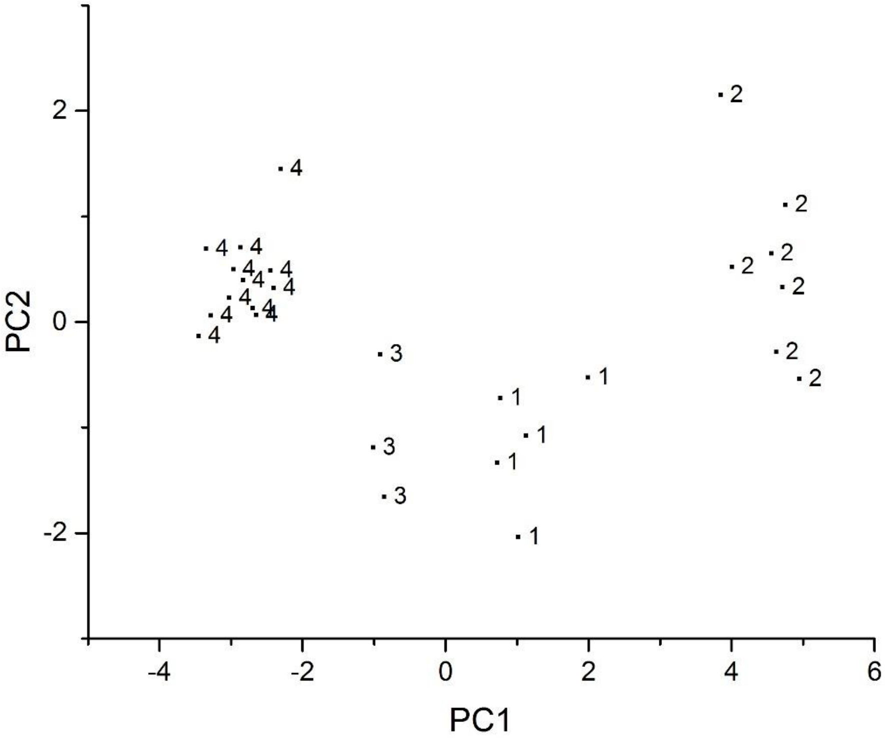


Figure 3.15. Plot of the two largest principal components of the plate 1a ion distribution profiles from the training set and the 12 wavelet coefficients identified by the pattern recognition GA. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

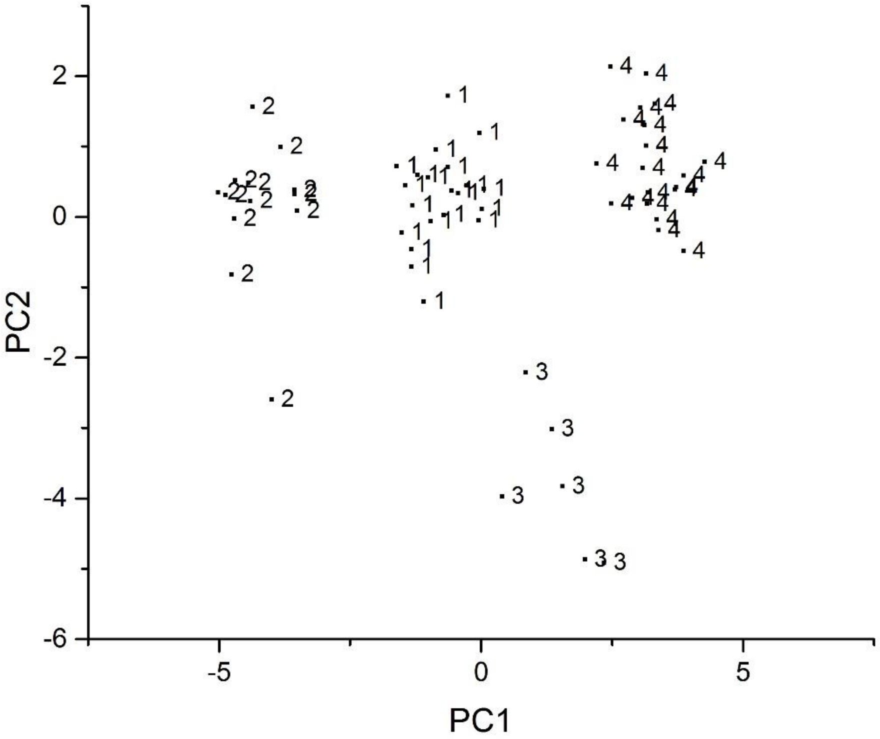


Figure 3.16. Plot of the two largest principal components of the plate 1b, 2a, and 2b ion distribution profiles from the training set and the 14 wavelet coefficients identified by the pattern recognition GA. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

The truncated training set of 87 ion distribution profiles and 2696 wavelet coefficients was analyzed by the pattern recognition GA to identify informative coefficients (correlated to disease phenotype) by sampling key feature subsets and scoring their PC plots. After 200 generations, the boosting routine of the pattern recognition GA steered the population to an optimal solution.

The capability of the set of features identified by the pattern recognition GA to delineate between NC, BE, HGD and EAC phenotypes was assessed using principal component analysis. Figure 3.17 shows a plot of the two largest principal components of the 15 wavelet coefficients identified by the pattern recognition GA. All four phenotypes are clearly delineated.

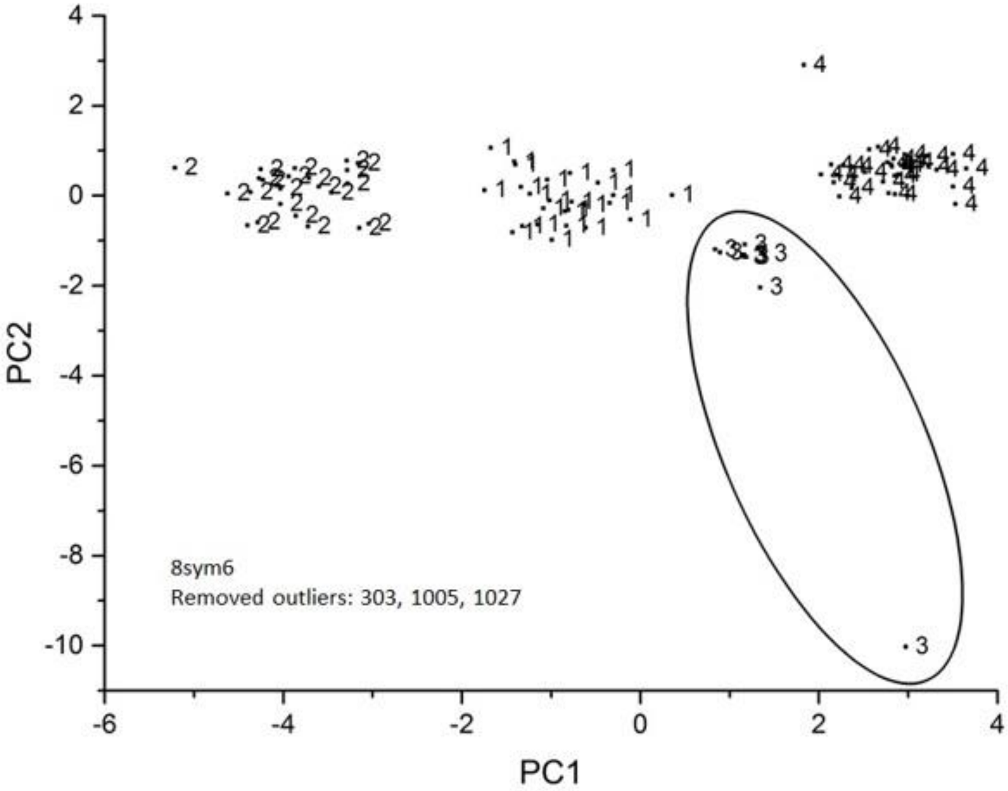


Figure 3.17. Plot of the two largest principal components of the 87 spectra and the 15 wavelet coefficients identified by the GA. 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC.

The predictive ability of the 15 wavelet coefficients was assessed by projecting the 26 blind samples onto the principal component plot of the data developed from the 87 training set samples and 15 wavelet coefficients identified by the pattern recognition GA (see Figure 3.18). All 26 blind samples fall within a given phenotype. Furthermore, only three blind samples were misclassified. These three samples (U_16, U_19 and U_21) were also misclassified by the discriminant developed from 24 features directly extracted from the drift bins of the ion mobility distributions of the concatenated 9 glycans without wavelet preprocessing.

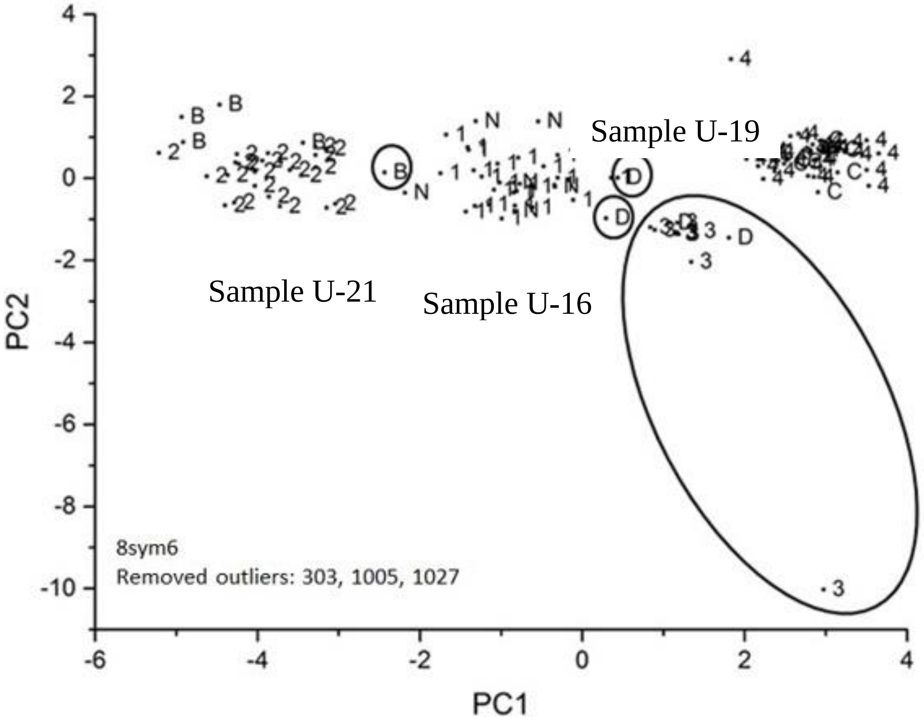


Figure 3.18. Blind samples projected onto the principal component plot defined by the 87 training set samples and 15 wavelet coefficients identified by the pattern recognition GA. For the training set, 1 = NC, 2 = BE, 3 = HGD, and 4 = EAC. For the blinds, N = Normal Controls, B = Barrett's Esophagus, H = High-grade dysplasia, C = Esophageal adenocarcinoma. Circled blinds are incorrectly predicted by the plot.

The predictions for the 26 blinds are also summarized in Table 3.5. There were no false positives, and the number of false negatives was 3. Within the false negative predictions, two samples were predicted as NC instead of HGD and one as NC instead of BE. The 15 wavelet coefficients from which the discriminant (principal component plot) was developed yielded 85% sensitivity and 100% specificity.

Table 3.5. Phenotype Prediction Results

Blind Sample	Phenotype	Prediction
U_1	BE	BE
U_2	BE	BE
U_3	BE	BE
U_4	EAC	EAC
U_5	EAC	EAC
U_6	EAC	EAC
U_7	HGD	HGD
U_8	HGD	HGD
U_9	HGD	HGD
U_10	NC	NC
U_11	NC	NC
U_12	EAC	EAC
U_13	BE	BE
U_14	BE	BE
U_15	EAC	EAC
U_16	HGD	NC
U_17	NC	NC
U_18	EAC	EAC
U_19	HGD	NC
U_20	BE	BE
U_21	BE	NC
U_22	EAC	EAC
U_23	NC	NC
U_24	BE	BE
U_25	EAC	EAC
U_26	NC	NC

The training set and validation set results for the wavelet transformed concatenated distribution profiles support the hypothesis that isomers of the 9 glycans (which are crucial for the full delineation of the disease phenotypes of esophageal cancer investigated in this study) cannot be completely separated by 2-dimensional mass spectrometry alone. However, multivariate and pattern recognition analysis offers the possibility of accessing information about these isomers from the mass spectral images for complete phenotype discrimination.

3.5 CONCLUSION

Serum N-linked glycans extracted from patients diagnosed with BE, HGD, EAC and NC were analyzed by MALDI-IMS-MS. A close examination of mobility profiles for the glycan ions $[S_1H_5N_4+Na]^+$, $[F_1H_5N_4+Na]^+$, and $[S_1F_1H_5N_4+Na]^+$ revealed that in some cases, variations across different phenotypes are immediately noticeable. Because of the number of samples and ions examined within each sample**,** a pattern recognition based approach methodology utilizing variable selection was implemented in order to assess the capability of the dataset for disease phenotype delineation. To perform this task, mobility distributions for nine N-linked glycan ions ($F_1H_3N_4$, $F_1H_4N_4$, $F_1H_5N_4$, $S_1H_5N_4$, $S_1F_1H_5N_4$, $S_1H_5N_5$, $S_2H_5N_4$, $S_1F_1H_5N_5$ and $S_2F_1H_5N_4$) were extracted from the mass spectral data and combined into a composite IMS distribution. Noticeably, NC, BE, HGD and EAC phenotypes were unambiguously differentiated. Among the nine N-linked glycan ions selected for this analysis, the major contributors for distinguishing phenotypes are $S_1F_1H_5N_4$, $F_1H_4N_4$, $F_1H_5N_4$), $S_1H_5N_4$ and $S_2H_5N_4$. Overall, this study demonstrates the capability of the combination of MALDI-IMS-MS and pattern recognition techniques for disease phenotype delineation.

REFERENCES

3.1 Hayeck TJ, Kong C, Spechler S.J, Gazelle G S, Hur C, The prevalence of Barrett's esophagus in the US: estimates from a simulation model confirmed by SEER data, Dis Esophagus. 2010, 23(6): 451-457.

3.2 http://www.mlo-online.com/articles/201404/barretts-esophagus-and-the-need-for-improved-diagnostic-and-prognostic-testing.php

3.3 Tischoff I, Tannapfel A, Barrett's esophagus: can biomarkers predict progression to malignancy, Expert Rev Gastroenterol Hepatol 2008, 2: 653–63

3.4 Farrow CD., Vaughan LT., Determinants of survival following the diagnosis of esophageal adenocarcinoma (United States), Cancer Causes Control, 1996 7: 322–32

3.5 Brown CS, Ujiki MB, Risk factors affecting the Barrett's metaplasia-dysplasia-neoplasia sequence, World J Gastrointest Endosc. 2015 May 16; 7(5): 438–445

3.6 Kadri S, Lao-Sirieix P, Fitzgerald RC, Developing a nonendoscopic screening test for Barrett's esophagus, Biomark Med, 2011, 5: 397–404

3.7 Reid BJ, Haggitt RC, Rubin CE, et al. Observer variation in the diagnosis of dysplasia in Barrett's esophagus, Hum Pathol, 1988, 19(2): 166-178

3.8 DeVault KR, Castell DO., Updated guidelines for the diagnosis and treatment of gastroesophageal reflux disease. Am J Gastroenterol.2005, 100: 190-200

3.9 Adamczyk B, Tharmalingam T, Rudd PM, Glycans as cancer biomarkers, Biochim Biophys Acta., 2012, 1820(9):1347-53

3.10 Zane T. H, Yehia M, Ahmed H, Slavka B, Min Zh, Kenneth A. K, and Milos V, Comparative glycomic profiling in esophageal adenocarcinoma, J Thorac Cardiovasc Surg, 2010, 139: 1216-23

3.11 Mann B, Madera M, Klouckova I, Mechref Y, Dobrolecki LE, HickeyRJ, et al. A quantitative investigation of fucosylated serum glycoproteins with application to esophageal adenocarcinoma, Electrophoresis, 2010, 31: 1833–41

3.12 Indranil M, Zexi Zh, Yuening Zh, Chuan-Yih Y, Zane T H, Haixu T, Yehia M, and Stephen C J, N-Glycan profiling by microchip electrophoresis to differentiate disease states related to esophageal adenocarcinoma, Anal, Chem., 2012, 84: 3621−3627

3.13 Lee, S., Valentine, S. J., Reilly, J. P., Clemmer, D. E., Analyzing a mixture of disaccharides by IMS-VUVPD-MS. Int. J. Mass. Spectrom., 2012, 309: 161–167.

3.14 Gaye MM, Valentine SJ, Hu Y, Mirjankar N, Hammoud ZT, Mechref Y, Lavine BK, Clemmer DE, Ion mobility-mass spectrometry analysis of serum N-linked glycans from esophageal adenocarcinoma phenotypes, J Proteome Res 2012, 11: 6102–10

3.15 Gaye MM, Ding T, Shion H, Hussein A, HU, S Zhou, Hammoud T., Lavine BK, Mechref Y, Gebler JC, Clemmer DE, Delineation of Disease Phenotypes Associated with Esophageal Adenocarcinoma by MALDI-IMS-MS Analysis of Serum N-linked Glycans, J. Proteome Research, submitted.

3.16 Hoaglund, C. S., Valentine, S. J., Sporleder, C. R.; Reilly, J. P., Clemmer, D. E. Three dimensional ion mobility/TOFMS analysis of electrosprayed biomolecules. Anal. Chem., 1998, 70: 2236–2242.

3.17 Tang, K., Shvartsburg, A. A., Lee, H.-N., Prior, D. C., Buschbach, M. A., Li, F., Tolmachev, A. V., Anderson, G. A., Smith, R. D., High-sensitivity ion mobility spectrometry/mass spectrometry using electrodynamic ion funnel interfaces. Anal. Chem., 2005, 77: 3330–3339

3.18 Zucker, S.M., Lee, S., Webber, N., Valentine, S.J., Reilly, J.P. and Clemmer, D. E., An ion mobility/ion trap/photodissociation instrument for characterization of ion structure. J. Am. Soc. Mass Spectrom., 2011, 22: 1477–1485.

3.19 Lee, S., Valentine, S. J., Zucker, S. M., Webber, N., Reilly, J. P., Clemmer, D. E., Extracted fragment ion mobility distributions: a new method for complex mixture analysis. Intl. J. Mass Spectrom., 2012, 309: 154–160.

3.20 Clemmer DE, Jarrold MF, Ion mobility measurements and their applications to clusters and biomolecules, Journal of mass spectrometry, 1997, 32(6): 577-592.

3.21 Hoaglund CS, Valentine SJ, Sporleder RC, Reilly JP, and Clemmer DE, Three-dimensional ion mobility/TOFMS analysis of electrosprayed biomolecules, Anal. Chem., 1998, 70: 2236-2242

3.22 Kang P.; Mechref Y.; Klouckova I, and Novotny MV, Solid-phase permethylation of glycans for mass spectrometric analysis, Rapid Commun. Mass Spectrom., 2005, 19: 3421-3428

3.23 Isailovic D, Kurulugama RT, Plasencia MD, Stokes, ST, Kyselova, Z, Goldman, R, Mechref Y, Novotny, MV, Clemmer DE, Profiling of human serum glycans associated with liver cancer and cirrhosis by IMS–MS, J. Proteome Res., 2008, 7: 1109–1117

3.24 Isailovic D, Plasencia M, Gaye M, Stokes S, Kurulugama R, Pungpapong V, Zhang M, Kyselova Z, Goldman R, Mechref Y, Novotny MV, Clemmer DE, Delineating diseases by IMS–MS profiling of serum N–linked glycans, J. Proteome Res., 2012, 11: 576–585

CHAPTER IV

SEARCH PREFILTERS FOR THE FORENSIC ANALYSIS OF AUTOMOTIVE PAINTS: GENERAL MOTORS, TOYOTA, NISSAN AND HONDA

4.1 INTRODUCTION

Applying Fourier infrared spectroscopies (FTIR) to a forensic paint sample analysis can be traced back to thirty five years ago, Royal Canadian Mounted Police (RCMP) found that vehicles could be identified by comparing the color, the layer sequence and the chemical composition of each individual layer of a paint sample. [1] They developed the system called Paint data query (PDQ) for classifying, storing and retrieving evidential paint information. PDQ database provides information in physical attributes of a paint sample, the chemical composition of each layer of the original manufacturer's paint system via IR spectra. To better understand the chemical analysis of automotive paint samples by using FTIR, it is necessary to understand the paint components of each layer and the paint layer construction of a model vehicle. Modern automotive paint systems [2] consist of four layers; from the top to the bottom, they are the clear coat, the color coat, the surface-primer coat and the e-coat primer (see Figure 4.1). [3] The clear coat and the color coat are called topcoat while the surfacer and the e-coat primer are called undercoat.

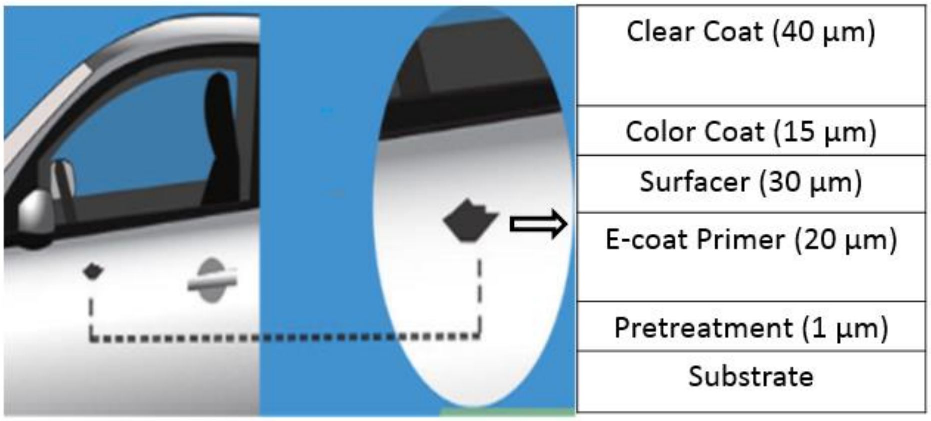


Figure 4.1. Scheme of the multilayer coating of cars

The chemical components in a clear coat are paint resins and binders without color pigments, while other layers have pigments, fillers, polymers and binders. In United States, there are two clear coat formulations, one is acrylic melamine topcoat (see Figure4.2); the other is carbamate melamine (see Figure 4.3); Carbamate polyurethane (see Figure 4.4) commonly exists in European automotive paint formulations. Meanwhile β-hydroxyl polyesters is the main chemicals applied on Japanese cars (see Figure 4.5). [3] As a backbone material, acrylic polymers are modified by styrene (see Figure 4.6). Since FTIR is sensitive to molecular functional groups, transmission bands in a FTIR spectrum will characterize the chemical compositions of automotive top coat. For example, melamine triazine ring (C_3N_3) should have the band at 1550 cm^{-1} and a small non-diagnostic shoulder at 1450 cm^{-1}.The 815 cm^{-1} sharp band reflects out-of-plane triazine ring vibration. Acrylic resins should show transmission bands of carbonyl at 1732 cm^{-1} and C-O stretch in 1100-1310 cm$^{-1.}$ The hydroxyl functional acrylic carbonyl bands move to 1689 cm^{-1} due to a hydrogen bond. Bands at1293 cm^{-1} and 1169 cm^{-1} are bending bands of the aliphatic ester.

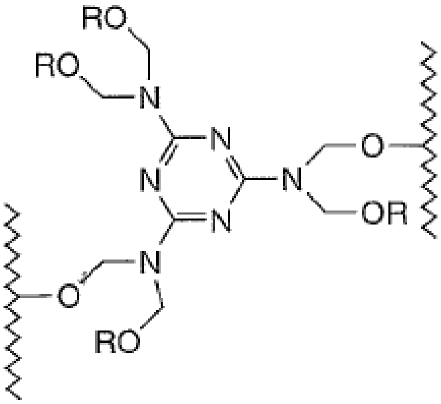


Figure 4.2. The chemical structure of acrylic melamine

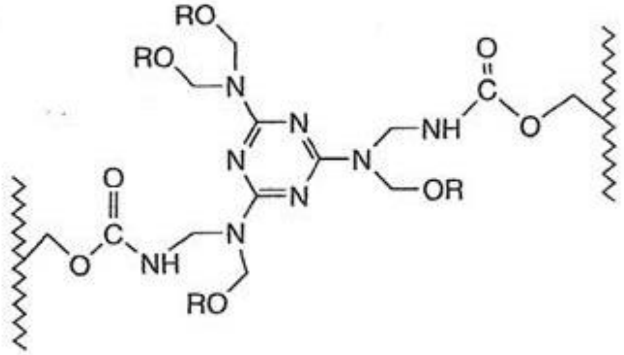


Figure 4.3. The chemical structure of carbamate melamine

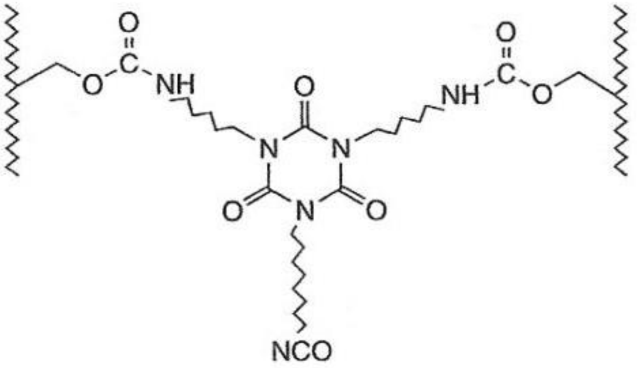


Figure 4.4. The chemical structure of carbamate polyurethane

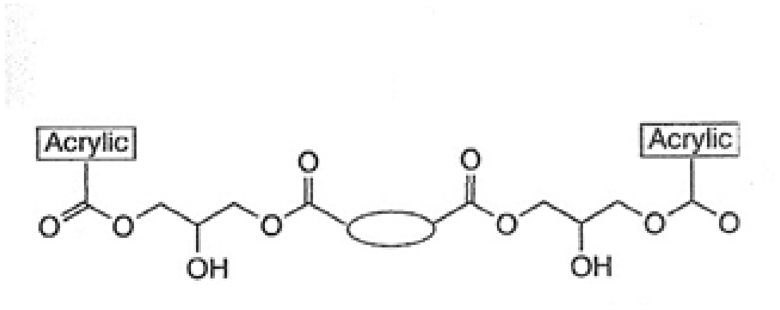


Figure 4.5. The chemical structure of β-hydroxyl polyesters

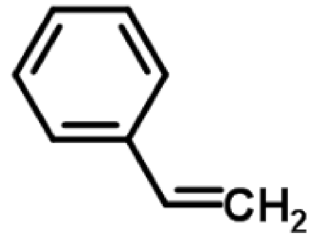


Figure 4.6. The chemical structure of styrene

The color coat is also called base coat. Color is valuable to discriminate an automobile paint. The paint in this layer is composed of a pigment portion and a vehicle portion. Pigment portion (transparent extender pigment and opaque coloring pigment) dominates the discrimination ability in the sample comparisons. For color coat, micro spectrophotometry (MSP) combined with X-ray spectrometry and scanning electron microscope (SEM) are more sensitive than FTIR instruments in the detection of heavy metals. Even if TiO2, ZnO, BaSO4 and CaCO3 have a visible peak in FTIR, since most pigment compounds have broad or weak bands in a FTIR spectrum and their bands range is from 400-1000 cm^{-1} [5] (signals from FTIR-ATR microscopy are not stable when the wavenumber is less than 675 cm^{-1} in the transmission mode); in addition, high concentration of inorganic components in this layer hinds FTIR in detecting paint binders. FTIR is not suitable to examine the chemical information of a trace paint sample. This layer is not considered to be used for developing a prefilter in this study.

The surface-primer coat is also called a “filler” or “middle” coat, chemical components are isophthalic alkyd (polyester), melamine, strontium chromate ($SrCrO_4$), Kaolin ($Al_2Si_2O_5(OH)_4$), barium sulfate ($BaSO_4$). [8] Polyester as a resin component exists in almost all types of primer surfacers. In IR spectra, ester group should have C=O stretch band in 1650-1750 cm $^{-1}$, C-O-C bending band in around 1250 or 1120 cm $^{-1.}$ Substitute benzene ring will show a band around 800 cm^{-1}, aromatic ring stretch in 1510-1615 cm $^{-1}$, aromatic C-H in-plate bend 950-1225 cm $^{-1}$, aromatic C-H out-of-plate band in 670-900 cm $^{-1}$. [6] Color in this layer normally will match to the color coat.

The e-coat primers are applied to protect automotive substrate and adhere the surface-primer layers. Common chemical components in this layer are epoxy, polyurethane, kaolin, titanium, and dioxide· [6] The structure of epoxy resins is showed in Figure 4.7. Typically, FTIR characterizing epoxy resins with C=O band around 1730 cm^{-1}, C-O-C band in 1285cm^{-}1 and 1122 cm^{-1}, CH2─or CH3 ─ bending band in 1376,1467 cm^{-1} and aromatic ring bending in 706 cm^{-1}. Polyurethane FTIR characterizing bands are N-H bending bands in 1468 cm^{-1} and 1522 cm^{-1}, C-H bending band in 1380 cm^{-1}, and a broad band in 1254 cm^{-1} representing C-N-H vibration and C-O bending. The most important mention in this book reference is “many of the parts coming into an assembly plant are pre-primed and may receive common final coats”. [6] The primer layers may have similar chemical compositions for different automotive manufacturers.

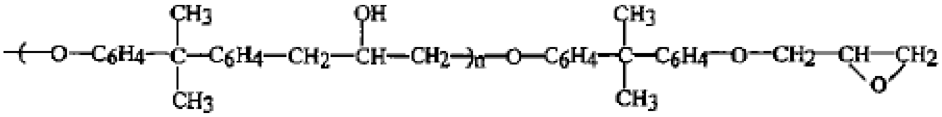


Figure 4.7. The chemical structure of epoxy polymer

FTIR provides molecular structure information characterizing organic and inorganic constituents of paint layers. Based on this knowledge, paint data query (PDQ) was developed in the mid-1970s [7] as an automotive paint standard examination database. In PDQ, each layer was coded with symbols. For example, the clear coat is marked as OT2; OT1 represents the color coat; the surface-primer coat is coded in OU1 and OU2 denotes the e-coat primers. FTIR spectra were available by separating each paint layer of an automotive sample and placed between two diamond anvils to measure. Up to now, PDQ contains 21000 samples and total over 84000 individual paint layers. [8] PDQ is the largest international forensic automotive searchable database of chemical and color information of automotive paints, it is used in the United States, Japan, Australia, New Zealand, Singapore, and the European Union. [9] PDQ contains a large number of transmission spectra and are often used as references for the purpose of qualitative analysis of paint fragment samples left at crime scenes. Automotive paint coatings are consisted of four or five layers, [3] by comparing the infrared spectrum of each paint layer of paint fragments left at hit-and-run to the relative infrared spectra of PDQ, police can narrow the search for unknown vehicles by using database match. Nevertheless, text-code PDQ system are unable to give accurate automotive information, because text based codes cannot uniquely characterize the IR spectrum of each paint layer. Multi-layered automotive paint fragments brought difficulties in the forensic examination of the composition of an automotive paint due to similar chemical paint compositions and small sample size. Unspecific PDQ search produces a large number of hits and increases the workload and difficulty for forensic investigation. [10] In addition, the color layer may be too thin or small to be compared with manufacturer's paint color standards in the PDQ system. [8] Broad and undefined features, scatter effects challenge directly using paint IR spectra to differentiate assembly plants. The new technology is needed to improve the accuracy of the PDQ library search.

Pattern recognition was introduced to solve the above mentioned problem. Pattern recognition is able to find the pattern embed in automotive paint sample FTIR data and interpret large data objectively and visually. Principle component analysis maximizes the difference between paint chemical compositions and then improve the discriminative capacity of the trance paint sample analysis. Many researches were carried out for developing supervised classification model as a prefilter to predict an unknown paint sample. [8, 10, 11, 12-19]. The prefilter using pattern recognition conjugated PDQ and a cross correlation library searching algorithm were successfully developed. First, the prefilter was developed from the clear coat to determine suspect car related to a single manufacturer: General Motors (GM) or Chrysler. Since the major chemical compositions of the top layer coats are composed of either acrylic melamine styrene or acrylic melamine styrene polyurethane in PDQ, the prefilter developed from the IR spectra of the clear coat paint samples lacked the differential ability when multi-manufacturer existence. The clear coat, surfacer-primer, and e-coat layers of each paint sample in PQD were conjugated together for developing a more powerful prefilter, which can be used to identify the make, line and model of an automotive vehicle from three manufacturers: General Motors (GM), Chrysler and Ford. However, the PDQ includes another three automotive manufacturers: Honda, Nissan and Toyota. The further study should carry out to develop a prediction model based on all manufacturers included in PDQ.

My research aims to develop robust search prefilters to determine the make, line and model of an automobile based on its paint fragment in a limited year range (2000-2006) involves GM, Chrysler, Ford, Honda, Nissan, and Toyota. The hypothesis is that each assembly plant should have a unique paint formulation for an individual paint layer, the difference in the paint formulation is enough to be characterized by FTIR to discriminate the unknown paint fragment from the same source as PDQ. In this study, two prefilters were developed; Three-layer prefilters (The clear coat, surfacer-primer, and e-coat layers) were compared with two-layer prefilters (The clear coat plus surfacer-primer). The three-layer prefilters did not show a significant improvement in an unknown paint sample prediction. Compared with the previous prefilters, the six-manufacturer prefilters significantly improved the discrimination capability and scope of automotive manufacturers. It narrows down a suspect automotive list for the further forensic trace evidence investigation and greatly improved the work efficiency by its function - simultaneously multi unknown sample predictions. This method can also applied for the discrimination of trance paint samples whose spectra is collected by ATR microscope reflectance or transmission mode. This method enables a forensic scientist to draw a more accurate conclusion between the evidence and comparative materials. Two different methods were explored for three-layer prefilters.

4.2 FOURIER TRANSFORM INFRARED SPECTROSCOPY (FTIR)

Fourier transform infrared spectroscopy (FTIR) is widely used for material identification, since its spectra disclose the unique combination of atoms making up a molecular. The bands in FTIR are generated by the vibration between the bonds of the atoms in a molecular. Based on this theory, FTIR spectra are able to carry chemical compositions of paint samples, and are used for the identification of a car make and model based on the paint sample left at hit-a-run crime scene. In PDQ database, the paint fragment spectra were collected by a Fourier transform infrared spectroscopy (FTIR) transmission mode. Therefore, we need understand the basic theoretical principle of a FTIR spectrometer and the fundamental knowledge of FTIR. The instrumentation of the Thermo Nicolet FTIR spectrometer is in the Figure 4.8. [20] A FTIR spectrometer consists of an IR light source, an interferometer, a sample compartment and a detector. Nernst glowers are used in all infrared spectrometers as light sources. Deuterated triglycerine sulfate and mercury cadmium telluride are the most two commonly used detectors. The Michaelson Interferometer is the heart of a FTIR spectrometer, by employing an interferometer (see Figure4.9) [21], a FTIR spectrometry greatly reduces the sample scanning time comparing with other types of IR instruments. This is an advantage of FTIR. The beamsplitter is a KBr plate with a thin coating of germanium, which reflects the half of incident light and transmits the remaining half incidence of IR through two separate optical paths. One optical path is from a fixed mirror, and the other is from a moving mirror. The sum of these two beams arrived the detector to generate the signal called an interferogram. When the interferogram is measured, all frequencies are being measured simultaneously, this is the reason why an interferometer can fast the measuring time. The signal embedded in an interferogram is decoded by Fourier transformation (see Equation 4-1) to form a FTIR spectrum. Since diamond cells are transparent to IR radiation except in the region of 2400 cm-1 to 1700 cm-1, diamond anvils in a sample compartment are used to hold a sample to obtain a transmission FTIR spectrum. The transmittance is measured by the Equation 4-2 [21], where T is transmittance, I is the intensity of incident light transmitted by a sample, I_0 is the intensity of incident light reaching the sample; d is the thickness of a sample and α is the absorption coefficient.

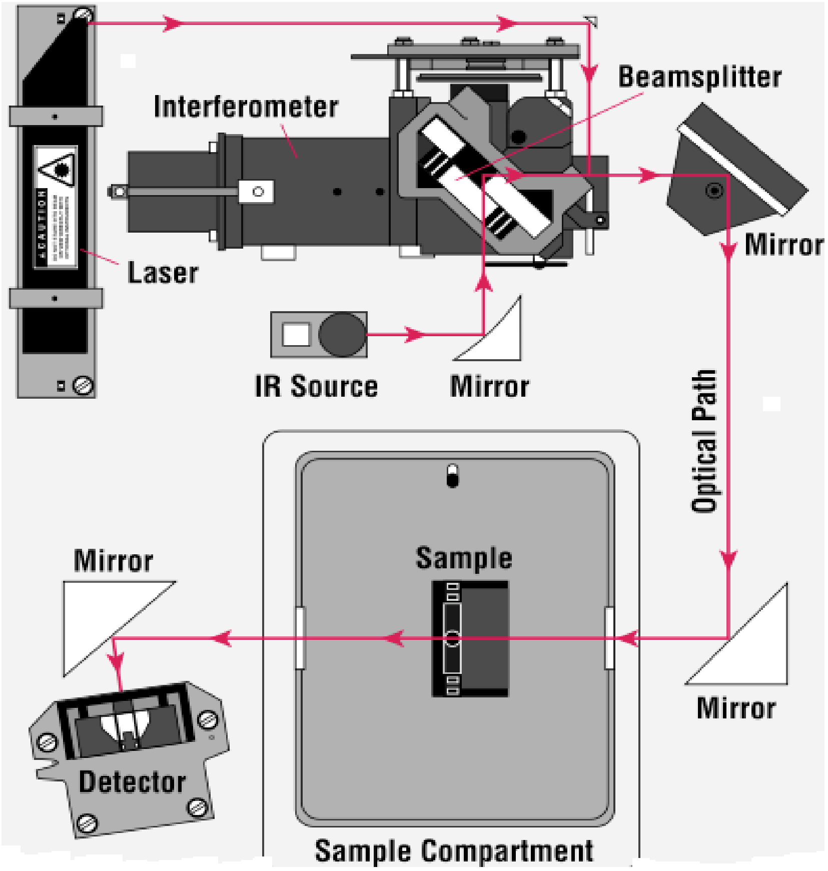


Figure 4.8. The schematic diagram of a FTIR

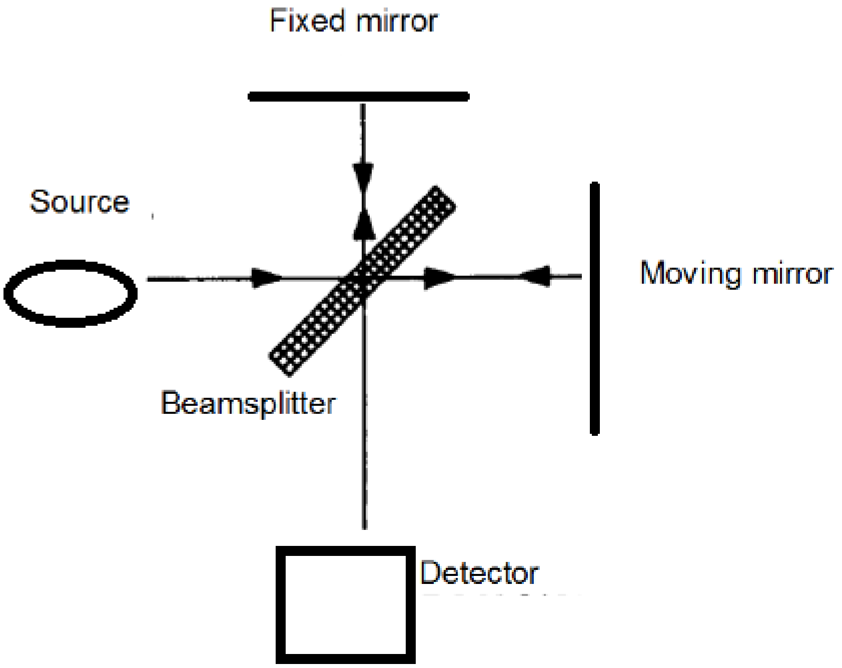


Figure 4.9. A schematic diagram of a Michelson Interferometer

$$F\ (\omega) = \int_{-\infty}^{+\infty} f(X)e^{i\omega X}dX \qquad (4.1)$$

F (ω): FTIR spectrum

f (X): Interferogram

ω: Angular frequency

X: Optical path difference

$$T = \frac{I}{I_0} = e^{-\alpha d} \qquad (4.2)$$

4.3 METHOD

4.3.1 Experiment and Materials

4.3.1.1 Materials

All paint sample IR spectra were provided by RCMP with records of the make, line, model, year, substrate, plant, vehicle type, PDQ number and automotive manufacturers Etc... In this study, sample collection consisted of six manufacturers: GM, Chrysler, Ford, Honda, Nissan, and Toyota. The production year of those vehicles spanned from 2000 to 2006. The total qualified sample population was 1773 paint samples with intact four layers, major sample information was listed in Table 4.1-Table4.7. Sub plants were identified by the PCA plot of the samples in a particular manufacturing plant. Each assembly plant in PDQ database with less than five samples was not chosen for this study, whose information was listed in Figure 4.10.

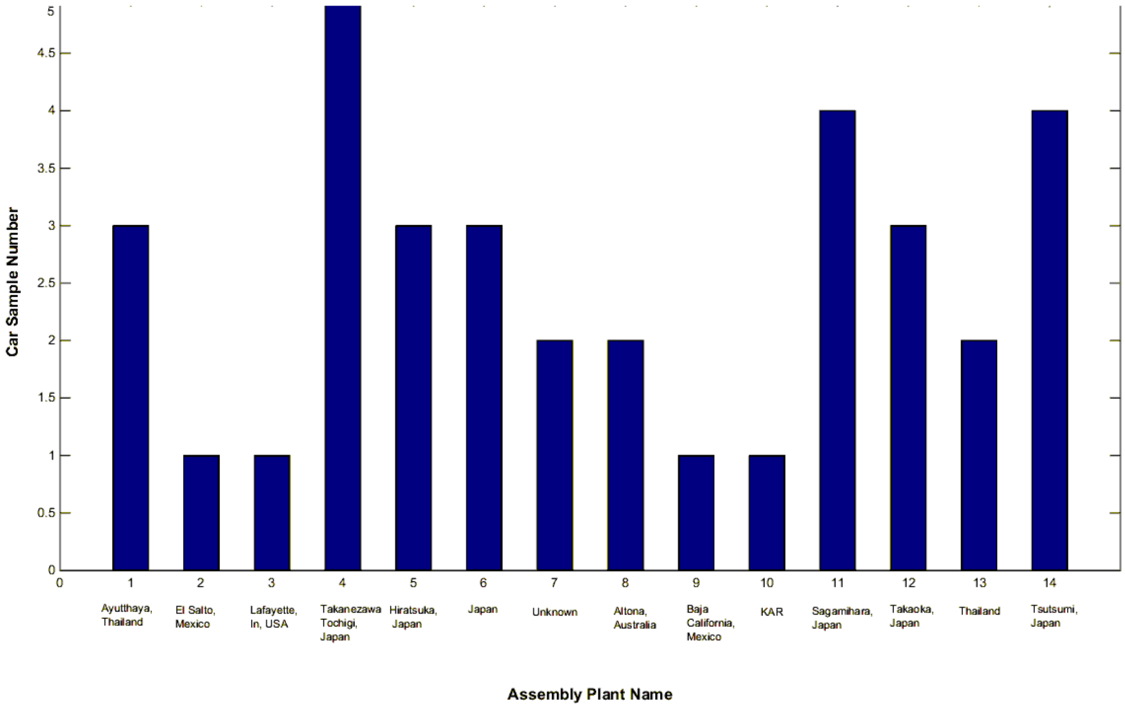


Figure 4.10 The histogram of assembly plants with less 5 samples

Table 4.1 Doublet sample assembly plants used to develop the search prefilter

Manufacturer	Plant name	Plant ID	Sample number
GM	Baltimore (BAL)	2	8
	Hamtramck (HAM)	10	18
	Orion (ORI)	21	12
	Ramos Arizpe (RAM)	24	25
	Silao (SIL)	26	19
	Spring Hill (SPH)	27	9
	Saint Therese (THE)	28	7
	Wentzville (WEN)	29	9
	Wilmington (WIL)	30	7
	Lansing (LAN)	114	8
Chrysler	Jefferson North (JFN)	1004	24
	Newark (NEW)	1006	23
	Jefferson North (JFN)	1104	13
	Newark (NEW)	1106	12
Ford	Wixom (WIX)	2017	10
	Saint Thomas-Talbotsville (STT)	2114	5
	Wixom (WIX)	2217	9
Honda	East Liberty, OH, USA	3102	15
	Marysville, OH, USA	3106	19

Table 4.2 Singlet sample assembly plants from GM used to develop the search prefilter

Manufacturer	Plant name	Plant ID	Sample number
GM	Arlington (ARL)	1	20
	Doraville (DOR)	4	26
	Fairfax (FAI)	5	28
	Flint (FLI)	6	8
	Fort Wayne (FOR)	8	15
	Fremont (FRE)	9	12
	Ingersoll (INE)	11	10
	Janesville (JAN)	12	16
	LAF*	13	1
	Lansing (LAN)	14	32
	Linden (LIN)	16	15
	Lordstown (LRD)	17	39
	Moraine (MOR)	18	29
	Oklahoma City (OKL)	20	7
	Oshawa (OSH)	22	19
	Pontiac (PON)	23	13

	Shreveport (SHR)	25	19
	Oklahoma City (OKL)	120	5
	Oshawa (OSH)	122	22
	Oshawa (OSH)	222	17

Table 4.3 Singlet sample assembly plants from Chrysler used to develop the search prefilter

Manufacturer	Plant name	Plant ID	Sample number
Chrysler	Belvidere (BEL)	1000	36
	Bloomington (BLO)	1001	7
	Bramalea/Brampton (BRA/BRP)	1002	14
	Dodge Main (DOD)	1003	19
	Saltillo (SAL)	1007	29
	Sterling Heights (STH)	1008	22
	St. Louis (STL)	1009	21
	Toledo (TOL)	1010	15
	Toluca (TOU)	1011	28
	Windsor (WIN)	1012	27
	Bramalea/Brampton (BRA/BRP)	1102	43
	Dodge Main (DOD)	1103	21
	Sterling Heights (STH)	1108	8
	St. Louis (STL)	1109	32
	Toledo (TOL)	1110	27

Table 4.4 Singlet sample assembly plants from Honda used to develop the search prefilter

Honda	Alliston, ON, Canada	3000	44
	East Liberty, OH, USA	3002	9
	Lincoln, Alabama	3005	8
	Marysville, OH, USA	3006	23
	Sayama (Saitama) , Japan	3007	22
	Suzuka, Japan	3008	10

Table 4.5 Singlet sample assembly plants from Ford used to develop the search prefilter

Manufacturer	Plant name	Plant ID	Sample number
Ford	Atlanta (ATL)	2000	16
	Chicago (CHI)	2002	21
	Dearborn (DEA)	2003	21
	Flat Rock (FLA)	2005	20
	Hermosillo (HER)	2006	15
	Kansas City (KAN)	2007	23
	Kentucky Truck (KTR)	2008	29
	Lorain (LOR)	2009	6
	Louisville (LOU)	2010	16
	Norfolk (NOR)	2011	13
	Oakville (OAK)	2012	17
	Saint Louis (STL)	2013	6
	Saint Thomas-Talbotsville (STT)	2014	14
	Twin Cities-Saint Paul	2015	12
	Wayne (WAY)	2016	61
	Dearborn (DEA)	2103	7
	Hermosillo (HER)	2106	5
	Kansas City (KAN)	2107	21
	Louisville (LOU)	2110	16
	Norfolk (NOR)	2111	6
	Saint Louis (STL)	2113	8
	Twin Cities-Saint Paul	2115	8
	Wayne (WAY)	2116	11
	Hermosillo (HER)	2206	4

Table 4.6 Singlet sample assembly plants from Nissan used to develop the search prefilter

Manufacturer	Plant name	Plant ID	Sample number
Nissan	Aguascalientes, Mexico	4000	11
	Canton, MS	4001	23
	Kyushu #1,2,3, Japan	4004	6
	Oppama #1,2, Japan	4005	9
	Smyrna, TN, USA	4006	30
	Tochigi #1,2,3, Japan	4007	9
	Aguascalientes, Mexico	4100	7
	Kyushu #1,2,3, Japan	4104	13
	Oppama #1,2, Japan	4105	8
	Smyrna, TN, USA	4106	21

Table 4.7 Singlet sample assembly plants from Toyota used to develop the search prefilter

Manufacturer	Plant name	Plant ID	Sample number
Toyota	Cambridge, ON, Canada	5002	31
	Fremont, CA, USA	5003	16
	Georgetown, KY, USA	5004	27
	Japan	5005	79
	Princeton, IN (Evansville)	5007	22
	Fremont, CA, USA	5103	12
	Georgetown, KY, USA	5104	13
	Japan	5105	5

4.3.1.2 Experimental

The IR transmission spectra of 1773 automotive paint samples from the PDQ database were collected by a Bio-Rad 40A, Bio-Rad 60A or Thermo-Nicolet 6700FTIR spectrometers, the qualified sample must have intact four layers: clear coat, color coat, surface and primer. FTIR spectrometers used to collect IR spectra for PDQ database were equipped with a DTGS detector. FTIR operation resolution was 4 cm^{-1} with apodization Happ-Genzel. Spectra were collected between 400 cm^{-1}-4000 cm^{-1}. The spectrum from each layer of a sample comprised 1869 points. The details about the sampling conditions were described in elsewhere. [8]

4.3.2 Pattern Recognition Method

4.3.2.1 Data Preprocessing

Pattern recognition prefilter is developed to determine the similarity or dissimilarity between an unknown sample IR spectrum and the spectra from PDQ. Before data preprocessing, each spectrum was normalized by frequency in order to achieve an integrated frequencies for the all samples by using the OMNIC software. Previous study[8] showed the fingerprint region of IR spectra from 667-1640 cm $^{-1}$ contained information to discriminate assembly plants and while carbonyl stretch (1650-1750 cm $^{-1}$) was not ; Nevertheless, carbonyl stretch bands were useful to determine plant groups by either doublet (acrylic melamine styrene polyurethane) or singlet (acrylic melamine styrene) in clear coat. IR spectrum range from 2100-2500 cm^{-1} are C-H stretching band existing commonly in paint polymers and also contaminated by diamond anvil cell IR absorption. However, considering about prediction samples from an IR-ATR microscopy source, the IR spectra in the training set were truncated from 680-1641cm^{-1}. All qualified samples were divided into two datasets (doublets/singlet data set) according to the numbers of the carbonyl bands in the clear coat of each sample. The IR truncated data from each layer was scaled by vector normalization. To minimize noises and magnify signals, the data was further preprocessed by "8sym6"wavelet decomposition described in our previous study [10]. The discriminative ability of classifiers were compared by using Savitzky-Golay smoothed IR data and unsmoothed IR data. The IR spectra from the individual layer were smoothed by Savitzky-Golay method. The comparison involved both the single clear coat data and the three layers one (the clear coat horizontally concatenating the two undercoat layers).

4.3.2.2 Data Analysis

The carbonyl stretch bands from IR spectra of the paint clear coat in each assembly plant were visually carefully checked. Two separate datasheets were constituted either by single carbonyl stretch band or doublet carbonyl stretch bands for further data analysis. Before using Hierarchical Cluster Analysis (HCA) for either the singlet sheet or the doublet sheet, unsupervised principal component analysis (PCA) was applied to each assembly plant to assess its class structure (subgroups).The initial sub plant information of a manufacturer was achieved based on the clear coat (OT2) IR truncated data. HCA produced a dendrogram to gather similar paint formulation assembly plants to a cluster and provided initial class subgroups information for the first search prefilter development by using the average OT2 IR spectra of assembly plants in a dataset. Genetic algorithm (GA) identified wavelet coefficients from OT2 “8sym6”wavelet preprocess data to build up the first prefilter, which contains information to pattern samples into different assembly plant groups. The second prefilter was developed to differentiate the manufacturers of a sample locating in the same class of the first prefilter. “8sym6”wavelet preprocess data from the clear coat horizontally concatenated “8sym6”wavelet preprocess data from two undercoats, whose wavelet coefficients were identified by a GA as classifiers in the second prefilter. The detail data fusion technical is described in the previous research at our lab. [22] The final step prefilter was done by GA for pattern recognition analysis of an assembly plant from the manufacturer identified by the second prefilter. The data analysis process was described in Figure 4.11. “8sym6”wavelet preprocess data from the clear coat horizontally concatenated “8sym6”wavelet preprocess data from the surfacer-primer coat were also used to develop the prefilters to assist infrared library searching system. [12]

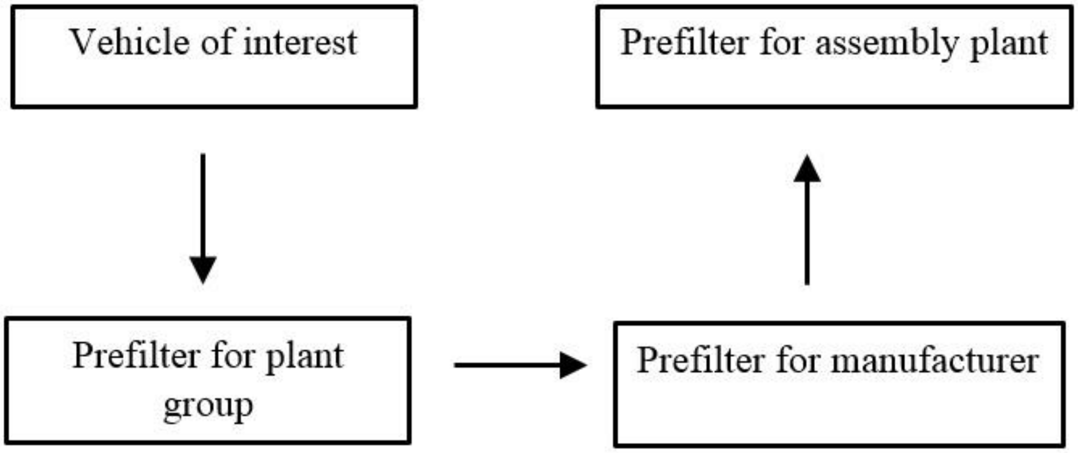

Vehicle of interest
Prefilter for assembly plant
Prefilter for plant group
Prefilter for manufacturer

Figure.4.11 Block diagram of the vehicle classification process using pattern recognition prefilter

4.4 RESULTS AND DISCUSSION

4.4.1 Subplant and Carbonyl Band Information

Sub plant and carbonyl band information of the assembly plants from GA, Chrysler and Ford came from the previous research at our lab and was listed in the Table 4.1-Table 4.2 and Table 4.4-Table 4.5. [22] The split assembly plants from Honda, Nissan and Toyota were showed in Figure 4.12 – Figure 4.16 by the two largest principle components; the split assembly plants with the doublet of carbonyl band were showed in Figure 4.17- Figure 4.18. Their sub plant and carbonyl band information were concluded in Table 4.1, Table 4.3, Table 4.6- Table 4.7.

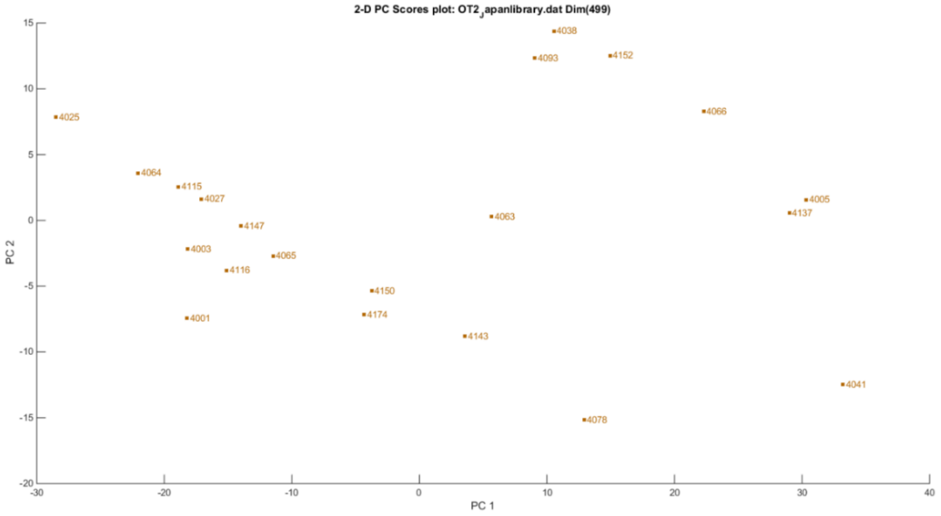


Figure 4.12. 2-PC plot of the samples from assembly plant Kyushu (Nissan)

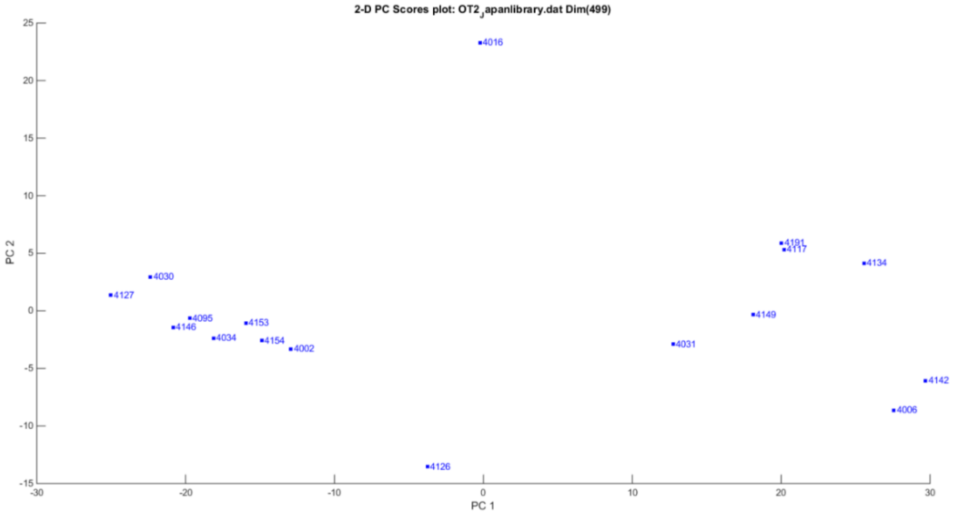


Figure 4.13. 2-PC plot of the samples from assembly plant Oppama (Nissan)

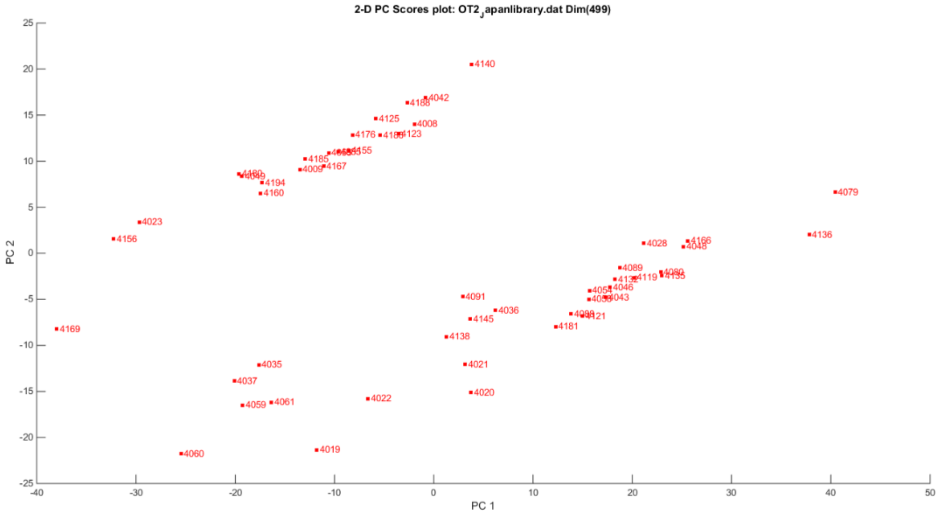


Figure 4.14. 2-PC plot of the samples from assembly plant Smyrna (Nissan)

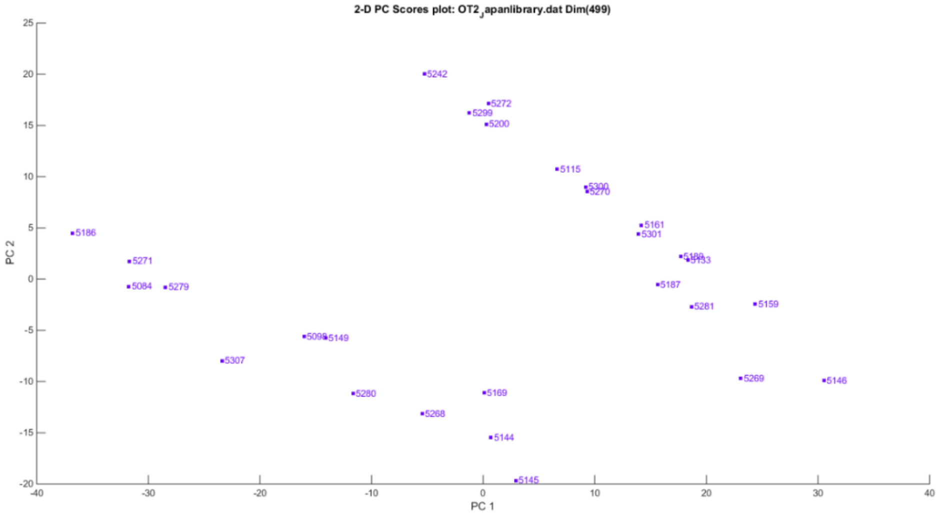


Figure 4.15. 2-PC plot of the samples from assembly plant Fremont (Toyota)

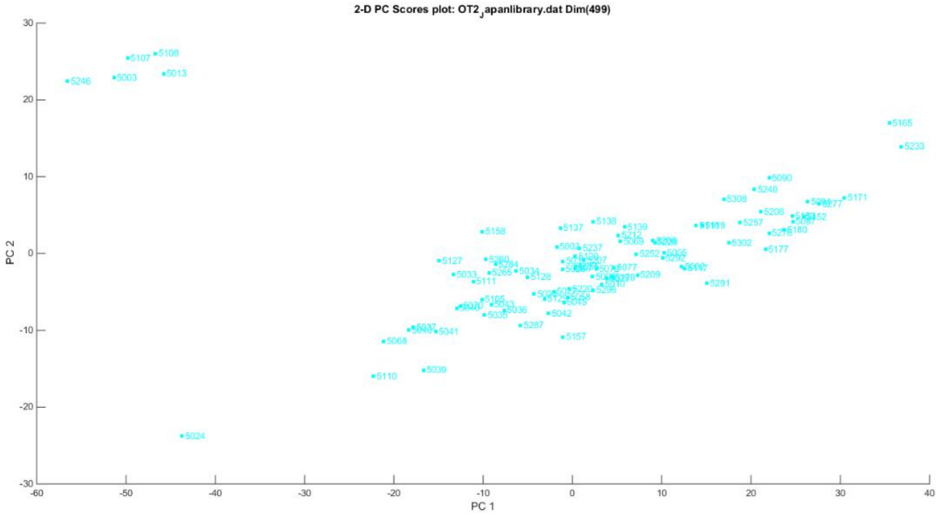


Figure 4.16. 2-PC plot of the samples from assembly plant Japan (Toyota)

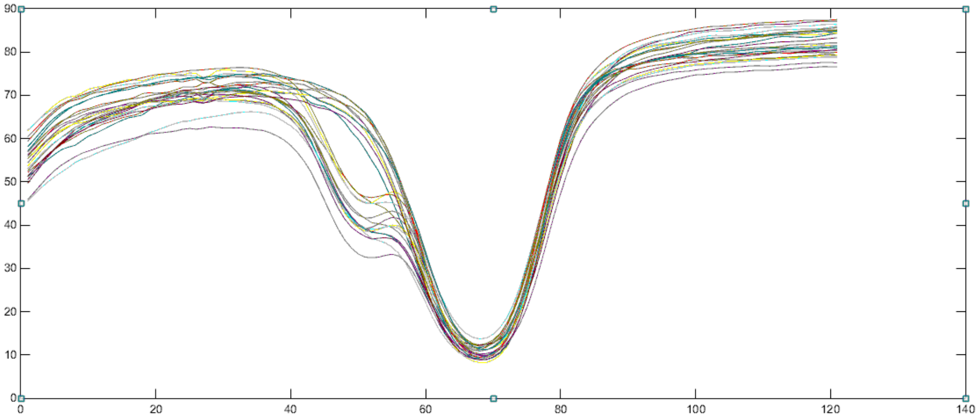


Figure 4.17. The carbonyl IR bands of samples from assembly plant East Liberty (Honda)

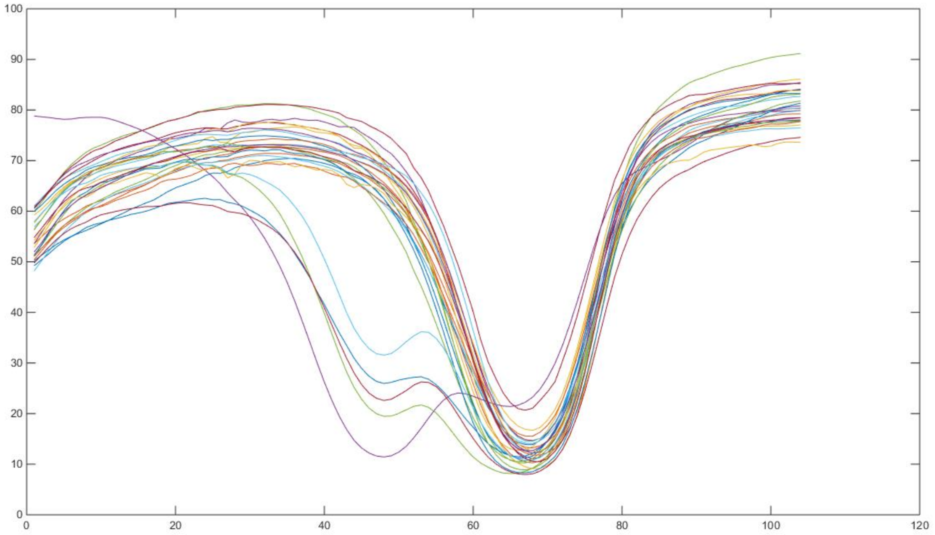


Figure 4.18. The carbonyl IR bands of samples from assembly plant Marysville (Honda)

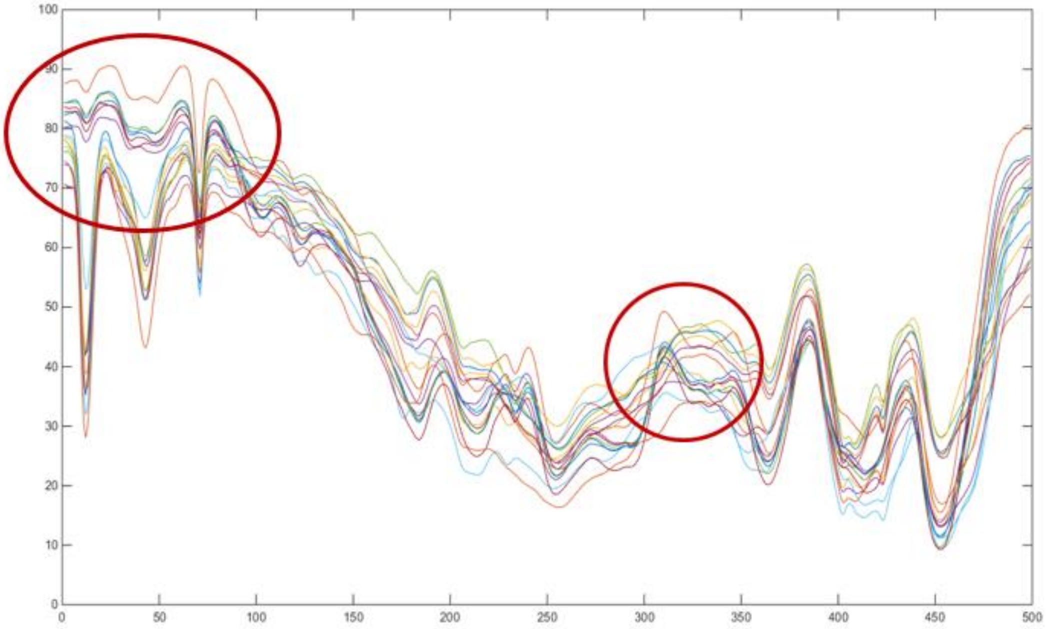


Figure 4.19. IR spectra of the samples from assembly plant Aguascalientes (Nissan)

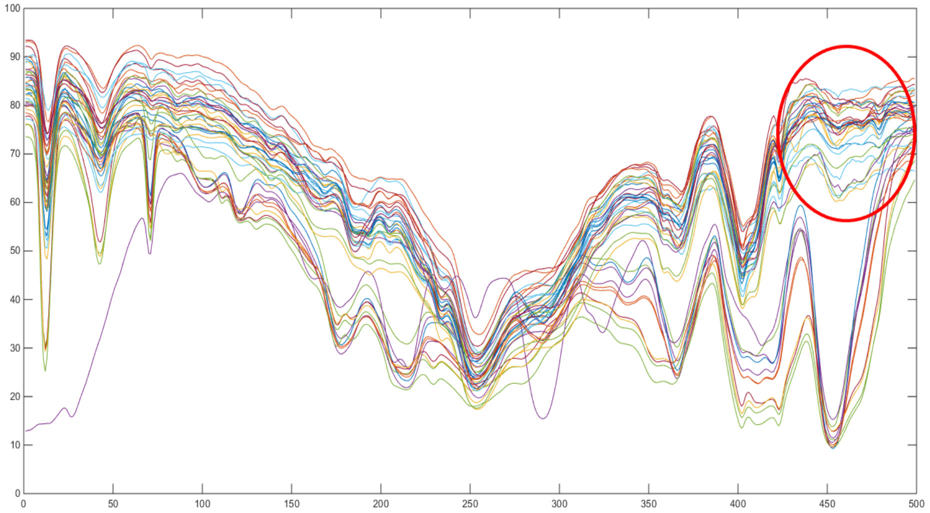


Figure. 4.20 IR spectra of the samples from assembly plant Georgetown (Toyota)

Even if the PCA plots of the samples from assembly plant Georgetown and Aguascalientes did not show obvious split, the IR spectra from the same assembly plant are different, which indicated these two assembly plants would have sub plants representing different clear coat formulations. The IR spectra split in Aguascalientes assembly plant may be caused by paint composition change in 2005. The IR spectra split in Georgetown assembly plant were unclear. The split of the both two assembly plants was not caused by the paint color.

We assumed the PCA split in one assembly plant due to the change of the clear coat paint formulation. Samples from Nissan Kyusha split into two sub plants (PID4004 and 4104) indicated that this assembly plant might use two different clear coat paint composition. Samples from Nissan Oppama split into PID4105 (most from make Nissan) and PID4005 (most from make Infiniti). But they did not have strict quality control for paint composition. Samples from Nissan Smyrna split caused by cars (PID 4006) or trucks

(PID 4106) using two different paints. Samples from Toyota Fremont split because paint formulations were changed in year 2000 – 2002 (PID 5003) and year 2005-2006 (PID5103). Most time Toyota assembly plant Japan used the same clear coat paint except the truck produced before 2002 in line 4Runner and line Prado. Acrylic melamine styrene was the common clear coat paint composition in most assembly plants belong to automotive manufacturers: Honda, Nissan and Toyota.

4.4.2 Hierarchical Cluster Analysis and Principal Component Analysis

4.4.2.1 Doublets

A hierarchical classification scheme and principal component analysis were employed to identify the assembly plant of an automotive paint sample with double carbonyl bands in a clear coat layer. Their results were shown in Figure 4.21 – Figure 4.22. For assembly plants whose polymer formulation for the clear coat is acrylic melamine styrene polyurethane (double carbonyl band), the results of the principal component analysis and the hierarchical cluster analysis suggested to group these ninteen assembly plants and sub plants into five plant groups (see Table 4.8), each plant group was assumed that the chemical composition of a clear coat was similar.

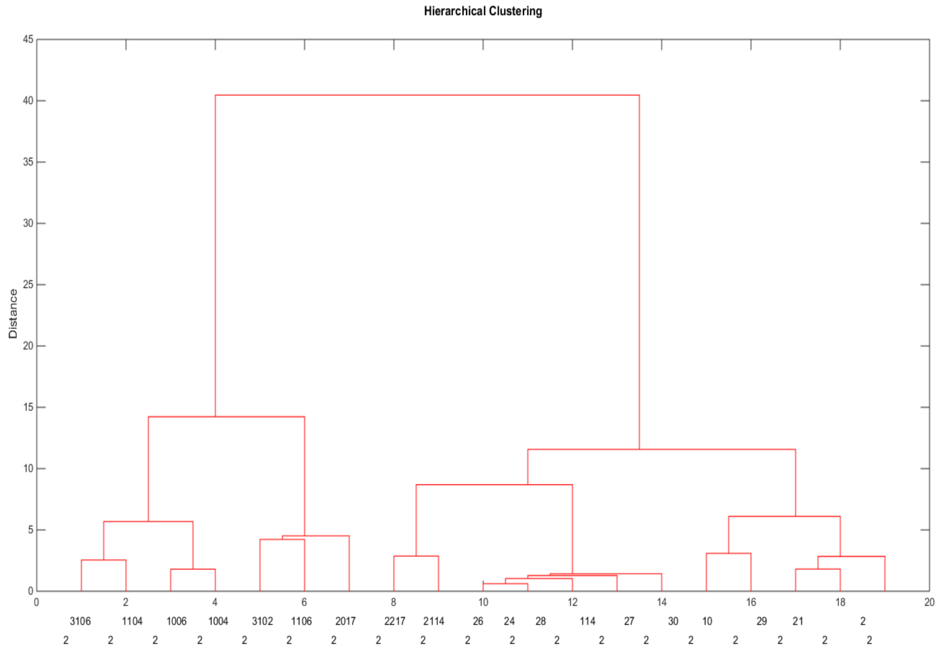


Figure 4.21. All doublet assembly plants hierarchical cluster analysis of the average IR spectrum

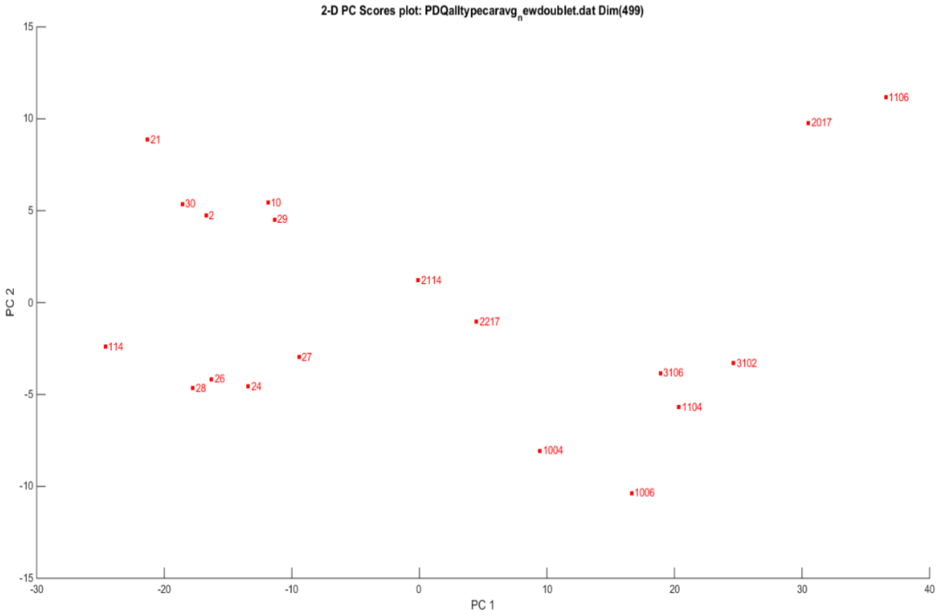


Figure 4.22. All doublet assembly plants principal component analysis of the average IR spectrum

Table 4.8 Double carbonyl bands plant group assignment

Assigned Group Plant Number	Manufacturer Name	Plant Numbers
1	GM	2,10,21,29,30
2	GM	24,26,27,28,114
3	GM + Honda	1106,3102
4	Chrysler + Honda	1004,1006,1104,3106
8	Ford	2017,2114,2217

4.4.2.2 Singlet

A hiearchical classification scheme and principal component analysis were employed to identify the assembly plant of an automotive paint sample with single carbonyl band in a clear coat layer. Their results were shown in Figure 4.23 – Figure 4.24. For assembly plants whose polymer formulation for the clear coat is acrylic melamine styrene (single carbonyl band), the results of the principal component analysis and the hierarchical cluster analysis suggested to group these eighty assembly plants and sub plants into six groups. However, the grouping method was unsucessful as the one applied on doublets. Some assembly plants kept on splitting into sub plants by continuously checking IR spectra and GA run results . Assembly plants from Alliston (Honda), Marysville (Honda), Cambridge (Toyota), Fremount (Toyota), Georgetwon (Toyota) continuously split into further sub plants based on 2-PC plots and the final group information was listed on Table 4.9. Each plant group was assumed that the chemical composition of a clear coat was similar.

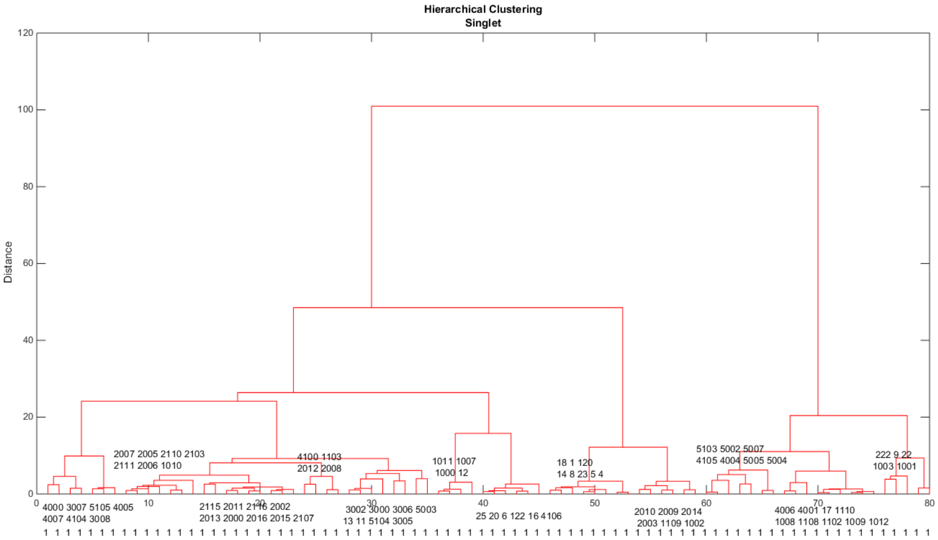


Figure 4.23. All Singlet assembly plants hierarchical cluster analysis of the average IR spectrum

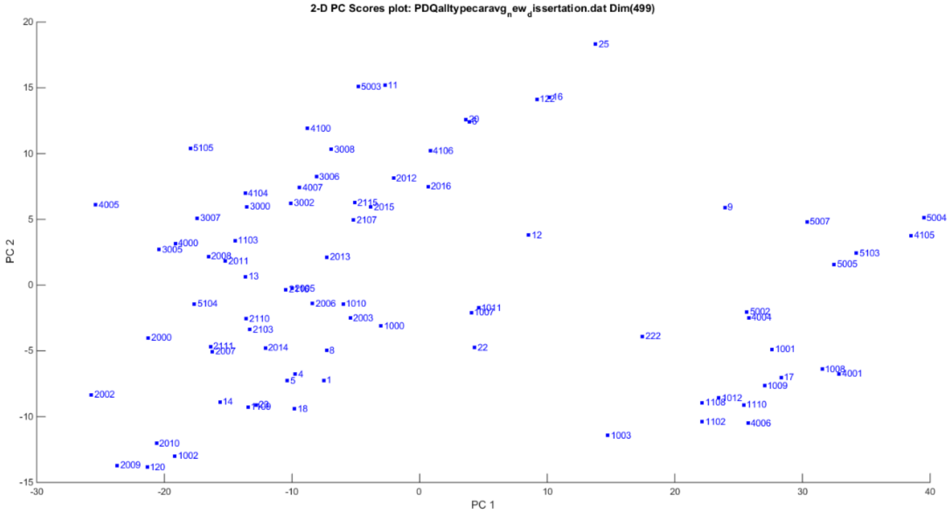


Figure 4.24. All singlet assembly plants principal component analysis of the average IR

Table 4.9 single carbonyl bands plant group assignment

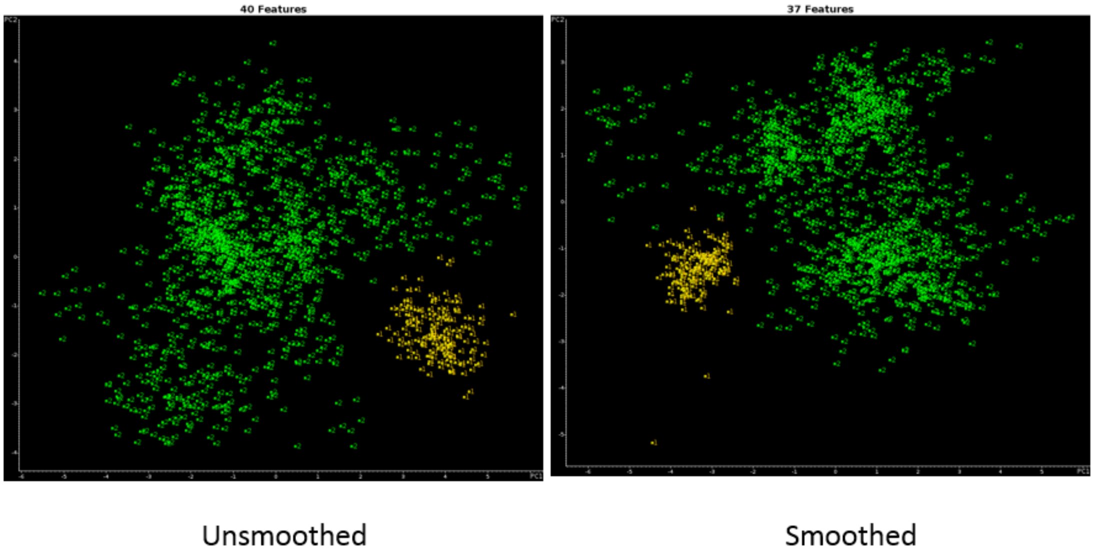
Assigned Plant Group Number	Manufacturer Name & Plant ID’s

1	GM: 1,4,5,6,8,9,11,12,13,14,16,17,18,20, 22,23, 25, 120,122,22
2	Nissan: 4004, 4105 Toyota: 5004,5005,5007, 5102,5103
3	Chrysler: 1000,1001,1003,1007,1008,1009,1011,1012,1102,1108,1110 Nissan: 4001, 4006 Toyota: 5002,5203
4	Chrysler: 1002,1010,1103,1109 Ford: 2007,2106,2014,2006,2005,2013,2011,2002,2010,2110, 2003,2008,2012,2016,2015,2107 Honda: 3106,3100 Nissan: 4100,4106
5	Ford: 2116,2113,2111,2103,2206,2009,2115 Honda: 3000,3002,3005,3006 Toyota: 5003,5104,5303
6	Honda: 3007,3008,3200 Nissan: 4000,4005,4007,4104 Toyota: 5105,5204

4.4.3 Smoothed Data versus Unsmoothed Data

The Spectra in the fingerprint region from the clear coat and two undercoats were smoothed using Savitzky-Golay filter (4th order polynomial, 17 point window) respectively. The smoothed IR spectra of each paint layer were vector normalized and then transformed by using 8Sym6 individually. [22] The wavelet coefficients from OT2, OU1 and OU2 of a sample were horizontally concatenated into a single vector by the method described in our previous study. [22] The same processes were applied on unsmoothed IR spectra. Smoothed data and unsmoothed data were compared by using the same GA classification process. All experimental results (Figure 4.25 – Figure 4.28) disclosed Savitzky-Golay smooth data did not improve the GA classifications significantly than unsmooth data. For the first prefilter, the 2-PC plots of GA pattern recognition have no significant difference between the smoothed data and the unsmoothed data (see Figure

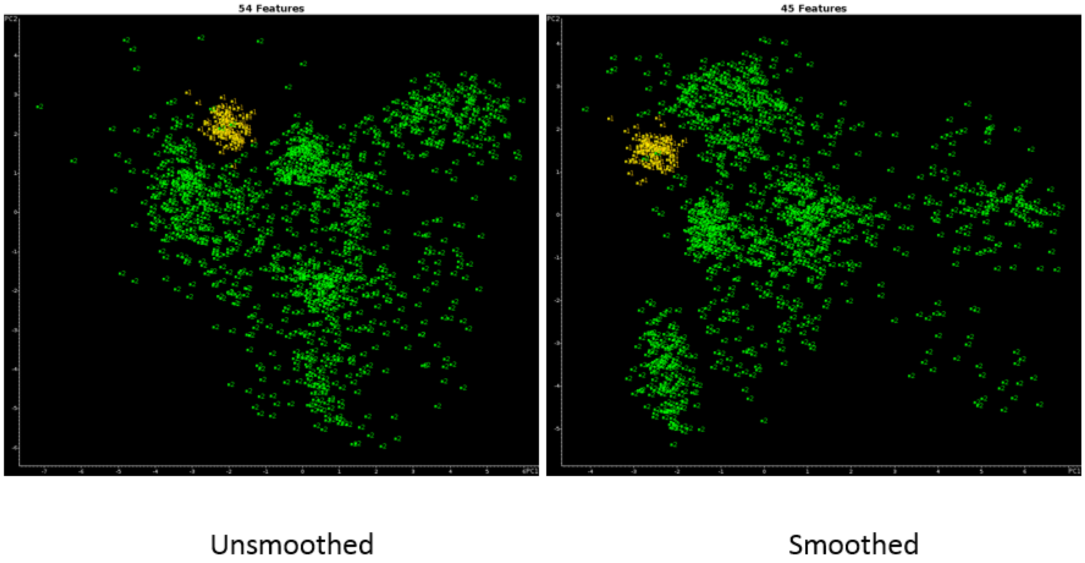
4.29). Therefore, in this study, prefilters were developed from the IR spectra data without Savitzky-Golay smooth.



Unsmoothed

Smoothed

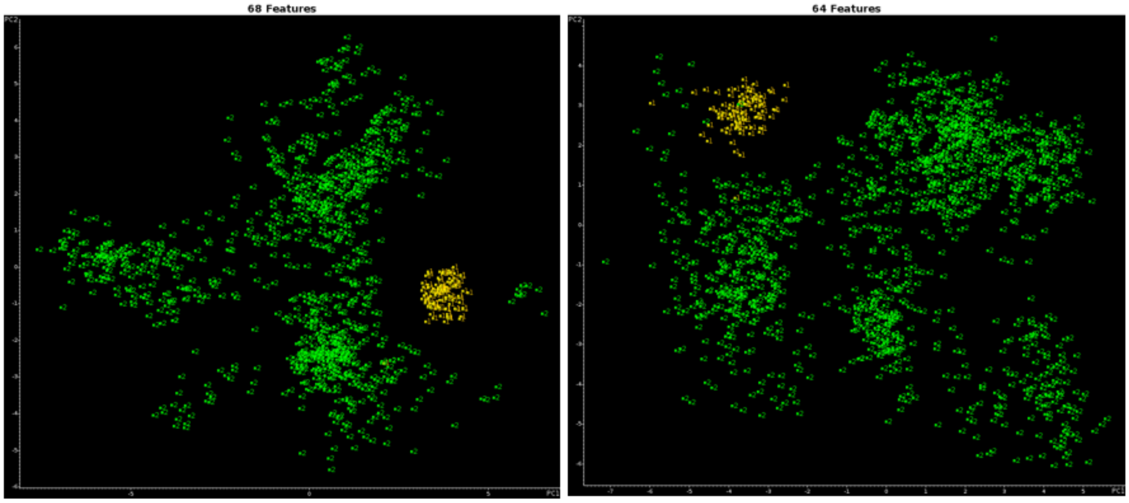
Figure 4.25. 2-PC plot of the 1373 paint samples with 3426 wavelet coefficients comprising the training set data (1 = Toyoda, 2 = GM + Chrysler + Ford + Honda +Nissan)



Unsmoothed

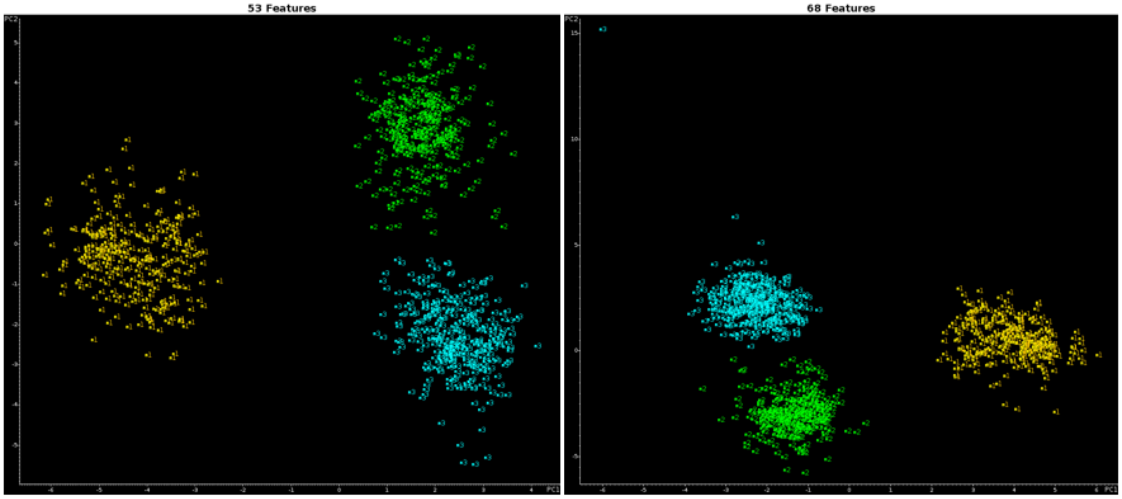
Smoothed

Figure 4.26. 2-PC plot of the 1373 paint samples with 3426 wavelet coefficients comprising the training set data (1 = Nissan, 2 = GM + Chrysler + Ford + Honda +Toyota)



Unsmoothed Smoothed

Figure 4.27. 2-PC plot of the 1373 paint samples with 3426 wavelet coefficients comprising the training set data (1 = Honda, 2 = GM + Chrysler + Ford + Nissan +Toyota)



Unsmoothed Smoothed

Figure 4.28. 2-PC plot of the 976 paint samples with 3426 wavelet coefficients comprising the training set data (1 = GM, 2 = Chrysler, 3 = Ford)

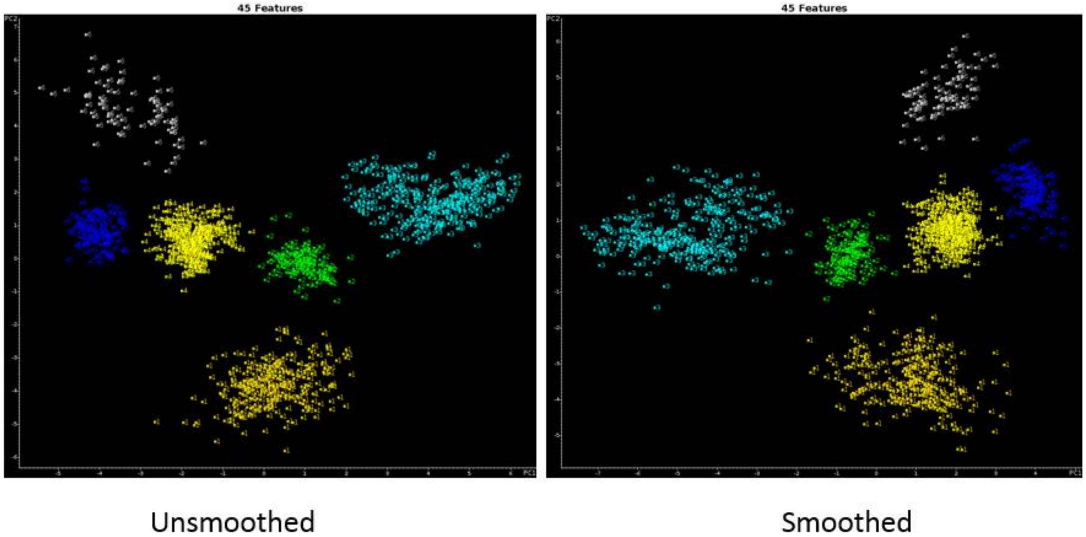


Figure 4.29. 2-PC plot of the 1377 paint samples with 1142 wavelet coefficients comprising the training set data (1=group1, 2=group2, 3=group3, 4=group4, 5=group5, 6=group6)

4.4.4 Search Prefilter for Assembly Plant Groups

A genetic algorithm for feature selection and pattern recognition analysis was applied on both single carbonyl band sample (see Table 4.10) data sheet and double carbonyl bands sample (see Table 4.11) data sheet to identify wavelet coefficients to characterize the similar paint formulation in a clear coat layer based on assembly plants. After 200 generations, the pattern recognition GA identified the certain numbers of wavelet coefficients whose 2-PC plot exhibited clustering of the clear coat IR spectra on the basis of assembly plants from six automotive manufacturers. The 2-PC plots for the singlet training and validation set were seen in Figure 4.30-Figure 4.31. However, the pattern recognition GA was unable to identify the wavelet coefficients of the clear coat IR spectra to cluster assembly plants on the basis of automotive manufacturers (See Figure 4.32). The 2-PC plots for the doublet training and validation set were seen in

Figure 4.35-Figure 4.36. This results suggests that information about automotive manufacturers cannot be directly obtained from the clear coat but this layer provides the information to narrow down the search scope of automotive manufacturers, since each group of assembly plants are consisted of specific automotive manufacturers.

4.4.4.1 Singlet

Table 4.10 The distribution of the training set and validation set for the first prefilter

Plant Group	Manufacturer	Training Samples	Validation Samples
1	GM	324	30
2	Nissan, Toyota	147	18
3	Chrysler, Nissan, Toyota	313	29
4	Chrysler, Ford, Honda, Nissan	414	37
5	Ford, Honda, Toyota	126	15
6	Honda, Nissan, Toyota	79	8

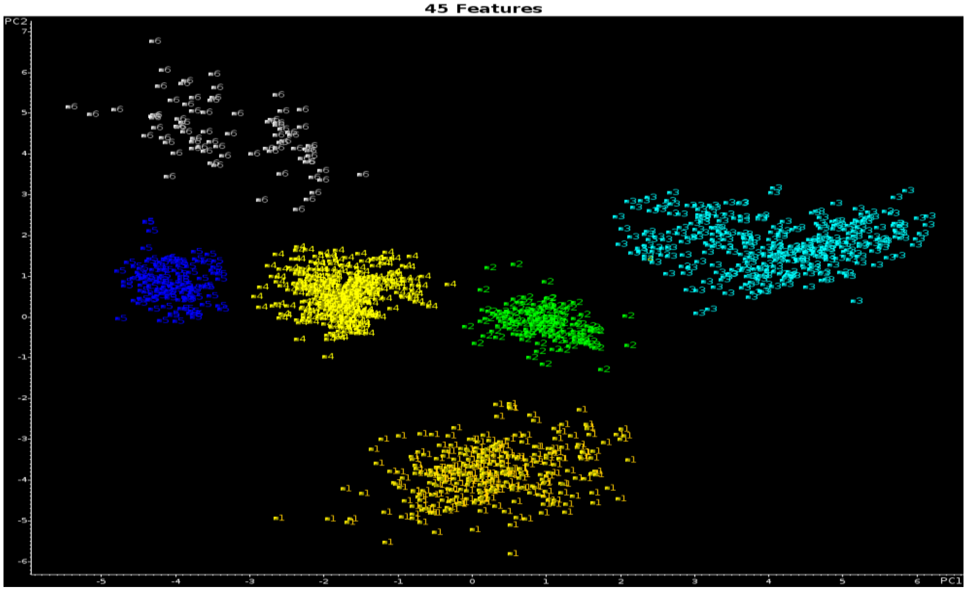


Figure 4.30. 2-PC plot of the 1377 training set samples and the 45 wavelet coefficients identified by the pattern recognition GA (1= group1, 2=group2, 3=group3, 4=group4, 5=group5, 6=group6)

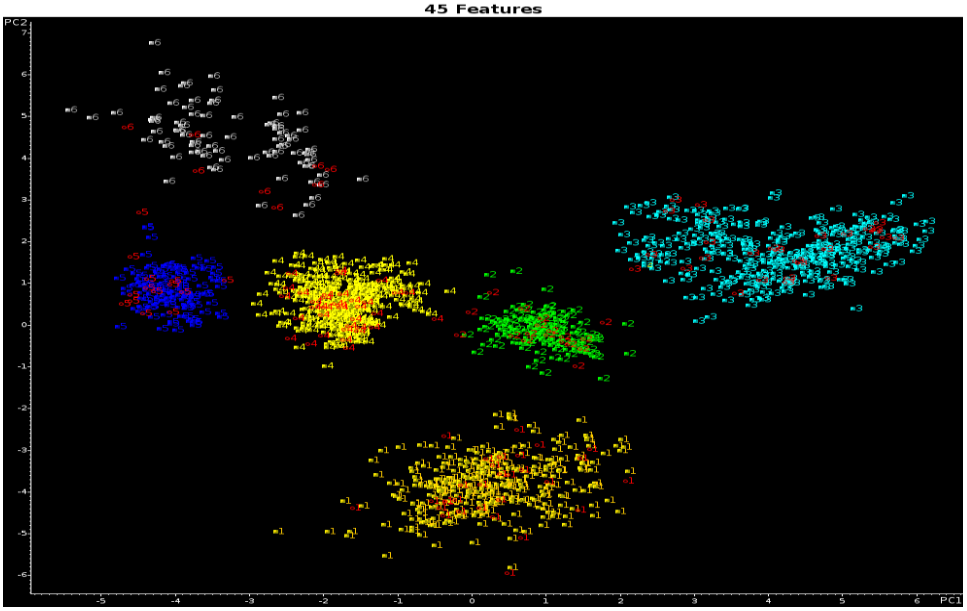


Figure 4.31. Projection of the 137 validation set samples onto the PC plot of the 1377 training set samples and the 45 wavelet coefficients identified by the pattern recognition GA(1= group1, 2=group2, 3=group3, 4=group4, 5=group5, 6=group6)

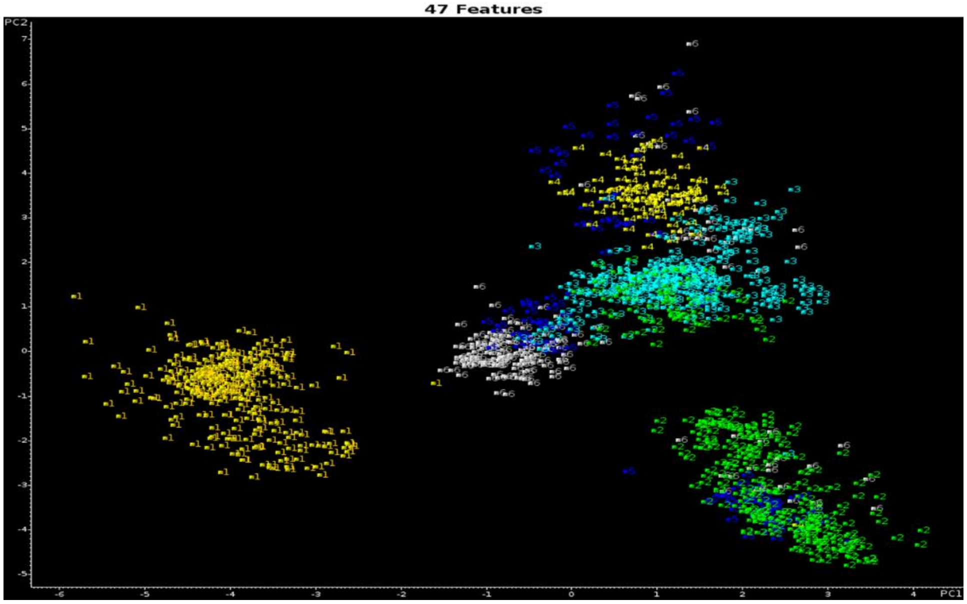


Figure 4.32. 2-PC plot of the 1377 training set samples and the 47 wavelet coefficients identified by the pattern recognition GA based on automotive manufacturer (1=GM, 2=Chrysler, 3=Ford, 4=Honda, 5=Nissan, 6=Toyota)

To build up the search prefilter for assembly plant groups whose samples have single carbonyl band, 26 outliers were removed. Outliers were identified by comparing the sample IR spectrum with the average IR spectrum of the assembly plant it came from (see Figure 4.33); or by comparing its IR spectrum with the IR spectra from the other samples in the same assembly plant (see Figure4.34). Samples in the plant Group1 for the first search prefilter are all from GM, this suggests that the clear coat paint formulation of GM is different from other 5 manufacturers. The first prefilter is able to predict an unknown sample produced from GM or not. It also can narrow down the manufacturers from the information provided by the plant group if an unknown sample is located out of plant group 1 (GM).

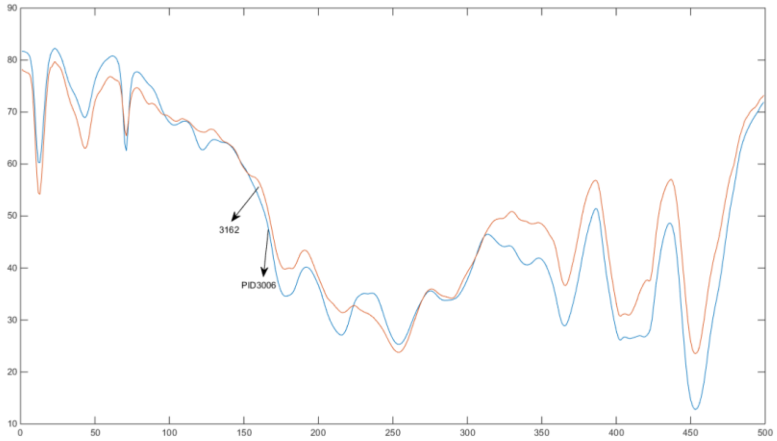


Figure 4.33. The OT2 IR spectrum of the sample vs the average sample OT2 IR spectrum of the assembly plant Marysville (Outlier: sample ID3162)

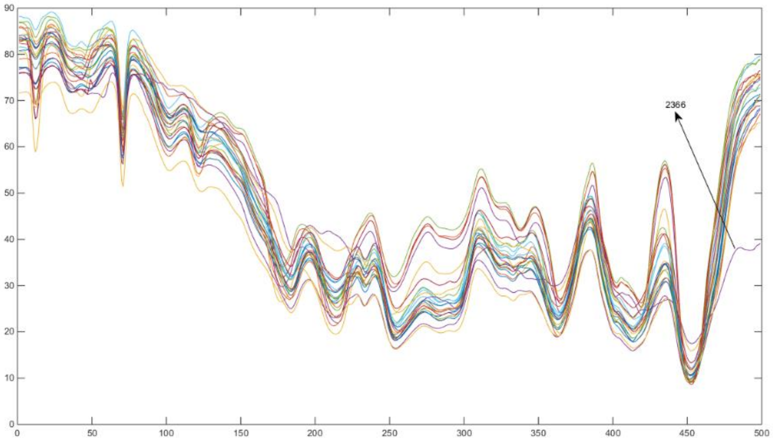


Figure 4.34. The OT2 IR spectra of the samples in the assembly plant Hemosillo (Outlier: sample ID2366)

4.4.4.2 Doublets

Table 4.11 The distribution of the training set and validation set for the first prefilter

Plant Group	Manufacturer	Training Samples	Validation Samples
1	GM	48	6
2	GM	61	7
3	Chrysler, Honda	24	2
4	Chrysler, Honda	59	6
8	Ford	21	3

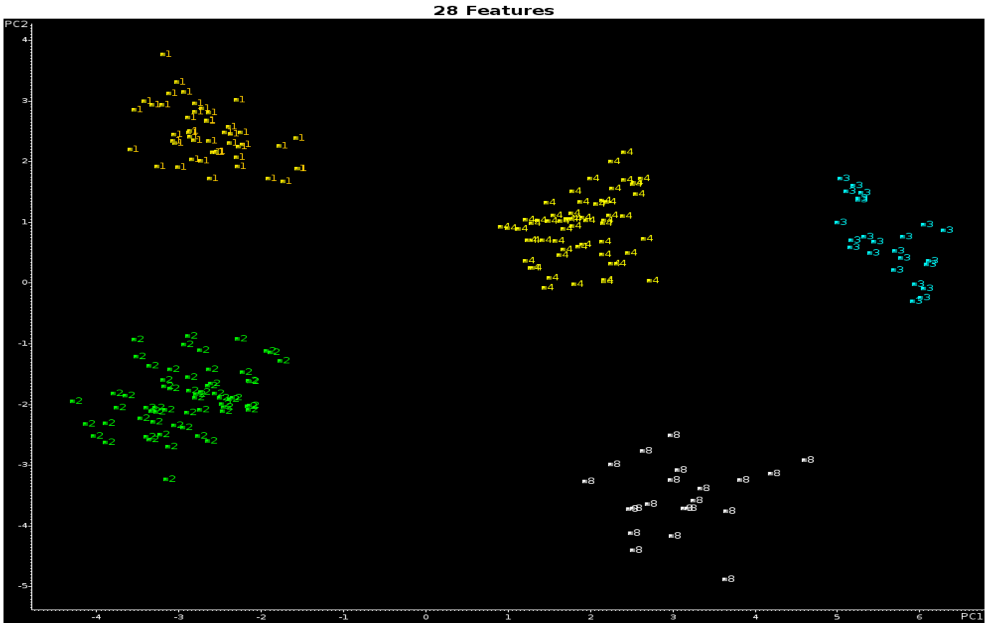


Figure 4.35. 2-PC plot of the 213 training set samples and the 28 wavelet coefficients identified by the pattern recognition GA (1= group1, 2=group2, 3=group3, 4=group4, 8=group5)

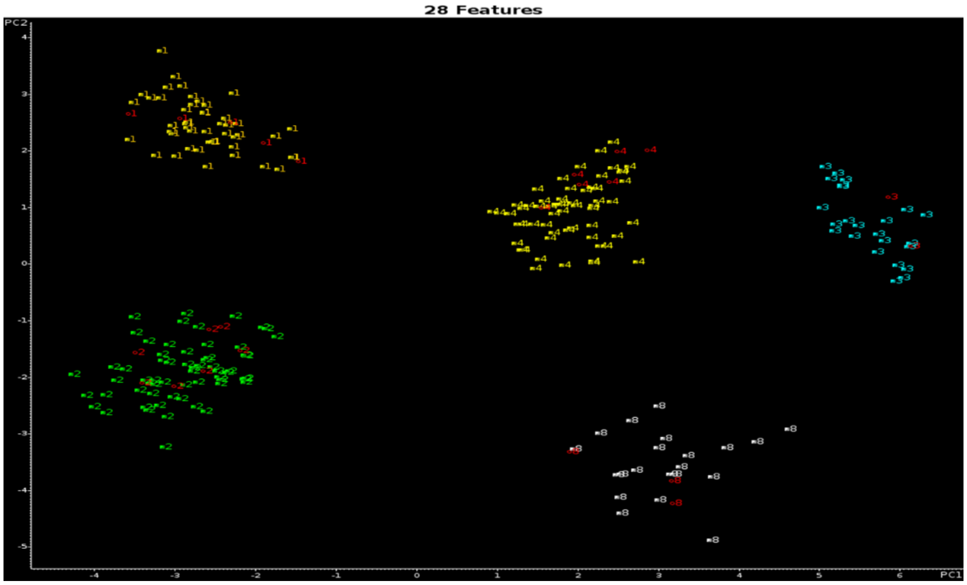


Figure 4.36. Projection of the 23 validation set samples onto the PC plot of the 213 training set samples and the 28 wavelet coefficients identified by the pattern recognition GA(1= group1, 2=group2, 3=group3, 4=group4, 8=group5)

For doublets, the first prefilter was build up without outlier remove. Nevertheless, one validation sample (SID172) was identified as an outlier and removed from the validation set (see Figure 4.37). Samples located in group 1 and group2 were from GM, while samples located in group 8 were produced by Ford. But samples in group3 and group4 were either from Chrysler or Honda. This suggests that a sample with acrylic melamine styrene polyurethane clear coat is easier to obtain the information of automotive manufacturer than a singlet one.

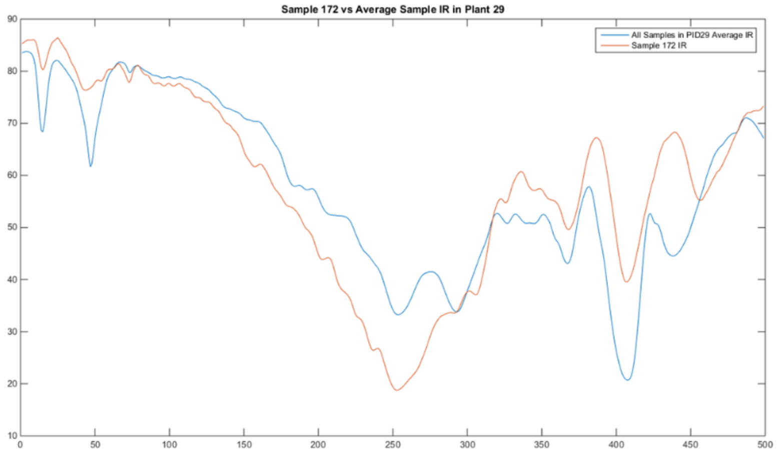


Figure 4.37. The OT2 IR spectrum of the sample vs the average sample OT2 IR spectrum of the assembly plant Wentzville (Outlier: sample ID172)

4.4.5 Search Prefilter For Manufacturers And Assembly Plants

Since the clear coat layer cannot provide enough information to differentiate the automotive manufacturer and the assembly plant, two under coat layers were expected to add discriminatory ability of assembly plants or subplants by horizontally concatenating with the clear coat. Doublets samples located in a specific group identified by the first prefilter only need one step to obtain the assembly plant information. Nevertheless, singlet samples located in a specific group classified by the first prefilter need one more step (search prefilter for manufacturers) to obtain the assembly plant information.

4.4.5.1 Doublets

Doublets group 1 comprises of 5 GM assembly plants (Baltimore, Hamtramck, Orion, Wentzville, and Wilmington). After 11 generations, pattern recognition GA identified 10 wavelet coefficients from three layers to discriminate the samples by assembly plant or sub plant, the results were listed in Figure 4.38- Figure 4.39. Sample (SID172) was previously identified as an outlier and was removed from assembly plant Wentzville in the validation set. Assembly plant Wilmington comprises 4 samples and no validation sample was set in this assembly plant by computer. The other validation samples were predicted correctly.

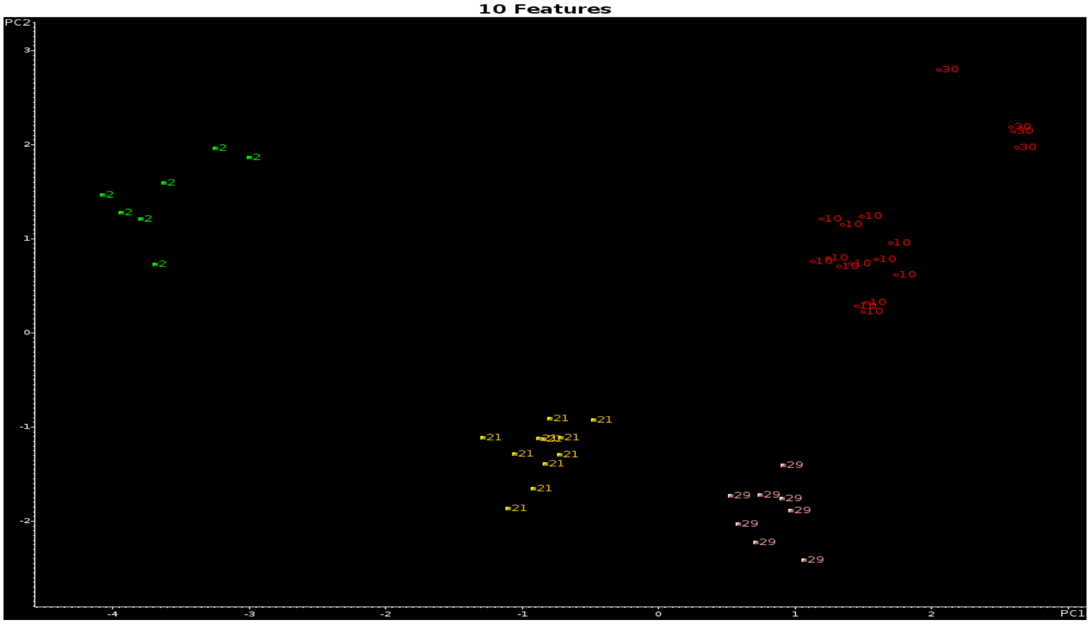


Figure 4.38. 2-PC plot of the 43 training set samples and the 10 wavelet coefficients identified by the pattern recognition GA (2 = Baltimore, 10=Hamtramck, 21=Orion, 29=Wentzville, 30=Wilmington)

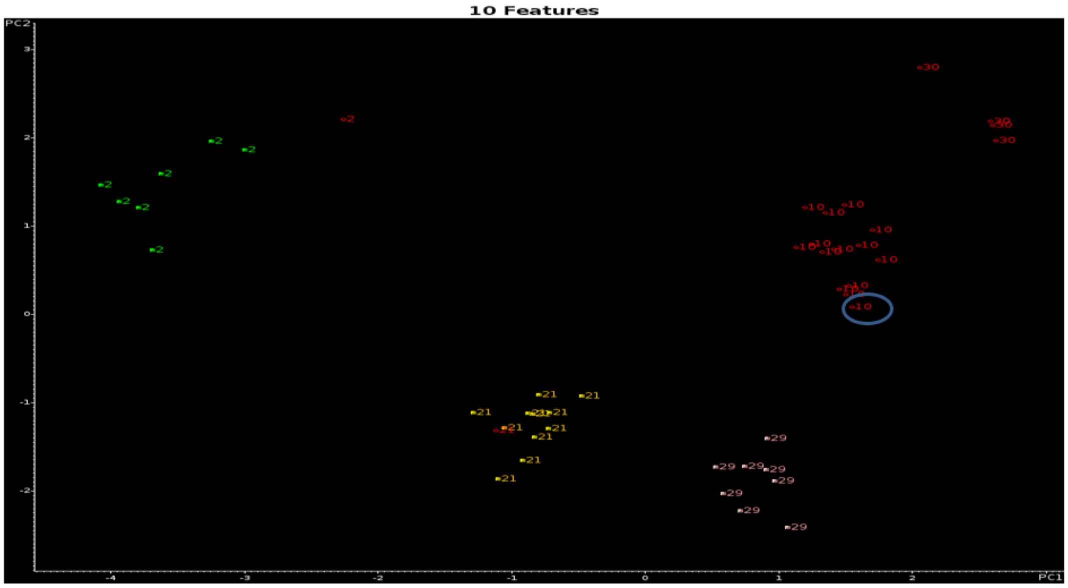


Figure 4.39. Projection of the 3 validation set samples onto the PC plot of the 43 training set samples and the 10 wavelet coefficients identified by the pattern recognition GA (2 = Baltimore, 10=Hamtramck, 21=Orion, 29=Wentzville, 30=Wilmington)

Doublets group 2 consists of 5 GM assembly plants (Ramos Arizpe, Silao, Spring Hill, Saint Therese and Lansing). Pattern recognition GA identified 26 wavelet coefficients from three layers after 66 generations to discriminate the samples by assembly plant or sub plant, the results were listed in Figure 4.40- Figure 4.41. Sample (SID 367) was identified as an outlier and removed from the training set. All the validation samples were predicted correctly.

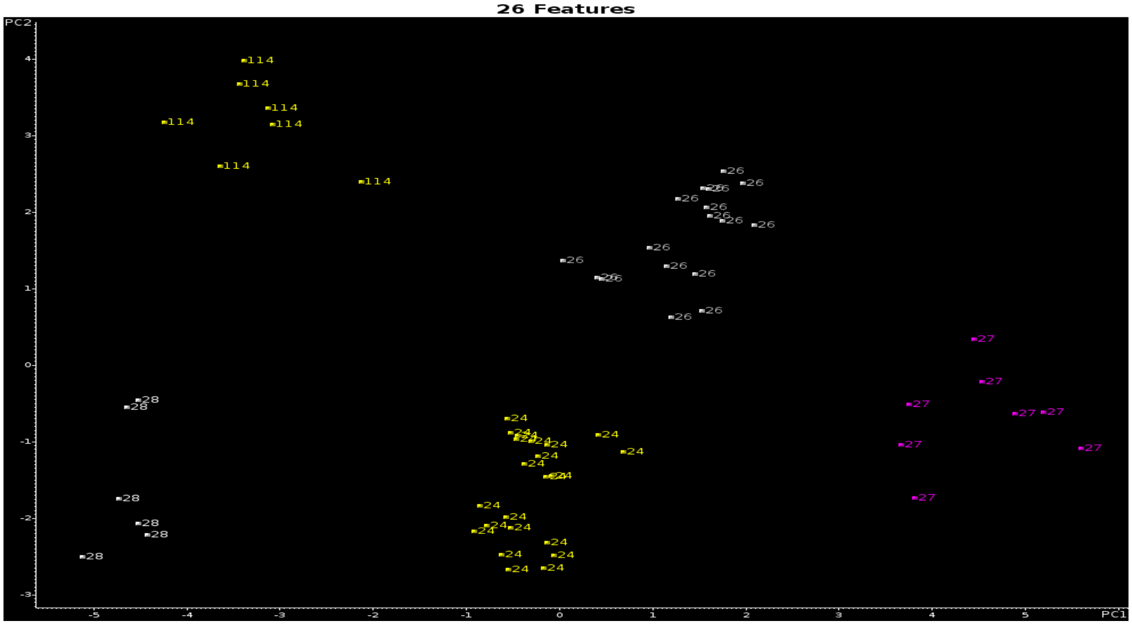


Figure 4.40. 2-PC plot of the 60 training set samples and the 26 wavelet coefficients identified by the pattern recognition GA (24 = Ramos Arizpe, 26= Silao, 27=Spring Hill, 28=Saint Therese, 114=Lansing)

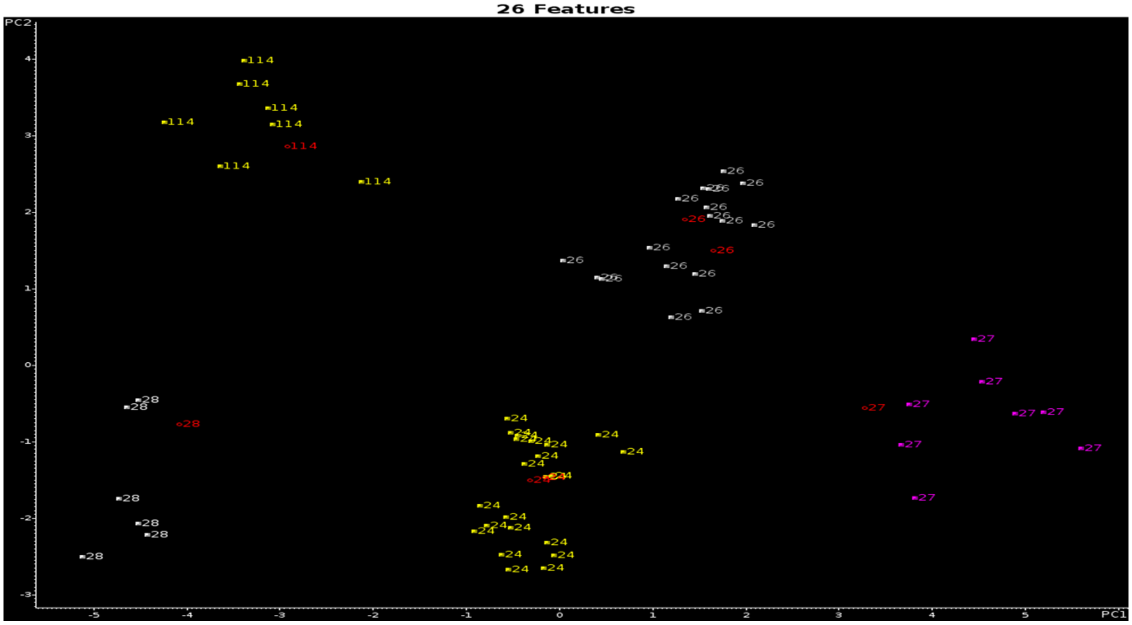


Figure 4.41. Projection of the 7 validation set samples onto the PC plot of the 60 training set samples and the 26 wavelet coefficients identified by the pattern recognition GA (24 = Ramos Arizpe, 26= Silao, 27=Spring Hill, 28=Saint Therese, 114=Lansing)

Doublets group 3 includes two sub plants (Newark and East Liberty). After 1 generations pattern recognition GA identified 2 wavelet coefficients from three layers to discriminate the samples by assembly plant or sub plant, the results were listed in Figure 4.42- Figure 4.43. The validation samples were in the region of their belonging cluster.

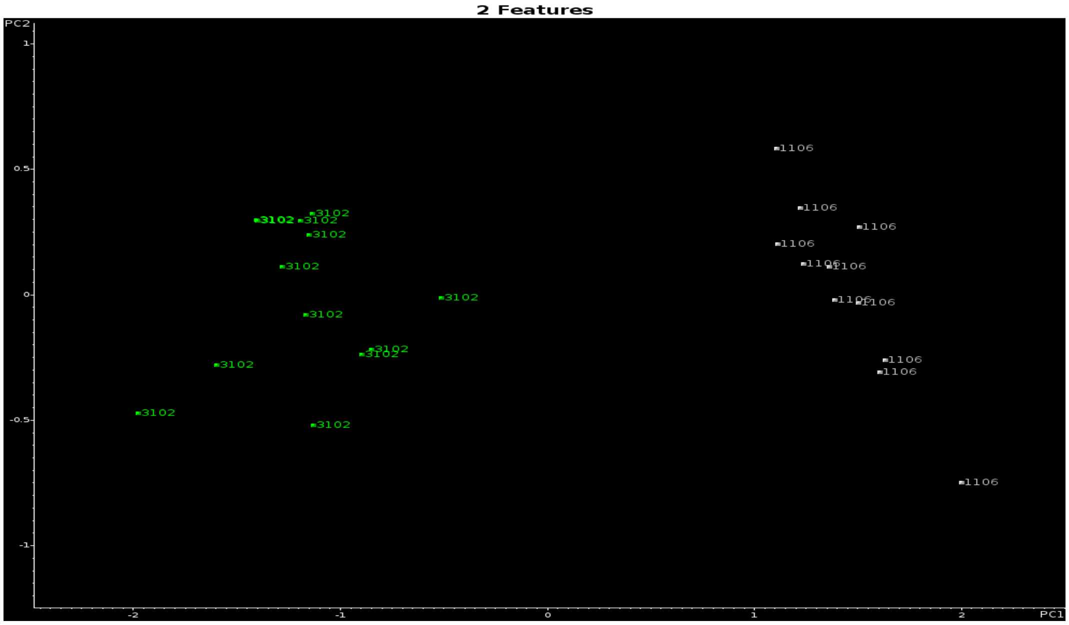


Figure 4.42. 2-PC plot of the 24 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (1106=Newark, 3102=East Liberty)

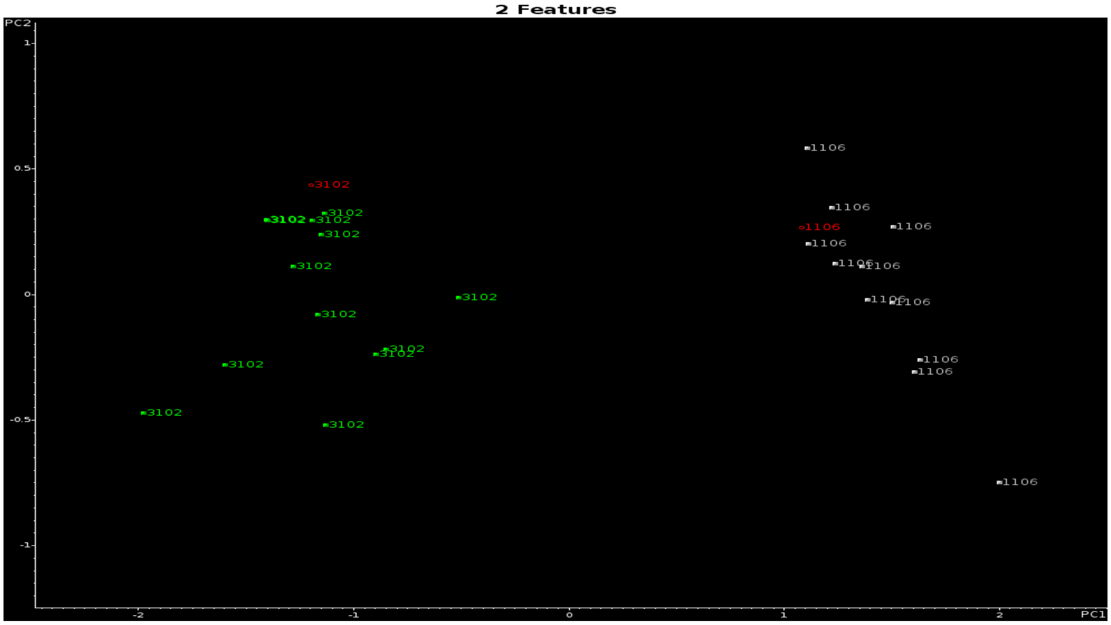


Figure 4.43. Projection of the 2 validation set samples onto the PC plot of the 24 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (1106=Newark, 3102=East Liberty)

1 Honda assembly plant Marysville, 2 Chrysler assembly plants (Jefferson North and Newark) and 1 sub plant (Jefferson North) comprises the doublets group 4. After 42 generations and remove one outlier (SID 1161), the pattern recognition GA identified 20 wavelet coefficients whose 2-PC plot showed clustering on the basis of assembly plant (see Figure 4.44). Each projected validation set sample was located in the region of the map with paint samples from the same assembly plant or sub plant (see Figure 4.45).

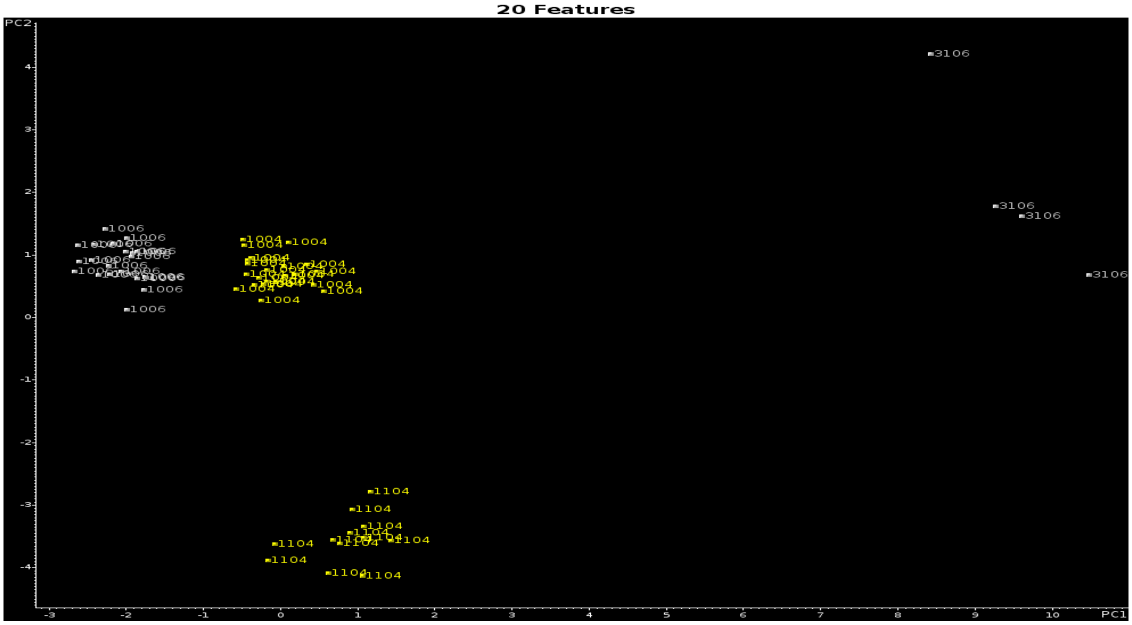


Figure 4.44. 2-PC plot of the 59 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (1004= Jefferson North, 1006=Newark, 1104= Jefferson North, 3106=Marysville)

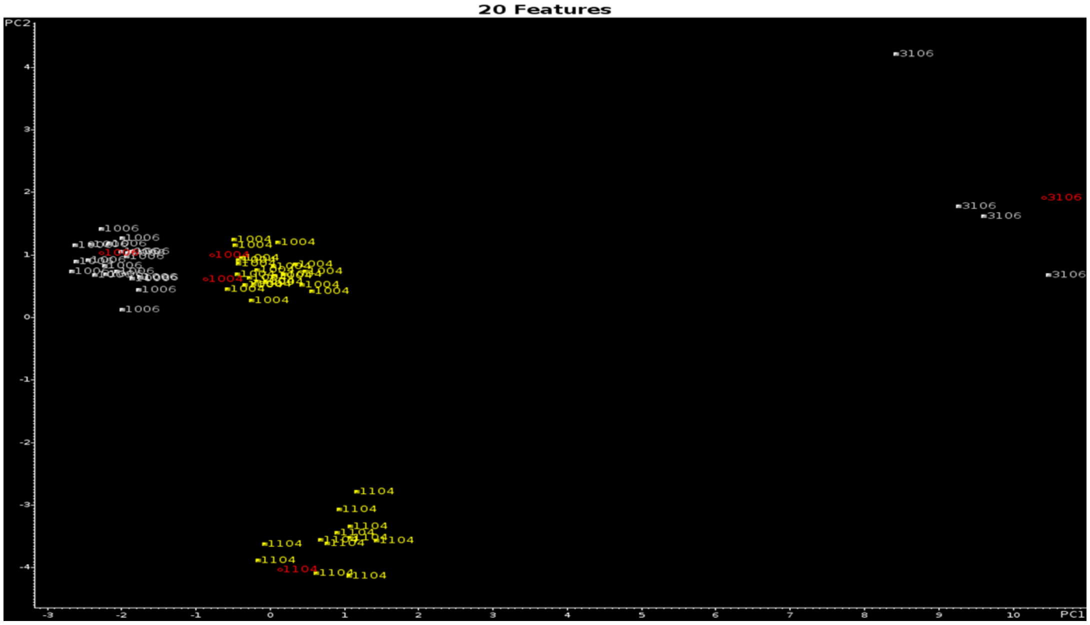


Figure 4.45. Projection of the 5 validation set samples onto the PC plot of the 59 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (1004= Jefferson North, 1006=Newark, 1104= Jefferson North, 3106=Marysville)

The assembly plants or sub plant of the doublets group 8 are all from Ford (Saint Thomas-Talbotsville and wixom). After 2 generations, the pattern recognition GA identified 6 wavelet coefficients whose 2-PC plot showed clustering on the basis of assembly plant (see Figure 4.46). Each projected validation set sample was located in the region of the map with paint samples from the same assembly plant or sub plant (see Figure 4.47).

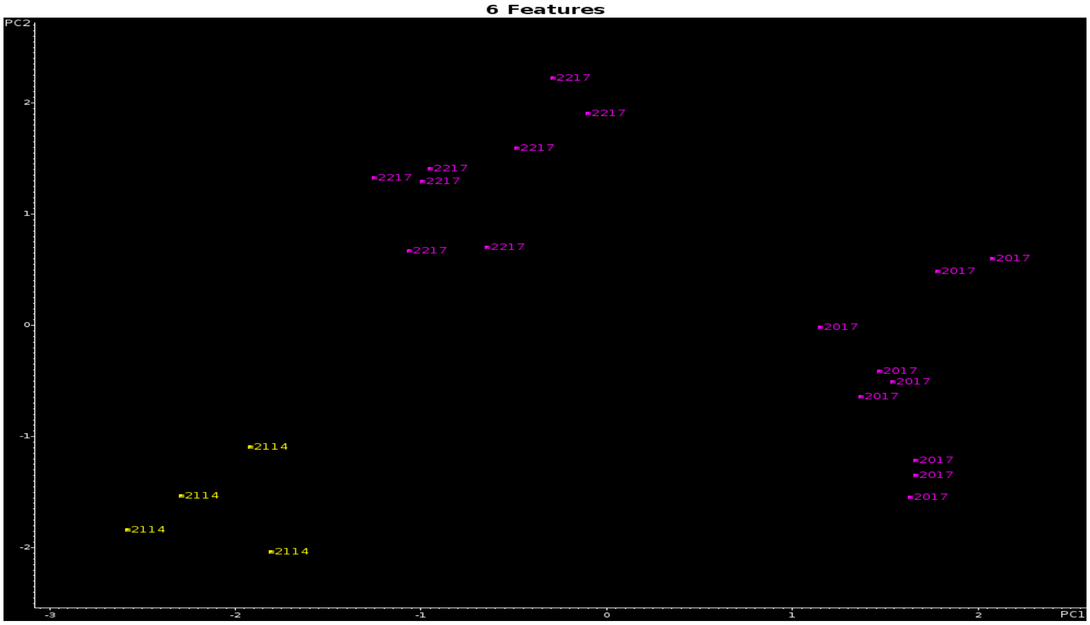


Figure 4.46. 2-PC plot of the 21 training set samples and the 6 wavelet coefficients identified by the pattern recognition GA (2017=Wixom, 2114= Saint Thomas-Talbotsville, 2217=Wixom)

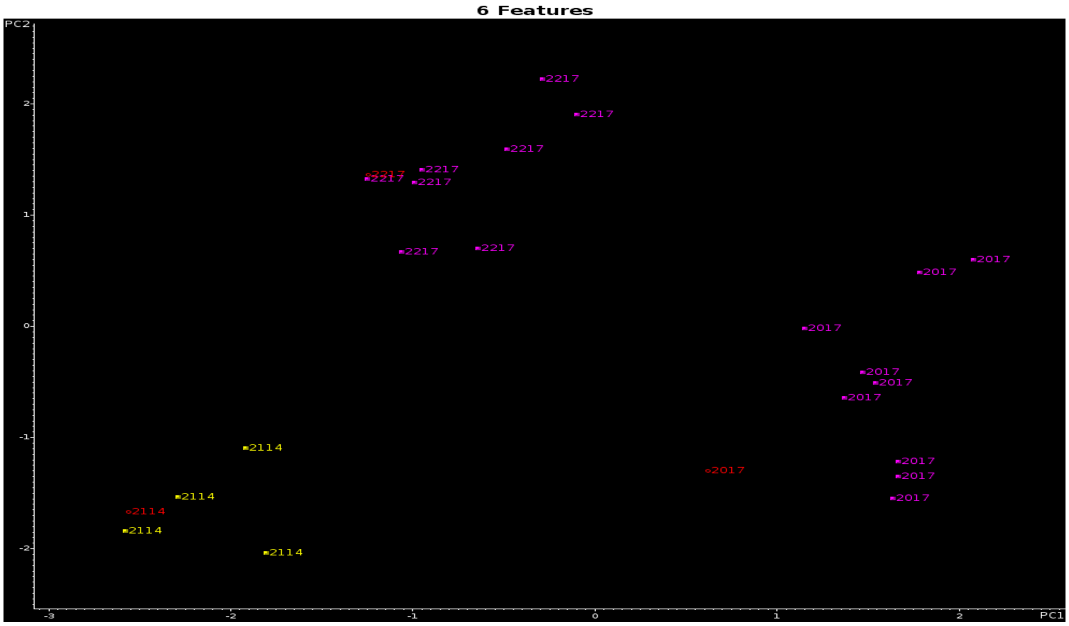


Figure 4.47. Projection of the 3 validation set samples onto the PC plot of the 21 training set samples and the 6 wavelet coefficients identified by the pattern recognition GA (2017=Wixom, 2114= Saint Thomas-Talbotsville, 2217=Wixom).

4.4.5.2 Singlet

After the membership of each plant group was ascertained in the first level (1142 wavelet coefficients), the second prefilter (3426 wavelet coefficients) was developed to distinguish the samples by manufacturers in each plant group. “8sym6” preprocessed data based on clear coat, surfacer – primer and e-coat layers were conjugated together to achieve this goal. The training and validation sets for manufacturer differentiation in each plant group were summarized in Table 4.12. Unlike the doublets, the second prefilter cannot achieve the assembly plant or sub plant information directly after conjugated three paint layers (3426 wavelet coefficients) , because the total numbers of assembly plant or sub plant in the second prefilter are beyond the limitation window space of pattern recognition GA. The third prefilter was developed following ascertained manufacturer information.

Each sample have total 3426 wavelet coefficients and GA identified wavelet coefficients characteristic of manufacturer information and then characteristic of plant information according to the method I described in Figure 4.11 (4.3.2.2 Data analysis). The experimental results were seen Figure 4. 49 – Figure 4.63.

The samples located in the plant group 1 are all from manufacturer GM. To obtain the information of assembly plant and sub plant, the method was taken from the previous study [22] as following (Figure 4.48). For paint samples from GM, the clear coat IR spectra had enough information to linearly differentiate GM from other manufacturers and could be used as manufacturer level prefilter. The sub manufacturer groups were classified by the comparison of average IR spectra of assembly plant or sub plant. The assembly plants or sub plants with similar OT2 IR spectra were defined as the same sub GM manufacturer ID numbers.

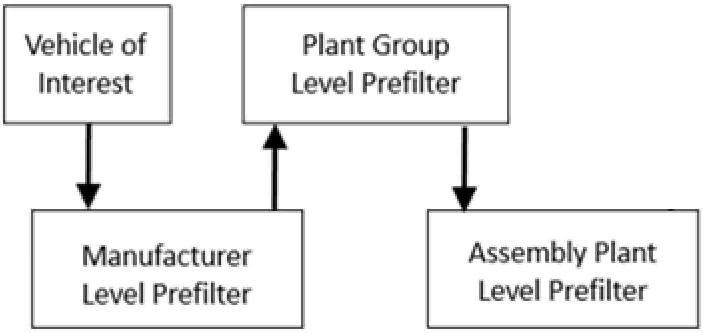

Vehicle of Interest
Plant Group Level Prefilter
Manufacturer Level Prefilter
Assembly Plant Level Prefilter

Figure 4.48. Block diagram of the vehicle classification process for GM

Table 4.12. Composition of the IR spectral data set in plant group 1 (GM)

Manufacturer	Manufacturer sub IDs	Plant IDs	Training set samples	Validation set samples
	1	1, 4, 14	68	9
	2	18, 120	31	3

GM	3	5, 8, 22, 23	70	6
	4	9, 12, 17, 222	71	8
	5	6, 11, 20, 122	40	5
	6	16, 25	31	3

After 200 generations, pattern recognition GA (Fitness function: Hopkins 0.1) identified 50 wavelet coefficients whose PC plot (see Figure 4.49) showed clustering of the fused IR spectra on the basis of their sub GM group. The 34 validation set samples were then projected onto the PC plot (see Figure 4.50) define by the 311 training set samples and the 50 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 sub GM group.

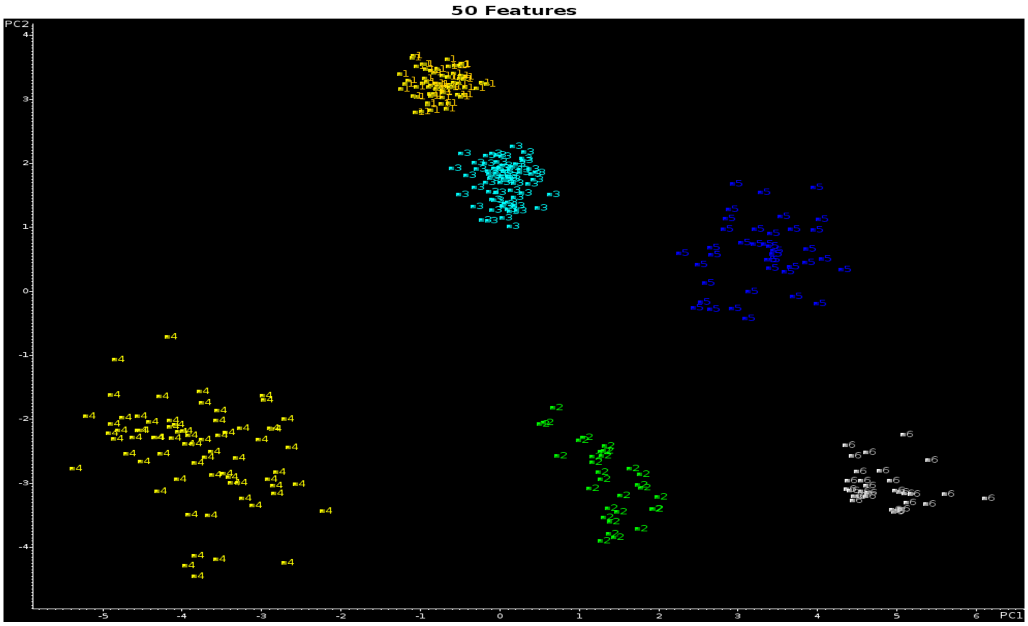


Figure 4.49. 2-PC plot of the 311 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA (1=GMsubgroup1, 2=GMsubgroup2, 3= GMsubgroup3, 4=GMsubgroup4, 5=GMsubgroup5, 6=GMsubgroup6)

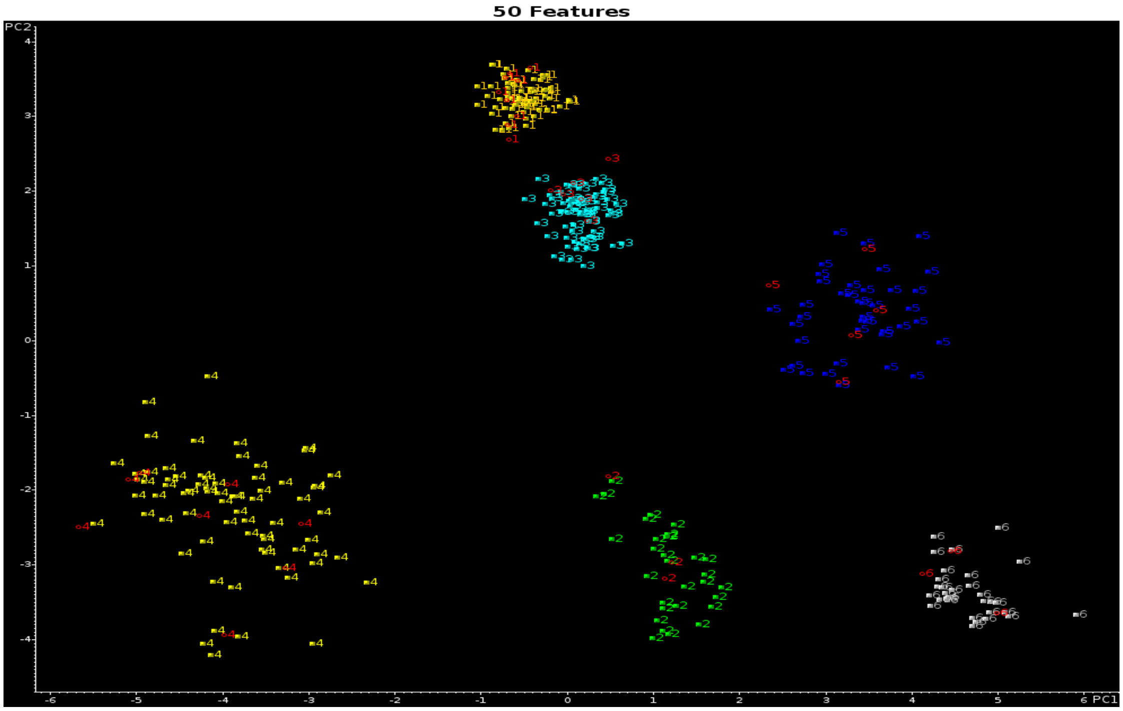


Figure 4.50. Projection of the 34 validation set samples onto the PC plot of the 311 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA (1=GMsubgroup1, 2=GMsubgroup2, 3= GMsubgroup3, 4=GMsubgroup4, 5=GMsubgroup5, 6=GMsubgroup6)

After 84 generations, Figure 4.51- Figure 4.52 showed the 2-PC plots of training set samples and the validation set samples in the GMsubgroup1 by using pattern recognition GA. All the validation samples were in the region of their belonging cluster.

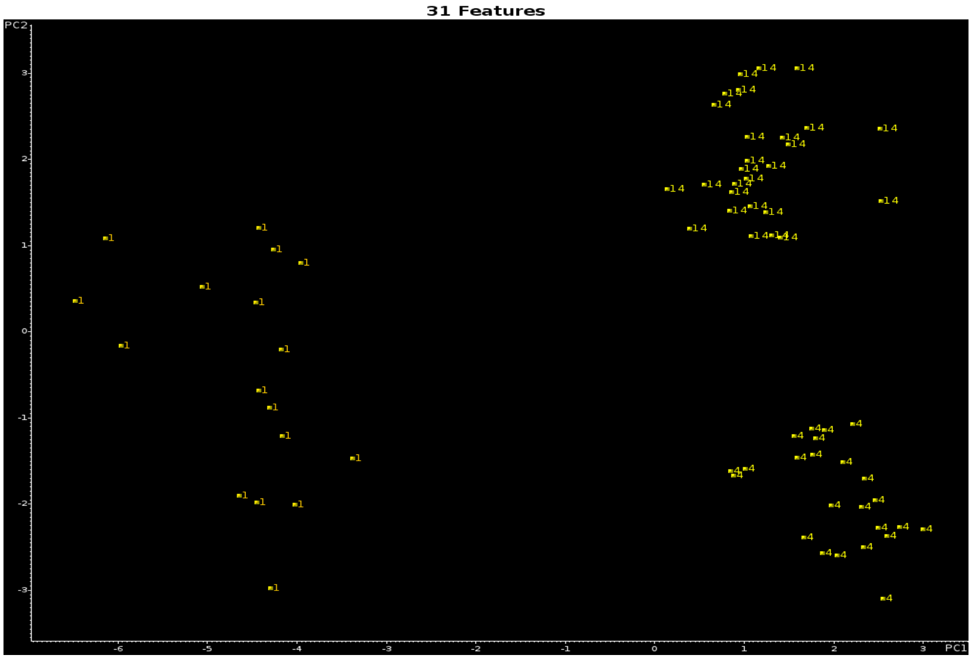


Figure 4.51. 2-PC plot of the 68 training set samples and the 31 wavelet coefficients identified by the pattern recognition GA (1=Arlington, 4=Doranlle, 14=Lansing)

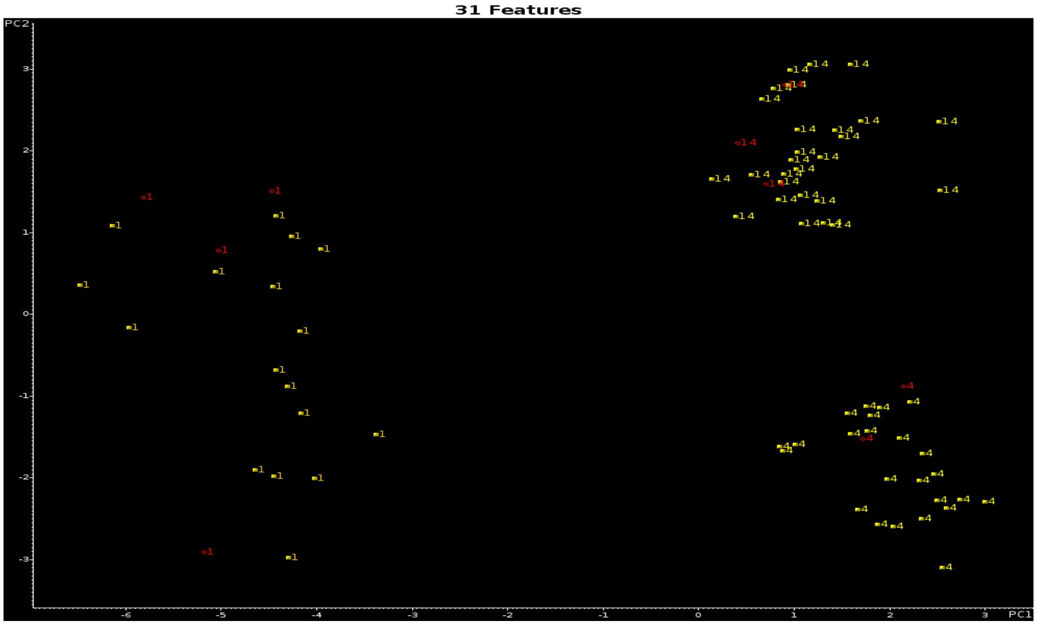


Figure 4.52. Projection of the 9 validation set samples onto the PC plot of the 68 training set samples and the 31 wavelet coefficients identified by the pattern recognition GA(1=Arlington, 4=Doraville, 14=Lansing)

After 5 generations, the pattern recognition GA identified 3 wavelet coefficients whose PC plot (see Figure 4.53) showed clustering of the spectra on the basis of GM sub group2.To assess the predictive ability of these 3 wavelet coefficients, a validation set of 2 paint samples were projected into 2-PC developed from the 32 training set and the wavelet coefficients identified by GA using the normal routine of the pattern recognition GA. The validation set samples were assigned to correct assembly plant (see Figure 4.54). The training set samples from Oklahoma City sub plant were less than ten and no validation sample was assigned by computer.

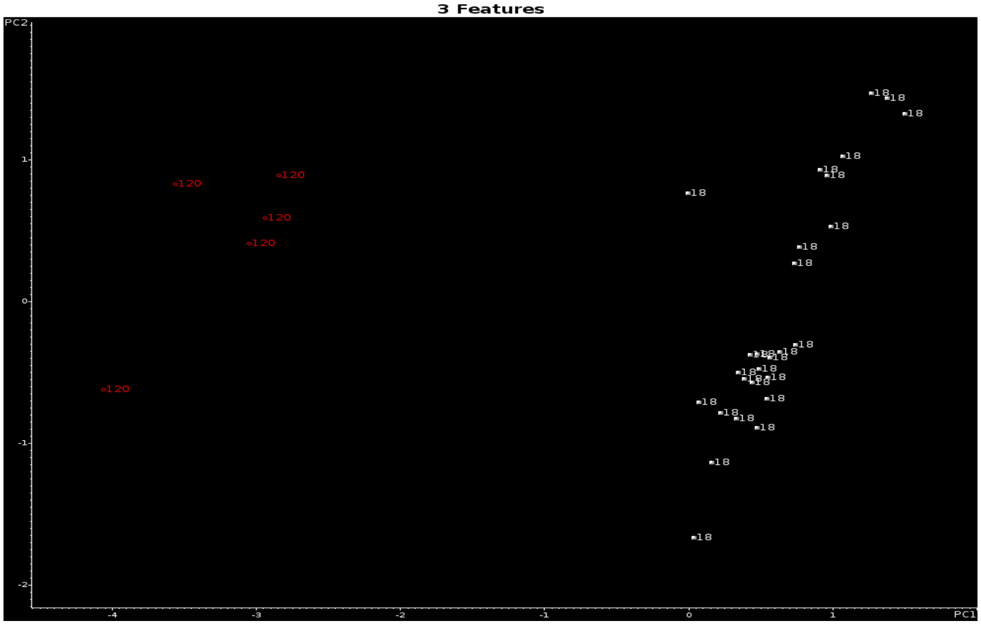


Figure 4.53. 2-PC plot of the 32 training set samples and the 3 wavelet coefficients identified by the pattern recognition GA(18= Moraine, 120=Oklahoma City)

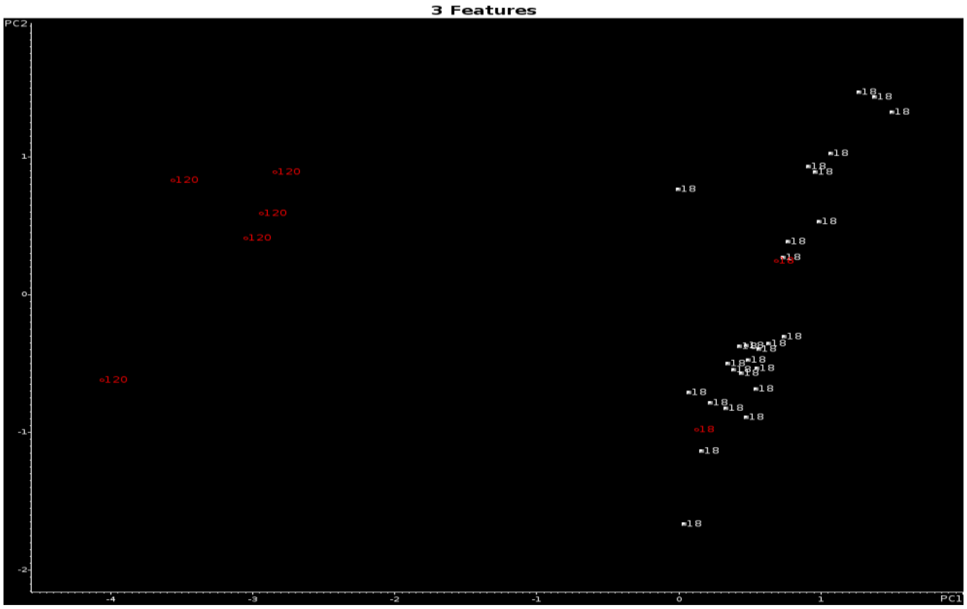


Figure 4.54. Projection of the 2 validation set samples onto the PC plot of the 32 training set samples and the 3 wavelet coefficients identified by the pattern recognition GA (18= Moraine, 120=Oklahoma City)

After 167 generations, the pattern recognition GA identified 44 wavelet coefficients whose PC plot (see Figure 4.55) showed clustering of the spectra on the basis of GM sub group3. The average IR spectra of clear coat, surfacer-primer and e-coat from assembly plant Fort Wayne and Pontiac are similar, so these two assembly plants merged together into a new assembly plant whose ID is 823(see Figure 4.57). To assess the predictive ability of these 44 wavelet coefficients, a validation set of 6 paint samples were projected into 2-PC developed from the 70 training set and the wavelet coefficients identified by GA using the Hopkin 0.1 of the pattern recognition GA. The validation set samples were assigned to a correct assembly plant (see Figure 4.56).

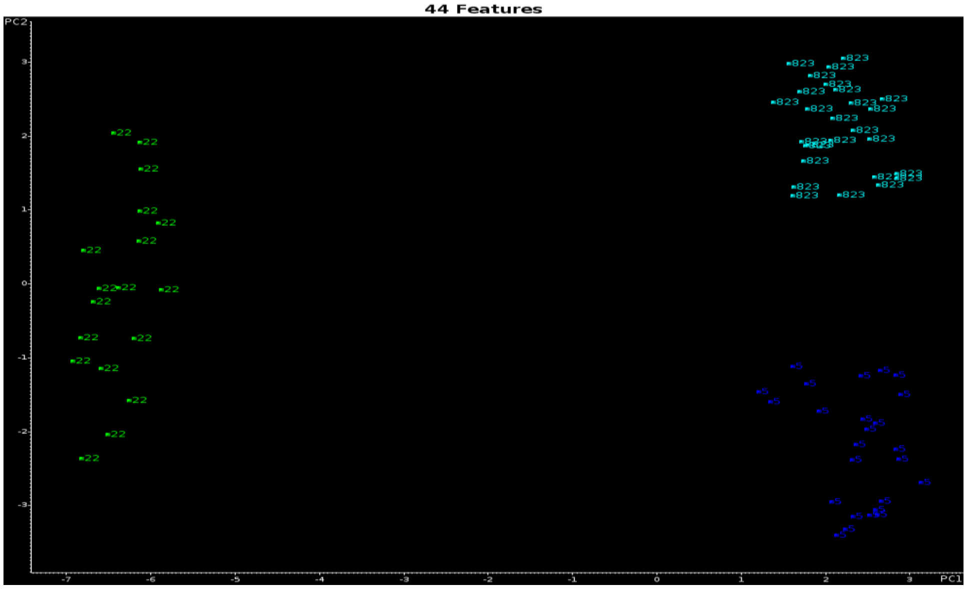


Figure 4.55. 2-PC plot of the 70 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA (5=Fairfax, 22=Oshawa, 823= Fort Wayne merged with Pontiac)

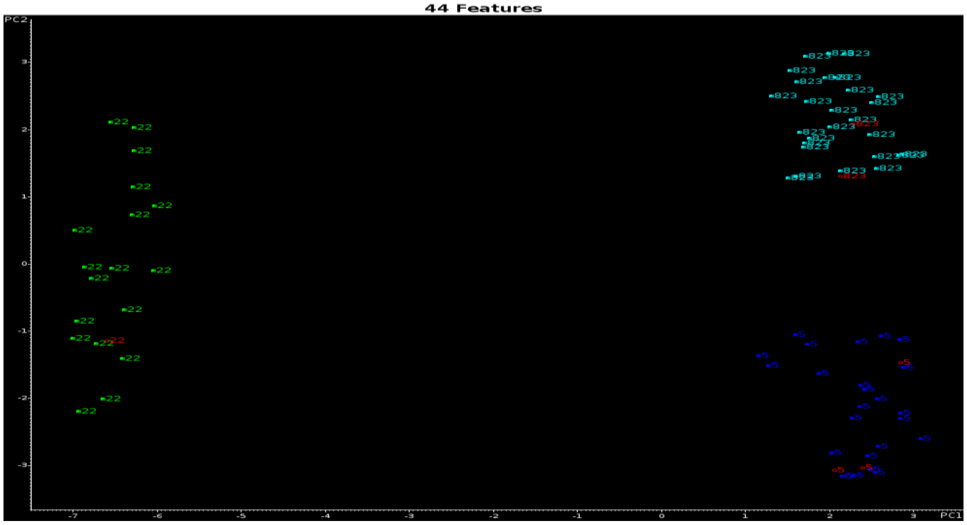


Figure 4.56. Projection of the 6 validation set samples onto the PC plot of the 70 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA (5=Fairfax, 22=Oshawa, 823= Fort Wayne merged with Pontiac)

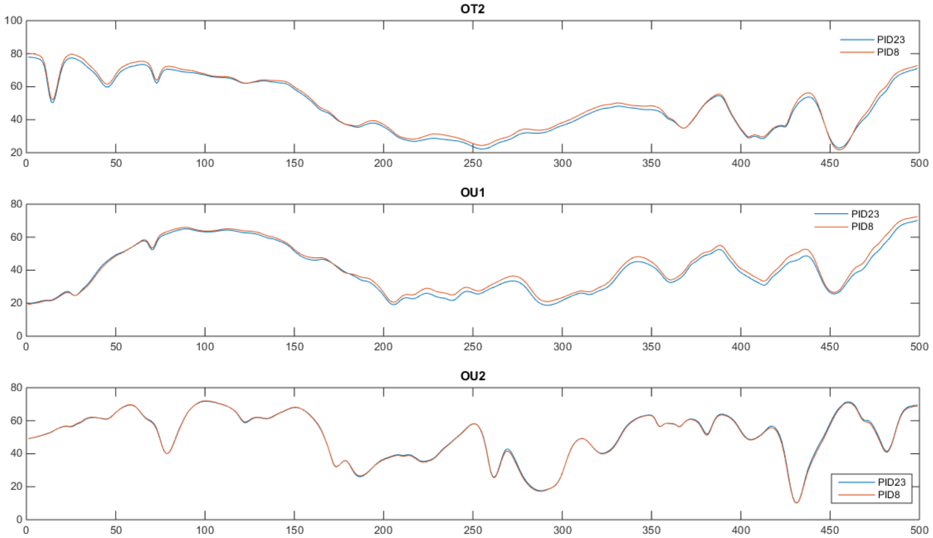


Figure 4.57. The average IR spectra comparison of assembly plant Fort Wayne and Pontiac

After 31 generations, the pattern recognition GA identified 14 wavelet coefficients whose PC plot (see Figure 4.58) showed clustering of the spectra on the basis of GM sub group 4. To assess the predictive ability of these 14 wavelet coefficients, a validation set of 8 paint samples were projected into 2-PC developed from the 73 training set and the wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA. The validation set samples were assigned to correct assembly plants (see Figure 4.59).

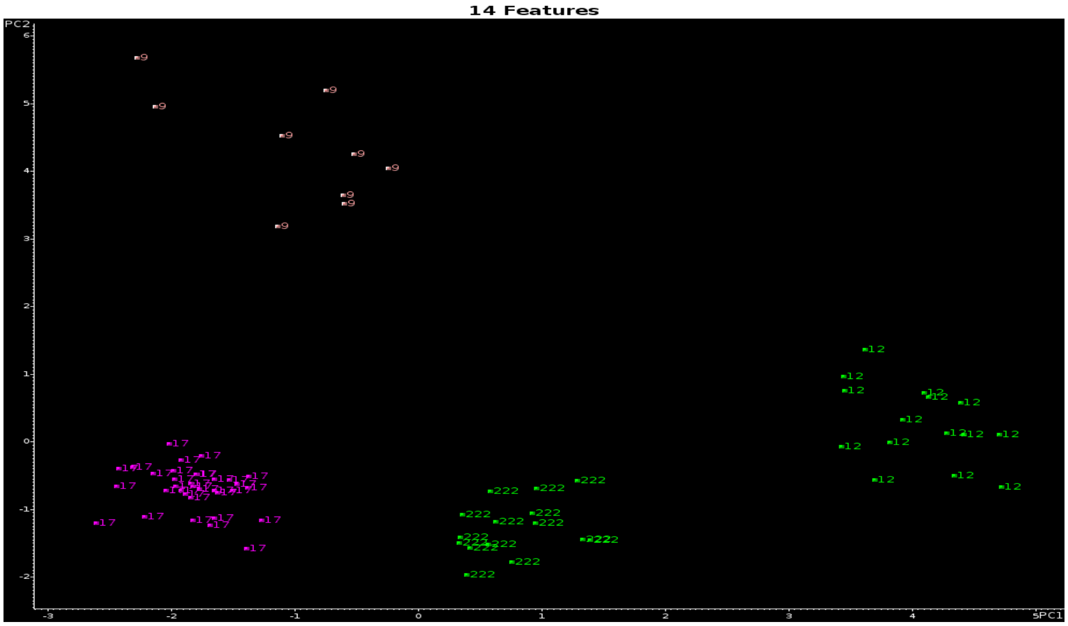


Figure 4.58. 2-PC plot of the 73 training set samples and the 14 wavelet coefficients identified by the pattern recognition GA (9=Fremont, 12=Janesville, 17=Lordstown, 222= Oshawa)

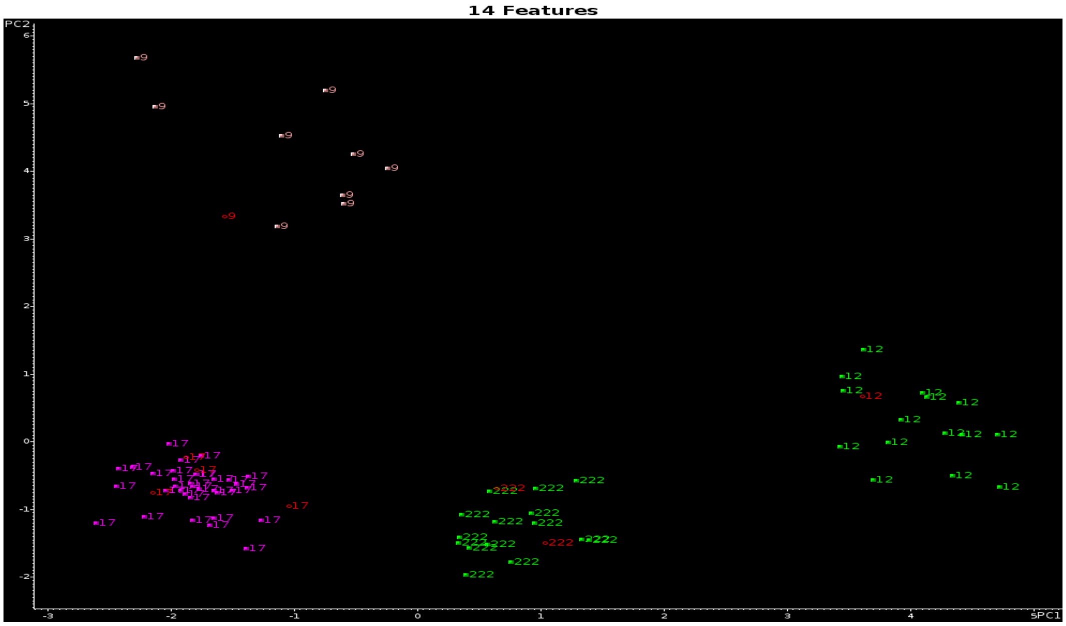


Figure 4.59. Projection of the 8 validation set samples onto the PC plot of the 73 training set samples and the 14 wavelet coefficients identified by the pattern recognition GA (9=Fremont, 12=Janesville, 17=Lordstown, 222= Oshawa)

After 200 generations, the pattern recognition GA identified 19 wavelet coefficients whose PC plot (see Figure 4.60) showed clustering of the spectra on the basis of GM sub group 5. To assess the predictive ability of these 19 wavelet coefficients, a validation set of 5 paint samples were projected into 2-PC developed from the 40 training set and the wavelet coefficients identified by GA using Mehual0.1fitness function of the pattern recognition GA. The validation set samples were assigned to correct assembly plants (see Figure 4.61).

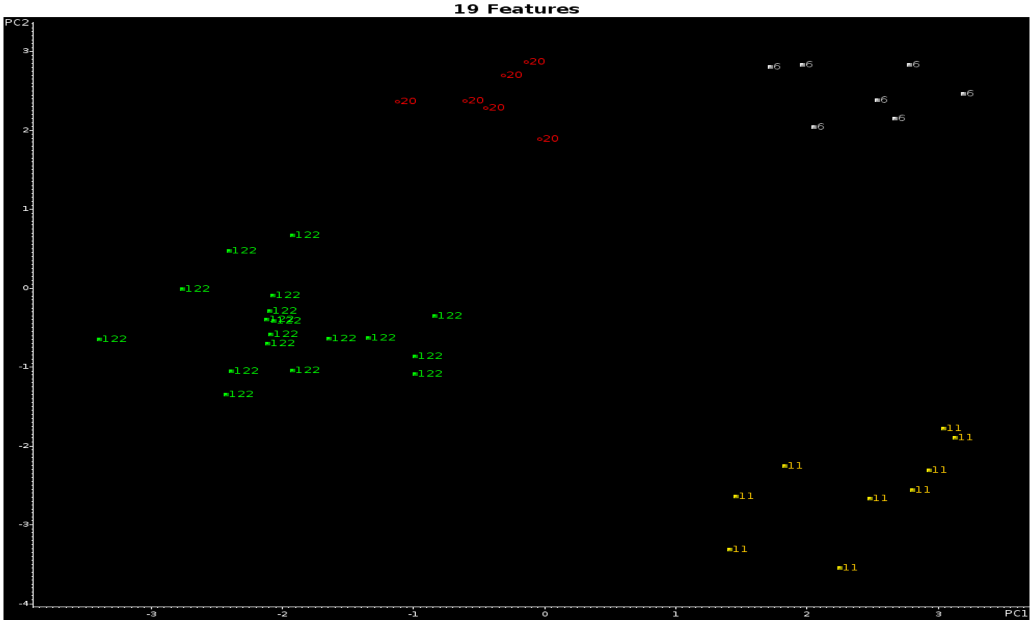


Figure 4.60. 2-PC plot of the 40 training set samples and the 19 wavelet coefficients identified by the pattern recognition GA (6=Flint, 11=Ingersoll, 20=Oklahoma City, 122=Oshawa)

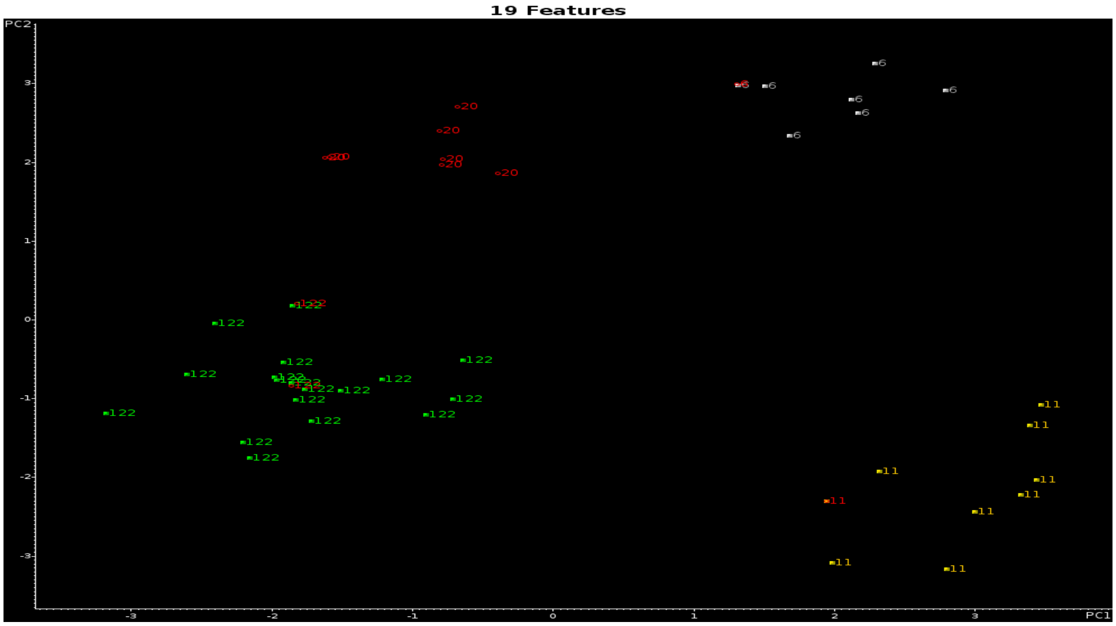


Figure 4.61. Projection of the 5 validation set samples onto the PC plot of the 40 training set samples and the 19 wavelet coefficients identified by the pattern recognition GA (6=Flint, 11=Ingersoll, 20=Oklahoma City, 122=Oshawa)

After 71 generations, the pattern recognition GA identified 20 wavelet coefficients whose PC plot (see Figure 4.62) showed clustering of the spectra on the basis of GM sub group 6. To assess the predictive ability of these 20 wavelet coefficients, a validation set of 3 paint samples were projected into 2-PC developed from the 29 training set and the wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA. The validation set samples were assigned to correct assembly plants (see Figure 4.63).

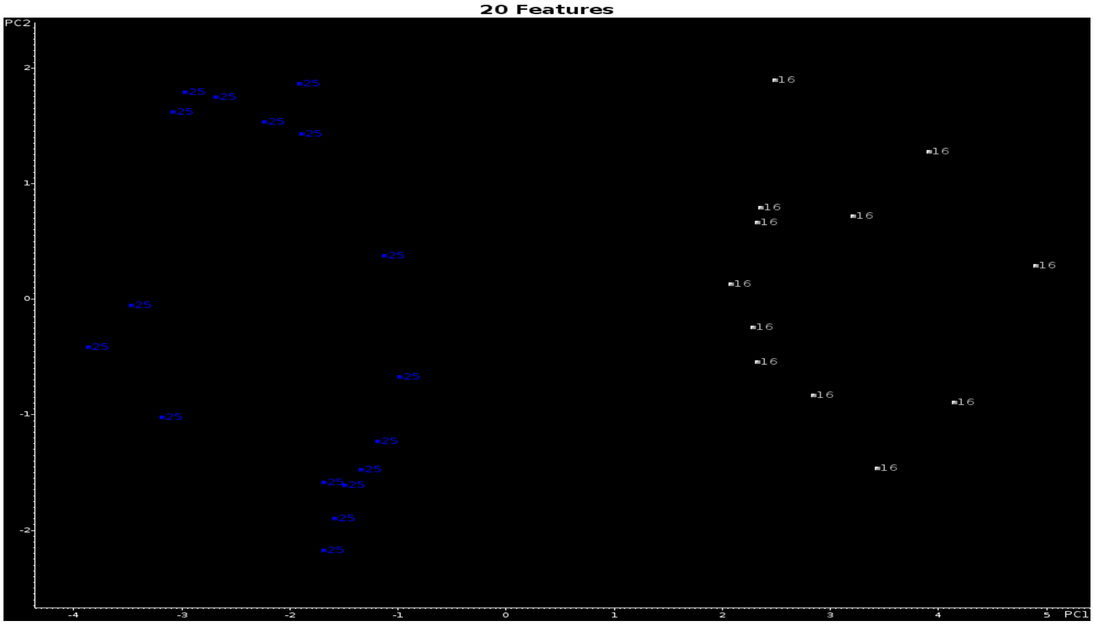


Figure 4.62. 2-PC plot of the 29 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (16=Linden, 25=shreveport)

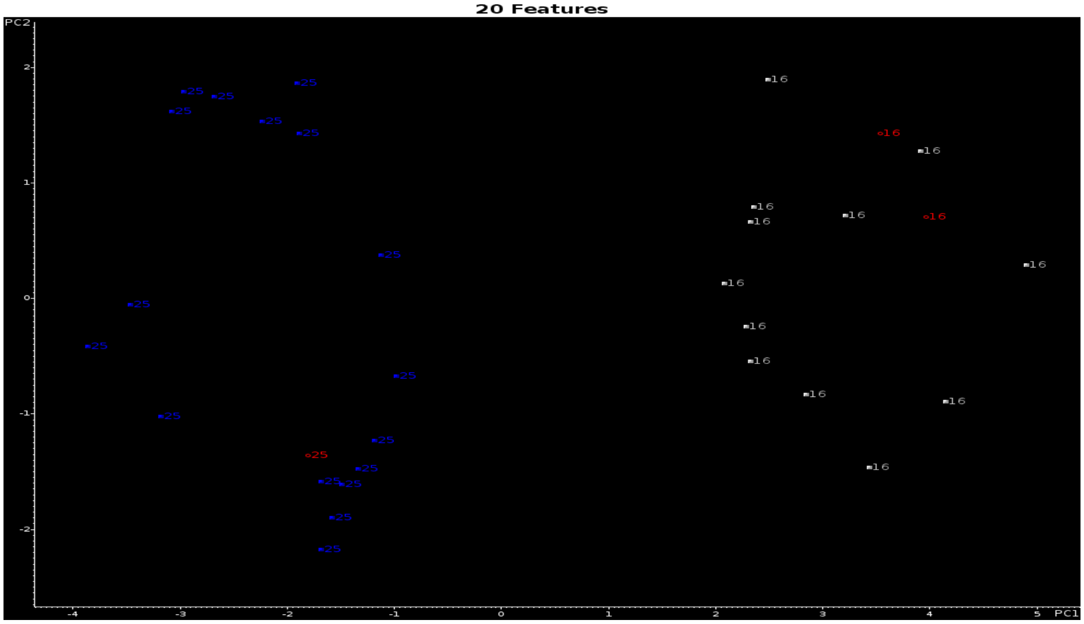


Figure 4.63. Projection of the 3 validation set samples onto the PC plot of the 29 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (16=Linden, 25=shreveport)

For the samples located in the plant group 2, the training and validation sets for manufacturer differentiation in plant group 2 were summarized in Table 4.13. After 30 generations, pattern recognition GA (Fitness function: normal) identified 10 wavelet coefficients whose PC plot (see Figure 4.64) showed clustering of the fused IR spectra on the basis of manufacturers in the plant group 2. The 18 validation set samples were then projected onto the PC plot (see Figure 4.65) define by the 164 training set samples and the 10 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 manufacturer.

Table 4.13. Composition of the IR spectral data set in plant group 2

Manufacturer	Manufacturer IDs	Plant IDs	Training set samples	Validation set samples
Nissan	4	4004, 4105	14	2
Toyota	5	5004,5005,5007,5102,5103	150	16

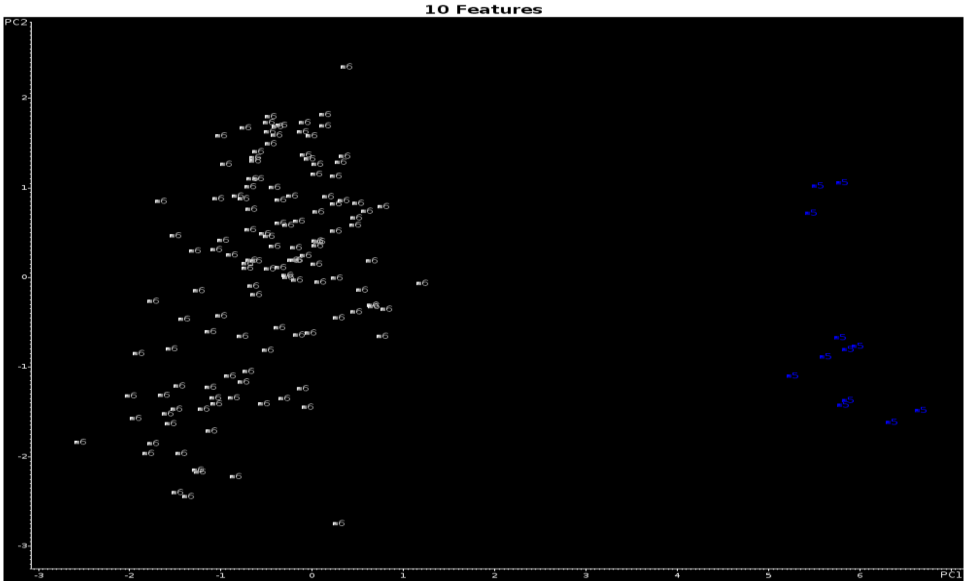


Figure 4.64. 2-PC plot of the 164 training set samples and the 10 wavelet coefficients identified by the pattern recognition GA (5=Nissan, 6=Toyota)

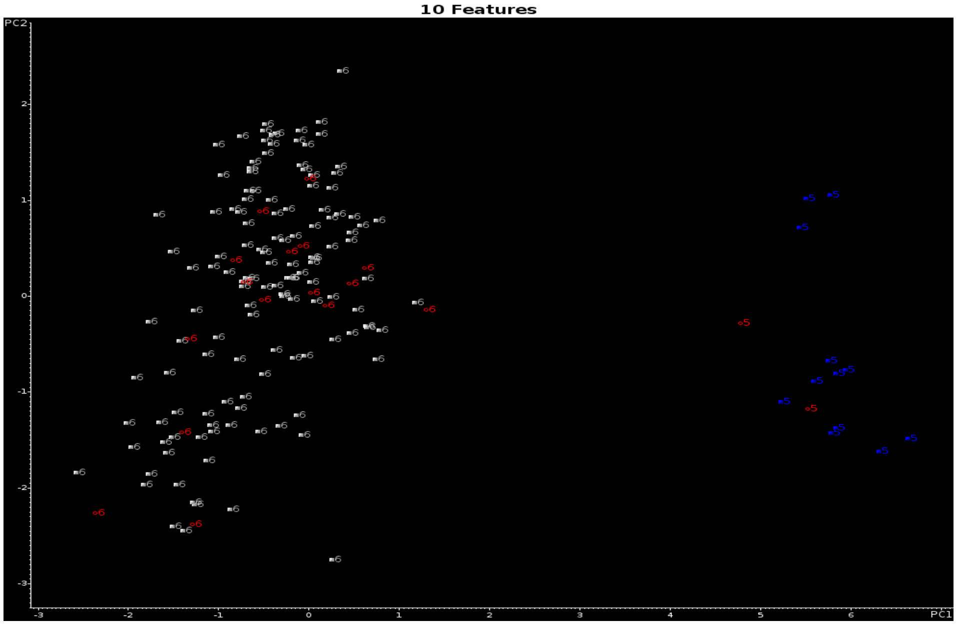


Figure 4.65. Projection of the 18 validation set samples onto the PC plot of the 164 training set samples and the 10 wavelet coefficients identified by the pattern recognition GA (5=Nissan, 6=Toyota)

Until the second prefilter found the manufacturer information in the basis of plant group 2, the third prefilter was developed to differentiate the assembly plant or sub plant information in the basis of manufacturer by using a genetic algorithm (GA) for features selection and pattern recognition. The pattern recognition GA identified 3 wavelet coefficients whose PC plot (see Figure 4.66) showed clustering of IR the spectra on the basis of assembly plants from Nissan after 2 generation run. Figure 4.68 showed the clustering of the IR spectra on the basis of assembly plants from Toyota after 200 generation run. To assess the predictive ability of these 3 wavelet coefficients, a validation set of 2 paint samples located in the Nissan region of the second prefilter were projected into 2-PC developed from the 12 training set and the wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA. The same method was applied to the validation paint samples located in Toyota region of the second prefilter. The validation set samples were assigned to correct assembly plants (see Figure 4.67). The assembly plant Fremont and Georgetown may use similar paint in all three layers, pattern recognition GA cannot discriminate the assembly plant of an unknown sample if it is projected in this region (see Figure 4.69). The assembly plant Fremont (PID5103) only had five training samples, the further explore for this assembly plant told us there were two sample IR spectra of OU1(see Figure 4.70) different from other samples. The samples in the training set were too less to predict the assembly plant Fremont, therefore, this assembly plant was removed from data sheet. The new results were show in Figure 4.71-Figure 4.72.

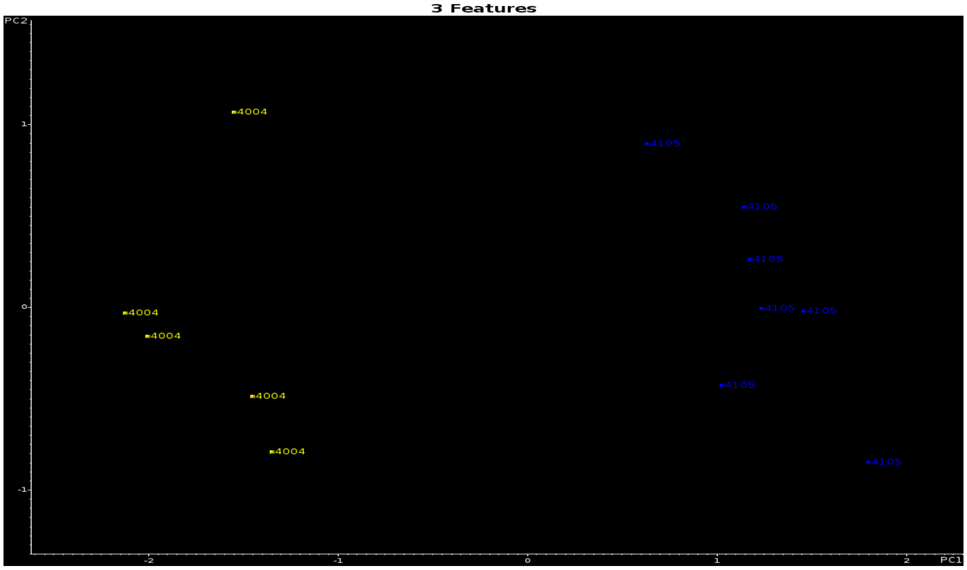


Figure 4.66. 2-PC plot of the 12 training set samples and the 3 wavelet coefficients identified by the pattern recognition GA (4004=Kyushu, 4105=Oppama)

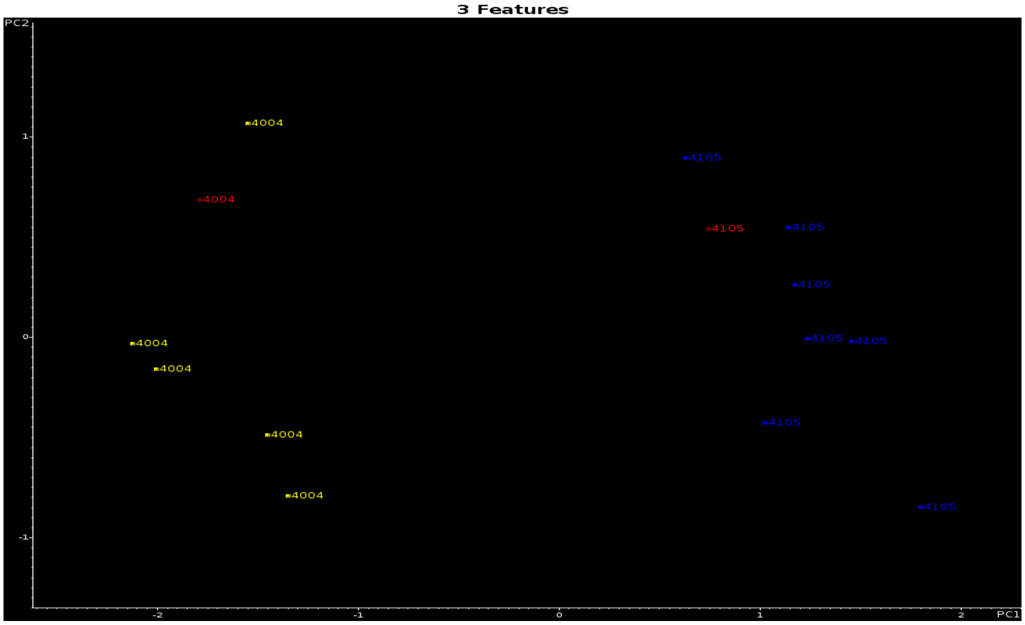


Figure 4.67. Projection of the 2 validation set samples onto the PC plot of the 12 training set samples and the 3 wavelet coefficients identified by the pattern recognition GA (4004=Kyushu, 4105=Oppama)

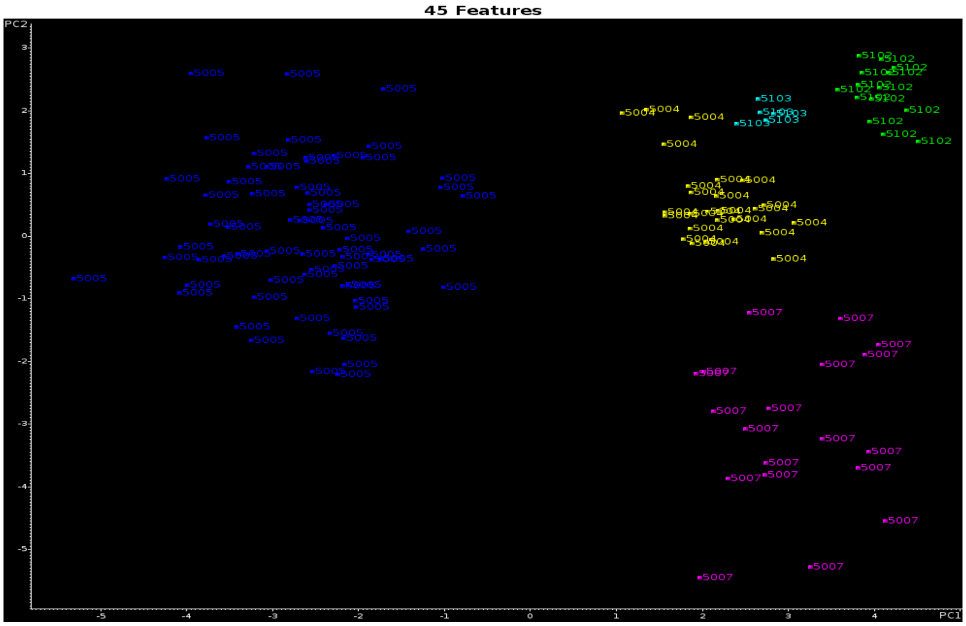


Figure 4.68. 2-PC plot of the 129 training set samples and the 45 wavelet coefficients identified by the pattern recognition GA (5004=Georgetown, 5005=Japan, 5007=Princeton, IN, 5102=Cambridge,ON,Canada, 5103=Fremont)

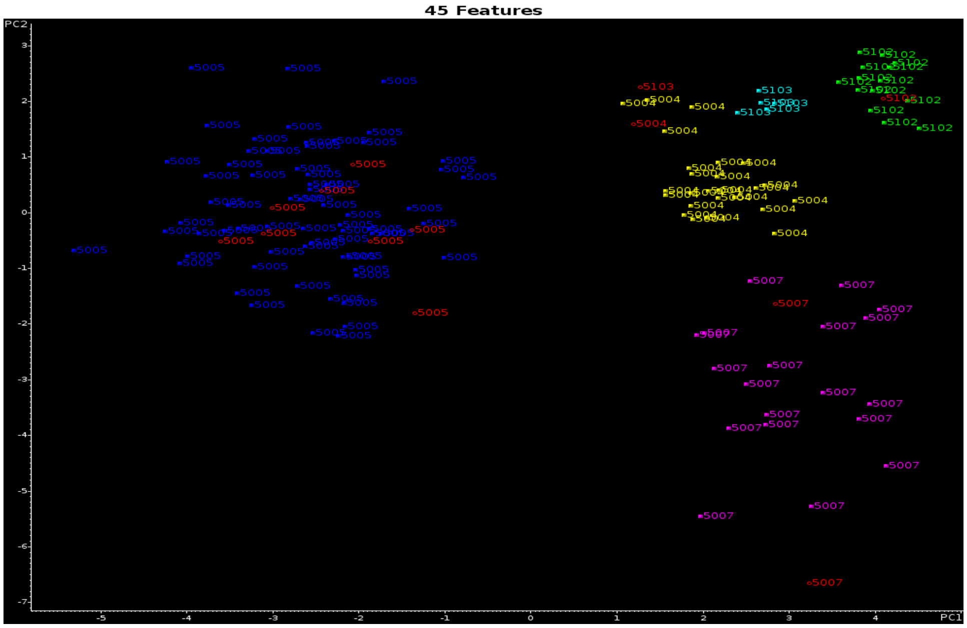


Figure 4.69. Projection of the 13 validation set samples onto the PC plot of the 129 training set samples and the 45 wavelet coefficients identified by the pattern recognition GA (5004=Georgetown, 5005=Japan, 5007=Princeton, IN, 5102=Cambridge,ON,Canada, 5103=Fremont)

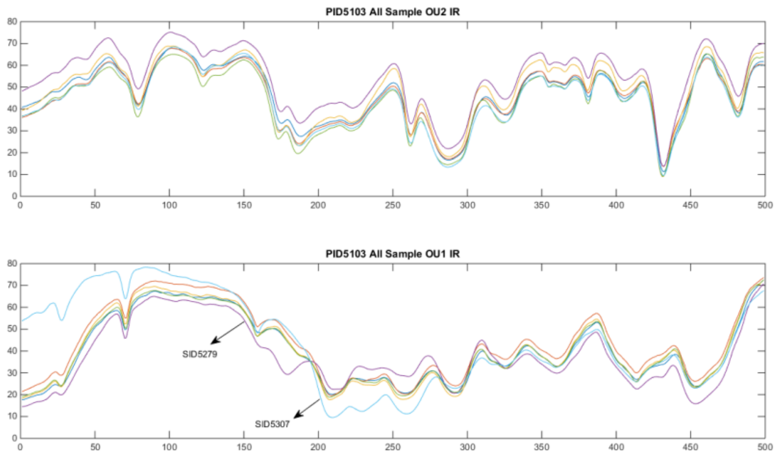


Figure 4.70. All samples in assembly plant Fremont (PID5103) undercoat IR spectra

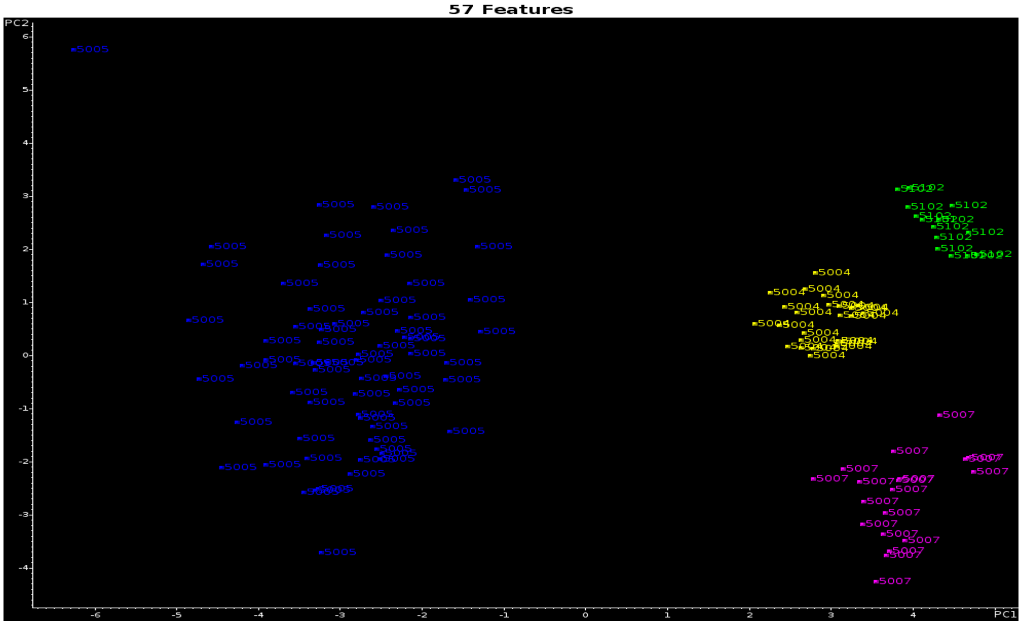


Figure 4.71. 2-PC plot of the 126 training set samples and the 57 wavelet coefficients identified by the pattern recognition GA (5004=Georgetown, 5005=Japan, 5007=Princeton, IN, 5102=Cambridge,ON,Canada)

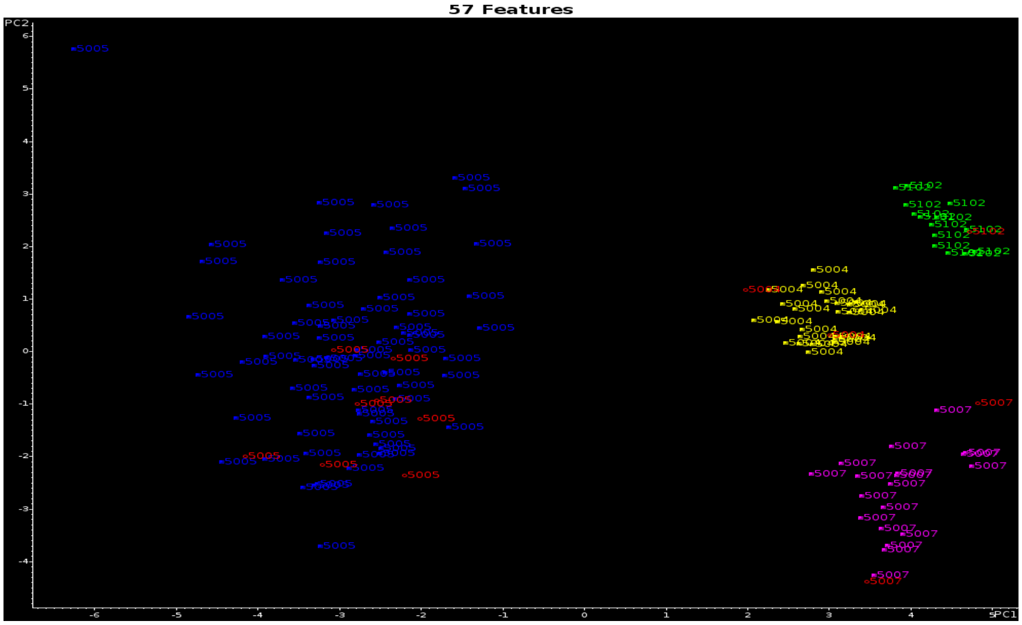


Figure 4.72. Projection of the 13 validation set samples onto the PC plot of the 126 training set samples and the 57 wavelet coefficients identified by the pattern recognition GA (5004=Georgetown, 5005=Japan, 5007=Princeton, IN, 5102=Cambridge,ON,Canada)

For the samples located in the plant group 3, the training and validation sets for manufacturer differentiation in plant group 3 were summarized in Table 4.14. After 200 generations, pattern recognition GA (Fitness function: normal) identified 44 wavelet coefficients whose PC plot (see Figure 4.73) showed clustering of the fused IR spectra on the basis of manufacturers in the plant group 3. The 29 validation set samples were then projected onto the PC plot (see Figure 4.74) define by the 311 training set samples and the 44 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 manufacturer.

Table 4.14. Composition of the IR spectral data set in plant group 3

Manufacturer	Manufacturer IDs	Plant IDs	Training set samples	Validation set samples
Chrysler	2	1000,1001,1003,1007,1008,1009, 1011,1012,1102,1108,1110	244	23
Nissan	5	4001, 4006	49	4
Toyota	6	5002, 5203	19	2

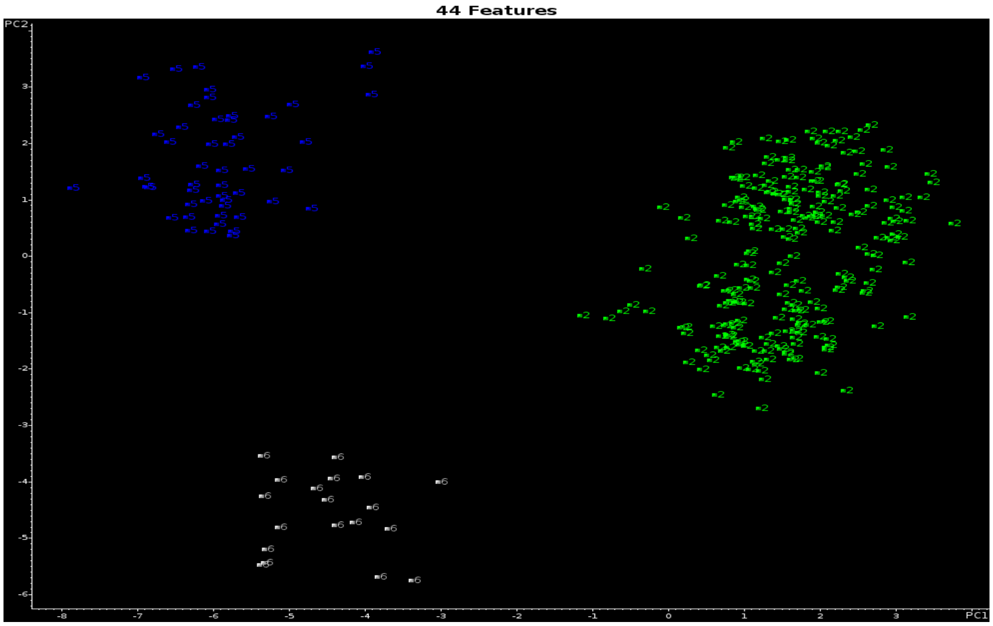


Figure 4.73. 2-PC plot of the 311 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA (2=Chrysler, 5=Nissan, 6=Toyota)

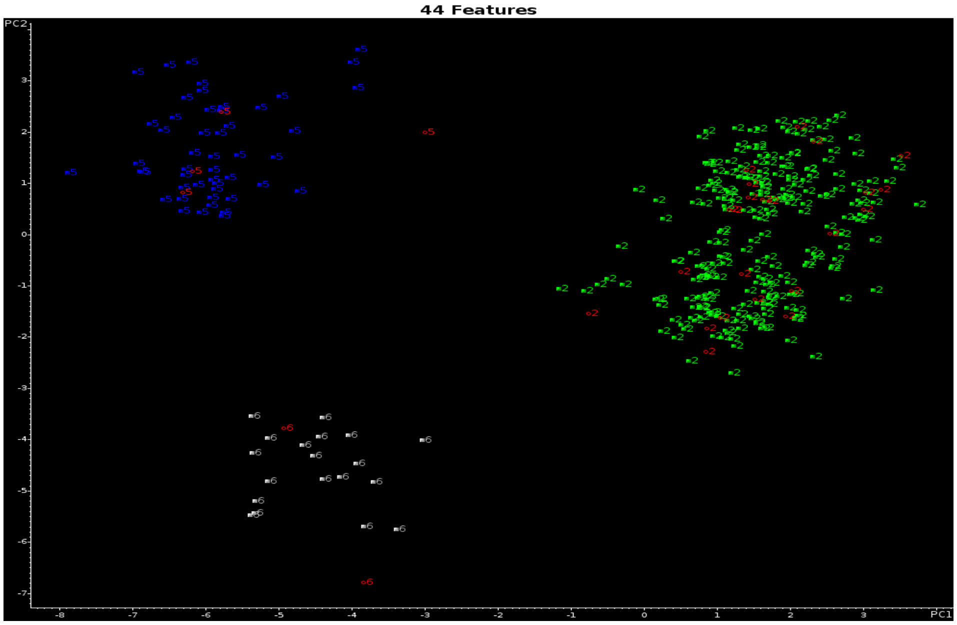


Figure 4.74. Projection of the 29 validation set samples onto the PC plot of the 311 training set samples and the 44 wavelet coefficients identified by the pattern recognition GA (2=Chrysler, 5=Nissan, 6=Toyota)

After ascertain the manufacturer information from the above prefilter, the third prefilter was developed to differentiate the assembly plant or sub plant information in the basis of manufacturer by using a genetic algorithm (GA) for features selection and pattern recognition. The pattern recognition GA identified 10 wavelet coefficients whose PC plot (see Figure 4.75) showed clustering of IR the spectra on the basis of assembly plants from Chrysler after 18 generation run. To assess the predictive ability of these 10 wavelet coefficients, a validation set of 23 paint samples located in the Chrysler region of the second prefilter were projected into 2-PC developed from the 242 training set and the 10 wavelet coefficients identified by GA using Hopkin 0.1 fitness function of the pattern recognition GA (see Figure 4.76). The IR spectra of assembly plant Sterling heights (PID1008), St. Louis (PID1009), Windsor (PID1012),sub Sterling heights (PID1108), Toledo (PID1110) in OT2,OU1 and OU2 paint layers were very similar and merged into new assembly plant with ID16892. The assembly plant Saltillo (PID1007) and Toluca (PID1011) were merged into a new assembly plant with ID1671 in the same way. The IR spectra of OT2, OU1 and OU2 were seen in Figure 4.77-Figure 4.79. The same method was applied to the validation paint samples located in Nissan or Toyota region of the second prefilter. Figure 4.80 and Figure 4.82 showed the clustering of the IR spectra on the basis of assembly plants from Nissan or Toyota. The validation set samples were assigned to correct assembly plants (see Figure 4.81-4.83).

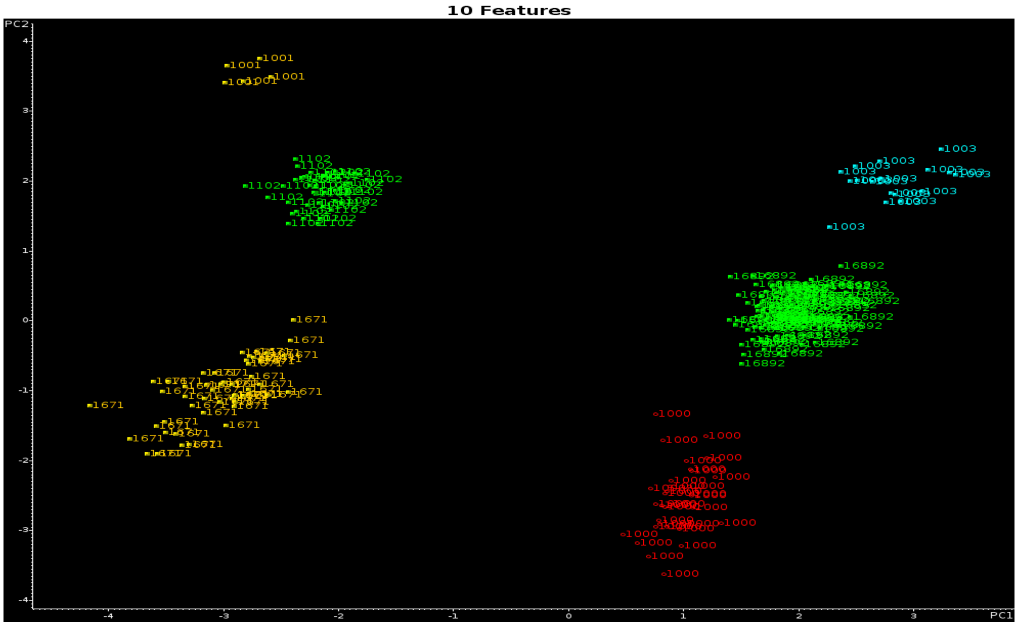


Figure 4.75. 2-PC plot of the 242 training set samples and the 10 wavelet coefficients identified by the pattern recognition GA (1000=Belvidere, 1001= Bloomington, 1003=Dodge Main, 1102= Bramalea, 1671=Saltillo and Toluca, 16892=Sterling heights, St. Louis, Windsor and Toledo)

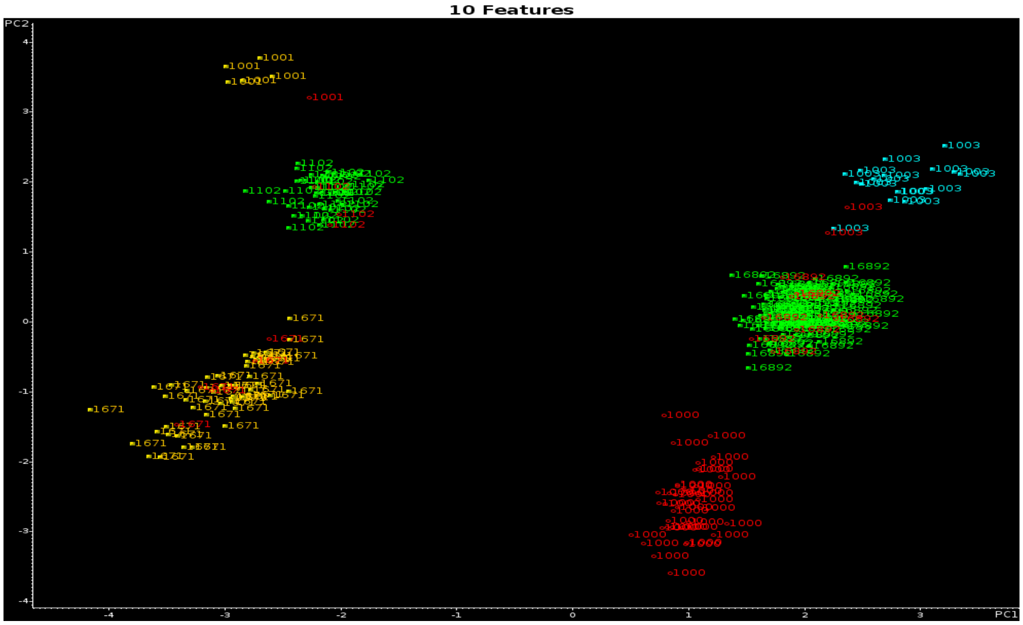


Figure 4.76. Projection of the 23 validation set samples onto the PC plot of the 242 training set samples and the 10 wavelet coefficients identified by the pattern recognition GA (1000=Belvidere, 1001= Bloomington, 1003=Dodge Main, 1102= Bramalea, 1671=Saltillo and Toluca, 16892=Sterling heights, St. Louis, Windsor and Toledo)

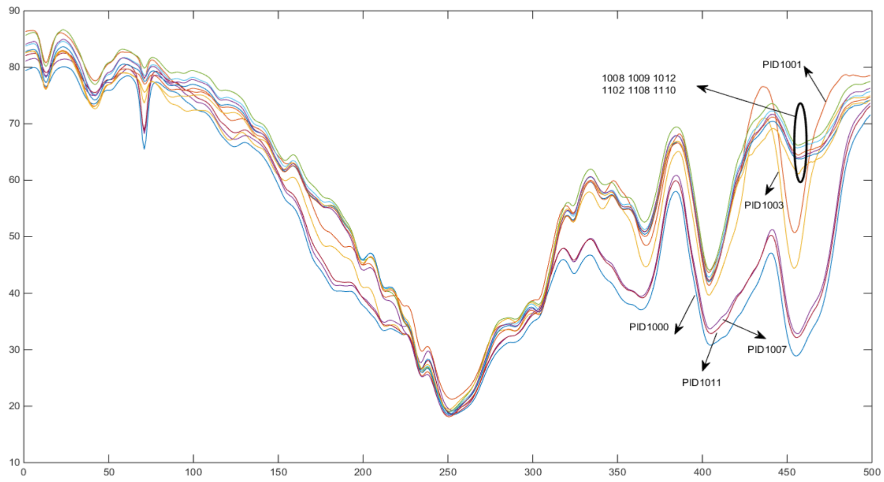


Figure 4.77. Average assembly plant OT2 IR spectra from Chrysler

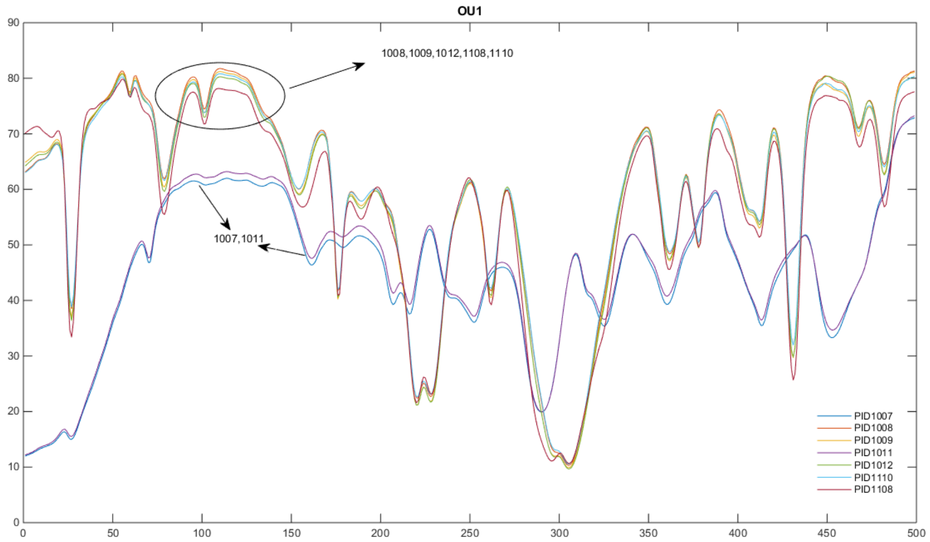


Figure 4.78. Average assembly plant OU1 IR spectra from Chrysler

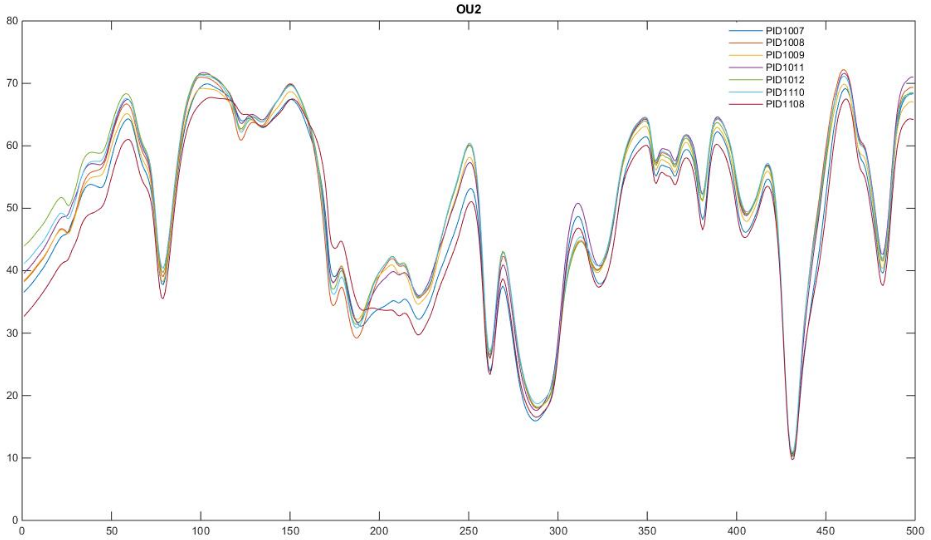


Figure 4.79. Average assembly plant OU2 IR spectra from Chrysler

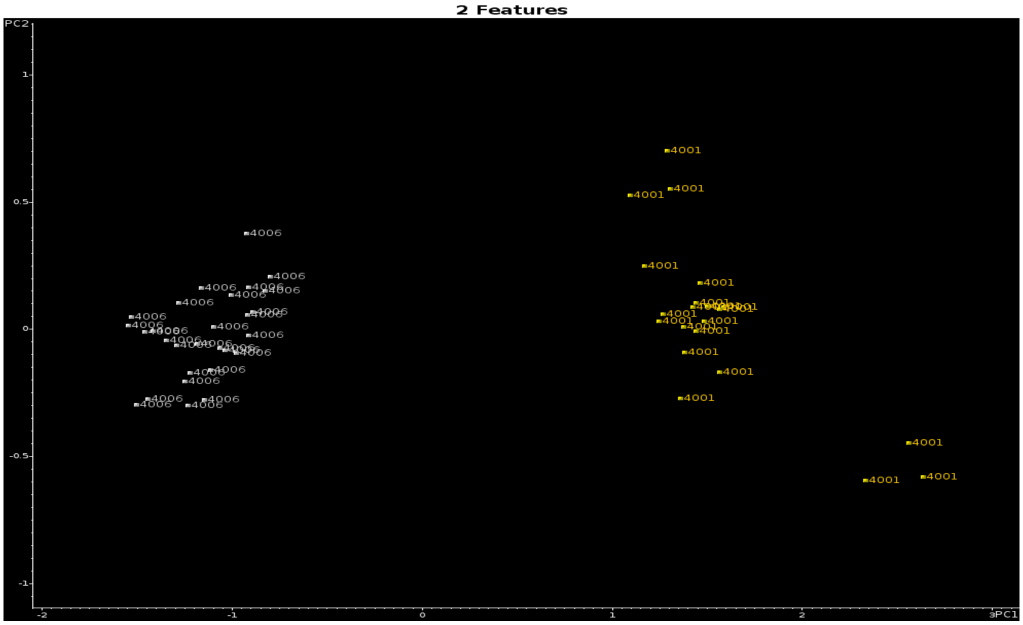


Figure 4.80. 2-PC plot of the 49 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4001=Canton, 4006=Smyrna)

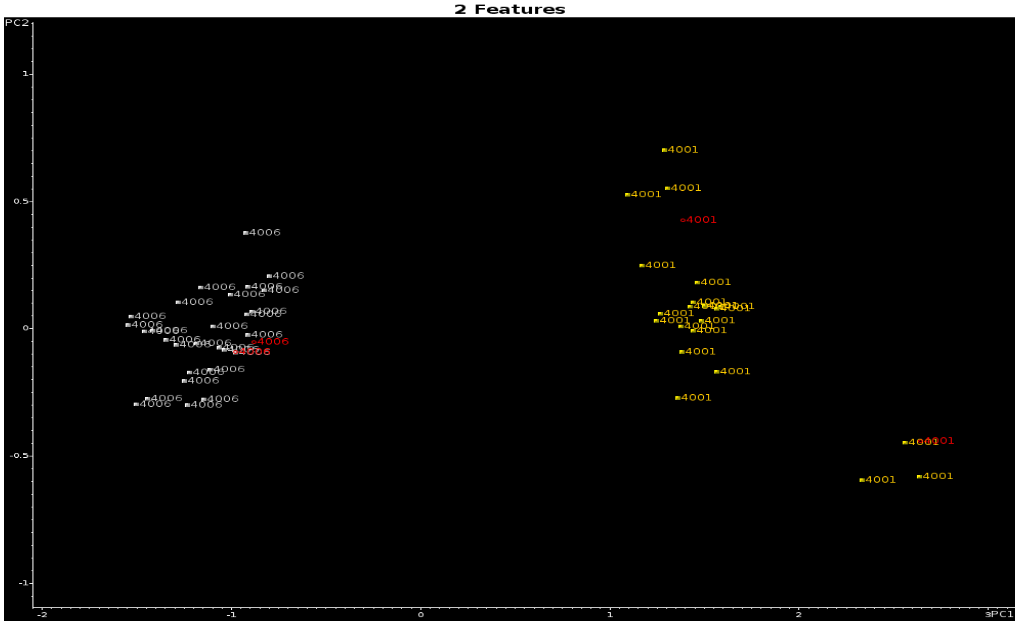


Figure 4.81. Projection of the 4 validation set samples onto the PC plot of the 49 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4001=Canton, 4006=Smyrna)

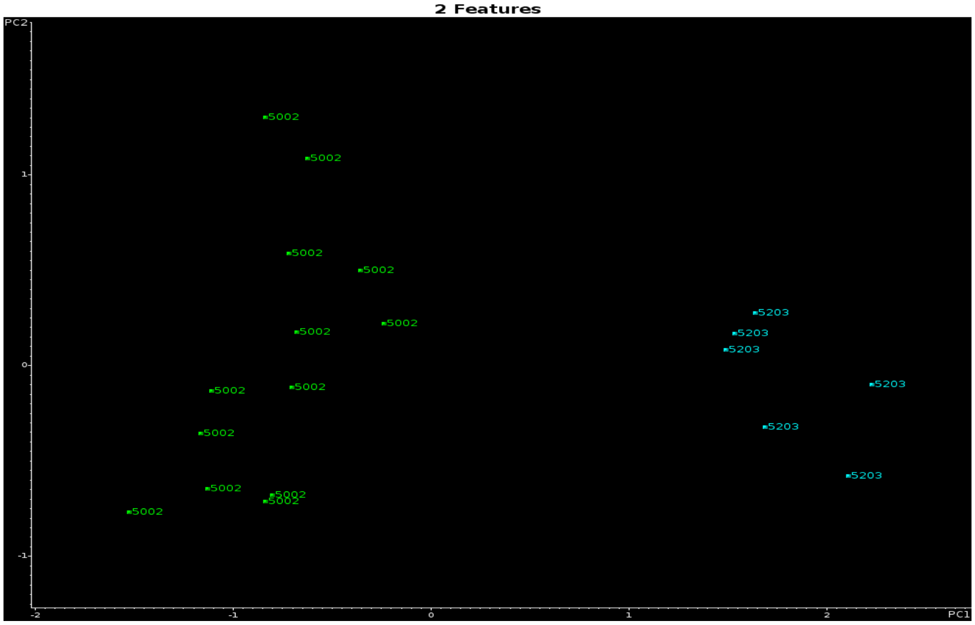


Figure 4.82. 2-PC plot of the 19 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (5002=Cambridge, ON, Canada, 5203= Fremont, CA)

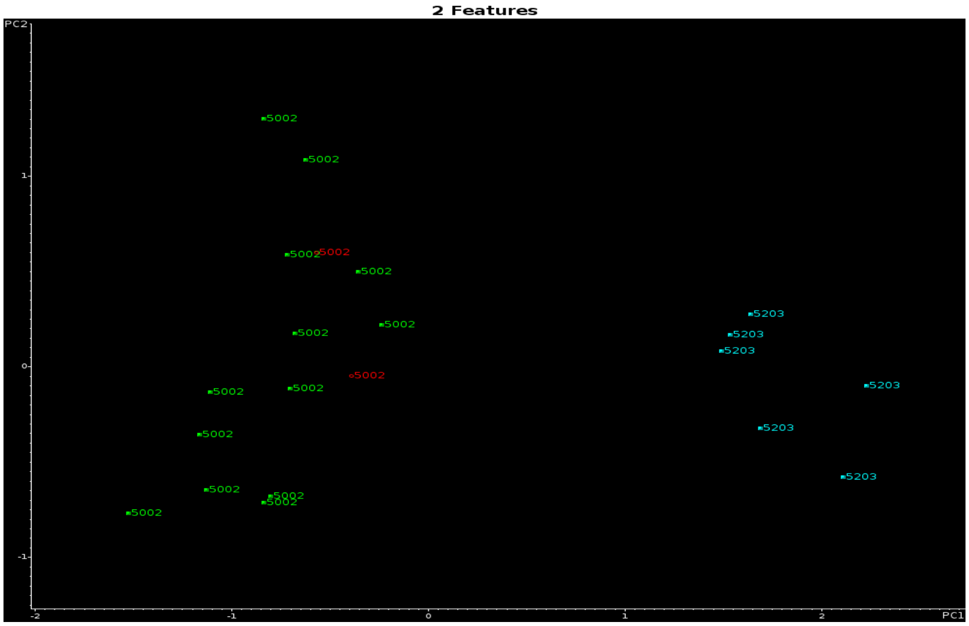


Figure 4.83. Projection of the 2 validation set samples onto the PC plot of the 19 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (5002=Cambridge, ON, Canada, 5203= Fremont, CA)

The samples located in the plant group 4 are the most difficult to discriminate manufacturer information. The training and validation sets for manufacturer differentiation in the plant group 4 were summarized in Table 4.15. After 200 generations, pattern recognition GA (Fitness function: normal) identified 50 wavelet coefficients whose PC plot (see Figure 4.84) showed clustering of the fused IR spectra on the basis of manufacturers and assembly plant in the plant group 4. The 37 validation set samples were then projected onto the PC plot (see Figure 4.85) define by the 398 training set samples and the 50 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 manufacturer.

Table 4.15. Composition of the IR spectral data set in plant group 4

Manufacturer	Manufacturer IDs	Plant IDs	Training set samples	Validation set samples
Chrysler	21	1010	12	1
	22	1103	18	2
	23	1109	27	3
	24	1002	13	1
Ford	31	2000,2002,2003,2008, 2011,2012,2015,2016	172	18
	32	2005,2006,2013,2106	40	4
	33	2007,2107,2014,2110	68	5
	34	2010	13	1
Honda	4	3100, 3106	9	1
Nissan	5	4100,4106	26	1

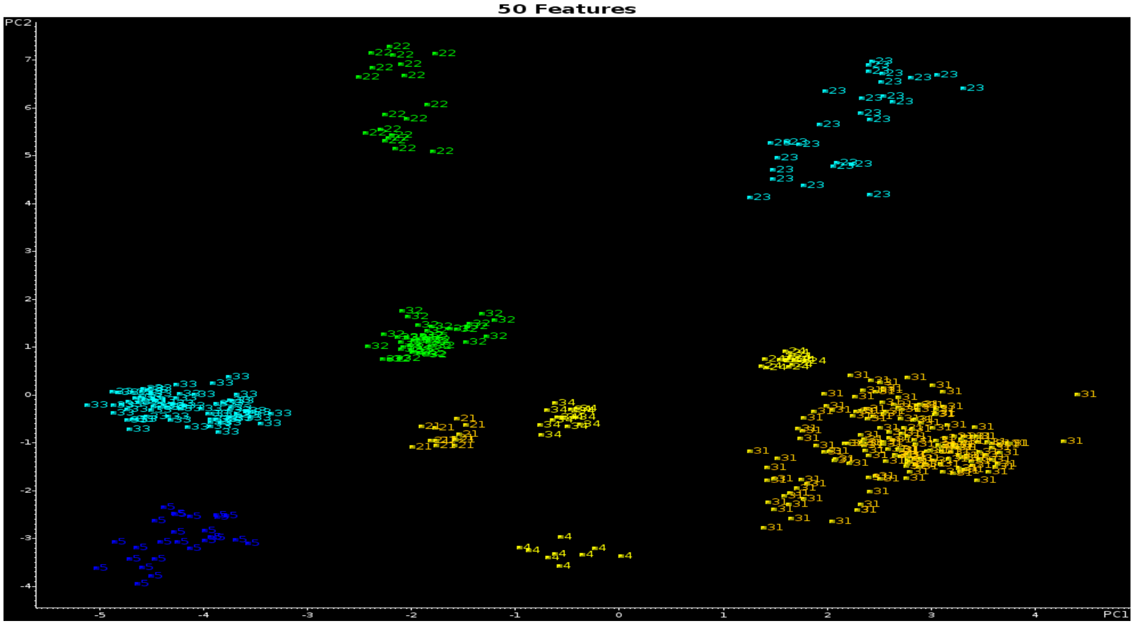


Figure 4.84. 2-PC plot of the 398 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA (21=Chrysler, 22=Chrysler, 23=Chrysler, 24=Chrysler, 31=Ford, 32=Ford, 33=Ford, 34=Ford, 4= Honda, 5=Nissan)

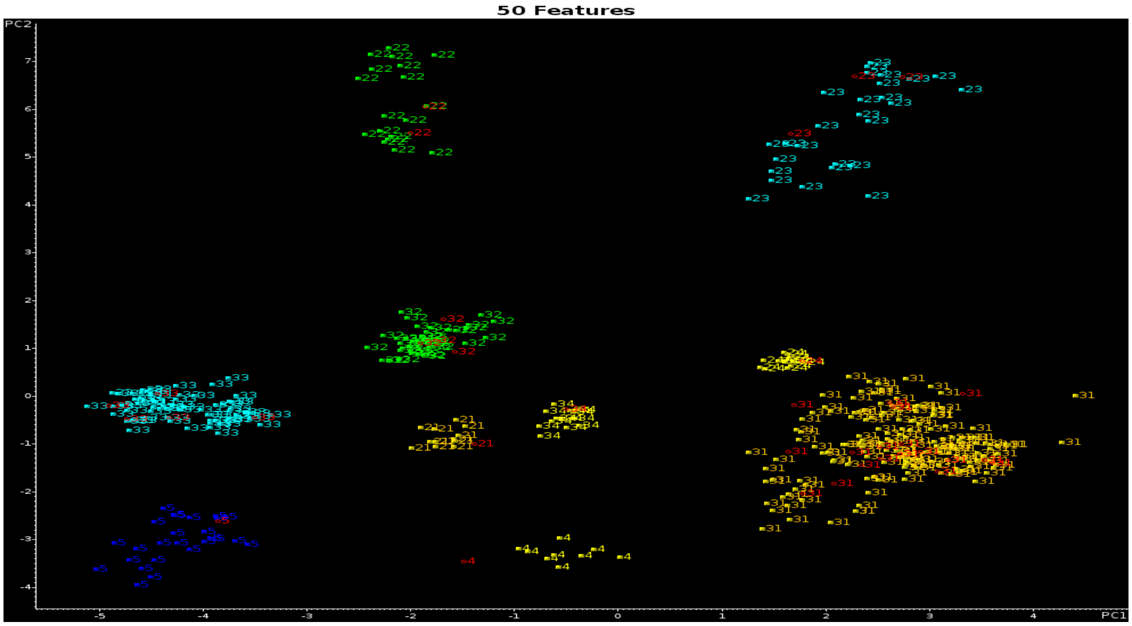


Figure 4.85. Projection of the 37 validation set samples onto the PC plot of the 398 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA(21=Chrysler, 22=Chrysler, 23=Chrysler, 24=Chrysler, 31=Ford, 32=Ford, 33=Ford, 34=Ford, 4= Honda, 5=Nissan)

Since the samples locate in the cluster21, cluster22, cluster23, cluster24 and cluster34 are from the identified assembly plants: Toledo, Dodge Main, St. Louis, Bramalea/Brampton and Louisville respectively; those samples are not necessary to do further investigation. Samples located in the manufacturer cluster 31, 32, 33, 4 and 5 need the third level prefilter to identify their assembly plants. The pattern recognition GA identified 50 wavelet coefficients whose PC plot (see Figure 4.86) showed clustering of IR the spectra on the basis of assembly plants from Ford after 200 generation run. To assess the predictive ability of these 50 wavelet coefficients, a validation set of 18 paint samples located in the Ford region 31 of the second prefilter were projected into 2-PC developed from the 167 training set and the 50 wavelet coefficients identified by GA using Hopkin 0.1 fitness function of the pattern recognition GA (see Figure 4.87). Samples from the assembly plants-Atlanta, Chicago, Norfolk, Oakville, Twin Cities-Saint Paul and Wayne merged together due to their similar IR spectra (see Figure 4.88-4.90).

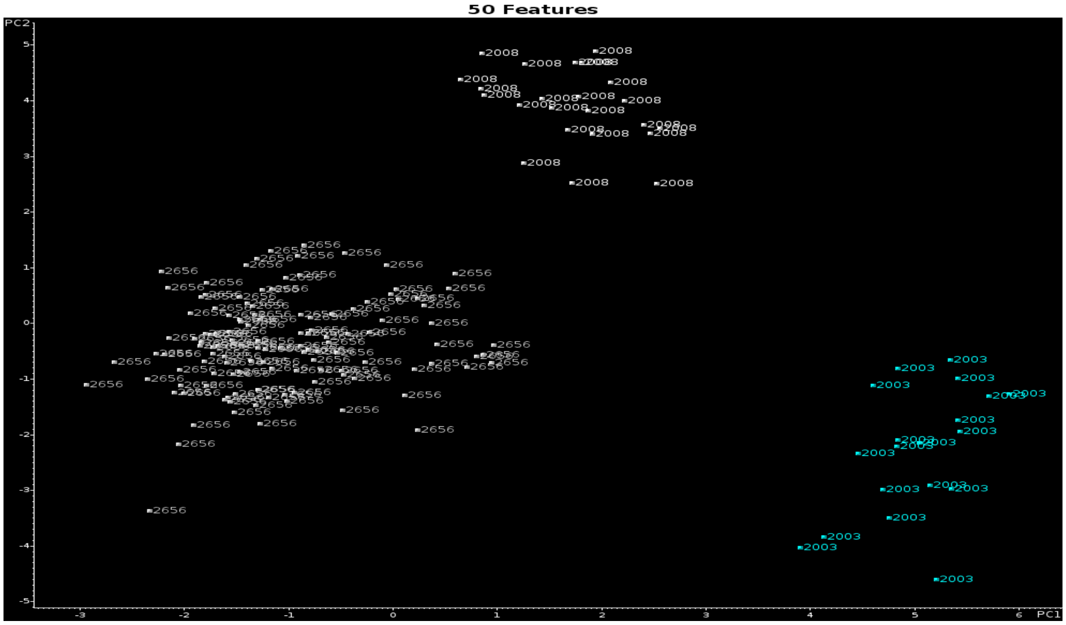


Figure 4.86. 2-PC plot of the 167 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA (2003=Dearborn, 2008=Kentucky Truck, 2656= Atlanta, Chicago, Norfolk, Oakville, Twin Cities-Saint Paul and Wayne)

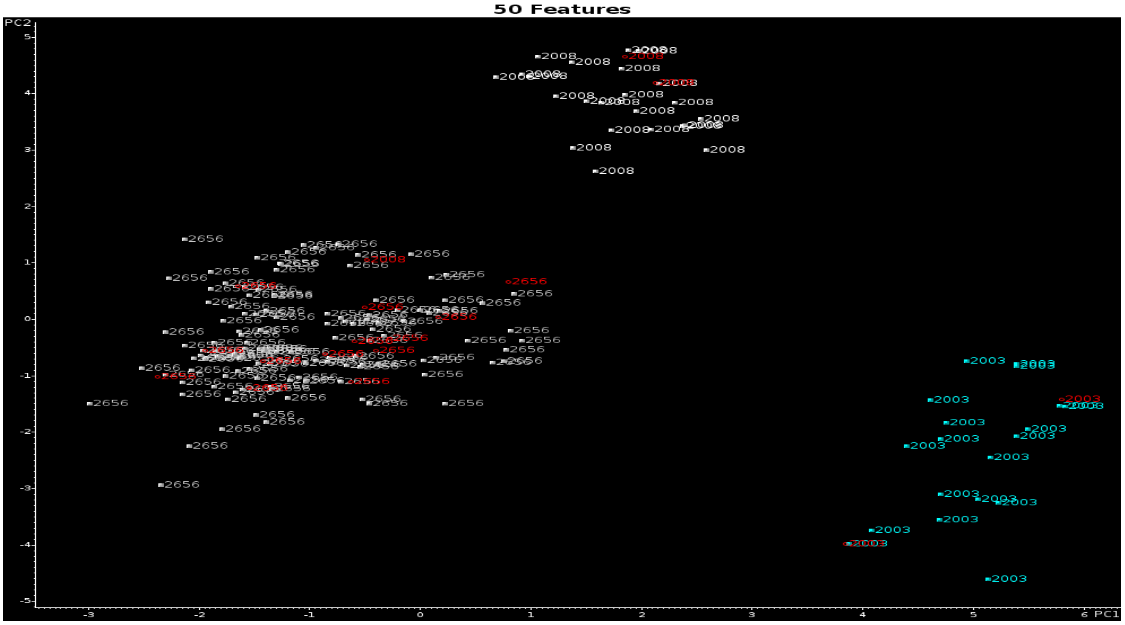


Figure 4.87. Projection of the 18 validation set samples onto the PC plot of the 167 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA (2003=Dearborn, 2008=Kentucky Truck, 2656= Atlanta, Chicago, Norfolk, Oakville, Twin Cities-Saint Paul and Wayne)

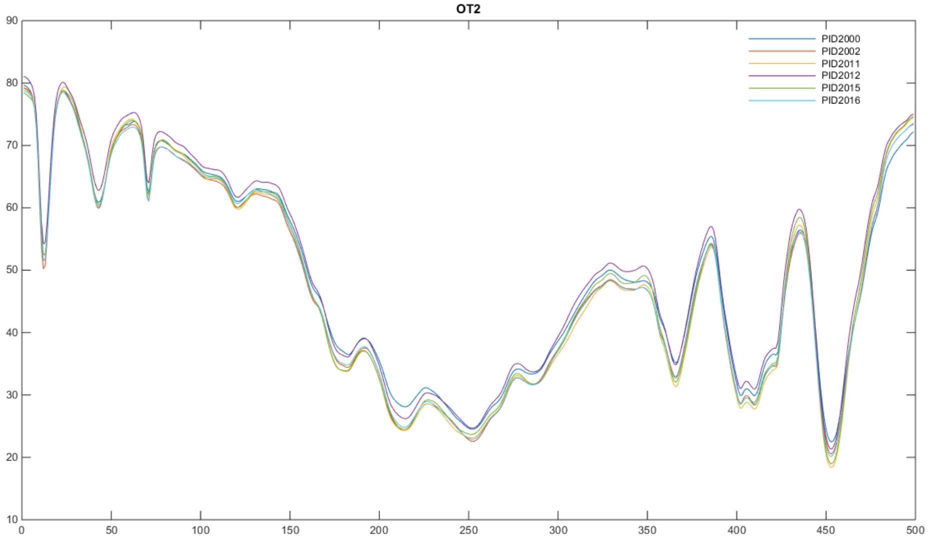


Figure 4.88. The average OT2 IR spectra from the assembly plants: Atlanta, Chicago, Norfolk, Oakville, Twin Cities-Saint Paul and Wayne

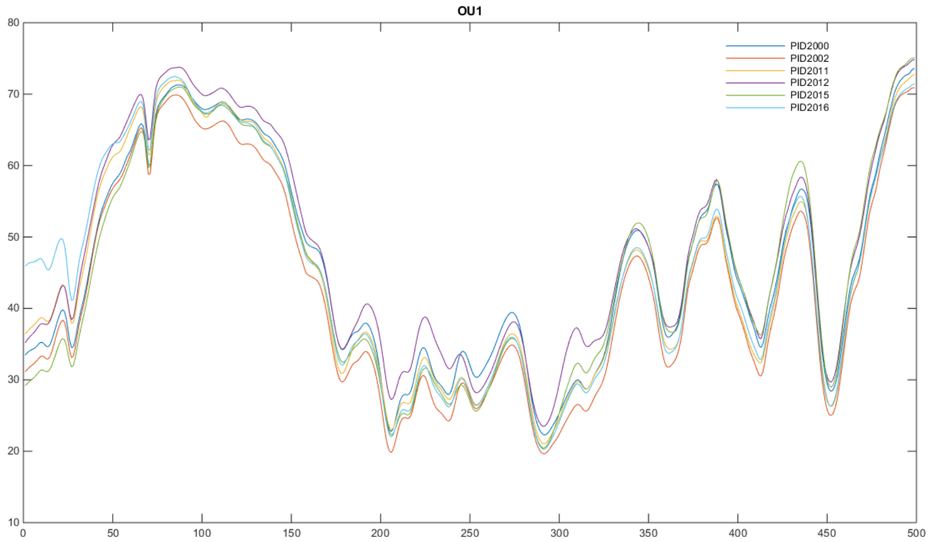


Figure 4.89. The average OU1 IR spectra from the assembly plants: Atlanta, Chicago, Norfolk, Oakville, Twin Cities-Saint Paul and Wayne

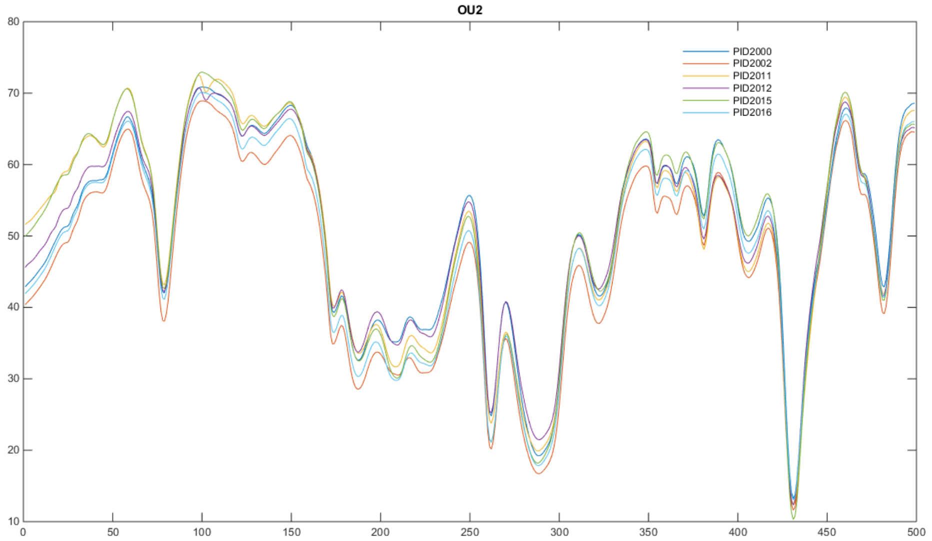


Figure 4.90. The average OU2 IR spectra from the assembly plants: Atlanta, Chicago, Norfolk, Oakville, Twin Cities-Saint Paul and Wayne

The pattern recognition GA identified 29 wavelet coefficients whose PC plot (see Figure 4.91) showed clustering of IR the spectra on the basis of assembly plants from Ford after 200 generation run. To assess the predictive ability of these 29 wavelet coefficients, a validation set of 3 paint samples located in the Ford region 32 of the second prefilter were projected into 2-PC developed from the 38 training set and the 29 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.92).

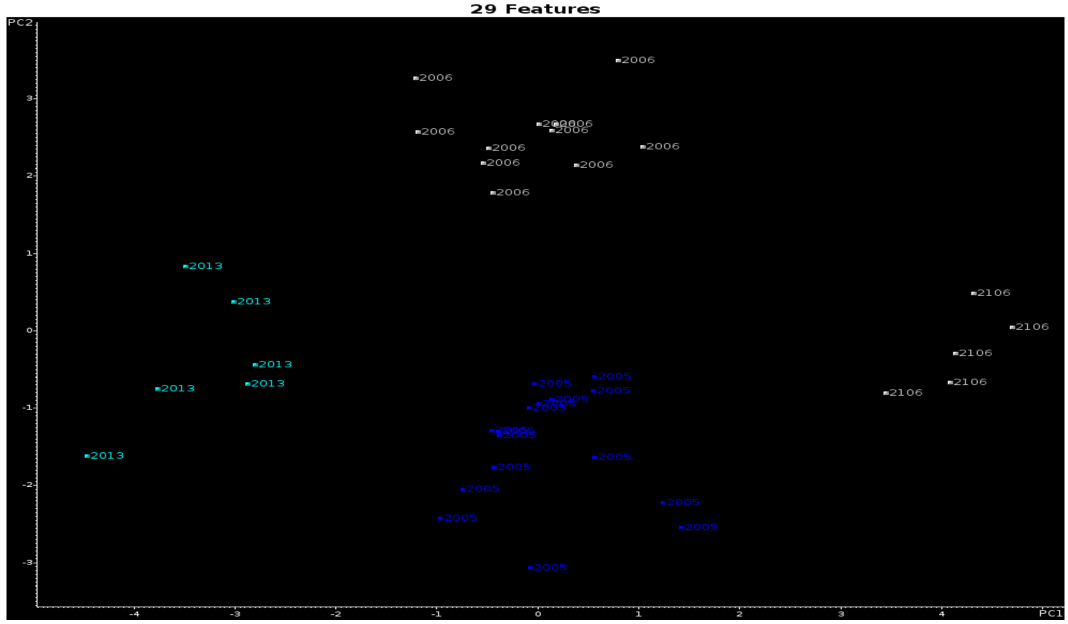


Figure 4.91. 2-PC plot of the 38 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (2005=Flat Rock, 2006=Hermosillo, 2013= Saint Louis, 2106 =Hermosillo)

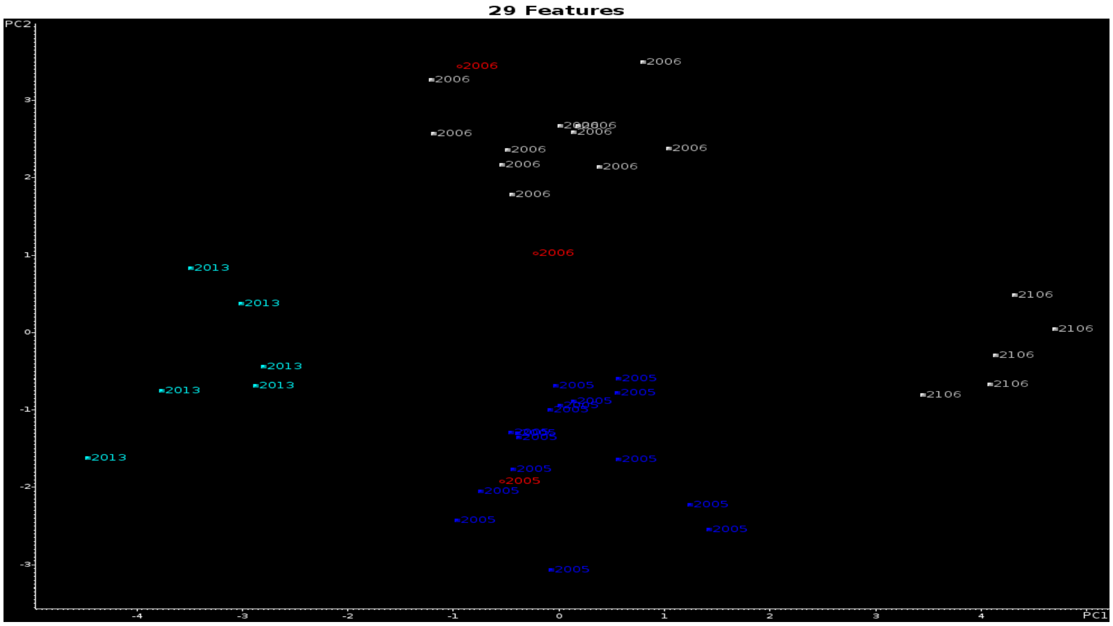


Figure 4.92 Projection of the 3 validation set samples onto the PC plot of the 38 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (2005=Flat Rock, 2006=Hermosillo, 2013= Saint Louis, 2106 =Hermosillo)

The pattern recognition GA identified 30 wavelet coefficients whose PC plot (see Figure 4.93) showed clustering of IR the spectra on the basis of assembly plants from Ford after 200 generation run. To assess the predictive ability of these 30 wavelet coefficients, a validation set of 4 paint samples located in the Ford region 33 of the second prefilter were projected into 2-PC developed from the 68 training set and the 30 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.94). Samples from the assembly plants- Kansas City and Louisville merged together due to their similar IR spectra (see Figure 4.95). However, a validation sample from assembly plant Saint Thomas-Talbotsville was misclassified.

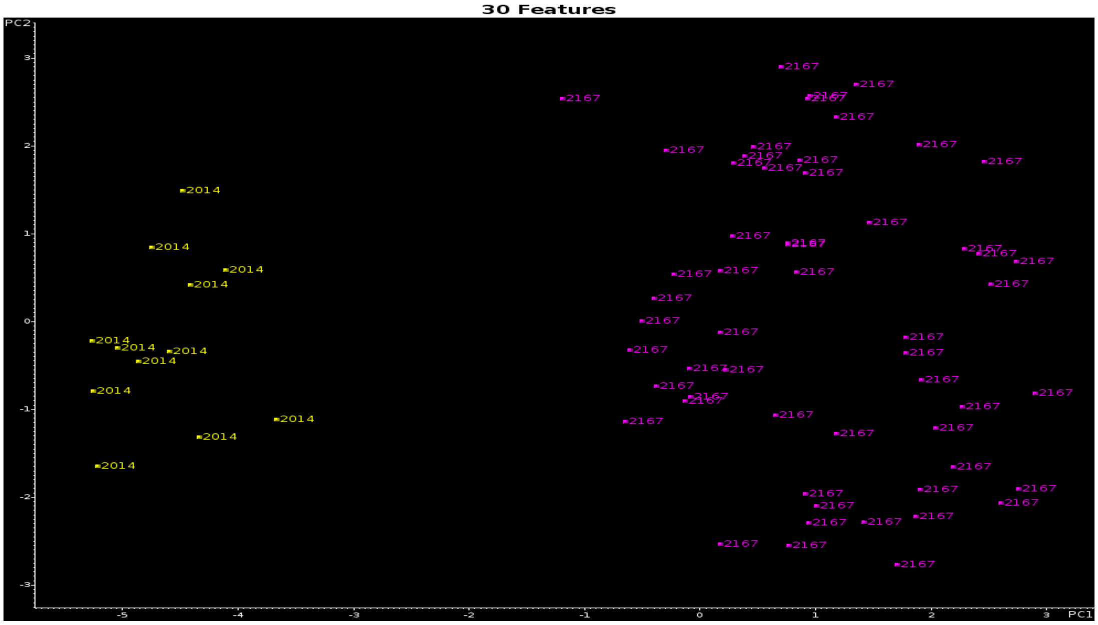


Figure 4.93. 2-PC plot of the 68 training set samples and the 30 wavelet coefficients identified by the pattern recognition GA (2014= Saint Thomas-Talbotsville, 2167= Kansas City, Louisville)

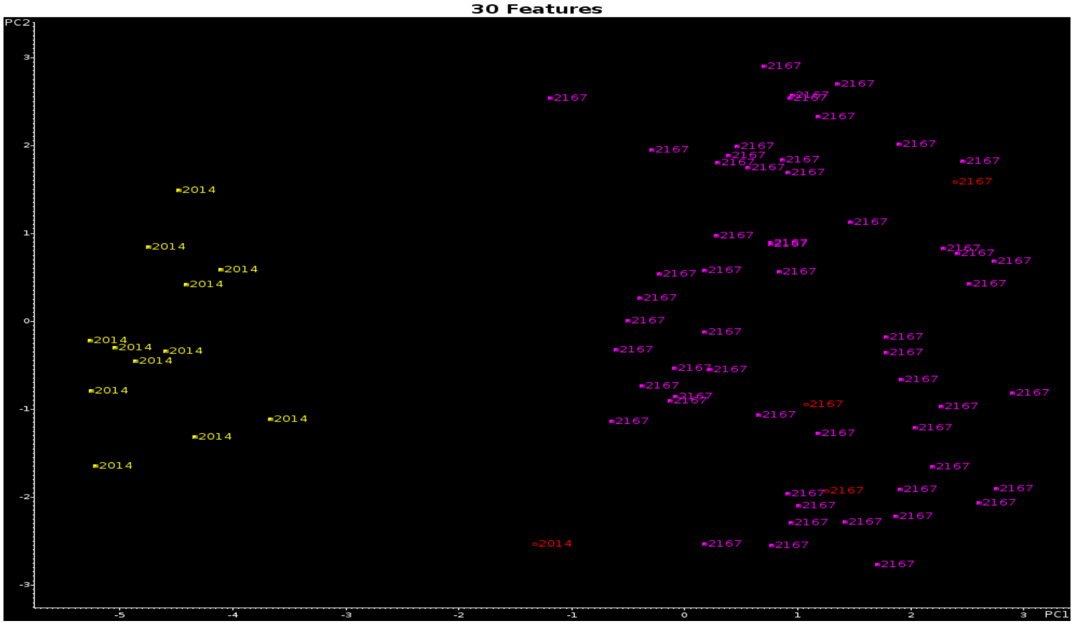


Figure 4.94. Projection of the 4 validation set samples onto the PC plot of the 68 training set samples and the 30 wavelet coefficients identified by the pattern recognition GA (2014= Saint Thomas-Talbotsville, 2167= Kansas City, Louisville)

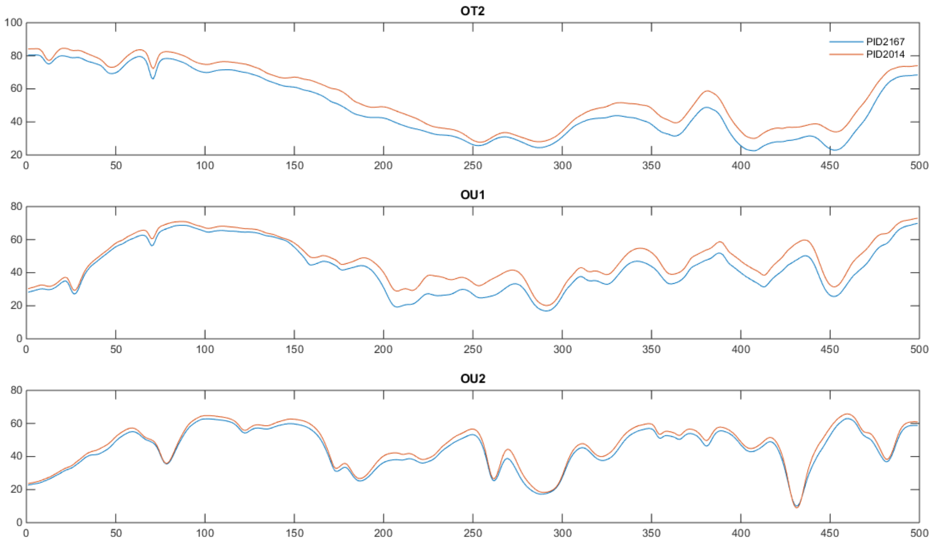


Figure 4.95. The average three-layer IR spectra from the assembly plants: Kansas City, Louisville

The pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.96) showed clustering of IR the spectra on the basis of assembly plants from Honda after 1 generation run. To assess the predictive ability of these 2 wavelet coefficients, a validation set of 1 paint samples located in the Honda region 4 of the second prefilter were projected into 2-PC developed from the 9 training set and the 2 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.97). Samples from the assembly plant Alliston are too less to be used for prediction.

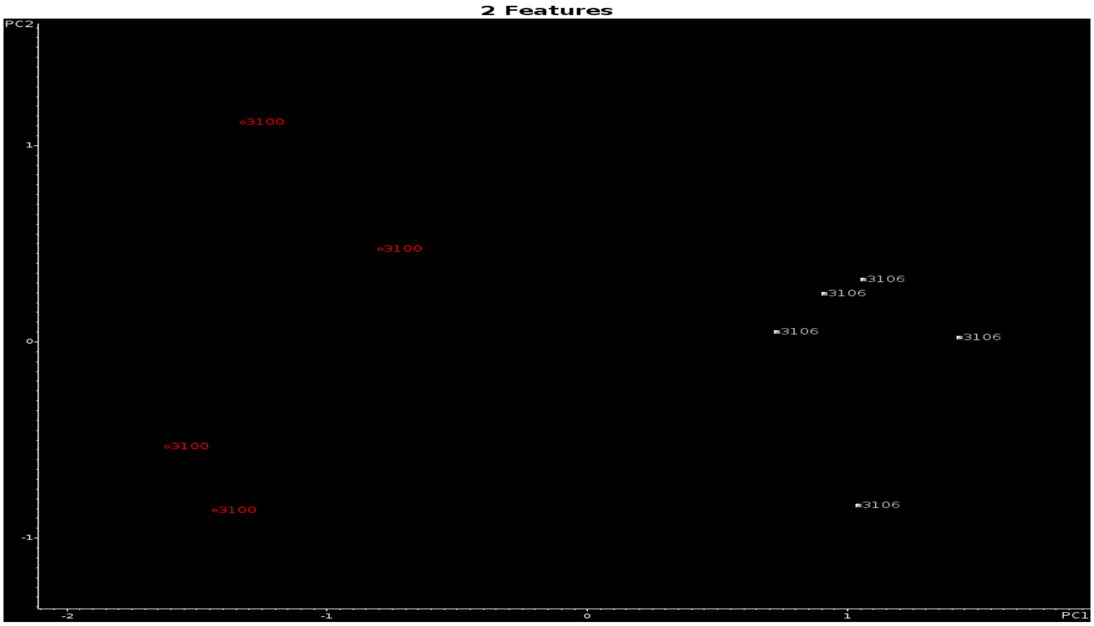


Figure 4.96. 2-PC plot of the 9 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (3100= Alliston, 3106= Marysville)

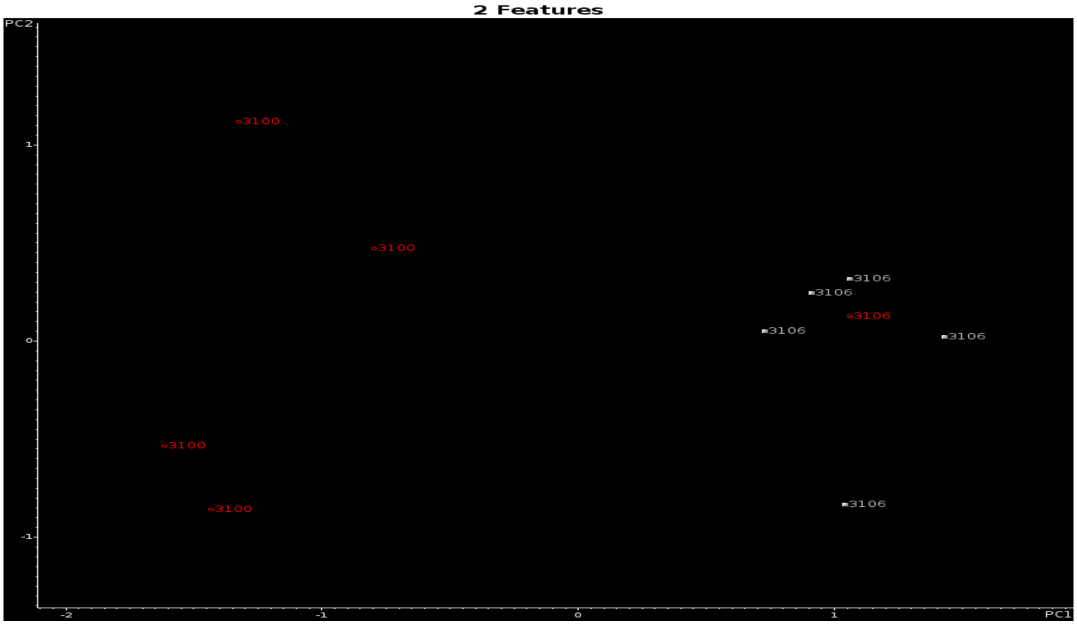


Figure 4.97. Projection of the 1 validation set samples onto the PC plot of the 9 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA(3100= Alliston, 3106= Marysville)

The pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.98) showed clustering of IR the spectra on the basis of assembly plants from Nissan after 1 generation run. To assess the predictive ability of these 2 wavelet coefficients, a validation set of 1 paint samples located in the Nissan region 5 of the second prefilter were projected into 2-PC developed from the 27 training set and the 2 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.99).

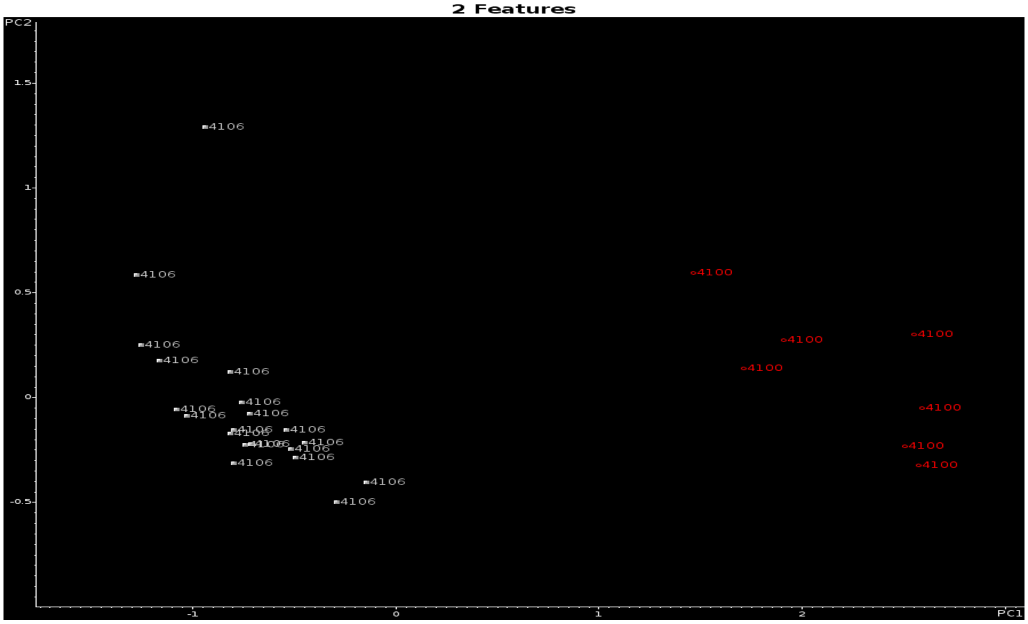


Figure 4.98. 2-PC plot of the 27 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4100= Aguascalientes, 4106= Smyrna)

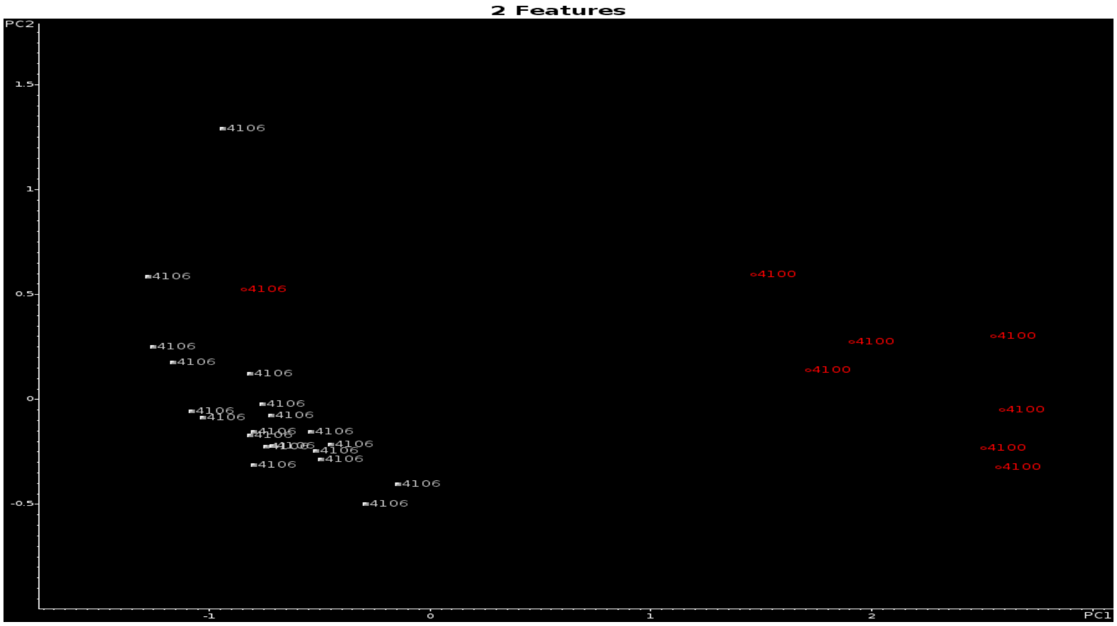


Figure 4.99. Projection of the 1 validation set samples onto the PC plot of the 27 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4100= Aguascalientes, 4106= Smyrna)

For the samples located in the plant group 5, the training and validation sets for manufacturer differentiation in plant group 5 were summarized in Table 4.16. After 200 generations, pattern recognition GA (Fitness function: normal) identified 35 wavelet coefficients whose PC plot (see Figure 4.100) showed clustering of the fused IR spectra on the basis of manufacturers in the plant group 5. The 14 validation set samples were then projected onto the PC plot (see Figure 4.101) define by the 119 training set samples and the 35 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 manufacturer.

Table 4.16. Composition of the IR spectral data set in plant group 5

Manufacturer	Manufacturer IDs	Plant IDs	Training set samples	Validation set samples
Ford	3	2009, 2103, 2111, 2113, 2115, 2116, 2206	46	4
Honda	4	3000, 3002, 3005, 3006	62	6
Toyota	6	5003, 5104, 5303	20	3

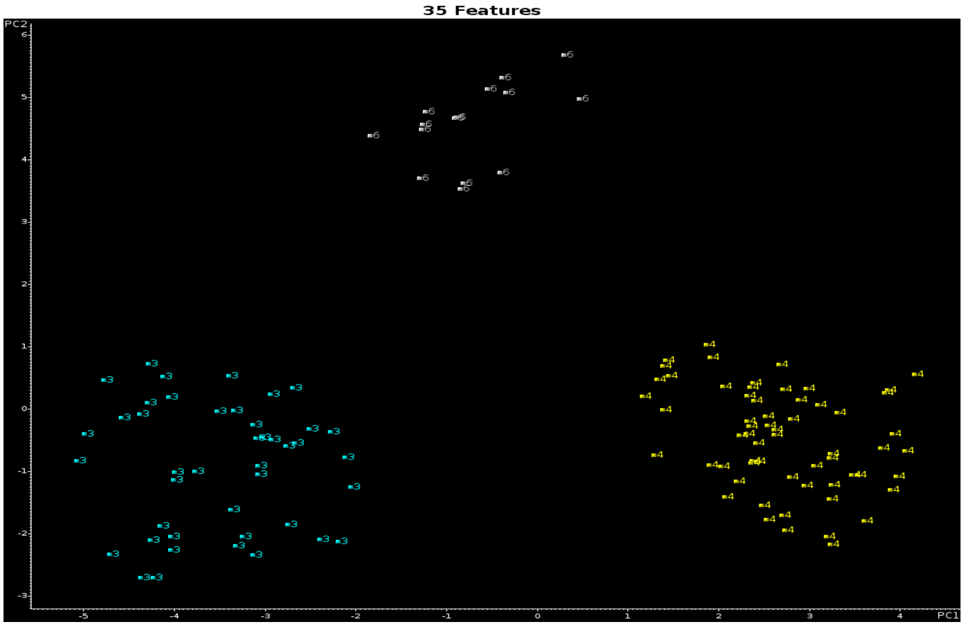


Figure 4.100. 2-PC plot of the 119 training set samples and the 35 wavelet coefficients identified by the pattern recognition GA (3= Ford, 4= Honda, 6=Toyota)

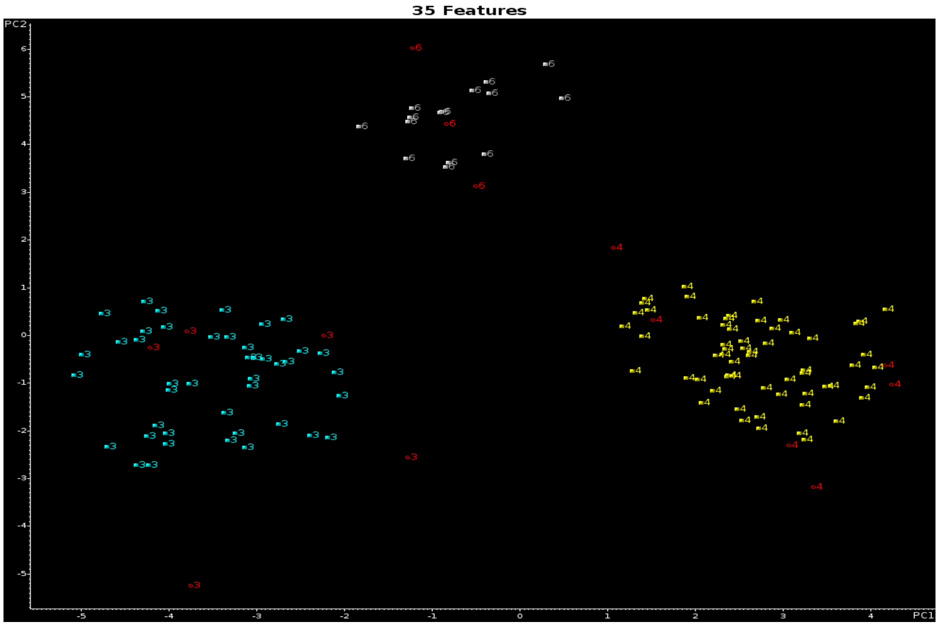


Figure 4.101. Projection of the 14 validation set samples onto the PC plot of the 119 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (3= Ford, 4= Honda, 6=Toyota)

For samples located in the manufacturer cluster 3 required the third level prefilter to identify their assembly plants. The pattern recognition GA identified 17 wavelet coefficients whose PC plot (see Figure 4.102) showed clustering of IR the spectra on the basis of assembly plants from Ford after 36 generation run. To assess the predictive ability of these 17 wavelet coefficients, a validation set of 4 paint samples located in the Ford region of the second prefilter were projected into 2-PC developed from the 43 training samples and the 17 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.103). The assembly plants Norfolk, Dearborn, Twin Cities-Saint Paul and Wayne were merged into one plant ID (26531) due to their similar IR spectra.

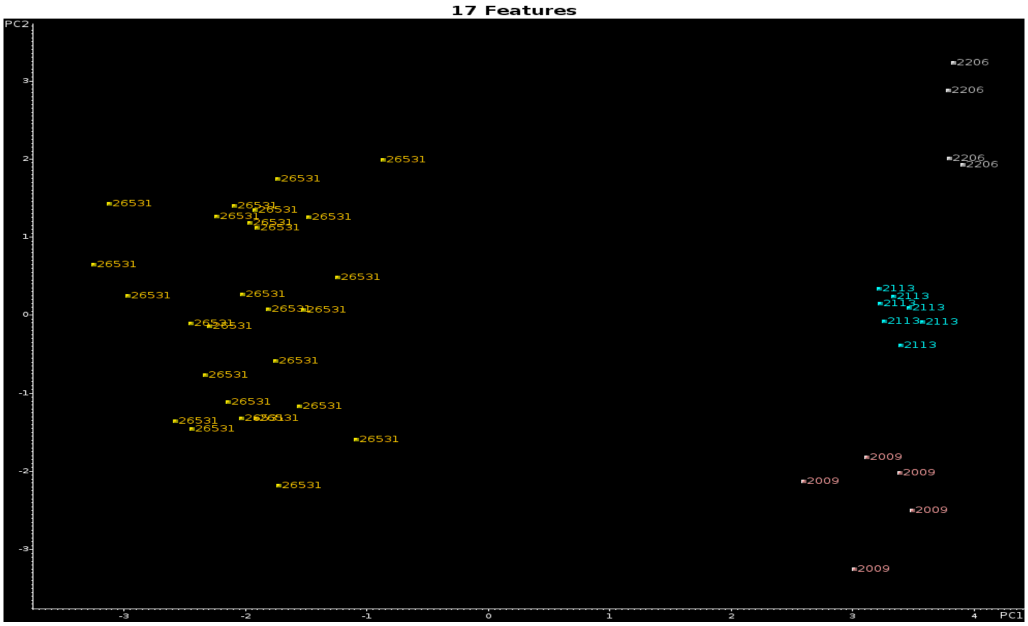


Figure 4.102. 2-PC plot of the 43 training set samples and the 17 wavelet coefficients identified by the pattern recognition GA (2009=Lorain, 2113=Saint Louis, 2206=Hermosillo,26531= Norfolk, Dearborn, Twin Cities-Saint Paul and Wayne)

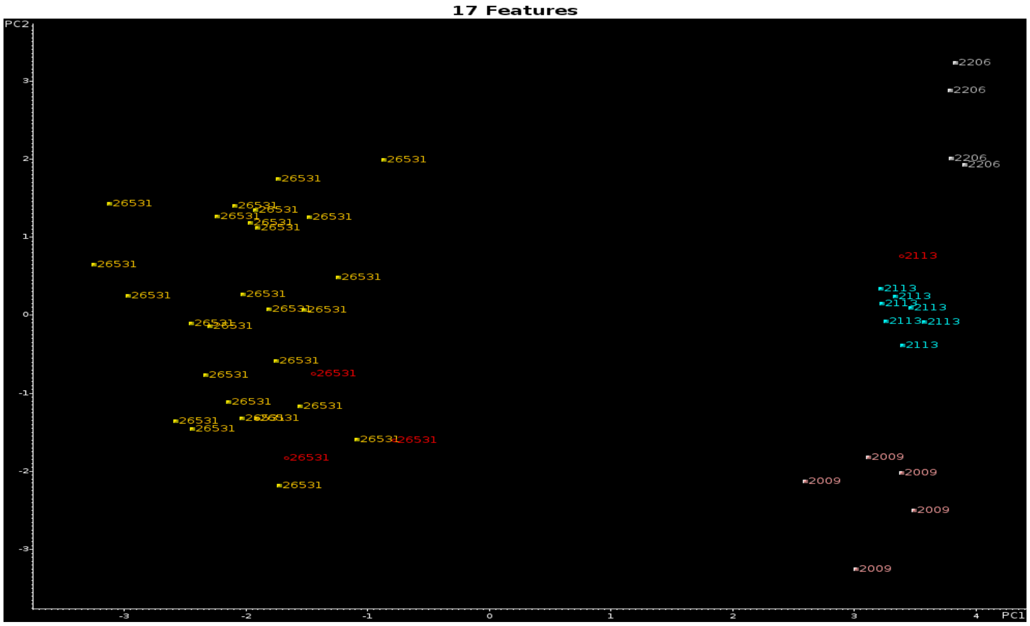


Figure 4.103. Projection of the 4 validation set samples onto the PC plot of the 43 training set samples and the 17 wavelet coefficients identified by the pattern recognition GA (2009=Lorain, 2113=Saint Louis, 2206=Hermosillo,26531= Norfolk, Dearborn, Twin Cities-Saint Paul,Wayne)

For samples located in the manufacturer cluster 4 required the third level prefilter to identify their assembly plants. The pattern recognition GA identified 43 wavelet coefficients whose PC plot (see Figure 4.104) showed clustering of IR the spectra on the basis of assembly plants from Honda after 200 generation run. To assess the predictive ability of these 43 wavelet coefficients, a validation set of 6 paint samples located in the Honda region of the second prefilter were projected into 2-PC developed from the 60 training samples and the 43 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.105). The assembly plants Allison and East Liberty were merged into one plant ID (3802) due to their similar IR spectra.

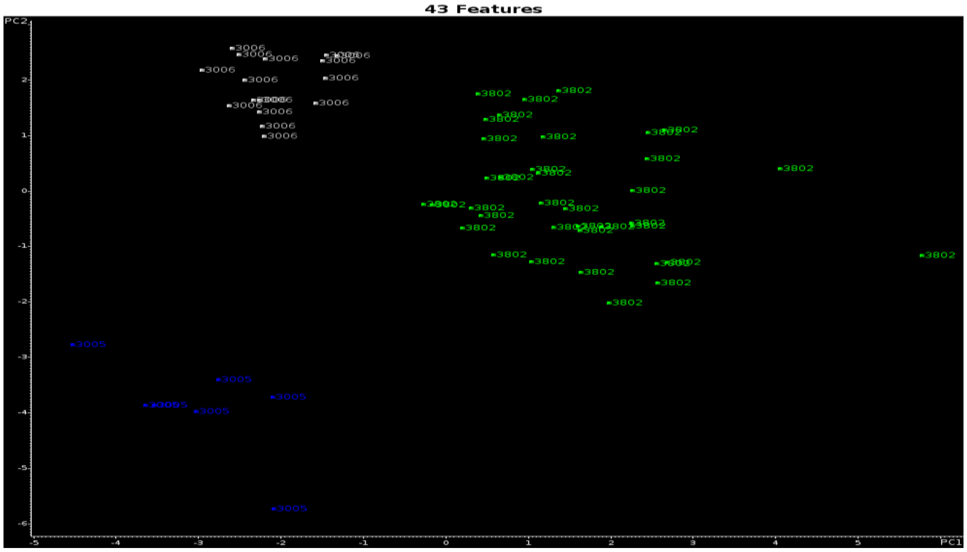


Figure 4.104. 2-PC plot of the 43 training set samples and the 17 wavelet coefficients identified by the pattern recognition GA (3005=Lincoln, 3006=Marysville, 3802=Allison, East Liberty)

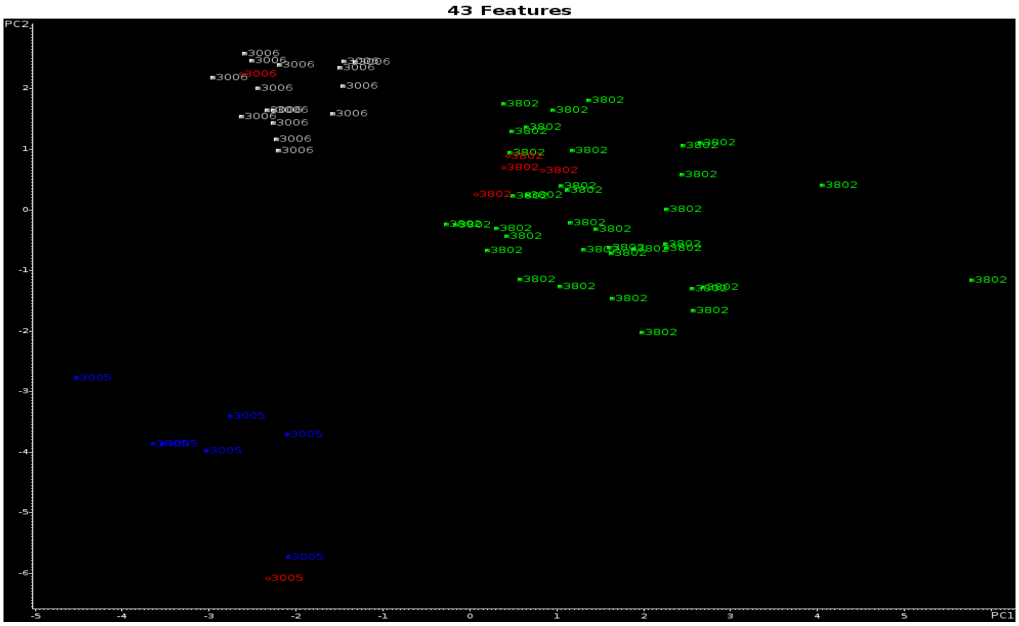


Figure 4.105. Projection of the 6 validation set samples onto the PC plot of the 60 training set samples and the 43 wavelet coefficients identified by the pattern recognition GA (3005=Lincoln, 3006=Marysville, 3802=Allison, East Liberty)

For samples located in the manufacturer cluster 6 required the third level prefilter to identify their assembly plants. The pattern recognition GA identified 5 wavelet coefficients whose PC plot (see Figure 4.106) showed clustering of IR the spectra on the basis of assembly plants from Toyota after 5 generation run. To assess the predictive ability of these 5 wavelet coefficients, a validation set of 3 paint samples located in the Toyota region of the second prefilter were projected into 2-PC developed from the 15 training samples and the 5 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.107).

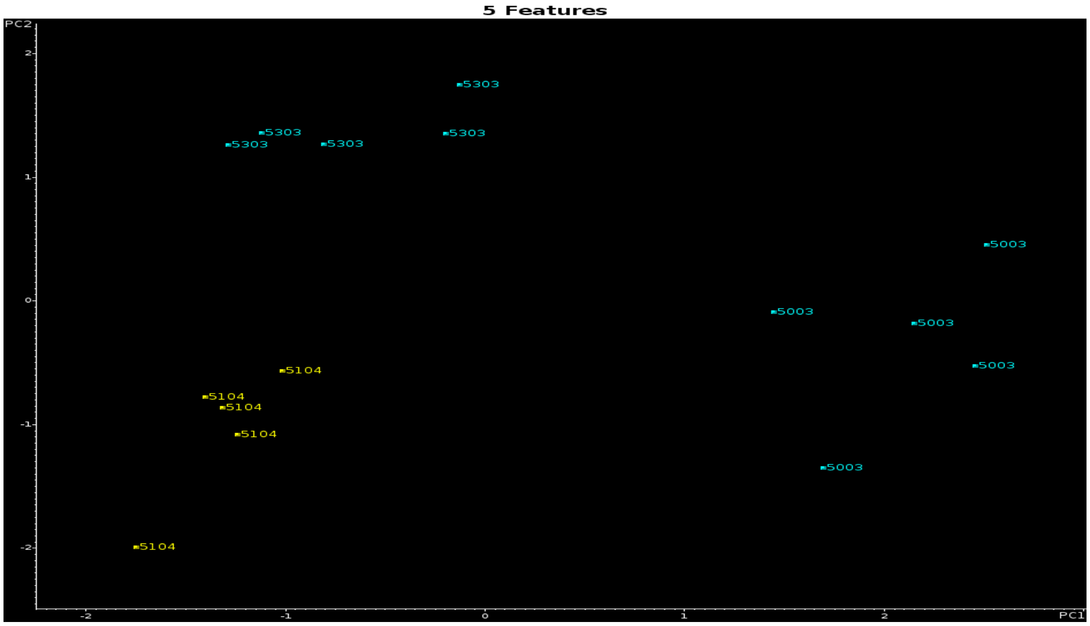


Figure 4.106. 2-PC plot of the 15 training set samples and the 5 wavelet coefficients identified by the pattern recognition GA (5003=Fremont, 5104=Georgetown, 5303=Fremont)

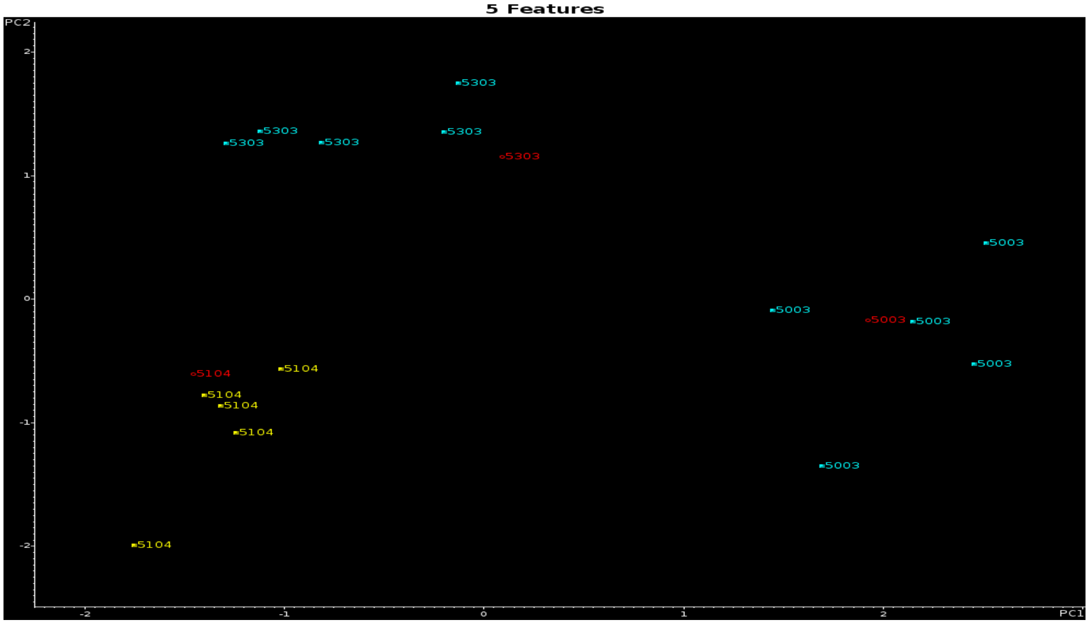


Figure 4.107. Projection of the 3 validation set samples onto the PC plot of the 15 training set samples and the 5 wavelet coefficients identified by the pattern recognition GA (5003=Fremont, 5104=Georgetown, 5303=Fremont)

For the samples located in the plant group 6, the training and validation sets for manufacturer differentiation in plant group 6 were summarized in Table 4.17. After 200 generations, pattern recognition GA (Fitness function: normal) identified 50 wavelet coefficients whose PC plot (see Figure 4.108) showed clustering of the fused IR spectra on the basis of manufacturers in the plant group 6. The 8 validation samples were then projected onto the PC plot (see Figure 4.109) define by the 77 training samples and the 50 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 manufacturer.

Table 4.17. Composition of the IR spectral data set in plant group 6

Manufacturer	Manufacturer IDs	Plant IDs	Training set samples	Validation set samples
Honda	4	3007, 3008, 3200	31	3
Nissan	5	4000, 4005, 4007, 4104	37	5
Toyota	6	5105, 5204	9	0

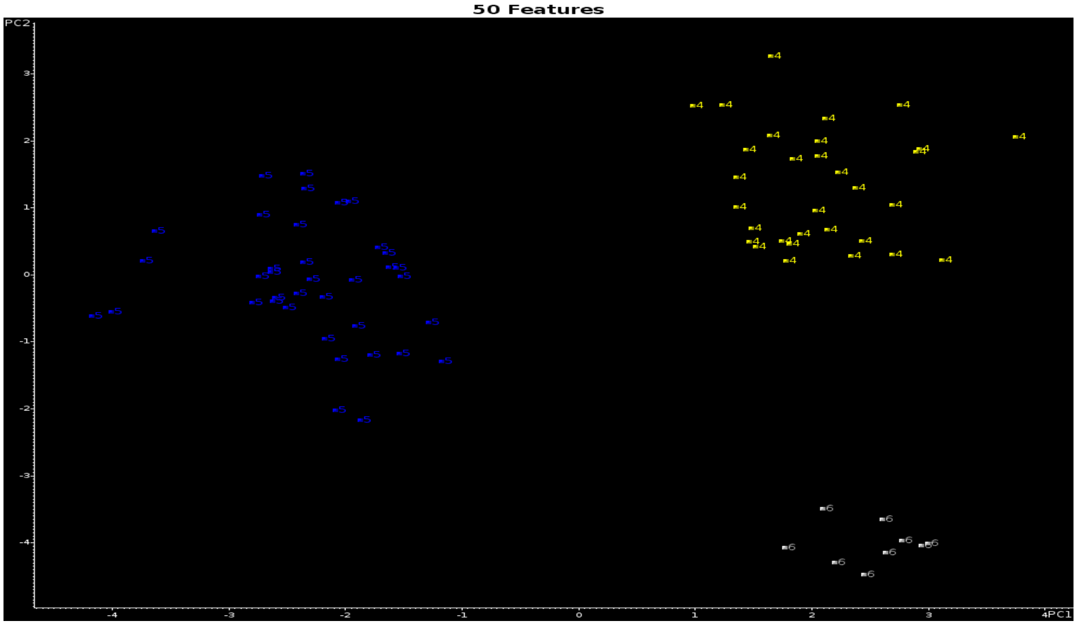


Figure 4.108 2-PC plot of the 77 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA (4=Honda, 5=Nissan, 6=Toyota)

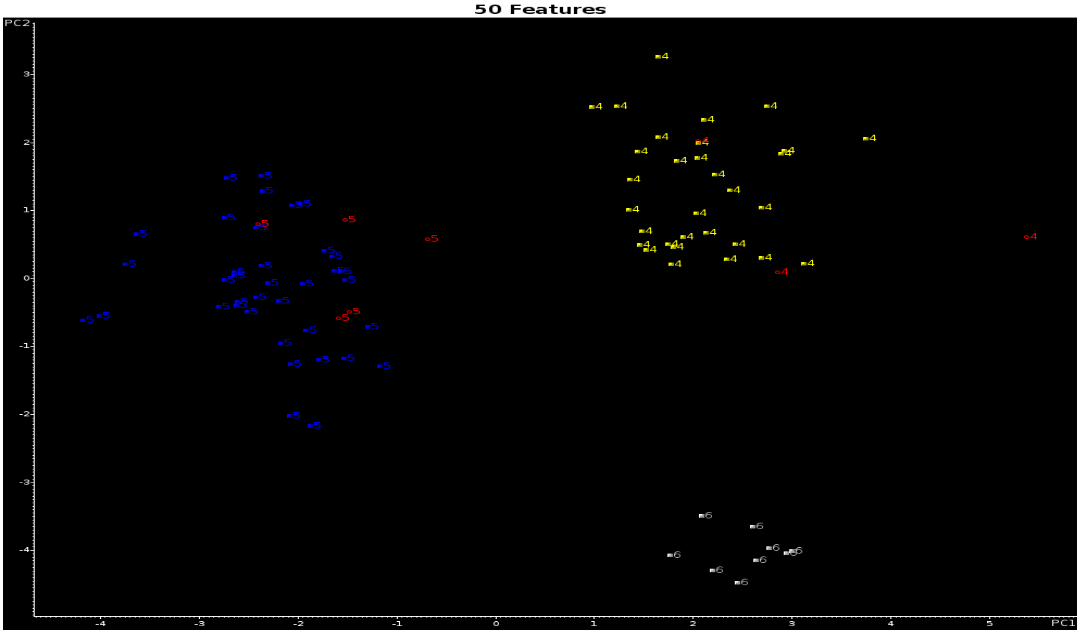


Figure 4.109 Projection of the 8 validation set samples onto the PC plot of the 77 training set samples and the 50 wavelet coefficients identified by the pattern recognition GA (4=Honda, 5=Nissan, 6=Toyota)

The pattern recognition GA identified 5 wavelet coefficients whose PC plot (see Figure 4.110) showed clustering of IR the spectra on the basis of assembly plants from Honda after 7 generations run. To assess the predictive ability of these 7 wavelet coefficients, a validation set of 3 paint samples located in the Honda region of the second prefilter were projected into 2-PC developed from the 31 training samples and the 7 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.111). The assembly plants Sayama and Suzuka were merged into the new assembly plant whose ID is 3087.

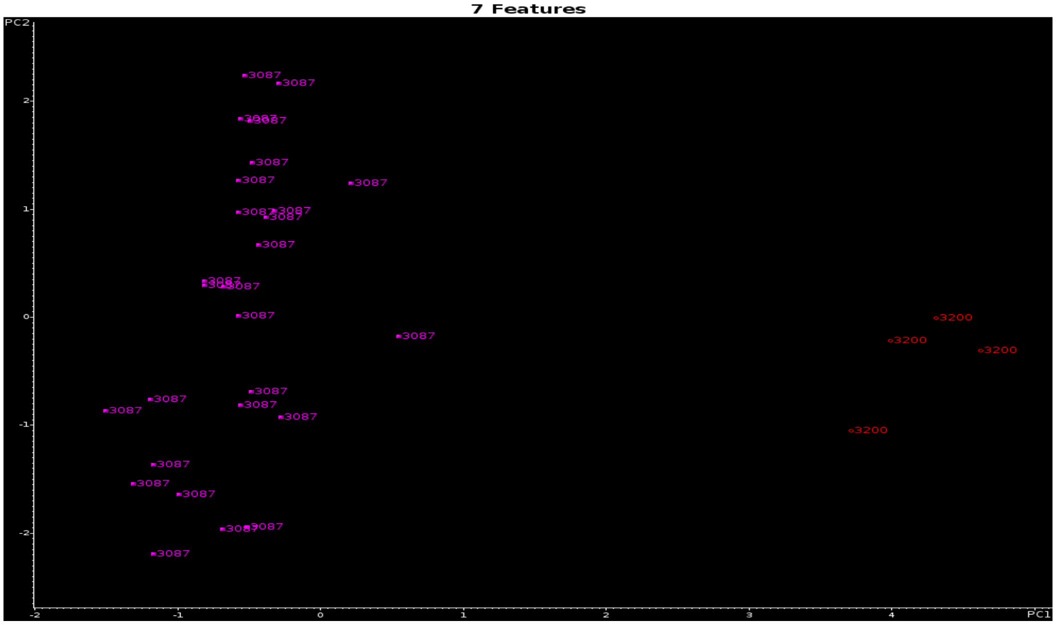


Figure 4.109 2-PC plot of the 31 training set samples and the 7 wavelet coefficients identified by the pattern recognition GA (3087=Sayama, Suzuka, 3200=Alliston)

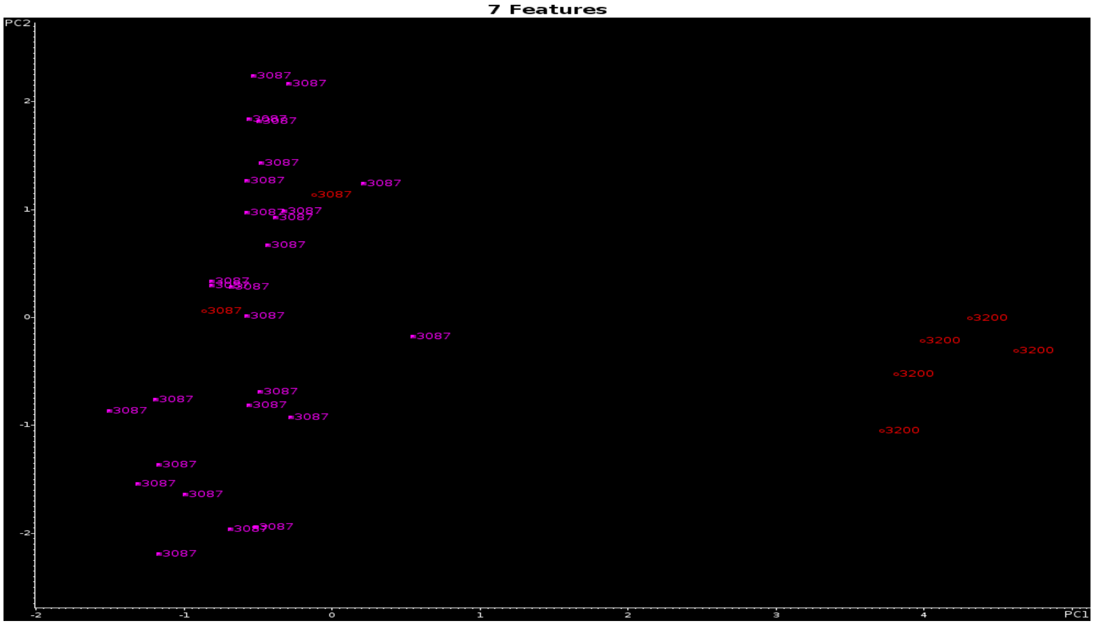


Figure 4.111 Projection of the 3 validation set samples onto the PC plot of the 31 training set samples and the 7 wavelet coefficients identified by the pattern recognition GA (3087=Sayama, Suzuka, 3200=Alliston)

The pattern recognition GA identified 16 wavelet coefficients whose PC plot (see Figure 4.112) showed clustering of IR the spectra on the basis of assembly plants from Nissan after 40 generations run. To assess the predictive ability of these 16 wavelet coefficients, a validation set of 5 paint samples located in the Nissan region of the second prefilter were projected into 2-PC developed from the 37 training samples and the 16 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.113).

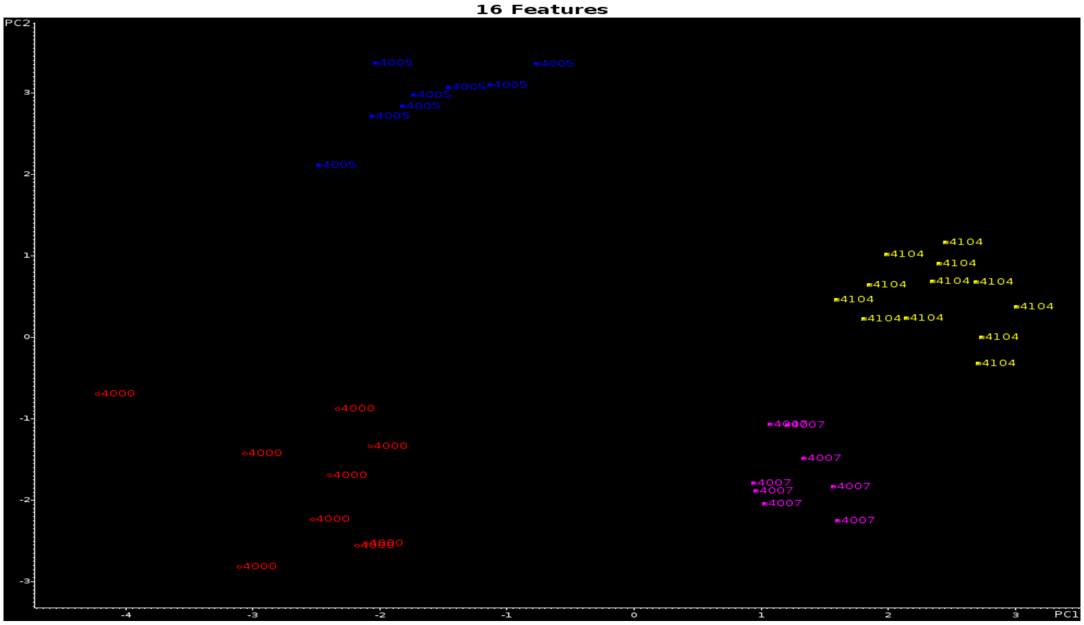


Figure 4.112 2-PC plot of the 37 training set samples and the 16 wavelet coefficients identified by the pattern recognition GA (4000= Aguascalientes, 4005=Oppama, 4007=Tochigi, 4104=Kyushu)

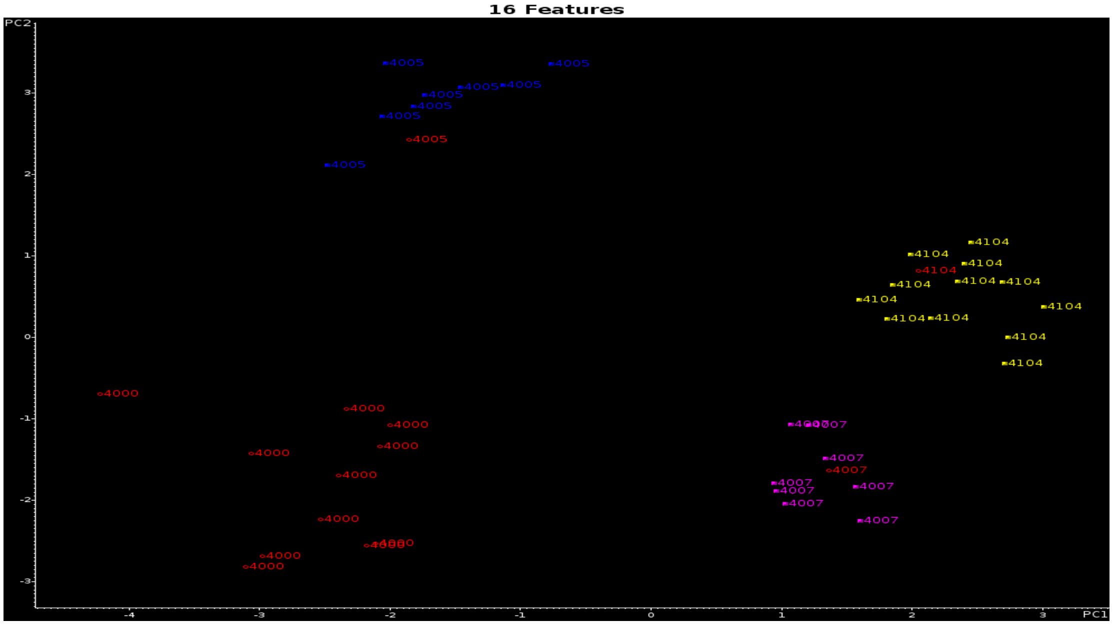


Figure 4.113 Projection of the 5 validation set samples onto the PC plot of the 37 training set samples and the 16 wavelet coefficients identified by the pattern recognition GA (4000= Aguascalientes, 4005=Oppama, 4007=Tochigi, 4104=Kyushu)

The pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.114) showed clustering of IR the spectra on the basis of assembly plants from Toyota after 1 generation run. Because the number of samples in the both two assembly plant is too less to assess the predictive ability of these 2 wavelet coefficients, there is no validation sample was set in this test.

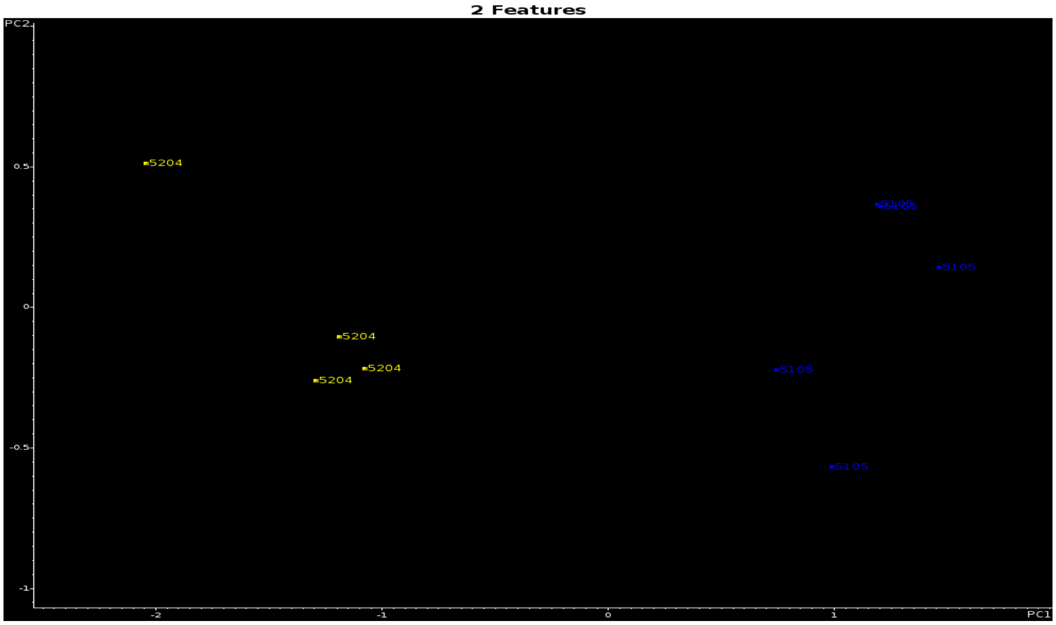


Figure 4.114 2-PC plot of the 9 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (5105=Japan, 5204=Georgetown)

4.4.6 Two-layer Search Prefilters

OT2 and OU1 were conjugated together to build up the two-layer prefilter, we got almost the same results except the following experiments, in which the two-layer prefilter performed worse than three-layer prefilter. The space between the two classes is bigger in three-layer prefilter than in the two-layer one (See Figure 4.115- Figure 4.116). For the assembly plants in the GM subgroup 5, the sample from the assembly plant Oklahoma City was predict wrong in the assembly plant Oshawa by using two-layer prefilter (See Figure 4.117- Figure 4.118); instead, it was predicted correctly by using three-layer prefilter in the same conditions. The assembly plants Saint Thomas-lalbotsville from the Ford region 33 in the group 4 cluster were always predicted wrong no matter using three-layer or two-layer prefilter. But the two-layer prefilter got worse prediction than the three-layer one (See Figure 4.119-Figure 4.120), the validation sample from Saint Thomas-lalbotsville totally mixed with the samples from Kansas City or Louisville. The IR spectra from OU1 for the both two assembly plants are very similar and lack of the discriminative ability of an assembly plant (Figure 4.121). Overall, two-layer prefilter has no problem to differentiate the paint fragment samples instead of three-layer prefilter except the above three situations.

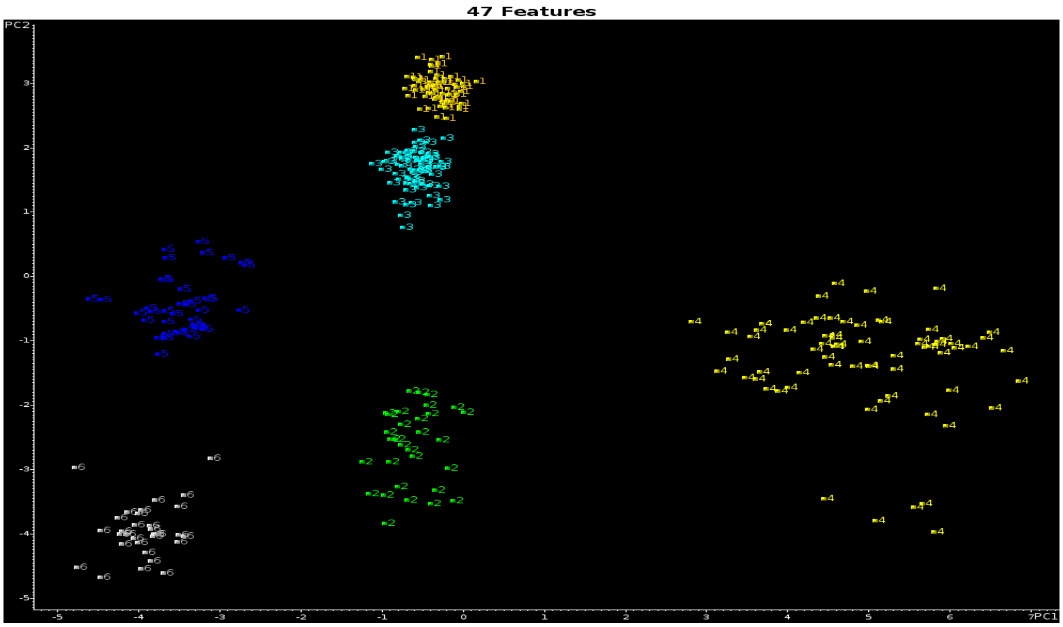


Figure 4.115. 2-PC plot of the 311 training set samples and the 47 wavelet coefficients identified by the pattern recognition GA (1=GMsubgroup1, 2=GMsubgroup2, 3= GMsubgroup3, 4=GMsubgroup4, 5=GMsubgroup5, 6=GMsubgroup6)

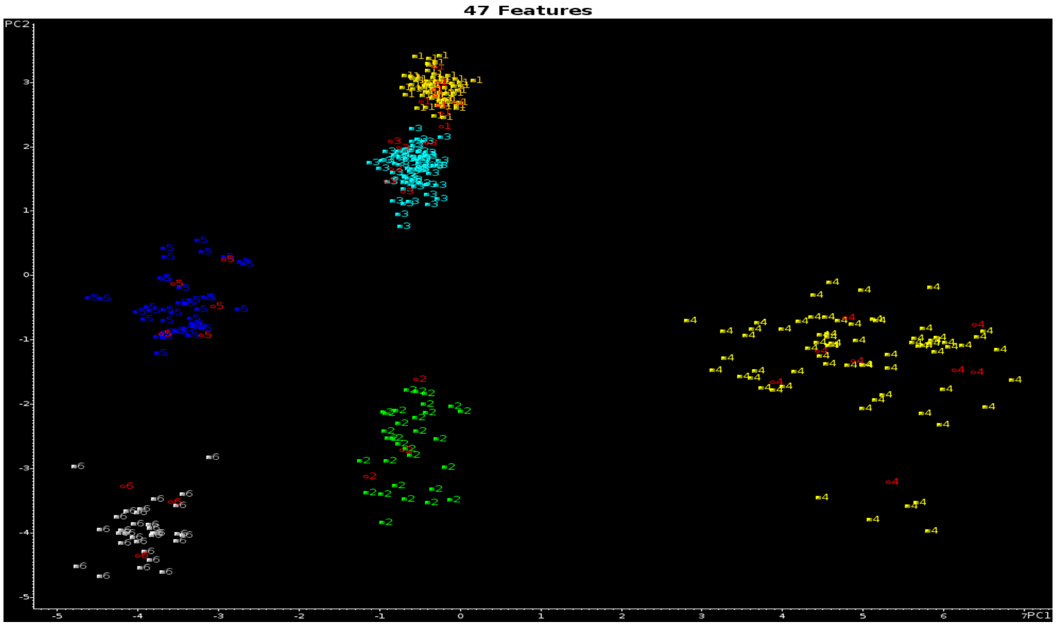


Figure 4.116 Projection of the 34 validation set samples onto the PC plot of the 311 training set samples and the 47 wavelet coefficients identified by the pattern recognition GA (1=GMsubgroup1, 2=GMsubgroup2, 3= GMsubgroup3, 4=GMsubgroup4, 5=GMsubgroup5, 6=GMsubgroup6)

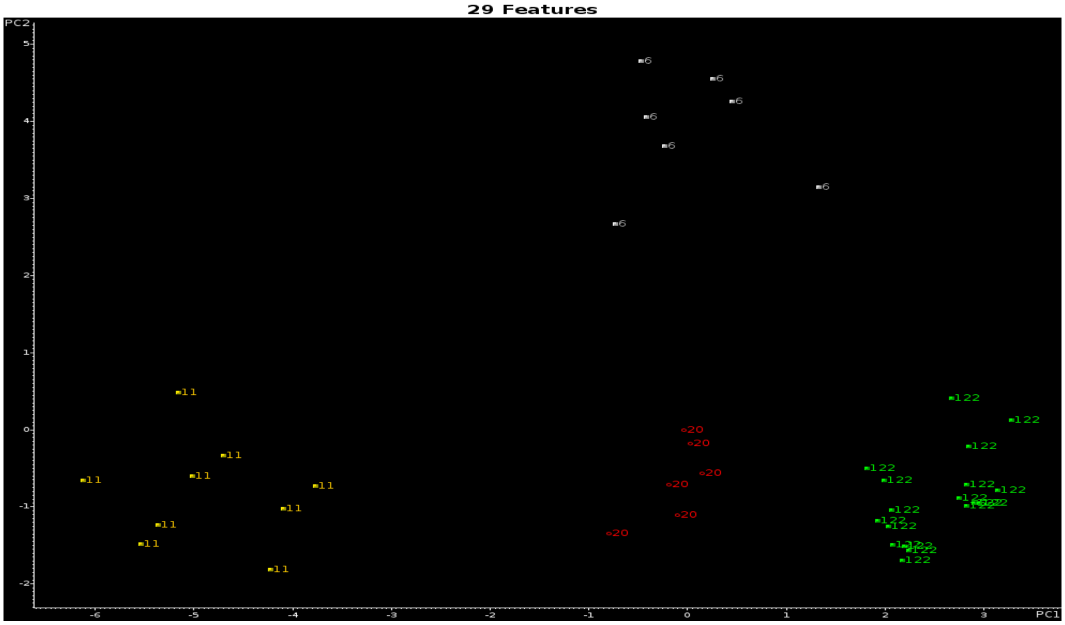


Figure 4.117 2-PC plot of the 40 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (6=Flint, 11=Ingersoll, 20=Oklahoma City, 122=Oshawa)

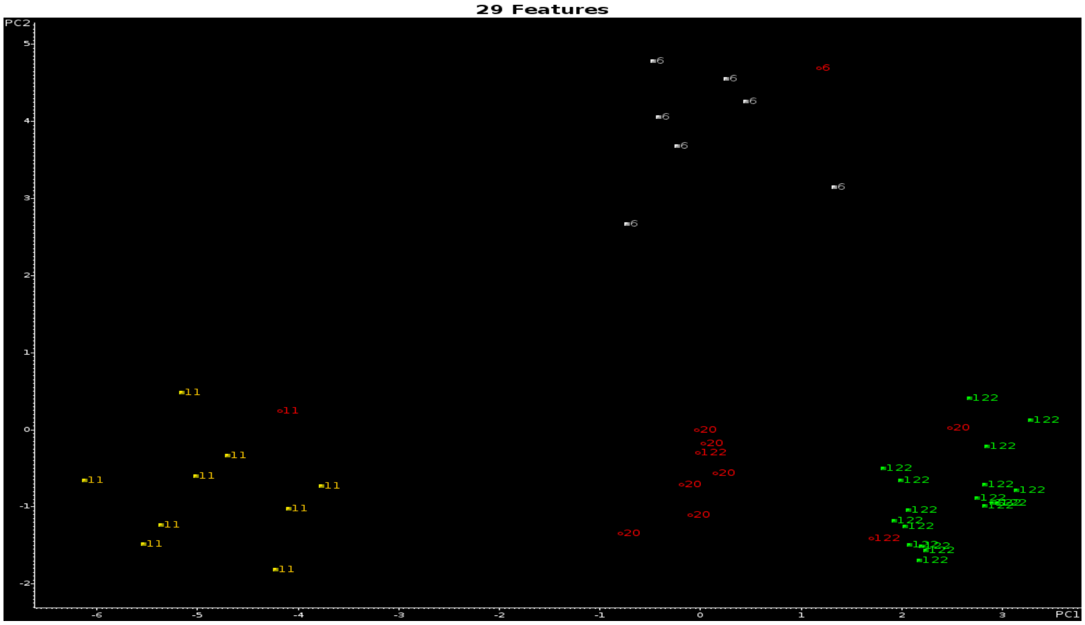


Figure 4.118. Projection of the 5 validation set samples onto the PC plot of the 40 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (6=Flint, 11=Ingersoll, 20=Oklahoma City, 122=Oshawa)

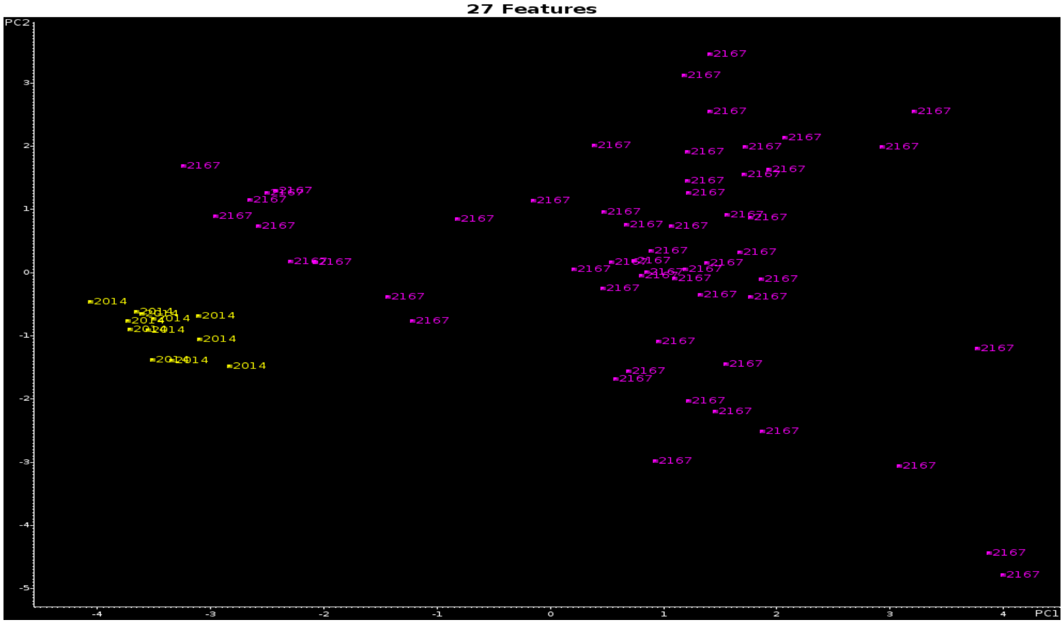


Figure 4.119. 2-PC plot of the 68 training set samples and the 30 wavelet coefficients identified by the pattern recognition GA (2014= Saint Thomas-Talbotsville, 2167= Kansas City, Louisville)

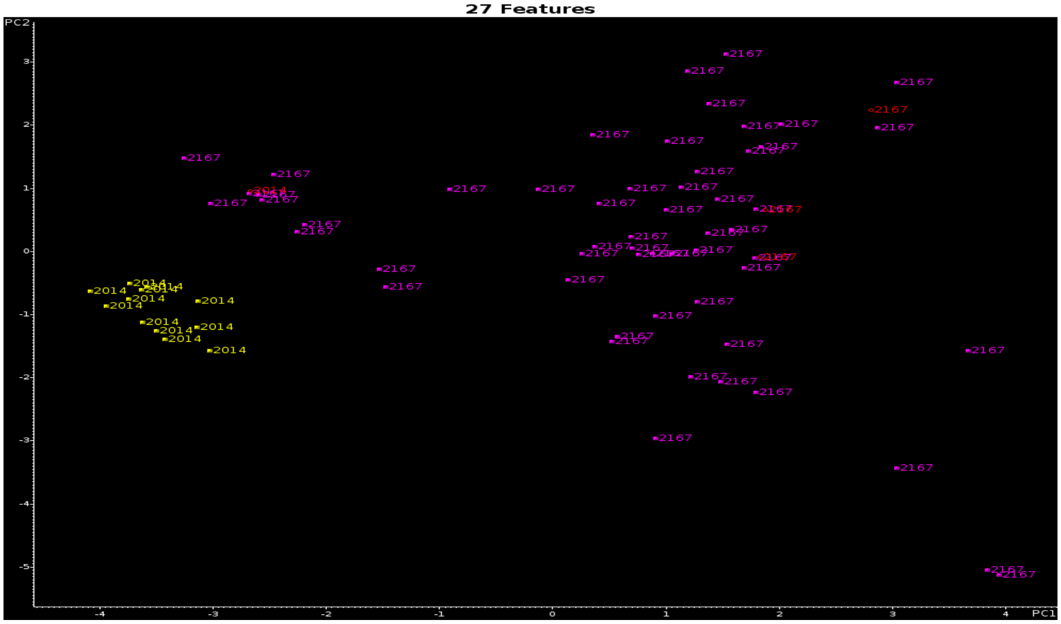


Figure 4.120. Projection of the 4 validation set samples onto the PC plot of the 68 training set samples and the 27 wavelet coefficients identified by the pattern recognition GA (2014= Saint Thomas-Talbotsville, 2167= Kansas City, Louisville)

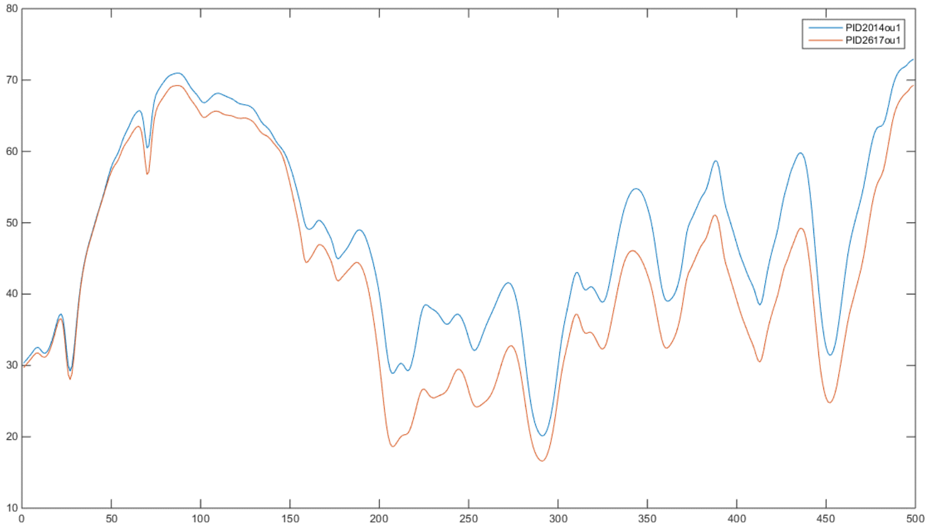


Figure 4.121. The comparison of the average IR spectra of the assembly plant Thomas-lalbotsville vs the assembly plant Kansas City or Louisville

4.4.7 Three-layer Search Prefilters Based on Manufacturer

For the sample with a single carbonyl band, the previous study discovered that it was impossible to assign an unknown paint sample in the basis of six manufacturers because the relationship between the manufacturers are not linear. The training samples from GA is different from ones made at other five manufacturers. This method aims to develop the first prefilter for singlet sample prediction basis of manufacturer, the second prefilter basis of plant group in a particular manufacturer and the third one for the prediction of assembly plant. A block diagram of the vehicle sample classification process used in the prototype pattern recognition assisted library search system for the PDQ database is summarized in Figure 4.122.

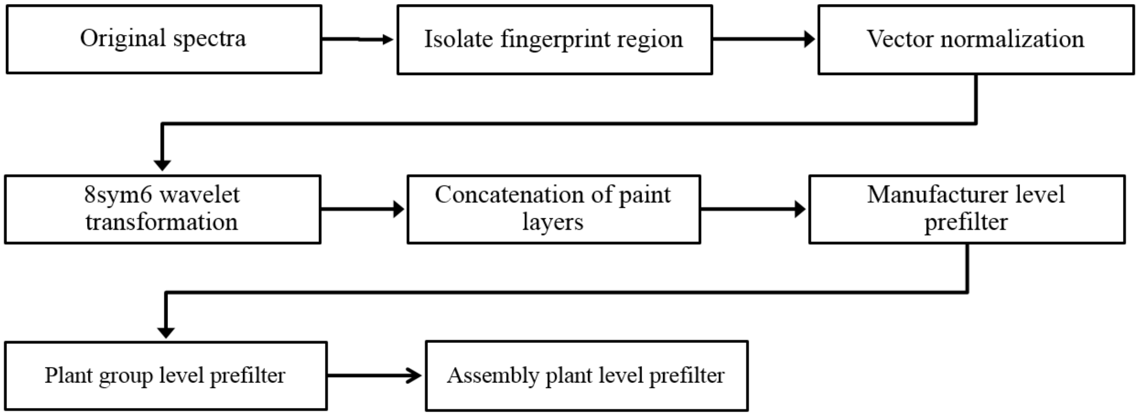

Original spectra
Isolate fingerprint region
Vector normalization
8sym6 wavelet transformation
Concatenation of paint layers
Manufacturer level prefilter
Plant group level prefilter
Assembly plant level prefilter

Figure 4.122. Block diagram of the vehicle classification process used in the prototype pattern recognition driven library search system for the PDQ database.

4.4.7.1 Manufacturer Search Prefilters

A block diagram of the developing process of the manufacturer search prefilter is summarized in Figure 4.123. The detail GA running results are showed in Figure 4.124-Figure 4.138.

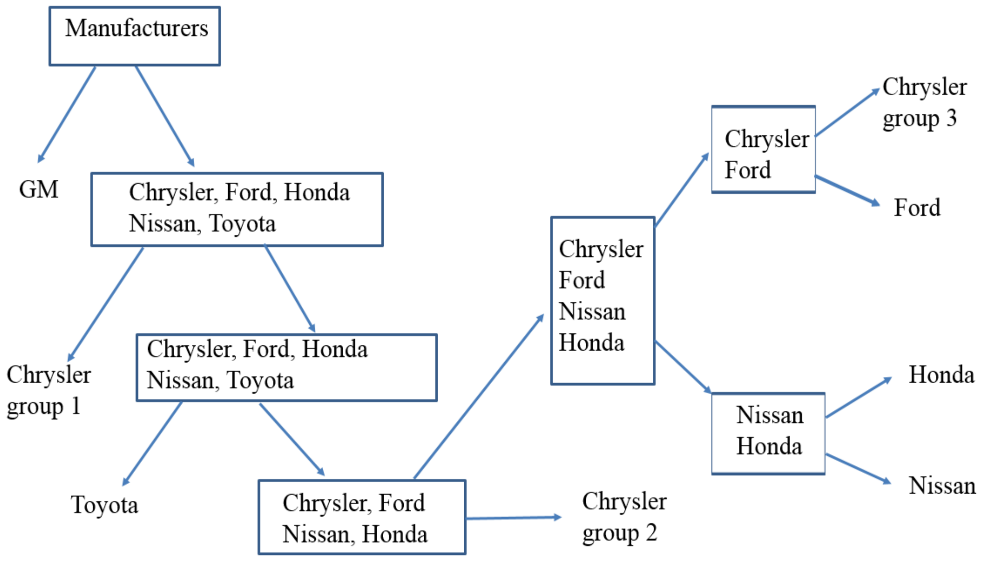

Manufacturers
GM
Chrysler, Ford, Honda Nissan, Toyota
Chrysler group 1
Chrysler, Ford, Honda Nissan, Toyota
Toyota
Chrysler, Ford Nissan, Honda
Chrysler group 2
Chrysler Ford Nissan Honda
Chrysler Ford
Chrysler group 3
Ford
Nissan Honda
Honda
Nissan

Figure 4.123. Block diagram of the manufacturer search prefilter developing process

First, GM was expected to be separated from the other five manufacturers. 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 154 generations, the pattern recognition GA identified 29 wavelet coefficients whose PC plot showed clustering of the fused IR spectra on the basis of GM and the automotive manufacturer comprising of Chrysler, Ford, Honda, Nissan and Toyota (see Figure 4.124), the fitness function was Normal (PCKaNN). To assess the predictive ability of the 29 wavelet coefficients identified by the pattern recognition GA, a validation set of 136 paint samples was employed. In Figure 4.125, the validation set samples are projected onto the PC plot of the data defined by the 1374 wavelet transformed fused IR spectra and the 29 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 whether the automotive manufacturer is GM or not can be extracted from the wavelet transformed fused IR spectrum of an unknown paint sample.

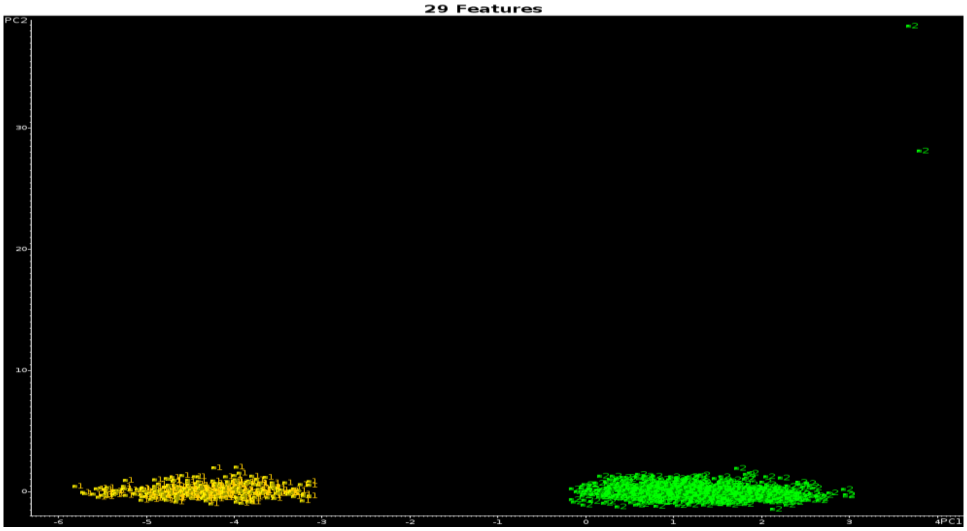


Figure 4.124. 2-PC plot of the 1374 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (1= GM , 2= Chrysler, Ford, Honda, Nissan and Toyota)

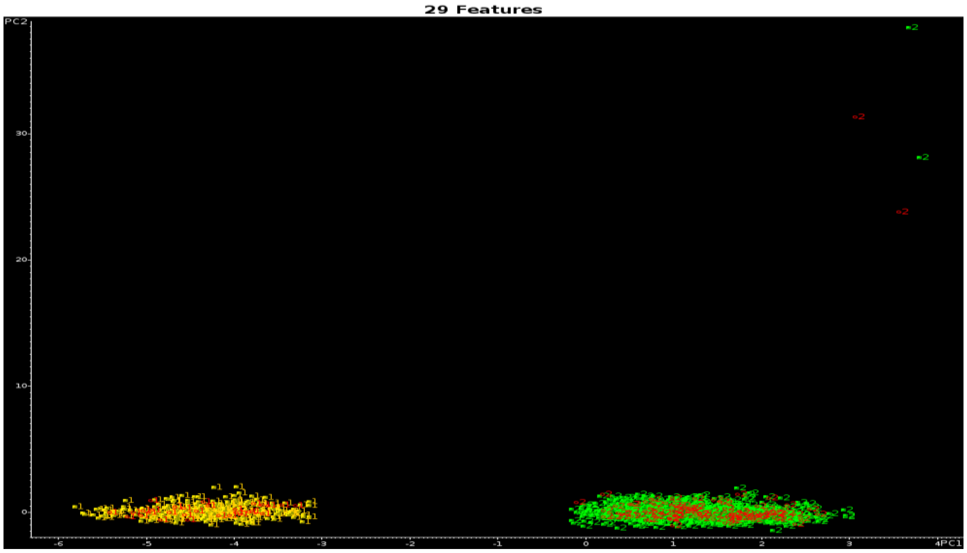


Figure 4.125. Projection of the 136 validation set samples onto the PC plot of the 1374 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (1= GM , 2= Chrysler, Ford, Honda, Nissan and Toyota)

If an unknown sample falls in the cluster1, and this sample will go to the GM prefilter; otherwise, this sample should continue to explore its manufacturer. The second step is to discriminate this sample belonging to the Chrysler or not. Nevertheless, the training samples in the Chrysler are not homogenous and divided into three groups in the all feature 2-PC plot (Figure 4.126), Table 4.18 listed the detail composition of each Chrysler group.

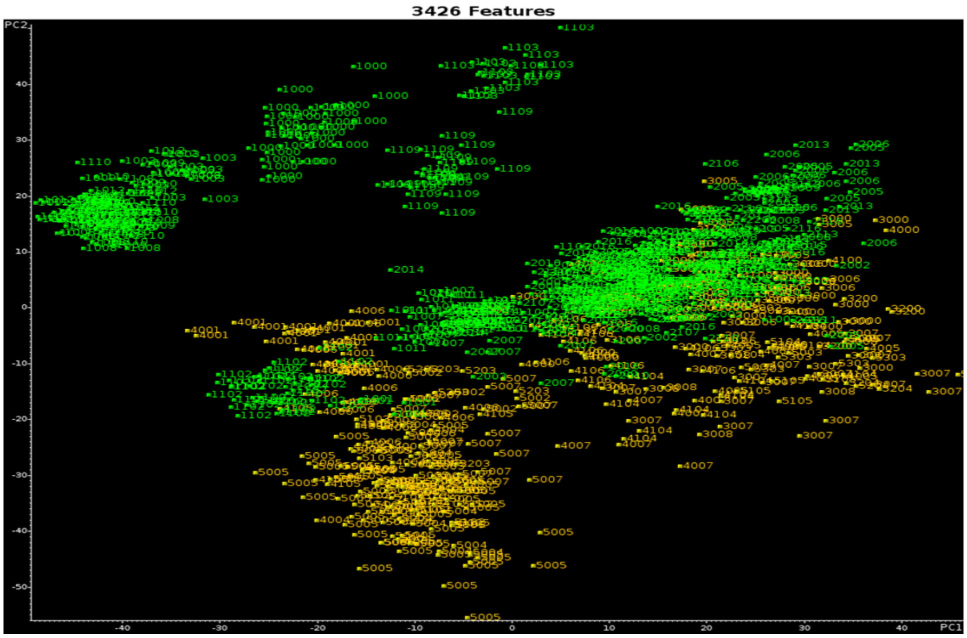


Figure 4.126. 2-PC plot of the assembly plants from Chrysler, Ford, Honda, Nissan and Toyota

Table 4.18 The Chrysler group composition in the basis of assembly plants

Chrysler Group #	PIDs
1	1000,1003,1008,1009,1012,1103,1108,1109,1110
2(1)	1007,1011
2(2)	1001,1102
3	1002,1010

The pattern recognition GA identified 78 wavelet coefficients whose PC plot (see Figure 4.127) showed clustering of IR the spectra on the basis of assembly plants from Chrysler after 200 generations run. To assess the predictive ability of these 78 wavelet coefficients, a validation set of 106 paint samples were projected into 2-PC developed from the 1054 training samples and the 78 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.128).

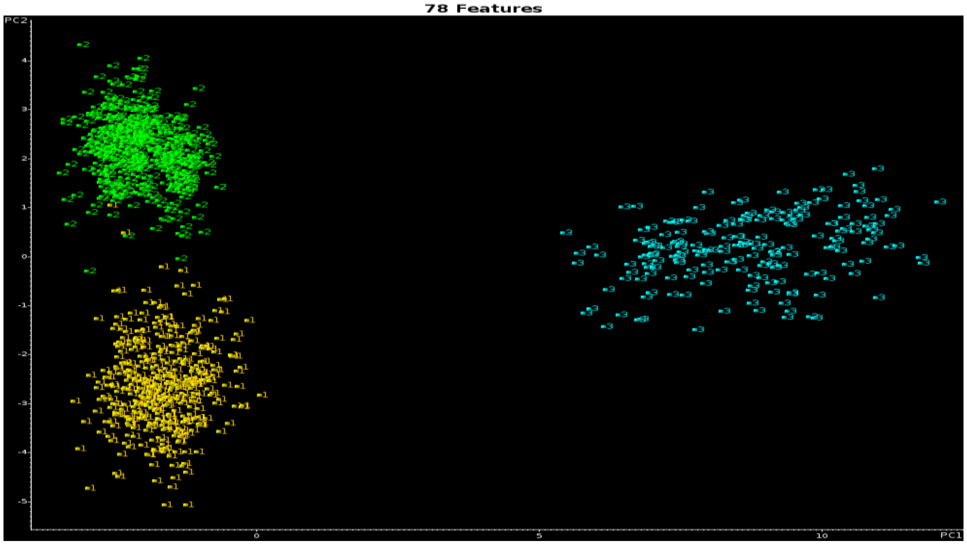


Figure 4.127. 2-PC plot of the 1054 training set samples and the 78 wavelet coefficients identified by the pattern recognition GA (1= Honda, Nissan and Toyota , 2= Chrysler group2 and group3, Ford, 3 = Chrysler group1)

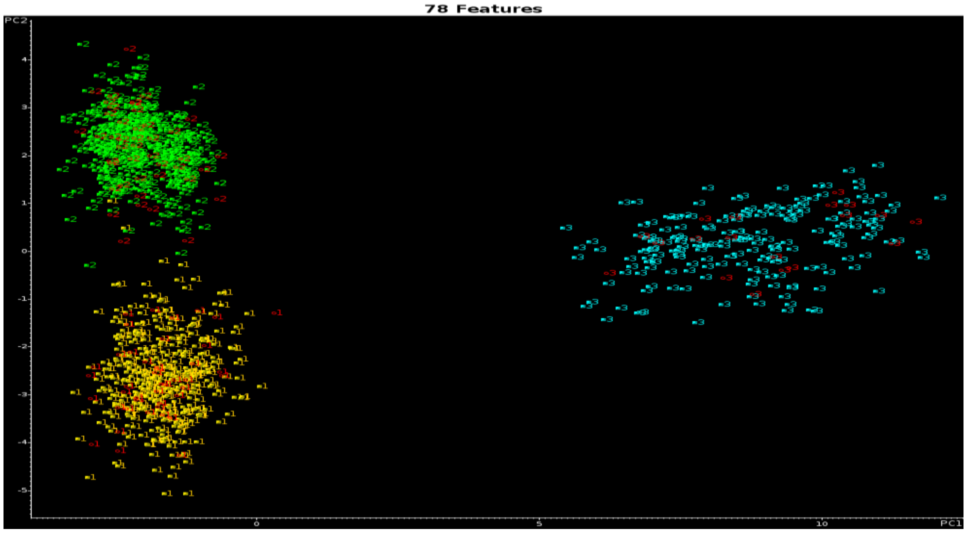


Figure 4.128. Projection of the 106 validation set samples onto the PC plot of the 1054 training set samples and the 78 wavelet coefficients identified by the pattern recognition GA (1= Honda, Nissan and Toyota , 2= Chrysler group2 and group3, Ford, 3 = Chrysler group1)

The next step was to identify an unknown sample from Toyota or not if this sample did not fall in the first Chrysler group. The pattern recognition GA identified 55 wavelet coefficients whose PC plot (see Figure 4.129) showed clustering of IR the spectra on the basis of assembly plants from Toyota after 200 generations run. To assess the predictive ability of these 55 wavelet coefficients, a validation set of 87 paint samples were projected into 2-PC developed from the 865 training samples and the 55 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.130).

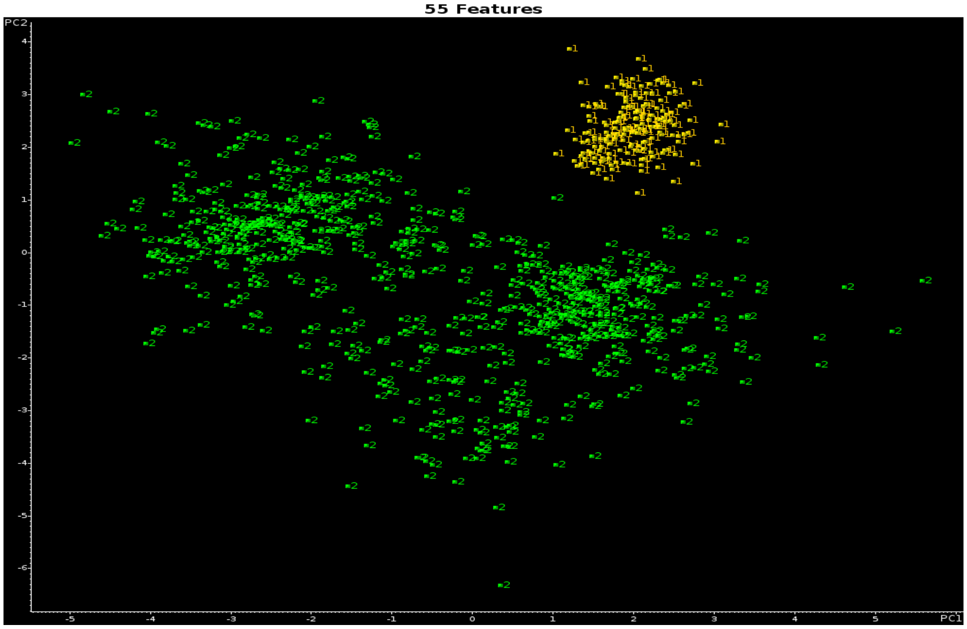


Figure 4.129. 2-PC plot of the 865 training set samples and the 55 wavelet coefficients identified by the pattern recognition GA (1= Toyota , 2= Honda, Nissan, Chrysler group2 and group3, Ford)

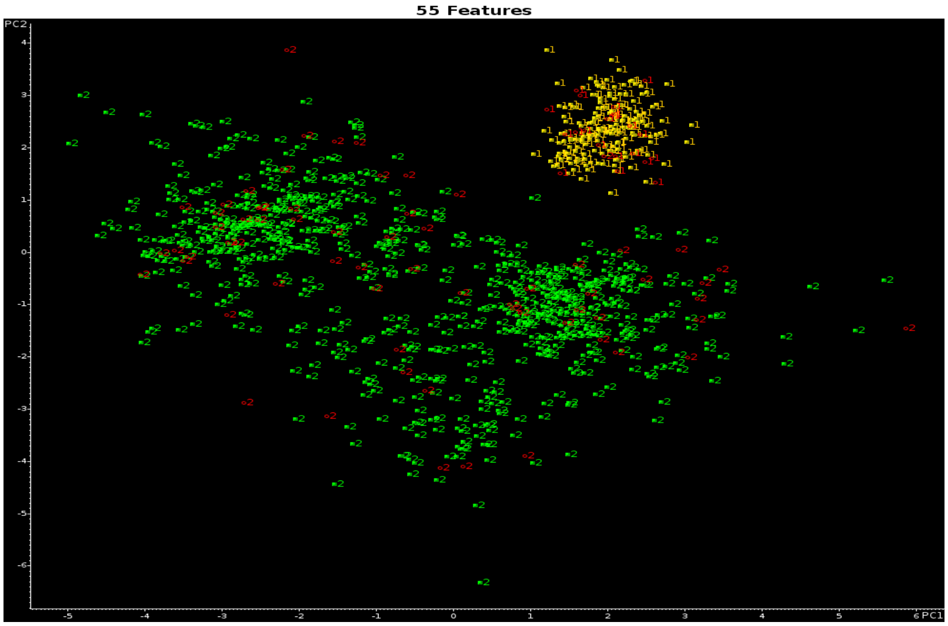


Figure 4.130. Projection of the 87 validation set samples onto the PC plot of the 865 training set samples and the 55 wavelet coefficients identified by the pattern recognition GA (1= Toyota , 2= Honda, Nissan, Chrysler group2 and group3, Ford)

The following step was to identify an unknown sample from Chrysler group 2 or not if this sample did not fall in the previous Toyota cluster. The pattern recognition GA identified 29 wavelet coefficients whose PC plot (see Figure 4.131) showed clustering of IR the spectra on the basis of assembly plants from Chrysler group 2 after 162 generations run. To assess the predictive ability of these 29 wavelet coefficients, a validation set of 66 paint samples were projected into 2-PC developed from the 689 training samples and the 29 wavelet coefficients identified by GA using Hopkin 0.1 fitness function of the pattern recognition GA (see Figure 4.132).

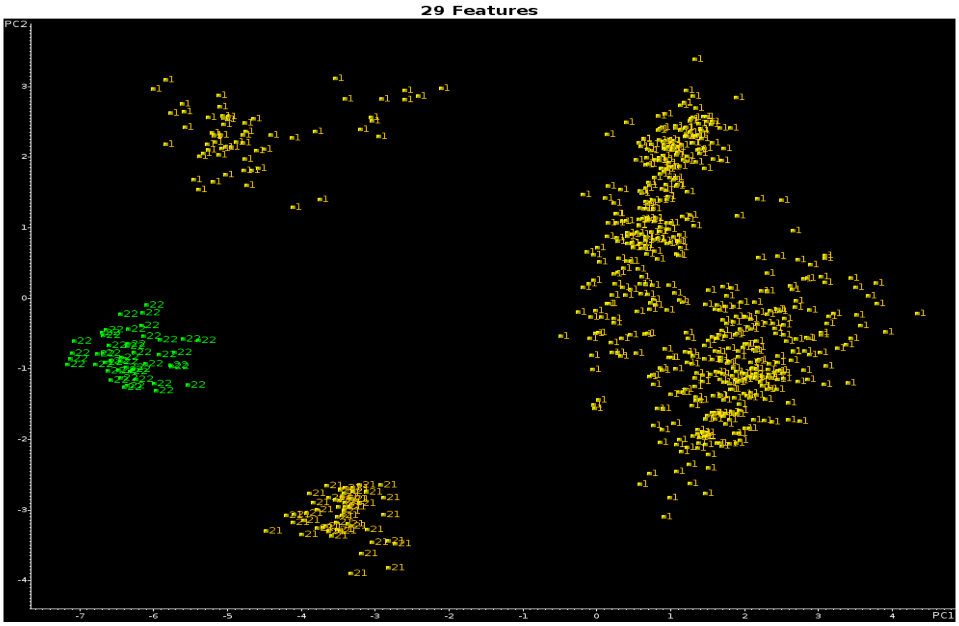


Figure 4.131. 2-PC plot of the 689 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (1= Honda, Nissan, Chrysler group3, Ford, 21= Chrysler group 2(1), 22 = Chrysler group 2(2))

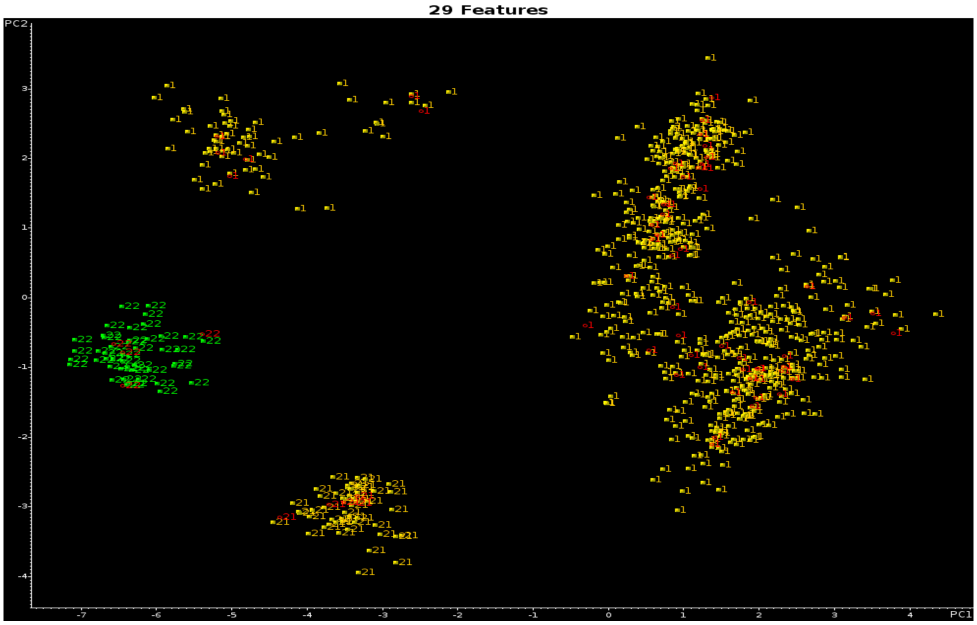


Figure 4.132. Projection of the 66 validation set samples onto the PC plot of the 689 training set samples and the 29 wavelet coefficients identified by the pattern recognition GA (1= Honda, Nissan, Chrysler group3, Ford, 21= Chrysler group 2(1), 22 = Chrysler group 2(2))

This step was to identify an unknown sample from either Honda and Nissan or Ford and Chrysler group 3 if this sample did not fall in the previous Chrysler group 2 cluster. The pattern recognition GA identified 67 wavelet coefficients whose PC plot (see Figure 4.133) showed clustering of IR the spectra on the basis of assembly plants from Honda and Nissan or Ford and Chrysler after 198 generations run. To assess the predictive ability of these 67 wavelet coefficients, a validation set of 57 paint samples were projected into 2-PC developed from the 592 training samples and the 67 wavelet coefficients identified by GA using Hopkin 0.1 fitness function of the pattern recognition GA (see Figure 4.134).

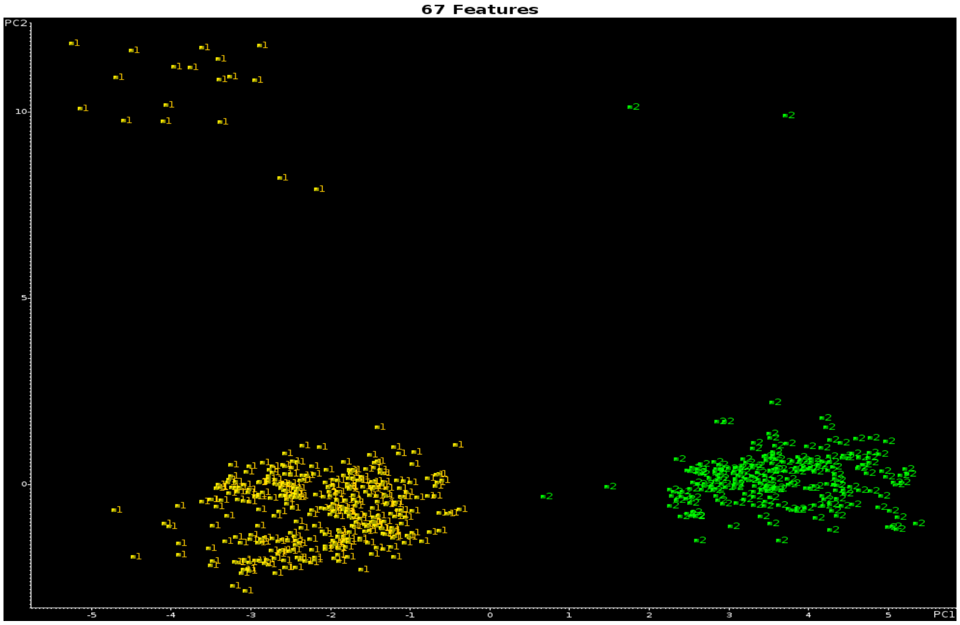


Figure 4.133. 2-PC plot of the 592 training set samples and the 67 wavelet coefficients identified by the pattern recognition GA (1= Chrysler group3, Ford, 2= Honda, Nissan)

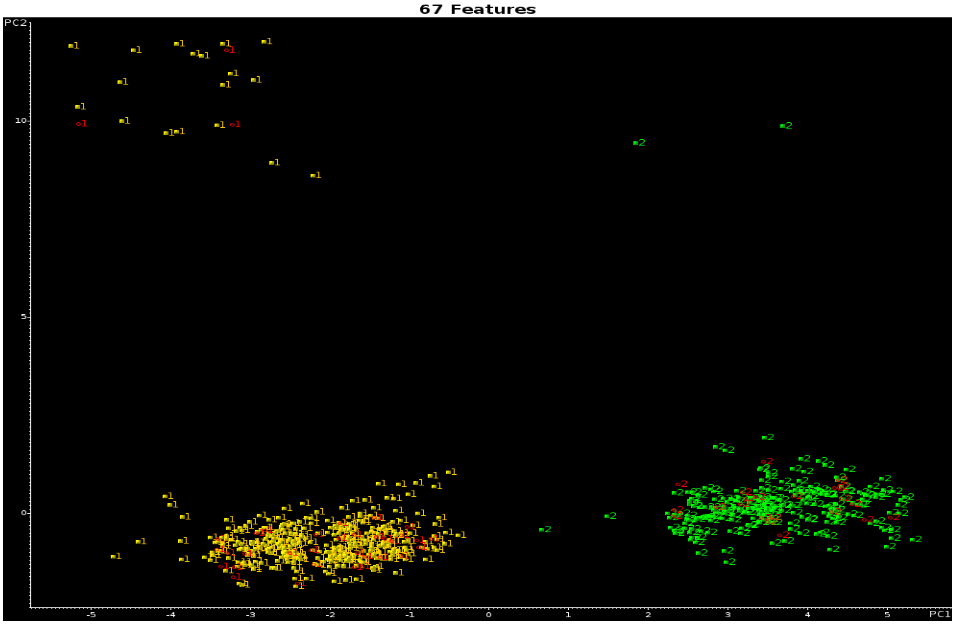


Figure 4.134. Projection of the 57 validation set samples onto the PC plot of the 592 training set samples and the 67 wavelet coefficients identified by the pattern recognition GA (1= Chrysler group3, Ford, 2= Honda, Nissan)

If the unknown sample fell in the cluster comprising of Honda and Nissan, it went to the final step to identify its manufacturer. The pattern recognition GA identified 35 wavelet coefficients whose PC plot (see Figure 4.135) showed clustering of IR the spectra on the basis of assembly plants from Honda or Nissan after 117 generations run. To assess the predictive ability of these 35 wavelet coefficients, a validation set of 22 paint samples were projected into 2-PC developed from the 225 training samples and the 35 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.136).

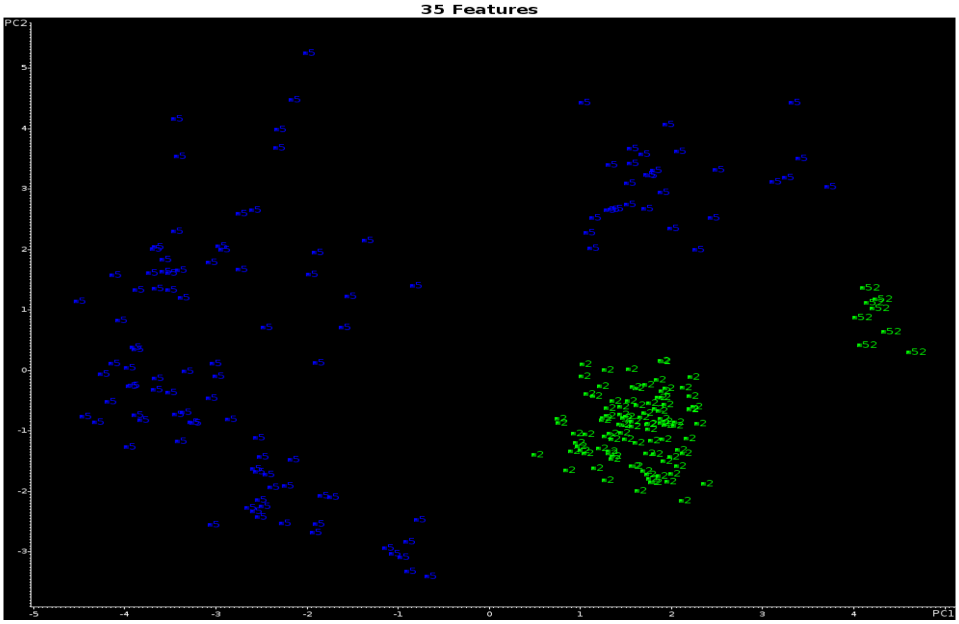


Figure 4.135. 2-PC plot of the 225 training set samples and the 35 wavelet coefficients identified by the pattern recognition GA (2= Honda, 5= Nissan, 52=Nissan)

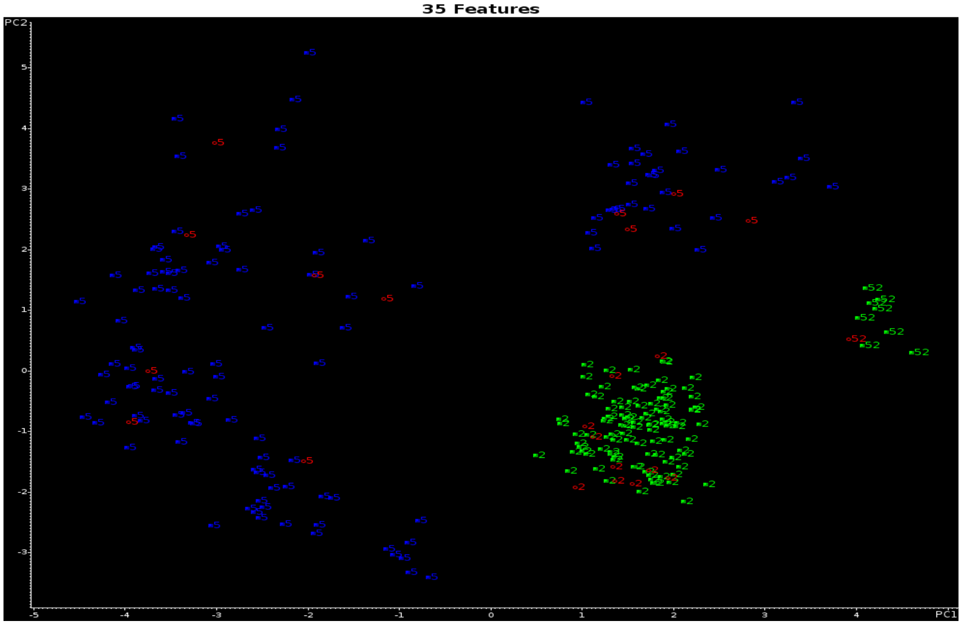


Figure 4.136. Projection of the 22 validation set samples onto the PC plot of the 225 training set samples and the 35 wavelet coefficients identified by the pattern recognition GA (2= Honda, 5= Nissan, 52=Nissan)

If the unknown sample fell in the Ford and Chrysler group 3 cluster, it went to the last step to discern its manufacturer. The pattern recognition GA identified 18 wavelet coefficients whose PC plot (see Figure 4.137) showed clustering of IR the spectra on the basis of assembly plants from Ford or Chrysler after 38 generations run. To assess the predictive ability of these 18 wavelet coefficients, a validation set of 35 paint samples were projected into 2-PC developed from the 367 training samples and the 18 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.138).

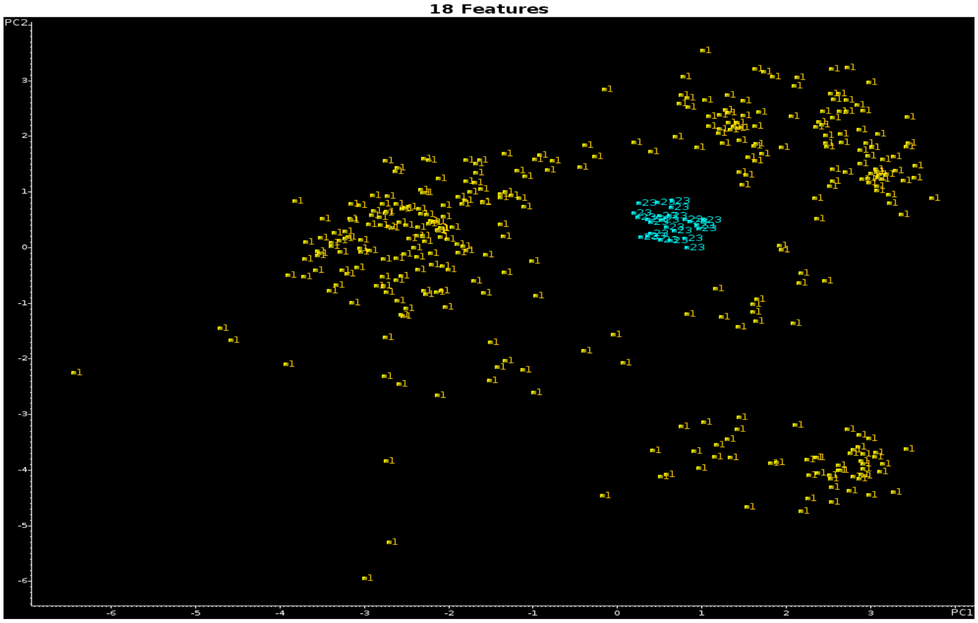


Figure 4.137. 2-PC plot of the 367 training set samples and the 18 wavelet coefficients identified by the pattern recognition GA (1= Ford, 23= Chrysler group 3)

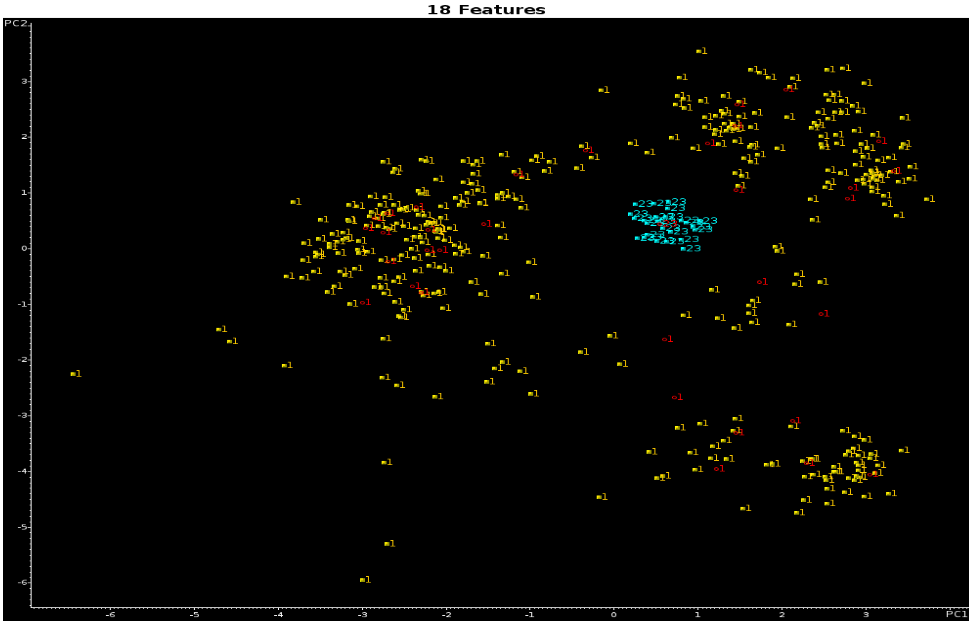


Figure 4.138. Projection of the 35 validation set samples onto the PC plot of the 367 training set samples and the 18 wavelet coefficients identified by the pattern recognition GA(1= Ford, 23= Chrysler group 3)

4.4.7.2 Honda, Nissan and Toyota Plant Group Level Prefilters

Manufacturer assembly plant level search prefilters are needed to differentiate the assembly plant of an unknown sample lying in a particular manufacturer cluster. To fulfil this task, a hierarchical cluster analysis and principal component analysis were employed to identify an automotive paint sample by assembly plant group. Average assembly plant clear coat IR spectra from an individual manufacturer were chosen to do hierarchical cluster analysis and principal component analysis, the results were shown in Figure 4.139 – Figure 4.144. Although the result of hierarchical cluster analysis showed Toyota assembly plant 5004, 5005, 5102 and 5103 in the same plant group cluster; however, the principle component analysis and the average IR spectra from Toyota assembly plant 5004, 5005, 5102, 5103, 5002, 5203 and 5007 (Figure 4.145) suggested that the assembly plant

5007 should group with Toyota assembly plant 5004, 5005, 5102 and 5103. The results of the principal component analysis and the hierarchical cluster analysis suggested to group 9 assembly plants and sub plants of Honda into three plant groups (see Table 4.1). By the same way, 10 Nissan assembly plants and sub assembly plants were divided into 4 assembly plant groups. Each plant group was assumed that the chemical composition of a clear coat was similar. The plant group information of Nissan and Toyota are also listed in Table 4.1.

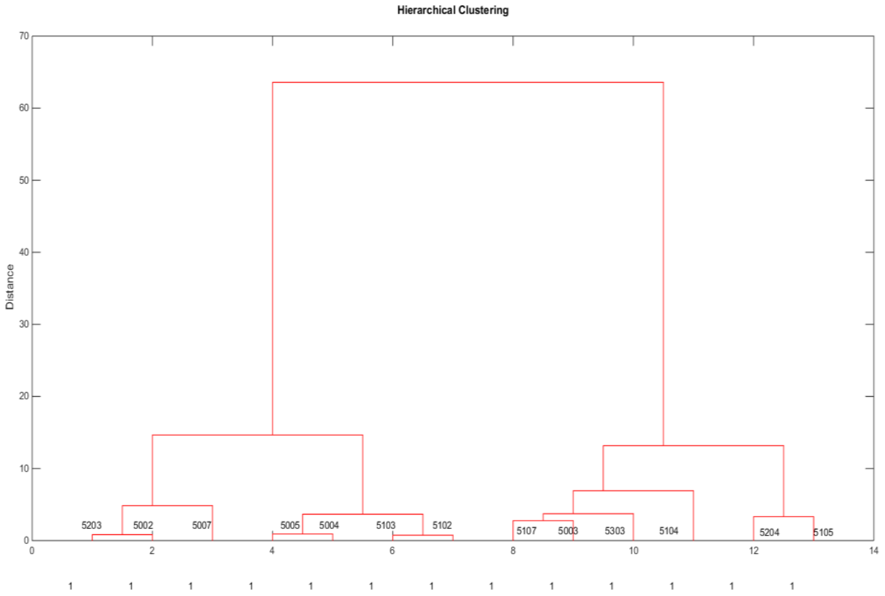


Figure 4.139. Toyota assembly plant hierarchical cluster analysis

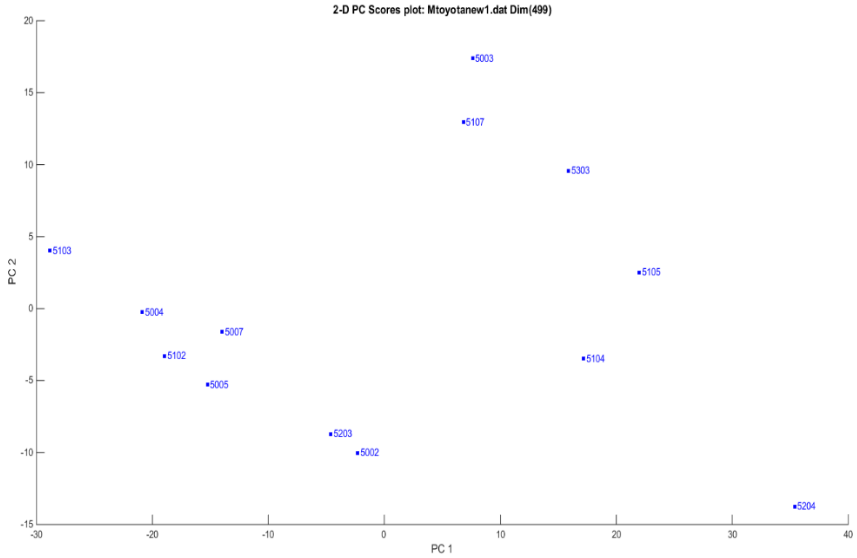


Figure 4.140. Toyota assembly plant principal component analysis

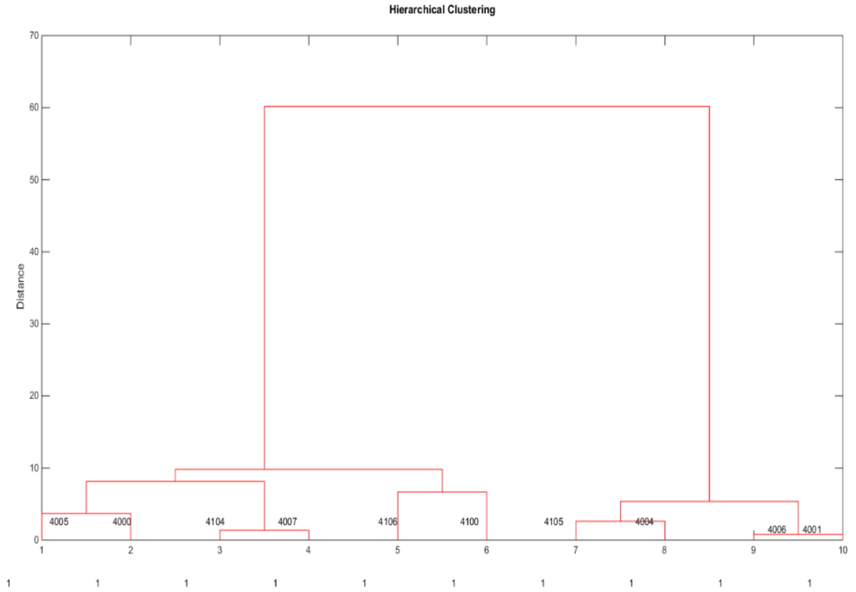


Figure 4.141. Nissan assembly plant hierarchical cluster analysis

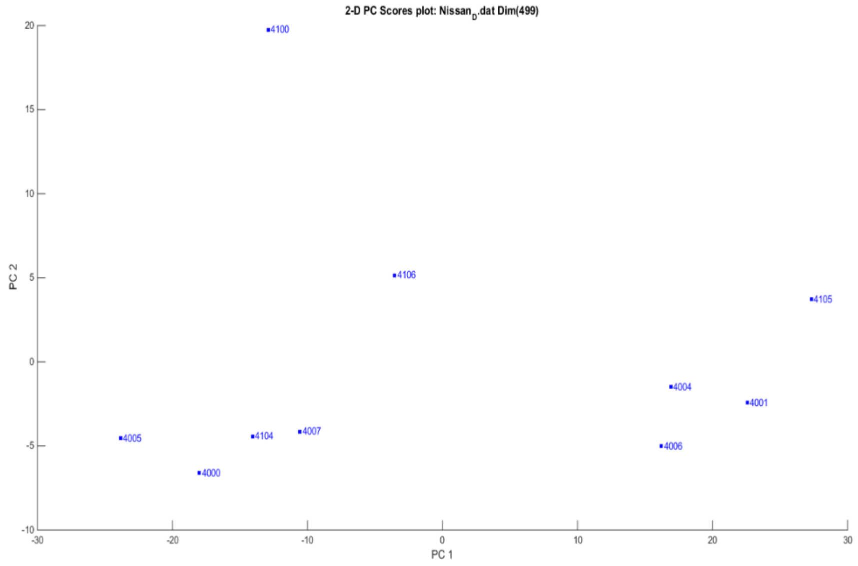


Figure 4.142. Nissan assembly plant principal component analysis

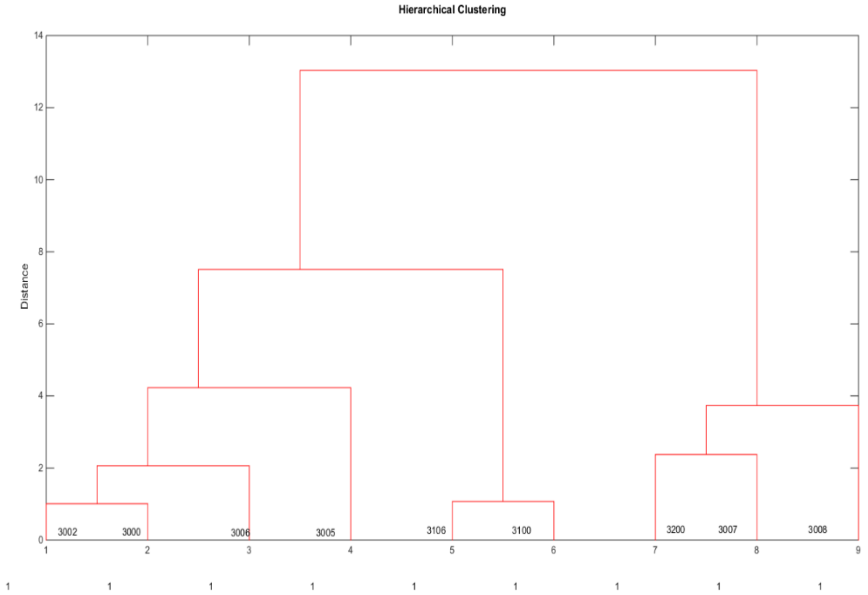


Figure 4.143. Honda assembly plant hierarchical cluster analysis

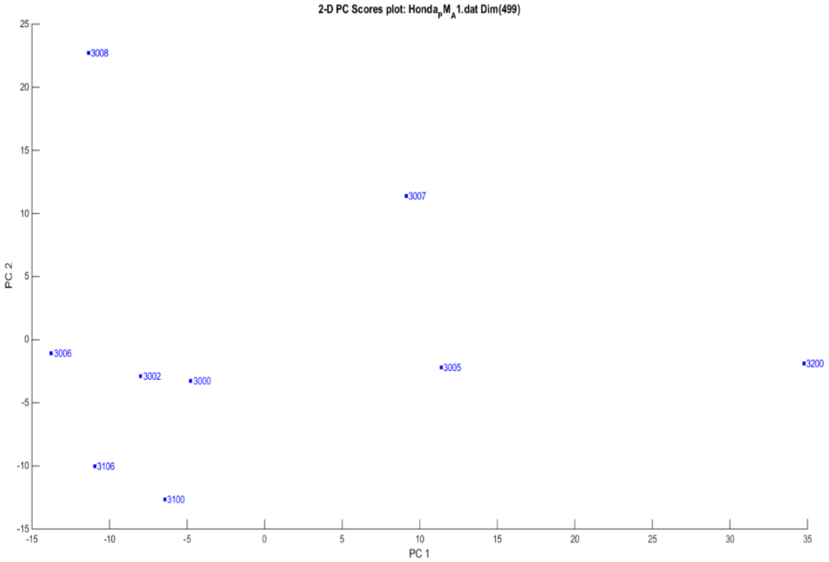


Figure 4.144. Honda assembly plant principle component analysis

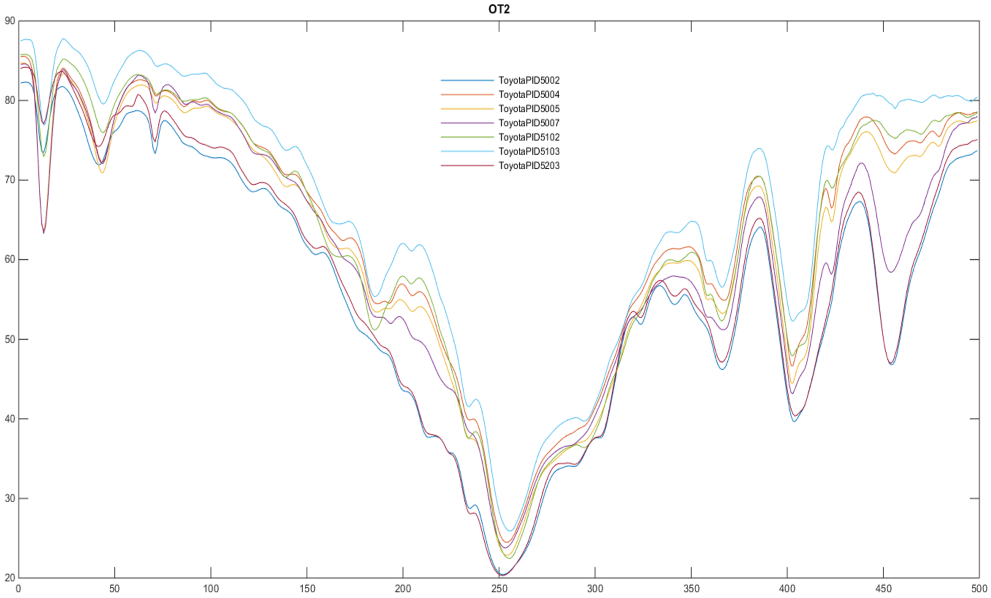


Figure 4.145. The average IR spectra of Toyota assembly plants

Table 4.19 Assembly plant group information in the basis of manufacturer

Manufacturer	Assigned group number	Plant IDs
Honda	1	3000,3002,3005,3006
	2	3100,3106
	3	3007,3008,3200
Nissan	1	4000,4005,4007,4104
	2	4100,4106
	3	4004,4105
	4	4001,4006
Toyota	1	5004,5005,5007,5102,5103
	2	5002,5203
	3	5003,5104,5303
	4	5105,5204

A genetic algorithm for feature selection and pattern recognition analysis was applied to identify assembly plant groups for single carbonyl band samples in the basis of manufacturer Honda, Nissan and Toyota. After 102 generations, the pattern recognition GA identified 24 wavelet coefficients whose 2-PC plot exhibited clustering of the clear coat IR spectra on the basis of assembly plant group of Toyota by using Hopkins 0.1 fitness function and removing outliers whose sample ID are 5146, 5238, 5265, 5281, 5298 and 5304. 175 training set samples and the 24 wavelet coefficients identified by the pattern recognition GA. To assess the predictive ability of these 24 wavelet coefficients, a validation set of 23 paint samples were projected into 2-PC developed from the 175 training set and the wavelet coefficients identified by GA. Each validation set sample lies in a correct region of the PC plot associate to Toyota plant group. The 2-PC plots for the training and validation set of Toyota were seen in Figure 4.146-Figure 4.147.

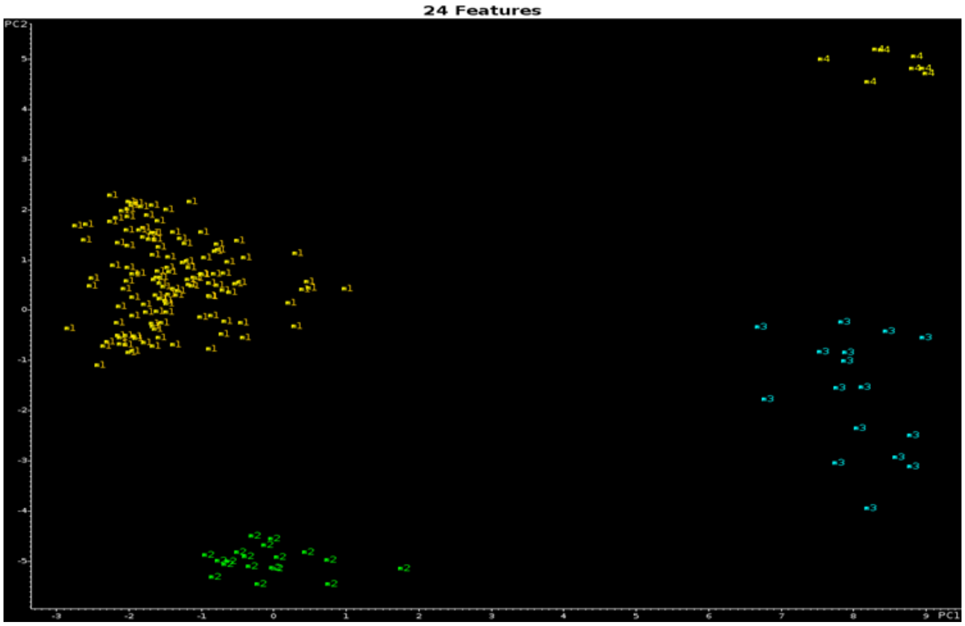


Figure 4.146. 2-PC plot of the 175 training set samples and the 24 wavelet coefficients identified by the pattern recognition GA (Toyota: 1= group1, 2=group2, 3=group3, 4=group4)

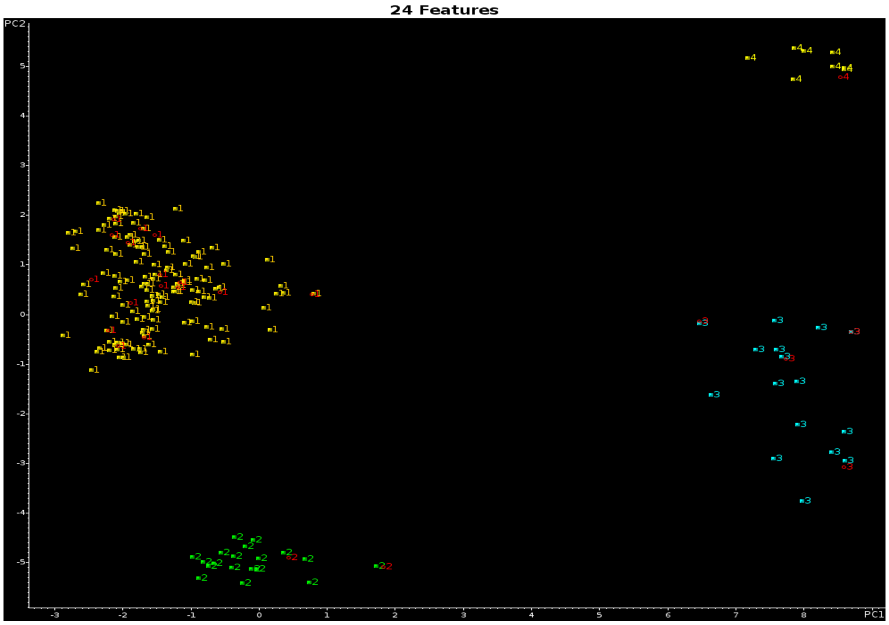


Figure 4.147. Projection of the 23 validation set samples onto the PC plot of the 175 training set samples and the 24 wavelet coefficients identified by the pattern recognition GA (Toyota: 1= group1, 2=group2, 3=group3, 4=group4)

After 10 generations, the pattern recognition GA identified 8 wavelet coefficients whose 2-PC plot exhibited clustering of the clear coat IR spectra on the basis of assembly plant group of Nissan. 122 training set samples and the 8 wavelet coefficients identified by the pattern recognition GA using the Normal fitness function and remove one outlier (SID: 4169). To assess the predictive ability of these 8 wavelet coefficients, a validation set of 14 paint samples were projected into 2-PC developed from the 122 training set and the wavelet coefficients identified by GA. The 2-PC plots for the training and validation set of Nissan were seen in Figure 4.148-Figure 4.149.

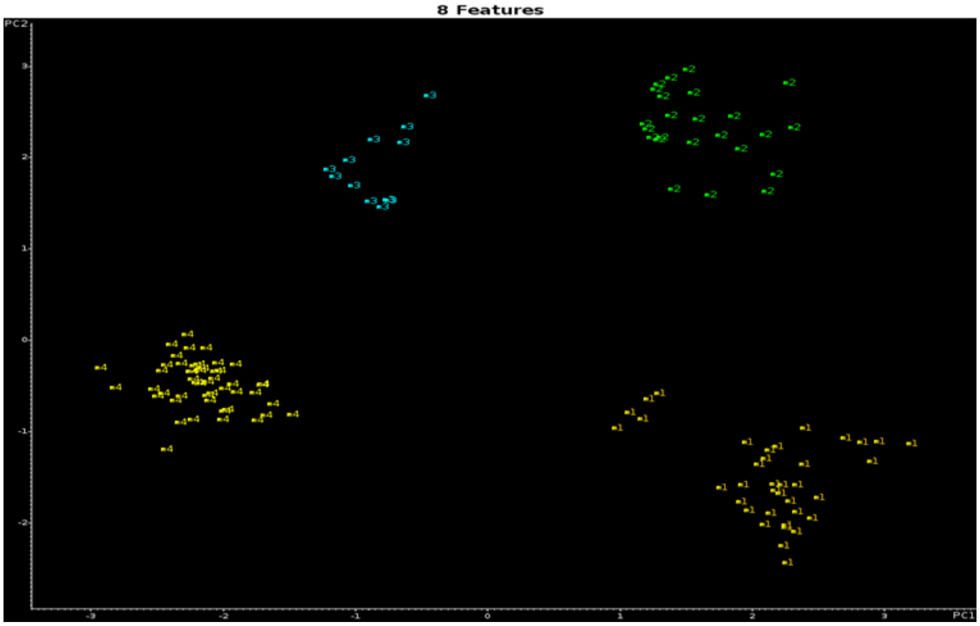


Figure 4.148. 2-PC plot of the 122 training set samples and the 8 wavelet coefficients identified by the pattern recognition GA (Nissan: 1= group1, 2=group2, 3=group3, 4=group4)

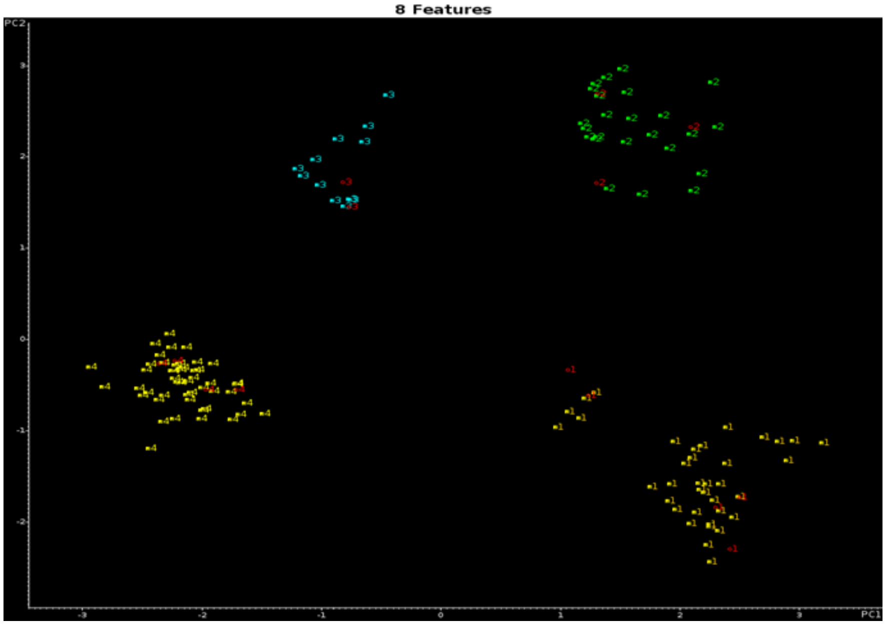


Figure 4.149. Projection of the 14 validation set samples onto the PC plot of the 122 training set samples and the 8 wavelet coefficients identified by the pattern recognition GA (Nissan: 1= group1, 2=group2, 3=group3, 4=group4)

For the prefilter used to differentiate the assembly plant group of Honda, the pattern recognition GA identified 20 wavelet coefficients whose 2-PC plot exhibited clustering of the clear coat IR spectra on the basis of Honda assembly plant group by removing an outlier (SID: 3126) after 40 generations,. To assess the predictive ability of these 20 wavelet coefficients, a validation set of 12 paint samples were projected into 2-PC developed from the 99 training set and the wavelet coefficients identified by GA, the 2-PC plots for the training and validation set of Honda were seen in Figure 4.150-Figure 4.151. This results suggests that some assembly plants of a specific automotive manufacturer have the similar paint formulation in a clear coat.

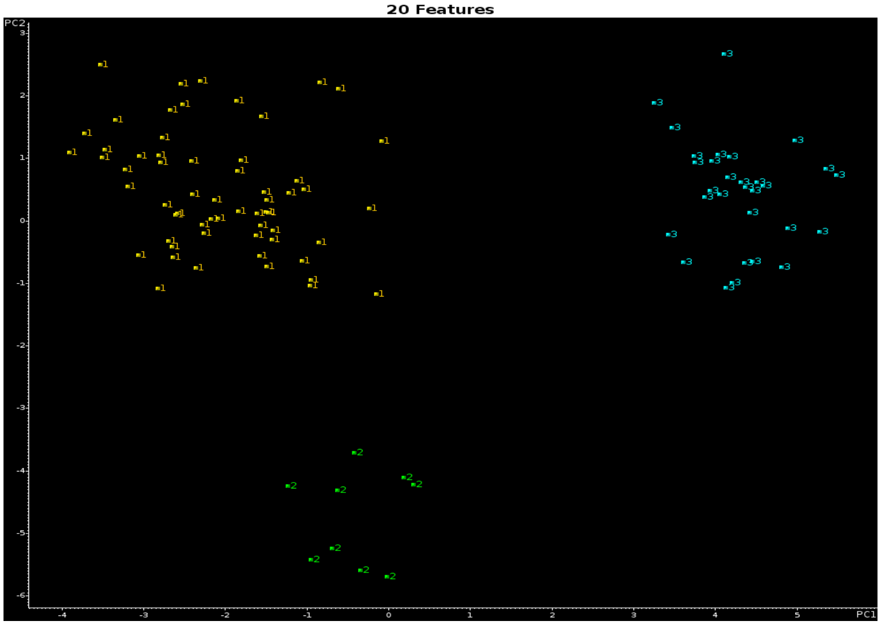


Figure 4.150. 2-PC plot of the 99 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (Honda: 1= group1, 2=group2, 3=group3)

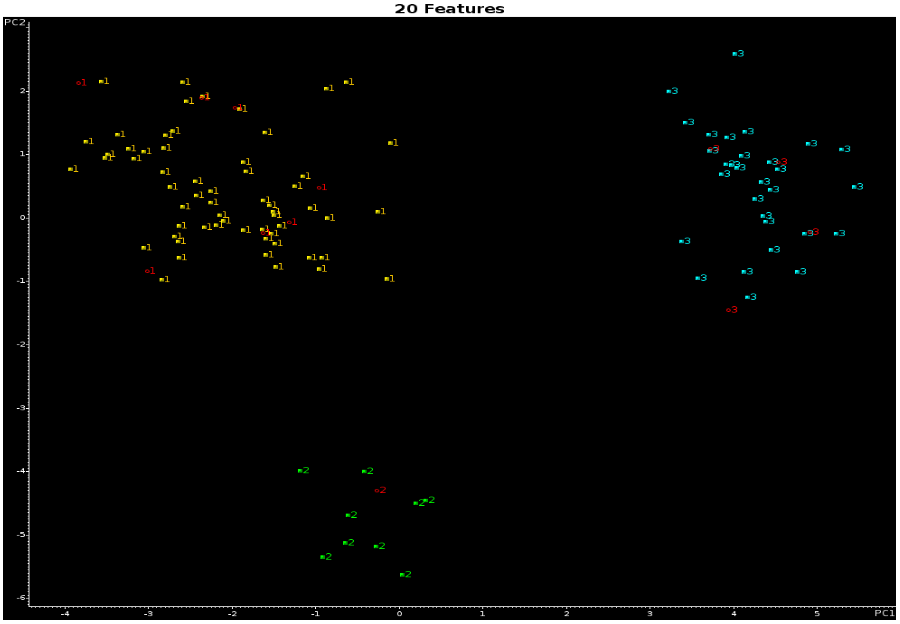


Figure 4.151. Projection of the 12 validation set samples onto the PC plot of the 99 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (Honda: 1= group1, 2=group2, 3=group3)

4.4.7.3 Honda, Nissan and Toyota Assembly Plant Level Prefilter

After assigning a sample with manufacturer plant group membership, the next step will find the membership of assembly plant for this sample. . “8sym6”wavelet preprocessing data from the clear coat horizontally concatenated “8sym6”wavelet preprocessing data from both two undercoats. This fused IR data were used for developing assembly plant level prefilters in the basis of each manufacturer.

4.4.7.3.1 Honda Assembly Plant Level Prefilters

To identify whether the unknown is from assembly plant 3002 or from the rest all assembly plants (PID 3856: merging PID 3000, 3005 and 3006), pattern recognition GA (Fitness function: normal) identified 7 wavelet coefficients whose PC plot (see Figure 4.152) showed clustering of the fused IR spectra on the basis of assembly plant 3002 in Honda plant group 1 after 18 generation runs. The 7 validation samples were projected onto the PCs (see Figure 4.153) define by the 59 training samples and the 7 wavelet coefficients identified by the pattern recognition GA. Each validation set sample lies in a right assembly plant 3002 region of the PC plot.

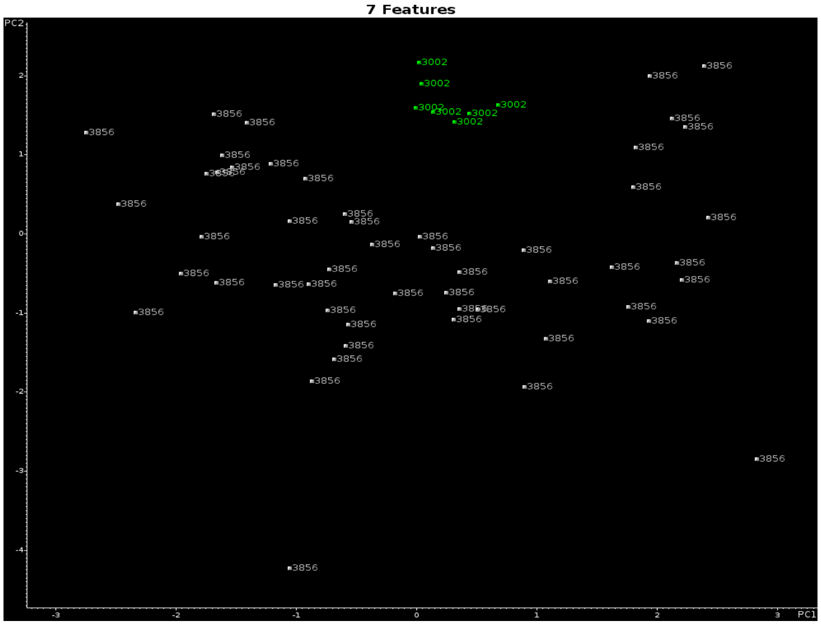


Figure 4.152. 2-PC plot of the 59 training set samples and the 7 wavelet coefficients identified by the pattern recognition GA (3002= East Liberty, OH, USA, 3856= Alliston, ON, Canada; Lincoln, Alabama; Marysville, OH, USA)

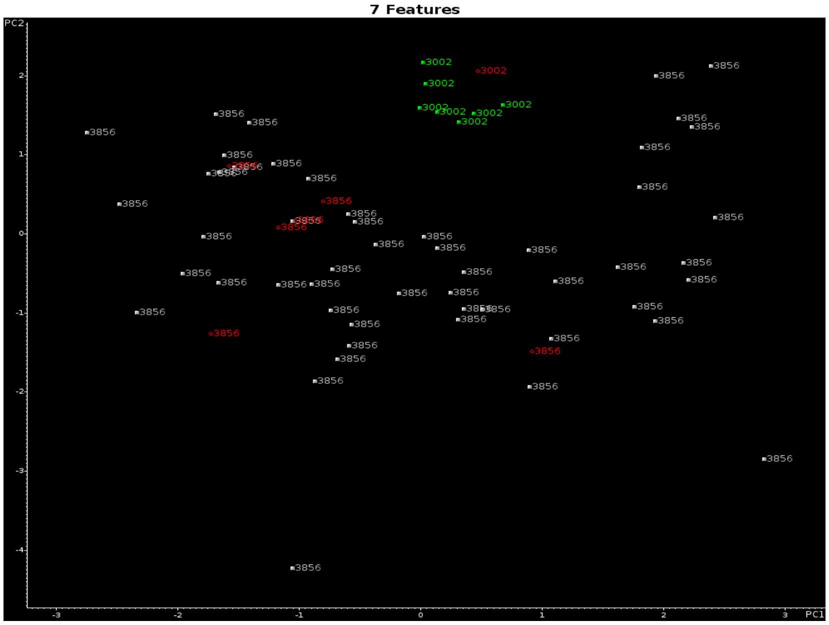


Figure 4.153. Projection of the 7 validation set samples onto the PC plot of the 59 training set samples and the 7 wavelet coefficients identified by the pattern recognition GA (3002= East Liberty, OH, USA, 3856= Alliston, ON, Canada; Lincoln, Alabama; Marysville, OH, USA)

If the validation sample falls out of the assembly plant 3002 region, pattern recognition GA will be used to further identify the assembly plant of this sample in Honda plant group 1. After 200 generations, pattern recognition GA (Fitness function: normal) identified 46 wavelet coefficients whose PCs (see Figure 4.154) showed clustering of the fused IR spectra on the basis of assembly plant in Honda plant group 1 except the assembly plant 3002. The 6 validation samples were then projected onto the PCs (see Figure 4.155) define by the 52 training samples and the 46 wavelet coefficients identified by the pattern recognition GA. Each validation set sample lies in a right assembly plant region of the PC plot.

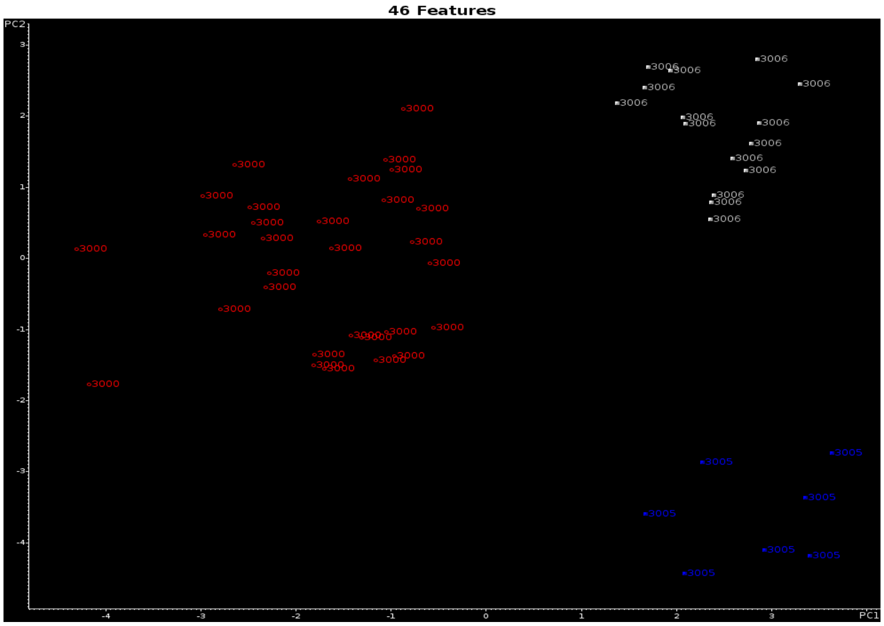


Figure 4.154. 2-PC plot of the 52 training set samples and the 46 wavelet coefficients identified by the pattern recognition GA (3000= Alliston, ON, Canada, USA, 3005= Lincoln, Alabama, 3006= Marysville, OH, USA)

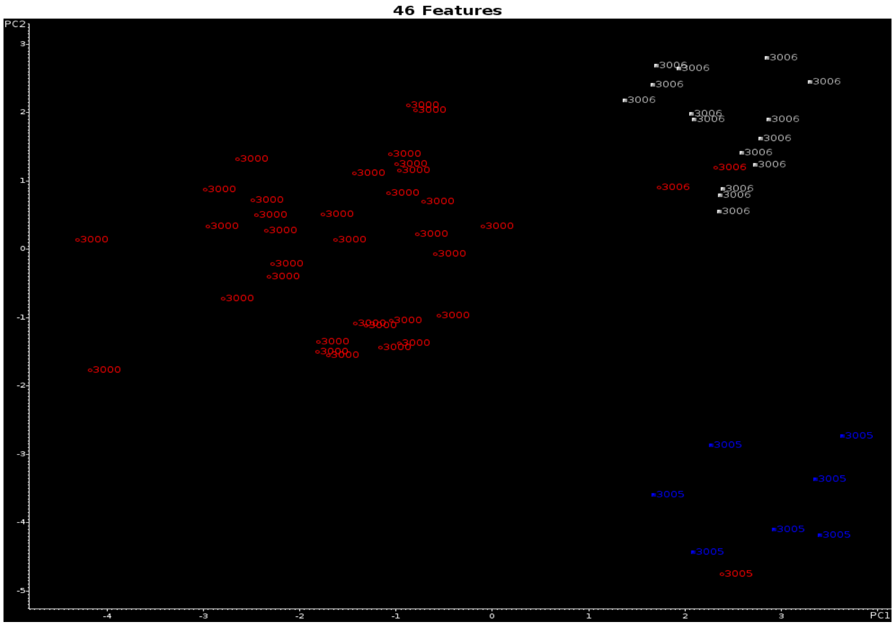


Figure 4.155. Projection of the 6 validation set samples onto the PC plot of the 52 training set samples and the 46 wavelet coefficients identified by the pattern recognition GA (3000= Alliston, ON, Canada, USA, 3005= Lincoln, Alabama, 3006= Marysville, OH, USA)

The pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.156) showed clustering of IR the spectra on the basis of assembly plants from Honda group 2 after 1 generation run. To assess the predictive ability of these 2 wavelet coefficients, a validation set of 1 paint sample were projected into 2-PCs developed from the 9 training set and the 2 wavelet coefficients identified by GA. The pattern recognition GA ran by using Normal fitness function (see Figure 4.157). Samples from Honda assembly plant Alliston are too less to be used for prediction.

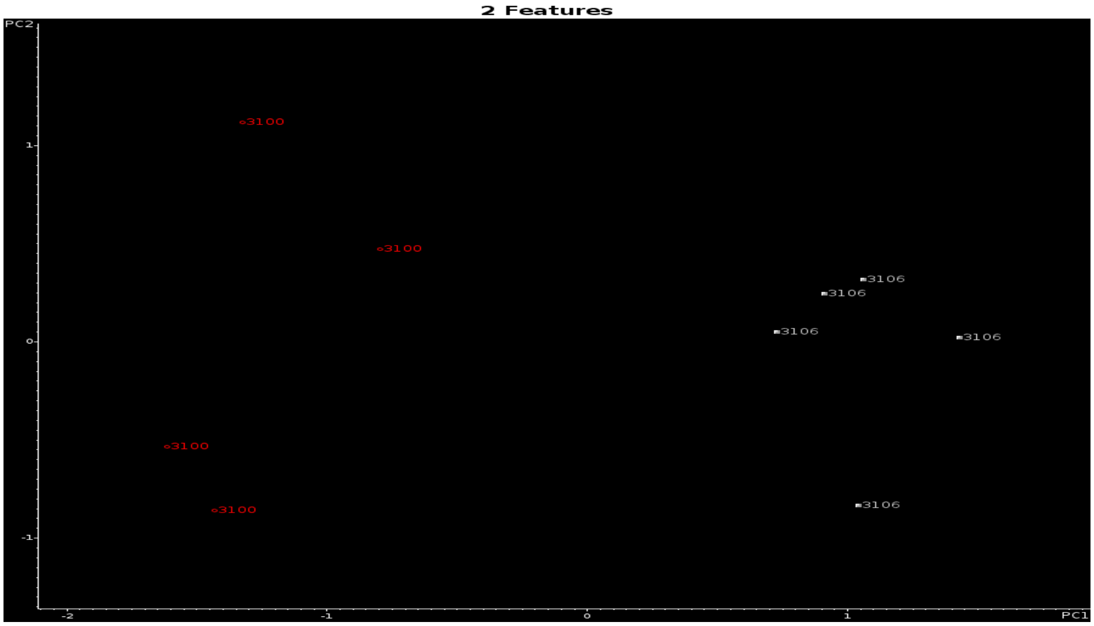


Figure 4.156. 2-PC plot of the 9 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (3100= Alliston, 3106= Marysville)

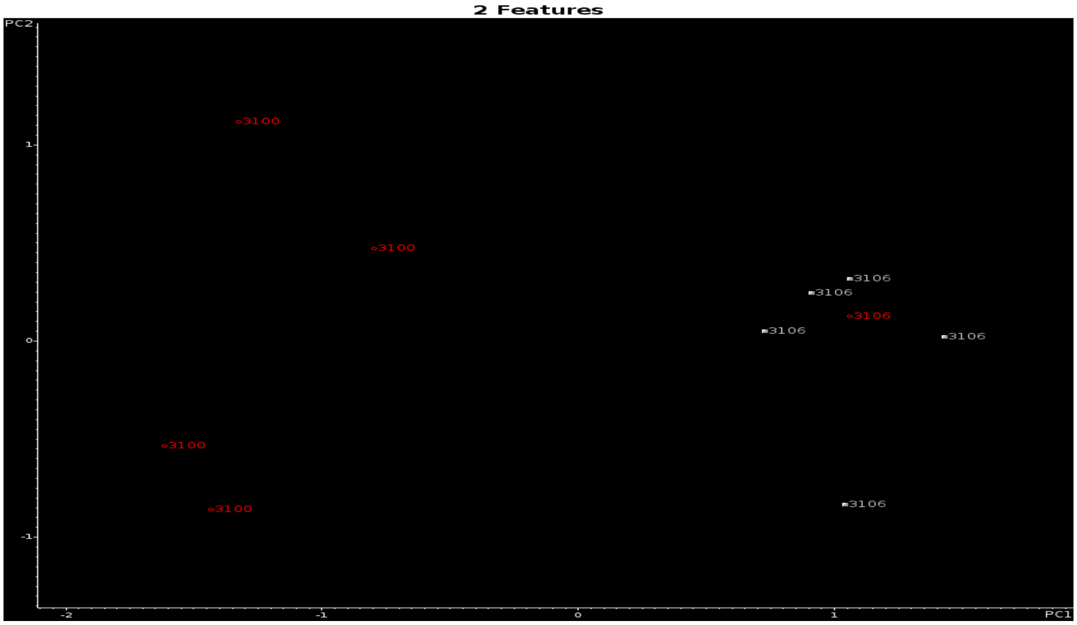


Figure 4.157. Projection of the 1 validation set samples onto the PC plot of the 9 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA(3100= Alliston, 3106= Marysville)

For the prediction of a sample falling in Honda assembly plant group 3, the situation was complicated. After 52 generations, the pattern recognition GA identified 20 wavelet coefficients whose PC plot (see Figure 4.158) showed clustering of the spectra on the basis of assembly plant in Honda plant group 3. To assess the predictive ability of these 20 wavelet coefficients, a validation set of 4 paint samples were projected into 2-PCs developed from the 30 training set and the wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA. The validation set sample from Honda assembly plant 3007 was close to the one belonging to the assembly plant 3008 (see Figure 4.159). Even if the fitness function is switched to Hopkins 0.1 and Mehual 0.1, the model failed in predicting samples between the assembly plant 3007 and 3008. The individual average IR spectra from the assembly plant 3007 and the assembly plant 3008 were compared (Figure 4.160), and the result suggested to merge both two assembly plants together. After 4 generations, the pattern recognition GA identified 20 wavelet coefficients whose PC plot (see Figure 4.161) showed clustering of the spectra on the basis of assembly plant in Honda plant group 3. To assess the predictive ability of these 20 wavelet coefficients, a validation set of 4 paint samples were projected into 2-PC developed from the 30 training set and the wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA. The validation set sample from the assembly plant 3007 was close to the one belonging to the assembly plant 3008 (see Figure 4.162).

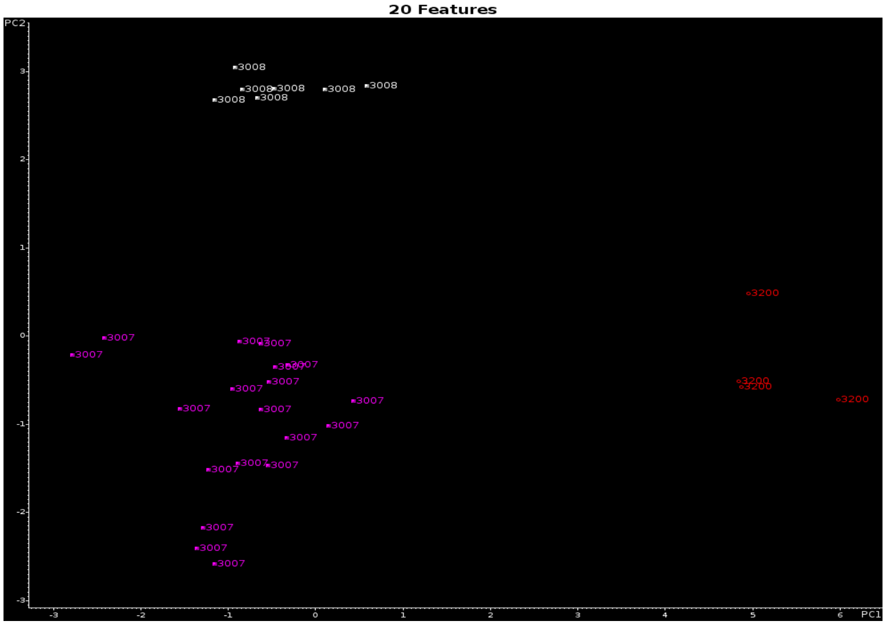


Figure 4.158. 2-PC plot of the 30 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (3007=Sayama, 3008=Suzuka, 3200=Alliston)

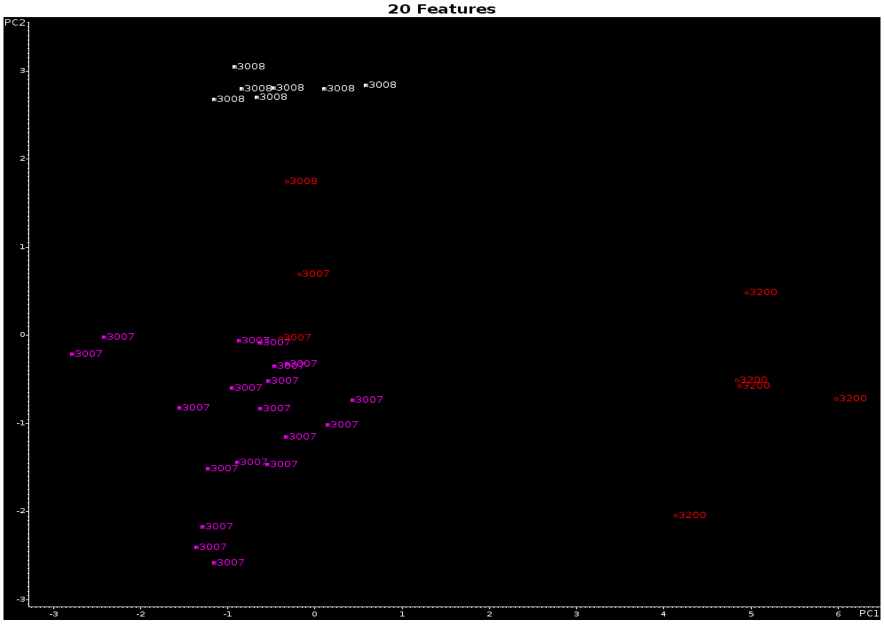


Figure 4.159. Projection of the 4 validation set samples onto the PC plot of the 30 training set samples and the 20 wavelet coefficients identified by the pattern recognition GA (3007=Sayama, 3008=Suzuka, 3200=Alliston)

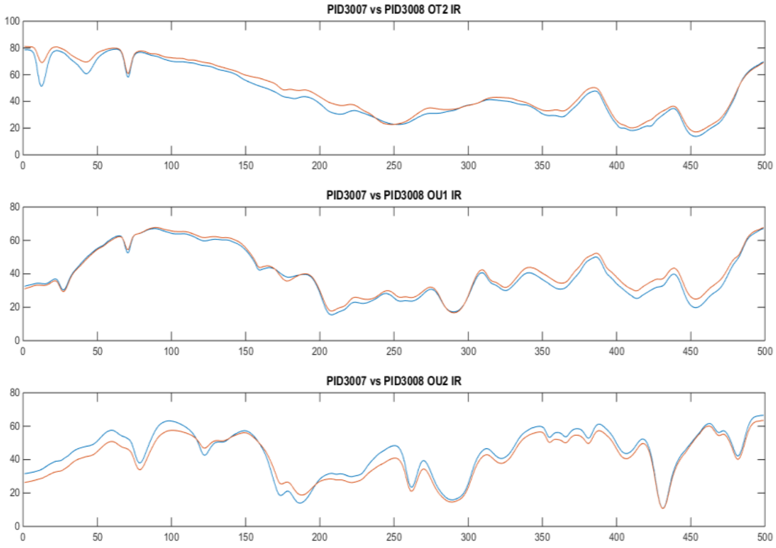


Figure 4.160. The comparison of the average IR spectra of the assembly plant Sayama (3007) vs the assembly plant Suzuka (3008)

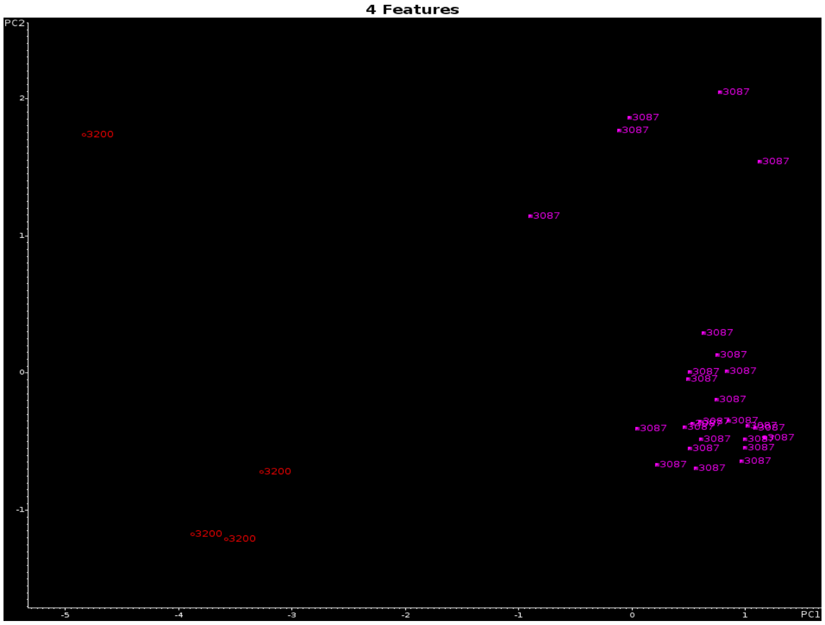


Figure 4.161. 2-PC plot of the 30 training set samples and the 4 wavelet coefficients identified by the pattern recognition GA (3087=Sayama, Suzuka, 3200=Alliston)

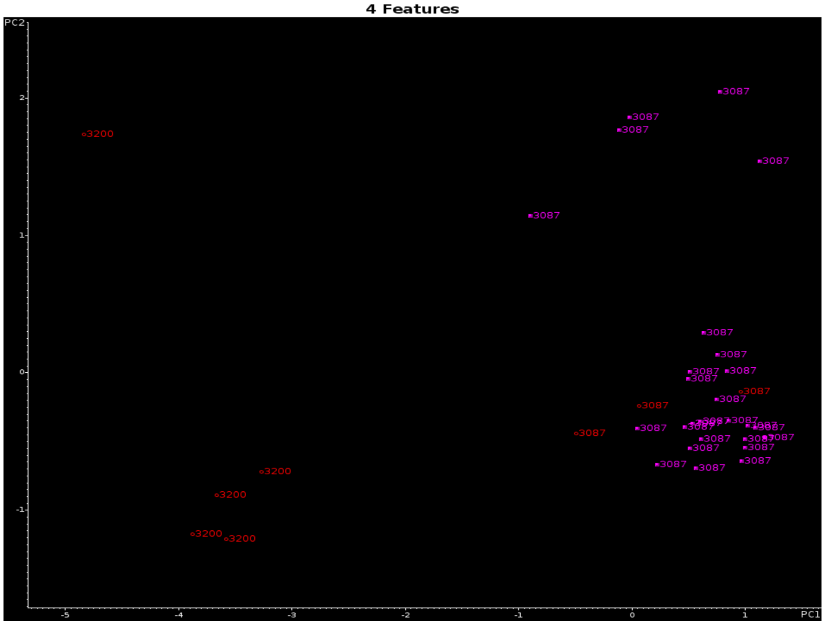


Figure 4.162. Projection of the 4 validation set samples onto the PC plot of the 30 training set samples and the 4 wavelet coefficients identified by the pattern recognition GA (3087=Sayama, Suzuka, 3200=Alliston)

4.4.7.3.2 Nissan Assembly Plant Level Prefilters

To identify the unknown sample membership in Nissan assembly plant group 1, pattern recognition GA identified 15 wavelet coefficients whose PC plot (see Figure 4.163) showed clustering of the fused IR spectra on the basis of assembly plant in Nissan plant group 1 after 41 generations. The 5 validation samples were then projected onto the PCs (see Figure 4.164) define by the 37 training samples and the 15 wavelet coefficients identified by the pattern recognition GA. The GA fitness function is Normal.

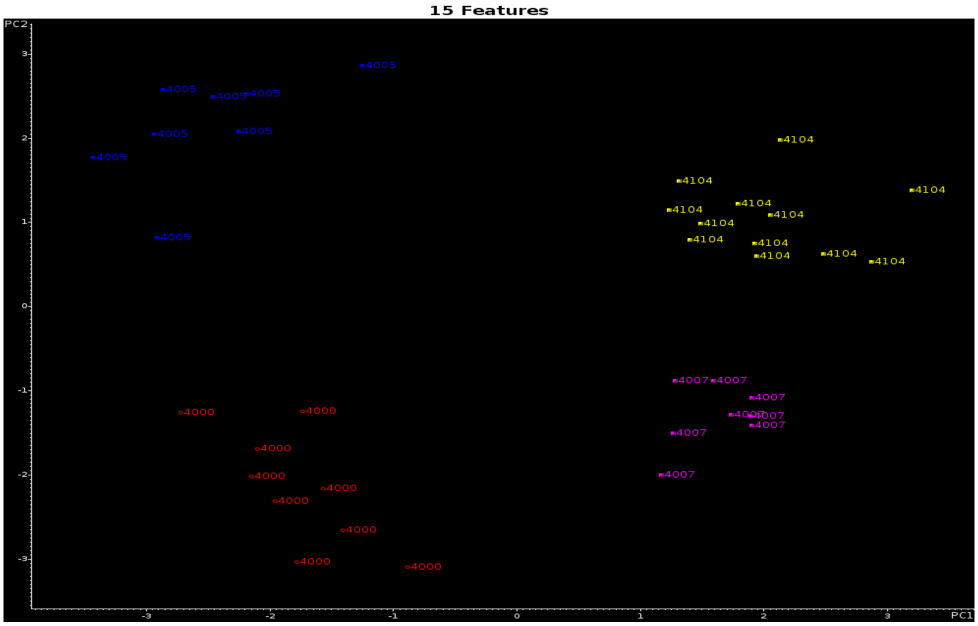


Figure 4.163 2-PC plot of the 37 training set samples and the 15 wavelet coefficients identified by the pattern recognition GA (4000= Aguascalientes, 4005=Oppama, 4007=Tochigi, 4104=Kyushu)

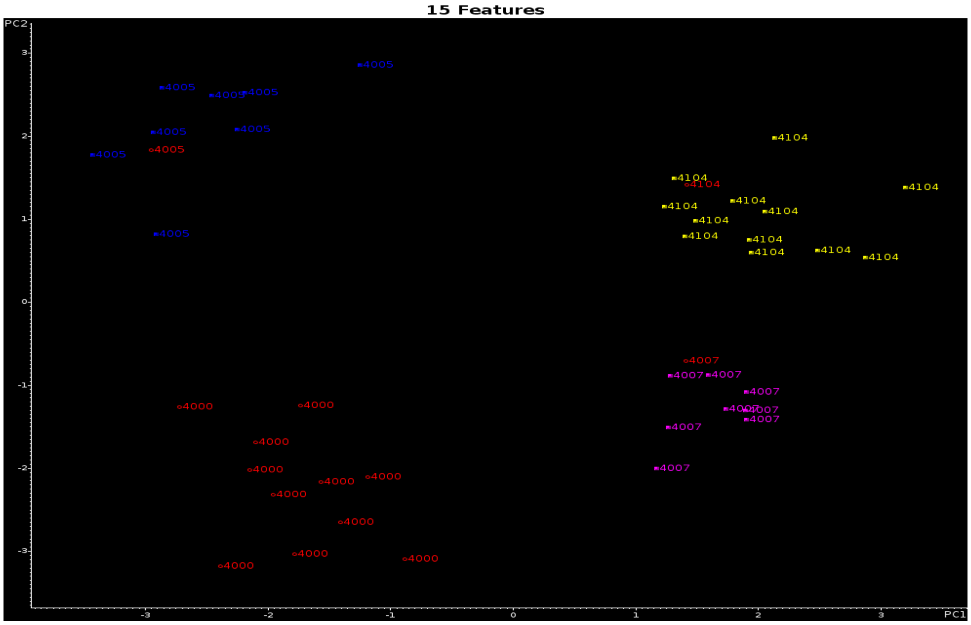


Figure 4.164 Projection of the 5 validation set samples onto the PC plot of the 37 training set samples and the 15 wavelet coefficients identified by the pattern recognition GA (4000= Aguascalientes, 4005=Oppama, 4007=Tochigi, 4104=Kyushu)

To identify the unknown sample membership in Nissan assembly plant group 2, pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.165) showed clustering of the fused IR spectra on the basis of assembly plant in Nissan plant group 2 after 2 generations. The 3 validation samples were projected onto the PCs (see Figure 4.166) define by the 24 training samples and the 2 wavelet coefficients identified by the pattern recognition GA. The fitness function of the pattern recognition GA is Normal.

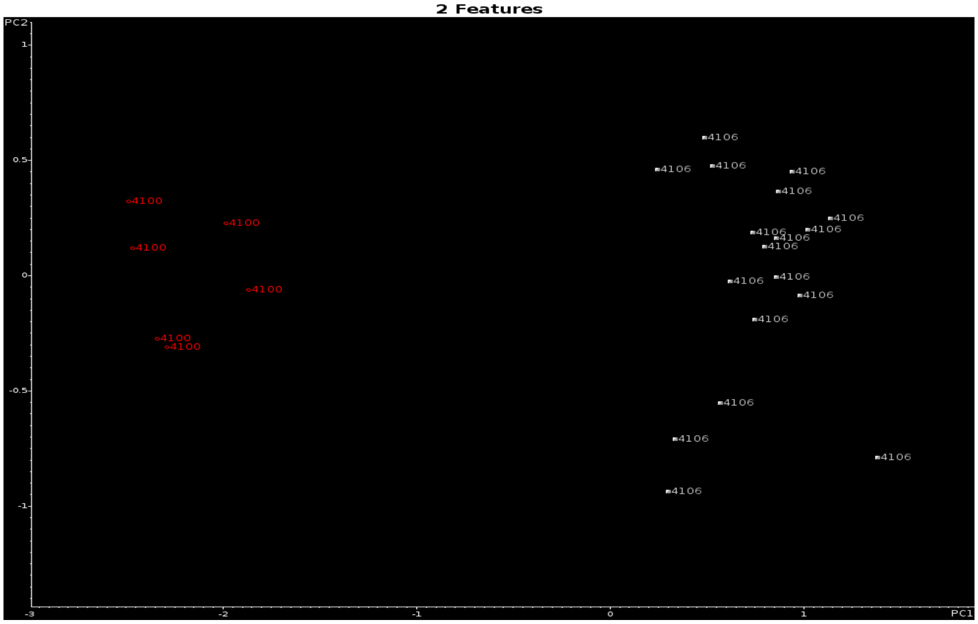


Figure 4.165 2-PC plot of the 24 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4100= Aguascalientes, 4106= Smyrna)

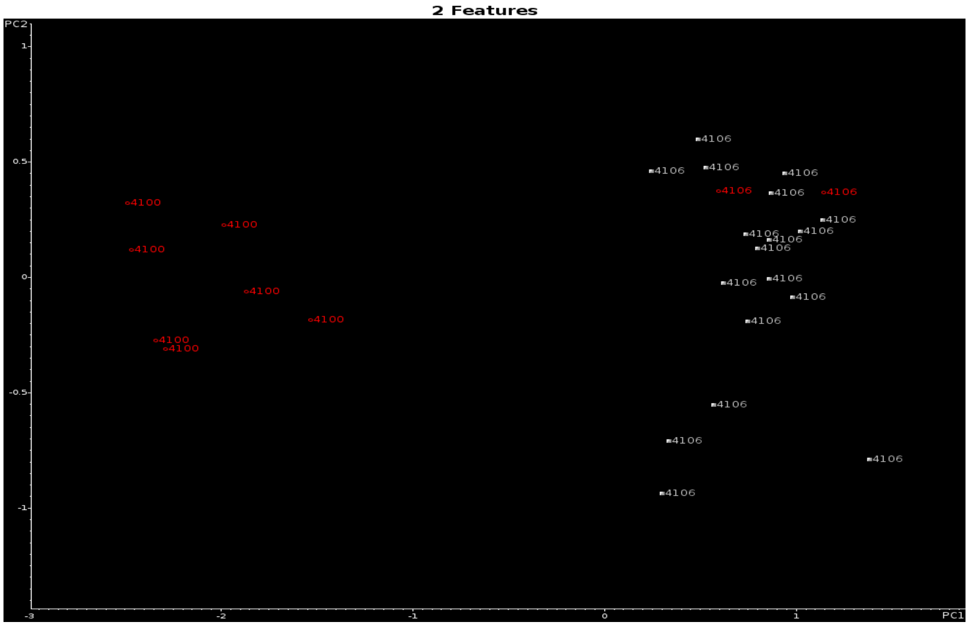


Figure 4.166 Projection of the 3 validation set samples onto the PC plot of the 24 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4100= Aguascalientes, 4106= Smyrna)

For samples located in the Nissan group 3, the pattern recognition GA identified 3 wavelet coefficients whose PC plot (see Figure 4.167) showed clustering of IR the spectra on the basis of assembly plants from Nissan group 3 after 2 generation run. To assess the predictive ability of these 3 wavelet coefficients, a validation set of 2 paint samples were projected into 2-PCs developed from the 12 training samples and the 3 wavelet coefficients identified by GA using normal fitness function of the pattern recognition GA (see Figure 4.168).

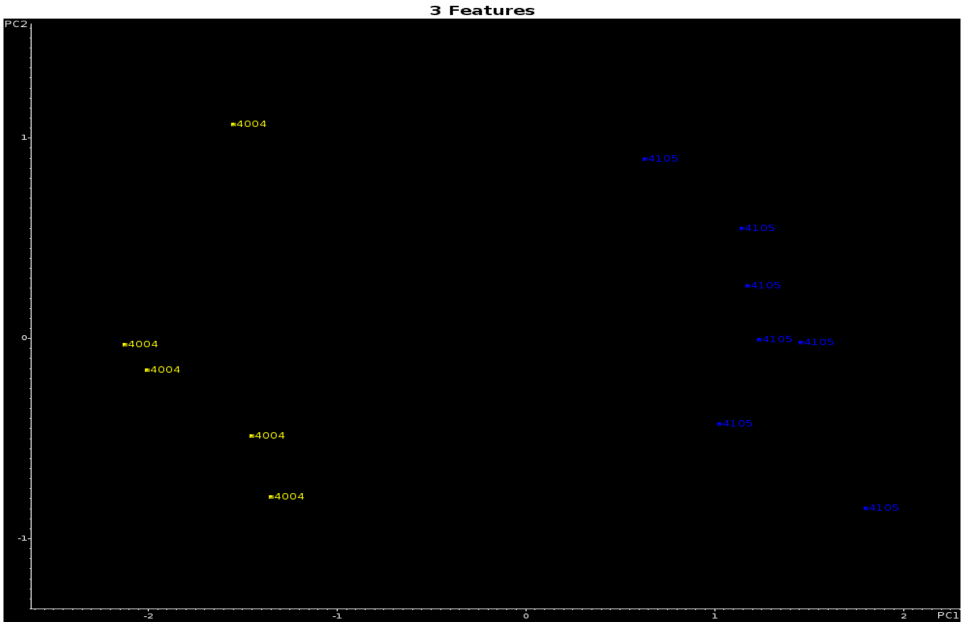


Figure 4.167 2-PC plot of the 12 training set samples and the 3 wavelet coefficients identified by the pattern recognition GA (4004=Kyushu, 4105=Oppama)

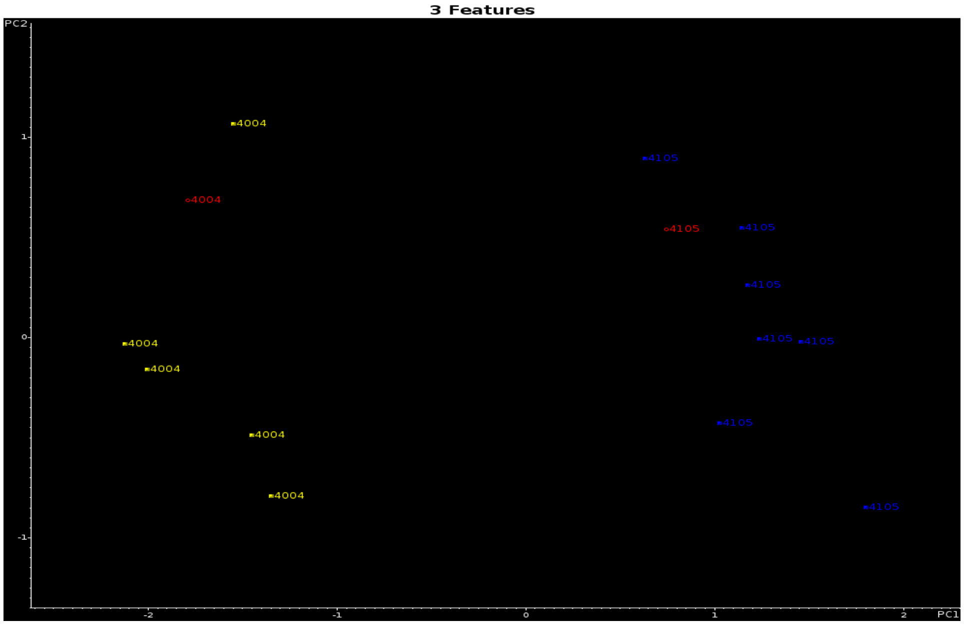


Figure 4.168 Projection of the 2 validation set samples onto the PC plot of the 12 training set samples and the 3 wavelet coefficients identified by the pattern recognition GA (4004=Kyushu, 4105=Oppama)

The pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.169) showed clustering of IR the spectra on the basis of assembly plants from Nissan group 4 after 2 generation runs. To assess the predictive ability of these 2 wavelet coefficients, a validation set of 4 paint samples were projected into 2-PCs developed from the 49 training set and the 2 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.170). The validation set samples were assigned to the correct assembly plants.

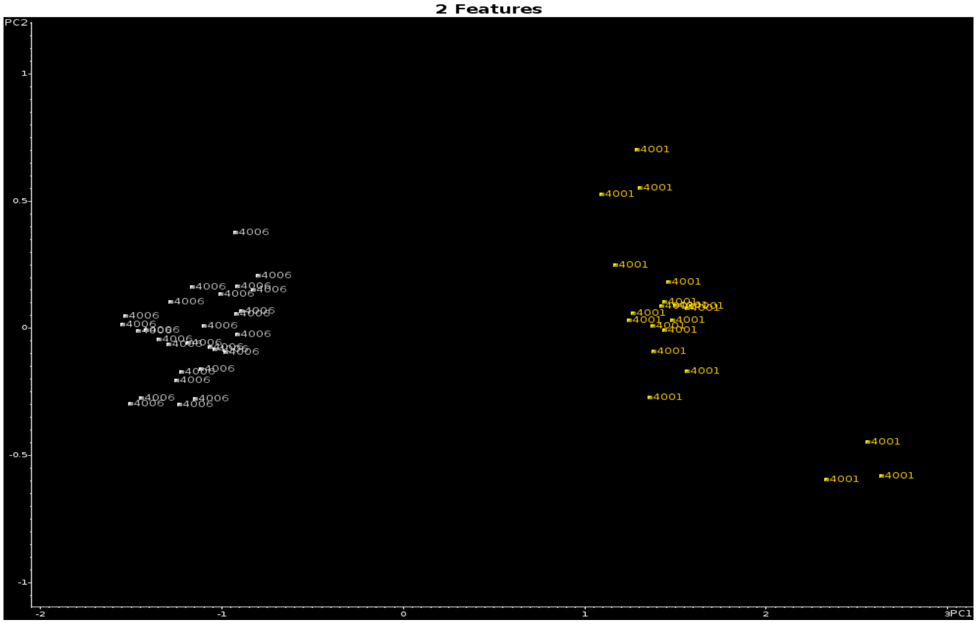


Figure 4.169 2-PC plot of the 49 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4001=Canton, 4006=Smyrna)

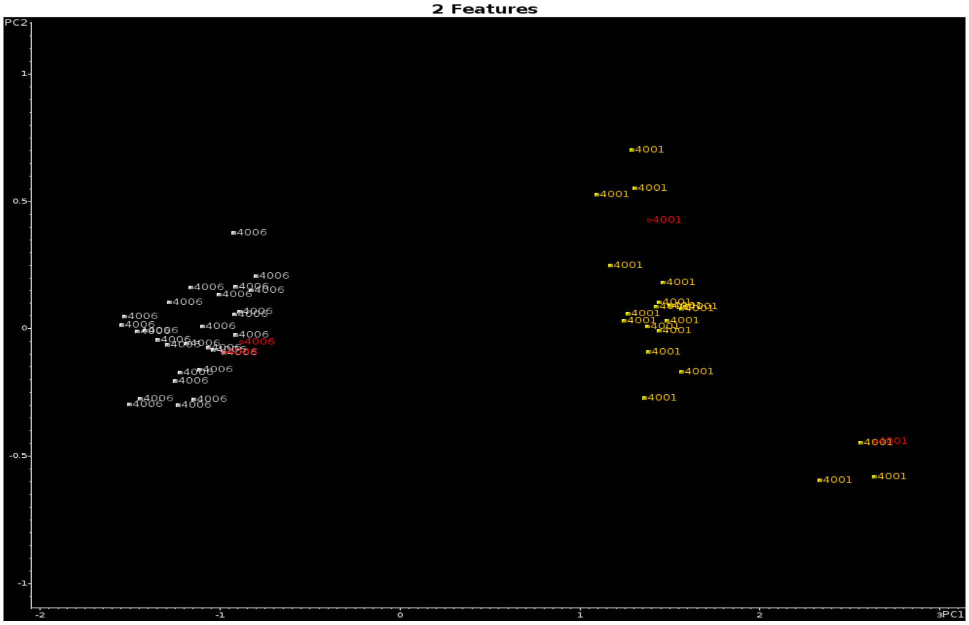


Figure 4.170 Projection of the 4 validation set samples onto the PC plot of the 49 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (4001=Canton, 4006=Smyrna)

4.4.7.3.3 Toyota Assembly Plant Level Prefilters

After the membership of an unknown sample was ascertained in Toyota group 1, the second prefilter was developed to distinguish the sample by Toyota assembly plant. To obtain the information of assembly plant and sub plant, the pattern recognition GA was applied to identify 36 wavelet coefficients whose PC plot (see Figure 4.171) showed clustering of the spectra on the basis of Toyota plant group 1 after 127 generation runs. Toyota assembly plant 5004, 5007 and 5103 were so close, even if the validation set samples were assigned to a certain correct assembly plant (see Figure 4.172), however, the model is unreliable to predict unknown samples. Examining the average IR spectra from these three assembly plant for each individual layer (OT2, OU1 and OU2), Figure 4.173-Figure 4.175 show the assembly plant 5004 and 5103 are very similar in OT2, and three assembly plants are similar in both OU1 and OU2. In addition, the OU1 IR spectrum of an individual sample varies in the same assembly plant might be the reason for misclassification, see Figure 4.176-Figure 4.177. The above testing results suggested to merge these three assembly plants together. After 112 generations, the pattern recognition GA identified 30 wavelet coefficients whose PC plot (see Figure 4.178) showed clustering of the spectra on the basis of assembly plant of Toyota plant group 1. To assess the predictive ability of these 30 wavelet coefficients, a validation set of 16 paint samples were projected into 2-PCs developed from the 131 training set and the wavelet coefficients identified by GA. The fitness function of the pattern recognition GA is Normal. The validation set samples were assigned to the correct assembly plants (see Figure 4.179).

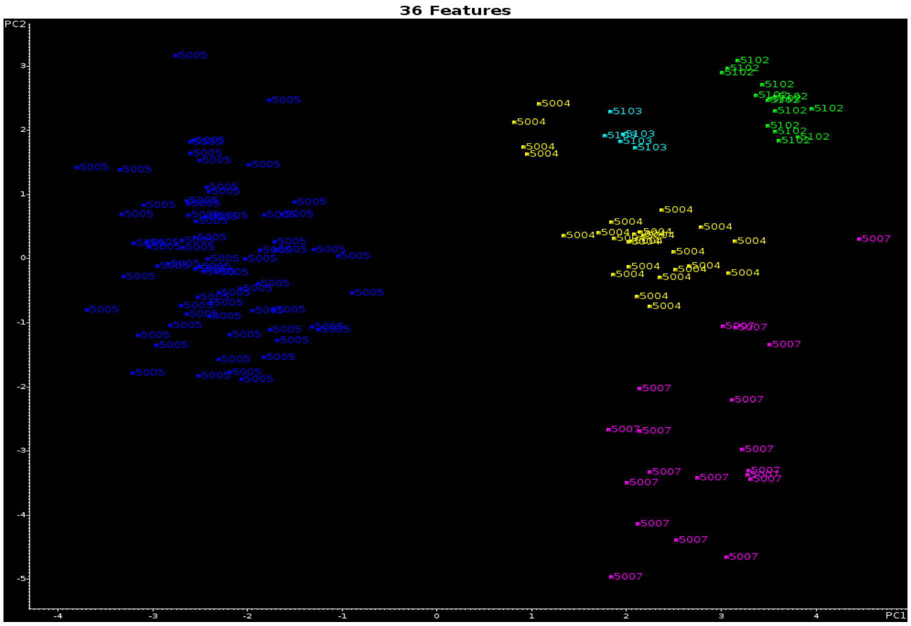


Figure 4.171 2-PC plot of the 131 training set samples and the 36 wavelet coefficients identified by the pattern recognition GA (5004=Georgetown, 5005=Japan, 5007=Princeton, 5102=Cambridge, 5103=Fremont)

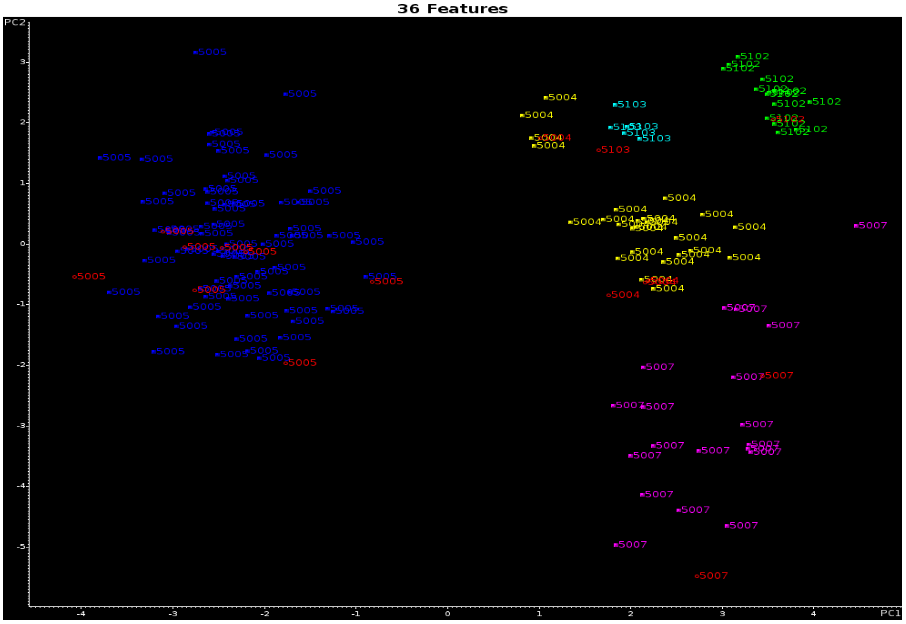


Figure 4.172 Projection of the 16 validation set samples onto the PC plot of the 131 training set samples and the 36 wavelet coefficients identified by the pattern recognition GA (5004=Georgetown, 5005=Japan, 5007=Princeton, 5102=Cambridge, 5103=Fremont)

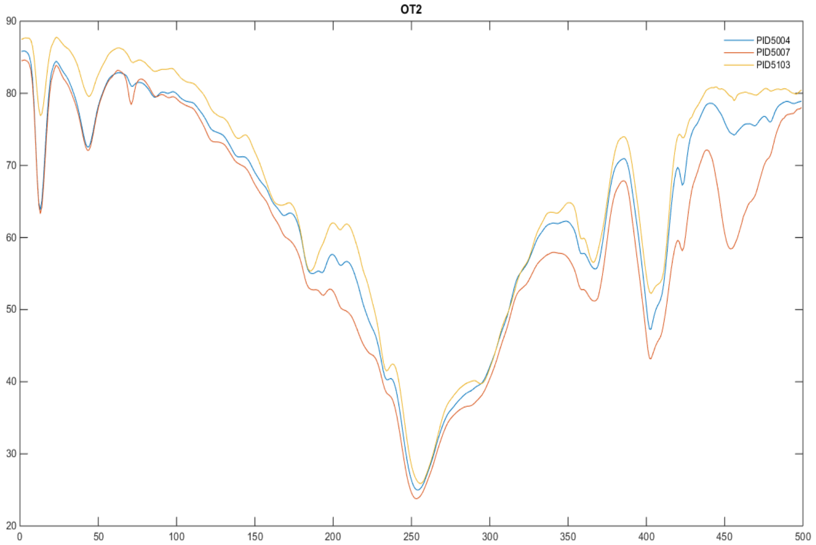


Figure 4.173 The average clear coat IR spectra comparison of assembly plant Georgetown, Princeton and Fremont

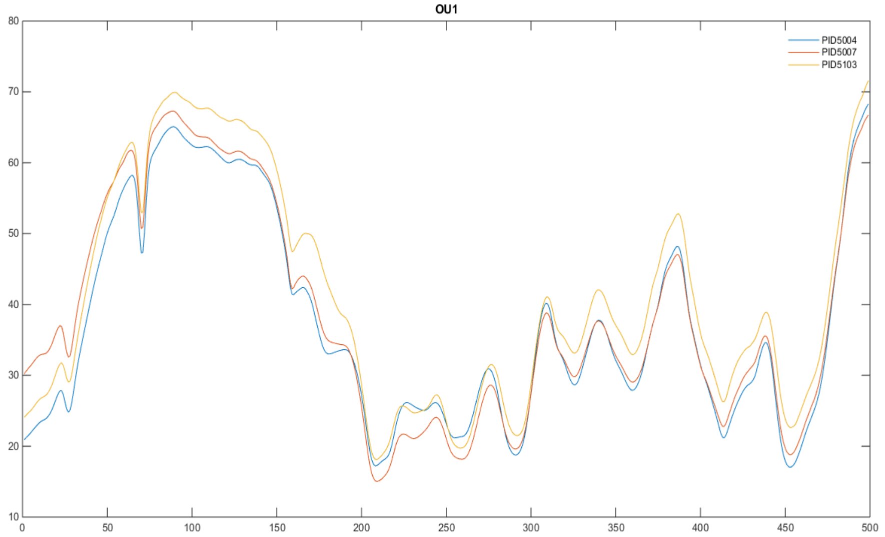


Figure 4.174 The average surfacer IR spectra comparison of assembly plant Georgetown, Princeton and Fremont

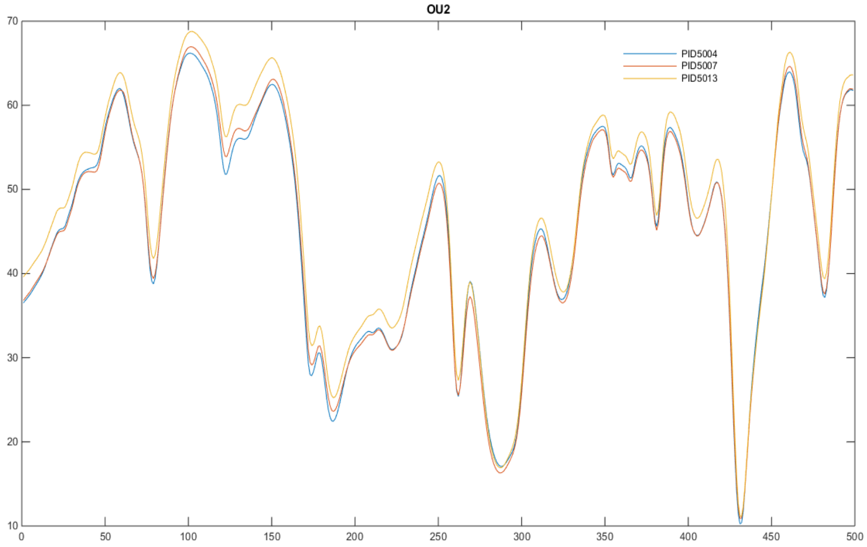


Figure 4.175 The average e-coat primer IR spectra comparison of assembly plant Georgetown, Princeton and Fremont

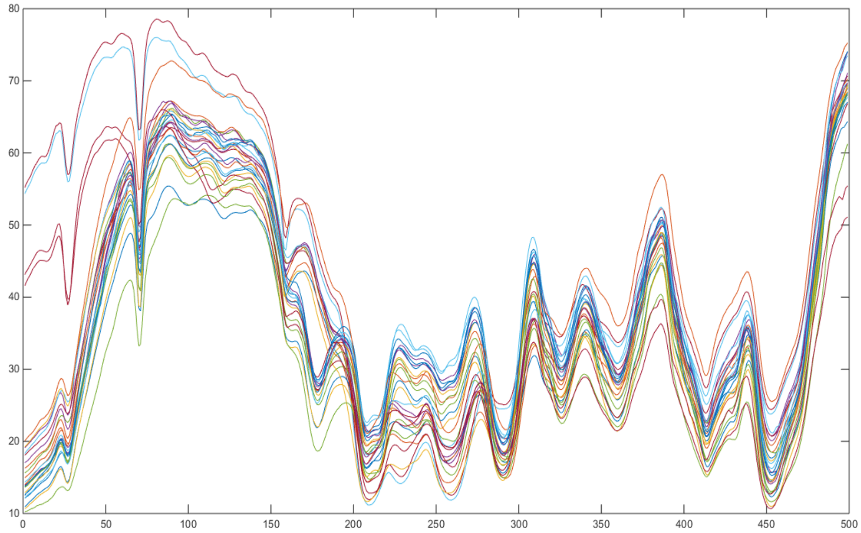


Figure 4.176 The average surfacer IR spectra comparison of assembly plant Georgetown

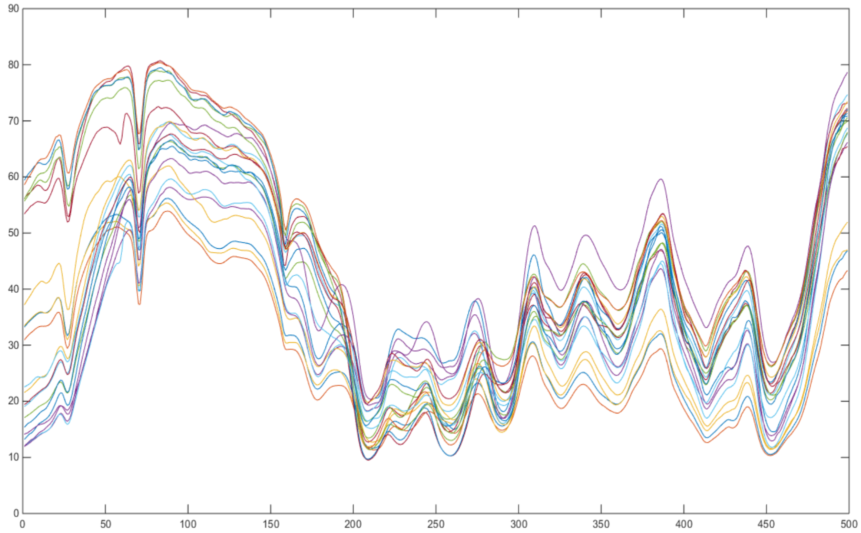


Figure 4.177 The average surfacer IR spectra comparison of assembly plant Princeton

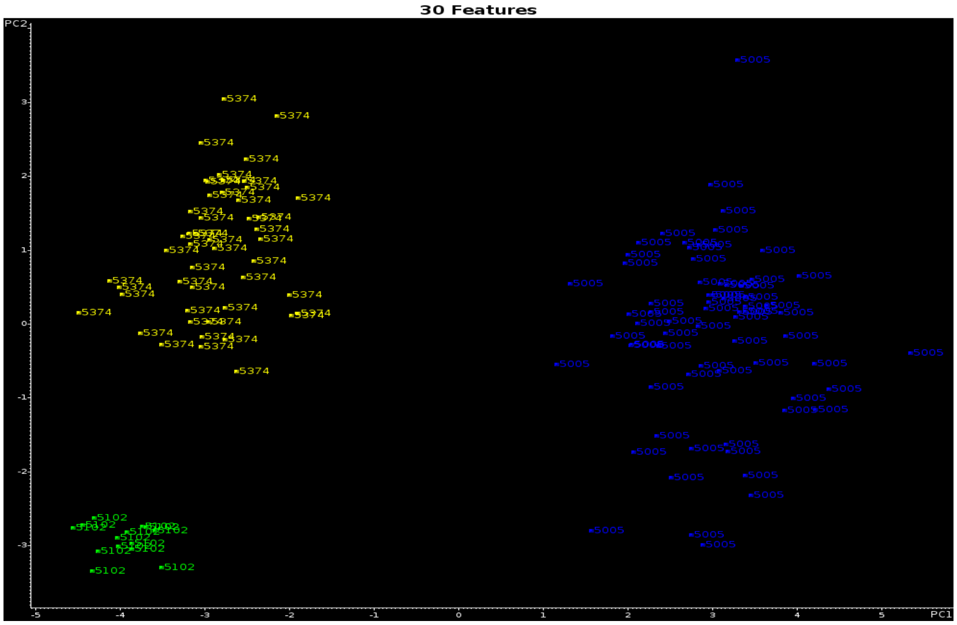


Figure 4.178 2-PC plot of the 131 training set samples and the 30 wavelet coefficients identified by the pattern recognition GA (5005=Japan, 5102=Cambridge, 5374= Georgetown, Princeton Fremont)

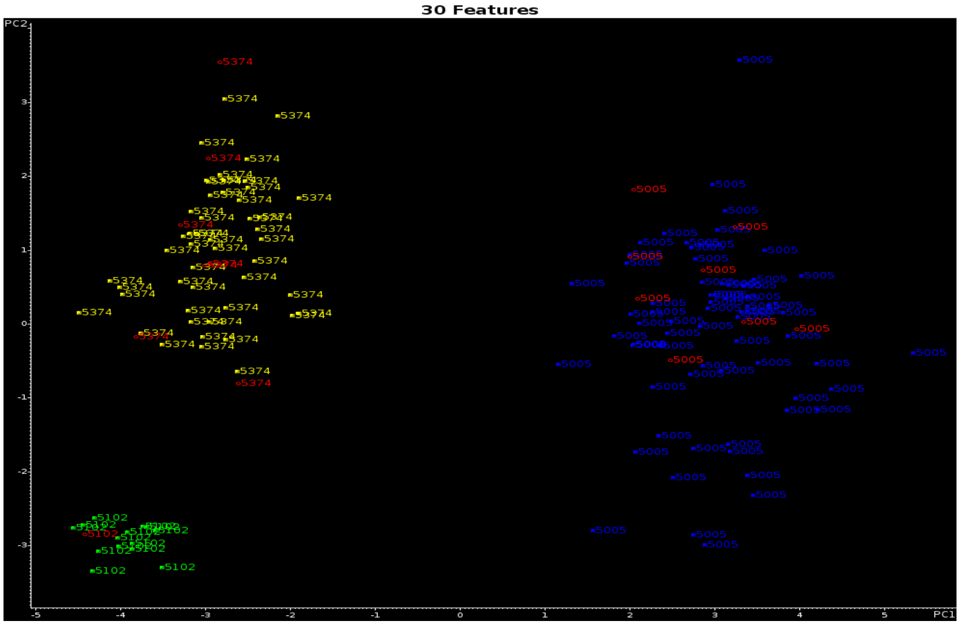


Figure 4.179 Projection of the 16 validation set samples onto the PC plot of the 131 training set samples and the 30 wavelet coefficients identified by the pattern recognition GA(5005=Japan, 5102=Cambridge, 5374= Georgetown, Princeton Fremont)

For the assembly plants in Toyota plant group 2, the pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.180) showed clustering of IR the spectra on the basis of assembly plants from Toyota plant group 2 after 2 generation runs. To assess the predictive ability of these 2 wavelet coefficients, a validation set of 2 paint samples were projected into 2-PCs developed from the 19 training samples and the 2 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA (see Figure 4.181).

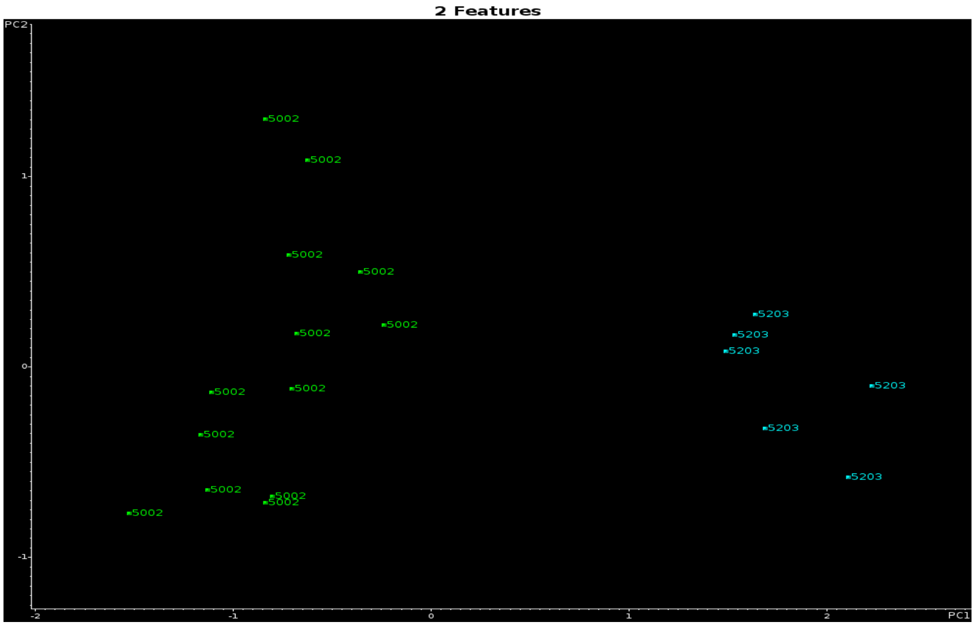


Figure 4.180 2-PC plot of the 19 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (5002=Cambridge, ON, Canada, 5203= Fremont, CA)

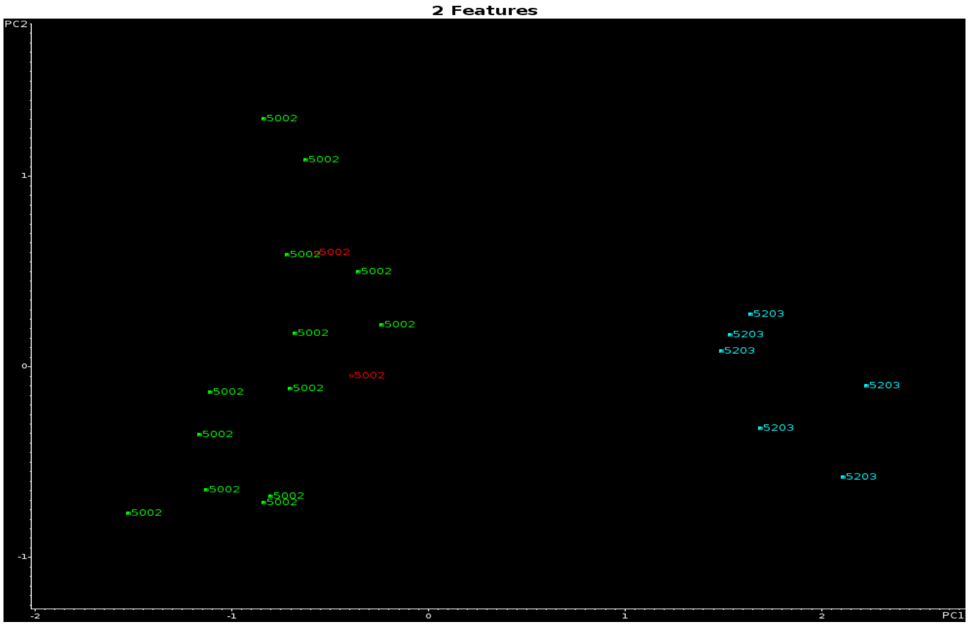


Figure 4.181 Projection of the 2 validation set samples onto the PC plot of the 19 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (5002=Cambridge, ON, Canada, 5203= Fremont, CA)

For the prefilter used for identifying the assembly plants or sub plants of Toyota plant group 3, the pattern recognition GA identified 5 wavelet coefficients whose PC plot (see Figure 4.182) showed clustering of IR the spectra on the basis of assembly plant in Toyota plant group 3 after 4 generation runs. To assess the predictive ability of these 5 wavelet coefficients, a validation set of 3 paint samples were projected into 2-PCs developed from the 15 training set and the 5 wavelet coefficients identified by GA using Normal fitness function of the pattern recognition GA. The validation set samples were assigned to the correct assembly plants (see Figure 4.183).

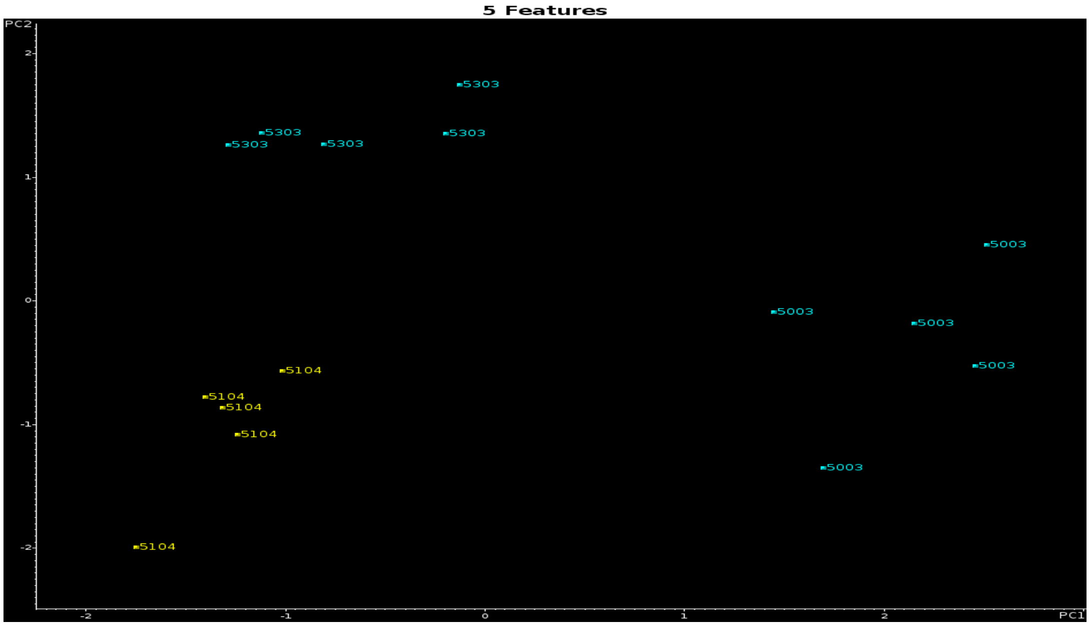


Figure 4.182 2-PC plot of the 15 training set samples and the 5 wavelet coefficients identified by the pattern recognition GA (5003=Fremont, 5104=Georgetown, 5303=Fremont)

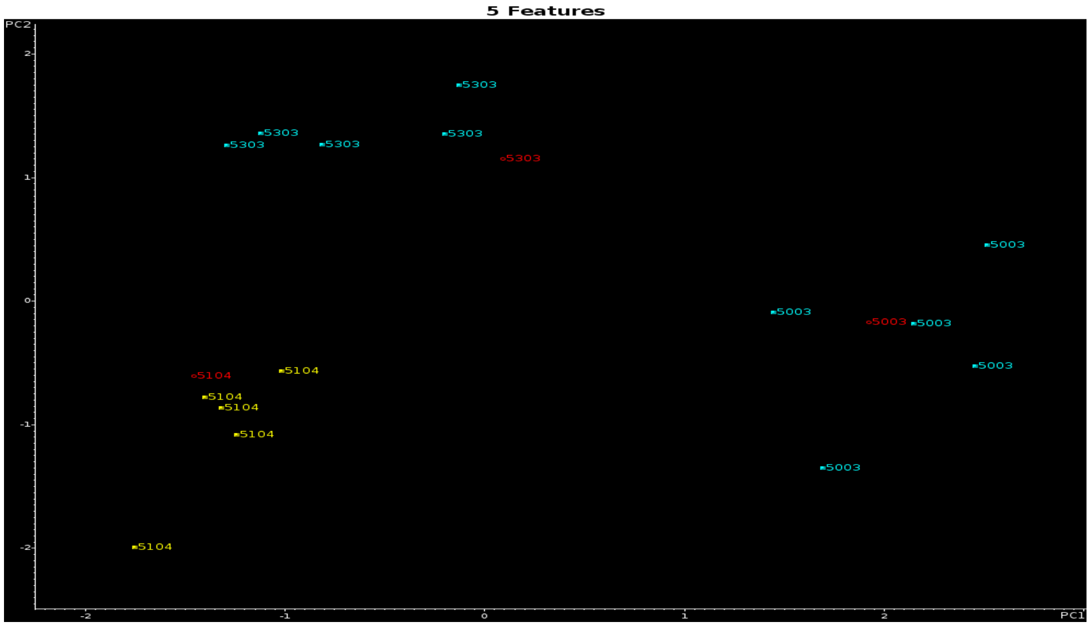


Figure 4.183 Projection of the 3 validation set samples onto the PC plot of the 15 training set samples and the 5 wavelet coefficients identified by the pattern recognition GA (5003=Fremont, 5104=Georgetown, 5303=Fremont)

To develop the assembly plant prefilter for the samples falling in the Toyota plant group4, the pattern recognition GA identified 2 wavelet coefficients whose PC plot (see Figure 4.184) showed clustering of IR the spectra on the basis of assembly plants of the Toyota plant group after 1 generation run. Because the number of samples in the both two assembly plant is too less to assess the predictive ability of these 2 wavelet coefficients, there is no validation sample was set in this test.

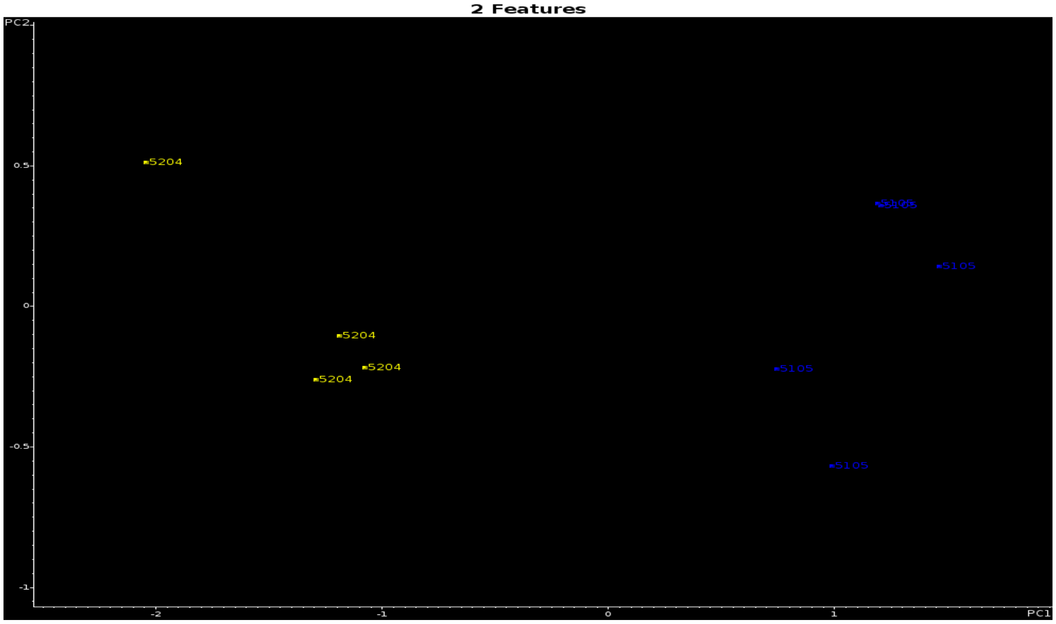


Figure 4.184 2-PC plot of the 9 training set samples and the 2 wavelet coefficients identified by the pattern recognition GA (5105=Japan, 5204=Georgetown)

4.4.7.3.4 GM Search Prefilter

See the results in 4.4.5.2 Singlet GM manufacturer prefilter.

REFERENCES

1. Paint data query. http://www.rcmp-grc.gc.ca/fsis-ssji/paint-peinture-eng.htm

2. Fettis, G Automotive Paints and Coatings; VCH: New York, 1995

3. Streitberger, H.J. Dossel, K.F. Automotive Paints and Coatings, 2nd Ed; NY: Willy, 2008

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

5. Physical-chemical study of car paint coats.
http://www.forensicscience.pl/pfs/39_nieznanska.pdf

6. Caddy B. Forensic Examination of Glass and Paint: Analysis and Interpretation, London: Taylor&Francis, 2002, pp183-220

7. Buckle, J. L., Macdougall, D. A., & Grant, R. R. (1997). PDQ—Paint Data Queries: The History and Technology Behind the Development of the Royal Canadian Mounted Police Forensic Laboratory Services Automotive Paint Database. Canadian Society of Forensic Science Journal, 30(4), 199–212. http://doi.org/10.1080/00085030.1997.10757099

8. Lavine, B. K.; White, C. G.; Allen, M. D.; Weakley, A., Pattern Recognition-Assisted Infrared Library Searching of the Paint Data Query Database to Enhance Lead Information from Automotive Paint Trace Evidence. Appl Spectrosc 2016.

9. Paint data query. http://www.rcmp-grc.gc.ca/fs-fd/pdfs/pdq-eng.pdf

10. Lavine, B. K.; White, C. G.; Allen, M. D.; Fasasi, A.; Weakley, A., Evidential significance of automotive paint trace evidence using a pattern recognition based infrared library search engine for the Paint Data Query Forensic Database. Talanta 2016, 159, 317-329.

12. Lavine, B. K.; White, C.; Allen, M., Forensic analysis of automotive paints using a pattern recognition assisted infrared library searching system: Ford (2000–2006). Microchemical Journal 2016, 129, 173-183.

13. Lavine, B. K.; Mirjankar, N.; Ryland, S.; Sandercock, M., Wavelets and genetic algorithms applied to search prefilters for spectral library matching in forensics. Talanta 2011, 87, 46-52.

14. Lavine, B. K.; Fasasi, A.; Mirjankar, N.; White, C.; Mehta, J., Search prefilters for library matching of infrared spectra in the PDQ database using the autocorrelation transformation. Microchemical Journal 2014, 113, 30-35.

15. Lavine, B. K.; Fasasi, A.; Sandercock, M., Improving PDQ database search strategies to enhance investigative lead information for automotive paints. *Microchemical Journal* **2014,** *117*, 133-137.

16. Lavine, B. K.; Fasasi, A.; Mirjankar, N.; Sandercock, M.; Brown, S. D., Search prefilters for mid-infrared absorbance spectra of clear coat automotive paint smears using stacked and linear classifiers. Journal of Chemometrics 2014, 28 (5), 385-394.

17. Lavine, B. K.; Fasasi, A.; Mirjankar, N.; Sandercock, M., Development of search prefilters for infrared library searching of clear coat paint smears. Talanta 2014, 119, 331-340.

18. Lavine, B. K.; Fasasi, A.; Mirjankar, N.; White, C.; Sandercock, M., Search prefilters to assist in library searching of infrared spectra of automotive clear coats. Talanta 2015, 132, 182-190.

19. Fasasi, A.; Mirjankar, N.; Stoian, R. I.; White, C.; Allen, M.; Sandercock, M. P.; Lavine, B. K., Pattern recognition-assisted infrared library searching of automotive clear coats. Appl Spectrosc 2015, 69 (1), 84-94.

20. Chapter-3: CharacterizationTechniques and Instrumentation
http://shodhganga.inflibnet.ac.in/bitstream/10603/83415/8/08_chapter3.pdf

21. Fourier transform infrared FTIR spectroscopy
http://www.physics.nus.edu.sg/~L3000/Level3manuals/FTIR.pdf

22. Collin G. White Variable selection to improve classification in structure-activity studies and spectroscopic analysis, Oklahoma State University, Ph.D. Dissertation, Stillwater, OK, 2016

.

CHAPTER V

CONCLUSION

In the preceding chapters, a basic methodology for analyzing underdetermined and redundant spectroscopic data sets which utilized variable selection for model development was described. An IR spectrum or an ion mobility distribution profile was represented as a point in a high dimensional measurement space. The discrete wavelet transform was applied to each sample data vector to resolve overlapping spectral bands. To identify the wavelet coefficients containing signal, a genetic algorithm for variable selection and classification was applied to the data to identify wavelet coefficients that optimize the separation of the classes in a plot of the two or three largest principal components of the data. A good principal component plot can only be generated using coefficients that contain information about the class membership of the samples comprising the data set. Wavelet coefficients that maximize the ratio of between-class to within-class variance are selected by the pattern recognition GA.

The proposed methodology for underdetermined and redundant data sets has been validated on a wide range of data. In one study, search prefilters to identify the make and model of an automobile from which a paint chip originated were developed from the fingerprint region of IR spectra of automotive paints to facilitate searching of IR spectra in the PDQ database. In another study, discriminants developed from ion mobility distribution profiles N-linked glycans extracted from sera and analyzed by MALDI-IMS-MS differentiated individuals diagnosed with Barrett's esophagus, high-grade dysplasia, esophageal adenocarcinoma and disease-free controls. In both studies, the combination of wavelet preprocessing and variable selection using a genetic algorithm as a general solution to problems in the field of spectral pattern recognition was demonstrated.

Pattern recognition methods operate with well-defined criteria and attempt to extract useful information from raw data. If the limitations of the methods are not fully understood, the danger of misinterpretation or misuse of costly measurements is significant. The dramatic increase in the number and sophistication of chemical instruments has triggered interest in the development of new data analysis techniques that can extract information from the large arrays of chemical data routinely generated in laboratories. Evaluating data and extracting information from it is a task that is always changing as the sophistication and methodology of modern instruments increases. For these reasons, new pattern recognition techniques that need to be developed to analyze these new streams of data should focus on extending the ability of human pattern recognition. Hence, the approach used in the research described in this dissertation relied heavily on graphics for the presentation of results. Although the computer can assimilate more numbers at any given time than can a scientist, it is the scientist, who in the end, must make the decisions and judgements.

VITA

Tao Ding

Candidate for the Degree of

Doctor of Philosophy

Thesis: Pattern Recognition Studies of Complex Spectroscopic Data Sets

Major Field: Chemistry

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Chemistry at Oklahoma State University, Stillwater, Oklahoma in December, 2016.

Completed the requirements for the Master of Science in Chemistry at Oklahoma State University, Stillwater, Oklahoma in 2009.

Completed the requirements for the Bachelor of Science in Pharmacy at Shenyang Pharmaceutical University, Shenyang, China in 1992.

Experience: Teaching/Research Assistant, Department of Chemistry, Oklahoma State University (2013-2016), Senior Research Support Specialist, Department of Pharmacy, SUNY at Buffalo (2010), Teaching/Research Assistant, Department of Chemistry, Oklahoma State University (2007-2009), Regulatory Investigator, the Center for Drug Evaluation and Research, Sichuan Food and Drug Administration (2005-2006), Associate Professor, the Division of Pharmaceutical Formulation, Sichuan Industrial Institute of Antibiotics (2003-2007), Research Scientist, the Division of Pharmaceutical Formulation, Sichuan Industrial Institute of Antibiotics (1996-2003), Research Assistant, the Division of Pharmaceutical Formulation, Sichuan Industrial Institute of Antibiotics (1992-1996).

Professional Memberships: Society for Applied Spectroscopy