Talanta 119 (2014) 331–340

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Development of search prefilters for infrared library searching of clear coat paint smears

Barry K. Lavine [a,*], Ayuba Fasasi [a], Nikhil Mirjankar [a], Mark Sandercock [b]

[a] Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, USA
[b] Royal Canadian Mounted Police Forensic Laboratory, 15707-118th Avenue, Edmonton, Alberta, T5V 1B7, Canada

ARTICLE INFO

Article history:
Received 17 September 2013
Received in revised form
30 October 2013
Accepted 31 October 2013
Available online 12 November 2013

Keywords:
Pattern recognition
Feature selection
Classification transfer
Genetic algorithms
Forensic analysis
Search prefilters

ABSTRACT

Search prefilters developed from spectral data collected on two 6700 Thermo-Nicolet FTIR spectrometers were able to identify the respective manufacturing plant and the production line of an automotive vehicle from its clear coat paint smear using IR transmission spectra collected on a Bio-Rad 40A or Bio-Rad 60 FTIR spectrometer. All four spectrometers were equipped with DTGS detectors. An approach based on instrumental line functions was used to transfer the classification model between the Thermo-Nicolet and Bio-Rad instruments. In this study, 209 IR spectra of clear coat paint smears comprising the training set were collected using two Thermo-Nicolet 6700 IR spectrometers, whereas the validation set consisted of 242 IR spectra of clear coats obtained using two Bio-Rad FTIR instruments.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Automotive paints consist of multiple layers of paint: original and repaint layers, topcoats and primers [1,2]. An original factory applied paint system for a motor vehicle has a typical layer sequence of substrate, primer, primer–surfacer, color coat and clear coat. A forensic database for automotive paints, known as the Paint Data Query (PDQ) database, was created by the Royal Canadian Mounted Police (RCMP) Forensic Laboratory Services to identify the model, line, and production year range of a vehicle from a paint sample recovered at a crime scene [3,4]. PDQ contains information about the physical attributes, the chemical composition and the infrared (IR) spectrum of each layer of the original paint system for a motor vehicle. If the original automotive paint is present in a paint sample, PDQ can assist in identifying the specific manufacturer and year of manufacture of the motor vehicle. Currently, PDQ is the largest international automotive paint database in existence, and is being used by forensic scientists working in Canada, United States, Australia, New Zealand, Singapore, Japan, South Africa, the Middle East, and many European countries.

PDQ was designed as a general text-based search and retrieval system [5,6]. The text-based search of both the physical and chemical characteristics of each layer of automotive paint can serve as a potent pre-screen to a manual infrared spectral search of materials that tend to be chemically very similar to one another. However, pattern recognition software coupled with this database has the potential for more specific searches by relying less on subjective text-based characteristics. PDQ contains information on the complete topcoat and undercoat paint layers of most domestic and foreign vehicles marketed or imported into North America since the mid-1970s.

The major problem with PDQ is its use of text to code chemical information about each paint layer. Searches of the PDQ database require the user to code their IR spectrum according to predetermined guidelines, and to search these codes against the codes in the database. Direct searching of IR spectra in the database does not exist, and commercial library search algorithms cannot distinguish subtle differences between IR spectra from one model to the next. The coding used in PDQ is generic, which can impair both the accuracy of the search and lead to non-specific search criteria that result in a large number of hits that a scientist must then work through and eliminate. Furthermore, the text based system of PDQ does not allow for searching of clear coats. All modern clear coats applied to automotive components have only one of two possible formulations. They are coded as either acrylic melamine styrene or acrylic melamine styrene polyurethane. With the exception of the clear coats, each paint layer contains color pigments and fillers. As there are no inorganic fillers or color with which to further discriminate a clear coat, paint samples that do not contain the color coat layer or at least one of the undercoat (primer) layers

* Corresponding author.
E-mail address: bklab@chem.okstate.edu (B.K. Lavine).

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.10.066

cannot be searched for in PDQ because the text based search system relies on the relatively large variations of color and chemical formulations present in the color coat or undercoat layers. The inability to accurately search IR spectra in PDQ and the inability to identify clear coats are significant limitations to the current text-based PDQ database.

Our research group has used pattern recognition techniques to search the IR spectral libraries of the PDQ database in an effort to differentiate between similar but nonidentical IR paint spectra and to correctly identify an unknown paint sample as to the assembly plant, model, and line of the vehicle [7,8]. Searches with commercial library search algorithms have met with only limited success as automotive paint libraries are composed of a large number of similar spectra and commercial search algorithms have not proven to be sufficiently sensitive at distinguishing shoulders and minor peaks which can be crucial for identifying specific models and lines. By applying wavelets [9,10], subtle but significant features in the IR spectra of clear coats can be enhanced by decomposing each spectrum into wavelet coefficients which represent the sample's constituent frequency. A genetic algorithm (GA) for pattern recognition analysis [11,12] is used to identify wavelet coefficients characteristic of the assembly plant of the automobile from which the clear coat paint sample was obtained. Even in challenging trials where the samples evaluated were all the same make (Chrysler) with a limited production year range (1999–2000), the respective assembly plants and line of the motor vehicles could be correctly identified [13,14] using search prefilters (i.e., discriminants) developed from wavelet coefficients identified by the pattern recognition GA.

The use of search prefilters generates fewer hits and increases accuracy, translating into a significant time savings for the forensic scientist. Information derived from the proposed pattern recognition searches also serves to quantify the general discrimination power of original automotive paint comparisons encountered in casework, and will further efforts to succinctly communicate the significance of the evidence to the courts. Addressing these concerns is a direct response to Recommendation 3 of the National Academies' February 2009 report [15], “Strengthening Forensic Science in the United States: A Path Forward.”

Analogous to the situation found in multivariate calibration, a classification model (i.e., a search prefilter) developed from IR spectra measured on one instrument is often not valid when applied to the prediction of spectra collected on a second instrument. As a result of the large number and variety of FTIR spectrometers sold, the ability to transfer multivariate classification models between FTIR spectrometers is crucial for the successful application of the search prefilters developed using IR data from the PDQ database. Therefore, the transfer of multivariate classification models between laboratory spectrometers has been investigated as part of this study. Algorithms previously developed to implement classification transfer can be divided into two groups. In the first group, a set of standards common to both the primary and secondary instruments is used to correct for unwanted instrumental variation. Of these approaches, the most popular are piecewise direct standardization [16] and orthogonal signal correction [17]. In the second group, signal preprocessing and data transformation techniques are used to correct for unwanted instrumental variation, e.g., finite impulse response filtering [18] and slope and bias correction [19], as transfer standards are unavailable.

Although there are discussions, reviews, and algorithm comparisons published on this subject, fundamental and first principle derivations are lacking. Often, practitioners are confronted with the situation of applying a potpourri of algorithms to their data to empirically determine what “works” best for their own application. They end up searching for models with fewer factors and smaller standard errors of prediction (without confidence limits or tolerance verification) rather than applying a thorough understanding and rigor to seek a more fundamental solution to the problem of instrument transfer.

In this study, classification transfer was accomplished by matching spectral line shapes of different instruments using convolution and deconvolution functions implemented with Nicolet's OMNIC software system. This allowed the spectra from one instrument appear to have been collected on a second instrument. The success of transforming spectral lines between spectrometers will enable new pattern recognition techniques developed for spectral library searching of the PDQ database to be implemented in a large number of forensic laboratories regardless of the spectrometer used to collect the data.

2. Experimental

IR spectra of clear coats in the PDQ library were collected using four different FT-IR spectrometers: Bio-Rad 40A, Bio-Rad 60A, and two Thermo-Nicolet 6700 FTIR spectrometers. All IR spectrometers were run at 4 cm^{-1} resolution. Each spectrometer was equipped with a DTGS detector. All clear coat paint samples were between 3 and 4 micrograms. Each paint sample was run using diamond windows [20,21]. Additional details about the sampling conditions used to generate the IR data in this study can be found elsewhere [22].

Using OMNIC, all IR spectra (in both the training set and validation set) were aligned. The number of points collected in the wavelength range interrogated by the Thermo-Nicolet instruments varied from 1878 points to 1958 points whereas all spectra collected on the two Bio-Rad instruments for the same wavelength range and resolution were represented by 1944 points. Band shifting was also observed in spectra collected on both the Thermo Nicolet and Bio-Rad instruments. These problems were resolved using Nicolet's OMNIC software as an editor to process the Bio-Rad spectra and the spectra from the Thermo Nicolet instruments using an appropriate estimate of the spectral line function of the two Thermo-Nicolet instruments.

Spectral line shapes between instruments were matched using convolution and deconvolution functions developed with Nicolet's OMNIC software system. An instrumental line function representative of the two Thermo Nicolet instruments and developed by OMNIC was applied to the Bio-Rad spectra to ensure that all measurements made by the Bio-Rad instruments were comparable to IR spectra collected on the two Thermo-Nicolet instruments. To authenticate wavelength alignment along the x-axis for all clear coat spectra of GM automobiles between the years 2000–2006, IR spectra of similar clear coat samples collected on both the BioRad and Thermo Nicolet instruments were subtracted after performing the alignment procedure. The subtraction yielded zero at each point.

To further improve spectral alignment, we focused on the region from 600 cm^{-1} to 1500 cm^{-1}. Each IR spectrum selected for processing was normalized to the helium neon laser frequency of 15798.0 cm^{-1}. The laser frequency value was set to that measured at the aperature setting. This makes the sample peak positions independent of aperature setting. This also results in a change in data point spacing and the resulting data point locations. Although the default laser frequency of the spectrometer is 15798.3 cm^{-1}, 15798.0 cm^{-1} was used as it also solved problems when importing spectra from GRAMS and other instruments. This ensured proper spectral alignment along the x-axis for imported Bio-Rad spectra to the Thermo-Nicolet instrument.

For alignment along the y-axis (transmittance) of the spectra (600–1500 cm^{-1}), we ensured that all spectra started from the same transmittance value. The quality of the diamond cell transmission spectra in the PDQ library (e.g., no sloping baseline or

baseline offsets, and the value of the carbonyl absorbance peak in all library spectra being unity) proved pivotal in the successful alignment of these spectra along the *y*-axis.

3. Data analysis

Spectral features characteristic of the manufacturer plant or model of the vehicle were identified by a genetic algorithm (GA) for pattern recognition analysis that utilized supervised learning to identify coefficients that optimize separation of the IR spectra by manufacturing plant in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by the features selected by the pattern recognition GA is about differences between the different classes (assembly plants) in the data. A principal component plot that shows separation of the data by class can only be generated using features whose variance or information is primarily about differences between the classes. This criterion used in the fitness function of the pattern recognition GA will reduce the size of the search space. To minimize convergence to a local optimum, the pattern recognition GA focuses on specific classes and/or samples difficult to classify by boosting the relative importance of these classes and/or samples in the calculation for fitness during training. The pattern recognition GA learns its optimal parameters in a manner similar to a neural network while simultaneously integrating aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection and classification.

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

The fitness function of the pattern recognition GA (which is called PCKaNN) emulates human pattern recognition through machine learning to score the principal component plots and thereby identify a set of features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. To facilitate the tracking and scoring of the principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see Eqs. (1) and (2)) where CW(c) is the weight of class c (with c varying from 1 to the total number of classes in the data set). $SW_c(s)$ is the weight of sample s in class c. The class weights sum to 100, and the sample weights for the objects comprising a particular class sum to a value equal to the class weight of the class in question.

$$\mathrm{CW}(c) = 100\frac{\mathrm{CW}(c)}{\Sigma_c \mathrm{CW}(c)} \tag{1}$$

$$\mathrm{SW(s)} = \mathrm{CW}(c)\frac{\mathrm{SW(s)}}{\Sigma_{s\in c}\mathrm{SW(s)}} \tag{2}$$

The scoring of a feature subset by the fitness function of the pattern recognition GA can be understood by considering the following binary classification problem. Each class in the data set is assigned equal weights. The number of samples in Class 1 is 50, and the number of samples in Class 2 is 10. All samples in a given class have the same weight during generation 0. Therefore, all samples in Class 1 have as their sample weight 1, and each sample in Class 2 has a weight of 5. If a sample from class 2 has 8 class one samples as its nearest neighbors, SHC/K will equal 0.8, and $(\mathrm{SHC/K}) \times \mathrm{SW} = 0.8 \times 5$ or 4. By summing $(\mathrm{SHC/K_c}) \times \mathrm{SW}$ for each sample, each principal component plot is scored (see Eq. (3)). An obvious advantage of using this scoring procedure for the principal component plots is that a class containing a large number of samples will not dominate the calculation.

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

By changing (i.e., boosting) the class and sample weights, the fitness function of the pattern recognition GA is able to focus on samples and classes that are difficult to classify. To perform boosting, the sample-hit rate (SHR) and the class-hit rate (CHR) must be computed. SHR is the mean value of $\mathrm{SHC/K_c}$ over all feature subsets formulated in a particular generation (see Eq. (4)) and CHR is the mean sample hit rate of all samples in a class (see Eq. (5)). ϕ in Eq. (4) is the number of chromosomes in the population, and AVG in Eq. (5) is the average or mean value. Class and sample weights are adjusted in each generation using a perceptron (see Eqs. (6) and (7)). The momentum term, *P*, of the perceptron is set by the user – $g+1$ in Eqs. (6) and (7) refer to the current generation, and *g* is the previous generation. Classes with a lower CHR and samples with a lower SHR are boosted more heavily than classes and samples that score well.

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

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

$$\mathrm{CW}_{g+1}(\mathrm{s}) = \mathrm{CW_g(s)} + P(1-\mathrm{CHR_g}(s)) \tag{6}$$

$$\mathrm{SW}_{g+1}(\mathrm{s}) = \mathrm{SW_g(s)} + P(1-\mathrm{SHR_g}(s)) \tag{7}$$

Boosting is crucial to ensure the successful operation of the pattern recognition GA because it modifies the fitness landscape by adjusting the values of the class and sample weights. This helps to minimize the problem of convergence to a local optimum. Hence, the fitness function of the pattern recognition GA is continually changing using information from previous generations as the population is evolving towards a solution. Further details about the genetic algorithm used for pattern recognition analysis and feature selection can be found elsewhere [23–31].

For pattern recognition analysis, all clear coat IR transmittance spectra were normalized to unit length. Each IR spectrum was initially represented as a data vector, $x=(x_1, x_2, x_3,\ldots x_j,\ldots\ldots x_{611})$ where x_{611} is the transmittance of the clear coat paint sample for the 611th point. The spectral region from 2000 $\mathrm{cm^{-1}}$ to 600 $\mathrm{cm^{-1}}$, used to develop the search prefilters for plant group, was represented by 611 points. All spectral features in this region were autoscaled to ensure that each measurement has a mean of zero and a standard deviation of one throughout all spectra. Autoscaling removed any inadvertent weighing of the data that otherwise would occur due to differences in the magnitude among the measurement variables comprising each IR spectrum.

4. Results and discussion

In this study, 209 IR spectra of clear coat paint smears that comprised the training set were collected using Thermo-Nicolet 6700 IR spectrometers, whereas the validation set consisted of 242 IR spectra of clear coats obtained using two Bio-Rad IR instruments. The clear coat paint spectra used in this study were obtained from paint samples collected from automobiles manufactured by General Motors (GM) in 21 North American plants between 2000 and 2006. This made the classification problem challenging because the samples evaluated were all from the same

manufacturer (General Motors) with a limited production year range (2000–2006). Only clear coat paint spectra from metallic automobile components were used to develop the search prefilters. IR spectra of clear coats from bumpers and other plastic substrates were excluded as these components are often not painted in the same plant where the vehicle is assembled. Table 1 lists the 21 g manufacturing plants that were investigated in this study.

A hierarchical classification scheme formulated from a visual inspection of the data was used to develop the search prefilters. The 21 g manufacturing plants were divided into five major plant groups (see Table 2). The spectra were initially divided into two categories based on the carbonyl band at 1729 cm^{-1}. In one category, the carbonyl band in each spectrum is a singlet (Plant Groups 1, 3, and 4), and in the other category the carbonyl band is

Table 1
General motors plants investigated in this study.

Plant ID	Plant	Make	Line
1	ARLINGTON	CADILLAC, CHEVROLET, GMC	SUBURBAN, YUKON, ESCALADE,CTA
3	BOWLING	CADILLAC, CHEVROLET	CORVETTE, XLR
4	DORAVILLE	PONTIAC	VENTURE, SILHOUETTE, MONTANA, UPLANDER,
5	FAIRFAX	CHEVROLET, PONTIAC, OLDSMOBILE,	MALIBU, INTRIGUE
6	FLINT	CHEVROLET, GMC	SILVERADO, SIERRA
8	FORT WAYNE	CHEVROLET, GMC	SILVERADO, SIERRA
9	FREMONT	GENERAL MOTORS	VIBE, PRIZM
10	HAMTRAMCK	BUICK, CADILLAC, PONTIAC	DEVILLE, LUCERNE, LESABRE, ELDORADO
12	JANESVILLE	GMC	TAHOE, SUBURBAN, YUKON
14	LANSING	PONTIAC	STS
16	LINDEN	CHEVROLET, GMC	BLAZER, JIMMY,S10
17	LORDSTOWN	PONTIAC	SUNFIRE, CAVALIER, COBOLT, PURSUIT
18	MORAINE	CHEVROLET, GMC, SAA	JIMMY, ENVOY,9S7, BLAZER, TRAIL BLAZER
20	OKLAHOMA CITY	CHEVROLET, GMC	MALIBU, TRAIL BLAZER, ENVOY, EQUIPE, XUV
21	ORION	PONTIAC, BUICK	BONNEVILLE, LESABRE, AURORA, PARK AVENUE
22	OSHAWA	GMC, PONTIAC	ALLURE, REGAL
23	PONTIAC	CHEVROLET,GMC	SILVERADO, SIERRA
24	RAMOS ARIZPE	BUICK, CHEVROLET, PONTIAC	CAVALIER, SUNFIRE, RENDEZVOUS, AZTEK
25	SHREVEPORT	CHEVROLET,GMC	S10, COLOGNE, SONATA
26	SILAO	CHEVROLET,GMC,SAAB	SUBURBAN, YUKON XL
27	SPRING HILL	STARLET	SSL,ION,SC1,SC2,SL1,VUE

Table 2
Manufacturing plants comprising each plant group.

Plant group	Plant ID number	Manufacturing plant
1	1, 4, 5, 8, 14, 18, 23	ARLINGTON, DORAVILLE, FAIRFAX, FORT WAYNE, LANSING, MORAINE, PONTIAC
2	3, 10, 21	BOWLING, HAMTRAMCK, ORION
3	6, 9, 16, 17, 20, 22, 25	FLINT, FREMONT, LINDEN, LORDSTOWN, OKLAHOMA STATE, OSHAWA, SHREVEPORT
4	12	JANESVILLE
5	24, 26, 27	RAMOS ARIZPE, SILAO, SPRING HILL

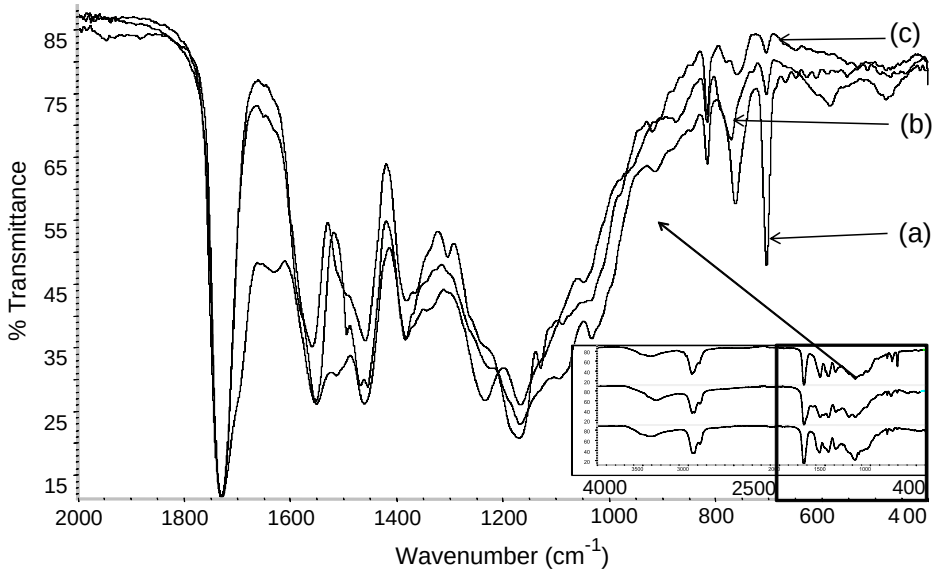

Fig. 1. Average spectra of plant groups 1 (a), 3 (b), and 4 (c) overlayed with the region 2000–400 cm^{-1} expanded.

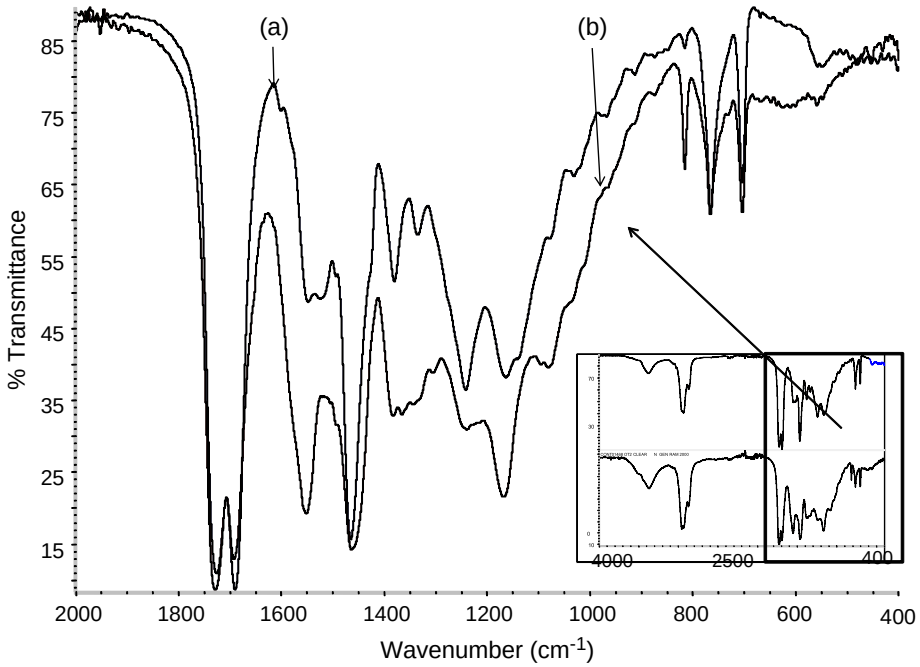

Fig. 2. Average spectra of groups 2 (a) and 5 (b) overlayed with the region 2000–400 cm^{-1} expanded.

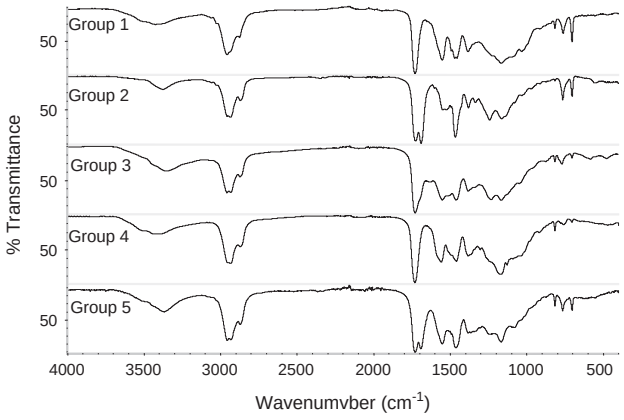

Fig. 3. IR spectra of clear coat paint smears from the PDQ library for GM automobiles in the United States and Canada (2000–2006) could be assigned to one of five plant groups.

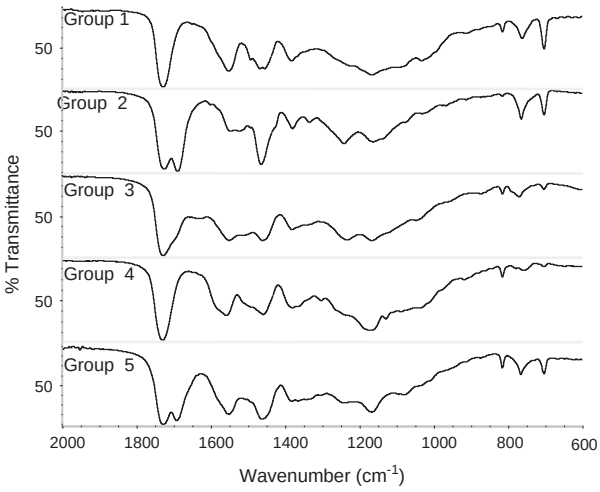

Fig. 4. The fingerprint region of the five plant-groups characteristic of GM clear coats (2000–2006).

a doublet (Plant Groups 2 and 5) due to the presence of polyurethane. Spectra representative of these two categories are shown in Figs. 1 and 2. An examination of the expanded fingerprint region (2000–400 cm^{-1}) reveals five distinct spectral patterns with each pattern designated as a specific Plant Group.

IR spectra of clear coats from the 21 g plants (2000–2006) were then assigned to one of the five plant groups, i.e., groups of manufacturing plants (see Fig. 3). The spectral region used to differentiate the five plant groups includes both the fingerprint region and the carbonyl band of the clear coat. Initially, the carbonyl band was not excluded from the study because it contributed to discrimination of the spectra among the five plant groups. The spectral region from 4000 cm^{-1} to 2000 cm^{-1} included the C–H stretch, which is common to all organic samples, and noise associated with the diamond transmission cell. As this spectral region would not be expected to contain information characteristic of the manufacturing plant of the paint sample, it was not used in the development of the search prefilters.

A search prefilter was developed to classify spectra into one of the five plant groups. Each plant group was further divided into individual manufacturing plants or into subgroups of manufacturing plants using a search prefilter to classify the individual spectra within each plant group. For the development of the search prefilter for plant groups, the spectral region of each clear coat paint sample was limited to the extended fingerprint region (2000–600 cm^{-1}), see Fig. 4. Search prefilters for individual assembly plants were developed from the fingerprint region (1500–600 cm^{-1}). We chose to limit ourselves to the fingerprint region for individual plants to exclude the carbonyl band as it was not sufficiently discriminating for this level of classification.

For the development of the search prefilters, transmittance spectra, not absorbance spectra were used. The crucial issue for the development of the search prefilters was deconvolving overlapping spectral responses using wavelets, not removing noise associated with variations in the optical path length of each sample which was obviated by adjusting the thickness (amount) of the sample and the pressure applied by the diamond

Table 3
Training set and validation set for plant group search prefilter.

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

transmission cell. This ensured that an absorbance of one was obtained for the carbonyl band, which possessed the highest intensity, in all clear coat paint spectra.

The initial focus of this study was to develop a search prefilter to classify the IR spectra by plant group. In this phase of the study, the pattern recognition GA was directly applied to the standardized IR spectra. The training set of 209 IR spectra (Thermo Nicolet) was divided into 5 classes by plant group (see Table 3). The first step in the study was to apply principal component analysis (PCA) to the normalized and autoscaled IR spectral data. PCA is a powerful method for uncovering hidden relationships in complex multivariate data sets. Using this procedure is tantamount to developing a new coordinate system that is better at displaying the information present in the data than axes defined by the original measurement variables. Fig. 5 shows a principal component (PC) plot of the two largest principal components of the 209 IR spectra and the 611 features from each IR spectrum. Each paint sample is represented as a point in the PC map of the data (1=Plant Group 1, 2=Plant Group 2, 3=Plant Group 3, 4=Plant Group 4, and 5=Plant Group 5). The overlap of the clear coats from each plant group in the map of the data is evident.

The next step was feature selection. A genetic algorithm for pattern recognition analysis was used in the study to identify spectral features characteristic of the profile of each plant group. The pattern recognition GA identified features by sampling key feature subsets, scoring their PC plots, and tracking those clear coat paint samples or plant groups that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the GA identified 12 spectral features (i.e., transmittance values at 12 specified wavelengths) whose PC plot showed clustering of the data on the basis of Plant Group membership (see Fig. 6).

A validation set of 242 IR spectra of clear coats from two Bio-Rad instruments was employed to assess the predictive ability of the 12 spectral features identified by the pattern recognition GA and the efficacy of the alignment procedure used to transfer the search prefilters for use by another instrument. Fig. 7 shows the validation set samples projected onto the PC plot defined by the 209 IR spectra (Thermo-Nicolet) and the 12 spectral features identified by the pattern recognition GA. Each projected sample lies in a region of the map with paint samples from the same plant group. This result alone suggests that information about the manufacturing plant is contained in the IR spectrum of the clear coat paint smears.

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

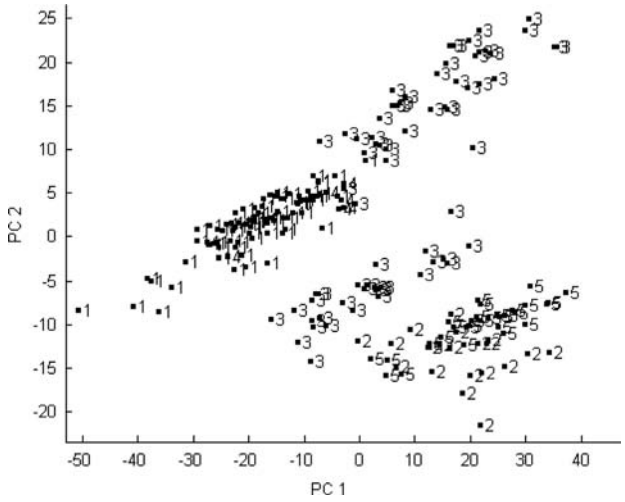

Fig. 5. A PC plot of the two largest principal components of the 209 IR spectra and the 611 features of each spectrum is shown. Each paint sample is represented as a point in the PC plot of the data (1=Plant Group 1, 2=Plant Group 2, 3=Plant Group 3, 4=Plant Group 4, and 5=Plant Group 5).

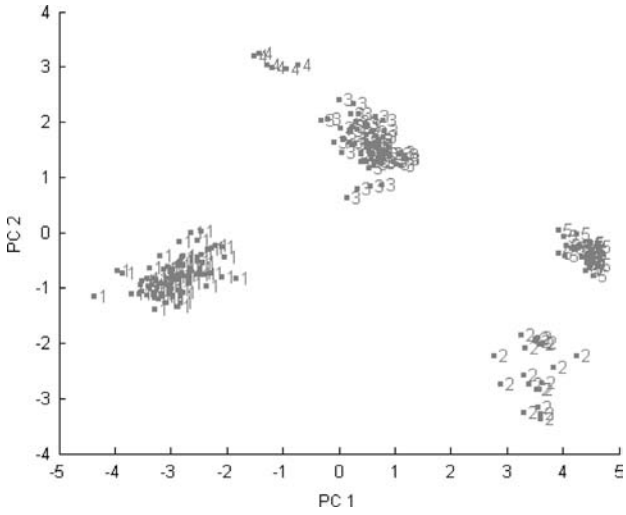

Fig. 6. A PC plot of the two largest principal components of the 209 IR spectra and the 12 spectral features identified by the pattern recognition GA is shown. Each paint sample is represented as a point in the PC plot of the data (1=Plant Group 1, 2=Plant Group 2, 3=Plant Group 3, 4=Plant Group 4, and 5=Plant Group 5).

for the IR spectra in the validation set. The results from the LDA study, which are summarized in Table 4, are consistent with the results obtained using PCA.

The next step in this study was to develop search prefilters to classify paint spectra by manufacturing plant of the paint sample. For each plant group, a search prefilter was developed to discriminate the spectra by manufacturing plant within a plant group. In this phase of the study, the spectral region used to formulate discriminants was from 1500 cm^{-1} to 600 cm^{-1}. The carbonyl band, which was useful for discriminating the IR spectra by plant group, did not prove to be informative for discriminating spectra by manufacturing plant within a plant group due to the similarity of the shape and the intensity of the carbonyl band for manufacturing plants within a plant group.

Because of the similarity of the IR spectra within a plant group, more powerful preprocessing methods were needed to extract information about manufacturing plant from the IR spectra of the clear coats. For this reason, the wavelet packet transform was applied to each normalized IR spectrum using the MATLAB Wavelet toolbox 3.0.4 (MathWorks, Natick, MA). Each IR spectrum was

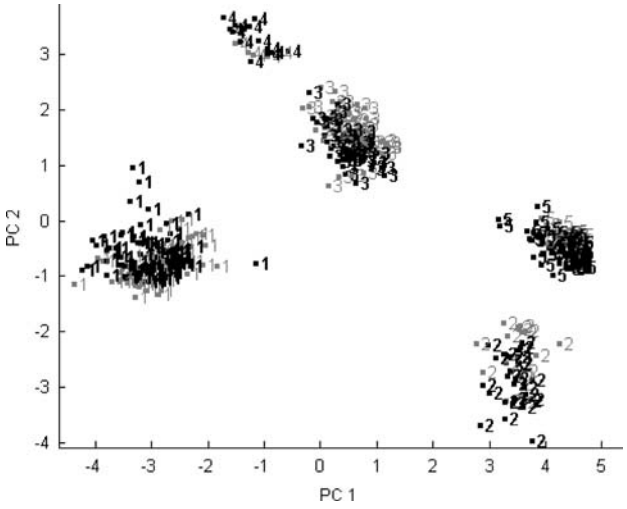

Fig. 7. Validation set samples (black) projected onto the PC plot of the Thermo-Nicolet training set samples defined by 12 spectral features identified by the pattern recognition GA.

Table 4
LDA analysis of 12 spectral features for plant group.

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

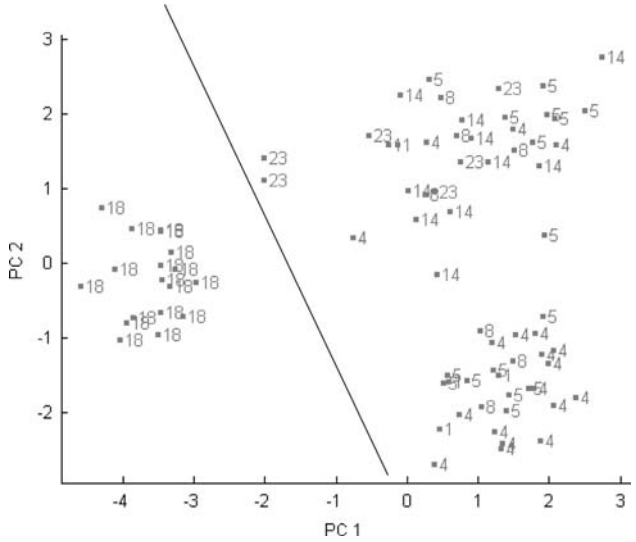

Fig. 8. A plot of the two largest principal components of the 19 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group is shown. Each IR spectrum is represented as a point in the PC plot. 18=Moraine OH, 1=Arlington TX, 4=Doraville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

Table 5
Training set and validation set for plant group 1.

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

iteratively passed through pairs of wavelet filters, which are scaled wavelet functions, at different levels of decomposition. A wavelet filter extracts either the high or low frequency signal components from the IR spectrum. These components are expressed as wavelet coefficients with each coefficient representing the similarity (i.e., dot product) between a scaled wavelet and a section of the IR spectrum. Higher-frequency signal components are extracted using a compressed wavelet with rapidly changing features (i.e., low scaled wavelet). This constitutes the high-pass wavelet filter. The low-pass wavelet filter is a stretched out smoother wavelet (i.e., high scaled wavelet) which extracts lower frequency signal components. For each level of filtering, the spectrum is broken down into a high frequency packet and low frequency packet using a pair of high-pass and low-pass filters. Each packet in turn is further broken down at the next level of decomposition using another pair of high-pass and low-pass wavelet filters. This process will continue until the required level of decomposition has been achieved.

The wavelet coefficients for each clear coat paint spectrum were organized as a data vector. Each coefficient was autoscaled. For this phase of the study, each IR spectrum was represented by 1080 wavelet coefficients using the Symlet6 mother wavelet at the 8th level of decomposition (i.e., 8Sym6) to denoise and deconvolute each IR spectrum. This mother wavelet was selected based on its ability to extract information about assembly plant from the IR spectra. We observed a decrease in the ability of the pattern recognition GA to correctly classify the IR spectra when other mother wavelets were used to transform the spectral data. The efficacy of 8Sym6 for this particular application can be explained by a well known empirical rule used by workers in the field to guide the selection of suitable wavelets for preprocessing their data. If the spectrum contains very sharp peaks, the Haar or other compact wavelets would be indicated for denoising, whereas the Daubachies, which is a smoother wavelet, is recommended for spectra containing broader peaks. For spectral peaks that lie between these two extremes, such as mid-IR spectral data, the Symlet 4 through Symlet 8 mother wavelets are expected to give good results.

Fig. 8 shows a plot of the two largest principal components of the 19 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group (see Table 5). Each IR spectrum is represented as a point in the PC plot of the data. Plant 18 (Moraine OH) is well separated from the other manufacturing plants in the PC plot. Although the pattern recognition GA was parameterized to search for wavelet coefficients to separate all assembly plants, the class structure of the data detected by the GA when performing feature selection indicated that only a single assembly plant (Plant 18, Moraine OH) could be identified among the 7 assembly plants that constitute this plant group. A visual examination of the spectra from the other 6 manufacturing plants (Arlington TX, Doraville GA, Fairfax KS, Fort Wayne IN, Lansing MI, and Pontiac MI) revealed that they were super-imposable, which prevented further discrimination by assembly plant of these clear coats.

A validation set of 90 IR spectra (see Table 5) was employed to assess the predictive ability of the 19 wavelet coefficients identified by the pattern recognition GA. We chose to map the 90 spectra directly onto the PC map defined by the 81 spectra of the training set and the 19 wavelet coefficients identified by the pattern recognition GA. Fig. 9 shows the validation set samples projected onto the PC map developed from the training set data. Each projected sample lies in a region of the map with paint samples that have the same class label: either plant 18 or plants 1, 4, 5, 8, 14, and 23. Evidently, the pattern GA can identify wavelet

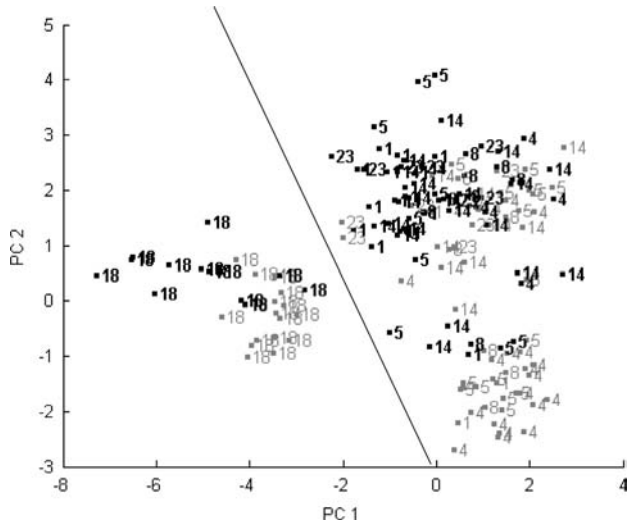

Fig. 9. Validation set samples (black) projected onto a plot of the two largest principal components of the 19 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the first plant group is shown. Each IR spectrum is represented as a point in the PC plot. 18=Moraine OH, 1=Arlington TX, 4=Doraville GA, 5=Fairfax KS, 8=Fort Wayne IN, 14=Lansing MI, and 23=Pontiac MI.

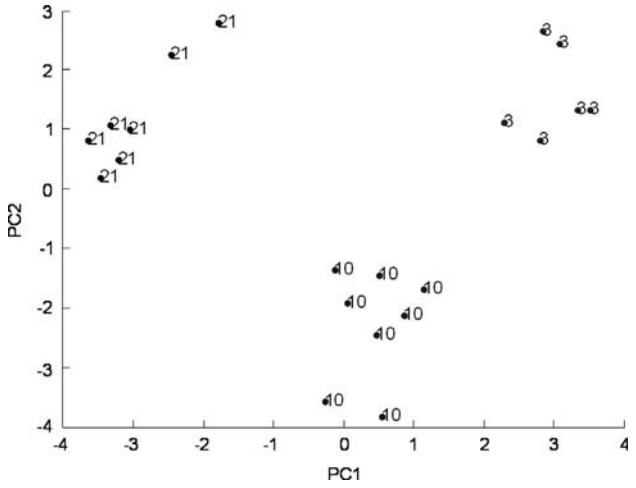

Fig. 10. A plot of the two largest principal components of the 23 wavelet coefficients identified by the pattern recognition GA for the manufacturing plants comprising the second plant group is shown. 3=Bowling Green KY, 10=Hamtramck MI, 21=Orion MI.

Table 6
Training set and validation set for plant group 2.

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

coefficients characteristic of the manufacturing plant of a clear coat paint smear. This suggests that search prefilters developed from IR spectra of clear coats can be used to characterize paint smears by manufacturing plant or can identify a limited number of manufacturing plants associated with the clear coat paint layer.

Fig. 10 shows a plot of the two largest principal components of the 22 IR spectra of the training set and the 23 wavelet coefficients identified by the pattern recognition GA for the manufacturing

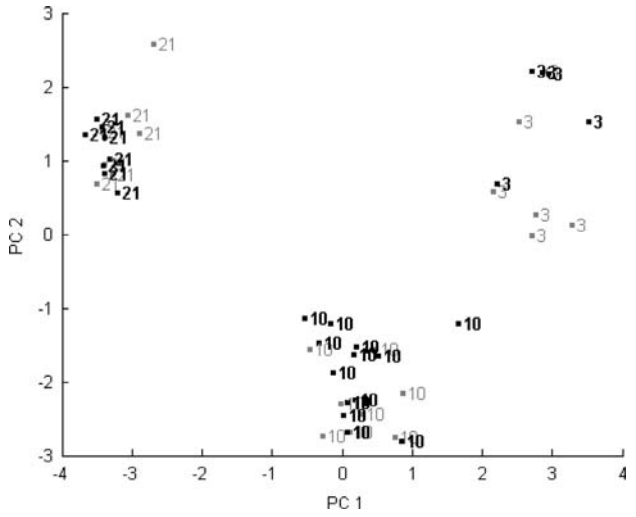

Fig. 11. PC plot of the validations set samples (black) projected onto the PC map developed from the 21 IR spectra and 23 wavelet coefficients identified by the pattern recognition GA. 3=Bowling Green KY, 10=Hamtramck MI, and 21=Orion MI.

Table 7
LDA results for plant group 2.

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

plants comprising the second plant group (see Table 6). Each IR spectrum is represented as a point in the plot. All 3 manufacturing plants (Bowling Green KY, Hamtramck MI, and Orion MI) are well separated from each other in the PC plot of the data. Fig. 11 shows the validation set samples (see Table 6) projected onto the PC map developed from the 22 IR spectra of the training set and the 23 wavelet coefficients identified by the pattern recognition GA. Each projected sample lies in a region of the map with paint samples that have the same class label.

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

Fig. 12 shows a plot of the two largest principal components of the 82 training set samples and the 9 wavelet coefficients identified by the pattern recognition GA for manufacturing plants comprising the third plant group (see Table 8). IR spectra from two manufacturing plants represented as 9 and 17 (Fremont CA and Lordstown OH) and trucks from Oshawa Ontario (Plant 22) cluster in distinct regions of the PC map of the data. GMC vehicles from Oklahoma City (Plant 20) cluster in the same region of the map with Buicks from Oshawa Ontario (Plant 22). Vehicles from Linden NJ (Plant 16), trucks from Flint MI (Plant 6), trucks from Shreveport

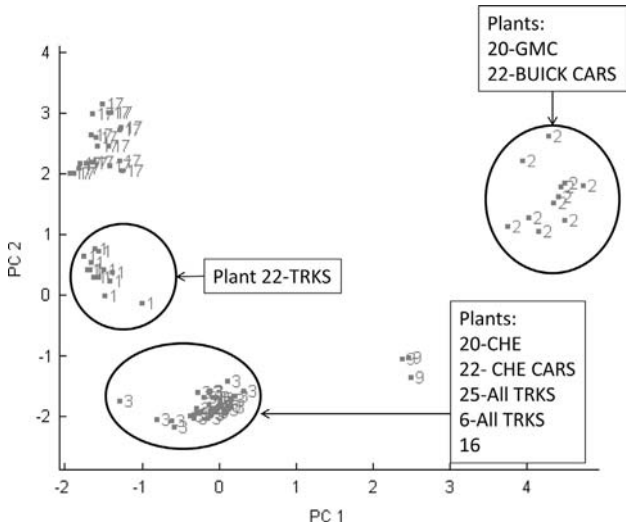

Fig. 12. A plot of the two largest principal components of the 9 wavelet coefficients identified by the pattern recognition GA for manufacturing plants comprising the third plant group is shown. 2=GMC (Oklahoma City) and Buick (Oshawa Ontario); 3=Trucks (Flint MI), CHE and GMC (Linden NJ), Chevrolet (Oklahoma City), Chevrolet (Oshawa Ontario), and GMC (Shreveport LA); 9=Fremont CA, 17=Lordstown OH.

Table 8
Training set and validation set for manufacturing plants from plant group 3.

Plants	Training set samples (Thermo-Nicolet)	Validation set samples (Bio-Rad)
9	3	2
17	19	16
1 (Plant 22 trucks)	13	9
2 (Plants 20-General Motors Corporation, 22-Buick Cars)	11	8
3 (Plants 6-all Truck,16-all, 20-Chevrolet, 22-Chevrolet cars, 25 all Trucks)	36	37
Total	82	72

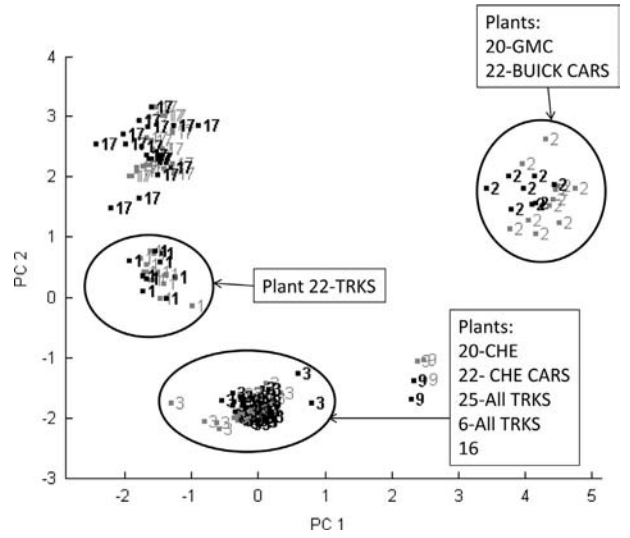

Fig. 13. PC plot of the validation set samples (black) projected onto the PC map developed from the 82 IR spectra and 9 wavelet coefficients identified by the pattern recognition GA. 2=GMC (Oklahoma City) and Buick (Oshawa Ontario); 3=Trucks (Flint MI), CHE and GMC (Linden NJ), Chevrolet (Oklahoma City), Chevrolet (Oshawa Ontario), and GMC (Shreveport LA); 9=Fremont CA, 17=Lordstown OH.

LA (Plant 25), Chevrolet cars from Oshawa Ontario (Plant 22), and Chevrolet trucks and cars from Oklahoma City (Plant 20) also form a distinct sample cluster. Two manufacturing plants, Oshawa Ontario (Plant 22), and Oklahoma City (Plant 20), have clear coat spectra that are distinct for specific models and lines. The net result is multiple clusters for automobiles or trucks from these two manufacturing plants.

Fig. 13 shows the validations set samples (see Table 8) projected onto the PC map developed from the 82 IR spectra of the training set and the 9 wavelet coefficients identified by the pattern recognition GA. Each projected paint sample lies in a region of the map with other paint samples that have the same class label.

In this study, paint samples from Plant Group 3, which were not part of the original study, were added to both the training set and validation set. These samples had been previously excluded from this study because they were identified as outliers due to excessive instrument noise. After rerunning these samples, our data for Plant Group 3 was updated and the results were compared to a previous study which did not include these samples. As the results were the same, this supported our conclusion about the quality of this data and justified our decision to rerun these samples.

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

5. Conclusions

In this study, a two step procedure for spectral library matching of clear coat paint smears from the PDQ database is proposed. First, search prefilters are employed to divide the IR spectra of the clear coats into plant groups. A genetic algorithm for pattern recognition analysis is used to identify discriminating wavelengths characteristic of each plant group. Second, search prefilters are developed for the IR spectra from each plant group to identify the specific manufacturing plant or the set of manufacturing plants that possess similar IR spectra to the unknown. In this phase of the study, wavelets are used to preprocess the data. The wavelets decompose each spectrum into wavelet coefficients which represent both the high and low frequency components of the signal. Wavelet coefficients that contain information about the manufacturing plat of the paint samples are identified using the genetic algorithm for pattern recognition analysis and feature selection. This two step procedure is able to develop search prefilters independent of the IR spectrometer used to generate the data as the IR spectra are preprocessed using the instrumental line function of the master instrument. The search prefilters have the potential to facilitate spectral library searching in the PDQ database as the size of the library is culled to those paint samples obtained from the same assembly plant as the unknown.

Acknowledgments

The authors acknowledge the work of Tamara Hodgins, Andrew Ho, and Edmund Leung, who collected, coded and entered in the PDQ database the FTIR spectra used in this study. This research was supported by Award no. 2010-DN-BX-K17 awarded by the National Institute of Justice, Office of Justice Programs, the United States Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/

exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

References

[1] A. Beveridge, T. Fung, D. MacDougall, in: B. Caddy (Ed.), Forensic Examination of Glass and Paint: Analysis and Interpretation, Taylor and Francis, NY, 2001, pp. 220–233.
[2] G.A. Bishea, J.L. Buckle, and S.G. Ryland, International forensic automotive paint database, in: Proceedings – Investigation and Forensic Science Technologies; International Society of Optical Engineering (SPIE), vol. 3576, February 1999, p. 73.
[3] J.L. Buckle, D.A. MacDougal, R.R. Grant, Can. Soc. Forensic Sci. J. 30 (1997) 199–212.
[4] N.S. Cartwright, P.G. Rodgers, Can. Soc. Forensic Sci. J. 9 (1976) 145–154.
[5] B.B. Christy, The use of the PDQ (Paint Data Query) database along with other resources to provide vehicle information for hit and run fatalities within Virginia, in: Proceedings of the Trace Evidence Symposium, August 13–16, Clearwater Beach, FL, 2007.
[6] A. Hobbs, Sifting through the layers: the application of forensic databases to tape and paint analyses, in: Proceedings of the Trace Evidence Symposium, August 13–16, Clearwater Beach, FL, 2007.
[7] B.K. Lavine, Chemometrics in the 21st Century, International Forum on Process Analytical Chemistry, Baltimore, MD, January 25 2013.
[8] B.K. Lavine, Pattern Recognition Based Library Searching Techniques for IR Spectra of Clear Coat Paint Smears, American Academy of Forensic Sciences, Washington DC, 2013. (February 19).
[9] F. Chau, Y. Liang, J. Gao, X. Shao, Chemometrics – From Basics to Wavelet Transform, John Wiley & Sons, NY, 2004.
[10] J.S. Walker, Primer on Wavelets and Their Scientific ApplicationsChapman & Hall/CRC, Boca Raton, FL, 1999.
[11] J. Karasinski, S. Andreescu, O.A Sadik, B. Lavine, M.N. Vora, Anal. Chem. 77 (2005) 7941–7949.
[12] B.K. Lavine, C.E. Davidson, W.T. Rayens, Comb. Chem. High Thru. Screen. 7 (2004) 115–131.
[13] B.K. Lavine, N. Mirjankar, Wavelets and Genetic Algorithms Applied to Spectral Pattern Recognition in Forensics, Federation of Analytical Chemistry & Spectroscopy Societies, Memphis, TN, October 16 2007.
[14] B.K. Lavine, N. Mirjankar, S. Ryland, M. Sandercock, Talanta 87 (2011) 46–52.
[15] Strengthening Forensic Science in the United States: A Path Forward, National Research Council of the National Academies Press, Washington, DC, February 2009.
[16] Y.-D. Wang, D.J. Veltkamp, B.R. Kowalski, Anal. Chem. 63 (1991) 2750–2756.
[17] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[18] T.B. Blank, S.T. Sum, S.D. Brown, Anal. Chem. 68 (1996) 2987–2995.
[19] A.J. Myles, T.A. Zimmerman, S.D. Brown, Appl. Spectrosc. 60 (2006) 1198–1203.
[20] P.G. Rodgers, R. Cameron, N.S. Cartwright, W.H. Clark, J.S. Deak, E.W. W. Norman, Can. Soc. Forensic Sci. J. 9 (1976) 1–14.
[21] P.G. Rodgers, R. Cameron, N.S. Cartwright, W.H. Clark, J.S. Deak, E.W. W. Norman, Can. Soc. Forensic Sci. J. 9 (1976) 49–68.
[22] P.G. Rodgers, R. Cameron, N.S. Cartwright, W.H. Clark, J.S. DeakE.W.W. Norman, Can. Soc. Forensic Sci. J. 9 (1976) 103–111.
[23] B.K. Lavine, A.J. Moores, L.K. Helfend, J. Anal. Appl. Pyrolysis 50 (1999) 47–62.
[24] B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, H.T. Mayfield, Anal. Chem. 72 (2000) 423–431.
[25] B.K. Lavine, D. Brzozowski, A.J. Moores, C.E. Davidson, H.T. Mayfield, Anal. Chim. Acta 437 (2001) 233–246.
[26] B.K. Lavine, C.E. Davidson, A.J. Moores, P.R. Griffiths, Appl. Spectrosc. 55 (2001) 960–966.
[27] B.K. Lavine, C.E. Davidson, A.J. Moores, Chemom. Intell. Lab. Syst. 60 (2002) 161–171.
[28] B.K. Lavine, C.E. Davidson, A.J. Moores, Vib. Spectrosc. 28 (2002) 83–95.
[29] B.K. Lavine, C.E. Davidson, C. Breneman, W. Katt, J. Chem. Inf. Sci. 43 (2003) 1890–1905.
[30] B.K. Lavine, C.E. Davidson, D.J. Westover, J. Chem. Inf. Comput. Sci. 44 (2004) 1056–1064.
[31] G.A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F.K. Tadjimukhamedov, B.K. Lavine, N. Mirjankar, Anal. Chim. Acta 579 (2006) 1–10.