On the effect of data fusion methods for the classification of colored automotive paint mixtures using visible microspectrophotometry and Raman spectroscopy

Morgan Carpenter*, B.S.; Patrick Buzzini, Ph.D.

Department of Forensic Science, Sam Houston State University, Huntsville, TX 77340

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

INTRODUCTION

Comparative examinations of paint typically involve the use of an analytical sequence of microscopical examinations and various instrumental analysis methods.

In the event of a non-differentiation, and in order to address the question of a common source between compared sets, the expert should estimate the chance of observing in the population of interest another coated object with the properties detected with method A *and* the properties detected with method B *and* the properties detected with method C *and* so on.

Exploring data fusion methods for this purpose is essential because:

- They offer the potential to use the detected properties simultaneously to address the question of interest.
- They can detect and manage redundancy and dependency of some of the detected features (i.e., color and pigments).

In this study, the interest is in exploring if the way to combine high dimensional data collected from automotive paint samples of known binary pigments compositions and proportions, using two different analytical techniques, impacts the classification accuracy of a classical chemometric method.

The question arises because repeated measurements with both methods are not taken on the same area of the specimen simultaneously. Hence, it is hypothesized that a sequential data acquisition process constitutes a source of variation due to the heterogeneity of automotive paint.

MATERIALS AND METHODS

Samples:

- Known proportions of Standox Standohyd Basecoat Automotive Refinish Paint from the base colors red (C.I. *Pigment Red 254*) and yellow (C.I. *Pigment Yellow 184*), blue (C.I. *Pigment Blue 15*) and yellow and and blue and green (C.I. *Pigment Green 36*) were mixed in the following proportions: 70%-30%; 50%-50% and 30%-70% by weight.

Analytical methods:

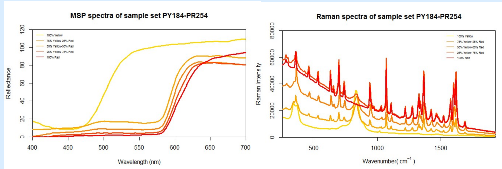
- Visible MSP analysis (400-700nm) in reflectance mode using the Video Spectral Comparator 6000 by Foster & Freeman.
- Raman analyses using an XploRA Raman microscope (50x obj.) from Horiba Scientific using a NIR laser source at 785nm and a spectral range of 2000-250cm^{-1}.
- Seven replicates were taken at random points on each sample.

Data Analysis:

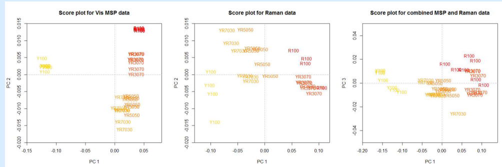
- 5 datasets of MSP and Raman from each sample set were combined as follows:
  - 1 set by side-by-side binding (Comb0)
  - 4 sets by randomly assigning replicates within each proportion (CombR1 through R4)
- Due to differing scales between the MSP and Raman signals the normalization method of area unity was used.
- Principal component analysis (PCA) was used as a data reduction technique and to verify group separations.
- Linear discriminant analysis (LDA) was conducted on the 3 first principal components as new variables; 4 replicates for each mixture set were extracted as a training set and the 3 remaining ones were used as a testing set. Assignments of training sets and testing sets was kept the same for all compared datasets.
- Random assignments for datasets binding, PCA and LDA were computed using R Statistical software (1) with the integrated MASS package (2).

RESULTS

C.I. *Pigment Yellow 184* – C.I. *Pigment Red 254* Sample Sets



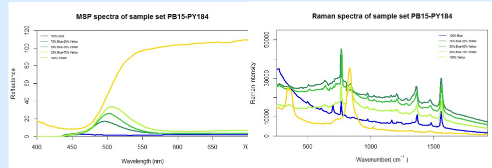
- Vis MSP spectra show high contribution of the red pigment even if present at lower proportions compared with the yellow pigment. However, the 5 sets of spectra exhibit variation around 470nm and 600-650nm.
- Like MSP, Raman spectra also show high contribution of the red pigment. However, note the prominent band of PY 184 around 850cm^{-1} present in all mixtures.



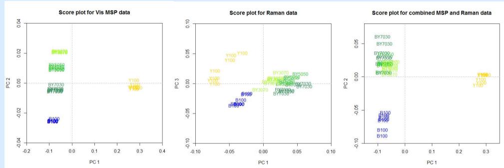
- The PC1-PC2 score plots of the Vis MSP data show distinct separation between the mixed sets except for YR 70:30 and 50:50. Raman data exhibit more overlapping between the 100% red pigment and the the YR 30:70 proportion which is in line with the full spectra above. These overlaps are present in the PC1-PC3 plot for the side-by-side combined datasets (Comb0).
- Although an expected lower correct classification rate observed for Raman, the way to combine the datasets of the two techniques did not prevent a correct classification for all spectra.

PCA+LDA classification prediction accuracy						
MSP	Raman	Comb0	CombR1	CombR2	CombR3	CombR4
100%	73%	100%	100%	100%	100%	100%

C.I. *Pigment Blue 15* – C.I. *Pigment Yellow 184* Sample Sets



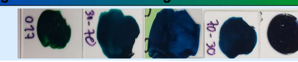
- Vis MSP spectra of these 5 sets show clear differences. Note the low reflectance percentage from the blue pigment which is indeed very dark.
- The major bands of the blue phthalocyanine pigment dominate Raman spectra at all mixed proportions. The dominant band of PY 184 is still perceptible in the spectra of mixed paints.

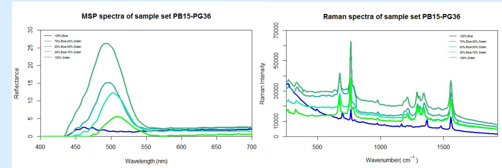


- The PCs plots of the MSP data show distinct groups. Raman data exhibit closeness between the mixed group in the projected plane, however the groups are still discernible. This latter observation applies to the side-by-side combined datasets (Comb0) where a more accentuated overlapping can be observed.
- For this sample set, different accuracy rates are observed for two of the four randomly combined datasets. Note that the highest accuracy is observed for the side-by-side combined datasets.

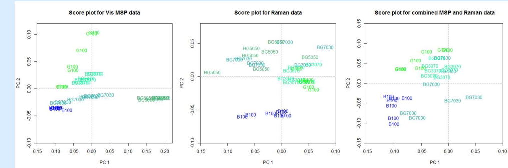
PCA+LDA classification prediction accuracy						
MSP	Raman	Comb0	CombR1	CombR2	CombR3	CombR4
100%	100%	100%	100%	100%	87%	93%

C.I. *Pigment Blue 15* – C.I. *Pigment Green 36* Sample Sets





- Vis MSP spectra of the 5 sets can be easily differentiated. The overall low reflectance percentages from both the dark blue and green pigments can be noticed.
- Both the detected blue and green pigments are phthalocyanines which generate few bands in common. However, the green pigment is more predominant at lower concentrations and produces a larger number of Raman bands.



- The PCs plots of the MSP data show again distinct groups. Overlaps are detected for Raman data between spectra of mixtures BG 50:50 and 70:30 as well as 100% G and BG 30:70 confirming the high contribution of the green component. Such overlaps are present within the side-by-side combined datasets (Comb0).
- The same accuracy rates are observed for the four randomly combined datasets which are slightly lower than the correctly classified side-by-side combined datasets. The low accuracy observed for Raman data is in line with the observed spectral features.

PCA+LDA classification prediction accuracy						
MSP	Raman	Comb0	CombR1	CombR2	CombR3	CombR4
100%	60%	100%	93%	93%	93%	93%

DISCUSSION

- Visible MSP and Raman were chosen due to the dependencies of the detected analytical information related to the pigment components. A disparity of the number of variables existed between the datasets of the two methods: 300 variables for MSP and 1607 for Raman data. In this case the number of variables was not a factor contributing to inter-source variations.
- The chemometric approach (PCA+LDA) was not optimized to obtain a fine-tuned accuracy rate (i.e., no crossvalidation); the goal was to avoid the effect of differential random selection of training and testing sets. A further phase of this project will investigate a more rigorous fine-tuning approach.
- The choice of the sample sets (i.e., mixtures) was intended to span different levels of variations between the pure colors and the mixed proportions: this aspect was more accentuated for MSP in the sense that the resonance effects observed in Raman produce spectra of mixed components that are dominated by the signal of a single pigment even if present in lower proportions.
- Mixture proportions were chosen to avoid challenging the sensitivity of the chosen classifier to assign groups of mixed paints at closer proportions (i.e., including 90%-10% and 80%-20% between 100% and 70%-30% used in this study).
- The response on the way to combine datasets from the two analytical methods differed between the three colored sample sets: in the present setting, the classifications of MSP data and of side-by-side combined datasets have always resulted in 100% accuracy for all sample sets.
- In this study only five sample sets of combined data were tested: a higher number of randomly assigned combinations could be tested, and different number of replicate measurements could be considered as well (e.g., 5 and 10 instead of the 7 used in this study).

CONCLUSIONS

- This study was concerned with the way that datasets obtained from the same samples using two different analytical techniques should be combined for a simultaneous use while addressing questions on source attributions in the context of comparative examinations of paints.
- Out of comparisons involving 5 combined datasets – 1 side-by-side combined dataset and 4 randomly assigned combinations of datasets – differences were observed between the three chosen sample sets consisting of mixtures of pairs of colored paint of known composition and known proportions.
- Future studies envision the introduction of a third technique (e.g., element ratios from SEM-EDS data) which is expected to be useful while combined with the two techniques used in this study: it would be useful to combine methods with assumed lower discriminating abilities like Raman and SEM-EDS to avoid that the higher discriminating ability that MSP has demonstrated in the present study could eclipse the effect of combining datasets.

REFERENCES

1) R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.
2) Venables WN, Ripley BD. Modern Applied Statistics with S (4th Ed.). Springer, New York, NY (2002).