Estimation of hydrocarbon types in light gas oils and diesel fuels by ultraviolet absorption spectroscopy and multivariate calibration

Peter D. Wentzell, Darren T. Andrews, Jennifer M. Walsh, Jean M. Cooley, and Paige Spencer

Abstract: Ultraviolet absorption spectra are used in conjunction with multivariate calibration methods to predict the percent content of saturates (61–99%), monoaromatics (1–34%), diaromatics (0–5%), and polyaromatics (0–1%) in light gas oil and diesel fuel samples. A total of 114 samples taken from three pilot plant studies between 1992 and 1994 were used to develop the calibration models. Supercritical fluid chromatography with flame ionization detection was employed as the reference method for quantifying hydrocarbon types. Several multivariate calibration methods (classical least squares, multiple linear regression, principal components regression (PCR), and partial least squares (PLS)) were examined. PCR and PLS gave the best overall performance, with root-mean-squared errors of prediction (absolute) of 0.9% (saturates), 0.8% (monoaromatics), 0.2% (diaromatics), and 0.05% (polyaromatics) based on 44 prediction samples. Relative errors (absolute × 100/range) for these same species were 2.4%, 2.4%, 3.2%, and 5.3%.

Key words: light gas oil, diesel, hydrocarbon content, UV spectroscopy, multivariate calibration, aromatics, petroleum.

Résumé : On a utilisé conjointement les spectres d’absorption ultraviolets et des méthodes de calibration à multivariables pour prédire le pourcentage de produits saturés (61–99%), monoaromatiques (1–34%), diaromatiques (0–5%) et polyaromatiques (0–1%) contenus dans des échantillons de diesel et de gazole léger. Pour développer les méthodes de calibration, on a utilisé cent quatorze (114) échantillons recueillis à partir de trois usines pilotes entre 1992 et 1994. On a fait appel à la chromatographie fluide superfliclue avec détection à ionisation de flamme comme méthode de référence pour la quantification des divers types d’hydrocarbures. On a étudié plusieurs méthodes de calibration à multivariables (moindres carrés classiques, régression linéaire multiple, régression des principaux composants (RPC) et moindres carrés partiels (MCP). Les méthodes RCP et PLS donnent les meilleures performances globales avec des erreurs quadratiques moyennes de prédiction (absolues) de 0,9% (saturées), 0,8% (monoaromatiques), 0,2% (diaromatiques) et 0,05% (polyaromatiques) sur la base de quarante-quatre (44) échantillons de prédiction. Pour les mêmes espèces, les erreurs relatives (absolues × 100/écart) sont respectivement 2,4%, 2,4%, 3,2% et 5,3%.

Mots clés : gazole léger, contenu en hydrocarbures, spectroscopie UV, calibration à multivariables, aromatiques, pétrole.

[Traduit par la Rédaction]

1. Introduction

Transportation fuels such as gasoline, diesel, and jet fuels must meet detailed specifications for their quality. Such specifications ensure, for example, that the fuel will be suitable for the engine for which it is intended, will be stable for a reasonable time frame, and will meet certain environmental standards. Some of the analytical tests developed to monitor adherence to these specifications are rapid, simple, and require very little sample, while other assays require a more significant amount of fuel and specialized test equipment. Obviously, it is desirable to replace the latter test with the former wherever possible, not only to reduce costs and delays in analysis, but also to facilitate more efficient quality control at the processing stage.

One specification for transportation fuels that is becoming more important is the quantity of aromatic compounds present. This is often measured by fluorescent indicator analysis (FIA) (ASTM D1319 or UOP 501), a classical technique that is time consuming and, for the higher boiling samples (e.g., diesel), is difficult to perform reliably. A more recently developed method to measure aromatic content is based on analysis by supercritical fluid chromatography (SFC) (1, 2). In addition to being more rapid and reliable, this technique has the advantage of being able to speciate aromatics into mono-, di- and polyaromatics. However, it requires more...
complex instrumentation than techniques like FIA and is not available in many labs. Other techniques, such as high-performance liquid chromatography, mass spectrometry, and nuclear magnetic resonance spectroscopy, have also been used to determine aromatic content, but these methods face similar drawbacks.

An alternative approach to predict the aromatic content, as well as other properties of refinery products, is the use of spectroscopic techniques, and such methods have been widely reported in the literature. Near-infrared. Fourier transform infrared (FTIR), and Raman spectroscopy have been used most extensively for this purpose, especially with gasoline, aviation fuel, and other light products (3–10). Typically, these spectroscopic methods are used in conjunction with a variety of multivariate calibration techniques that can extract the desired information from the mixtures. The general consensus of these studies is that spectroscopy can be used as a rapid and simple means to predict certain properties of hydrocarbon mixtures provided that a training set of samples that is representative of future samples is available.

Fewer studies on the spectroscopic prediction of properties from heavier refinery products such as diesels have been carried out (11, 12). Quantitation of the hydrocarbon types in diesel fuels is becoming more important since future regulations on diesel properties may make it necessary to reduce the aromatic content in order to meet fuel specifications. A few studies have examined the application of ultraviolet absorption spectroscopy to the prediction of diesel and gas oil properties (13–15). These studies have been limited in scope and often restricted to a few wavelengths, without taking full advantage of multivariate calibration methods.

In this paper, a method for the spectroscopic determination of hydrocarbon types in light gas oils and diesel fuels is evaluated. The present study can be distinguished from many of the earlier works in three ways: (i) it utilizes ultraviolet (UV) absorption measurements as opposed to the more commonly employed IR or NIR measurements, (ii) it is applied to middle distillates as opposed to gasoline, (iii) aromatic content is subdivided into mono-, di-, and polyaromatics (saturated hydrocarbon content is also determined), and (iv) a large number of chemically diverse samples with a wide range of aromatic content were available for calibration. Five methods of multivariate calibration are applied to the spectroscopic data and compared for their ability to predict the four properties of interest.

2. Multivariate calibration techniques

Multivariate calibration techniques are now widely used in many practical applications, not the least of which is the petroleum industry. Whereas traditional univariate calibration relies on a single scalar measurement for each sample, multivariate calibration methods utilize a vector (or matrix) of data for each sample, such as a spectrum, which permits analysis in the presence of multiple interferences. In general, the object is to obtain an \( nx1 \) vector of regression coefficients, \( b \), such that

\[
\hat{y}_{\text{unk}} = x^T b
\]

where \( \hat{y}_{\text{unk}} \) is the estimated value of the property of interest (e.g., % monoaromatics) for an unknown sample and \( x \) is an \( nx1 \) vector of responses for the sample measured at \( n \) channels (e.g., the spectrum measured at \( n \) wavelength channels). (In this paper, scalars will be indicated by lower case italic symbols, vectors by boldface lower case symbols, and matrices with boldface upper case symbols. The superscript "\( T \)" will be used to indicate the transpose of a matrix or vector.) A wide range of calibration techniques are now available, with the main differences being the methods that they employ to determine the regression vector, \( b \).

In most cases, the calibration procedure is as follows. First, a set of samples is obtained that is representative of the range of compositions that are likely to be encountered in practice. Spectra (or other vectorial measurements) are then recorded for each sample. The property of interest is also measured for each of these samples by an independent reference method (in this work, supercritical fluid chromatography). The samples are then divided into two groups, one that is used to build the calibration model (the training or calibration set) and another that is used to independently assess the performance of the model (the validation or prediction set).

This study focuses on the evaluation of a few key multivariate calibration methods for the problem of interest. These are: (i) classical least squares (CLS), (ii) forward selection multiple linear regression (FS-MLR), (iii) stepwise selection multiple linear regression (SS-MLR), (iv) principal components regression (PCR), and (v) partial least-squares regression (PLS). These techniques have been reviewed in the literature (16–21), but will be described briefly here in the context of the present study.

2.1. Classical least squares

Classical least squares (CLS), also known as the K-matrix method, assumes that the observed responses are linearly related to the component concentrations and that the concentrations of all components in the mixtures are known for the calibration set. The most common example of this is the multivariate form of Beer's Law, which relates a matrix of spectra, \( X \), to a matrix of concentrations, \( Y \), through:

\[
[2] \quad X = KY
\]

Here, \( X \) is an \( m \times n \) matrix of mixture spectra (\( m \) samples measured at \( n \) wavelength channels), \( Y \) is an \( m \times p \) matrix of component concentrations (\( p \) components for each of \( m \) samples), and \( K \) is the \( p \times n \) matrix of pure component spectra (molar absorptivities, assuming a unit path length). Unless pure component spectra are known, \( K \) is normally calculated from the calibration data by solving the system of linear equations. From there it is a straightforward procedure to obtain the regression vectors for each component.

CLS can be an effective method when all components in the mixture contributing to the spectra have been identified and can be quantitated, but it is less useful when these conditions are not met because the model is incomplete. Furthermore, the model really only applies when the property of interest (e.g., concentration) can be associated with a component with a clearly defined spectrum, and its utility for more general properties (e.g., cetane number) is questionable. In the present study, it might be argued that the properties measured (saturates and aromatics) approximate a
complete set of "components" for UV spectroscopy, so the potential of CLS was explored.

2.2. Multiple linear regression

Perhaps the most serious deficiency of CLS is that it requires a knowledge of all of the components in the mixture, while there may be only a few properties of interest. Multiple linear regression (MLR), sometimes called the P-matrix method, circumvents this problem by assuming a model of the form:

\[
Y = XP
\]

where \( Y \) is the \( m \times p \) matrix of properties to be estimated (\( p \) properties for \( m \) samples), \( X \) is the \( m \times n \) matrix of spectra, and \( P \) is the \( n \times p \) matrix of regression coefficients. This is sometimes referred to as the "inverse model," since \( X \) and \( Y \) are reversed from the CLS model in eq. [2], and MLR is sometimes referred to as inverse least squares. A distinct advantage of MLR is that it permits the determination of the regression vector \( (b \) in eq. [1]) for a desired component or property without requiring a complete knowledge of all components. It is therefore possible to apply this equation using only reference values for the property of interest. Unfortunately, reliable estimation requires that \( m \), the number of samples in the calibration set, is substantially greater than \( n \), the number of wavelength channels. Since modern spectrometers can have hundreds to thousands of channels, this condition is seldom met and it is necessary to select a much smaller subset of the wavelength channels. The channel selection is key, and in this work two common methods of variable selection were used: forward selection (FS) and stepwise selection (SS).

Forward selection is a screening procedure that selects the best variables (wavelength channels in this case) and adds them to the regression model one at a time until no additional improvement in model performance is found (17). This point is often determined using a standard F-test, but with a large number of wavelength channels, this test breaks down due to the statistical problem of multiple testing (17, 22). In this work, the alternative approach of internal validation was used. In this procedure, the calibration set is further subdivided into two parts, one that is used to build the models (calibration subset) and the other that is used to determine when to stop adding variables (prediction subset). Note that the latter is not the same as the prediction set described earlier, since the calibration procedure must remain blind to the data set used for validation. Variables are added to the calibration model by FS using the calibration subset as described above. For each new set of variables, the regression vector is calculated and the prediction subset is used to calculate the root-mean-squared standard error of prediction (rmsep) from:

\[
\text{rmsep} = \left\{ \frac{\sum_{i=1}^{n_{\text{pred}}} (y_{\text{pred}} - x_{\text{pred}}b)^2}{n_{\text{pred}}} \right\}^{1/2}
\]

where \( n_{\text{pred}} \) is the number of samples in the prediction subset. Note that in this case, the rows of \( X \) do not contain the full spectra of the prediction samples, only those variables selected by the FS procedure. The rmsep is a measure of the predictive ability of the model and will generally pass through a minimum when plotted against the number of variables added to the model, indicating that additional variables are overfitting. The minimum is used to determine the optimum number of wavelength channels. The FS procedure is then repeated using the full calibration set and terminating at this number of variables.

The technique of stepwise selection is fundamentally the same as FS for the addition of variables, but it allows one to check the significance of previously added variables after a new one is added (17). As with FS-MLR, the internal validation procedure was used to select the optimum number of variables for SS-MLR.

2.3. Principal component regression

Principal component regression (PCR) has emerged as one of the tools of choice for the multivariate calibration problems in chemistry. In PCR, the overall dimensionality of the data is reduced to only a few variables, thereby addressing the rank deficiency problem inherent in MLR. Unlike the variable selection procedures used for MLR, variable reduction in PCR occurs by combining those variables that exhibit strong correlations into latent variables, so information is not lost through the elimination of wavelength channels. The first step in PCR is to decompose the \( m \times n \) matrix of calibration responses, \( X \), using principal component analysis (PCA) to give scores and loadings matrices:

\[
X = TL
\]

Assuming that \( m < n \), PCA returns the scores matrix, \( T \) (\( m \times m \)), and the loadings matrix, \( L \) (\( m \times n \)). In order to retain \( p \) latent variables for the regression, these matrices are then truncated by removing the right-hand columns and bottom rows to give \( \tilde{T} \) (\( m \times p \)) and \( \tilde{L} \) (\( p \times n \)). The latent variables selected in this way form an orthogonal set and are the linear combinations of the original variables that account for the largest amount of variance in the data. Because of this, these variables are likely to be useful in developing a model. The matrix \( \tilde{T} \) is analogous to the matrix \( X \), except that it has fewer columns, making it more readily employed in MLR. The corresponding regression model is:

\[
y = \tilde{T}q
\]

where \( q \) is the \( p \times 1 \) vector of regression coefficients. Solution of eq. [6] for \( q \), the regression vector for the scores, and its transformation to \( b \), the regression vector in terms of the original variables, is straightforward.

An important consideration in PCR is the number of latent variables used in the regression. In this work, this parameter was determined by leave-one-out cross-validation. In this procedure, the calibration is carried out using all but one of the calibration samples. This sample is then predicted using the model determined and the difference between the predicted and reference values of \( y \) is determined. This process is repeated \( m \) times, each time leaving out a different calibration sample. All of the prediction errors are then combined to give the root-mean-squared error of cross-validation (rmsecv):

\[\text{rmsecv} = \text{sqrt}(\frac{\sum_{i=1}^{m} (y_{\text{cal}} - y_{\text{pred}})^2}{m})\]
Here $y_i$ is the observed value of the property of interest for calibration sample $i$, $\hat{y}_i$ is the predicted value of the property, and $m$ is the number of calibration samples. A plot of the $\text{rmsecv}$ vs. the number of latent variables should pass through a minimum at the optimum number of latent variables. This approach is similar to the internal validation procedure used for variable selection in MLR, but is often considered more reliable because more samples are used to build the model. For MLR, cross-validation cannot generally be used for variable selection, since a new group of variables may be selected on each iteration of the process.

### 2.4. Partial least-squares regression

One weakness of PCR is that it extracts the latent variables, which are linear combinations of the original variables, on the basis of the amount of variance in $X$ that they encompass. Therefore, wavelength channels with large variations in response over the calibration set tend to be weighted most heavily. However, these wavelengths are not necessarily the ones that are most useful in predicting the desired property. To mitigate this problem, the technique of partial least-squares regression (PLS) can be used (16). In recent years, PLS has become one of the most widely used regression techniques in chemistry. In many instances, PCR and PLS produce comparable results, but in certain cases PLS will exhibit superior performance and often can produce a model with fewer latent variables.

Regression with PLS is very similar to PCR, with latent variables being extracted from the calibration data and used in the regression model. However, while PCR extracts the latent variables to maximize the variance in $X$, PLS generates the latent variables by maximizing the covariance between $X$ and $y$. In this way, variables that have good predictive abilities are given a greater weight. The latent variables, or scores, in PLS can be calculated by a variety of algorithms (16), the details of which will not be discussed here. As with PCR, it is necessary to determine the optimum number of latent variables by cross-validation.

### 3. Experimental

#### 3.1. Data

There were a total of 115 samples from pilot plant studies available for study in this work. These consisted of two sets of light gas oil samples (boiling range 195–343°C) and one set of diesel samples (boiling range 177–343°C). Fifty-nine light gas oil samples were obtained from a pilot plant program conducted at Syncrude Canada Ltd. to evaluate three different catalysts. The project consisted of 20 mass balances and included temperatures ranging from 320 to 400°C, pressures ranging from 6.9 to 13.8 MPa, and liquid hourly space velocities ranging from 0.7 to 1.3 mL L⁻¹ min⁻¹. A further 26 samples were obtained from a subsequent program, which used a different feed as well as different catalysts. Finally, 30 diesel samples were obtained from a third pilot study that involved very severe treating of light gas oil samples with an aim of significantly reducing aromatic content.

The use of three separate pilot plant studies ensured chemical diversity in the samples studied.

The reference values for percentages of saturates, monoaromatics, diaromatics, and polyaromatics were determined by supercritical fluid chromatography (1) using an HP model G1205A supercritical fluid chromatograph (Hewlett-Packard, Palo Alto, Calif.) equipped with an autosampler. The oils were separated neat on a 250 x 4.6 mm HP hydrocarbon group separation column using carbon dioxide as the carrier (2 mL min⁻¹). The oven temperature was 35°C and the outlet pressure was 150 bar. Quantitation was performed with a flame ionization detector (FID). To obtain the weight percentages of each of the four groups, the FID signal was integrated over ranges specified by the CGSB for each group. This area was then ratioed to the total area for the chromatogram. The ranges obtained for the samples were as follows: saturates, 60.6–99.0%; monoaromatics, 0.8–34.1%; diaromatics, 0–5.0%; polyaromatics, 0–0.9%. The wide range of concentrations is an asset in the construction of reliable models.

For the spectral analysis, approximately 0.12 g of each sample was accurately weighed into a 25 mL volumetric flask and diluted with HPLC grade cyclohexane. This solution was further diluted by a factor of 1/200 with cyclohexane before running the spectra. (In preliminary studies, higher concentrations had also been employed, but the lower concentrations were found to be more useful.) Spectra were obtained on a Cary 3 UV–visible spectrophotometer (Varian Instruments, San Fernando, Calif.) in a 1 cm flow-through cell. Scanning parameters were: spectral band width, 0.5 nm; signal averaging time, 0.1 s; data interval, 0.35 nm; scan rate, 21 nm min⁻¹; range, 200–400 nm. There were a total of 572 points acquired for each spectrum. Each spectrum was normalized to an initial concentration of 0.1 g/25 mL by multiplying the absorbance readings by 0.1/(sample mass).

#### 3.2. Data pretreatment

The data for the original 115 samples were first analyzed to identify the presence of outliers using the procedure outlined by Martens and Naes (16). In this procedure, samples that are highly influential (high leverage) and have large studentized residuals are flagged as outliers. Only one sample was identified as an outlier and eliminated from subsequent analysis. This corresponded to a feed sample that was expected to have a considerably different composition from the other samples. The UV spectra for the remaining 114 samples used in this work are shown in Fig. 1.

Prior to regression analysis, the spectra were adjusted by subtracting the mean absorbance at each wavelength from all of the measurements at that wavelength. This preprocessing step, called column mean-centering, is commonly employed with spectroscopic data to eliminate intercept terms from the model. Likewise, each of the four groups of reference measurements was centered about its mean value. In the final prediction, a correction for the offset was made.

For the regression analysis, it was necessary to divide the full data set into a calibration set and a prediction set for purposes of validation. For this work, the sizes of the calibration and prediction sets were chosen to be 70 and 44 samples, respectively. To ensure that each set is representative of the other, the subset selection procedure of Kennard

© 1999 NRC Canada

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
and Stone was used (23) to split the data. Visual verification of the representativity was also made by examining two-dimensional projections of the samples after PCA. For FS-MLR and SS-MLR, further splitting of the calibration set was necessary for internal validation, giving a calibration subset of 50 samples and a prediction subset of 20 samples. The same method was used for obtaining these subsets.

3.3. Computational aspects

All of the results presented here were obtained on a Pentium-based personal computer using programs written in Matlab v4.2x (The Math Works, Natick, Mass.). All programs were written by the authors, except for the PLS routines, which were taken from the PLS Toolbox for Matlab (Eigenvector Research Inc., Manson, Wash.). Computation times for building the models varied with the regression technique, but all were under an hour. Once the regression coefficients are computed, calculation of results for new samples is computationally trivial.

4. Results and discussion

Initially, the feasibility of using UV spectroscopic data to predict aromatic content was evaluated by calculating the correlation coefficient between the absorbance at each wavelength and the reference value for each of the four properties to be predicted. The data were mean-centered for these calculations. Figure 2 shows the correlation coefficient plotted as a function of wavelength for the four properties. Each of the properties exhibits regions where there is a high degree of positive or negative correlation, indicating that these data may be useful in the development of a regression model. For the three classes of aromatics, correlation is high at short wavelengths and the maximum shifts to longer wavelengths with increasing aromaticity. This is consistent with the observed transitions for aromatic compounds and the bathochromic shift that occurs with increasing conjugation. For saturates, there is a strong negative correlation in this region, since these compounds do not absorb in this part of the spectrum and the content is essentially correlated with the absence of aromatics. For all constituents, the degree of correlation falls off with increasing wavelength until it is effectively zero above 350 nm.

Despite the high degree of correlation found, attempts to use single wavelengths to model the properties of interest were unsuccessful, generally resulting in nonlinear plots. This led to the application of the multivariate methods described above.

Table 1 summarizes the results obtained by the various calibration methods. The table is divided by constituent and calibration method and gives the prediction errors in each case, along with other relevant parameters. The errors presented in the table are rmsep values obtained using the prediction set of 44 samples, which were completely excluded from the calibration process. Therefore, these values should be good estimates of the true predictive ability of each model in this sample domain. Both absolute errors (% content of constituent) and relative errors (% of range of predicted value) are shown. The relative error is often a better means of evaluating the predictive ability of the model since it takes into account the range of values to be predicted. Also given in the table are the number of wavelength channels selected, the corresponding wavelengths (MLR), and the number of latent variables used (PCR and PLS).
Fig. 2. Correlation of properties to be estimated with absorbance at each wavelength: (a) % saturates, (b) % monoaromatics, (c) % diaromatics, (d) % polyaromatics. Data were column mean-centered prior to calculation.

Table 1. Comparison of calibration methods for the prediction set (44 samples) for each of the compound classes.

<table>
<thead>
<tr>
<th>Property</th>
<th>Method*</th>
<th>Absolute error*</th>
<th>% Relative error*</th>
<th>Wavelengths used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturates</td>
<td>CLS</td>
<td>1.24</td>
<td>3.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FS-MLR (12)</td>
<td>1.27</td>
<td>3.32</td>
<td>204 211 327 346 360 371 373 380 385 394 398</td>
</tr>
<tr>
<td></td>
<td>SS-MLR (12)</td>
<td>1.27</td>
<td>3.32</td>
<td>As for FS-MLR</td>
</tr>
<tr>
<td></td>
<td>PCR (10)</td>
<td>0.94</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLS (6)</td>
<td>0.93</td>
<td>2.42</td>
<td></td>
</tr>
<tr>
<td>Monoaromatics</td>
<td>CLS</td>
<td>1.35</td>
<td>4.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FS-MLR (2)</td>
<td>0.78</td>
<td>2.33</td>
<td>201 319</td>
</tr>
<tr>
<td></td>
<td>SS-MLR (2)</td>
<td>0.78</td>
<td>2.33</td>
<td>201 319</td>
</tr>
<tr>
<td></td>
<td>PCR (10)</td>
<td>0.8</td>
<td>2.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLS (9)</td>
<td>0.81</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>Diaromatics</td>
<td>CLS</td>
<td>0.226</td>
<td>4.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FS-MLR (9)</td>
<td>0.184</td>
<td>3.67</td>
<td>200 204 208 223 246 261 270 283 308</td>
</tr>
<tr>
<td></td>
<td>SS-MLR (4)</td>
<td>0.292</td>
<td>5.85</td>
<td>246 261 270 308</td>
</tr>
<tr>
<td></td>
<td>PCR (6)</td>
<td>0.159</td>
<td>3.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLS (4)</td>
<td>0.16</td>
<td>3.21</td>
<td></td>
</tr>
<tr>
<td>Polyaromatics</td>
<td>CLS</td>
<td>0.134</td>
<td>14.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FS-MLR (1)</td>
<td>0.058</td>
<td>6.4</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>SS-MLR (1)</td>
<td>0.058</td>
<td>6.4</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>PCR (5)</td>
<td>0.051</td>
<td>5.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLS (5)</td>
<td>0.048</td>
<td>5.34</td>
<td></td>
</tr>
</tbody>
</table>

*Numbers in parentheses indicate the number of wavelength channels selected (MLR) or the number of latent variables used (PCR and PLS).
*Values given are the root-mean-squared errors of prediction (rmsep) for % content.
*Calculated from (rmsep/range) × 100.

© 1999 NRC Canada

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
wavelengths have been rounded to the nearest nanometre to save space. The predictive abilities are also summarized in the bar chart in Fig. 3 for easy comparison.

All of the methods used were able to quantitate the four hydrocarbon types reasonably well, with the possible exception of the polyaromatic content, which was below 1% in all of the samples examined. For the other three groups of compounds, all of the relative prediction errors are below 6% and often below 4%. CLS performed better than expected, probably because the aromatic classes approximate a complete set of absorbers in the UV. FS-MLR and SS-MLR yielded the same wavelength subsets (and hence the same predictive abilities) in all cases except the diaromatics, with SS generating a subset of the FS case with poorer predictive ability. Only two wavelengths were selected for the monoaromatics and a single wavelength was found to be sufficient for the polyaromatics. While it is tempting to try to interpret the selected wavelengths from a chemical perspective, this is difficult to do because the wavelengths with the best predictive ability do not necessarily correspond to the peak maximum for a given class of compounds in a complex mixture.

In general, PCR and PLS performed the best of the five methods tested, but the improvement was only marginal. These two methods gave similar results but, as is common, PLS often required fewer latent variables. The regression vectors obtained by PCR and PLS can often yield information about important regions of the spectrum and suggest critical measurement regions for lower resolution instruments. These regression vectors are plotted for each of the hydrocarbon classes in Fig. 4. There appears to be good agreement in the general shape of the regression vectors (with the possible exception of the monoaromatics), which explains their similar performance. These plots suggest that shorter wavelengths are more important, as expected. Beyond this, it is difficult to say much about the shape of the curves, but it is likely that the oscillating structure of the regression coefficients is the result of regularly spaced bands for the aromatic species present. It is possible that restricting the wavelength range to shorter wavelengths prior to calibration may further improve the results, but this was not attempted in this work.

Finally, as a typical example, Fig. 5 shows results for predicted vs. measured properties for the PLS method. The plot distinguishes among calibration and prediction samples and the three different data subsets in order to convey information about the distribution of samples. Note that the correlation between measured and predicted values is very good for the first three constituents and there is no suggestion of bias. The agreement is not as good for the polyaromatics, but the figure makes it clear that one of the reasons for this is digitization error in the reference method arising from the low concentrations of these species within the samples, which is anticipated for light gas oils.

5. Conclusions

The results presented here demonstrate that UV spectroscopic measurements, in conjunction with multivariate calibration methods, can be used to predict the aromatic content of intermediate petroleum distillates with good reliability. The models were constructed and validated using two differ-
Fig. 4. PCR and PLS regression vectors: (a) and (b) % saturates, (c) and (d) % monoaromatics, (e) and (f) % diaromatics, (g) and (h) % polyaromatics.
Fig. 5. Comparison of predicted vs. measured properties using the PLS regression model. △ 1992 pilot plant data, □ 1993 pilot plant data, ○ 1994 pilot plant data. Filled symbols indicate prediction samples, open symbols indicate calibration samples. Note that all of the observed measurements were rounded to the nearest 0.1%.

ent sets of light gas oil samples and one set of diesel samples. In addition to different boiling ranges, the samples are also chemically different due to severe hydrotreating/ hydrocracking of the diesel samples. Therefore, this data set should represent a fairly rigorous test of the multivariate modeling methods. Furthermore, unlike some earlier studies that only quantified total aromatics, this work was able to differentiate among three different classes of aromatic compounds. With the exception of polyaromatic content, which suffered from imprecision in the reference method, relative prediction errors for PCR and PLS were around 3%, which is quite good for this type of analysis and comparable to current methods for diesel aromatics. The absolute errors were similar to the repeatability specifications for the SFC method, suggesting that uncertainty in the reference method may be a significant contributor to the prediction error.

The use of an ultraviolet spectrometer to measure aromatic content has potential advantages over SFC and other techniques. UV spectrometers are relatively inexpensive, rugged, and simple to use, making them more suitable for on-site measurements. In this work, spectroscopic measurements were carefully made on a high-quality instrument to ensure reliable data. It is likely, however, that comparable results can be obtained with faster scans on lower resolution instruments. Of course, steps would need to be taken to ensure that the calibration models exhibit good long-term reliability with changes in plant operating conditions and feedstocks. Nevertheless, this study has shown the feasibility of estimating hydrocarbon content through UV spectroscopy.

Acknowledgments
The authors gratefully acknowledge the support of Natural Sciences and Engineering Research Council of Canada. D.T.A. was partially supported by a fellowship from the Sumner Foundation. P.S. was employed at Syncrude in conjunction with the Chemistry Industrial Internship Program at the University of Alberta. The authors thank Syncrude Canada Ltd. for permission to publish this work.

References
1. Methods of testing petroleum and associated products: standard test method for the determination of aromatic hydrocarbons by their ring number in petroleum middle distillates by supercritical fluid chromatography Canadian General Standards Board CAN/CGSB-3.0, No. 15.0-94 (1994).
2. Standard test method for determination of the aromatic content and polynuclear aromatic content of diesel fuels and aviation turbine fuels by supercritical fluid chromatography. ASTM D5186-96.