energy&fuels

ARTICLE

pubs.acs.org/EF

Mid-infrared Attenuated Total Reflectance (MIR-ATR) Predictive Models for Asphaltene Contents in Vacuum Residua: Asphaltene Structure−Functionality Correlations Based on Partial Least-Squares Regression (PLS-R)

Jorge A. Orrego-Ruiz,*[,†] Alexander Guzmán,[‡] Daniel Molina,[†] and Enrique Mejía-Ospino[†]

[†]Laboratorio de Espectroscopia Atómica y Molecular, Universidad Industrial de Santander, A.A. 678, Bucaramanga, Colombia

[‡]Ecopetrol, Instituto Colombiano del Petróleo, Piedecuesta, Colombia

ABSTRACT: In this research, an original way to obtain models able to predict the asphaltene content in vacuum residua by mid-infrared attenuated total reflectance (MIR-ATR) spectroscopy by reduction of *X* variables was developed. Partial least-squares regression (PLS-R) was used to reach this goal. A total of 69 samples for calibration and 18 samples for external prediction were used. It was demonstrated that dimensional reduction of the dependent variables greatly improves the prediction power of models. This methodology was evaluated in three processes of modeling, and 18 predictive models are reported. The model with better predictability was constructed with 35 spectral intensities, in which errors of calibration, validation, and prediction were 1.354, 2.095, and 1.24, respectively, and in which regression coefficients of calibration, validation, and prediction were 0.9799, 0.9720, and 0.9838, respectively. The spectral intensities in the best model are part of nine clearly differentiated functionalities. On the basis of this, it was possible to infer structural characteristics of the asphaltenes that are part of the studied vacuum residua; it means that structure−functionality correlations were reached.

INTRODUCTION

Many works have been published concerning the potential of chemometric methods to perform physicochemical analysis of oil and its fractions using spectroscopic methods. However, such investigations have been focused on obtaining valid statistical prediction models without an appropriate discussion about the mathematical structure of models. Concerning infrared (IR) spectroscopy, authors were limited to choose indiscriminately huge spectral regions, which have signals of several functional groups.[1−5] Nevertheless, some physicochemical properties, such as asphaltene content, should be closely related to functional groups, whose signals appear in particular IR regions. Understanding these relationships was the motivation of the present work.

Mid-infrared (MIR) spectra of heavy oil fractions show a similar profile in general terms. However, if a deconvolution is performed, the relative intensities of bands are very different.[6] For example, the region from 3000 to 2850 cm^{-1}, where the stretching mode of CH_3, CH_2, and CH appears, is taken in partial least-squares (PLS) analysis by authors without considering that some specific functional groups can affect more than others some physicochemical properties of samples, such as vacuum residua. American Petroleum Institute (API) gravity is affected by the length of aliphatic chains and the aromaticity of molecules that compose oil fractions.[7,8] Broadly, aromatic hydrocarbons have API values lower than paraphinic ones. For example, benzene has an API value of 28.72, and *n*-hexane has a value of 81.25. On the other hand, the microcarbon residue (MCR) depends upon the presence of polycondensed aromatic hydrocarbons. The results by Hassan et al. suggest that (nC7) asphaltenes are key components contributing to the MCR.[9] As it is well-known, asphaltenes are composed of aromatic clusters bonded to aliphatic chains, with the aromatic moieties being responsible for MCR.[10] The preceding suggests that an appropriated discrimination of spectral regions that consider the contribution of aliphatic and aromatic groups will allow for the building of predictive models on properties of vacuum residua. Nevertheless, because of the dramatic structural complexity of vacuum residua, the handy selection of adequate intensities is impossible but may be performed by chemometry tools. Partial least-squares regression (PLS-R) can help to choose spectral intensities.

In a recent paper, Parisotto et al. determined the total acid number in crude oil distillation residua using attenuated total reflection−Fourier transform infrared (ATR-FTIR) spectroscopy and a variable selection by chemometric methods.[11] They concluded that the use of methods for variable selection did not provide models with root-mean-square error of prediction (RMSEP) values significantly better than those obtained by the full-spectrum PLS model. However, in this work, we demonstrate that it is possible.

Determining how many variables should be used for modeling is a critical step during the development of a predictive model. Unfortunately, reliable and fast rules to make this decision are missing at present. In general, if just a few variables are used, a less precise model will result. On the contrary, if too many variables are used, the estimates from the model may be unstable; that is, small changes in the spectrum within the order of the spectral noise may produce statistically significant changes in the estimates. The maximum number of variables that should be used for

Received: June 6, 2011
Revised: July 15, 2011
Published: July 22, 2011

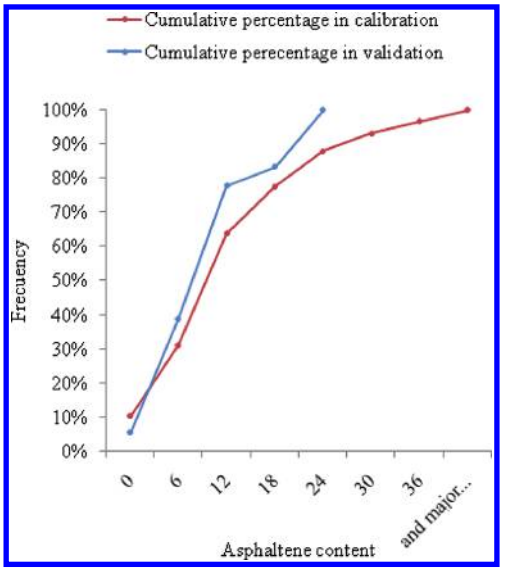

Figure 1. Distribution of samples for calibration and validation.

the development of a multivariate calibration model is related to the number of detectable, spectrally distinguishable components (or functionalities) that are present in the calibration set. Components (or functionalities) are spectrally distinguishable if they give rise to absorptions that are not linearly correlated among the calibration samples and if the change in the absorptions among the calibration spectra is larger than the spectral noise.[12]

The present research initiated the construction of models with intensities from IR spectra of assured repeatability. Next, IR intensities with weighted coefficients lower than 10% relative to the maximum coefficient were rejected. The main reason is that *X* variables with a small influence in the model generate uncertainty in the prediction. These variables represent low variances associated with the *Y* variable, and they have the lowest weighted coefficients. The former procedure provided models with a minimum number of predictive variables. Models with a significant increased prediction ability in terms of root-mean-square reduction of error and an increased regression coefficient (R^2) for calibration, validation, and prediction were achieved. To prove that the reduction of *X* variables always improved the parameters of calibration and validation, the same procedure was carried out for three processes of modeling. This work was performed with 87 vacuum residua and molecular distillation fractions divided into two sets: calibration and prediction. In the calibration set, two outliers were detected, while in the external prediction set, just one outlier was detected.

EXPERIMENTAL SECTION

Samples. A total of 87 samples of vacuum residua of representative Colombian crude oils and molecular distillation fractions were used in the present work. The IP-143 precipitation method was used to measure the asphaltene content using *n*-heptane (nC7) as the solvent. Vacuum residua of typical Colombian crude oils were obtained following the standard distillation methods ASTM 2892 and ASTM 5236-03. All vacuum residua are >500 °C. Molecular distillation fractions were obtained using a wiped-film molecular distillation unit model KD-6-1S from Chemtech Services, Inc. The samples were divided into two groups: 69 of them for calibration and 18 for external prediction. An ideal calibration sample set should contain a sufficient number of samples to statistically define the relationships between the spectral variables and the component concentrations or properties to be modeled. For complex mixtures as vacuum residues, where there are thousands of molecules interacting between them, obtaining an ideal calibration set is difficult, if not impossible. The number of samples that are required for calibration depends upon the complexity of the target samples. If a large number of components vary in the analyzed samples, then a larger number of calibration samples are required for the development of the model.[12] That was the main reason for using 80% of the samples for calibration and the remaining 20% for validation. We chose prediction samples, so that they were in the calibration range, which was between 0.00 and 41.96% asphaltene, as presented in Figure 1. It is important to note that condensed fractions from molecular distillation are lighter than their corresponding parent vacuum residua because they are enriched in saturate and aromatic compounds and depleted in resins and asphaltenes.[13]

Acquisition of MIR Spectra. MIR spectra were recorded on a Shimadzu IR-Prestidge 21 spectrometer with a spectral resolution of 8 cm^{-1} over the range of 4000−650 cm^{-1} and 32 scans. We use this resolution despite the fact that, in the majority of works, the acquisition of spectra is reported with 4 cm^{-1}. The aim was to reduce the acquisition time to one-half, an important aspect for analyzing a high number of samples. The spectrometer is equipped with a deuterated triglycine sulfate (DTGS) detector and a Pike Miracle attenuated total reflectance (ATR) diamond cell with single reflection and angle set up to 45°. An adjustable pressure system was used for assuring the contact between the sample and the ATR crystal. After measurement of the spectra, the remaining sample over the crystal was removed with a soft solvent wet paper. The acquisition time for 32 scans was approximately 15 s. The background spectrum of air was collected for every sample before the analysis in the single-beam mode. The spectral files were transformed to ASCII format using the IR Solution software and exported to The Unscrambler, version X.

Data Analysis. PLS-R was used to establish the relationship between IR spectral intensities and the percentage of asphaltenes in samples. PLS-R is a useful procedure to find a common structure between one or more characteristic properties of the samples, called independent variables or simply *Y* variables, and the spectral intensities, called dependent variables or *X* variables. In our case, the content of asphaltenes in the samples is the *Y* variable, while the IR spectral intensities are *X* variables. The linear PLS model finds a few variables, which are estimates of the latent variables (LVs) or their rotations. These variables are called *X* scores, and they are predictors of *Y* and also model *X*; i.e., both *Y* and *X* are assumed to be, at least partly, modeled by the same LVs. *X* scores are estimated as linear combinations of the original matrix of *X* variables with the weighted coefficients, $W_{ka}{}^*$ ($a = 1, 2, ..., A$). This could be more clearly seen in eq 1

$$t_{ia} = \sum_k W^*_{ka} X_{ik} \quad (1)$$

where *k* denotes the number of *X* variables or spectral intensities and *i* is the number of samples.[14]

The regression model can be written according the eq 2

$$Y = b_0 + b_1 X_1 + \ldots + b_k X \quad (2)$$

meaning that the observed response values are approximated by a linear combination of the values of the predictors. The coefficients of that combination are called regression coefficients or *b* coefficients.

These coefficients are calculated from *X* scores and *Y* score. A deeper analysis about the relationship between scores is beyond the scope of this paper, but it is correctly explained by Wold et al.[14] Regression coefficients summarize the relationship between all predictors and a given response. For PLS, the regression coefficients can be computed for any number of components; i.e., for every latent variable, there are different *b* coefficients. When *X* variables have been centered, all predictors are brought back to the same scale; the coefficients show the relative

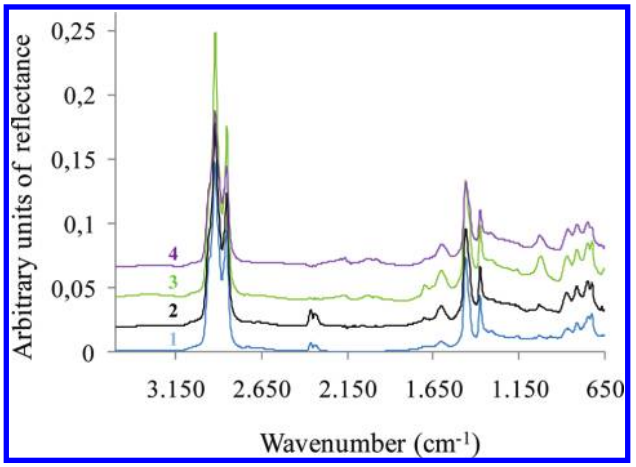

Figure 2. Raw MIR spectra of samples.

importance of the *X* variables in the model. Predictors with a large regression coefficient play an important role in the regression model; a positive coefficient shows a positive link with the response, and a negative coefficient shows a negative link. Predictors with a small coefficient are negligible.

One spectrum acquired with the mentioned resolution gives as a result 870 intensities. All spectral signals were first derived and then normalized. Several pretreatment strategies were evaluated, with those that showed the best prediction results being discussed. Subsequently, *X* variables were mean-centered. Almost as a rule, the first stage in multivariate modeling using projection methods is to subtract the average from each variable. This operation, called mean-centering, ensures that all results will be interpretable in terms of variation around the mean.

Cross-validation at random with two samples per segment was used to find the optimal number of latent variables and perform internal validation of the calibration model. The quality of models was evaluated according to values of root-mean-square error of calibration (RMSEC), validation (RMSEV), and prediction (RMSEP).[12,14] To ensure the confidence in the methodology, three processes of modeling were performed. Each one of them allowed for the building of six PLS models with a gradual reduction of *X* variables.

RESULTS AND DISCUSSION

Spectral Features and Data Pretreatment. In general terms, vacuum residua have very similar spectra to other petroleum fractions. In Figure 2, it is possible to differentiate between the spectra of four samples containing 0.00, 9.44, 19.1, and 31.3 wt % asphaltenes. Sample 1 (no asphaltenes) corresponds to a fraction from molecular distillation, and samples 2, 3, and 4 correspond to vacuum residua. Even though it is possible to differentiate them through three bands located at 1800−1650, 1150−900, and 730−720 cm^{-1}, only after modeling is it possible to know if these differences are associated with the *Y* variable. Table 1 shows the assignation of IR signals to samples.[6,15,16] As seen in Table 1, the majority of signals correspond to carbon−hydrogen bonds because nearly 92−94% of this kind of sample is composed of carbon and hydrogen.[8] Some of them appear simultaneously in several spectral regions according to their active IR vibrational modes. For instance, methyl groups can be detected at 2954 and 2894 cm^{-1} because of asymmetric and symmetric stretching and also at 1463 and 1376 cm^{-1} because of asymmetric and symmetric bending. At the same time, methylene groups can be

Table 1. Band Assignments Observed on MIR-ATR Spectra of Vacuum Residua

range of band (cm^{-1})	functional groups
3625−3100	hydrogen-bond interactions
3090−3030	$(C-H)_{ar}$ stretching
2954−2943	$R-CH_3$ asymmetric methyl stretching
2924−2916	R_2CH_2- asymmetric methylene stretching
2900−2881	$R_3(C-H)_{al}$ methyne stretching
2877−2858	$R-CH_3$ symmetric methyl stretching
2858−2846	R_2CH_2- symmetric methylene stretching
1632−1700	(C═O) stretching
1578−1628	$(C{=}C)_{ar}$ stretching
1473−1500	R_2CH_2- symmetric deformation bending
1423−1469	RCH_3 symmetric deformation bending
1395−1370	RCH_3 asymmetric deformation bending
1150−1070	$(C-O-C)_{al}$ stretching
1275−1200	$(C-O-C)_{ar}$ stretching
1060−970	(S−O) sulfoxide stretching
844−898	$C_{ar}-H$ (1H) isolated hydrogen bending out of plane
783−841	$C_{ar}-H$ (2H or 3H) two or three adjacent hydrogens bending out of plane
740−780	$C_{ar}-H$ (4H) four adjacent hydrogens bending out of plane
720−736	$R(CH_2)_n-R$ rocking when $n > 3$

observed at 2912 and 2850 cm^{-1} because of asymmetric and symmetric stretching and also at 1469 cm^{-1} because of bending.

In the region 950−700 cm^{-1}, bands from out of plane bending for the C−H aromatic bond are observed. The band at ∼870 cm^{-1} suggests the presence of aromatic rings with one isolated hydrogen (1H), i.e., penta-substituted rings. The band with a maximum at 812 cm^{-1} may be attributed to systems containing two or three adjacent aromatic hydrogens, i.e., tri- and tetra-substituted rings. The band at about 750 cm^{-1} may be attributed to systems containing four adjacent aromatic hydrogens, i.e., *ortho* substitution of the aromatic rings.[17−19] Finally, a sharp band responsible for rocking of chains with more than three contiguous methylene groups is detected at 720 cm^{-1}. A suitable analysis of this region can provide important conclusions about condensation of aromatic rings in asphaltenes.

PLS Analysis. The strategy employed in this work started by modeling with those *X* variables whose repeatability was assured. To achieve this, a set of 14 vacuum residua with variable chemical composition were chosen. Relative standard deviations (RSDs) on the intensities of 10 acquired spectra for every sample by the same analyst were calculated.[20] Intensities with RSD > 10% were rejected, and then the number of variables was reduced from 870 to 166, from which a preliminary PLS model was found. An ensuing model was recalculated without the *X* variables, whose weighted coefficients were lower than 10% relative to the maximum coefficient. This procedure was iterated until six models were obtained, for which it was possible to establish significant differences in terms of prediction errors. Three processes of modeling were performed to evaluate this methodology. Table 2 summarizes the principal statistical parameters to assess the predictability of all models. In all cases, six latent variables were used, with just two samples in the calibration set eliminated as outliers and one sample in the validation set eliminated as an outlier. Two types of outliers can be identified

Table 2. PLS Calibration, Validation, and Prediction Results for Models

		R^2		RMSE		
	number of *X* variables	calibration	validation	calibration	validation	prediction
modeling 1	166	96.751	92.201	2.2244	3.4750	1.6386
	75	97.240	95.642	2.0603	2.6177	1.5301
	43	98.021	96.876	1.7526	2.2249	1.3051
	37	98.340	97.488	1.6180	2.0091	1.5989
	32	98.589	98.132	1.5034	1.7769	1.9763
	27	98.502	98.091	1.5490	1.8174	2.0677
modeling 2	166	96.557	94.247	2.3010	3.0025	1.5869
	80	97.410	95.688	1.9957	2.6077	1.4418
	41	97.946	96.957	1.7772	2.1831	1.2715
	35	98.000	97.201	1.3538	2.0946	1.2421
	24	98.764	98.226	1.4112	1.7370	1.3729
	20	98.806	98.430	1.3869	1.6199	1.2824
modeling 3	166	96.751	92.194	2.2244	3.4624	1.6386
	83	97.453	95.459	1.9937	2.6590	1.6835
	59	97.504	96.221	1.9737	2.5268	1.6278
	41	98.226	97.348	1.6778	2.1255	1.6751
	24	98.676	98.029	1.4564	1.8037	1.7680
	20	98.575	98.000	1.5110	1.7889	1.8636

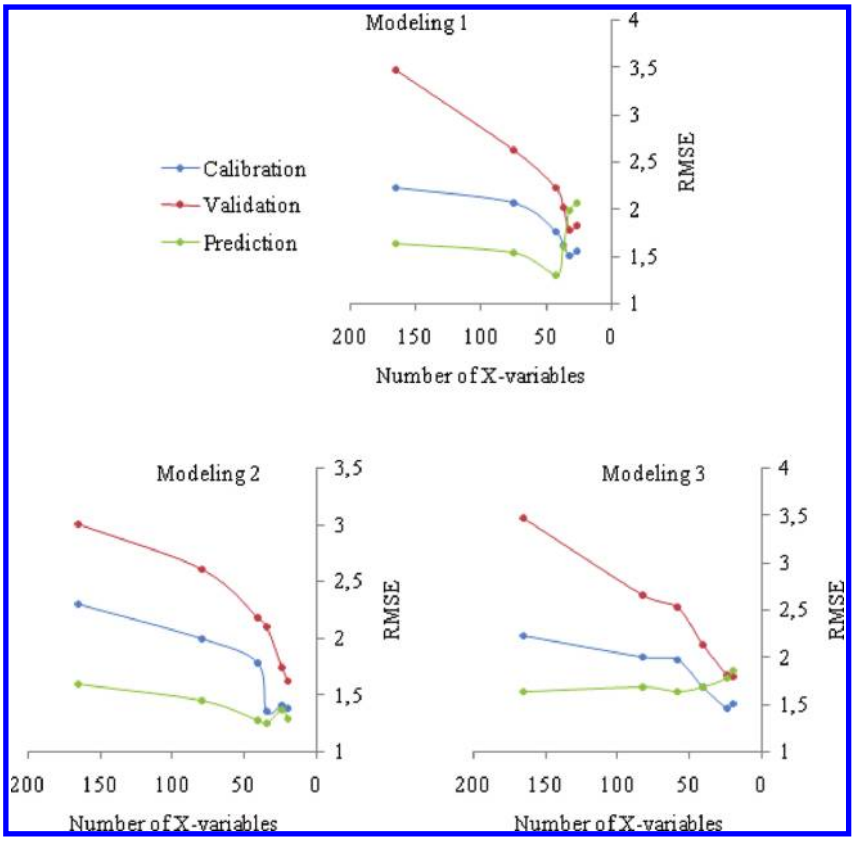

Figure 3. Errors of calibration, validation, and prediction with respect to the number of *X* variables.

during the calibration procedures. The first type is a sample that represents an extreme composition relative to the remainder of the calibration set. The second type of outlier is one for which the estimated value differs from the reference value by a statistically significant amount. Such outliers could indicate an error in the reference measurement, an error in the spectral measurement, a clerical error in sample attribution or reference value transcription, or a failure of the model.

dx.doi.org/10.1021/ef200834x |*Energy Fuels* 2011, 25, 3678–3686

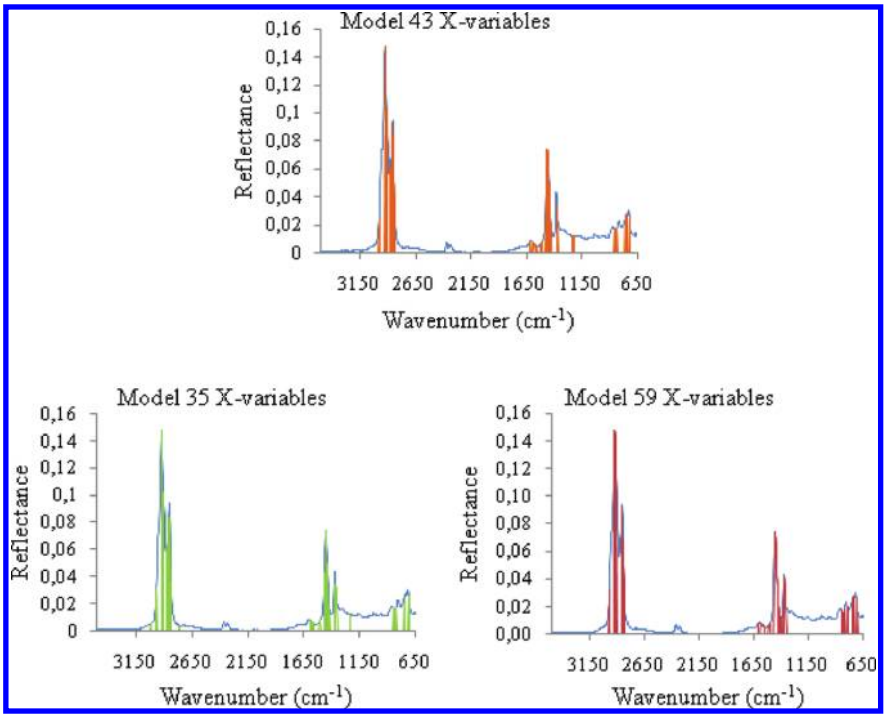

Figure 4. Spectral regions used to predict asphaltene contents superimposed on the spectrum of sample 1.

In this last type of outlier, we classify the samples that were discarded.

Figure 3 shows the variation of RMSEC, RMSEV, and RMSEP in relation to the number of *X* variables in the modeling of asphaltene contents. It is clearly seen that this procedure improves the predictability of models if it is considered that that regression coefficient is increased as well. According to Table 2, the models built with 43, 35, and 59 *X* variables were the best. Nevertheless, the majority of *X* variables are common in the three final models. The same general spectral regions were chosen after reducing *X* variables. In Figure 4, the chosen spectral regions to predict the asphaltene contents can be seen in color. These regions represent mainly aliphatic and aromatic functionalities. Why is it possible to explain the models only with these functionalities? As known, asphaltenes are not exclusively represented by C−H aromatic and aliphatic functionalities. The MIR region of the spectrum does not show many features of S, O, and especially N compounds present in asphaltenes. However, the presence of these heteroatoms in aliphatic chains or in aromatic rings can generate chemical shifts. These little variations can be detected during the PLS-R, where the variances between spectral intensities (data are mean-centered) are associated with the variances in *Y* variables. On the other hand, according to final coefficients of regression for calibration, the whole variance in *Y* is not explained. Indeed, the best model, in terms of RMSEP, explained 98% of *Y* variance (Table 2). The remaining could be associated with N, O, and S functional groups, because the regions where these bonds appeared are not considered at the beginning of the analysis. It means many of these signals appear in noise zones.

Figure 5 shows the plots of prediction for the three principal models with their regression coefficients. According to Figure 5, there is dispersion in the errors for the complete range of validations for the three models. Figure 1 shows the non-homogeneity in the distribution of validation samples. Nearly 40% of validation samples have asphaltene contents below 6%, but obtaining a proper set of validations is almost impossible in this kind of sample. On the other hand, in the three cases, the *Y* intercept is different from zero. This is because the models explain up to 98% of the total variance of the data.

After assessing the statistically calculated parameters, a question arises as to what criterion governs the selection of the best model? The outcomes presented in Table 2 show clearly that RMSEP has a minimum value for an intermediate number of variables. In general, if just a few variables are used, the highest RMSEP will be obtained. If too many variables are used, the estimates of the model may be unstable; it means that small changes in the spectrum of the order of the spectral noise may produce statistically significant changes in the estimates. When 166 *X* variables were used in the three models, the highest errors of cross-validation (above 3%) were determined (Table 2). Considering that the most important parameter is RMSEP, even though the model considering 20 variables seems a valid alternative, the model with 35 *X* variables was chosen as the best.

Structure−Functionality Correlations. According to previous analysis in this paper, the model with 35 *X* variables was chosen to thoroughly assess structure−functionality correlations. A total of 69 samples with an important structural variety were used in calibration. For that reason, it was possible to infer that the asphaltene chemical structure is closely related to some functional groups that are contained in the IR spectrum. In the work by Ibrahim et al.,[21] it was shown that all asphaltenes were composed of similar functional groups. However, the intensities for the characteristic functional groups were different for different asphaltenes. Additionally, they calculated several structural parameters from nuclear magnetic resonance (NMR) and FTIR spectroscopies to correlate asphaltene precipitation behavior and inhibitor effectiveness for four crude oils. Some of these parameters were R_{CH_2/CH_3}, $S_{1H/4H}$, CH_2/CH_3, and I_{1600}/I_{2923}, which are molar ratios of CH_2/CH_3 groups, degrees of substitution and condensation in polyaromatic rings, alkyl side-chain length, and

dx.doi.org/10.1021/ef200834x |*Energy Fuels* 2011, 25, 3678–3686

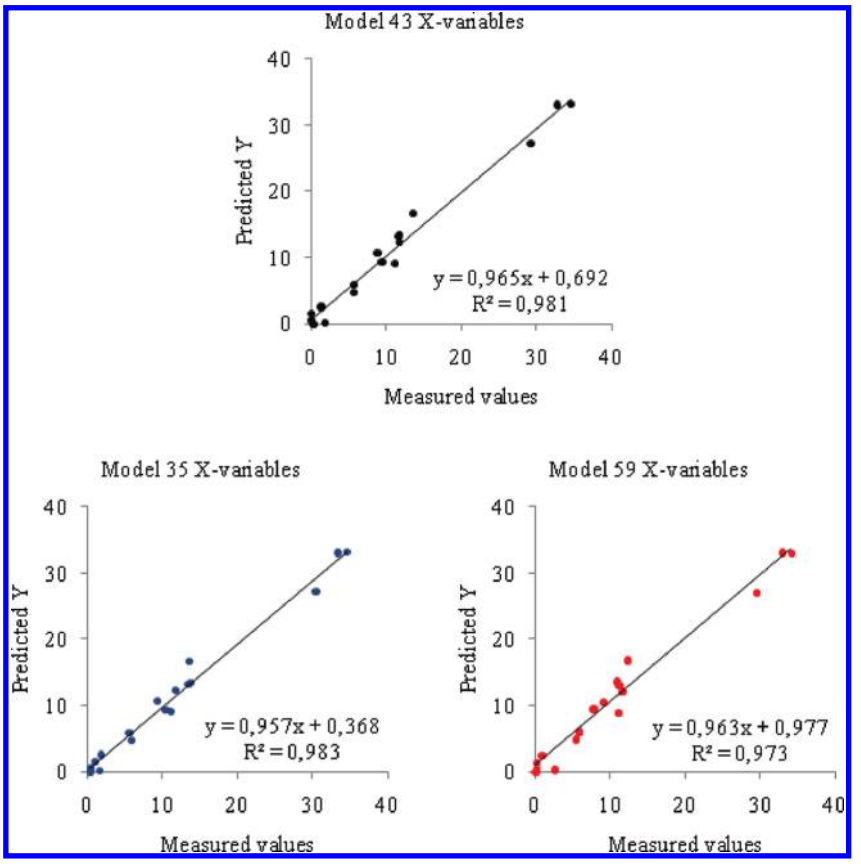

Figure 5. Plots of the correlation of predicted versus measured values for the three principal models studied in this work.

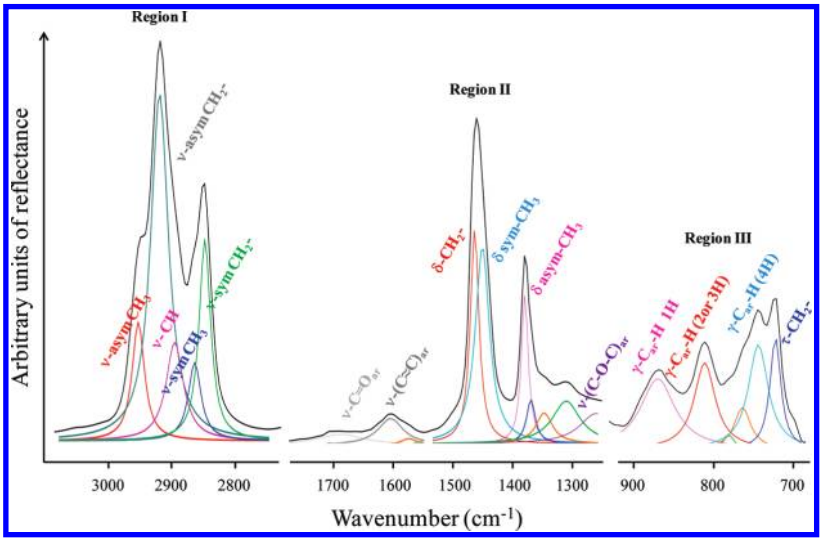

Figure 6. Deconvolution of IR spectra of sample 1.

aromatic/aliphatic ratio, respectively. Alvarez and co-workers[22] used the *ortho* substitution and the aromaticity index to characterize and correlate the pyrolysis behavior of a set of pitches prepared from anthracene oil. The *ortho* substitution was calculated by thoroughly assessing the 900−700 cm^{-1} region.

To find meaningful spectral features in IR spectra, deconvolution of the three regions was performed to sample 1 and shown in Figure 6, together with the assignments for every resulting band. The number and position of the single components used in the fitting procedure have been evaluated by means of visual inspection, finding agreement with the data reported in the literature for coals.[23,24] In this procedure, Lorentzian functions were employed. The region between 900 and 740 cm^{-1} (region I in Figure 6) corresponds to C_{ar}−H bonds of aromatic di-, tri-, tetra-, and penta-substituted rings, which means with 4, 3, 2, and 1 hydrogen, respectively. The orange band was not assigned

dx.doi.org/10.1021/ef200834x |*Energy Fuels* 2011, 25, 3678–3686

Table 3. *X* Scores for Asphaltene Contents of Four Different Samples

wavenumber (cm^{-1})	assignations	sample 1	sample 2	sample 3	sample 4
718	τ-CH_2−	8.31	7.06	6.42	9.25
756	γ-C_{ar}−H (4H)	−1.40	−1.95	−1.43	−2.34
829	γ-C_{ar}−H (2H or 3H)	−2.22	−2.93	−2.46	−4.12
845	γ-C_{ar}−H (1H)	0.88	0.31	0.51	1.79
853		0.36	0.96	0.75	1.94
1242	ν-$(C-O-C)_{ar}$	0.09	0.03	0.03	0.12
1358	δ asym-CH_3	−3.84	−2.83	−2.57	−4.66
1362		−6.53	−5.01	−4.35	−7.71
1366		−7.69	−6.70	−5.28	−9.41
1369		−16.07	−13.29	−10.12	−17.64
1423	non-assignable (minima)	8.27	8.75	7.90	16.07
1439	δ sym-CH_3	−20.78	−17.77	−13.09	−21.50
1443		−25.45	−19.15	−14.62	−23.36
1454		−9.87	−4.63	−4.62	−2.68
1458	δ-CH_2−	0.92	1.86	1.08	7.84
1470		46.30	34.99	31.55	52.57
1504	non-assignable (minima)	−0.64	−0.66	−0.54	−0.91
1508		−0.51	−0.57	−0.48	−0.73
1512		−0.45	−0.58	−0.38	−0.61
1516		−0.42	−0.59	−0.34	−0.60
1570	$(C=C)_{ar}$ stretching	0.69	1.54	1.57	2.48
1574		0.68	1.62	1.58	2.38
1582		0.60	1.52	1.36	2.04
1585		0.53	1.25	1.11	1.75
1589		0.43	1.00	0.80	1.29
2766	non-assignable	−0.12	−0.16	−0.14	−0.14
2835	ν-sym CH_2−	37.21	36.02	30.58	38.89
2847		45.10	41.17	40.85	42.67
2859	ν-sym CH_3	−23.12	−24.09	−25.08	−17.54
2874	mixture of signals CH, CH_2, and CH_3	2.40	1.69	1.33	2.27
2878		2.12	0.53	0.22	1.57
2905	ν-asym CH_2−	13.89	14.94	14.45	14.54
2920	ν-asym CH_2− (maximum)	−0.70	6.53	10.98	−6.18
2970	ν-asym CH_2−	8.21	6.64	5.58	9.25
3021	(ν-C_{ar}−H + ν-asym CH_3)	−0.20	−0.16	−0.13	−0.14

because any intensity in this region does not explain any variance in *Y* to the final model. It is difficult to assign the bands in the fingerprint region (between 1350 and 950 cm^{-1}) because of the multiple bands that are detected for oxygenated groups, such as S−O and C−O−C types. However, for the purpose of this research, only the fraction between 1350 and 1230 cm^{-1} was considered, because the region between 1230 and 950 cm^{-1} was discarded for its inadequate repeatability. One band near 1250 cm^{-1} was necessary to fit the deconvoluted region II. This band was assigned to $(C-O-C)_{ar}$ stretching. The green, orange, and dark blue bands in region II of Figure 6 were found to fit this region in a proper way, but they were not assigned because they do not affect the final analysis because none of the 35 *X* variables appears in these regions of the IR spectrum. The spectral range with the highest intensities corresponds to region I, where three functional groups are assigned: aliphatic stretching of CH_3, CH_2, and CH. Although five bands were assigned, asymmetric stretching provides the same structural information as symmetric stretching for methyl and methylene groups.

Table 3 shows the scores for samples 1, 2, 3, and 4 for each *X* variable of the 35 *X* variable model. The negative values "subtract" the amount of asphaltenes in samples, and the positive values "add" to this parameter. A negative value for a specific sample does not mean that this specific functionality is not present in the asphaltene structure. It means that this particular vacuum residue has a lower concentration of this functional group compared to the others, which was associated with asphaltene contents. Finally, the sum of every score in each column plus the residual equals −57.038 for the 35 *X* variable model, which gives the estimated asphaltene for every sample. This residual is the *Y* intercept in a 35th dimensional space.

As discussed before, not all assignments correspond to different functional groups. For instance, although the intensities 1358−1369 and 1454−1504 cm^{-1} correspond to asymmetric and symmetric R−CH_3 bending deformation, respectively, they describe just R−CH_3 bending deformation in aliphatic chains. These two assignments negatively affect asphaltene contents, although they are separated bands. The same occurs for the

2859 cm^{-1} band corresponding to the symmetric stretching of CH_3. Despite the separation in the IR spectrum, this signal also negatively affects the model.

If a detailed revision is performed of the results included in Table 3, it is possible to infer that aliphatic groups have the highest relevance in the estimated value of asphaltenes: scores for 2859 and 2847 cm^{-1} bands from sample 1 are −23.12 and 45.10, respectively. This does not contradict the image of polycondensed rings with aliphatic chains chemically bonded to aromatic structures to the asphaltenes. It just means that there are a high number of possibilities of bonding in aliphatic structures in relation to aromatic structures. While one aliphatic chain of six carbons could or could not have methyl or methine groups, considering all of its isomers and including a naftenic ring, an aromatic ring of six carbons has just one possibility. Thus, the variability of aliphatic regions, which is correlated with *Y* property, is highest concerning aromatic regions. In addition, the extinction coefficient is higher in C−H bonds than C−C bonds. IR signals do not only depend upon the concentration of functional groups but also depend upon the probability of transition between vibrational states. The extinction coefficient is directly correlated with the change of the electric dipole between vibrational states; i.e., it depends upon the electric dipole integral. This change is higher in C−H (different atoms) than C−C (same atoms). This is the main reason by which the intensity of $(C{=}C)_{ar}$ stretching (band in 1600 cm^{-1}) is lower than C−H stretching in aliphatic groups in the region between 3000 and 2800 cm^{-1}. Relative intensities cannot be compared unless the extinction coefficients are known for every vibrational mode.

At present, it is relevant to emphasize that spectra are acquired in a complex system with thousands of different molecules and MIR spectra are averages of all absorbances produced by chemical bonds, which are present in not only asphaltenes but also saturate, aromatic, and resin fractions. On the other hand, some regions mentioned in Table 3 were not assigned because they correspond to spectral minima and their relative intensities depend upon the bands surrounding them. We do not disclaim the limitations of IR spectroscopy, but we do not diminish its potential. Finally, determined structure−functionality relationships based on PLS-R are as follows: (1) Asphaltenes are composed of polycondensed aromatic molecules with aliphatic chains bonded to them. This can be inferred if it is considered that ν-$(C{=}C)_{ar}$, which is particular of polycondensed rings, and γ-C_{ar}−H (1H) are positively affecting the model, while γ-C_{ar}−H (4H) and γ-C_{ar}−H (2H or 3H) are negatively affecting the model. The ν-$(C{=}C)_{ar}$ stretching around 1600 cm^{-1} is particular of polycondensed aromatic rings, and signals that appear in 950−740 cm^{-1} correspond to di-, tri-, tetra-, and penta-substituted aromatic rings. Therefore, if the presence of C_{ar}−H bonds with adjacent hydrogens are negatively affecting the model, the periphery of clusters is barely bonded to aliphatic chains. Penta-substituted rings are typical of the aromatic clusters, and they are positively affecting the model. (2) The aromatic region from 1500 to 1627 cm^{-1} was considered without influence of the C=O stretching (Figure 4). It means that C=O functional groups (polar bonds) are not related to asphaltenes. According to the work by Carbognani and Buenrostro-Gonzalez,[25] resins have the greatest contribution of polar bands compared to asphaltene fractions. They demonstrated that typical oxygenated bands (carbonyls and sulfoxides) are not observed in the asphaltenes, while C−O−C types are observed in fewer cases. This assertion is supported by considering that the $-(C{-}O{-}C)_{ar}$ stretching mode positively affects the model. The scores at 1242 cm^{-1} are lowest for all samples in Table 3, indicating that this functional group has little variation between the samples. (3) Aliphatic chains with more than three $-CH_2-$ contiguous units positively affect the model (τ-CH_2- in Figure 6), and methyl groups negatively affect the model. This signal is related to the length and degree of condensation of aliphatic chains. This allows for the proposal that naphthenic rings are anchored to the polyaromatic hydrocarbon (PAH) system and/or that aliphatic chains are connecting aromatic clusters. Long alkyl appendages are characteristic of asphaltenes, as reported by Strausz and Lown for Athabasca samples.[26]

CONCLUSION

A methodology for obtaining models to predict the asphaltene contents in vacuum residua by MIR-ATR spectroscopy was developed. The procedure of reducing *X* variables improves the predictability of models in terms of increased regression coefficients and reduction of errors of calibration, validation, and prediction, in contrast with a recently published work by Parisotto et al.[11] PLS-R was used to find the most important *X* variables correlated with asphaltene structures from vacuum residua, making it possible to obtain structure−functionality relationships for the studied nC7 asphaltenes. The asphaltene chemical structure is closely related to specific functional groups that are contained in three specific regions of the IR spectra: 3000−2800, 1600−1250, and 870−715 cm^{-1}.

AUTHOR INFORMATION

Corresponding Author

*Telephone: +57-76349069. Fax: +57-76349069. E-mail: jorgecarbon2001@yahoo.es.

ACKNOWLEDGMENT

The authors are grateful to Ecopetrol S.A. for its financial support through the "Convenio de Cooperación Tecnológica 003 de 2007" between the Instituto Colombiano del Petróleo and the Universidad Industrial de Santander and to Colciencias for its Scholarship "Francisco José de Caldas" given to Jorge A. Orrego-Ruiz.

REFERENCES

(1) Hongfu, Y.; Xiaoli, C.; Haoran, L.; Yupeng, X. *Fuel* **2006**, *85*, 1720–1728.

(2) Brian, K.; William, W.; Welch, T.; Graham, J. R. *Energy Fuels* **1998**, *12*, 1008–1012.

(3) Satya, S.; Roehner, R. M.; Deo, M. D.; Hanson, F. V. *Energy Fuels* **2007**, *21*, 998–1005.

(4) Bezerra De Lira, L.; Cruz de Vasconcelos, F. V.; Fernandes-Pereira, C.; Silveira-Paim, A. P.; Stragevitch, L.; Pimentel, M. F. *Fuel* **2010**, *89*, 405–409.

(5) Al-Ghouti, M.; Al-Degs, Y.; Mustafa, F. *Fuel* **2010**, *89*, 193–201.

(6) Chiaberge, S.; Guglielmetti, G.; Montanari, L.; Salvalaggio, M.; Santolini, L.; Spera, S.; Cesti, P. *Energy Fuels* **2009**, *23*, 4486–4495.

(7) Riazi, M. R. *Characterization and Properties of Petroleum Fractions*, 1st ed.; American Society for Testing and Materials (ASTM) International: West Conshohocken, PA, 2005; ASTM Manual Series: MNL 50, pp 1−149.

(8) Speight, J. G. *The Chemistry and Technology of Petroleum*, 4th ed.; CRC Press (Taylor and Francis Group): Boca Raton, FL, 2007.

(9) Hassan, A.; Carbognani, L.; Pereira-Almao, P. *Fuel* **2008**, *87*, 3631–3639.
(10) Mullins, O. C. *Energy Fuels* **2010**, *24*, 2179–2207.
(11) Parisotto, G.; Ferrao, M. F.; Muller, A. L. H.; Muller, E. I.; Santos, M. F. P.; Guimaraes, R. C. L.; Dias, J. C. M.; Flores, E. M. M. *Energy Fuels* **2010**, *24*, 5474–5478.
(12) American Society for Testing and Materials (ASTM) International. *ASTM E1655. Standard Practices for Infrared Multivariate Quantitative Analysis*; ASTM International: West Conshohocken, PA, 2005.
(13) Orrego, J.; Contreras, Z.; Mejía, E.; Guzmán, A.; Molina, D. *Proceedings of the XXIX Latin American Congress of Chemistry—CLAQ 2010*; Cartagena, Colombia, Sept 27−Oct 1, 2010; PRI 26.
(14) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 109–130.
(15) Orrego-Ruiz, J. A.; Cabanzo, R.; Mejía-Ospino, E. *Int. J. Coal Geol.* **2011**, *85*, 307–310.
(16) Buenrostro-Gonzalez, E.; Andersen, S. I.; Garcia-Martinez, J. A.; Lira-Galeana, C. *Energy Fuels* **2002**, *16*, 732–741.
(17) Guillen, M. D.; Iglesias, M. J.; Dominguez, A.; Blanco, C. G. *Fuel* **1995**, *74* (11), 1595–1598.
(18) Guillen, M. D.; Iglesias, M. J.; Dominguez, A.; Blanco, C. G. *Energy Fuels* **1992**, *6*, 518–525.
(19) Borrego, A. G.; Blanco, C. G.; Prado, J. G.; Díaz, C.; Guillén, M. D. *Energy Fuels* **1996**, *10* (1), 77–84.
(20) Chung, H.; Ku, M.; Lee, J. S. *Appl. Spectrosc.* **2000**, *54*, 239–245.
(21) Ibrahim, H. H.; Idem, R. O. *Energy Fuels* **2004**, *18*, 1354–1369.
(22) Álvarez, P.; Granda, M.; Sutil, J.; Menendez, R.; Fernandez, J. J.; Viña, J. A.; Morgan, T. J.; Millan, M.; Herod, A. A.; Kandiyoti, R. *Energy Fuels* **2008**, *22*, 4077–4086.
(23) Orrego-Ruiz, J. A.; Cabanzo, R.; Mejía-Ospino, E. *Rev. Mex. Fis.* **2010**, *56*, 251–254.
(24) Sobkowiak, M.; Painter, P. *Fuel* **1992**, *71*, 1105–1125.
(25) Carbognani, L.; Buenrostro-Gonzalez, E. *Energy Fuels* **2006**, *20*, 1137–1144.
(26) Strausz, O. P.; Lown, E. M. *The Chemistry of Alberta Oil Sands, Bitumen and Heavy Oils*, 1st ed.; Alberta Energy Research Institute: Calgary, Alberta, Canada, 2003; pp 575−584.