Research Article

Magnetic Resonance in Chemistry

Received: 13 June 2012 Revised: 26 July 2012 Accepted: 21 August 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI 10.1002/mrc.3872

Prediction of physical–chemical properties of crude oils by ^{1}H NMR analysis of neat samples and chemometrics

Alice Masili,[a] Sonia Puligheddu,[b]* Lorenzo Sassu,[b] Paola Scano[a] and Adolfo Lai[a]

In this work, we report the feasibility study to predict the properties of neat crude oil samples from 300-MHz NMR spectral data and partial least squares (PLS) regression models. The study was carried out on 64 crude oil samples obtained from 28 different extraction fields and aims at developing a rapid and reliable method for characterizing the crude oil in a fast and cost-effective way. The main properties generally employed for evaluating crudes' quality and behavior during refining were measured and used for calibration and testing of the PLS models. Among these, the UOP characterization factor *K* (K_{UOP}) used to classify crude oils in terms of composition, density (*D*), total acidity number (TAN), sulfur content (*S*), and true boiling point (TBP) distillation yields were investigated. Test set validation with an independent set of data was used to evaluate model performance on the basis of standard error of prediction (SEP) statistics. Model performances are particularly good for K_{UOP} factor, TAN, and TPB distillation yields, whose standard error of calibration and SEP values match the analytical method precision, while the results obtained for *D* and *S* are less accurate but still useful for predictions. Furthermore, a strategy that reduces spectral data preprocessing and sample preparation procedures has been adopted. The models developed with such an ample crude oil set demonstrate that this methodology can be applied with success to modern refining process requirements. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: NMR; ^{1}H; crude oil; chemometrics; PLS; property prediction

Introduction

Crude oil is a mixture of a very large number of different hydrocarbons: the most commonly found molecules are alkanes (linear or branched), cycloalkanes, aromatic hydrocarbons, and more complicated chemicals such as asphaltenes. Each petroleum variety has a unique mix of molecules, which define its physical and chemical properties and ultimately its behavior during refining.[1]

Currently, the refiners have to cope with an increasing demand of high quality distillates of tighter specifications, while facing the need to process increasingly heavier and poorer quality crudes. To make both ends meet, refiners are looking at optimizing production and improving selection of crudes and crude mixtures to be processed that can only be achieved through a detailed knowledge of the composition and quality of feed and finished products.

Generally, the assessment of the crude properties is carried out by standard analytical methods.[2] These methods are often time-consuming, elaborate, and expensive and require a large amount of sample and solvents. As a fast and viable alternative, NMR spectroscopy combined with chemometrics has been proposed.[3–5] In comparison with other spectroscopic techniques,[6–9] NMR has the advantage to provide directly the sample molecular details that determine its physical–chemical properties at a macroscopic level.[10] In fact, the NMR spectrum contains information in terms of the molecular functional groups and, if recorded with the proper resolution, may allow characterizing a sample at the molecular level.[11,12] Unfortunately, the NMR spectra of crude oils are very complex and contain a great number of signals reflecting their great chemical complexity.[13–15] Therefore, to exploit the full information content of the NMR data acquired on complex systems, various multivariate data analysis methods have been developed.[16] In quantitative analysis chemometric regression techniques as partial least squares (PLS) and principal component regression (PCR) are used to highlight the correlations among the NMR spectra and the properties of interest.[17]

Although industrial applications of NMR and chemometrics to predict petroleum distillate properties already exist, only few running industrial applications are reported for crudes; in these cases, performance data are often confidential, and no experimental details are given.[18] On the other hand, scientific literature mostly refers to the laboratory analysis of petroleum cuts, namely synthetic hydrocarbon mixtures,[19,20] heavy fuel oil,[21] decant oil,[22] gasoline,[23] and adulterated gasoline,[24] and to our knowledge, only few publications deal specifically with crude oils.[3–5] Generally, laboratory applications use NMR

* *Correspondence to: Sonia Puligheddu, Saras Ricerche e Tecnologie S.p.A., Traversa C, 5° Strada Ovest, Z.I. Macchiareddu, C. P. 237, I-09032 Assemini, CA, Italy. E-mail: sonia.puligheddu@sartec.it*

a *Dipartimento di Scienze Chimiche, Università degli Studi di Cagliari, Cittadella Universitaria di Monserrato, S.S. 554 – Bivio per Sestu, 09042 Cagliari, Italy*

b *Saras Ricerche e Tecnologie S.p.A., Traversa C, 5° Strada Ovest, Z.I. Macchiareddu, C. P. 237, I-09032 Assemini, CA, Italy*

spectral data of diluted samples acquired at high resolution, whereas online applications run with 60-MHz spectrometers on neat samples. In the first case, the solvent–solute interactions can generate annoying signal shifts and spectral artifacts with detrimental effects for the subsequent statistical analysis,[25–30] whereas in the latter, the small chemical shift drift corrections are less important and data pretreatment is easier, at the expense of reduced resolution.

In order to exploit the advantages of each of the above-mentioned approaches, we recorded the proton NMR spectra of neat crude samples with a 300-MHz laboratory spectrometer. PLS regression models for prediction were built using analytical data obtained in our laboratory: we focused our attention on properties such as *D*, *S*, TAN, TBP distillation yields, and K_{UOP} factor. These parameters are needed for process monitoring and are particularly important during scheduling refining operation. Because each given set of different crudes on site-specific context determines the feasibility on the application, we tested an ample set of data including crudes originating from different geographical areas, spanning a wide compositional range.

Materials and Methods

Samples

Sixty-four samples of different crudes, kindly provided by SARAS Refinery (SARAS S.p.A., Sarroch, Italy), were collected and analyzed. Because crude oils can degrade with time, samples were stored at low temperature (5 °C). In order to verify crude oil stability under these conditions, the ^{1}H NMR spectra of a light crude was acquired on a monthly basis for 1 year. The spectral profile and the normalized area of meaningful peaks (cyclohexane, cyclopentane, and benzene) were monitored and compared. Results show negligible month to month variation and no specific trend in the data. In order to obtain representative samples and to assure the acquisition of reproducible high quality data,[31] all crude specimens were properly homogenized above the pour point temperature prior to sampling. The bulk sample was used for the determination of physical and chemical properties according to the reference methods,[2] while a small aliquot was taken for the NMR spectroscopy experiments.

The crude oil samples were collected over a 3-year period from 28 different extraction fields located worldwide. Crude oils may show distinct, broadly variable characteristics according to their geographical origin, and within the same macro region, they may differ by a large extent depending on the extraction field. Furthermore, well age may also affect crude properties. To capture the possible differences among crudes, greater variability ones were collected in larger number (up to five samples), while stable crudes were acquired at least twice to account for wells production changes. For a limited number of crude type, only one sample was available.

NMR experiments

^{1}H NMR spectra were recorded at 25 °C on a Varian Unity Inova 300 spectrometer (Varian Inc., Palo Alto, CA (USA)), operating at 299.905 MHz, using the following conditions: spectral width 4382 Hz, pulse width 6.9 μs (pi/2), relaxation delay 30 s, number of transients 128. Data were processed with a zero filling of 32 K points to improve the digital resolution for a better spectra alignment. The longitudinal relaxation times (T_1) of crude oil protons were measured applying the inversion recovery sequence: a 30-s delay was chosen to assure complete relaxation of protons for all samples. Under these experimental conditions, a very good signal-to-noise (S/N) ratio was achieved. The NMR spectra were acquired using a coaxial tube. The crude oil, spiked with TMS for 0 ppm referencing, was introduced into the internal capillary tube, whereas the external one was filled with deuterated trifluoroacetic acid for lock referencing. This procedure allows for removing any interactions between the solvent and the sample. Phase and baseline were manually adjusted and corrected.

Data preprocessing

Prior to multivariate modeling, spectral area from 0.22 to 9.00 ppm was divided in constant width segments of 0.04 ppm, called bins or buckets, whereas the data points between 4.20 and 5.90 ppm were omitted to avoid including the water signal and a wide noise region into the models. The bucketing procedure was preferred to the alternative one that uses the original data points to reduce the dimension of the data matrix (64 samples × 178 buckets). The total area was then normalized. Next, the data matrix was mean centered implying that the average spectrum is subtracted from each of the individual spectra.

Multivariate analysis

Principal components analysis (PCA) is an unsupervised explorative analysis technique, and as such, it requires no information about class membership and just looks for inherent variation in the dataset. PCA transforms the variables in a data matrix *X* into a smaller number of new latent variables called principal components (PCs). These new variables are linear combinations of the original variables but highlight the variance within the dataset and remove redundancies. The PCs are uncorrelated with each other and are calculated in order of decreasing contribution to the total variance of the original dataset. The dimensionality reduction of the data is obtained excluding the PCs containing less information. Observations (i.e., samples) are assigned scores according to the variation along individual PCs. When displayed graphically, samples with similar scores cluster together and separate out from other groups with differing scores. The explanation of what each PC represents in relation to the original variables may then be assessed using the loadings. These are a set of weights associated to each of the original variable.[32]

Partial least squares regression is a multivariate latent-variable method used to find the fundamental relations between two data matrices (*X* and *Y*). It relates one dependent variable (*y*) to a set of independent variables (*X*), combining the features of PCA and multiple regression. A PLS model looks for the multidimensional direction in the *X* space that explains the maximum multidimensional variance direction in the *Y* space. Thus, the *y* vector is used in obtaining the decomposition of *X* into its principal components, and these, in turn, are used as regressors on *Y*. PLS regression is particularly suited when the *X* matrix has more variables than observations and when there is multicollinearity among the X values. For such a reason, it is the preferred choice when dealing with NMR spectral data.[33] Although the term PLS has been used to describe various mathematical algorithms, the version used in

Table 1. Selected physical and chemical properties of crude oils

Sample	Density at15 °C [a] (kg/m^3)	K_{UOP}[b]	TAN[c] (mg_{KOH}/g)	Sulfur[d] (%w)
1	839.1	12.09	0.254	0.124
2	838.2	12.45	0.113	0.139
3	896.8	11.73	0.140	1.930
4	884.0	11.82	0.526	0.670
5	837.8	—	—	0.127
6	879.0	11.82	0.130	2.252
7	860.2	11.89	0.100	—
8	873.7	11.68	0.200	—
9	870.3	11.94	0.237	1.280
10	842.1	12.09	0.276	0.121
11	875.0	11.59	0.415	0.169
12	873.5	11.82	0.148	1.840
13	889.7	11.91	0.811	0.410
14	871.5	11.91	0.182	1.641
15	870.6	11.91	0.780	—
16	933.0	11.55	2.233	0.913
17	851.3	12.00	0.087	0.592
18	890.1	11.50	0.363	0.159
19	937.9	11.63	0.629	3.750
20	826.0	12.27	0.079	0.046
21	889.5	11.77	0.431	0.830
22	899.0	11.86	0.733	0.620
23	870.0	12.40	0.487	0.048
24	810.1	12.60	0.032	0.027
25	875.3	11.86	0.334	2.350
26	872.8	11.95	0.234	1.486
27	843.2	12.09	0.266	0.137
28	863.0	11.95	0.123	2.730
29	825.0	12.05	0.105	0.498
30	879.9	11.95	0.539	1.178
31	915.0	11.66	0.219	3.890
32	893.7	11.73	0.148	1.850
33	849.5	12.09	0.369	0.149
34	866.5	11.95	0.115	2.700
35	913.4	11.70	1.512	0.544
36	935.9	11.59	1.561	1.200
37	915.1	11.64	0.266	4.224
38	841.9	—	0.305	0.151
39	878.1	11.82	0.163	2.168
40	849.0	12.05	0.423	0.141
41	871.2	11.95	0.251	1.278
42	874.5	11.77	0.142	1.789
43	939.4	11.50	2.156	0.837
44	895.2	11.73	0.762	1.499
45	873.3	—	—	1.370
46	868.5	11.86	0.074	1.353
47	896.3	11.68	0.209	1.665
48	837.4	—	—	0.150
49	878.8	11.82	0.136	2.288
50	844.9	12.05	0.276	0.130
51	871.8	12.43	0.609	0.059
52	914.7	11.68	1.638	0.560
53	808.2	12.60	0.028	0.021
54	868.7	11.86	0.106	2.599
55	896.0	11.73	0.343	0.916
56	871.6	11.79	0.118	1.923
57	939.2	11.50	2.121	0.913
58	876.6	12.00	0.456	1.089
59	912.9	11.64	0.237	4.054
60	838.6	12.09	0.278	0.149
61	870.3	11.95	0.220	1.296
62	896.5	11.73	0.190	1.729
63	838.9	12.05	0.351	0.143
64	936.7	11.59	1.602	1.155

[a]ASTM D1298-99 (2005).
[b]UOP 375-07.
[c]ASTM D664-11a.
[d]ASTM D2622-10.

this work uses the PLS-1 algorithm and deals with only one set of reference values at a time.

In this work, PCA and PLS modeling were performed as implemented in The Unscrambler X® software (CAMO, Oslo, Norway).

Results and Discussion

In this work, 64 samples of neat crude oils obtained from different geographical areas and from different fields were studied. Physical–chemical properties were measured using standard analytical methods, and the data are reported in Tables 1 and 2. Here, it can be seen that the sample set encompasses light and heavy crudes with low to high sulfur content and paraffinic to naphthenic chemical composition. The ^{1}H NMR spectra of neat crude oils were recorded, and as an example, the spectra of two typical crude oils with different viscosity are shown in Fig. 1. The chemical shifts of the main functional groups of the constituent molecules are also reported. Despite using the neat samples, at 300 MHz, the NMR spectral resolution and quality are in general good, and in some cases (light crudes), they are also surprisingly high. Furthermore, some samples, from high to low viscosity, were tested with a 500-MHz spectrometer; no appreciable gain in spectral resolution and quality were observed (spectra not reported).

It is worth point out that the great majority of the works in the NMR literature on viscous matter adopt the common practice to dilute the analyte in a deuterated solvent in order to assure good signal resolution, adequate signal to noise ratio, and substantial lock level. Unfortunately, though, this *modus operandi* brings about solvent overlap issues, sample–solvent interactions (e.g., unwanted signal shifts).[34] This experimental procedure, if applied to complex crude mixtures, compromises the possibility to obtain repeatable data. In fact, any observed variation in the NMR data should be related to the sample intrinsic composition and not to the changes in chemical shifts, line-widths, baseline, or artifacts due to sample preparation. To minimize these effects and to simplify data handling, many different techniques have been proposed.[25–29,34] A disregarded alternative at the laboratory level is to analyze the sample without prior dilution. In fact, although this is *the practice* in process applications, in laboratory, one tends not to renounce the higher resolution reachable working in solution. In our samples, many of the spectral features obtainable in solution would in any

Table 2. TBP distillation yields (%w) according to ASTM D2892-11a

Sample	Boiling point range (°C)											
	Pi[a]-nC4[b]	iC5[c]-70	70–90	90–155	155–175	165–230	175–230	230–250	250–350	350–370	230–370	370+
1	1.27	3.12	1.90	11.42	2.57	11.77	10.28	4.01	22.04	3.97	30.02	39.42
2	1.05	2.8	1.43	9.04	2.83	9.64	8.20	3.35	17.61	3.51	24.47	50.41
3	0.64	2.00	1.04	6.42	2.47	8.47	7.29	3.20	16.98	3.58	23.76	56.38
4	1.43	2.79	1.36	8.47	2.73	8.99	7.75	3.38	16.48	3.22	23.08	52.39
5	1.10	2.90	1.60	9.40	2.80	9.70	8.20	3.30	17.70	3.50	24.50	49.50
6	1.40	3.40	1.80	9.06	3.02	9.58	7.96	3.15	16.28	3.15	22.58	50.76
7	—	—	—	—	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—	—	—	—	—
9	1.19	2.95	1.19	8.46	2.76	9.53	8.17	3.52	18.55	3.52	25.59	49.69
10	1.26	3.11	2.06	10.12	3.24	11.25	9.65	4.27	22.21	4.05	30.53	40.03
11	0.84	2.28	1.81	8.39	2.58	11.94	10.62	5.51	30.05	4.13	39.69	33.79
12	0.68	2.36	1.38	9.06	3.34	11.34	9.56	3.73	19.11	3.88	26.72	46.95
13	1.17	2.00	1.66	6.17	2.15	7.45	6.37	3.09	16.35	3.42	22.86	57.62
14	1.03	3.07	1.65	9.47	3.10	10.12	8.63	3.52	17.86	3.47	24.85	48.23
15	0.76	1.97	1.41	7.84	2.35	9.59	8.39	3.94	21.33	3.90	29.17	48.12
16	0.31	0.53	0.30	1.92	1.03	5.67	5.10	3.12	20.10	4.02	27.25	63.54
17	1.55	3.51	1.79	9.99	3.10	10.67	9.17	3.72	19.61	3.88	27.21	43.68
18	1.21	2.41	1.43	7.28	1.50	8.22	7.43	4.75	31.05	4.77	40.57	38.17
19	0.86	1.95	0.87	4.69	1.60	5.77	5.01	2.18	12.55	2.80	17.53	67.50
20	1.43	5.21	1.99	12.02	3.51	11.28	9.49	3.84	17.69	3.67	25.21	41.14
21	0.92	2.97	1.62	8.09	2.39	8.16	6.97	3.48	19.50	3.93	26.91	50.12
22	0.29	1.11	0.81	5.38	1.60	6.97	6.18	3.10	17.05	3.62	23.77	60.86
23	0.20	0.84	0.73	3.73	1.64	7.37	14.55	7.18	19.12	3.41	—	66.35
24	0.50	2.30	1.32	8.99	3.09	12.31	22.70	10.39	11.34	4.19	—	45.57
25	1.13	3.56	1.79	9.73	3.31	9.86	8.20	3.30	16.30	3.23	22.83	49.45
26	1.22	2.75	1.27	7.92	2.69	8.98	7.64	3.52	18.11	3.65	25.28	51.23
27	1.25	2.81	1.70	9.34	2.93	13.39	24.39	11.00	9.02	5.08	—	43.48
28	1.70	5.20	1.49	11.14	3.30	9.32	7.70	2.84	14.44	3.33	20.61	48.86
29	2.89	6.30	2.81	14.32	4.15	12.02	9.85	3.90	17.88	3.36	25.14	34.54
30	0.92	2.59	1.49	7.47	2.66	8.18	6.85	3.07	16.64	3.52	23.23	54.79
31	1.12	3.05	0.85	6.82	2.44	7.89	6.61	2.51	14.23	3.15	19.89	59.22
32	0.64	1.95	0.97	6.87	2.61	8.50	7.22	3.01	16.71	3.79	23.51	56.23
33	1.09	2.43	1.37	9.14	3.18	10.75	9.09	4.12	22.68	3.68	30.48	43.22
34	1.46	4.87	1.29	10.20	3.41	9.59	7.73	2.84	14.54	3.12	20.50	50.54
35	0.17	0.35	0.45	3.54	1.58	7.28	6.46	3.25	19.63	4.01	26.89	60.56
36	0.29	0.10	0.44	0.48	0.81	4.44	4.00	2.61	18.45	4.26	25.32	68.56
37	1.18	3.10	0.87	6.88	2.48	7.74	6.41	2.58	13.89	3.03	19.50	59.58
38	1.41	3.71	12.14	11.88	3.28	15.90	—	—	—	—	—	40.57
39	1.85	3.27	1.28	6.35	2.63	9.27	7.97	3.26	18.58	3.95	25.79	50.86
40	1.23	2.23	1.32	9.09	3.23	10.66	8.88	4.13	22.76	3.92	30.81	43.21
41	1.21	2.72	1.11	8.23	2.70	9.09	7.65	3.45	18.22	3.76	25.43	50.95
42	0.61	2.37	1.13	8.93	3.48	11.09	9.24	3.78	19.12	3.72	26.62	47.62
43	0.27	0.31	0.28	2.13	0.96	5.27	4.77	2.94	19.68	4.31	26.93	64.35
44	0.61	1.87	0.88	6.67	2.57	9.18	7.85	3.44	19.54	3.44	26.42	53.13
45	1.32	2.95	1.22	8.34	2.69	9.20	7.73	3.19	17.32	3.49	24.00	51.75
46	1.74	3.48	1.29	9.11	2.76	9.95	8.35	3.33	17.56	3.38	24.27	49.00
47	0.57	1.98	1.22	6.54	2.49	8.26	6.92	3.10	17.06	3.61	23.77	56.51
48	1.12	2.97	1.45	8.97	2.77	9.92	8.46	3.44	17.64	3.34	24.42	49.84
49	1.48	3.85	1.45	9.07	2.98	9.53	8.10	3.26	16.06	3.36	22.65	50.43
50	—	—	—	—	—	—	—	—	—	—	—	—
51	0.27	0.97	1.26	3.04	1.56	7.29	14.40	7.11	19.09	—	—	66.52
52	0.11	0.42	0.58	3.32	1.64	7.40	6.58	3.46	19.55	3.96	26.97	60.38
53	0.63	2.46	1.29	8.27	2.96	13.04	23.26	10.22	8.02	5.26	—	42.25
54	1.73	4.64	1.43	9.89	3.09	9.21	7.64	2.86	14.80	3.62	21.28	50.30
55	0.83	2.61	1.31	7.64	2.42	7.96	6.69	3.22	19.13	4.27	26.62	51.88
56	0.94	2.50	1.18	8.83	3.27	11.18	9.35	3.75	18.67	3.54	25.96	47.97

(Continues)

Table 2. *(Continued)*

Sample	Boiling point range (°C)											
	Pi[a]-nC4[b]	iC5[c]-70	70–90	90–155	155–175	165–230	175–230	230–250	250–350	350–370	230–370	370+
57	0.47	0.58	0.84	1.56	0.95	5.31	4.78	3.07	19.30	4.46	26.83	63.99
58	1.05	2.51	1.31	8.21	2.63	7.93	6.61	2.98	16.74	3.40	23.12	54.56
59	1.20	3.19	0.95	7.47	2.58	7.83	6.42	2.59	13.86	2.82	19.27	58.92
60	1.57	3.72	1.63	10.49	3.44	11.09	9.16	4.13	22.34	4.28	30.75	39.24
61	1.28	2.86	1.17	8.40	2.78	9.32	7.75	3.40	17.97	3.73	25.10	50.66
62	0.53	1.88	0.90	7.10	2.19	8.27	7.09	3.12	16.90	3.67	23.69	56.62
63	1.44	3.34	1.84	10.42	3.65	11.07	9.17	4.16	22.45	3.76	30.37	39.77
64	0.11	0.12	0.05	0.83	0.83	4.55	4.09	2.62	19.04	4.16	25.82	68.15

[a]Initial boiling point.
[b]BP (°C) of *n*-butane.
[c]BP (°C) of *i*-pentane.

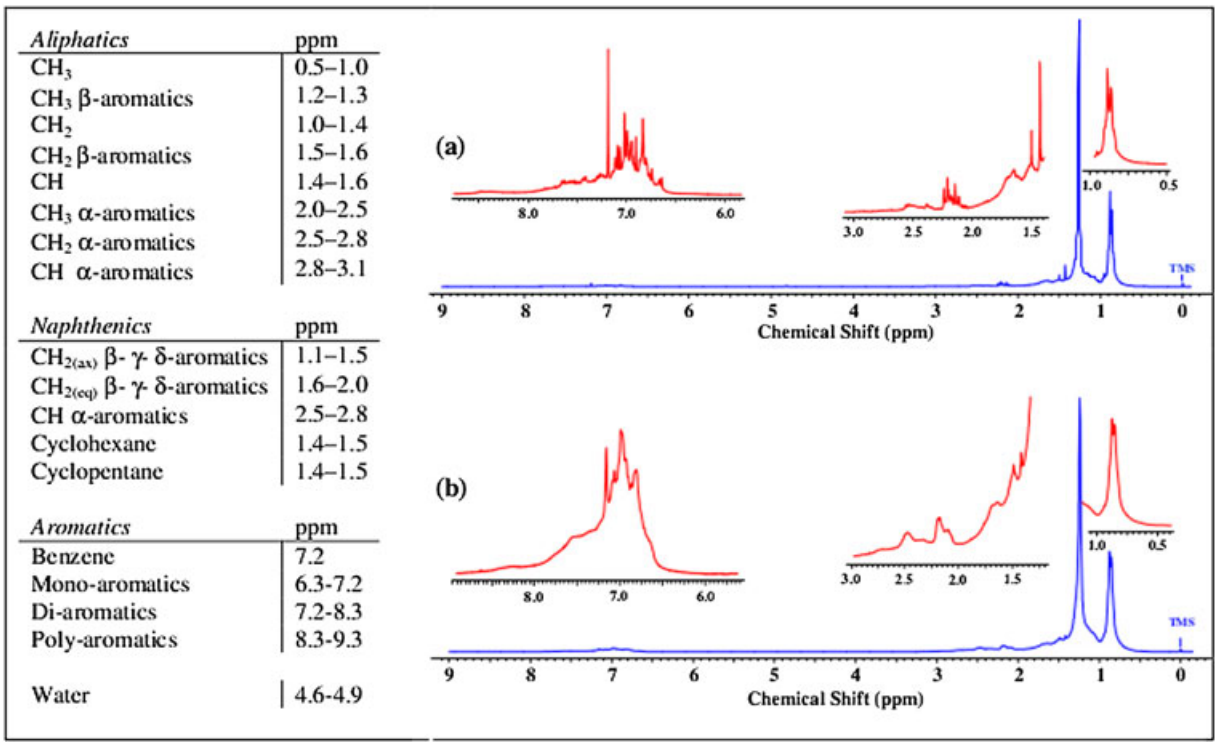

Figure 1. ^{1}H NMR spectra of two crude oils: (a) low viscosity and (b) high viscosity samples. The most important chemical shift regions are also reported.

case be lost because of the binning procedure (see Materials and Methods); thus, the loss of information going from solution to neat is negligible for this application.

The obtained spectral data were submitted to the multivariate analysis to correlate the relative intensity and position of NMR resonance peaks to laboratory data of physical and chemical properties determined by reference analytical methods. In that, the expected quality of a multivariate NMR model is limited to the precision of the reference values.

In general, the correlation between the analytical data and the spectral NMR data is obtained via multivariate calibration methods such as multilinear regression, PCR and PLS regression. The latter method is particularly suited for constructing predictive models when the variables are many and highly collinear as in the case of ^{1}H NMR spectra of crude oils.[35] It is generally not known *a priori* whether a multivariate model of adequate precision can be developed; in this case, feasibility studies have to be performed, and if successful, they can be expanded. Choosing adequate data pretreatment, setting the proper number of samples for each data set, and assessing the dimensionality of the optimal model are all issues to be considered. Finally, the resulting multivariate calibration models are applied to NMR spectra of unknown samples to provide an estimate of the physical and chemical property values.

The spectral data were submitted to the PLS analysis. After a preliminary calibration on the entire data set (not reported), the prediction ability of the models was tested using external test set validation. Therefore, the normalized spectral data of the 64 samples (Tables 1 and 2) were divided into a calibration and a validation set (samples from 1 to 47 and from 48 to 64, respectively) by analyzing the sample distribution with a PCA. In particular, in order to keep as much as possible the variability of the original set, the validation set was selected by a visual inspection of the PCA score plot shown in Fig. 2, verifying that one sample representative of each extraction field was present, and if only one sample of a crude typology was available, it was included in the calibration set; the validation set thus contained less samples, although equally spread in variance. When the spectra were projected in the space of the first and second components (using their scores as coordinates), we obtained natural clustering of samples into different groups, according to their spectral similarity and physical and chemical properties. For example, on the right of the plot, we find light samples, on the left the heavy ones (Fig. 2). After a careful analysis of the spectra, we can say that PC1 and PC2 are related to chemical shift and peak width variations, respectively. The first one explains 71% of the variance and is linked to the chemical composition of each crude oil typology, whereas peak width explains 24% of the variance and reflects the different *D* and viscosity of the samples. One of the strengths of PCA is to provide a quick view of the sample distribution and, thereby, to identify samples that exhibit deviating features (outliers) or discover trends and groups. As previously mentioned, the PCA of the NMR spectra explains 95% of the variance with the first two principal components and no one sample fell outside the 95% (Hotelling's T^2) confidence ellipse, indicating the absence of potential outliers.

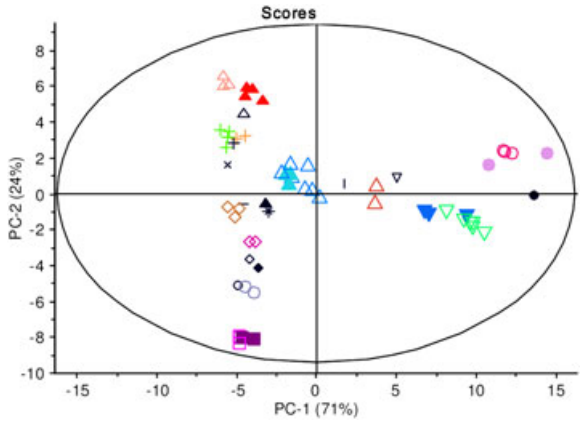

Figure 2. PC1 versus PC2 score plot of NMR spectral data for 64 samples of neat crude oil. Variance is reported in brackets; each symbol identifies a crude type.

Table 3. Results of PLS modeling and prediction of the crude oil properties based on ^{1}H NMR spectroscopy

Property	Value	σ_R	No. of factors	R^2	SEC	SEP
D (kg/m^3)	808.2–939.4	0.5	7	0.991	2.8	3.0
K_{UOP} (–)	11.50–12.60	>0.5	7	0.960	0.05	0.06
Pi-C4 (%w)	0.11–2.89	0.47	7	0.872	0.19	0.23
C5-70 (%w)	0.10–6.30	0.47	6	0.930	0.34	0.30
70–90 (%w)	0.05–12.14	0.47	6	0.852	0.19	0.20
90–155 (%w)	0.48–14.32	0.47	6	0.974	0.44	0.67
155–175 (%w)	0.81–4.15	0.47	6	0.940	0.17	0.24
165–230 (%w)	4.44–15.90	0.47	7	0.941	0.44	0.65
175–230 (%w)	4.00–24.39	0.47	6	0.920	0.41	0.61
230–250 (%w)	2.18–11.00	0.47	6	0.880	0.16	0.19
250–350 (%w)	8.02–31.05	0.47	6	0.830	0.98	0.88
350–370 (%w)	2.80–5.26	0.47	6	0.709	0.19	0.26
230–370 (%w)	17.53–40.57	0.47	6	0.839	1.21	1.09
370+ (%w)	33.79–68.56	0.47	7	0.970	1.31	1.74

To assess the feasibility of the application, a PLS model was built for each property. Calibration was based on 'leave-one-out' cross-validation, whereas validation was based on the predictions of the test set (see Section on Materials and Methods). Tables 3 and 4 show the mean value for each property, and the results are presented in terms of number of factors used, correlation coefficient (R^2), standard error of calibration (SEC), and standard error of prediction (SEP). SEC and SEP are the standard deviation for the differences between measured and NMR estimated values for samples within the calibration and validation set, respectively, as defined in Eqn (1):

$$\text{SEC},\ \ \text{SEP} = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(\hat{y}_i - y_i - \text{Bias})^2} \qquad (1)$$

where I is the number of samples, $\hat{y}_i$ is the predicted value (either in calibration or validation), and y_i is the measured value. They measure the total residual error due to the particular regression equation to which it applies. The SEC statistic is a useful estimate of the theoretical 'best' accuracy obtainable for a specified set of variables used to develop a calibration model. This provides an indication of the goodness of the model when compared against the standard deviation calculated from the reference method reproducibility (σ_R). When the precision of method is constant across the range of reference values used, to demonstrate the agreement between the model and the

Table 4. Results of PLS modeling and prediction of the crude oils properties based on ^{1}H NMR spectroscopy

Property	Value	Reproducibility	No. of factors	% within ± reproducibility[a]	R^2	SEC	SEP
TAN (mg_{KOH}/g)	0.03–2.23	0.141–0.456	7	94	0.878	0.18	0.17
S (%w)	0.02–4.22	0.003–0.218	8	23	0.967	0.20	0.25

[a]Percentage of the residual values between the reference and predicted values lying within plus or minus the reproducibility.

reference methods, the SEP was compared with the σ_R, as reported in Table 3. On the other cases, the agreement is demonstrated if the percentage of the residual values between the reference and predicted values lies within plus or minus the reproducibility at any given confidence interval (Table 4). As shown in Table 3, the number of factors needed to develop the models is always compatible with the required sample size.[2] Results are also depicted as prediction plots in Figs 3, 4, and 5.

In most cases, prediction of crude properties through the use of NMR spectra resulted to be in line with the corresponding ASTM methods. Performances are particularly good for K_{UOP} factor, TAN, and TPB distillation yields, whose SEC and SEP values match the ASTM precision (Tables 3 and 4). K_{UOP} (Fig. 3b) is a compositional index used to classify crudes in terms of their paraffinic, naphthenic, and aromatic content; the direct link to the compositional information contained in the NMR spectra makes it particularly well suited for the NMR approach, and indeed, the model standard error is exceptionally low. On the contrary, the TAN (Fig. 3(c)), which measures the total acid content, displays a relatively low correlation coefficient (0.878). Furthermore, the TAN experimental data determined on the 64 crudes cover a wide range, although most of the samples are concentrated in the low value region. In this case, the high TAN value samples show very high leverage that indicates their influence on the model. Nevertheless, and despite first sight appearance of the prediction plot, a closer inspection of the single sample residuals shows very good agreement with ASTM in 94% of the cases. As depicted in the predicted versus measured plots in Figs 3, 4, and 5, equally performing models may differ greatly.

Moreover, distillation yields number among those physical properties of primary importance in establishing the economic value of crude and in determining their processing route and, therefore, the accuracy level in the prediction model may make a difference. A fast and reliable method alternative to the time-consuming TBP would be particularly welcome by refinery professionals. In general, the PLS models for the yields perform within desired limits. The SEC are within the reference precision in most cases, and correlation coefficients vary from 0.709 to 0.974 (Tables 3 and 4). Coefficients above 0.6 indicate a good correlation among the spectral data and properties. Generally, the absolute error (i.e., SEC) increases with the percentage yield so that, for example, the absolute error on the 230–370 °C cut, on average, is higher than that on the 165–230 °C cut. On the contrary, the relative error seems constant (Table 3). Yields are also peculiar in the fact that, by definition, they must add up to 100%, and thus, an error in the laboratory analysis for any single cut is carried on to the other distillation points. Because each yield percentage is modeled independently from the others, predicted data may not add up to 100%. On the other hand, experimental errors may average out, thus leading to better overall prediction. As to *D*, this parameter is off the required ASTM performance level. This result is not surprising, because *D* is one of the toughest physical parameters to be modeled by indirect methods, given the very high intrinsic precision of the laboratory analysis. Nonetheless, the results obtained show that the performance is still acceptable for these models to be used for indicative estimates of crude quality and *D* group type.[1] Finally, the model for determining the percentage of *S* represents a singular case. Figure 3(d) clearly shows the very wide

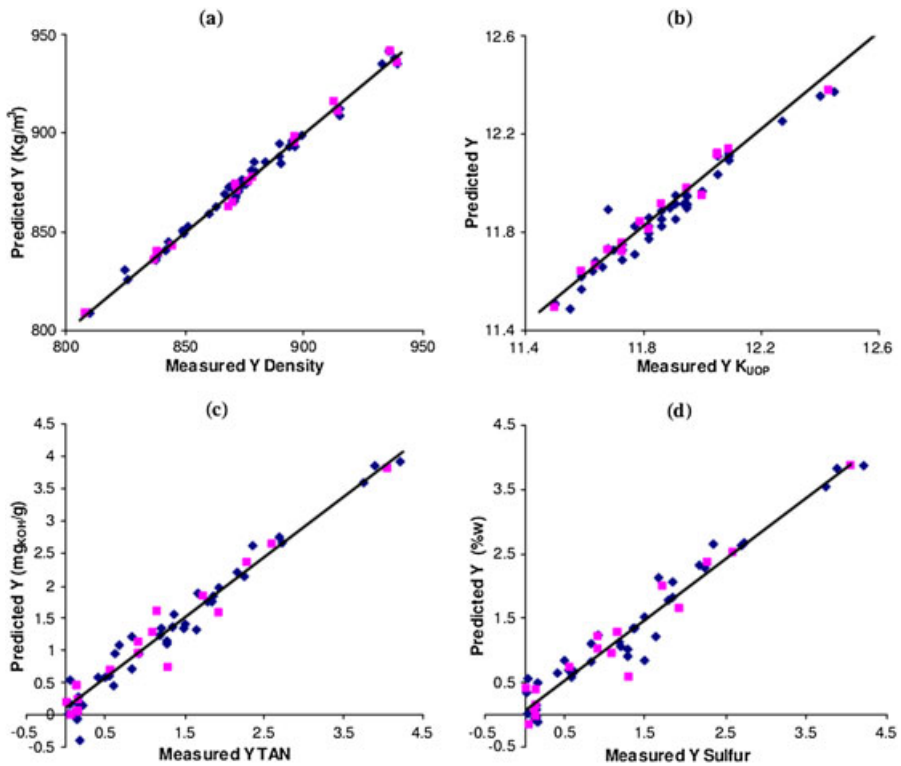

Figure 3. Prediction plots of PLS modeling of (a) density, (b) K_{UOP} factor, (c) TAN, and (d) sulfur content. Models are based on a calibration set of 47 spectra (♦) and on a validation set of 17 spectra (■).

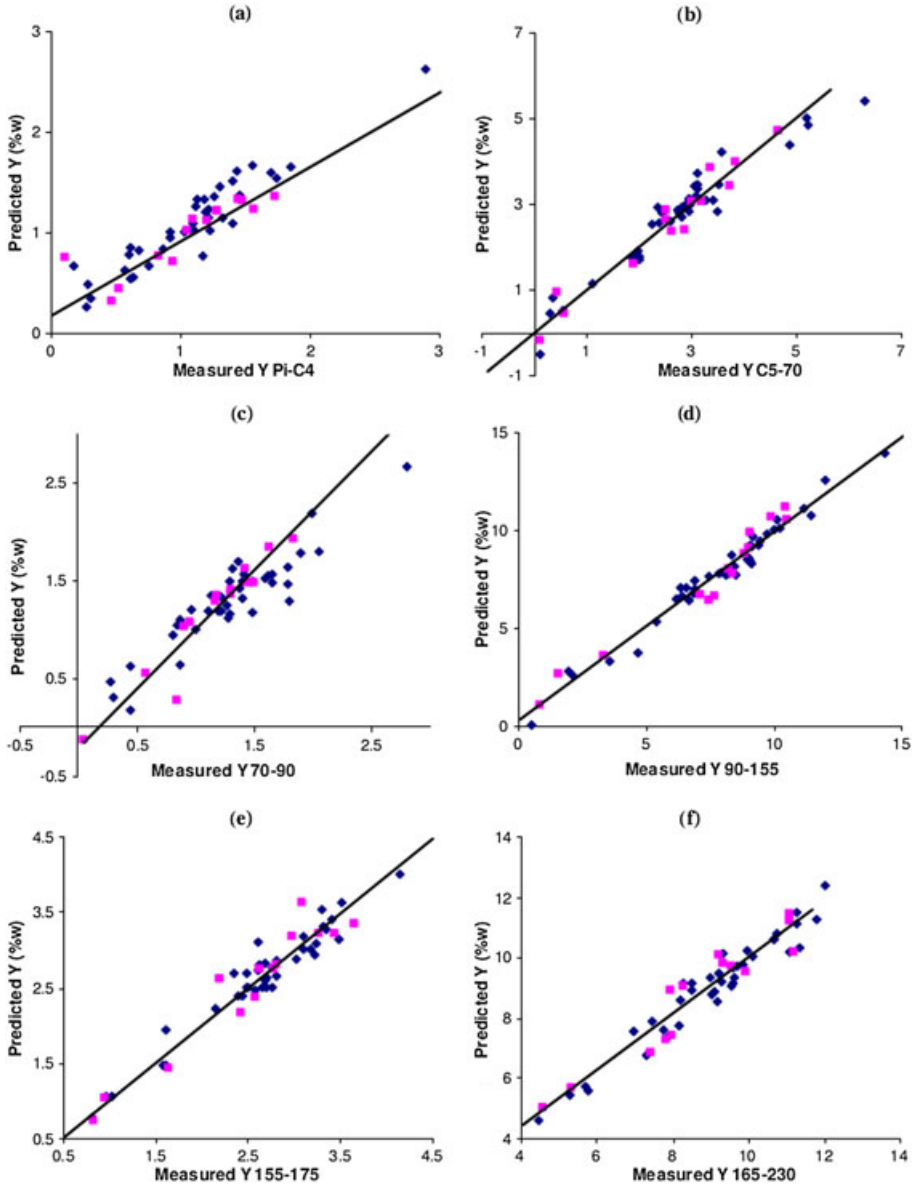

Figure 4. Prediction plots of PLS modeling of TBP yields (%w) for each distillation range in °C: (a) Pi-C4, (b) C5-70, (c) 70–90, (d) 90–155, (e) 155–175, and (f) 175–230. Models are based on a calibration set of 47 spectra (♦) and on a validation set of 17 spectra (■).

distribution in the data ranging from 0.02% to 4.2% *S*, and it can observed that all the samples with *S* below 0.4% are predicted very poorly. The model is evidently leveraged on the high *S* samples. The SEC and SEP in this case are much larger than the matching ASTM reproducibility.

Choosing to model together many types of crudes with very different spectral characteristics and spanning a wide range in most of the properties pose a serious challenge to obtaining quality predictions. Although as a whole, the data set pictures a situation compatible with what one would expect to find in complex refining sites in terms of crude variability and, therefore, is very well suited to confirm the value and broadness of the results obtained herein. On the other side of the coin, because of specificity of the data set, in some cases we had to deal with models spanning few orders of magnitude in the property of interest, as in the case of *S* and TAN, or models in which the data were not uniformly distributed throughout the calibration range (e.g., TAN). The former case might require development of range-specific models, but this would call for additional samples. The latter issue concerns the possible onset of inliers when predicting unknown samples. Again, if models should have adherence to the real world, this is an issue one has to cope with. Most importantly, provided that the unknown samples remain of the same types of crude included in the calibration, this issue looses importance.

Conclusions

In this work, we demonstrated that the use of ^{1}H NMR spectroscopy on neat crude oil samples combined with PLS modeling has high potential to predict crude oil properties, even in the cases of complex refining sites operating with a variable and

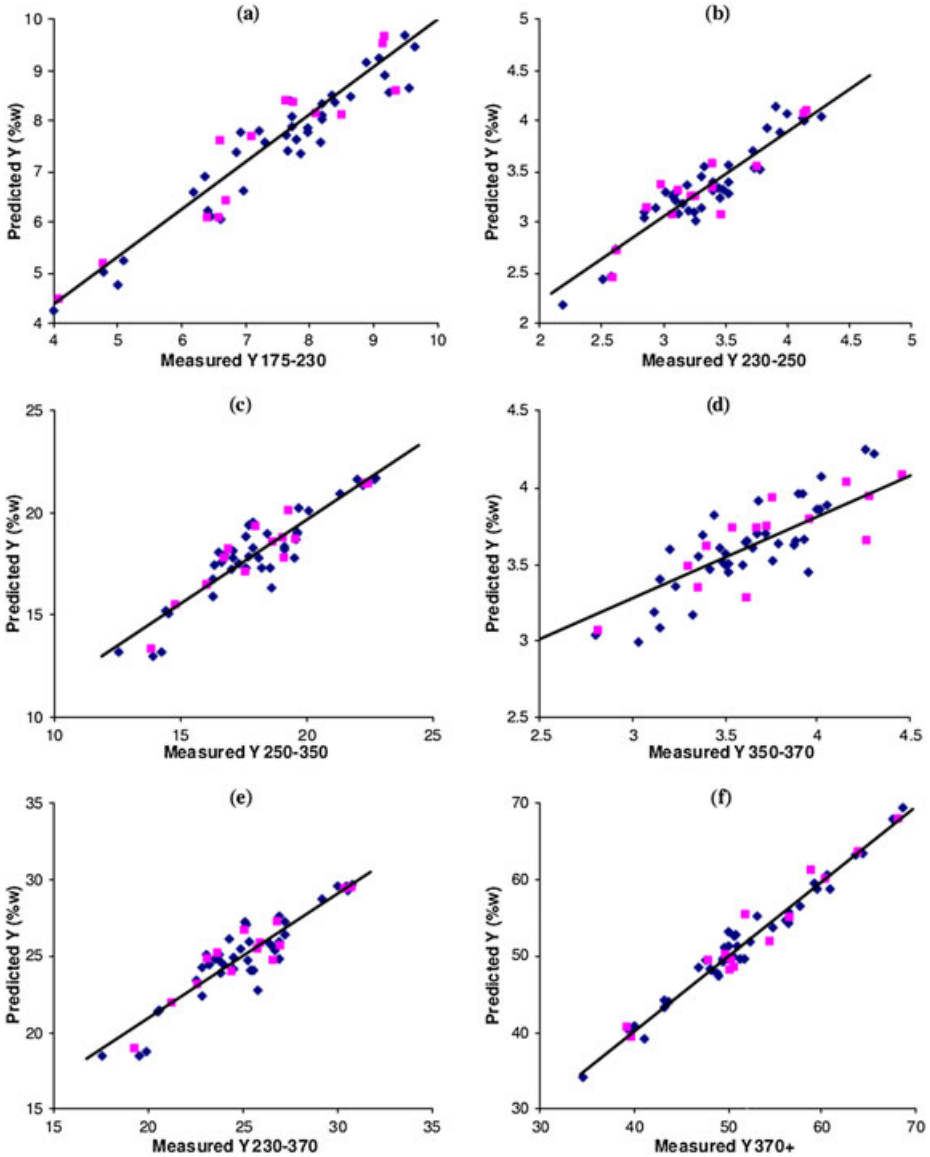

Figure 5. Prediction plots of PLS modeling of TBP yields (%w) for each distillation range in °C: (a) 165–230, (b) 230–250, (c) 250–350, (d) 350–370, (e) 230–370, and (f) 370+. Models are based on a calibration set of 47 spectra (♦) and on a validation set of 17 spectra (■).

wide crude slate. This method offers a very fast alternative to the time-consuming laboratory methods, avoiding tedious sample preparation and elaborate spectral data pretreatment. The prediction plots and SEP values obtained for the validation set show, in most cases, equivalent precision between the PLS models from ^{1}H NMR data and the corresponding ASTM method. The data set covers a very wide range of crudes from super light to super heavy, from very low to high sulfur content, suggesting that when increasing the number of samples and possibly developing range-specific models a further improvement in accuracy and robustness can be obtained. Despite this, the obtained results are already more accurate than those reported in the literature regarding more homogeneous and less variable sample sets.

Acknowledgements

A. Masili gratefully acknowledges Sartec S.p.A. for her doctoral grant. The authors wish to thank the laboratory at SARAS Refinery (Sarroch, Italy) for kindly supplying crude oil samples and Sartec staff for the crude assay results used in this study.

References

[1] J. Speight, *The Chemistry and Technology of Petroleum*, Marcel Dekker Inc., New York, **1999**.
[2] Analytical methods (http://www.astm.org/)
- ASTM E1655-05, "Standard Practices for Infrared Multivariate Quantitative Analysis".
- ASTM D1298-99, "Density, Relative Density (Specific Gravity), or API Gravity of Crude Petroleum and Liquid Petroleum Products by Hydrometer Method".
- UOP375-07, "Calculation of UOP Characterization Factor and Estimation of Molecular Weight of Petroleum Oil".
- ASTM D2892-11a, "Distillation of Crude Petroleum (15-Theoretical Plate Column)".
- ASTM D664-11a, "Acid Number of Petroleum Products by Potentiometric Titration".
- ASTM D2622-10, "Sulfur in Petroleum Products by Wavelength Dispersive X-ray Fluorescence Spectrometry".

[3] V. D. Molina, U. N. Uribe, J. Murgich, E. Petro, *Energy Fuel* **2007**, *21*, 1674–1680.
[4] P. De Peinder, T. Visser, D. D. Petrauskas, F. Salvatori, F. Soulimani, B. M. Weckhuysen, *Energy Fuel* **2009**, *23*, 2164–2168.
[5] S. L. Silva, A. M. S. Silva, J. C. Ribeiro, F. G. Martins, F. A. Da Silva, C. M. Silva. *Anal. Chim. Acta* **2011**, *707*, 18–37.
[6] P. De Peinder, D. D. Petrauskas, F. Singelenberg, F. Salvatori, T. Visser, F. Soulimani, B. M. Weckhuysen, *Appl. Spectrosc.* **2008**, *62*, 414–422.
[7] S. Satya, R. M. Roehner, M. D. Deo, F. V. Hanson, *Energy Fuel* **2007**, *55*, 1444–1451.
[8] J. Jehlička, O. Urban, J. Pokorný, *Spectrochim. Acta, Part A* **2003**, *59*, 2341–2352. doi:10.1016/S1386-1425(03)00077-5.
[9] J. Sjoblom, H. Kullberg, S. Wold, *Chemom. Intell. Lab. Syst.* **1998**, *44*, 229–244.
[10] K. A. Lintelmann, *Anal. Chem.* **1995**, *67*, 327–331.
[11] L. Petrakis, D. Allen, *Ind. Eng. Chem. Process Des. Dev.* **1987**, *22*, 298–305.
[12] B. Behera, S. S. Ray, I. D. Singh, *Fuel* **2008**, *87*, 2322–2333. doi:10.1016/j.fuel.2008.01.001.
[13] P. L. Gupta, P. V. Dogra, R. K. Kuchhal, P. Kumar, *Fuel* **1986**, *65*, 515–519.
[14] D. F. Kushnarev, T. V. Afonina, G. A. Kalabin, R. N. Presnova, N. I. Bogdanova, *Petrol. Chem. U.S.R.R.* **1990**, *29*, 149–159.
[15] M. U. Hasan, M. F. Ali, A. Bukhari. *Fuel* **1983**, *62*, 518–523.
[16] T. M. Alam, M. K. Alam, *Annu. Rep. NMR Spectrosc.* **2005**, *54*, 42–80. doi:10.1016/S0066-4103(04)54002-4.
[17] J. Burri, R. Crockett, R. Hany, D. Rentsch. *Fuel* **2004**, *83*, 187–193. doi:10.1016/S0016-2361(03)00261-8.
[18] http://www.process-nmr.com/
[19] T. Brekke, O. M. Kvalheim, E. Sletten, *Anal. Chim. Acta* **1989**, *223*, 123–134.
[20] H. Witjes, M. Van den Brink, W. J. Melssen, L. M. C. Buydens, *Chemom. Intell. Lab. Syst.* **2000**, *52*, 105–116. doi:10.1016/S0169-7439(00)00085-X.
[21] K. E. Nielsen, J. Dittmer, A. Malmendal, N. C. Nielsen, *Energy Fuel* **2008**, *22*, 4070–4076.
[22] A. T. Castro, *J. Braz. Chem. Soc.* **2006**, *17*, 1181–1185.
[23] R. Meusinger, R. Moros, *Fuel* **2001**, *80*, 613–621.
[24] C. R. Kaiser, J. L. Borges, A. R. dos Santos, D. A. Azevedo, L. A. D'Avila, *Fuel* **2010**, *89*, 99–104. doi:10.1016/j.fuel.2009.06.023.
[25] J. T. W. E. Vogels, A. C. Tas, J. Venekamp, J. Van Der Greef. *J. Chemom.* **1996**, *10*, 425–438.
[26] H. Witjes, W. J. Melssen, H. J. A. in 't Zandt, M. Van Der Graaf, A. Heerschap, L. M. C. Buydensm, *J. Magn. Reson.* **2000**, *44*, 35–44. doi:10.1006/jmre.2000.2021.
[27] J. Forshed, I. Schuppe-Koistinen, S. P. Jacobsson, *Anal. Chim. Acta* **2003**, *487*, 189–199.
[28] J. Forshed, R. J. O. Torgrip, K. M. Åberg, B. Karlberg, J. Lindberg, S. P. Jacobsson, *J. Pharm. Biomed. Anal.* **2005**, *38*, 824–832.
[29] R. Stoyanova, T. R. Brown, *J. Magn. Reson.* **2002**, *154*, 163–175.
[30] T. R. Brown, R. Stoyanova, *J. Magn. Reson. B* **1996**, *112*, 32–43.
[31] D. V. Molina, U. N. Uribe, J. Murgich, *Fuel* **2010**, *89*, 185–192. doi:10.1016/j.fuel.2009.07.021.
[32] S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
[33] S. Wold, P. Geladi, K. Esbensen, J. Öhman, *J. Chemom.* **1987**, *1*, 41–56. doi:10.1002/cem.1180010107.
[34] A. Mitra, P. J. Seaton, R. A. Assarpour, T. Williamson, *Tetrahedron* **1998**, *54*, 15489–15498.
[35] P. Geladi, B. R. Kowalski, *Anal. Chim. Acta* **1996**, *185*, 1–17.