

Deconvolving the absorbance of methyl and methylene groups in the FT-IR 3000-2800 cm⁻¹ band of petroleum fractions

Nikos Pasadakis^{1,*}, Giorgos Livanos² and Michalis Zervakis²

¹Mineral Resources Engineering Department, ²Electronic and Computer Engineering Department, Technical University of Crete, Chania, Greece

ABSTRACT

A new algorithm is proposed to deconvolve the infrared spectrum of complex hydrocarbon mixtures in the 3000-2800 cm⁻¹ region. The algorithm enables the accurate estimation of the contribution of methyl and methylene groups in petroleum samples, which is highly characteristic for their composition. The algorithm is developed based on the analysis of FT-IR spectra of seventy oil fractions, practically covering the whole range of a petroleum refinery intermediate and final products. The experimentally derived spectra are deconvolved by fitting three Lorentzian and one asymmetric Gaussian distributions, corresponding to methyl and methylene asymmetric and symmetric stretching vibrations. Molar absorptivities for these peaks are estimated from the FT-IR spectra of pure n-alkanes and alkyl-aromatics. The curve fitting procedure is implemented in Sequential Quadratic Programming (SQP) utilizing linear and non-linear constraints to incorporate chemical information, including the absorbance band positions and their molar absorptivity values. The developed methodology manages to reconstruct efficiently the FT-IR spectra of petroleum fractions, as indicated by the Mean Square Error (MSE) metric. The correctness of the selected peaks (position, amplitude) is further demonstrated by the practically constant ratios of the peak areas obtained for the asymmetric and symmetric methyl and methylene absorption bands, respectively of the

whole data set. The algorithm facilitates the spectra modeling and the accurate estimation of the fitted methyl and methylene peak areas, which can be used for calculating specific compositional parameters of oil samples instead of the usually employed peak heights. Such modeling is extremely important for heavy petroleum fractions, where detailed compositional information is difficult to be obtained.

KEYWORDS: infrared spectroscopy, spectrum deconvolution, curve fitting, nonlinear optimization, petroleum composition

1. INTRODUCTION

Detailed compositional data of petroleum fractions is appropriate in multiple tasks like identification, characterization and process optimization. Due to the inherent compositional complexity of petroleum, even when dealing with light or middle fractions, an exhaustive description of all compounds present is not possible due to analytical and/or economical limitations. Therefore, bulk compositional characteristics expressed as structural parameters, are often employed. Especially, in the case of heavy oil fractions or compound groups like resins and asphaltenes, this approach is practically the only available characterization scheme today.

Infrared (IR) spectroscopy certainly constitutes one of the most widely employed methods in petroleum analysis due to its ability to handle nearly any sample in any physical state, providing valuable information about its molecular structure in a rapid

*Corresponding author: pasadaki@mred.tuc.gr

and inexpensive manner. An additional advantage of this technique is its ability to reflect sample composition without influencing its “internal equilibrium”, a feature usually unavoidable when dynamic methods, like chromatography, are utilized. The latter is of great importance in the analysis of heavy petroleum fractions that contain a vast quantity of compounds in a more or less unknown physicochemical equilibrium status.

The infrared spectrum of a molecule, being its unique characteristic signature, can be used as a sensitive fingerprint for identification purposes by comparison to a reference spectrum. IR spectroscopy is employed for light petroleum fractions for fingerprinting and identification purposes, even if the presence of a large number of individual components leads to a complex spectrum with a strong overlap between different absorption bands, which obstructs the identification of any single compound. Nevertheless, the spectrum remains a unique characteristic of the specific oil sample and can be used effectively to define differences or similarities with other fractions. For middle and heavy cuts, IR spectrum is fundamental for understanding functional properties, since a more detailed compositional characterization is usually not available.

C-H vibrations of methyl, methylene and methyne groups are dominant for molecules containing aliphatic fragments, as most of the petroleum constituents do. The presence of carbon atoms in different structural units, (methyl terminated aliphatic chains, bridges between aromatic units, methylene chains, naphthenic units) may be approximated using ^{13}C NMR techniques as it has been reported in many studies [1, 2]. Nevertheless, NMR analysis is still far from being a routine experimental technique in petroleum laboratories. Therefore, the use of alternative available analytical methods enabling structural characterization of the samples, such as IR spectroscopy is of great importance [3, 4].

In the $3000\text{--}2800\text{ cm}^{-1}$ region of the IR spectrum, asymmetric and symmetric methyl and methylene group vibrations give rise to absorption bands in the $2970\text{--}2950$, $2880\text{--}2860$, $2935\text{--}2915$, $2865\text{--}2845\text{ cm}^{-1}$ ranges, respectively. Their relative abundance reflects characteristic structural features of petroleum samples (e.g. alkyl chains length).

Peak heights corresponding to these bands have been used to calculate empirical compositional parameters related to the paraffinic character of samples [5]. Coelho *et al.* [6] showed through the analysis of alkyl-benzenes that the $2927/2957\text{ cm}^{-1}$ absorbance height ratio is linearly related to the relative concentration of methylene to methyl groups in asphaltene fractions and may be used to calculate the length of the terminal aliphatic chains. In the same work, a detailed survey on IR applications in structural characterization of petroleum is provided. In other works [7, 8], methyl/methylene height ratios are demonstrated to correlate well with geochemical parameters of kerogen, asphaltene and bitumens.

The heights of absorbance peaks in the $3000\text{--}2800\text{ cm}^{-1}$ region of petroleum spectra are commonly used as qualitative parameters, since the existing strong overlap of methyl/methylene absorptions prevents the accurate quantitative estimation of the corresponding areas. Although the height of a peak depends on the concentration and the molar absorptivity of the specific group, peak broadening and displacement are often experienced, contributing to significant inaccuracies. Distortions in both peak shape and maximum position of specific groups in a mixture spectrum are caused by the surrounding molecules. Complex distributions combining characteristics of both Gaussian and Lorentzian functions are usually employed to simulate absorption bands in liquid mixtures like petroleum and its fractions. Peak area constitutes a representative quantitative indicator, because the final peak profile is the sum of all the individual elements contributing to the absorption. As a result, if correctly calculated, the peak areas remain constant independently of possible shape distortions.

The deconvolution of overlapping peaks in IR spectra is usually carried out using Fourier Self-Deconvolution (FSD), Partial Least squares (PLS) and the classical curve-fitting methods [9]. The goal of the latter is to mathematically create individual distributions under the experimental signal, the sum of which should match as precisely as possible with the original experimental spectrum. A least-square minimization of error between the experimental and the reconstructed spectra is often employed. In most commercially available deconvolution software packages, the selection of the appropriate number of distributions to be

fitted, their position and width are usually specified by the user so that the experimental spectrum could be reconstructed with a minimum error. In order to meet convergence criteria, the algorithm may ignore existing peaks, add additional absorption bands, or deform peak shapes, leading to physically meaningless results as the fitted distributions cannot be associated with real components existing in the sample. Moreover, the characteristics of the curves (center, amplitude, and variance) are computed in a way that guarantees similarity to the experimental spectrum, and do not correspond to the absorbance of underlying components (functional group in our case). As a result, these characteristics cannot be safely used in quantitative calculations. Thus, the introduction of complementary information about the nature of the sample becomes necessary for obtaining substantial and valid results.

Following this direction, in the current work we present a novel curve-fitting algorithm, which effectively determines the peak shapes of the absorption bands corresponding to the methyl and methylene groups in the $3000\text{--}2800\text{ cm}^{-1}$ region of the IR spectrum of hydrocarbon mixtures. The spectrum is deconvolved by modelling the original signal as sum of three Lorentzian and one modified asymmetric Gaussian distributions. The novelty in our approach lies on the fact that the convergence of the deconvolution algorithm is guided by an additional structural constraint, implying that the peak areas of the symmetric and asymmetric vibrations of both methyl and methylene groups should be equal, if weighted by their specific molar absorptivity, since they rise from the same population in each sample. The molar absorptivities are experimentally determined from the IR spectra of pure hydrocarbons (alkanes, cycloalkanes, alkyl aromatics), for which the quantities of the methyl and methylene groups are known. Another feature of the proposed methodology is the fact that the optimization procedure can be enriched with any meaningful constraint defined by the user, thus coupling appropriate chemical information with the curve-fitting procedure and the mathematical model adopted. In addition, no prior knowledge or assumption on the nature of the data is required in the optimization process. The deconvolution methodology was tested on an extended set of petroleum fractions covering the range from volatile

naphthas to solid asphaltic residues. Given that no other constraints (e.g. forced convergence limitations, amplitude and variance range of the calculated distributions) were applied to the algorithm, the obtained quality of fit and the stability of the peak maxima positions indicate the effectiveness of the proposed methodology. The fact that the deconvolution was applied to samples with significant compositional differences and, therefore, significantly different spectral patterns, ensures the robustness of the algorithm and its applicability to a wide range of petroleum hydrocarbon mixtures. The use of our methodology facilitates the reliable determination of methyl and methylene peak areas to be used instead of peak heights, thus leading to more accurate compositional calculations.

2. MATHEMATICAL ALGORITHM

A characteristic of a large class of constrained-optimization methods is the transformation of the constrained criterion into a basic unconstrained problem penalizing solutions that are near or beyond the constraint boundary. In this way, the constrained problem is solved using a sequence of parameterized unconstrained optimization stages, which at the limit (of the sequence) converge to the constrained problem. More recent and efficient methods are focusing on the solution of the Karush-Kuhn-Tucker (KKT) equations as necessary optimality conditions for a constrained-optimization problem. The Sequential (or Successive) Quadratic Programming (SQP) algorithm has been established as one of the most successful methods for solving non-linear constrained optimization problems. It constitutes a powerful tool for minimizing functions of several parameters subject to well defined equality and/or inequality constraints. A nonlinear programming problem is described as

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && f(\mathbf{x}) \\ & \text{subject to} && h_i(\mathbf{x}) = 0 \quad , i = 1, \dots, m \\ & && g_j(\mathbf{x}) \leq 0 \quad , j = 1, \dots, n \end{aligned} \quad (1)$$

where $f(\mathbf{x})$ is the objective function to be optimized, \mathbf{x} is the parameters vector, $h(\mathbf{x})$ is the equality constraint function and $g(\mathbf{x})$ the inequality one. Inequality constraints of the form $g'(\mathbf{x}) \geq 0$ can be rewritten as $g(\mathbf{x}) = -g'(\mathbf{x})$ and a maximization procedure could be applied if we use the formula $f'(\mathbf{x}) = -f(\mathbf{x})$. The equality and inequality constraints

may be linear, expressed in the form $\mathbf{A}^*\mathbf{x} - \mathbf{b} = \mathbf{0}$ and $\mathbf{C}^*\mathbf{x} - \mathbf{d} \leq \mathbf{0}$, or non-linear in general.

The formulation through Lagrangian functions achieves the linearization of both inequality and equality constraints. The Lagrangian of the problem in equation (1) is expressed as:

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) - \sum_{i=1}^m \lambda_i h_i(\mathbf{x}) - \sum_{i=1}^m \mu_i g_i(\mathbf{x}), \quad (2)$$

where λ , μ are the Lagrange multipliers of constraint functions h and g , respectively. The idea of SQP is to model the quantities described above at the current point \mathbf{x}_k (k is the current number of the iteration state) by a quadratic sub-problem and to use its solution to find the new point \mathbf{x}_{k+1} . Adopting the Lagrangian L as the objective functional, the original optimization problem is transformed to [10, 11]:

$$\begin{aligned} & \text{minimize} \quad \nabla L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)^T d(\mathbf{x}) + \frac{1}{2} d(\mathbf{x})^T HL(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k) d(\mathbf{x}) \\ & \text{over} \quad d(\mathbf{x}) \in \mathfrak{R}^n \\ & \text{subject to} \quad h(\mathbf{x}^k) + \nabla h(\mathbf{x}^k)^T d(\mathbf{x}) = 0 \\ & \quad \quad \quad g(\mathbf{x}^k) + \nabla g(\mathbf{x}^k)^T d(\mathbf{x}) \leq 0 \end{aligned} \quad (3)$$

where $d(\mathbf{x}) = \mathbf{x} - \mathbf{x}^k$ is the parameters vector displacement, $L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)$ is the Lagrangian of the objective function $f(\mathbf{x})$, $\boldsymbol{\lambda}^k, \boldsymbol{\mu}^k$ are the Lagrange multipliers vectors of the constraint functions $h(\mathbf{x})$, $g(\mathbf{x})$, respectively at iteration step k and $HL(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)$ is the Hessian (matrix of second partial derivatives) of the Lagrangian L , defined as

$$HL(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k) = \frac{\partial^2 L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)}{\partial \mathbf{x}^k \partial \mathbf{x}^k} \quad (4)$$

The SQP implementation consists of three main stages. At the first stage, the Hessian matrix of the Lagrangian is updated, while remaining positive definite. The second phase includes the quadratic-programming procedure where the estimate of the active constraints at the solution point and the “search direction” \hat{d}_k are updated (via minimization of the objective function while remaining on any active constraint boundaries). After the new search direction is found, the next iteration point \mathbf{x}_{k+1} and the Lagrange multipliers $\boldsymbol{\lambda}_{k+1}$ and $\boldsymbol{\mu}_{k+1}$ are extracted. The final stage involves the line search and merit function calculation, which determine the global convergence of the algorithm among the computed local minima by means of appropriately selected merit functions M . At the $(k+1)$ -st iteration, having determined the Newton

increment $d(\mathbf{x})$, suitable step length α is computed such that $M(\mathbf{x}^k + \alpha d) < M(\mathbf{x}^k)$. The most popular choices of merit functions are augmented Lagrangian and l_p -norms, $p \geq 1$ [12].

In our model, the objective function $f(\mathbf{x})$ is the Mean Square Error (MSE) between the approximated spectrum through the curve fitting algorithm and the original experimental FT-IR spectrum, defined as

$$MSE(\mathbf{x}) = \frac{1}{n} \bullet \sum_{i=1}^n \left(S_{i-\text{estimated}}(\mathbf{x}) - S_{i-\text{original}}(\mathbf{x}) \right)^2, \quad (5)$$

where n is the number of data points in the spectrum and $S_{i-\text{estimated}}(\mathbf{x})$ and $S_{i-\text{original}}(\mathbf{x})$ form the spectrum values at point i of the approximated and original signals, respectively. As mentioned earlier, we model the overall signal as the sum of three symmetric and one asymmetric distributions, i.e. Gaussian-based, Lorentzians and Extreme-value curves, which have been tested in the representation of spectroscopic peaks [13]. These distributions, referred to as d1-d4, correspond to the absorbance bands at 2865-2845, 2880-2860, 2935-2915, 2970-2950 cm^{-1} , respectively. Thus, the modeled spectrum is defined in the following form:

$$S_{estimated}(z) = \underbrace{\frac{a_1}{1 + \left(\frac{z - \mu_1}{\sigma_1}\right)^2}}_{\text{Lorentzian function}} + \underbrace{\frac{a_2}{1 + \left(\frac{z - \mu_2}{\sigma_2}\right)^2}}_{\text{Lorentzian function}} + \underbrace{\alpha_3 \exp\left(-\exp\left(-\frac{z - \mu_3}{\sigma_3}\right) - \frac{z - \mu_3}{\sigma_3} + 1\right)}_{\text{Extreme Value function}} + \underbrace{\frac{a_4}{1 + \left(\frac{z - \mu_4}{\sigma_4}\right)^2}}_{\text{Lorentzian function}} \quad (6)$$

where z is the vector of the wavelengths (cm^{-1}) values, $i = 1, 2, 3, 4$ defines the number of the approached distribution, α_i , μ_i , σ_i are the amplitude, the mean value and the standard deviation of the mathematical distribution d_i , respectively, representing the height (α_i), the position parameter (μ_i) and the scale parameter (σ_i) of each curve. A characteristic estimated spectrum extracted as the sum of four distributions is depicted in Figure 1. The parameters vector \mathbf{x} over which the objective function (MSE) will be minimized is constructed by the 12 shape characteristics (3 shape parameters for each of the four distributions) [α_1 , μ_1 , σ_1 , α_2 , μ_2 , σ_2 , α_3 , μ_3 , σ_3 , α_4 , μ_4 , σ_4]. In the proposed modeling, there are no linear constraints on the variables except that they are strictly positive (relaxed constraint in order to obtain chemically valid results as negative spectrum values have no real meaning) and the centers (μ_i) of the peaks have a distance tolerance of $\pm 10 \text{ cm}^{-1}$ from the corresponding starting values. The key idea of this optimization scheme is to guide the minimization procedure through a carefully selected initial state and then “relax” the constraint values. In this way, the algorithm selects the optimal solution from a large variety of possible directions confirming with the assumptions. The non-linear limitations are described in detail in “Results and Discussion”.

3. MATERIALS AND METHODS

The FT-IR spectra of seventy petroleum samples were acquired over the $4000\text{--}650 \text{ cm}^{-1}$ range. The sample set covers practically all the commercially significant intermediate and final products in a modern petroleum refinery, from light naphthas to solid asphalt. Briefly, the sample set contains 19 naphthas, 10 vacuum gas-oils, 10 residual fractions, 7 fuel oils, 3 atmospheric gas-oils, 4 kerosines, 5 diesel fuels, 2 light solvent mixtures, 2 white spirits, 2 gasolines and one sample of wax,

alkylate, reformat, isomate, light cycle oil, and asphalt fractions, respectively. The samples were collected from three refineries, which process crude oils of different origin. Additionally, a sample set consisting of thirteen pure hydrocarbons (analytical grade purity solvents) was also used. Namely eight normal alkanes (C_6 , C_7 , C_{12} , C_{13} , C_{14} , C_{15} , C_{16} , C_{17}), four alkyl-benzenes (butyl, hexyl, oktyl, decyl) and cyclohexane were analyzed by FT-IR.

The spectroscopic analysis was carried out on a Perkin-Elmer Spectrum 1000 FT-IR with a deuterated triglycine sulfate (DTGS) detector. Liquid samples were introduced using a horizontal attenuated total reflectance HATR (PIKE Technologies) cell with a zinc-selenide (ZnSe) crystal. For the solid samples (wax, asphalt), a thin film of the substance was formed on the IR cell crystal by introducing the sample dissolved in n-hexane or chloroform, respectively, with subsequent evaporation of the solvent under a nitrogen stream. The spectra were acquired in absorbance mode as 20 co-added scans within the range $4000\text{--}650 \text{ cm}^{-1}$ at a resolution of 2 cm^{-1} . All the obtained spectra were digitized with a step of 2 cm^{-1} . The $3150\text{--}2750 \text{ cm}^{-1}$ band, after baseline subtraction (straight line), was used in the subsequent calculations, which were developed in Matlab R2010a environment. Characteristic FT-IR spectra of the samples under study are presented in Figure 2 (naphtha straight-run, diesel, distillate vacuum, residue atm).

4. RESULTS AND DISCUSSION

As described above, the aim of this work is to investigate the ability to fit a suite of distributions (statistically defined peaks) to the FT-IR spectrum of petroleum samples in the region of $3000\text{--}2800 \text{ cm}^{-1}$, in such a way that they quantitatively reflect the presence of methyl and

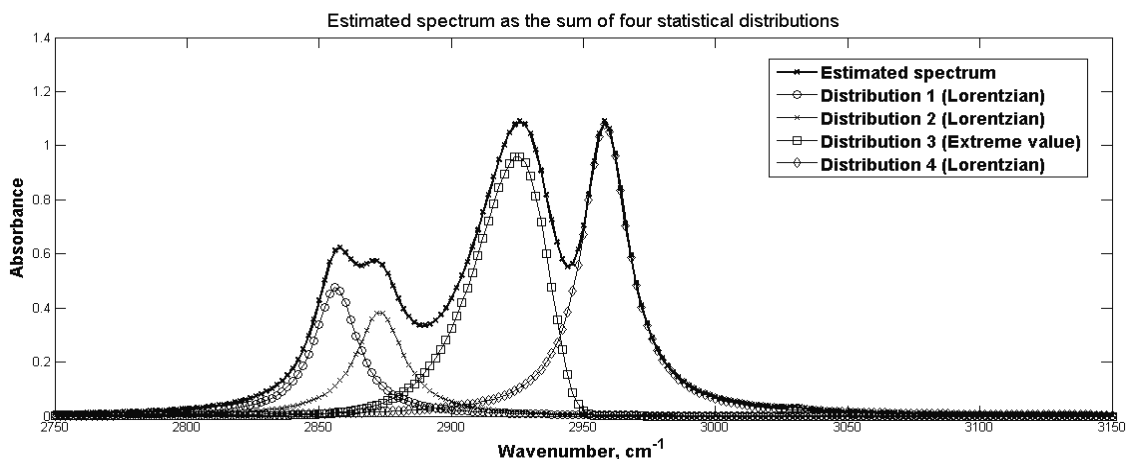


Figure 1. Estimated spectrum as the sum of the four proposed statistical distributions (Three Lorentzian distributions and one Extreme Value).

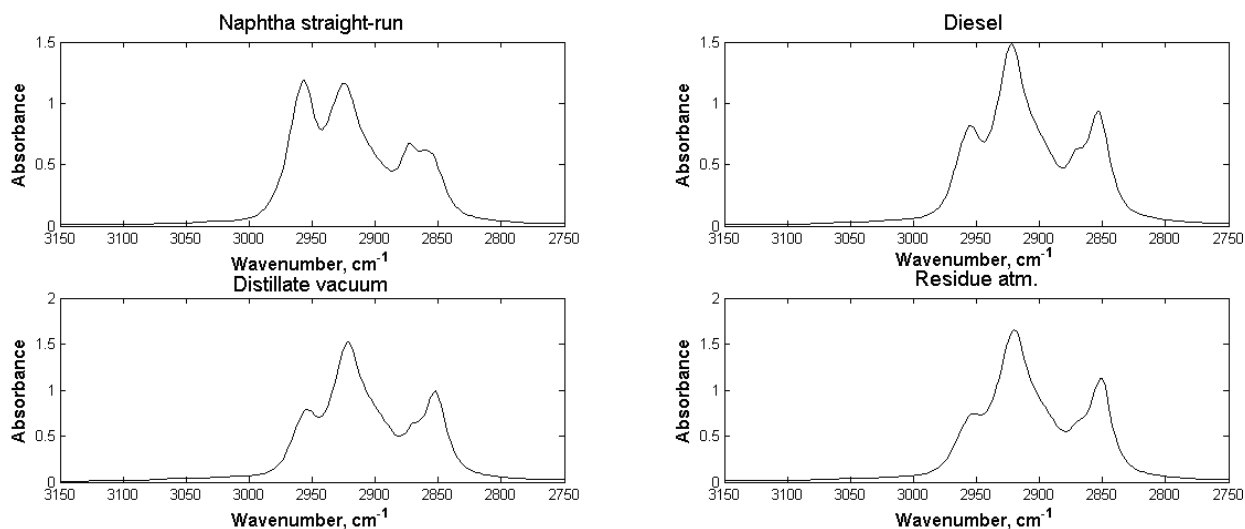


Figure 2. FT-IR spectra of characteristic petroleum fractions (naphtha straight-run, diesel, distillate vacuum, residue atm.).

methylene functional groups. In this region, four main absorbance bands exist at 2865–2845, 2880–2860, 2935–2915 and 2970–2950 cm^{-1} . The pair of the peaks 2865–2845 and 2935–2915 cm^{-1} correspond to symmetric and asymmetric vibrations of the methylene groups while 2880–2860 and 2970–2950 cm^{-1} correspond to the respective vibrations of methyl groups present in the sample. The absorbance of these groups (and the areas of the fitted peaks respectively) in each pair are formed actually from the same population of functional groups and therefore their ratio should be constant,

assuming that the corresponding peak shapes have been correctly determined. Moreover, this ratio should be equal to unity if the areas could be weighted by their specific molar absorptivity coefficient. These absorptivity values can be determined based on the experimental FT-IR signals of pure hydrocarbons. For these model compounds, the number of methyl and methylene groups is known and, therefore, a molar absorptivity (A_i) value for each group can be calculated, taking into account the corresponding densities and molecular weights of the samples. Assuming that the penetration

depth in the ATR cell is of the same magnitude for all the analyzed samples, A_i can be calculated following the equation (7):

$$A_i = \frac{S_i * MW}{d * V * n} \quad (7)$$

where: i – the index of each fitted peak ($i = 1, 2, 3, 4$)
 S_i - area of the peak i

MW - molecular weight of the model compound

d - density of the model compound

n – number of methyl or methylene groups present in the compound's molecule

V – the sample volume penetrated in the ATR cell

Although the volume V is not known, it is assumed to be constant for all analyzed samples and therefore can be omitted in the calculations. Thus, equation (7) results in a pseudo-molar absorptivity value $A'_i = A_i * V$, which is obviously instrument-specific. This fact does not limit the applicability of the proposed algorithm in a different spectrometer, since the A'_i coefficients may be readily recomputed.

The deconvolution methodology was developed following three subsequent steps:

- **Step 1:** selection of the appropriate mathematical distributions to fit each absorbance band
- **Step 2:** calculation of the molar absorptivity coefficients A'_i based on model compounds spectra
- **Step 3:** implementation of the deconvolution procedure through the application of the proposed non-linear constrained optimization algorithm to the IR spectra of petroleum fractions.

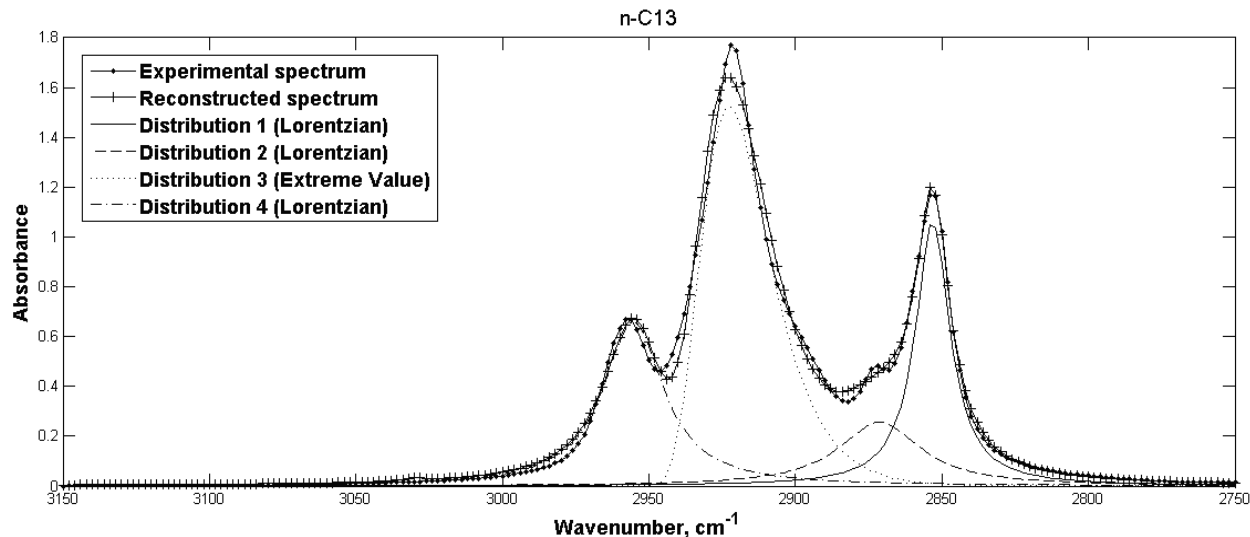
Initially, the entire data set consisting of the IR spectra of the analyzed pure hydrocarbons, as well as those of petroleum fractions, was treated using the above described deconvolution algorithm. The basic set of distributions was selected through a common set of mathematical functions used in chromatography [13] and consists of Gaussian, Lorentzian and Extreme Value distribution functions, as this combination produced the best quality metrics value (Mean Squared Error) in the proposed modeling. At the first stage, the algorithm is implemented constrained only by the fitted peak positions and allowing a shift of the peak maxima $\pm 10 \text{ cm}^{-1}$ from the anticipated wavenumber. As starting points for the $[\alpha 1, \mu 1, \sigma 1, \alpha 2, \mu 2, \sigma 2, \alpha 3,$

$\mu 3, \sigma 3, \alpha 4, \mu 4, \sigma 4]$ the specific values [0.7, 2855, 5, 0.5, 2870, 5, 1.2, 2925, 5, 0.75, 2960, 5] are used. It is concluded that three Lorentzian curves for peaks d1, d2, and d4 and one Extreme Value distribution for peak d3 provided the best reconstruction of the spectra as indicated by the mean squared error (MSE). The selection of the curve d3 shape (asymmetric methylene stretching) was additionally motivated by previous molecular modelling studies [6], where it was shown that methylenes in the methyl terminated aliphatic chains (MTAC) overlap significantly with methylene groups associated with naphthenic rings or with aromatic units (MBA). Since, in general there is no prior information on the extent of this overlap for the tested petroleum fractions, the Extreme Value (EV) distribution curve is justified for exhibiting asymmetry towards smaller wavenumbers.

At the second step, using the above selected distributions the areas of the four methyl and methylene absorption bands of the pure hydrocarbons were determined and the respective A'_i coefficients were calculated through a least-squares minimization procedure. These values were inserted in the deconvolution model and new areas of the methyl and methylene peaks were determined. At this stage, a non-linear constraint was introduced in the model, based on the assumption that the ratio of the weighted areas of distributions d1 and d3, corresponding to the absorbance of methyl groups, should differ no more than 20% from each other. The same should hold for the weighted areas of distributions d2 and d4, corresponding to the absorbances of methylene groups. The weighted area of each distribution was calculated by dividing the area below the curve with the corresponding absorptivity coefficient A'_i ($i = 1, 2, 3, 4$). The algorithm was applied in an iterative way on the model compounds data until no significant changes were observed in both the MSE of the reconstructed signals as well as in the values of A'_i coefficients. The finally computed peak areas for the methyl and methylene groups were found to provide ratios for the distributions d1/d3 and d2/d4 of mean value 1.02 and 1.07 with a standard deviation of 0.07 and 0.13, respectively. The determined A'_i coefficients and the model parameters are shown in Table 1.

Table 1. Algorithmic setup.

| Algorithm parameters and calibration | | | | | | | | | | | |
|--|--|------|------|------|--|-------------------|------|------|------|--|--|
| Objective function $f(x)$ | MSE(x) | | | | | | | | | | |
| Variables vector x | $[\alpha_1, \mu_1, \sigma_1, \alpha_2, \mu_2, \sigma_2, \alpha_3, \mu_3, \sigma_3, \alpha_4, \mu_4, \sigma_4]$ | | | | | | | | | | |
| Initial state x_0 | $[0.7 \ 2855 \ 5 \ 0.5 \ 2870 \ 5 \ 1.2 \ 2925 \ 5 \ 0.75 \ 2960 \ 5]$ | | | | | | | | | | |
| Shape parameters | Lower bound | | | | | Upper bound | | | | | |
| $\alpha_i, i = 1, 2, 3, 4$ | 0 | | | | | Infinite | | | | | |
| $\mu_i, i = 1, 2, 3, 4$ | 2845 | 2860 | 2915 | 2950 | | 2965 | 2880 | 2935 | 2970 | | |
| $\sigma_i, i = 1, 2, 3, 4$ | 0.01 | | | | | Infinite | | | | | |
| Linear constraints $h(x)$ | None | | | | | | | | | | |
| Absorptivity coefficients A_i' | 515 1200 1128 2763 | | | | | | | | | | |
| Nonlinear constraints $g(x)$ | $(area_d1/A'_1)/(area_d3/A'_3) - 1.2$ | | | | | $g_1(x) \leq 1.2$ | | | | | |
| | $0.8-(area_d1/A'_1)/(area_d3/A'_3)$ | | | | | $0.8 \leq g_2(x)$ | | | | | |
| | $(area_d2/A'_2)/(area_d4/A'_4) - 1.2$ | | | | | $g_3(x) \leq 1.2$ | | | | | |
| | $0.8-(area_d2/A'_2)/(area_d4/A'_4)$ | | | | | $0.8 \leq g_4(x)$ | | | | | |
| X vector tolerance (difference between successive points) | 0.000001 | | | | | | | | | | |
| Objective function tolerance | 0.00001 | | | | | | | | | | |
| Constraints tolerance | 0.001 | | | | | | | | | | |
| Maximum number of iterations | 1000000000 | | | | | | | | | | |

**Figure 3.** The experimental and the reconstructed spectrum of the model compound used (n-tridecane (n-C₁₃)) for the determination of coefficients A'_i .

As it is demonstrated in this table, each parameter was updated by a tolerance of $\pm 10^{-6}$ from its previous state and the constraint reached a tolerance of 10^{-3} from the target value. The calculation of

each local minimum was stopped if its absolute difference from the corresponding value of the previous iteration was less than 10^{-5} (termination criterion); the maximum number of iterations was

selected as 10^9 . In Figure 3, the experimental and the reconstructed spectrum of n-tridecane ($n\text{-C}_{13}$) obtained through the deconvolution algorithm are depicted.

Finally, at the third step the deconvolution model was applied to the IR spectra of the petroleum fractions, attempting to approximate the original spectra with the curve originated as the sum of the four statistical distributions. The original and estimated spectra of a representative sample (VGO light) are depicted in Figure 4. It was found

that in all cases the experimental signals were efficiently reconstructed, as demonstrated by the small values of the MSE in Figure 5. Furthermore, the centers of the determined four peaks fitted under the experimental spectrum exhibit consistent localization in wavenumber, as indicated by the tight, non-overlapping bounds of the corresponding Box plots in Figure 6. In addition, the algorithm constraints were satisfied in every case.

The above considerations demonstrate that the algorithm fits the selected peaks at consistent

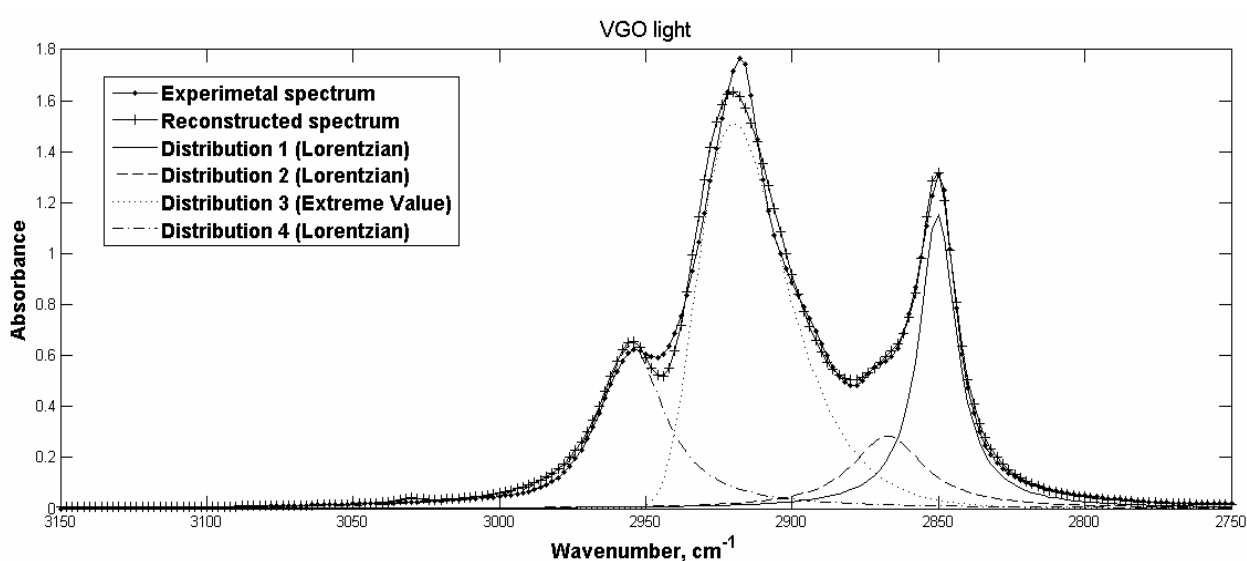


Figure 4. Original and estimated spectrum of a representative sample.

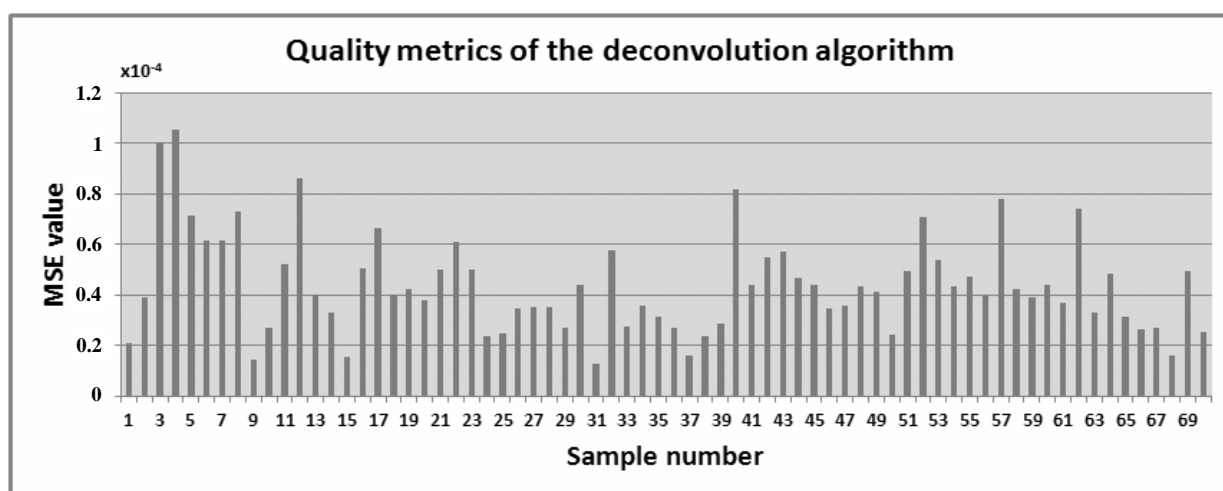


Figure 5. Quality metrics (MSE) of the optimization results.

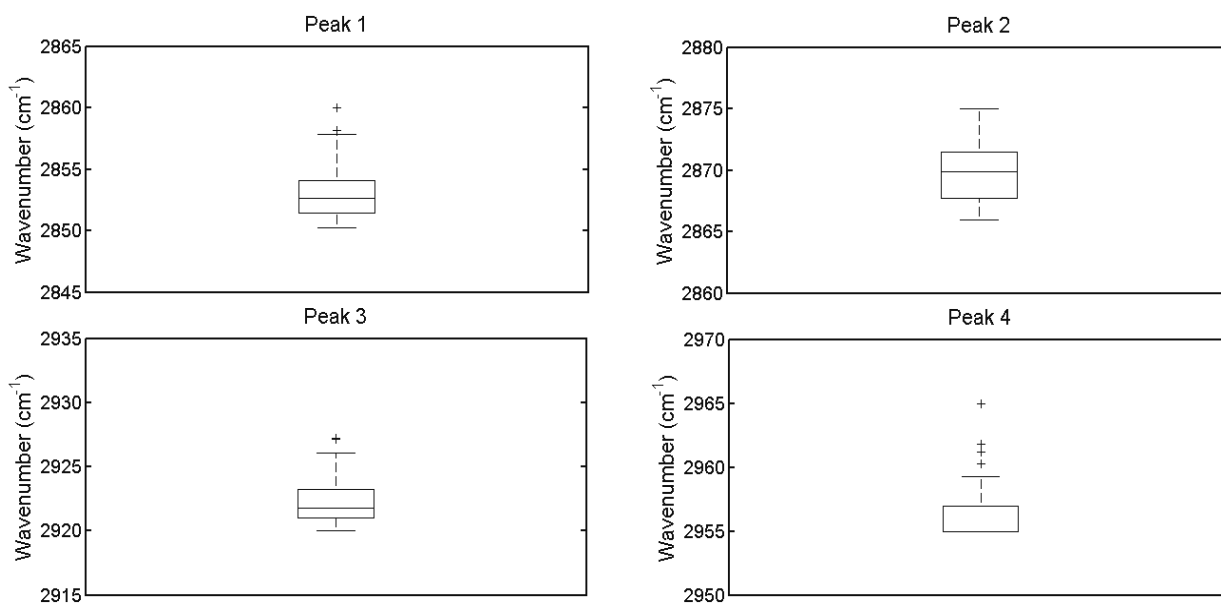


Figure 6. Box plot of the centers of the estimated distributions.

positions, keeping at the same time constant the ratio between the areas corresponding to the methyl and methylene groups of petroleum samples. These are strong indicators that the selected distributions reflect accurately the existence of methyl and methylene groups in petroleum samples. Furthermore, the calculated areas of these distributions are a sensitive quantitative characteristic, which will improve the accuracy of empirical structural models, especially for heavy hydrocarbon mixtures. The developed methodology presented in this work may be further improved by including additional bands of the FT-IR spectrum corresponding to methyl/methylene groups or other functional groups of interest for petroleum samples.

5. CONCLUSIONS

A new method was developed, enabling the reliable spectrum deconvolution and identification of the contribution of methyl and methylene groups in the 3000-2800 range of IR spectrum of petroleum fractions. The novelty of the method lies on the fact that the convergence of the deconvolution algorithm is guided by an additional structural constraint, implying that the peak areas of the symmetric and asymmetric vibrations of both methyl and methylene groups

should be equal, weighted by their specific molar absorptivity, since they rise from the same population in each sample. The molar absorptivities of these groups are experimentally determined from the IR spectra of pure hydrocarbons (alkanes, cycloalkanes, alkyl aromatics), for which the quantities of the methyl and methylene groups are known. The particular advantage of the proposed modeling scheme is that it enables the accurate reconstruction of the experimental signal, while preserving equal areas under the spectrum for the symmetric and asymmetric vibrations of methyl and methylene groups weighted by their specific molar absorptivities. The robustness of the developed deconvolution scheme was verified by examining the identified methyl and methylene peak areas in an extensive set of petroleum samples. The method will be useful in compositional calculations of petroleum fractions, where the use of peak areas instead of less accurate peak heights will improve their accuracy.

REFERENCES

1. Kidenas, K., Murata, S. and Nomura, M. 1996, *Energy & Fuels*, 10, 672.
2. Artok, L., Su, Y., Hirose, Y., Hosokawa, M., Murata, S. and Nomura, M. 1999, *Energy & Fuels*, 13, 287.

3. Brown, J. K. and Ladner, W. R. 1960, Fuel, 39, 87.
4. A-Liang, W., Que, G., Chen, Y. and Liy, C. 2000, Chemical Composition and Characteristics of Residues of Chinese Crude Oils, Chapter 10, Development in Petroleum Science, T. F. Yen and G. V. Chilingarian, (Eds.), Elsevier, New York, 40B 2, 287.
5. Siddiqui, M. N. 2008, Petroleum Science and Technology, 26, 2048.
6. Coelho, R. R., Hovell, I., De Mello-Monte, M. B., Middea, A. and de Souza, A. L. 2006, Fuel Processing Technology, 87, 325.
7. Ganz, H. and Kalkreuth, W. 1987, Fuel, 66, 708.
8. Coelho, R. R., Hovell, I. and Raja-gopal, K. 2012, Fuel Processing Technology, 97, 85.
9. Stuart, B. H. 2004, Infrared Spectroscopy Fundamentals and Applications, John Wiley & Sons, Ltd., West Sussex, 54.
10. Robinson, S. M. 1974, Math. Programming, 7, 1.
11. Powell, M. J. D. 1978, Numerical Analysis, G. A. Watson (Ed.), Lecture Notes in Mathematics, Springer Verlag, Vol. 630, 144.
12. Boggs, P. T. and Tolle, J. W. 2000, Journal of Computational and Applied Mathematics, 124, 123.
13. Di Marco, V. B. and Bombi, G. 2001, Journal of Chromatography A, 931, 1.