Oil Shale, 2022, Vol. 39, No. 1, pp. 20–36
doi: https://doi.org/10.3176/oil.2022.1.02
ISSN 0208-189X

Multivariate models based on infrared spectra as a substitute for oil property correlations to predict thermodynamic properties: evaluated on the basis of the narrow-boiling fractions of Kukersite retort oil

Zachariah S. Baird, Vahur Oja*

Received 4 May 2021, accepted 25 January 2022, available online 10 March 2022

Department of Energy Technology, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia

Abstract. *This article investigates a potential for using models based on infrared spectra to predict basic thermodynamic properties of narrow boiling range oil fractions or pseudocomponents. The work took advantage of the simultaneous availability of a property database of narrow boiling range fractions of Kukersite oil shale retort oil (from the industrial retorting process) together with infrared spectra of these fractions. The work was based on the hypothesis that the models based on infrared spectra could potentially be used to reduce experimental data when developing other predictive methods, or even as a substitute for other prediction methods. In this study four basic oil properties, which are often used to predict other thermodynamic properties, were predicted from infrared spectra using support vector regression. These were specific gravity, refractive index parameter, average boiling point and molecular weight. According to bulk property prediction approach these selected properties can be grouped into energy parameters (two former) and size parameters (two latter). It was found that, for distillation fractions with varying compositions, both the energy parameters (specific gravity, refractive index) as well as the size parameters (molecular weight, average boiling point) can be predicted from Fourier transform infrared (FTIR) spectra, and that the accuracy of the predictions based on infrared spectra was comparable with the accuracies of petroleum bulk property correlations. Thus, infrared spectra can provide a convenient alternative in the thermodynamic property prediction field because they can be easily measured and correlated to a wide variety of properties.*

Keywords: *property prediction, FTIR spectroscopy, oil, liquid fuel, shale oil, chemometrics.*

* Corresponding author: e-mail *vahur.keemteh@outlook.com*

1. Introduction

The thermodynamic and physical properties of oils are required so that chemical-physical processes can be designed or evaluated, both in terms of the plant and in the environment [1–3]. For pure compounds and simple mixtures, for which the complete composition can be known, the properties for each compound can be specifically designated. However, for complex undefined mixtures of unknown composition, such as oils from various resources like petroleum, biomass, coal, or oil shale, simplifications are required to various degrees. For these materials empirical correlations which are based on bulk properties are historically applied [4–6]. These correlations are usually based on commonly measured characteristic properties of the oil's narrow boiling range fractions (distillation cuts), which are referred to as pseudocomponents, such as their specific gravity, viscosity, molecular weight, or average boiling points from distillation curves. The specific property of a pseudocomponent can be predicted, depending upon system complexity, from one or more other properties using suitable regression equations. For conventional oil cuts with average boiling points below 350 °C regression equations with following general forms are often proposed [4]:

$$\theta = a{\theta_1}^e{\theta_2}^f, \tag{1}$$

$$\theta = aexp(b\theta_1 + c\theta_2 + d\theta_1\theta_2){\theta_1}^e{\theta_2}^f. \tag{2}$$

Equations (1) and (2) contain the property θ that is to be predicted, with θ_1 and θ_2 being the two input parameters (or properties from which the property θ is to be predicted), and a to f are empirically derived regression constants. For heavier or more polar substances these two parameter equations may not be suitable [4]. Therefore those correlations which supply accurate results for a range of oils (for oil cuts with boiling points of up to 350 °C), and which are used in process simulators, are usually based on at least two input parameters: preferably one describing molecular size (such as carbon number, molecular weight, average boiling point), and the other describing molecular energy (such as specific gravity, refractive index, hydrogen-carbon ratio) [4]. While the deviation of properties into molecular size and molecular energy parameters is quite tentative, the approach still serves to emphasise the fact that molecules which are of a similar size (described, for example, by properties such as carbon number, molecular weight, or average boiling point) can involve components of various structure classes (such as, for example, in the case of conventional petroleum, a grouping into paraffins, naphthenes, or aromatics is often used). Therefore a variation exists in terms of property values. To be able to develop these bulk property correlations a large amount of experimental data are needed; however, experimental measurements are often time consuming and expensive.

Our own laboratory's work with Kukersite oil shale retort oil was what initially led us to investigate infrared-based prediction methods. The Kukersite oil shale retort oil is a synthetic crude oil which is produced from Estonian Kukersite oil shale by pyrolysis or retorting [7, 8]. As with many alternative liquid fuel sources, shale oils manufactured from different sources have compositions that are more or less different from those of most conventional petroleum crudes [9]. As an example, the shale oil that is manufactured via retorting from Kukersite oil shale has a high content of oxygen-containing compounds, with the largest portion being phenolic compounds [10, 11]. For this reason physical/thermodynamic property correlations, which have been developed based on petroleum fuels, may give worse results for Kukersite shale oil than would be required in applications.

In the process of finding approaches that provide the desired accuracy, we began investigating the potential for using correlations that are based on infrared spectra to support the development of bulk property prediction methods for the thermodynamic and/or physicochemical properties of oil cuts. The initial practical idea was to use the Fourier transform infrared (FTIR) method to measure and/or predict structural characteristics (especially the amount/concentration of phenolic OH groups [11]); however, it was later seen as a convenient tool for detecting random experimental measurement errors (identifying outliers) for all measured properties or to help reduce the amount of experimental data that would otherwise be needed to develop predictive bulk property correlations. The current paper is the third on this topic to have been issued from our laboratory. Application options in regard to the FTIR-based multilinear regression approach in order to be able to determine structure characteristics (such as hydroxyl concentrations in narrow boiling shale oil cuts [11]), and to determine temperature-dependent properties with linear temperature dependence (such as the density temperature dependence of narrow boiling shale oil cuts [12]), have been presented in earlier articles that have been published by this laboratory. Although the use of FTIR together with multilinear regression is a common tool for the property evaluation of various materials [13–15], using this approach as a thermodynamic property prediction tool – the current area of interest – has never previously been emphasised to our knowledge. The most likely reason is the unavailability of a suitable database which simultaneously involves quantitative information on thermodynamic properties and FTIR spectra for oil distillation cuts (narrow boiling range fractions). There is also an additional restriction that could have reduced interest in the wider scientific community when it came to predicting thermodynamic properties that can be obtained from FTIR spectra. This means that some form of standardisation or calibration transfer is needed to be able to use correlations on another spectrometer and, therefore, permit them to be used by other teams.

In this paper we take experimental property data for over two hundred Kukersite shale oil fractions, together with their FTIR spectra, and investigate

the use of FTIR-based models to predict the basic temperature-independent thermodynamic properties, with an emphasis on predicting so-called 'size parameters'. In this article we focus on four basic properties that are commonly used in characterising oils from the point of view of thermodynamic property prediction: the specific gravity, the refractive index parameter (which is calculable from the refractive index [16]), the average boiling point, and the average molecular weight. Although the specific gravity and the refractive index are temperature-dependent properties, they are often measured at a single standard temperature and are used as a characteristic parameter. In this sense, these properties at a specified temperature are temperature-independent properties. As infrared spectra contain information about the molecular structure of the sample and do not directly contain information about the size of the molecules in the sample, the current work was driven by our initial interest to evaluate whether at all, or how well, FTIR-based models can predict so-called 'molecular size' parameters such as molecular weight and average boiling point. Application for FTIR-based models when it comes to density and refractive index as properties of fuels (here grouped into energy parameters) can be found in the available literature [12].

2. Experimental methods

2.1. Sample preparation

The oil shale retort oils used for this study were obtained from Eesti Energia's Narva Oil Plant (Narva, Estonia). This plant uses the solid heat carrier retorting method (called the Galoter process) [17, 18]. Some additional information on the Kukersite oil shale, the processes occurring during pyrolysis and characteristics of the resulting oil can be found from the literature [10, 11, 19–25]. At the plant oil is separated into wide technical fractions as a product (currently typically into shale gasoline, fuel oil, and heavy oil). Mainly gasoline and fuel oil samples (technical fractions) from the plant were used for this study. The wide technical fractions from the plant were further separated into narrow boiling fractions via distillation, either by simple distillation or rectification, at our laboratory. However, most distillations were simple batch distillations that were carried out either at atmospheric pressure (using an Engler distillation [26]), and/or in a vacuum. Additional information on experimental settings and procedures can be found in: rectification [22, 27] and simple distillation [23, 28] of gasoline; rectification [27, 29] and simple distillation [28] of fuel oil.

To increase diversity, wide technical fractions were obtained from different plants that use different oil shale processing regimes and were taken at multiple times over the course of three years. Additionally, to map/screen trends, some fuel oil technical fractions (or their distillation cuts) were artificially adjusted via

extraction and/or mixing. For this purpose, the samples were separated into phenolic and neutral fractions using extraction with a 10% NaOH solution [10, 11]. In this manner additional fuel oil samples were created that had lower and higher contents of phenolic compounds than the original samples themselves (with hydroxyl contents ranging from about zero to 10 wt% OH).

The number of samples, mostly narrow boiling range fractions (or cuts) that were used in this FTIR-based models study, amounted to 355 for specific gravity, 327 for refractive index parameter, 229 for average boiling point, and 277 for number average molecular weight. It should be noted that although the property data that was used in this study covers a wide range of property values (such as boiling points between 350–670 K or 80–400 °C; a refractive index parameter between 0.34–0.45; specific gravity between 0.7–1.10; and a molecular weight between 70–450 g/mol), not all data would be reliable for the development bulk property correlations of desired accuracy. On one side, this is due to the observation that, during sample preparation to the narrow boiling range cuts, in the case of higher boiling fractions the applied temperature-time history of distillation could have resulted in a thermal decomposition-based chemical alteration (i.e. resulting in systematic anomalies). In addition, on the other side, artificially adjusted samples may not be the best choice for developing reliable bulk property correlations as the artificial adjustment of the oil's nature could have resulted in unreliable changes taking place (i.e. causing some systematic anomalies). However, in order to increase diversity, we have included in database for the development of FTIR-based models the properties of the aforementioned fractions of somewhat questionable representative quality. Although not all of the data can be used for developing or evaluating bulk property correlations, they are still valuable for the purposes of this study – to evaluate the potential of applying FTIR-based methods.

2.2. Property measurements

The methods and devices that have been used for measuring the properties (density, refractive index, average boiling point, average molecular weight), together with estimated standard uncertainties for the purpose of this study, are summarised in Table 1 and given in more detail below.

2.2.1. Density

The density at 20 °C was measured using an oscillating tube density meter (DMA 5000 M, Anton Paar GmbH) equipped with a heating attachment that heats the sample at the unit's inlet to lower the viscosity. The performance of the device was checked using distilled water and air. Based on repeat measurements of selected narrow boiling range fractions the standard uncertainty was estimated to be roughly 0.00015 g/cm^3. The uncertainties for the heavy samples may be slightly greater than those of the lighter fractions. For heavier samples, due to their higher viscosity, several densities were

Table 1. Methods used for measuring the properties of shale oil samples

Property	Measurement method (device)	Estimated standard uncertainty
Density	Oscillating tube density meter (Anton Paar DMA 5000 M)	0.00015 g/cm^3
Refractive index	Refractometer (Anton Paar Abbemat HT)	0.0011
Average boiling point	Thermogravimetric analyser based method, the method developed in-house (Du Pont 951)	2.1 °C
Average molecular weight	Cryoscopy, ASTM D2224 standard (device built in house) Vapor pressure osmometer (Osmomat 070 or Knauer K-7000)	7 g/mol

measured at higher temperatures, and a density at 20 °C was then calculated from the linear temperature dependence [12].

2.2.2. Refractive index parameter

The refractive index at 20 °C was measured at 589.592 nm using an Abbemat HT refractometer (Anton Paar GmbH). Performance was checked before and after each set of measurements and was carried out using distilled water. From repeat measurements of selected narrow boiling range fractions the standard uncertainty was estimated to be 0.0011 (with an expanded uncertainty of 0.0021 at a level of 95%). The refractive index parameter was calculated from the refractive index at 20 °C using the equation given by Huang [16]:

$$I = \frac{n^2-1}{n^2+1}, \tag{3}$$

where I is the refractive index parameter and n is the refractive index at 20 °C.

2.2.3. Average boiling point

The average boiling points for the samples were measured by means of a thermogravimetric analyser based method [27–29]. The accuracy of this method was evaluated using measured oil narrow boiling range fractions that had been obtained by Rannaveski et al. [27] according to the ASTM D2892 standard. Based on these fractions, the standard method uncertainty of 2.1 °C (with an expanded measurement uncertainty of 4.3 °C at 95%) is used here [27].

2.2.4. Molecular weight

The average molecular weight was measured using mainly two different methods: cryoscopy (built in-house, as in the ASTM D2224 standard) and vapour pressure osmometry (Osmomat 070, Gonotec GmbH or later in the project Knauer K-7000, Knauer GmbH). Benzene was used as a solvent for the cryoscopy and was also used mainly in osmometric measurements. For both methods calibration was carried out using solutions of benzyl with known concentrations. Standard uncertainties were calculated based both upon the accuracy of the calibration and tests with pure compounds. The relative expanded uncertainty (at the 95% level) was determined to be between ±6 and ±7%, for a single method (device). As for fractions, the uncertainty was smaller for fractions with lower molecular weights and larger for heavier fractions. When taken on average, the absolute standard uncertainty of 7 g/mol (an absolute expanded uncertainty of 14 g/mol at the level of 95%) can be used here.

2.3. Infrared spectral measurements

Infrared spectra were measured using a Fourier transform infrared spectrometer that was fitted with an attenuated total reflection (ATR) measurement accessory. A single reflection ZnSe crystal was used. The spectrometer was an Interspec 301-X portable mid-infrared spectrometer (Interspectrum OÜ). The spectra were measured over the range of 700 to 4000 cm^{-1} at a resolution of 1 cm^{-1}. A cosine apodisation was used ($\cos(0.5\cdot\pi\cdot x)\cdot(\cos(0.5\cdot\pi\cdot x))^2$). Ten scans were taken and averaged together to produce the spectrum. Baseline correction was carried out by fitting a third order polynomial to regions in which shale oil does not absorb (2000–2200 and 3700–4000 cm^{-1}).

2.4. Multivariate regression

Regression was carried out using support vector regression, which was implemented in Python (version 2.7) using the Scikit-learn package (version 0.15) [30]. A mixed kernel was used, which combined the polynomial and radial basis function kernels using a single weighting parameter [31]. The regression parameters were optimised by minimising the five-fold cross validation error using the SciPy differential evolution solver [32, 33].

To make spectra more comparable to those from different instruments, with the hope of creating models that could be used on a wider range of instruments, the spectra were pre-processed. First, the spectra were transformed to remove the wavelength dependence that is inherent in ATR spectra, and therefore to make them more like a transmission spectrum. This was done using the algorithm that was presented by Bertie et al. [34] and Bertie and Lan [35]. Then, because the first half of the spectra contained most of the chemical information, the region above 1800 cm^{-1} was removed. After this the standard

normal variate transformation was applied to the spectra, in which each spectral data point is offset by the mean of the spectrum and is then divided by the standard deviation of the variation in absorbance values. Regression was carried out using these pre-processed spectra.

The residuals of the models were used to detect outliers. If it was possible, then infrared spectra and the properties of outlying samples were re-measured, and this allowed some measurement errors to be identified and corrected. Some samples could not be re-measured due to the limited volume of the sample, so errors in the experimental data could also account for some of the outliers that were observed.

2.5. Error statistics

For FTIR models, four error statistics were calculated from the cross validation values in order to evaluate performance of models: root mean squared error (RMSE), average absolute deviation (AAD), average absolute relative deviation (%AAD), and the Pearson correlation coefficient squared (R^2):

$$RMSE = \sqrt{\frac{\Sigma(\theta_{pred}-\theta_{actual})^2}{n}}, \tag{4}$$

$$AAD = \frac{\Sigma|\theta_{pred}-\theta_{actual}|}{n}, \tag{5}$$

$$\%AAD = \frac{\Sigma\frac{|\theta_{pred}-\theta_{actual}|}{\theta_{actual}}\cdot 100}{n}, \tag{6}$$

$$R^2 = \left(\frac{\Sigma(\theta_{actual}-\theta'_{actual})(\theta_{pred}-\theta'_{pred})}{\sqrt{\Sigma(\theta_{actual}-\theta'_{actual})^2\,\Sigma(\theta_{pred}-\theta'_{pred})^2}}\right)^2, \tag{7}$$

where, in Equations (4) to (7), the value θ_{pred} is the predicted property value (found during cross validation), θ_{actual} is the actual property value and n is the number of data points.

3. Results and discussion

The accuracy of the predictions for each of the four parameters can be visualised from Figure 1, which shows the residual (the difference between the measured value and the predicted value) for each property of each sample when it was

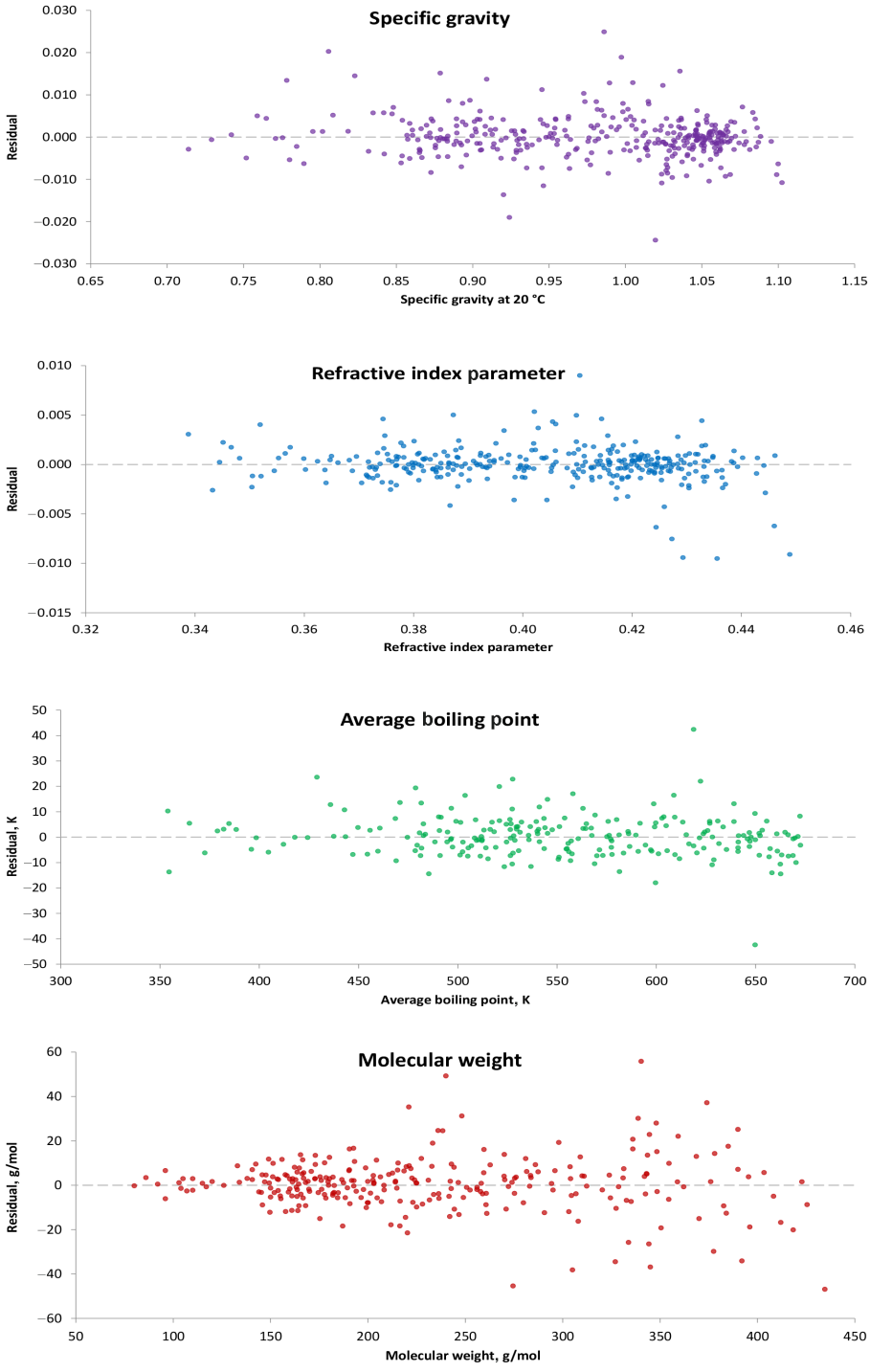

Fig. 1. Residuals of the predicted values for samples as part of the cross validation set.

part of the cross validation set. Figure 1 indicates that while there is quite a random fluctuation of residuals of the predicted values (quite a symmetrical distribution with no clear trends), the fluctuations pattern varies somewhat between those properties that were investigated. As the same infrared spectra were used for all of the property models, then it is reasonable to expect that the effect of inaccuracies in the infrared spectra likely had a similar impact on the residuals of predicted values across the different properties. Therefore the experimental property accuracy (measurement accuracy) and the strength of the correlation between spectra and the property are the most likely factors that could lead to any differences. Figure 1 also shows that several samples have quite large residuals (points that are much farther from other points). The majority of these outliers were more of the property specific type, but a minority of the outliers had large residuals across all of the four properties, which suggests that the sample preparation may have resulted in fractions with less common chemical compositions. The residuals for average boiling point provide a good example of the latter: there are two samples with boiling points of about 600 K that have residuals of more than 30 K. These two samples, and about five or six others, had consistently large residuals across all of the four properties.

More details about the models are given in Table 2, including error statistics such as root mean squared errors, average absolute deviations, and relative mean

Table 2. Errors and model parameters for each of the multivariate models created based on infrared spectra

		Specific gravity (20/20)	Refractive index parameter	Boiling point	Molar mass
RMSE		0.00506	0.00186	8.1 K	12.2 g/mol
AAD		0.00344	0.00117	5.8 K	8.24 g/mol
%AAD		0.36 %	0.29%	1.07%	3.42 %
R^2		0.9963	0.9938	0.9879	0.9769
Property range		0.714–1.102	0.339–0.449	350–670 K	80–435
Number of samples		355	327	229	277
Regression parameters	C	88.27799	0.2313152	0.5626782	30.87789
	ε	0.0001976886	0.02136228	0.007404886	0.01527662
	γ	0.2634851	0.6102149	0.9954894	0.007116409
	Degree	3	3	2	2
	Zero coefficient	2.134564	1.871514	3.419010	0.8006127
	Mixing coefficient	0.8034127	0.9270210	0.8509057	0.8523829

deviation for each property. Table 2 reveals that somewhat better predictions are obtained for the specific gravity and the refractive index parameter than for the average boiling point and the average molecular mass. This makes sense because the specific gravity and the refractive index parameter are quantitatively more closely related to the types of bonds (functional groups) in the mixture (which is the information that an infrared spectrum gives). At the same time, the measurement-related standard uncertainties in Table 1 indicate that the experimental data for these properties were more accurate than was the experimental data for average boiling point and molecular weight. The uncertainty ratios for predicted/measured values (predicted as RMSE and measured as standard uncertainty) for these four properties were as follows: 31 for specific gravity, 2.9 for refractive index parameter, 3.3 for average boiling point, and 1.7 for molecular weight. The comparison of uncertainty ratios for the predicted values and measured values, especially those of the refractive index parameter and the average boiling point, indicate that model accuracy was not only limited by the measurement accuracy of the experimental data, but was also dependent upon property, more or less, by factors that served to influence correlations between infrared spectra and properties. For example, the measurement method for density had a very low level of uncertainty, and the high ratio here suggests that accuracy was not limited by the accuracy of the experimental data, but instead by other factors.

As infrared spectra contain information about the molecular structure of the sample and do not directly contain information about the size of molecules in the sample, the current work was driven by our initial interest in evaluating how well, if at all, FTIR-based models can predict 'molecular size' parameters for narrow boiling range oil fractions (or pseudocomponents) that are prepared by distillation. As can be seen from Figure 1 and Table 2, when it comes to the distillation fractions, FTIR models can reliably predict 'molecular size parameters' (parameters that are more strongly related to the size of the molecules rather than to the types of bonds or functional groups in the mixture). Moreover, their values are quite accurately predicted. Therefore, in order to be able to accurately predict molecular size-related properties, as seen in this work, there should exist some form of indirect relation between these properties and infrared spectra. In this regard, it was observed that there are systematic changes between the infrared spectra [11, 36] in the collected series of fractions with narrow boiling ranges (i.e. occurring from the first fraction to the last fraction collected). These systematic changes that accompany changes in the boiling point of the samples are likely to be what supplies the additional information that is necessary for predicting molecular size properties.

The performance of the models from a practical point of view can also be checked by viewing the results for sequential fractions from a single simple batch distillation. For most of the fractions the difference between model and measured values is smaller than the difference between subsequent fractions. That is, the infrared models, both for 'energy parameters' and 'size parameters',

can generally distinguish between two fractions. This is illustrated in Figure 2 where specific gravity (as an energy parameter) and average boiling point (as a size parameter) are evaluated for a selected distillation. However, Figure 2 shows one additional performance-related indication between 'energy parameters' and 'size parameters' (a tendency that was generally more or less observable). In this exemplary distillation, at the point at which the distillation pressure was reduced from atmospheric pressure to low pressure, there is an inflection point in the overall trend (the drop in property values). It can be seen in Figure 2 that the average boiling point model had larger errors for these two samples at the inflection point, but the density model could better account for this anomaly. This makes sense when supporting the view that the average

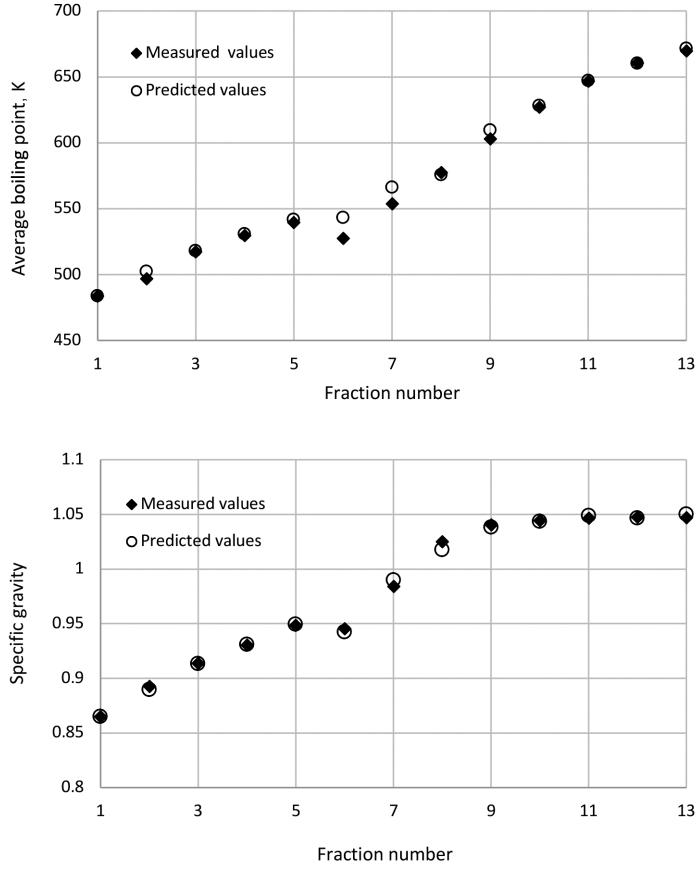

Fig. 2. Comparison of the performance of the average boiling point model and the specific gravity model for fractions from a typical distillation. For the average boiling point, the experimental error bars are data point sizes. For the specific gravity, experimental error bars are smaller than the data point.

boiling point prediction from FTIR spectra could for the most part rely on systematic changes in FTIR spectra from fraction to fraction (something which could be related to changes in the types and concentration of bonds in the sample), but specific gravity predictions from FTIR spectra could rely both on systematic changes from spectra to spectra and the structure-related information of the specific spectra. Therefore, the correlation between a property and spectra could be more direct for some properties than others, but the spectra of distillation cuts can both directly and indirectly contain significant amounts of information which will help to predict the property.

Finally, whether or not the accuracy of the FTIR-based models (brought out in Table 2) is sufficient depends upon the desired application, and for calculations with a small tolerance for error it may still be preferable to measure the value experimentally. However, based on values for AAD and %AAD, the accuracies are at the same level as those that have been stated for petroleum correlations (bulk property correlations useable in process simulators) [4]. It is to note here that the accuracy of the current FTIR study and those petroleum bulk property correlations that are currently available are not directly comparable in terms of the number of data points, the variability of samples, and the chemical nature of samples being used for correlations. Still, for indicative purposes alone, the fact can be highlighted here that, for a specific gravity, the current FTIR model had an AAD that was about half that of the bulk property correlations for conventional oils. However, at the same time, some of the petroleum bulk correlations for molecular weight had better levels of accuracy than did the current FTIR model, but these correlations were for a narrower range of molecular weights. At the same time, the molecular weight correlations that included the heavier fractions (as the regression that was based on FTIR spectra did) had significantly higher AADs. In addition, in a literature review of the use of multivariate regression for fuel property prediction [4], we found that other researchers have also created models for some of these parameters for other fuels and have achieved similar or better levels of accuracy. So it appears that in specific cases multivariate models have been shown to have levels of accuracy that are as good (or even better) than corresponding petroleum correlations that are based on physical properties. Finally, it should be noted that this article only looked at the implementation of the FTIR based models, however, data fusion from different analytical sources, such as FTIR and NMR, has shown an improvement in the ability to estimate some physico-chemical properties of crude oils when compared to a single analytical technique [37].

4. Conclusions

In this study an investigation was carried out into the potential for using predictive models based on infrared spectra to predict four basic oil parameters

(specific gravity, refractive index parameter, average boiling point, and average molecular weight) in terms of narrow boiling range distillation fractions (or pseudocomponents). It was found that, for batch distillation fractions from Kukersite oil shale oil of varying compositions, in a way that is similar to energy parameters (parameters that are more closely related to the types of bonds and functional groups, with these here being the specific gravity and the refractive index parameter) the size parameters (parameters that are more closely related to the size of the molecules, with these here being the molecular weight and the average boiling point) can be reliably predicted from FTIR spectra. Therefore the spectra can both directly and indirectly contain significant amounts of information which will help to predict the property. However, it was also seen that the models gave somewhat better results for physical parameters that were related more to the molecular structure than for those that were more closely related to the molecular size. For generalization, FTIR-based models could be a useful "tool" for both determining/predicting various thermodynamic properties and for detecting random experimental errors (identifying outliers) of the measured data during a measurement project (if the FTIR analysis is included in the project).

Although not directly comparable, the comparison with performance parameters of the more commonly used bulk property correlations (which have been developed for conventional oils) suggests that predictive models that are based on infrared spectra could be used to reduce the experimental data required when developing other predictive methods, or even to use as a substitute for other prediction methods. Although issues do exist such as, for example, over-fitting the data, the problematic transfer of correlations to another spectrometer, or the questionable application to samples that are not included in the calibration set, thermodynamic and property correlations which are based on infrared spectra would still be advantageous to current prediction methods in some situations and applications.

Acknowledgements

The authors gratefully acknowledge financial support provided by the Estonian Ministry of Education and Research, under target financing SF0140022s10, and by the Estonian National R&D program Energy under project AR10129 "Examination of the Thermodynamic Properties of Relevance to the Future of the Oil Shale Industry" (P.I. Prof. V. Oja). On the FTIR-based model side, the authors are very grateful to PhD Madis Listak from Tallinn University of Technology for technical assistance. Regarding the availability of the section of the database on physicochemical/thermodynamic properties used, the authors thank the following participants from Tallinn University of Technology for their valuable contributions to the project AR10129: PhD Alfred Elenurm, PhD Alexei Yanchilin, PhD Oliver Järvik, PhD Rivo Rannaveski, PhD Eduard

Tearo, PhD Jelena Hruljova, PhD Natalja Savest, Mrs Ilme Rohtla, Mr Einart Lindaru, MSc Sven Kamenev, MSc Pamela Puidak, MSc Ruth Rooleht and Mrs Hanna Ennomäe.

REFERENCES

1. De Hemptinne, J. C., Behar, E. Thermodynamic modelling of petroleum fluids. *Oil Gas Sci. Technol.,* 2006, **61**(3), 303–317.
2. Riazi, M. R., Al-Enezi, G. A. Modelling of the rate of oil spill disappearance from seawater for Kuwaiti crude and its products. *Chem. Eng. J.,* 1999, **73**,161–172.
3. Riazi, M. R., Daubert, T. E. Characterization parameters for petroleum fractions. *Ind. Eng. Chem. Res*., 1987, **26**(4), 755–759.
4. Riazi, M. R. *Characterization and Properties of Petroleum Fractions*. ASTM International, 2005.
5. Kollerov, D. K. *Physicochemical Properties of Oil Shale and Coal Liquids.* Moscow, 1951 (in Russian).
6. Tsonopoulos, C., Heidman, J. L., Hwang, S. C. *Thermodynamic and Transport Properties of Coal Liquids.* John Wiley & Sons, 1986.
7. Oja, V., Suuberg, E. M. Oil shale processing, chemistry and technology. In: *Encyclopedia of Sustainability Science and Technology* (Meyers, R. A., ed.), Springer, 2012, 7457−7491.
8. Lee, S. *Oil Shale Technology*. CRC Press, 1990.
9. Urov, K., Sumberg, A. Characteristics of oil shales and shale-like rocks of known deposits and outcrops: monograph. *Oil Shale*, 1999, **16**(3 SPECIAL), 1–64.
10. Kogerman, P. N. *On the Chemistry of the Estonian Oil Shale "Kukersite"*. Tartu, Estonia, Oil Shale Research Laboratory, 1931.
11. Baird, Z. S., Oja, V., Järvik, O. Distribution of hydroxyl groups in kukersite shale oil: quantitative determination using Fourier transform infrared (FT-IR) spectroscopy. *Appl. Spectrosc*., 2015, **69**(5), 555–562.
12. Baird, Z. S., Oja, V. Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density. *Chemom. Intell. Lab. Syst.*, 2016, **158**, 41–47.
13. Pavoni, B., Rado, N., Piazza, R., Frignani, S. FT-IR spectroscopy and chemometrics as a useful approach for determining chemical-physical properties of gasoline, by minimizing analytical times and sample handling. *Annali di Chimica,* 2004, **94**(7–8), 521–532.
14. Pasquini, C., Bueno, A. F. Characterization of petroleum using near-infrared spectroscopy: Quantitative modeling for the true boiling point curve and specific gravity. *Fuel*, 2007, **86**(12–13), 1927–1934.
15. Cooper, J. B., Wise, K. L., Groves, J., Welch, W. T. Determination of octane numbers and Reid vapor pressure of commercial petroleum fuels using FT-Raman spectroscopy and partial least-squares regression analysis. *Anal. Chem.,* 1995, **67**(22), 4096–4100.

16. Huang, P. K. *Characterization and Thermodynamic Correlations for Undefined Hydrocarbon Mixtures*. Ph.D. Thesis, Pennsylvania State Univ., University Park, 1977.
17. Golubev, N. Solid oil shale heat carrier technology for oil shale retorting. *Oil Shale,* 2003, **20**(3 SPECIAL), 324–332.
18. Elenurm, A., Oja, V., Tali, E., Tearo, E., Yanchilin, A. Thermal processing of dictyonema argillite and kukersite oil shale: Transformation and distribution of sulfur compounds in pilot-scale Galoter process. *Oil Shale,* 2008, **25**(3), 328−333.
19. Lille, Ü., Heinmaa, I., Pehk, T. Molecular model of Estonian kukersite kerogen evaluated by ^{13}C MAS NMR spectra. *Fuel*, 2003, **82**(7),799–804.
20. Derenne, S., Largeau, C., Casadevall, E., Sinninghe Damsté, J. S., Tegelaar, E. W., de Leeuw, J. W. Characterization of Estonian Kukersite by spectroscopy and pyrolysis: Evidence for abundant alkyl phenolic moieties in an Ordovician, marine, type II/I kerogen. *Org. Geochem*., 1990, **16**(4–6), 873–888.
21. Oja, V., Rooleht, R., Baird, S. Z. Physical and thermodynamic properties of kukersite pyrolysis shale oil: literature review. *Oil Shale*, 2016, **33**(2), 184−197.
22. Mozaffari, P., Baird, Z. S., Listak, M., Oja, V. Vapor pressures of narrow gasoline fractions of oil from industrial retorting of Kukersite oil shale. *Oil Shale,* 2020, **37**(4), 288–303.
23. Siitsman, C., Oja, V. Extension of the DSC method to measuring vapor pressures of narrow boiling range oil cuts. *Thermochim. Acta*, 2015, **622**, 31−37.
24. Oja, V. Examination of molecular weight distributions of primary pyrolysis oils from three different oil shales via direct pyrolysis Field Ionization Spectrometry. *Fuel*, 2015, **159**, 759−765.
25. Oja, V. Is it time to improve the status of oil shale science? *Oil Shale*, 2007, **24**(2), 97–99.
26. ASTM D86. *Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure*. ASTM International, West Conshohocken, PA, USA, 2012.
27. Rannaveski, R., Järvik, O., Oja, V. A new method for determining average boiling points of oils using a thermogravimetric analyser. *J. Therm. Anal. Calorim.,* 2016, **126**, 1679–1688.
28. Rannaveski, R., Listak, M., Oja, V. ASTM D86 distillation in the context of average boiling points as thermodynamic property of narrow boiling range oil fractions. *Oil Shale*, 2018, **35**(3), 254–264.
29. Rannaveski, R., Oja, V. A new thermogravimetric application for determination of vapour pressure curve corresponding to average boiling points of oil fractions with narrow boiling ranges. *Thermochim. Acta,* 2020, **683**, 178468.
30. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.,* 2011, **12**, 2825–2830.

31. Smits, G. F., Jordaan, E. M. Improved SVM regression using mixtures of kernels. In: *Neural Networks,* IJCNN'02., 2002, 2785–2790.
32. Storn, R., Price, K. Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.,* 1997, **11**, 341–359.
33. Jones, E., Oliphant, T., Peterson, P. *SciPy: Open Source Scientific Tools for Python*, 2001. http://www.scipy.org
34. Bertie, J. E., Zhang, S. L., Keefe, C. D. Measurement and use of absolute infrared absorption intensities of neat liquids. *Vib. Spectrosc.,* 1995, **8**(2), 215–229.
35. Bertie, J. E., Lan, Z. An accurate modified Kramers–Kronig transformation from reflectance to phase shift on attenuated total reflection. *J. Chem. Phys.,* 1996, **105**(19), 8502–8514.
36. Oja, V. Characterization of tars from Estonian Kukersite oil shale based on their volatility. *J. Anal. Appl. Pyrolysis,* 2005, **74**(1–2), 55–60.
37. Moro, M. K., Neto, A. C., Lacerda, V., Romão, W., Chinelatto, L. S., Castro, E. V. R., Filgueiras, P. R. FTIR, 1H and 13C NMR data fusion to predict crude oils properties. *Fuel*, 2020, **263**, 116721.