THESIS ON CHEMISTRY AND CHEMICAL ENGINEERING G49

Predicting Fuel Properties from Infrared Spectra

ZACHARIAH STEVEN BAIRD

TTÜ
PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Engineering
Department of Energy Technology

This dissertation was accepted for the defense of the degree of Doctor of Philosophy in Chemical and Materials Technology on July 31th, 2017.

Supervisor: Prof. Vahur Oja
Department of Energy Technology, TTÜ

Opponents: Prof. Dr. Selhan Karagoz
Karabuk University, Turkey

Prof. Dr. Mati Karelson
Tartu University, Estonia

Defense of the thesis: September 15, 2017

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for a doctoral degree at Tallinn University of Technology, has not been submitted for any other academic degree.

ISSN 1406-4774
ISBN 978-9949-83-145-6 (publication)
ISBN 978-9949-83-146-3 (PDF)

KEEMIA JA KEEMIATEHNIKA G49

Kütuse omaduste hindamine infrapunaspektrilt

ZACHARIAH STEVEN BAIRD

TTÜ KIRJASTUS

TALLINNA TEHNIKAÜLIKOOL
Inseneriteaduskond
Energiatehnoloogia instituut

Doktoritöö on lubatud kaitsmisele filosoofiadoktori kraadi taotlemiseks keemia- ja materjalitehnoloogia erialal 31. juulil 2017.

Juhendaja: Prof. Vahur Oja
Energiatehnoloogia instituut, TTÜ

Oponendid: Prof. Dr. Selhan Karagoz
Karabuk University, Türki

Prof. Dr. Mati Karelson
Tartu Ülikool, Eesti

Doktortöö kaitsmise aeg: 15. september, 2017

Deklaratsioon:
Deklareerin, et see doktoritöö, mis on minu iseseisva töö tulemus, on esitatud Tallinna Tehnikaülikooli doktorikraadi taotlemiseks ja selle alusel ei ole varem taotletud doktori- või sellega võrdsustatud teaduskraadi.

ISSN 1406-4774
ISBN 978-9949-83-145-6 (publication)
ISBN 978-9949-83-146-3 (PDF)

TABLE OF CONTENTS

LIST OF PUBLICATIONS

Z. S. Baird, V. Oja. "Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density," Chemometrics and Intelligent Laboratory Systems, vol. 158, pp. 41–47, 2016.

V. Oja, R. Rooleht, and Z. S. Baird, "Physical and Thermodynamic Properties of Kukersite Pyrolysis Shale Oil: Literature Review," Oil Shale, vol. 33, no. 2, pp. 184–197, Apr. 2016.

Z. S. Baird, V. Oja, and O. Järvik, "Distribution of Hydroxyl Groups in Kukersite Shale Oil: Quantitative Determination Using Fourier Transform Infrared (FT-IR) Spectroscopy," Applied Spectroscopy, vol. 69, no. 5, pp. 555–562, May 2015.

LIST OF ABBREVIATIONS

%AAD	Average absolute relative deviation
%RMSE	Root mean squared relative error
AAD	Average absolute deviation
GC	Gas chromatograph
H/C	Hydrogen-carbon ratio
MS	Mass spectrometer
PLS	Partial least squares
R	Pearson correlation coefficient
RMSE	Root mean squared error
SVR	Support vector regression
NMR	Nuclear magnetic resonance
PNA	Paraffins, naphthalenes, aromatics
SARA	Saturates, aromatics, resins, asphaltenes

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Current correlations for many fuel properties are based on bulk properties or detailed chemical information, both of which can be time consuming and costly to measure [1], [2]. Infrared spectra provide a convenient alternative because they can be easily measured and correlated to a wide variety of properties. Although predictive models based on infrared spectra have long been used with fuels, the vast majority of earlier work has been focused on quality parameters of fuels. A review of research done in this area for fuel property prediction is given in Chapter 2 (see also [3], which is included here as Article 1). In this thesis it is instead proposed that models based on infrared spectra could be used as a substitute for existing thermodynamic correlations to provide detailed information about the behavior of a fuel. Based on the review in Chapter 2 and Article 1 [3], no one has tried such an approach before.

When attempting to use multivariate infrared models for predicting thermodynamic properties three main questions arise:

- Which properties are more difficult to predict and why?
- Can infrared models perform at a level equivalent to that of the bulk property correlations commonly used?
- Is it possible to develop infrared models as a complete solution for thermodynamic property prediction?

This thesis addresses these questions.

The theory underlying property prediction using chemometric models, especially models based on infrared spectra, is detailed in Chapter 3. Because infrared spectroscopy is the most often used input data for these chemometric models, Chapter 4 gives an overview of the various infrared measurement systems that exist.

In this PhD research the feasibility of predicting thermodynamic data from infrared spectra was investigated by using experimental data for Kukersite shale oil samples. An overview of the database containing this experimental data is given in Chapter 5, along with information about sample preparation and the measurement techniques used. Kukersite shale oil was used because our research group was performing a project to measure and predict thermodynamic properties of this oil (Project AR10129 “Examination of the Thermodynamic Properties of Relevance to the Future of the Oil Shale Industry”). Additionally, because Kukersite shale oil contains large quantities of polar phenolic compounds, many conventional correlations give poor results for Kukersite shale oil. Therefore, it provides a good test case because many other alternative fuels, such as bio-oils and coal liquids, contain polar compounds. Based on the experimental data and infrared spectra (overviewed in Chapter 6), multivariate models were created using partial least squares regression and support vector regression, and the method for doing so is described in Chapter 0.

Models for simpler temperature independent properties were created first, and the results are presented in Chapter 8. Properties related to molecular structure (i.e. density and refractive index parameter), molecular size (i.e. molecular weight and average boiling point) and chemical composition were investigated, along with a few other properties. Although predicting a thermodynamic property at one temperature does provide good information, because many of these properties change with temperature, the real need is to predict these properties over a range of temperatures. Chapter 9 outlines an approach for doing so, and shows results for two temperature dependent thermodynamic properties.

The performance of the models is then assessed in Chapter 10 to determine which properties can be predicted well and whether or not models based on infrared spectra can indeed be used as a substitute for conventional thermodynamic correlations for fuels. Based on the experience gained while performing this PhD research, some obstacles in using models based on infrared spectra were identified. Chapter 11 gives more information on these issues, and these can be directions for future research.

1. FUEL PROPERTY PREDICTION – A REVIEW

The need for property prediction

We can gather information through observation, but as soon as we begin to interact with the world around us predictions are required. Predictions are our attempts to use observed information to estimate what occurs in situations we have never observed. As we analyze the information we already have we start to see underlying patterns and causes, and our predictions evolve from mere guesses towards sophisticated models and theories. In this way, predictions provide an outlet for our knowledge and lead to a practical result, and at the same time provide motivation for increasing our understanding of the world around us. Indeed, prediction is the connection between knowledge and action.

For chemical systems in particular, property prediction allows a material's properties to be estimated when experimental data is unavailable and can reduce costs by decreasing the number of measurements required. Accurate predictions essentially give us additional knowledge about a system, which helps us make better decisions and achieve better results. Also, in our attempts to predict properties we often gain a more fundamental understanding of the laws of nature that determine what properties a material will have.

Approaches to fuel property prediction

Prediction is finding a relationship between parameters so as to estimate one or more parameters from the others. The process can be broken into three phases: (1) collecting input data, (2) transforming the input using a data abstraction method, (3) modeling the relationship between the data and the property.

Collecting input data

For fuels experimental data is often used as the input to the models, and the main types of experimental data used are [1], [2]

- Bulk, or average, properties (density, average boiling point, distillation curve, molecular weight, etc.)
- Simple chemical composition information (PNA composition, SARA composition, NMR spectra, infrared spectra)
- Detailed molecular information (data from chromatography coupled with NMR or MS)

Although a detailed chemical description is readily available for pure compounds and simple mixtures, the majority of fuels are complex mixtures for which detection of all the compounds present is infeasible. Instead, the experimental data usually only enables identification of groups of compounds (such as paraffins, naphthalenes and aromatics (PNA)) or average parameters describing the molecular structures found in the sample (such as types of bonds or functional groups). Modern analytical techniques, such as gas chromatography coupled with mass spectrometry (GC-MS), can indeed provide a large amount of information, but even these instruments do not provide complete separation. Very similar compounds, such as isomers, will still be indistinguishable [4]. Samples for which information about the chemical composition is known are termed defined mixtures. When this information is not known the samples are called undefined mixtures, and these mixtures must be characterized using their bulk physical properties, such as density, average boiling point, viscosity or refractive index.

Sometimes the measured data is also converted to another input type. For example, detailed data from GC-MS systems can be used to calculate simpler parameters, like the PNA ratio or density and distillation curve. This is often done so that correlations can be used that are based on those input parameters. Also, some methods have been developed that attempt to determined compositional information from simple bulk properties [2], [5].

The optimum type of input data depends on the cost of obtaining the data and the relative benefit of having more information. For situations where detailed experimental data is difficult to obtain, such as for reservoir fluids, methods based on bulk properties are often used [6]. In large refineries, where improved optimization can lead to higher profits, detailed molecular information may be used [4]. However, because such detailed data can be expensive to obtain [2], some compromise between the extremes is usually found that gives enough information to obtain reasonably accurate predictions for a good range of different fuel compositions.

Data abstraction

Although predictions can be made directly from the input data, often more robust models can be obtained using some form of data abstraction. Data abstraction is used to transform input data to a form that makes it easier to get good predictions for many different fuel compositions. Generally, data abstraction is not used with simple correlations based on bulk properties (although something like the Watson K characterization factor [7] could be considered a form of data abstraction), but complex data such as chromatograms or spectra often need data abstraction for predictions to be accurate because such data can vary significantly between fuel samples or measurement device, which can make it difficult for regression methods to find a good predictive correlation [8]. For fuels data abstraction is often accomplished by converting compositional data into one of the following forms:

- Pseudocomponents [1]
- Molecular type and homologous series matrix representation [2]
- Average molecular structures
- Molecular components (large collection of individual representative molecules) [2], [4]

These different data types all attempt to do the same thing: represent the mixture in terms of a simplified composition by lumping together similar compounds. A pseudocomponent is a group of similar molecules that are lumped together and considered a single component [1]. Similarly, a molecular type and homologous series matrix divides the mixture by molecular class (alkane, alkene, cycloalkane, aromatic, etc.) and size (carbon number or average boiling point) [2]. The mixture can also be considered a mixture of functional groups or small structures, characterized using average molecular structural parameters (such as the average number of carbon atoms per molecule, weight percent of CH_2 groups or hydrogen/carbon ratio) [2]. The most detailed form is to use a large number of representative molecules [2], [4], [8], for which individual compounds or groups of compounds (often isomers) in the mixture are identified.

Modeling the relationship

Property prediction is usually accomplished using one of these types of correlation:

- Empirical equations with a few variables
- Empirical equations with many variables
- Equations based on theory (equations of state, group contribution methods, etc.)

Many equations based on bulk properties are simple empirical correlations that take the form of an algebraic equation containing a few variables. With the advent of computers, it is now also feasible to create multivariate equations based on a large number of variables, which can use the larger datasets provided by modern analytical techniques. These multivariate correlations, or models, are usually found using advanced statistical and mathematical regression methods, including partial least squares regression, neural networks and support vector regression.

Often, empirical equations are simply used to determine the parameters for an existing thermodynamic relationship, such as an equation of state or a group contribution method. Then important physical properties and phase behavior can be calculated from these thermodynamic relationships, which reduces the number of empirical correlations needed and can extend the prediction range.

Shale oil

This research investigation used data for Kukersite shale oil samples. Shale oil is a liquid fuel produced through pyrolysis of oil shale [9], [10]. Oil shale is a solid fossil fuel that is found in great abundance around the world [11], and it has been estimated that about 4,700 billion barrels of shale oil could be produced from these reserves [12]. The largest oil shale deposit is the Green River Formation in the United States, and other major deposits are found in Australia, Brazil, China, Congo, Estonia, Italy, Jordan, Morocco and Russia [12]. Although large reserves are available, oil shale has seen only limited use due to technological and environmental challenges that have kept it from being an economically viable fuel in most situations [13]. However, several countries still have active oil shale industries, including Estonia, China and Brazil, and several countries are investigating the use of their oil shale resources as a way to reduce dependence on foreign energy imports [12]. In Estonia, the oil shale industry has operated for about a century, and about 2.3 million barrels of shale oil were produced in 2008 [14].

To produce oil, the oil shale is heated to high temperatures in an inert environment and pyrolysis reactions start to occur [9], [10]. During pyrolysis, the solid organic macromolecular structure (called kerogen) starts to decompose into smaller molecules and oil and gas are produced. A variety of technologies have been developed to carry out oil shale pyrolysis. So far all commercial production has taken place in above ground plants (also called retorts), but many investigations have also been made into producing oil underground directly in the oil shale formation (termed in situ retorting). The composition and properties of the resulting shale oil depend not only on the raw oil shale used, but also on the retorting technology [15].

Property prediction for shale oil

Most property correlations for oil shale oil are simple algebraic formulas based on bulk properties of the oil. This is due to the lack of experimental data for creating correlations. The small amount of data on important physical and thermodynamic properties found in the literature for shale oil is for undefined fractions, and often only a small amount of characterization data is presented (see Article 2 [16]). Therefore, only the simplest correlations for undefined mixtures can be developed.

As was reviewed in Article 2 [16], most of the data and correlations are given in three sources [17]–[19]. These correlations are presented either as simple algebraic equations of one or two variables or as figures and tables. Most of the data is for lighter fractions with average boiling points below 350 °C, and the data in these sources was measured during the period from 1920 to 1950. A larger database, containing samples with a wider range of compositions and more detailed characteristic data, is needed to support development of more advanced predictive correlations.

2. CHEMOMETRICS FOR FUEL PROPERTY PREDICTION

Overview of developments

Many analytical techniques, such as infrared spectroscopy, gas chromatography and mass spectrometry, result in a large amount of data about a sample. With the advent of computers, more complex mathematical and statistical tools became widely available for analyzing these data sets, and a field emerged that was related to their use for extracting chemical information from data: chemometrics [20].

Beyond just chemical concentrations, it has long been shown that other properties can be predicted from chemical information through the use of multivariate (chemometric) methods. Some early studies were done on wheat to predict properties like the protein and moisture contents (1982) [21], and it was shown that quality parameters of tape could also be predicted (1984) [22]. The earliest study found about predicting fuel properties was a study on peat from 1986 in which the calorific value, carbon content, quantity of hydrolysable material and volume weight were predicted [23]. Other studies soon followed, and it appears that there has been increasing interest in this area, as judged by the fact that the number of papers published on the subject per year has been continually increasing. In performing this review (published in Article 1 [3]) more than 300 papers and patents were found in which fuel properties were predicted using chemometric methods.

So, the basic idea of predicting properties from data from analytical methods has long been shown to be possible. Over time the range of fuels studied and number of properties predicted has increased. Different regression or machine learning methods have also been used over time as they are developed by the wider research community, including methods for calibration transfer. However, many of the papers are very similar, and most offer little in terms of new insights or developments.

There are, however, a few noteworthy papers that stand out. Balabin and Smirnov [24] compared different regression/machine learning methods on their performance when extrapolation or interpolation is required. A research group from the U.S. Naval Research Laboratory has also published a couple thorough papers [8], [25]. Both have a wide scope, predicting many properties for a wide range of fuels from around the globe, and they show how difficult it can be to create a model that can be extended to cover the wide range of fuels used around the world. Additionally, in their later paper [8] they introduce a data abstraction scheme to try to address this difficulty of predicting properties for fuels with compositions that do not fall within the calibration range, which is a major problem that limits the usefulness of chemometric models for fuel prediction (see also Chapter 11). Some papers have had a more general aim of introducing improved or new chemometric methods, and have simply used fuel property prediction as a test case.

To get a better overview of work done with fuel property prediction, the reviewer examined the properties predicted, the types of fuel used, the analytical methods used for gathering input data, the accuracies of the models and the regression methods used. The results of this review were initially published in

[3], and this article is included as a supplement to this dissertation (Article 1). For ease of reading, the main points are reiterated here.

Properties predicted

A wide range of fuel properties have been predicted. In total, 104 different properties were identified in this review, and most certainly there have been others predicted. A list of all the properties that were predicted in at least 3 sources is shown in Table 1 of [3] (article included here as Article 1), which also gives some statistics about the range of accuracies for models for each property.

This analysis showed that almost all of the properties that have been predicted are fuel quality parameters. Some important thermodynamic and physical properties are also quality parameters, and have therefore been predicted, although temperature dependent properties have only been predicted at a single given standard temperature.

Fuel types used

Models have been created for many different fuels, but petroleum and liquid biofuels (biodiesel and ethanol) have been investigated most. The proportion of articles investigating each type of fuel is shown in Figure 2-1. No studies were found that predicted properties for shale oil besides those published as part of this PhD work.

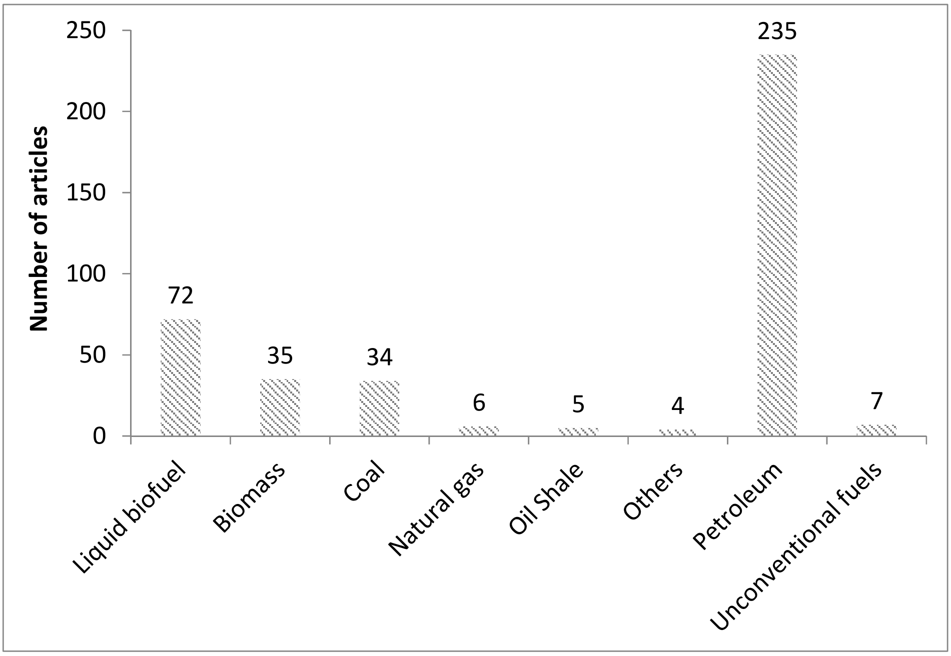

Figure 2-1. Fuel types studied in articles which predicted fuel properties using chemometric methods. The Unconventional fuels category includes coal liquids, shale oil and Fischer-Tropsch fuels. The Others category includes charcoal and rocket fuel. (Figure 1, Article 1)

Analytical method used for measuring input data

The most popular analytical method has been infrared spectroscopy (see Figure 2-2), which is probably due to the fact that it is gives quick measurements without any need for sample preparation. It is also applicable to many different types of samples, and correlates with many properties. Installation costs are also lower than some other methods because the measurement accessory can often be inserted directly into the process stream and using fiber optics with a near infrared spectrometer allows measurements at multiple process locations simultaneously. Many other methods have also been used, although less frequently, and the Other category includes a wide range of different methods, such as an electronic nose and a thermal wave interferometer.

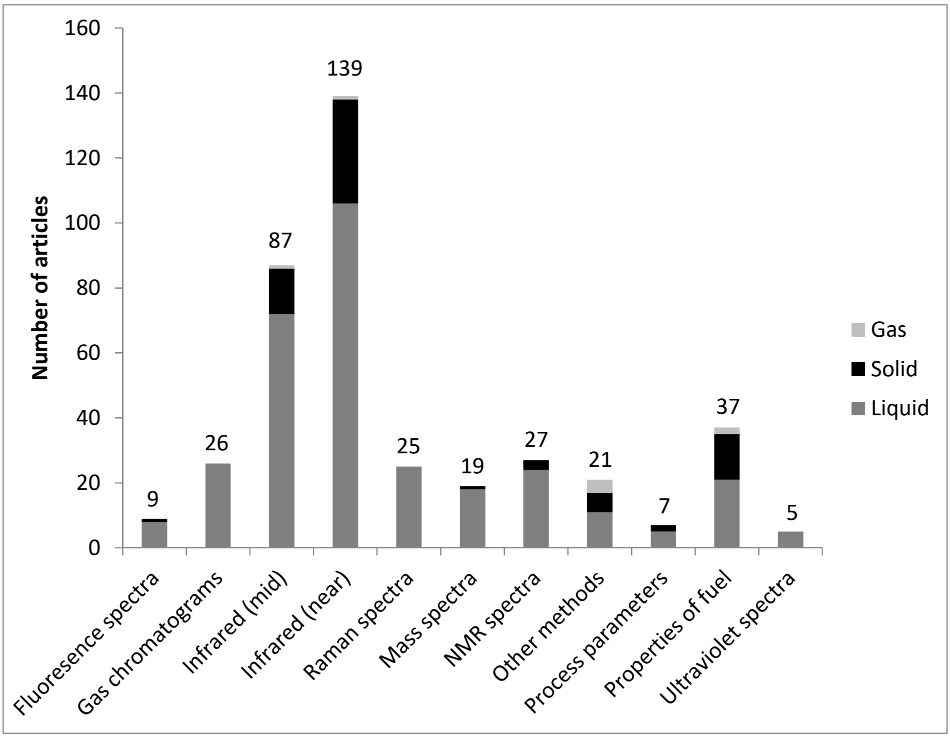

Figure 2-2. Types of input data used for predicting fuel properties using chemometric methods. The proportion from different types of fuels (solid, liquid or gas) is also shown. (Figure 3, Article 1)

Accuracy of the models

The accuracy of a model depends on a variety of factors, including things like the repeatability of the spectrometer, the accuracy of the reference data, the presence of outliers, the range of samples used and the strength of the correlation between the input data and property to be predicted [26]. Therefore, a full comparison of the performance of different models would require more information and a more detailed analysis. However, a few observations can be made simply by comparing the accuracies of the models given in the literature reviewed. Summary statistics for many different properties are given in Table 1 of [3] (article included here as Article 1).

More information can be found by looking at the distribution of accuracies for different models of the same property. This distribution is shown for liquid density in Figure 2-3. The distributions examined in this review appeared to have a left skewed distribution, as shown by the distribution for liquid density. The model that stands out as having the largest errors (0.028 g/cm^3) was for crude oil residual fractions [27], which are more difficult to handle and measure. There was a general correlation between the types of fuels examined and the accuracy. Models with larger errors were mostly for crude oils, residual oils or fuels from a broad range of sources (including nonpetroleum sources). The models with the smallest errors were for fuels like diesel, gasoline and biodiesel, for which density measurements can be more precise and which cover a smaller range of compositions. Some studies also only investigated a small sampling of fuels, and a narrower range of compositions would also likely enabled tighter fits and better accuracies.

Again, more information and analysis would be needed for any further investigation as to the causes of difference in performance. However, it is evident that different models for the same property can have a fairly wide range of accuracies, and some of difference is inherent to the situation being investigated and not affected by the quality of the analysis. Thus, a model that has larger errors than other models may still be a well-developed model if the samples used or other situational factors keep the accuracy from being any better.

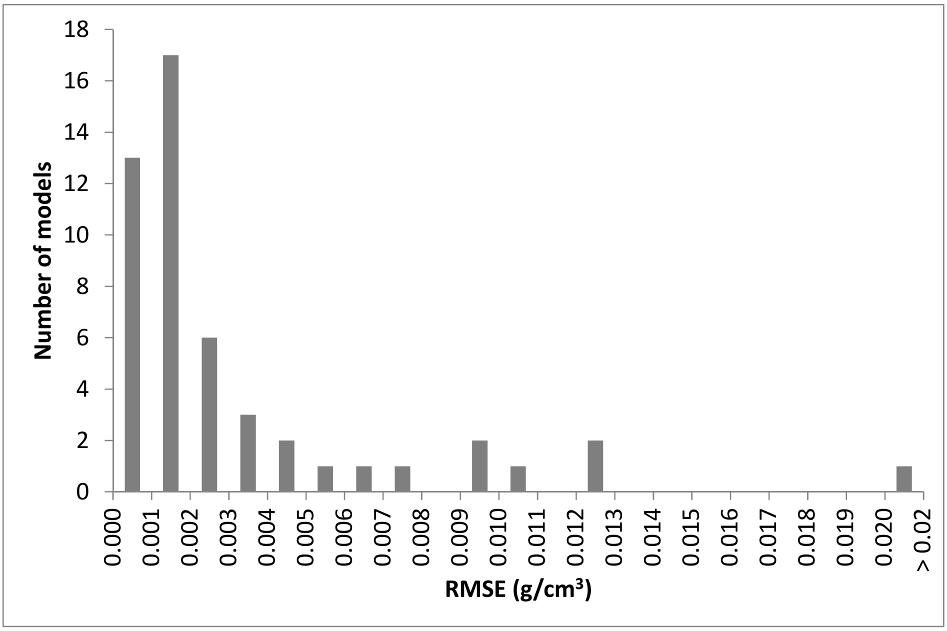

Figure 2-3. Distribution of models for fuel liquid density according to their root mean squared error. (Figure 2, Article 1)

Regression methods used

Partial least squares (PLS) regression is by far the most often used regression technique, which could be expected because it has been the main method used since even the early beginnings chemometrics. Over time there has been a slow transition towards using newer methods. Early on principal component regression and multiple linear regression (ordinary least squares regression) were used, but these are now used less often. Instead, techniques that can model nonlinearity are receiving more attention, such as nonlinear versions of PLS (poly-PLS, spline PLS, kernel PLS), artificial neural networks and support vector regression. Also, note that within a given group of regression methods there are often many different algorithms and modifications have been used.

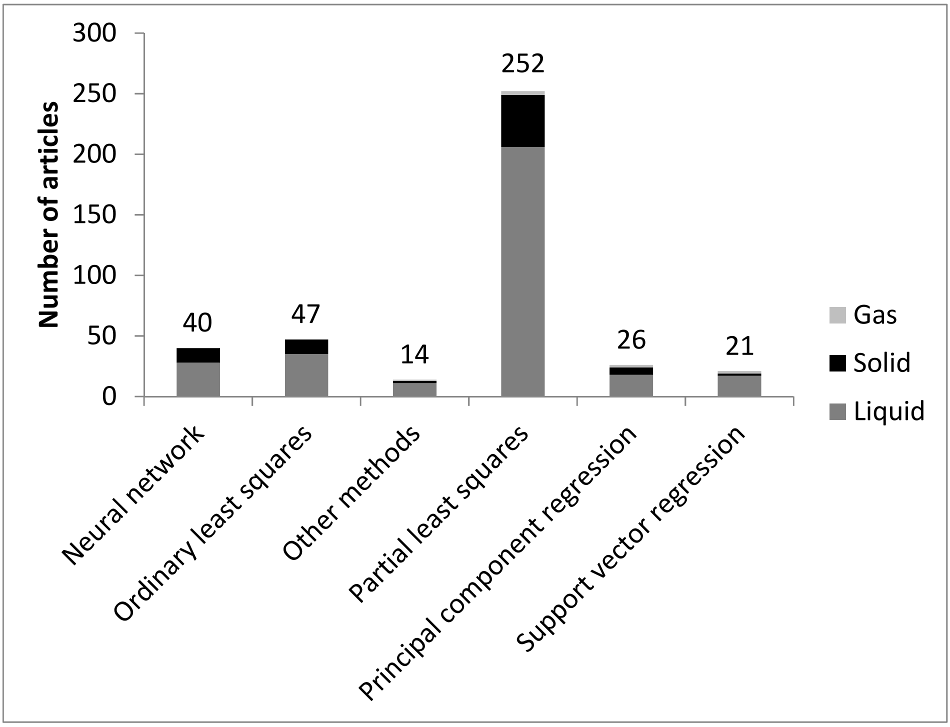

Figure 2-4. Regression methods used in articles that predicted fuel properties using chemometric methods. (Figure 4, Article 1)

3. THEORY

The basis for predicting properties from chemical information

If there is a compound or class of compounds that produces a distinct peak on a spectrum or chromatogram, then their concentration can often be directly measured because the peak area is a function of the concentration of the compound. Most fuel properties, however, are not directly measured by the analytical method (spectrometer, chromatograph, etc.). For this reason a chemometric system for determining such properties is often called a “soft sensor” because it does not directly measure the property like a real sensor would. The system is just a type of prediction method.

What even enables such predictions to be possible? As with any prediction method, a relationship must exist between the input information and the parameters to be predicted. The properties of a fuel are, naturally, related to the chemical composition of the fuel, and these relationships allow properties to be predicted from information about the chemical composition. Of course, the relationships between composition and properties can be quite complex, which is why empirical tools, such as chemometric methods, are often used for analyzing and approximating the relationships. Also, for complex mixtures, which most fuels are, current analytical methods do not give the exact composition. Instead, they give some form of simplified information about the sample, and thus, the real relationship is between this simplified compositional data and the properties.

Because infrared spectroscopy is the main analytical method used for chemometric predictions of fuel properties, it deserves a more detailed analysis. Molecules absorb electromagnetic radiation in the mid or near infrared ranges (wavelengths of about 0.8 to 25 μm) if a photon contains the right amount of energy to cause a molecular vibration. The energy level, or wavelength, at which a vibration occurs depends in large part on the strength of the bond and on the molecular structures surrounding it. The absorption is registered by the spectrometer and a corresponding peak occurs on the spectrum. [28], [29] So, infrared spectroscopy essentially gives information about the strength and local environment of bonds in a sample.

Some important limitations arise due to the nature of the information contained in infrared spectra. First, an infrared spectrum is related to the types and concentration of bonds in the sample, but the exact relationship between spectrum and chemical composition is often complicated. This means empirical relationships are generally needed, even when predicting parameters related to chemical composition. Second, infrared spectra mostly give information about local bond structure, and do not directly give information about overall molecular structure. As a result, different compounds or mixtures of compounds can give identical spectra. This also implies that infrared spectra do not directly give information about the molecular weights of the molecules in the mixture. [30]

Based on these limitations, the correlation between a property and infrared spectra should theoretically be more direct for some properties than others. Trygstad et al. [31] emphasized this, and gave sulfur content and Reid vapor pressure as two examples of properties for which the correlation is tenuous. For both, the concentrations of compounds that directly affect these properties are quite low, and they cannot be detected from the infrared spectrum of the fuel. According to this line of reasoning, spectra give information that is directly related to the concentration of bonds in a sample, and therefore, parameters such as carbon or hydrogen content should be more directly related to the spectra. Conversely, properties dependent on molecular size, such as average boiling point or average molecular weight, should have a weaker indirect relationship to infrared spectra.

In reality though, there are still often strong correlations between infrared spectra and properties such as molecular weight and sulfur content. In these cases there is apparently some underlying cause that leads to simultaneous changes in both the compounds affecting the property and some other functional group which is observable in the infrared spectra. For example, this may occur when the processing method used to remove sulfur also affects other compounds in the sample, or when the types of bonds and functional groups change as the molecular weights of samples increase. In other words, spectra can indirectly contain significant amounts of information about these properties that theoretically should only be weakly related to infrared spectra.

Overview of multivariate regression methods

Multivariate regression methods are used with data sets that contain more than two variables. During the past decades the size of data sets has grown, which has led to increased interest in multivariate statistics and regression methods. Here we give a brief overview of some of the main concepts related to multivariate regression.

Likely the simplest method is multiple linear regression, which is patterned after ordinary linear regression and finds a single linear coefficient for each input variable. The coefficients are found directly from the input data, often by finding coefficients that minimize the sum of the squares of the model residuals. However, this method can quickly become ineffective as the number of variables increases.

A common problem with multivariate regression methods is having fewer data points than variables. Another is that relationships often exist between the input variables (called collinearity). In both cases multiple linear regression will find several sets of coefficients for which a minimum occurs, but cannot provide any distinction as to which solutions will actually provide good predictions. For this reason, methods like principal component regression, ridge regression and partial least squares (PLS) regression were developed.

Almost all multivariate regression methods seek to deal with the problem of sparse data by identifying a smaller number of underlying structures in the data that can be used as preferred directions for regression. By defining these underlying metavariables, the data set is simplified and a stable regression can be achieved. Principal component regression and PLS regression identify what are called components, which are related to the eigenvectors of the data [32], [33]. Often large data sets can be described by just a handful of components, which reduces the amount of data needed and leads to a regression model with better performance.

However, these methods are still linear methods. Often neural networks or support vector regression are used, which can model nonlinearity more effectively. These describe the underlying structure of the data in terms of neurons or support vectors, and can provide a closer fit than linear regression methods in many situations.

Multivariate regression methods can be quite powerful, and one danger is that they can over fit the data. This occurs when they start to model noise in the data and fit it so closely that the model's prediction accuracy actually decreases. Therefore, when performing regression, or training, the error of the model must be determined from data points that are left out of the training set (the set of points used for regression, or training). This second set, or test set, can give a good estimate of the actual accuracy of the model's predictions and can indicate at which point added model complexity does not increase accuracy. Parameters to be used by the regression methods can then be optimized based on this prediction error.

Setting data aside as a test set can be problematic, however, because it requires more data to be measured to compensate for those points that are left out. Additionally, the points left out have to be chosen in such a way as to give an accurate representation of the error that could be expected for model predictions. To avoid these problems cross validation is used, which involves performing regression multiple times with different samples left out each time. Often the data set is split into groups, or folds, and each group is sequentially left out. If only one data point is left out at a time, it is called leave-one-out cross validation. The estimated error is then obtained from the residuals of each data point when it was left out of the regression. After estimating the error this way, all the data points are used in one final regression to get the regression model.

The remaining sections of this chapter give more details about the two methods used in this work, PLS regression and support vector regression, and also describe

Partial least squares regression

Partial least squares regression is based on finding underlying components that describe most of the variation in the input data and also correlate well with the property to be predicted [32]. The components are then summed together to give the overall predictive model. The final model takes a linear form, which is given as Equation 3-1

$$Y = X \cdot B + C \tag{3-1}$$

where X and Y are the matrices of input and predicted property values, B is the matrix of model coefficients and C is a constant.

The use of components simplifies the data and provides a way to handle the collinearity that can occur in high dimensional datasets. The procedure for finding the components, or factors, is similar to that used in principal component analysis. In principal component analysis a vector is chosen that points in the direction of maximum variance (which is an eigenvector of the data). That vector is taken as a component, and it can be represented as a linear combination of the input variables. For PLS regression this vector is found for both the input and predicted variables, but in choosing the vector the scores of the input and predicted matrices are switched. This allows information about the relationships between input and predicted variables to be used for finding the vector, and this rotates the vector to point more towards those variables that are more closely related to the variables to be predicted.

The components are calculated sequentially. After a component is calculated, the residuals are calculated by subtracting the component from the data, and the residuals are used as the input for the next PLS cycle. Each additional component adds complexity to the model, and the number of components used in the final model is usually chosen so that it minimizes the error in the prediction, which can be estimated using cross validation. At a certain point additional components start to model the noise in the data, and prediction accuracy will decrease if these components are included. To find the final model, all the component weights for each input variable are summed up to give a single regression coefficient for each variable (the coefficient matrix B in Equation 3-1), and the regression constant C is calculated.

PLS regression generally provides better accuracy than multiple linear regression because PLS regression can take into account the collinearity that exists between variables. This has made it a popular choice for multivariate regression, especially in the fields of chemistry and chemical engineering. It has also been used in industry for process monitoring and control [34]. Additionally, because PLS is a linear method it generally gives better results for extrapolation than neural network methods [24].

However, since PLS regression is a linear method it generally does not give as good of an accuracy for nonlinear data sets [24], [35]. It can still approximate some degree of nonlinearity because it incorporates multiple components, but generally other methods such as neural networks or support vector regression will give better accuracy. For this reason, some nonlinear variations of PLS have been proposed which seek to account for this disadvantage, including Poly-PLS, Spline PLS and Kernel PLS [34], [36].

Support vector regression

Support vector regression attempts to find a function that fits the data closely, but is also as smooth as possible [37]. The regression equation takes a linear form. The cost function used to describe the two goals of the regression is given by Equation 3-2.

$$minimize \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \left(\xi_i + \xi_i \ \right) \tag{3-2}$$

Here w is the set of weights used in the linear regression equation, and taking the 2-norm of the weights effectively smooths out the resulting regression equation. The second term is related to the error of the fit, and uses slack variables (ξ_i and ξ_i^*) to make the regression problem feasible for a wider range of data sets. The slack variables essentially determine the maximum allowable error for a given observation, and minimizing that term reduces the overall error of the fit. C is the parameter that controls the tradeoff between the two objectives. If C is larger, then the regression algorithm will search for a closer fit at the expense of the smoothness of the function. The minimization is performed subject to the constraints given as Equation 3-3

$$\begin{gathered} y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i \\ \xi_i, \xi_i \ \geq 0 \end{gathered} \tag{3-3}$$

where x and y are the input data. ε is an error limit. If an observation has a residual that is higher than ε then it is termed a support vector, while an observation with a residual less than ε has slack variables equal to zero and does not impact the resulting regression. If ε is not used (or is set to zero), then the method is termed least squares support vector regression. [38]

The optimization problem is generally solved in its dual formulation because it is usually considered easier to do so than solving the primal problem [37]. In optimization, the dual of a problem is formed by creating a Lagrangian that takes into account both the original cost function and the constraints [39]. The dual problem can then be solved using quadratic programming techniques.

To obtain the best solution both C and epsilon need to be optimized. This can be done by finding the values that minimize the cross validation error. As mentioned, this involves leaving some data points out each time the regression is performed, and then using another optimization routine to change C and epsilon until a minimum cross validation value is obtained.

Although support vector regression finds a linear function that fits the data, it can also be extended to model nonlinear relationships by using kernels. A kernel is a function that maps, or transforms, input data into a new space, and so kernels can be used to embed nonlinear input data in a feature space in which it becomes linear [40]. Two popular kernels are the polynomial and radial basis function kernels. A polynomial kernel, as might be expected, transforms the data using a polynomial function. The radial basis function uses Gaussian functions centered at each of the observations, which means the function fits the shape of the data. Kernels usually have additional parameters that must be optimized. For the polynomial function, for instance, the degree and zero coefficient of the polynomial must be chosen, and these can also be optimized along with C and epsilon. [37], [40]

The choice of kernel often depends on the problem at hand, and domain knowledge is often useful in choosing. A kernel like the radial basis function follows the shape of the data and has the ability to provide better local fitting than linear or polynomial kernels. Indeed, when using a kernel like the radial basis function, support vector machines become universal approximators [41], [42]. However because of this, radial basis functions rarely provide good estimates when used for extrapolation. A good way to take advantage of the strengths of two kernels at the same time is to combine them by creating a mixed kernel. For example, by combining the polynomial and radial basis function kernels a kernel can be created that is mostly a polynomial and provides good extrapolation, but also has some of the radial basis function character to enable better adherence to the data. [37], [43]

Balabin et al. [24] compared SVR to other multivariate regression methods. They found that SVR performed well on chemometric problems where extrapolation and interpolation were required, while neural networks did not appear to have this ability. At the same time, SVR was also able to model nonlinear datasets, while more linear methods like PLS regression were not able to do this. So, SVR appears to be able to give better overall fits than many other methods: that is, it can give a solution that is complex and nonlinear, but that also has structure to it to allow extrapolation. Additionally, SVR leads to a global model, which is another advantage it has compared to neural networks, which can get stuck in local minima [37], [44], [45].

However, because SVR uses quadratic programming to find a solution, it takes more time than many other methods, such as PLS regression. This drawback becomes more important as the size of the dataset increases, which means SVR can become impractical for large datasets. So, for large datasets some sort of data reduction would likely be needed to obtain an answer in a reasonable timeframe.

4. INFRARED MEASUREMENT SYSTEMS

As noted, infrared spectra are the most widely used input data for multivariate models for fuel property prediction. Therefore, a few words about infrared measurement systems are in order. A variety of different infrared measurement systems exist, and an understanding of their differences can aid in selecting the best setup for a given application and in analyzing the resulting data.

Mid or near infrared

One way to categorize infrared measurement instruments is by their wavelength range. The range of wavelengths is determined by the radiation source. Usually the source emits in either the mid-infrared (2.5 to 25 μm) or near infrared (0.8 to 2.5 μm) region, but there are some sources that extend over both the mid and near infrared regions. Often these extended sources combine mid and near infrared sources in order to do so. There have been some studies that have attempted to compare the accuracy of models using the mid or near infrared regions, but in general they give comparable results because they contain related information. The peaks are more resolved in the mid-infrared region, however, and so qualitative analysis often uses this region. Near infrared spectrometers have the advantage of being more stable, and they usually have a better repeatability than mid-infrared spectrometers.

Types of accessories

Infrared instruments can also be categorized by measurement accessory. The measurement accessory used has a large effect on the resulting spectrum, and certain accessories may be preferred depending on the type of system being measured. The main types of infrared measurements are transmission, attenuated total reflection and reflection.

Transmission

With transmission measurements the infrared beam simply passes through the sample. For liquids a sample cuvette is often used that is made of a material that allows the infrared radiation to pass through. For measurements in a process transmission probes are available which have a small gap through which the sample can flow to be measured. For liquids the path length should be small enough that 100% absorption (saturation) does not occur.

For gases transmission measurements can be made either in a separate gas cell or directly at the place of interest using an open path measurement. With gas measurements a longer path length is usually desired because gases do not absorb very strongly. An increased path length increases the sensitivity of the measurements, and concentration measurements at the level of parts per million can be attained. Open path measurements are often used for monitoring, such as part of a safety/warning system in a plant. Open path measurements are also used for atmospheric measurements, where the sun is the infrared source.

Attenuated total reflection

Attenuated total reflection (ATR) was developed later than transmission methods, and is now a popular choice of measurement accessory. ATR accessories use a phenomenon called total internal reflection, which as the name suggests, means that all the radiation stays inside the measurement crystal and none passes through the sample. When the beam hits the crystal/sample interface an effervescent wave is created which propagates a small distance into the sample (usually no more than a millimeter or two). Energy from that wave can be absorbed by the sample, which creates the infrared spectrum.

Some advantages of ATR accessories include that very small sample sizes can be analyzed. With a small single reflection crystal, a single drop of a liquid may be sufficient. Another advantage is that solids and strongly absorbing materials can be easily analyzed using this method because the beam does not need to pass through the sample, as it does with transmission measurements. Many ATR accessories also come with pressure attachments that can be used to ensure good contact between the solid and ATR crystal, and can also be used to compact powders to achieve better measurements.

The range of an instrument can be somewhat limited by the crystal chosen because at certain wavelengths the absorbance of the crystal itself can become significant. ZnSe is a popular choice because it has a high refractive index (which means it can be used to measure samples with a higher refractive index) and does not dissolve in water. Diamond crystals are often used with abrasive or corrosive samples.

ATR accessories also come in different configurations. The most common is to have the crystal placed flat in the accessory, and then to have the sample placed on top of it. Another type is a horizontal ATR. This is usually used with liquids, and generally consists of a small container that holds the liquid up against the side of the crystal. Circle ATR cells also exist, which consist of a cylindrical crystal placed in a small container that is filled with the liquid to be measured. In addition, crystals can be designed so that the infrared beam reflects of the crystal/sample interface multiple times, which increases the effective path length of the measurement, and therefore, the absorbance.

Reflection

Reflection measurements are generally performed on solids. It would be difficult to measure solids using transmission because a thin slice would be needed to allow the beam to pass through. Therefore, the spectrum is collected instead based on the radiation that is reflected off the sample. The advantage over ATR accessories is that the sample does not need to be positioned and pressed tightly against the ATR crystal. For this reason reflection measurements are often used in processes that handle solid materials, such as wheat or coal, because the infrared sensor can be placed above a conveyor belt to continuously scan the material. In laboratory instruments the samples are often ground into small pieces and mixed with a material that does absorb infrared radiation, such as KBr powder.

Reflection spectra are generally more complicated than transmission or ATR spectra because the reflection spectra also depend on parameters such as the size of the sample particles and the distance between the infrared receptor and the sample. This can make analyzing and using these spectra more complicated.

Implementation in industrial processes

For use in continuous monitoring of a process, a probe can either be placed directly in the process or the process can be constructed to have a cell through which transmission measurements can take place. ATR and transmission probes exist, and can often be placed directly into a process stream. A flow through cell can also be added to a process, which has a window that allows the infrared beam to pass through, but adding this can be expensive and sampling errors may also occur. A separate flow through measurement cell may also be used to allow control of the temperature at the point of measurement.

Generally, industrial systems use fiber optics to direct the infrared beam to the measurement probe. For this reason near infrared spectrometers have a clear advantage over mid infrared spectrometers for industrial applications. Many common materials can be used to construct fibers for the near infrared range, and due to their low absorbance in the near infrared range, the fibers can even be 100 m long. This allows one near infrared device to monitor several points in the process simultaneously. For the mid-infrared range, however, there are only a few fiber materials that can be used. Usually polymer materials are used, and even these absorb enough of the mid-infrared beam that the fibers usually cannot be much longer than 1 m.

And as mentioned, reflection measurement systems can also be used, and for these the infrared receptor is generally placed above a conveyor belt that moves the material. For these systems, the positioning of the infrared receptor, the speed of the conveyor and the number of scans averaged together can all have an effect on the resulting spectra, and optimal parameters may need to be selected to ensure good model performance [46].

Wavelength separation methods

In recent years, new methods of separating an infrared beam by wavelength have been developed that allow small spectrometers and sensors to be developed. Early spectrometers would often use a prism or different grating slits to allow specific wavelengths to be isolated and measured. Then Fourier transform spectrometers were developed, which allowed rapid and more precise measurements.

Fourier transform spectrometers use what is called an interferometer to separate the infrared beam by wavelength. One of the most common types is a Michelson interferometer, which uses a beamsplitter and a moving mirror to retrieve information at specific wavelengths from the overall infrared beam. In the Michelson interferometer the infrared source is split into two beams by the beamsplitter. One of the beams passes through the sample, where absorption takes place. The two beams are then recombined, which causes interference to occur in which the beams either add together or cancel each other out (constructive or destructive interference). The proportion of constructive or destructive interference that occurs depends on the wavelength of the radiation and the difference in the distance covered by the two split beams (called the optical path difference). By moving a mirror placed in the path of one of the beams, the optical path difference can be changed, which changes how much each wavelength contributes to the overall beam sent to the detector. This interferogram is then transformed to give information according to wavelength (a spectrum) using a Fourier transform. [47]

By using an interferometer and a Fourier transform information can be gathered about multiple wavelengths simultaneously, and the overall measurement time decreases. With older prism or grating instruments it would often take a couple hours to measure a spectrum, but with Fourier transform spectrometers a spectrum can be measured in less than a minute. Many interferometers, however, are large. More recent research has led to new wavelength separation methods that allow spectrometers to be smaller and cheaper.

One example is linear array detectors. For these a small chip is created that has infrared detectors placed on it as tiny pixels. Then a layer is applied to the top of the chip to separate the infrared light by wavelength. This is usually done by applying the layer with a varying thickness, and then the layer works like a prism. Different pixels in the array are then exposed to only a specific wavelength of radiation, and the responses for all the pixels can then be combined to give information at multiple wavelengths, i.e. a spectrum.

Another interesting development is micromechanical interferometers. These use micro machining techniques to create an interferometer that is small in size. One type creates a micro scale Fabry-Perot Interferometer, which works by placing two reflecting mirrors close together and varying the distance between the mirrors to select different wavelengths of radiation [48], [49]. The result is a small infrared sensor, which can be used to create smaller and cheaper spectrometers.

These new developments in spectrometer construction can enable measurements to be performed more cheaply and in situations where it would be difficult to use the larger spectrometers that are most commonly used right now. Many of these new infrared sensors are meant to be cheap, and therefore, produce low quality data. However, there are also several high end products available with performance that rivals that of even laboratory spectrometers.

5. CREATION OF PROPERTY DATABASE

Samples used

This work was performed as part of a larger project to measure the thermodynamic properties of Kukersite shale oil. The shale oil for this project was obtained from Estonian Energy's Narva Oil Plant (Narva, Estonia). This plant uses solid heat carrier technology [50] (sometimes called the Galoter process), and uses Kukersite oil shale (Estonia). Even for one specific process the production regime can change, which then can result in compositional differences in the products. Other factors, such as natural variation in the raw materials or processing conditions, lead to variations in the resulting products. To be able to take these fluctuations into account, samples were taken at multiple different times over the course of the project (2013 to 2015). Additionally, samples were obtained from both the older Enefit 140 plant (also known as UTT 3000) and the newer Enefit 280 plant. In the plant the crude shale oil is separated into three wide technical fractions: shale gasoline, fuel oil and heavy oil. Gasoline and fuel oil samples were used for the models developed for this PhD research.

Sample preparation

Distillation

The wide fractions from the plant were further separated into narrow boiling fractions using distillation. Most of the distillations were simple batch distillations at either atmospheric pressure (an Engler distillation, ASTM D86 standard [51]) or in a vacuum. The experimental setup for the simple distillations involved heating a glass bulb containing the sample, and collecting the vaporized compounds in a glass condenser that was cooled to about -40 °C using a thermostat. From there the condensed oil flowed down into a flask for collection. In general, fractions were taken at about 5 to 10 °C intervals. The ASTM D86 standard specifies using 100 ml of the oil, but due to the need to measure many properties of the oil fractions a larger bulb and more oil were used (generally around 300 ml).

Because the shale oil samples are unstable at higher temperatures, distillations could only be performed up to about 350 or 400 °C before the sample started to decompose. For the gasoline samples atmospheric distillation was sufficient, but for the fuel oil only about 40 wt% of the initial oil could be distilled at atmospheric pressure before reaching decomposition temperatures. Therefore, the project used vacuum distillation to recover the higher boiling fractions. For these vacuum distillations the lightest portion of the fuel oil was first distilled at atmospheric pressure. Then the system was evacuated and distillation was continued to recover heavier fractions. The setup used was the same except that the flasks for collecting the fractions were placed in a sealed container which was connected to a vacuum pump. The vacuum was usually maintained at about 20 mmHg. Using vacuum distillation about 90 wt% of the initial fuel oil could be recovered.

A few distillations were also performed using a rectification column. Two different rectification columns were used: a plate rectification column (4.2 theoretical plates, reflux ratio of 6:1) and a packed rectification column (15 theoretical plates, reflux ratio of 5:1). The packing material used was wire spirals with a length of 3 mm and an external diameter of 2.5 mm. The wire had a diameter of 0.24 mm. The height of the packed column was 0.86 m and the diameter was 3.5 cm. Once again, the rectification column could only be used up to a temperature of about 350 to 400 °C to avoid decomposition of the sample.

After distillation the mass recovered in each fraction was measured. Then the sample containers were closed in a nitrogen environment to reduce the amount of oxygen in the sample vials, and the samples were stored in a refrigerator.

The wide fractions from the plant were also included in the study. Also, a sample approximating the original crude oil was created by mixing the gasoline, fuel oil and heavy oil in the ratio given by the plant's design documents (1:3:1). The crude oil could not be obtained straight from the plant because the vapors from the retort are sent directly to the rectification column.

Extraction

To develop more robust models the composition of some samples was artificially adjusted via extraction and/or mixing. This was an important part of the project because oil from different production regimes can have different compositions (i.e. different amounts of polar phenolic compounds). Extraction was performed with a 10% NaOH solution to separate out different compounds from the original shale oil samples. The samples were mixed several times with the NaOH solution in separation funnels to remove most of the polar compounds in the oil. To remove NaOH from the oil the samples were then washed multiple times with water until the water removed had a pH of 7. The oil was then dried in a vacuum rotary evaporator at about 80 °C. More details about the extraction procedure are given in Article 3 [52] and by Kogerman [53]. Using this procedure samples were obtained that had lower and higher contents of phenolic compounds than the fuel oil itself (respectively, dephenolated and phenol rich samples).

After extraction, some of the samples were also remixed with the original oil to give a composition that was between that of the extracted samples and the original oil. Some of these modified samples were then also separated into narrow fractions using distillation. Some of the samples were already narrow boiling fractions, and therefore, distillation was not performed for these samples.

Measurement methods used

Table 5-1 gives the method used to measure each of the properties included in the database. Most of the properties were measured using commercial devices or according to ASTM standards, and for the remaining properties references are included that give more details about the measurement method. More information about each measurement method is also given later in this section.

Table 5-1. Measurement methods used for the properties measured for the shale oil database.

Property	Method	Device	Estimated standard uncertainty	Estimated expanded uncertainty (95% level)	Ref.
Density	Oscillating tube	Anton Paar DMA 5000M	0.00015 g/cm3	0.0003 g/cm3	
Refractive Index (n_D)	Abbe refractometer	Anton Paar Abbemat HT refractometer	0.0011	0.0021	
Average boiling point	Thermogravimetric analysis	Modified Du Pont 951 thermogravimetric analyzer	2.1 °C	4.3 °C	[54]
Molecular weight	Cryoscopy or vapor pressure osmometry	Cryoscopy setup or Knauer K-7000	7 g/mol	14 g/mol	[55], [56]
Hydrogen content	Combustion analysis	Exeter CE-440 element analyzer	0.084 wt%	0.17 wt%	
Carbon content	Combustion analysis	Exeter CE-440 element analyzer	0.36 wt%	0.72 wt%	
Sulfur content	X-ray fluorescence (ASTM D4294)	Lab-X 3500 Benchtop XRF Analyser	0.032 wt%	0.063 wt%	
Hydroxyl content	Acetylation reaction and titration	Common chemistry glassware and titration setup	0.19 wt%	0.40 wt%	[52], [57]
Pour point	ASTM D97	Common pour point setup, but with smaller sample vial	3 °C	6 °C	[58]
Viscosity	Oscillating piston method (ASTM D7483) or rotational rheometry	VISCOlab 3000 viscometer or Bohlin Gemini 2 rheometer	VISCOlab 3000: 6.4%; Gemini 2: roughly 10%	VISCOlab 3000: 10.8%; Gemini 2: roughly 20%	
Infrared spectrum	ATR-FTIR	Interspec 301-X mid-IR spectrometer	0.00053 absorbance units (on average)	0.0012 absorbance units (on average)	

Density

Density was measured using an oscillating tube density meter (DMA 5000M, Anton Paar GmbH, Graz, Austria). Before and after each set of measurements the performance of the device was checked using distilled water and air, and the deviation between the measured and reference values was ensured to be less than

0.00005 g/cm^3 before starting measurements (usually it was about 0.00001 g/cm^3).

However, oil samples are more difficult to measure than water or air. The oil samples were generally quite viscous, and with the very viscous heavy fractions it is possible that bubbles remain in the sample. To avoid this the DMA 5000M in our laboratory is equipped with a heating attachment that heats the sample at the inlet to the device to keep the sample hot and the viscosity low. Also, the heavy samples were heated before loading them into the syringe for measurements. Once they were in the syringe, they were again heated to lower the viscosity and allow any bubbles to be removed. This heated sample was then injected into the heated density meter to avoid bubble formation. Some samples were remeasured when there was a suspicion that the initial measurements were inaccurate, and a few densities were adjusted accordingly. Therefore, the uncertainties for the heavy samples may be slightly larger than those of the lighter fractions.

To assess the uncertainty of the density measurements repeat measurements were made with oil samples. Based on these measurements the standard uncertainty was roughly estimated to be 0.00015 g/cm^3.

Densities were measured between 15.6 and 90 °C, and the thermal expansion coefficient and specific gravity at 20 °C for each sample was calculated from the experimental data.

Refractive index

Using an Abbemat HT refractometer (Anton Paar GmbH, Graz, Austria) the refractive index was measured at 589.592 nm. The values reported were measured at 20 °C. The performance of the device was checked every day before and after measurements using distilled water, and the difference between the measured and reference values was kept below 0.00002. To assess the uncertainty for oil samples, repeat measurements were made with oil samples to estimate the reproducibility. From these measurements the standard uncertainty was estimated to be 0.0011 (expanded uncertainty of 0.0021 at the 95% level).

Average boiling point

The method used for measuring the average boiling points of the samples was developed in our laboratory and is described in detail by Rannaveski et al. [54]. With this method the sample is placed in a small metal pan that is then closed with a lid. The lid has a small pinhole in the top. The sample is then heated in a TGA, and the compounds in the sample vaporize and escape through the pinhole. This mass loss data gives information about the boiling points of different portions of the sample, and from this data the average boiling point can be calculated.

The accuracy of this method was checked by measuring oil samples that had been distilled according to the ASTM D2892 standard [59]. The ASTM D2892 standard distillation allows the boiling point of each fraction to be calculated. These standard values were then compared to the results from the TGA method, and from this data the standard uncertainty of the method was calculated to be 2.1 °C (expanded uncertainty of 4.3 °C at the 95% level).

This method could not reliably be used for samples with very high boiling points because at higher temperatures the compounds in shale oil start to decompose. Therefore, only samples with average boiling points below 400 °C were used for this study.

Molecular weight

The number average molecular weight was measured using two different methods: vapor pressure osmometry (Knauer K-7000, Wissenschaftliche Gerätebau Dr. Ing. Herbert KNAUER GmbH, Germany) and cryoscopy (device built in house, method described in ASTM D2224 standard [55]). Benzene was used as the solvent for both methods. Two different methods were used because it was found that vapor pressure osmometry gave better results for samples with higher molecular weights and cryoscopy performed better for low molecular weight samples. This was determined by performing repeated measurements with both methods for a large portion of the shale oil samples.

For both methods calibration was performed using solutions of benzyl with known concentrations. Standard uncertainties were calculated based on both the accuracy of the calibration and tests with pure compounds. This estimation was described in more detail by Järvik and Oja [56]. The relative expanded uncertainty (at the 95% level) was determined to be between ±6 and ±7%. The uncertainty was smaller for fractions with lower molecular weights and larger for heavier samples. When taken on average, the absolute standard uncertainty was 7 g/mol (absolute expanded uncertainty of 14 g/mol at the 95% level).

Hydrogen content

A CE-440 elemental analyzer (Exeter Analytical Inc., North Chelmsford, MA, USA) was used to determine the hydrogen contents of the samples. The estimated standard uncertainty of the hydrogen content was 0.084 wt% (estimated expanded uncertainty of 0.17 wt% at the 95% level). This was determined from measurements made with acetanilide, cyclohexanone, resorcinol and pyrene. The performance of the device was checked with acetanilide before each set of measurements.

Carbon content

Carbon content was also determined using the CE-440 elemental analyzer (Exeter Analytical Inc., North Chelmsford, MA, USA). The estimated standard uncertainty of the carbon content was 0.36 wt% (estimated expanded uncertainty of 0.72 wt% at the 95% level). Again, the uncertainty was estimated by measuring acetanilide, cyclohexanone, resorcinol and pyrene, and performance checks were completed using acetanilide.

Sulfur content

Sulfur content was measured using a Lab-X 3500 Benchtop XRF Analyser (Oxford Instruments, Abingdon, United Kingdom), which uses the ASTM D4294 method. The standard deviation of parallel measurements was 0.0024 wt%. The ASTM D4294 standard gave a higher reproducibility though, which can be calculated using Equation 6 in the standard. Using this equation the expanded uncertainty at the 95% level was calculated to be 0.063 wt%, and the standard uncertainty was then estimated to be 0.032 wt% (half of the expanded uncertainty).

Hydroxyl content

Hydroxyl group content was measured using a titration method developed by Aarna and Paluoja [57], which has been used with shale oils for decades. This method involves acetylation of the hydroxyl groups using an excess of acetic anhydride. The amount of acetic anhydride remaining is determined by titration with KOH. A blank sample is also used, and the amount of hydroxyl groups is calculated from the difference in acetic anhydride between the blank sample and oil sample. This method also measures primary and secondary alcohols and that organic acids also affect the titration, but the majority of hydroxyl groups in Kukersite shale oil are phenols and there are not significant quantities of organic acids [60]. More details about the method are given in Article 3 [52]. The standard uncertainty of this method was evaluated by taking the standard deviation of parallel measurements, and was calculated to be 0.19 wt%OH (expanded uncertainty of 0.40 wt% at the 95% level). One measurement was also made using a pure compound (4-nonylphenol) and the deviation for this measurement was 0.17 wt%.

Pour point

For measuring the pour point the ASTM D97 standard [58] was used. Because only a small amount of each sample was available, the standard was modified by using smaller vials that had internal diameters of either 1.6 or 2.5 mm. The sample was kept at approximately 10 to 20 mm. The standard also states that the sample vial should be placed in a metal cylinder, not directly in the cooling bath. However, we found that this configuration gave a slow cooling rate and a minimum temperature that was significantly higher than that of the cooling bath. Therefore, some of the samples were placed directly in the cooling bath. Measurements were made to confirm that these modifications did not affect the results. First, a shale fuel oil sample was measured in a small vial and in a larger vial comparable to that specified by the standard. The difference in the pour point was only 0.4 °C. Second, two shale fuel oil samples were measured directly in the cooling bath and then in a metal cylinder. When placing directly in the cooling bath the temperature of the sample was continuously monitored, which showed that the cooling rate was still slow enough to enable good temperature resolution. The pour point was 0.6 and 1.6 °C lower when placed directly in the cooling bath. Note that all of these differences are smaller than the 3 °C step size specified in the ASTM D97 standard, and can therefore be considered insignificant.

Repeat measurements were not made, but the standard states that the reproducibility of this method is 6 °C (that is, only 1 measurement in 20 would deviate by more than 6 °C). This was take to be the expanded uncertainty of the pour point measurements at the 95% level. Half of this, or 3 °C, was take to be the standard uncertainty.

Viscosity

Two different instruments were used for measuring viscosity: a VISCOlab 3000 viscometer (Cambridge Viscosity, PAC L.P., Houston, Texas, USA) and a Bohlin Gemini 2 rotational rheometer (Malvern Instruments Ltd, Malvern, United Kingdom). The VISCOlab 3000 uses the oscillating piston method [61]. Measurements with the rheometer were performed using a stainless steel cone and plate geometry (either a 40 mm/4° cone or a 25 mm/2.5° cone). Viscosity standards were used to check the performance of instruments. From these measurements the standard uncertainty of the VISCOlab 3000 was found to be 6.4%. At 37.8 °C the relative deviations with the Gemini 2 were 2%, 2% and 4% for the three standards. At higher temperatures fluctuations in the temperature occurred that affected the accuracy of the Gemini 2. We estimate that, on average, the standard uncertainty at higher temperatures is about 10 to 15%. Some of the samples were measured on both instruments, and the measured viscosities generally agreed to within 10%. 69 of the samples were measured at multiple temperatures between 17 and 151 °C.

Infrared measurement procedure

Because the infrared spectra are a fundamental part of this work, a more detailed description of the measurement procedure is warranted. The device used was an Interspec 301-X portable spectrometer (Interspectrum OÜ, Tõravere, Estonia). It measured in the mid-IR region (wavelength range from 400 to 7000 cm^{-1}). It is a Fourier transform infrared (FTIR) spectrometer, and for measuring the oil samples it was fitted with a single bounce, ZnSe, attenuated total reflection (ATR) measurement accessory. The angle of incidence was 45 degrees. Also, because of the design of the instrument, the infrared beam passed through the open laboratory air. Therefore, some additional noise occurred in the spectra in the absorption regions of water vapor in the air.

Measurements were taken at a resolution of 1 cm^{-1} over the range 600-4000 cm^{-1}. A cosine apodization was used (more specifically, $\cos(0.5\cdot\pi\cdot x)\cdot(\cos(0.5\cdot\pi\cdot x))^2$). The recorded spectrum was taken as the average of 10 scans. The device was switched on at least a couple of hours before beginning measurements to allow the temperature of the IR source to stabilize. At the beginning of a set of measurements, the spectrum of distilled water was measured to check the performance of the device. This ensured that the ATR crystal was clean and that the response of the spectrometer had not changed. Between samples the crystal was cleaned using toluene and isopropyl alcohol.

Then a spectrum was recorded with no sample to ensure that the crystal was clean.

The spectra were preprocessed by removing the data below 700 cm^{-1}, removing the carbon dioxide peaks between 2300 and 2400 cm^{-1} and applying a baseline correction. The baseline correction was performed by fitting a 3rd order polynomial to the points in the ranges 2000-2200 cm^{-1} and 3700-4000 cm^{-1}. Also, when comparing a measured spectrum for toluene with a spectrum calculated from standard optical constant spectra [62], it was found that the wavenumbers given by the Interspec 301-X spectrometer were slightly shifted. The magnitude of the shift was determined by comparing the peak locations from the Interspec spectrum with those of the standard spectrum. Then, a linear line was fit to the data (Equation 5-1) and this equation was used to correct the wavenumbers of the spectra.

$$\nu_{corrected} = -0.0005950127533633 \cdot \nu + 0.009786571607383 \qquad (5\text{-}1)$$

To estimate the uncertainty of the infrared spectra, 10 parallel measurements were made with an oil fraction. These were preprocessed using the same method used for all the other spectra. The standard deviation of the measurements at each wavelength was taken to be the standard uncertainty of the measurements, and the expanded uncertainty at the 95% level was calculated from this. To make these uncertainties more understandable, they are expressed as relative uncertainties. The relative uncertainties were calculated in reference to the average oil spectrum (the mean of all measured oil spectra), and these relative uncertainties are shown in Figure 5-5. The median standard uncertainty was 1.8%. When looking at Figure 5-5 it can be seen that regions with lower values on the average spectrum generally have higher relative uncertainties, as would be expected. However, the largest errors can be seen in the water vapor absorption regions, especially between 1400 and 1700 cm^{-1} and above 3100 cm^{-1}.

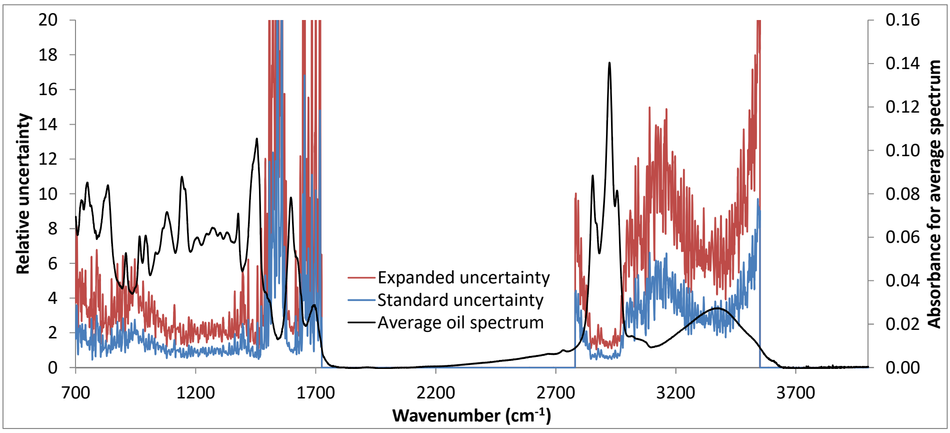

Figure 5-5. Relative uncertainties of the infrared spectra at each wavenumber. Expanded uncertainties were calculated at the 95% level.

6. INFRARED SPECTRA

Some general molecular structures present in the oil can be observed from infrared spectra. These relationships between the spectra and the bonds and functional groups contained in the samples help to explain how oil properties can be predicted from infrared spectra.

Figure 6-6 and Figure 6-7 show spectra of a whole fuel oil sample and a whole shale gasoline sample, and also point out some of the features that can be linked to specific functional groups or bonds. These group assignments were mainly based on the group frequencies given by Coates [28]. Notice the absence of the features related to hydroxyl groups in the shale gasoline spectrum.

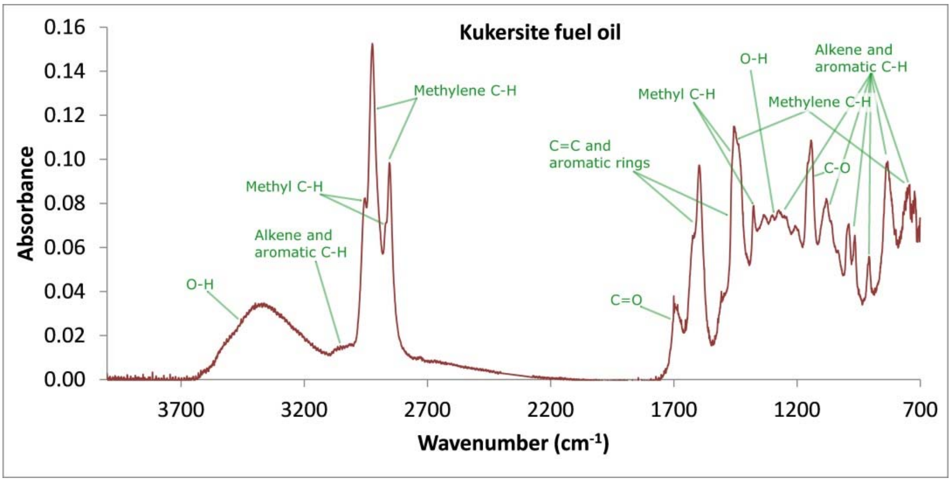

Figure 6-6. Infrared spectrum of fuel oil from the Enefit 140 solid heat carrier retort.

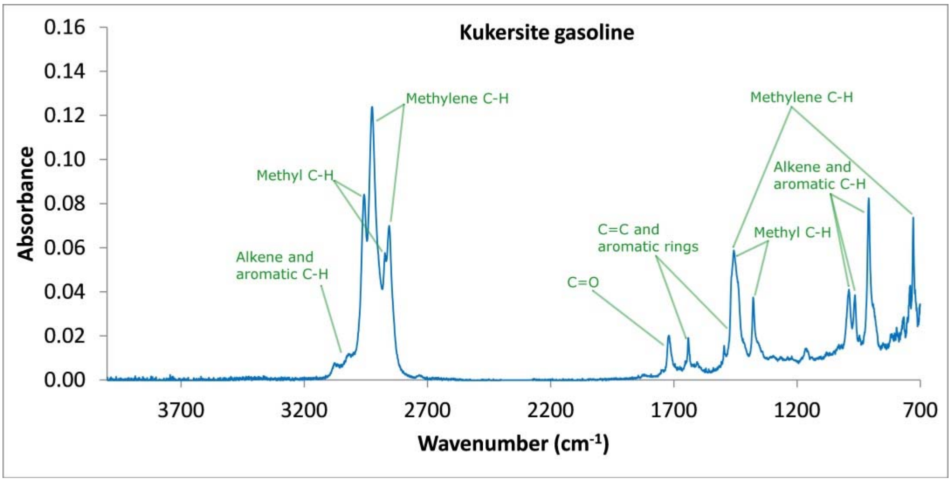

Figure 6-7. Infrared spectrum of shale gasoline from a solid heat carrier retort.

The changes with boiling point and hydroxyl group content can better be seen from Figure 6-8, Figure 6-9, Figure 6-10 and Figure 6-15. The figures showing the effect of boiling point were created by plotting the fractions from one gasoline distillation and one fuel oil distillation. The trends are hard to completely grasp at first glance, and looking at the zoomed in views of Figure 6-9 and Figure 6-10 helps. At first, absorbance in most regions of the spectrum increases with boiling point. Once the fuel oil region is reached, at about 200 °C, then peaks caused by hydroxyl groups become prominent. At about 320 °C then the trend changes. The peaks related to hydroxyl groups and aromatic structures start to decrease, and at the same time the peaks due to aliphatic groups (saturated C-H bonds) continue to increase. This shift occurs because in the heavier fractions the size of the molecules increases. Although most molecules still contain a hydroxyl group, their relative proportion in the mixture begins to decrease. Instead, the straight aliphatic side chains attached to the aromatic core increase in length, which leads to increases in the spectral areas related to methyl and methylene C-H groups.

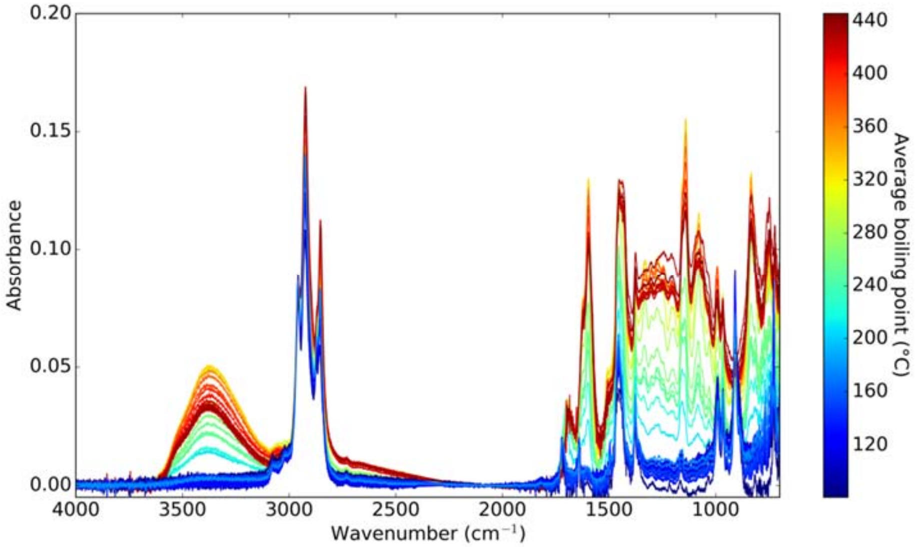

Figure 6-8. Changes in the infrared spectra of shale oil fractions with progressively higher boiling points.

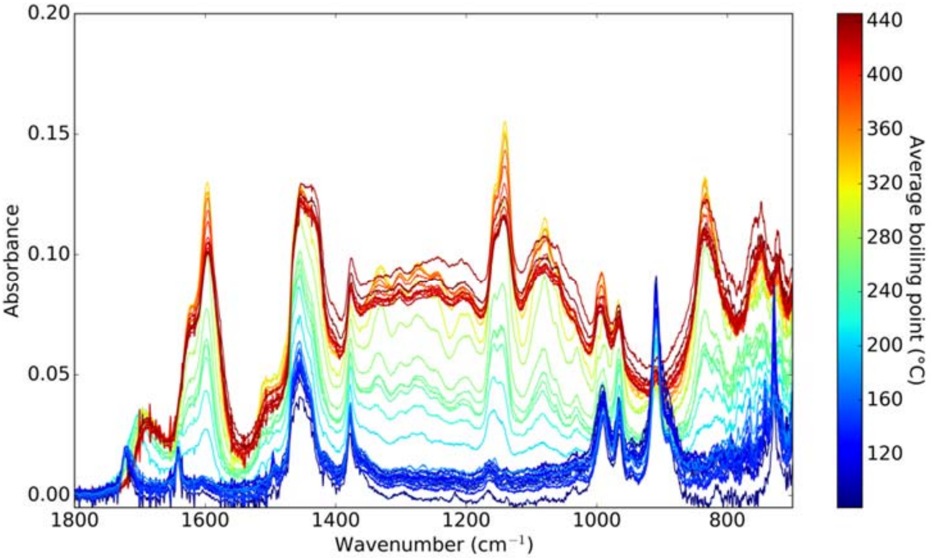

Figure 6-9. Changes in the infrared spectra of shale oil fractions over the range of 700 to 1800 cm^{-1}.

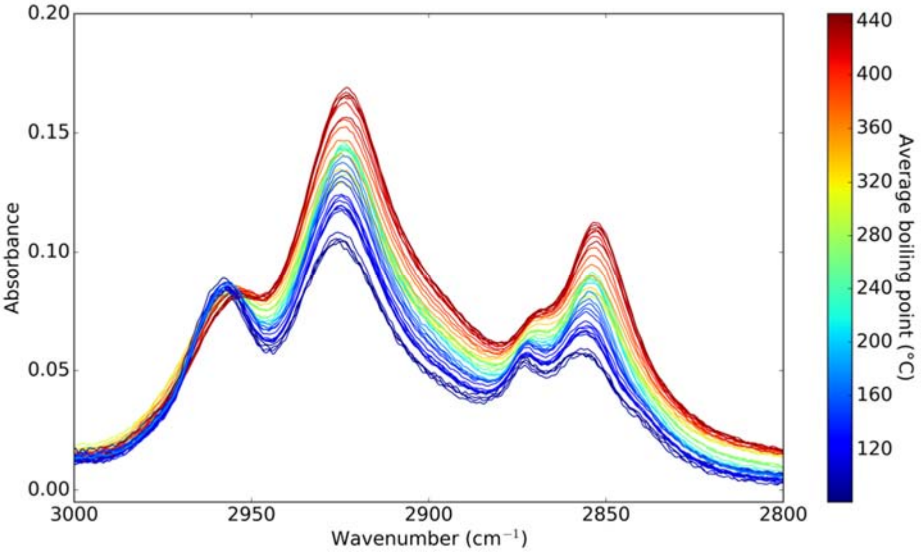

Figure 6-10. Changes in the infrared spectra of shale oil fractions over the range of 2800 to 3000 cm^{-1}.

To better illustrate how peaks in the spectra change as the boiling point of a narrow fraction increases, the absorbance at 4 different wavenumbers was plotted versus the average boiling point of the sample (Figure 6-11, Figure 6-12, Figure 6-13 and Figure 6-14). The figures only show similar samples obtained using simple (Engler) distillation to emphasize the trend that generally occurs among fractions from a single distillation. Although the absorbance values cannot directly be equated with the concentration of the given functional group, the plots do show how systematic structural changes between the fractions can be observed on infrared spectra.

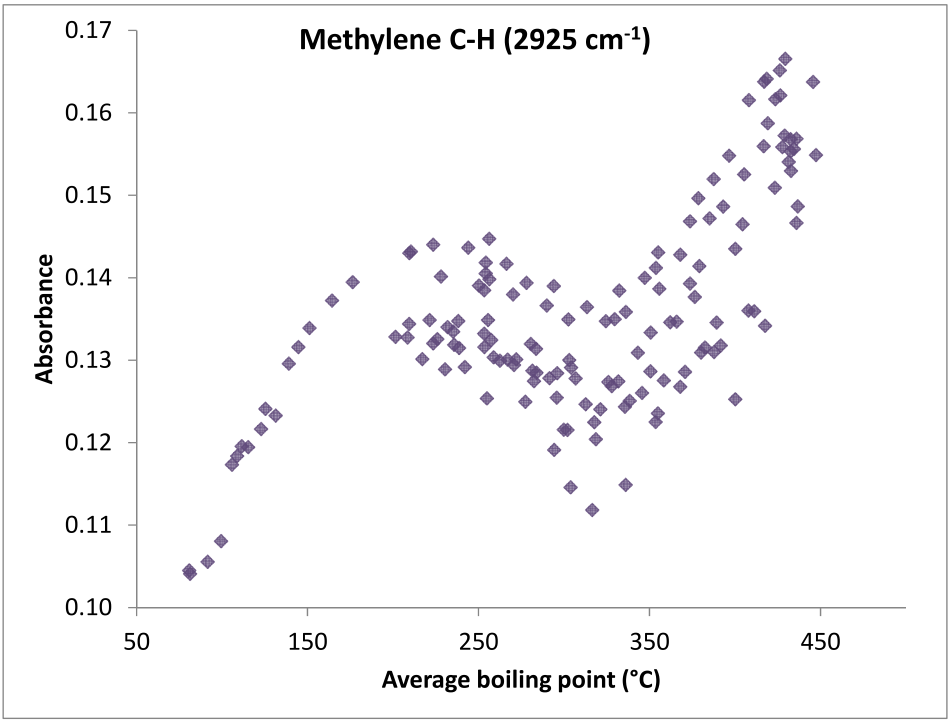

Figure 6-11. How absorbance related to methylene C-H bonds changes as a function of boiling point.

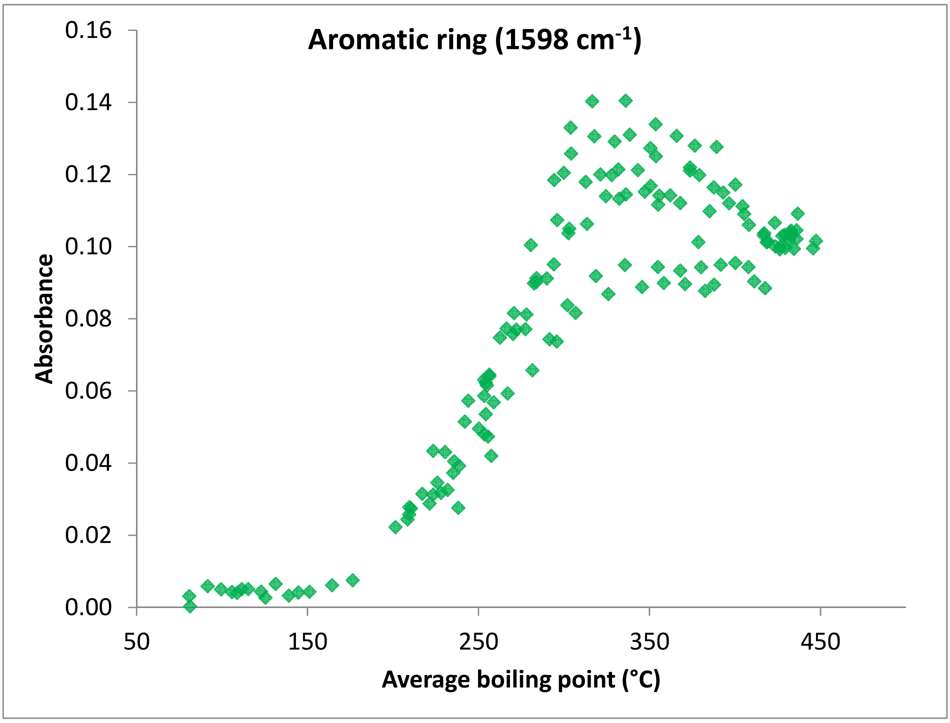

Figure 6-12. How absorbance related to aromatic rings changes as a function of boiling point.

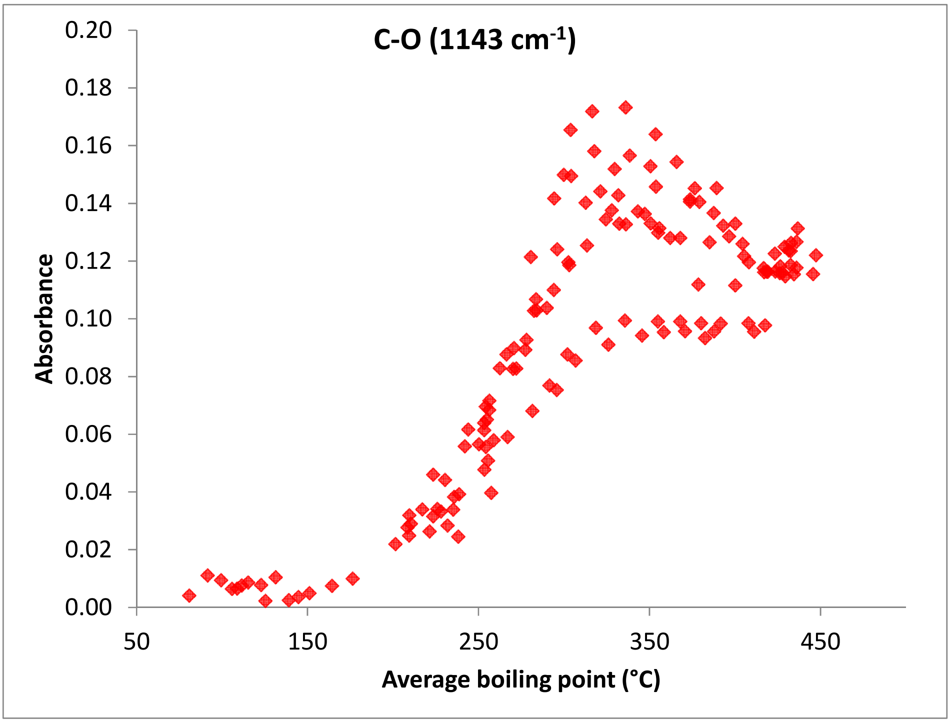

Figure 6-13. How absorbance related to C-O bonds changes as a function of boiling point.

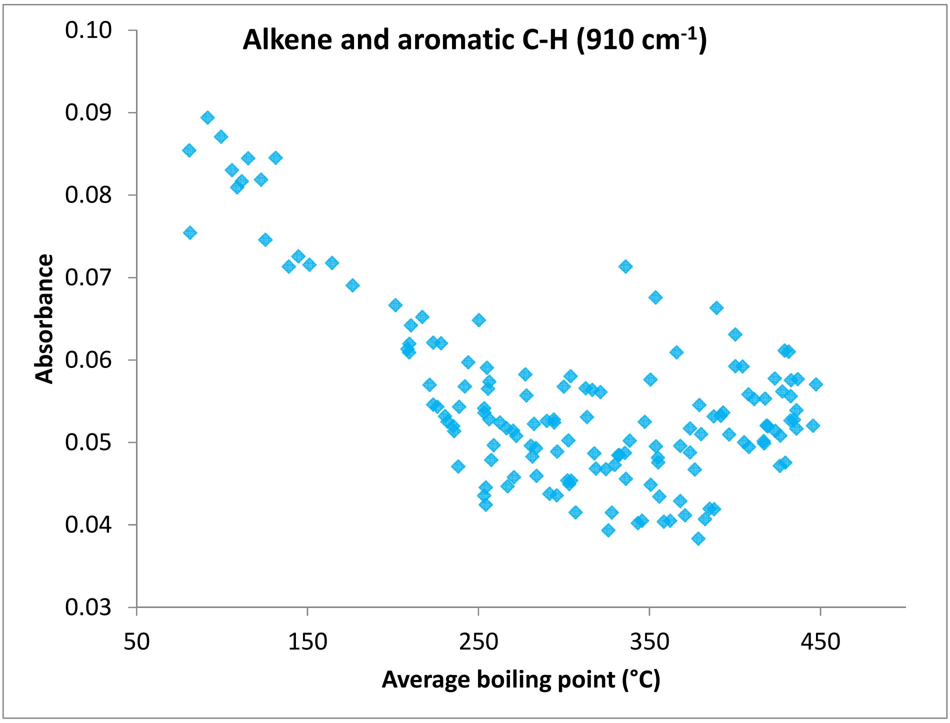

Figure 6-14. How absorbance related to alkene and aromatic C-H bonds changes as a function of boiling point.

Although trends can be seen, these plots of absorbance versus boiling point all show large amounts of scatter. Some of this is caused because peaks overlap. For instance, the height of the C-O stretch peak is also influenced by other peaks around it that are related to functional groups such as O-H and C-H bonds. Noise and uncertainty in the measurements also add to the scatter. But beyond these reasons, samples with similar boiling points often do have significantly different chemical compositions. Natural variation occurs in the products from the plant, which then carries through to the fractions. One significant example is variations in the amount of phenolic compounds, and these differences in phenolic concentration are a main cause of the large scatter seen for samples with boiling points between about 250 and 400 °C.

The effect caused by variation in the amount of phenolic compounds was isolated by plotting 7 samples with essentially the same average boiling points (boiling points between 310 and 320 °C). This plot is shown as Figure 6-15. The effect is quite pronounced. The fraction that had the lowest hydroxyl group content is completely separate from the other fractions. Even among the other fractions significant differences can be noticed. This emphasizes the fact that infrared spectra measure functional groups, and changes in the structural composition are much easier to detect than changes in the size of molecules in the fuel. Indeed, the spectra of the heaviest fractions are quite similar because the composition of functional groups in the fuel remains largely the same, even though the average boiling points range over about 100 °C.

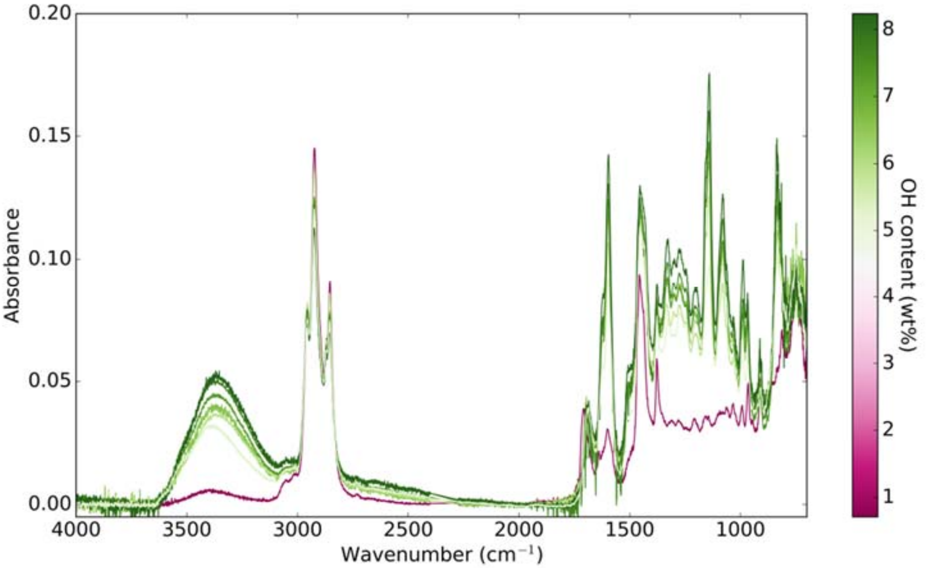

Figure 6-15. Changes in the infrared spectra caused by changes in the quantity of phenolic compounds. The fractions displayed have essentially the same average boiling point (between 310 and 320 °C).

The systematic changes shown here that accompany changes in the boiling point of the samples are likely what give the information necessary for predicting molecular size properties since infrared spectra do not directly contain information about molecular size. However, because the spectra of heavier fractions are quite similar, this may make it more difficult to predict the properties of these samples. This problem could also occur if samples were prepared in a different manner, for instance if a light and heavy fraction were mixed the resulting spectrum might resemble a middle range fraction. This spectrum might still give good results for some average properties, but for properties for which a linear mixing rule does not apply (e.g. viscosity or vapor pressure) there may not be enough information in the spectrum to enable a good prediction.

7. MODEL CREATION PROCESS

A predictive model is always limited by the quality of the data used for constructing the model and the strength of the relationship between the input and predicted properties. However, inaccuracies in the model itself also affect the model performance, and therefore, when creating a model the goal is to do so in such a way that the model is as true to the underlying relationship as possible. This chapter discusses some of the aspects of model creation that need to be taken into account to get performance that is close to the limit determined by the data and relationship. The specific methods used for this study are also given. First though, the parameters used in this study to estimate the error of a model are presented.

Error estimation

As mentioned, for regression when the number of variables is larger than the number of observations the error of the model must be estimated using separate test observations that were not used in the regression. This is because it is easy for a model to over fit the data, which would result in great accuracy with the regression set, but would actually cause the model to give large errors when predicting values for new observations.

The method used here for error estimation was cross validation, which consists of dividing the samples into groups and sequentially leaving one group out of the regression while calculating the model. The deviation for each sample when it is left out is then used to estimate the error of the model. Commonly, the RMSE of these cross validation values is then presented as the error of the model.

As noted, the regression parameters were also optimized using the cross validation error. To ensure that the error estimate was not influenced by the optimization process, a second outer cross validation loop was used for estimating the error. So all together, the regression was performed by dividing the data set into 15 groups for the outer 15 fold cross validation loop, and then leaving each group out once, a model was found by minimizing an inner 5 fold cross validation of the included samples. The error was then estimated using the value predicted for each sample when it was left out as part of the 15 fold outer loop. After the error had been estimated in this manner, the final regression model was found using all of the samples in the data set.

Four error statistics were calculated from the cross validation values to help characterize the performance of each model: root mean squared error (RMSE), average absolute deviation (AAD), average absolute relative deviation (%AAD) and the Pearson correlation coefficient squared (R^2). These were calculated using Equations 0-1 through 0-4.

$$RMSE = \sqrt{\frac{\left(\theta_{pred}-\theta_{actual}\right)^2}{n}} \tag{0-1}$$

$$AAD = \frac{\left|\theta_{pred}-\theta_{actual}\right|}{n} \tag{0-2}$$

$$\%AAD = \frac{\frac{\left|\theta_{pred}-\theta_{actual}\right|}{\theta_{actual}} \cdot 100}{n} \tag{0-3}$$

$$R^2 = \left(\frac{\Sigma\left(\theta_{actual}-\theta'_{actual}\right)\left(\theta_{pred}-\theta'_{pred}\right)}{\sqrt{\Sigma\left(\theta_{actual}-\theta'_{actual}\right)^2 \Sigma\left(\theta_{pred}-\theta'_{pred}\right)^2}} \right)^2 \tag{0-4}$$

In Equations 0-1 through 0-4 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.

The distribution of residuals obtained through cross validation was also used to estimate the 95% prediction interval for each model. To do so, the standard deviation of the residuals was multiplied by the t-statistic. For machine learning regression methods the degrees of freedom is generally not equal to the number of model parameters [63], and instead it must be estimated. The simple solution used here was to fit the t distribution to the residuals by adjusting the effective degrees of freedom. This procedure gave reasonable values for the degrees of freedom for most of the properties. However, for a few of the parameters with small sample sizes the fitted value was unrealistically large, and instead a conservative value of 5 was used for those properties.

Outlier removal

Because the quality of the reference database affects the accuracy of the resulting model, it is important to check the data. A large amount of research has been done on outlier detection, and here we do not attempt to make any summary or comparison of the many different methods. Rather, we describe the procedure we have used.

Some outliers can be easily detected as having an extreme value for one of the data variables, and we started by investigating samples that had extreme property values. As an example, some of the extracted samples had densities that were much higher than all the other samples, and a more accurate model was obtained when these samples were left out.

It is more difficult to detect outliers that are not extremes for any one variable, but have a combination of variable values that does not fall in line with the rest of the samples. For investigating samples based on combinations of variables, we calculated the pairwise Euclidean distances between each of the samples based on the input data. Then we examined the distributions of the distances and selected those that had extreme mean distances for further investigation. We also viewed samples based on the value for 3% of the cumulative distribution, that is the distance which was larger than 3% of the pairwise distances for the sample. This enabled us to investigate how well the sample fit with other samples that were similar to it, rather than with all the data as a whole.

The residuals of the model also provide a good way to detect outliers. Those samples with large residuals often are different from the main group of samples, and those differences keep the model from being able to describe these outlying samples.

Those samples that had been identified as outliers were scrutinized to determine if they should be left out. If needed, some samples were remeasured to determine whether or not a measurement error had occurred. 16 samples were completely removed from the database because their infrared spectra contained significant noise. All but one of these samples had been measured at the very start of the project before the infrared measurement procedure had been completely worked out. 10 of the extracted samples were only used for the hydroxyl content model, and were left out of all other models. These samples had a significantly different chemical composition due to the way they were prepared, and this resulted in infrared spectra that stood out from the rest of the samples. Additionally, they had extreme values for many of the properties. These abnormalities limited the accuracy of most of the models. But because they had such large amounts of phenolic compounds, they actually improved the performance of the model for hydroxyl content. Besides these general removals, additional samples were only removed as outliers if it was clear there was an error with the measured property value.

A few more outlying samples were also identified when comparing the residuals for different properties for each sample. For instance, the two samples with the largest residuals for average boiling point also had some of the largest residuals for many of the other properties. Consistently large residuals across different models indicates that there was a problem in the spectra or preparation done for these samples. However, because no anomalies could be detected in the spectra, this assumption could not be confirmed, and these samples were not removed from the database. In total, 8 of these samples were identified, most from different distillations. Given that there were several hundred samples in the database, these 8 outliers constitute only about 2 to 3% of all samples.

Sample selection

A related step is selection of the calibration samples. For best results, the calibration samples often need to be well spaced throughout the range of sample variation for which the model is expected to work [26], [64], [65]. Chung [26] points out that often problems arise when the full range of expected variations is not taken into account, and further states that it is usually not useful to include a disproportionally large number of samples that have similar compositions. Circumstances do not always allow such an optimal selection of calibration samples, as was the case with the research done for this PhD dissertation. Weighting certain samples in the regression process to even out their relative effect may also help correct for an uneven distribution. For the research done here, the sample preparation process was done as part of a larger research project, and therefore, the samples were not created with the goal of ensuring an even distribution among composition and properties.

To determine how many samples are needed to obtain an accurate model, a test was run using specific gravity and average boiling point, which had been measured for most of the samples. First, one-third of the samples were separated out to be used as the test set for calculating the accuracy of the model. Then, some of the remaining samples were randomly selected to be used as the calibration set. The number of samples used was varied. Then a support vector regression model (using a mixed kernel [43]) was created and the RMSE of the model was calculated using the test set. This was repeated multiple times, and the samples used in the test set were also changed to avoid any bias that may have resulted due to the samples chosen for the test set. The results of this test are shown in Figure 7-16.

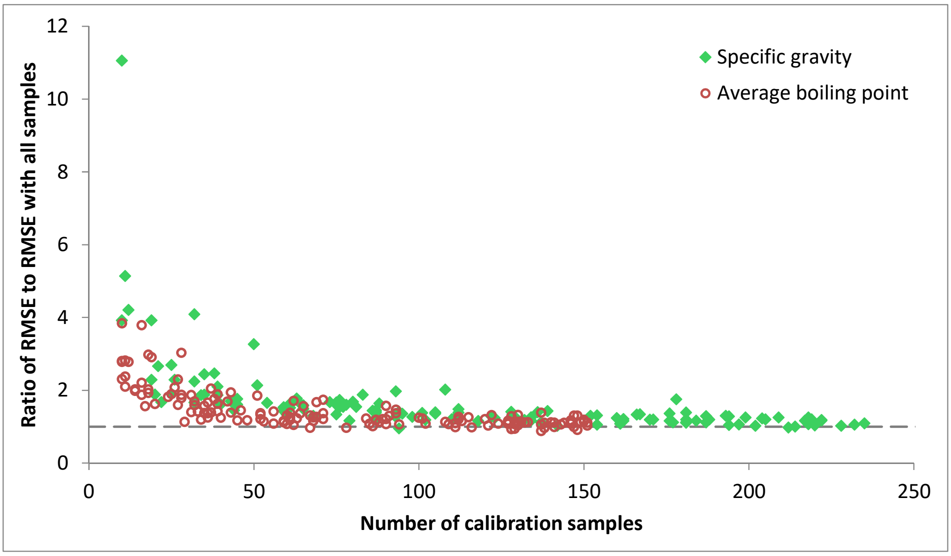

Figure 7-16. The RMSE of a support vector regression model as a function of the number of calibration samples used. The orange line represents the RMSE obtained when using all the samples.

As Figure 7-16 shows, having too few samples in the calibration set clearly affects the accuracy of the model. However, when using about 40 samples the error can already be at the same order of magnitude as that of a model using all the samples (about 300 samples). Using more than about 100 samples makes essentially no difference. It should be noted that these numbers are specific to this problem, and the exact effect of sample size likely depends on the given dataset.

The scatter seen in Figure 7-16 shows that other factors also influence the RMSE of the model. One such factor is which samples are included in the calibration set. For example, if the calibration set did not contain a certain type of sample (e.g. gasoline samples or heavy samples), then the resulting RMSE of the model would be larger than for a calibration set in which the full range of samples was well represented. This helps to explain the difference between models that used the same number of calibration samples. What can also be seen is that the spread of the scatter decreases as more calibration samples are used. This leads to a significant conclusion: if only a smaller number of calibration samples is going to be measured, then it is important to pay attention to which types of samples are included. When a large number of calibration samples can be measured it may not be so important to make sure that the calibration set is well balanced because there is a greater chance that all the different types of samples will be well represented, at least as long as the calibration samples span the range of samples expected.

Therefore, how the samples are prepared is important. Although 40 samples could be quickly made by performing two distillations of the same wide industrial fraction, the model would be more accurate if the 40 samples were taken from a variety of industrial fractions. This is because variation naturally occurs in raw materials and processing conditions. Using samples gathered over a longer time would allow the model to take these variations into account. To test how many distillations might be needed to get good results, models were created using only some of the distillations. Table 7-2 and Table 7-3 show the results for specific gravity and average boiling point. The error statistics given in the tables only take into account the fuel oil and gasoline samples because the extracted samples have compositions that fall outside of the range that could be expected in the plant. When using 5 distillations the accuracy of the model was close to that of the final model that included all samples. Performing 5 distillations would require a significant amount of work, but based on Figure 7-16 not all of the fractions would need to be measured.

Table 7-2. How the accuracy of the model for specific gravity is affected by the number of different distillations the calibration samples are taken from. Each cell contains the RMSE and the 95th percentile (given in italics) of the residuals for the fuel oil and gasoline samples.

		Number of distillations with extracted samples			
		0	**1**	**2**	**6**
Number of distillations	**3**	0.00741 *0.01379*	0.01035 *0.01936*		
	5	0.00487 *0.00987*		0.00568 *0.01175*	
	7	0.00401 *0.00871*	0.00412 *0.00826*	0.00454 *0.00948*	
	22				0.00364 *0.00758*

Table 7-3. How the accuracy of the model for average boiling point is affected by the number of different distillations the calibration samples are taken from. Each cell contains the RMSE and the 95th percentile (given in italics) of the residuals for the fuel oil and gasoline (units are kelvin).

		Number of distillations with extracted samples			
Number of distillations		0	1	2	6
	3	9.4 *19.6*	8.3 *16.7*		
	5	7.7 *16.0*		7.3 *13.6*	
	7	5.8 *10.9*	6.4 *11.6*	7.1 *14.3*	
	19				5.6 *10.9*

As mentioned, extraction had been used to prepare some of the samples with the hope that the wider range of compositions would make predictions more robust towards sample variations that naturally occur in the plant. The effect of using extracted samples was also investigated by using extracted samples in place of some of the fuel oil distillations. However, it was difficult to make conclusions based on the results. A few points showed better results when using extracted samples, a few showed poorer results and many were about the same. This may indicate that there is no advantage in using the extracted samples, but it is also possible that the set of samples in the database did not vary enough. For more extreme samples, that occur less frequently in the plant, including extracted samples may be important for obtaining good results.

So, machine learning techniques are often sensitive to the distribution of calibration samples chosen, and it is best to select samples that cover the full range of expected variations. To get an idea of how the models created here might perform when used beyond the range of the calibration samples, a model for hydroxyl content was created with all the dephenolated and gasoline samples left out. Doing so removed most of the samples with low hydroxyl group contents. The difference between the values predicted using this model and the model based on all the samples is shown in Figure 7-17, and shows a consistent positive bias for samples with low hydroxyl contents. These samples have hydroxyl contents between 0 and 3 wt%, so a consistent positive error of about 0.4 wt% is relatively large. So, extending a model to predict samples with low hydroxyl contents would give larger errors if these types of samples were not included in the calibration set. Removing the dephenolated samples also seems to have affected the results for the phenol rich samples that had the highest concentration of hydroxyl groups. All of these samples were created from the same distillation and extraction as the dephenolated samples, and therefore, the spectra of the dephenolated samples may contain information that is also helpful in predicting values for these phenol rich fractions.

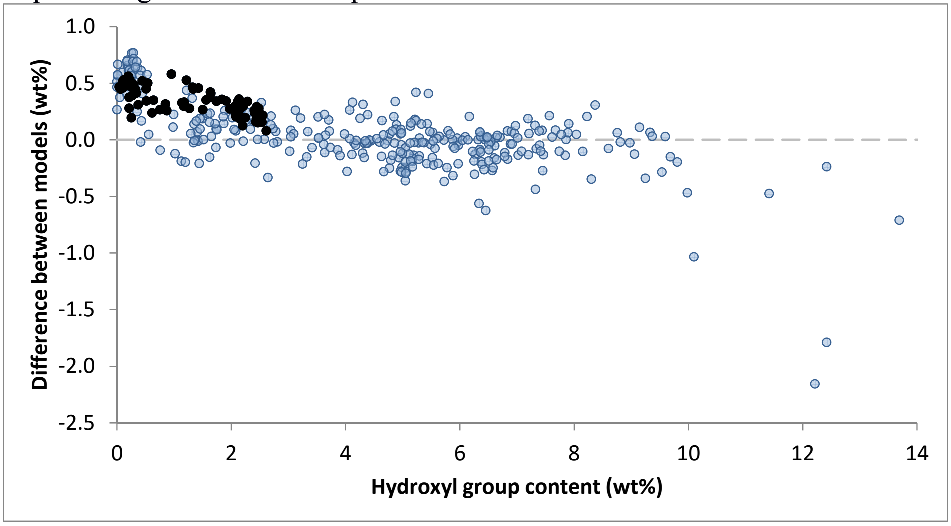

Figure 7-17. Hydroxyl contents predicted using a model that did not include dephenolated samples minus the predictions from one that included all samples. Black points indicate dephenolated samples.

A second, similar test was performed by creating models using only 5 distillations, and using those models to predict 3 other distillations. 1 of the 5 was the gasoline distillation, and the other 4 were fuel oil vacuum distillations (simple batch distillations). The other distillations were prepared using either a rectification column or a two-step batch distillation. For the two-step distillation a first batch distillation was performed to collect only about the first 30 to 40% of the fuel oil, and then this lighter sample was distilled again using batch distillation. Therefore, fractions from these distillations had a somewhat different composition than those from the vacuum distillations. In this way the ability of the model to predict samples created using different distillation methods was assessed.

Models were created for specific gravity and average boiling point, and the results of this test are shown in Figure 7-18 and Figure 7-19. A similar trend emerges, where samples with the lowest specific gravities and boiling points have the largest residuals. From there the residuals decreased and were generally within the RMSE of the models that included all the samples (RMSE of 0.00467 for specific gravity and 6.95 K for boiling point). So, the models did not extrapolate well for those lightest fractions, but they seemed to provide accurate predictions for the heavier samples that were more similar to the calibration samples, even though these samples were produced using a different distillation method.

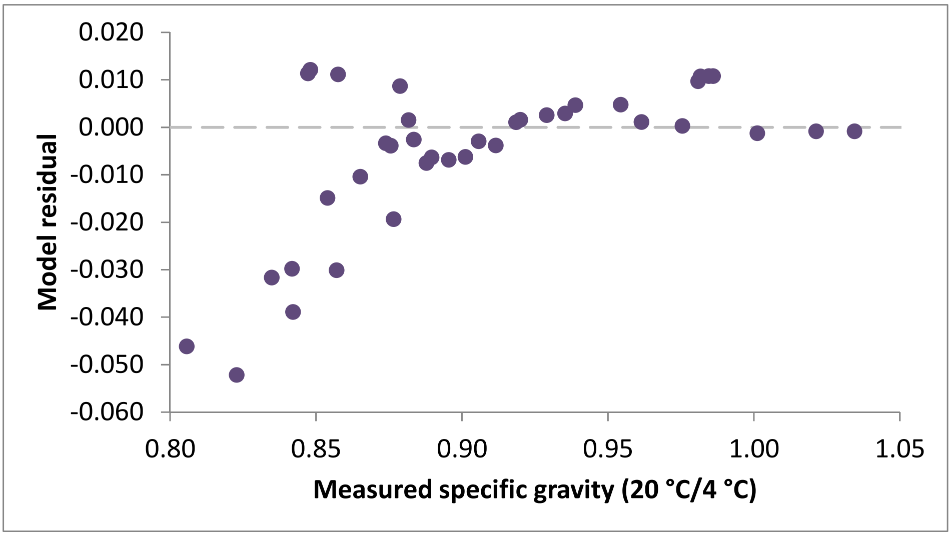

Figure 7-18. Residuals that occurred when using a model based on 5 distillations to predict specific gravities for samples that were produced using a different distillation method.

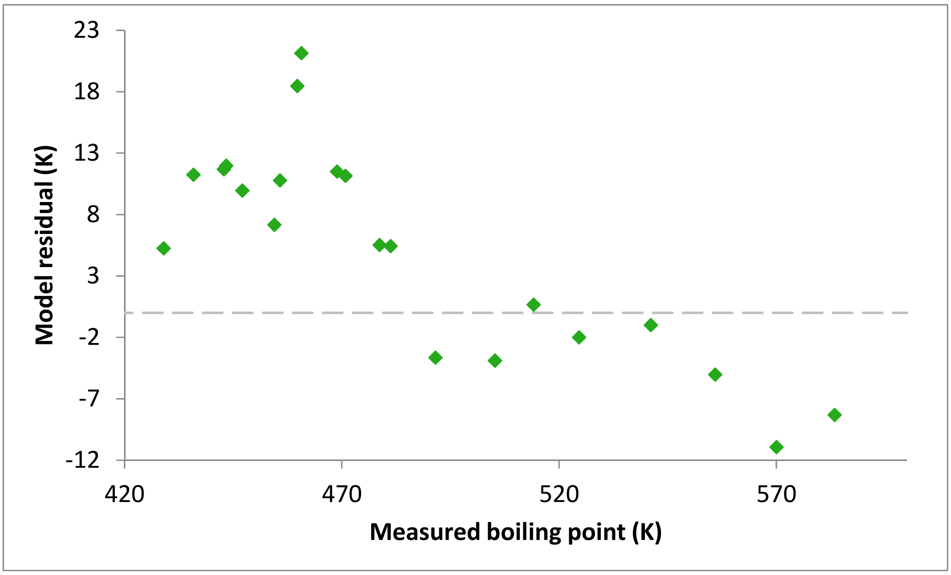

Figure 7-19. Residuals that occurred when using a model based on 5 distillations to predict average boiling points for samples that were produced using a different distillation method.

Because using more samples was seen to increase the chances of obtaining a model with good accuracy, for this study all available samples were used. For some of the properties only a limited number of samples had been measured, which made it even more important to use all the samples. Table 7-4 shows the number of samples used for each property. The "Fuel oil (batch)" category refers to fuel oil fractions obtained using simple batch distillation (Engler distillation or a similar method at low pressure), and the "Fuel oil (other)" category includes samples obtained using a rectification column or two-step batch distillation. "Wide samples" refers to the original wide industrial fractions obtained from the plant and crude shale oil samples.

Table 7-4. Number and type of samples used in creating the model for each property.

	Total	Shale gasoline	Fuel oil (batch)	Fuel oil (other)	Dephenolated	Phenol rich	Wide samples
Specific gravity	355	16	171	38	75	45	10
Refractive index param.	327	14	150	37	74	45	7
Average boiling point	229	16	103	29	52	29	0
Molecular weight	277	16	135	9	71	44	2
Carbon content	257	16	123	18	49	43	8
Hydrogen content	258	16	123	18	49	43	9
Sulfur content	59	16	40	0	0	0	3
Hydroxyl content	57	0	19	9	6	21	2
Pour point	68	0	28	0	19	19	2
Thermal expansion coef.	319	16	138	38	73	45	9
Viscosity at 37.8 °C	115	0	29	2	44	38	2

Variable selection

Large input data sets, such as spectra, can contain hundreds or thousands of variables. Many of these variables can be closely related and contain similar information, that is, they are redundant. Also, there can be variables that do not correlate very strongly, if at all, with the parameter to be predicted. Removing uninformative or redundant variables can make a model simpler. This in turn can make the model more robust to noise and fluctuations, can improve the accuracy and can significantly reduce the computing time needed for calculating the model.

Many different methods for variable selection have been developed. Mehmood et al. [66] gave a review of methods used with partial least squares regression. They categorized the methods into three groups: filter, wrapper and embedded methods. These can be used to describe variable selection methods for other regression techniques as well.

Filter methods rank variables based on a criterion and then remove the variables for which the criterion is below an arbitrary cutoff. Often correlation coefficients are used, such as the Pearson correlation coefficient (R) or some type of mutual information parameter. For example, the correlation coefficient is calculated between each wavelength in a spectrum and the property to be predicted, and then those that do not correlate well enough (as defined by the user as a cutoff value) are removed from the data set. Other common criteria include the weight of the variable in the regression model, the Akaike information criterion and the RMSE of a small group, or window, of the variables (which is termed the moving window method [67]).

Like filter methods, wrapper methods use a criterion to rank and separate variables, but the cutoff is determined by optimization of the error of the model. That is, instead of just assigning a set cutoff, different cutoffs are tried with the goal of finding the cutoff that minimizes the error. These methods usually give better performance than filter methods, but also cost more computationally because the regression must usually be performed many times.

Embedded methods work by incorporating variable selection directly into the regression algorithm. Rather than calculating the whole model before choosing variables to remove, as with wrapper methods, a step is added to the regression algorithm in which certain variables are removed or down weighted. These methods are generally not as universal as other methods because they are closely related to the method algorithm. Also, they are usually more complex to implement because they involve modifying the underlying algorithm of the regression method. However, they can often yield good results in a shorter period of time than wrapper methods.

In addition to all these methods, some variables can be eliminated just by knowledge of the problem at hand. For example, large sections of the spectrum can be removed where infrared absorption does not occur. Also, if only certain peaks are known to be of interest or related to the parameter to be predicted, then the spectrum can be reduced down to that region of interest.

Many different methods were tested during this research, but in the end most of the spectrum was used. Wrapper methods initially seemed to give good results, but they took a long time to carry out and any improvements in accuracy did not appear to be pronounced. This was the case with the searching combination moving window partial least squares method [68] and when using a genetic algorithm for variable selection. With filter methods often large sections of the spectrum were ranked fairly highly, and the best results were generally achieved when using most of the absorbing area of the spectrum. The Joint Mutual Information Maximization method [69] seemed to be one of the more promising filter type methods, but initial tests with it did not yield any improvements in accuracy. So, it was decided to simply use the entire absorbing area (700 – 1800 and 2750 – 3600 cm^{-1}). The one exception was the model for hydroxyl group content. This model was created first and it had been found that a somewhat narrower range (907 – 1464 cm^{-1}) gave a better RMSE of 0.350, as opposed to 0.411 with the full range.

Regression method

As discussed in Chapter 3, different multivariate regression methods use different underlying structures. Some methods, such as a neural network, can be called universal approximators. That is, they have no predefined structure and can fit the data regardless of the shape the data takes. Other methods, such as PLS regression, have a linear structure to them. PLS methods have also been developed that have other structures, such as poly-PLS, which uses polynomials instead of linear fits [34]. Although at first it may seem that a universal approximator would be the best choice, when this type of model is extrapolated to regions where training data was not available then large errors can occur because the underlying structure is only determined by the data [24]. Therefore, the choice of which model type to use depends on the given problem.

For this study three different model types were tested: PLS regression and support vector regression with two different types of kernels (radial basis function and a mixed kernel that combined the polynomial and radial basis function kernels). These were chosen to investigate the tradeoff between structure and flexibility. PLS regression is the most rigid because it is a linear method. When using the radial basis function kernel with support vector regression then it becomes a universal approximator that can match the structure of any shape of data. When using a mixed kernel, as suggested by Smits and Jordaan [43], support vector regression can be given properties between these two extremes. For this study, we used a mixed kernel that combined the polynomial and radial basis function kernels. The kernel was mostly weighted towards the polynomial behavior, and the radial basis function portion was only a small amount of the overall kernel (always less than 20%).

A comparison of the RMSEs of the different model types is given in Figure 7-20. For each property, the models were created using the same samples and wavenumbers. To determine if the differences between the different methods were significant, each regression method was repeated at least 15 times while randomly changing the order in which samples were left out during cross validation. The Mann-Whitney test [70] was then used with these sets of 15 values to determine if the differences between the best performing method and the other methods were significant at the 95% level. All of the differences were significant except for with molecular weight when comparing the mixed and radial basis function methods.

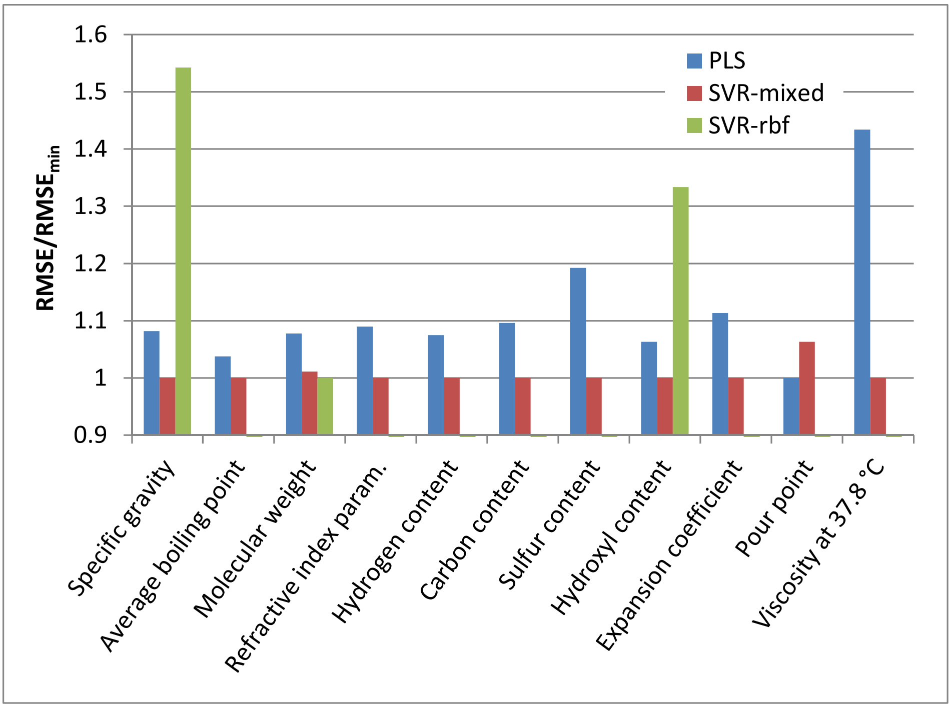

Figure 7-20. Performance of different model types for each parameter.

In addition to comparing the RMSEs, the spread was compared using the 95th percentile of the residuals for each model. This comparison is shown in Figure 7-21. The results are similar to those given in Figure 7-20 for RMSE: the PLS model for pour point gives the smallest spread, but for the other properties support vector regression has a lower 95th percentile or the difference between the two is small and probably within the uncertainty of the metric.

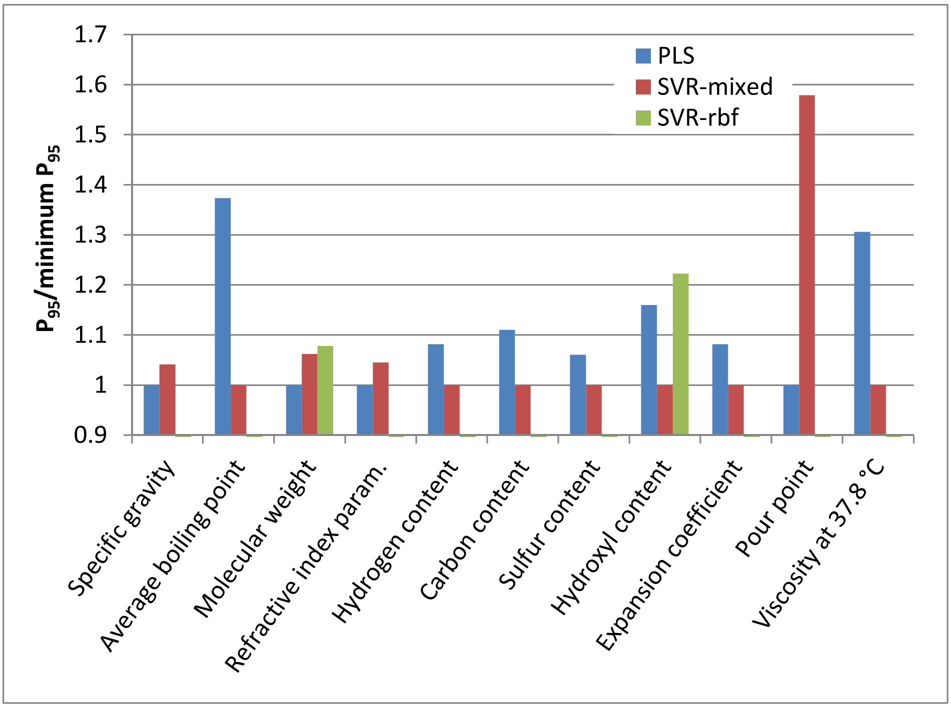

Figure 7-21. Comparison of the 95th percentile of the residuals for different model types.

Support vector regression with the mixed kernel generally gave the best results. This suggests that the relationships between many of the properties and the spectral values were somewhat nonlinear. And indeed, nonlinearity was observed when visually examining the relationship between selected wavelengths and specific gravity and molecular weight. Notably, the PLS model was actually more accurate for pour point, and therefore, PLS regression was used to create the final model for this property.

Therefore, the main model type used in this study was support vector regression with a mixed kernel [43]. The mixed kernel combined the polynomial and radial basis function kernels. The mixing coefficient was used to weight the polynomial portion and one minus the mixing coefficient was used as the weight for the radial basis function portion. The only property for which this model type was not used was pour point. For pour point PLS regression was used instead.

Model parameter optimization

For PLS regression optimization was simple because only the number of components (or factors) needed to be selected. Models were created using 1 to 24 components, and the smallest number of factors that had a RMSE within 2% of the minimum was selected. This was done because sometimes several components had RMSEs close to the minimum value, and in this case a simpler model (i.e. with fewer components) was preferred due to the expectation that a simpler model is more robust.

Finding regression parameters for support vector regression was more difficult because 6 different parameters needed to be determined. Also, support vector regression takes more time than PLS regression, which reduces the number of parameter combinations that can be tried in a reasonable timeframe. Unfortunately, simpler gradient based techniques did not give satisfactory results, likely because the solution space contained multiple local minima. Global optimization techniques gave the best results because they could search over a wider range of possible values. The differential evolution solver [71] implemented in the SciPy package [72] was selected because it was a global technique that resulted in a more accurate model than other techniques, but could also converge in a reasonable amount of time. The optimization criterion was the RMSE of the 5 fold cross validation samples. The one exception was when predicting the viscosity at 37.8 °C for which the root mean squared relative error of cross validation was used because relative error gives a more uniform estimate of the error for viscosity.

8. MODELS FOR TEMPERATURE INDEPENDENT PROPERTIES

Structure parameters

The two most commonly used parameters that are related to molecular structure are specific gravity (or density) and the refractive index. Here the refractive index parameter is used instead of the measured refractive index. The refractive index parameter is calculated using the equation given by Huang [73], which is shown here as Equation 8-1, where *I* is the refractive index parameter and *n* is the refractive index.

$$I = \frac{n^2-1}{n^2+1} \tag{8-1}$$

Of course, specific gravity and the refractive index are temperature dependent properties, but they are often measured at a single standard temperature and used as a characteristic parameter. In this sense, these properties at a specified temperature are temperature independent properties. The hydrogen-carbon ratio is also sometimes used as an energy parameter, but this parameter will be discussed later in this chapter in the section about chemical composition.

The accuracies and regression parameters for the models for the specific gravity and the refractive index parameter are given in Table 8-5. The specific gravity and refractive index parameter at 20 °C were used, and the specific gravity was calculated using the density of water at 20 °C as the reference. The residuals for each fraction can be seen in Figure 8-22 and Figure 8-23. In all of the residual plots in this chapter different markers have been used to distinguish between different types of samples. Those marked “Fuel oil (batch)” were produced using simple batch distillation (similar to an Engler distillation) and those marked “Fuel oil (other)” were produced using some type of multistage distillation (either a rectification column or a two-step batch distillation).

Table 8-5. Accuracies and regression parameters for the multivariate models for predicting energy parameters.

<table>
<tr><th colspan="2"></th><th>Specific gravity (20/20)</th><th>Refractive index parameter</th></tr>
<tr><td colspan="2">RMSE</td><td>0.00467</td><td>0.00174</td></tr>
<tr><td colspan="2">AAD</td><td>0.00308</td><td>0.00112</td></tr>
<tr><td colspan="2">%AAD</td><td>0.32%</td><td>0.28%</td></tr>
<tr><td colspan="2">R^2</td><td>0.997</td><td>0.995</td></tr>
<tr><td colspan="2">95% pred. interval</td><td>±0.011</td><td>±0.0042</td></tr>
<tr><td colspan="2">Range</td><td>0.714 – 1.102</td><td>0.339 – 0.449</td></tr>
<tr><td colspan="2">Number of samples</td><td>355</td><td>327</td></tr>
<tr><td colspan="2">Std. uncertainty</td><td>0.00015</td><td>0.000599</td></tr>
<tr><td rowspan="6">Regression parameters</td><td>C</td><td>1.272</td><td>1.073</td></tr>
<tr><td>ε</td><td>0.003501</td><td>0.003880</td></tr>
<tr><td>γ</td><td>0.8605</td><td>0.7904</td></tr>
<tr><td>Degree</td><td>2</td><td>2</td></tr>
<tr><td>Zero coefficient</td><td>4.166</td><td>4.769</td></tr>
<tr><td>Mixing coefficient</td><td>0.9983</td><td>0.9727</td></tr>
</table>

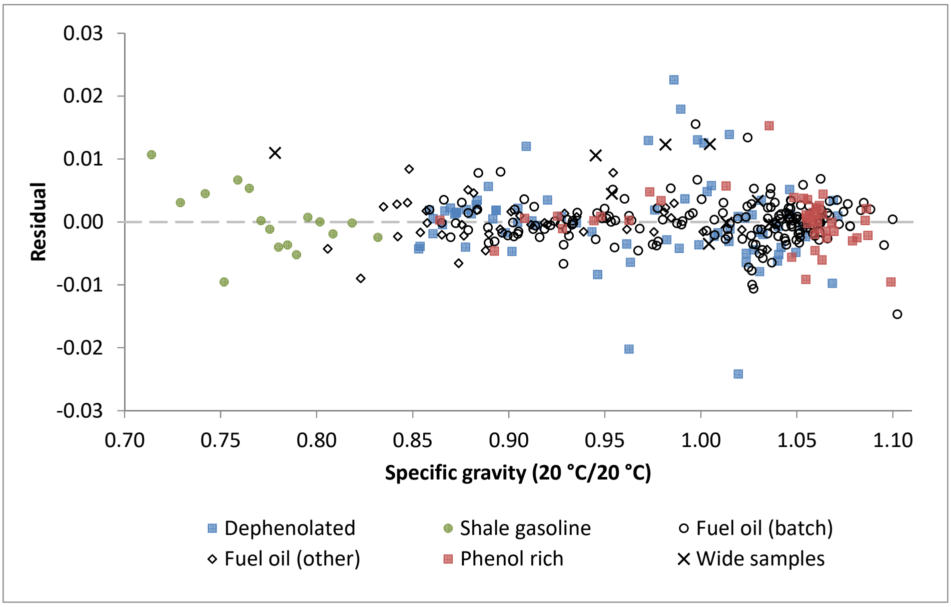

Figure 8-22. Residuals for the model for specific gravity at 20 °C.

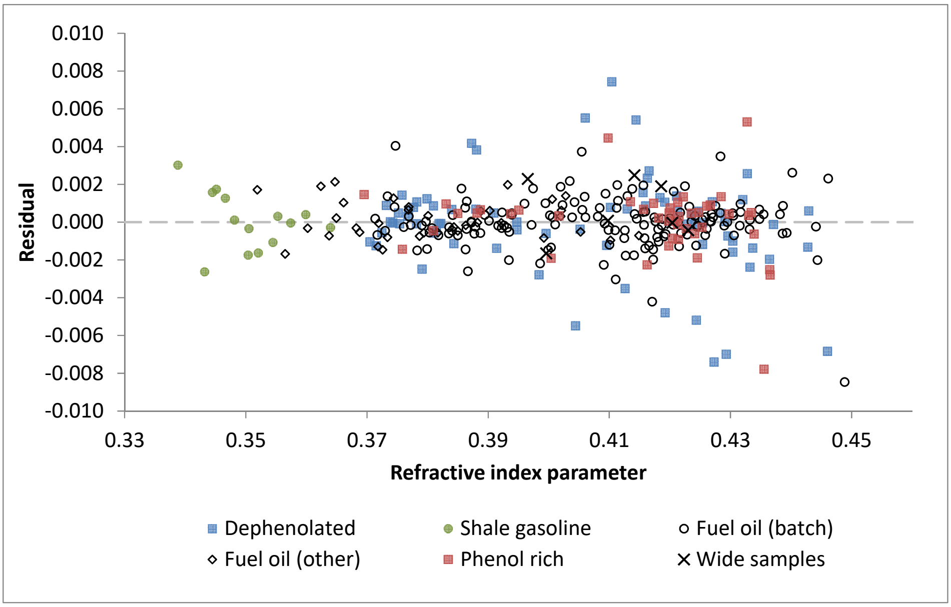

Figure 8-23. Residuals for the model for the refractive index parameter.

The models for these parameters give average relative deviations (%AAD) that are less than 1% and R^2 values greater than 0.994, which shows the models provide a good fit. Some larger residuals can be seen in Figure 8-22 and Figure 8-23. Some of these samples were remeasured and their values corrected, but some of the samples could not be remeasured. Even though the models fit the data better than for most other properties, the RMSE for the specific gravity model was still significantly higher than the measurement uncertainty of the experimental data. This anomaly is further discussed in Chapter 10.

Size parameters

The main size parameters used are average boiling point and molecular weight. A large number of samples also had data for these properties, but the measurement uncertainty for these properties is higher than for specific gravity or the refractive index parameter. Also, note that only data for boiling points up to 400 °C were used because the measurement method could not reliably measure samples with higher boiling points. This is one reason the measurement uncertainty, and resulting model accuracy, are better for boiling point than for molar mass. The model parameters and resulting accuracies can be seen in Table 8-6, and the residuals are plotted in Figure 8-24 and Figure 8-25.

Table 8-6. Accuracies and regression parameters for the models for molecular size parameters.

		Boiling point	**Molar mass**
RMSE		6.95 K	11.8 g/mol
AAD		4.96 K	7.96 g/mol
%AAD		0.90%	3.37%
R^2		0.991	0.978
95% pred. interval		±16 K	±27 g/mol
Range		353 – 673 K	80 – 435 g/mol
Number of samples		229	277
Std. uncertainty		2.1 K	7 g/mol
Regression parameters	**C**	17.79	1.022
	ε	0.01543	0.003438
	γ	0.4687	0.4383
	Degree	1	2
	Zero coefficient	1.055	3.780
	Mixing coefficient	0.9992	0.8412

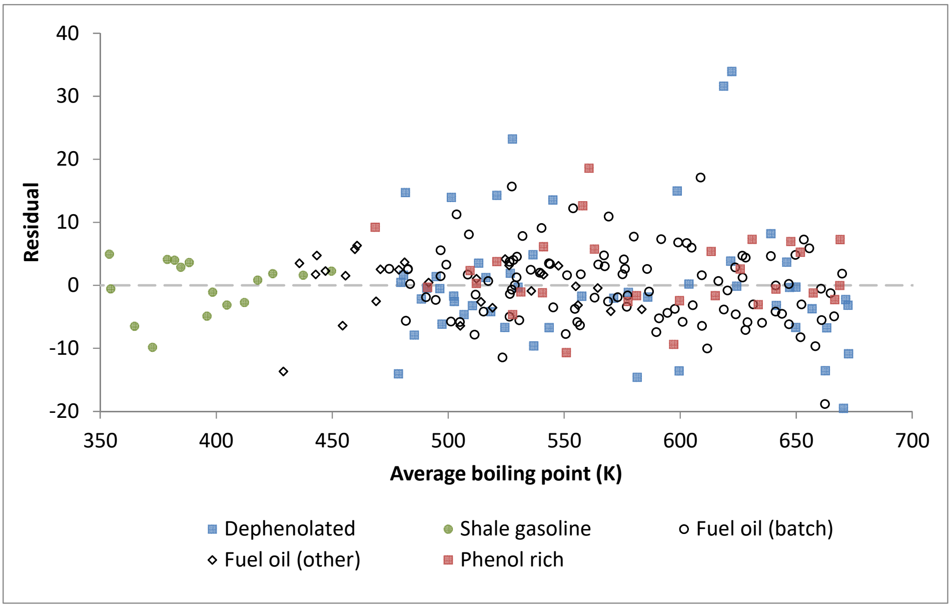

Figure 8-24. Residuals for model for predicting average boiling point.

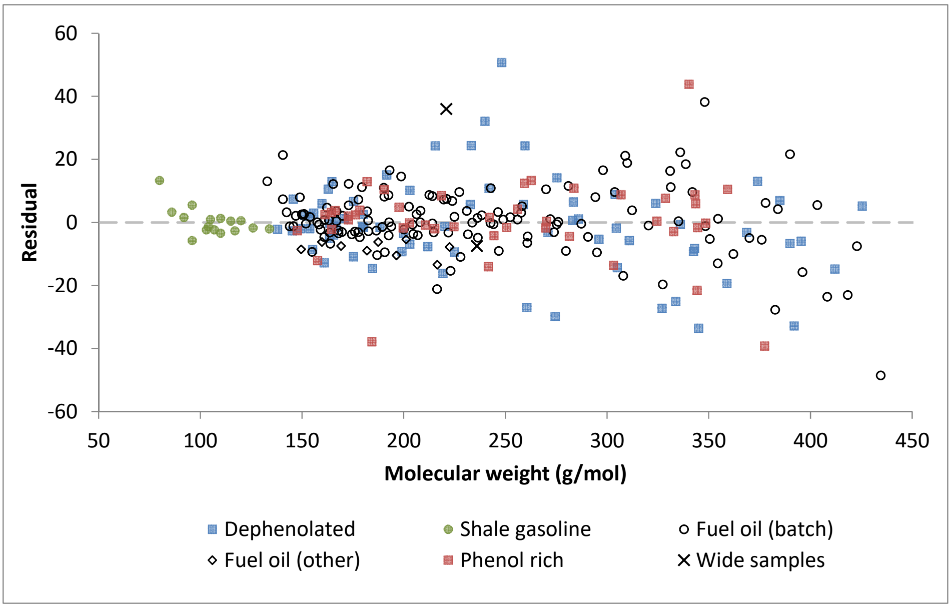

Figure 8-25. Residuals for the model for molecular weight.

Both models have an RMSE that is relatively close to the measurement uncertainty, although the model for molecular weight had a lower R^2 value of 0.978. The lower accuracy for molecular weight seems to be caused by the heavier samples because Figure 8-25 shows that the largest residuals were for these heavy samples. It seems likely that the experimental measurements were least accurate for those samples with high molecular weights, which would lead to the higher residuals seen with these samples. The model for molecular weight also showed greater complexity, which is indicated by the lower value for the mixing coefficient (0.8412). This means the model had more of the nonlinear radial basis function character to it. Again, this greater complexity seems to be caused by the heavier fractions because when using only molecular weights up to 400 g/mol then even a linear PLS model could achieve a better RMSE (about 9.4 g/mol).

Two samples clearly had larger residuals than the rest for the average boiling point model. These two samples also had some of the largest residuals for the other properties. This indicates that there may have been a problem with the infrared spectra or in sample preparation. However, the infrared spectra were visually inspected and no distortions could be seen. Because it could not be confirmed that an error had occurred with these samples, they were left in the database.

Chemical composition

Elemental composition is an important set of data that is regularly measured for fuel samples, and models were also developed to predict elemental composition for Kukersite shale oil. Sulfur content is especially of interest because sulfur form pollutants during combustion, and there are often regulations regarding their concentrations or the resulting pollutant amounts. However, sulfur is usually present in such small amounts that peaks related to its functional groups usually do not appear on infrared spectra. This was also true for the samples in this study: peaks caused by sulfur groups could not be visibly identified from the spectra. For this reason, Trygstad et al. [31] suggested that any relationships between infrared spectra and sulfur content is weak.

Another important parameter for Kukersite shale oil is the content of hydroxyl groups, since these polar functional groups affect the physical and thermodynamic properties of the fuel. Therefore, a model was also created to predict this parameter.

The models for these parameters are described in Table 8-7 and Table 8-8, and Figure 8-26, Figure 8-27, Figure 8-28 and Figure 8-29 show how the residuals are distributed.

Table 8-7. Accuracies and regression parameters for the models for carbon and hydrogen content.

		Carbon	**Hydrogen**
RMSE		0.564 wt%	0.133 wt%
AAD		0.376 wt%	0.102 wt%
%AAD		0.46%	1.05%
R^2		0.916	0.988
95% pred. interval		±1.3 wt%	±0.28 wt%
Range		76.4 – 85.6 wt%	7.8 – 13.2 wt%
Number of samples		257	258
Std. uncertainty		0.36 wt%	0.085 wt%
Regression parameters	**C**	21.68	0.5429
	ε	0.06146	0.003568
	γ	0.001211	0.5297
	Degree	1	2
	Zero coefficient	3.789	3.615
	Mixing coefficient	0.9874	0.9612

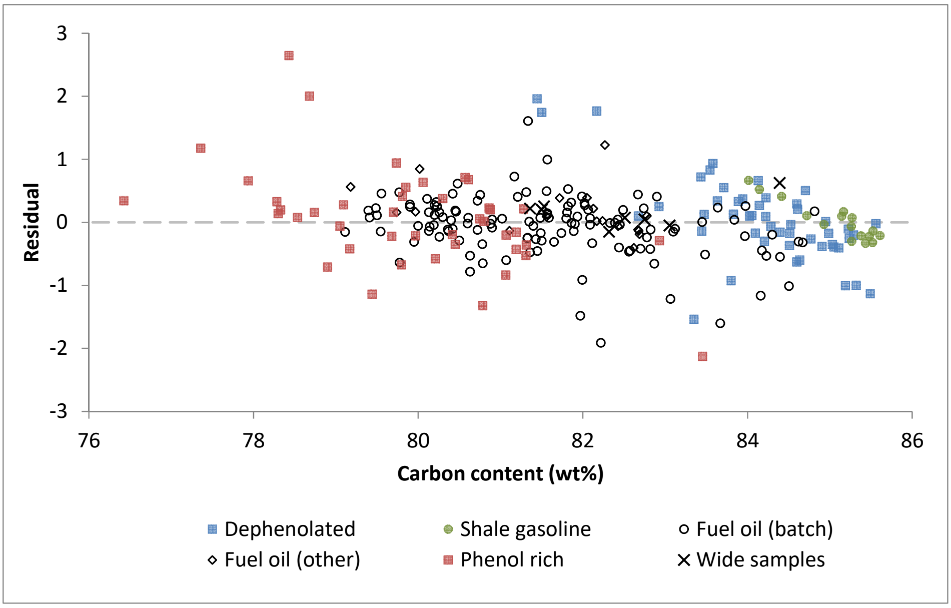

Figure 8-26. Residuals for model for carbon content.

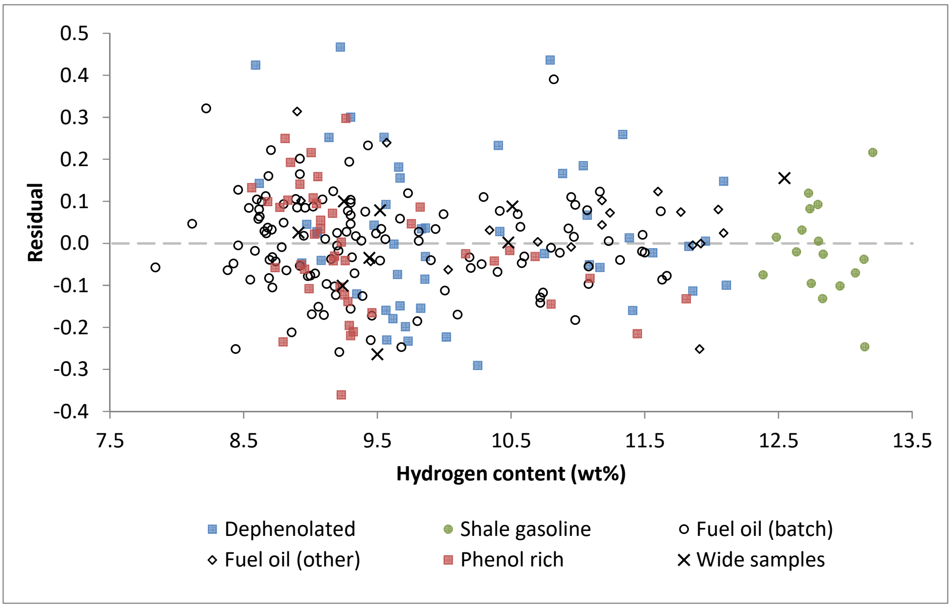

Figure 8-27. Residuals for the model for hydrogen content.

Table 8-8. Accuracies and regression parameters for the models for sulfur and hydroxyl group content.

		Sulfur	**Hydroxyl groups**
RMSE		0.0791 wt%	0.350 wt%
AAD		0.0511 wt%	0.262 wt%
%AAD		5.86%	7.34%
R^2		0.941	0.989
95% pred. interval		±0.21 wt%	±0.72 wt%
Range		0.52 – 1.97 wt%	0.28 – 13.65 wt%
Number of samples		59	57
Std. uncertainty		0.032 wt%	0.19 wt%
Regression parameters	**C**	8.477	12.12
	ε	0.03850	0.01164
	γ	0.6808	0.9968
	Degree	2	1
	Zero coefficient	1.250	4.326
	Mixing coefficient	0.9998	0.9996

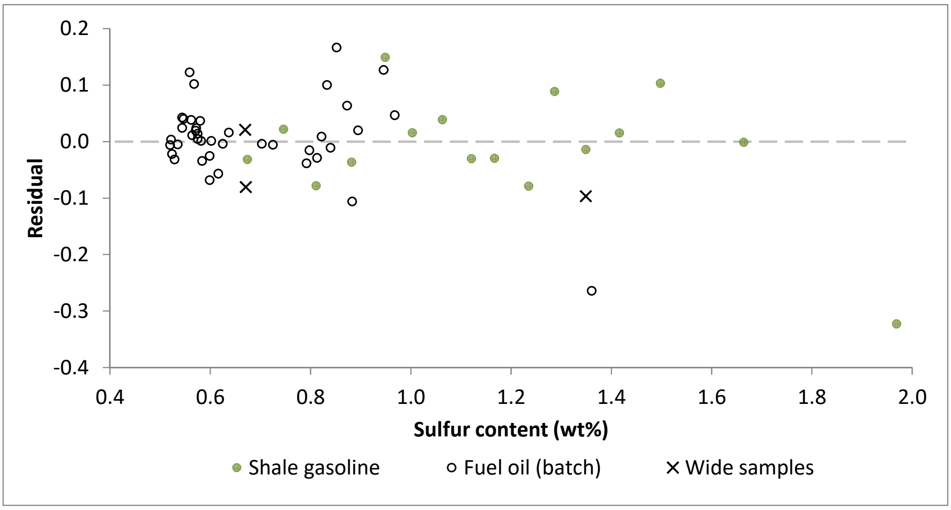

Figure 8-28. Residuals for the model for sulfur content.

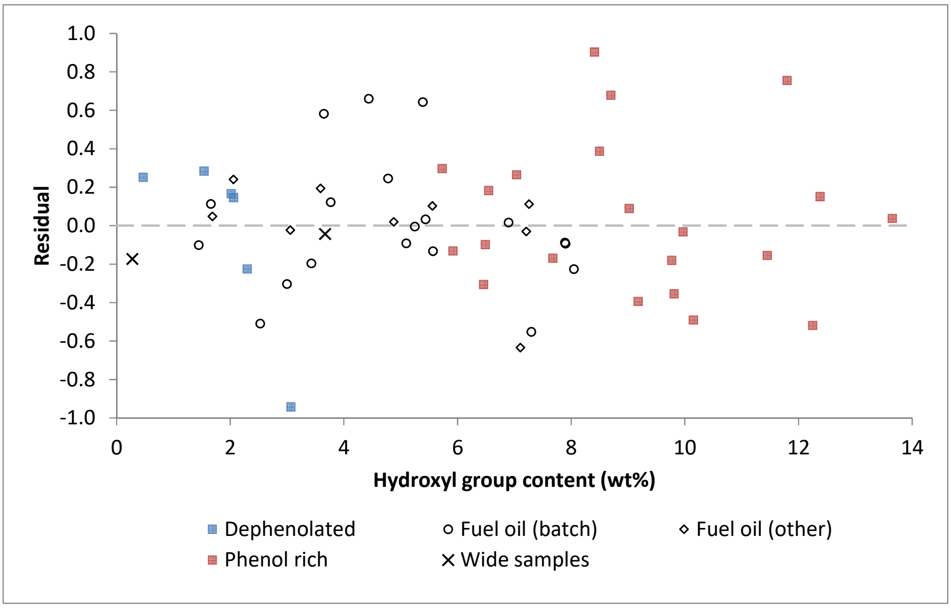

Figure 8-29. Residuals for the model for hydroxyl group content.

Each of these chemical composition properties could be predicted with RMSEs close to the standard uncertainties of the measured data. The R^2 values, however, were relatively low for some of the properties, which was a little surprising because close fits were expected for the compositional parameters. However, because the RMSEs were close to measurement uncertainty, it seems likely that the experimental data itself exhibited a larger amount of scatter than for other properties such as specific gravity.

The lone point with the highest sulfur content also had a large residual, and this probably because there were no other similar samples in the data set. To estimate the residual for a sample, it must be left out of the regression (see discussion of cross validation in Chapter 0). However, for this lone sample a model based on the other samples had a hard time describing this sample, which led to the larger residual. This also appears to be the explanation behind the other point with a large residual because it had the highest sulfur concentration of any of the fuel oil fractions.

Other properties

Three other temperature independent properties were also predicted: the pour point, the thermal expansion coefficient at 20 °C and viscosity at 37.8 °C. More about these models can be found from Table 8-9, and also from Figure 8-30, Figure 8-31 and Figure 8-32. When creating the model for viscosity the data was first transformed by taking the natural logarithm of the viscosity, which helped to improve the performance of the model. Whenever deviations from experimental values were calculated the values were transformed back and relative deviations were calculated.

Table 8-9. Accuracies and regression parameters for the models for pour point, the thermal expansion coefficient at 20 °C and the viscosity at 37.8 °C.

		Pour point	**Thermal expansion coefficient at 20 °C**	**Viscosity at 37.8 °C**
RMSE		5.4 °C	$1.06 \cdot 10^{-5}$ °C^{-1}	-
%RMSE			-	64.3%
AAD		4.6 °C	$0.723 \cdot 10^{-5}$ °C^{-1}	-
%AAD		-	0.94%	43.2%
R^2		0.950	0.993	-
95% pred. interval		±14 °C	$\pm 2.4 \cdot 10^{-5}$ °C^{-1}	approx. -75 to 200%
Range		-46.2 – 47.8 °C	$6.04 \cdot 10^{-4} – 13.6 \cdot 10^{-4}$ °C^{-1}	1.56 – 969000 mPa·s
Number of samples		68	319	115
Std. uncertainty		3 °C	$1.32 \cdot 10^{-6}$	5%
Regression parameters	**PLS factors**	12	-	-
	C	-	52.09	3.818
	ε	-	0.001825	0.04320
	γ	-	0.2822	0.4121
	Degree	-	3	2
	Zero coefficient	-	3.388	0.4178
	Mixing coefficient	-	0.8226	0.9963

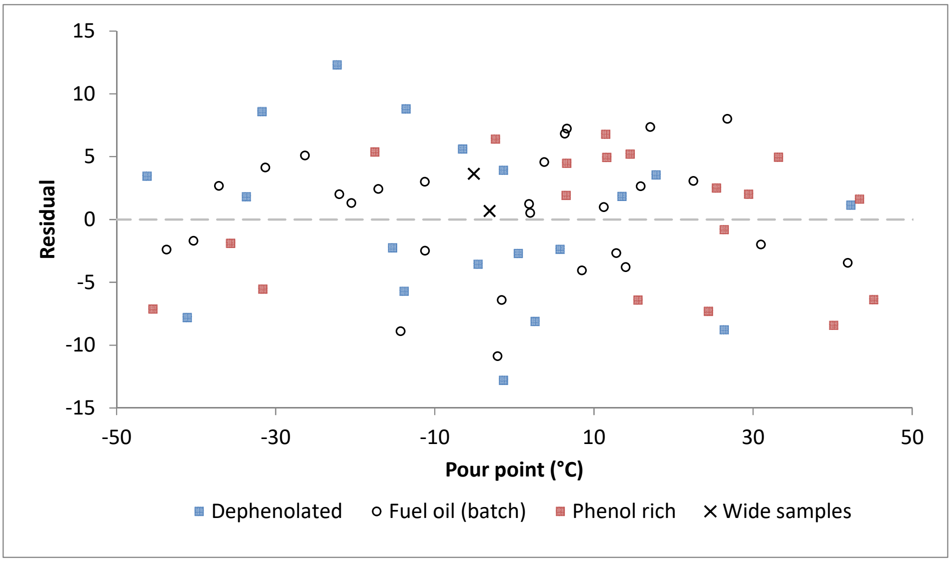

Figure 8-30. Residuals for the model for predicting pour point.

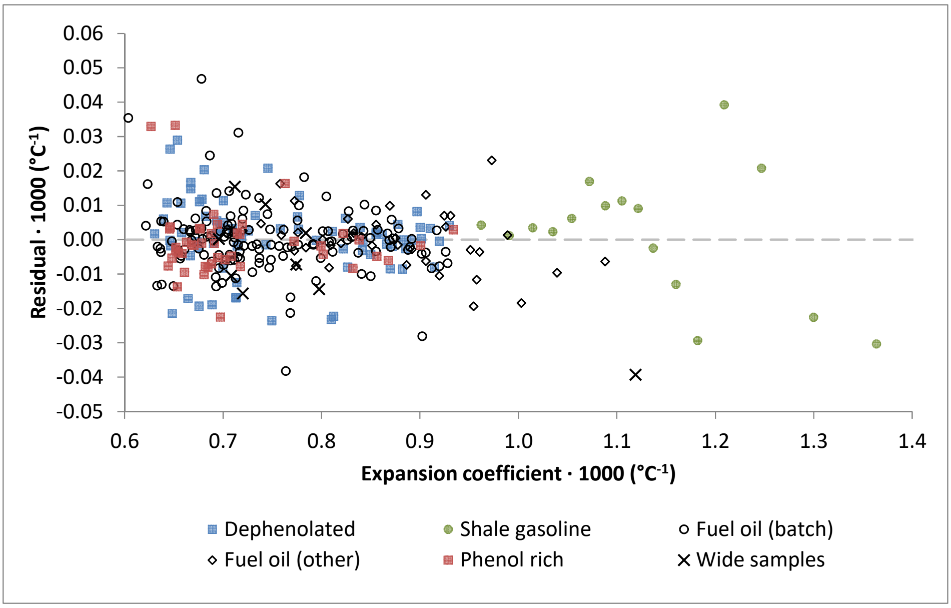

Figure 8-31. Residuals for the model for thermal expansion coefficient at 20 °C.

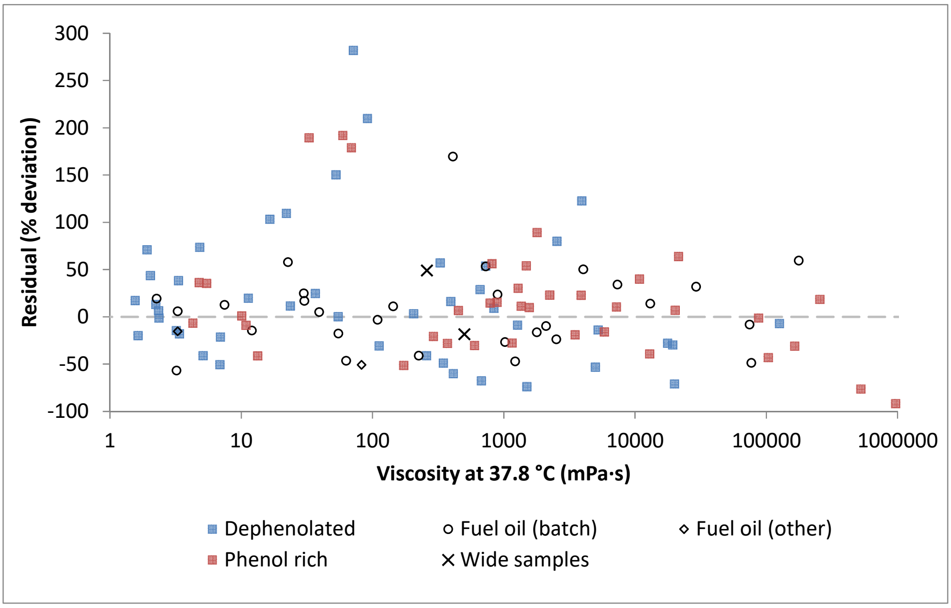

Figure 8-32. Residuals for the model for viscosity at 37.8 °C.

The model for viscosity had a large %RMSE of 64.3%. And yet, viscosity is a difficult property to predict, and errors of 20 – 50% are common for fuel viscosity correlations [1]. Also, the range of viscosities is quite large as the database includes samples with average boiling points above 400 °C. A model was also created using only samples with viscosities less than 1000 mPa·s at 37.8 °C, but only a minor improvement in the accuracy was seen (%RMSE of 62.7%). This suggests that the large range of viscosities did not have a significant impact on the accuracy, and a large portion of the errors may simply be inherent to the property and samples.

9. MODELS FOR TEMPERATURE DEPENDENT PROPERTIES

General approach

Many important physical and thermodynamic fuel properties cannot be fully described by a single value. Some examples are distributions (such as the molecular weight distribution or boiling point distribution) and temperature dependent properties. When reviewing the literature no studies were found that attempted to predict these more complex properties, such as temperature dependent properties over a range of temperatures (see [3], which is included here as Article 1). If a predicted property was temperature dependent, then it was measured and predicted only at specified temperatures. This is likely because the focus so far has been on quality parameters of fuels, and quality parameters are temperature independent. Therefore, an important part of this work was to investigate the potential for predicting temperature dependent properties over a range of temperatures. Attempting to predict these temperature dependent properties is also a step towards a larger goal of modeling distributions and eventually of predicting parameters for equations of state.

The approach used here was to model the temperature dependence using an algebraic equation, and then to predict the coefficients of the algebraic equation from infrared spectra. Once the coefficients have been determined, then the property can be predicted at various temperatures using the algebraic equation. This approach could also be extended to predict distributions, such as the molecular weight distribution, and is similar to the larger goal of predicting parameters for equations of state. Predicting such parameters would allow infrared models to be a more comprehensive solution for thermodynamic modeling: in addition to single parameters, the behavior of fuels at various conditions could be estimated.

In this research two properties were investigated: density and viscosity. The results for density were published in [3] (see Article 1).

Density

The density exhibits a temperature dependence that is essentially linear at moderate temperatures (temperatures below the boiling region of the sample) for many different fuels, including petroleum [74]–[76], biofuels [77], shale oil [19] and coal liquids [19], [78]. Therefore, the density-temperature relationship was modelled using a linear relationship given here as Equation 9-1

$$\rho_T = \rho_{20} - \gamma(T - 20) \tag{9-1}$$

where ρ_T is the density (g/cm^3) at temperature T (T in °C), ρ_{20} is the density at the reference temperature of 20 °C and γ is a constant that describes the slope of the density-temperature relationship. γ is also equal to the density at the reference temperature multiplied by the thermal expansion coefficient at 20 °C.

Equation 9-1 fit the experimental data well, with a RMSE of only 0.0001054 g/cm^3.

The density at 20 °C and γ were then predicted for the different shale oil samples from their infrared spectra. The results for these models are shown in Table 9-10, and show that good levels of performance can be obtained for both of these parameters, as judged by the RMSE and R^2 values.

When using these coefficients to then predict the density at other temperatures the error was at the same level as that for the density at 20 °C (RMSE of 0.004660 g/cm^3 for predictions at the same temperatures as the measured data). Also, there was no noticeable increase in error with temperature. The accuracy appeared to be mostly determined by the error in the estimate for the density at 20 °C, which can be explained by the fact that the slope is relatively small compared to the value of the density. So, the effect of any errors in γ is reduced.

Table 9-10. Performance statistics and regression parameters for the models for predicting the density-temperature relationship.

		γ (slope)	**Density at 20 °C**
RMSE		$7.95 \cdot 10^{-6}$ g/cm^3/°C	0.00463 g/cm^3
AAD		$4.76 \cdot 10^{-6}$ g/cm^3/°C	0.00312 g/cm^3
%AAD		0.63%	0.13%
R^2		0.973	0.997
Number of samples		327	327
Range		$6.65 \cdot 10^{-4} – 9.72 \cdot 10^{-4}$ g/cm^3/°C	0.712 – 1.10 g/cm^3
Regression parameters	**C**	48.03	7.247
	ε	$4.168 \cdot 10^{-6}$	0.0003936
	γ	$1.021 \cdot 10^{-5}$	$3.347 \cdot 10^{-5}$

Viscosity

Viscosity displays a more complex nonlinear temperature dependence. Two different equations were compared when trying to predict the viscosity at different temperatures: a double logarithm equation given by Seeton [79] and an exponential equation called the VFT (Vogel-Fulcher-Tammann) equation [80]. Most other proposed equations take a form similar to one of these two equations. The equation given by Seeton, for instance, is just an improvement of the more widely known Wright equation [81], which was used to create the ASTM viscosity chart. Seeton's equation is given as Equation 9-3

$$\begin{array}{c}\left(ln\left(\mu + 0,7 + e^{-\mu}K_0(\mu + 1.244067)\right)\right) = A - B * ln(T) \\ ln\end{array} \qquad (9\text{-}3)$$

where μ is the kinematic viscosity in centistokes at temperature T (in Kelvin), K_0 is the zero order modified Bessel function of the second kind and A and B are constants. Note that μ + 1.244067 is the input to the Bessel function. The VFT equation is shown as Equation 9-4

$$v = A * e^{\left(\frac{B}{T-C}\right)} \tag{9-4}$$

where v is the dynamic viscosity in mPa/s at temperature T (in Kelvin) and A, B and C are constants. Seeton's equation fit the experimental data to an RMSE of 4.19% (AAD of 2.62%), and the VFT equation had an RMSE of 2.93% (AAD of 1.11%).

The models for predicting the coefficients of Seeton's equation are described in Table 9-11, and it can be seen that relatively good predictions are achieved with a %AAD of about 3% and R^2 values around 0.9. However, the viscosities predicted using these coefficients were very inaccurate. Only 11% of the predicted values had absolute relative deviations of less than 30%. Overall, the median relative deviation was 94%. Figure 9-33 shows the distribution of errors graphically. The poor performance was surprising because the coefficients were predicted fairly accurately.

Table 9-11. Performance statistics and regression parameters for the models for the coefficients used in Seeton's viscosity equation.

		A (intercept)	**B (slope)**
RMSE		0.957	0.185
AAD		0.748	0.156
%AAD		2.63%	3.31%
R^2		0.928	0.883
Number of samples		69	69
Range		19.2 – 35.4	3.35 – 5.85
Regression parameters	**C**	42.42	29.61
	ε	0.1277	0.2783
	γ	0.001672	0.7273
	Degree	2	4
	Zero coef.	4.850	4.797
	Mixing coef.	0.8140	0.8774

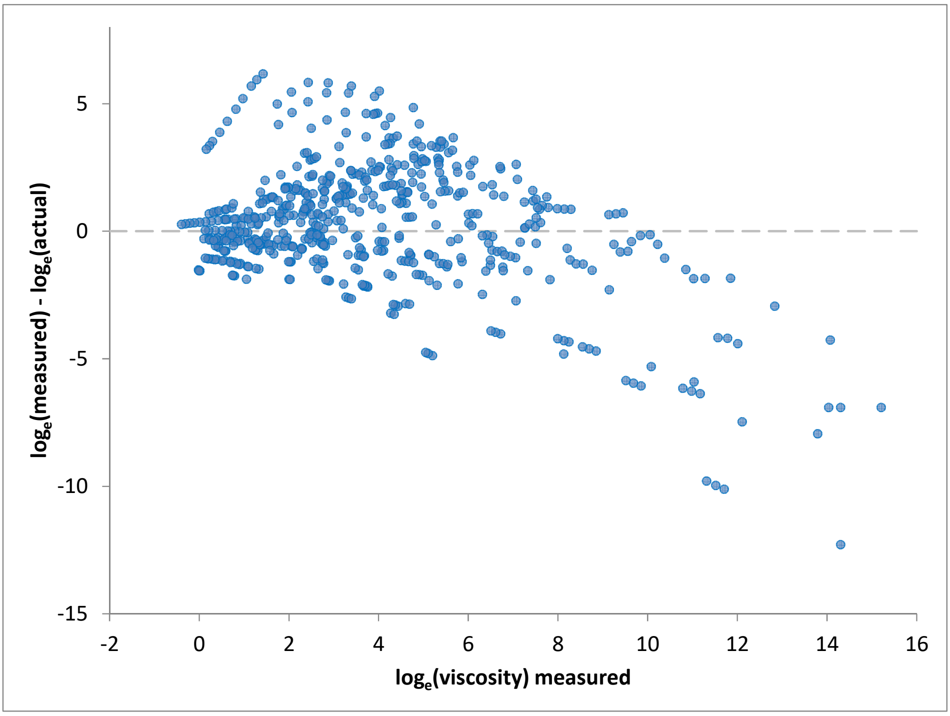

Figure 9-33. Deviation of predicted viscosities from experimentally measured values. Viscosities were predicted using Seeton's equation and the coefficients predicted from infrared spectra.

The coefficients for the VFT equation could not be predicted as accurately as those for Seeton's equation, as seen by comparing their %AAD and R^2 values (see Table 9-12). What was interesting though is that the predicted viscosities were actually more accurate than those calculated by using Seeton's equation. This can be seen visually by comparing Figure 9-33 and Figure 9-34. 35% of the predicted values had absolute relative deviations of less than 30% (compared with 11% when using Seeton's equation), and the median relative deviation was 48%.

Table 9-12. Performance statistics and regression parameters for the models for the coefficients used in the VFT viscosity equation.

		A	B	C
RMSE		0.0383	156	13.3
AAD		0.0255	119	9.62
%AAD		67.5%	17.0%	5.0%
R^2		0.330	0.703	0.689
Number of samples		69	69	69
Range		0.00476 – 0.264	262 – 1670	133 – 241
Regression parameters	**C**	3.032	72.26	4.866
	ε	0.5548	0.4777	0.1965
	γ	0.6208	0.9440	0.03572
	Degree	2	2	1
	Zero coef.	0.3466	3.897	-0.2781
	Mixing coef.	0.9987	0.8059	0.9102

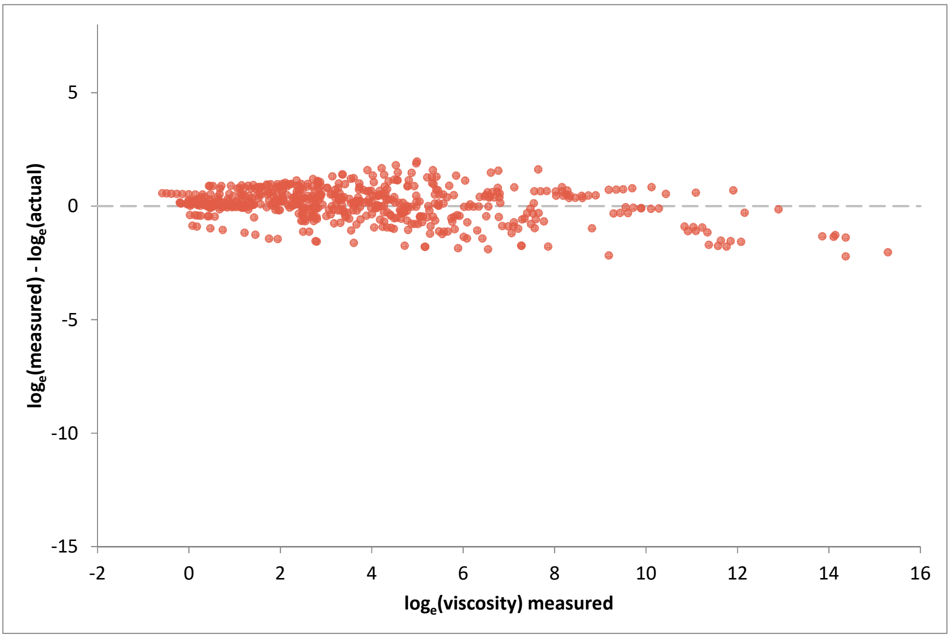

Figure 9-34. Deviation of predicted viscosities from experimentally measured values. Viscosities were predicted using the VFT equation and the coefficients predicted from infrared spectra.

Upon closer comparison of the viscosity equations it was found that the VFT equation is not as sensitive to deviations in the coefficients as Seeton's equation is. A quick calculation to confirm this was done by taking the coefficients for one sample and randomly adding a deviation of up to 3% to the coefficients. Then the viscosity was predicted at a given temperature using both the initial coefficient values and the coefficients with random deviations. By repeating this several hundred times an estimate of the distribution of deviations was obtained. Seeton's equation had a wider range of deviations, with a standard deviation of 157% and a maximum relative deviation of 1200%. The VFT equation had a narrower standard deviation of only 7.2%. The maximum relative deviation was 22%. The difference in sensitivity could also be expected simply by observing the forms of the equations: Seeton's equation uses a double logarithm, and this amplifies deviations more than the single exponential term in the VFT equation. Due to the lower sensitivity to errors in the coefficients, the viscosities predicted using the VFT equation were more accurate, despite the fact that the predicted coefficients were less accurate than those for Seeton's equation.

In conclusion, the results of the two viscosity equations highlight that the choice of equation can play a significant role in the performance of the model for temperature dependent properties. For temperature dependencies that are linear or close to it, the choice of equation is simple, but with complex nonlinear temperature dependencies more attention needs to be paid to the choice of equation. The same effect would likely be seen if using the same method to predict distributions, such as the molecular weight distribution or boiling point distribution.

As a side note, the accuracy of the VFT equation also needs to be put into context because a median deviation of 48% can still appear to be quite large. However, viscosity is "one of the most complex physical properties to predict" [1], and therefore, the accuracy of these predictions is actually relatively good. Note that the average accuracy of the measured viscosity data was about 5 to 10% (Table 5-1). The prediction accuracy for density was also several times larger than its measurement uncertainty, and so compared to the measurement accuracy, 48% is not unreasonably large. Part of the error comes simply from the fact that viscosity is a difficult property to measure and the reference values themselves contain greater uncertainty. Additionally, correlations for predicting the liquid viscosities of petroleum fuels can often give errors of 20 – 50% or more [1], so a median relative deviation of 48%, especially for the wide range of viscosities covered by this shale oil data, corresponds to the level of performance seen with other fuel viscosity prediction methods.

10.PERFORMANCE OF THE MODELS

Accuracies of the models

The performance of the models was assessed against both the actual data for each property and the uncertainty of the experimental method. For comparing how well the predictions matched the actual data Pearson's correlation coefficient was used (R^2), which is a convenient way to compare properties with different units. The R^2 values for the different property models are shown in Figure 10-35. Models for specific gravity and the refractive index parameter had the highest R^2, which shows they closely fit the experimental data. In general though, it is difficult to identify trends among the properties. Although it was expected that specify gravity and the refractive index parameter would have higher accuracies, properties such as carbon content, which should correlate well with infrared spectra, had low R^2 values. This is because the correlation coefficient does not give the full picture. The correlation coefficient is also affected by the measurement uncertainty of the initial reference data used to create the model.

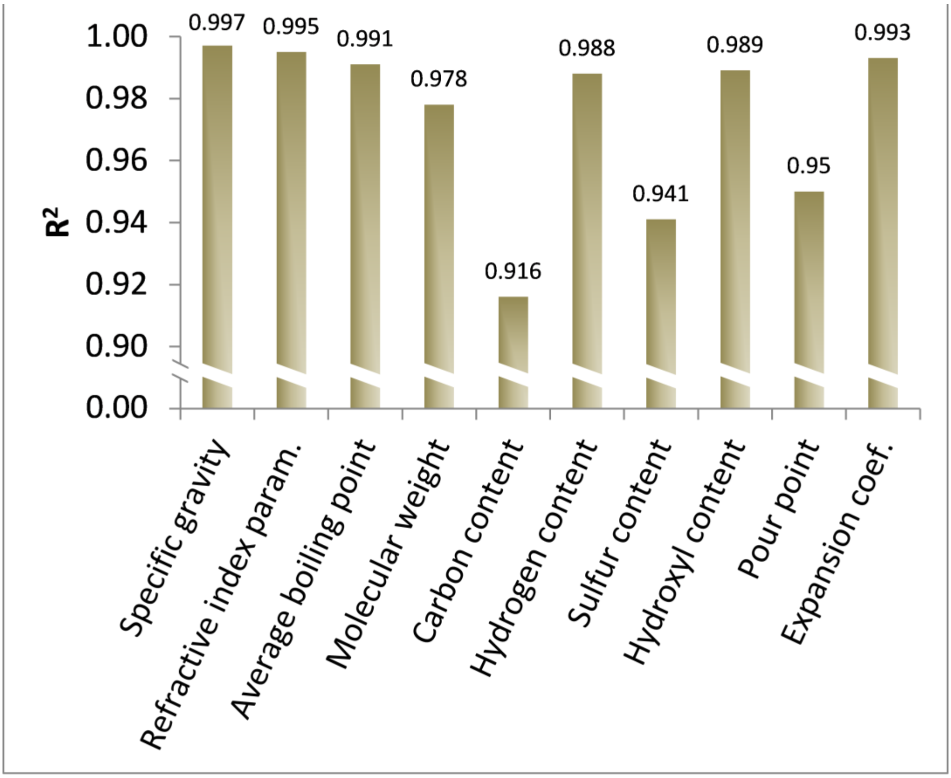

Figure 10-35. Comparison of the correlation coefficients (R^2) of predictive models for different properties.

For this reason performance was also evaluated by comparing the accuracy of the predictions with the measurement uncertainty of the experimental method. Figure 10-36 presents the ratio of these two values for different properties, and in many ways this figure is the opposite of Figure 10-35. Here we can see that although the R^2 values for models for carbon content and pour point were poor, their RMSE to measurement uncertainty ratios are low, which indicates that the poor fit for these properties is in large part a result of a larger uncertainty in the reference data. Specific gravity, however, had an RMSE that was much higher than the measurement uncertainty of the underlying data.

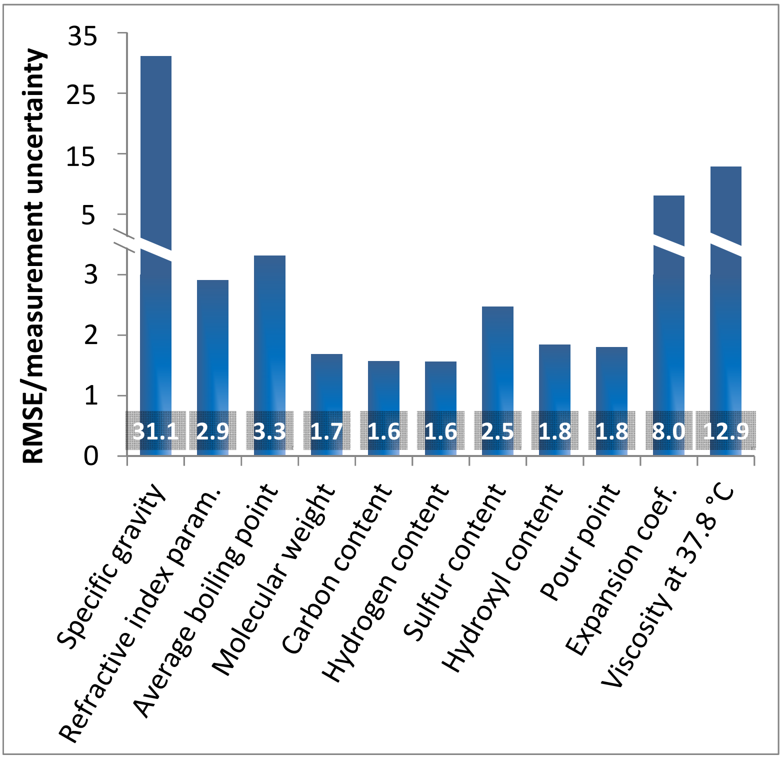

Figure 10-36. Ratio of the RMSE to the measurement uncertainty for each of the property models.

Causes underlying performance

Direct comparison of the models based on performance statistics, such as the R^2 value or ratio between RMSE and measurement uncertainty, did not provide enough information to understand the causes that led to these differences in performance. There are four main factors that can affect the final performance of the model:

- Quality of experimental data (measured property values)
- Quality of input data (infrared spectra)
- Strength of correlation between input and predicted property
- Ability of regression method to fit the shape of the data

Correct implementation of an appropriate regression method is certainly an important factor. As discussed in Chapter 0, extensive work was performed to test different methods to ensure that the models used performed close to the optimal level. Furthermore, Figure 10-36 showed that most of the models had RMSEs close to the measurement uncertainty of the corresponding experimental method, and those that did not had the highest R^2 values, indicating a model that gave a good fit. Therefore, it can be assumed that the performance of the models here was not limited by the regression method, but rather by one of the other three factors.

Additionally, the same infrared spectra were used for all of the models. Although noise and inaccuracies in the infrared spectra likely impacted model performance, it is reasonable to expect that the effect was similar across the different properties. Therefore, the experimental property data and the strength of the correlation between spectra and the property are the most likely factors leading to differences in performance.

Two metrics were calculated to quantify these two factors. The quality of the experimental property data was characterized using the ratio of the measurement uncertainty to the range of values for each property (expressed as a percentage). Using a ratio such as this allows comparison for properties that have different units, and standardizing against the range provides more consistent results than something like the mean. Using this metric, the quality of the experimental data for each property is compared in Figure 10-37. Note that viscosity is not included, and that is because the accuracy of the viscosity model was expressed using relative errors and this metric could not be calculated.

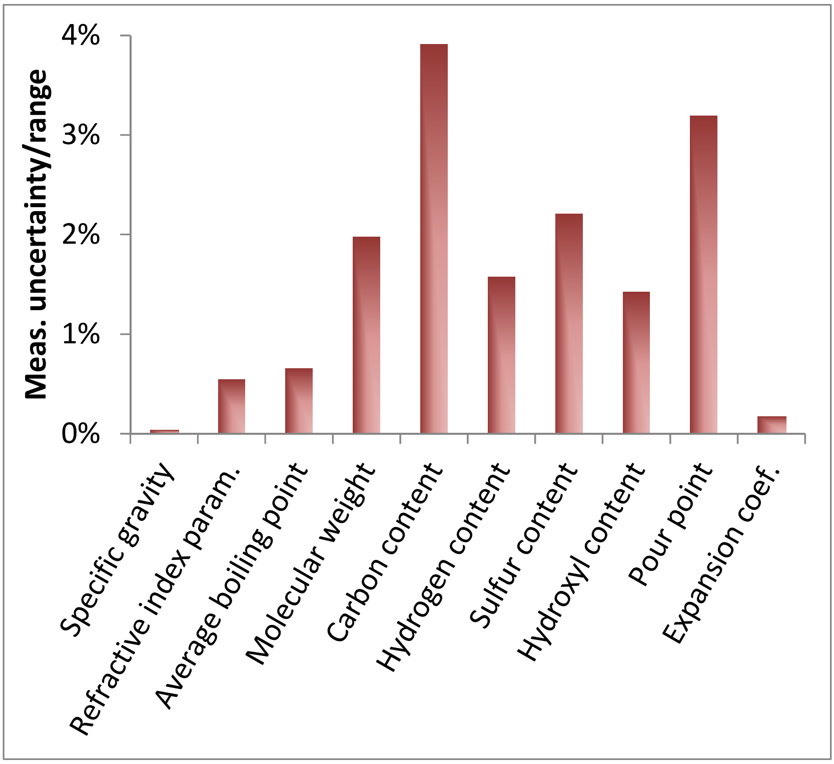

Figure 10-37. Comparison of the accuracy of the data for each of the properties. Here the accuracy is characterized as the relative size of the measurement uncertainty compared to range of property values that occurred.

The extent to which each property was correlated with different wavelengths in the spectra was evaluated using distance correlation [82], which was used because it can also detect nonlinear correlations between variables. Distance correlation was implemented using the distcorr.py script [83]. The code for this script was compared to the implementation in the R Energy package [84] to confirm that the underlying algorithm was the same. The results are presented as a box and whisker plot in Figure 10-38. The median distance correlation value was used to characterize how strongly each property was correlated with the infrared spectra.

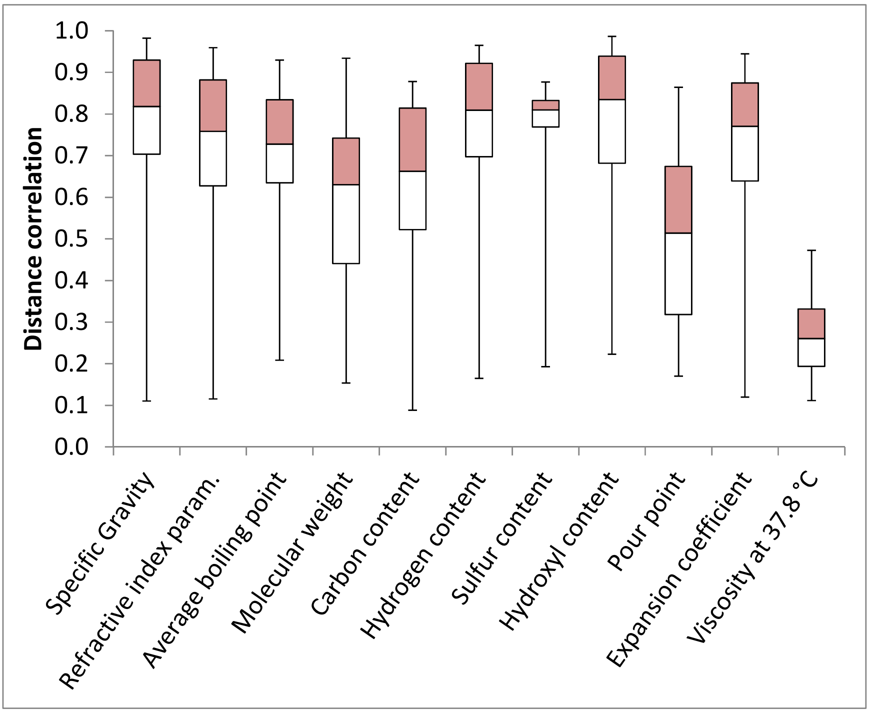

Figure 10-38. Box and whisker plot showing the distribution of distance correlation coefficients that occurred between each spectral wavelength and the given property.

Obviously, distance correlation is also affected by the quality of the property data. To investigate this relationship the distance correlation values were plotted versus the ratio of measurement uncertainty and range, and this plot is shown as Figure 10-39. Interestingly, two parallel trends emerge, with composition parameters falling along one line and the physical and thermodynamic parameters falling along a lower line. So the relationship between correlation strength and measurement uncertainty depends on the type of property being considered. Put another way, for properties that both have a similar accuracy the one related to chemical composition will be more strongly correlated to the spectra than the physical or thermodynamic property.

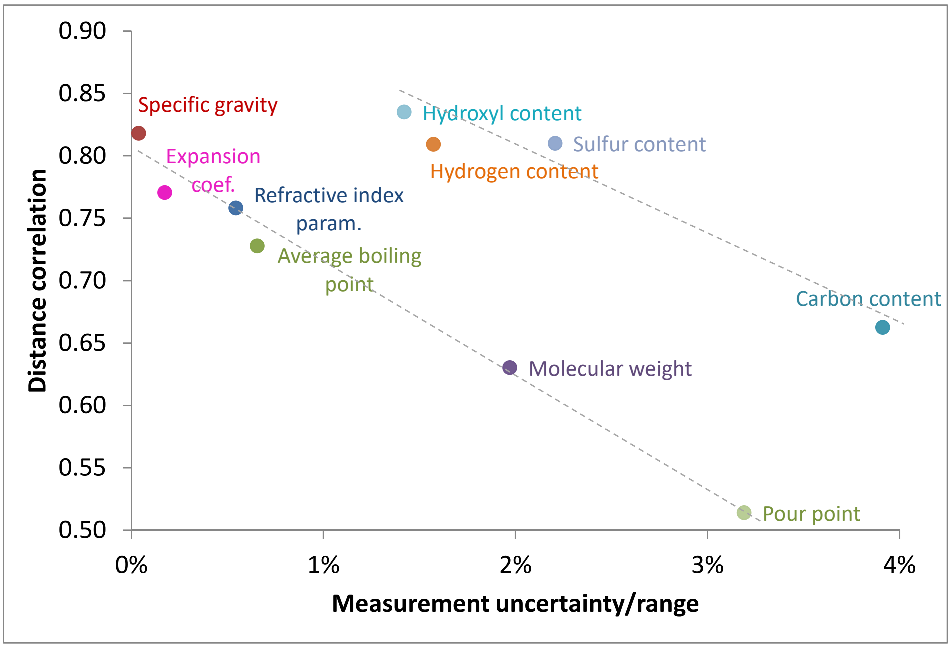

Figure 10-39. Relationship between the quality of the experimental property data and the median distance correlation between the spectra and a property.

This result was somewhat surprising. Although it is logical that certain types of properties correlate better with infrared spectra than others, the exact split seen here was unexpected. As Trygstad et al. [31] conceptualized it, some properties (such as those describing chemical composition) are more directly related to the spectra because many functional groups directly cause the peaks that appear in the spectra. This explains why the composition parameters were shown to be more strongly correlated in Figure 10-39. By contrast, physical and thermodynamic properties could be classified as having an indirect relationship. That is, they are related to chemical composition and are only indirectly related to the peaks in the spectra. And yet, it was expected that some of these properties would be more closely correlated than others. For example, it was hypothesized that structure parameters (specific gravity and the refractive index parameter) are more closely related to the functional groups present, and would therefore be more closely correlated with the spectra than size parameters (molecular weight and average boiling point). Figure 10-39 does not show a difference though, and it appears that any differences in model performance for physical and thermodynamic properties can be described solely by the differences in the accuracy of the experimental method used for measuring the data for model creation.

Another surprise was finding that sulfur content followed the same trend as the other composition parameters. Sulfur is only present in the oil in small concentrations, and peaks directly caused by sulfur functional groups were not noticed in the spectra. Trygstad et al. [31] suggested that for these types of composition parameters any correlation would be weak and indirect because it would rely only on links between other functional groups and the concentration of the trace compound. In this study though, no difference was observed. However, this may be because the sulfur data was for samples from only three of the distillations. It is possible that if more data points were measured that the connection between the sulfur content and the concentrations of other functional groups would be weakened.

With these metrics in hand, the next step is to investigate how these factors influence the resulting performance. Here performance was quantified by taking the ratio of the RMSE to the range of property values. When plotting this measure of performance versus the two metrics defined earlier, it turns out that the distance correlation value is not strongly correlated with model performance. When taking the effect of measurement uncertainty into account, it seems that the strength of the correlation between spectral and property values has essentially no impact on the performance of the resulting model. Instead, the quality of the property data used in building the model is the strongest predictor of the performance of the model. This can be seen from Figure 10-40, which plots model performance versus the ratio of measurement uncertainty to the range. Both composition parameters and physical and thermodynamic properties lie on the same line, indicating that the stronger correlation between composition parameters and the spectra did not impact the eventual performance of the models. The model for sulfur content fell furthest from the general trend, but this may be because the measurement uncertainty for this property was taken from the ASTM standard instead of being determined experimentally with standard compounds.

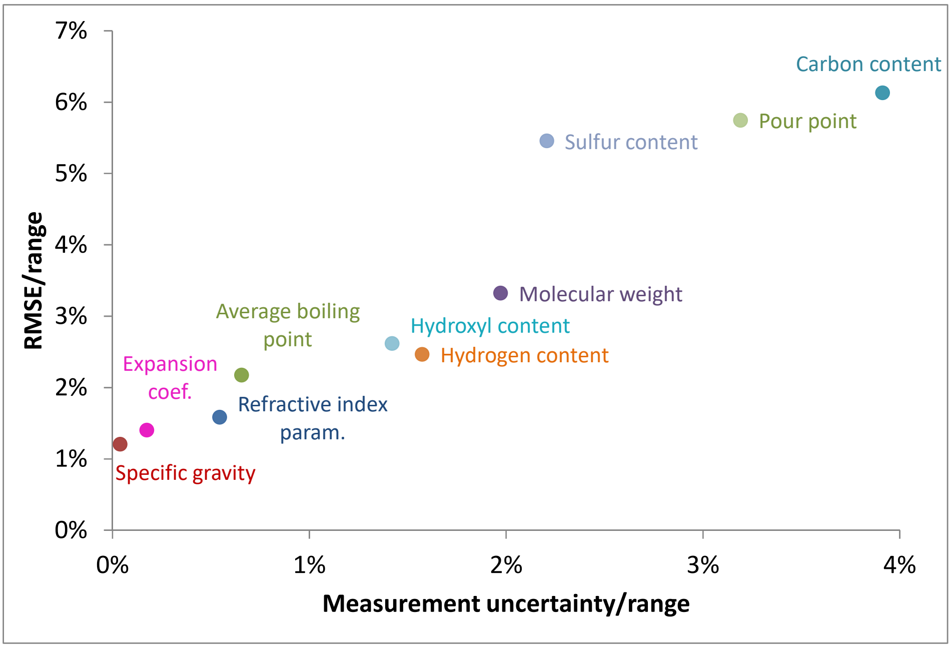

Figure 10-40. Effect of the quality of the experimental property data on the performance of the resulting multivariate model.

This brings up an interesting question, though. Why is it that model performance was not better for those composition parameters that are more strongly correlated with the infrared spectra? One explanation could be that the maximum distance correlation values are high for all of the samples except viscosity (see Figure 10-38). This indicates that although some properties are better correlated over the entire spectrum (as indicated by the medians), almost all of the properties still have at least some wavelengths that are strongly correlated with the property of interest. It is possible that these few wavelengths with a strong correlation can give all the information needed to obtain a fit for a physical or thermodynamic property that is as good as the fit for a compositional parameter.

Another important result gleaned from Figure 10-40 is that there is a lower limit to model performance. As the measurement uncertainty approaches zero, the RMSE does not. For this particular set of experiments it approaches 1% of the data range. This explains why the RMSE for specific gravity is so much larger than its measurement uncertainty (see Figure 10-36). This behavior also suggests there is a second factor that influenced the performance of these models and caused this lower limit to appear. Most likely it is the uncertainty in the input data: the infrared spectra.

As stated in Chapter 5, the infrared spectra were estimated to have a median relative standard uncertainty of 1.8% (see also Figure 5-5), and this uncertainty in the spectral values can effectively put a limit on the resolution that can be obtained by a model. For example, the relative standard uncertainty of the specific gravity measurements is 0.015% of the mean specific gravity value. This means that two samples with specific gravities that differ by only 0.1% can be distinguished using a density meter, but using the spectrometer any differences would likely be drowned out by the noise in the spectra.

Comparison with bulk property correlations

Bulk property correlations are commonly used for predicting fuel properties. They are often used for shale oil because more detailed data is usually not available. Therefore, comparing the accuracies of infrared models to those of bulk property correlations provides a way of assessing the performance of the infrared models.

Many bulk property correlations for petroleum can be found in the literature [1]. Figure 10-41 compares the stated accuracies for the best of these correlations to the accuracies obtained with the models given here. The correlations compared were taken from Riazi [1]. To emphasize, these literature correlations were not actually applied to the shale oil samples from this study. What is being compared is only the accuracy of the correlation as stated in the literature reference. Applying these literature correlations to shale oil directly usually gives large errors and would not provide a meaningful comparison. Also, the accuracy of the correlations for average boiling point, specific gravity, molecular weight and the refractive index parameter was assessed using representative pure compounds, not actual petroleum samples.

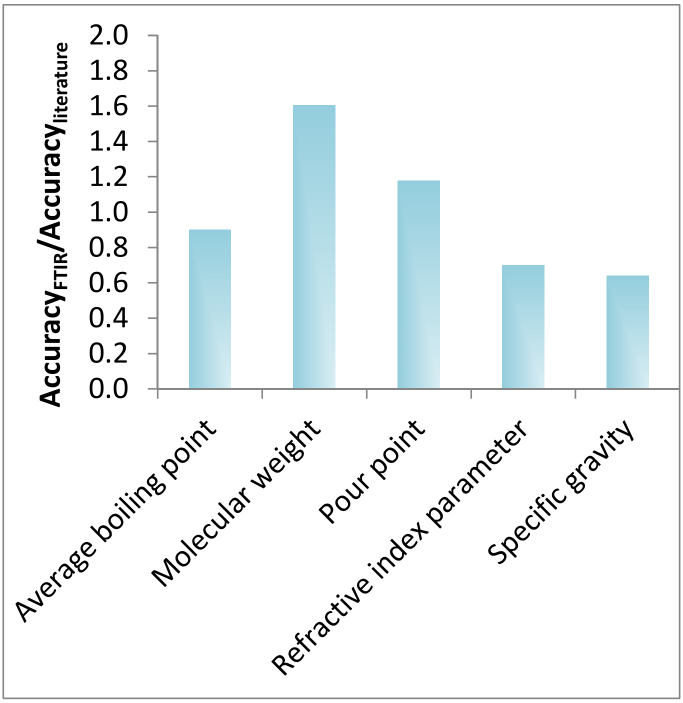

Figure 10-41. Comparison of the accuracy of the infrared models and bulk property correlations for petroleum given in the literature [1], [85].

Therefore, based on the comparison in Figure 10-41, the infrared models generally have an accuracy similar to or better than bulk property correlations found in the literature. However, this may partially be due to the fact that some of the literature correlations were created using either a wider range of fractions or only pure compounds.

To get a better direct comparison the infrared models were compared to bulk property correlations developed using the same shale oil database. The ratios of the RMSEs of the infrared and bulk property prediction methods are presented in Figure 10-42. The infrared models give similar or better results for all of the properties compared except viscosity, and in general, the infrared models are more accurate. The property with the lowest ratio is specific gravity. And yet, even the models for average boiling point and molecular weight had better accuracies than the corresponding bulk property correlations. This was surprising because infrared spectra do not directly measure molecular size, while the bulk property correlations actually do incorporate this information (by using either the average boiling point or average molecular weight to estimate the other).

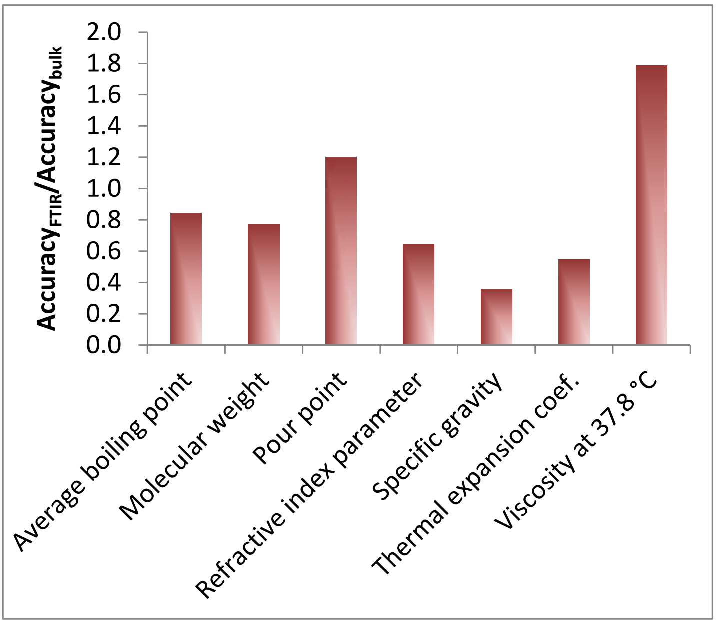

Figure 10-42. Comparison of the accuracies of the infrared models and bulk property correlations.

The accuracies of bulk property correlations improved if smaller property ranges were fit. For example, when fitting only the fractions with boiling points below 350 °C, then the bulk property correlation for molecular weight gave results that were better than the infrared correlation. When using the same range for average boiling point, then there is also an improvement in the accuracy, although the accuracy is still lower than that of the infrared model. So, part of the advantage of the infrared model may come because it can more closely fit the shape of the data than the simpler bulk property equation.

Bulk property correlations were also created for some temperature dependent properties. The density, for instance, was predicted by using the measured density at 20 °C in conjunction with the average boiling point or molecular weight, if available, to predict the slope of the density-temperature relationship (γ). This was then used with the same linear equation (Equation 9-1) to predict the density at other temperatures. Data and details about these bulk property correlations will be the subject of an article from our laboratory that will be published in the near future.

When using this method, the RMSE versus the measured density data was about 0.001 g/cm^3, which was several times smaller than that obtained using the model based on infrared spectra (0.00466 g/cm^3). However, upon further examination, it was noticed that most of the error for the infrared predictions came as a result of errors in predicting the density at 20 °C. Obviously, the measured density value used with the bulk property correlation is more exact than the predicted value used with the infrared method. The predicted slope, however, was more accurate when using the infrared method. When using the measured density at 20 °C with the slope predicted from infrared spectra then the accuracy of the predictions were better than using either the bulk property correlation or infrared models on their own.

In a similar manner, the coefficients for the VFT viscosity equation (Equation 9-4) can also be predicted from bulk properties. More specifically, correlations based on the density at 20 °C and the average boiling point were created, using the simple equation form given by Riazi and Daubert [86]. Again, specifics about the creation of these bulk property correlations will be given in a future article from our laboratory.

Using these bulk property correlations and the VFT equation, the viscosity could be predicted with a median relative deviation of 35% (compared to 48% with the constants predicted using infrared models). It appears that, at least in this case, the bulk property correlations allowed more accurate predictions.

Comparison by fraction

The performance of the models can also be checked by viewing the results for sequential fractions from a single distillation. A typical distillation was selected for this comparison, and both predicted and measured values were plotted. Figure 10-43 shows the results for specific gravity and Figure 10-44 shows the results for average boiling point. Error bars were also placed on the measured values to show the expanded uncertainty, but for specific gravity the expanded uncertainty was so small that the error bars are hidden by the marker.

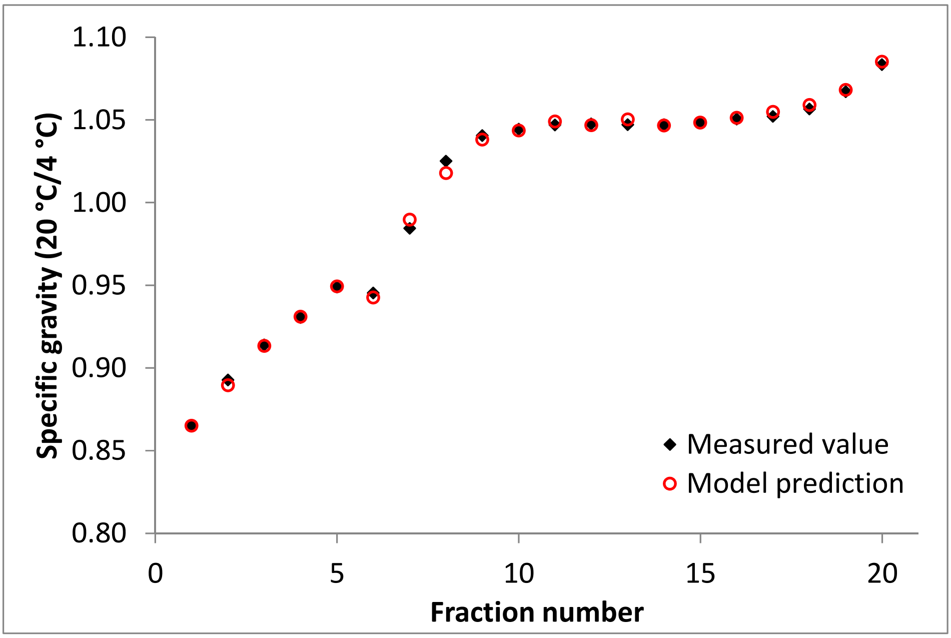

Figure 10-43. Comparison of the performance of the specific gravity model for fractions from a typical distillation.

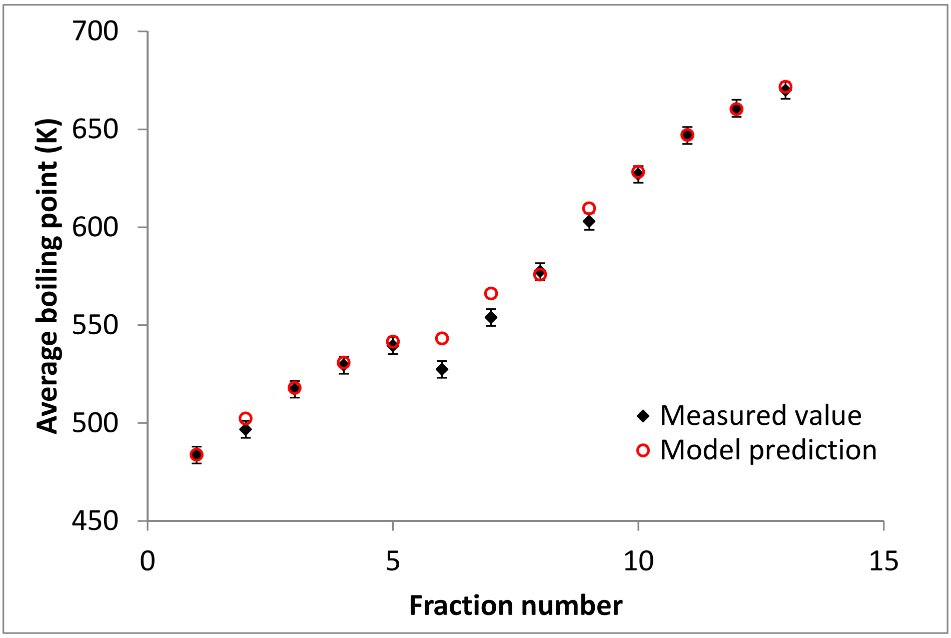

Figure 10-44. Comparison of the performance of the average boiling point model for fractions from a typical distillation.

For almost all the fractions in these figures the difference between model and measured values is smaller than the difference between subsequent fractions. That is, the infrared models can generally distinguish between two fractions. For higher boiling fractions the density is essentially constant, and it would be difficult to distinguish between fractions even when using measured density values. Note that fractions 6 and 7 are lower than the general trend given by the rest of the data. At that point in the distillation the pressure was reduced from atmospheric pressure to low pressure, and doing so required a pause in the distillation. This caused fractions 6 and 7 to have different compositions that did not follow the overall trend, leading to the drop in specific gravity and boiling point. The density model could account for this difference, but the average boiling point model had larger errors for these two samples.

11.ADDITIONAL ISSUES FOR WIDER APPLICATION

Transfer of a model for use on other instruments

Because infrared spectrometers generally give different responses for the same sample, instrument specific variations usually need to be taken into account before using the multivariate model on a new spectrometer. This is a large limitation because it means a model can only be used on the spectrometer it was created with, or a spectrometer that happens to give a very similar response. Also, if the spectrometer itself changes over time (due to aging or intentional modifications) then the model can easily be rendered useless [26]. For this reason, many methods have been proposed for transferring calibrations between instruments [87].

Unfortunately, current methods still require significant effort to reach a satisfactory level of accuracy. Feudale et al. [87] separated calibration transfer methods into three groups: methods used before model creation, standardization methods and preprocessing methods. Methods used before model creation, however, can increase the cost of the measurement system and model development, and unforeseen or uncontrollable variations are still likely to occur. With standardization methods many reference samples are usually needed to create an accurate transfer model, and because the same samples must often be measured on both instruments, the reference samples and/or spectrometers must be carefully maintained and transported. Preprocessing methods offer the advantage that no reference standards are needed, but they rely on having enough of an understanding of the sources of variation that these extraneous effects can be removed while still retaining the important chemical information.

For comparing different methods, it is helpful to think about the effort they require compared to the creation of an entirely new model. This idea is illustrated in Figure 11-45, which shows the relative amount of effort different methods require. Generally, the best method would be that which requires the least effort while still achieving the required accuracy. More details about the different types of methods, including common methods and references, can be found in Table 11-13.

Figure 11-45 has several illustrative curves to show how the accuracy of the transfer might increase as progressively more difficult methods are used. The shape of the curve mainly depends on how complex the differences between spectra are and the complexity of the prediction model. If the spectra are simply offset or are related by a set ratio, then a simple preprocessing method may give an accuracy that is almost as good as creating a whole new model. If differences are harder to model or a complex model is required to predict the parameter, then the easier methods may not yield much of an increase in accuracy. Additionally, the curve shape can be made more favorable by exerting more effort during the measurement or model creation steps, for example by using only spectrometers that give very similar responses or by removing temperature effects by controlling the temperature of the measurement system. In this way some effects are never even introduced into the spectra or are automatically accounted for by the model. Although as mentioned, these methods increase the cost of the system.

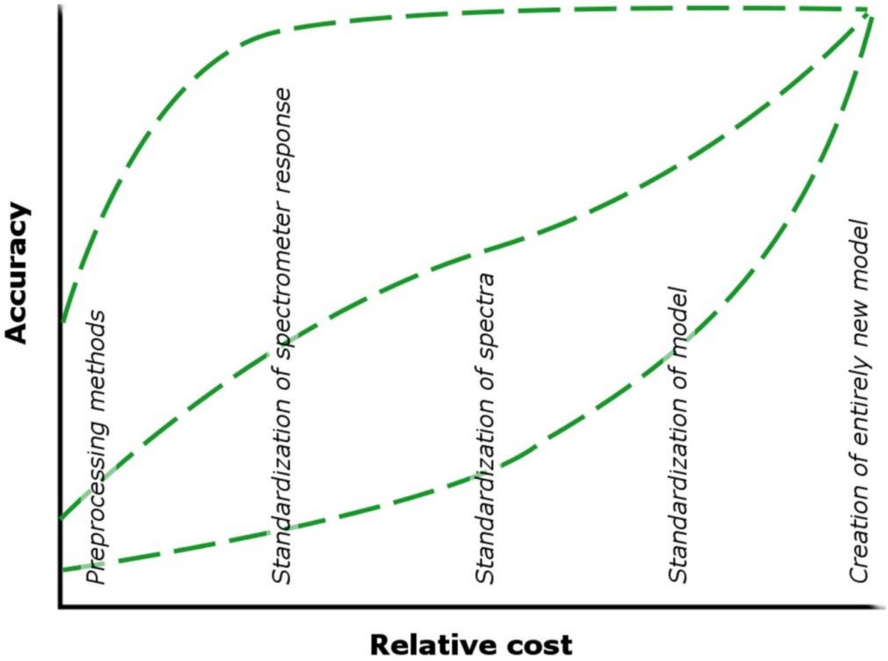

Figure 11-45. Illustration comparing the relative costs and potential accuracies of the main different types of model transfer methods. The different curves show that different behaviors can be expected depending on the problem at hand.

Table 11-13. Descriptions of the different types of model transfer methods, including examples of common methods.

Type of method	Characteristic features	Common methods
Preprocessing	The spectra are mathematically transformed with the goal of removing unimportant information and standardizing data. No calibration or additional data is required.	Standard normal variate, orthogonal signal correction, multiplicative signal correction, finite impulse response filtering [88], first or second derivatives
Standardization of spectrometer response	The response of a spectrometer is modelled and is used to correct spectra or the model. Usually only a few spectra for pure compounds are needed.	Virtual standards [89], [90]
Standardization of spectra	A separate model is created to transform the spectra from the new instrument to resemble those of the primary instrument. A number of representative samples must be measured on both instruments.	Direct standardization [91], piecewise direct standardization [91]
Standardization of model	The model for a specific parameter is adjusted to better apply to spectra from the new instrument. Many calibration standards (samples for which the parameter value is known) must be measured on the new instrument.	Slope-bias correction [92], creating transfer model using PLS [93], model updating

Preliminary investigations were performed to get an idea of how difficult it would be to transfer the multivariate models created in this study to other spectrometers. 19 oil samples were measured on a second spectrometer (a Nicolet IR100, Thermo Fischer Scientific, Madison, WI, USA), and then the models were applied to these spectra from the second spectrometer.

There was a clear difference between the response for the two spectrometers, and it proved to be difficult to account for. Good results were achieved when using more complex methods, such as piecewise direct standardization or model updating. With model updating, by introducing some spectra from the second spectrometer into the calibration data set, the accuracy of predictions for the second spectrometer reached the same level as those from the primary spectrometer. This test was performed with the model for density at 20 °C.

Piecewise direct standardization gave good results for the model for hydroxyl group content, although it also introduced artifacts into the transformed spectra. Other researchers have studied this drawback of piecewise direct standardization and have suggested ways to deal with it, so it is likely that these artifacts could be reduced or eliminated. It was also noticed that it was easier to transfer the model for hydroxyl group content than most other models. The hydroxyl group model was less complicated than most other models (for instance, fewer PLS components were needed to achieve a good accuracy for hydroxyl group content than for other properties), which seems to account for this difference. Thus, it is likely that for most physical and thermodynamic properties the models will have a higher complexity and will be more sensitive to variations in the spectra. This could make model transfer more difficult.

Because creating and measuring calibration samples on both spectrometers is burdensome, it would be preferred to have a method for transferring spectra that does not require samples to be measured on both spectrometers or the creation of calibration standards. A variety of methods were tried, including using the signal normal variate of the spectra, multiplying the spectra by a constant ratio (multiplicative signal correction) and using a wavelet transformation. The virtual standards method introduced by Cooper et al. [89] was also tried. These methods all gave improvements in accuracy over no correction, but the accuracy was still at least two or three times lower than that for spectra from the primary spectrometer. Attempts to better model the variation between the spectrometers did not give significant improvements over the other methods. A more thorough study would need to be done to better investigate model transfer, which would likely require more samples to be measured on multiple different instruments.

Better basis for extrapolation to fuels from other sources

Another problem is that current chemometric methods are not able to accurately predict values for types of fuel not included in the calibration set [8]. This problem arises due to the fact that different fuels can have very large differences in composition, some of which are present in the original sources of the fuels and some of which are caused by processing. The spectra or chromatograms often used as input for multivariate models are also quite complex, and therefore, the trends or structures identified in the calibration samples will likely not hold for fuels with compositions outside of the calibration range.

As of yet, there does not appear to be a good general solution to this problem. For some properties it may be possible to find a handful of key variables that still enable accurate predictions, and thus, the input data can be simplified, which would make it more likely that it could provide good results over a wider range of fuels. A more expensive solution would be to update the model every time a new fuel type is introduced by measuring the desired properties for samples of the new fuel and incorporating the data into the calibration data set. Neither of these solutions, however, solve the underlying problem.

A promising method would be to use some form of data abstraction: that is, to take the input data and calculate abstract, or meta, variables to be used in the prediction model. The abstraction step would essentially seek to find a basis that is common to a wide range of fuels, and thus, would enable the model to describe a wide range of fuels. Cramer et al. [8] attempted to do this by identifying hundreds of different groups of similar compounds in the fuels with the hope that different fuels would at least have some of the compounds in common, allowing more accurate predictions. They showed that extrapolation accuracy was improved when using this method, although overall the accuracy was still poorer for those samples outside the compositional range of the calibration set. Additionally, detailed GC-GC-MS data is required to obtain this level of information.

With infrared spectra the problem is that the location of peaks for the same functional group or bond can shift location and that peaks for different bonds can overlap. A logical data abstraction scheme for solving this problem would be to reduce spectra to the concentrations of different bonds in the sample. This might be accomplished, for instance, by taking one of the existing large databases of spectra for pure compounds and creating a model that can identify which peaks or regions are caused by a specific bond type. A neural network might be able to perform this type of analysis. Although fuels contain thousands of different compounds, they contain only a relatively small number of different bond types or functional groups, and this would be a basis that would apply to a wide range of fuels. This type of data abstraction would be analogous to group contribution methods already used in property prediction.

12.CONCLUSIONS

Thermodynamic and transport properties can usually be predicted from infrared spectra to comparable or better accuracy than correlations based on bulk, or average, properties. Additionally, almost all of the models had accuracies that were on the same order of magnitude as the uncertainty of the experimental measurements, and generally were within 1.5 to 3 times the measurement uncertainty. Therefore, the predictions based on infrared spectra could be used in many calculations and as the basis for many technical decisions. However, because the models are created using experimental data they generally cannot be more accurate than the measurement method itself, and so when the best accuracy is required experimental measurements would still be necessary.

The accuracy of a model varied depending on the property predicted. It was found though that the single most important factor in determining the accuracy of the model was the accuracy of the experimental property data which was used in regression. Significantly, how well a property was correlated with infrared spectra did not have a noticeable impact on the accuracy. Therefore, one important conclusion from this is that a wide variety of different properties could be predicted from infrared spectra. Additionally, this indicates that improving the quality of the regression data is one of the best ways to improve the performance of a model. The results also indicate that the accuracy of the input data (infrared spectra) can also set a lower limit on the accuracy that a model can obtain.

Through experiments on temperature dependent properties, it was found that properties can also be predicted at a range of different conditions from infrared spectra. This is an important step for property prediction because many important properties depend on temperature and pressure. The eventual goal would be to predict constants for equations of state using infrared spectra, and the work done here with temperature dependent properties indicates that this can be done. Distributions are also important for characterizing fuel samples (e.g. boiling point distributions and molecular weight distributions), and the method used here may also prove useful in predicting these distributions from infrared spectra.

It is important to remember that although the results of this research are promising, that some challenges need to be addressed to make it more economical to use these infrared predictive models more widely, including in industry. More specifically, a method needs to be identified to transfer models between instruments without the need for standard spectra, and methods need to be developed that will allow models to be extrapolated and give good results for samples with compositions different from the calibration samples. Both of these are topics that could be the subject of future investigations.

REFERENCES

[1] M. R. Riazi, *Characterization and Properties of Petroleum Fractions*. ASTM International, 2005.

[2] M. I. Ahmad, N. Zhang, and M. Jobson, “Molecular components-based representation of petroleum fractions,” *Chemical Engineering Research and Design*, vol. 89, no. 4, pp. 410–420, Apr. 2011.

[3] Z. S. Baird and V. Oja, “Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density,” *Chemometrics and Intelligent Laboratory Systems*, vol. 158, pp. 41–47, 2016.

[4] R. J. Quann, “Modeling the chemistry of complex petroleum mixtures.,” *Environ Health Perspect*, vol. 106, no. Suppl 6, pp. 1441–1448, Dec. 1998.

[5] J. Gomez-Prado, N. Zhang, and C. Theodoropoulos, “Characterisation of heavy petroleum fractions using modified molecular-type homologous series (MTHS) representation,” *Energy*, vol. 33, no. 6, pp. 974–987, Jun. 2008.

[6] P. Hosseinifar, M. Assareh, and C. Ghotbi, “Developing a new model for the determination of petroleum fraction PC-SAFT parameters to model reservoir fluids,” *Fluid Phase Equilibria*, vol. 412, pp. 145–157, Mar. 2016.

[7] K. M. Watson and E. F. Nelson, “Improved Methods for Approximating Critical and Thermal Properties of Petroleum Fractions,” *Industrial and Engineering Chemistry*, vol. 25, no. 8, pp. 880–887.

[8] J. A. Cramer, M. H. Hammond, K. M. Myers, T. N. Loegel, and R. E. Morris, “Novel Data Abstraction Strategy Utilizing Gas Chromatography–Mass Spectrometry Data for Fuel Property Modeling,” *Energy & Fuels*, vol. 28, no. 3, pp. 1781–1791, Mar. 2014.

[9] S. Lee, *Oil Shale Technology*. CRC Press, 1990.

[10] V. Oja and E. M. Suuberg, “Oil Shale Processing, Chemistry and Technology,” in *Fossil Energy*, R. Malhotra, Ed. Springer New York, 2013, pp. 99–148.

[11] K. Urov and A. Sumberg, *Characteristics of oil shales and shale-like rocks of known deposits and outcrops: monograph*. Tallinn: Estonian Acad. Publ, 1999.

[12] World Energy Council, *World Energy Resources*. London: World Energy Council, 2013.

[13] K. Brendow, “Global oil shale issues and perspectives (Synthesis of the Symposium on Oil Shale held in Tallinn (Estonia) on 18 and 19 November 2002),” *Oil Shale*, vol. 20, no. 1, pp. 81–92, 2003.

[14] “2010 Survey of Energy Resources,” World Energy Council, London, 2010.

[15] S. H. Guo, "The chemistry of shale oil and its refining," in *Coal, Oil Shale Natural Bitumen, Heavy Oil and Peat – Vol. II*, vol. 2, Publishers Company Limited, 2009, pp. 94–106.

[16] V. Oja, R. Rooleht, and Z. S. Baird, "Physical and Thermodynamic Properties of Kukersite Pyrolysis Shale Oil: Literature Review," *Oil Shale*, vol. 33, no. 2, pp. 184–197, Apr. 2016.

[17] P. N. Kogerman and A. Kõll, *Physical properties of Estonian shale oils*. Tartu, Estonia, 1930.

[18] K. Luts, *The Estonian Oil Shale Kukersite, its Chemistry, Technology and Analysis (Der estländische Brennschiefer-Kukersit, seine Chemie, Tehnologie und Analyse)*. Tartu, Estonia: K. Mattiesens Buchdruckerei Ant.-Ges., 1934.

[19] D. K. Kollerov, *Physicochemical properties of oil shale and coal liquids (Fiziko-khimicheskie svojstva zhidkikh slantsevykh i kamenougol'nykh produktov)*. Moscow, 1951.

[20] D. B. Hibbert, P. Minkkinen, N. M. Faber, and B. M. Wise, "IUPAC project: A glossary of concepts and terms in chemometrics," *Analytica Chimica Acta*, vol. 642, no. 1–2, pp. 3–5, May 2009.

[21] William R. Hruschka and H. Martens, "Principal Component Analysis Predicts Protein and Moisture Content from Near Infrared Spectra of Ground Wheat," Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy," presented at the Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, Atlantic City, NJ, USA, 1982, p. Paper No. 375.

[22] I. Frank, J. Feikema, N. Constantine, and B. Kowalski, "Prediction of Product Quality from Spectral Data Using the Partial Least-Squares Method," *J. Chem. Inf. Comput. Sci.*, vol. 24, no. 1, pp. 20–24, Feb. 1984.

[23] J. A. Persson, E. Johansson, and C. Albano, "Quantitative thermogravimetry of peat. A multivariate approach," *Anal. Chem.*, vol. 58, no. 6, pp. 1173–1178, May 1986.

[24] R. M. Balabin and S. V. Smirnov, "Interpolation and extrapolation problems of multivariate regression in analytical chemistry: benchmarking the robustness on near-infrared (NIR) spectroscopy data," *The Analyst*, vol. 137, no. 7, p. 1604, 2012.

[25] R. E. Morris *et al.*, "Rapid Fuel Quality Surveillance through Chemometric Modeling of Near-Infrared Spectra," *Energy Fuels*, vol. 23, no. 3, pp. 1610–1618, Mar. 2009.

[26] H. Chung, "Applications of Near-Infrared Spectroscopy in Refineries and Important Issues to Address," *Applied Spectroscopy Reviews*, vol. 42, no. 3, pp. 251–285, May 2007.

[27] S. Satya, R. M. Roehner, M. D. Deo, and F. V. Hanson, "Estimation of Properties of Crude Oil Residual Fractions Using Chemometrics," *Energy Fuels*, vol. 21, no. 2, pp. 998–1005, Mar. 2007.

[28] J. Coates, “Interpretation of Infrared Spectra, A Practical Approach,” in *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Ltd, 2006.

[29] N. Colthup, *Introduction to Infrared and Raman Spectroscopy*. Elsevier, 2012.

[30] W. M. Trygstad and D. Horgen, “Motor fuel property prediction by inferential spectrometry: Understanding conditions and limitations,” presented at the 59th Annual Symposium of the Analysis Division, Baton Rouge, LA, USA, 2014.

[31] W. M. Trygstad, R. Pell, and M. Roberto, “Motor fuel property prediction by inferential spectrometry 2. Overcoming limitations,” presented at the ISA 60th Analysis Division Symposium, Galveston, TX, USA, 2015.

[32] P. Geladi and B. R. Kowalski, “Partial least-squares regression: a tutorial,” *Analytica Chimica Acta*, vol. 185, pp. 1–17, Jan. 1986.

[33] J. H. Kalivas, “Interrelationships of multivariate regression methods using eigenvector basis sets,” *J. Chemometrics*, vol. 13, no. 2, pp. 111–132, Mar. 1999.

[34] R. D. Tobias, “An introduction to partial least squares regression,” in *Proceedings of the twentieth annual SAS users group international conference*, 1995, pp. 1250–1257.

[35] R. M. Balabin, E. I. Lomakina, and R. Z. Safieva, “Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy,” *Fuel*, vol. 90, no. 5, pp. 2007–2015, May 2011.

[36] F. Lindgren, P. Geladi, and S. Wold, “The kernel algorithm for PLS,” *J. Chemometrics*, vol. 7, no. 1, pp. 45–59, Jan. 1993.

[37] A. J. Smola and B. Schölkopf, “A Tutorial on Support Vector Regression,” Berlin, Germany, NC2-TR-1998-030, 1998.

[38] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[39] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[40] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[41] C. A. Micchelli, *Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions*. IBM Thomas J. Watson Research Division, 1984.

[42] N. Dyn, “INTERPOLATION OF SCATTERED DATA BY RADIAL FUNCTIONS,” in *Topics in Multivariate Approximation*, Elsevier, 1987, pp. 47–61.

[43] G. F. Smits and E. M. Jordaan, “Improved SVM regression using mixtures of kernels,” in *Neural Networks, 2002. IJCNN’02. Proceedings of the 2002 International Joint Conference on*, 2002, vol. 3, pp. 2785–2790.

[44] U. Thissen, M. Pepers, B. Üstün, W. J. Melssen, and L. M. C. Buydens, “Comparing support vector machines to PLS for spectral regression

applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 73, no. 2, pp. 169–179, Oct. 2004.

[45] U. Thissen, B. Üstün, W. J. Melssen, and L. M. Buydens, "Multivariate calibration with least-squares support vector machines," *Analytical chemistry*, vol. 76, no. 11, pp. 3099–3105, 2004.

[46] He Cheng, Yang Zengling, Chen Longjian, Huang Guangqun, Liao Na, and Han Lujia, "Influencing factors of on-line measurement of straw-coal blends using near infrared spectroscopy," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 30, no. 9, pp. 192–200, May 2014.

[47] P. R. Griffiths and J. A. D. Haseth, *Fourier Transform Infrared Spectrometry*. John Wiley & Sons, 2007.

[48] "Technology." [Online]. Available: http://www.spectralengines.com/learn-more/technology. [Accessed: 04-Nov-2016].

[49] M. Tuohiniemi, "Micromechanical tunable Fabry-Perot interferometer and a method for producing the same," US8913322 B2, dets-2014.

[50] N. Golubev, "Solid oil shale heat carrier technology for oil shale retorting," *Oil Shale*, vol. 20, no. 3, pp. 324–332, 2003.

[51] "ASTM D86, Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure," ASTM International, West Conshohocken, PA, USA, 2012.

[52] Z. S. Baird, V. Oja, and O. Järvik, "Distribution of Hydroxyl Groups in Kukersite Shale Oil: Quantitative Determination Using Fourier Transform Infrared (FT-IR) Spectroscopy," *Applied Spectroscopy*, vol. 69, no. 5, pp. 555–562, May 2015.

[53] P. N. Kogerman, *On the Chemistry of the Estonian Oil Shale "Kukersite."* Tartu, Estonia: Oil Shale Research Laboratory, 1931.

[54] R. Rannaveski, O. Järvik, and V. Oja, "A new method for determining average boiling points of oils using a thermogravimetric analyzer," *J Therm Anal Calorim*, pp. 1–10, Jun. 2016.

[55] "ASTM D2224 - Method of Test for Mean Molecular Weight of Mineral Insulating Oils by the Cryoscopic Method," ASTM International, West Conshohocken, PA, USA, 1983.

[56] O. Järvik and V. Oja, "Molecular weight distributions and average molecular weights of pyrolysis oils from oil shales: Literature data and measurements by SEC and ASAP MS for oils from four different deposits," *Energy Fuels*, Nov. 2016.

[57] A. Aarna and V. Paluoja, "Determination of Hydroxyl Groups in Shale Oil by the Acetylation Method," in *Analytical Methods for Oil Shale and Oil Shale Products*, Tallinn, Estonia, 1961, pp. 23–26.

[58] "ASTM D97-16, Standard Test Method for Pour Point of Petroleum Products," ASTM International, West Conshohocken, PA, USA, 2016.

[59] "ASTM D2892, Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column)," ASTM International, West Conshohocken, PA, USA, 2016.

[60] H. Luik, "Chemicals and Other Products From Shale Oil," in *Coal, Oil Shale Natural Bitumen, Heavy Oil and Peat*, vol. 2, EOLSS Publishers Company Limited, 2009, pp. 107–128.

[61] "ASTM D7483 - 13a, Standard Test Method for Determination of Dynamic Viscosity and Derived Kinematic Viscosity of Liquids by Oscillating Piston Viscometer," ASTM International, West Conshohocken, PA, USA, 2013.

[62] J. E. Bertie, R. N. Jones, Y. Apelblat, and C. D. Keefe, "Infrared Intensities of Liquids XIII: Accurate Optical Constants and Molar Absorption Coefficients Between 6500 and 435 cm^{-1} of Toluene at 25°C, from Spectra Recorded in Several Laboratories," *Appl. Spectrosc., AS*, vol. 48, no. 1, pp. 127–143, Jan. 1994.

[63] T. Gao and V. Jojic, "Degrees of Freedom in Deep Neural Networks," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States, 2016, pp. 232–241.

[64] S. Macho and M. S. Larrechi, "Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry," *TrAC Trends in Analytical Chemistry*, vol. 21, no. 12, pp. 799–806, Dec. 2002.

[65] B. Zadrozny, "Learning and Evaluating Classifiers Under Sample Selection Bias," in *Proceedings of the Twenty-first International Conference on Machine Learning*, New York, NY, USA, 2004, p. 114–.

[66] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, Aug. 2012.

[67] J.-H. Jiang, R. J. Berry, H. W. Siesler, and Y. Ozaki, "Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data," *Anal. Chem.*, vol. 74, no. 14, pp. 3555–3565, Jul. 2002.

[68] Y. P. Du, Y. Z. Liang, J. H. Jiang, R. J. Berry, and Y. Ozaki, "Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares," *Analytica Chimica Acta*, vol. 501, no. 2, pp. 183–191, Jan. 2004.

[69] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, Dec. 2015.

[70] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[71] R. Storn and K. Price, “Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces,” *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359.

[72] E. Jones, E. Oliphant, and P. Peterson, “SciPy: Open Source Scientific Tools for Python.” 2001.

[73] P. K. Huang, “Characterization and thermodynamic correlations for undefined hydrocarbon mixtures,” *Ph.D. Thesis*, 1977.

[74] H. W. Bearce and E. L. Peffer, “Density and thermal expansion of American Petroleum oils,” in *Technological Papers of the Bureau of Standards*, Washington D.C.: United States Department of Commerce, 1916.

[75] M. R. Lipkin and S. S. J. Kurtz, “Temperature Coefficient of Density and Refractive Index for Hydrocarbons in the Liquid State,” *Industrial and Engineering Chemistry*, vol. 13, pp. 291–294, 1941.

[76] Y. L. Rastorguev, “Methods of Assessing Fuel and Oil Quailty,” *Khimiya i Tekhnologiya Topliv i Masel*, no. 9, pp. 56–60, 1971.

[77] B. Esteban, J.-R. Riba, G. Baquero, A. Rius, and R. Puig, “Temperature dependence of density and viscosity of vegetable oils,” *Biomass and Bioenergy*, vol. 42, pp. 164–171, Jul. 2012.

[78] J. A. Gray, C. J. Brady, J. R. Cunningham, J. R. Freeman, and G. M. Wilson, “Thermophysical properties of coal liquids. 1. Selected physical, chemical, and thermodynamic properties of narrow boiling range coal liquids,” *Industrial & Engineering Chemistry Process Design and Development*, vol. 22, no. 3, pp. 410–424, 1983.

[79] C. J. Seeton, “Viscosity–temperature correlation for liquids,” *Tribol Lett*, vol. 22, no. 1, pp. 67–78, Jun. 2006.

[80] G. S. Fulcher, “Analysis of Recent Measurements of the Viscosity of Glasses,” *Journal of the American Ceramic Society*, vol. 8, no. 6, pp. 339–355, Jun. 1925.

[81] W. A. Wright, “AN IMPROVED VISCOSITY-TEMPERATURE CHART FOR HYDROCARBONS,” *Journal Materials*, vol. 4, no. 1, pp. 19–25, Mar. 1969.

[82] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, “Measuring and testing dependence by correlation of distances,” *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[83] “Distance Correlation in Python,” *Gist*. [Online]. Available: https://gist.github.com/satra/aa3d19a12b74e9ab7941. [Accessed: 15-May-2017].

[84] M. L. Rizzo and G. J. Szekely, *energy: E-Statistics: Multivariate Inference via the Energy of Data*. 2016.

[85] S. A. Channiwala and P. P. Parikh, “A unified correlation for estimating HHV of solid, liquid and gaseous fuels,” *Fuel*, vol. 81, no. 8, pp. 1051–1063, May 2002.

[86] M. R. Riazi and T. E. Daubert, “Characterization parameters for petroleum fractions,” *Ind. Eng. Chem. Res.*, vol. 26, no. 4, pp. 755–759, Apr. 1987.

[87] R. N. Feudale, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, and J. Ferré, “Transfer of multivariate calibration models: a review,” *Chemometrics and Intelligent Laboratory Systems*, vol. 64, no. 2, pp. 181–192, 2002.

[88] T. B. Blank, S. T. Sum, S. D. Brown, and S. L. Monfre, “Transfer of near-infrared multivariate calibrations without standards,” *Analytical chemistry*, vol. 68, no. 17, pp. 2987–2995, 1996.

[89] J. B. Cooper, C. M. Larkin, and M. F. Abdelkader, “Calibration transfer of near-IR partial least squares property models of fuels using virtual standards: Virtual standards calibration transfer,” *Journal of Chemometrics*, vol. 25, no. 9, pp. 496–505, Sep. 2011.

[90] M. F. Abdelkader, J. B. Cooper, and C. M. Larkin, “Calibration transfer of partial least squares jet fuel property models using a segmented virtual standards slope-bias correction method,” *Chemometrics and Intelligent Laboratory Systems*, vol. 110, no. 1, pp. 64–73, Jan. 2012.

[91] Y. Wang, D. J. Veltkamp, and B. R. Kowalski, “Multivariate instrument standardization,” *Anal. Chem.*, vol. 63, no. 23, pp. 2750–2756, Dec. 1991.

[92] E. Bouveresse, C. Hartmann, D. L. Massart, I. R. Last, and K. A. Prebble, “Standardization of Near-Infrared Spectrometric Instruments,” *Anal. Chem.*, vol. 68, no. 6, pp. 982–990, Jan. 1996.

[93] M. Forina *et al.*, “Transfer of calibration function in near-infrared spectroscopy,” *Chemometrics and Intelligent Laboratory Systems*, vol. 27, no. 2, pp. 189–203, Feb. 1995.

ACKNOWLEDGEMENTS

To understand these acknowledgements you have to understand one thing: this PhD thesis is, for me, just a footnote in the otherwise much more important story of life. And on the personal level, this thesis is stained in the emotions and experiences that have swirled around it. Success and joy, certainly, but also disappointment and disillusionment.

You see, for me, this thesis also symbolizes some of the things that are wrong in the world. I resent the fact that from now on I will be judged by a few letters next to my name or by a stupid piece of paper with a university seal rather than by who I am. And it has been disappointing to see how students are sometimes treated as just a number, just a product that needs to be made, packaged and shipped before the due date.

Now most things in life are not all bad and not all good, and I have also met those who took time to truly understand and care. And doctoral degrees are also connected to some great ideals, such as the process of thoroughly studying a question and making evidence based conclusions. I can acknowledge many people who have brought these sorts of good things into my life.

To all my colleagues at TTÜ, Oliver, Rivo, Sven, Inna, Ilme and many others, thank you for your help and friendship. I have enjoyed many good conversations with you that have made me a better person.

To my supervisor, Vahur, thank you for giving me the chance to come work at TTÜ and for showing me the value of doing thorough and accurate science.

I must also acknowledge the head of the Department of Energy Technology, Andres Siirde, for his help and for being an all-around good person.

Obviously, I owe much to my family. Thank you Mom and Dad for giving me a good start in life. Thank you Piret for being an awesome companion on this epic journey. And thank you kiddos for helping me slow down and see the garbage trucks and ants around us.

And above all, I must thank my Heavenly Father, who has done more to help me grow than anyone else.

ABSTRACT

Because experimentally measuring fuel properties is usually time consuming and costly, property values are often estimated using correlations. And yet, even when predicting these values input data must still be measured. This makes methods based on infrared spectra an attractive alternative because infrared spectra can be measured quickly and continuously, even within an existing process.

Although fuel properties have long been predicted from infrared spectra, the focus has been almost entirely on quality parameters of the fuel. Instead, this research focused specifically on determining the properties that are needed for modeling a fuel's behavior in industrial processes and in the environment. In this work the predictive models based on infrared spectra were created using machine learning techniques. Support vector regression was the main method used. Models were created for predicting 11 different properties: specific gravity, the refractive index parameter, average boiling point, average molecular weight, carbon content, hydrogen content, sulfur content, hydroxyl group content, pour point, the thermal expansion coefficient and viscosity. Experimental data for creating the models were measured as part of a larger project.

The performance of the infrared models was compared to that of conventional bulk (or average) property correlations commonly used. Predictions based on infrared spectra had a comparable or better accuracy for all the properties compared, except viscosity. This shows that predictive methods based on infrared spectra can perform well enough to be used as a substitute for conventional predictive methods. Further analysis also indicated that the main factor limiting the accuracy of the models was the accuracy of the experimental data used in regression.

Because many important fuel properties vary with temperature or pressure, and predicting properties at different temperatures was also investigated. To model temperature dependence, the coefficients of an algebraic equation were found using infrared spectra. The results showed that the models obtained can be used to predict temperature dependent properties over a wide range of temperatures, which allowed density and viscosity to be predicted with accuracies comparable to the models for predictions at a single temperature.

An eventual goal would be to predict parameters for equations of state from infrared spectra. As the results presented here indicate, this is possible. However, limitations discussed in this work need to be taken into account.

KOKKUVÕTE

Kuna kütuste omaduste katseline määramine on tihtipeale aeganõudev ja kulukas, siis saadakse omaduste väärtused tihti hinnanguliselt korrelatsioonidega. Samas, ka hinnanguliste väärtuste saamine nõuab lähteandmete mõõtmist. Seetõttu oleks alternatiivina kasulikud infrapunaspektrite põhised metoodikad, kuna infrapunaspektreid saab mõõta kiiresti ja pidevalt, ja seda isegi olemasolevas protsessis.

Kütuste omadusi on hinnatud infrapunaspektritel põhinevate mudelite abil juba kaua aega, kuid seni on keskendatud peaaegu täielikult kütuste kvaliteediparameetritele. Käesolev töö keskendub aga nende omaduste määramisele, mis on vajalikud kütuste käitumise modelleerimiseks tööstusprotsessides ja keskkonnas. Töös olevate infrapunaspektritel põhinevate mudelite loomiseks kasutati masinõppe meetodeid. Peamiseks kasutatud metoodiks oli tugivektorregressioon. Mudelid loodi 11 erineva omaduse määramiseks: suhteline tihedus, murdmisnäitaja parameeter, keskmine keemispunkt, keskmine molaarmass, süsinikusisaldus, vesinikusisaldus, väävlisisaldus, hüdroksüülrühmade sisaldus, hangumispunkt, paisumiskoefitsient ja viskoossus. Andmed mudelite saamiseks mõõdeti katseliselt suurema projekti raames.

Infrapunaspektritel põhinevate mudelite efektiivsust võrreldi üldiselt kasutatavate, keskmistel omadustel põhinevate korrelatsioonidega. Ilmnes, et infrapunaspektrite kaudu arvutatud erinevate omaduste väärtused olid sama täpsed või täpsemad kõikide omaduste puhul, v.a viskoossus. Seega, infrapunaspektri põhised mudelid võivad olla piisavalt head, et asendada korrelatsioonvõrrandeid. Sügavamale analüüsile tuginedes võib väita, et peamine, mis piirab mudelite täpsust, on mudelite koostamiseks kasutatud katseandmete täpsus.

Kuna paljud tähtsad kütuste omadused sõltuvad temperatuurist ja rõhust, siis uuriti käesolevas töös ka infrapunaspektritel põhinevate mudelite kasutamist omaduste temperatuursõltuvuse määramiseks. Temperatuursõltuvuse mudeli saamiseks leiti infrapunaspektri abil algebralisse võrrandisse koefitsiendid. Tulemused näitasid, et saadud mudelid on kasutatavad laias temperatuuripiirkonnas ka temperatuurist sõltuvate omaduste hindamiseks, võimaldades hinnata tihedust ja viskoosust sama täpsusega, kui mudelid, mille abil hinnati vastava omaduse väärtust kindlal temperatuuril.

Edasine eesmärk oleks rakendada infrapunaspektril põhinevaid mudeleid olekuvõrrandi parameetrite määramiseks. Nagu käesolevas töös saadud tulemused näitavad, on seda võimalik teha. Seejuures tuleb aga arvestada töös kirjeldatud piirangutega.

ARTICLE 1

Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density

Published in Chemometrics and Intelligent Laboratory Systems, vol. 158, pp. 41–47, 201.

Chemometrics and Intelligent Laboratory Systems 158 (2016) 41–47

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density

Zachariah Steven Baird, Vahur Oja *

Department of Chemical Engineering, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia

ARTICLE INFO

Article history:
Received 3 February 2016
Received in revised form 20 July 2016
Accepted 7 August 2016
Available online 10 August 2016

Keywords:
Density
Temperature dependence
Fourier-transform infrared spectroscopy (FTIR)
Oil shale
Phenol rich oil
Chemometrics

ABSTRACT

Although the use of chemometric methods to predict fuel quality properties has received wide attention over the past three decades, as seen from the review included with this article, no studies were found about predicting temperature dependent properties of fuels. Since our research is focused on determining thermodynamic properties, rather than quality properties, taking temperature dependencies into account became even more important. To determine if accurate predictions could be obtained over a range of temperatures, the densities of over 300 fuel samples (mostly narrow boiling range oil fractions, considered here as pseudocomponents) were measured and predicted. An alternative fuel (a phenol-rich oil shale oil) was studied because the property prediction methods developed for conventional petroleum samples often give poor results for this and other alternative fuels. The temperature dependence of density for these fuel samples was modelled using a linear equation based on the density at 20 °C and the slope of the density-temperature relationship. Support vector regression was used to predict these parameters for each sample from its infrared spectrum. Then these parameters were used to predict the densities at other temperatures. Densities spanned the range from 0.713 to 1.088 g/cm^3, and the root mean squared error of the predicted values was 0.004660 g/cm^3, which is a relative error of less than 1%. In addition to the experimental portion, a literature review is included, which contains an assessment of the accuracy of chemometric methods for predicting many fuel properties.

1. Introduction

Chemometric methods have been used for predicting fuel quality properties since at least 1986 [1], and there have been hundreds of articles on the subject. Chemometrics generally refers to the use of statistical or mathematical methods to extract information about a chemical system from measurements made on that system [2]. Usually large sets of measured data are used (for example, all the data points in a spectrum), which necessitate advanced, computer-based statistical or mathematical methods. The driving force behind the interest chemometric methods is that standard laboratory methods for assessing fuel quality are time consuming, and rapid methods for evaluating fuel properties allow fuel quality to be determined much more quickly and at a lower cost. Also, the same methods can be used for online measurements in refineries and blending operations to improve process optimization. Several patents have also been granted for these types of methods [3–7] and they have also been implemented in some petroleum production operations, which attest to the usefulness of these predictive techniques.

In order to more quantitatively summarize the previous research about using chemometrics for fuel property prediction, we have performed an analysis type review (containing graphs and tables) of the literature on the subject, which is presented in Section 2. To get this thorough overview, we reviewed 341 scientific articles and 5 patent applications which were published between 1986 and 2015. For each study we identified the properties predicted, the input data used, the fuel types examined, the regression method used and the error of the predictive models created. A spreadsheet containing all this information collected during the review is included as a supplement to this article. Although there are certainly more articles available on the subject, 346 is a large enough sample size to give a good representation of the basic trends of the research done in this field.

The overview indicates a growing interest in the topic, with the number of articles published on the topic showing an increasing trend. A wide range of fuels and properties have been examined, and although most studies have used infrared spectra as input data, many other analytical techniques have also been investigated. Models with a

* Corresponding author.
E-mail address: vahur.oja@ttu.ee (V. Oja).

http://dx.doi.org/10.1016/j.chemolab.2016.08.004
0169-7439/

good degree of accuracy have been created for many different properties (see Table 1). Overall though, developments in this field have been quite modest. The major finding, that a wide range of physical and quality properties can be reliably predicted from conventional chemical data (especially infrared spectra), was already established by the earliest articles, and had actually been shown earlier in studies on wheat [8] and tape [9]. Later articles contributed mostly by extending the same techniques to other types of fuels or to other properties, and there has been a fair amount of overlap between articles, with multiple articles investigating the same properties for similar types of fuels and using similar methods.

Nevertheless, a few noteworthy articles do stand out. Balabin et al. [10] investigated the performance of different regression methods when they are used to predict the properties of samples outside the range used for model calibration. Also, a group from the U.S. Naval Research Laboratory has produced a couple thorough articles about predicting fuel properties [11,12]. In one of their more recent articles [12] they attempted to create a model that was more robust to changes in fuel composition by basing the model on the content of representative compounds, as identified by gas chromatography-mass spectroscopy. Some studies have also contributed to addressing obstacles encountered with online industrial models, such as model stability and updating [13–16] and the effect of spectrometer positioning [17]. And, Cooper et al. [18,19] introduced a new model for transferring fuel property models between spectrometers. And as this subfield is connected to the larger field of chemometrics as a whole, some of the articles reviewed were directed more towards improving chemometric methods by presenting a new algorithm or technique.

So far research has mainly been focused on quality parameters, but chemometric methods could also be of wide interest for estimating other properties, including the thermodynamic and transport properties of liquid fuels. In this role, it could be an alternative to the

Table 1
The fuel properties most commonly predicted in the articles reviewed and statistics about the accuracy of the models created for these properties. The "number of articles" column shows the total number of articles found during the review that predicted that property. All error values are root mean squared (RMSE) errors.

Property	Number of articles	Models included in accuracy statistics	Min	1st Quartile	Median	3rd Quartile	Max	Unit
Acid number	15	8	0.003	0.016	0.137	0.208	0.32	mg KOH/g
Aniline point	3	–	–	–	–	–	–	
Api gravity	11	5	0.24	0.25	0.265	0.731	0.811	Degrees API
Aromatics	58	35	0.05	0.57	0.71	1.47	2.7	vol% or wt%
Ash content	22	16	0.01	0.56	0.91	1.94	3.8	wt%
Ash fusion temperature	5	3	70	–	77	–	90	°C
Asphaltenes	15	13	0.13	0.42	0.69	2.08	2.6	wt%
Boiling point	8	4	1.29	2.45	3.23	4.66	5	°C
Carbon content	16	10	0.8	1.19	1.9	2.41	4.1	wt%
Carbon residue	10	6	0.16	0.24	0.94	2	2	wt%
Cetane index	22	15	0.2	0.43	0.6	1.44	2	
Cetane number	24	16	0.27	0.49	1.01	1.95	2.11	
Chemical composition[a]	44	–	–	–	–	–	–	
Cloud point	8	4	2.92	2.98	3.05	3.1075	3.11	°C
Cold filter plugging point	3	2	0.77	–	0.89	–	1	°C
Conductivity	3	2	0.84	–	–	–	77	pS/m
Contaminant content	17	25	0.01	0.4	1.1	1.88	4.1	vol% or wt%
Density	79	50	0.0004	0.001	0.0018	0.003	0.028	g/cm^3
Distillation temperatures	68	124	0.08	1.72	3	5.3	23	°C
Fixed carbon	10	8	0.78	0.88	1.72	3.58	4.6	wt%
Flash point	38	28	0.69	1.96	3.37	5	10.44	°C
Fluid system icing inhibitor	4	3	0.007	–	0.009	–	0.01	vol%
Freezing point	24	14	1.26	1.57	2.18	2.99	7	°C
Grindability	4	2	3.84	–	4.16	–	4.47	
Heat of combustion	38	21	0.013	0.134	0.3	0.585	1.9	MJ/kg
Hydrogen content	26	14	0.016	0.073	0.12	0.245	0.92	wt%
Hydrogen/carbon ratio	4	2	0.048	–	0.069	–	0.089	
Iodine number	7	4	0.51	0.65	0.8	1.28	1.4	g I2/100 g
Isoparaffins	4	4	0.17	0.51	0.64	1	1.12	wt%
Lubricity	7	–	–	–	–	–	–	
Mixture composition[b]	38	45	0.01	0.2	0.52	1.32	10.2	vol% or wt%
Moisture content	17	13	0.09	0.41	0.69	0.99	2.5	wt%
Naphthenes	11	10	0.2	0.25	0.38	0.52	1.9	vol% or wt%
Nitrogen content	17	11	0.002	0.048	0.1	0.35	0.55	wt%
Octane number	40	25	0.07	0.2	0.31	0.51	1.6	
Olefins	18	12	0.07	0.21	0.3	2.44	4.3	vol% or wt%
Other properties	51	–	–	–	–	–	–	
Oxygen content	6	3	0.003	–	0.728	–	2.16	wt%
Paraffins	10	9	0.12	0.24	0.4	0.67	1.3	vol% or wt%
Pour point	14	9	2.06	2.4	4.65	9.1	9.2	°C
Refractive index	5	4	0.0003	0.0004	0.0008	0.0012	0.0012	nD
Resins	9	5	0.26	0.75	0.75	1.13	1.46	wt%
Saturates	24	16	0.56	0.63	0.9	1.69	1.7	vol% or wt%
Specific gravity	5	3	0.0004	–	0.0005	–	0.0008	
Sulfur content	45	30	0.00012	0.01483	0.034	0.25	1.6	wt%
Thermal stability	4	2	46	–	46.7	–	47.3	degF
Vapor pressure	11	6	1.04	1.77	3.27	4.6	5.99	kPa
Water content	15	8	0.003	0.005	0.229	0.57	0.78	vol% or wt%
Viscosity	52	31	0.026	0.098	0.152	0.242	22	cSt
Vitrinite reflectance	4	2	0.1	–	0.13	–	0.15	
Volatile matter	15	10	0.5	1	1.4	3.7	7	wt%
Yield	9	–	–	–	–	–	–	

[a] Chemical composition was used to categorize articles where the concentration of a specific chemical or family of chemicals was predicted.

[b] Mixture composition refers to articles where the proportions of different types of fuel in a blend were predicted.

commonly used undefined mixture approach, where simple correlative models based on easily measurable bulk properties are used to estimate physical, thermodynamic and transport properties [20]. Although some thermodynamic properties are also quality properties, and therefore, have been predicted, no articles were found that focused on predicting thermodynamic properties. Additionally little, if any, research has been done on predicting temperature dependent properties over a range of temperatures.

In this study, we sought to extend the use of chemometric prediction methods to temperature dependent thermodynamic properties. More specifically, we attempted to predict the density of fuel samples from their infrared spectra. Fourier transform infrared (FTIR) spectroscopy, is the most commonly used analytical method for these types of fuel property predictions (see Fig. 3). It is a convenient method for which devices are widely available, and infrared spectra correlate well with many properties.

We used oil shale oil as the fuel for this study, which is produced from oil shale by pyrolysis [21]. Additionally, most of the measurements were made on fractions with narrow boiling ranges, which were obtained from the initial wide fractions by distillation. These narrow fractions (often called pseudocomponents) are often used in thermodynamic studies of liquid fuels to be able to better understand the range of compounds present in the oil, and therefore, enable better predictions [20]. This fractionation process gave samples that characterize much of the variation in composition and properties that occur in the shale oil studied.

Density is a fundamental temperature dependent property that can be measured very accurately and has been predicted with good accuracy as a temperature independent property (i.e. density at one specified temperature, usually at 20 or 15.6 °C) using chemometric methods (see Fig. 2). Additionally, for most liquid fuels the temperature dependence of density is linear at moderate temperatures, which is easy to model. For these reasons density was chosen as the property investigated in this study. Additionally, infrared spectra correlate quite well with density, likely because density is closely related to the chemical structure of a sample.

Although PLS was shown to be the most popular regression method (see Fig. 4), we decided to use support vector regression (SVR). SVR has the advantage of being able to take into account nonlinearities through the use of kernel functions, and Balabin et al. [10,22] reported that SVR gives more accurate models. We created initial models with PLS and found that for our data SVR also gave superior results.

2. Analysis of literature about predicting fuel properties

2.1. Fuel types investigated

Fig. 1 indicates that a wide range of fuels have been investigated so far, including various conventional petroleum oils and refinery products, biodiesels and biodiesel-diesel blends, biomass samples, coal, Fischer Tropsch fuels, ethanol, oil shale, natural gas, coal liquids and even rocket fuel. The number of articles on each fuel type is shown in Fig. 1, and as seen, the majority of studies have focused on conventional petroleum samples. This is likely because diesel, gasoline and jet fuel are readily available and commercially important. Liquid biofuels (which essentially consists of biodiesel and ethanol) have also received a lot of attention, which probably mirrors the increasing interest in them in the research community at large.

2.2. Properties predicted

Essentially all of the research so far has been directed towards predicting fuel quality parameters, which is likely due to the immediate applicability of these models in industry. Table 1 gives an overview of the parameters measured, and shows that a wide range of properties have been predicted. In fact, Table 1 only includes the properties which were predicted in more than 2 sources. In total we identified 104 different properties that had been measured, and there are certainly more that have been measured in articles not reviewed for this paper. As could be expected, more attention has been given to the properties that are used most frequently and to those which are important for petroleum samples.

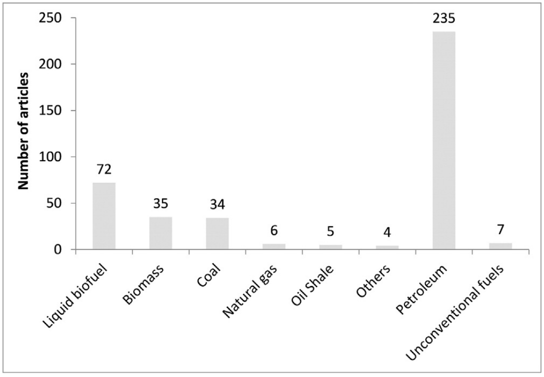

Fig. 1. The fuel types studied in the articles about predicting fuel properties using chemometric methods. The unconventional fuels category includes coal liquids, oil shale oil and Fischer-Tropsch fuels. The others category includes charcoal and rocket fuel.

2.3. Accuracy of the model

Some general observations about the accuracy of these predictive methods can also be made. Although a thorough investigation of model performance requires detailed information about the samples, measurement methods and regression techniques used, a quick glance at the error levels reported gives us some initial insights. Accuracy statistics for many properties are given in Table 1, and those statistics were calculated using only the models for which error data was found (Column 2 shows the total number of articles which predicted each property). Here we will look more closely at the distribution for liquid density. 47 articles were selected which gave the RMSE of the model for density. 3 articles gave results for 2 separate models, which gave a total of 50 density models. The number of models falling within given RMSE ranges is shown in Fig. 2.

Immediately one density model stood out as having large errors (0.028 g/cm^3), and it was noticed that this model was created for crude oil residual fractions [23] for which it is more difficult to measure the density. This indicates that the quality of the reference data can affect a model's accuracy. For the remaining models, there is also a rough

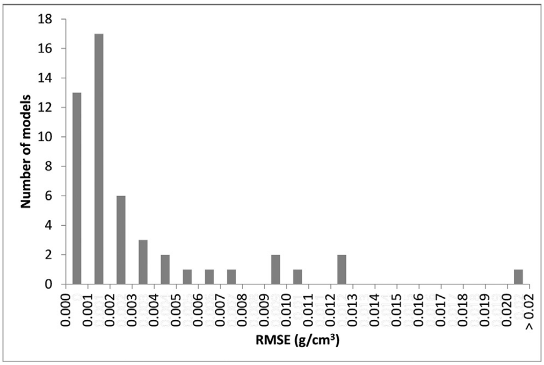

Fig. 2. The distribution of the prediction models for fuel liquid density compared in this review according to their root mean squared error.

correlation between the type of fuel and the accuracy of the model. The density models with the lowest errors tend to be for fuels like diesel, biodiesel and gasoline, which are generally easier to measure and cover a narrower range of densities and compositions. Those with the highest errors are for samples like crude oils, residual oils or a broad range of fuels (fuels from many sources worldwide, including from nonpetroleum sources). Without more detailed information, it is difficult to identify exactly what other factors led to the different accuracies seen here. The skill of the analyst, the type of data processing used and the specifics of how a regression method was implemented could all potentially affect the accuracy of the resulting model. However, it is evident that two models for the same property can have quite different accuracies.

2.4. Analytical method used for input data

When looking at the input data upon which the models are based, we see that most studies have chosen to use some form of infrared spectra, as shown in Fig. 3. This is probably due to the advantages of infrared spectroscopy, which include a quick measurement time, no need for sample preparation, applicability to a wide range of samples, good results for predicting most properties, easy extension to continuous online measurements and a cost that is lower than many other methods. Although infrared spectra are the most popular, many other types of data have been used. Only the most frequently used methods are shown in Fig. 3, and grouped into the other category is a whole list of methods, including using devices such as a thermal wave interferometer and an electronic nose.

2.5. Regression technique used

Fig. 4 shows that partial least squares (PLS) regression has been the most popular regression technique. The PLS was used more frequently than all the other methods combined. PLS has been the main method used in chemometrics since the early days of the field, which is probably why so many studies used it. As chemometric methods have developed, researchers have gradually switched to using newer methods. Initially, principal component regression and multiple linear regression (ordinary least squares regression) were used fairly frequently. More recently, techniques that can model nonlinearity are gaining favor, including nonlinear versions of PLS (poly-PLS, spline PLS, kernel PLS), artificial neural networks and support vector regression. It is also worth noting that even within each category a variety of algorithms and modifications have been used.

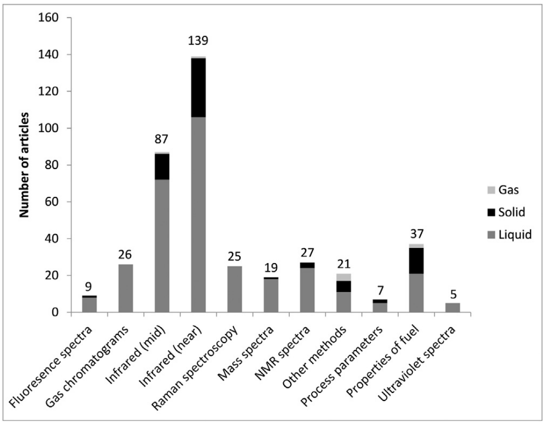

Fig. 3. The types of input data used in articles about predicting fuel properties using chemometric methods. Each type is further divided to show the proportion from different types of fuels (solid, liquid or gas).

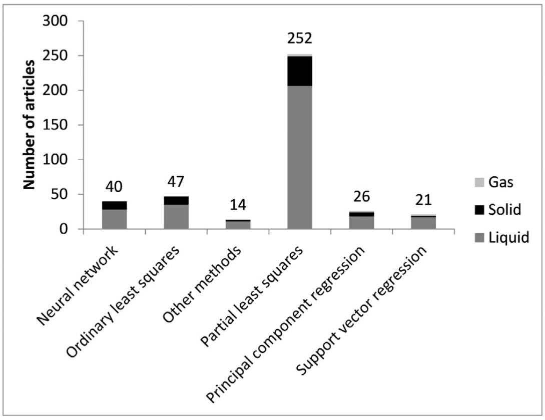

Fig. 4. The regression methods used in articles about predicting fuel properties using chemometric methods.

3. Experimental

3.1. Samples

Oil shale oil was used for this study, which is a synthetic crude oil created by pyrolysis of the solid organic matter in oil shale. Kukersite oil shale (Estonia) was the source oil shale for all of the samples used in this study. Due to the molecular structure of Kukersite, its oil has large quantities of phenolic compounds [24,25]. Almost all of the samples were obtained from Estonian Energy's Narva Oil Plant (Narva, Estonia). This plant uses a commercial-scale solid heat carrier retort (Galoter process) [26]. Two additional samples were created by retorting oil shale in a laboratory-scale Fischer retort [27] (which is a method commonly used in oil shale studies to evaluate the oil production potential of an oil shale).

The oil produced at the Narva Oil Plant is separated into three wide technical fractions: gasoline, fuel oil and heavy oil. The densities of these wide fractions were measured, and then these oils were further separated into narrow boiling range fractions by distillation. Most of the distillations were simple distillations performed at either atmospheric pressure (an Engler distillation [28]) or in a vacuum. Two distillations were also performed using a rectification column. The narrow fractions were generally taken at distillation intervals of 5–10 °C. One sample was also created to correspond to crude Galoter oil by mixing the three fractions (gasoline, fuel oil and heavy oil) according to their respective proportion in the crude oil (20:60:20, as given in the plant's design documents). This was done because it was not possible to obtain a crude oil sample directly from the process.

To further investigate the role of the polar phenolic compounds on the properties of the oil, phenols were extracted from several samples. The extraction was carried out using a 10% NaOH solution, and resulted in a phenol-rich sample and a neutral oil (dephenolated) sample [25]. 10 narrow fractions were separated in this manner, as were several fuel oil fractions. The phenolic and dephenolated fuel oil fractions were then separated into narrow fractions by distillation. Also, these phenolic and dephenolated fuel oils were mixed with original fuel oil to create samples with varying contents of phenolic compounds. All of this is, of course, very complicated, but the important points are that a wide variety of shale oil fractions were used and that the fractions had a wide range of hydroxyl group content (from essentially 0 to 10.2 wt% OH). In total, 327 samples were used for the current study. 13 of these were wide industrial fractions or crude shale oil samples, and the remaining were narrow boiling fractions.

3.2. Measurements

Densities as a function of temperature were measured using an Anton Paar DMA 5000 (Anton Paar GmbH, Graz, Austria). Densities were measured between 15.6 and 90 °C. Heavier fractions were measured at higher temperatures and then their densities at 20 °C were calculated by extrapolation, as is commonly done for samples that are highly viscous at 20 °C. Densities spanned the range from 0.713 to 1.088 g/cm^3. The manufacture states the accuracy of the devices as being ± 0.000005 g/cm^3, but for oil samples we expect the accuracy is a little bit poorer (about ± 0.00002 g/cm^3) due to small errors associated with sampling. Additionally, heavier fractions are opaque and viscous, so for them the likelihood that an air bubble is in the measurement cell increases. Before each day of measurements, the accuracy of the device was checked with water and air, and between each measurement the device was checked with air.

Infrared spectra were measured using an Interspec 301-X spectrometer fitted with an ATR accessory (Interspectrum OÜ, Tõravere, Estonia). The ATR accessory had a single reflection, ZnSe internal reflection element. Interspec for Windows software (version 3.40 Pro, Interspectrum OÜ, Tõravere, Estonia) was used to collect the spectra. Spectra were obtained between 600 and 4000 cm^{-1} at a resolution of 1 cm^{-1}. Each spectrum was an average of 5 to 10 scans (except for the heavy oil fractions, which were scanned only once). The baseline of each spectrum was then corrected using a cubic spline interpolation based on 4 points: 3999, 3796, 2200 and 1800 cm^{-1}. For a few spectra the correction was poor due to noise, but good baselines were obtained for these spectra by slightly shifting the points that were taken in order to avoid the noise. The baseline correction was performed using Essential FTIR software (version 3.10.016, Operant LLC, VA, USA).

3.3. Data analysis and model development

First, the data was examined for errors and outliers. A few samples were found to have measurement errors and were remeasured. Additionally, a group of phenolic samples was identified as outliers due to their higher than normal densities and their different composition. After creating an initial model, a couple more outliers were identified. These proved to be caused by measurement errors because upon remeasuring the residuals for these samples were significantly lower. There were three samples with higher residuals that could not be remeasured because those samples were no longer available, and thus, the data for these samples was left unchanged. In total, data from 327 samples was used to develop the models.

Oils generally display a linear density-temperature relationship at moderate temperatures (temperatures below the boiling region of the sample). Therefore, for this study the change in density with temperature was modelled with a simple linear equation shown as Eq. (1)

$$\rho_T = \rho_* - \gamma(T - T_*) \tag{1}$$

where ρ_T is the density at temperature T, ρ_* is the density at reference temperature T_* and γ is a constant that describes the slope of the density-temperature relationship. The reference temperature used in this study was 20 °C. This linear equation can be used for any liquid fuel that exhibits a linear temperature dependence for density over the temperature range of interest. This seems to apply, at moderate temperatures, for most liquid fuel types, including petroleum [29–31], biofuels [32], oil shale oil [33] and coal liquids [34]. Generally, the relationship becomes nonlinear as the sample gets close to its boiling region, so the linear model would likely work over a longer range for heavier fuels and a shorter range for light fuel products like gasoline. Therefore, this technique could also be applied to other fuels besides shale oil, if the experimental databased used to create the model includes samples from these other fuel types.

Table 2
Parameters used for support vector regression.

Property	Density model	Slope (γ) model
C	48.02915706	7.246754595
Gamma	$4.1683544258 \cdot 10^{-6}$	0.000393627
Epsilon	$1.02093930394 \cdot 10^{-5}$	$3.34174597874 \cdot 10^{-5}$
Kernel function	Radial basis function	Radial basis function

The data for each sample was fit using Eq. (1), and the γ coefficients (slopes) were obtained as a result. This linear equation fit the data quite well. The root mean squared error (RMSE) of Eq. (1) for the data in this study was only 0.0001054 g/cm^3.

Then models were created to predict the density at 20 °C and the slope (γ) using support vector regression (SVR). The parameters used for SVR are shown in Table 2. The data was mean centered and scaled according to the standard deviation of the data. SVR was implemented in Python (version 2.7) using the Scikit-learn package (version 0.15) [35]. The parameters were optimized based on the 5-fold cross validation value. Therefore, the accuracy of the models was estimated using a second, outer cross validation loop with 50 folds. For the density model the three regression parameters and the variables (wavelengths) used were optimized simultaneously using a genetic algorithm. The genetic algorithm was also performed in Python, and the DEAP package was used (version 1.0.2) [36]. The same variables used for the density model were also used for the slope model, and the regression parameters were optimized using the SciPy brute force algorithm [37].

4. Results and discussion

4.1. Model for density at 20 °C

The model for density had good accuracy, with an RMSE of 0.004628 g/cm^3. This accuracy, however, is still much lower than the

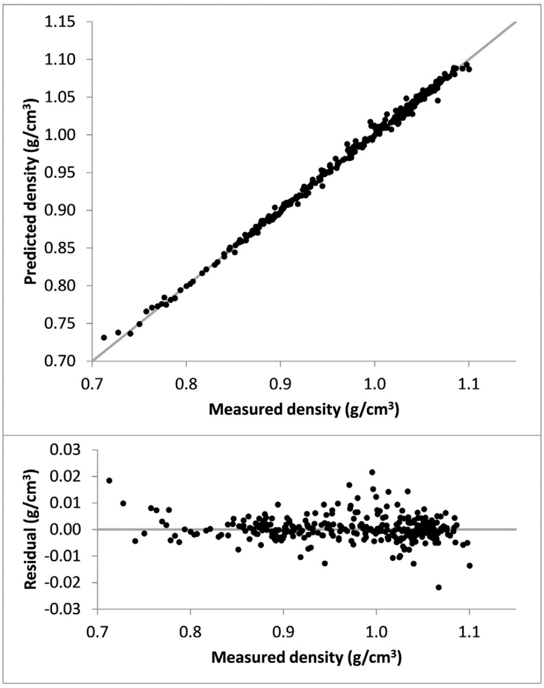

Fig. 5. Performance of the support vector regression model for density at 20 °C.

measurement accuracy of approximately ±0.00002 g/cm^3. The deviation for each sample is shown in Fig. 5. Although the density model is accurate, its RMSE is still higher than those reported in previous studies (shown in Fig. 2). The third quartile of the RMSEs given in previous studies was 0.003 g/cm^3 (see Table 1), and therefore, the accuracy of our model was lower than at least 75% of those earlier models. The larger error may be cause by the wider range of samples spanned by this model, which covers a density range much wider than those covered by many other studies. Also, since the densities at 20 °C for the heaviest fractions were calculated by extrapolation from data at higher temperatures, the density data for these fractions includes more uncertainty, which could also decrease the overall effectiveness of the model. Additionally, there are several samples that have relatively large residuals. The two lightest gasoline fractions likely show large residuals because when left out during cross validation then there were no other comparable fractions, which decreased the accuracy of the model for that end of the range. Also, it is suspected that a couple of the large residuals were due to measurement errors. Some samples were remeasured and their data corrected, but a couple of the samples had already been disposed of and could not be remeasured.

4.2. Model for a range of temperatures

As mentioned, the temperature dependence of density was described using a linear equation (Eq. (1)). To create a multi-temperature SVR model, a second SVR model was created to predict the slope coefficient (γ). The slope model had an RMSE of $7.569 \cdot 10^{-6}$ g/cm^3/°C. The residuals for each sample are shown in Fig. 6. The largest residual is for one of the samples that could not be remeasured, and it is likely that a measurement error occurred because it gives large residuals for both models. The large residual at the high end of the scale is for the lightest gasoline fraction. As mentioned before, this large residual is likely due to the fact that there were no other fractions with slopes as high as this.

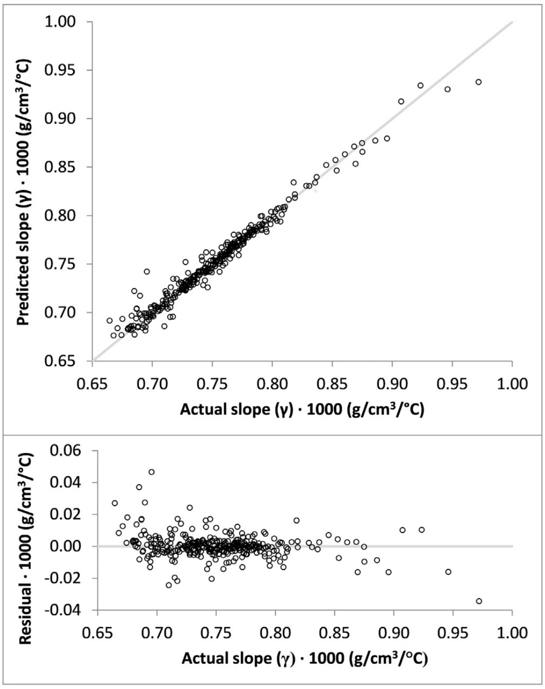

Fig. 6. Performance of the support vector regression model for the slope coefficient (γ).

By using the predicted density and slope values in Eq. (1), density values were predicted at temperatures for which experimental data had been measured. The RMSE between the predicted and actual density values was 0.004660 g/cm^3, which is an error of < 1%. There was also a small bias of 0.0001839 g/cm^3 in the predictions.

When looking at the RMSE by temperature (shown in Table 3), it can be seen that the error tends to be similar regardless of the temperature. The variation in RMSE from one temperature to another seems more related to the samples included than the temperature, and this variation occurs because the samples were not all measured at the same temperatures. We also looked at the results for some individual fractions, and again saw that the deviation from the experimental values are essentially the same regardless of the temperature. This can be explained by the fact that the change in density with temperature is relatively small compared to the value of density itself. For instance, over the temperature range from 20 to 80 °C, the change in density was less than 10% for all of the samples measured here. Thus, the relatively small influence of temperature over the measured range decreases the impact of the error from the slope model, and therefore, most of the prediction error comes from the model for density at 20 °C.

To see how the accuracy of the multi-temperature model compares to that of a model for only one specific temperature, we created two additional one-temperature models for 50 and 80 °C. As seen in Table 3, the one-temperature models and the multi-temperature model had comparable accuracies. The added source of error for the multi-temperature modelling method described here is the error associated with the equation used to characterize the temperature dependent behavior. If that underlying equation has an accuracy better than that of a single-temperature chemometrics model, then it is likely that the multi-temperature model will perform as accurately as a model for only one temperature. Additionally, the correlation between the input data and the equation coefficients is also important. Here we used density and the slope (which is closely related to the thermal expansion coefficient), and because these values are related to the chemical composition of a sample, a model based on chemical information (FTIR spectra) was able to provide good results.

Although density can be predicted quite well with the method used here, it is reasonable to expect that the same method may give poorer results for other properties. Predictive models cannot be expected to perform more accurately than the reference data on which they are based, and as mentioned, when using equation constants as the reference data, the accuracy of that equation also affects the accuracy of the reference data. It is expected that for properties with more complex temperature dependences (such as viscosity and vapor pressure) the accuracy will be lower.

5. Conclusions

Based on the results given here, it can be concluded that chemometrics can also be used to predict some temperature dependent properties over a range of temperatures. The method used here, i.e. using chemometric methods to estimate the coefficients of an equation that describes the temperature dependence of a property, gave results

Table 3
The root mean squared error of predicted values for density at a given temperature.

Temperature	Multi-temperature model	One-temperature models
°C	g/cm^3	g/cm^3
15.6	0.003966	
20	0.004628	0.004628
35	0.004716	
40	0.004345	
50	0.005211	0.00603
60	0.004286	
65	0.005036	
80	0.00485	0.005577

that were just as accurate as models developed for the property at one specific temperature. The experiment presented here involved predicting the density of fuel samples over a range of temperatures from infrared spectra. The resulting RMSE of the predictions was 0.004660 g/cm^3, which is an error of less than 1%. These results are, of course, specific to this experiment, but it is likely that some other temperature dependent properties could also be predicted using this method.

Acknowledgements

The authors gratefully acknowledge funding provided 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”, and by the Estonian Minister of Education and Research under target financing SF0140022s10.

Appendix A. Sources used for the literature review

The information collected during the literature review process is available online as a supplement to this article. Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.chemolab.2016.08.004.

References

[1] J. Persson, C. Albano, Quantitative Thermogravimetry on Peat, 1986 1173–1178.

[2] D.B. Hibbert, P. Minkkinen, N.M. Faber, B.M. Wise, IUPAC project: a glossary of concepts and terms in chemometrics, Anal. Chim. Acta 642 (2009) 3–5, http://dx.doi.org/10.1016/j.aca.2009.02.020.

[3] H.W. Chu, C. Lu, C.H. Huang, S.Y. Fu, Mobile fuel analysis apparatus and method thereof, US 20080272303 A1, 2008.

[4] S. Farquharson, W.W. Smith, Method and apparatus for determining properties of fuels, US 20100211329 A1, 2010.

[5] R.H. Clarke, Hydrocarbon analysis based on low resolution Raman spectral analysis, US 5139334 A, 1992.

[6] J.B. Cooper, R.R. Bledsoe, K.L. Wise, M.B. Sumner, W.T. Welch, B.K. Wilt, Process and apparatus for octane numbers and reid vapor pressure by Raman spectroscopy, US 5892228 A, 1999.

[7] L. May, J. Gonzalez, V. Sanchez, Use of NIR spectra for property prediction of bio-oils and fractions thereof, US 8911512 B2, 2014.

[8] W.R. Hruschka, H. Martens, Principal component analysis predicts protein and moisture content from near infrared spectra of ground wheat, Pittsbg. Conf. Anal. Chem. Appl. Spectrosc., Atlantic City, NJ, USA 1982, p. 375.

[9] I.E. Frank, J. Feikema, N. Constantine, B.R. Kowalski, Prediction of product quality from spectral data using the partial least-squares method, J. Chem. Inf. Model. 20–24 (1984).

[10] R.M. Balabin, S.V. Smirnov, Interpolation and extrapolation problems of multivariate regression in analytical chemistry: benchmarking the robustness on near-infrared (NIR) spectroscopy data, Analyst 137 (2012) 1604, http://dx.doi.org/10.1039/c2an15972d.

[11] R.E. Morris, M.H. Hammond, J.A. Cramer, K.J. Johnson, B.C. Giordano, K.E. Kramer, et al., Rapid fuel quality surveillance through chemometric modeling of near-infrared spectra, Energy Fuel 23 (2009) 1610–1618, http://dx.doi.org/10.1021/ef800869t.

[12] J.A. Cramer, M.H. Hammond, K.M. Myers, T.N. Loegel, R.E. Morris, Novel data abstraction strategy utilizing gas chromatography–mass spectrometry data for fuel property modeling, Energy Fuels 28 (2014) 1781–1791, http://dx.doi.org/10.1021/ef4021872.

[13] M. Garci-a-Menci-a, An empirical approach to update multivariate regression models intended for routine industrial use, Fuel 79 (2000) 1823–1832, http://dx.doi.org/10.1016/S0016–2361(00)00046–6.

[14] M. C., S. K., B. H., A unified recursive just-in-time approach with industrial near infrared spectroscopy application, Chemom. Intell. Lab. Syst. 135 (2014) 133–140, http://dx.doi.org/10.1016/j.chemolab.2014.04.007.

[15] K. He, H. Cheng, W. Du, F. Qian, Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy, Chemom. Intell. Lab. Syst. 134 (2014) 79–88, http://dx.doi.org/10.1016/j.chemolab.2014.03.007.

[16] K. He, F. Qian, H. Cheng, W. Du, A novel adaptive algorithm with near-infrared spectroscopy and its application in online gasoline blending processes, Chemom. Intell. Lab. Syst. 140 (2015) 117–125, http://dx.doi.org/10.1016/j.chemolab.2014.11.006.

[17] C. He, Z. Yang, L. Chen, G. Huang, N. Liao, L. Han, Influencing factors of on-line measurement of straw-coal blends using near infrared spectroscopy, Trans. Chin. Soc. Agric. Eng. 30 (2014) 192–200.

[18] J.B. Cooper, C.M. Larkin, M.F. Abdelkader, Calibration transfer of near-IR partial least squares property models of fuels using virtual standards, J. Chemom. 25 (2011) 496–505, http://dx.doi.org/10.1002/cem.1395.

[19] M.F. Abdelkader, J.B. Cooper, C.M. Larkin, Calibration transfer of partial least squares jet fuel property models using a segmented virtual standards slope-bias correction method, Chemom. Intell. Lab. Syst. 110 (2012) 64–73, http://dx.doi.org/10.1016/j.chemolab.2011.09.014.

[20] M.R. Riazi, Characterization and Properties of Petroleum Fractions, first ed. ASTM International, 2005.

[21] S. Lee, Oil Shale Technology, CRC Press, Boca Raton, FL, USA, 1991.

[22] R.M. Balabin, E.I. Lomakina, Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? Phys. Chem. Chem. Phys. 13 (2011) 11710–11718, http://dx.doi.org/10.1039/c1cp00051a.

[23] S. Satya, R.M. Roehner, M.D. Deo, F.V. Hanson, Estimation of properties of crude oil residual fractions using chemometrics, Energy Fuel 21 (2007) 998–1005, http://dx.doi.org/10.1021/ef0601420.

[24] S. Derenne, C. Largeau, E. Casadevall, J.S. Sinninghe-Damste, E.W. Tegelaar, J.W. de Leeuw, Characterization of Estonian Kukersite spectroscopy and pyrolysis: evidence for abundant alkyl phenolic moieties in an Ordivician, marine, type Il/l kerogen, Org. Geochem. 16 (1990) 873–888, http://dx.doi.org/10.1016/0146-6380(90)90124-I.

[25] Z.S. Baird, V. Oja, O. Järvik, Distribution of hydroxyl groups in kukersite shale oil: quantitative determination using Fourier Transform Infrared (FT-IR) Spectroscopy, Appl. Spectrosc. 69 (2015) 555–562, http://dx.doi.org/10.1366/14-07705.

[26] N. Golubev, Solid heat carrier technology for oil shale retorting, Oil Shale 20 (2003) 324–332.

[27] ISO 647:1974, Brown Coals and Lignites - Determination of the Yields of Tar, Water, Gas and Coke Residue by Low Temperature Distillation, 2009.

[28] ASTM, D86, Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure, 2012.

[29] H.W. Bearce, E.L. Peffer, Density and thermal expansion of American petroleum oils, Technol. Pap. Bur. Stand. United States Department of Commerce, Washington, D.C., USA, 1916.

[30] M.R. Lipkin, S.S.J. Kurtz, Temperature coefficient of density and refractive index for hydrocarbons in the liquid state, Ind. Eng. Chem. 13 (1941) 291–294.

[31] Y.L. Rastorguev, Methods of assessing fuel and oil quailty, Khimiya I Tekhnologiya Topl. I Masel. 56–60 (1971).

[32] B. Esteban, J.R. Riba, G. Baquero, A. Rius, R. Puig, Temperature dependence of density and viscosity of vegetable oils, Biomass Bioenergy 42 (2012) 164–171, http://dx.doi.org/10.1016/j.biombioe.2012.03.007.

[33] D.K. Kollerov, Physico-chemical Properties of Liquid Shale and Coal Products, St. Petersburg, Russia, 1951.

[34] J.A. Gray, C.J. Brady, R. Cunnlngham, R. James, G.M. Wllson, Thermophysical properties of coal liquids. 1. Selected physical, chemical, and thermodynamic properties of narrow boillng range coal liquids, Ind. Eng. Chem. Process. Des. Dev. 22 (1983) 410–424.

[35] F. Pedregosa, G. Varoquaux, Scikit-learn: machine learning in python, J. Mach. 12 (2011) 2825–2830 (http://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf.).

[36] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, C. Gagne, DEAP : evolutionary algorithms made easy, J. Mach. Learn. Res. 13 (2012) 2171–2175.

[37] E. Jones, E. Oliphant, P. Peterson, et al., SciPy: open source scientific tools for python, http://www.scipy.org2001 (accessed January 5, 2016).

ARTICLE 2

Physical and Thermodynamic Properties of Kukersite Pyrolysis Shale Oil:
Literature Review

Published in Oil Shale, vol. 33, no. 2, pp. 184–197, 2016.

Oil Shale, 2016, Vol. 33, No. 2, pp. 184–197
doi: 10.3176/oil.2016.2.06

ISSN 0208-189X

PHYSICAL AND THERMODYNAMIC PROPERTIES OF KUKERSITE PYROLYSIS SHALE OIL: LITERATURE REVIEW

VAHUR OJA[*], RUTH ROOLEHT, ZACHARIAH S. BAIRD

Department of Chemical Engineering, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia

Abstract. *This study presents a literature review of the physical and thermodynamic properties of kukersite oil shale oil (or "synthetic crude oil") as found in public literature. The work showed that although there is nearly a century-old history of shale oil production in Estonia, there are very few data about the thermodynamic properties and only a limited number of property prediction methods related to shale oil produced from kukersite. Publicly available information on the physical and thermodynamic properties of kukersite shale oil originates mainly from the period of 1930 to 1960. The data found are predominantly for the lighter part of the synthetic crude oil, i.e. the part for which the condensation temperatures of the atmospheric distillation curve (average atmospheric boiling points of the fractions) are lower than 300–350 °C. Data and studies can be found about several main physical and thermodynamic properties, such as specific gravity, atmospheric boiling point, molecular weight, enthalpy of vaporization at the boiling point, heat capacity, thermal conductivity, viscosity, surface tension and vapor pressure. But in general, this information is not a systematic set of data intended for determination of thermodynamic properties, but rather it lays out trends and supports the simplest approaches for calculating thermodynamic properties based on "undefined" pseudocomponents (a mixture of compounds that behave similarly).*

Keywords: *kukersite, retort oil, thermodynamic properties, physical properties.*

1. Introduction

Estimates of the depletion of global oil reserves have led to research into the potential for using various alternative resources. One alternative is crude oil produced from oil shale (i.e. a synthetic crude oil or synthetic petroleum), and it is estimated that oil shale resources are equivalent to 2.8–3.3 trillion

[*] Corresponding author: e-mail *vahur.oja@ttu.ee*

barrels of oil [1]. Thus, oil shale resources contain approximately three times more oil than conventional petroleum reserves (conventional oil reserves contain about 1.2 trillion barrels) [2]. The technologies for obtaining oil from oil shale are based on the thermal decomposition of kerogen (the cross-linked macromolecular organic matter in oil shale) [3]. Low temperature pyrolysis up to about 500 °C (also referred to as retorting, semicoking or low temperature carbonization) has historically been the preferred thermo-chemical conversion process for oil shales with high oil yield per organic matter. During the low temperature pyrolysis or retorting process the organic matter is converted to oil, gas and solid residue. Over the course of the development of shale oil production technologies several hundred different types of retorts (technologies) have been invented, including in situ (below ground) and ex situ (above ground) retorting technologies. However, throughout the long history of oil shale utilization above ground, i.e. ex situ, retorting technologies have been the only production methods used commercially for producing oil from oil shale [3].

The whole crude shale oil produced via oil shale retorting is a complex mixture of hydrocarbons and organic compounds containing heteroatoms, just like petroleum or coal liquids. Crude shale oil from commercial ex situ retorts, or so-called "synthetic crude oil", can be classified by API gravity as an average heavy crude oil. Crude shale oils from different oil shales have wide boiling distributions (generally less than 30% can be distilled below 300 ^{o}C at atmospheric pressure) and wide molecular weight distributions (extending up to 800–1000 g/mol) [4–7]. Usually about 50% of the oil can be used directly as fuel oil. Generally, crude shale oil from ex situ retorts has characteristics in the following ranges: hydrogen/carbon ratio 1.2–1.6, average molecular weight 190–310 daltons, specific gravity 0.8–1.04 (mostly < 1), 30–70% boils higher than 350 ^{o}C, groups of chemical compounds such as nonaromatic hydrocarbons up to 60%, aromatic hydrocarbons 10–50% and heteroatomic compounds 20–60% [3, 4]. Thus, based on its characteristics, crude shale oil is situated somewhere between crude petroleum and coal liquids. Shale oil is more aromatic than petroleum, but not as aromatic as coal liquids [8]. At the same time, shale oil usually contains more olefins than coal liquids and petroleum, and can also contain more heteroatomic organic compounds [4]. Additionally, because shale oil's composition is specific to a given oil shale deposit, shale oils also contain different amounts of heteroatoms, depending on the composition of heteroatoms in the organic matter of the parent oil shale. For example, kukersite shale oil (Baltic basin) contains more than 5% oxygen, El Lajjun shale oil (Jordan) contains up to 10% sulfur, Green River shale oil (USA) contains 2% nitrogen [4]. Therefore, due to the composition of shale oils, the accuracy of using petroleum-based empirical methods for determining the physical and thermodynamic properties of a given shale oil is questionable, at least without a corresponding evaluation of the method's applicability.

Shale oil has been produced in Estonia for almost a century and many different processes (technologies) have been used [9]. The technologies used in industry have been retort generators (Kiviter process) (since 1925), tunnel ovens (1926–ca 1980), Davidson rotating retorts (1931–1961), chamber ovens (1948–1987) and solid heat carrier retorts (Galoter process) (since 1963). From the different retorts and different process regimes kukersite shale oil with somewhat different parameters is obtained, in terms of both physical and chemical properties. Compared to the group composition of petroleum, kukersite contains more olefins and aromatic hydrocarbons, especially in lighter fractions (fractions with lower boiling points) [10–12]. The main component of petroleum is various paraffins, whereas shale oil generally contains few paraffins. One unique attribute of kukersite shale oil is the high content of phenolic compounds (over 30%) [13]. These differences in composition suggest that before using petroleum-based prediction methods it would be necessary to access the accuracy of these methods for kukersite shale oil. Thus, the present article is a literature review of publicly available information about the physical and thermodynamic properties of shale oil produced from kukersite.

2. Analysis and discussion

2.1. General overview of kukersite shale oil studies

The literature review indicated that the most systematic experimental data on the thermodynamic properties of kukersite shale oil and the physical properties necessary for predicting those properties was measured by Kogerman and Kõll at the beginning of the last century [14]. This assessment is made with the objective of predicting thermodynamic properties in mind, since for prediction a thermodynamic or physical property of interest should be related to at least two conveniently measurable properties (of which one describes preferably molecular size and another energy, or structure). The experimental data was presented in 1930 in the book "Physical properties of Estonian shale oils" [14]. Data was given for narrow boiling fractions that were taken at 25 °C intervals from the whole crude oil. The crude shale oil was produced with one specific retorting technology – a Kiviter-type experimental retort (Kohtla-Järve experimental generator). The data presented cover fractions with average boiling points in the range of 150–300 °C. This book provides average property data of the fractions, but not correlations and relationships. The data given in the book contains specific gravity and viscosity at different temperatures (20, 30, 40, 50, 60, 70 ^{o}C), molecular weight and specific heat and surface tension at 20 ^{o}C. And additionally, the fractions' average atmospheric boiling points, thermal expansion coefficients and heats of vaporization at the boiling point were found by calculation. No chemical characteristics, such as elemental composition, amounts of different functional groups or compound classes (paraffins, olefins, aromatics, etc.), were provided for these fractions. It is worth noting that Kogerman's and

Kõll's work was published before the beginning of the systematic characterization of the thermodynamic and physical properties of petroleum-based hydrocarbons. The beginning of that process could be considered the year 1933, when Watson and Nelson developed two empirical figures that contained the dependence of molecular weight on boiling temperature and the characterization parameter K_w or API gravity [15]. The data obtained by Kogerman and Kõll are partially or entirely given in later compilations, for example, in the appendix of Kogerman's own 1931 monograph "On the chemistry of the Estonian oil shale kukersite" (the appendix is titled "Physical properties of Estonian oil shale") [16]; Luts's 1934 book "Der estländische Brennschiefer-Kukersit, seine Chemie, Tehnologie und Analyse" (in German) [17] and Kollerov's compilation "Fiziko-khimicheskie svoistva zhidkikh slantsevykh i kamenougolnykh produktov" (1951, in Russian) [18].

In the book "Der estländische Brennschiefer-Kukersit, seine Chemie, Tehnologie und Analyse" published by Luts in 1934, one chapter is dedicated to the physical and thermodynamic properties of distillation products from Estonian shale oil [17]. In addition to some data taken from the book by Kogerman and Kõll [14], formulas are given for calculating the heat of combustion and hydrogen content from specific gravity. What could be considered the most important contribution is the correlation for molecular weight based on average boiling point for phenol-free wide technical fractions obtained in the boiling range of about 30–300 °C: light gasoline (34–185 °C), automobile gasoline (46–173 °C), heavy gasoline (132–193 °C), motor fuel (186–247 °C), diesel (243–297 °C). This is similar to the distribution of industrial distillation fractions for petroleum oils: naphtha (boiling range 60–100 °C), gasoline (boiling range 40–205 °C), kerosene (boiling range 175–325 °C), diesel fuel (boiling range 250–350 °C).

The majority of publically available kukersite shale oil data (also referred to more generally as Baltic basin shale oil data) is summarized in Kollerov's compilation "Fiziko-khimicheskie svoistva zhidkikh slantsevykh i kamenougolnykh produktov" [18]. This is a broader compilation on the physical and thermodynamic properties of oils obtained from solid fossil fuels in which data obtained by Kogerman and Kõll [14] is given along with data about kukersite shale oils produced in tunnel ovens, chamber ovens and retort generators (Kiviter process). This data is more abundant for wide technical fractions (gasoline, diesel and other wider oil fractions) than narrow fractions. Again, the fractions are characterized only by average properties and no information on or links to chemical characteristics are provided. Average physical and thermodynamic properties of kukersite shale oil fractions are presented in the book in tables and in many cases represented as graphical and/or equation based relationships. And yet, three-parameter relationships are given only as a few figures, and only for general liquid organic compounds. Because it is a compilation, data is not really presented systematically and the data is not supported with enough additional information to assess its quality. Data, graphs, equations and assessments are presented for kukersite (or Baltic basin) shale oil properties such as specific gravity, molecular mass, boiling point, thermal conductivity, heat capacity, enthalpy of vaporization, vapor pressure and surface tension. A chapter in the book "Khimiya i tekhnologiya produktov pererabotki slantsev", published a few years later, in 1954 (in Russian), gives additional data and linear empirical relationships for the temperature dependence of specific gravity, heat capacity and thermal conductivity of Baltic basin shale oil [19]. The experimental data are on four narrow boiling range fractions and one wider boiling range oil fraction of neutral oxygen-containing oil substances, and on two fractions from chamber oven oil. However, the boiling ranges of these fractions are not specified and the fractions are characterized only by average properties (molecular weight, average boiling point, specific gravity and kinematic viscosities at 20, 50, 75 ^{o}C). This book also contains two more chapters related to properties of general liquid organic compounds, one addressing the Bachinski relationship of viscosity and the other Kollerov's K factor [20, 21].

The most important subsequent overview could be a chapter in the book "Khimiya i tekhnologiya slantsevoj smoly" (1968, in Russian) which, based on Kollerov's book, presents both equations and graphs for predicting the physical and thermodynamic properties of shale oil [22]. Later experimental data on physical and thermodynamic properties, such as boiling points, specific gravities and molecular weights, can be found in a limited form in several works; however, these in and of themselves are not studies about thermodynamic properties, but parts of studies about the chemical composition of shale oil. Worth mentioning is also a later determination of viscosity for wide fractions [23–25].

In conclusion, searching the literature showed that publically available data on the physical and thermodynamic properties of shale oil produced from kukersite oil shale is mostly from the time period between 1930 and 1960. This was the age which was dominated by graphical relationships. Relatively little systematic experimental data was found for narrow boiling range fractions, or data for fractions with a boiling range smaller than 30 ^{o}C (about 50 °F). Experimental data can generally be found for the lighter portion of oil for which the condensation temperatures of the atmospheric distillation curve (the average atmospheric boiling points of the fractions) are lower than 300 ^{o}C. It must also be acknowledged that the respective studies/measurements have not historically been carried out with the development of prediction methods in mind. The data found can acceptably be used for correlations based on undefined fractions, or pseudocomponents described by average parameters, and this only in a relatively limited form.

The next subsection (section 2.2) gives a short overview of some basic aspects related to the prediction of the thermodynamic properties of oil fractions in the context of existing kukersite shale oil data. The following two subsections provide more specific information about two basic compilations of data. In subsection 2.3, a more detailed description of the experimental methods and the original data are presented from Kogerman's and Kõll's "Physical properties of Estonian shale oils" [14]. Subsection 2.4 gives some observations about Kollerov's compilation "Fiziko-khimicheskie svojstva zhidkikh slantsevykh i kamenougolnykh produktov" [18], and the main graphical relationships and equations related to kukersite shale oil are presented as a table.

2.2. Some considerations related to predicting kukersite shale oil properties using available data

As mentioned above, the physical and thermodynamic property data on kukersite shale oil fractions can be found primarily by means of average bulk properties. In connection with the fact that the actual composition of oils cannot generally be quantitatively described at the level of individual components, the use of a method for describing oil as a mixture of discrete pseudocomponents (a mixture of compounds that behave similarly) has been widely adopted [12]. The behavior of each individual pseudocomponent is considered as the behavior of a single compound [12]. Oil can be divided into pseudocomponents based on both molecular size (boiling temperature T_b, number of carbon atoms per molecule N_c) and the groups of compounds for a given molecular size (based on chemical characteristics, for instance n-paraffins, isoparaffins, olefins, naphthenes and aromatics). Therefore, there are essentially two approaches for characterizing a fraction for predicting thermodynamic properties: 1) the undefined mixture approach, or average parameter method, which views narrow boiling fractions (or cuts) as individual pseudocomponents that are described by the fraction's average parameters; 2) the defined mixture approach, which divides a narrow boiling fraction or cut, based on the type of compound, into pseudocomponent classes (for petroleum generally three classes are used: paraffins, naphthenes and aromatics). As mentioned earlier, and it is important to emphasize it again here, the publicly available kukersite shale oil data only support the use of undefined mixture, or average parameter, prediction methods.

For petroleum it has been found that to predict the thermodynamic properties of light petroleum fractions (molecular weight < 300 g/mol, boiling point T_b < 350 ^{o}C) using the average parameter method (undefined mixture method), at least two input parameters are needed (as one-parameter characterization is successful in special cases, for paraffinic crude oil fractions) [12]. It is recommended that these parameters describe the organic component's (single pseudocomponent's) molecular size and energy (or structure) [12]. When these are known, then pseudocomponents can be treated as components whose thermodynamic properties can be calculated using suitable existing equations and correlations. In most cases, average boiling point and specific gravity are used as the two input parameters. One of the simplest ways for dividing oil into undefined pseudocomponents for which the minimum necessary information is known (two known characteristics) is through the constant K factor hypothesis. The K factor characterizes a fuel fraction’s paraffinity and is an empirical relationship between specific gravity and boiling point according to the equation:

$$K_W = \frac{\sqrt[3]{T_b}}{s}. \quad (1)$$

In the most common form of the K factor, called the Universal Oil Products Company (UOP) or Watson characterization factor, T_b is the fraction’s average boiling point in °R and S is its specific gravity at 60 °F (15.5 °C). This K factor (also UOP factor, Watson K factor) was put into practice in 1933 by Watson and Nelson [15]. To use the constant K factor hypothesis it is necessary to know the oil’s boiling point distribution, found from the distillation curve (from this the oil’s mean average boiling point is calculated), and the oil’s overall specific gravity (the specific gravity of the whole crude oil). Based on this hypothesis it is possible to find every fraction’s (or cut’s) specific gravity from its average boiling point on the basis of the value of the K factor calculated from the whole crude oil’s average boiling point and average density (assuming that the K value is the same for all fractions).

To evaluate the constant K factor hypothesis for kukersite shale oil Figure was created. The Figure shows the change in the Watson K factor K_w with boiling point. It also demonstrates that for kukersite oil (data from [14, 18]) the K factor is not constant, rather it decreases rapidly for fractions in the average boiling point range from 50 to 350 °C. At the same time, for

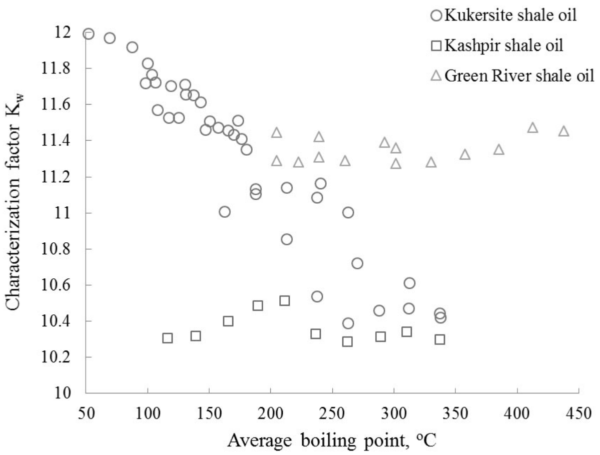

Fig. The change in Watson’s characterization factor K_w with temperature for kukersite oil. Shale oils from Kashpir (Russia, Volga basin) and Green River (USA) oil shales are shown for comparison.

Kashpir (Russia, Volga basin; data from [18]) and Green River (USA; data from [26]) oil shales the constant K factor hypothesis is a quite acceptable approach. This strongly nonconstant behavior could be attributable to the high content of phenolic compounds and their distribution among fractions [13]. Therefore, the minimum information required for describing kukersite oil using the two-parameter undefined pseudocomponent method would be the oil's distillation curve (boiling point curve) and specific gravity curve, or two other acceptable input property curves.

2.3. Shale oil data measured by Kogerman and Kõll

As mentioned earlier, the most systematic experimental data on the thermodynamic properties of kukersite shale oil, when property determination is the goal, was measured by Kogerman and Kõll [14]. To adequately use or interpret the data one needs sufficient information on the details of the experiments (the information not provided sufficiently in the review given in Kollerov's book). To perform their experiments (for determining physical and thermodynamic properties) Kogerman and Kõll used freshly distilled Kohtla shale oil that was produced in a vertical, heat carrier cross-flow Kiviter retort (Kohtla-Järve experimental generator). The shale oil studied was described as a dark brown liquid which had a green fluorescence and an unusual smell. The sample contained 1.19% water, the specific gravity at 18 °C was 1.008, the flame point (Martens-Pensky method) was 91 °C and the viscosity (Engler viscometer) was 6.4 Engler at 50 °C.

To separate the oil into narrow boiling range fractions (cuts) a 5 liter 26 cm long copper flask, which was equipped with a still head, was used. To avoid decomposition at higher temperatures a pressure of 50 mmHg was used. The distillation rate was two drops per second. 8 oil fractions were collected: the first at the initial boiling point up to 150 °C (contained 1.17% water) and then 7 fractions at 25 °C intervals. One significant shortcoming of the data given is the published temperatures of the fractions, which appear to be presented at atmospheric pressure. Neither the fraction temperatures at 50 mmHg nor the calculation of the temperatures at atmospheric pressure from those at vacuum pressures are presented or explained in Kogerman's and Kõll's book [14]. This shortcoming is significant because average boiling point is generally the first choice as an input parameter describing molecular size in property correlations. The data include initial and final condensation temperatures corresponding to fractionation at atmospheric pressure for each fraction, and from these the average boiling points were calculated as the arithmetic mean. After finding the average boiling point the following parameters were measured for every fraction: specific gravity, molecular weight, viscosity, specific heat and surface tension. The thermal coefficient of expansion and the heat of vaporization were also determined by calculation. Because it is necessary to know the measurement method to evaluate the accuracy of the data, these methods are presented in Table 1. In Tables 2–4 the original data from the book by Kogerman and Kõll are given.

Table 1. The measurement and calculation methods used by Kogerman and Kõll [14]

Parameter	Notes
Specific gravity	Measured using a pycnometer and Mohr’s balance. For thermoregulation a water thermostat was used.
Molecular weight	Average molecular weights were determined from the freezing point depression of a stearic acid solution with 2% oil compared to a pure stearic acid solution. The cryoscopic constant was measured using naphthalene and benzoic acid (the average value was 41.8).
Viscosity	Measured using an Ostwald viscometer. Distilled water was used as the reference compound.
Specific heat at 20 °C	Specific heat was determined using a Dewar’s flask that was equipped with a heating coil, mixer and thermometer (accuracy 0.1 °C). For every experiment 200 g of oil was used and the temperature was measured every 30 seconds. The heating period was 2 minutes. Distilled water was used as the standard compound.
Surface tension at 20 °C	Measured using the drop method in relation to air. Surface tension values were calculated using the following formula: $\sigma_{20} = 7.30 \cdot S \cdot \frac{AW}{A0}$, where σ_{20} is the surface tension (mg/mm); S is the specific gravity of oil at 20 °C; A_w is the number of drops of pure water; A_0 is the number of oil drops.
Thermal expansion coefficient	Calculated using the following formula: $a = \frac{a-b}{b \cdot (t'-t)}$, where a and b are the specific gravities at temperatures t' and t, respectively.
Heat of vaporization at the boiling point	Calculated by the Trouton equation using a constant value of 20 cal/mol K as the entropy of vaporization.

Table 2. Data for narrow shale oil fractions obtained from the Kiviter experimental plant (measured by Kogerman and Kõll [14])

Fraction, °C	T_b, °C	d_4^{20}	ΔH, cal/kg	MW, g/mol	σ^{20}, mg/mm	B, 1/°C	C_p^{20}, cal/g °C
150–175	162.5	0.8375	–	126	2.824	0.0009523	–
175–200	187.5	0.8459	69.4	132	2.818	0.0009393	0.548
200–225	212.5	0.8582	68.4	142	2.876	0.0009027	0.504
225–250	237.5	0.8770	60.4	169	2.931	0.0008677	0.507
250–275	262.5	0.8977	60.2	178	2.899	0.0008226	0.500
275-300	287.5	0.9257	56.4	199	2.868	0.0007716	0.502
Kohtla retort	–	–	–	–	3.380	0.0007190	–

Note: T_b – average boiling point; d_4^{20} – specific gravity; ΔH – heat of vaporization at the boiling temperature; MW – average molecular weight; σ^{20} – surface tension at 20 °C; β – average expansion coefficient at 20 °C; C_ρ^{20} – specific heat at 20 °C.

Table 3. The temperature dependence of dynamic viscosity for narrow shale oil fractions obtained from the Kiviter experimental retort (unit is cP) (measured by Kogerman and Kõll [14])

Temperature, °C	Fraction 150–175 °C	Fraction 175–200 °C	Fraction 200–225 °C	Fraction 225–250 °C	Fraction 250–275 °C	Fraction 275–300 °C
20	1.1300	1.3013	1.5960	2.2630	3.7990	8.7410
30	0.9889	1.1033	1.3310	1.8260	2.9240	1.1033
40	0.8557	0.9461	1.1360	1.5060	2.3030	0.9461
50	0.7545	0.8343	0.0977	1.2720	1.8530	8.3430
60	0.5973	0.7332	0.8560	1.0970	1.5540	2.6380
70	0.5973	0.6517	0.7554	0.9430	1.3200	2.1090

Table 4. The temperature dependence of specific gravity (d_4^{20}) for narrow shale oil fractions obtained from the Kiviter experimental retort (measured by Kogerman and Kõll [14])

Temperature, °C	Fraction 150–175 °C	Fraction 175–200 °C	Fraction 200–225 °C	Fraction 225–250 °C	Fraction 250–275 °C	Fraction 275–300 °C
20	0.8375	0.8459	0.8582	0.8770	0.8977	0.9257
30	0.8298	0.8382	0.8507	0.8694	0.8905	0.9187
40	0.822	0.8303	0.8430	0.8620	0.8330	0.9116
50	0.8139	0.8230	0.8356	0.8548	0.8758	0.9047
60	0.8066	0.8149	0.8280	0.8480	0.8694	0.8980
70	0.7987	0.8074	0.8212	0.8403	0.8620	0.8910

2.4. Overview of Kollerov's book

As mentioned earlier, Kollerov's book contains the most extensive information on the thermodynamic and transport properties of kukersite shale oil. More generally, Kollerov's 1951 book "Fiziko-khimicheskie svojstva zhidkih slantsevykh i kamenougolnykh produktov" [18] is a compilation which combined data about the physical and thermodynamic properties of both shale oils and coal oils that were in use in the scientific community in the Soviet Union. A note in connection with this is that in Kollerov's book, the data taken from the publication by Kogerman and Kõll are incorrectly associated with Davidson retort crude oil, but not with Kiviter type experimental retort crude oil. In Kollerov's book, in addition to Kogerman's and Kõll's data, a substantial amount of data is given for Estonian kukersite shale oils produced in industrial tunnel ovens, chamber ovens and Kiviter processes. Most of the data are given for wide boiling range fractions (technical fractions). Data are also provided for dephenolated fractions (fractions from which phenolic compounds have been removed) as well as the phenols. In addition to kukersite shale oil, there are also data given for Kaspir shale oil and coal pyrolysis tars. One shortcoming of the book is that because it is a compilation of data, then information concerning the measurement details of the data is not sufficiently provided. In addition to experimental data, Kollerov's book gives both graphical and equation based options (one-parameter correlations) for determining thermodynamic properties. The emphasis is on graphical relationships because from 1930 to 1960 prediction methods were mainly presented graphically. Table 5 gives the most important relationships from Kollerov's book for kukersite shale oil.

When using the data of the book, the reader should take note of the characterization factor, or the K factor, and the graphs based on it. For characterizing shale oils Kollerov used a K factor (general form of the equation given by Equation 1) where the boiling temperature was in degrees Kelvin and the specific gravity was d_4^{20}. Although the K factor used by

Table 5. An overview of the parameter relationships for kukersite shale oil given in Kollerov's book [18]

Relationship	Notes
Between specific gravity $\left(d_4^{20}\right)$ and average boiling point (T_b)	Graph $d_4^{20} = f(T_b)$. Empirical equation $T_b = f\left(d_4^{20}\right)$.
Between molecular weight (MW) and average boiling point (T_b)	Graph MW = $f(T_b)$. Empirical equations: 1) Luts's equation is given for calculating the molecular weight (for phenol-free oil) MW = T^2 / 1580, where T is the boiling temperature (K) that corresponds to 50 vol% distilled by Engler distillation. 2) For calculating the molecular weight of tunnel oven and retort generator shale oil fractions MW = 59.5 + 0.38*t + 0,0023*(t–0.95)$^{1.9}$, where t is the boiling temperature (°C).
Between specific gravity $\left(d_4^{20}\right)$ and molecular weight (MW)	Graphs $d_4^{20} = f(\mathrm{MW})$.
Between the temperature dependence of the heat capacity and specific gravity $\left(d_4^{20}\right)$	The constants from equation $C_t = C_o + bt$ given as graphs $C_o = f\left(d_4^{20}\right)$ and $b = f\left(d_4^{20}\right)$. Luts's equation is also given in the form C_t = a + 0.0011 (t–20), where t is temperature (°C).
Between enthalpy of vaporization (ΔH) and average boiling point (T_b)	Graphs $\Delta H = f(T_b)$.
Between kinematic viscosity at specific temperatures (ν) and specific gravity $\left(d_4^{20}\right)$	Graphs for viscosity at specific temperatures $\nu = f\left(d_4^{20}\right)$.
Between vapor pressure and temperature	Graphs for wide shale oil fractions: graphed as ln P = $f(1/T)$ lines for different vapor-liquid ratios.

Kollerov has the same general equation form as the Watson K factor (Kw), Watson's and Kollerov's K factors can differ by as much as 17%. This results from the fact that Watson's Kw is calculated using the Rankine temperature unit and the specific gravity at 60 °F (15.56 ^{o}C)

3. Conclusions

This review of data available in public literature shows that although there has been almost a century-long history in Estonia of research related to the production of oil from kukersite oil shale, the information on the thermodynamic properties of oil is quite poor. Although data can be found about basic physical and thermodynamic properties (such as the temperature dependence of specific gravity, atmospheric boiling point, molecular weight and enthalpy of vaporization at the boiling point or temperature dependent properties such as heat capacity, thermal conductivity, viscosity, specific gravity, surface tension and vapor pressure), the information is usually not systematic, when the intent is determining thermodynamic properties or evaluating the applicability of a petroleum based prediction method. Likewise, there are few shale oil based correlations and empirical prediction methods for calculating thermodynamic properties.

It is known that for the same oil shale the specific properties, composition and parameters of the oil depend on the retorting conditions used (the technology used): retorting temperature, duration, heating rate and size of the shale pieces. The current trend in Estonia is towards using the solid heat carrier (the heat carrier is the ash) retorting technology, or the Galoter process, for producing retort oil from oil shale. There is very little data about the physical properties of oils from solid heat carrier retorts. The main existing data is for oil from retort generators (Kiviter process), tunnel ovens and chamber ovens.

Thus, so far the publicly available information has been spotty and poor for evaluating the applicability of contemporary prediction methods – studies/measurements were not historically performed with that goal in mind. For using two-parameter correlations the situation is made more complex by the fact that the Watson characterization factor Kw is not constant over a broad distillation range. Therefore, to obtain the input parameters needed for determining the thermodynamic properties of an oil, both a boiling curve (distillation data) and specific gravity distribution are needed.

Acknowledgements

Support for the study was provided by the National R&D program "Energy" under the project AR10129 "Examination of the Thermodynamic Properties of Relevance to the Future of the Oil Shale Industry". The authors also acknowledge financial support provided by the Estonian Ministry of Education and Research, under target financing SF0140022s10 and under Estonian Scientific Foundation Grant 9297.

REFERENCES

1. Dyni, J. R. Geology and resources of some world oil-shale deposits. *Oil Shale,* 2003, **20**(3),193–252.
2. Owen, N. A., Inderwildi, O. R., King, D. A. The status of conventional world oil reserves – Hype or cause for concern? *Energ. Policy*, 2010, **38**(8), 4743–4749.
3. Oja, V., Suuberg, E. M. Oil shale processing, chemistry and technology. In: *Fossil Energy, Selected Entries from the Encyclopedia of Sustainability Science and Technology* (Malhotra, R., ed.). Springer Science + Businesss Media, New York, 2013, 99–148.
4. Urov, K., Sumberg, A. Characteristics of oil shales and shale-like rocks of known deposits and outcrops. *Oil Shale*, 1999, **16**(3 special: monograph), 1–64.
5. Oja, V. Characterization of tars from Estonian Kukersite oil shale based on their volatility. *J. Anal. Appl. Pyrol.,* 2005, **74**(1–2), 55–60.
6. Oja, V. Vaporization parameters of primary pyrolysis oil from kukersite oil shale. *Oil Shale,* 2015, **32**(2), 124–133.
7. Solomon, P. R., Carangelo, R. M., Horn, E. The effects of pyrolysis conditions on Israeli oil shale properties. *Fuel*, 1986, **65**(5), 650–662.
8. Tsonopoulos, C., Heidman, J. L., Hwang, S.-C. *Thermodynamic and Transport Properties of Coal Liquids*. *An Exxon Monograph*. John Wiley & Sons, 1986.
9. Oja, V. Is it time to improve the status of oil shale science? *Oil Shale*, 2007, **24**(2), 97–99.
10. Zelenin, N. I., Fainberg, V. S., Chernysheva, K. B. Chapter 5: Composition and properties of tars (Sostav i svojstva smoly). In: *Chemistry and technology of oil shale tars* (*Khimiya i tekhnologiya slantsevoj smoly).* Khimiya, Leningrad, 1968, 132–179 (in Russian).
11. Klesment, I., Hallik, E. Comparative charcterization of oil shale semicoking tars (Sravnitel'naya kharakteristika smol polukoksovaniya gorjuchikh slantsev). In: Chemistry and technology of oil shale and products of its processing (*Khimiya i tekhnologiya gorjuchikh slantsev i produktov ikh pererabotki*). (Petuhov, E. F., et al., eds.). Leningrad, 1963 (in Russian).
12. Riazi, M. R. *Characterization and Properties of Petroleum Fractions*. ASTM International, Philadelphia (USA), 2005.
13. 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.
14. Kogerman, P. N., Kõll, A. *Physical properties of Estonian shale oils*. Oil Shale Research Laboratory, Tartu, 1930.
15. Watson, K. M., Nelson, E. F. Improved methods for approximating critical and thermal porperties of petroleum fractions. *Ind. Eng. Chem.*, 1933, **25**(8), 880–887.
16. Kogerman, P. N. *On the chemistry of the Estonian oil shale "Kukersite": a monograph*. K. Matties, Ltd., Tartu, 1931.

17. Luts, K. *Der estländische Brennschiefer-Kukersit, seine Chemie, Tehnologie und Analyse.* Revaler Buchverlag G.M.B.H., Reval, 1944 (in German).
18. Kollerov, D. K. *Physicochemical properties of oil shale and coal liquids* (*Fiziko-khimicheskie svojstva zhidkikh slantsevykh i kamenougol'nykh produktov*). Moscow, 1951 (in Russian).
19. Skrynnikova, G. N. Experimental investigation of thermal conductivity coefficients of oil shale pyrolysis liquids (Eksperimental'noe issledovanie koéffitsienta teploprovodnosti zhidkikh slantsevykh produktov). In: *Chemistry and technology of products of oil shale processing* (*Khimiya i tekhnologiya produktov pererabotki slantsev*). (Kollerov, D. K., Zelenin, N. Ya., Garnovskaya, G. N., eds.). Leningrad, 1954, 242–268 (in Russian).
20. Kollerov, D. K. About Bachinski's relationship of viscosity (O zakone vyazkosti Bachinskogo). In: *Chemistry and technology of products of oil shale processing* (*Khimiya i tekhnologiya produktov pererabotki slantsev*). (Kollerov, D. K., Zelenin, N. J., Garnovskaya, G. N., eds.). Leningrad, 1954, 216–227 (in Russian).
21. Kollerov, D. K. Regarding the use of the characterization fractor K $=\frac{\sqrt[3]{T_k}}{d_4^{20}}$ in characterization of hydrocarbon liquids (K voprosu o fizicheskoj sushchnosti pokazatelya K= $\frac{\sqrt[3]{T_k}}{d_4^{20}}$ uglevodorodnykh zhidkostej). In: *Chemistry and technology of products of oil shale processing* (*Khimiya i tekhnologiya produktov pererabotki slantsev*). (Kollerov, D. K., Zelenin, N. J., Garnovskaya, G. N., eds.). Leningrad, 1954, 228–241 (in Russian).
22. Zelenin, N. I., Fainberg, V. S., Chernysheva, K. B. Chapter 5: Physicochemical properties (Fiziko-khimicheskie svojstva). In: *Chemistry and technology of oil shale tars* (*Khimiya i tekhnologiya slantsevoj smoly*). Khimiya, Leningrad, 1968, 132–179 (in Russian).
23. Mölder, L., Tamvelius, H., Tiikma, L., Tshuryumova, T. Viscosity of shale oil binary blends. *Oil Shale*, 1999, **16**(1), 42–50.
24. Mölder, L., Tamvelius, H., Tiikma, L. Viscosity of shale oil originated distillate oil – residual petroleum oil binary blends. *Oil Shale*, 1999, **16**(2), 133–140.
25. Mölder, L., Tamvelius, H., Tiikma, L. Viscosity and stability of distillate petroleum oil – residual petroleum oil and distillate petroleum oil - shale oil binary blends. *Oil Shale*, 1999, **16**(3), 239–248.
26. Lovell, P. F. Production of Utah shale oils by the Paraho DH and Union "B" retorting processes. In: *Eleventh Oil Shale Symposium Proceedings* (Gary, J. H., ed.). Colorado School of Mines Press, Golden, Colorado, 1978, 184–192.

Received July 8, 2015

ARTICLE 3

Distribution of Hydroxyl Groups in Kukersite Shale Oil: Quantitative Determination Using Fourier Transform Infrared (FT-IR) Spectroscopy

Published in Applied Spectroscopy, vol. 69, no. 5, pp. 555–562, 2015.

Distribution of Hydroxyl Groups in Kukersite Shale Oil: Quantitative Determination Using Fourier Transform Infrared (FT-IR) Spectroscopy

Zachariah Steven Baird, Vahur Oja,* Oliver Järvik

Department of Chemical Engineering, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia

This article describes the use of Fourier transform infrared (FT-IR) spectroscopy to quantitatively measure the hydroxyl concentrations among narrow boiling shale oil cuts. Shale oil samples were from an industrial solid heat carrier retort. Reference values were measured by titration and were used to create a partial least squares regression model from FT-IR data. The model had a root mean squared error (RMSE) of 0.44 wt% OH. This method was then used to study the distribution of hydroxyl groups among more than 100 shale oil cuts, which showed that hydroxyl content increased with the average boiling point of the cut up to about 350 °C and then leveled off and decreased.

Index Headings: **Phenols; Shale oil; Fourier transform infrared spectroscopy; FT-IR spectroscopy; Partial least squares regression; PLS regression; Hydroxyl group.**

INTRODUCTION

Oil shale is a solid fossil fuel that is found in great abundance around the world. It is has a high mineral content, and most of the organic matter in shale is dispersed in the rock as a macromolecular structure called kerogen. Oil shale resources are vast, but largely unused due to technical and environmental challenges that make extracting and refining it uneconomical.

To overcome these challenges and use oil shale resources well, it is important to understand the physical and chemical properties of the shale and the resulting shale oil. Conventional crude oils have been thoroughly researched, and detailed models have been created to predict the thermodynamic and transport properties of the oil.[1] However, even though oil shale deposits exist around the globe,[2] significantly less thermodynamic data is available for shale oils.

Shale oil is a synthetic crude produced by heating oil shale, which causes the solid organic matter to undergo pyrolytic decomposition. Oil produced from oil shale should not be confused with tight oil (also sometimes called shale oil), which is conventional oil that is found in shale and sandstone formations. Oil shale oil has been produced in Estonia for about a century. Many different retorting methods have been used, including generators, tunnel ovens, rotating retorts, chamber ovens, and solid heat carrier retorts.[3] The properties of the oil depend on the type of retort used.[4] This study used Estonian kukersite shale oil that was produced using the commercial solid heat carrier retorting process (also known as the Galoter process).[5] Currently, the trend in Estonia is toward this type of process, in large part because it is more environmentally friendly. The newest plants use, or will use, this technology.

Kukersite shale oil contains significant quantities of phenols, which affect the properties of the oil, and therefore, how the oil is refined and used. Most correlations for the thermodynamic and transport properties of narrow boiling point petroleum cuts up to a molecular weight of 300 g/mol (equivalent to an atmospheric boiling point of about 350 °C) are based on two parameters: one that describes the molecular energy of the mixture (e.g., specific gravity, refractive index, etc.) and one for the molecular size (e.g., boiling point, molecular weight, etc.).[1] However, because of the high content of polar phenols in kukersite shale oil, it may be necessary to include a third parameter that describes association forces to develop suitable correlations for this shale oil. Therefore, it is important to be able to conveniently measure phenolic content and understand the distribution of phenols in shale oil. The amount of phenols in a shale oil sample has traditionally been measured using a titration technique that requires several hours to complete. For this study, it was necessary to find another method that would save time and allow more measurements to be made.

Mid-infrared spectroscopy was used because it has been used extensively to measure the concentrations of a wide variety of compounds, from antioxidants in onions[6] to contaminants in soil.[7] Many early quantitative analyses performed on infrared (IR) spectra were based on a simple, linear relationship between the IR absorbance and concentration (known as Beer's law). The advent of computers allowed complex statistical methods to be applied to IR spectral analysis. Two of the most used methods are principal component analysis and partial least squares (PLS) regression. Principal component analysis reduces the input (spectral data) and output (properties to be measured) down to a handful of underlying factors that describe most of the variation between samples. PLS regression, like principal component analysis, finds a few factors describing the variation, but also creates a good predictive model by combining the input and output factors in a way that allows the model to predict the property of interest for new samples. The resulting PLS model takes the form of a linear equation where each input variable is multiplied by a coefficient and then summed together, along with a constant, to obtain the predicted value of the desired property.[8]

Received 28 August 2014; accepted 18 November 2014.

* Author to whom correspondence should be sent. E-mail: vahur.oja@ttu.ee.

DOI: 10.1366/14-07705

0003-7028/15/6905-0555/0

To create a PLS model to measure a property from IR spectra, spectra are first taken of reference compounds for which the desired property is already known. Usually only a portion of the IR spectrum is used as the basis for the PLS model because including portions of the spectrum where the compounds of interest do not absorb or portions that do not correlate well with the desired property leads to a less accurate model. The accuracy of the PLS model also depends on the number of factors used to create it. As more factors are added, the model fits the data more closely, but if too many factors are used, overfitting can occur. This means that factors are included which do not really correlate with the property to be measured. This decreases the predictive ability of the model.[9,10] For this reason, the accuracy of a model is usually validated by testing it on a different set of samples. The number of factors can also be determined by using the original set of data and simply calculating the model several times while rotating which samples are left out of the calculation. This is called cross validation.[8,11]

The use of PLS regression, and other techniques, has enabled IR spectroscopy to become a powerful quantitative tool. Infrared spectroscopy coupled with PLS regression has also been used for measuring crude oil compositions and recently has even been used to predict other physical properties such as density and viscosity.[12] This and similar methods could be used to quickly provide a wealth of information about shale oil and other liquid fuels derived from fossil fuels. This study sought to use IR spectroscopy to determine the hydroxyl content of kukersite shale oils, which would reduce measurement time to a few minutes. The data collected using this method was then used to study the distribution of phenols among narrow boiling shale oil cuts.

MATERIALS AND METHODS

Samples. The shale oil for this experiment was obtained from a commercial solid heat carrier (Galoter) process in Narva, Estonia. Most of the samples were from the fuel oil fraction of the distillation (also referred to as "middle oil fraction" in some older Galoter process-based publications), but a sample of heavy oil from the same plant was also studied. In total, six different oil samples from Narva were used (five fuel oil samples and one heavy oil sample), which were collected over a yearlong period. This shale oil was further separated into narrower boiling range cuts by distillation. Three of the distillations were performed under vacuum conditions, one was done in a rectification column, and the other eight were Engler distillations at atmospheric pressure according to the ASTM D86 standard.[13] The distillations cuts were volume based, so each narrow cut was taken to have roughly the same volume as the other cuts. Most of the cuts spanned distillation temperature intervals of about 5 to 10 °C. The cuts had average boiling points ranging from about 200 to 450 °C at atmospheric pressure. Additionally, in order to study the phenols more closely, phenols from ten of the vacuum distillation cuts were extracted. These ten phenol cuts, along with the corresponding dephenolated samples, were also measured as part of this study.

Devices. The IR spectra of the oil samples were obtained using an Interspec 301-X spectrometer fitted with an ATR accessory (Interspectrum OÜ, Tõravere, Estonia). The ATR accessory had a single reflection, ZnSe internal reflection element. Interspec for Windows software (version 3.40 Pro, Interspectrum OÜ, Tõravere, Estonia) was used to collect the spectra. Spectra were obtained between 600 and 4000 cm^{-1}, and a resolution of 4 cm^{-1} was used. For each cut, five to ten scans were taken and then averaged together (except for the heavy oil cuts, which were scanned only once). This improved the repeatability of the spectra and reduced noise. The baseline of each spectrum was then corrected using a cubic spline interpolation based on 4 points: 3999, 3796, 2200, and 1800 cm^{-1}. For a few spectra the correction was poor due to noise, but good baselines were obtained for these spectra by slightly shifting the points that were taken in order to avoid the noise. The baseline correction was performed using Essential FT-IR software (version 3.10.016, Operant LLC, Burke, VA).

Other characteristics of the oil cuts were also measured. Densities were measured using an Anton Paar DMA 5000M. Average boiling points were measured using a DuPont 951 thermogravimetric analyzer. This also allowed atmospheric boiling points to be found for the vacuum distillation cuts. Average molecular mass was measured using a cryoscopic method in which samples were dissolved in benzene.[14]

Methods. To measure the hydroxyl content for the calibration points, a titration method was used. The method has been used for decades to measure the hydroxyl content of shale oils.[15] It should be noted that this method also measures primary and secondary alcohols and that organic acids also affect the titration. However, in shale oil the majority of hydroxyl groups are phenols and there are not significant quantities of organic acids,[16] which is why this method has been used since the beginning of the last century to experimentally determine phenol content. The titration method used in this study was published in Russian, but the procedure is the same as the acetylation method described in Vogel's *Elementary Practical Organic Chemistry*[17] with the following modifications: about 0.5 g of the sample was used, only 2 ml of the acetylating reagent was used, the mixture was heated for 1.5 hours, and a pH meter was used for the titration instead of an indicator. To summarize the procedure, the hydroxyl groups in the sample were acetylated using an excess of acetic anhydride in pyridine. A parallel blank experiment was prepared at the same time. At the end of the reaction, water was added to the samples to convert any remaining acetic anhydride to acetic acid. Then, the acetic acid was titrated using potassium hydroxide, and the difference in titration volume between the sample and the blank was used to calculate the weight percent of hydroxyl groups in the sample.

As mentioned in the previous section, phenols were extracted from ten cuts. They were extracted using a 10% solution of sodium hydroxide. Sodium hydroxide was added five times, and then the samples were washed with distilled water three times. Any remaining water or solvent was then removed from the oil portion of the extraction with a rotary evaporator (at a temperature

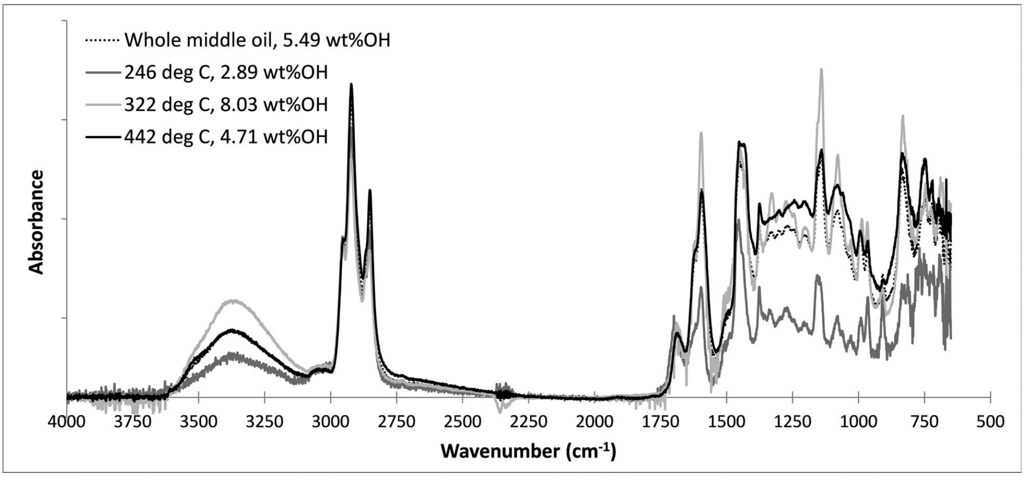

FIG. 1. Representative IR spectra from shale oil cuts. Temperatures in the legend are average boiling points.

of about 80 °C). These dephenolated oil cuts had a hydroxyl content of about 2 wt% OH. The sodium hydroxide portion was acidified, and then the phenols were extracted using diethyl ether. The solvent was then removed from the phenols using the rotary evaporator.

RESULTS AND DISCUSSION

Constructing the Partial Least Squares Model. The IR spectra of the shale oil contained several peaks that could be assigned to general functional groups and molecular structures. Figure 1 gives spectra from a few of the shale oil cuts. The large group of peaks between 2800 and 3000 cm^{-1} result from aliphatic C–H stretching. The large, rounded peak at about 3400 cm^{-1} is the O–H stretch peak, and shows that hydroxyl groups are present. The lower end of the spectrum contains a jumble of peaks caused by the numerous compounds in the oil cuts, but a few notable features can be picked out. The peak at about 1600 cm^{-1} is from aromatic ring stretching and thus shows the aromaticity of the sample. Peaks for O–H bending occur between 1260 and 1410 cm^{-1}, and peaks for C–O stretching occur between about 1050 and 1200 cm^{-1}, all of which give a measure of hydroxyl content.[18] Lots of information about the molecular structure of the oil samples is contained in this lower range of the spectrum.

Partial least squares (PLS) regression was used to develop a model to predict hydroxyl content. The PLS analysis was performed using a self-made Python program that used the Scikit-learn PLS function.[19] The Scikit-learn function uses the non-linear iterative partial least squares (NIPALS) PLS algorithm.

The region from 800 to 1640 cm^{-1} was used because, after trying several other regions of the spectrum as well as the entire spectrum, it was found that this region gave the best results. When data points from the 3100 and 3700 cm^{-1} range were included, the accuracy of the model decreased, and thus, this region was not used. The searching combination moving window technique proposed by Du et al.[20] was also used to search for the best spectral ranges, but the region from 800 to 1640 cm^{-1} still gave better results.

For the analysis, a resolution of 4 cm^{-1} was used. Here, 48 cuts with known hydroxyl content, as found by titration, were used as the calibrating substances. Further, 11 factors were used for the regression because this number gave the model with the lowest error. The accuracy of the resulting model was estimated using leave-one-out cross-validation, and the average root mean square error (RMSE) of the model was 0.44 wt% OH. This error is the result of noise or other distortions in the spectra, error in the titration method used, and error in the PLS model. The titration method itself had an approximate error of 0.29 wt% OH (the average standard deviation for parallel measurements was 0.145 wt% OH). In Fig. 2, the actual hydroxyl content is plotted versus the predicted values. The predicted values are the values obtained from cross validation.

It should also be noted that the PLS model was designed to give hydroxyl concentration in units of molar volume (mol/cm^3). These are the units that most directly correspond to the way the FT-IR works since spectral features are affected by the number of hydroxyl groups in the measured sample space. Therefore, the reference values were converted to moles per cubic centimeter to create the PLS model, and then, the model results were converted to weight percent OH.

The coefficients resulting from the PLS model allowed hydroxyl content to be found for all the fuel oil and heavy oil cuts distilled and for the original whole fuel oil samples. Thus, all that is needed to find the hydroxyl content of shale oil is its IR spectrum, if a PLS model appropriate for a specific shale oil is available.

Distribution of Hydroxyl Groups Among Narrow Boiling Point Cuts. The data collected using this method are shown in Figs. 3–5. Figure 3 shows OH content in a narrow cut versus the average boiling point of the narrow cut. The narrow cuts were taken over distillation temperature ranges of 5 to 10 °C. The samples were

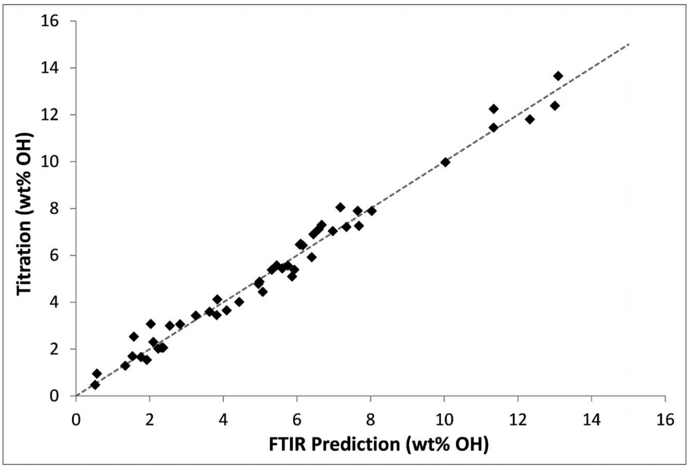

Fig. 2. Performance of the partial least squares model. The predicted values were obtained from cross validation.

collected over a year-long period from an industrial retort run under the same operating conditions. Figures 3–5 show that hydroxyl content in the shale oil cuts ranged from about 0 to 9 wt%. The five whole fuel oil cuts measured had an average of 5.2 wt% OH. The general trend is that hydroxyl content increases with boiling point up to about 350 °C, and then for the heavier cuts, the hydroxyl content decreases and levels off. Similar behavior is also seen as the molar mass of the cuts increases, as shown in Fig. 5.

Only a few sets of data for the phenolic content of kukersite shale oil are publicly available, and these were mostly based on wide boiling cuts that were taken over distillation ranges of about 100 °C. These datasets are not as detailed as the data obtained from narrow boiling cuts in this study (which were taken at 5 to 10 °C intervals). Also, the data shown was measured using different methods and was based on oil from several different retorts, and thus, the data varies widely. However, this literature data does show some similar trends.

Figure 6 shows these literature values and separates them according to the type of retort used to produce the oil. Data for the solid heat carrier retort was taken from the literature[21] as well as from measurements made at Tallinn University of Technology in 1999. Data for the gas-generator retorts was taken from these references.[22–24] Most of the data was measured for fractions boiling below

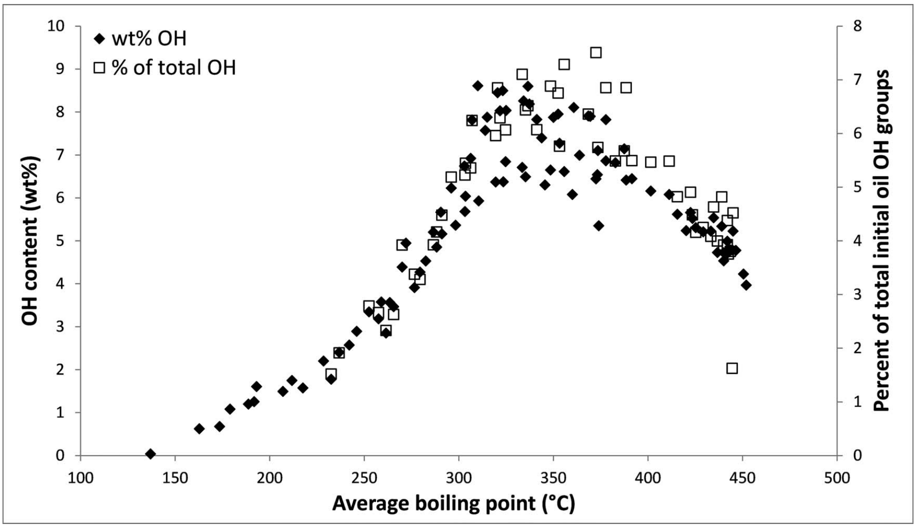

Fig. 3. OH content versus average boiling point for narrow boiling point shale oil cuts. The data is shown in units of weight percent OH on the left axis and as the percent of the total OH groups in the initial oil on the right axis.

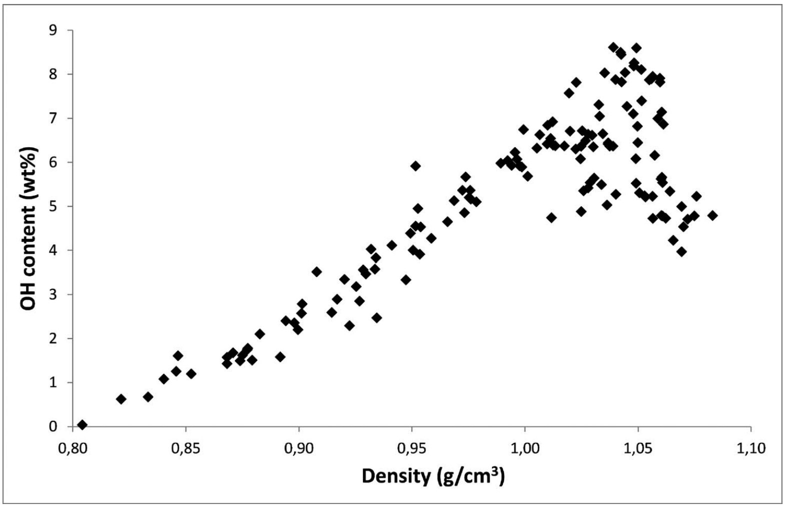

FIG. 4. OH content versus the density of shale oil cuts.

370 °C, and therefore, they only show an increase in phenol content. Two oil samples were also fractionated using vacuum distillation. The atmospheric distillation temperatures of these fractions were roughly estimated using the correlation presented on page 106 in Riazi,[1] and the resulting data is shown as the black points in Fig. 6. Data from these vacuum distillations also include cuts with higher boiling points and show a peak in phenol content and then a decrease, which corresponds to the behavior observed in this study.

In the vacuum distillations, about 90% of the initial fuel oil was distilled, and about 90% of the hydroxyl groups in the initial fuel oil were recovered. This means 10% of the hydroxyl groups remained in the residue, which indicates that even the heaviest part of kukersite shale oil still contains significant quantities of hydroxyl groups.

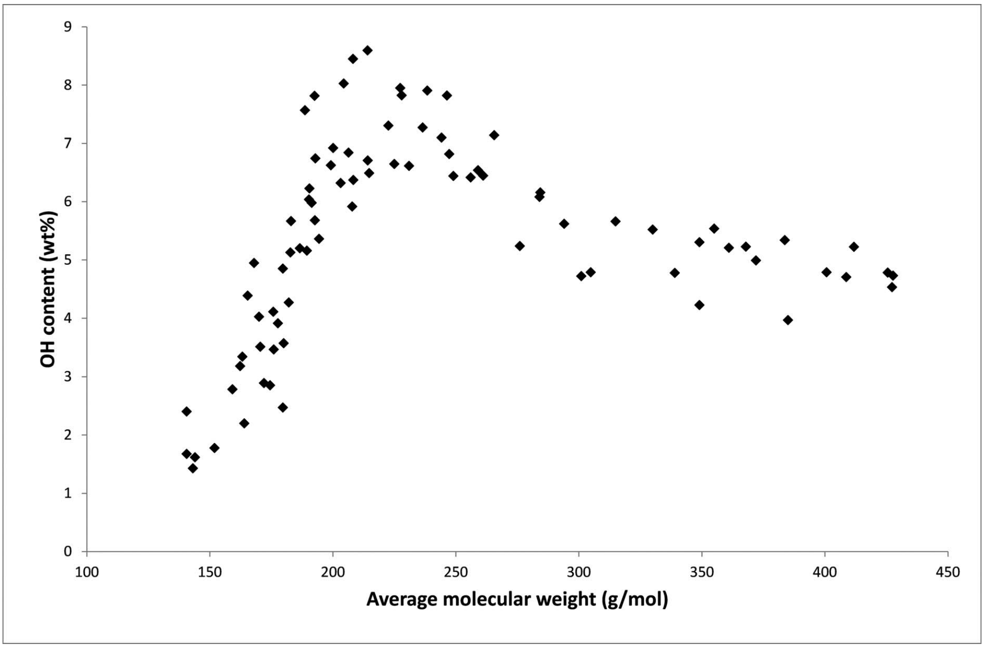

FIG. 5. OH content versus the average molecular weight of shale oil cuts.

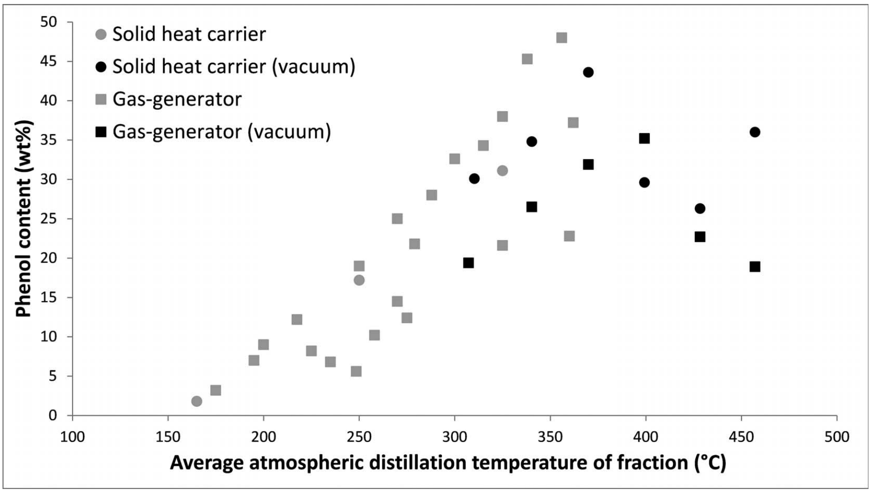

Fig. 6. Previously measured data for phenol content of kukersite shale oil from two types of refining processes.

This is supported by observations from an earlier study on kukersite tar.[25] In the experiment, kukersite primary pyrolysis tar, which is an intermediate in the pyrolysis of oil shale, was fractionated by sublimation up to a temperature of 220 °C in a vacuum of 10^{-5} torr. Infrared spectra of the fractions and the residue all showed a significant OH peak. In Engler distillations, only about 40% of the hydroxyl groups were recovered, which also explains why measurements made only on lower boiling fractions do not completely describe the distribution of hydroxyl groups in shale oil. The heavy oil distillation fractions (which is the heaviest fraction from the commercial plant) also had significant quantities of hydroxyl groups.

Structure of the Shale Oil Phenols. As mentioned before, the majority of hydroxyl groups in kukersite shale oil are phenolic, including resorcinol and naphthol derivatives. Some general information about the structure of phenols in shale oil can be taken from the data in this study and from past studies. Knowledge of the structure also helps to explain the peak and subsequent decrease in phenolic content that occurs as the boiling point of shale oil cuts increases. Figure 7 shows that the number of OH groups per molecule increases with boiling point up to about 350 °C, which is approximately the boiling point of the cuts with the highest OH content. Beyond that point, the ratio levels off, even though the boiling point, and therefore molar mass, of the cuts is increasing. This means that the phenols in heavier cuts

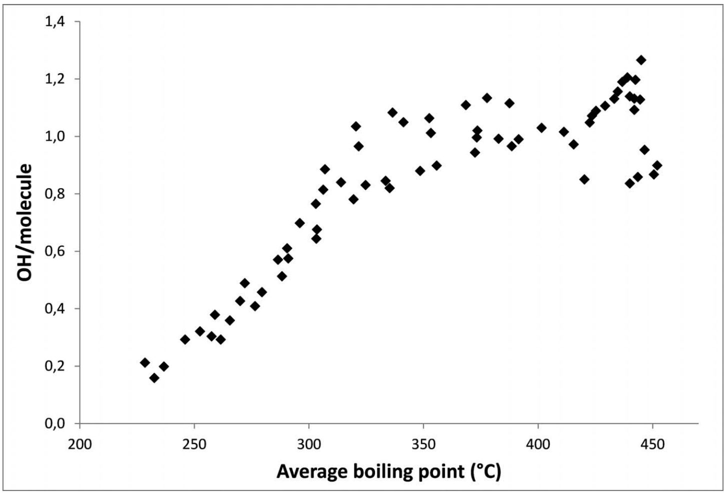

Fig. 7. Number of OH groups per molecule for shale oil cuts as a function of average boiling point.

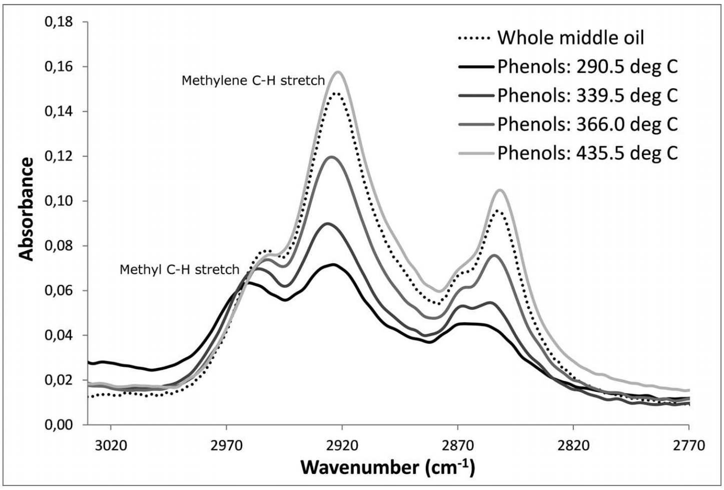

FIG. 8. Aliphatic peaks from IR spectra of shale oil phenols. The middle peak (about 2925 cm^{-1}) is from methylene C–H stretching, and the peak on the left (about 2955 cm^{-1}) is from methyl C–H stretching. The phenol spectra here are labeled according to their average boiling point.

are larger, leading to a decrease in the mass percent of hydroxyl groups, even though the mole percent of phenols in general remains roughly constant. Similarly, Zelenin and Vassiliev[24] also noticed that the bulk of shale oil phenols are larger phenols that appear in high-boiling fractions. Additionally, Luts[22] made the observation that the average molecular weight of phenols in the crude shale oil is about 300, of which only one third is the basic phenol core (which has a molecular weight of 94). Kukersite shale oil also contains naphthol and resorcinol derivatives. The basic naphthol and resorcinol cores have molecular weights of 144 and 110, respectively, which are still much smaller than the average shale oil phenolic compound.

Infrared spectra from the extracted phenols also support this explanation. The peak at 2925 cm^{-1} is generally assigned as antisymmetric methylene C–H stretching, and the peak at 2955 cm^{-1} as antisymmetric methyl C–H stretching.[18] The spectra for all the shale oil cuts in this study show that the methylene peak is significantly higher than the methyl peak, which indicates that shale oil contains more straight, long, aliphatic chains than branched chains and methyl groups. However, for the phenols extracted from the lighter cuts, the methylene peak is the same height as the methyl peak. For phenols from successively heavier cuts the methylene peak becomes larger, as shown in Fig. 8. These observations indicate that the phenols distilled first are smaller and lack the long, aliphatic chains, and as the temperature increases, distilled phenols contain successively longer aliphatic chains. In this study, the phenols with long chains started to distil at about 340 °C, and by 370 °C, the phenols had a methylene–methyl C–H ratio similar to that of shale oil in general. This is the same boiling point range where a maximum in hydroxyl content occurs.

Zelenin and Vassiliev,[24] in an earlier study on shale oil phenols, separated phenolic molecules from four fractions of kukersite shale oil taken between 180 and 350 °C. Some of their results are shown in Fig. 9, which show that naphthol and resorcinol derivatives make up a large portion of the phenolic compounds in shale oil that boils below 350 °C. They also found that, for the phenols, 20 to 40% of the carbon atoms were aliphatic, which means that there are 2 to 4.4 carbon side chains on the phenols on average.[24] However, this study does not give any information on the phenols present in higher boiling fractions. Also, Zelenin and Vassiliev[24] extracted phenols using a 10% alkali solution. This type of solution was also used for extracting phenols in this study and, after extraction the oil phase, still contained about 2 wt% OH. Thus, many phenols may not have been extracted or examined by Zelenin and Vassiliev, but their research still gives insight into the types of phenolic compounds contained in kukersite shale oil.

CONCLUSIONS

This study shows that IR spectroscopy can be used to measure the hydroxyl content of shale oils with a RMSE of 0.44 wt% OH. The model used to do this was found using partial least squares regression. Using IR spectra greatly reduces the time needed to obtain information about the hydroxyl content of oils. Infrared spectroscopy could also likely be extended to measure other compositional properties for shale oil, as well as conventional oil.

The data from this study also provides new information about the distribution of hydroxyl groups in kukersite shale oil. The distributions according to average boiling point, molecular weight, and density were shown. Normal Engler distillations generally only show the increasing hydroxyl trend because Engler distillations are only done up to boiling points of around 300 °C. At higher temperatures, the oil samples start to degrade. By performing distillations under vacuum conditions, samples with higher boiling points were also collected, which enabled collection of about 90% of the hydroxyl

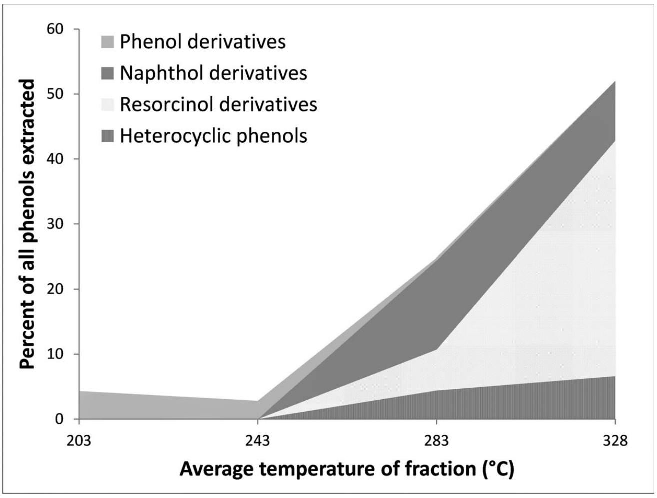

FIG. 9. Types of phenols recovered from kukersite shale oil fractions taken between 180–350 °C.

groups in the fuel oil fraction. These samples showed the peak and subsequent decrease in hydroxyl content described in this study.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding provided 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”. The authors also thank Ilme Rohtla and Einart Lindaru for their help in carrying out the experiments.

1. M.R. Riazi. Characterization and Properties of Petroleum Fractions. West Conshohocken, PA: ASTM International, 2005.
2. K. Urov, A. Sumberg. “Characteristics of Oil Shales and Shale-like Rocks of Known Deposits and Outcrops”. Oil Shale. 1999. 16(3): 1-64.
3. V. Oja. “Is it Time to Improve Oil Shale Science?”. Oil Shale. 2007. 24(2): 97-99.
4. S.H. Guo. “The Chemistry of Shale Oil and its Refining”. In: G. Jinsheng, editor. Coal, Oil Shale Natural Bitumen, Heavy Oil and Peat–Vol. II. EOLSS Publishers Company Limited, 2009. Pp. 94-106.
5. N. Golubev. “Solid Oil Shale Heat Carrier Technology for Oil Shale Retorting”. Oil Shale. 2003. 20(3 Special)): 324-332.
6. X. Lu, J. Wang, H. Al-Qadiri, C. Ross, J. Powers, J. Tang, B. Rasco. “Determination of Total Phenolic Content and Antioxidant Capacity of Onion (Allium cepa) and Shallot (Allium oschaninii) Using Infrared Spectroscopy”. Food Chem. 2011. 129(2): 637-644. doi:10.1016/j.foodchem.2011.04.105.
7. R. Okparanma, A. Mouazen. “Visible, Near-Infrared Spectroscopy Analysis of a Polycyclic Aromatic Hydrocarbon in Soils”. Sci. World. J. 2013. doi:10.1155/2013/160360.
8. R. Tobias. “An Introduction to Partial Least Squares Regression”. SUGI Proceedings. 1995.
9. P. Geladi, B. Kowalski. “Partial Least Squares Regression: A Tutorial”. Anal. Chim. Acta. 1986. 185: 1-17.
10. S. Wold, C. Albano, W.J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, et al. “Multivariate Data Analysis in Chemistry”. In: B.R. Kowalski. editor. Chemometrics: Mathematics and Statistics in Chemistry. Dordrecht, Holland: Springer, 1984. Pp. 17-96. 1st ed.
11. B. Mevik, H.R. Cederkvist. “Mean Squared Error of Prediction (MSEP) Estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR)”. J. Chemom. 2004. 18(9): 422-429. doi:10.1002/cem.887.
12. P. de Peinder, T. Visser, D. Petrauskas, F. Salvatori, F. Soulimani, B. Weckhuysen. “Partial Least Squares Modeling of Combined Infrared, 1H NMR and 13C NMR Spectra to Predict Long Residue Properties of Crude Oils”. Vib. Spectrosc. 2009. 51: 205-212. doi:10.1016/j.vibspec.2009.04.009.
13. American Society for Testing and Material. ASTM Standard D86. “Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure”. West Conshohocken, PA: ASTM International, 2012. doi:10.1520/D0086-12.
14. American Society for Testing and Materials. ASTM Standard D2224. “Method of Test for Mean Molecular Weight of Mineral Insulating Oils by the Cryoscopic Method”. West Conshohocken, PA: ASTM International, 1983.
15. A. Aarna, V. Paluoja. “Determination of Hydroxyl Groups in Shale Oil by the Acetylation Method”. In: Analytical Methods for Oil Shale and Oil Shale Products. Tallinn, Estonia: 1961. Pp. 23-26 (in Russian).
16. H. Luik. “Chemicals and Other Products From Shale Oil”. In: G. Jinsheng, editor. Coal, Oil Shale, Natural Bitumen, Heavy Oil and Peat–Vol. II. EOLSS Publishers Company Limited, 2009. Pp. 107-128.
17. A. Vogel. “Elementary Practical Organic Chemistry: Quantitative Organic Analysis Part 3”. London: Longman, 1958.
18. J. Coates. “Interpretation of Infrared Spectra, A Practical Approach”. In: R.A. Meyers, editor. Encyclopedia of Analytical Chemistry. Chichester, UK: John Wiley and Sons Ltd, 2000. Pp. 10815-10837.
19. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. “Scikit-learn: Machine Learning in Python”. J. Mach. Learn. Res. 2011. 12: 2825-2830. 2011.
20. Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki. “Spectral Regions Selection to Improve Prediction Ability of PLS Models by Changeable Size Moving Window Partial Least Squares and Searching Combination Moving Window Partial Least Squares”. Anal. Chim. Acta. 2004. 501(2): 183-191. doi:10.1016/j.aca.2003.09.041.
21. N.I. Zelenin, V.L. Fainberg, K.B. Kernosheva. Oil Shale Chemistry and Technology, Chemistry. Leningrad, Russia: 1968. (in Russian).
22. K. Luts. “The Estonian Oil Shale Kukersite, its Chemistry, Technology and Analysis”. Tartu, Estonia: 1934 (in German).
23. V.M. Yefimov, T.M. Volkov, E.F. Petukhov, I.K. Rooks. “Thermal Processing of Lump Oil Shale: the Kiviter Process”. In: V.D. Allred, editor. Oil Shale Processing Technology. East Brunswick, NJ: The Center for Professional Advancement, 1982. Pp. 67-81.
24. N.I. Zelenin, M.L. Vassiliev. “Oil Shale Phenols and Ways of Their Utilization”. United Nations Symposium on the Development and Utilization of Oil Shale Resources Section 111. Tallinn, Estonia: August 26–September 4, 1968.
25. V. Oja. “Characterization of Tars From Estonian Kukersite Oil Shale Based on Their Volatility”. J. Anal. Appl. Pyrolysis. 2005. 74(1-2): 55-60. doi:10.1016/j.jaap.2004.11.032.

CURRICULUM VITAE

Work Experience

Junior Researcher

09.2015 - present *Tallinn University of Technology* Tallinn, Estonia

Engineer

09.2013 - 09.2015 *Tallinn University of Technology* Tallinn, Estonia

Research Assistant

04.2012 - 08.2012 *BYU Catalysis Lab* Provo, UT, USA

Intern

05.2012 - 08.2012 *RPM Revenue Drivers* Orem, UT, USA

Instructor

06.2011 - 04.2013 *Missionary Training Center* Provo, UT, USA

Factory Worker

05.2011 - 06.2011 *Bridgeport National Bindery* Agawam, MA, USA

Data Entry Worker

09.2008 - 04.2009 *Harold B. Lee Library* Provo, UT, USA

Analyst

05.2008 - 08.2008 *MassMutual Financial Group*Springfield, MA, USA

Education

Ph.D. Chemical and Materials Technology

09.2015 – 11.2017 *Tallinn University of Technology* Tallinn, Estonia

M.S. Chemical and Environmental Technology

09.2014 - 06.2015 *Tallinn University of Technology* Tallinn, Estonia

B.S. Chemical Engineering

08.2007 - 04.2013 *Brigham Young University* Provo, UT, USA

Languages

English - native language
Estonian - fluent
Finnish - basic conversational ability

ELULOOKIRJELDUS

Töökogemus

Nooremteadur
09.2015 - present *Tallinna Tehnikaülikool* Tallinn, Eesti

Insener
09.2013 - 09.2015 *Tallinna Tehnikaülikool* Tallinn, Eesti

Uurimisassistent
04.2012 - 08.2012 *BYU Catalysis Lab* Provo, UT, USA

Praktikant
05.2012 - 08.2012 *RPM Revenue Drivers* Orem, UT, USA

Õpetaja
06.2011 - 04.2013 *Missionary Training Center* Provo, UT, USA

Tehase töötaja
05.2011 - 06.2011 *Bridgeport National Bindery* Agawam, MA, USA

Andmete sisestaja
09.2008 - 04.2009 *Harold B. Lee Library* Provo, UT, USA

Analüütik
05.2008 - 08.2008 *MassMutual Financial Group*Springfield, MA, USA

Haridus

Ph.D. Keemia- ja materjalitehnoloogia
09.2015 – 11.2017 *Tallinna Tehnikaülikool* Tallinn, Eesti

M.S. Keemia- ja keskkonnatehnoloogia
09.2014 - 06.2015 *Tallinna Tehnikaülikool* Tallinn, Eesti

B.S. Keemiatehnika
08.2007 - 04.2013 *Brigham Young University* Provo, UT, USA

Keele oskused

Inglise keel – emakeel
Eesti keel – kodune keel
Soome keel – algtase

DISSERTATIONS DEFENDED AT TALLINN UNIVERSITY OF TECHNOLOGY ON *CHEMISTRY AND CHEMICAL ENGINEERING*

1. **Endel Piiroja**. Oxidation and Destruction of Polyethylene. 1993.

2. **Meili Rei**. Lihatehnoloogia teaduslikud alused. Fundamentals of Food Technology. 1995.

3. **Meeme Põldme**. Phase Transformations in Hydrothermal Sintering Processing of Phosphate Rock. 1995.

4. **Kaia Tõnsuaadu**. Thermophosphates from Kovdor and Siilinjärvi Apatites. 1995.

5. **Anu Hamburg**. The Influence of Food Processing and Storage on the N-Nitrosamines Formation and Content in Some Estonian Foodstuffs. 1995.

6. **Ruth Kuldvee**. Computerized Sampling in Ion Chromatography and in Capillary Electrophoresis. 1999.

7. **Külliki Varvas**. Enzymatic Oxidation of Arachidonic Acid in the Coral *Gersemia fruticosa*. 1999.

8. **Marina Kudrjašova**. Application of Factor Analysis to Thermochromatography and Promotion Studies. 2000.

9. **Viia Lepane**. Characterization of Aquatic Humic Substances by Size Exclusion Chromatography and Capillary Electrophoresis. 2001.

10. **Andres Trikkel**. Estonian Calcareous Rocks and Oil Shale Ash as Sorbents for SO_2. 2001.

11. **Marina Kritševskaja**. Photocatalytic Oxidation of Organic Pollutants in Aqueous and Gaseous Phases. 2003.

12. **Inna Kamenev**. Aerobic Bio-Oxidation with Ozonation in Recalcitrant Wastewater Treatment. 2003.

13. **Janek Reinik**. Methods for Purification of Xylidine-Polluted Water. 2003.

14. **Andres Krumme**. Crystallisation Behaviour of High Density Polyethylene Blends with Bimodal Molar Mass Distribution. 2003.

15. **Anna Goi**. Advanced Oxidation Processes for Water Purification and Soil Remediation. 2005.

16. **Pille Meier**. Influence of Aqueous Solutions of Organic Substances on Structure and Properties of Pinewood (*Pinus sylvestris*). 2007.

17. **Kristjan Kruusement**. Water Conversion of Oil Shales and Biomass. 2007.

18. **Niina Kulik**. The Application of Fenton-Based Processes for Wastewater and Soil Treatment. 2008.

19. **Raul Järviste**. The Study of the Changes of Diesel Fuel Properties a its Long Term Storage. 2008.

20. **Mai Uibu**. Abatement of CO_2 Emissions in Estonian Oil Shale-Based Power Production. 2008.

21. **Valeri Gorkunov**. Calcium-Aluminothermal Production of Niobium and Utilization of Wastes. 2008.

22. **Elina Portjanskaja**. Photocatalytic Oxidation of Natural Polymers in Aqueous Solutions. 2009.

23. **Karin Reinhold**. Workplace Assessment: Determination of Hazards Profile using a Flexible Risk Assessment Method. 2009.

24. **Natalja Savest**. Solvent Swelling of Estonian Oil Shales: Low Temperature Thermochemical Conversion Caused Changes in Swelling. 2010.

25. **Triin Märtson**. Methodology and Equipment for Optical Studies of Fast Crystallizing Polymers. 2010.

26. **Deniss Klauson**. Aqueous Photocatalytic Oxidation of Non-Biodegradable Pollutants. 2010.

27. **Oliver Järvik**. Intensification of Activated Sludge Process – the Impact of Ozone and Activated Carbon. 2011.

28. **Triinu Poltimäe**. Thermal Analysis of Crystallization Behaviour of Polyethylene Copolymers and Their Blends. 2011.

29. **Mariliis Sihtmäe**. (Eco)toxicological Information on REACH-Relevant Chemicals: Contribution of Alternative Methods to *in vivo* Approaches. 2011.

30. **Olga Velts**. Oil Shale Ash as a Source of Calcium for Calcium Carbonate: Process Feasibility, Mechanism and Modeling. 2011.

31. **Svetlana Jõks**. Gas-Phase Photocatalytic Oxidation of Organic Air Pollutants. 2012.

32. **Aleksandr Dulov**. Advanced Oxidation Processes for the Treatment of Water and Wastewater Contaminated with Refractory Organic Compounds. 2012.

33. **Aleksei Zaidentsal**. Investigation of Estonian Oil Shale Thermo-bituminization in Open and Closed System. 2012.

34. **Dmitri Šumigin**. Composites of Low-Density Polyethylene and Poly(Lactic Acid) With Cellulose and Its Derivatives. 2014.

35. **Aleksandr Käkinen**. The Role of Physico-chemical Properties and Test Environment on Biological Effects of Copper and Silver Nanoparticles. 2014.

36. **Ada Traumann**. Improvement of Work Environment through Modelling the Prevention of Health Risks Focusing on Indoor Pollutants. 2014.

37. **Marika Viisimaa**. Peroxygen Compounds and New Integrated Processes for Chlorinated Hydrocarbons Degradation in Contaminated Soil. 2014.

38. **Olga Budarnaja**. Visible-light-sensitive Photocatalysts for Oxidation of Organic Pollutants and Hydrogen Generation. 2014.

39. **Jelena Hruljova**. Role of Specifically Interacting Solvents in Solvent Swelling of Kukersite Oil Shale Kerogen. 2014.

40. **Irina Klimova**. Modification of Ammonium Nitrate Fertilizer. 2014.

41. **Julia Krasulina**. Upgrading of Liquid Products from Estonian Kukersite Oil Shale by Catalytic Hydrogenation. 2015.

42. **Irina Epold**. Degradation of Pharmaceuticals by Advanced Oxidation Technologies in Aqueous Matrices. 2015.

43. **Kadriann Tamm**. Leaching of the Water-Soluble Calcium Components of Oil Shale Waste Ash. 2016.

44. **Galina Sharajeva**. Thermochemical Destruction of Graptolite Argillite. 2016.

45. **Juri Bolobajev**. Effects of Organic Reducing Agents on the Fenton-like Degradation of Contaminants in Water with a Ferric Sludge Reuse. 2016.

46. **Can Rüstü Yörük**. Experimental and Modeling Studies of Oil Shale Oxy-fuel Combustion. 2016.

47. **Liina Kanarbik**. Ecotoxicological Evaluation of Shale Fuel Oils, Metal-Based Nanoparticles and Glyphosate Formulations. 2017.

48. **Natalja Pronina**. Degradation of Persistent Micropollutants in Suspended-Bed Reactor by Photocatalytic Oxidation and Combination of Biological Treatment with Photocatalysis. 2017.