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 ^a, Vahur Oja ^{a,*}

Originally published in Chemometrics and Intelligent Laboratory Systems, vol. 158, pp. [41–47](#), 2016.

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/cm³, and the root mean squared error of the predicted values was 0.004660 g/cm³, which is a relative error of less than 1%. In addition to the experimental portion, a literature review^b is included, which contains an assessment of the accuracy of chemometric methods for predicting many fuel properties.

Keywords: density, temperature dependence, Fourier-transform infrared spectroscopy (FTIR), oil shale, phenol rich oil, chemometrics

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

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

^b The information collected from the review is included as a supplement in the form of a spreadsheet.

* Corresponding author: vahur.oja@ttu.ee

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 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 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 Figure 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 Figure 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 Figure 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 also 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

Figure 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 Figure 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.

Figure 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.

2.2. 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.

[Table 1 is at the end of the article]

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 are shown in Figure 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 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.

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

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 Figure 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 Figure 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.

Figure 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).

2.5. Regression technique used

Figure 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.

Figure 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/cm³. The manufacture states the accuracy of the devices as being ± 0.000005 g/cm³, but for oil samples we expect the accuracy is a little bit poorer (about ± 0.00002 g/cm³) 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 Equation 1

$$\rho_T = \rho_* - \gamma(T - T_*) \quad (1)$$

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.

The data for each sample was fit using Equation 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 Equation 1 for the data in this study was only 0.0001054 g/cm³.

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]

Table 2. Parameters used for support vector regression.

Property	Density model	Slope (γ) model
C	48.0291570616	7.24675459478
Gamma	$4.1683544258 \cdot 10^{-6}$	0.000393626518193
Epsilon	$1.02093930394 \cdot 10^{-5}$	$3.34174597874 \cdot 10^{-5}$
Kernel function	Radial basis function	Radial basis function

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/cm³. This accuracy, however, is still much lower than the measurement accuracy of approximately ± 0.00002 g/cm³. The deviation for each sample is shown in Figure 5. Although the density model is accurate, its RMSE is still higher than those reported in previous studies (shown in Figure 2). The third quartile of the RMSEs given in previous studies was 0.003 g/cm³ (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 caused 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.

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

4.2. Model for a range of temperatures

As mentioned, the temperature dependence of density was described using a linear equation (Equation 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 $7.569 \cdot 10^{-6}$

$\text{g/cm}^3/^{\circ}\text{C}$. The residuals for each sample are shown in Figure 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.

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

By using the predicted density and slope values in Equation 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/cm³, which is an error of less than 1%. There was also a small bias of 0.0001839 g/cm³ 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.

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

Temperature	RMSE	
	Multi-temperature model	One-temperature models
15.6	0.003966	
20	0.004628	0.004628
35	0.004716	
40	0.004345	
50	0.005211	0.006030
60	0.004286	
65	0.005036	
80	0.004850	0.005577

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 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.

This technique could also be used with other types of fuel because many other liquid fuels exhibit a density-temperature relationship that is close to linear at moderate temperatures (temperatures below their boiling region). To obtain good results for a given fuel, it would likely be necessary to include reference samples with similar chemical compositions in the calibration set.

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”.

Supplementary data

The information collected during the literature review process is available online as a supplement to this article.

References

- [1] J. Persson and C. Albano, “Quantitative Thermogravimetry on Peat.,” pp. 1173–1178, 1986.
- [2] D. B. Hibbert, P. Minkinen, 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, 2009.
- [3] H. W. Chu, C. Lu, C. H. Huang, and S. Y. Fu, “Mobile fuel analysis apparatus and method thereof,” US 20080272303 A1, 2008.
- [4] S. Farquharson and 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, and 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, and V. Sanchez, “Use of NIR spectra for property prediction of bio-oils and fractions thereof,” US 8911512 B2, 2014.
- [8] W. R. Hruschka and H. Martens, “Principal Component Analysis Predicts Protein and Moisture Content from Near Infrared Spectra of Ground Wheat,” in *Pittsburg Conference on Analytical Chemistry and Applied Spectroscopy*, 1982, p. 375.
- [9] I. E. Frank, J. Feikema, N. Constantine, and B. R. Kowalski, “Prediction of product quality from spectral data using the partial least-squares method,” *Journal of Chemical Information and Modeling*, pp. 20–24, 1984.

- [10] 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.
- [11] R. E. Morris *et al.*, "Rapid fuel quality surveillance through chemometric modeling of near-infrared spectra," *Energy and Fuels*, vol. 23, no. 3, pp. 1610–1618, 2009.
- [12] 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, 2014.
- [13] M. Garci-a-Menci-a, "An empirical approach to update multivariate regression models intended for routine industrial use," *Fuel*, vol. 79, no. 14, pp. 1823–1832, 2000.
- [14] C. M., K. S., and H. B., "A unified recursive just-in-time approach with industrial near infrared spectroscopy application," *Chemometrics and Intelligent Laboratory Systems*, vol. 135, pp. 133–140, 2014.
- [15] K. He, H. Cheng, W. Du, and F. Qian, "Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy," *Chemometrics and Intelligent Laboratory Systems*, vol. 134, pp. 79–88, 2014.
- [16] K. He, F. Qian, H. Cheng, and W. Du, "A novel adaptive algorithm with near-infrared spectroscopy and its application in online gasoline blending processes," *Chemometrics and Intelligent Laboratory Systems*, vol. 140, no. 0, pp. 117–125, 2015.
- [17] C. He, Z. Yang, L. Chen, G. Huang, N. Liao, and L. Han, "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, 2014.
- [18] 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," *Journal of Chemometrics*, vol. 25, no. 9, pp. 496–505, 2011.
- [19] 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, 2012.
- [20] M. R. Riazi, *Characterization and Properties of Petroleum Fractions*, 1st ed. ASTM International, 2005.
- [21] S. Lee, *Oil Shale Technology*. Boca Raton, FL, USA: CRC Press, 1991.
- [22] R. M. Balabin and E. I. Lomakina, "Support vector machine regression (LS-SVM)--an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data?," *Physical Chemistry Chemical Physics : PCCP*, vol. 13, no. 24, pp. 11710–11718, 2011.
- [23] 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, 2007.
- [24] S. Derenne, C. Largeau, E. Casadevall, J. S. Sinninghe-Damste, E. W. Tegelaar, and J. W. de Leeuw, "Characterization of Estonian Kukersite spectroscopy and pyrolysis: evidence for abundant alkyl phenolic moieties in an Ordovician, marine, type II/I kerogen," *Organic Geochemistry*, vol. 16, no. 4–6, pp. 873–888, 1990.
- [25] Z. S. Baird, V. Oja, and O. Järvi, "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, 2015.
- [26] N. Golubev, "Solid Heat Carrier Technology for Oil Shale Retorting," *Oil Shale*, vol. 20, no. 3, pp. 324–332, 2003.
- [27] "ISO 647:1974 - Brown coals and lignites - Determination of the yields of tar, water, gas and coke residue by low temperature distillation." ISO, 2009.

- [28] “ASTM D86, Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure.” ASTM International, West Conshohocken, PA, USA, 2012.
- [29] 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., USA: United States Department of Commerce, 1916.
- [30] 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, no. 5, pp. 291–294, 1941.
- [31] Y. L. Rastorguev, “Methods of Assessing Fuel and Oil Quality,” *Khimiya i Tekhnologiya Topliv i Masel*, no. 9, pp. 56–60, 1971.
- [32] 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, 2012.
- [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. Cunnlgham, R. James, and G. M. Willson, “Thermophysical Properties of Coal Liquids. 1. Selected Physical, Chemical, and Thermodynamic Properties of Narrow Boilling Range Coal Liquids,” *Industrial and Engineering Chemistry Process Design and Development*, vol. 22, no. 3, pp. 410–424, 1983.
- [35] F. Pedregosa and G. Varoquaux, “Scikit-learn: Machine Learning in Python,” *Journal of Machine ...*, vol. 12, pp. 2825–2830, 2011.
- [36] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagne, “DEAP : Evolutionary Algorithms Made Easy,” *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, 2012.
- [37] E. Jones, E. Oliphant, P. Peterson, and et al., “SciPy: Open Source Scientific Tools for Python,” 2001. [Online]. Available: <http://www.scipy.org/>. [Accessed: 05-Jan-2016].

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	Accuracy statistics					Unit
			Min	1st Quartile	Median	3rd Quartile	Max	
acid number	15	8	0.003	0.016	0.137	0.208	0.320	mg KOH/g
aniline point	3	-	-	-	-	-	-	
api gravity	11	5	0.240	0.250	0.265	0.731	0.811	degrees API
aromatics	58	35	0.05	0.57	0.71	1.47	2.70	vol% or wt%
ash content	22	16	0.01	0.56	0.91	1.94	3.80	wt%
ash fusion temperature	5	3	70	-	77	-	90	°C
asphaltenes	15	13	0.13	0.42	0.69	2.08	2.60	wt%
boiling point	8	4	1.29	2.45	3.23	4.66	5.00	°C
carbon content	16	10	0.80	1.19	1.90	2.41	4.10	wt%
carbon residue	10	6	0.16	0.24	0.94	2.00	2.00	wt%
cetane index	22	15	0.20	0.43	0.60	1.44	2.00	
cetane number	24	16	0.27	0.49	1.01	1.95	2.11	
chemical composition*	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	-	1.00	°C
conductivity	3	2	0.84	-	-	-	77.00	pS/m
contaminant content	17	25	0.01	0.40	1.10	1.88	4.10	vol% or wt%
density	79	50	0.0004	0.0010	0.0018	0.0030	0.0280	g/cm ³
distillation temperatures	68	124	0.08	1.72	3.00	5.30	23.00	°C
fixed carbon	10	8	0.78	0.88	1.72	3.58	4.60	wt%
flash point	38	28	0.69	1.96	3.37	5.00	10.44	°C

fluid system icing inhibitor	4	3	0.007	-	0.009	-	0.010	vol%
freezing point	24	14	1.26	1.57	2.18	2.99	7.00	°C
grindability	4	2	3.84	-	4.16	-	4.47	
heat of combustion	38	21	0.013	0.134	0.300	0.585	1.900	MJ/kg
hydrogen content	26	14	0.016	0.073	0.120	0.245	0.920	wt%
hydrogen/carbon ratio	4	2	0.048	-	0.069	-	0.089	
iodine number	7	4	0.51	0.65	0.80	1.28	1.40	g I2/100 g
isoparaffins	4	4	0.17	0.51	0.64	1.00	1.12	wt%
lubricity	7	-	-	-	-	-	-	
mixture composition**	38	45	0.01	0.20	0.52	1.32	10.20	vol% or wt%
moisture content	17	13	0.09	0.41	0.69	0.99	2.50	wt%
naphthenes	11	10	0.20	0.25	0.38	0.52	1.90	vol% or wt%
nitrogen content	17	11	0.002	0.048	0.100	0.350	0.550	wt%
octane number	40	25	0.07	0.20	0.31	0.51	1.60	
olefins	18	12	0.07	0.21	0.30	2.44	4.30	vol% or wt%
other properties	51	-	-	-	-	-	-	
oxygen content	6	3	0.003	-	0.728	-	2.160	wt%
paraffins	10	9	0.12	0.24	0.4	0.67	1.3	vol% or wt%
pour point	14	9	2.06	2.40	4.65	9.10	9.20	°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	0.03400	0.25000	1.60000	wt%
thermal stability	4	2	46.0	-	46.7	-	47.3	degF
vapor pressure	11	6	1.04	1.77	3.27	4.60	5.99	kPa
water content	15	8	0.003	0.005	0.229	0.570	0.780	vol% or wt%
viscosity	52	31	0.026	0.098	0.152	0.242	22.000	cSt
vitronite reflectance	4	2	0.10	-	0.13	-	0.15	
volatile matter	15	10	0.5	1.0	1.4	3.7	7.0	wt%

Yield	9	-	-	-	-	-	-
-------	---	---	---	---	---	---	---

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

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