Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation

Werickson Fortunato de Carvalho Rocha [1] and David A. Sheen[2]

[1] National Institute of Metrology, Quality, and Technology-INMETRO, Division of Chemical Metrology,
25250-020 Duque de Caxias, RJ, Brazil,

[2] Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Keywords: Uncertainty analysis; bootstrap; partial least squares; support vector machine; physical-chemical properties; fuel

Abstract

The physicochemical properties of a substance, such as a fuel, can vary significantly with composition. Determining these properties with ASTM standard methods is both expensive and time-consuming, which has led to a desire to use chemometric modeling as an alternative. In this study, we compare the accuracy and robustness of two chemometric models, partial least squares (PLS) regression and support vector machine (SVM) with uncertainty estimation to determine how the physicochemical properties depend on the composition. A set of hydrocarbon mixtures, including crude oil, oil, gasoline, and biofuel/biodiesel, were collected. GC-MS data were taken, and physicochemical properties were measured for these mixtures using ASTM standard methods. PLS and SVM were used to develop predictive models of the physicochemical properties. Uncertainty in the estimated property values was estimated using a bootstrapping technique. With this uncertainty estimate, it is possible to assess the trustworthiness of any prediction, which ensures that the chemometric models can be applied for general purposes. SVM was found to be generally better for predicting the physicochemical properties, although we expect that with a more comprehensive data set the performance of the PLS models can be improved. We show in this work that PLS and SVM can be used to generate a predictive model of physicochemical properties based on GC-MS data. Combined with uncertainty analysis, these models provide robust predictions that can be used for regulatory, economic, and safety purposes.

1 Introduction

Alternative fuels have been identified as a critical need for the transportation industry, as detailed by the United States Federal Aviation Administration (FAA). In the future, it is expected that feedstocks will transition from entirely conventional petroleum sources to blends of conventional and unconventional sources (e.g. shale oil) and biomass. Fuels generated from these different feedstocks differ greatly from each other and from conventional fuels, meaning that their suitability may not be easily determined.

Certifying a fuel for aviation purposes is an onerous process that consumes millions of liters. Although alternative technologies to internal combustion exist in other transportation areas, it is not anticipated that these technologies can be easily applied to aviation in the same way. The most promising technology, the lithium-air battery, still faces challenges [1], which suggests that liquid fuels will be with us for the forseeable future. This places aviation in an auspicious position to be an industry leader in alternative fuels and applications.

Fuels are currently defined by their physicochemical properties such as viscosity and elemental composition, as ASTM D4814 for gasoline, ASTM D975 for petroleum diesels and ASTM D7467 for biodiesel blends [2-4]. In our previous work [5], we proposed using chemometric methods combined with a library of analytical profiles (such as chromatograms) for known fuels. A new fuel's suitability could be assessed by comparing its profile with fuels already in the library. Here, we propose using the same idea to develop a regression model for fuels' physicochemical properties. The relevant properties are measured for the substances in the fuels library and a regression model is fit to those measurements. This model can then be used to predict the physicochemical properties of any other fuel. Therefore, if a proposed fuel can be synthesized in quantities large enough to be used for an analytical technique such as chromatography, all of its physicochemical properties can be estimated immediately and thereby its potential range of suitability as a fuel can be determined.

Methods for determining the viscosity of hydrocarbon mixtures has been the subject of past research [6-9], although the relations proposed in the literature are empirical relations for known mixtures. Petroleum derivatives have a composition that is composed of hundreds or thousands of compounds, and modeling a relationship between composition and physicochemical behavior for mixtures of that complexity would require a similar number of correlation coefficients. Chemometric methods offer an alternative means of finding a relation between composition and physicochemical properties without needing to explicitly determine the composition. In this work, we use the ASTM standard methods for determining physicochemical properties and use those methods, along with profiles taken using gas chromatography, to generate regression models to predict the properties of an unknown substance. The goal of the study is not to replace the ASTM methods with the chemometric models, but rather to provide a model that can be used as a screening test to determine whether a fuel should be produced in quantity.

Chemometric methods of analysis have been successfully used to quantify properties [10-16] and monitor quality of fuels [17-19]. Palou et al [20] used calibration sets selection strategy for the construction of robust partial least squares (PLS) models for simultaneous determination of seven properties of biodiesel/diesel blends: density, cetane index, fatty acid methyl esters (FAME) content, cloud point, boiling point at 95% of recovery, flash point and Sulphur using near-infrared (NIR) spectroscopy. Pinto et al [19] used nuclear magnetic resonance spectroscopy of hydrogen (H-1 NMR) along with principal component analysis (PCA) and soft independent modelling of class analogies (SIMCA) to differentiate common and additive gasolines. Morales-Medina and Gusman [21] used principal components regression and PLS to predict viscosity and density of crude oils. Some authors [5, 22-29] have used non-linear methods, such as support vector machine (SVM) to study fuels. Rocha et al [30] used SVM for exploratory analysis of the different biodiesel samples. In that work, SVM could give the best classification results. Alves et al [31, 32] used SVM to determine biodiesel content in diesel fuel blends with more effective, accurate and appropriate results than for PLS. Balabin et al [23, 33] analyzed gasoline, ethanol-gasoline (bioethanol), and diesel fuel data with different chemometric methods. SVM models showed the best results to solve classical regression and interpolation/extrapolation tasks. In these studies, however, the substances used covered a fairly narrow range of potential fuel types, for instance gasoline only. There has been one study, by Cramer et al [34], which used PLS to predict many properties of fuels including viscosity, density, and cetane index and which developed a single model for a diverse set of fuels ranging from kerosene to marine diesel.

In this paper, we extend the proven work on chemometric regression models by adding an uncertainty analysis component. We employ the bootstrap-based uncertainty estimation algorithm used in our previous work [35] to place confidence bounds on the property estimates obtained using the chemometric models. This allows a more robust comparison between the ASTM-determined values and the chemometric model.

In Fig. 1, we show a schematic representation of the logic used in this work. The samples have their physical properties measured using the ASTM methods and likewise their chromatograms taken by GC-MS. This information is used as an input to the chemometric analysis (PLS and SVM). Then, based on how well the models perform with respect to the experimental data, we

Deleted: to the problem of estimating physicochemical properties across a large library of fuels. The work we perform here is new in that it requires a model that will work for such diverse substances as crude oils, marine diesels, jet fuels and kerosene, and biodiesel fuels. To our knowledge, such a study has not been performed before.¶

make a decision about which, if either, should be recommended. It should be noted that, since the models are trained against the ASTM data, they cannot be better than it; the question to be addressed is whether the models are a suitable surrogate.

1.1 Chemometric methods used

In this work, we study the use of partial least squares (PLS) and support vector machine (SVM) regression with uncertainty estimation to predict the physicochemical properties of the fuels. PLS and SVM, like all regression, finds a relationship between a set of independent variables **X** and dependent variables **y**. In this work, the **X** values are the GC-MS profiles obtained for the various substances and the **y** values are the physicochemical property values.

1.1.1 Partial least squares

Partial least squares (PLS) [30, 36, 37] is a linear regression technique that finds a relationship between **X** and **y** such that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. The method relies on a decomposition similar to principal components analysis which will find latent structures within **X** that are most closely correlated with variability in **y**.

1.1.2 Support vector machine

Support vector machines [38-41] are a class of learning algorithms that determine hypersurfaces that separate the data. In the linear case, these surfaces are hyperplanes. For a classification problem, the SVM finds a single surface that separates the groups of samples. For a regression problem such as that posed here, the SVM essentially finds the contour surfaces of **y** within **X**.

1.2 Bootstrap uncertainty estimation

Bootstrapping is used to estimate uncertainty in a statistical model in terms of confidence limits. The procedure involves creating many artificial data sets by random sampling to get an estimate of the uncertainty. Using bootstrapping, it is possible to calculate statistics in model outputs such as variances and confidence intervals that may be used to represent the uncertainty in the model outputs. In this work, the regression model is calculated as

$$\mathbf{y} = \mathbf{M}\left(\mathbf{X}\right) + \mathbf{F} \tag{1}$$

where **M** is the regression model, $\mathbf{y}_{\text{pred}} = \mathbf{M}(\mathbf{X})$ is the prediction of that model, and **F** is the residual in the model predictions. Here, we use the residual bootstrapping technique developed by Almeida, et al. [42] To conduct the bootstrap, the weighted residual $\mathbf{F}_{\text{w}}$ is calculated using

$$\mathbf{F}_{\text{w}} = \frac{\mathbf{F}}{\sqrt{1 - D_f / K}} \tag{2}$$

where K is the number of independent variables and D_f is the number of pseudo-degrees of freedom [43]. A new properties vector **y*** is generated by drawing with replacement from $\mathbf{F}_{\text{w}}$, that is, $\mathbf{y}^* = \mathbf{M}\left(\mathbf{X}\right) + \text{bootstrap}\left(\mathbf{F}_{\text{w}}\right)$ where bootstrap(**A**) means sampling with replacement from the elements of **A**. A population of **y*** vectors can be generated by executing bootstrap many times, and then calculating a new model **M*** for each **y***, along with a corresponding $\mathbf{y}^*_{\text{pred}} = \mathbf{M}^*\left(\mathbf{X}\right)$, very similar to a standard bootstrap. Confidence limits are calculated using percentiles of $\mathbf{y}^*_{\text{pred}}$; for the 95% CI, the limits are the 2.5 and 97.5 percentiles.

2 Methods

2.1 Experimental data

The experimental fuel samples used in this work consisted of a set of NIST Standard Reference Materials (SRMs), augmented with military aviation fuels obtained from the Air Force Research Laboratory and with commercially available gasoline. This sample set is the same as used in our previous work [5] and is detailed in Table 1. GC-MS chromatograms were obtained for these samples in triplicate, and the details can be found in our previous work [5]. In that study, retention times were obtained up to 150 minutes and m/z values from 50 to 300 Da. For this study, a set of physicochemical property measurements typical for heavier fuels was measured, namely the density ρ at two temperatures, the kinematic viscosity ν at two temperatures, the pour point T_{p}, and the total acid and base numbers N_{A} and N_{B}. The properties and the relevant ASTM test methods used to calculate them are shown in Table 1. The values for the properties are shown in Table 2.

As shown in Table 2, some density and viscosity measurements for the commercial gasoline as well as the pour point value for SRM 2723b were removed from the data set before regression. Including these measurements resulted in highly degraded performance for the PLS models (and sometimes the SVM models), which suggested there was an issue either with the chromatograms or with the property value measurements.

2.2 Multivariate analysis

The GC-MS data, which form the **X** matrix, are arranged into as a three-way array with dimension $k \times i \times j$, representing the samples, GC retention times, and mass spectra respectively. This array was unfolded into a matrix with dimension k x ij. For each physicochemical property, there is a **y** vector consisting of the k measured property values. Because many of the GC-MS array elements are zero, principal components analysis was used to reduce the dimensionality of the GC-MS arrays. Principal components explaining more than 1 % of the variance were retained, and the largest loadings for each array element were identified. Array elements were retained if the largest loading was greater than a specified cutoff, which was chosen as 0.003; this cutoff resulted in about 65 000 GC-MS array elements retained for the analysis.

Partial least squares regression and support vector regression with the radial basis function kernel are used. The PLS equation is

$$\mathbf{y} = \mathbf{X}'\mathbf{W}\mathbf{q} \tag{3}$$

where X**′** is a matrix of spectra whose properties are being predicted by the model, of size k' x ij. **W** is a transformation matrix that is of size $ij \times l$, where l is the number of latent variables, a free parameter, and q is a vector of size $l \times 1$. The SVM equation is

$$\mathbf{y}\left(\mathbf{X}'\right) = \left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{*}\right)\mathbf{K}\left(\mathbf{X}, \mathbf{X}'\right) + b \tag{4}$$

where $\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{*}\right)$ is the vector of dual coefficients whose nonzero values identify the support vectors and **K** is the $k \times k$ kernel matrix, whose elements are $K_{ij} = \exp\left(-\gamma \left|\mathbf{x}_i - \mathbf{x}'_j\right|^2\right)$ for the radial basis function. The parameter γ can be reinterpreted as a characteristic distance $\sigma = 1/\sqrt{2\gamma}$, which describes how far the influence of each sample extends within feature space.

As mentioned above, there are three separate GC-MS chromatograms and three property values for each of the samples. When constructing the **X** and **y** arrays for regression, we use these three values as representative of the measurement uncertainty in these processes. Each experimental sample is represented by nine "virtual" measurements, representing each possible chromatogram and physicochemical property value pair. This allows a representation of the measurement uncertainty in the chemometric model. Consequently, *k* is approximately 150, although this varies from property to property since some properties could not be measured for some substances.

Before fitting the model, the **X** and **y** arrays are split into a calibration and validation set using a random shuffle procedure. For each property, approximately 30% of the samples are held out in the validation set. In addition, this splitting procedure is stratified to ensure that, for a given substance, roughly the same proportion of its virtual measurements appear in the calibration and validation sets. This is not ideal; the reason is because there is considerable variability from chromatogram to chromatogram for a substance and also between substances. For instance, the animal and plant biodiesel fuels are different from each other and from the other fuels, meaning that if either of these substances were only in the validation set, then suddenly they would be outside the range of applicability of the model. As this work is a pilot study with a small set of fuels, we feel that these concerns are relatively minor. A large study to thoroughly investigate the relationships between GC-MS data and physical properties would have more data that would cover the space of possible fuels more densely and thus not be likely to have these problems.

2.3 *Software implementation*

The PLS and SVM models were implemented using scikit-learn 0.19.1, run in Anaconda 4.0.0 with Python 3.6.0. Uncertainty in the PLS and SVM models was calculated using the bootstrap machine-learning uncertainty module developed in our previous work [35].

3 Results and discussion

3.1 *Analysis of the chromatograms and selected variables*

The petroleum derivatives samples are all hydrocarbon mixtures of similar origin and their compositions are expected to be qualitatively similar. On the other hand, biodiesel samples consist basically of fatty acid methyl esters (FAME). These substances from petroleum derivatives and biodiesel samples can be seen in the variables which are selected using chemometric analysis, shown in Fig. 2. In addition, we show in Fig. 3 the chromatograms for SRM 2773 (biodiesel), SRM 1624d (Diesel), the commercial gasoline, and SRM 2779 (crude oil) as representative samples. Note that, in these figures, we do not display data for times greater than 110 minutes and m/z values greater than 175 Da because few significant variables were identified outside this range. Several groups of significant features can be identified, associated with characteristic compounds of the various fuels. At low retention times (< 10 minutes) there are clusters of features that can be associated with alkanes and olefins of up to 9 carbons, as well as benzene, xylene, trimethyl benzene, and isomers. These substances are common in gasolines and the peaks show prominently in the commercial gasoline chromatogram. At retention times of up to 100 minutes, there are broad features associated with alkanes, olefins, and cycloalkanes of between four and seven carbons, as well as features that are consistent with both small polycyclic aromatic hydrocarbons and alkanes

Deleted: in Fig. 3

and olefins of 12 or fewer carbons, all present in diesel and crude oils. Additionally, a cluster of features can be seen between 60 and 62 minutes and between 69 and 73 minutes of retention time, with *m*/*z* values consistent with carboxylic acids, and fatty acid methyl esters. These features prominently identify the biodiesels, which are almost exclusively composed of these substances.

3.2 Chemometric models and overall precision

Individual multivariate calibration models were developed for the quantification of physicochemical properties of fuels using PLS and SVM models with uncertainty estimation. PLS and SVM both have free parameters whose best values must be estimated. The best parameters are found by using a 5-fold cross-validation grid search. This kind of analysis consists in breaking the calibration set into five partitions, with the model calibrated against four of the five folds and then validated against the fifth. The procedure is then repeated for each partition. The optimal parameters are those with the smallest root mean squared error of cross-validation (RMSECV). For PLS, the free parameter is the number of latent variables (LVs), while for SVM, the free parameters are the cost parameter *C*, the kernel coefficient γ, and the threshold ε. The number of LVs (for PLS) and the values of *C*, γ, and ε (for SVM) can be viewed as a measure of the complexity of the model, determining the tradeoff between minimizing the training error and minimizing the model complexity [44, 45].

The parameters obtained for construction of PLS and SVM models for each property are presented in Table 3. As mentioned above, as these parameters increase in value, the model becomes more complex. This can represent increased chemical complexity in the system, that is, that more material components of the mixture are interacting in more complex ways. For large parameter values, this could indicate overfitting. Conversely, small parameter values can indicate model form errors, such that the relationship between the physicochemical property and the GC-MS chromatogram cannot be captured by this kind of model. The PLS models all have between 20 and 30 LVs, except for the viscosity v_{60}, with 13. In the case of SVM, γ (for the RBF) indicates how much influence measurements have on nearby regions of GC-MS space; larger values mean less influence, which allows for a more complex and varied model but at the risk of overfitting. C indicates how large a penalty must be used to fit the measurements to the SVM, which also increases complexity of the model at the risk of overfitting. ε defines minimum residual at which the penalty starts to be applied. All of the models have a C of around 10^6 and ε values that are negligibly small. Because the γ values describe a distance over which each sample's influence extend, one metric for the model is how many other samples are within that distance. For v_{60}, and T_p, that number is nearly half the total number of samples, and for v_{40} it is about 5% of them. For the other properties, it is about 10 %. Values close to 100 % would suggest that the model is close to linear, while smaller values indicate a more nonlinear and possibly overfit model. It should be noted that, since there are 9 virtual measurements for each substance, 10 % is approximately the threshold at which all chromatograms for one substance are within each other's distance of influence. This suggests that most of the SVM model results should be taken with a great deal of care.

Deleted: 6

When there are multiple models of varying complexity, it is necessary to have some means of choosing between those models. The SVM model with the RBF kernel is more internally complex than the PLS model, which allows it to more effectively match a given calibration set, but at the expense of overfitting to that calibration set and ruining the model's predictive power. We consider two comparisons between the PLS and SVM models, both of which rely on comparing the models' performance on the validation set. Traditionally, this comparison is done by calculating the root mean squared error of prediction (RMSEP) for the two models and using an F-test. The $\mathfrak{F}$ value is calculated using

$$\mathfrak{F} = \left(\frac{e_{\mathrm{SVM}}}{e_{\mathrm{PLS}}} \right)^2 , \tag{3}$$

where e is the RMSEP value. The SVM more accurately represents the data, to a $p > 0.05$ level of confidence, if $\mathfrak{F}$ is greater than about 1.65 for these data sets. The critical value will vary a small amount because not all properties have the same number of measured values.

In addition, since there is an uncertainty on each prediction, we can assess whether the PLS and SVM models predict the experimental measurement value to within any arbitrary confidence interval. We describe this as a hit function H, defined to be $H_i = 1$ if y_i is within the confidence interval of $\mathbf{M}(\mathbf{X}_i)$, and 0 otherwise. The number of hits is then sum of H. By comparing the number of hits in the validation set, we can obtain an uncertainty-based estimate of the models' relative performance.

The RMSEP and number of misses is presented in Table 4, along with the $\mathfrak{F}$ values from Eq. 3, which can be used to gauge the relative performance of the PLS and SVM models. In addition, shown for comparison are the root-mean-squared error of calibration (RMSEC) and the Pearson R^2 values (denoted R_c^2 and R_p^2 for the calibration and validation sets). The $\mathfrak{F}$ values show that the SVM model conclusively ($p < 0.05$) outperforms the PLS model for all the properties except for v_{60} and T_p,. Interestingly, these are the same properties for which the SVM produces the least-nonlinear models, which suggests that, for this set of data, a linear model is the best that can be done for those properties. The relative number of misses shows that the SVM models for ρ_{60} and v_{40} conclusively outperform the PLS models, but it is difficult to decide between the others. We can say, then, that there is a statistical basis for selecting the SVM model over the PLS model for all properties except v_{60} and T_p. The R^2 values for the SVM models are almost all very close to 1, however, so there is some concern that the models are overfit.

3.3 Predictions of the chemometrics models

The F-test results are only valid for a population of measurements, so to understand any individual measurement it is necessary to plot the measurements and predictions individually. We show the model predictions of the PLS and SVM models for ρ_{60}, v_{60}, T_p, and N_B in Fig. 4. These properties can be viewed as representative examples; we also show all predictions in Fig. S1. We also present the performance of each model on each target substance; the PLS performance is shown in Fig. S2 and the SVM performance in Fig. S3.

As a rule, the PLS models have larger uncertainty than the SVM models (e.g. 1 kg/m^3 vs. 20 kg/m^3 for the density measurements). Furthermore, there are four substances with which the PLS models have difficulties, namely SRMs 2722 (a crude oil), 2771 (a Diesel) and SRM 2723b (also a Diesel) and the commercial gasoline. Property values are accurately reproduced for all other substances. Conversely, the SVM models predict all the measurements, sometimes with a few misses per substance. Given that there are many Diesel and crude oils in our sample library, the fact that the PLS models have difficulty with these examples suggests there is something unusual in the chromatograms.

3.4 Physical interpretation of the chemometric models

Given the ability of both the PLS and SVM models, in general, to predict the physicochemical properties, it is useful to interrogate them to gauge what spectral features, and therefore which components, are correlated with increasing or decreasing which properties. In the case of PLS, the model can be directly interrogated with the linear model coefficients **WQ** from Eq 3. The SVM models can be interrogated with the *p*-vector [46], which is defined by

$$\mathbf{p} = \left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{*}\right)\mathbf{X}. \tag{5}$$

This vector is a highly qualitative description of the impact each spectral feature has on the SVM result, but it has been used with some success in petroleomics to interpret SVM models [47]. We present the coefficients of the PLS models, as well as the *p*-vectors of the SVM models for ρ_{60}, v_{60}, T_{p}, and N_{B}, in Fig. 5. The remaining properties show patterns like one of these four, and their respective coefficients and *p*-vectors are shown respectively in Figs. S4 and S5. As we said earlier, this study is intended to be a proof-of-concept, and so we do not attempt to attach a specific chemical identity to the spectral features; indeed, as the samples are mostly petroleum derivatives, this would be an unwieldy task. Instead, we merely show that the models produce results that make some sense when viewed from a chemical perspective.

For all of the properties we investigated, the spectral features with the largest coefficients are relatively short alkane chains at long scan times, suggesting that the properties are most strongly affected by heavy, low-volatility species that elute poorly. In addition, there is varying influence from lighter substances. In the PLS model for v_{60}, for instance, the viscosity is decreased strongly by fast-eluting single-ring aromatic ions and increased strongly by polycyclic compounds.

Furthermore, if we look at the SVM *p*-vectors in Fig S5, we can see that many of properties show wildly varying influence from all available spectral features, suggesting that the models are "reaching" so that they can reproduce as many substances' properties as possible but in a way that is not generalizable. From this perspective, the SVM models for ρ_{60}, v_{60}, T_p, and N_A are the most trustworthy, which is the same conclusion drawn from Table 3.

By themselves, the coefficients and *p*-vectors can be difficult to interpret, so we can also interrogate the influence of each spectral feature on each substance's property value using,

$$I_{ik} = \begin{cases} (\mathbf{WQ})_i X_{ik} & \text{PLS} \\ p_i X_{ik} & \text{SVM} \end{cases}, \tag{6}$$

where I_{ik} is the influence of spectral feature *i* on the property value for substance *k*. Taking the sum of *I* over all substances, for PLS, would then yield the property value for that substance; for SVM, the interpretation of *I* is more holistic but still valuable. We show these influence values on ρ_{60} for four representative substances in Fig. 6, and on v_{60} in Fig. 7; the influence values for all substances are shown in Figs. S6 and S7. Here, we can see that the long-elution-time features that are so prominent in Fig. 5, are mostly associated with the crude oil and the residual oil. It also becomes clearer that the properties of SRMs 2723b and 2771, as well as the commercial gasoline, are poorly reproduced by the PLS models because the influence plots for those fuels closely resemble those for the crude oils, despite the widely varying physicochemical properties. That is to say, the features identified by the PLS regression as significant for each property are similar in these chromatograms, and so PLS cannot distinguish these substances. Furthermore, the influence vectors of the SVM models again suggest that many of the models are reaching to reproduce measurements but doing so in a way that is not easily generalized.

Deleted: crude

Deleted:

3.5 Recommendation for choice of model

Based on the results from Table 3 and Fig. S5, we can conclude that the SVM models for ρ_{15}, v_{40}, and N_B are likely overfit and should not be used. From Table 4, we can conclude that the PLS models for v_{60} and T_p are statistically comparable to the corresponding SVM models. Consequently, we would recommend use of PLS for ρ_{15}, v_{40}, v_{60}, T_p, and N_B, and SVM for the remaining parameters.

4 Conclusions

In this work, linear and non-linear methods with uncertainty were used to estimate physicochemical properties of fuels using GC-MS data. The properties were measured using ASTM standard methods, and partial least squares and support vector machines were used to determine a relation between the GC-MS data and the measured properties.

The fuels examined spanned a wide range, from lighter fuels such as gasoline, through jet fuels, kerosenes, crude oils, to heavy marine diesels. This ensured that the models could cover a wide range of potential fuels, without being restricted to any one class. The size of the fuel sample set was not meant to represent a realistic fuels library, but rather to be representative; a real library would have dozens or hundreds of samples from a much broader range of available fuels. Properties investigated included density and viscosity at several temperatures as well as pour point and acid and base content. Generally, the support vector machine was best able to reproduce the physical properties using the GC-MS data, although partial least squares proved significantly better for viscosity at higher temperature, and somewhat better for the pour point. However, after examining the support vector machine models for evidence of overfitting, it was evident that the partial least squares models were most trustworthy for most of the physical properties.

This paper has shown that using chemometric regression models can lead to useful results, even when used on a wide library of fuels that one would normally expect not to be comparable. Much work remains to be done, however. To be practically useful, we would need a fuels library that covers the available fuel space more densely than that used here. Furthermore, much of the uncertainty in the model predictions appears to come from variability in the GC-MS data, and future iterations of this project would need to take steps to minimize this variability.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology or the National Institute of Metrology, Quality and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Acknowledgements

The authors thank Dr. Katrice Lippa and Dr. Pam Chu at NIST for productive discussions.

References

[1] K.-N. Jung, J. Kim, Y. Yamauchi, M.-S. Park, J.-W. Lee, and J.H. Kim, *Rechargeable lithium-air batteries: a perspective on the development of oxygen electrodes*, Journal of Materials Chemistry A 4 (2016), pp. 14050-14068.

[2] *ASTM D4814-17 Standard Specification for Automotive Spark-Ignition Engine Fuel*, ASTM International, West Conshohocken, PA, 2017.

[3] *ASTM D975-17 Standard Specification for Diesel Fuel Oils*, ASTM International, West Conshohocken, PA, 2017.

[4] *ASTM D7467-17 Standard Specification for Diesel Fuel Oil, Biodiesel Blend (B6 to B20)*, ASTM International, West Conshohocken, PA, 2017.

[5] W.F.D. Rocha, M.M. Schantz, D.A. Sheen, P.M. Chu, and K.A. Lippa, *Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data*, Fuel 197 (2017), pp. 248-258.

[6] R.M. Díaz, M.I. Bernardo, A.M. Fernández, and M.B. Folgueras, *Prediction of the viscosity of lubricating oil blends at any temperature*, Fuel 75 (1996), pp. 574-578.

[7] R.E. Maples, *Petroleum Refinery Process Economics*, PennWell, Tulsa, OK, 2000.

[8] A. Naseri, M. Nikazar, and S.A. Mousavi Dehghani, *A correlation approach for prediction of crude oil viscosities*, Journal of Petroleum Science and Engineering 47 (2005), pp. 163-174.

[9] P. Saxena, S. Jawale, and M.H. Joshipura, *A Review on Prediction of Properties of Biodiesel and Blends of Biodiesel*, Procedia Engineering 51 (2013), pp. 395-402.

[10] G. Mendes, and P.J.S. Barbeira, *Detection and quantification of adulterants in gasoline using distillation curves and multivariate methods*, Fuel 112 (2013), pp. 163-171.

[11] R.D.M. Scafutto, and C.R. de Souza, *Quantitative characterization of crude oils and fuels in mineral substrates using reflectance spectroscopy: Implications for remote sensing*, Int. J. Appl. Earth Obs. Geoinf. 50 (2016), pp. 221-242.

[12] C.E. Freye, B.D. Fitz, M.C. Billingsley, and R.E. Synovec, *Partial least squares analysis of rocket propulsion fuel data using diaphragm valve-based comprehensive two-dimensional gas chromatography coupled with flame ionization detection*, Talanta 153 (2016), pp. 203-210.

[13] N. Dupuy, Z. Brahem, S. Amat, and J. Kister, *Near-Infrared Spectroscopy Analysis of Heavy Fuel Oils Using a New Diffusing Support*, Appl. Spectrosc. 69 (2015), pp. 1137-1143.

[14] B. Kehimkar, B.A. Parsons, J.C. Hoggard, M.C. Billingsley, T.J. Bruno, and R.E. Synovec, *Modeling RP-1 fuel advanced distillation data using comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry and partial least squares analysis*, Anal. Bioanal. Chem. 407 (2015), pp. 321-330.

[15] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, and R.E. Synovec, *Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis*, J. Chromatogr. A 1327 (2014), pp. 132-140.

[16] D.A. Saldana, L. Starck, P. Mougin, B. Rousseau, and B. Creton, *On the rational formulation of alternative fuels: melting point and net heat of combustion predictions for fuel compounds using machine learning methods*, SAR QSAR Environ. Res. 24 (2013), pp. 525-543.

[17] G.T. Tanaka, F.D. Ferreira, C.E.F. da Silva, D.L. Flumignan, and J.E. de Oliveira, *Chemometrics in fuel science: demonstration of the feasibility of chemometrics analyses applied to physicochemical parameters to screen solvent tracers in Brazilian commercial gasoline*, J. Chemometr. 25 (2011), pp. 487-495.

[18] A.C. Neto, E.C.S. Oliveira, V. Lacerda, E.V.R. Castro, W. Romao, R.C. Silva, R.G. Pereira, T. Sten, P.R. Filgueiras, and R.J. Poppi, *Quality control of ethanol fuel: Assessment of adulteration with methanol using H-1 NMR*, Fuel 135 (2014), pp. 387-392.

[19] V.S. Pinto, F.F. Gambarra-Neto, I.S. Flores, M.R. Monteiro, and L.M. Liao, *Use of H-1 NMR and chemometrics to detect additives present in the Brazilian commercial gasoline*, Fuel 182 (2016), pp. 27-33.

[20] A. Palou, A. Miró, M. Blanco, R. Larraz, J.F. Gómez, T. Martínez, J.M. González, and M. Alcalà, *Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 180 (2017), pp. 119-126.

[21] G. Morales-Medina, and A. Guzmán, *Prediction of Density and Viscosity of Colombian Crude Oils from Chromatographic Data*, CT&F - Ciencia, Tecnología y Futuro 4 (2012), pp. 57-73.

[22] C.E.C. Galhardo, and W.F.D. Rocha, *Exploratory analysis of biodiesel/diesel blends by Kohonen neural networks and infrared spectroscopy*, Anal. Methods 7 (2015), pp. 3512-3520.

[23] R.M. Balabin, and S.V. Smirnov, *Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data*, Anal. Chim. Acta 692 (2011), pp. 63-72.

[24] R. Fernandez-Varela, J.M. Andrade, S. Muniategui, D. Prada, and F. Ramirez-Villalobos, *Identification of fuel samples from the Prestige wreckage by pattern recognition methods*, Mar. Pollut. Bull. 56 (2008), pp. 335-347.

[25] R.M. Balabin, and E.I. Lomakina, *Support vector machine regression (SVR/LS-SVM)-an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data*, Analyst 136 (2011), pp. 1703-1712.

[26] M.A. Oloso, M.G. Hassan, M.B. Bader-El-Den, and J.M. Buick, *Ensemble SVM for characterisation of crude oil viscosity*, Journal of Petroleum Exploration and Production Technology (2017).

[27] R. Piloto-Rodríguez, Y. Sánchez-Borroto, M. Lapuerta, L. Goyos-Pérez, and S. Verhelst, *Prediction of the cetane number of biodiesel using artificial neural networks and multiple linear regression*, Energy Conversion and Management 65 (2013), pp. 255-261.

[28] C.I. Rocabruno-Valdés, L.F. Ramírez-Verduzco, and J.A. Hernández, *Artificial neural network models to predict density, dynamic viscosity, and cetane number of biodiesel*, Fuel 147 (2015), pp. 9-17.

[29] H. Yang, Z. Ring, Y. Briker, N. McLean, W. Friesen, and C. Fairbridge, *Neural network prediction of cetane number and density of diesel fuel from its chemical composition determined by LC and GC–MS*, Fuel 81 (2002), pp. 65-74.

[30] W.F.C. Rocha, B.G. Vaz, G.F. Sarmanho, L.H.C. Leal, R. Nogueira, V.F. Silva, and C.N. Borges, *CHEMOMETRIC TECHNIQUES APPLIED FOR CLASSIFICATION AND QUANTIFICATION OF BINARY BIODIESEL/DIESEL BLENDS*, Anal. Lett. 45 (2012), pp. 2398-2411.

[31] J.C.L. Alves, and R.J. Poppi, *Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration*, Fuel 165 (2016), pp. 379-388.

[32] J.C.L. Alves, and R.J. Poppi, *Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM)*, Talanta 104 (2013), pp. 155-161.

[33] 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*, Analyst 137 (2012), pp. 1604-1610.

[34] 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 28 (2014), pp. 1781-1791.

[35] W.F.C. Rocha, and D.A. Sheen, *Classification of biodegradable materials using QSAR modelling with uncertainty estimation*, SAR QSAR Environ. Res. 27 (2016), pp. 799-811.

[36] S. Wold, M. Sjostrom, and L. Eriksson, *PLS-regression: a basic tool of chemometrics*, Chemometr Intell Lab 58 (2001), pp. 109-130.

[37] W.F.D. Rocha, R. Nogueira, and B.G. Vaz, *Validation of model of multivariate calibration: an application to the determination of biodiesel blend levels in diesel by near-infrared spectroscopy*, J. Chemometr. 26 (2012), pp. 456-461.

[38] S.S. Qiu, and J. Wang, *The prediction of food additives in the fruit juice based on electronic nose with chemometrics*, Food Chem. 230 (2017), pp. 208-214.

[39] R.E. Stern, J. Song, and D.B. Work, *Accelerated Monte Carlo system reliability analysis through machine learning-based surrogate models of network connectivity*, Reliab Eng Syst Safe 164 (2017), pp. 1-9.

[40] M. Maalouf, and M. Abutayeh, *Improved Modeling of Solar Flash Desalination Using Support Vector Regression*, J Energ Eng 143 (2017).

[41] J.S. Chou, and A.D. Pham, *Nature-inspired metaheuristic optimization in least squares support vector regression for obtaining bridge scour information*, Inform Sciences 399 (2017), pp. 64-80.

[42] M.R. de Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi, and R.J. Poppi, *Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation*, Microchem J 109 (2013), pp. 170-177.

[43] H. van der Voet, *Pseudo-degrees of freedom for complex predictive models: the example of partial least squares*, J. Chemometr. 13 (1999), pp. 195-208.

[44] F. Liu, Y. Jiang, and Y. He, *Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer*, Anal. Chim. Acta 635 (2009), pp. 45-52.

[45] A.S. Luna, I.C.A. Lima, W.F.C. Rocha, J.R. Araujo, A. Kuznetsov, E.H.M. Ferreira, R. Boque, and J. Ferre, *Classification of soil samples based on Raman spectroscopy and X-ray fluorescence spectrometry combined with chemometric methods and variable selection*, Anal. Methods 6 (2014), pp. 8930-8939.

[46] B. Üstün, W.J. Melssen, and L.M.C. Buydens, *Visualisation and interpretation of Support Vector Regression models*, Anal. Chim. Acta 595 (2007), pp. 299-309.

[47] P.R. Filgueiras, C.M.S. Sad, A.R. Loureiro, M.F.P. Santos, E.V.R. Castro, J.C.M. Dias, and R.J. Poppi, *Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration*, Fuel 116 (2014), pp. 123-130.