2. Experimental

2.1. Sample sets

Eighteen (18) different sample sets were used in this study (Table 1). These sets include eight sets of gasoline data, including density, benzene content, octane number, and fractional composition/boiling points; two sets of ethanol-gasoline biofuel data, including density and ethanol content; three sets of diesel fuel data, including total sulfur content, flash point, and viscosity; three sets of petroleum macromolecular data, including weight percentage of asphaltenes, resins, and paraffins in toluene; one set of petroleum resins data, the resin content in benzene; and one set of biodiesel data, the diesel content in biodiesel fuel [41]. In all cases, NIR spectra (Table 1) were used to build calibration models to predict the desired system property. See refs.[12-15,41] for detailed descriptions of the data sets used. Table 1 summarizes the main parameters of interest for all 18 data sets. See refs.[12-15] for a detailed discussion of the reference data collection for each particular case.

2.2. NIR apparatus and experimental parameters

All NIR spectra (except those for diesel and biodiesel) were acquired with an NIR FT Spectrometer InfraLUM FT-10 (LUMEX, Russia) fitted with a special sampler for liquids. See Table 2 in the previous publication by Balabin et al. [12] for detailed spectrometer parameters. The spectra were acquired at room temperature (20-23 °C). Background spectra were recorded before and after each measurement to compensate for the absence of thermostating. The averaged background spectrum was subtracted from the sample spectrum before all pre-processing procedures. This resulted in an analytical signal with satisfactory accuracy and precision. The instrument calibration for wavelength and transmittance was performed using four pure hydrocarbons: toluene (C_7H_8), normal hexane (*n*-C_6H_{14}), benzene (C_6H_6), and isooctane (*iso*-C_8H_{18}). This calibration was repeated at least once per day to ensure stability of the experimental setup and data accuracy and reproducibility.

The NIR spectra of diesel fuel and biodiesel were collected using a MPA Multi Purpose FT-NIR Analyzer (Bruker) at room temperature. The MPA NIR spectrometer was calibrated with benzene and cyclohexane (*c*-C_6H_{12}) at least twice per day to minimize the influence of variable laboratory conditions. The spectral range between 11,000-4000 cm^{-1} (909-2500 nm) was scanned with a resolution of 8 cm^{-1}. Sixty-four scans were averaged for each spectrum. A background spectrum was measured every 45 min. A cylindrical glass cell with an 8 mm optical path was used throughout this study. Approximately 1 mL of diesel sample was required for each NIR measurement, which is much less than the 200 mL needed for

distillation analysis, for example, to determine fractional composition [41]. The NIR spectrum collection was repeated five times with cell rotation inside the spectrometer between repetitions to minimize the interference from the cell or glass defects. Measurement of one sample took less than five minutes. The averaged and background-corrected spectrum was used for subsequent data pre-processing.

See refs.[12-15] for experimental spectra examples and discussion.

2.3. Model efficiency estimation of interpolation and extrapolation

To characterize the prediction ability and efficiency of the created regression model, the root mean squared error of prediction (RMSEP) was calculated for each case. A validation set was constructed as one-fifth of all samples from every sample set (i.e., 19 out of 95 gasoline samples, 24 out of 120 diesel fuels, etc.). The validation set consisted of samples from the entire property range [41].

The interpolation task was performed in the *Y*-space by deleting the central 20% of the samples (Figures 2 & 3, top) and using the remaining data as an independent prediction set. This method allows us to create an artificial gap inside the calibration range where no training samples were present. Note, there was a genuine gap around approximately 715 kg m^{-3} in gasoline density (Figure 2; see the above discussion about the necessity for interpolation in real-world experimental data).

The extrapolation problem (Figures 2 & 3, bottom) was created in the *Y*-space by deleting the samples with the lowest (10%) and the highest (10%) *Y*-values, 20% in total. This type of a problem is expected to be much more complicated for multivariate methods [47]. The choice of 20% as a fraction of samples for prediction is the usual choice in many MDA projects [5-20].

The mean average percentage error (MAPE) was also calculated to estimate the relative accuracy of each calibration model. This is especially important for properties with a large range, such as sulfur in diesel fuel or petroleum resin content in benzene. See refs.[12-14] for the exact formulas and additional discussion.

Five-fold or ten-fold cross-validation was used to optimize the model's parameters based on the root mean squared error of cross-validation (RMSECV). The cross-validation set consisted of samples from the entire property range. Other variants of the cross-validation procedure, e.g., 7-fold cross-validation and leave-one-out cross-validation (LOOCV), were checked and found to produce almost identical results.

2.4. NIR spectra pre-processing and outlier detection

Different types of spectra pre-processing (pre-treatment) methods were used, including normalization, magnitude normalization, multiplicative scatter correction (MSC), linearization, Savitzky-Golay differentiation, Savitzky-Golay double differentiation, autoscaling, and range scaling at different intervals. The best pre-processing technique was found for each calibration method and for each chemical system property. See refs.[12-15,41] for detailed discussions of each particular system.

2.5. Spectra reduction and feature selection

In order to create an effective and robust regression model, the spectra data, which has up to 10^4 independent variables, should be reduced. Two common data reduction techniques, spectra averaging and principal component analysis (PCA), were used to achieve this goal for the LS-SVM, SVR, and ANN methods. Note that PLS-based techniques have an intrinsic data reduction ability through the specification of the latent variables. The PCA results are reported because this technique was found to produce the best results and the lowest errors in all cases. The other methods of feature selection (wavelets, UVE-PLS, etc. [52,53]) are out of the scope of the current study.

2.6. Multivariate regression methods

We refer the readers to refs.[12-15,41] and references therein for detailed descriptions of all of the regression methods used. To conserve space, we have not repeated all of the descriptions and formulas here.

2.7. Method optimization

To compare the different classification models, the best results from each model need to be obtained; otherwise, the comparison is useless. The results from each model depend on the model parameters. We have used a wide range of model parameters to achieve the best results. RMSECV minimization was used for optimization in all cases and for all models.

These parameters and the corresponding model are as follows:

- PLS: the number of latent variables (LV);

- Poly-PLS: LV and degree of polynomial (n);

- ANN/MLP: number of input neurons (IN; equal to the number of principal components, PC), number of hidden neurons (HN), and transfer function of hidden layer: $f(\mathbf{x})$ = {*logsig*; *tansig/tanh*}.

Detailed procedures for ANN training can be found in ref.[12]; see, for example, Table 4 in ref.[12] for the ANN training procedure for gasoline data.

- SVR: the error weight (C), maximal error value (ε), and kernel-related parameters. The same set of kernels (linear, polynomial, and radial basis function (RBF)) was used for SVR and LS-SVM model building. See ref.[41] for the parameter definitions and other clarifications.

- LS-SVM/LSSVM: the regularization parameter (γ), which determines the trade-off between the fitting error minimization and the smoothness of the estimated function, and the kernel-related parameters (e.g., σ or σ^2 for the RBF kernel, Table 2). See ref.[41] for the parameter definitions and other clarifications.

The regression methods were optimized based on a cross-validation procedure and tested using fully independent test (validation) sets (see also above).

2.8. Software and computing

MATLAB 2008b and 2011a were used as the standard software for multivariate method realization. The following toolboxes were used: MATLAB Statistics Toolbox, MATLAB Support Vector Machine Toolbox, MATLAB Neural Network Toolbox, N-way Toolbox for MATLAB, and PLS_Toolbox Version 4.0. For the SVR calculations, a MATLAB toolbox developed and described by Gunn was used [54]. The LS-SVM regression model was built using the LS-SVMlab1.5 MATLAB toolbox [55]. Refs.[54,55] contain detailed descriptions of the algorithms and procedures. The standard programs of these toolboxes were modified and extended by BRM (see also refs.[12,14]).

2.9. Chemical systems: their quality and representativeness

The current study deals with five chemical systems of petroleum and one of vegetable origin. The first group consists of gasoline, a classical sample for analytical chemistry in general and chemometrics in particular [12]; ethanol-gasoline biofuel, an increasingly popular type of motor fuel that is partly produced from renewable sources that may have a colloid (dispersed) structure [41]; diesel fuel, a product of petroleum refining with a higher boiling range than gasoline due to a more complicated mixture of hydrocarbons and heteroatomic compounds [41]; a solution of all three classes of petroleum macromolecules, asphaltenes (the molecules responsible for the colloid structure formation in crude oil [41]), resins, and paraffins, in an aromatic solvent, such as toluene, in which each macromolecule class is an extremely complicated mixture; and a petroleum resin solution in benzene,

which is used to calibrate the NIR setup for adsorption studies [15]. The biodiesel sample set consists of vegetable oil- or animal fat-based diesel fuels consisting of alkyl (methyl) esters. Biodiesel was made by chemically reacting lipids (vegetable oil) with an alcohol (methanol) [56,57]. Details about some of these systems have been published during the last 5 years by Balabin and co-workers [12-15,28-30,39,41,43,47].

The systems presented here greatly range in composition, properties, and behavior. While low-molecular weight substances with 6-12 carbon atoms and low intermolecular forces, such as *n*-hexane, heptane isomers, and isooctane, form gasoline [41], heavy (above 500 Da) molecules with a high tendency for aggregation and phase separation, like resins and asphaltenes, are found in four systems [41]. The number of effective components ranges from one in petroleum resins to millions. Therefore, rather general conclusions about algorithm behavior can be made based on the systems studied.

2.10. Sample sets

The properties of the six petroleum/biofuel systems described above form eighteen sample sets that are very different in nature (Table 1). For gasoline, these include the density at 20 °C; the fractional composition, including the initial boiling point (IB), end boiling points 10%, 50%, and 90% v/v (T10, T50, and T90, respectively), and the final boiling point (FB); and the benzene content. For ethanol-gasoline fuel, these sample sets include the density at 20 °C and the ethanol content [EtOH]. For diesel fuel, the sample sets are based on the total sulfur content [Sulfur], flash point (FP) and viscosity. For the petroleum macromolecules, the sets are asphaltene content [A], resin content [R], and paraffin content [P]. For the petroleum resins, the relevant sample set is the resin concentration in benzene [Res]. Finally, the diesel fuel content in the biodiesel sample, [Diesel], is the last sample set. See refs.[12-17,41,58] for detailed discussions of the industrial importance of the quality parameters discussed.

Note that the quality, in terms of the accuracy, repeatability, and reproducibility, of the reference data ranges greatly from one property to another (Table 1). It is important to estimate the effect of the initial data quality on the final prediction results. The same can be said about the property ranges, some of which are rather limited (e.g., T50), and others are very broad (e.g., [Sulfur] or [R]). In industrial applications, it is usually impossible to model the quality (in terms of either the accuracy or range) of the data sets. Therefore, the machine learning algorithms that show very good, even brilliant, results for the model systems do not always show the same results when applied to real-world problems

[45-47]. In this work, we tried to use a wide-range of reference data qualities to help make our conclusions as general as possible.

The spectroscopic information for most sample sets (Table 1) was recorded in the short-wave part of the NIR region (above 8000 cm^{-1}). This is the region with the second to fifth overtones of the characteristic molecular vibrations observed by standard IR and Raman techniques [41]. The only exclusion is the diesel/biodiesel sample set because its spectrum lies in the 4000-11,000 cm^{-1} region. In this particular case, it was important to obtain information from the long-wave part of the NIR spectrum to predict the sulfur concentration in diesel samples [22].

The number of samples in the sample sets ranged from 57 to 186 (Table 1). Because the number of samples can influence the quality of the multivariate model prediction, we tried to ensure that sample set saturation or complete sample set (CSS/CoSSt) was observed, at least in the case of the simplest (PLS) method, which is similar to the basis set limit (BSL) or the complete basis set (CBS) methods in quantum chemistry.

Table S1. The root means squared error of prediction ($RMSEP_{int}$) for all eighteen (18) NIR data sets in an "interpolation mode" (see Figure 4).

Petroleum system	Property	Unit	$RMSEP_{int}$				
			PLS	Poly-PLS	ANN	SVR	LS-SVM
Gasoline[a]	Density at 20 °C	$kg\,m^{-3}$	5.9	7.3	5.8	4.3	4.0
	Initial boiling point (IB)	°C	3.7	4.7	5.0	2.8	2.7
	End boiling point 10% v/v (T10)	°C	4.4	5.7	5.3	3.5	3.0
	End boiling point 50% v/v (T50)	°C	5.0	7.0	5.1	3.1	2.8
	End boiling point 90% v/v (T90)	°C	4.3	7.4	6.7	4.7	3.6
	Final boiling point (FB)	°C	6.6	6.5	5.0	3.7	4.4
	Octane number (ON)	a.u.	3.45	6.07	3.05	2.49	2.04
	Benzene content[b]	% w/w	1.86	2.51	2.22	1.02	1.10
Biofuel: ethanol-gasoline[b]	Density at 20 °C	$kg\,m^{-3}$	5.69	7.98	6.87	3.66	3.64
	Ethanol content[b]	% w/w	0.45	0.62	0.51	0.32	0.29
Diesel fuel	Total sulfur content	ppm	741	1229	471	361	269
	Flash point (FP)	°C	18.25	23.90	7.90	4.59	4.33
	Viscosity	$cSt\,/\,mm^2\,s^{-1}$	0.300	0.297	0.156	0.041	0.043
Petroleum macromolecules[c]	Asphaltene content	% w/w	0.72	0.58	0.51	0.25	0.22
	Resin content	% w/w	1.32	1.88	0.79	0.50	0.42
	Paraffin content	% w/w	0.61	0.84	0.25	0.21	0.19
Petroleum resins in benzene[d]	Resin content	$mg\,L^{-1}$	2.6	4.1	4.8	3.5	2.5
Biodiesel	Diesel content	% v/v	0.276	0.150	0.206	0.066	0.065

Table S2. The root means squared error of prediction ($RMSEP_{ext}$) for all eighteen (18) NIR data sets in an "extrapolation mode" (see Figure 5).

Petroleum system	Property	Unit	$RMSEP_{ext}$				
			PLS	Poly-PLS	ANN	SVR	LS-SVM
Gasoline[a]	Density at 20 °C	$kg\,m^{-3}$	11.5	21.4	22.7	8.4	5.9
	Initial boiling point (IB)	°C	7.2	15.1	39.0	4.7	4.0
	End boiling point 10% v/v (T10)	°C	6.0	27.2	50.5	4.5	6.1
	End boiling point 50% v/v (T50)	°C	6.3	28.0	50.3	4.6	3.8
	End boiling point 90% v/v (T90)	°C	11.9	24.3	57.1	7.0	5.5
	Final boiling point (FB)	°C	6.3	30.5	48.5	5.6	5.0
	Octane number (ON)	a.u.	3.34	11.91	9.01	2.76	1.87
	Benzene content[b]	% w/w	3.38	8.14	7.54	1.75	1.78
Biofuel: ethanol-gasoline[b]	Density at 20 °C	$kg\,m^{-3}$	8.43	28.85	19.25	6.06	5.77
	Ethanol content[b]	% w/w	0.60	3.02	1.11	0.50	0.48
Diesel fuel	Total sulfur content	ppm	736	3181	2021	416	441
	Flash point (FP)	°C	28.8	48.30	43.3	10.2	7.81
	Viscosity	$cSt\,/\,mm^2\,s^{-1}$	0.305	0.581	0.451	0.098	0.084
Petroleum macromolecules[c]	Asphaltene content	% w/w	1.43	2.41	2.24	1.09	1.00
	Resin content	% w/w	1.98	6.31	2.27	0.99	0.75
	Paraffin content	% w/w	1.48	3.21	2.47	0.83	0.48
Petroleum resins in benzene[d]	Resin content	$mg\,L^{-1}$	9.0	20.0	26.6	9.9	5.8
Biodiesel	Diesel content	% v/v	0.699	0.931	1.106	0.166	0.097