Available online at www.sciencedirect.com
ScienceDirect
Talanta 70 (2006) 344–352
Talanta
www.elsevier.com/locate/talanta

Assessment of infrared spectroscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricating oils

Arnobio Roberto Caneca [a], M. Fernanda Pimentel [a,*], Roberto Kawakami Harrop Galvão [b], Cláudia Eliane da Matta [c,d], Florival Rodrigues de Carvalho [a], Ivo M. Raimundo Jr. [e], Celio Pasquini [e], Jarbas J.R. Rohwedder [e]

[a] *Departamento de Engenharia Química, Universidade Federal de Pernambuco (UFPE), Av. Prof. Artur de Sá S/N, Cidade Universitária, 50740-521 Recife, PE, Brazil*
[b] *Divisão de Engenharia Eletrônica, Instituto Tecnológico de Aeronáutica (ITA), Brazil*
[c] *Divisão de Ciência da Computação, Instituto Tecnológico de Aeronáutica (ITA), Brazil*
[d] *Centro Universitário Salesiano de São Paulo, Unidade de Lorena, Brazil*
[e] *Instituto de Química, Universidade Estadual de Campinas (UNICAMP), Brazil*

Received 5 November 2005; received in revised form 18 February 2006; accepted 18 February 2006
Available online 17 April 2006

Abstract

This paper presents two methodologies for monitoring the service condition of diesel-engine lubricating oils on the basis of infrared spectra. In the first approach, oils samples are discriminated into three groups, each one associated to a given wear stage. An algorithm is proposed to select spectral variables with good discriminant power and small collinearity for the purpose of discriminant analysis classification. As a result, a classification accuracy of 93% was obtained both in the middle (MIR) and near-infrared (NIR) ranges. The second approach employs multivariate calibration methods to predict the viscosity of the lubricant. In this case, the use of absorbance measurements in the NIR spectral range was not successful, because of experimental difficulties associated to the presence of particulate matter. Such a problem was circumvented by the use of attenuated total reflectance (ATR) measurements in the MIR spectral range, in which an RMSEP of 3.8 cSt and a relative average error of 3.2% were attained.

Keywords: Viscosity; Lubricating oil; Infrared spectroscopy; Multivariate calibration; Discriminant analysis; SPA

1. Introduction

Lubricants play a key role in extending the working life of rotating machines. In order to maintain a proper lubrication, it is important not only to use oils with suitable properties but also to monitor their state of degradation in a periodic manner. In fact, contaminations may compromise the lubricating capability of an oil, which increases the wear of the machine components, as well as the risk of mechanical collapse.

The tests currently employed to assess lubricant properties are time-consuming and require specific equipments for the determination of each parameter of interest (for instance, kinematic viscosity and flash point) [1,2]. In this context, the use of spectroscopy in conjunction with multivariate calibration techniques has been proposed as a multi-parametric alternative to the present methods. In particular, middle (MIR) and near (NIR) infrared spectroscopy offers several advantages for this type of application, such as high sample throughput, non-destructiveness and low cost [3]. Moreover, compact instruments in the MIR and NIR ranges can be realized for field use.

Much research has been conducted on the analysis of oil products by IR spectroscopy, for both classification and calibration purposes [3–5]. As regards lubricants, Lima et al. [6] studied the correlation between NIR spectra and the carcinogenic potential of basic oils employing principal component regression (PCR). Sastry et al. [7] used MIR spectroscopy and partial least squares regression (PLS) to determine the chemical composition (paraffins, isoparaffins, naftenes, aromatics and heteroaromatics) and

* Corresponding author. Tel.: +55 81 21267291; fax: +55 81 21267235.
E-mail addresses: mfp@ufpe.br, mf-pimentel@uol.com.br (M.F. Pimentel), kawakami@ele.ita.br (R.K.H. Galvão), ivo@iqm.unicamp.br (I.M. Raimundo Jr.), pasquini@iqm.unicamp.br (C. Pasquini), jarbas@iqm.unicamp.br (J.J.R. Rohwedder).

0039-9140/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2006.02.054

its influence on the physico-chemical properties (viscosity and viscosity index) in lubricants of mineral basis. MIR [8–11] and NIR [12] spectroscopy have been employed for the prediction of contaminants, degradation products and additives employing PCR, PLS and interval-PLS. The potentiality of MIR spectroscopy for the prediction of viscosity in lubricating oils for locomotives [10] and diesel engines [13] was assessed in a small-size set of samples (20 and 40 samples, respectively) by PCR and interval-PLS.

This paper proposes two strategies for monitoring the condition of lubricating oils being used in diesel engines by means of near and middle infrared spectroscopy. The first strategy is a qualitative approach formulated in the context of pattern classification. In this case, the samples are categorized in three classes according to their stage of use (short, medium, and long-term use) and classification is performed by discriminant analysis. In order to circumvent ill-conditioning problems, the dimensionality of the problem is reduced by using a variable selection algorithm. This algorithm is aimed at maximizing the discriminability of the spectral variables included in the model while avoiding multicollinearity problems. For comparison, a conventional KNN (K-nearest neighbours) classifier is also employed.

The second strategy is a quantitative approach that employs IR spectroscopy and multivariate calibration techniques in order to predict viscosity, which is the main control parameter for lubricants in service. In this case, MLR (multiple linear regression), PCR, and PLS techniques are employed in the calibration. The effect of different pre-processing procedures, as well as the utility of variable selection, is assessed by means of a factorial design study.

2. Background and theory

2.1. Qualitative analysis: Classification with respect to the stage of use

The classification method adopted in this work is based on the classic discriminant analysis technique, which assumes that the objects follow a gaussian distribution within each class. Under this assumption, the probability density function $p_j(\mathbf{x})$ for the objects $x = [x_1\ x_2\ \ldots\ x_d]^T$ belonging to the jth class is of the form:

$$p_j(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}_j)}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right] \quad (1)$$

where $\boldsymbol{\mu}_j$ $(d \times 1)$ and $\boldsymbol{\Sigma}_j$ $(d \times d)$ are the mean vector and the covariance matrix, respectively, which can be estimated from a set of training objects of known classification [14,15]. Variable x_i corresponds to the absorbance measured at the ith wavelength monitored by the spectrometer. Henceforth, with a slight abuse of language, the terms variable/wavelength and object/spectrum will be used with the same meaning.

In a problem involving C classes with equal a priori probabilities, the classification of a given object $\mathbf{x}$ is performed by calculating $p_j(x)$, $j = 1, 2, \ldots, C$ and by taking class j for which $p_j(x)$ is maximum. Such a classification rule is known as quadratic discriminant analysis (QDA) [15] because the decision boundaries defined by $p_{j1}(x) = p_{j2}(x)$, $j1 \neq j2$, are quadratic surfaces.

Simpler boundaries can be realized by adopting the regularization hypothesis that the covariance matrices are equal, that is $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_C = \boldsymbol{\Sigma}$. In this case, the decision surfaces are hyperplanes and the resulting classification rule is known as linear discriminant analysis (LDA) [15].

Both QDA and LDA usually benefit from a convenient selection of spectral variables [16]. In fact, if the number of variables employed in the classification model is large as compared to the number of training objects, the decision boundaries may be subject to overfitting and the resulting classifier is likely to have a poor generalization ability. Such a problem is aggravated in the presence of significant collinearity between the classification variables [17]. In the present work, a stepwise selection algorithm that takes into account both the discriminant power of each variable and the collinearity between variables is proposed.

2.2. Proposed variable selection algorithm for qualitative analysis

The proposed algorithm evaluates the individual value of each spectral variable according to its discriminability (as defined in Appendix A) with respect to the classes under consideration. At each step, the variable x_i with the largest discriminability D_i is selected, a *leave-one-out* cross-validation procedure is performed [14], and the number of errors is noted. Before the next step, the variables that are highly correlated with those already selected are discarded in order to avoid collinearity problems. The algorithm stops when no more variables are available. The set of variables that resulted in the smallest number of cross-validation errors is then presented to the analyst. Such a selection strategy can be summarized as follows.

Let A and B be the index sets for the variables already selected and those still available, respectively. Moreover let L be a correlation threshold $(0 < L < 1)$ established by the analyst. In what follows, N is a counter that indicates the number of variables already selected.

Step 0 (initialization). $A = \{\}$, $B = \{1, \ldots, d\}$, $N = 0$.

Step 1. Calculate D_i for $1 \leq i \leq d$.

Step 2. $i^* = \arg\max D_i,\ i \in B$.

Step 3. Move i^* from B to A. Let $N = N + 1$.

Step 4. Perform a leave-one-out cross-validation procedure using the variables with indexes in A. Let ECV(N) be the number of resulting cross-validation errors.

Step 5. Calculate the coefficient of multiple correlation r_i of each variable x_i with index in B with respect to the variables with indexes in A.

Step 6. Exclude from B the indexes of variables with coefficient of multiple correlation larger than L.

Step 7. If $B \neq \{\}$, return to Step 2.

Step 8. The optimum number n^* of variables is obtained from the minimum of ECV(n), $n = 1, \ldots, N$. The selected variables correspond to the first n^* indexes in A.

Remark 1. The coefficient of multiple correlation employed in Step 5 is defined, for each variable x_i, as:

$$r_i = \frac{\sigma(\hat{x}_i)}{\sigma(x_i)} \tag{2}$$

where $\sigma(.)$ denotes the standard deviation calculated in the training set and $\hat{x}_i$ is an estimate of x_i obtained by multiple linear regression from the variables already selected. If r_i is close to one, it can be concluded that the inclusion of x_i does not bring additional information into the classification model, because the values of x_i can be predicted from the variables already selected.

Remark 2. If there are several values of n associated to the minimum number of cross-validation errors ECV(n), the smallest n is chosen in Step 8. Such a choice is based on the Parsimony Principle [14], which states that, given classification models with similar prediction ability, the simplest one (smallest number of variables) should be favoured.

Remark 3. The selection procedure is performed both for LDA and QDA. If an LDA and a QDA model lead to the same number of cross-validation errors, the model with the smallest number of variables is favoured, as discussed above. If both models have the same number of variables, LDA is favoured because it employs simpler decision surfaces.

2.3. Quantitative analysis: multivariate regression for viscosity prediction

The multivariate regression methods most frequently used in infrared spectroscopy are multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS). PCR uses principal components provided by principal component analysis (PCA) to perform regression. PCA finds directions of greatest variability by considering spectral information, whereas PLS uses both spectral and target-property information. PCR and PLS have the ability to overcome problems common to IR data, such as collinearity, band overlap and interactions. MLR is the simplest quantitative multivariate analysis method, yielding models which are simpler and easier to interpret than PCR and PLS, since these calibration techniques perform regression on latent variables, which may not be amenable to a straightforward physical interpretation. On the other hand, MLR is more sensitive to collinearity problems, and usually requires a judicious choice of spectral variables.

Selecting from the full spectrum the wavelengths that result in the maximum accuracy for regression models is still a challenging task. Several approaches have been proposed to select optimal sets of variables for multivariate calibration, such as the use of mutual information [18], simulated annealing [19,20], genetic algorithms [21–23], artificial noise introduction in PLS modelling [24], hybrid linear analysis [25], cyclic subspace regression [26], iterative predictor weighting PLS [27], and discriminant PLS [28]. Among these different variable selection strategies, genetic algorithms (GAs) have become very popular owing to their simplicity and flexibility. GAs are guided random search techniques inspired on natural selection mechanisms, which explore the solution space in an efficient manner and are suitable for parallel processing implementations.

A recently proposed wavelength selection strategy for MLR calibration, the successive projections algorithm (SPA), was specifically designed to remove collinearity from the data set in order to improve numerical conditioning and reduce noise propagation [29,30]. SPA has been successfully employed for variable selection in UV–vis [29], ICP-OES [30] and NIR [31] spectrometry. In all those applications, SPA led to MLR models with better predictive ability than PLS or PCR models employing the full spectrum. Moreover, the results reported in Refs. [30,31] provided evidence that SPA yields MLR models with better prediction performance than a genetic algorithm.

SPA works on the basis of a calibration (cal) and a validation (val) sets, consisting of instrumental response data ($\mathbf{X}$) and parameter values measured by a reference method ($\mathbf{y}$). The main operations in SPA consist of algebraic manipulations carried out on matrix $\mathbf{X}_{\text{cal}}$ (Kc × J), whose rows and columns correspond to Kc calibration samples and J spectral variables, respectively. Starting from a column $\mathbf{x}_0$ (which is associated to the initial variable of the selection), SPA determines which of the remaining columns has the largest projection on the subspace S_0 orthogonal to $\mathbf{x}_0$. This column, denoted by $\mathbf{x}_1$, can be regarded as the one that contains the largest amount of information not included in $\mathbf{x}_0$. At the next iteration, SPA restricts the analysis to subspace S_0, taking $\mathbf{x}_1$ as the new reference column, and proceeds with the steps described above. Thus, the selection criterion in SPA favours the minimization of collinearity between the variables. Moreover, no more than Kc variables can be selected in this manner. In fact, after each projection operation, the dimension of the column space of $\mathbf{X}_{\text{cal}}$ is reduced by one (that is, one degree of freedom is removed). Thus, after Kc projection operations all the column vectors of $\mathbf{X}_{\text{cal}}$ will have been projected onto the origin of the space, that is, $\mathbf{X}_{\text{cal}}$ will have become a null matrix.

The determination of the best initial variable (column of $\mathbf{X}_{\text{cal}}$) and the optimum number N of variables is carried out as follows. If N is fixed, J subsets of N variables can be selected, using each of the J available variables as a starting point for SPA. For each of those variable subsets, an MLR model is calibrated and the root-mean-square error of prediction in the validation set (RMSEV) is calculated as

$$\text{RMSEV} = \sqrt{\frac{1}{\text{Kv}}\sum_{k=1}^{\text{Kv}}(y_{\text{v}}^k - \hat{y}_{\text{v}}^k)^2} \tag{3}$$

where y_{v}^k and $\hat{y}_{\text{v}}^k$ are the reference and predicted values of the parameter of interest in the kth validation sample and Kv is the number of validation samples. The smallest RMSEV thus obtained is denoted by RMSEV$^*(N)$, where the star is used to indicate the best result for subsets of N variables. By repeating this procedure for $N = 1, 2, \ldots,$ Kc (note that N cannot be larger than Kc, as explained above) the optimum N can be obtained from the minimum of the RMSEV$^*(N)$ curve. To

Table 1
Partitioning of the samples in training and test sets

Set	Class			Total
	1	2	3	
Training	20	39	26	85
Test	9	10	10	29

save computational time, the analyst may interrupt the procedure before N reaches Kc if the reduction in RMSEV$^*(N)$ after a corner point is minor or if the curve starts to increase after a local minimum point.

3. Experimental

A set of 114 samples of lubricating oil (TURBO 15W40, Petrobras) for diesel engines in different stages of use was employed in this work. The samples were collected from an urban transportation company that operates in the city of Recife, Brazil.

Used lubricating oils for diesel engines display a very dark colour and a substantial amount of particulated matter, which prevents direct determinations in the NIR range. Attempts at minimizing this problem by means of centrifugation and filtration were unsuccessful. Attempts at reducing the optical path length were also unsuccessful. The samples were then diluted with toluene at the proportion of 1:5 (v/v). In order to inspect the spectra with minimum solvent influence, representative samples were also diluted in carbon tetrachloride. However, such an option would not be practical for routine use, because of its toxicity.

The NIR spectra were acquired in the 3996–14,000 cm^{-1} range (714–2500 nm) with an ABB Bomem MB 160D spectrophotometer fitted with a Hellma transflectance probe. A spectral resolution of 8 cm^{-1} and an optical path length of 2 mm were employed. The reference spectra were obtained with toluene or tetrachloride, according to the solvent used for dilution of the sample.

In the middle infrared, the spectra were acquired in the 650–4000 cm^{-1} range (2500–14,000 nm) with an FT-IR Perkin Elmer Spectrum GX spectrophotometer fitted with an ATR probe. A spectral resolution of 8 cm^{-1} was employed and air was used as reference.

All acquisitions, both in the NIR and MIR range, were carried out at room temperature (25 ± 1 °C).

3.1. Qualitative analysis

For classification purposes, the oil samples were grouped in three classes according to their stage of use: class 1 (short-term use—less than 5000 km), class 2 (medium-term use—from 5000 to 20,000 km) and class 3 (long-term use—more than 20,000 km). The samples were divided in training and test sets as shown in Table 1.

Classification was performed both with original and derivative spectra. The derivative spectra were obtained after smoothing by a Savitzky–Golay filter with a second-order polynomial and a 5-point window. Moreover, a preliminary elimination of variables with low signal-to-noise ratio was carried out by discarding the variables for which the maximum signal intensity over all derivative spectra did not exceed 10% of the maximum signal intensity in the entire data set. Furthermore, in the NIR derivative spectra, the spectral range closer to visible (10,130–14,000 cm^{-1}) was discarded because of the high level of noise caused by scattering.

3.2. Quantitative analysis

Determinations of kinematic viscosity at 40 °C were carried out according to the ASTM D445 method [2]. Two out of the 114 oil samples were deemed outliers because of abnormal viscosity values, which were ascribed to errors in the viscosity determination procedures.

The remaining 112 samples were divided into calibration, validation, and prediction sets with 64, 25, and 23 samples, respectively. The validation set was employed for the selection of PLS/PCR factors by external validation, and for the selection of wavelengths in SPA and GA. The prediction set was used for the final assessment and comparison of the models. The adopted figure of merit was the root-mean-square error in the prediction set (RMSEP).

The GA employed standard binary chromosomes with length equal to the number of wavelengths in the spectrum (a "1" gene indicates a selected wavelength) [22]. The fitness of each individual was taken as the inverse of the RMSEV (Eq. (3)) calculated by using the wavelengths coded in the chromosome. The probability of a given individual being selected for the mating pool was proportional to its fitness (roulette method). One-point crossover and mutation operators were employed with probabilities of 60 and 10%, respectively. Population size was kept constant, each generation being completely replaced by its descendants. The GA was carried out for 150 generations with 80 chromosomes each. Moreover, the algorithm was repeated 20 times, starting from different random initial populations. The best solution resulting from the 20 realizations of the GA was adopted.

A 2^3 factorial design was employed to assess the influence of pre-processing and variable selection procedures in the predictive ability of the resulting model. The factors under consideration were spectrum differentiation, smoothing (5-point Savitsky–Golay with second-order polynomial) and variable selection. For PLS and PCR, the low and high design levels for variable selection were no selection at all (i.e., use of full spectrum) and GA selection, respectively. For MLR, the low and high levels were SPA and GA selections, respectively.

4. Results and discussion

Fig. 1a presents spectra of toluene-diluted lubricating oils in three different stages of use. It is worth noting that the absorption of toluene, used as reference, is stronger than that of the oil, and therefore negative peaks are observed in the spectra. For comparison, Fig. 1b depicts the spectra of the same samples

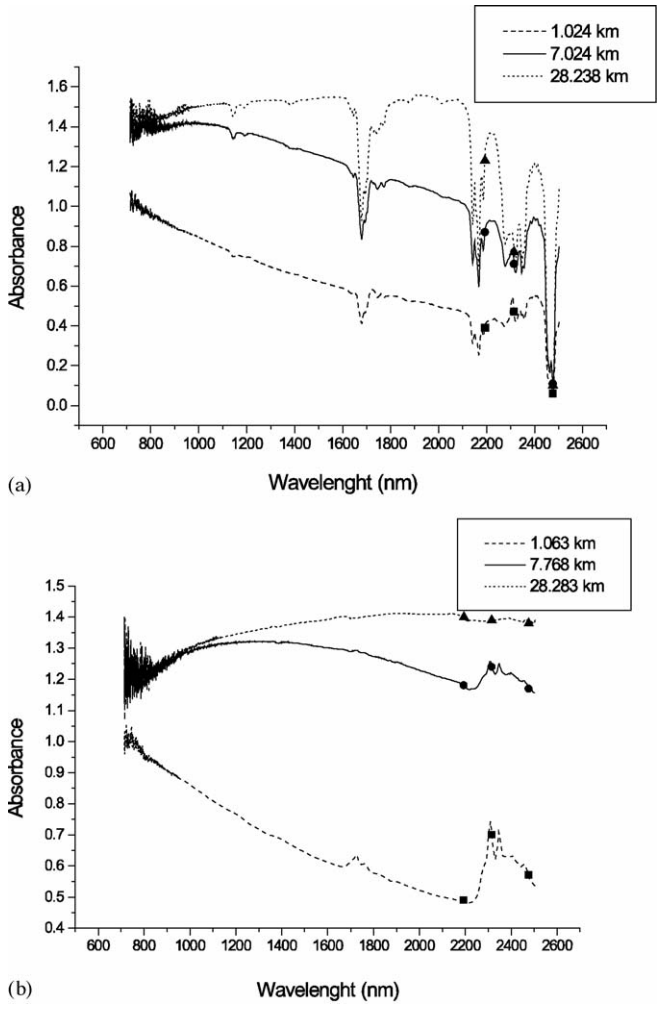

Fig. 1. Typical NIR spectra of (a) toluene-diluted and (b) carbon tetrachloride-diluted lubricating oils in different stages of use. The markers indicate the three wavelengths selected for qualitative analysis by QDA after pre-processing.

diluted with carbon tetrachloride, in which positive absorption bands appear, since carbon tetrachloride does not absorb in this region. As can be seen in both Fig. 1a and b, a positive baseline shift is associated with an increasing wear of the lubricant. Such a finding may be ascribed to the panchromatic absorption of particulate matter [32]. A related effect, also described in Ref. [32], consists of a decrease in the size of the absorption bands in Fig. 1b.

The spectra of the same lubricating oils in the MIR range are displayed in Fig. 2. By comparing Figs. 1 and 2, it can be seen that in the MIR range the peaks are narrower and more intense and that the baseline shift caused by particulate matter is less noticeable than in the NIR spectra. The region between 3000 and 4000 nm comprises low-intensity bands associated to the ring deformation of C–H in the aromatic ring superimposed to the high-intensity bands of ring deformations of CH_3-, CH_2- groups. The bands close to 6500 nm are distinctive of aromatic groups present in the samples. Bands of symmetric angular deformation of CH_2- groups, including CH_2–S bonds, are found in the region near 7500 nm. The bands ascribed to ring vibration of C–C bonds are weak and appear in the

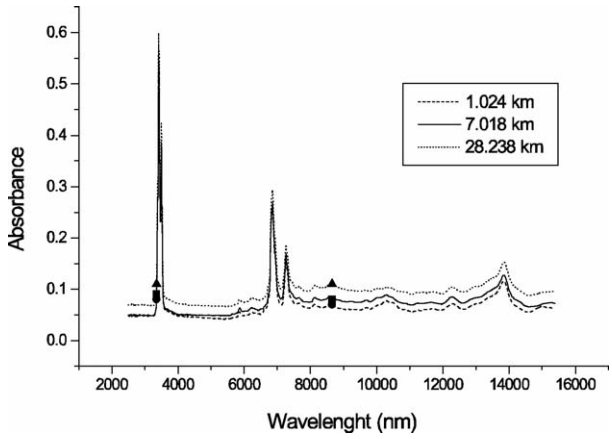

Fig. 2. Typical MIR spectra of lubricating oils in different stages of use. The markers indicate the two wavelengths selected for qualitative analysis by QDA.

Table 2
Number of classification errors in the MIR data set

	Original	Pre-processed
DA	2 (QDA, 2 wavelengths)	12 (LDA, 4 wavelengths)
KNN	6 ($k=28$)	14 ($k=1$)

For DA, the discriminant type (QDA or LDA) and the number of wavelengths are indicated, whereas for KNN, the number k of nearest neighbours is given.

region between 8300 and 12,500 nm. Finally, the region between 12,500 and 15,000 nm encompasses the absorption of several functional groups, including polynuclear aromatics, other aromatic groups and alkenes.

4.1. Qualitative analysis

Table 2 presents the classification results for DA and KNN in the MIR range. As can be seen, the best result was obtained with QDA by employing two wavelengths without pre-processing. In this case, only 2 out of 29 test objects were incorrectly classified, leading to a classification accuracy of 93%. The errors consisted of a class 2 and a class 3 objects included in classes 1 and 2, respectively. It is worth noting that, if the pre-processing is carried out, the number of errors increases both for DA and KNN. It could be argued that, in the MIR range, baseline information, which is removed by the derivative procedure, is important to discriminate the classes.

Table 3 presents the classification results in the NIR range. In this case, the best results for DA and KNN were obtained after the pre-processing procedures, unlike in the MIR range. It could be argued that baseline fluctuations in the NIR spectra cause a large within-class variability compared to the between-class

Table 3
Classification results for the NIR data set

	Original	Pre-processed
DA	8 (LDA, 1 wavelength)	2 (QDA, 3 wavelengths)
KNN	7 ($k=13$)	5 ($k=9$)

For DA, the discriminant type (quadratic or linear) and the number of wavelengths are indicated, whereas for KNN, the number k of nearest neighbours is given.

variability, which should be removed to improve the separation between the classes. The best result was obtained with QDA, as in the MIR data set, by employing three wavelengths. Again, only 2 out of 29 test objects were incorrectly classified (two class 3 objects included in class 2).

4.2. Quantitative analysis

The prediction results for kinematic viscosity at 40 °C employing NIR spectra and PLS, PCR, or MLR models were not satisfactory. Nonlinear calibration attempts using neural networks [33–35] were also unsuccessful. It could be argued that the radiation scattering and absorption by particulate matter, or possibly the sample dilution in toluene, may have masked the spectral features related to viscosity.

The prediction results obtained with MIR spectra and PCR/PLS models are shown in Table 4. On the overall, the PLS model predictions were more accurate when compared to PCR. Therefore, the discussion will be henceforth restricted to the PLS results.

A Pareto diagram for the effects calculated from Table 4 (PLS calibration) is presented in Fig. 3a. It is worth noting that the effects of interaction between factors are considerable. In order to obtain a better interpretation of the effects, a cube representation for the factorial design is presented in Fig. 3b.

An inspection of Fig. 3b reveals that a change from the low to the high level of the factors leads to an RMSEP increase in most cases. An exception that should be pointed out is the use of derivative and variable selection, in which case an RMSEP reduction of 4.4 cSt is observed when smoothing is not performed. The best result (smallest RMSEP) is obtained by using low levels for the three factors (4.2 cSt). Such a finding is in line with the conclusions of the classification study, in which the best results for the MIR data were also obtained without pre-processing.

Table 4
Factorial design matrix and PLS/PCR results for the MIR prediction of 40 °C kinematic viscosity

Factors				Levels	
				–	+
Derivative				No	Yes
Smoothing				No	Yes
Variable selection by GA				No	Yes
Trial	Factors			RMSEP (cSt)	
	1	2	3	PLS	PCR
1	–	–	–	4.2 (5)	4.7 (6)
2	+	–	–	5.6 (5)	6.7 (8)
3	–	+	–	4.3 (5)	4.8 (6)
4	+	+	–	4.8 (6)	6.4 (9)
5	–	–	+	4.9 (7)	4.4 (8)
6	+	–	+	8.9 (4)	10.3 (5)
7	–	+	+	5.6 (7)	5.7 (10)
8	+	+	+	4.5 (6)	4.6 (9)

The number of latent variables employed in each model is shown in parenthesis.

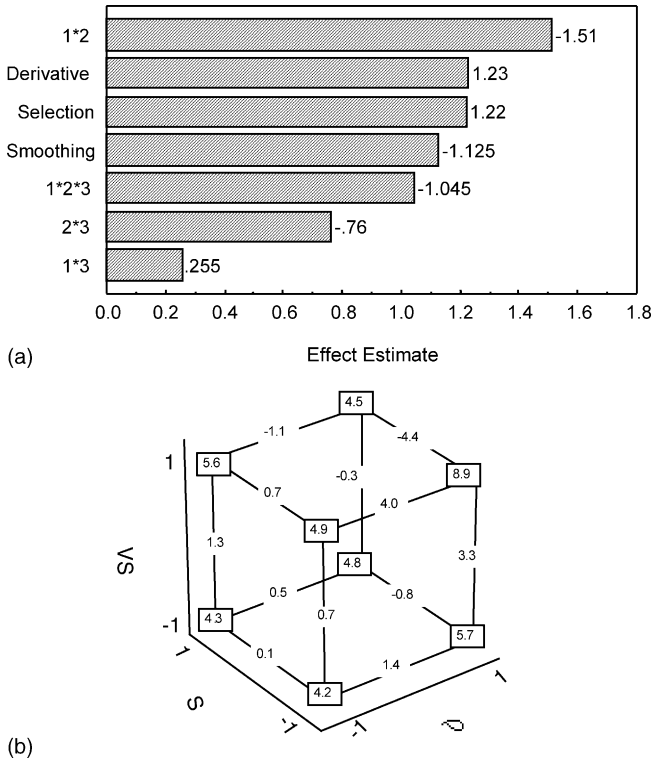

Fig. 3. (a) Pareto effect diagram for PLS results in the 2^3 factorial design: RMSEP values for the MIR prediction of 40 °C kinematic viscosity. (b) Cube representation for the 2^3 factorial design involving derivative (D), smoothing (S), and variable selection (VS) on the PLS results. The effects are expressed in terms of RMSEP values for the MIR prediction of 40 °C kinematic viscosity.

For the best settings of the PLS calibration, the graph of predicted versus observed values for the prediction samples is presented in Fig. 4.

Table 5 presents the factorial design performed for MLR calibration. The only difference from the design in Table 4 consists of the levels for the third factor. In this case, the selection of variables was carried out either by SPA (low level) or by GA (high level). It is worth noting that MLR cannot be directly applied to the full spectrum without variable selection because of ill-conditioning problems.

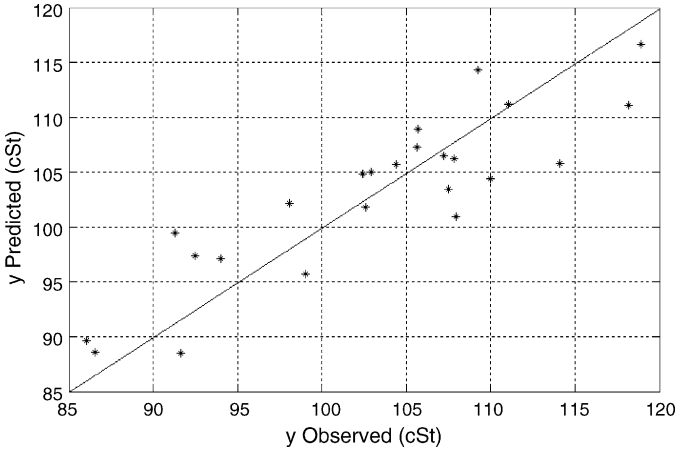

Fig. 4. PLS-MIR (without pre-processing) predictions of 40 °C kinematic viscosity versus reference values. A straight line was drawn to indicate the bisectrice of the quadrant.

Table 5
Factorial design matrix and MLR results for the MIR prediction of 40 °C kinematic viscosity

Factors	Levels	
	–	+
Derivative	No	Yes
Smoothing	No	Yes
Variable selection	SPA	GA

Trial	Factors			RMSEP (cSt)
	1	2	3	
1	–	–	–	5.1 (24)
2	+	–	–	7.6 (7)
3	–	+	–	3.8 (28)
4	+	+	–	5.9 (18)
5	–	–	+	4.9 (25)
6	+	–	+	5.4 (25)
7	–	+	+	5.0 (25)
8	+	+	+	6.1 (24)

The number of wavelengths employed in each model is shown in parenthesis.

A Pareto diagram and a cube representation for the effects calculated from Table 5 (MLR calibration) are presented in Fig. 5a and b, respectively. As in the PLS case, the effects of interaction between factors are substantial.

On average, the use of derivative increases the RMSEP by 1.6 cSt. Such an effect is more prominent when variable selection is performed by SPA (2.3 cSt average increase in RMSEP). It might be argued that the SPA result is more compromised by the use of derivative than the GA result because the SPA policy of selecting variables which are weakly correlated may favour the selection of noisy variables, a problem that is aggravated by the derivative calculation. In fact, the effect of changing the variable selection algorithm from SPA to GA decreases the RMSEP by 2.2 cSt, when the derivative is used without smoothing, which is the situation in which noise is maximally amplified by the pre-processing procedures. In the opposite situation (smoothing employed without the derivative), in which noise is maximally attenuated, the variable selection effect is also the opposite, that is, a 1.1 cSt increase in the RMSEP is observed when SPA is replaced with GA.

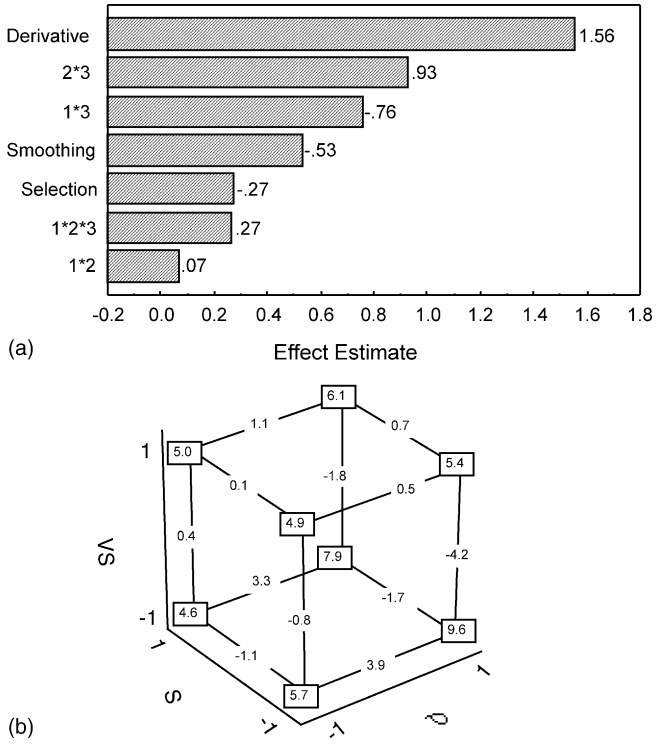

Fig. 5. (a) Pareto effect diagram for MLR results in the 2^3 factorial design: RMSEP values for the MIR prediction of 40 °C kinematic viscosity. (b) Cube representation for the 2^3 factorial design involving derivative (D), smoothing (S), and variable selection (VS) on the MLR results. The effects are expressed in terms of RMSEP values for the MIR prediction of 40 °C kinematic viscosity.

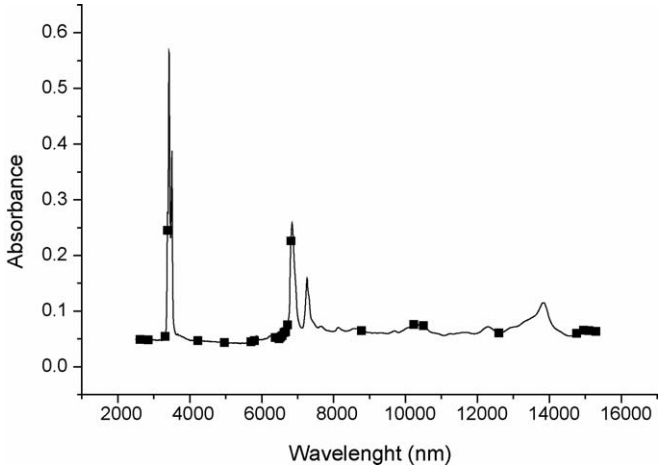

Fig. 6. Wavelengths selected by SPA for the MIR prediction of 40 °C kinematic viscosity.

Fig. 6 indicates the wavelengths that led to the best MLR result (RMSEP of 3.8 cSt), which was obtained without derivative, with smoothing, and with SPA variable selection. Such an outcome is slightly better than the best PLS result (RMSEP of 4.2 cSt), but the difference is not significant according to an *F*-test at 95% confidence level. Neural network models were also employed in an attempt to achieve better predictions. However,

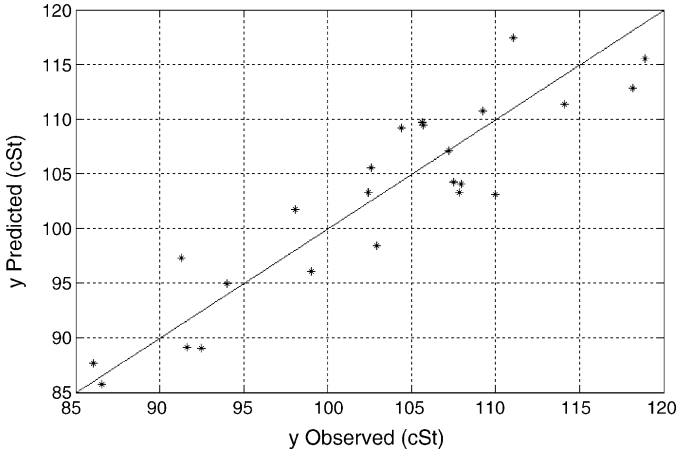

Fig. 7. APS-MLR (with smoothing) predictions of 40 °C kinematic viscosity vs. reference values. A straight line was drawn to indicate the bisectrice of the quadrant.

the results were similar to those yielded by the linear methods under study. For the best settings of the MLR calibration, the graph of predicted versus observed values for the prediction samples is presented in Fig. 7.

As the ASTM D445 norm does not present the reprodutibility for the determination of 40 °C kinematic viscosity for oils in service, the reference method repeatability was estimated in our laboratory. A relative standard deviation of 3.3% was obtained, which is comparable to the relative average errors obtained with the best PLS (3.5%) and MLR (3.2%) models.

5. Conclusions

This work presented two proposals for monitoring the service condition of diesel-engine lubricating oils by using infrared spectroscopy. In the first approach, the oil spectra were classified into three groups according to the stage of use. For this purpose, a variable selection algorithm was proposed to allow the use of simple discriminant analysis models. In this case, a classification accuracy of 93% was obtained both in the MIR and NIR ranges.

The second approach employed multivariate calibration methods to predict viscosity, which is the main control parameter for lubricants in service. In this case, the use of the NIR range was not successful regardless of the modelling method. Such a problem may be ascribed to the experimental methodology employed for spectra acquisition, which required the dilution of the samples because of the presence of particulated matter. This difficulty was circumvented by use of attenuated total reflectance (ATR) measurements in the MIR spectral range, in which an RMSEP of 3.8 cSt and a relative average error of 3.2% were attained. Those values can be considered satisfactory for monitoring the condition of lubricants in service.

The proposed methodologies may lead to substantial gains for companies that operate a large number of diesel engines, by allowing a more efficient condition-based replacement of the lubricating oil.

Acknowledgments

This work was supported by CAPES/PROCAD, FINEP/CTPETRO and CNPq (PRONEX grant and research fellowships). The authors also acknowledge the collaboration of Borborema Imperial Transportes Ltda, which provided the lubricant samples for this study.

Appendix A. Discriminability

In classification problems, the variables can be ranked on the basis of their ability to discriminate the classes under consideration. According to Duda et al. [14], the discriminability D_i of variable x_i can be quantified as:

$$D_i = \frac{S_{Bi}}{S_{Wi}} \tag{A.1}$$

where S_{Wi} and S_{Bi} are measures of the within-class and between-class dispersions for variable x_i, respectively. The within-class dispersion S_{Wi} is defined as

$$S_{Wi} = \sum_{j=1}^{C} s_{ij} \tag{A.2}$$

where s_{ij} is the dispersion of x_i in class j, calculated as

$$s_{ij} = \sum_{k \in I_j} [x_i^k - m_{ij}]^2 \tag{A.3}$$

where x_i^k denotes the value of x_i in the kth object and m_{ij} is the mean value of x_i in class j, that is:

$$m_{ij} = \frac{1}{n_j} \sum_{k \in I_j} x_i^k \tag{A.4}$$

The between-class dispersion S_{Bi} is defined as

$$S_{Bi} = \sum_{j=1}^{C} n_j [m_{ij} - m_i]^2 \tag{A.5}$$

where m_i is the average of x_i over all training objects.

References

[1] Standard Test Method for Flash and Fire Points by Cleveland Open Cup, D 92, ASTM (American Society of Testing Materials), 2001.
[2] Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (the Calculation of Dynamic Viscosity), D 445, ASTM (American Society of Testing Materials), 1994.
[3] C. Pasquini, J. Brazil. Chem. Soc. 14 (2003) 198.
[4] S. Macho, M.S. Larrechi, Trends Anal. Chem. 21 (2002) 799.
[5] M. Blanco, J. Pagès, Anal. Chim. Acta 463 (2002) 295.
[6] F.S.G. Lima, M.A.S. Araújo, L.E.P. Borges, Tribol. Int. 36 (2003) 691.
[7] M.I.S. Sastry, A. Chopra, A.S. Sarpal, S.K. Jain, S.P. Srivastava, A.K. Bhatnagar, Energy Fuels 12 (1998) 304.
[8] F.R. Van De Voort, J. Sedman, V. Yaylayan, Appl. Spectrosc. 58 (2003) 193.
[9] J. Dong, F.R. Van De Voort, V. Yaylayan, A.A. Ismail, D. Pinchuk, A. Brazeau, Lubr. Eng. 45 (2000) 30.
[10] A.D. Stuart, S.M. Trotman, K.J. Doolan, P.M. Fredericks, Appl. Spectrosc. 43 (1989) 55.
[11] A. Borin, R.J. Poppi, Vibr. Spectrosc. 37 (2005) 27.
[12] J. Paschoal, F.D. Barboza, R.J. Poppi, J. Near Infrared Spectrosc. 11 (2003) 211.
[13] A. Borin, Aplicação de quimiometria e espectroscopia no infravermelho no controle de qualidade de lubrificantes, Universidade Estadual de Campinas, Campinas, SP, 2003.
[14] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., John Wiley, New York, 2001.
[15] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Chem. Intell. Lab. Syst. 329 (1996) 257.
[16] M. Kudo, J. Sklansky, Pattern Recogn. 33 (2000) 25.
[17] T. Naes, B.H. Mevik, J. Chem. 15 (2001) 413.
[18] N. Benoudjit, D. François, M. Meurens, M. Verleysen, Chem. Intell. Lab. Syst. 74 (2004) 243.
[19] U. Horchner, J.H. Kalivas, Anal. Chim. Acta 311 (1995) 1.
[20] J.H. Kalivas, N. Roberts, J.M. Sutter, Anal. Chem. 61 (1989) 2024.
[21] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135.
[22] R. Leardi, J. Chem. 15 (2000) 559.
[23] D.L. Massart, D. Jouan-Rimbaud, R. Leardi, O.E. De Noord, Anal. Chem. 67 (1995) 4295.
[24] V. Centner, D.L. Massart, O.E. deNoord, S. Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.

[25] H.C. Goicoechea, A.C. Olivieri, Analyst 124 (1989) 725.
[26] G.A. Bakken, T.P. Houghton, J.H. Kalivas, Chem. Intell. Lab. Syst. 45 (1999) 225.
[27] M. Forina, C. Casolino, C.P. Millan, J. Chem. 13 (1999) 165.
[28] B.K. Alsberg, D.B. Kell, R. Goodacre, Anal. Chem. 70 (1998) 4126.
[29] M.C.U. Araujo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, Chem. Intell. Lab. Syst. 57 (2001) 65.
[30] R.K.H. Galvão, M.F. Pimentel, M.C. Araújo, U. Yoneyama, T.V. Visani, Anal. Chim. Acta 443 (2001) 107.
[31] M.C. Breitkreitz, I.M. Raimundo Jr., J.J.R. Rohwedder, C. Pasquini, H.A. Dantas Filho, G.E. José, M.C.U. Araújo, Analyst 128 (2003) 1204.
[32] J. McClelland, R.W. Jones, Lubr. Eng. 57 (2001) 17.
[33] Z. Ramadan, P.K. Hopke, M.J. Johnson, K.M. Scow, Chem. Intell. Lab. Syst. 75 (2005) 23.
[34] F. Estienne, F. Despagne, B. Walczak, O.E. de Noord, D.L. Massart, Chem. Intell. Lab. Syst. 73 (2004) 207.
[35] C. Ruckebusch, L. Duponchel, J.P. Huvenne, Chem. Intell. Lab. Syst. 62 (2002) 189–198.