FULL PAPER
THE ANALYST
www.rsc.org/analyst

Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration

Márcia C. Breitkreitz,[a] Ivo M. Raimundo, Jr,*[a] Jarbas J. R. Rohwedder,[a] Celio Pasquini,[a] Heronides A. Dantas Filho,[b] Gledson E. José[b] and Mário C. U. Araújo[b]

[a] *Institute of Chemistry, UNICAMP, CP 6154, CEP 13084-971, Campinas, Brazil*
[b] *Department of Chemistry, UFPB, João Pessoa, Brazil*

Received 12th May 2003, Accepted 2nd July 2003
First published as an Advance Article on the web 18th July 2003

A method for sulfur determination in diesel fuel employing near infrared spectroscopy, variable selection and multivariate calibration is described. The performances of principal component regression (PCR) and partial least square (PLS) chemometric methods were compared with those shown by multiple linear regression (MLR), performed after variable selection based on the genetic algorithm (GA) or the successive projection algorithm (SPA). Ninety seven diesel samples were divided into three sets (41 for calibration, 30 for internal validation and 26 for external validation), each of them covering the full range of sulfur concentrations (from 0.07 to 0.33% w/w). Transflectance measurements were performed from 850 to 1800 nm. Although principal component analysis identified the presence of three groups, PLS, PCR and MLR provided models whose predicting capabilities were independent of the diesel type. Calibration with PLS and PCR employing all the 454 wavelengths provided root mean square errors of prediction (RMSEP) of 0.036% and 0.043% for the validation set, respectively. The use of GA and SPA for variable selection provided calibration models based on 19 and 9 wavelengths, with a RMSEP of 0.031% (PLS-GA), 0.022% (MLR-SPA) and 0.034% (MLR-GA). As the ASTM 4294 method allows a reproducibility of 0.05%, it can be concluded that a method based on NIR spectroscopy and multivariate calibration can be employed for the determination of sulfur in diesel fuels. Furthermore, the selection of variables can provide more robust calibration models and SPA provided more parsimonious models than GA.

Introduction

The sulfur content in fuels depends on the origin of the petroleum, the cracking process used and the fuel pre-treatment. Diesel fuel usually has higher concentrations of sulfur than gasoline, this element being found as mercaptans, sulfides, disulfides and heterocyclic compounds. These compounds, due to the combustion of the fuel in the engine, can be converted into SO_2, which is a very polluting gas and produces acidic rain. Besides being harmful to the environment, sulfur dioxide can cause a pronounced wearing of the engine as, at low temperatures, humidity can be condensed inside the engine, producing sulfurous and sulfuric acids after reacting with sulfur dioxide and trioxide, respectively.

The commercialisation of fuels in Brazil is regulated by the National Petroleum Agency, which classifies automotive diesel fuels in two categories, type B (countryside diesel) and type D (metropolitan diesel), whose maximum sulfur contents are 0.35% (w/w) and 0.20% (w/w), respectively.

The determination of sulfur compounds in diesel fuel has been carried out by the lamp method,[1] coulometry,[2] X-ray fluorescence,[3,4] gas chromatography[5,6] and chemiluminescence.[7,8] Nowadays, optical methods for fuel analysis have increasingly been used, as they allow direct determinations without sample pre-treatment, are non-destructive and provide high sample throughput. Near infrared (NIR) is the region of the electromagnetic spectra which comprises radiation from 780 to 2500 nm (4000 to 12,820 cm^{-1}). The energies associated with this spectral range correspond mainly to vibrational transitions due superior harmonics (overtones) and combination bands. NIR spectroscopy has been employed for the determination of physical properties and chemical compositions of several petroleum derivatives. Thus, several contributions can be found in the literature regarding applications of NIR spectroscopy for simultaneous determination of saturated hydrocarbons, aromatics and olefins in gasoline;[9] for determination of octane number and vapour pressure,[10,11] benzene, toluene, ethylbenzene and xylene (BTEX) in fuels[12] and for characterisation of crude oil.[13]

Despite the increasing use of NIR spectroscopy, the spectrum in this region cannot be interpreted in a straightforward way as with the mid infrared region, because the sharp peaks found in the mid region are almost strictly related to fundamental vibrational transitions, while, in the near region, wide bands occur, as a result of overtones and combination bands of fundamental vibrations.[14] Due to this reason, NIR spectral data are frequently treated by means of chemometric methods, such as partial least square regression (PLS), principal component regression (PCR), multiple linear regression (MLR) and principal components analysis (PCA). Multivariate calibration combines values of analytical measurements with mathematical algorithms in order to obtain models well adjusted to the experimental data. The mathematical model obtained must be robust, that is, it should be able to perform predictions (*e.g.*, concentration of a unknown sample) with acceptable precision and accuracy during long term use under variable conditions regarding the instrument environment and under some variable sample physical-chemical characteristics. Several strategies have been employed to improve the robustness of multivariate calibration models, such as pre-treatment of data and variable selection. While the pre-treatment of data (for example, smoothing and use of first derivative) is a procedure that has been practically integrated to ordinary multivariate calibration methods, the selection of variables is still a choice, aimed at excluding colinearity, redundancies and noise from the whole set of data. Among the various strategies, the genetic algorithm (GA) has frequently been employed,[15–17] providing more robust multivariate calibration models. The GA is a non-deterministic algorithm based on the Darwin natural selection theory, which states that individuals more adapted to the environment have

 DOI: 10.1039/b305265f

higher probability of surviving and reproducing. Therefore, the algorithm works by adjusting the number of individuals (chromosomes) of the initial population, number of generations, probability of crossover and mutation. Each chromosome is composed of a set of genes, which are binary coded spectral variables. As the algorithm run, the variables are set as "one" or "zero" (select or not select) and the chromosomes that produce models (PLS or MLR) with lower values of RMSEP (root mean square error of prediction) for each generation are chosen, as those represent the more adapted individuals. New generations are produced by combining the more adapted individuals of the previous generation. Crossover and mutation are introduced in each new generation, which help to overcome local optimisation. Despite its simplicity and efficiency, due to its stochastic nature, the selection of variables by GA may not be reproducible.

Recently, the successive projections algorithm (SPA) has been proposed for variable selection in multivariate calibration.[18] SPA is a forward selection method, which employs vector projections (column vectors constituted by absorbance of the samples at each wavelength) for selection of the wavelengths that produce the lowest RMSEP in the prediction of the parameter of interest through a MLR model. In this way, in a space of *n*-dimensions (where *n* is the number of original variables), a start vector is randomly chosen. Subsequently, in a orthogonal sub-space, the vector of higher projection is selected, becoming the new starting vector. Therefore, one wavelength is incorporated at each iteration until a pre-set number is reached. The choice of the orthogonal sub-space at each iteration is made in order to select only the non-collinear variables. Afterwards, each set of selected variables is employed by the algorithm to construct models based on MLR, aimed at finding the set which provides the lowest RMSEP. Finally, the wavelengths selected by the algorithm are employed to construct the definitive calibration model, based on MLR, which is applied to predict the parameter of interest for an external set of samples employed for final model validation. Due to its deterministic characteristic, SPA provides more reproducible results, performing the selection in a time interval usually shorter than with GA.

In this work, a method for determination of total sulfur in diesel fuel employing NIR spectroscopy, variable selection and multivariate calibration is proposed. The performances of principal component regression (PCR) and partial least square (PLS) chemometric methods were compared with those obtained by multiple linear regression (MLR), carried out after variable selection based on the genetic algorithm (GA) or the successive projection algorithm (SPA).

Experimental

Calibration and validation samples

Ninety seven diesel samples (64 type B, 24 type D and 9 with addition of cetane improver) were divided arbitrarily into three sets (41 for calibration, 30 for internal validation and 26 for external validation), each of them covering the full range of sulfur concentrations (from 0.07% to 0.33% w/w). The concentrations of sulfur in these samples were determined by energy-dispersive X-ray fluorescence, employing a Spectro Titan spectrophotometer (current of 400 μA, tube voltage of 5,5 kV and irradiation time of 300 s), according to the ASTM 4294 standard method.[19]

Instrumentation and procedure

A Luminar 2000 Brimrose NIR spectrophotometer, equipped with a 4 mm optical path transflectance probe, was employed for acquisition of spectral data from 850 to 1800 nm, in steps of 2 nm. For measurements, *ca.* 3 mL of diesel sample were put into a test tube and the probe was introduced, avoiding air bubbles in the optical path. Spectra were obtained as the average of 300 runs. Before starting measurements, a spectrum of the probe in air was always obtained and employed for background correction.

Data pre-treatment and calculations

The first derivatives of the transflectance spectra were always employed in the calculations. Multivariate calibration (PLS, PCR and MLR) as well as principal component analysis were performed with Unscrambler 7.5 software (CAMO). Variable selection was performed with Matlab 6.1 software, based on the GA and SPA mathematical algorithms.

Results and discussion

Preliminary studies showed that background correction is necessary in order to overcome some drawbacks that arise from the use of the probe, such as irreversible adsorption of some interfering species on the surfaces of the optical fibre tip or the reflective mirror. These interfering compounds cause almost imperceptible changes in the spectra, impairing the prediction capability of a model for long term use if the probe is continuously in use. On the other hand, when the background correction is performed, this problem is eliminated, leading to more reliable models. Therefore, before a series of measurements, a spectrum with the probe in air was always run for background correction and the sample spectrum was obtained as the ratio of the transflectance spectrum to the background and, as a consequence, an absorbance-like spectrum was recorded.

Despite the background correction, the spectra obtained were not immune to baseline shifts, as shown in Fig. 1. These shifts were corrected by applying a first derivative to the spectra, which was further smoothed by a 2nd order Savitzky-Golay polynomial procedure. The resulting spectra, which were employed in the multivariate calibration, are shown in Fig. 2.

Initially, an exploratory analysis of the sample set was carried out, employing principal component analysis (PCA). Fig. 3 shows the graphic of scores of the first two principal components (PC) obtained for the whole sample set. As can be noted, there are three distinct groups, one composed by diesel with cetane improver (central group) and another two, which have both type B and type D diesels. The occurrence of two distinct groups containing the two types of diesel fuel cannot be

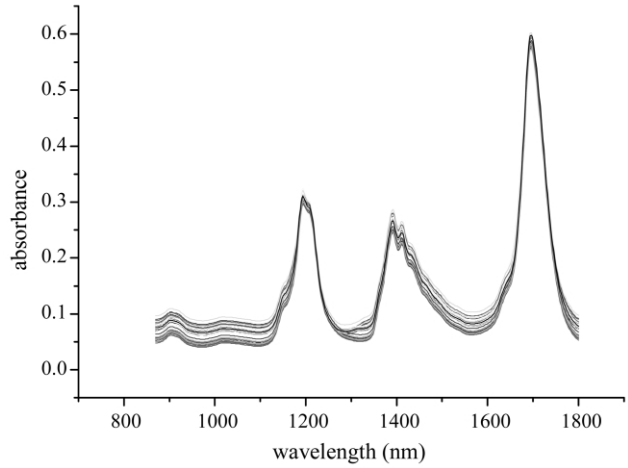

Fig. 1 Spectra of 97 diesel samples obtained with background correction.

straightforwardly explained although it is probably a consequence of the origin of the petroleum and/or of the cracking process used. However, it is worth noting that diesel samples with cetane improver belong to a particular group, which can be easily distinguished from the other two, indicating that PCA can be applied to verify the authenticity of a diesel with cetane improver. The graphic of loadings as a function of the first PC (which allows distinction of the three groups) indicated that variables regarding the wavelengths from 1170 to 1240 nm (second overtone of C–H bond) and from 1642 to 1800 nm (first overtone of C–H and S–H bonds) are the most important in the identification of the groups.

Although PCA has identified three classes of diesel fuel, previous calculation employing PLS demonstrated that there are no significant differences among the prediction capabilities of the models constructed with sample sets containing or not diesel with cetane improver. Therefore, all the results shown below were obtained without any distinctions among the types of diesel samples.

As described in the Experimental section, the 97 diesel samples were divided in three groups (calibration, 41; internal validation, 30; and external validation, 26), covering the full range of sulfur concentration in the samples (from 0.07% to 0.33% w/w). The external validation set was constituted by diesel samples which were not employed for the construction of the chemometric model, allowing its true assessment. In this sense, the performances of the models constructed employing different strategies were compared based on the RMSEP (root mean square errors of prediction) for the external validation set.

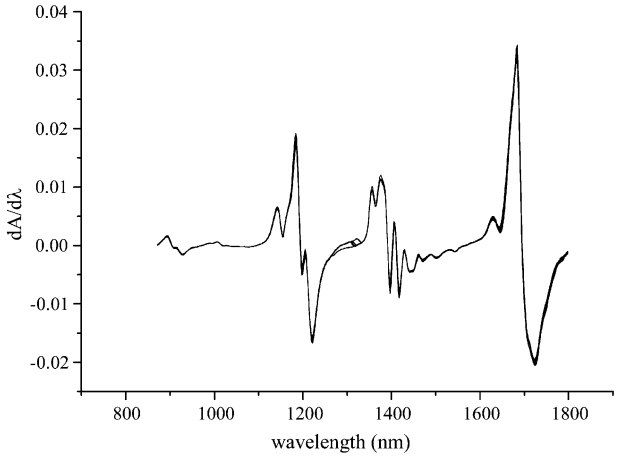

Fig. 2 First derivative spectra of 97 diesel samples (Fig. 1) after smoothing with the Savitzky-Golay procedure.

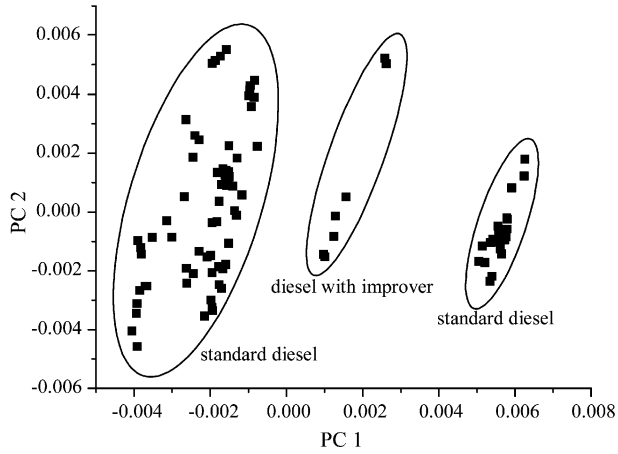

Fig. 3 Graphic of the scores obtained by PCA for standard (types B, left, and D, right) and with improver diesel fuel samples.

Initially, two calibration models were constructed with PLS and PCR algorithms, employing the signals referred to all the 454 wavelengths of the spectra, without any variable selection. It was not possible to apply MLR in this case, as the number of variables was higher than the number of samples. After that, variable selection procedures were performed employing SPA or GA. In both algorithms, the best number of wavelengths was searched considering a minimum of 1 and a maximum of 20, and the RMSEP of the internal validation set was used as a criterion for the choice. In the GA procedure, a population of 250 individuals, a crossover of 0.60, mutation of 0.10 and 100 generations were also used as criteria for variable selection. Due to its stochastic nature, software was run 10 times and the selection of wavelengths that provided the lowest value of RMSEP for the internal validation set was employed in the further calculations.

Table 1 summarises the results obtained in these calibrations that were based on different strategies. As can be seen, the best results can be obtained after selection of variables by either GA or SPA. The use of SPA for variable selection reduced the number of wavelengths from 454 to 9, while for GA the number was reduced to 19. The results shown in Table 1 demonstrate that the prediction capabilities of these models are better, as the RMSEP values are lower, and a comparison between the values found by the model and the reference values shows a lower bias and better correlation coefficients than those obtained for calibrations without variable selection. Although the RMSEP and regression coefficients values provided by SPA and GA seem to be contradictory, it is worth mentioning that differences are not significant considering the deviations of the calibration models. Therefore, it can be concluded that GA and SPA produce more robust calibration models.

The wavelengths chosen by GA and SPA to perform multivariate calibrations are specified in Table 2. Considering the first derivative spectra shown in Fig. 2, although the absorbance intensities at 878, 910 and 1028 nm show slight changes, these wavelengths have been chosen by GA. In

Table 1 Results obtained in the determination of sulfur in diesel fuel, employing different strategies

Model	NW[a]	PC[b]	RMSEP (%)[c]	Intercept/slope[d]	R[e]
PLS	454	6	0.036	0.043/0.789	0.877
PCR	454	10	0.043	0.051/0.754	0.805
PLS/GA	19	2	0.031	−0.008/1,020	0.912
MLR/GA	19	—	0.034	0.007/0.972	0.986
MLR/SPA	9	—	0.022	0.003/0.895	0.946

[a] Number of wavelengths. [b] Number of principal components determined by software. [c] % of sulfur, w/w, absolute value. [d] Intercept and slope of the expected *vs.* predicted values curve. [e] Correlation coefficient of the expected values *vs.* predicted values curve.

Table 2 Wavelengths (in nm) selected by GA and SPA for multivariate calibrations

GA	878	910	1028	1146	1332	1384	1424	1454	1476	1490
	1546	1552	1614	1632	1672	1742	1758	1760	1784	—
SPA	1178	1184	1194	1222	1402	1472	1654	1672	1686	—

addition, 1742, 1758 and 1760 nm, wavelengths that are related to S–H vibrations, have been also chosen by GA. The other wavelengths chosen by both algorithms can be ascribed to C–H overtones and combination bands. Although both algorithms are aimed at minimising the RMSEP value, they select different wavelengths. This difference arises from the fact that GA can select wavelengths that does not carry any chemical or physical information although they decrease the RMSEP, while SPA considers the vectors that contain more information. In addition, different solutions are allowed to solve a multivariate problem, such as the sulfur determination in diesel fuel treated in this work. Finally, as the number of wavelengths chosen by SPA is lower than those indicated by GA, the former algorithm is more parsimonious than the latter, a result similar to a previous study.[19]

Conclusions

The results obtained in this work indicate that near infrared spectroscopy can be employed to determine sulfur in diesel fuel samples. Besides being non-destructive, non-pollutant, simple and fast, NIR spectroscopy, in conjunction with multivariate calibration methods, such as PLS, PCR and MRL, can allow the simultaneous determination of several parameters of quality of diesel and other fuels, which demonstrate its advantage over X-ray diffraction spectroscopy, frequently employed by reference methods. Considering that the ASTM 4294–90[3] method for determination of sulfur in diesel fuel accepts a reproducibility of 0.05% (w/w, absolute value), all the calibration models obtained in this work fulfil this requirement, as the worst RMSEP value was 0.043%. Furthermore, both AG and SPA can be employed for selection of variables, leading to more robust models. In the present work, SPA spent a longer time than GA for variable selection, as a consequence of the size of the data set. However, SPA provides more parsimonious models than GA, as it is able to select a lower number of variables for the construction of the models.

Acknowledgements

Authors are grateful to CTPETRO/FINEP and PROCAD/CAPES for financial support. MCB, HADF and GEJ thank PIBIC/CNPq and CAPES for the fellowships. Professor C. H. Collins is kindly acknowledged for manuscript revision.

References

1 American Society for Testing and Materials (ASTM), D1266–91.
2 American Society for Testing and Materials (ASTM), D3120–92.
3 American Society for Testing and Materials (ASTM), D4294–90.
4 R. A. Jones, *Anal. Chem.*, 1961, **33**, 71.
5 T. G. Albro, P. A. Dreifuss and J. Wormsbecher, *J. High Resolut. Chromatogr.*, 1993, **16**, 13.
6 D. A. Clay, C. H. Rogers and R. H. Jungers, *Anal. Chem.*, 1977, **49**, 126.
7 M. J. Navas and A. M. Jimenez, *Crit. Rev. Anal. Chem.*, 2000, **30**, 153.
8 F. P. Di Sanzo, W. Bray and B. Chawla, *J. High Resolut. Chromatogr.*, 1994, **4**, 225.
9 J. J. Kelly and J. B. Callis, *Anal. Chem.*, 1990, **62**, 1444.
10 I. Litano-Barzilai, I. Sela, V. Bulatov and I. Zimmerman, *Anal. Chim. Acta*, 1997, **339**, 193.
11 G. Bohacs, Z. Ovadi and A. Salgo, *J. Near Infrared Spectrosc.*, 1998, **6**, 341.
12 J. B. Cooper, K. L. Wise, W. T. Welch, M. B. Summer, B. K. Wilt and R. R. Bledsoe, *Appl. Spectrosc.*, 1997, **51**, 1613.
13 K. Hidajat and S. M. Chong, *J. Near Infrared Spectrosc.*, 2000, **8**, 53.
14 L. Bokobza, *J. Near Infrared Spectrosc.*, 1998, **6**, 1998.
15 R. Leardi, M. B. Seasholtz and R. J. Pell, *Anal. Chim. Acta*, 2002, **461**, 189.
16 J. M. Roger and V. Bellon-Maurel, *Appl. Spectrosc.*, 2000, **54**, 1313.
17 A. Herrero and M. C. Ortiz, *Anal. Chim. Acta*, 1999, **378**, 245.
18 M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame and V. Visani, *Chemom. Intell. Lab. Syst.*, 2001, **57**, 65.
19 R. K. H. Galvão, M. F. Pimentel, M. C. U. Araújo, T. Yoneyama and V. Visani, *Anal. Chim. Acta*, 2001, **443**, 107.