Talanta 104 (2013) 128–134

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Short communication

Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear discriminant analysis

Mohammadreza Khanmohammadi [a,*], Amir Bagheri Garmarudi [a,b], Miguel de la Guardia [c]

[a] Chemistry Department, Faculty of Science, IKIU, Qazvin, Iran
[b] Department of Chemistry & Polymer Laboratories, Engineering Research Institute, Tehran, Iran
[c] Department of Analytical Chemistry, University of Valencia, 50 Dr. Moliner Street, E-46100 Burjassot, Valencia, Spain

ARTICLE INFO

Article history:
Received 14 August 2012
Received in revised form
11 November 2012
Accepted 12 November 2012
Available online 20 November 2012

Keywords:
Diesel
ATR-FTIR
Chemometrics
Wavelength selection
Classification

ABSTRACT

A rapid approach has been developed for the characterization of diesel quality, based on attenuated total reflectance – Fourier transform infrared (ATR-FTIR) spectrometry, which could be useful for diagnosing the sample quality condition. As a supervised technique, linear discriminant analysis (LDA) was employed to process the spectrometric data. The role of variable selection methods was also evaluated. Successive projection algorithm (SPA) and genetic algorithm (GA) feature selection techniques were applied prior to the discriminative procedure. It was aimed to compare the effect of feature selection procedures on classification capability of IR spectrometry for the diesel samples according to their quality passed or quality failed situation. Predictive capability of LDA was compared with that obtained by GA-LDA and SPA-LDA. Results showed 91.1%, 93.3% and 95.6% of accuracy for LDA, GA-LDA and SPA-LDA respectively. Thus SPA-LDA together with ATR-FTIR spectrometry was proposed as a fast screening analytical test for the evaluation of quality passed/failed situation in diesel samples.

1. Introduction

Diesel (petrodiesel) is a specific fraction of petroleum distillation process, which is consumed as fuel in diesel engines. It is a formulated mixture of different chemicals, such as linear chain hydrocarbons, naphthenic and aromatic compounds. Quality of diesel fuels is affected by several parameters while the most important factor is the relative proportion of each type of hydrocarbon in the formulation. Consistency in the quality of diesel is of high importance for commercial aims, customer rights, business competition and environmental risks. This concern would confirm the need of a careful quality control in the refining and fuel manufacturing industrial systems. Accordingly, each refinery industrial plant maintains a strategic department of quality control (QC), equipped with appropriate instruments and experimental expert human resources which would monitor the quality of fuel products in different stages of refining-manufacturing process. All the QC tests must be performed based on standard test procedures, which have been validated in the petroleum industry previously. These methods are utilized to evaluate the physico-chemical characteristics of fuel samples. Environmental, economic and industrial reasons encouraged the specialists for the development of rapid, precise and reliable experimental monitoring approaches to evaluate the fuel quality inside the refineries. There are several reports, dealing with analytical proposals for quality screening of diesel and biodiesel products [1]. Chromatographic methods, i.e. thin layer (TLC) [2], high performance liquid (HPLC) [3], gas (GC) [4] and gel permeation (GPC) [5] are the most common analysis approaches. However, most of these well-known methods are destructive, tedious and labor, consuming huge amount of chemical reagents to be performed. On the other hand, IR spectroscopy based methodologies are getting consolidated as rapid, reliable and non-destructive ones, which are free of sample preparation steps [6], thus offering green alternatives for the determination of sample parameters and characteristics without any previous sample pretreatment [7].

Research efforts concerning the application of IR spectrometry in diesel analysis involve the quantitative determination of some parameters in diesel samples [8,9] and pattern recognition strategies to assess them. In case of pattern recognition of diesel samples by IR spectrometry, reports are mostly related to recognizing the adulteration in diesel or its blend with biodiesel [10,11]. Most of these reports are based on chemometric processing of the spectrometric data and there are many efforts to improve the mathematical treatment of data in order to obtain as accurate as possible information while the analyst is enabled to

* Correspondence to: Chemistry Department, Faculty of Science, IKIU, P.O. Box: 34149-1-6818, Qazvin, Iran. Tel./Fax: +98 281 3780040.
E-mail address: mrkhanmohammadi@gmail.com (M. Khanmohammadi).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.11.032

increase the quality of results. An approach for this aim is to improve the previous methods by variable selection techniques and thus, in this work, we have evaluated the improvement in the capability of a supervised pattern recognition technique (LDA) in discrimination between quality passed and quality failed diesel samples from their ATR-FTIR spectra by incorporating two different variable selection methods, successive projection algorithm (SPA) and genetic algorithm (GA) in order to make a clear comparison between the outputs.

2. Experimental

2.1. Samples, apparatus and software

Commercial diesel samples (a total number of 90 samples) were obtained from the quality control laboratory of Shahid Tondguyan Oil Refining Company (Tehran, Iran) collected during 3 months of manufacturing. All the samples were kept refrigerated prior to IR spectrum recording, avoiding the probable loss of volatile compounds. Using a Tensor-27 Bruker FT-IR spectrometer, mid-IR spectra were obtained in triplicate for each sample and the average of spectra was used to be processed. In all cases, spectra were recorded at room temperature (22 ± 2 °C) using a horizontal, fixed path ATR cell (ZnSe, 45°, and single reflection by Pike®), a Ge–KBr beam splitter, a DTGS detector and Beer–Norton apodization. The spectral resolution was fixed at 8 cm^{-1} and 64 scans were accumulated over the range from 600 to 4000 cm^{-1} for each spectrum. Chemometric data processing was performed by MATLAB Ver. 8.0

2.2. Reference quality assessment of diesel by standard experimental procedures

The standard quality situation of diesel samples was established from the results obtained by ASTM reference test methods. The main monitored parameters were specific gravity at 60 F [12], cetan number [13], initial and final boiling temperatures [14], 10–90% distillation temperatures (within 10% intervals) [15], kinematic viscosity [16], cloud point [17], pour point [18], water and sediment content [19,20] and color [21]. All the aforementioned tests were made on each sample by triplicate and the average was considered as the analysis output. Considering the quality limits of the refinery, after performing the standard test experiments, the QC decision was made according to the defined values. Obtained samples were separated into two different groups: QC passed (60 samples) and QC failed (30 samples).

3. Chemometric methods employed

3.1. Linear discriminant analysis (LDA)

LDA is a powerful classification method, widely employed in analytical studies [22]. As a pattern recognition technique, LDA maximizes the ratio of between-class variance to the within-class variance in any particular data set in order to obtain a maximum discrimination. It does not make any change in the location of objects but only tries to obtain clear class discrimination, making a decision between the given classes. This would also help for understanding the distribution of the featured data. LDA is capable of handling the classification cases where the within-class frequencies are unequal and the performances have been examined on randomly generated test data. Investigating an unknown sample by LDA, a critical parameter is prior probability of class k (π_k), which is usually estimated simply by empirical frequencies of the training set, being evaluated through the use of Eq. (1)

$$\sum_{k=1}^{K}\pi_k = 1k \tag{1}$$

being π defined in (2)

$$\hat{\pi} = \frac{\text{number of samples in class } k}{\text{total number of samples}} \tag{2}$$

The class-conditional density of a sample of X in the class is defined in (3)

$$G = k \quad \text{is} \quad f_k(x) \tag{3}$$

and posterior probability of this classification can be evaluated as (4)

$$\Pr\left(G = k \mid X = x\right) = \frac{f_k(x)\,\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \tag{4}$$

3.2. Successive projection algorithm

SPA is a powerful algorithm for variable selection by minimizing the redundant information content of obtained signals and resolving the problems caused by collinearity [23–25]. The prediction capability of SPA coupled to a classification model is very critical.

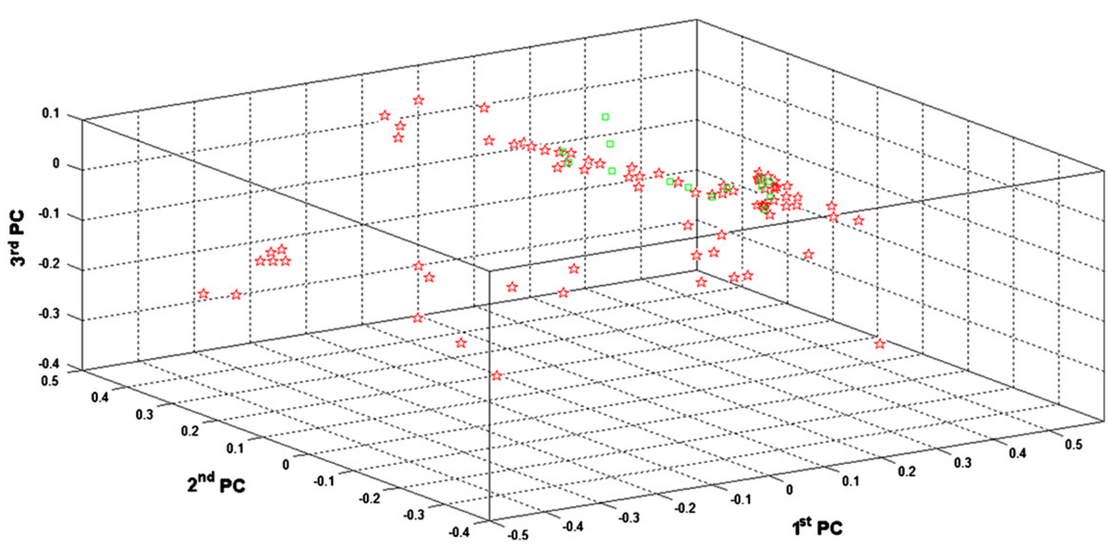

Fig. 1. PCA score pot obtained for quality based analysis of ATR-FTIR spectra of diesel samples.