Fuel 116 (2014) 151–157

Contents lists available at ScienceDirect

Fuel

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

Classification of gasoline as with or without dispersant and detergent additives using infrared spectroscopy and multivariate classification

Michelle Patrícia Ferreira da Silva [a], Lívia Rodrigues e Brito [a], Fernanda Araújo Honorato [b], Ana Paula Silveira Paim [a], Celio Pasquini [c], Maria Fernanda Pimentel [b,*]

[a] Universidade Federal de Pernambuco, Departamento de Química Fundamental, Recife, PE, Brazil
[b] Universidade Federal de Pernambuco, Departamento de Engenharia Química, Recife, PE, Brazil
[c] Universidade Estadual de Campinas, Instituto de Química, Campinas, SP, Brazil

HIGHLIGHTS

- Method based on IR spectra and classification methods to detect additives in gasoline.
- Three classification methods (LDA, PLS-DA, SVM) were compared.
- Three algorithms for selection of variables for LDA were evaluated – SW, GA, SPA.
- The best results were obtained for the MIR region using the LDA/SPA and LDA/GA.
- 100% correct classification was achieved for the test set.

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form 25 July 2013
Accepted 26 July 2013
Available online 13 August 2013

Keywords:
Gasoline
Detergents and dispersants additives
Infrared spectroscopy
Classification

ABSTRACT

Gasoline may contain additives which can minimize the amount of pollutants emitted to the atmosphere. Detergents and dispersants added to gasoline can reduce gas emissions towards atmosphere and the formation of deposits in engines. The Brazilian Agency of Petroleum, Natural Gas and Biofuel (ANP) has established that Brazilian commercial gasoline must contain detergent and dispersant additives, thus requiring the development of methods for their identification in commercial gasoline. This work proposes a methodology which uses infrared spectra in the medium and near region (MIR and NIR) of the residue of distillation for classification of gasoline samples into two groups: with or without detergent/dispersant additives. The performances of three types of classification methods were compared: linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA) and Support Vector Machines (SVM). Different algorithms for selection of spectral variables for LDA were evaluated: stepwise (SW), genetic algorithm (GA) and successive projections algorithm (SPA). The best results were obtained using LDA/GA or SPA/LDA for MIR region.

1. Introduction

Gasoline is a petroleum-derived fuel which is a complex mixture of several compounds, the majority of which are classified within the group of hydrocarbons. In Brazil, the gasoline produced at refineries or petrochemical plants is called "Type A gasoline". However, to be marketed, Brazilian laws establish that anhydrous ethanol fuel must be added to obtain "Type C gasoline" or "Common gasoline" [1]. The ethanol content in gasoline differs in various countries, in general varying from 5% to 10% (v/v) [2]. In Brazil, this content may vary from 18% (v/v) to 25% (v/v), depending on the availability and price of the ethanol [3].

Besides anhydrous ethanol, gasoline may contain additives, which are compounds added to improve fuel performance, avoid motor problems and minimize the amount of pollutants emitted to the atmosphere. There are several types of additives for gasoline, such as detergents, dispersants, antiknock agents and antioxidants. Dispersant additives prevent fuel stratification during storage at low temperatures [4]. Detergent additives reduce gas emissions to the atmosphere as well the formation of deposits in engines [4]. The chemical composition of detergent and dispersant additives is not exactly known due to the fact that it is kept secret by the manufacturing industries. However, it is well-known that these additives must exhibit good thermal stability up to 300 °C, which is the observed highest temperature of an internal combustion engine [5]. Although the literature presents several works reporting

* Corresponding author. Address: Universidade Federal de Pernambuco, Departamento de Engenharia Química, Laboratório de Combustíveis (LAC), CEP 50740-521 Recife, PE, Brazil. Tel./fax: +55 81 2126 7235.
E-mail addresses: mfernanda.pimentel@gmail.com, mfp@ufpe.br (M.F. Pimentel).

0016-2361/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.fuel.2013.07.110

on the benefits of oxygenated additives and on they determination, no work could be found concerning detergent and dispersant additives [6–10].

Quality control of fuels has become increasingly rigorous, not only because of economic and social issues, but also due environmental concerning. The Brazilian Agency of Petroleum, Natural Gas and Biofuel (ANP) has established that after January 1st of 2014 all Brazilian commercial gasoline should contain detergent and dispersant additives [11]. The concentration depends on the formulation of the additives and its effectiveness. Currently, additives may be mixed with Type C gasoline at concentrations ranging from 100 to 5000 mg kg^{-1} and the gasoline is labeled as "Additive Type C gasoline". The compulsory use of detergent and dispersant requires the development of analytical methods capable of identifying the presence of these additives in gasoline samples, in order to verify whether commercial gasoline accomplishes with the norms established by ANP [11].

The literature presents many examples of application of near infrared (NIR) and middle infrared (MIR) spectroscopy associated with multivariate analysis for quality control of fuels [12–16]. Infrared (IR) spectroscopic methods show several advantages, such as the use of relatively low cost equipment that allows field analysis; minimum or no sample treatment; expeditious analysis; causing no sample destruction, and demanding no reagents.

Perhaps, the main drawback of IR based analytical methods is in its relatively low detectability. As dispersant and detergent additives are typically added to Brazilian gasoline in concentrations ranging from 200 to 500 mg kg^{-1} it is not possible to determine these species directly in gasoline using IR spectroscopy [17]. To concentrate the additives up to a concentration level capable to be detected by IR spectroscopy, gasoline samples have been submitted to distillation, following the ASTM D86-11b standard procedure [18], before analysis. As dispersant and detergent additives presents low volatility, they are concentrated in the distillation final residue by a factor of about 100 or more, allowing the use of IR spectroscopy, associated to supervised pattern recognition methods, to identify their presence in gasoline samples

Pattern recognition methods such as principal components analysis (PCA), linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLS-DA) are extensively used in classification problems [19]. PCA is an unsupervised pattern recognition method used to detect similarities among samples [19]. Different from this, LDA and PLS-DA are supervised pattern recognition methods and require a training set of known groups [18,19]. These last two methods are similar because the samples may only be classified in one of the classes.

PLS-DA is based on the PLS algorithm for calibration, but instead of using concentration information, it uses class labels as dependent y vector [12,19]. In PLS-DA, for example, a value close to one indicates that the sample belongs to the class under consideration and a value close to zero indicates that it does not. The threshold used in PLS-DA may vary, but for a two class problem it is common to use a threshold of 0.5.

LDA employs the Mahalanobis distance and seeks a linear combination function that maximizes the between-class variance relative to the within-class variance [13,20]. In LDA the number of training samples must be larger than the number of variables to be included in the LDA model [12]. When using spectral data, in which there are a great number of variables, LDA requires the use of variable selection algorithms, such as: stepwise algorithm (SW), successive projections algorithm (SPA) and genetic algorithm (GA).

The SW algorithm evaluates the individual value of each spectral variable according to its discriminability with respect to the classes under consideration [21]. The variable with the largest discriminability is selected and a leave-one-out cross-validation procedure is performed using LDA [21]. To avoid collinearity problems, the remaining variables that are highly correlated with those already selected are discarded [21]. For the decision about which variables need to be discarded, a threshold value for the coefficient of multiple correlation must be defined by the operator. The algorithm repeats this procedure until no more variables are available for selection. The set of variables that resulted in the smallest number of cross-validation errors is then adopted.

The goal of SPA is to select variables with minimum multicollinearity and maximum information [13]. In SPA the validation set is used to choose the best subset of variables by minimizing the cost function (Eq. (1)) defined as the average risk of misclassification by LDA [22].

$$G = \frac{1}{kV}\sum_{k=1}^{kV} g_k \tag{1}$$

Where g_k (risk of misclassification of the kth validation object x_k) is defined as

$$g_k = \frac{r^2(x_k, \mu_{1k})}{\min_{1j\neq 1k} r^2(x_k, \mu_{1j})} \tag{2}$$

The numerator in Eq. (2) is the squared Mahalanobis distance between object x_k (of class index Ik) and the sample mean μ_{Ik} of its true class (both row vectors). The denominator in Eq. (2) is the squared Mahalanobis distance between object x_k and the center of the closest incorrect class. The object x_k should be close to the center of its true class and distant from the centers of all other classes [22].

The GA is a stochastic algorithm invented to imitate the evolutionary process of living species [23]. Given two distinct classes, the algorithm can be used as an optimization procedure to determine the key variables that maximize the separation between the two classes [23]. As in SPA the GA uses a function to evaluate the quality of the variable selection. In this case, this function is called the fitness function, which is defined as the inverse of the cost function (Eq. (1)) [12]. For more details on GA see Refs. [23,24].

Support Vector Machine (SVM) is a machine learning technique derived from two foundations: Statistical Learning Theory and Mathematical Optimization, applied for classification, regression, and other learning tasks showing high performance in practical applications [25–28]. The support vector machine determines the limits of a decision, producing a great separation between classes by minimizing the errors. For this, SVM implements two basic mathematical operations: non-linear mapping of input vectors in a high dimensional feature space (kernels), and constructing a maximum margin hyperplane in the feature space. The construction of this hyperplane is performed in accordance with the principle of structural risk minimization (SRM). For more details on SVM see Refs. [19,25,26].

The present paper proposes an analytical method to classify gasoline as with or without dispersant and detergent additives based on infrared spectroscopy and supervised pattern recognition methods. The spectra of the distillation residues, which contain the pre-concentrated additives, were employed to overcome the lack of detectability of the IR spectroscopy.

2. Materials and methods

2.1. Samples

A total of 125 samples of gasoline type C were acquired in fuel stations from the states of Pernambuco, Alagoas and Sergipe in the Northeast of Brazil. So as to encompass the variability of the gasoline composition, the samples were collected from different producers and distributors. From these 125 samples, 49 were