Fuel 87 (2008) 3664–3668

Contents lists available at ScienceDirect

Fuel

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

Determination of gasoline origin by distillation curves and multivariate analysis

Helga G. Aleme [a], Letícia M. Costa [b], Paulo J.S. Barbeira [a,*]

[a] Laboratório de Combustíveis, Departamento de Química, ICEx, UFMG, Av. Antonio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
[b] Grupo de Espectrometria Atômica e Preparo de Amostras, Departamento de Química, ICEx, UFMG, Av. Antonio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil

ARTICLE INFO

Article history:
Received 25 September 2007
Received in revised form 19 June 2008
Accepted 20 June 2008
Available online 16 July 2008

Keywords:
Gasoline origin
Chemometrics
Pattern recognition
Linear discriminant analysis

ABSTRACT

Pattern recognition (PCA – principal component analysis) and classification (LDA – linear discriminant analysis) were employed to determine the origin of various samples of gasoline commercialized in the state of Minas Gerais, Brazil. With this in view, distillation curves were performed following ASTM D86 standard method, and PCA demonstrated that a small number of variables dominate the total data variability since the first three principal components (PCs) accounted for 87% of total variability. LDA was constructed using the origin declared in the invoices and the distances between groups were used to determine the similarity of the samples. Refineries REPLAN/REVAP presented the lowest distance value and REDUC/REGAP, the highest. About 80% of the samples, whose origins were not declared, were classified as belonging to the REGAP group, with 95% probability of correct classification.

1. Introduction

The Brazilian market commercializes many automotive fuels such as gasoline, hydrated ethanol, diesel oil, and natural gas. In order to increase efficiency and expand activities through new public and private sector investments, in 1997 the Brazilian government sanctioned the Petroleum Law, which softened the monopoly exercised by Petrobras. This opening stimulated competition which, despite expectations, gave rise in the area to several formal complaints of tax evasion, fraud, and commercialization of adulterated fuel – mainly automotive fuel.

Gasoline and industrial solvents have different taxations that vary according to the state where they are commercialized, thus market forces can greatly influence fuel adulteration and quality turning adulteration a profitable factor in the fuel market. As Minas Gerais is among the states with the lowest taxation rates in Brazil [1,2], large volumes of gasoline from various refineries and different states are commercialized in this state. Therefore, the study of the origin of fuel samples helps address the urgent need to reduce tax evasion in the commercialization of gasoline.

Gasoline is a mixture of hydrocarbons with chains that vary from 4 to 12 carbon atoms, boiling points between 30 and 220 °C, and specific gravities between 0.72 and 0.78 g m L^{-1} [3]. Brazil adds anhydrous ethanol to automotive gasoline at a variable rate of 20–25% (v/v), depending on alcohol market supply. Twenty-four billion liters of gasoline were commercialized in 2007, and presumably 10% of that fuel was adulterated with industrial solvents [4]. Quality assessment of gasoline is carried out through a series of assays recommended by the National Petroleum Agency (ANP) [4].

Studies carried out at Laboratório de Ensaios de Combustíveis (LEC-UFMG) showed that in 2007 approximately 20% of the analyzed gasoline were considered atypical [5]. Atypical samples are those that show a different behavior, although they are within the parameters established by the ANP. This lack of consistency may be attributed to the careful adulteration of the fuel with the addition of solvents, so that the final product is within legal specifications, or to the fact that the fuel is from different refineries [6].

Multivariate methods are useful tools to detect gasoline adulteration, as shown in recent literature [7–11]. Wiedemann et al. [7] discovered four groups of samples, according to the quantity/solvent ratio used for gasoline adulteration in REDUC and Manguinhos refineries (Brazil), using physicochemical and chromatographic parameters applied to hierarchical cluster analysis (HCA).

Skrobot et al. [8] produced a group of samples of adulterated gasoline for analysis via GC–MS by mixing gasoline with variable quantities of four different organic solvents. The HCA method was used in this research to visualize the distribution of the samples according to the solvent added, and the *k*-nearest neighbor (KNN) method was used to construct a classification diagram that can distinguish samples of pure gasoline from mixtures.

The KNN method was applied to nuclear magnetic resonance data by Kowalski and Bender [9] to classify substances according to different structural groups. The linear discriminant analysis (LDA) and the principal components analysis (PCA) associated to FTIR allowed Pereira et al. [10] to analyze the adulteration of

[*] Corresponding author. Tel.: +55 31 34995767.
E-mail address: barbeira@ufmg.br (P.J.S. Barbeira).

0016-2361/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.fuel.2008.06.015

gasoline with four different organic solvents (thinner, turpentine, paraffin oil, and rubber solvent).

Oliveira et al. [11] applied methods of multivariate analysis to detect adulterations in gasoline samples using distillation curves. Samples considered to be both within and without the limits established by the ANP were used for this study. Applying the SIMCA chemometric method (Soft Independent Modeling of Class Analogy) and starting from those observations, the authors constructed a group of training samples with samples previously considered within tolerances, and a group of test samples composed of conforming and non-conforming samples. As observed, 92% of the samples were classified correctly when the SIMCA method was applied. Three tests would be necessary to prove if the samples are within the established limits (90% and final boiling point of the distillation and alcoholic content) if a different test were used.

Other studies showed that gasoline from diverse origins has different physicochemical properties [12,13]. Moreira et al. [12] observed that the samples from REDUC and Manguinhos refineries (Brazil) have distinguishable behaviors with GC–MS analysis. Barbeira et al. [13] showed that gasoline from different refineries have diverse behaviors using results from different assays and multivariate analysis methods PCA, HCA, and LDA. Therefore, gasoline samples can be separated according to their origin using LDA with 97% efficiency.

In order to simplify and speed up the process of identification of the origin of gasoline during inspection procedures, the goal of this study was to evaluate the origin of gasoline samples commercialized in gas stations of Minas Gerais state using a multivariate analysis that combines pattern recognition (PCA) and classification (LDA), to contribute to the identification of samples from different Brazilian refineries. For this, the results obtained in the distillation curves, according to the American Society for Testing and Materials (ASTM D86), were used to construct the data sample.

ASTM D86 [14] describes the method for distillation of several petroleum products, at atmospheric pressure, in order to determine volatility features. This is carried out by checking if the light and heavy proportions of the fuel produced are appropriate for good performance during combustion and to detect adulterations with other products. For Brazilian automotive gasoline, the ANP establishes maximum temperature values for 10%, 50%, and 90% of the recovered volume beyond final boiling point and residue volume [15].

2. Experimental

2.1. Equipment and materials

A total of 3347 samples of regular and special gasoline were collected from approximately 1800 gas stations in 550 districts in the east of the state of Minas Gerais, over the course of 12 months. Of these samples, 2148 were used to construct PCA and LDA models, since 35.8% (1199 samples) were of unknown origin. All the samples used for the models were within the limits established by ANP [15].

In Brazil, special gasolines are products obtained by adding additives to regular gasoline. The small concentrations of these additives, detergents, and dispersants do not affect the physical-chemistry properties of the gasoline, making possible its use in the construction of the models. Premium gasolines were not included in this study because of their distinct composition (higher aromatic hydrocarbons content).

The origin of the samples was determined by taking into consideration data contained in the invoices issued by the gas stations when collecting the samples. The samples were stored in appropriate polyethylene flasks, and sealed and kept in cold storage (8–15 °C) until distillation analysis, to avoid loss of volatile components.

Gasoline samples were distilled in a Herzog HDA 627 Automatic Distillation Analyzer, according to ASTM D86 [14].

2.2. Experimental procedure

One hundred milliliters of gasoline were transferred to a specific distillation flask, equipped with a thermocouple sensor, and heated to keep the distillation rate between 4 and 5 mL per min. The distilled and condensed steam was collected in a cooled test tube (13–18 °C) and the recovered volume was measured with a digital volume sensor. Distillation curves (distillation temperature according to the recovered volume) were obtained after correcting temperatures to an atmospheric pressure of 760 mmHg and after correcting volume loss by measuring residue volume, according to ASTM D86 [14].

2.3. Pattern recognition and classification tools

LDA is a method used to classify the elements of a sample or population. To apply it, the groups in which each sample element may be classified have to be previously defined, taking into account its general features. This criterion helps to define the classification or discrimination rule, which is used to classify new samples in the

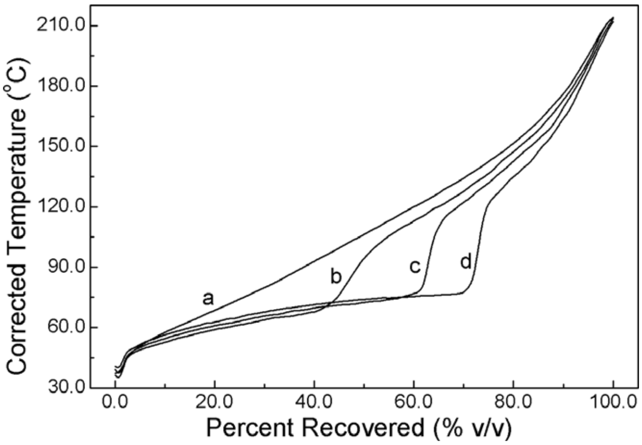

Fig. 1. Distillation curves obtained from gasoline samples doped in different concentrations of anhydrous ethanol. (a) 0; (b) 10; (c) 20; (d) 30% (v/v).

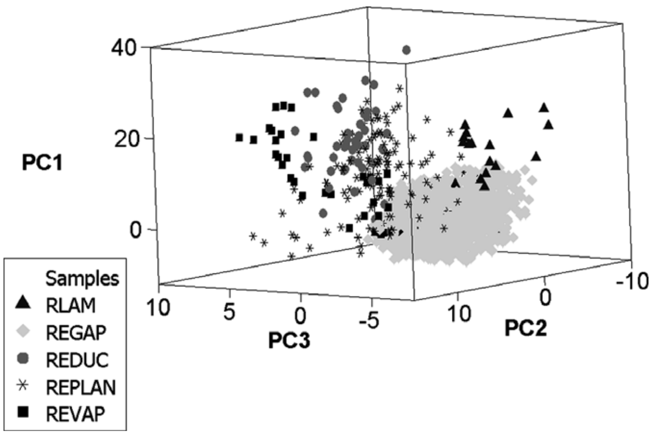

Fig. 2. Evaluation of the formation of clusters in gasoline samples from RLAM, REGAP, REPLAN, REVAP, and REDUC refineries using PCA with correlation matrix.

existing groups. According to this method, the classification or discrimination rule is based on probability theory, and on the principle of maximum verisimilitude [16].

The PCA aims at explaining the structure of variance and covariance of a random vector, composed of *p*-random variables, to construct linear combinations of the original variables [16]. In PCA the data matrix is decomposed into score and loading matrices. The score vectors describe the relationship between samples in the subspace model and the loading vectors describe the importance of each variable in the model. It can graphically represent intersample and intervariable relationships and can provide a way to reduce the dimensionality of the data [17].

Minitab Release (version 14 for Windows) was used to construct PCA and LDA models.

3. Results and discussion

The temperature in the distillation curve of gasoline with no alcohol added varies from 30 to 220 °C. Changes in the shape of the distillation curves may be observed when different quantities of anhydrous ethanol are added (Fig. 1) and an alteration can be viewed close to 70 °C. This alteration may be caused by different azeotropes formed in an ethanol–gasoline mixture. The behavior of these azeotropes depend on the gasoline components, that is, the behavior of gasoline with higher content of light fractions (isoparaffins and naphthalenes) will differ from gasoline with higher content of heavy fractions (isoparaffins and aromatics) [18].

To reduce the influence of ethanol on the distillation curve temperatures, the values ranging from 60% to 84% (v/v) were removed,

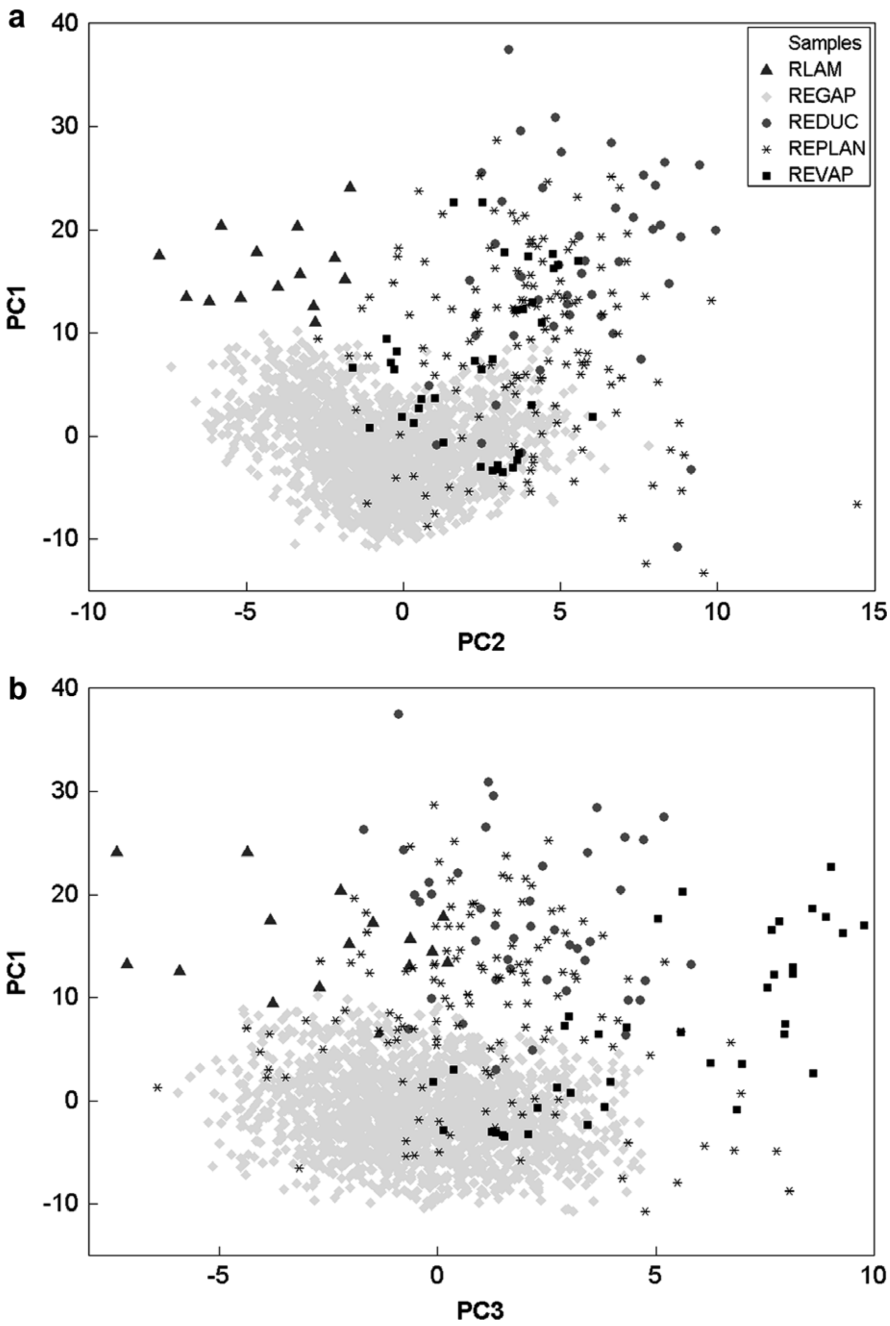

Fig. 3. Discrimination of gasoline origin by PCA. (a) PC1 × PC2; (b) PC1 × PC3.

as this alteration in the distillation curve provides poor reproducible values in this range.

Distillation curves in different ranges (from 0% to 59% (v/v) and 85% to 93% (v/v)), together with the final distillation point, were transformed in a data matrix composed of boiling temperatures and distilled volumes as lines and columns, respectively.

Before applying the PCA method, the data were previously mean centered and dimensionally reduced through linear combinations of the original variables, thus giving rise to score and loading matrices. Using the PCA method the number of variables were reduced, but still show high variance. In order to use the LDA method, the origin of the data was assessed and, applying the discrimination rule based on the theory of probabilities, the probability of each element of the sample being of the selected group was checked. It was also observed that a larger number of samples than of variables were needed to apply this method [13].

Figs. 2 and 3 show the grouping of samples according to the origin stated in the invoice. PCA demonstrated that a small number of variables dominate the total data variability, as the three first principal components (PCs) accounted for 87% of the total variability. The first principal component is responsible for 66% of the entire information, the second accounts for 12%, and the third explains 9% of the total information. Fig. 2 shows the data with the three principal components and clearly differentiated groups. In Fig. 3a (PC1 × PC2), the first principal component is responsible for the separation of RLAM and REGAP refineries, as well as for the separation of the samples from REDUC and REGAP. The second principal component PC2 separates RLAM from REPLAN, REVAP, and REDUC. In Fig. 3b (PC1 × PC3), the third principal component separates the set of samples from REDUC from the REVAP samples. A distinct visual clustering appears when the data were displayed with respect to the first two principal components, which was not surprising since the first principal component accounts for the maximum possible one-dimensional projection of the total variation of the individual data points.

The samples from the REGAP refinery are the largest group of the data set and have quite similar features, compared to the other groups. The first principal component showed that variable 31% (v/v) distilled volume is the most important for the separation of origins, while variable 93% (v/v) is the least important to construct the model. The release of the light fraction of gasoline occurs mainly at point 31% (v/v) of the distillation curve. This may be associated to the light aliphatic hydrocarbons (as hexanes) and to a small amount of olefinic hydrocarbons (as hexenes) [11,19]. Variables 10%, 50%, and 90% (v/v) of the distilled volume, beyond final boiling point, whose limits are established by the ANP [15], presented high weights in the first principal component and are not only important in quality evaluation of fuel, but in the determination of its origin as well. According to the invoice, samples were correctly classified, even after removing the ethanol influence from the distillation curves.

Table 1
Average distances between groups in the LDA model with cross-validation

Origin	Average distance
RLAM/REDUC	33.0
RLAM/REGAP	23.8
RLAM/REVAP	33.1
RLAM/REPLAN	35.6
REDUC/REPLAN	29.3
REDUC/REGAP	57.6
REDUC/REVAP	22.5
REGAP/REPLAN	21.9
REGAP/REVAP	26.6
REPLAN/REVAP	12.1

Table 2
Summary of classifications with cross-validation in the determination of the origins of the 2148 gasoline samples used in the LDA model

Origin	Correct classification (%)
REGAP	97.8
REPLAN	71.9
REVAP	73.1
REDUC	75.9
RLAM	88.2

Table 3
Origins of the 3347 gasoline samples according to invoices, and the origins of the 1199 unknown samples after the chemometric analysis

Origin	Origin according to invoices (%)	Chemometric analysis of unknown origin (%)
Unknown	35.8	–
REGAP	56.5	79.6
REPLAN	4.4	9.5
REVAP	1.2	4.4
REDUC	1.6	3.6
RLAM	0.5	2.9

Model LDA with cross-validation was obtained from the same set of samples of the PCA model. The distances between groups composed by different refineries can be observed in Table 1 and the summary of classification in Table 2. In the LDA model, the separation of groups was carried out using the origin stated as classification parameter, as presented in Table 3 (invoice column).

As shown in Table 1, the groups composed by the refineries with smaller distances present similar compositions. The REPLAN/REVAP pair presented a smaller distance, while the REDUC/REGAP pair presented the longest one. This means that the samples from REPLAN and REVAP have similar behavior in the distillation curve, while samples from REDUC and REGAP present different behaviors. In that model, the percentage of successful cross-validations was 95.1%.

Table 2 shows that the highest percentage of correct classifications corresponds to the REGAP samples. RLAM presented the second highest percentage, whereas for REDUC, REVAP and REPLAN the percentages were over 70%. The high percentage of correct classifications of the REGAP samples may be due to the large number of grouped samples in the model, as shown in Fig. 2. The high percentage shown for RLAM is related to the fact that this group is the farthest from the others, as shown in Table 1.

Even under these conditions, the model produced an acceptable percentage of correct separations between groups. Similar results were presented by Barbeira et al. [13], using five principal components, to describe the model with the results of physical-chemistry assays.

The origin of the set of unknown samples (1199 samples) was obtained based on the LDA model created previously, with 95.1% of correct classifications, as shown in Table 3. The chemometric analysis showed that most of the unknown samples are originally from REGAP, as the samples used in the model. However, the percentages of the other origins of the unknown samples differ greatly from those used to construct the model. This shows that the amount of samples in each set of the refineries did not cause a tendentious classification of the unknown samples.

4. Conclusions

PCA and LDA multivariate methods applied to the distillation curve data allowed the separation of gasoline samples, commercialized in the state of Minas Gerais (Brazil) from different refineries, according to their origin.

The number of variables of the set sample may be reduced and a high total variance can be maintained using PCA pattern recognition method. It was also evident that this method was efficient to visualize the separation of groups according to the refinery of origin with only three PCs accumulating 87% of variability.

Moreover, it was observed that by using the LDA classification method the distance between refineries was similar, except for the case of REGAP/REDUC refineries that presented the greatest distance, and for REVAP/REPLAN, with the shortest distance. The origin of unknown samples can be predicted with a 95% chance of correct classification.

Multivariate analysis associated to the distillation curve proved to be an important tool for the classification of gasoline samples according to their origin, and may be used to avoid fraud in the commercialization of this fuel.

References

[1] <www.sef.rj.gov.br> (accessed August 2007).

[2] <www.fazenda.mg.gov.br> (accessed August 2007).

[3] Campos AC, Leontsinis E. Petróleo e Derivados. São Paulo: Editora Técnica Ltda; 1990.

[4] <www.anp.gov.br> (accessed August 2007).

[5] Relatório Mensal, Programa de Monitoramento da Qualidade dos Combustíveis – ANP, maio; 2007.

[6] Barbeira PJS. Engenharia Térmica 2002;2:48–50.

[7] Wiedemann LSM, d'Avila LA, Azevedo DA. Fuel 2005;84(4):467–73.

[8] Skrobot VL, Castro EVR, Pereira RCC, Pasa VMD, Fortes ICP. Energy Fuels 2005;19(6):2350–6.

[9] Kowalski BR, Bender CF. Anal Chem 1972;44:1405.

[10] Pereira RCC, Skrobot VL, Castro EVR, Fortes ICP, Pasa VMD. Energy Fuels 2006;20(3):1097–102.

[11] Oliveira FS, Teixeira LSG, Araujo MCU, Korn M. Fuel 2004;83(7–8):917–23.

[12] Moreira LS, d'Avila LA, Azevedo DA. Chromatographia 2003;58(7/8):501–5.

[13] Barbeira PJS, Pereira RCC, Corgozinho CNC. Energy Fuels 2007;21(4):2212–5.

[14] American Society for Testing and Materials (ASTM). Standard test for distillation of petroleum products at atmospheric pressure, ASTM-D86, West Conshohocken PA: ASTM Committee of Standards; 2007.

[15] Portaria no 239, 12 de Janeiro de 2001, Agência Nacional do Petróleo.

[16] Mingoti SA. Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada. Belo Horizonte: Editora UFMG; 2005.

[17] Fernandes AP, Santos MC, Lemos SG, Ferreira MMC, Nogueira ARA, Nóbrega JA. Spectrochim Acta B 2005;60:717–24.

[18] Balabin RM, Syunyaev RZ, Karpov SA. Energy Fuels 2007;21(4):2460–5.

[19] Line DR. CRC – handbook of chemistry and physics. 84th ed. CRC Press Inc.; 2004.