ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/231272262

Use of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in Gas Chromatographic (GC) Data in the Investigation of Gasoline Adulteration

ARTICLE *in* ENERGY & FUELS · SEPTEMBER 2007

Impact Factor: 2.79 · DOI: 10.1021/ef0701337

CITATIONS
26

READS
38

5 AUTHORS, INCLUDING:

Vinicius L. Skrobot
Brazilian Regulatory Agency of Petroleum, …
4 PUBLICATIONS **102** CITATIONS
SEE PROFILE

Eustáquio Vinicius Ribeiro de Castro
Universidade Federal do Espírito Santo
87 PUBLICATIONS **851** CITATIONS
SEE PROFILE

Vânya M. D. Pasa
Federal University of Minas Gerais
54 PUBLICATIONS **527** CITATIONS
SEE PROFILE

Isabel C. P. Fortes
Federal University of Minas Gerais
12 PUBLICATIONS **219** CITATIONS
SEE PROFILE

Available from: Vânya M. D. Pasa
Retrieved on: 31 January 2016

Use of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in Gas Chromatographic (GC) Data in the Investigation of Gasoline Adulteration

Vinicius L. Skrobot,*,† Eustáquio V. R. Castro,‡ Rita C. C. Pereira,† Vânya M. D. Pasa,† and Isabel C. P. Fortes†

Laboratório de Ensaio de Combustíveis, Departamento de Química, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Avenida Antônio Carlos, 6627, Campus Pampulha, CEP 31270-901, Belo Horizonte, Minas Gerais, Brazil, and Departamento de Química, Centro de Ciências Exatas, Universidade Federal do Espírito Santo, Avenida Fernando Ferrari, s/n CEP 29060-900, Vitória, Espírito Santo, Brazil

Received March 14, 2007. Revised Manuscript Received August 15, 2007

Chemometric data analysis was applied to chromatographic data as a modeling tool to identify the presence of solvents in gasoline obtained at gas stations in the Minas Gerais state. As a training set, 75 samples were formulated by mixing pure gasolines with varying concentrations of four solvents and analyzed by gas chromatography–mass spectrometry. Selected chromatographic peak areas were used in chemometric analysis. Sample distribution patterns were investigated with principal component analysis (PCA). Score graphics revealed a clear sample agglomeration according to the solvents added. Classification models were created with linear discriminant analysis (LDA). Because gasoline presents a very complex profile and the chromatographic data contains too many variables, two approaches were tested to reduce the dimensionality of the data before LDA. Fisher weights were used as an exclusion criterion of lesser variables, and the original variables were substituted for a few principal components obtained from the covariance matrix. To test the quality of the models, a test set with a total of 31 new samples was prepared using certified gasolines mixed with the same solvents used in the training set. Both models indicated the presence of solvent in gasoline effectively, failing only for samples whose solvent concentrations were low. The PCA plus LDA model was more efficient in signaling solvent-free samples, which reduced the number of false positive cases.

1. Introduction

In Brazil, oil refining and fuel distribution stopped being the monopoly of state companies in 1995.[1] Since then, the role played by state and private companies in the Brazilian fuel market has changed greatly and thousands of new fuel distribution companies and gas stations have been opened. This opening market not only increased competition and lowered prices but also gave a chance to some distribution companies to raise revenue by adulterating fuel. The addition of solvents is one of the most common adulteration practices because of the large difference in the taxation of gasoline and solvents. This kind of adulteration causes environmental pollution, poor engine performance, and tax revenue losses.[2,4] Recently, to overcome this problem, the Brazilian government has developed and implemented a program that determines the use of solvent markers to facilitate their identification in gasoline.[5] However, this has a high cost to the country, and only few laboratories are able to analyze the presence of the markers in gasoline. The development of analytical methodologies to identify the presence of solvents in fuel has been the subject of academic and forensic research.[2,3,6]

Automotive gasoline is a complex mixture comprised basically of hundreds of different hydrocarbons ranging from C_4 to C_{12}. Because most solvents used in the adulteration are petrochemical derivatives, the identification of their presence in gasoline is a challenging task. Because the proper performance of this fuel depends upon a balanced combination of its compounds, a number of physicochemical tests are currently applied to evaluate its quality.[7–10] However, because these tests are not intended to identify adulteration and some solvents are quite similar to gasoline, they usually are not efficient in flagging illegal additions.[6]

Gas chromatography (GC) has been widely used to evaluate the quality of gasoline.[6,11–13] Moreira et al., for example, have used GC with flame ionization detector (FID) and mass spectrometry (MS) detection to evaluate modifications resulting from the addition of some solvents to gasoline.[6] However, the richness of information obtained by GC makes the evaluation of the quality of gasoline by this technique extremely complex. A visual comparison of reference gasoline chromatograms to

* To whom correspondence should be addressed. Telephone/Fax: 55-02131-3499-6650. E-mail: vinicius_skrobot@hotmail.com.

† Universidade Federal de Minas Gerais.

‡ Universidade Federal do Espírito Santo.

(1) Brazilian National Petroleum Agency. Dois anos/ANP; A Agência: Rio de Janeiro, 2000, *67*.

(2) Oliveira, F. S.; Teixeira, L. S. G.; Araújo, M. C. U.; Korn, M. *Fuel* **2004**, *83*, 917–923.

(3) Kaligeros, S.; Zannikos, F.; Stournas, S.; Lois, E. *Energy* **2003**, *28*, 15–26.

(4) *Gasolina Automotiva*; Refinaria Gabriel Passos: Brazil, 2000.

(5) Brazilian National Petroleum Agency. Portaria Agência Nacional do Petróleo 274 de 1/11/2001.

(6) Moreira, L. S.; Ávila, L. A.; Azevedo, D. A. *Chromatographia* **2003**, *58*, 501–505.

(7) American Society for Testing and Materials. D4052, Standard test method for density of liquids by digital density meter. 2002.

(8) American Society for Testing and Materials. D86, Standard test method for distillation of petroleum products at atmospheric pressure. 2005.

(9) American Society for Testing and Materials. D6277, Standard test method for determination of benzene in spark-ignition engine fuels using mid-infrared spectroscopy. 2006.

(10) Brazilian Association for Technical Standards. NBR 13992, Automotive gasoline. Method for stablish alcohol content.

10.1021/ef0701337 CCC: $37.00 © 2007 American Chemical Society
Published on Web 09/21/2007

Table 1. Composition of the Solvents[19,20]

solvent	composition
(A) naphtha	aliphatic hydrocarbons that distillate between 151 and 254 °C
(B) light naphtha	aliphatic, naftenic, and aromatic hydrocarbons that distillate between 52 and 120 °C
(C) thinner	aromatic hydrocarbons (toluene and xylenes), acetates, and alcohols
(D) kerosene	aliphatic, naftenic, and aromatic hydrocarbons that distillate between 150 and 239 °C

those of different gasoline samples is cumbersome and ineffective because changes in the oil feedstock, refining process, and aging causes modifications in the gasoline chromatographic profile, which do not necessarily mean quality deterioration. The development of methodologies to evaluate the gasoline quality must take into account these composition variations.

A number of studies have demonstrated the usefulness of the application of chemometric tools to gasoline analysis. Oliveira et al.[2] gathered data from gasoline distillation tests to create a model by soft independent modeling of class analogy (SIMCA) capable of indicating samples that did not meet Brazilian specifications. Sandercock et al.[14] used principal component analysis (PCA) and linear discriminant analysis (LDA) to establish the origin of gasoline samples with three different grades based on chromatographic data.

Application of the LDA chemometric tool using chromatographic data has already been employed to detect adulteration in whisky samples.[15] Tan et al. have used PCA and SIMCA to create classification models to identify petroleum-based accelerants in fire debris.[16]

In this paper, a set of samples of pure gasoline and their mixtures with four different solvents in various concentrations were analyzed by GC–MS, and the data obtained were analyzed by chemometric tools. Chromatographic data were analyzed by PCA to look for patterns in chromatographic data, and LDA was used to create models to classify the samples in classes determined by the solvent added or its absence. Because of the large number of variables, two tools were tested to reduce them without losing important information: Fisher weights and PCA. In the former approach, the variables with less value for sample segregation were deleted. In the latter, the original variables were substituted for a few principal components obtained by PCA. The combination of PCA and LDA is commonly applied to problematic databases.[17,18] One of the concerns in this work was related to the representativity of the variability of the samples used to create the classification models. If the training set did not capture the normal variation in gasoline composition, the statistical model could flag pure gasoline as adulterated because of normal variations in composition. To evaluate the quality of the classification models, they were tested on new samples prepared with other gasolines (external samples). The working hypothesis of this research was that, despite the complexity of solvents and gasoline and variations in its composition, chromatographic data would supply enough information to identify the subtle differences between pure gasoline and samples containing solvents. Moreover, it is conceivable that univariate statistical analysis would not suffice to identify the differences between samples because of the complexity of the mixtures, and only multivariate tools could deal with this problem. The objective of this study was to create statistical models to identify the presence of solvents in gasoline based on chromatographic data. As far as we are concerned, such use of chemometrics has not been tried before now. The proposed methodologies herein are readily applicable to GC using FID, because the identification of the individual compounds is not a crucial step of the process; the comparison between peaks of the same retention times is the only requirement. The use of MS, nevertheless, offered important information to evaluate the quality of the proposed methodologies at this stage of the research. In opposition to the technology applied nowadays, this approach to identify adulteration does not rely on the expensive (and passive to fraud) use of chemical markers.

Table 2. Preparation of the Training Set

pure gasolines	pure samples plus solvent[a] (gasoline)(solvent)-[solvent concentration in % (v/v)]			
	solvent A	solvent B	solvent C	solvent D
G1[b]	G1A-2[b]	G1B-2[b]	G1C-2[b]	G1D-2[b]
G2[b]	G2A-4[c]	G2B-4[c]	G2C-4[c]	G2D-4[c]
G3[c]	G3A-6[b]	G3B-6[b]	G3C-6[b]	G3D-6[b]
G4[b]	G4A-8[b]	G4B-8[b]	G4C-8[b]	G4D-8[b]
G5[b]	G5A-10[c]	G5B-10[c]	G5C-10[c]	G5D-10[c]
G6[c]	G6A-12[b]	G6B-12[b]	G6C-12[b]	G6D-12[b]
G7[b]	G7A-14[b]	G7B-14[b]	G7C-14[b]	G7D-14[b]
G8[b]	G8A-16[c]	G8B-16[c]	G8C-16[c]	G8D-16[c]
G9[b]	G9A-18[b]	G9B-18[b]	G9C-18[b]	G9D-18[b]
G10[c]	G10A-20[b]	G10B-20[b]	G10C-20[b]	G10D-20[b]
G11[c]	G11A-22[c]	G11B-22[c]	G11C-22[c]	G11D-22[c]
G12[b]	G12A-24[b]	G12B-24[b]	G12C-24[b]	G12D-24[b]
G13[b]	G13A-26[b]	G13B-26[b]	G13C-26[b]	G13D-26[b]
G14[b]	G14A-28[c]	G14B-28[c]	G14C-28[c]	G14D-28[c]
G15[c]	G15A-30[b]	G15B-30[b]	G15C-30[b]	G15D-30[b]

[a] For example, the code G13B-26 corresponds to the mixture of pure gasoline 13 with solvent B in a concentration of 26% (v/v). [b] Samples used in the reduced training set. [c] Samples used in the reduced test set.

Table 3. Preparation of the Test Set

certified gasolines	pure samples plus solvent[a] (gasoline)(solvent)-[solvent concentration in % (v/v)]			
	solvent A	solvent B	solvent C	solvent D
G′1	G′1A-2	G′3B-2	G′1C-2	G′1D-2
G′2	G′1A-5	G′3B-5	G′1C-5	G′1D-5
G′3	G′1A-10	G′3B-10	G′1C-10	G′1D-10
	G′1A-15	G′3B-15	G′2C-15	G′1D-15
	G′1A-20	G′3B-20	G′2C-20	G′1D-20
	G′1A-25	G′3B-25	G′2C-25	G′1D-25
	G′1A-30	G′3B-30	G′2C-30	G′1D-30

[a] For example, the code G13B-26 corresponds to the mixture of pure gasoline 13 with solvent B in a concentration of 26% (v/v).

2. Experimental Section

2.1. Samples. The samples were prepared from certified gasolines supplied by Petrobras (Petróleo Brasileiro SA), collected in gas stations, and four solvents described in Table 1. All solvents and certified gasoline samples were supplied by Petrobras, except for one, thinner, which was commercially available (Dissolminas-3500). Gasoline

(11) Blomberg, J.; Schoenmakers, P. J.; Brinkman, U. A. *J. Chromatogr., A* **2002**, *972*, 137–173.
(12) Philp, R. P.; Mansuy, L. *Energy Fuels* **1997**, *11*, 749–760.
(13) Sojak, L.; Addova, G.; Kubinec, R.; Kraus, A.; Bohac, A. *J. Chromatogr., A* **2004**, *1025*, 237–253.
(14) Sandercock, P. M. L.; Pasquier, E. *Forensic Sci. Int.* **2003**, *134*, 1–10.
(15) Saxberg, B. E.; Duewer, D. L.; Booker, J. L. *Anal. Chim. Acta* **1978**, *103*, 201–212.
(16) Tan, B.; Hardy, J. K.; Snavely, R. E. *Anal. Chim. Acta* **2000**, *422*, 37–46.
(17) Seregély, Z.; Deák, T.; Bistray, G. D. *Chemom. Intell. Lab. Syst.* **2004**, *72*, 195–203.
(18) Brereton, R. G. *Chemometrics—Data Analysis for the Laboratory and Chemical Plant*, 1st ed.; John Wiley and Sons, Inc.: U.K., 2003.
(19) Petrobras Distribuidora S.A. http://www.br.com.br/ (accessed December 2004).
(20) Ipiranga Química. http://ipirangaquimica.ipiranga.com.br/ (accessed December 2004).

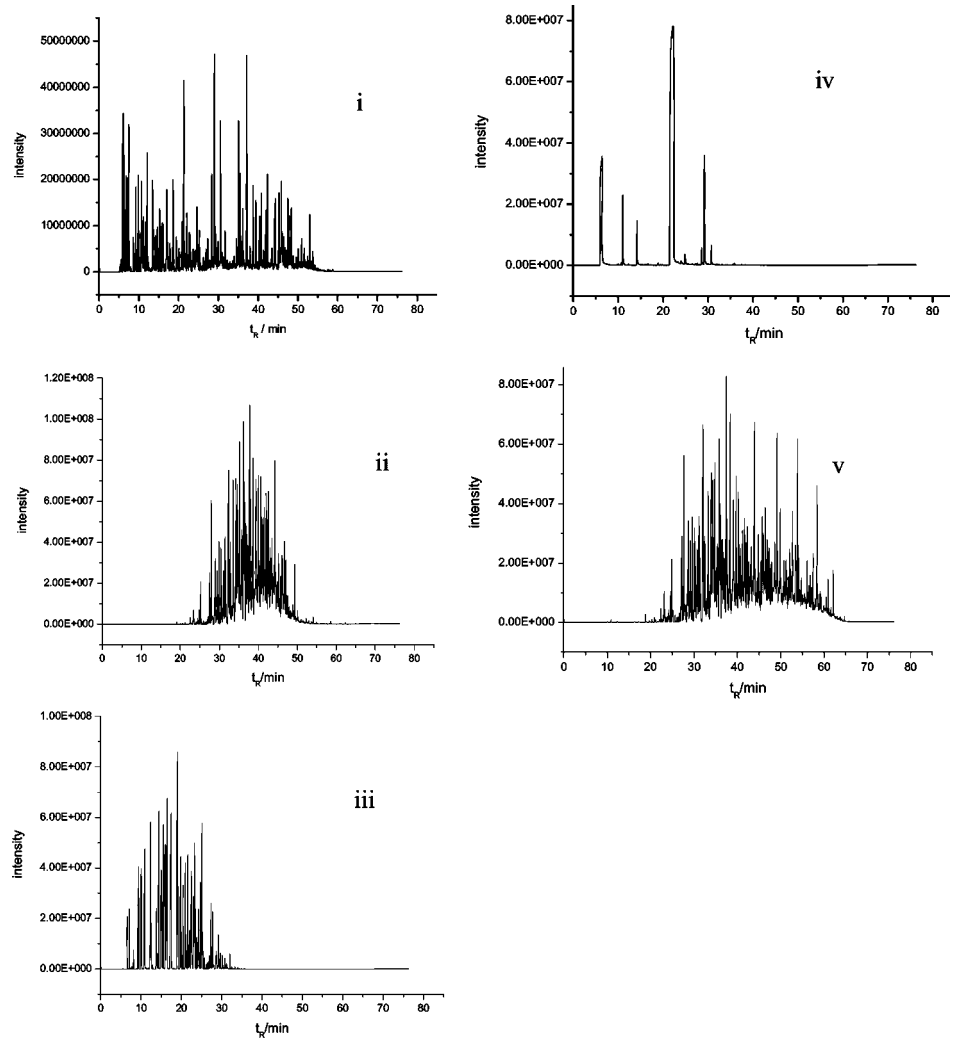

Figure 1. TICs of reference gasoline (i) and solvents. Naphtha (ii), light naphtha (iii), thinner (iv), and kerosene (v) obtained as described in section 2.3.1.

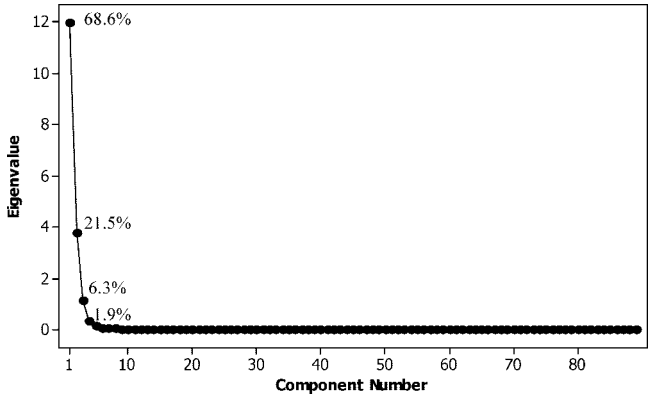

Figure 2. Scree plot of the 89 principal components with the variance captured for the first four components.

samples were analyzed according to standard methods established by the Brazilian Government Petroleum Agency (ANP).

A total of 15 samples of regular gasoline were collected from different gas stations belonging to eight different distributors for 1 month. The samples were analyzed and approved according to standard methods.

To simulate as many adulteration conditions as possible, the gasolines collected at gas stations were mixed with the four solvents in concentrations ranging from 2 to 30% (v/v) and were labeled according to Table 2. In Brazil, regular gasoline has ethanol in its composition in concentrations varying from 13% (1990) to 26% (1999)

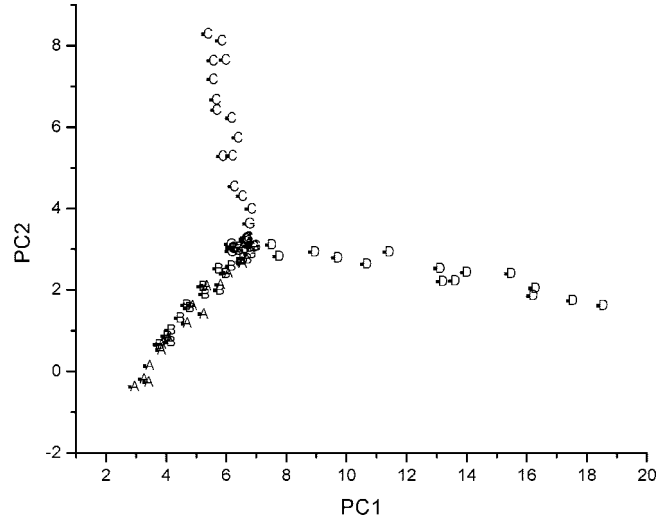

Figure 3. Score plot of PC1 and PC2 for the 75 samples of pure and mixed gasolines. A, naphtha; B, light naphtha; C, thinner; D, kerosene; and G, gasoline.

in volume, depending upon the sugar price (ethanol and sugar are both obtained from sugar cane in Brazil). At present, its concentration is established as 25% in volume.[21] The ethanol concentration was corrected after the addition of the solvents with ethanol P.A. (synthetic). This group of mixtures was called the training set because they were used to create the classification model.

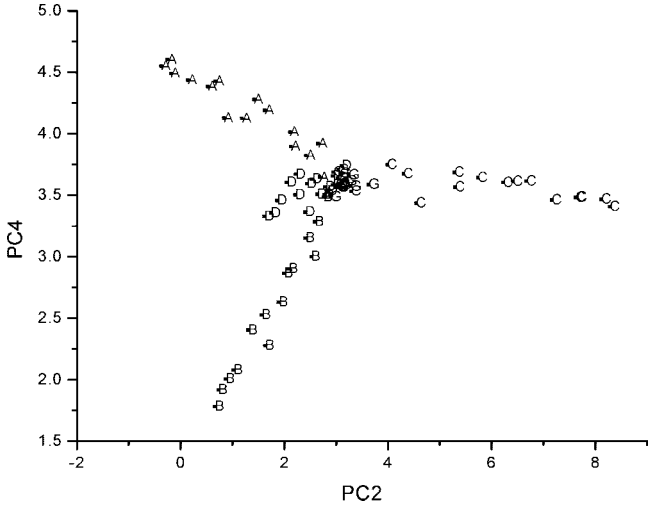

Figure 4. Score plot of PC2 and PC4 for 75 samples of pure and mixed gasolines. A, naphtha; B, light naphtha; C, thinner; D, kerosene; and G, gasoline.

Table 4. Exclusion Progress of the Variables by Fisher Weight

Fisher weight	number of variables used	percentage of correctly classified samples (from the model)	percentage of correctly classified samples (cross-validation)
1.500	53	100.0	72.0
1.750	48	100.0	78.7
2.000	35	98.7	84.0
2.250	29	97.3	88.0
2.500	23	94.7	86.7
2.625	18	96.0	86.7
2.750	16	93.3	86.7
3.000	5	84.0	78.7

The group called the test set comprising certified samples mixed with the same solvents as shown in Table 3 was used to test the quality of the classification performed by the statistic model created.

2.2. Gas Chromatography–Mass Spectrometry (GC–MS) Analysis. General profiles of all samples were obtained using electron impact (EI)/MS. Analyses were conducted on an automated GC–MS Shimadzu equipment model GC-17A/QP-5050A using a fused silica capillary column (50 m × 0.2 mm i.d. × 0.5 μm, PONA, HP), with polymethylsiloxane as the stationary phase and helium as the carrier gas, at a constant flow rate of 0.4 mL/min. Sample aliquots of 0.5 μL were injected in the split mode (1:24) without solvent delay. Analyses were performed under the following conditions: the column was kept at 34 °C for 5 min and then heated to 60 °C at 2 °C/min. After that, the temperature was increased at a rate of 3 °C/min up to 185 °C and finally to 250 °C at 10 °C/min. The final temperature was kept constant for 10 min.

The mass spectrometer working in electron ionization mode at 70 eV was operated in full-scan mode (*m*/*z* 45–350) with a sampling rate of 2 scans/s.

The presence of different classes of compounds in samples was confirmed using the total ion chromatogram (TIC) in addition to fragmentation patterns and library matching (Wiley Class 5000, 6th edition).

2.3. Data Analysis. *2.3.1. GC–MS.* Chromatograms were normalized to the unit area to minimize variations because of fluctuations in equipment response. Two criteria were used to select chromatographic peaks for statistical analysis: (a) Only peaks with areas larger than 0.9% of the total (normalized area) were considered. (b) After compound automatic identification by the software, compounds with mass spectra less than 90% similar to the reference spectra were discarded. Note that the identification algorithm compares each of the mass peak found in the analysed compound to a databank with thousands of common fuel substances.

Table 5. Chromatographic Peaks Selected To Create the Discriminant Model

variable	compound	retention time (min)
P1	*n*-hexane	10.67
P2	3-methylheptane	14.69
P3	1,3-dimethylcyclopentane (cis)	15.98
P4	*n*-heptane	17.06
P5	ethylcyclopentane	19.49
P6	1,2,4-trimethylcyclopentane	20.13
P7	3,5-dimethyloctane	22.14
P8	3-methylheptane	22.71
P9	1,2-dimethylbenzene	29.02
P10	*n*-nonane	31.70
P11	1-ethyl-3-methylbenzene	35.19
P12	1,2,4-trimethylbenzene	35.54
P13	1-ethyl-4-methylbenzene	36.15
P14	*n*-decane	37.96
P15	1,3,5-trimethylbenzene	38.75
P16	1,2,3,4-tetramethylbenzene	46.03
P17	2,3-dihydro-4,7-dimethyl-1*H*-indane	51.62
P18	1-methylnaftalene	53.02

Only well-separated peaks detectable in all gasoline and mixture chromatograms were considered because of the characteristics of the chemometric tools used. That is to say that the matrices under study must be complete; i.e., all samples must have values for all variables.

2.3.2. Chemometrics. PCA relies on the linear transformation of the original set of measurements into a substantially smaller set of orthogonal variables while retaining as much information present in the original data set as possible.[22] The original data set is substituted for two matrices that contain information about the weight of the original variable in the PC space (loading matrix) and the scattering of the samples in this space (score matrix). In this work, PCA was performed with S-PLUS (Mathsoft, Inc.). Because all variables considered in this study had the same scale (normalized area), PCs were obtained from the covariance matrix.

Fisher weights, *W*, allow for the evaluation of how useful a variable is to discriminate the samples between groups. This tool uses the variance and the difference between averages for each variable for a group of representative samples (training set) to calculate a score related to the ability of the variable to indicate differences between groups.[23] Fisher weights were calculated with Microsoft Excel 2000 (Microsoft Corporation).

LDA is a classification tool that consists of calculating linear combinations (classification functions) of the original variables that have the property of maximizing differences between groups and minimizing differences within groups. These functions are obtained from a group of samples whose classes are known (the training set). Minitab (Minitab, Inc.) was used to perform the calculations. The algorithm used by this software is a modification of the Fisher discriminant functions.[23]

The use of LDA required a reduction of the number of variables. As mentioned before, this was accomplished by elimination of the lesser variables (according to Fisher weights) and by PCA.

The use of PCA as a variable reduction tool before LDA requires a careful evaluation of the number of components to be used. In this work, this was accomplished by dividing the training set into two groups; the reduced training set and the reduced test set, which contained two-thirds and one-third of the original training set, respectively (Table 2). The ideal number of PCs was evaluated by applying the classification model obtained with the reduced training set in the reduced test set.

3. Results and Discussion

3.1. GC–MS. With an illustration of the complexity of the systems under study, the chromatograms of a reference gasoline and the solvents used in this study are presented in Figure 1.

(21) Santos, A. S.; Valle, M. L. M.; Giannini, R. G. *Economia e Energia*, 2000, *19*.

(22) Héberger, K.; Csomós, E.; Simon-Sarkadi, L. *J. Agric. Food Chem.* **2003**, *51*, 8055–8060.

Table 6. Test Set Classification by Fisher Weights plus LDA

pure gasolines		gasoline plus solvent A		gasoline plus solvent B		gasoline plus solvent C		gasoline plus solvent D	
sample	attribution	sample	attribution	sample	attribution	sample	attribution	sample	attribution
G′1	G	G′1A-2	G	G′3B-2	B	G′1C-2	C	G′1D-2	C
G′2	C	G′1A-5	B	G′3B-5	B	G′1C-5	C	G′1D-5	B
G′3	G	G′1A-10	B	G′3B-10	B	G′1C-10	C	G′1D-10	B
		G′1A-15	A	G′3B-15	B	G′2C-15	C	G′1D-15	G
		G′1A-20	A	G′3B-20	B	G′2C-20	C	G′1D-20	D
		G′1A-25	A	G′3B-25	B	G′2C-25	C	G′1D-25	D
		G′1A-30	A	G′3B-30	B	G′2C-30	C	G′1D-30	D

Table 7. Percentage of Samples Correctly Classified by PCA plus LDA

	percentage of samples correctly classified	
number of PCs	reduced training set	reduced test set
1	50	64
2	68	72
3	68	72
4	80	84
5	86	84
6	84	84
7	88	84
8	88	84
9	88	84
10	88	84
11	90	84
12	90	84

As one can notice, many of the compounds present in the solvents also belong to gasoline. This can be confirmed by the similar retention times and identifications in mass spectra. Naphtha is comprised of intermediate compounds that come out between 20 and 50 min of the chromatographic run. Light naphtha has in its composition mainly compounds with high vapor pressure and low molecular weight. The third solvent, thinner, is the least complex one and presents only a few compounds. The fourth solvent used, kerosene, is formed mainly by compounds with relatively low vapor pressure and high molecular weights because they come out after 30 min.

Because of the complex feature of gasoline and the coincidence with compounds present in solvents, any attempt to identify the presence of solvents in gasoline must consider many peaks.

Thus, 89 peaks from the gasolines were selected for the statistical calculations according to the criteria described in section 2.3.1. The visual inspection of the chromatograms of each sample showed a considerable amount of noise, which stressed the need for using a large number of peaks to achieve a stable classification model.

3.2. PCA. Figure 2 presents the scree plot of the variance captured by each PC obtained from the covariance matrix. It shows that the total variability of the data can be concentrated in just a few new coordinates. The first four principal components captured around 98% of the whole data variance, with the fourth component still representing 1.9% of the variance. It suggests that there are four independent variation sources, suggesting that each solvent contributes with a different modification to the chromatographic data.

To check for visual clustering of all 75 samples analyzed, different combinations of PC scores were performed. Score plot of PC1 versus PC2 is given in Figure 3. Samples were identified according to the letters given in Table 1, and pure gasoline samples were labeled only as "G". One can notice that the first PC alone is enough to separate the samples containing solvent "D" (kerosene) from the rest, while the second PC allows for the discrimination of samples with solvent "C" (thinner). Detailed inspection of the distribution of the samples in the PC space also revealed that they tend to group according to the solvent concentration and that the less concentrated ones tend to group in the center of the three-arm pattern.

As shown in Figure 4, among all score PC combinations, PC2 versus PC4 yielded the best sample clustering. In this figure, it is possible to distinguish the samples containing solvents "A" (naphtha), "B" (light naphtha), and "C" (thinner) clearly.

Moreover, Figures 3 and 4 confirmed that the chromatographic data really contains enough information to aggregate the samples according to the solvent added. However, for samples containing a low solvent concentration [less than 6% (v/v)], the results were not very good because the samples are more similar to each other.

3.3. LDA. *3.3.1. Fisher Weight and LDA.* To find the ideal number of variables to exclude, increasing Fisher weight values were used as "cut off" values, starting from the value that equals the number of variables and samples (75). Nevertheless, the matrix obtained was singular (making impossible to calculate the discriminant function). Table 4 presents the practical Fisher weight values associated with the corresponding number of variables used and the respective percentage of correctly classified samples by the model and cross-validation.

It can be seen that the classification quality reaches a maximum, for both model samples and cross-validation, when the Fisher weight is 2.625 and diminishes afterward. Accordingly, the 18 most important variables were used to construct a LDA classification model.

The 18 most important variables (peaks) are shown in Table 5. It can be seen that the important peaks are well-spread in the chromatogram; that is, they present important peaks all over the retention time range, which is in accordance with the characteristics of the solvent chromatographic profiles (Figure 1).

Table 6 presents the classification obtained for the test set when these equations were applied. The results show that the model performed quite well for the samples containing light naphtha and thinner, because samples containing concentrations as low as 2% (v/v) were correctly classified. On the other hand, for samples containing naphtha, the samples with a solvent concentration lower than 10% (v/v) were classified wrongly, and for samples containing kerosene, this concentration was even higher [15% (v/v)]. It can be said that naphtha and kerosene promoted subtler changes in the chromatographic profiles than light naphtha and thinner, at least considering the data selected. It can be seen that the wrong classifications were not random but tended to attribute samples to the class of samples containing light naphtha.

3.3.2. PCA and LDA. The ideal number of PCs is that which produces the best classification with the smallest number of PCs without overfitting. Table 7 presents the percentage of samples of the reduced training and test sets for various numbers of PCs as LDA variables correctly classified.

As described before in the literature,[25] the number of samples correctly classified in the training set tended to increase steadily as the number of PCs used in the model increased and the correctness in the classification of the test set tended to stabilize or even decrease after a certain number of PCs was reached.

Table 8. Most Important Substances for Each PC Based on the Loading

order of importance	PC1	PC2	PC3	PC4
1st	methylbenzene	methylciclohexene	1,2-dimethylbenzene	decane
2nd	1,2-dimethylbenzene	*n*-heptane	1,4-dimethylbenzene	1,2,3-trimethylbenzene
3rd	ethanol	cyclohexane	ethanol	NI[a]
4th	decane	hexane	1,2,3-trimethylbenzene	1,3,5-trimethylbenzene
5th	1,4-dimethylbenzene	methylcyclopentane	1,2-dimethylciclopropane	1-ethyl-3,5-dimethylbenzene
6th	nonane	isopropylcyclobutane	1-ethyl-2-methylbenzene	1,2,4-trimethylbenzene
7th	ethylbenzene	3-methylhexane	1-methylciclopentene	1-methyl-3-propylbenzene

[a] NI = compound not identified.

Table 9. Test Set Classification by PCA plus LDA

pure gasolines		gasoline plus solvent A		gasoline plus solvent B		gasoline plus solvent C		gasoline plus solvent D	
sample	attribution	sample	attribution	sample	attribution	sample	attribution	sample	attribution
G′1	G	G′1A-2	G	G′3B-2	G	G′1C-2	G	G′1D-2	G
G′2	G	G′1A-5	G	G′3B-5	G	G′1C-5	C	G′1D-5	G
G′3	G	G′1A-10	G	G′3B-10	B	G′1C-10	C	G′1D-10	D
		G′1A-15	A	G′3B-15	B	G′2C-15	C	G′1D-15	D
		G′1A-20	A	G′3B-20	B	G′2C-20	C	G′1D-20	D
		G′1A-25	A	G′3B-25	B	G′2C-25	C	G′1D-25	D
		G′1A-30	A	G′3B-30	B	G′2C-30	C	G′1D-30	D

Over this number, the model tends to overfit. Accordingly, the ideal number of PCs was found to be four, which, not coincidentally, was considered the dimensionality of the system in section 3.2.

Table 8 depicts an overview of the relative importance of the original parameters (chromatographic peaks) based on the loadings of the four most important PCs. The substances were identified by mass spectrometry as described in section 2.3.1.

A detailed evaluation of the chromatographic peaks allowed us to establish the relationship between PCs and solvents. More specifically, the first PC is dominated by loadings corresponding to the compounds predominant in naphtha. The second PC corresponds to light naphtha; the third PC corresponds to thinner; and the fourth PC corresponds to kerosene.

Equations 1–5 are the classification equations calculated with four PCs. Here, it becomes clear the importance of the fourth PC in the discrimination process, despite the risk of interpreting the coefficients given by these discriminating equations as discussed by Klecka.[24] A simple analysis of the variance of each of the variables in the matrix containing the training set revealed that the highest loadings for the fourth PC correspond to variables that contain relevant variability for the four groups of samples (corresponding to the four solvents). This is a clear demonstration of the power of LDA as a discriminant tool.

$$h_G = -90.27 + 2.44PC1 + 2.60PC2 + 3.59PC3 + 38.63PC4 \quad (1)$$

$$h_A = -104.88 + 1.72PC1 + 0.55PC2 + 0.70PC3 + 47.47PC4 \quad (2)$$

$$h_B = -48.90 + 1.90PC1 + 1.33PC2 + 2.87PC3 + 28.38PC4 \quad (3)$$

$$h_C = -96.44 + 2.20PC1 + 6.19PC2 + 0.74PC3 + 39.39PC4 \quad (4)$$

$$h_D = -104.41 + 4.51PC1 + 2.02PC2 + 2.17PC3 + 39.16PC4 \quad (5)$$

Table 9 presents the classification by applying these equations to the test set. In general, the classification was satisfactory for every class. For samples containing thinner, the model indicated the adulteration at 5% (v/v), and for naphtha, light naphtha, and kerosene, concentrations were 15, 10, and 10% (v/v), respectively. An interesting aspect of misclassification was that they were all attributed to the pure gasoline class. It seems quite reasonable, because the diluted samples were expected to be more similar to pure gasoline. Additionally, one can notice that the classification with the actual test set and the reduced one was not significantly different because neither was able to segregate samples with low solvent concentrations.

4. Conclusions

An investigation to develop an analytical methodology to identify the presence of solvents in gasoline based on chemometric analysis of chromatographic data was conducted. Simulated adulterations with four complex solvents were prepared. The use of mass spectrometry was important to assure that the variation in the concentration of the same compounds was evaluated for each sample.

PCA obtained from the covariance matrix showed that chromatographic data could contribute to differentiate samples according to the solvent present.

The LDA classification model obtained after the exclusion of less informative variables based on Fisher weights was capable of identifying the solvent added to gasoline even in low concentrations [2% (v/v)] of light naphtha and thinner, while for kerosene, this concentration was around 20% (v/v). Moreover, this model classified two pure gasoline samples out of three correctly. On the other hand, the use of the first four PCs obtained from the covariance matrix as LDA variables yielded a model capable of identifying the presence of solvents added to gasoline at 5% (v/v) for thinner and 15% (v/v) for naphtha. Additionally, the three pure gasoline samples were correctly classified.

It is worthwhile noting that, considering our experience in monitoring fuels, it is rare for the occurrence of adulteration in concentrations as low as 2% (v/v).

The PCA plus LDA classification model presented a trend toward classifying mixed samples with a lower amount of solvents in the test set as pure. Because the differences in

(23) Bruns, R. E.; Faigle, J. F. G. *Quimiometria*; Química Nova: Brazil, April, 1985.
(24) Klecka, W. R. *Discriminant Analysis*; Sage: Beverly Hills, 1980.
(25) Keemley, E. K. *Chemom. Intell. Lab. Syst.* **1996**, *33*, 47–61.

chromatograms between pure gasolines and their mixtures with low solvent concentrations are small, those misclassifications were expected. Considering the correct classification of the three pure gasolines in the test set, this tool is a strong candidate for substituting the current methodology used to identify the presence of solvents in gasoline.

Further attempts to improve the ability of the statistical models to classify pure samples are currently under way, mainly by extending the training set. The good results obtained after exclusion of irrelevant variables encouraged adaptations of the chromatographic elution program. The chromatographic program could be improved to emphasize only the important peaks and thus save time. Another improvement that can be considered if required is the automation of the process in such a way that the software that performs the chemometric calculus can access the data from the one that acquires the chromatograms.

Acknowledgment. The authors are grateful to the National Agency of Petroleum (ANP) for the financial support (CTPETRO), to CNPq and CAPES for fellowships, and to the Laboratório de Ensaio de Combustíveis (LEC) staff, without whom this work would not be possible.

EF0701337