Identification of Adulteration of Gasoline Applying Multivariate Data Analysis Techniques HCA and KNN in Chromatographic Data

ARTICLE *in* ENERGY & FUELS · OCTOBER 2005

Impact Factor: 2.79 · DOI: 10.1021/ef050031l

Identification of Adulteration of Gasoline Applying Multivariate Data Analysis Techniques HCA and KNN in Chromatographic Data

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, Depto de Química, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Campus Pampulha, CEP 31270-901, Belo Horizonte, Minas Gerais, Brazil, and Depto de Química, Instituto de Ciências Exatas, Universidade Federal do Espírito Santo, Av. Fernando Ferrari, s/n°, CEP 29060-900, Vitória, Espírito Santo, Brazil

Received January 29, 2005. Revised Manuscript Received September 15, 2005

Chemometric data analysis tools were applied to chromatographic data to identify the presence of solvents in gasoline samples from gas stations in Minas Gerais state, Brazil. A training set of 75 samples was formulated by mixing pure gasoline with various concentrations of four complex solvents. The samples were analyzed by GC-MS, and the selected peaks were used in chemometric studies. Hierarchical cluster analysis, HCA, was used to search for sample distribution patterns according to the solvent added. *K*-nearest neighbor (KNN) was used to create a classification scheme to differentiate pure and mixed samples and to indicate the type of solvent present. HCA revealed a clear clustering tendency of samples containing the same solvent. However, only after the exclusion of lesser variables (peaks) by means of Fisher weights was it possible to separate samples with low solvent concentrations. After optimization of the KNN algorithm, it was possible to classify 88% of the samples of the training set correctly. To check the quality of the model, another group of samples was prepared with certified gasoline and the same solvents. The algorithm classified the great majority of the samples correctly once again.

1. Introduction

In Brazil, oil refining and fuel distribution ceased being a state monopoly in 1995.[1] Since then, the roles played by state and private companies in the Brazilian fuel market have changed greatly and thousands of new distribution companies and gas stations have been opened. This market opening 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 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 developed and implemented a program determining 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 gasoline for the presence of markers. 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}. Since most solvents used in gasoline adulteration are petrochemical derivatives, the identification of their presence in gasoline is a challenging task. As the proper performance of gasoline fuel depends on a balanced combination of its compounds, a number of physicochemical tests are currently applied to evaluate its quality.[7] However, as these tests were not created to identify adulteration and as some solvents are quite similar to gasoline, they usually are not efficient in flagging adulteration.[6,8]

Gas chromatography, along with different detection methods, has been widely used to evaluate the quality of gasoline.[6,9−11] Moreira et al., for example, used GC with flame ionization and MS detection to evaluate

* To whom correspondence should be addressed: Tel.: +55 31 3499-6651. E-mail: skrobot@ufmg.br.
† Universidade Federal de Minas Gerais.
‡ Universidade Federal de Espírito Santo.

(1) Agência Nacional do Petróleo, Dois anos/ANP, Rio de Janeiro: A Agência 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: Betim, Brazil, 2000.
(5) Portaria Agência Nacional do Petróleo 274 of 11/01/2001.
(6) Moreira, L. S.; Ávila, L. A.; Azevedo, D. A. *Chromatographia* **2003**, *58*, 501−505.
(7) Portaria Agência Nacional do Petróleo 309 of 12/27/2001.
(8) Skrobot, V. L. Estudo quimiométrico dos perfis cromatográficos e propriedades físico-químicas determinadas por ensaios regulares de gasolinas e suas misturas com solventes. Dissertation, Universidade Federal de Minas Gerais, Minas Gerais, Brazil, 2004.
(9) Blomberg, J.; Schoenmakers, P. J.; Brinkman, U. A. *J. Chromatogr., A* **2002**, *972*, 137−173.
(10) Philp, R. P.; Mansuy, L. *Energy Fuels* **1997**, *11*, 749−760.

10.1021/ef050031l CCC: $30.25
Published on Web 10/07/2005

modifications resulting from the addition of some solvents to gasoline.[6] A more detailed profile of this fuel is obtained when a mass spectrometer (GC-MS) is used as detector. However, the richness of information makes the evaluation of the quality of gasoline by this technique extremely complex. Visual comparison of reference gasoline chromatograms to those of different gasoline samples is cumbersome and ineffective since changes in oil feedstock, refining processes, and aging cause modifications in the gasoline chromatographic profile that do not necessarily mean quality deterioration. The development of methodologies to evaluate gasoline quality must take into account these variations in composition.

A number of studies have demonstrated the usefulness of the application of chemometric techniques 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 and Pasquier[12] used principal component analysis (PCA) and linear discriminant analysis (LDA) to establish the origin of gasoline samples with three different grades on the basis of chromatographic data. The chemometric technique, LDA, was also employed to detect adulteration in whisky samples using chromatographic data.[13] Tan et al. used PCA and SIMCA to create classification models to identify petroleum-based accelerants in fire debris.[14] Kowalski and Bender have applied *K*-nearest neighbor (KNN) to nuclear magnetic resonance spectral data[15] to classify substances according to subtly different structural groups.

In the present work, a set of pure gasoline samples and their mixtures with four different solvents in various concentrations were analyzed by GC-MS, and the data obtained were submitted to two chemometric techniques, hierarchical cluster analysis (HCA) KNN. From GC-MS data, 89 chromatographic peaks were selected to make up a multivariate data set. Visual sample grouping was checked with HCA. After tests with distance measures between groups and linkage criteria, it was observed that the original data set could not be properly used without a pretreatment. Thus, aiming to obtain a better separation in the dendrogram, we used Fisher weights[16] as a criterion to exclude lesser variables (peaks). Then, the KNN algorithm was used to create a model able to identify the solvent present in gasoline in a new sample set. As far as we know, this is the first application of these chemometric techniques to chromatographic data of fuel.

2. Experimental Section

2.1. Samples. The samples were prepared from certified gasoline supplied by Petrobras (Petróleo Brasileiro SA, Minas Gerais, Brazil), gasoline collected in gas stations, and the 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, Minas Gerais, Brazil). Gasoline samples were analyzed and approved according to standard methods established by the Brazilian Government Petroleum Agency (ANP). These methods include alcohol quantification, anti-knocking properties (which were estimated by infrared spectrometry), and others according to international standards applied worldwide.[19,20]

Table 1. Composition of the Solvents[17,18]

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

Fifteen regular gasoline samples were collected from different gas stations belonging to eight different distributors over one month. Samples were transported in poly(ethylene terephthalate) flasks and in boxes containing ice to avoid volatilization and were analyzed and approved according to the same standard methods mentioned before.

To simulate as many adulteration conditions as possible, the gasoline samples collected were mixed with the four solvents in concentrations ranging from 2 to 30% in volume. These samples were labeled according to Table 2. In Brazil, regular gasoline has ethanol in its composition in concentrations varying from 13% (1990) to 26% in volume (1999) depending on sugar market prices, as ethanol and sugar are both obtained from sugar cane in Brazil. At present, its concentration is established as 25% in volume.[20] Ethanol concentration was corrected after the addition of the solvents with ethanol P. A. (Synth). This group of mixtures, called the *training set*, was used to create the classification model.

The *test set* group, 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 by KNN.

2.2. Chromatographic Analysis. General profiles of all samples were obtained using EI/MS. Analyses were conducted on automated GC-MS Shimadzu equipment model GC-17A/QP-5050A using a fused capillary column (50 m × 0.2 mm i.d. × 0.5 μm; PONA, HP) with poly(methylsiloxane) as the stationary phase and helium as the carrier gas at a constant flow rate of 0.4 mL min^{-1}. Sample aliquots of 0.5 μL were injected in 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 was operated in electron ionization mode at 70 eV, full-scan mode (45−350 *m/z*) with a sampling rate of 2 scans/s.

The individual compounds were identified by automatic comparison of the fragmentation patterns with the Wiley MS (Wiley Class 5000, sixth edition) database using the GCMSolution software.

(11) Sojak, L.; Addova, G.; Kubinec, R.; Kraus, A.; Bohac, A. *J. Chromatogr., A* **2004**, *1025* (2), 237−253.
(12) Sandercock, P. M. L.; Pasquier, E. *Forensic Sci. Int.* **2003**, *134*, 1−10.
(13) Saxberg, B. E.; Duewer, D. L.; Booker, J. L. *Anal. Chim. Acta* **1978**, *103*, 201−212.
(14) Tan, B.; Hardy, J. K.; Snavely, R. E. *Anal. Chim. Acta* **2000**, *422*, 37−46.
(15) Kowalski, B. R.; Bender, C. F. *Anal. Chem.* **1972**, *44* (8), 1405−1411.
(16) Bruns, R. E.; Faigle, J. F. G. *Quím. Nova* **1985**, *8*.
(17) Petrobras Distribuidora S. A. http://www.br.com.br/ (accessed in December 2004).
(18) Ipiranga Química. http://ipirangaquimica.ipiranga.com.br/ (accessed in June 2005).
(19) ASTM D 86—Distillation of petroleum products
(20) ASTM D 4052—Density and relative density of liquids by digital density meter.

Table 2. Preparation of the Training Set

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

[a] Example, G13B-26 corresponds to the mixture of pure gasoline 13 with solvent B at a concentration of 26% in volume.

Table 3. Preparation of the Test Set

certified gasoline	pure samples + solvent [(gasoline)(solvent)-(solvent concentration in % in volume)]			
	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

2.3. Data Analysis. *2.3.1. GC-MS.* Chromatographic peaks were normalized to unit area to minimize variations due to fluctuations in equipment response. Normalization consisted of dividing the area of each peak by the summation of the areas of every peak in the chromatogram. 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.

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 matrixes under study must be complete, that is, all samples must have values for all variables.

2.3.2. Chemometrics. 2.3.2.1. Fisher Weights. Fisher weights, W, allow the evaluation of how useful a variable is at discriminating 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 according to eq 1.

$$W_{pq}(i) = \frac{[\bar{X}_i(p) - \bar{X}_i(q)]^2}{S_i^2(p) + S_i^2(q)} \quad (1)$$

where the average, $\bar{X}$, and the variance, S^2, of variable i are evaluated for samples from groups p and q. The Fisher weight for variable i is the average for every group.

The Fisher weights for each variable were calculated using Microsoft Excel 2000 (Microsoft Corporation).

2.3.2.2. Hierarchical Cluster Analysis. HCA is a statistical method for finding relatively homogeneous sample clusters on the basis of measured characteristics (variables).[22] It starts with each sample in a separate cluster, and then, the clusters are combined sequentially, reducing their number at each step until only one cluster is left. Ergo, when there are N cases, this involves $N - 1$ clustering steps. The usual way to represent the clustering progression is by means of a dendrogram, which depicts the clustering steps and the distances between the samples or groups. Two criteria must be chosen to perform HCA, the distance measurement between samples or groups and the criterion to link samples and groups. Among the most popular distance measurements are Euclidean

$$d_{kl} = \sqrt{\sum_{j=1}^{J}(x_{kj} - x_{lj})^2} \quad (2)$$

and Manhattan distances

$$d_{kl} = \sum_{j=1}^{J}|x_{kj} - x_{lj}| \quad (3)$$

where the distance between objects k and l is evaluated in J dimensions.

The linkage criteria state how close the groups are. This information is used to decide how each pair of a group will be joined in each step of the formation of the dendrograms. The criteria most commonly used are simple, average, complete, and Ward. The simple criterion states that the distance between two groups is the smallest distance between an object from one group to an object in the other group. In the average criterion, the distance between groups is the average of all distances between objects from the groups. The complete criterion states that the distance between two groups is the maximum distance between an object from one group to an object in the other group. In the Ward criterion, the groups are linked in such way that each step of joining causes a minimum "loss of information"; this loss is defined in terms of the sum of squared errors (more detailed information can be found in ref 21).

(21) Santos, A. S.; Valle, M. L. M.; Giannini, R. G. *Econ. Energia* **2000**, *19*, 84−99.

(22) Sharma, S. *Applied Multivariate Techniques*, 1st ed.; John Willey & Sons Inc.: New York, 1996.

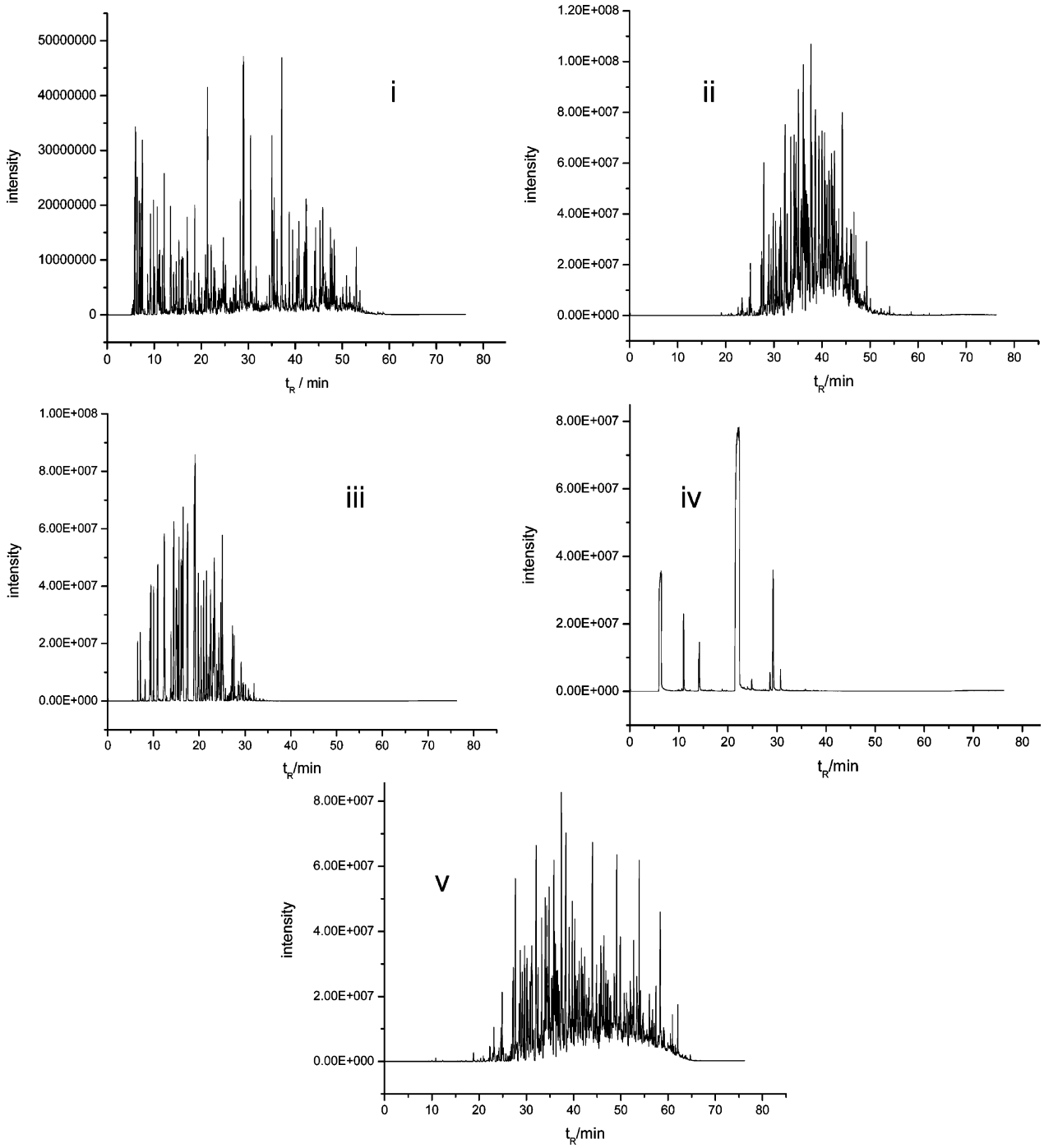

Figure 1. Chromatograms of reference gasoline (i) and solvents naphtha (ii), light naphtha (iii), thinner (iv), and kerosene (v) obtained as described in Section 2.2.

Table 4. Agglomerative Coefficients Obtained with Combinations of Distance Measurements and Linkage Criteria

	distance	
linkage	Euclidean	Manhattan
simple	0.600	0.475
average	0.911	0.867
complete	0.951	0.925
Ward	0.979	0.969

To select the best combination of distance measure and linkage criteria, dendrograms with every combination of these parameters were performed and the agglomerative coefficients,[24] AC, were calculated according to eq 4. This tool allows the evaluation of the clustering that accomplished the clearest segregation between groups of samples.

$$\mathrm{AC} = \frac{\sum(d_{\max} - d_{ij})}{N - 1} \tag{4}$$

where $d_{\max}$ is the distance between the farthest groups, d_{ij} is the distance between objects i and j, and N is the total number of samples. AC is close to 1 when a clear structuring has been found; when it is close to zero, the algorithm has not found a natural structure.

(23) Johnson, R. A.; Wichern, S. W. *Applied Multivariate Statistical Analysis*, 2nd ed.; Prentice Hall: New York, 1988.
(24) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data*, 1st ed.; John Willey & Sons Inc.: New York, 1990.

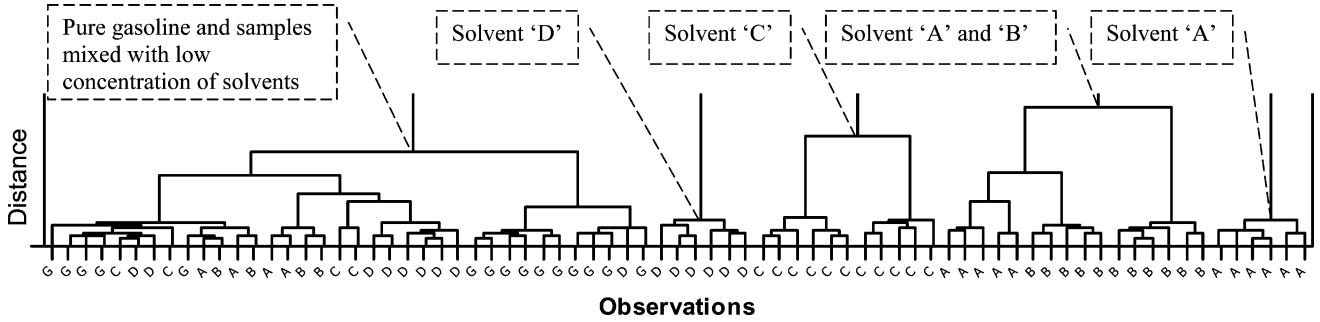

Figure 2. Partial dendrogram of gasoline samples and their mixtures with solvents based on 89 chromatographic peaks.

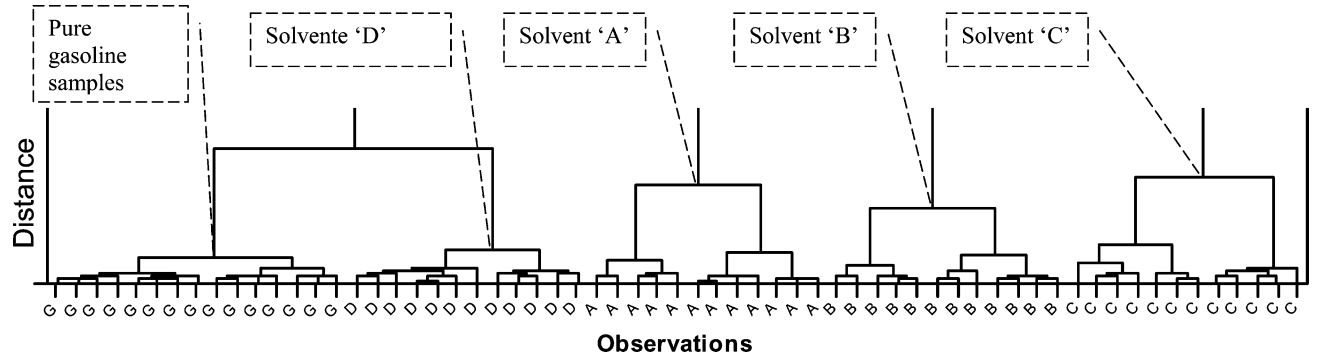

Figure 3. Partial dendrogram of gasoline samples and their mixtures (only 6% in volume or more) based on the 22 most important variables.

The hierarchical cluster analysis was performed with the software Minitab-14 (Minitab Inc.), and the agglomerative coefficients were calculated with S-Plus 4.5 (MathSoft, Inc.).

2.3.2.3. K-Nearest Neighbor (KNN). KNN is a nonparametric classification method that does not form a mathematical model based on the calibration data set. The "calibration model" consists of the calibration set and the criteria to measure distances between samples.[25−27] KNN attempts to categorize an unknown sample on the basis of its proximity to *K* samples already placed in categories (the training set).[25] The most common proximity measure is the Euclidean distance. More specifically, KNN categorizes the unknown sample in the class that contains more samples among the *K* samples. An important step in the creation of the KNN "model" is the choice of an appropriate *K* value. In this work, it was done by leave-one-out cross validation. Cross validation consists of taking part of the samples (one sample in the leave-one-out case) and creating the model on the basis of the rest of the samples using a chosen criterion. Next, this sample is classified by the model. This procedure is repeated until each sample is left out once. The number of samples correctly classified by the model is a measure of the quality of those criteria, and then, it can be compared to others.

KNN calculations were performed with Microsoft Excel 2000 (Microsoft Corporation).

3. Results and Discussion

3.1. GC-MS. Illustrating 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.

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 at between 20 and 50 min of the chromatographic run. Light naphtha has, in its composition, mainly compounds with high vapor pressures and low molecular weights. 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 pressures and high molecular weights as they come out after 30 min.

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

A total of 89 peaks from each sample were selected according to the criteria described in Section 2.3.1. The normalized areas of these peaks were used in the following statistical calculations. It is worthwhile to mention that the visual inspection of the chromatograms of each sample showed a considerable amount of noise, which stressed the need of using a large number of peaks to achieve a stable classification model.

3.2. HCA. According to Table 4, the highest agglomerative coefficient is that obtained by the combination of the Euclidean distance and the Ward linkage. As discussed before, it indicates that they give the best segregation among samples in the dendrogram. Moreover, the visual inspection of the dendrograms obtained with every combination of distance and linkage criteria confirmed that the above combination is the best. However, one can notice that a poor sample separation was obtained when all variables were used (89 peaks), as shown in Figure 2. In this figure, the samples were identified only by the letter corresponding to the solvent added and the pure gasoline samples were identified by

(25) Alsberg, B. K.; Goodacre, R.; Rowland, J. J.; Kell, D. B. *Anal. Chim. Acta* **1997**, *348*, 389−407.
(26) Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1972**, *16*, 5632−5639.
(27) Pirouette's help content. *Pirouette*, version 3.11; Infometrix, Inc.: Bothell, WA.

Table 5. Fisher Weights for Chromatographic Peaks

n	variable compound	Fisher weight	*n*	variable compound	Fisher weight
1	2-butene	0.44	46	1,4-dimethylcyclohexane	2.34
2	1-butene	0.47	47	1-ethyl-3-methylcyclopentane (cis/trans)	2.13
3	ethanol	1.24	48	1-ethyl-3-methylcyclopentane (cis/trans)	1.84
4	3-methylpentane	0.30	49	1-ethyl-2-methylcyclopentane	2.20
5	2-methyl-1-butene	0.65	50	1,2,3,4-tetramethylcyclobutene	1.63
6	1,2-dimethylcyclopropane (cis)	0.52	51	1,2-dimethylcyclohexene (trans)	2.53
7	3-ethyl-2,2-dimethylpentane	1.77	52	octane	1.75
8	1,1-dimethylcyclopropane	0.53	53	1,3-dimethylcyclohexane	1.93
9	2-methyl-1-butene	0.59	54	2,3,3-trimethyl-1,4-pentadiene	1.62
10	1,2-dimethylcyclopropane (trans)	0.44	55	2,7-dimethyloctane	1.02
11	cyclopentene	0.70	56	ethylcyclohexane	1.09
12	2-methyl-2-butene	0.88	57	1,1,2-trimethylcyclohexane	1.72
13	2-methylpentane	2.57	58	ethylbenzene	1.97
14	3-methylpentane	2.18	59	1,2-dimethylbenzene	2.79
15	2-methyl-1-pentene	1.10	60	3-methyloctane	1.81
16	hexane	2.96	61	1,4-dimethylbenzene	2.38
17	3-methyl-1-pentene	1.53	62	nonane	2.83
18	2-hexene	1.11	63	propylbenzene	1.93
19	2-methyl-2-pentene	0.95	64	1-ethyl-2-methylbenzene	1.11
20	2-hexene	0.99	65	1-ethyl-3-methylbenzene	3.54
21	3-methyl-2-pentene	1.10	66	1,2,4-trimethylbenzene	3.03
22	methylcyclopentane	2.44	67	1-ethyl-4-methylbenzene	2.83
23	1-methylcyclopentene	1.04	68	1,2,3-trimethylbenzene	0.98
24	benzene	1.77	69	decane	2.76
25	cyclohexane	1.94	70	1,3,5-trimethylbenzene	2.78
26	3-methylheptane	3.01	71	2,3-dihydro-1H-indane	1.47
27	2,3-dimethylpentane	2.36	72	1,2-diethylbenzene	1.76
28	3-methylhexane	2.61	73	1-methyl-3-propylbenzene	1.90
29	1,3-dimethylcyclopentane (trans)	2.59	74	1,3-diethylbenzene	1.48
30	1,3-dimethylcyclopentane (cis)	2.68	75	1-ethyl-3,5-dimethylbenzene	1.48
31	isopropylcyclobutane	2.31	76	4-ethyl-1,2-dimethylbenzene	2.02
32	heptane	2.86	77	1-methyl-4-(methylethyl)benzene	0.84
33	3-methyl-3-hexene	0.65	78	2-ethyl-1,4-dimethylbenzene	1.29
34	4,4-dimethylcyclopentene	1.39	79	1,2,4,5-tetramethylbenzene	1.41
35	methylcyclohexane	2.50	80	1,2,3,5-tetramethylbenzene	1.28
36	1,1,3-trimethylcyclopentane	2.08	81	2,3-dihydro-5-methyl-1H-indane	0.86
37	ethylcyclopentane	2.72	82	2,3-dihydro-4-methyl-1H-indane	0.98
38	2,4-dimethylhexane	0.36	83	1,2,3,4-tetramethylbenzene	2.84
39	1,2,4-trimethylcyclopentane	2.66	84	naphthalene	1.38
40	1,2,3-trimethylcyclopentane	1.81	85	2,3-dihydro-4,6-dimethyl-1H-indane	1.27
41	3-methyl-2,4-hexadiene	0.75	86	2,3-dihydro-1,6-dimethyl-1H-indane	1.95
42	methylbenzene	2.12	87	2,3-dihydro-4,7-dimethyl-1H-indane	3.20
43	3,5-dimethyloctane	3.32	88	1-methylnaphthalene	2.95
44	3-methylheptane	2.90	89	tridecane	1.60
45	1,2-dimethylcyclohexane (cis)	2.32			

Table 6. Percentage of Samples Correctly Classified, %CC, by KNN Cross Validation for $K = 1-7$

K	%CC	*K*	%CC
1	86.7	5	88.0
2	88.0	6	88.0
3	89.3	7	85.3
4	88.0		

the letter "G". This poor classification can be attributed to the great number of variables (peaks) that do not contribute to the differentiation of samples according to the solvent added. Nonetheless, a more detailed inspection of the dendrograms reveals that there is a remarkable difference between samples with high and low solvent concentrations. While the former are well-gathered in clusters, the latter tend to group with the pure gasoline samples.

Table 5 presents Fisher weights calculated for each of the 89 variables considered. The values show a big inhomogeneity in the importance of the variables. It was observed that many compounds presented low correlation in their variation with the group (solvent added), particularly the lighter ones. In an iterative process, variables with progressively higher Fisher weights were excluded and new dendrograms were obtained. The best sample separation was obtained when all samples with a Fisher weight lower than 2.50 were excluded (Figure 3). Specifically, all samples with a solvent concentration higher than 6% in volume were grouped in well-defined clusters in the dendrogram, and so were pure gasoline samples. It was concluded that, when the proper variables are selected, the chromatographic data, in fact, contain the necessary information to segregate gasoline samples according to the solvent added. Additionally, one can notice that the samples containing thinner were the farthest from the pure gasoline group, which indicates that this solvent is the most different from the others used, as confirmed by the chromatogram profile shown in Figure 1.

3.3. *K*-Nearest Neighbor. To determine the best K value (the number of neighbors), the Euclidian distances between each pair of samples were calculated using the 89 variables and the class of every sample was determined as if they were external samples (leave-one-out cross validation). The K values tested ranged from 1 to 7, which resulted in the percentage of corrected classified samples given in Table 6. For $K = 3$, the classification is given in Table 7. It can be noticed that only the samples with low solvent concentrations were misclassified. Moreover, samples containing 2% in volume of the solvent were all attributed to the pure gasoline group. Only samples with 6% in volume of the solvent or less were misclassified. Namely, when

Table 7. Classification of the Training Set Samples by KNN with $K = 3$

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

Table 8. Classification of the Test Set Samples by KNN with $K = 3$

pure gasoline		gasoline + solvent A		gasoline + solvent B		gasoline + solvent C		gasoline + solvent D	
sample	attribution	sample	attribution	sample	attribution	sample	attribution	sample	attribution
G′1	C	G′1A-2	C	G′3B-2	C	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	D
G′3	G	G′1A-10	B	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

$K = 3$, samples G1A-2, G2A-4, G3A-6, G1B-2, G2B-4, G1C-2, G1D-2, and G6D-12 were misclassified. Sample G6D-12 was misclassified even when more diluted samples were correctly classified, suggesting that it is an outlier.

The final "model" consisted of only correctly classified samples. Applying these criteria to classify the test set samples led to the results given in Table 8. The model misclassified samples containing naphtha at 15% in volume or less, while for samples with light naphtha and kerosene, the model failed only for the most diluted ones. Finally, every sample containing thinner was classified correctly.

Two out of three samples in the pure sample group were misclassified. It can be attributed to the fact that this is the group with the smallest variability and, therefore, with the biggest probability of misattributing samples as a result of minor variations in their characteristics.

4. Conclusions

An investigation has been conducted to develop an analytical methodology based on chemometric analyses of chromatographic data to identify the presence of solvents in gasoline. Adulterations were simulated with four different solvents. Mass spectrometry was used to identify the composition of each solvent and gasoline. This technique was important to determine variations in concentration of the same compounds from sample to sample.

HCA has shown that chromatographic data do contain information that enables the differentiation of samples according to the solvent present, mainly if unimportant variables (peaks) are excluded.

A KNN model was developed using Euclidian distances, and the best K value was found to be 3. Applying the model to the test set, 77% of the samples were correctly classified. The model failed to classify only the samples containing low solvent concentrations and did not perform well with pure gasoline samples, classifying 2 out of 3 samples incorrectly. In the best case, every test set sample containing thinner was classified correctly, and in the worst case, samples containing 10% in volume or less of naphtha were misclassified.

Further investigations to improve the model's ability to classify pure samples are underway. Namely, chemometric techniques are being tested to overcome the problem of heteroscedasticity between classes in the training set. For now, the KNN classification must be seen as suspect whenever it indicates a sample as pure.

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

EF050031L