^{1}H NMR Fingerprinting of Brazilian Commercial Gasoline: Pattern-Recognition Analyses for Origin Authentication Purposes

Tainara Rodrigues Maia Rigo,[†,‡] Danilo Luiz Flumignan,[†,‡] Nivaldo Boralle,[‡] and José Eduardo de Oliveira*[,†,‡]

Center for Monitoring and Research of the Quality of Fuels, Biofuels, Crude Oil and Derivatives, CEMPEQC, and Organic Chemistry Department, Institute of Chemistry, São Paulo State University, UNESP, Rua Prof. Francisco Degni s/n, Quitandinha, 14800-900, Araraquara, SP, Brazil

Received December 15, 2008. Revised Manuscript Received June 15, 2009

In this work, the combination of hydrogen nuclear magnetic resonance (^{1}H NMR) fingerprinting of gasoline with pattern-recognition analyses provides an approach to distinguish Brazilian commercial gasoline, processed in different states of Brazil. Hierarchical cluster analyses (HCA) and principal component analyses (PCA) were carried out on chemical shifts in order to observe any natural grouping feature, while soft independent modeling of class analogy (SIMCA) was performed to classify external samples into previously origin-defined classes. PCA demonstrated that a small number of variables dominate the total data variability since the first three principal components (PCs) accounted for 64.9% of total variability; whereas a HCA dendrogram shows five natural cluster grouping features. Following optimized ^{1}H NMR-SIMCA algorithm, sensitivity values in the training set with leave-one-out cross-validation (86.0%) and external prediction set (77.3%) were obtained. Governmental laboratories could employ this method as a rapid screening analysis for origin authentication related to tax evasion purposes.

1. Introduction

In the last years, the Brazilian government's recent suspension of the state monopoly of fuel production and distribution has given rise to significant changes in the fuel market in Brazil. This event has opened up enormous opportunities both for established oil companies and for new comers to this market. Moreover, the removal of barriers in the retail sector by allowing "white flag" service stations (i.e., those not operating under the trademark of a particular distributor) and the liberalizing of resale margins have produced a significant increase in the number of fuel dealers and gas stations operated by national and foreign companies. The ensuing stronger competition for retail business has led to a substantial increase in the variation of the price of fuel, while the quality of the product has not necessarily been guaranteed.[1]

The Brazilian fuel market commercializes four types of automotive fuels: gasoline [with a mixture of 20−25% (v/v) of anhydrous ethanol], diesel oil, natural gas, and hydrated alcohol. In Brazil, refineries and petrochemical plants produce gasoline and sell it to distributors, who in turn add anhydrous ethanol and sell it to gas stations. Furthermore, each state has its own particular taxation but allows gas stations to buy fuel from any dealer. This means that gasoline can be produced in one specific refinery and then be commercialized in various geographical regions of Brazil. In internal state operations, all collected taxes are maintained in the respective state. In interstate operations, the taxes are greater in southeastern and south states than in North, Northeast, and Center-West states; in this case, the tax is collected in the producing state, but the taxes differences between interstate and internal operations are collected by consuming/importing states. Hence, the tax differences not collected configure fiscal evasion, a situation that needs to be controlled.[2]

Because of the necessity to supply the consumption of all regions, the national refinery park, constituted by 12 refineries, was strategically distributed throughout Brazil (1 in the northern region, 1 in northeastern region, 3 in southern region, and 7 in southeastern region - Figure 1).[3] The southeast region (color yellow), having the greatest population, has the great majority of the refineries, especially São Paulo state, which has a total of four refineries. On the other hand, the Center-West region (color orange) does not have a refinery and the southeast refineries supply its needs, especially São Paulo state refineries, due its territorial proximity and large production. A preliminary study concerning the origin of the gasoline based on information found on the invoices of 72 samples randomly collected for the Brazilian National Petroleum Agency Monitoring Program in all over Brazilian regions was carried out. Strategically, the collected sample (n) was proportional and representative of the geographic distribution of the Brazilian refineries ($n_{\text{south}} = 8$; $n_{\text{center-west}} = 10$; $n_{\text{north}} = 15$; $n_{\text{northeast}} = 18$; and $n_{\text{southeast}} = 21$).

* To whom correspondence should be addressed. Phone: +55 16 3301-6666. Fax: +55 16 3301-6693. E-mail: jeduardo.unesp@yahoo.com.br.
† CEMPEQC.
‡ Organic Chemistry Department, Institute of Chemistry.

(1) Flumignan, D. L.; Ferreira, F. O.; Tininis, A. G.; de Oliveira, J. E. *Anal. Chim. Acta* **2007**, *595*, 128–135.
(2) Vainsencher, A. Redução de ICMS impulsiona vendas, Valor Econômico [Online]; São Paulo, Feb. 16, 2004; Available at http://infoener.iee.usp.br/infoener/hemeroteca/imagens/77482.htm (accessed March, 2009).
(3) Adriana Costa Soares. Dissertação: Diagnóstico e modelagem da rede de distribuição de derivados de petróleo no Brasil; PUC-Rio, 2003.
(4) Berrueta, L. A.; Alonso-Salces, R. M.; H'eberger, K. *J. Chromatogr. A* **2007**, *1158*, 196–214.
(5) *Pirouette 3.11—Multivariate Data Analysis for IBM PC Systems*; Infometrix Co.: Woodinville, WA, 2003. http://www.infometrix.com.
(6) Balabin, R. M.; Safieva, R. Z. *Fuel* **2008**, *87*, 1096–1101.
(7) Culp, R. A.; Noakes, J. E. *J. Agric. Food Chem.* **1992**, *40*, 1892–1897.

10.1021/ef8010977 CCC: $40.75
Published on Web 07/13/2009

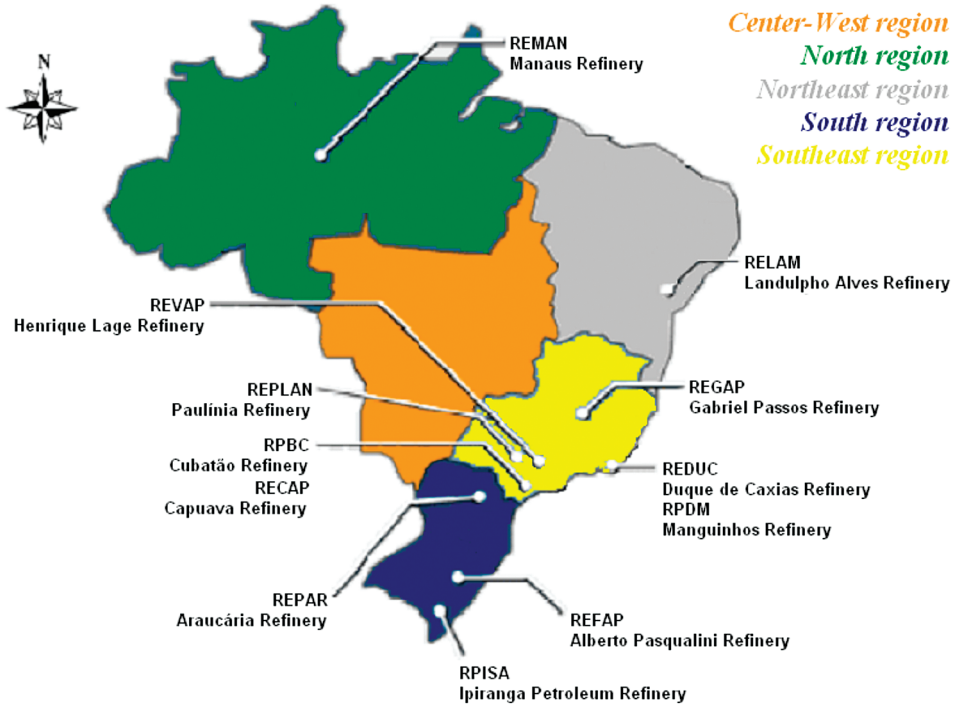

Figure 1. Spatial geographical distribution of the Brazilian refineries and their respective acronyms, as well as the region locations and the representative amount of gasoline samples randomly collected: North (green color; 15 samples), Northeast (gray color; 18 samples), Center-West (orange color; 10 samples), South (blue color; 8 samples), and Southeast (yellow color; 21 samples).

Nowadays, modern analytical instruments allow producing great amounts of information (variables or features) for a large number of samples (objects) that can be analyzed in relatively short time.[4,5] Exploratory analyses or unsupervised pattern recognition techniques, such as hierarchical cluster analysis (HCA) and principal component analysis (PCA), provide a quality check on the data to determine its information content and pinpoint key measurements. It is commonly used to simplify and gain better knowledge of data sets and also establish its viability with regard to classification model-building.[4,5] On the other hand, supervised pattern recognition techniques, such as soft independent modeling of class analogy (SIMCA), aims to establish a classification model based on experimental data in order to assign unknown samples to a previously defined sample class based on its pattern of measured features. This technique requires a training set with objects of known categories to derive a model for the identification of unknown samples.[4,5] Both techniques, unsupervised and supervised, have been applied to a wide variety of chemical data (chromatographic, spectroscopic, etc.) with diverse purposes (fingerprinting, authentication, detection of adulteration, quality control, data interpretation, etc.), in special, in gasoline, diesel oil, and jet fuel.[6−19] In regard to Brazilian commercial gasoline, many authors have studied its adulteration with the addition of petrochemical solvents.[10−17]

Table 1. Summary of Classification of the Training Set Samples by Leave-one-out Cross-validation in the SIMCA Model

	training set by leave-one-out cross-validation					
actual	North	Center-West	Southeast	Northeast	South	total samples
North	10					10
Center-West		7				7
Southeast		2	10	3		15
Northeast			2	10		12
South					6	6
correctly classified samples	10	7	10	10	6	**43/50**
sensitivity (%)			**86.0**			

Balabin and Safieva[6] examined different approaches to classify Russian gasolines according to their source (refinery or process) and type, based on near-infrared spectroscopy data. Gasoline identification by source is an important factor for both quality control and identification of gasoline adulteration, while gasoline type is needed for classification of gasolines by quality and price. All approaches were effective for gasoline classification purposes. Barbeira et al.[10] used histograms and normal distribution curves in association with physicochemical parameters to detect Minas Gerais state commercial gasoline having dissimilar compositions. This atypical behavior may have two possible causes: samples from different origins (e.g., different refineries) or careful adulteration from the controlled addition

(8) Suri, S. K.; Prasad, K.; Ahluwalia, J. C.; Rogers, D. W. *Talanta* **1981**, *28*, 281–286.
(9) Dhole, V. R.; Ghosal, G. K. *J. Liq. Chromatogr.* **1995**, *18*, 2475–2488.
(10) Barbeira, P. J. S.; Pereira, R. C. C.; Corgozinho, C. N. C. *Energy Fuels* **2007**, *21*, 2212–2215.
(11) Aleme, H. G.; Costa, L. M.; Barbeira, P. J. S. *Fuel* **2008**, *87*, 3664–3668.
(12) de Oliveira, F. S.; Teixeira, S. G.; Araújo, M. C. U.; Korn, M. *Fuel* **2004**, *83*, 917–923.
(13) Moreira, L. S.; d'Avila, L. A.; Azevedo, D. D. *Chromatographia* **2003**, *58*, 501–505.
(14) Wiedemann, L. S. M.; d'Avila, L. A.; Azevedo, D. D. *Fuel* **2005**, *84*, 467–473.
(15) Flumignan, D. L.; Anaia, G. C.; Ferreira, F. O.; Tininis, A. G.; de Oliveira, J. E. *Chromatographia* **2007**, *65*, 617–623.
(16) Flumignan, D. L.; Ferreira, F. O.; Tininis, A. G.; de Oliveira, J. E. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 53–60.
(17) Monteiro, M. R.; Ambrozin, A. R. P.; Lião, L. M.; Boffo, E. F.; Tavares, L. A.; Ferreira, M. M. C.; Ferreira, A. G. *Energy Fuels* **2009**, *23* (1), 272–279.
(18) Johnson, K. J.; Wright, B. W.; Jarman, K. H.; Synovec, R. E. *J. Chromatogr. A* **2003**, *996*, 141–155.
(19) Johnson, K. J.; Synovec, R. E. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 225–237.

Table 2. Summary of Classification of the External Prediction Set Samples in the SIMCA Model

	External Prediction Set Samples					
actual	North	Center-West	Southeast	Northeast	South	Total Samples
North	4	1				5
Center-West		2	1			3
Southeast			4	2		6
Northeast			1	5		6
South					2	2
correctly classified samples	**4**	**2**	**4**	**5**	**2**	**17/22**
sensitivity (%)			**77.3**			

of solvents. Within this way, Aleme et al.[11] employed unsupervised and supervised pattern-recognition to determine the refinery origin of several atypical gasoline samples commercialized in Minas Gerais state based on its distillation curve. Recently, our lab successfully employed pattern-recognition analyses to screen the quality of commercial Brazilian gasoline via chromatographic fingerprints.[15,16] Monteiro et al.[17] also describes the usefulness application of pattern-recognition analysis in NMR spectra as a screening method to determine quality control of Brazilian commercial gasoline.

Likewise, the importance of developing alternative approaches to origin authentication purposes of Brazilian commercial gasoline processed in different states of Brazil becomes essential. So, this preliminary work aims to make efficient use of pattern-recognition analyses to identify the origin of gasoline commercialized in gas stations all over Brazil, especially where this information is not available because the retailer does not have the purchase invoice.

2. Experimental Section

2.1. Sample Collection. The gasoline samples used in this work, all in agreement with the Brazilian Government Petroleum, Natural Gas and Biofuels Agency − ANP Regulation 309, were provided by the National Petroleum Agency Monitoring Program Central Laboratory for the Quality Control of Automotive Fuels (CPT), in particular, gasoline, ethanol, and diesel oil. Thus, our sampling universe does not belong to a well-defined distribution. In fact, the distribution of all possible samples in the sampling universe is unknown. Since the samples' distribution is unknown, the natural choice would be to use random sampling.

Therefore, 72 commercial gasoline samples were randomly collected, directly from gas stations, in 1 L polyethylene terephthalate (PET) amber flasks with sealing caps obtained from all of the geographical regions in Brazil. Strategically, the amount of collected sample (n) was proportional and representative of the geographic distribution of the Brazilian refineries ($n_{south} = 8$; $n_{center\text{-}west} = 10$; $n_{north} = 15$; $n_{northeast} = 18$; and $n_{southeast} = 21$, Figure 1). Samples were transported to the lab below 10 °C in refrigerated boxes following official ANP procedures.[20,21] When arriving at the lab, 90 mL samples were immediately collected in 100 mL amber PET flasks with sealing caps. These samples were stored in a freezer at temperatures below 0 °C to avoid volatilization and to keep their integrity until analysis.

2.2. Hydrogen Nuclear Magnetic Resonance (^{1}H NMR) Analyses. All ^{1}H NMR spectroscopic fingerprintings were acquired on a Varian INOVA 500 MHz (Palo Alto, CA, USA) instrument for proton observation using a 5 mm single cell ^{1}H/^{13}C inverse detection flow probe. For each sample, 30 μL of gasoline were diluted in 600 μL of deuterated chloroform ($CDCl_3$). ^{1}H NMR spectra were registered at a temperature of 300 K using 64 000 data points with standard pulse (s2pul) for proton sequence. Transients (32) were accumulated over a 4725 Hz spectral width with a relaxation frequency pulse (4.10 μs) and recycle delay (0.904 ms). Thirty-two scans were accumulated for each spectrum. The spectral profiles of gasoline were acquired in 2 min under the experimental conditions. The FIDs were zero filled and Fourier transformed. The phase and baseline were automatically corrected in all spectra. ^{1}H NMR chemical shifts are reported in parts per million (ppm) relative to residual proton signals of $CDCl_3$ at 7.24 ppm, which was used as reference chemical shift. Additionally, fingerprinting spectra were normalized to 1-norm (the area under the sample profile is set equal to one) for compensation of baseline distortions and bucket-width integrated (0.02 ppm) for more effective compensation of peak-shifts. At last, such spectra profile was exported as ASCII files and transferred to a PC for data analysis.

2.3. Pattern-recognition Analyses. Pattern-recognition analyses were used to evaluate the possibility of differentiating the commercial gasoline samples according to five geographical Brazilian regions, based on the value of (some of) the 26 713 variables (chemical shifts) acquired in the ^{1}H NMR analysis. So, multivariate analyses were applied to the resulting data matrix (26 713 × 72; chemical shifts × gasoline samples) using Pirouette software version 3.11 (Infometrix Inc., Tulsa, OK, USA).[22] Each line in the matrix constitutes a sample, and the columns represent the number values obtained from the chemical shifts and intensities of the peaks. Pattern-recognition analyses were applied to whole data set for exploring the data, feature selection, eliminate noises, and avoid overfitting. The ethanol peaks (CH_2 in the 3.7−3.6 ppm range and CH_3 in 1.2−1.1 ppm) were also selected for the construction of the data set.

HCA and PCA, unsupervised pattern-recognition analysis, provides a quality check on the data to determine its information content and pinpoint key measurements, giving complementary information about the similarities and groupings of the samples considered. If a trend exists, it is worthwhile evaluating the possibility of classifying the samples. In conjunction, SIMCA, a well-known supervised pattern-recognition method, constructs models using samples preassigned to a category, that is, in this case, it is a geographical-region discrimination approach.[1,4,5,23] SIMCA is the most used of the class-modeling techniques. In SIMCA, each category is independently modeled using PCA and can be described by a different number of principal components. The number of principal components for each class in the training set is determined by cross-validation.[4]

Several pretreatments and preprocessing were applied to the entire data matrix. In HCA, Euclidean distances among samples were calculated and transformed into similarity indices ranging from 0 to 1 using the incremental linkage method. In this criterion the groups are linked causing a minimum "loss of information".[5,23] In PCA and SIMCA, the data set was autoscale preprocessed, where each variable is meancentered and scaled to unity variance, to give each variable equal weight, and therefore, large and small peaks were treated with equal emphasis. Additionally, the data set was leave-one-out cross-validated processed and 95% probability threshold applied (as confidence level).

Finally, in SIMCA, the data set was divided into two sets: training set (50 gasoline samples) and external prediction set (22 gasoline samples). The training set provides spectral information that was used to determine principal components boundaries that characterize the samples into Southeast, Northeast, North, Center-West, or South geographical-regions classes. The external prediction set was randomly and proportionally selected, in which the distribution ended up as 6 samples from Southeast, 6 from Northeast, 5 from North, 3 from Center-West, and 2 from South. SIMCA performance was evaluated by sensitivity, calculated by the ratio of the sum of the diagonal values in the actual versus predicted portion to the sum of all of the actual rows. Another approach considers that each

(20) American Society for Testing and Materials. *Annual Book of ASTM Standards*; ASTM: Philadelphia, PA, 1995; Vol*05.02*, D4057.

(21) Associação Brasileira de Normas Técnicas. *Normas Técnicas Brasileiras, ONS34:000.02−012: 2002*; ABNT: Rio de Janeiro, RJ, 2002; No. 14883.

class is bounded by a region of space, which represents a percentage of confidence level (usually 95%) that a particular object belongs to a class.

3. Results and Discussion

3.1. ^{1}H NMR Fingerprinting Profiles of the Gasoline Samples. In Brazil, there are five geographical regions (Figure 1): the South, Center-West, North, Northeast, and Southeast regions. Each region has its own refinery park, which is strategically located to supply the local demand. On the other hand, each refinery has its own typical refining process. In this sense, Barbeira and co-workers[10] used PCA, HCA, and LDA techniques associated with physicochemical parameters to detect Minas Gerais state (a restricted sampling area located in Southwest geographical region) commercial gasoline having dissimilar compositions. This atypical behavior observed was attributed to two possible causes: samples from different origins (e.g., different refineries - Rlam, Regap, Replan, Revap, and Reduc) or careful adulteration from the controlled addition of solvents. In another work,[11] the same authors successfully employ pattern-recognition techniques to determine the refinery origin of these atypical gasoline samples commercialized in Minas Gerais state (again a restricted sampling area) based on its distillation curve.

In our case, the gasoline samples collected in gas stations from all the Brazilian geographical regions were in agreement to the patterns established by the Brazilian Government Petroleum, Natural Gas and Biofuels Agency − ANP Regulation 309, consequently the differences between them could not be observed or discriminated via those physicochemical assays. Our expertise in NMR spectrometry encouraged us to develop a work aiming the combination of ^{1}H NMR fingerprinting of gasoline associated with pattern-recognition analyses for origin authentication purposes. The ^{1}H NMR spectrum of gasoline is very complex, showing peaks almost in all spectral regions. Figure 2a presents typical NMR spectra of Brazilian gasoline and illustrates all chemical shifts, which were used in the data set. In general, classes of compounds (not individual ones) are associated with specific spectral regions. For example, aromatic compounds can be associated with peaks at 6.7−8.0, 3.7−3.6, and 1.2−1.1 ppm to ethanol peaks, and the 0.5−2.0 ppm region contains signals mainly due to cycloalkanes (naphthenes) and normal- and isoparaffins. As can be seen in Figure 2a, NMR fingerprinting spectra were very similar to the naked eye because the basic refinery processes are quite similar.[24] Clearly, a simple visual inspection of spectral profiles of commercial gasoline proceeding from different geographic regions (Figure 2b−f) is not enough to distinguish them. Therefore, multivariate statistical approach is a very useful tool and is often employed for gasoline discrimination.

3.2. Patter-recognition Analyses. In these analyses, we chose the entire ^{1}H NMR spectrum (except noises and $CDCl_3$ region) for the statistical analysis, because the choice of a large number of peaks allows us to achieve a more reliable classification models. Initially, all spectral data were converted into a data matrix with the 72 gasoline samples (n_{south} = 8; $n_{center\text{-}west}$ = 10; n_{north} = 15; $n_{northeast}$ = 18; and $n_{southeast}$ = 21) in the rows and the chemical shifts (26 713 variables) in the columns.

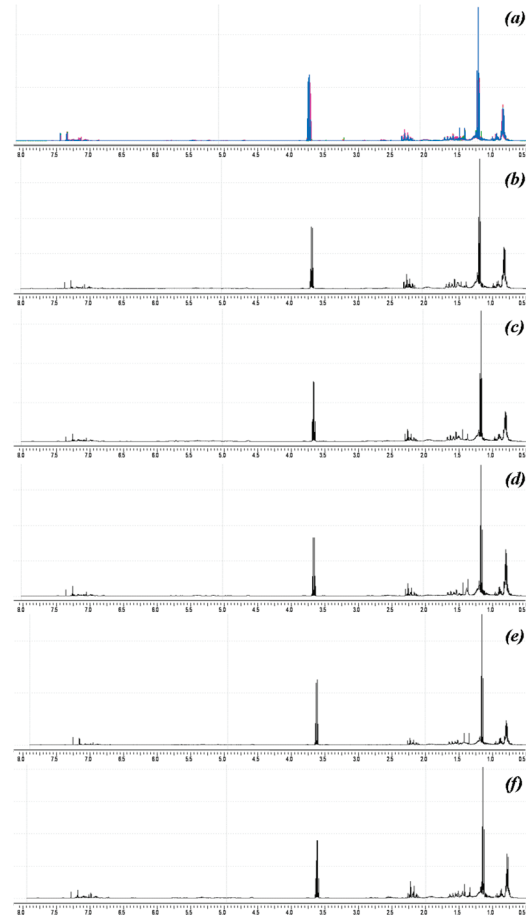

Figure 2. Typical NMR spectra of all Brazilian commercial gasoline samples (a) and the same proceeding from different geographical regions: (b) Southeast, (c) Northeast, (d) North, (e) Center-West, and (f) South.

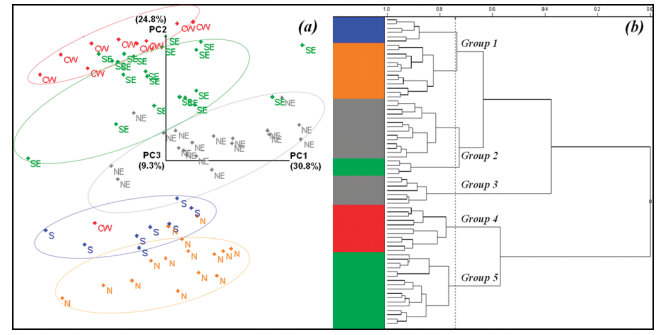

Figure 3. Unsupervised pattern-recognition analysis to evaluate the cluster-forming tendency of the Brazilian commercial gasoline samples from different geographical regions: (a) PCA 3D scores plot, using autoscale and logarithmical transformed data; (b) HCA dendrogram, using Euclidean distance and incremental linkage criteria. Note: North (N), Center-West (CW), Southeast (SE), Northeast (NE), and South (S).

In HCA and PCA analyses, all gasoline samples (72 samples) were used to construct the models to facilitate the visualization of the cluster tendencies in the graphs (Figure 3). However, in SIMCA analysis, five geographical regions (as class, Figure 4) were established, and 50 gasoline samples were used to compose the training set and 22 samples were used in the prediction set. Moreover, best pattern-recognition analyses were obtained using autoscale preprocessing and logarithmic pretreatment applied

(22) *Pirouette 3.11*, Register Software; Infometrix Inc.: Woodinville, WA, 2003; available at http://www.infometrix.com.

(23) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: a textbook*; Elsevier: Amsterdam, NE, 1988; pp. 319−412.

(24) Andre Vanzelote Barquette. Dissertação: Avaliação da melhor Localização do Sistema de Mistura em Linha de Diesel da REDUC; PUC-Rio, 2008.

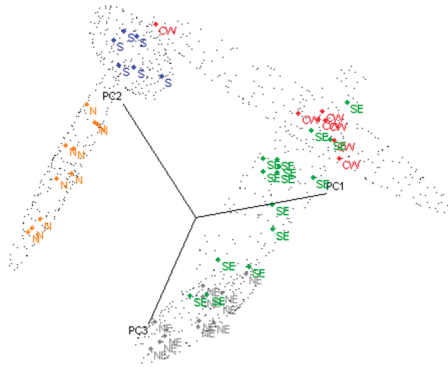

Figure 4. SIMCA-class 3D projections of samples in the training set on score plots. Note: CW (red points) represents Center-West samples, S (blue points) represents South samples, N (orange points) represents North samples, SE (green points) represents Southeast samples, and NE (gray points) represents Northeast samples.

to the ^{1}H NMR fingerprinting profiles. Autoscaling, in which each variable is meancentered and scaled to unity variance, was applied to give each variable equal weight for all spectral regions, and therefore, large and small peaks were treated with equal emphasis. The ethanol peaks were included in the data for the construction of the matrix.

3.2.1. HCA Analysis. In HCA analysis, the Euclidean distance was used as a metric, and an incremental linkage method was employed in the best separate clusters tendency. In the incremental criterion, the groups are linked in such way that each step of joining causes a minimum "loss of information". Moreover, the incremental link works better than other methods in instances where two groups of samples differ only slightly, which is the case of this study. Each group, illustrated in the dendrogram (Figure 3b), was constituted of similar samples in relation to their spectral data. With a similarity index, approximately 0.75, the samples were discriminated in accordance to their geographical region. This dendrogram was useful to obtain preselected profiles of high similarity. Figure 3b shows five main clusters, which represent five distinct groups (South: blue color bar; North: orange color bar; Northeast: gray color bar; Southeast: green color bar; and Center-West: red color bar), separated for the high similarity of their ^{1}H NMR fingerprinting profiles. As can be seen: Group 1 (19 samples), clearly classified two nodes constituted by South (blue color bar, 6 samples) and North (orange color bar, 13 samples) representatives; Group 2 (18 samples), constituted predominantly by Northeast representatives (gray color bar, 14 samples), except for a few Southeast representatives (green color bar, 4 samples); and Groups 3, 4, and 5, which were constituted only by Northeast representatives (gray color bar, 7 samples), Center-West representatives (red color bar, 11 samples), and Southeast representatives (green color bar, 17 samples), respectively.

3.2.2. PCA Analysis. Similarly to HCA, PCA analysis allowed the distinction between geographical origin groups. The best clustering was obtained using autoscale and logarithmical transformed data. The PCA scores plot (Figure 3a), obtained from the first three principal components (PC1, PC2, and PC3), indicated similarity among the samples; similar samples tended to form clusters. In these 3D score plot (Figure 3a), PC1 described 30.8% of the total variance, while PC2 described 24.8%, and PC3 described 9.3%; therefore, the three PCs together express 64.9% of the original information. In this way, five clear clusters, obtained from ^{1}H NMR fingerprinting profiles of all commercial gasoline, can be observed (Figure 3a): Center-West (CW, red points), Southeast (SE, green points), Northeast (NE, gray points), South (S, blue points), and North (N, orange points). Moreover, the Southeast cluster is not homogeneous, because the high number of refineries and slightly overlapping between (a) CW and SE samples and (b) S and N samples. These overlapping can also be interpreted in two ways: (1) the inexistence of a refinery in the Center-West region, where demand is supplied by the Southeastern region, due its territorial neighborhood and large production; and (2) the refining processes of North and South refineries being similar. Figure 2 shows that both intensity and peaks chemical shifts, especially those related to aromatics, cycloalkanes (naphthenes), and normal- and isoparaffins compounds, allowed the discrimination between geographical regions.

3.2.3. SIMCA Method. To carry out the SIMCA algorithm, all samples were divided into two sets: training set (50 samples) and external prediction set (22 samples). The external prediction set was randomly and proportionally selected. Additionally, the data set was leave-one-out cross-validated processed. Another approach considers that each class is bounded by a region of space, which represents a percentage of confidence level (usually 95%) that a particular object belongs to a class. Figure 4 shows the score points for five categories and the corresponding confidence intervals (represented by the single pixel black spots). The coordinates of a bounding ellipse (based on the standard deviations of the scores in each PC direction) for each category are projected into this SIMCA-class 3D projections (Figure 4); they form a confidence interval for the distribution of the category. In these SIMCA-class 3D projections (Figure 4), the first three principal components (PC1, PC2, and PC3) describe 66.5% of the total within-set variance, similar to the PCA method, and also gave (as determined by visual inspection) very good segregation between the samples classes. Rotation visualization of SIMCA-class 3D projections plot (Figure 4) reveals light overlapping of categories as well as training set samples lying beyond the confidence boundary of the corresponding class model.

So, four zones are defined on the 3D plot: North class (N, orange points), South class (S, blue points), Southeast class (SE, green points), and Northeast class (NE, gray points). In particular, the Southeast class is not homogeneous, probably because its high number of refineries. Also a slightly overlapping between Center-West (CW, red points) and Northeast (NE, gray points) classes with Southeast ones are shown. After leave-one-out cross-validation processed in training set, SIMCA performance reveals high sensitivity (86.0%, Table 1) using autoscale and logarithmical transformed data. Lastly, when extending this performance evaluation to the entire external prediction set, results showed a slightly decrease sensitivity (77.3%, Table 2), which is, however, uncritical.

4. Conclusion

The preliminary results shown in this study indicate that ^{1}H NMR spectroscopy coupled to pattern-recognition analyses, such as HCA, PCA, and SIMCA, is an appropriate technique to distinguish Brazilian commercial gasoline, processed in different states of Brazil. Therefore, the ^{1}H NMR-HCA, ^{1}H NMR-PCA, and ^{1}H NMR-SIMCA models are quite useful techniques for authentication purposes. Particularly, the ^{1}H NMR-SIMCA algorithm sensitivity values in the training set with leave-one-out cross-validation (86.0%) and in the external prediction set (77.3%) were obtained. This

correct authentication may allow reliable quality control by government laboratories in order to discourage state tax evasion.

Acknowledgment. Part of this manuscript was T. R. M. R.'s diploma work. The authors wish to thank the Agência Nacional do Petróleo, Gás Natural e Biocombustíveis - ANP and FUNDUNESP for financial support; CAPES, CNPq, and FUNDUNESP for the provision of scholarships; CEMPEQC (Centro de Monitoramento e Pesquisa na Qualidade de Combustíveis, Biocombustíveis, Petróleo e Derivados - Center for Monitoring and Research of the Quality of Fuels, Biofuels, Crude Oil and Derivatives) for offering its infrastructure and its staff for the development of this work; and CPT-ANP (Centro de Pesquisas e Análises Tecnológicas - Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) for providing the gasoline samples.

EF8010977