Talanta 84 (2011) 963–970

Contents lists available at ScienceDirect

Talanta

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

Classification of gasoline by octane number and light gas condensate fractions by origin with using dielectric or gas-chromatographic data and chemometrics tools

Vasiliy A. Rudnev[a], Alexander P. Boichenko[b,*], Pavel V. Karnozhytskiy[a]

[a] Department of Fuel and Carbon Materials, Kharkov National University "Kharkiv Polytechnic Institute", Frunze str., 21, 61002 Kharkiv, Ukraine
[b] Department of Chemical Metrology, Kharkov V.N. Karazin National University, Svoboda sq. 4, 61077 Kharkiv, Ukraine

ARTICLE INFO

Article history:
Received 2 December 2010
Received in revised form 21 February 2011
Accepted 25 February 2011
Available online 4 March 2011

Keywords:
Gasoline
Classification
Octane number
Cluster analysis
Principal component analysis
Neural network
Light gas condensate fraction

ABSTRACT

The approach for classification of gasoline by octane number and light gas condensate fractions by origin with using dielectric permeability data has been proposed and compared with classification of same samples on the basis of gas-chromatographic data. The precision of dielectric permeability measurements was investigated by using ANOVA. The relative standard deviation of dielectric permeability was in the range from 0.3 to 0.5% for the range of dielectric permeability from 1.8 to 4.4. The application of exploratory chemometrics tools (cluster analysis and principal component analysis) allow to explicitly differentiate the gasoline and light gas condensate fractions into groups of samples related to specific octane number or origin. The neural networks allow to perfectly classifying the gasoline and light gas condensate fractions.

1. Introduction

Discrimination of gasoline types and origin of feedstock for gasoline producing is of primary importance for quality control, forensic science, ecology monitoring etc. The Ukrainian gasoline market includes 4 main types of automobile fuels differentiated by octane number: A-76(80); A-92; A-95; and A-98. The quality of gasoline must satisfy to national standard requirements [1,2], which meets the technical requirements and control methods of EN 228:2004. However, the completing the full analysis according to [3] occupies a lot of time. As a result the fast identifying of the octane number adulteration of gasoline that is commonly encountered problem of Ukrainian market becomes a very hard task.

The high portion of Ukrainian gasoline is produced by processing the light gas condensate. The main gas pools are situated in the East of Ukraine, viz. in Kharkov, Poltava and Dnepropetrovsk regions. The largest depositions of gas in East of Ukraine are situated near such towns as Shebelinka, Krestyshchivka, Yulivka, Nova Vodolaga, Krasnokuts'k, Magdalinivka and Perescepene. It is clear that the properties of mined gas condensate should be related with their geographical origin. To identify the origin of gas condensate is necessary for forensic and environmental analysis, when the pollution or thievery source must be quickly discovered.

The search for alternative method based on chemical or physical data for determination of gasoline octane number or identification of oil source began more than 35 years ago [4–20]. Myers et al. have derived the multiple linear relationships between the octane numbers and isoparafin index determined by NMR, aromatic, lead and sulfur content with R^2 equals 0.90 [5]. The similar approach have been applied by Muhl and Srica to ^{1}H NMR data and fluid catalytic cracking gasoline octane number [21]. The adsorption spectra of gasoline in short wavelength near infrared region have been used by Kelly et al. for octane number prediction by multiple linear and partial least squares regressions [22].

The revolution in computation engineering made possible rapid growth of publications comprising the application of multi-signal analysis and chemometrics tools for identification and classification of crude oil products. The most popular analytical methods for classification of gasoline and other crude oil products are gas-chromatography coupled with classical flame-ionization detector (FID) [4–6,13,17,23–40] or MS detector [41–49] and FTIR spectrometry [22,50–58]. The chemometrics tools used for treatment of experimental involves as well-known principal component analysis (PCA) [25,30,34,37,39–41,43–47,49,51,56,57,59–63], discriminant analysis (DA) [25,35,50,51,60,62,64], artificial neural networks (ANN) [31,36,41,50–52,57–59,61,65,66], soft independent modeling of class analogy (SIMCA) [34,49–51,53,56,58,67,68], partial least squares (PLS) [22,32,51,53–57,69], genetic algorithms

* Corresponding author. Tel.: +380509151791.
E-mail addresses: vasiliy-rudnev@mail.ru (V.A. Rudnev), boichenko@univer.kharkov.ua (A.P. Boichenko).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.02.049

(GA) [26–28], support vector machine (SVM) [51,52,70] as well as less common fuzzy rule-building expert systems [71], parallel factor analysis (PARAFAC) [42,48], projection pursuit regression [30]. Recently Balabin et al. have compared several chemometrics algorithms for motor oil classification [72] and biodiesel analysis [73] based on NIR spectroscopy data.

Despite the successful results obtained with using GC-FID, GC-MS or FTIR several alternative methods that in some cases are cheaper or faster have been proposed for collection of experimental data and discrimination of crude oil or oil products. Ichikawa et al. have used the DA, PCA, k-nearest neighbor method and minimal spanning tree method for discrimination between regular and premium gasoline based on mass-spectrometric data [64]. The backpropagation ANN have been trained by Andrews and Lieberman for identification of gasolines, diesel fuels and oils on the basis of laser induced fluorescent spectra of samples [65]. The results of classification were satisfactory (90% of test samples were successfully identified), but the experimental method was developed in authors laboratory and is not generally accessible [65]. McCarrick et al. [66] have constructed apparatus based on vapor-sensitive array of detectors for classification of aviation fuels by ANN. However, the developed procedure was time-consuming [66]. ^{1}H NMR spectroscopic fingerprints analyzed by SIMCA were applied by Flumignan et al. for screening the Brazilian commercial gasoline quality (92% of samples were correctly classified) [67]. Oliviera et al. have proposed to apply the distillation curves data and SIMCA method for detection adulterations in Brazilian gasoline [68]. Recently, the same experimental method has been used by Aleme et al. for determination of gasoline origin with using LDA and PCA (only 70% of gasoline samples were correctly identified) [60]. The data from an array of conducting sensors were treated with PCA and multi-layer perceptron ANN for classification of 40 gasoline samples by Ozaki (90% of samples were identified correctly) [61]. Dispersive fiber-optic Raman spectroscopy was used by Flecher et al. for prediction of octane numbers of gasoline, but the errors were slightly higher in comparison with near-IR analysis [69]. Sobanski et al. developed the gas sensitive electronic nose for qualification of automobile fuels, which however needs further investigations [59]. The successful system based on electronic nose measurement system in cooperation with SVM have been proposed by Brudzewski et al. [70]. Gold nanoparticle chemiresistor sensor array that allow to differentiates between hydrocarbon fuels dissolved in artificial seawater was developed by Cooper et al. [62]. Recently, Guan et al. have applied dielectric spectroscopy analyzer specially designed for petroleum analysis for gasoline [74] and engine lubricating oil [75] classification by using PLS [75] and SVM [74].

In this work, an approach based on measurement of dielectric permeability of samples boiling fractions is proposed for classification of gasoline and light gas condensate fractions mined in Ukraine. The PCA, cluster analysis and several types of ANN were used in the work for successful discrimination of commercial gasoline and feedstock and comparing of classification results with classification observed with using GC data.

2. Experimental and precision of dielectric permeability estimated by "top–down" approach

2.1. Reagents and materials

The 32 samples of automobile gasoline attributed to 4 different brands A-76(80), A-92, A-95, A-98, and 48 samples of light gas condensate fractions collected at 7 different mines have been used in this work. The conformity of gasoline samples to Ukraine national standard DSTU 4063-2001 have been proved [1]. The towns near the mines of light gas condensate fractions, brands of gasoline, corresponding number of samples and abbreviations are presented in Table 1. Analytically grade 1-pentan, 1-hexan, 1-heptan, 1-octane, 1-undecan, toluene from Ukrainian suppliers and benzene from Merck have been used for estimation of repeatability and intermediate precision of dielectric measurements.

Table 1
Samples and their abbreviations.

No.	Type of gasoline and name of town near the light gas condensate are mined	Abbreviation	Number of samples
1	A-98	d	5
2	A-95	c	9
3	A-92	b	8
4	A-76 (A-80)	a	10
5	Krasnokuts'k	S	9
6	Yuliivka	J	8
7	Nova Vodolaga	EN	5
8	Shebelinka	Sh	7
9	Krestyshchivka	K	7
10	Magdalinivka	L	7
11	Perescepene	Y	5

2.2. Equipment and procedure of dielectric permeability calculation

The dielectric permeability of samples has been measured by using Q-meter type E4-4 with characteristics in keeping with current Ukraine national standard [76].

The principle of dielectric permeability measurement is based on the measurement of the cell capacity, which is filled with sample, capacity of empty cell. Then the relative capacity is calculated as ration of capacity of filled and empty cell, correspondingly. The measurements have been conducted after thermostating of the cell under standard conditions. The effect of parasitic capacitance has been taking into account for the calculation of dielectric permeability. The measurement of parasitic capacitance has been done on the basis of measured capacity of cell filled with reference sample with known dielectric permeability and capacity of empty cell. The benzene with dielectric permeability equals 2.273 [77] under standard conditions have been used for parasitic capacitance calculation. The calculations have been conducted on the basis of Eqs. (1) and (2):

$$\varepsilon = \frac{C_1 - C_p}{C_0 - C_p} \quad (1)$$

where ε: dielectric permeability; c_0: capacity (picofarad) of empty cell; c_1: capacity of cell with sample; and c_p: parasitic capacitance.

The parasitic capacitance has been calculated with the following equation:

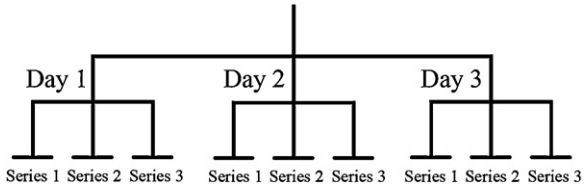
$$c_p = \frac{c_0\varepsilon_k - c_k}{\varepsilon_k - 1} \quad (2)$$

where c_0: capacity (picofarad) of empty cell; c_k: capacity of cell filled with standard solvent; and ε_k: the dielectric permeability of standard solvent.

Some theoretical considerations concerning the theory and measurement of dielectric permeability in oil can be found in [78,79].

2.3. Repeatability and intermediate precision of dielectric permeability measurements

The repeatability (the set of measurements conducted after cell calibration during the short interval of time) and intermediate precision (between sets of measurements and between days) have been investigated by analysis of variance (ANOVA), which was



Scheme 1.

successfully applied in our previous work [80]. In this work the hierarchical experimental design (Scheme 1) has been used for investigation the significance of between sets and between days components of variance. According to the two-factor ANOVA the result of measurement can be presented by following equation:

$$X_{ijt} = \xi + \gamma_t + \delta_{jk} + \varepsilon_{ijt} \quad (3)$$

where ξ: total mean; γ_t: the mean deviation with respect to ξ for k^{th} day; δ_{jt}–mean deviation of ith set with respect to intraday value ($\xi + \gamma_t$); and ε_{ijt}: repeatability error.

On the basis of the model (3) the variance s_3^2 (characterized between days variance), s_2^2 (characterized between sets variance in one day) and s_1^2 (characterized repeatability) were calculated. The s_1^2 value with $km(n-1)$ degrees of freedom is the estimation of repeatability variance $\sigma^2_{\text{rep.}}$. The s_2^2 value with $k(m-1)$ degrees of freedom is the estimation of general variance $n\sigma^2_{\text{rep.}} + \sigma^2_{\text{rep.}}$, where σ^2_{set} is the variance stipulated by between sets component, related with recalibration of equipment. The s_3^2 is the estimation of variance $nm\sigma^2_{\text{days}} + n\sigma^2_{\text{set}} + \sigma^2_{\text{rep.}}$, where σ^2_{days}: between days component of variance. k: number of days (in this work $k = 3$); m: number of sets during each day (in this work $m = 3$); and n: number of measurements in each set (in this work $n = 6$). Variances σ^2_{set} and σ^2_{days} are unequal to zero, if the Fisher criteria $F = s_2^2/s_1^2$ и $F = s_3^2/s_2^2$ are statistically significant. If the $F = s_2^2/s_1^2$ is statistically insignificant the mean value of s_1^2 and s_2^2 are calculated for $\sigma^2_{\text{rep.}}$ estimation:

$$\overline{s^2} = \frac{f_1 s_1^2 + f_2 s_2^2}{f_1 + f_2} \quad (4)$$

where f_1 and f_2 corresponds to degrees of freedom related to s_1^2 and s_2^2.

The homogeneity of variances had been tested by using F-criterion, Kohren criterion and Bartlett criterion. After the proving the homogeneity of variances in sets and insignificance if between sets component of variance the all experiments conducted during one day have been combined for exploring between days component of variance.

On the basis of F-criterion, Kohren criterion and Bartlett criterion the homogeneity of variance in each set has been proved for 5% significance level (Table 1 in Supplementary material). The results of ANOVA are presented in Table 2. It was observed that the between sets component of variance is insignificant and mean variance was calculated by Eq. (4). The calculated values of $F = s_3^2/\overline{s^2}$ are insignificant for 1% significance level for all solvents except 1-pentane, which has lowest boiling temperature in comparison with other solvents used for precision investigation.

The calculated standard deviation of dielectric permeability increases with increasing of absolute value of dielectric permeability (Fig. 1A in Supplementary material), however the relative standard deviation of dielectric permeability is practically unchanged (Fig. 1B in Supplementary material) and equals 0.3–0.5%. This value of uncertainty can be attributed to the dielectric permeability measurements in the range from 1.8 to 4.4. This range is common for gasoline without additives of ethanol and feedstock (light gas condensate fractions) for its producing.

2.4. Chromatographic conditions

The gas–liquid chromatography (GC) with gas chromatograph «Kristall-2000M» (Chromatek, Russian Federation) has been used for component analysis of gasoline and light gas condensate fractions samples. The volume of sample was 1 μL); gas-carrier-nitrogen; detector-flame-ionization; column-Quadrex 007–1, size 50 m × 0.25 mm (Quadrex Corporation, USA) with polydimethylsiloxane stationary phase; detector temperature was 210 °C, temperature of evaporator 180 °C; velocity of gas-carrier was 20 ml min^{-1}. The temperature of column was changed according to the following of scheme: initial temperature of column 35 °C, isothermal durability was 20 min; heating rate was 2 °C min^{-1}; final temperature was 200 °C. The duration of one analysis equals 110 min. The identification of peaks and quantitative analysis was performed by using program «Chromatek-Analytik 2.5» (Chromatek, Russian Federation, http://www.chromatec.ru) and program «Gasoline» (Chromatek, Russian Federation).

2.5. Collection of gasoline and feedstock boiling fractions

The temperatures of boiling fractions collection have been chosen on the basis of GC data about the gasoline composition, and literature data about boiling temperatures of gasoline components and their dielectric permeability [77,81]. The limit temperatures of boiling fractions collection have been specified as 75 °C, 120 °C and 195 °C. The first boiling fraction consisted mainly from the alkanes, the second fraction contains more complex hydrocarbons, toluene and benzene, and the third boiling fraction was enriched with other aromatic hydrocrabons. However, the differentiating of compounds between boiling fractions is only approximate due to a number of azeotropes that are formed in gasolines [82]. The standard procedure according to EN ISO 3405:2000 of boiling fractions of gasoline collection has been used. The difference between the dielectric permeability

Table 2
The analysis of variance of dielectric permeability measurements.

No.	Solute	Variance (×10^5)					$F = s_3^2/\overline{s^2}$ $F_{\text{table, }0.05,2,51} = 3.18$ $F_{\text{table, }0.01,2,51} = 5.05$
		s_1^2	s_2^2	s_3^2	$\overline{s^2}$	$F = s_2^2/s_1^2 F_{\text{table., }0.05,6,45} = 5.05$	
1	1-pentane	1.17	2.34	7.54	1.31	2.00	5.77
2	1-hexane	1.27	1.19	3.46	1.26	0.94	2.74
3	1-heptane	1.45	1.27	5.59	1.43	0.88	3.91
4	1-octane	1.40	1.09	4.35	1.36	0.78	3.19
5	1-undecane	1.87	2.01	4.74	1.89	1.07	2.51
6	Toluene	2.12	2.39	6.76	2.15	1.13	3.14
7	Methyltretbutylether	4.11	1.87	18.4	3.85	0.45	4.78

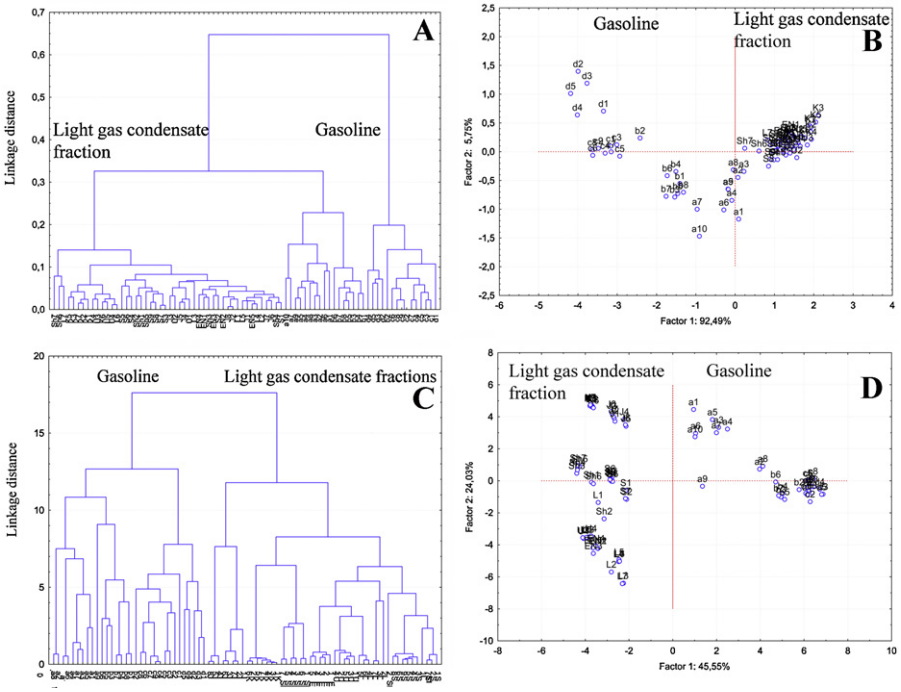


Fig. 1. The results of CA of gasoline and light gas condensate fractions by using values of dielectric permeability of original sample and three collected boiling fractions (A) and chromatographic data (C) and results of PCA of dielectric permeability data (B) and chromatographic data (D) of the same samples.

of original sample and collected boiling fractions were statistically significant, due to high precision of dielectric permeability measurements (see Section 2.3), e.g. next values of dielectric permeability were obtained for one of A-95 gasoline: 2.293; 2.229; 2.288; and 2.329.

2.6. Data sources, software and processing

The statistical analyses were performed with Statistica 8.0, data analysis software system (Statsoft Inc., http://www.statsoft.com) and Microsoft Excel (2003, Microsoft, http://office.microsoft.com).

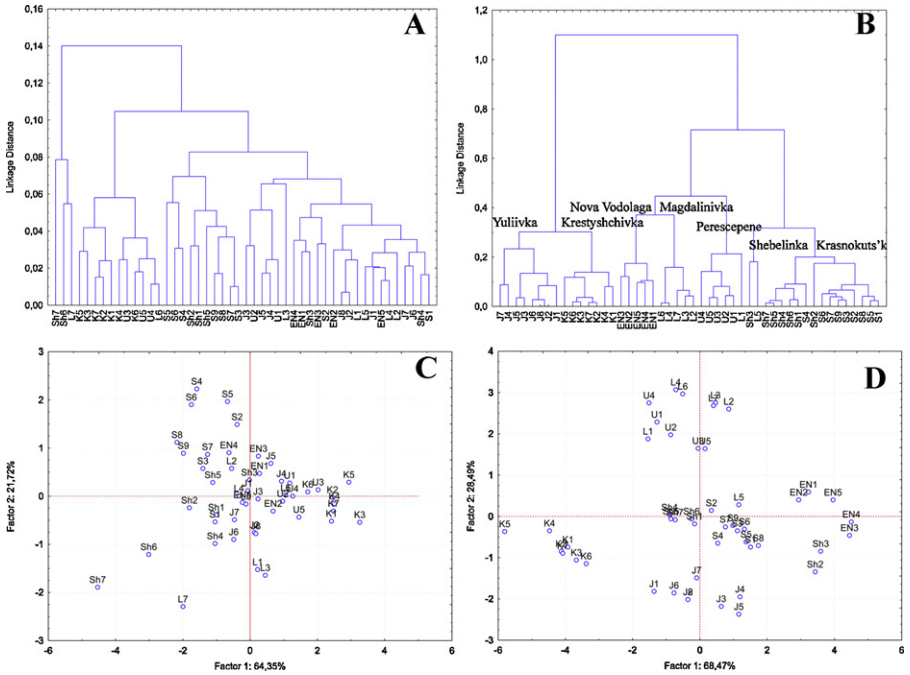


Fig. 2. The results of CA (A and B) and PCA (C and D) with using dielectric permeability data (A, and C) or dielectric permeability data, density of original sample and volumes of collected boiling fractions (B and D).

Table 3
Neural networks for gasoline and light gas condensate fractions classification.

No.	Network type and number of neurons in hidden layer	Training perf. (%)	Test perf. (%)	Training algorithm	Error function	Hidden layer activation function	Output layer activation function
	Gasoline classification based on dielectric permission data						
1	MLP, 5	100	100	Quasi-Newton	SOS[b]	Logistic	Tanh
2	MLP, 5	92	100	Quasi-Newton	CE[c]	Tanh[d]	Softmax[e]
3	RBF, 8	96	100	RBFT[a]	SOS	Gaussian	Identity[f]
4	MLP, 6	96	100	Quasi-Newton	CE	Exponential	Softmax
5	MLP, 7	96	100	Quasi-Newton	CE	Identity	Softmax
	Gasoline classification based on GC data						
1	MLP, 9	100	100	Quasi-Newton	SOS	Identity	Exponential
2	MLP, 23	100	100	Quasi-Newton	CE	Logistic	Softmax
3	MLP, 9	92	100	Quasi-Newton	CE	Logistic	Softmax
4	MLP, 19	96	100	Quasi-Newton	CE	Logistic	Softmax
5	MLP, 16	96	100	Quasi-Newton	SOS	Identity	Logistic
	Light gas condensate fractions classification based on dielectric permission data						
1	MLP, 5	100	100	Quasi-Newton	CE	Tanh	Softmax
2	MLP, 6	84	100	Quasi-Newton	CE	Identity	Softmax
3	RBF, 11	84	100	RBFT	CE	Gaussian	Softmax
4	MLP, 10	82	100	Quasi-Newton	CE	Tanh	Softmax
5	MLP, 5	92	100	Quasi-Newton	CE	Logistic	Softmax
	Light gas condensate fractions classification based on GC data						
1	MLP, 11	90	100	Quasi-Newton	CE	Identity	Softmax
2	MLP, 25	92	100	Quasi-Newton	SOS	Exponential	Tanh
3	MLP, 14	94	100	Quasi-Newton	CE	Tanh	Softmax
4	MLP, 22	94	100	Quasi-Newton	SOS	Exponential	Tanh
5	MLP, 11	100	100	Quasi-Newton	CE	Logistic	Softmax

[a] RBFT: RBF training algorithm.
[b] SOS: sum-of-squares.
[c] CE: cross entropy.
[d] Tanh: hyperbolic tangent function.
[e] Softmax: specialized activation function for one-of-N encoded classification networks. It performs a normalized exponential (i.e., the outputs add up to 1) [86].
[f] Identity–identity function, in which the activation level is passed on directly as the output.

3. Results and discussion

The discrimination of complex mixtures such as gasoline or light gas condensate fraction on the basis of results obtained by non-selective methods, e.g. dielectric permeability, density, conductance etc. cannot be done without the application of chemometrics tools for data treatment. As was mentioned in Section 2.5 the dielectric permeability of quite different by octane number gasolines and feedstocks from collected at different places could be similar due to integral properties of dielectric permeability. Thus, we have tried to increase the information about each sample by measurement of dielectric permeability of original sample and three boiling fractions. Additionally, the data on density of original sample and volumes of each collected boiling fraction have been tested for improving of discrimination. As a result, each tested sample was characterized maximum by 8 values: 4 values of dielectric permeability, density of original sample and volumes of 3 collected boiling fractions. The collected data for 80 samples of gasoline and light gas condensate fractions has been treated by using cluster analysis (CA), PCA and ANN. The observed results on samples discrimination have been compared with results of discrimination obtained on the basis of component analysis by GC of each sample by using same chemometrics tools.

3.1. Exploratory classification of gasoline and light gas condensate fractions by using cluster analysis and principal component analysis

The application of CA and PCA is important for exploratory analysis of data, because they allow visualizing obtained results. Another important advantage of CA and PCA is the absence of necessity of beforehand defining the number of clusters or groups and opportunity to investigate as final result of classification as well as

distances between groups, which combined in one cluster or group [83].

3.1.1. Basic ideas of cluster analysis and choosing of distance measure and clusters linking method

The CA can be divided into two steps. At the first step of the hierarchical CA the distances between all samples in multi-dimensional space is calculated by using one of the following equations: Euclidean $d(i,k) = \left(\sum_{j=1}^{n}(x_{i,j} - x_{k,j})^2\right)^{0.5}$, square Euclidean $d(i,k) = \sum_{j=1}^{n}(x_{i,j} - x_{k,j})^2$, Manhattan $d(i,k) = \sum_{j=1}^{n}\left|x_{i,j} - x_{k,j}\right|$, Chebyshev $d(i,k) = \max\left|x_{i,j} - x_{k,j}\right|$ and Power distance $d(i,k) = \left(\sum_{j=1}^{n}\left|x_{i,j} - x_{k,j}\right|^{p}\right)^{1/r}$ [84]. At the second step the joining of samples and clusters is performed. The several methods are often used for estimation of links: (i) single linkage; (ii) complete linkage; (iii) unweighted pair-group average; (iv) weighted pair-group average; (v) unweighted pair-group centroid; and (vi) weighted pair-group centroid (median). The basis of these methods are well described in the literature [84]. The result of hierarchical CA is presented as dendrogram, which can be used for step by step investigation of samples linkage.

In this work we have tested all types of distance calculation and all linking methods described above. The best results have been observed by using unweighted pair-group average method for clusters joining and Euclidean distance for calculation the distances between clusters. Thus, all classifications presented below have been performed with calculation of Euclidean distance and unweighted pair-group average method, in which the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. The data on the dielectric permeability have been used as received for CA. The content of each component in the samples determined by GC were also used for CA as received. In the case of involving the characteristics with different dimension (density of original sample and volumes of collected boiling fractions) the data have been standardized by common standardization method, which transforms all the data to have zero mean and unit standard deviation.

3.1.2. Basic ideas of principal component analysis for data classification

The main goal of PCA is to reduce the dimensionality of the data while retaining as much as possible variation in the dataset. Several well documented methods can be used for PCA: singular value decomposition, non-linear iterative partial least squares, covariance method etc. In the case of applying of PCA for dielectric permeability data (4 values for each sample) the main object was to obtain orthogonal components for exploring the relationships between groups, and in the case of chromatographic data treatment, where the 38 components have been used for characterization of each sample, to reduce the data dimensionality. The dielectric permeability and chromatographic data have been standardized before the PCA.

3.1.3. Comparison of samples classification with using CA and PCA to chromatographic and dielectric permeability data

The application of CA and PCA for classification of gasoline and light gas condensate fractions by using values of dielectric permeability of original sample and three collected boiling fractions are presented in Fig. 1A and B and the results of CA and PCA of chromatographic data are presented in Fig. 1C and D. It is clear that gasoline samples and samples of light gas condensate can be easily separated one from another on the basis of CA and PCA. Also it can be concluded that the gasoline samples with different octane number have less similarities then light gas condensate fractions collected in different feedstock.

On the basis of these results the gasoline and light gas condensate fractions have been separated into two groups to provide classification of gasoline by octane number and light gas condensate fractions by origin with CA and PCA. In the Figs. 2 and 3 of Supplementary material the hierarchical dendrogram of gasoline classification by CA and 2D graph of first and second principal components obtained by using dielectric permeability data (Figs. 2A and 3A in Supplementary material) and chromatographic data (Figs. 2B and 3B in Supplementary material) are presented.

The CA based on dielectric permeability and chromatographic data results in samples separation in two groups: with low (A-76(80) and A-92) and high octane number (A-95 and A-98). The comparison of classification based on dielectric permeability data and chromatographic gives next conclusions: (i) all samples of A-76(80) gasoline were correctly classified with using dielectric permeability data and one A-76(80) sample was attached to A-92 cluster with using chromatographic data; (ii) all samples of A-98 gasoline were correctly classified with using chromatographic data and one A-98 was added to A-95 cluster with using dielectric permeability data; and (iii) one sample of A-92 gasoline was incorrectly included in A-95 cluster in both cases. In general, the percent of incorrect classification of gasoline samples is equal 6.3% in both cases. However, it should be noted that incorrect classification of one of A-92 sample as A-95 gasoline could be due to really higher content of methyl-tret-butyl-ether (additive results in increasing of octane number) that was proved by quantitative GC analysis.

The projections of the first and second principal components calculated on the basis dielectric permeability and chromatographic data could be used for discrimination of four gasoline groups (Fig. 3 in Supplementary material). However, some of the samples cannot be clearly attributed to one of the group.

The application of CA and PCA with using only dielectric permeability data for classification of light gas condensate fractions by the place of origin was unsuccessful (Fig. 2A and C). The great improving of results was observed, when the density of original sample and volumes of each boiling fraction have been used in common with dielectric permeability data. These parameters are easily determined in the same experiment on samples fractionation. The Fig. 2B and D presents the results of CA and PCA on the basis of combined set of experimental data. The samples collected near the Yuliivka, Krestyshchyvka, Nova Vodolaga and Perescepene are combined into clusters without mistakes. Some samples from Krasnokuts'k, Shebelinka and Magdalinivka were erroneously classified that results in the around 10% of mistakes.

In the case of CA and PCA of chromatographic data the samples are more clearly form the clusters and groups, however, some samples were also erroneously classified, e.g. samples collected near Shebelinka and Magdalinivka were attributed to the samples, collected near Krasnokuts'k (Fig. 4 in Supplementary material). The error of classification in the case of GC data is around 4%.

3.2. Artificial neural networks for gasoline and light gas condensate fractions classification

The interest to ANN has grown rapidly over the past two decades. The application of ANN to chemistry problems is also increases, even despite the absence of clear relationships between dependent and independent variables. The classification is one of the main fields of ANN application. The neural networks are divided into two main groups: trained neural networks and self-organized Kohonen network. The first group of ANN has been used in our work for classification purposes. Multilayer Perceptron Neural Networks

(MLPN) and Radial Basis Function Neural Networks (RBFNN) are two types of trained ANN. The description of MLPN and RBFNN can be found in [85]. Briefly, the ANN consists from input and outputs neurons, which number, in the case of classification purpose, corresponds to the number of values, characterized each sample, and number of groups, respectively. Between the input and output layer the hidden layer of neurons is located. MLPN and RBFNN are feedforward neural network architecture with uni-directional full connections between successive layers. The neurons of a network have activation functions which transform the incoming signals from the neurons of the previous layer using a mathematical function, e.g. logistic, sigmoid, hyperbolic tangent, exponential, and Gaussian. The last one is solely used in hidden neurons of RBFNN.

Several algorithms (Conjugate Gradient Descent, Broyden–Fletcher–Goldfarb–Shanno, etc.) available in Statistica 8.0 software has been used for optimization of neural network architecture. The train sample size was 80% and test sample size was 20%. In Table 3 the results of Automated ANN algorithm application to dielectric permeability and GC data are presented. The number of output neurons was 4 and 7 for classification of gasoline and light gas condensate fractions, respectively. According to the number of values attributed to each sample the number of input neurons was 4 and 38 for classification of gasoline based on dielectric permeability and GC data, respectively, and 8 and 37 neurons for classification of light gas condensate fractions. The optimized architecture of ANN includes up to 10 hidden neurons in most cases that is agreed with previously published data on optimization of ANN for motor oil classification [58,72], however in some cases the optimal number of neurons in hidden layer is more than 20 that can be related with complex relationship between the optimal number of hidden neurons and quality of prediction.

As can be seen from Table 2 the ANNs provide in some cases excellent classification of gasoline and light gas condensate fractions. The errors in classification are related with mistakes observed for the same sample as for classification using CA. The classification of samples was good as by using dielectric permission as well as by using GC data. This confirms the applicability of proposed approach for routine use in laboratories of forensic and environmental analysis.

4. Conclusions

In this work new, fast, reliable and cheap approach for the classification of petrochemicals (gasoline by octane number and light gas condensate fractions by origin) on the basis of physical–chemical data and chemometric tools was proposed. The dielectric measurements are precise enough to establish the differences between the dielectric permeability of collected boiling fractions of gasoline and light gas condensate fractions used as feedstock for gasoline producing. The errors in results of samples classification are in agreement with best examples of gasoline classification by using FTIR and chromatographic data. However, it should be noted that the equipment for dielectric permeability measurements and collection of boiling fractions is more than thirty times lower in comparison with GC or FTIR equipment. Moreover several samples can be fractionated and studied simultaneously that allow to classify larger numbers of samples during the same time in comparison with GC or HPLC [86].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.talanta.2011.02.049.

References

[1] DSTU4063-2001, National Standard of Ukraine (in Ukrainian): Kiev, Ukraine, (2001) 9.
[2] DSTU4839-2007, National Standard of Ukraine (in Ukrainian): Kiev, Ukraine, (2007) 14.
[3] ASTMD2700, (2009) 54.
[4] H.A. Clark, P.C. Jurs, Anal. Chem. 47 (1975) 374–378.
[5] M.E. Myers, J. Stollsteimer, A.M. Wims, Anal. Chem. 47 (1975) 2301–2304.
[6] O.C. Zafiriou, Anal. Chem. 45 (1973) 952–956.
[7] M. Anbar, W.H. Aberth, Anal. Chem. 46 (1974) 59A–64a.
[8] C.W. Brown, P.F. Lynch, M. Ahmadjian, Environ. Sci. Technol. 8 (1974) 669–670.
[9] M.E. Garza Jr., J. Muth, Environ. Sci. Technol. 8 (1974) 249–255.
[10] A. Novikov Yu, A.V. Vikhlyantzev, O. Porksheyan Kh, Sudebno-Meditsinskaya Ekspertiza 17 (1974) 27–29.
[11] R. Dell'Acqua, J.A. Egan, B. Bush, Environ. Sci. Technol. 9 (1975) 38–41.
[12] D.L. Duewer, B.R. Kowalski, T.F. Schatzki, Anal. Chem. 47 (1975) 1573–1583.
[13] J.G. Pym, J.E. Ray, G.W. Smith, E.V. Whitehead, Anal. Chem. 47 (1975) 1617–1622.
[14] A.P. Bentz, Anal. Chem. 48 (1976) 454A–472A.
[15] C.W. Brown, P.F. Lynch, Anal. Chem. 48 (1976) 191–195.
[16] F.K. Kawahara, Environ. Sci. Technol. 10 (1976) 761–765.
[17] D.V. Rasmussen, Anal. Chem. 48 (1976) 1562–1566.
[18] J.S. Mattson, C.S. Mattson, M.J. Spencer, S.A. Starks, Anal. Chem. 49 (1977) 297–302.
[19] A.P. Bentz, Anal. Chem. 50 (1978) 655A–658A.
[20] S.H. Fortier, D. Eastwood, Anal. Chem. 50 (1978) 334–338.
[21] J. Muhl, V. Srica, Fuel 66 (1987) 1146–1149.
[22] J.J. Kelly, C.H. Barlow, T.M. Jinguji, J.B. Callis, Anal. Chem. 61 (1989) 313–320.
[23] X. Zhu, Fenxi Huaxue 30 (2002) 18–25.
[24] X. Zhu, L. Zhang, X. Che, L. Wang, Chemom. Intell. Lab. Syst. 45 (1999) 147–155.
[25] B.K. Lavine, H. Mayfield, P.R. Kromann, A. Faruque, Anal. Chem. 67 (1995) 3846–3852.
[26] B.K. Lavine, A.J. Moores, H. Mayfield, A. Faruque, Microchem. J. 61 (1999) 69–78.
[27] B.K. Lavine, A.J. Moores, H.T. Mayfield, A. Faruque, Anal. Lett. 31 (1998) 2805–2822.
[28] B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, H.T. Mayfield, Anal. Chem. 72 (2000) 423–431.
[29] B.K. Lavine, A. Vesanen, D.M. Brzozowski, H.T. Mayfield, Anal. Lett. 34 (2001) 281–293.
[30] J.A. van Leeuwen, R.J. Jonker, R. Gill, Chemom. Intell. Lab. Syst. 25 (1994) 325–340.
[31] A.M. Fonseca, J.L. Biscaya, J. Aires-de-Sousa, A.M. Lobo, Anal. Chim. Acta 556 (2006) 374–382.
[32] Y. Lu, P.B. Harrington, Anal. Chem. 79 (2007) 6752–6759.
[33] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, Anal. Chem. 77 (2005) 7735–7743.
[34] E.S. Bodle, J.K. Hardy, Anal. Chim. Acta 589 (2007) 247–254.
[35] H.A. Clark, P.C. Jurs, Anal. Chem. 51 (1979) 616–623.
[36] J.R. Long, H.T. Mayfield, M.V. Henley, P.R. Kromann, Anal. Chem. 63 (1991) 1256–1261.
[37] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101–110.
[38] G. Wang, J. Karnes, C.E. Bunker, M. Lei Geng, J. Mol. Struct. 799 (2006) 247–252.
[39] Z. Wang, Energy Sources A: Recov. Util. Environ. Effects 25 (2003) 491–508.
[40] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225–237.
[41] P. Doble, M. Sandercock, E. Du Pasquier, P. Petocz, C. Roux, M. Dawson, Forensic Sci. Int. 132 (2003) 26–39.
[42] D. Ebrahimi, J. Li, D.B. Hibbert, J. Chromatogr. A 1166 (2007) 163–170.
[43] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 134 (2003) 1–10.
[44] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 140 (2004) 43–59.
[45] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 140 (2004) 71–77.
[46] M.E. Sigman, M.R. Williams, J.A. Castelbuono, J.G. Colca, C.D. Clark, Instrum. Sci. Technol. 36 (2008) 375–393.
[47] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1129 (2006) 111–118.
[48] K.J. Johnson, S.L. Rose-Pehrsson, R.E. Morris, Pet. Sci. Technol. 24 (2006) 1175–1186.
[49] B. Tan, J.K. Hardy, R.E. Snavely, Anal. Chim. Acta 422 (2000) 37–46.
[50] R.M. Balabin, R.Z. Safieva, Fuel 87 (2008) 1096–1101.
[51] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Anal. Chim. Acta 671 (2010) 27–35.
[52] K. Brudzewski, A. Kesik, K. Kołodziejczyk, U. Zborowska, J. Ulaczyk, Fuel 85 (2006) 553–558.
[53] L.S.G. Teixeira, F.S. Oliveira, H.C. dos Santos, P.W.L. Cordeiro, S.Q. Almeida, Fuel 87 (2008) 346–352.
[54] M.A. Al-Ghouti, Y.S. Al-Degs, M. Amer, Talanta 76 (2008) 1105–1112.
[55] W. Bao, R. Zhou, J. Yang, D. Yu, N. Li, Mech. Syst. Signal Process. 23 (2009) 1458–1473.
[56] J.F. Padilha, R.W.S. Pessoa, J.G.A. Pacheco, P.R.B. Guimarães, in, (2009), pp. 753-758.
[57] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Chemom. Intell. Lab. Syst. 88 (2007) 183–188.
[58] R.M. Balabin, R.Z. Safieva, Fuel 87 (2008) 2745–2752.
[59] T. Sobanski, A. Szczurek, K. Nitsch, B.W. Licznerski, W. Radwan, Sens. Actuators B: Chem. 116 (2006) 207–212.
[60] H.G. Aleme, L.M. Costa, P.J.S. Barbeira, Fuel 87 (2008) 3664–3668.

[61] S.T.R. Ozaki, N.K.L. Wiziack, L.G. Paterno, F.J. Fonseca, in: M. Pardo, G. Sberveglieri (Eds.), AIP, Brescia (Italy), 2009, pp. 525–526.
[62] J.S. Cooper, B. Raguse, E. Chow, L. Hubble, K.-H. Müller, L. Wieczorek, Anal. Chem. 82 (2010) 3788–3795.
[63] N.A. Sinkov, J.J. Harynuk, Talanta 83 (2011) 1079–1087.
[64] M. Ichikawa, N. Nonaka, I. Takada, S. Ishimori, Anal. Sci. 9 (1993) 261–266.
[65] J.M. Andrews, S.H. Lieberman, Anal. Chim. Acta 285 (1994) 237–246.
[66] C.W. McCarrick, D.T. Ohmer, L.A. Gilliland, P.A. Edwards, H.T. Mayfield, Anal. Chem. 68 (1996) 4264–4269.
[67] D.L. Flumignan, N. Boralle, J.E.d. Oliveira, Talanta 82 (2010) 99–105.
[68] F.S. de Oliveira, L.S. Gomes Teixeira, M.C. Ugulino Araujo, M. Korn, Fuel 83 (2004) 917–923.
[69] P.E. Flecher, W.T. Welch, S. Albin, J.B. Cooper, Spectrochim. Acta A: Mol. Biomol. Spectrosc. 53 (1997) 199–206.
[70] K. Brudzewski, S. Osowski, T. Markiewicz, J. Ulaczyk, Sens. Actuators, B 113 (2006) 135–141.
[71] P. Rearden, P.B. Harrington, J.J. Karnes, C.E. Bunker, Anal. Chem. 79 (2007) 1485–1491.
[72] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Microchem. J. 98 (2011) 121.
[73] R.M. Balabin, E.I. Lomakina, R.Z. Safieva, Fuel 90 (2011) 2007.
[74] L. Guan, X.L. Feng, Z.C. Li, G.M. Lin, Fuel 88 (2009) 1453–1459.
[75] L. Guan, X.L. Feng, G. Xiong, Anal. Chim. Acta 628 (2008) 117–120.
[76] GOST6581-75, National Committee of USSR on standards (in Russian): Moscow (1988) 23.
[77] Y.Y. Ahadov, Dielectric properties of pure liquids. (in Russian), Izdatelstvo standartov, Moscow, 1972.
[78] R.Z. Syunyaev, R.M. Balabin, J. Dispersion Sci. Technol. 28 (2007) 419–424.
[79] H. n. Vrålstad, Ø.C. Spets, d. Lesaint, L. Lundgaard, J. Sjöblom, Energy Fuels 23 (2009) 5596–5602.
[80] A.P. Boichenko, A.L. Iwashchenko, L.P. Loginova, A.U. Kulikov, Anal. Chim. Acta 576 (2006) 229–238.
[81] F. Buckley, A. Maryott, Tables of Dielectric Dispersion Data for Pure Liquids and Dilute Solutions, U.S. EPO, Washington, DC, 1958.
[82] R.M. Balabin, R.Z. Syunyaev, S.A. Karpov, Energy Fuels 21 (2007) 2460–2465.
[83] G. Gan, C. Ma, J. Wu, Data Clustering Theory, Algorithms, and Applications SIAM, Alexandria, 2007.
[84] A.K. Jain, R.C. Dubes, Algorithms for clustering data, Prentice Hall, Englewood Cliffs, New Jersey, 1988.
[85] S. Haykin, Neural Networks: A Comprehensive Foundation, Macmillan, 1994.
[86] J.S. Bridle, Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition, in: F.F. Soulie, J. Herault (Eds.), Neurocomputing: Algorithms, Architectures and Applications, Springer-Verlag, Berlin, 1990.