Bulgarian Journal of Chemistry Volume 3 Issue 2

Article history:
Received: 5 September 2014
Revised: 3 October 2014
Accepted: 10 October 2014
Available online: 31 October 2014

Chemometric recognition of biodiesel fuels using their fatty acid methyl esters profiles

PAPER

Zylia Mustafa[1], Rumyana Milina[1], Pavlina Simeonova[2], Vasil Simeonov[3]*

[1] *Central Research Laboratory, Faculty of Natural Science, University "Prof. A. Zlatarov", Prof. Yakimov Blvd. 1, 8010 Bourgas*
[2] *Laboratory of Environmental Physics, Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences, Tzarigradsko Chaussee Blvd. 72, 1178 Sofia*
[3] *Laboratory of Chemometrics and Environmetrics, Chair of Analytical Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia "St. Kl. Okhridski", J. Bourchier Blvd. 1, 1164 Sofia*
[4] *Laboratory of Chemometrics and Environmetrics, Chair of Analytical Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia "St. Kl. Okhridski", 1164 Sofia, J. Bourchier Blvd. 1, Bulgaria*
** Corresponding author. E-mail: VSimeonov@chem.uni-sofia.bg*

The present study makes an initial effort to apply a chemometric expertise to a data set from chromatographic analysis of different plant sources of biofuel using their fatty acid methyl esters profiles (FAME). Cluster analysis and principal components analysis are used to create specific patterns for each one of the plants used in the study. Three major classes of biofuel profiles were proven and respective effort is made to correctly classify originally unknown samples by their chromatograms to one or another class of biofuel sources.

Keywords: Biodiesel; chemometrics; cluster analysis; fatty acid; methyl ester; principal components analysis

Introduction

Biodiesel is a potentially renewable substitute for diesel oil. A "green" fuel, biodiesel is biodegradable, nontoxic and is essentially free of metals, sulfur, carcinogenic aromatics and generates less greenhouse effect than fossil fuels [1-3].

Several processes and feedstock have been reported for the production of biodiesel [4-6]. Transesterification of vegetable oils and animal fats with alcohol (in most cases methanol) is currently one of the most attractive and widely accepted methodology. That is why biodiesel is often defined as a mixture of monoalkyl (methyl) esters of long-chain fatty acids.

At present, the dominant feedstock (about 80%) is vegetable oils, namely soybean oil in USA, rapeseed and sunflower oil in Europe and palm oil in Southeast Asia. Other feedstock having real or potential commercial interest are animal fats, non-edible and waste oils. Feedstock availability for biodiesel production varies according to geography, climate and economic of the countries.

Since biodiesel is a mixture of fatty acid methyl esters (FAME), its properties depend on the chemical structure of the individual FAME and their contents (FAME profile). FAME profiles of biodiesel are influenced by the stocks and origin of the oils used [1,7]. So, FAME profile may be used as an approach for selection of feedstock [7,8], for investigations [9,10], for fuel spillage and remedial actions in the environment [11].

Biodiesel FAME profiles can be obtained by chromatographic methods, providing valuable multi-component in-

formation. Gas chromatography-flame ionization detector (GC-FID), gas chromatography-mass spectrometry (GC-MS) [12-15] and high performance liquid chromatography [16] have been frequently used of FAME analysis.

However, visual evaluation of chromatograms is difficult and not reliable especially to compare a large number of specimens. The problem could be significantly overcome through the application of chemometric methods and intelligent data analysis for classification, modeling and interpretation of large data sets [17,18].

The aim of this study is to perform classification of biodiesels from different feedstock using their FAME profiles by chemometric expertise and assessment. Available literature data on FAME profiles of biodiesels from several types of oils were used [1,7,8,17,19]. Additionally, own data of FAME of Bulgarian biodiesels and samples produced by trans esterification of sunflower and rapeseed oils were utilized. Only those, meeting the requirement of EN 14214 [20], were included in our study.

Experimental

Materials

Totally 96 samples were used, indicated in *Table S1* with a number in respective class of sources. Class 1 is sunflower oil (totally 33 samples), class 2 is rapeseed oil (totally 19 samples), class 3 is maize oil (totally 8 samples), class 4 is soya oil (totally 19 samples), class 5 is palm oil (totally 6 samples), class 6 is peanut oil (totally 6 samples) and class 7 consists of 5 samples with unknown origin.

Certified reference materials: Fatty Acid Methyl Ester (B100), Methyl heptadecanoate (C17:0) were purchased from Spex CertiPrep; F.A.M.E. Mix Standard Rapeseed oil (cat № 18917) – from Supelco; Fatty acid methyl esters – myristate (C14:0), palmitate (C16:0), palmitoleate (C16:1), stearate (C18:0), oleate (C18:1), linoleate (C18:2), linolenate (C18:3), arachidate (C20:0), *cis*-11-eicosenoate (C20:1), behenate (C22:0), *cis*-13-docosanoate (C22:1), tetracosanoate (C24:0), *cis*-15-tetracosanoate (C24:1) – from Sigma-Aldrich.

Reagents of recognized analytical grade were used.

The analysis of the samples was performed by the use of gas chromatography. All GC analyses were performed on a GC system Agilent Technologies 7890A equipped with FID, split/splitless injector and Agilent 7693A automated liquid sampler. The fussed silica capillary column HP-INNOWAX, 30 m x 0.32 mm ID and 0.25 µm film thickness was used. Helium was used as a carrier gas, column flow was 1.5 ml/min. Hydrogen and air flows were set to 40 ml/min and 400 ml/min, respectively, makeup gas (nitrogen) 40 ml/min. The injection volume was 1 µl and split ratio was 1:80. The temperatures of the injector and the detector were 250 °C and 300 °C, respectively. The temperature program of the oven was initial temperature 210 °C for 9 minutes and then to 230 °C at 20 °C/min and hold there for 10 minutes.

ChemStation for GC (Agilent Technologies) was used for instrumental control, data acquisition and data analysis.

GC-MS analyses were carried out using GC system Agilent Technologies 7890A combined with MSD 5975 C Inert XL EI/CI, electron impact ionization (70 eV) mass range 30 – 500 m/z and the same chromatographic conditions. The components of biodiesels were identified by injection of standards (2.1) and by comparison of mass spectra with those of a NIST MS computer library.

Sample 0.3 µl were injected with split ratio 1:100. GC chromatograms of FAME from sunflower and rapeseed oil are shown in *Fig S1.* and *Fig S2*. Peaks are as follows: C14:0, C16:0, C17:0 (IS), C18:0, *cis*9 C18:1, *cis*9*cis*12 C18:2, *cis*9*cis*12*cis*15 C18:3, C20:0, *cis*11 C20:1, C22:0, *cis*13 C22:1, C24:0, *cis*15 C24:0. The content of each methyl ester was calculated by the method of internal standard (IS).

In *Figs. S1* and *S2* two typical chromatograms for rapeseed (*Fig. S1*) and sunflower oil (*Fig. S2*) are presented. Similar chromatograms were used for data collection.

Chemometric methods

In the present study two well-described multivariate statistical approaches for data interpretation were used – cluster analysis [21] and principal components analysis [22].

Cluster analysis is a widely used approach for environmetrical purposes. In order to cluster objects characterized by a set of variables (e.g. oil samples by FAME concentrations), one has to determine their similarity. A preliminary step of data scaling is necessary (e.g. autoscaling or z – transform, range scaling, logarithmic transformation) where normalized dimensionless numbers replaces the real data values. Thus, even serious differences in absolute (concentration) values are scaled to similar ranges. Then, the similarity between the objects in the variable space can be determined. Very often the Euclidean distance is used for clustering purposes. Thus, from the input matrix (raw data) a similarity matrix is calculated. Typical methods for linkage of similar objects into clusters include the single linkage, the complete linkage, the average linkage methods or, very often the Ward's method. The representation of the results of the cluster analysis is performed either by a tree-like scheme called dendrogram comprising a hierarchical structure.

Principal components analysis (PCA) is a typical display method, which allows estimating the internal relations in the data set. There are different variants of PCA but basically, their common feature is that they produce linear combination of the original columns in the data matrix (data set) responsible for the description of the variables characterizing the objects of observation. These linear combinations represent a type of abstract measurements (factors, principal components, PC) being better descriptors of the data struc-

ture (data pattern) than the original measurements. Usually, the new abstract variables are called latent factors and they differ from the original ones. It is a common finding that just a few of the latent variables account for a large part of the data set variation. Thus, the data structure in a reduced space can be observed and studied.

Generally, when analysing a data set consisting of n objects for which m variables have been measured, PCA can extract m principal components PCs (factors or latent variables) where $m < n$. The first PC represents the direction in the data, containing the largest variation. PC 2 is orthogonal to PC 1 and represents the direction of the largest residual variation around PC 1. PC 3 is orthogonal to the first two and represents the direction of the highest, residual variation around the plane formed by PC 1 and PC 2. The projections of the data on the plane of PC 1 and PC 2 can be computed and shown as a plot (score plot). In such a plot it is possible to distinguish similarity groups.

According to the theory of PCA the data matrix is decompose to factor score matrix and factor loadings matrix. The scores on the PCs are the new coordinates of the data space are a weighted sum of the original variables (e.g. chemical concentrations):

Score (value of object *I* along a PC *p*) = $\gamma_{1p}Y_1 + \gamma_{2p}Y_2 + ... + \gamma_{kp}Y_k$

where Y is indication of the variable value (e.g. concentration) and γ are the weights (called loadings). The information hidden in the loadings can also be displayed in loadings plots indicating the grouping of variables. It is important to note that PCA requires very often scaling of the input raw data to eliminate dependence on the scale of the original values.

Results and discussion

The input data matrix was of dimension [96 x 20]. Totally 96 analytical results for 7 different classes of biofuel plants: class 1- sunflower (№ 1-33); class 2 – rapeseed (№ 34- 52); class 3 – corn (№ 53-60); class 4 –soybean (№ 61-79); class 5 – palm (№ 80-85); class 6 - peanut (№ 86-91), class 7 – unknown (№ 92-96) were used as objects of the classification and multivariate statistical interpretation. Twenty variables (biofuel features) were used for the chemometric expertise:

Twelve concentration levels for the different FAME (*C16:0; C16:1; C18:0; C18:1; C18:2; C18:3; C20:0; C20:1; C22:0; C22:1; C24:0; C24:1*)

SAT – sum of esters of saturated fatty acids

MUNS – sum of esters of mono unsaturated fatty acids

PUNS – sum of esters of poly unsaturated fatty acids

FDU – full degree of unsaturation [8]

CN – cetane number [8]

IN – iodine number [8]

LCSF(A) – Long Chain Saturated Factor (A), parameter was calculated from the composition of saturated fatty acids and their corresponding melting point [8]

LCSF(B) – Long Chain Saturated Factor (B), taking into account the composition of saturated fatty acids and lending more weight to the composition of fatty acids with a long chain [8].

Cluster analysis (CA) and principal components analysis (PCA) were applied for data classification and interpretation. The software package STATISTICA 8.0 was used for calculations.

Cluster analysis of biofuel parameters

In *Fig. 1* the hierarchical dendrogram for linkage of variables (z-transformation of input data, squared Euclidean distance as similarity measure and Ward's method of linkage) is presented.

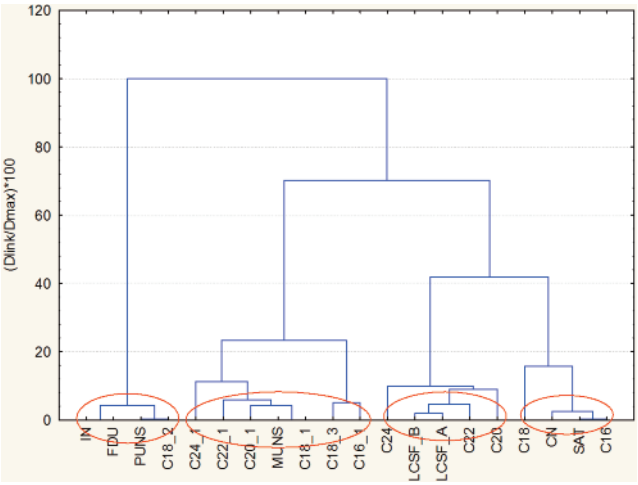

Figure 1. Hierarchical dendrogram for 20 variables

Four clusters are formed as follows:

K1: *IN FDU PUNS C18:2*

K2: *C24:1 C22:1 C20:1 MUNS C18:1 C18:3 C16:1*

K3: *C24:0 C22:0 C20:0 LCSF(A) LCSF(B)*

K4: *C18:0 CN SAT C16:0*

Cluster 1 includes iodine number, level of unsaturation, sum of esters of poly unsaturated acids, FAME concentration of C18:2. It may be stated that this is group of "*polyunsaturation*" factors having a specific impact on the quality of the biofuel.

Cluster 2, respectively, reflects the "*unsaturation*" factor and logically includes the FAME concentrations of the unsaturated acids and the sum of their esters.

Cluster 3 is closely related to cluster 4. Both of them indicate the impact of the "*saturation*" factor which involves the FAME concentrations of the saturated acids, the sum of

their esters, the length of chain and the CN.

It could be stated that three major factors are responsible for the overall quality of biofuels studied: the level (concentration) of saturated acids, the level of unsaturated acids and the level of polyunsaturated acids.

PCA for latent factors identification

The results above could be confirmed or completed by the application of PCA. It is proven by the scree plot (*Fig.2*) that 4 latent factors are responsible for the data structure.

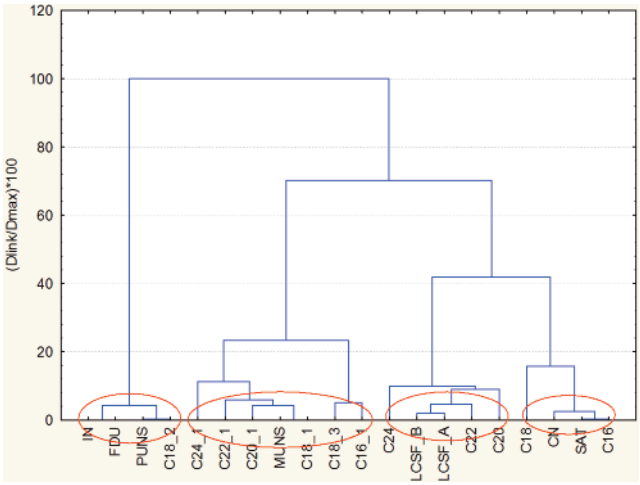

Figure 2. Plot of eigenvalues

In *Table 1* the factor loadings of the major latent factors explaining over 80% of the total variance are presented. The significant loadings are marked.

Table 1. Factor loadings

Variable	PC 1	PC 2	PC 3	PC 4
C16:0	-0.21	**0.96**	-0.005	0.05
C16:1	0.06	0.18	-0.29	**0.87**
C18:0	**-0.74**	0.31	0.03	-0.05
C18:1	**0.94**	0.20	0.11	0.09
C18:2	**-0.75**	-0.59	-0.09	-0.22
C18:3	0.47	-0.18	-0.20	**0.78**
C20:0	0.28	-0.01	**0.76**	0.08
C20:1	**0.83**	-0.05	0.32	0.14
C22:0	0.25	0.02	**0.85**	-0.19
C22:1	**0.85**	-0.11	-0.25	0.23
C24:0	0.07	-0.01	**0.68**	-0.13
C24:1	**0.60**	0.17	0.11	-0.07
SAT	-0.26	**0.94**	0.12	0.02
MUNS	**0.94**	0.19	0.11	0.09
PUNS	**-0.71**	**-0.67**	-0.14	-0.07
FDU	-0.39	**-0.88**	-0.17	0.06
CN	0.14	**0.95**	0.01	0.05
IN	-0.37	**-0.89**	-0.17	0.08
LCSF_A	-0.22	0.20	**0.90**	-0.16
LCSF_B	-0.08	0.50	**0.85**	-0.16
Expl. Var.%	30.50	28.60	18.60	8.40

Again, four latent factors explain the significant part of the total variance of the system (over 85% of the total variance). This number corresponds to the number of clusters found by CA.

The first latent factor PC1 explains 30.5% of the total variance. It indicates strong positive correlation between C18:1, C20:1, C22:1, C24:1 and MUNS and reflects the "*unsaturated*" impact of the variables on the biofuel quality. In this respect it coincides with K2 from CA. The strong negative correlation between C18:0, C18:2, PUNS includes components from cluster K1 being completed by negative correlations of PUNS, FDU, IN in latent factor PC2. Thus, the "*polyunsaturated*" impact from the cluster classification is represented in two latent factors in PCA. PC3 indicates convincingly the "saturated" pattern of variables affecting the biofuel characteristics (strong correlation between C20:0, C22:0, C24:0, LCSF(A) and LCSF(B).

The variable C18:3 could be attributed both to PC1 and PC4, indicating in this way the specific role of this acid to the product quality. It could be stressed that C16:0 and C16:1 also have more complicated position in the analysis and require more detailed interpretation, e.g. C16:1 is stronger correlated to the saturated acids (instead to unsaturated) and C16:0 – to unsaturated rather than to the saturated. Probably, the shorter chain plays a specific role.

These results could be represented in a typical biplot (*Fig. 3*).

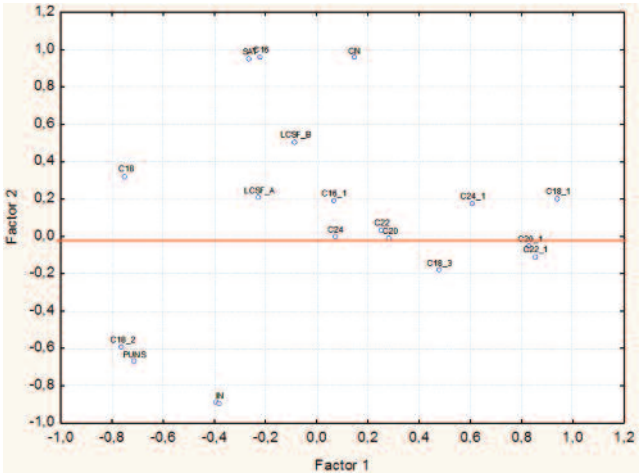

Figure 3. Biplot of factor loadings (PC1 vs. PC2)

In principle, CA and PCA give the same results concerning the factors determining the data structure (or biofuel quality aspects): "*polyunsaturated*", "*unsaturated*" and "*saturated*" factor.

In the next stage of the chemometric analysis it was of interest to classify the different samples (objects) in order to understand the role of the plant specification and to create different patterns of biofuel sources and their specific characteristics (source fingerprints). In *Fig. 4* the hierarchical

dendrogram for all 96 samples is presented (same conditions for hiererchical clustering as in variables clustering).

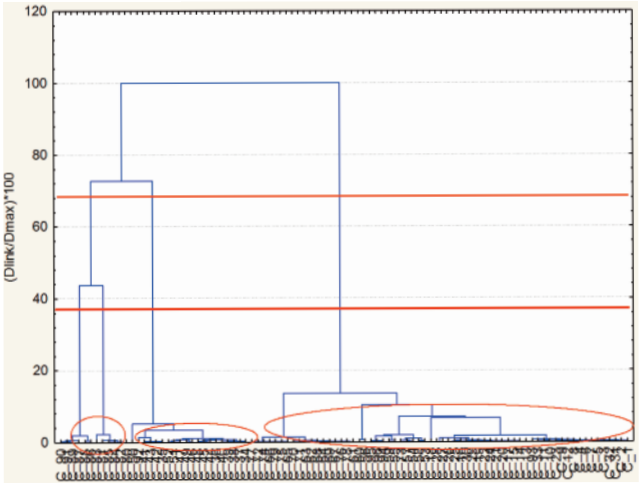

Figure 4. Hierarchical dendrogram for 96 samples (objects)

All samples could be divided into three major clusters. The first one (from the left side) consists of 14 objects (all peanut samples, all palm samples, one rapeseed sample and one from the samples with unknown origin). This cluster forms a **"saturated" pattern** for biofuel since it shows the highest levels of concentration (average values for the cluster) of C16:0, C18:0, C20:0, C22:0, C24:0 and also maximal values of SAT, LCSF(A), LCSF(B) and CN.

Eighteen samples belong to cluster 2 (in the middle of the dendrogram). All of them are rapeseed samples. The cluster is characterized by maximal levels of concentrations of unsaturated acids C16:1, C18:1, C20:1, C22:1 and C18:3 as well as MUNS. It could be concluded that this cluster is representative for a **"unsaturated" pattern** of biofuel.

The rest of the samples (all sunflower, soya and maize samples as well as the other three unknown samples) are included in cluster 3. Maximal values show C18:0, C18:2, PUNS, FDU, IN and in such a way a **"poly unsaturated" pattern** is formed.

It is important to note that the levels of C18:0 (averages for the respective cluster) are statistically equal for cluster 1 and cluster 3.

Our classification made it possible to conclude that probably one of the sample is either peanut or palm since the other four belong to the unspecific group of sunflower – maize – soya sources.

This classification makes it possible to distinguish the origin of the unknown samples: one of them (sample 92) is either peanut or palm. The other four (93, 94, 95, 96) belong to the sunflower – corn – soya pattern.

Conclusion

The classification performed by the use of only these two classical chemometric methods does not allow very specific identification of samples with unknown origin. Very often such samples are complex mixtures of various materials and the precise identification is not possible. Besides, the separation achieved by the chemometric expertise reveals quite unspecific links, eg. sunflower, soya and maize samples for a group of similarity. The same holds true for palm and peanut samples. Only rapeseed samples define a specific pattern. Probably, the choice of variables has to be further optimized to achieve higher selectivity. The selection of optimal number of variables requires application of more advanced chemometric approaches.

Supporting Information: *Supplementary data related to this article can be found at* www.http://chimexpert.com/bjc/

PAPER

References

[1] S.K. Hoekman, A. Broch, C. Robbins, E. Ceniceros, M. *Natarajan, Ren. Sust. Energy Rev.* 16 (**2012**) 143.
[2] Y.C. Sharma, B. Singh, S.N. Upadhyay, *Fuel* 87 (**2008**) 2355.
[3] A.E. Atabania, A.S. Silitonga, I. A. Badruddina, T.M.I. Mahlia, H.H. Masjuki, S. Mekhilef, *Ren. Sust. Energy Rev.* 16 (**2012**) 2070.
[4] D. Leung , X. Wu, M.K.H. Leung, *Appl. Energy* 87 (**2010**) 1083.
[5] A. Demirbas, *Energy Conv. Manag.* 50 (**2009**) 14.
[6] M. Balat, *Energy Conv. Manag* 52 (**2011**) 1479.
[7] M.J. Ramos, C. M. Fernandez, A. Casas, L. Rodriguez, A. Perez, *Biores. Technol.*100 (**2009**) 261.
[8] B.R. Moser, S. F. Vaughn, *Biomass Bioener.* 37 (**2012**) 31.
[9] D.Y.C. Leung, B.C.P. Koo, Y. Guo, *Biores. Technol.* 97 (**2006**) 250.
[10] G. Knothe, *Fuel Process. Technol.* 88 (**2007**) 669.
[11] R.R. Khoury, D. Ebrahimi, L. Hejazy, M.P. Bucknall, R. Pickford, D.B. Hibbert, *Fuel* 90 (**2011**) 2677.
[12] C.P. Prados, D.R. Rezende, L.R. Batista, M.I.R. Alves, N.R.A. Filho, *Fuel* 96 (**2011**) 476.
[13] E.D. Dodds, M.R. McCoy, L.D. Rea, M.J. Kennish, *Lipids* 40 (**2005**) 419.
[14] L. Mondello, A. Casilli, P.Q. Tranchida, R. Costa, B. Chiofalo, P. Dugo, *J. Chromatogr. A* 1035 (**2004**) 237.
[15] F. Adam, F. Bertoncini, V. Coupard, N. Charon, D. Thiebaut, D. Espinat, M.C. Henion, *J. Chromatogr. A* 1186 (**2008**) 236.
[16] E. Bravy, G. Perrety, L. Montanary, *J. Chromatogr. A* 1134 (**2006**) 210.
[17] D. Brodnjak-Voncina, Z.C. Kodba, M. Novich, Chemomet. *Intell. Lab. Syst.* 75 (**2005**) 31.
[18] M. Jalali-Heravi, H. Parastar, H. Sereshti, *Anal. Chim. Acta* 623 (**2008**) 11.
[19] E.G. Giakoumis, *Ren. Energy* 50 (**2013**) 858.
[20] EN 14214:2008 + A1, Automotive fuels – Fatty acid methyl esters (FAME) for diesel engines – Requirements and test methods.
[21] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis. J. Wiley & Sons, New York, **1983**.
[22] G.M. Vandeginste, D.L. Massart. L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers – Verbeke, Handbook of Chemometrics and Qualimetrics; Data Handling in Science and Technology, Elsevier, Amsterdam, **1997**.

Bulg. J. Chem. 3 (**2014**) 95-100

PAPER

Хемометрично разпознаване на биогорива чрез използване на профилите на метилови естери на мастните им киселини

Зилия Мустафа[1], Румяна Милина[1], Павлина Симеонова[2], Васил Симеонов[3]*

[1] Централна Изследователска лаборатория, Факултет по природни науки, Университет „проф. д-р А. Златаров". Бул. Проф. Якимов 1, 8010 Бургас
[2] Лаборатория по физика на околната среда, Институт по физика на твърдото тяло „Акад. Г. Наджаков", БАН, бул. Цариградско шосе 72, 1178 София
[3] Лаборатория по хемометрия и екометрия, Катедра Аналитична химия, Факултет по химия и фармация, Софийски университет „Св. Кл. Охридски, бул. Дж. Баучер 1, 1164 София
** Автор за кореспонденция. E-mail: tadjer@chem.uni-sofia.bg*

Получена: 5 септември 2014
Рецензирана: 1 октомври 2014
Приета: 3 октомври 2014
Излязла online: 31 октомври 2014

Настоящото изследване е начален опит за приложение на хемометрична експертиза към набор от данни от хроматографски анализ на различни източници на биогорива, при използване на профилите на метиловите естери на техните мастни киселини. Използвани са кластерен анализ и анализ на главни компоненти за създаване на специфични образци за всяко едно от растенията, използвани в изследването. Доказани са три основни профила на биогорива, след което е направен опит за коректно разпознаване на принадлежността на неизвестни по произход проби биогориво към някой от получените класове.

Ключови думи: анализ на главни компоненти; биодизел; кластерен анализ; мастни киселини; метилов естер; хемометрия