Available online at www.sciencedirect.com

SCIENCE DIRECT

Chemometrics and Intelligent Laboratory Systems 75 (2005) 31–43

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids

Darinka Brodnjak-Vončina[a,*], Zdenka Cencič Kodba[b], Marjana Novič[c]

[a] *Faculty of Chemistry and Chemical Engineering, University of Maribor, Smetanova 17, SI 2000 Maribor, Slovenia*
[b] *Public Health Institute, Environmental Protection Institute, Prvomajska 1, 2000 Maribor, Slovenia*
[c] *National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia*

Received 3 November 2003; received in revised form 1 April 2004; accepted 22 April 2004
Available online 2 July 2004

Abstract

A novel method for fast and efficient determination of classes of oil samples in routine analyses performed in food control laboratories is proposed. It is based on counterpropagation neural network, which offers a possibility for automatic classification. The fatty acid composition of 132 samples of different edible vegetable oils from the market, including pumpkin, sunflower, peanut, olive, soybean, rapeseed, corn and some mixed oils has been determined. Gas chromatography (GC) was used for qualitative and quantitative determination of methyl esters of palmitic, stearic, oleic, linoleic, linolenic, eicosanoic and eicosenoic acids. The resulting fatty acid composition was evaluated by different chemometrics methods: (i) principal component analysis (PCA), (ii) the clustering method based on the Kohonen neural network, and (iii) the counterpropagation neural network method applied for the classification of oil samples. The first two methods were both successful to distinguish clusters of different oil samples. However, they are not suitable for an automatic prediction of oil classes, because they require a visual inspection of resulting classes and consequently the final decision by an expert. The counterpropagation neural network method was found to be a good model for the classification of different edible vegetable oils on the basis of their composition regarding seven fatty acids. After training with 95 oil samples, the network was able to ascribe correct classes to all samples. The assessment of the model by the leave-one-out cross validation procedure demonstrated 94 correct predictions (i.e. 98.95%) for 95 samples of known type of oil. For a comparison, the linear discriminant analysis (LDA) was performed, which resulted in 92 (i.e. 96.84%) correct predictions. The remaining 37 samples of mixed oils and oils of unknown origin were used to test the model for its ability to predict the composition of mixtures and to show the implementation of the model for predicting the most probable class of an unknown sample.

Keywords: Edible vegetable oils; Fatty acids; Gas chromatography; Principal component analysis; Classification; Kohonen and counterpropagation neural networks

1. Introduction

Chemometrics methods have been often used for the classification and comparison of different vegetable oils. Some examples are, for instance, the differentiation of commercial edible vegetable oils by fatty acids composition [1], differentiation by the signals generated by an electronic nose [2–4], differentiation of olive oils by long-chain alcohols, triterpenes and acids composition [5], and differentiation by headspace-mass spectrometry analysis of volatile compounds [6], characterization of olive oils based on nuclear magnetic resonance determination [7,8], characterization of Calabrian olive oils [9], differentiation of sesame oils [10], and extra virgin olive oils by triacylglycerol composition [11], differentiation of fish oils and fats [12], classification of oil and fat by FT-Raman spectroscopy [13], chemometric resolution of essential oils in Cortex Cinnamomi [14], of essential oils of individual plants in *Hyptis suaveolens* [15,16], of essential oils of *Cistus parviflorus* L. [17], differentiation of redfish species based on tissue fatty acids composition [18], classification of wheat samples [19], and many others.

In the control analytical laboratories, monitoring the food quality one often runs over the problem of screening of various products regarding their composition. The main purpose is to detect samples that deviate from the pre-

* Corresponding author. Tel.: +38-6-2-229-44-32; fax: +38-6-2-252-77-74.
E-mail address: darinka.brodnjak@uni-mb.si (D. Brodnjak-Vončina).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.04.011

scribed parameter values, which are regulated by the state. Here, we propose a fast screening method to control the quality of oil samples present on the market. The quality of the oils was followed by the analysis of fatty acids composition. Gas chromatography was used for qualitative and quantitative determination of methyl esters of palmitic, stearic, oleic, linoleic, linolenic, eicosanoic and eicosenoic acids [20–22]. The ranging of oil samples into seven classes according to the vegetable source was done on the basis of the information from the producers: (1) pumpkin oils, (2) sunflower oils, (3) peanut oils, (4) olive oils, (5) soybean oils, (6) rapeseed oils, and (7) corn oils. Ninety-five oil samples of known declared origin were used to build the model. With different chemometrics methods, the remaining 37 samples of mixed oils and oils of unknown origin were classified in particular classes, regarding fatty acid analyses.

2. Experimental

2.1. Materials

Samples of commercial edible vegetable oils including pumpkin (1), sunflower (2), peanut (3), olive (4), soybean (5), rapeseed (6), corn (7) and some mixed oils were obtained from the market. The FAME Mix Standard, suitable for the determination of rapeseed oil, was obtained from Supelco (Cat. No. 07756). Reagents of recognised analytical grade were used.

2.2. Procedure

The ISO 5508/9 boron trifluoride method was used for the preparation of FAME [20]. The test portion of about 100 mg of the oil sample was introduced into a 50-ml flask. Four milliliters of methanolic sodium hydroxide solution was added for the saponification of glycerides. A reflux condenser was attached to the flask and heated to about 190 °C until the fat droplets disappeared. This procedure usually takes about 5 to 10 min. Five milliliters of boron trifluoride, methanolic solution, was added through the condenser and the boiling continued for some minutes. Three milliliters of heptane was added to the boiling mixture through the top of the condenser. Afterwards, the flask was removed from the heater and the condenser was removed. The saturated sodium chloride solution was added. After shaking it for about 15 s, we allowed the two phases to separate at room temperature. An aliquot of the heptane layer was transferred into a vial and a small amount of anhydrous sodium sulphate was added. The so-obtained heptane solution was prepared for direct injection into gas chromatography (GC).

The ISO 5508/9 method was used to determine the qualitative and quantitative composition of the mixture of FAME.

2.3. Gas chromatography

All GC analyses were performed on a HP6890 (Hewlett-Packard, USA) model equipped with a split/splitless injector and FID. The fused silica capillary column coated with a 0.25-μm film of HP-225, 30 m and 0.25 mm i.d. was used. Nitrogen 6.0 was used as a carrier and makeup gas. Hydrogen flow was set to 30 ml min^{-1} and the air flow to 300 ml min^{-1}. The amount injected was 1 μl and the split ratio was 1:60.

The temperature of the injector and the detector were 230 and 300 °C, respectively. The oven temperature was held at 200 °C for 8 min, and then programmed to 240 °C at 1.3 °C min^{-1} and held there for 0.1 min.

The total amount of palmitic (16:0), stearic (18:0), oleic (18:1), linoleic (18:2), linolenic (18:3), eicosanoic (20:0) and eicosenoic (20:1) acids was calculated by determining the percentage corresponded to the area of FAME. No correction factors or internal standards were used. The mean values of two replicates were accepted for the resulting concentrations of the oil samples (Table 1). The precision of the results is given in the description of the ISO 5508/9 standard method. It says that the difference between the values of two determinations, carried out in a rapid successions by the same operator using the same apparatus on the same test sample and for constituents present in excess of 5% (m/m), shall not exceed 3% (relative) of the determined value, with a maximum of 1% (m/m). For constituents present in smaller quantities, the difference should not exceed value of 0.2% (m/m). Consequently, the precision of the results is limited to one digit. A representative GC chromatogram of fatty acid methyl esters (FAME) from rapeseed oil on the HP 225 fused silica capillary column is shown in Fig. 1. Peaks in ascending retention time order are as follows: C14:0, C16:0, C18:0, C18:1 c9, C18:2 c9c12, C18:3 c9c12c15, C20:0, C20:1 c11, C22:0, C22:1 c13, C24:0.

2.4. Data analysis

The 132 samples of vegetable oils are characterized by the composition of seven fatty acids, namely palmitic, stearic, oleic, linoleic, linolenic, and eicosanoic and eicosenoic acids (see Table 1). The percentage levels of seven fatty acids represent seven variables, components of the vector representation of each sample, further used in chemometrics analysis [23]. The mutual correlation coefficients between all fatty acid percentage levels were calculated. The PCA [23,24] and self-organising artificial neural networks [25] were applied for grouping of oil samples according to the measured fatty acid composition. The Kohonen self organising maps [26] were the most suitable for clustering, while the counterpropagation artificial neural networks (CP ANNs) were used as a modelling and classification method [25,27–30]. All the calculations and plots in the following (PCA) section were done with the Teach/Me software [24]

Table 1
Fatty acid composition (percentage levels) of 132 oil samples

ID	Class[a]	Palmitic	Stearic	Oleic	Linoleic	Linolenic	Eicosanoic	Eicosenoic
1	1	9.7	5.2	31.0	52.7	0.4	0.4	0.1
2	1	11.1	5.0	32.9	49.8	0.3	0.4	0.1
3	1	11.5	5.2	35.0	47.2	0.2	0.4	0.1
4	1	10.0	4.8	30.4	53.5	0.3	0.4	0.1
5	1	12.2	5.0	31.1	50.5	0.3	0.4	0.1
6	1	9.8	4.2	43.0	39.2	2.4	0.4	0.5
7	1	10.5	5.0	31.8	51.3	0.4	0.4	0.1
8	1	10.5	5.0	31.8	51.3	0.4	0.4	0.1
9	1	11.5	5.2	35.0	47.2	0.2	0.4	0.1
10	1	10.0	4.8	30.4	53.5	0.3	0.4	0.1
11	1	11.1	5.0	32.9	49.8	0.3	0.4	0.1
12	1	9.3	4.4	43.3	39.2	2.3	0.4	0.5
13	1	12.2	5.0	31.1	50.5	0.3	0.4	0.1
14	1	9.7	5.6	35.2	47.8	0.5	0.4	0.2
15	1	11.1	4.6	28.3	50.6	4.2	0.4	0.2
16	1	11.0	4.7	29.8	49.6	3.4	0.4	0.2
17	1	11.6	6.0	35.4	45.9	0.2	0.4	0.1
18	1	9.8	5.3	31.7	51.3	0.8	0.4	0.2
19	2	6.1	4.1	24.0	64.3	0.1	0.3	0.1
20	2	6.2	3.9	27.1	59.9	1.3	0.3	0.3
21	2	6.0	4.9	22.8	64.4	0.3	0.3	0.2
22	0	6.7	3.4	35.2	49.3	3.6	0.4	0.5
23	0	3.8	3.9	77.8	12.4	0.2	0.3	0.2
24	0	10.1	4.7	26.1	52.6	5.2	0.4	0.2
25	0	4.9	2.2	58.2	25.2	7.1	0.6	1.2
26	0	6.1	2.7	45.7	37.4	5.9	0.5	0.8
27	2	7.2	4.5	25.6	61.1	0.2	0.3	0.2
28	0	5.9	2.4	55.5	27.7	6.2	0.6	1.1
29	0	10.9	3.8	25.3	53.0	5.6	0.4	0.2
30	2	6.2	4.1	26.8	61.4	0.1	0.3	0.2
31	2	6.1	4.0	25.2	63.2	0.1	0.2	0.2
32	5	10.9	3.6	26.0	52.6	5.5	0.4	0.2
33	5	9.6	3.5	30.3	49.2	5.9	0.4	0.3
34	7	10.0	2.3	36.9	47.1	2.2	0.5	0.5
35	6	5.1	2.3	55.9	27.4	6.8	0.5	0.5
36	3	9.7	3.4	59.3	20.5	0.1	1.5	1.2
37	0	5.2	2.0	59.3	22.7	8.2	0.6	1.3
38	0	8.3	3.5	33.6	45.6	6.3	0.4	0.6
39	2	6.1	4.1	26.7	61.0	0.6	0.3	0.2
40	2	6.2	4.0	25.8	62.2	0.4	0.3	0.2
41	2	6.0	3.8	29.6	57.7	1.2	0.3	0.3
42	2	6.5	2.8	23.2	66.1	0.1	0.2	0.2
43	2	7.0	3.4	23.2	64.9	0.1	0.2	0.2
44	2	6.0	4.0	28.3	60.1	0.1	0.3	0.2
45	1	11.2	6.2	37.7	43.6	0.2	0.5	0.2
46	1	7.6	4.7	29.1	56.1	0.7	0.4	0.2
47	1	7.7	4.7	31.1	54.3	0.8	0.4	0.2
48	1	9.9	5.2	37.4	44.2	2.1	0.4	0.3
49	1	11.5	5.1	27.8	54.5	0.2	0.4	0.1
50	1	11.3	5.8	35.2	46.7	0.2	0.4	0.1
51	4	14.9	2.6	68.2	12.8	0.6	0.4	0.3
52	4	9.3	2.8	65.0	17.0	3.9	0.5	0.7
53	5	10.5	4.2	25.5	52.0	7.8	<0.1	<0.1
54	5	10.0	4.2	24.9	53.2	6.9	0.4	<0.1
55	5	10.4	4.2	25.9	50.8	7.5	0.4	0.4
56	5	10.5	4.2	24.4	52.1	7.5	0.4	<0.1
57	5	10.5	4.3	24.6	53.1	7.6	<0.1	<0.1
58	5	10.2	4.0	23.1	55.1	7.1	0.5	<0.1
59	5	10.9	3.8	27.2	49.5	6.4	0.7	0.9
60	5	11.9	3.8	25.7	52.7	5.8	<0.1	<0.1
61	0	5.1	2.0	57.2	26.1	6.9	0.7	1.4
62	0	4.8	1.9	56.6	25.7	8.1	0.8	1.5

(continued on next page)

Table 1 (*continued*)

ID	Class[a]	Palmitic	Stearic	Oleic	Linoleic	Linolenic	Eicosanoic	Eicosenoic
63	0	10.9	3.2	25.3	53.2	5.8	0.4	0.3
64	0	11.1	3.8	24.4	52.8	6.0	0.4	0.3
65	0	10.4	4.8	24.5	51.8	5.7	0.5	0.5
66	0	11.0	4.3	22.9	54.1	7.6	<0.1	<0.1
67	0	10.8	4.3	22.7	54.6	7.7	<0.1	<0.1
68	0	10.8	4.3	23.1	54.1	7.6	<0.1	<0.1
69	0	8.0	4.3	32.1	44.8	6.4	1.5	1.8
70	0	5.9	5.2	24.7	62.7	0.5	1.1	<0.1
71	0	4.1	4.3	26.1	62.8	0.5	<0.1	<0.1
72	0	3.9	3.3	80.6	11.3	0.1	<0.1	<0.1
73	0	5.1	2.0	58.0	25.0	6.9	0.7	1.4
74	0	5.0	1.8	59.1	23.2	8.0	0.7	1.4
75	4	10.9	2.7	76.7	7.9	0.8	<0.1	<0.1
76	4	10.5	2.8	75.8	8.0	0.7	<0.1	<0.1
77	4	12.0	2.7	75.1	8.5	0.8	<0.1	<0.1
78	4	11.7	2.9	74.6	10.1	0.6	<0.1	<0.1
79	4	11.4	3.0	73.0	10.6	0.7	<0.1	<0.1
80	6	4.8	1.8	62.6	20.0	9.5	<0.1	1.4
81	6	5.5	1.7	59.0	21.3	9.3	0.6	1.5
82	6	5.1	1.9	59.2	22.3	9.3	0.7	1.6
83	6	4.8	1.9	61.6	20.9	8.0	0.8	1.5
84	0	6.2	4.1	31.1	56.7	0.8	0.3	<0.1
85	0	6.2	4.1	31.1	57.3	0.5	0.8	<0.1
86	0	6.7	5.5	24.1	60.8	0.9	0.7	0.7
87	0	8.3	4.0	34.0	52.4	0.3	0.4	<0.1
88	3	10.0	3.3	60.0	21.3	0.2	1.5	1.3
89	3	9.6	3.3	57.7	20.7	0.2	1.5	1.8
90	1	12.2	5.4	29.4	53.0	1.0	<0.1	<0.1
91	1	11.0	5.3	35.0	45.2	1.3	1.3	0.7
92	1	11.9	5.6	33.6	48.9	1.0	<0.1	<0.1
93	1	11.4	5.8	34.5	48.3	1.0	<0.1	<0.1
94	1	10.7	5.4	39.3	43.2	1.4	<0.1	<0.1
95	1	11.2	6.2	35.8	44.1	0.7	1.1	<0.1
96	1	11.4	5.8	33.9	48.1	0.8	0.8	<0.1
97	1	11.5	6.2	39.7	42.6	0.8	<0.1	<0.1
98	1	13.0	6.2	25.8	55.0	0.8	<0.1	<0.1
99	1	13.0	6.7	29.6	50.8	0.5	<0.1	<0.1
100	1	13.1	6.3	26.5	53.6	0.5	0.5	<0.1
101	1	11.6	6.5	38.4	42.8	0.5	0.7	<0.1
102	1	13.1	5.7	31.7	49.5	0.6	<0.1	<0.1
103	2	6.1	4.1	25.1	63.5	0.5	<0.1	<0.1
104	0	5.4	4.2	41.4	48.2	0.5	<0.1	<0.1
105	0	4.6	1.7	61.8	20.3	9.3	0.7	1.5
106	2	6.3	4.2	27.4	61.4	0.8	<0.1	<0.1
107	2	6.2	4.2	27.1	61.8	0.8	<0.1	<0.1
108	2	6.2	4.2	27.0	60.9	0.5	0.3	<0.1
109	2	6.2	4.0	28.3	59.7	0.9	<0.1	<0.1
110	2	5.6	4.2	25.7	58.9	1.7	2.8	0.9
111	2	6.4	3.9	26.0	63.7	0.5	<0.1	<0.1
112	2	6.8	4.3	26.4	60.5	1.3	<0.1	<0.1
113	2	6.0	4.4	27.1	60.9	0.9	<0.1	<0.1
114	2	6.4	4.8	25.3	61.8	1.0	<0.1	<0.1
115	2	5.9	4.5	24.1	61.7	0.9	0.6	0.6
116	2	6.2	4.1	29.9	57.8	1.3	<0.1	<0.1
117	2	6.6	4.7	24.5	62.8	0.3	0.4	<0.1
118	2	6.4	4.4	24.4	63.7	0.4	0.4	<0.1
119	6	5.4	2.0	53.2	28.9	7.3	0.6	1.3
120	6	5.1	1.9	59.2	22.4	9.3	0.6	1.5
121	7	10.7	1.8	30.2	55.5	0.9	0.5	0.3
122	0	7.5	5.1	29.5	56.3	0.1	<0.1	<0.1
123	0	8.0	3.1	40.1	39.6	6.9	0.6	1.0
124	0	8.2	3.4	34.4	45.8	5.7	0.6	0.8
125	0	11.0	3.8	24.6	52.5	6.1	0.5	0.4
126	0	10.5	4.1	22.3	53.6	8.3	0.4	<0.1

Table 1 (*continued*)

ID	Class[a]	Palmitic	Stearic	Oleic	Linoleic	Linolenic	Eicosanoic	Eicosenoic
127	0	9.4	4.7	34.9	49.1	0.2	<0.1	<0.1
128	0	9.8	4.9	34.8	48.4	1.7	0.5	<0.1
129	6	4.5	1.7	64.9	18.6	8.3	<0.1	<0.1
130	6	5.7	2.1	54.6	26.8	8.0	<0.1	<0.1
131	6	6.2	2.2	52.2	29.0	8.0	<0.1	<0.1
132	5	9.7	3.9	25.1	54.2	5.9	<0.1	<0.1

[a] Classes are enumerated as: pumpkin (1), sunflower (2), peanut (3), olive (4), soybean (5), rapeseed (6), corn (7), mixed or unknown type (0) of oil.

using Teach/Me Data Analysis option, which is one of the applications of the Teach/Me system, providing very flexible tools for most fields of data analysis.

3. Results and discussion

3.1. Statistical screening of data

The mutual correlation was sought for all measured variables, i.e. percentage levels of fatty acids. The cross-correlation matrix between the quantities of methyl esters of palmitic, stearic, oleic, linoleic, linolenic, eicosanoic and eicosenoic acid was calculated, see Table 2. The maximal negative correlation coefficient (-0.97) was found between the quantity of oleic and linoleic acid. The relationship between the percentage levels of two highly correlated fatty acids is shown in Fig. 2.

From Fig. 2, it can be seen that the correlation between the oleic (18:1) and linoleic (18:2) acids is high ($r = -0.97$). Already on the basis of one of these two fatty acids, the groups of different types of oils are distinguished. The sunflower and soybean oils, classes 2 and 5, contain the lowest percentage of oleic acid. They can be found considerably above (class 2) and below (class 5) the regression line shown in Fig. 2. Three samples of unknown oil type, marked in Fig. 2 with arrows and the ID numbers, are also significantly above the regression line. It was found in further data analysis that these three samples deviate from all types of oils.

3.2. Principal component analysis (PCA)

PCA [23,24] was performed in order to get an overall impression about the correlation of 132 oil samples described with fatty acids composition. PCA was applied on the matrix composed of 132×7 elements. In 132 rows commercial edible vegetable oils including pumpkin, sunflower, peanut, olive, soybean, rapeseed, corn, and mixed oil samples represented by seven variables are given. The data were additionally preprocessed by "Column centering" (Mean centering), which means that the mean value of each column was subtracted from individual (132) elements. In the first two principal components of transformed data, 97.8% of variance was explained as shown in Table 3.

It is evident from Table 3 that we do not lose a considerable amount of information by keeping only the

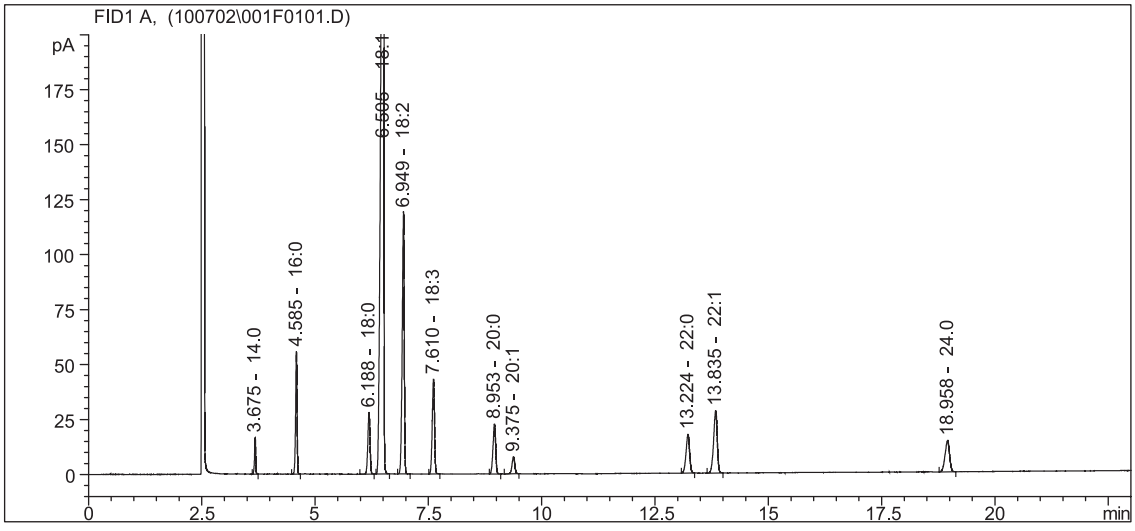

Fig. 1. Representative GC chromatogram of fatty acid methyl esters (FAME) from rapeseed oil (Supelco, cat. no. 07756) on the HP 225 fused silica capillary column. Peaks in ascending retention time order are as follows: C14:0, C16:0, C18:0, C18:1 c9, C18:2 c9c12, C18:3 c9c12c15, C20:0, C20:1 c11, C22:0, C22:1 c13, C24:0.

Table 2
Correlation coefficients between percentage levels of seven fatty acids for 132 samples

	Palmitic	Stearic	Oleic	Linoleic	Linolenic	Eicosanoic	Eicosenoic
Palmitic	1.00						
Stearic	0.53	1.00					
Oleic	− 0.20	− 0.60	1.00				
Linoleic	0.05	0.58	− 0.97	1.00			
Linolenic	− 0.20	− 0.59	0.20	− 0.34	1.00		
Eicosanoic	− 0.09	− 0.09	0.09	− 0.14	0.07	1.00	
Eicosenoic	− 0.38	− 0.58	0.45	− 0.50	0.51	0.60	1.00

first two principal components. The majority of information of 132 oil samples is gathered in the first two principal components. Fig. 3 shows the biplot (PC1 vs. PC2) resulting from PCA of the oil samples represented with seven variables.

It can be seen from Fig. 3 that the first component, PC1, is associated with variables 3 and 4, oleic and linoleic acids. The second component, PC2, represents mainly the linolenic and, to a smaller extent, the palmitic acid (variables no. 5 and 1, respectively). It is evident from Fig. 3 that the clusters are formed, corresponding to seven different oil classes. The samples of mixed or unknown oil type labeled with "0" are distributed into seven classes except for three samples (see arrows on the plot) that are separated from all others. These are samples 23, 72, and 104, which were already found to deviate from the regression line in Fig. 2. For further evaluation of the oil samples, other chemometrical methods, such as Kohonen and counterpropagation neural networks were implemented.

3.3. The Kohonen neural network

(Koh-NN) has been used as a nonlinear mapping method (see Appendix A). Seven-dimensional neurons were arranged in a rectangular grid. Each component of the neuron corresponds to one variable determining the oil samples, i.e. one of the seven fatty acids. Different Koh-NN architectures with 100, 144, and 400 neurons ($10 \times 10 \times 7$, $12 \times 12 \times 7$, $20 \times 20 \times 7$) were constructed. All were checked for the conflicts, which means that two samples of different type fall into the same neuron. No conflicts were observed in any of these networks. The analysis of formed clusters shows, similar to PCA, that the samples are separated into seven distinct groups according

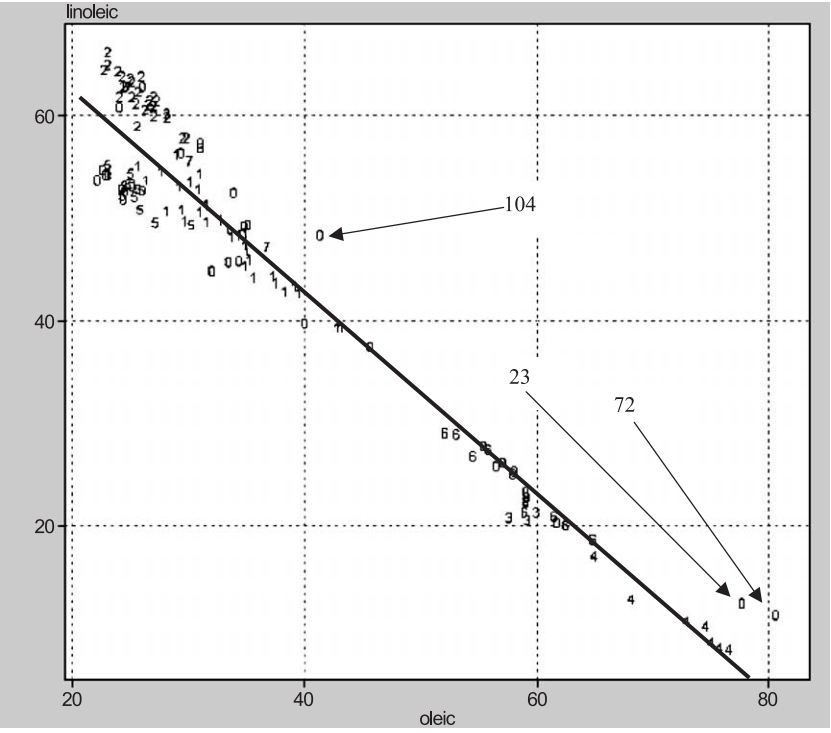

Fig. 2. Relationship between percentage levels of oleic (18:1) and linoleic (18:2) acid in 132 oil samples. Individual samples are indicated by the class numbers as defined in Table 1.

Table 3
The variances in seven principal components obtained after the PCA of 132 samples normalised by mean-centering

PC	%Var	Total
1	95.09	95.09
2	2.72	97.80
3	1.98	99.78
4	0.12	99.90
5	0.07	99.98
6	0.01	99.99
7	0.01	100.00

to seven main vegetable oil classes as defined in Table 1. To illustrate the cluster formation coinciding with the known origin of oil samples, the network of the dimension $12 \times 12 \times 7$ is shown in Fig. 4.

The weights in the *i*-th level of the Kohonen neural network [27,29] correspond to the *i*-th variable of the sample representation vector (Table 1). The normalized (min = 0.0, max = 1.0) weights' surfaces of seven levels in $12 \times 12 \times 7$ dimensional Koh-NN is shown in Fig. 5a–g. The distribution of weights in individual levels coincides with individual clusters, which is evident from the comparison of Fig. 4 with Fig. 5a–g.

The first map (Fig. 5a) represents the distribution of the weights assigned to the palmitic acid. High values are distributed in the direction from the upper left towards the lower right corner of the map, while the two regions

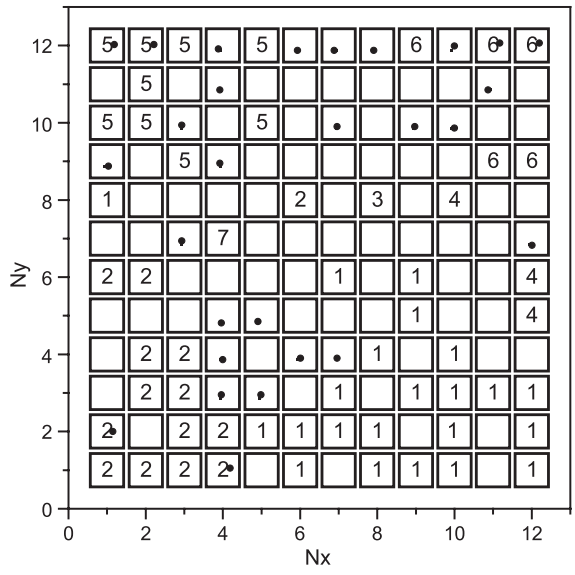

Fig. 4. Top map of the Koh-NN trained with 132 oil samples. The labels "1"–"7" correspond to pre-defined classes of oil samples given in Table 1, while the mixed oil samples and samples of unknown origin are represented by the dots.

of low values are notable in the lower left and upper right corners. Comparing this map with Fig. 4, we can see that the first level of the Koh-NN separates well the sunflower (label 2) and rapeseed (label 6) oils. The distribution of the weights assigned to the stearic acid

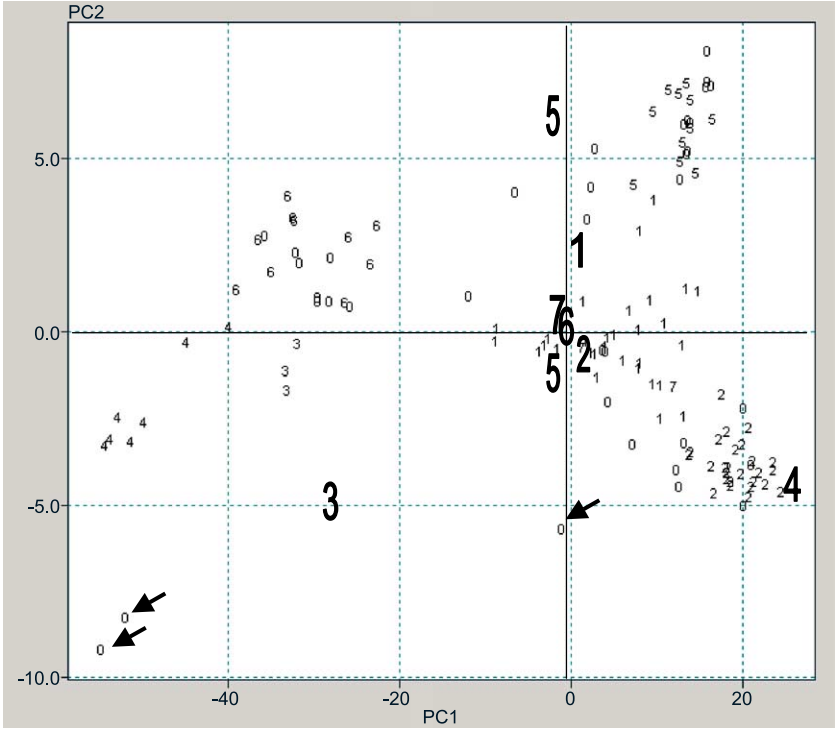

Fig. 3. Biplot (scores and loadings) of 132 samples and 7 variables in the PC1–PC2 co-ordinate system for oil samples labeled with class numbers given in Table 1. Loadings, i.e. seven fatty acids, are printed in bold.

(Fig. 5b) shows a local maximum in the lower-right corner, which coincides with the cluster of pumpkin oils (label 1 in Fig. 4). The third map (Fig. 5c) shows a local maximum in the middle of the right edge, which covers the olive oil samples (label 4 in Fig. 4). Fig. 5d shows a local minimum in the middle of the right edge, which embraces the olive oil samples (label 4), while the high values region coincides with the sunflower oil samples (label 2 in Fig. 4). The fifth level of the Koh-NN (Fig. 5e) separates soybean and rapeseed oils (labels 5 and 6 in Fig. 4) from the rest of the oil samples. The last two maps, representing low concentration eicosanoic and eicosenoic acids, form a local maximum approximately in the center of the map and seem to correlate with the positions of few samples of corn and peanut oils (labels 7 and 3 in Fig. 4, respectively). Taking into account the combination of information contained in all seven maps, i.e. levels of the Koh-NN, we assumed that the model based on the Kohonen and counterpropagation neural network algorithm described in Appendix A will be able to classify the olive oil samples.

3.4. Prediction of determined classes

Among 132 samples that were analysed for fatty acids composition, 95 samples of known origin were classified into seven main vegetable oil classes given in Table 1. Classes "3" and "7", peanut and corn oil were poorly represented (only three samples and two samples, respectively). It has to be stressed that each sample represented a different product and these two sorts of oil (peanut and corn) are rare on Slovenian market. With 95 samples of known origin, the counterpropagation artificial neural network (CP ANN, see Appendix A) [26,29] model was built. The seven-dimensional neurons were placed in the $N \times N$ network. The dimension N varied from 10

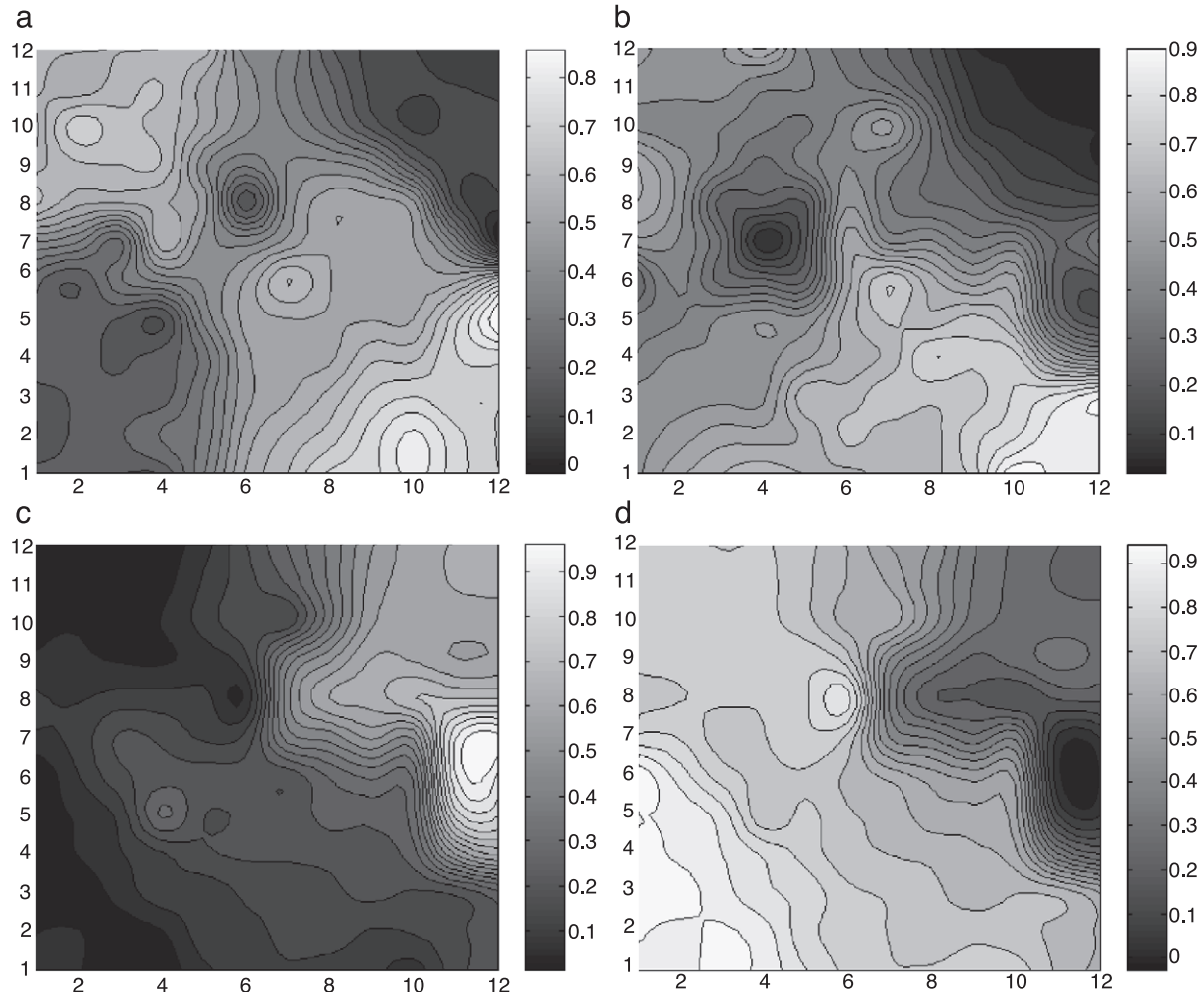

Fig. 5. Seven levels of Koh-NN. Each of the seven maps (a–g) represents the distribution of weights corresponding to one of the seven variables (percentage level of a fatty acid, see Table 1); (a) var. 1, palmitic acid; (b) var. 2, stearic acid; (c) var. 3, oleic acid; (d) var. 4, linoleic acid; (e) var. 5, linolenic acid; (f) var. 6, eicosanoic acid; (g) var. 7, eicosenoic acid.

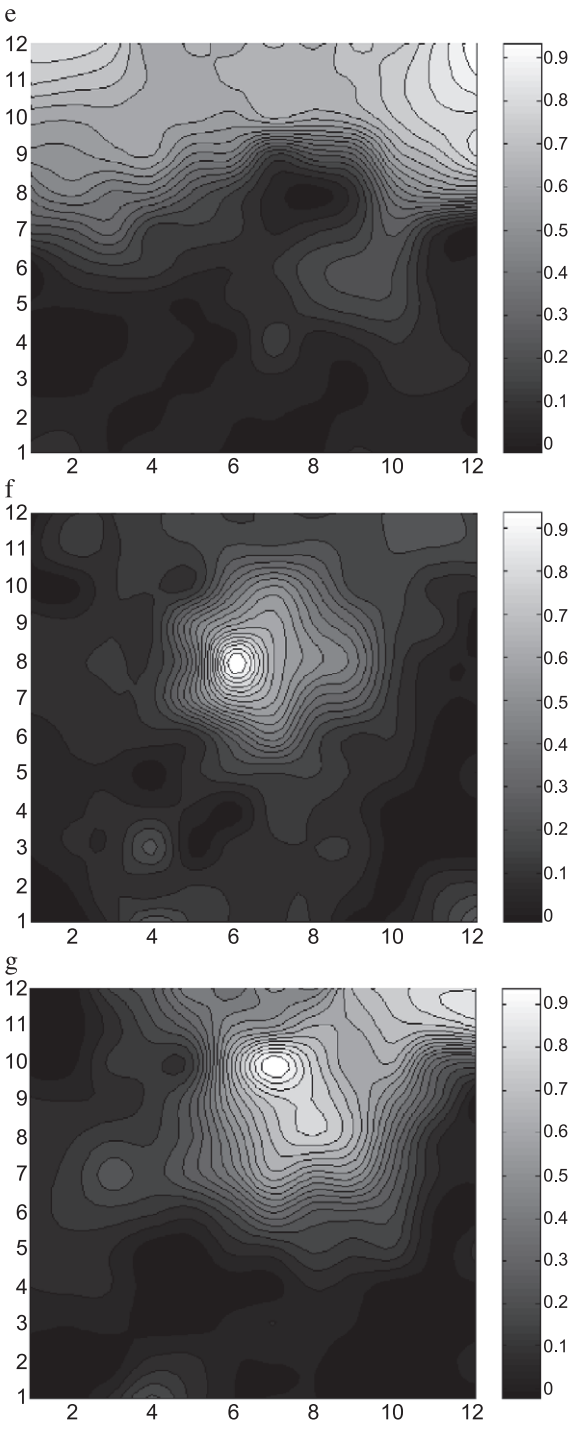

Fig. 5 (*continued*).

to 20. The CP ANNs were trained for 200 epochs, which was sufficient for a satisfactory recognition of the training samples. The seven components of each sample's vector representation are fatty acids composition described in Experimental. The maximal and minimal correction factors in the modelling procedure were 0.5 and 0.01, respectively.

For one particular sample, for example the sample with ID 16 (pumpkin oil, class "1", see Table 1), the CP ANN model ($20 \times 20 \times 7$) gave seven probability values for seven classes: 0.920, 0.001, 0.000, 0.000, 0.079, 0.000, and 0.001, classes "1" to "7", respectively. The class predicted with the highest probability is accepted as the classification result. For the sample with ID 16, the class "1" was correctly predicted, as the value 0.920 is the largest among seven probability values. Evidently, the sum of all probabilities obtained for one sample is 1.0.

The cumulative probability values of the 95 samples obtained by the leave-one-out cross validation of three networks, $10 \times 10 \times 7$, $12 \times 12 \times 7$, and $20 \times 20 \times 7$ are shown in Table 4. Seven probability values are obtained for each sample. The prediction for individual sample is treated as a correct one, if the highest probability value is obtained for the class, to which the sample belongs, as described in the above example for sample no. 16. For 94 out of 95 samples the correct classes were predicted, which is an overall 98.95% correct prediction result. The number of samples belonging to the class "3" and "7" is small, nevertheless, these results were included in Table 4 and show that the CP ANN classification is capable to discriminate between classes also if trained with only one sample. However, the prediction results of these two classes cannot be statistically evaluated.

To compare the classification results obtained with CP ANN, an alternative method, namely Linear Discriminant Analysis (LDA) was applied to 95 samples of known vegetable oil classes. The LDA model was developed to discriminate between seven classes with five predictor variables. The five discriminating functions with *P*-values less than 0.05 were statistically significant at the 95% confidence level. The correct predictions were obtained for 92 oil samples, i.e. 96.84%. With 98.95% correct predictions, the CP ANN classification model was selected for further routine analyses.

Thirty-seven samples of mixed oils and oils of unknown origin, for which the class was neither known nor determined experimentally, were presented to the constructed CP ANN model and the predicted classes were then analysed (see Table 5). Two samples were declared by the manufacturer as sunflower oils, (ID 104, 105), but appeared as an outlier in PCA analysis or produced conflicting situation in Kohonen network analysis, respectively. Since there is no exact information about oil classes for these 37 samples, the quality of the prediction results cannot be confirmed. However, the information about the compositions of mixtures (usually oils of two different classes) correlates well with the model predictions, because the model offers the categorization of one sample into several classes with different probabilities. The cumulative probabilities of seven classes for one sample are equal to 1.0. A detailed study of the predictions from Table 5 was made. Among 37 samples, 6

Table 4
Probability values of 95 known oil samples obtained by leave-one-out cross validation of three different CP ANN models

Class probability value									Correctly predicted samples[a]
Net	Class	1	2	3	4	5	6	7	
10 × 10	1	36.707	0.108	0	0	0.154	0.03	0	37
	2	1.069	23.913	0	0	0	0.018	0	24
	3	0	0	2.908	0.092	0	0	0	3
	4	0.002	0.002	0.485	6.017	0	0.493	0	6
	5	0.214	0.21	0	0	10.554	0.022	0	11
	6	0	0	0	0	0.255	9.603	0.142	10
	7	0.071	0	0	0	0	0	1.928	2
									Σ 93, 97.89%
12 × 12	1	36.555	0.067	0	0	0.37	0	0.001	37
	2	0.672	24.206	0.121	0	0.001	0	0	24
	3		0	2.969	0.016	0.009	0	0.006	3
	4	0	0	0.273	6.141	0	0.586	0	6
	5	0.356	0	0	0	10.641	0	0	11
	6	0.336	0.074	0	0.153	0.091	9.349	0	10
	7	0.188	0.114	0	0.007	0.005	0.049	1.642	2
									Σ 93, 97.89%
20 × 20	1	36.198	0.147	0.251	0	0.4	0	0.01	37
	2	0.706	24.261	0.01	0	0.001	0	0.024	24
	3	0.003	0	2.945	0.042	0	0	0.001	3
	4	0	0	0.029	6.72	0	0.251	0.002	7
	5	0.25	0	0	0	10.734	0.022	0	11
	6	0.033	0	0	0.075	0.212	9.672	0	10
	7	0.29	0	0	0	0.017	0.057	1.636	2
									Σ 94, 98.95%

A percentage of correct predictions is obtained considering 95 known oil samples in the data set.
[a] Number of correctly predicted samples is obtained from the class probability value obtained by LOO CV of individual sample.

were known to contain from 10% to 80% of pumpkin oil (ID 37, 122, 125, 126, 127, 128). Four of them were recognized by the model as pumpkin oils (class 1) with the probabilities from 0.24 to 0.98. The remaining predicted classes with non-zero probabilities of these four samples were "2", "5", and "6", i.e. sunflower, soybean, and rapeseed oils. The sample with ID 124, a mixture of soybean and rapeseed oils, was predicted as class "5" and "6" with probabilities 0.73 and 0.20, respectively, which is in good agreement with the declared composition.

4. Conclusions

The goal of the research presented in this work was to develop and implement an automated method for classification of oil samples in routine food control laboratories. The resulting model is based on the correlation found between the sort of oil (class) and the fatty acid composition. The Principal Component Analysis was used for screening of the data. It was shown that it is necessary to use mean centering of variables. From the results, it was concluded that the PCA method is discriminant enough. 97.8% variance was explained in the first two principal components. The analysis has shown that the variables with the greatest discriminating power were the percentage levels of the oleic and the linoleic acids. A high correlation between these two variables was found for all oil samples. However, slightly different slopes of the regression lines could be assigned to the classes containing the lowest percentage of oleic acid (sunflower and soybean oils). From the PCA and correlation analysis, three outliers were distinguished. These were the oils of an unknown sort (ID 23, 72 and 104), deviating in composition from all other oils.

The artificial neural networks were implemented as the method for clustering of oil samples using Kohonen neural network. For the prediction of oil classes, the counter-propagation neural network, which offers a possibility for automatic classification, was used.

The experience-based CP ANN model was built using the oil samples for which the origin was known. Except for the oil classes "3"and "7", which were poorly represented by three and two samples, respectively, the statistical evaluation of classification results was satisfactory. By the constructed model, the mixed oils and the oils of unknown origin were classified. The model yields the categorization of one sample into several classes with different probabilities, which enables the prediction of mixture composition. The obtained predictions correlate well with the available information of mixed oil samples. The overview of predicted results indicates that the proposed model is of significant value for the determination of unknown oil samples and could be implemented as a fast and efficient method in routine

Table 5
Predictions of 37 unknown oil samples obtained from 20 × 20 × 7 CP ANN model

Description of oil origin	No.	ID	"1"	"2"	"3"	"4"	"5"	"6"	"7"
Vegetable	1	22	0.029	0.741	0.000	0.000	0.230	0.000	0.000
Vegetable	2	23	0.000	0.000	0.000	1.000	0.000	0.000	0.000
Vegetable	3	24	0.730	0.000	0.000	0.000	0.270	0.000	0.000
Vegetable	4	25	0.000	0.000	0.000	0.000	0.020	0.980	0.000
Vegetable	5	26	0.121	0.000	0.000	0.003	0.000	0.571	0.304
Vegetable	6	27	0.195	0.805	0.000	0.000	0.000	0.000	0.000
Vegetable unrefined	7	28	0.006	0.000	0.000	0.000	0.002	0.804	0.187
Vegetable unrefined	8	29	0.014	0.000	0.000	0.000	0.987	0.000	0.000
Mixed pumpkin, vegetable	9	37	0.000	0.000	0.000	0.000	0.000	1.000	0.000
Vegetable	10	38	0.000	0.000	0.000	0.000	0.816	0.178	0.007
Vegetable	11	61	0.002	0.000	0.000	0.000	0.001	0.980	0.017
Vegetable	12	62	0.023	0.000	0.000	0.001	0.000	0.958	0.019
Vegetable	13	63	0.014	0.000	0.000	0.000	0.987	0.000	0.000
Vegetable	14	64	0.124	0.000	0.000	0.000	0.877	0.000	0.000
Vegetable	15	65	0.382	0.000	0.000	0.000	0.611	0.002	0.005
Vegetable	16	66	0.000	0.000	0.000	0.000	1.000	0.000	0.000
Vegetable	17	67	0.000	0.000	0.000	0.000	1.000	0.000	0.000
Vegetable	18	68	0.000	0.000	0.000	0.000	1.000	0.000	0.000
Vegetable	19	69	0.000	0.000	0.000	0.000	0.746	0.216	0.039
Vegetable	20	70	0.030	0.970	0.000	0.000	0.000	0.000	0.000
Vegetable	21	71	0.000	1.000	0.000	0.000	0.000	0.000	0.000
Vegetable	22	72	0.000	0.000	0.000	1.000	0.000	0.000	0.000
Vegetable	23	73	0.066	0.000	0.000	0.040	0.000	0.870	0.023
Vegetable	24	74	0.023	0.000	0.000	0.001	0.000	0.958	0.019
Oil for frying	25	84	0.000	1.000	0.000	0.000	0.000	0.000	0.000
Unrefined	26	85	0.195	0.805	0.000	0.000	0.000	0.000	0.000
Oil for frying	27	86	0.000	1.000	0.000	0.000	0.000	0.000	0.000
Oil for frying	28	87	0.974	0.026	0.000	0.000	0.000	0.000	0.000
Sunflower	29	104	0.000	1.000	0.000	0.000	0.000	0.000	0.000
Sunflower	30	105	0.000	0.000	0.000	0.001	0.000	0.999	0.000
Mixed 10% pumpkin, 90% vegetable	31	122	0.319	0.681	0.000	0.000	0.000	0.000	0.000
Mixed 50% soybean, 50% sunflower	32	123	0.000	0.000	0.000	0.000	0.438	0.524	0.038
Mixed 70% soybean, 30% rapeseed	33	124	0.002	0.000	0.000	0.000	0.732	0.202	0.065
Mixed 80% pumpkin, 20% sunflower	34	125	0.240	0.000	0.000	0.000	0.760	0.000	0.000
Mixed 20% pumpkin, 80% vegetable	35	126	0.000	0.000	0.000	0.000	1.000	0.000	0.000
Mixed 50% pumpkin, 50% vegetable	36	127	0.883	0.117	0.000	0.000	0.000	0.000	0.000
Mixed 50% pumpkin, 50% vegetable	37	128	0.977	0.000	0.000	0.000	0.000	0.000	0.023

analysis. It is worth mentioning that oil samples of suspicious origin were easily detected as outliers.

Acknowledgements

The authors thank the Ministry of Education, Science and Sport of Republic of Slovenia through the project grants P1-017 and PO-501 for financial support.

Appendix A. Kohonen and counterpropagation neural networks

Kohonen neural network (Koh-NN) [25,26] with its architecture and learning strategy imitates the structure of the brains, so called biological neuron network. Koh-NN is based on a single layer of neurons arranged in a two-dimensional plane having a well-defined topology of neurons with a defined neighbourhood structure. In present research the Koh-NN with 8-neighbours structure is used [25], the N^2 neurons produce a 2D top-map of dimension $N \times N$. All the neurons obtain the same multidimensional input, however, the response is localized to a small number of neurons in an area of topological neighbourhood. This so-called local feed-back makes the Koh-NN similar to the biological neural network. The training of Koh-NN is based on competitive learning, 'the-winner-takes-all' strategy. It means that input presented to the network activates only one neuron from the entire net of neurons which is then stimulated by an appropriate correction of weights. The neurons are competing with each other to gain the stimulation.

The selection of the winning neuron W_c (c for central) is made on the basis of the weight vector W_j (w_{j1},w_{j2},...w_{jm}) most similar to the input signal X_s (x_{s1},x_{s2},...x_{sm}):

$$\delta_j = \sum_{i=1}^{m} (x_{si} - w_{ji})^2 \quad \delta_c = \min(\delta_1, \delta_2, \ldots \delta_j, \ldots \delta_n,) \Rightarrow W_c \tag{A-1}$$

The neuron c fulfils the selected criterion of the best match between the neuron's weights and the input vector components. The weights of the c-th neuron, w_{ci} are then corrected in such a way that the response will be even closer to the optimal one ($\delta_j = 0$) for the given criterion in particular Koh-NN application. Besides the correction of the winning (central) neuron, also the weights of the neighbouring neurons are corrected. The following equation defines the corrections:

$$w_{ji}^{\text{new}} = w_{ji}^{\text{old}} + \eta(t) \times b(d_c - d_j) \times (x_{si} - w_{ji}^{\text{old}}) \qquad (A-2)$$

Parameter η determines the rate of learning; it is maximal at the beginning ($t = 1$, $\eta = a_{\max}$) and minimal at the end of the Koh-NN learning procedure ($t = t_{\max}$, $\eta = a_{\min}$). The function $b(.)$ in Eq. (A-2) describes how the correction of the weights w_{ji} decreases with increasing topological distance between the central neuron and the neuron being corrected. Index j specifies individual neuron and runs from 1 to n. Topological distance of the j-th neuron from the central one is defined according to the topology used for the distribution of neurons in the plane: in the rectangle net the central neuron has eight first neighbours ($d_c - d_j = 1$), 16 second neighbours ($d_c - d_j = 2$), etc. The minimal distance is zero ($j = c$, $d_c - d_j = 0$), which corresponds to the maximal correction function ($b = 1$). The maximal distance ($d_c - d_{\max}$) to which the correction is applied is shrinking during the learning procedure. The correction function at maximal distance is minimal ($b = 0$). At the beginning, the $d_c - d_{\max}$ covers the entire network, while at the end, at $t = t_{\max}$, it is limited only to the central neuron.

Counterpropagation Neural Network (CP ANN) [31,32] modelling is based on two-step learning procedure, which is unsupervised in the first step. The first step corresponds to the mapping of objects into the input or so called Kohonen layer, because it is identical to the Koh-NN learning procedure described above. The second step of the learning is supervised, which means that for the learning procedure the response or target value is required for each input. The network is thus trained with a set of input-target pairs $\{X_s, T_s\}$. For the implementation of CP ANN in the chemometrical treatment of oil samples, the input $X_s = (x_{s1}, x_{s2}, \ldots x_{si} \ldots x_{sm})$ represents the fatty acid composition of the s-th oil sample described by m fatty acid percentage levels or variables. In the application of CP ANN for the classification of oil samples, the corresponding target $T_s = (t_{s1})$ is seven-component binary vector indicating one of the seven possible classes of the s-th sample. The training of the network means adjusting the weights of the neurons in such a way that for each input sample X_s from the training set the network would respond with the output Out_s identical to the target T_s. The training is an iterative procedure involving the feeding of all input–output pairs $\{X_s, T_s\}$ to the network and correcting the weights of the neurons according to the differences between the targets and current outputs ($T_s - \text{Out}_s$). As already stressed, the targets are needed only in the last part of each iterative learning step. The unsupervised element in the CP ANN learning procedure is the mapping of the oil sample vectors into the Kohonen layer of the CP ANN which is based solely on the fatty acid composition. For this step, no knowledge about the target vector (oil class) is needed. Once the position (central neuron c) of the input vector is defined, the weights of the input and output layer of the CP ANN are corrected accordingly. The following equation (Eq. (A-3)) defines the corrections in the output layer, while the corrections of the weights in the input layer were given above, in Eq. (A-2).

$$\text{out}_{ji}^{\text{new}} = \text{out}_{ji}^{\text{old}} + \eta(t) \times b(d_c - d_j) \times (T_{si} - \text{out}_{ji}^{\text{old}}) \qquad (A-3)$$

References

[1] D.S. Lee, B.S. Noh, S.Y. Bae, K. Kim, Anal. Chim. Acta 358 (1998) 163–175.
[2] Y.G. Martin, J.L.P. Pavon, B.M. Cordero, C.G. Pinto, Anal. Chim. Acta 384 (1999) 83–94.
[3] Y.G. Martin, M.C.C. Oliveros, J.L.P. Pavon, C.G. Pinto, B.M. Cordero, Anal. Chim. Acta 449 (2001) 69–80.
[4] M.C.C. Oliveros, J.L.P. Pavon, C.G. Pinto, M.E.F. Laespada, B.M. Cordero, M. Forina, Anal. Chim. Acta 459 (2002) 219–228.
[5] G. Bianchi, L. Giansante, A. Shaw, D.B. Kell, Eur. J. Lipid Sci. Technol. 103 (2001) 141–150.
[6] I.M. Lorenzo, J.L.P. Pavon, M.E.F. Laespada, C.G. Pinto, B.M. Cordero, J. Chromatogr., A 945 (2002) 221–230.
[7] A. Sacco, M.A. Brescia, V. Liuzzi, F. Reniero, C. Guillou, S. Ghelli, P. van der Meer, J. Am. Oil Chem. Soc. 77 (2000) 619–625.
[8] A.D. Shaw, A. diCamillo, G. Vlahov, A. Jones, G. Bianchi, J. Rowland, D.B. Kell, Anal. Chim. Acta 348 (1997) 357–374.
[9] S. Lanteri, C. Armanino, E. Perri, A. Palopoli, Food Chem. 76 (2002) 501–507.
[10] D.S. Lee, E.S. Lee, H.J. Kim, S.O. Kim, K. Kim, Anal. Chim. Acta 429 (2001) 321–330.
[11] P. Damiani, L. Cossignani, M.S. Simonetti, B. Campisi, L. Favretto, L.G. Favretto, J. Chromatogr., A 758 (1997) 109–116.
[12] S. Dejong, Mikrochim. Acta 2 (1991) 93–101.
[13] V. Baeten, P. Hourant, M.T. Morales, R. Aparicio, J. Agric. Food Chem. 46 (1998) 2638–2646.
[14] F. Gong, L.Z. Liang, Q.S. Xu, F.T. Chau, J. Chromatogr., A 905 (2001) 193–205.
[15] N.R. Azevedo, I.F.P. Campos, H.D. Ferreira, T.A. Portes, J.C. Seraphin, de Paula, J.R. de Paula, S.C. Santos, P.H. Ferri, Biochem. Syst. Ecol. 30 (2002) 205–216.
[16] N.R. Azevedo, I.F.P. Campos, H.D. Ferreira, T.A. Portes, S.C. Santos, J.C. Seraphin, J.R. Paula, P.H. Ferri, Phytochemistry 57 (2001) 733–736.
[17] D. Angelopoulou, C. Demetzos, D. Perdetzoglou, Biochem. Syst. Ecol. 29 (2001) 405–415.
[18] H. Joensen, O.G. Nielsen, Comp. Biochem. Phys., B 129 (2001) 73–85.
[19] C. Armanino, R. De Acutis, M.R. Festa, Anal. Chim. Acta 454 (2002) 315–326.
[20] Animal and vegetable fats and oils, ISO 12228: 1999 (E).
[21] S. de Koning, B. van der Meer, G. Alkema, H.G. Janssen, U.A.T. Brinkman, J. Chromatogr., A 922 (2001) 391–397.
[22] M.B. Oliveira, M.R. Alves, M.A. Ferreira, J. Chemom. 15 (2001) 71–84.
[23] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997, pp. 519–556.

[24] Teach/Me, SDL—Software Development Lohninger; Teach/Me Data-Lab 2.002 © 1999 Springer, Berlin, Developed by H. Lohninger and the Teach/Me people.
[25] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, 2nd ed., Wiley-VCH, Weinheim, 1999, etc.
[26] T. Kohonen, Self-Organization and Associative Memory, Springer-Verlag, Berlin, 1988.
[27] J. Lozano, M. Novič, F.X. Rius, J. Zupan, Chemometr. Intell. Lab. Syst. 28 (1995) 61–72.
[28] J. Zupan, M. Novič, I. Ruisanchez, Chemometr. Intell. Lab. Syst. 38 (1997) 1–23.
[29] J. Zupan, M. Novič, J. Gasteiger, Chemometr. Intell. Lab. Syst. 27 (1995) 175–187.
[30] N. Majcen, K. Rajer-Kanduč, M. Novič, J. Zupan, Anal. Chem. 67 (1995) 2154–2161.
[31] R. Hecht-Nielsen, Counterpropagation networks, Appl. Opt. 26 (1987) 4979–4984.
[32] J. Dayhof, Neural Network Architectures, An Introduction, Van Nostrand Reinhold, New York, 1990, p. 192.