Contents lists available at ScienceDirect

Food Chemistry

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

Analytical Methods

Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques

I. Stanimirova [a], B. Üstün [b], T. Cajka [c], K. Riddelova [c], J. Hajslova [c], L.M.C. Buydens [b], B. Walczak [a,*]

[a] *Institute of Chemistry, Department of Chemometrics, Silesian University, 9 Szkolna Street, Katowice 40-006, Poland*
[b] *Institute of Molecules and Materials, Department of Analytical Chemistry, Radboud University Nijmegen, Heyendaalseweg 135, 6525AJ Nijmegen, The Netherlands*
[c] *Institute of Chemical Technology, Prague, Faculty of Food and Biochemical Technology, Department of Food Chemistry and Analysis, Technicka 3, 166 28 Prague 6, Czech Republic*

ARTICLE INFO

Article history:
Received 3 March 2008
Received in revised form 13 November 2008
Accepted 21 April 2009

Keywords:
Food analysis
Solid phase microextraction
Comprehensive two-dimensional gas chromatography
Time-of-flight mass spectrometry
Honey volatiles
Classification
LDA
DPLS
SIMCA
SVM

ABSTRACT

The goal of this study was to examine the possibility of verifying the geographical origin of honeys based on the profiles of volatile compounds. A head-space solid phase microextraction (SPME) combined with comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC–TOF-MS) was used to analyze the volatiles in honeys with various geographical and floral origins. Once the analytical data were collected, supervised pattern recognition techniques were applied to construct classification/discrimination rules to predict the origin of samples on the basis of their profiles of volatile compounds. Specifically, linear discriminant analysis (LDA), soft independent modeling of class analogies (SIMCA), discriminant partial least squares (DPLS) and support vector machines (SVM) with the recently proposed Pearson VII universal kernel (PUK) were used in our study to discriminate between Corsican and non-Corsican honeys. Although DPLS and LDA provided models with high sensitivities and specificities, the best performance was achieved by the SVM using PUK. The results of this study demonstrated that GC × GC–TOFMS combined with methods like LDA, DPLS and SVM can be successfully applied to detect mislabeling of Corsican honeys.

1. Introduction

Nowadays, the identification of the origin of food is one of the most important issues in food chemistry and in food quality control. Recently the European Commission has decided to introduce regulations for labeling food commodities. The origin of food, together with its main ingredients, should be available to the consumer. These regulations aim to guarantee product quality, safety, authenticity and to protect the rights of consumers. In light of the new regulations, there is an urgent demand to deliver cost-effective procedures for detecting fraudulent products and checking compliance with the quality specifications. This is essentially the goal of the EU-funded project "TRACE" (Tracing food commodities in Europe, www.trace.eu.org). This paper describes a part of the research done within the TRACE project on the commodity of honey.

Honey is a natural product, the consumption of which has increased in recent years. The traditional approach to recognizing the botanical origin of honey relies on a microscopic examination of its pollen (melissopalynology) (Anklam, 1998; Soria, González, de Lorenzo, Martínez-Castro, & Sanz, 2004). Melissopalynology can also be used to identify the geographical origin of honey if the pollen is specific enough in the area of interest. However, this method is expensive, time-consuming and strongly dependent on the qualifications and judgment of the analyst. Therefore, there is a tendency to replace pollen analysis by finding analytical and/or physicochemical markers for honey discrimination. Minerals and trace elements (Fernández-Torres et al., 2005; Hernández, Fraga, Jiménez, Jiménez, & Arias, 2005; Latorre et al., 1999), volatile compounds (Guyot, Scheirman, & Collin, 1999; Radovic et al., 2001; Soria et al., 2004), the protein pattern, flavonoids, physicochemical parameters like electrical conductivity, pH, total acidity and water activity (Acquarone, Buera, & Elizalde, 2007; Corbella & Cozzolino, 2006; Devillers, Morlot, Pham-Delegue, & Dore, 2004; Marini, Magrì, Balestieri, Fabretti, & Marini, 2004) are some of the parameters that have been extensively examined for the recognition of the floral and geographical origin of honeys.

Although the volatile composition has mainly been used for the characterization of the floral source of honey, the aim of this work was to evaluate whether the volatile profile also allowed for the identification of the geographical origin of honey. Specifically, its

[*] Corresponding author. Tel.: +48 32 359 2115; fax: +48 32 259 9978.
E-mail address: beata@us.edu.pl (B. Walczak).

0308-8146/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodchem.2009.04.079

use for discriminating between Corsican honeys and honeys from other geographical regions was studied. The reason for choosing Corsican honeys was to demonstrate the usefulness of the volatile profiles for tracing the origin of a PDO (protected designation of origin) commodity. The PDO label guarantees that the product consumers purchase is prepared, processed and produced in a specific region and consequently possesses unique properties. However, an emerging authenticity problem is the labeling of a non-PDO product as PDO. In other words, the aim of our study is to develop a reliable methodology to detect such a fraud in Corsican honey, i.e. a strategy for the successful differentiation of Corsican honeys from honeys originating from other regions (PDO or non-PDO) in Europe.

To meet the goal of the study, a large number of honey samples were collected in a two-year campaign (2006 and 2007) and their volatile constituents were analyzed using head-space solid phase microextraction (SPME) sampling followed by the separation/identification step employing comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC–TOFMS). This approach offers a quick (19 min GC run vs. 30–90 min using one-dimensional GC) and comprehensive analysis of honey volatiles, while minimizing the risk of their erroneous identification, which cannot be avoided in one-dimensional GC separation (Čajka, Hajšlová, Cochran, Holadová, & Klimánková, 2007).

A verification of the origin of food and/or checking the authenticity can be facilitated using chemometric approaches. The goal of supervised pattern recognition techniques is to create classification/discrimination rules using a set of training samples of a known origin and then using the created rules to predict the belongingness of new samples of an unknown origin to the available classes. Since the choice of a classification/discrimination method is strongly data dependent, performances of several pattern recognition methods (Berrueta, Alonso-Salces, & Héberger, 2007; González, 2007), which cover different aspects of data investigation, were evaluated for the two-class problem studied. These were linear discriminant analysis (LDA) (Fisher, 1936; Lebart, Morineau, & Warwick, 1984), discriminant partial least squares (DPLS) (Næs, Isaksson, Fearn, & Davies, 2002), soft independent modeling of class analogy (SIMCA) (Wold, 1976) and support vector machines (SVMs) (Vapnik, 1995) equipped with the recently proposed Pearson VII universal kernel (PUK) (Üstün, Melssen, & Buydens, 2006). The results of this study will help to confirm or reject the hypothesis that the volatile profiles of Corsican honeys can be used for tracing products that do not comply with the label information.

The following section describes the theoretical basis of LDA, DPLS, SIMCA and SVM. Next the experimental conditions and data are presented in Section 3. In Section 4, the results obtained from all the methods applied to the data collected in a two-year sampling campaign are discussed. Finally, the conclusions of the study are given in Section 5.

2. Theory

2.1. Linear discriminant analysis (LDA)

In LDA, it is assumed that the data of each class follow the normal distribution; the classes are linearly separable and the class variance–covariance matrices are equal. The objective of LDA is to find linear combinations of explanatory variables called discriminant functions that maximize the between-classes variance and at the same time minimize the within-classes variance. This is the well-known Fisher criterion (Fisher, 1936). The number of discriminant functions found is equal to the number of classes minus one, if the number of variables is larger than the number of classes. A disadvantage of LDA is that it is appropriate only for data in which the total number of objects is considerably larger than the number of variables. When this condition is not fulfilled, compression of data by means of PCA, regularized LDA, stepwise LDA or feature reduction methods (Wu et al., 1996) should be considered.

2.2. Discriminant partial least squares (DPLS)

The partial least squares model expresses a linear relationship between a response variable $\mathbf{y}$ $(n \times 1)$ and a set of p explanatory variables $\mathbf{X}$ $(n \times p)$. It can be seen as an extension of multiple linear regression that can deal with multicollinearity in the data by constructing new latent factors, $\mathbf{T}$ $(n \times f)$, which maximize the covariance between $\mathbf{X}$ and $\mathbf{y}$. In discriminant PLS, the elements of $\mathbf{y}$ are usually coded as 0 and 1 (a binary variable) or as -1 and 1 (a bipolar variable) that provide information about the belongingness of n objects to two defined classes.

The optimal number of factors, f, is usually selected through the use of a cross-validation procedure and the model selected has a complexity for which the smallest root mean square error of cross-validation (RMSCV) is observed. The prediction power of a model is scored by a root mean square error (RMSEP) calculated for an external set of samples (test set) that is not used during the development of the model. Similar to LDA, an optimal DPLS model can be constructed when the training set is balanced i.e. it contains two classes with an equal number of samples and a comparable variance (Brereton, 2006).

2.3. Soft independent modeling of class analogy (SIMCA)

The main idea in SIMCA (Wold, 1976) is to build a confidence limit for each class with the help of principal component analysis (PCA) and then to project the unclassified samples into each principal components space and to assign them to the class in which they fit best.

Selection of the optimal number of PCs (f) is a key point in SIMCA and is usually determined using a leave-one object out cross-validation (Vandeginste et al., 1998). In our study, the classification of samples to a given category was done by means of the distance–distance plot (Daszykowski, Kaczmarek, Stanimirova, Vander Heyden, & Walczak, 2007). The Mahalanobis distance in the score space and the orthogonal distance from the PCA model constructed were calculated for each object. The default cut-off value for each of the absolute-centered distances (Mahalanobis or orthogonal) obtained using the cross-validated score values was determined at three times its standard deviation. By setting the cut-off value in this way, it was assumed that 99.90% of the centered distances for samples could be found within the interval of three times the standard deviation. One can make a distinction among regular samples and three categories of outlying observations. In general, each sample is fitted to each of the PCA models constructed and is assigned to the model for which its z-scores for the Mahalanobis and orthogonal distances are smaller than three.

2.4. Support vector machines (SVMs)

Support vector machine (SVM) is a binary classification tool that performs classification by constructing an optimal separating hyperplane (OSH) (Vapnik, 1995). The optimal separating hyperplane is defined as the one that maximizes the distance between the objects of the two classes (margin). Consider a data set $\{(x_1, y_1),\ldots, (x_n, y_n)\}$ with input data $\mathbf{x}_i \in \mathrm{R}^{\mathrm{d}}$ (d-dimensional input space) and output data y with binary class labeling $\mathbf{y} \in \{-1, +1\}$. SVM tries to separate the given set of binary-labeled data (two-class problem) with a hyperplane that is maximally distant from them. The objects lying on the margin are called support vectors that control the discrimination power of the hyperplane. Moreover, SVM has a regularization constant *C* that controls the trade-off between errors of the SVM on the training data and the margin maximization. Using a non-linear kernel function, SVM can deal with complex, e.g. non-linear relationships in data. The kernel function transforms the original data into a high dimensional feature space, where non-linear relationships can be present in a linear form.

Some commonly used kernel functions in the literature are the inner-product based linear and polynomial kernels and the Euclidean distance-based Gaussian (Radial Basis Function (RBF)) kernel (Schölkopf & Smola, 2002). The particular choice of a kernel function greatly depends on the nature of the data, i.e. which kind of underlying relationship needs to be estimated to relate the input data to the desired output. The nature of the data is usually unknown, so the kernel function will be determined experimentally by applying and validating various kernel functions and selecting the one with the highest generalization performance. In this paper we used the recently proposed Pearson VII universal kernel (PUK) (Üstün et al., 2006). The PUK function has the flexibility to change easily from a Gaussian into a Lorentzian shape and into intermediate shapes as well. This flexibility results in a higher mapping power for PUK in comparison to the commonly used linear, polynomial and RBF kernel functions.

The PUK kernel contains two parameters, namely σ and ω. The parameters σ and ω control the width (also named Pearson width) and the actual shape (tailing behavior) of the Pearson VII function. In order to use SVM, both the kernel parameters and the SVM regularization constant, *C*, need to be defined by the user.

In this paper, those parameters were defined by means of a grid search optimisation. To decrease the number of parameter possibilities (time saving) an internal scaling factor β was introduced, according to (Melssen, Üstün, & Buydens, 2006).

3. Experimental

3.1. Honey samples

In total, 374 honey samples were collected within the framework of the EU TRACE project (www.trace.eu.org). The samples were obtained directly from the producers, which guarantees their authenticity. In 2006 (first harvest), 111 Corsican, 18 non-Corsican-French, 15 Italian, 18 Austrian, 2 Irish, and 18 German honey samples were collected. During 2007 (second harvest), 108 Corsican, 28 non-Corsican-French, 15 Italian, 23 Austrian, and 18 German samples were collected. Before distribution, each honey sample was incubated at 40 °C overnight in an air oven, then manually stirred, and adjusted with distilled water to a content of solids of 70° Brix. Prior to analysis, the honey sample (2 g) was placed into a 10 mL vial for SPME; after adding 2 mL of distilled water, the vial was sealed with a magnetic cap with PTFE/silicon septum and vortexed until complete homogenisation was achieved.

As regards the Corsican samples, products collected in spring and autumn, both with specified origin (maquis, bushes, sweet chesnut, strawberry-tree and Clémentinier – Citrus Reticulata) and non-specified were included in the experimental set.

3.2. Chemicals and materials

The SPME fibre 50/30 μm divinylbenzene/carboxen/polydimethylsiloxane, (DVB/CAR/PDMS) used for the sampling of honey volatiles was supplied by Supelco (Bellefonte, PA, USA). Prior to use, the fibre was conditioned following the manufacturer's recommendations.

The system used for GC × GC experiments comprised a DB-5 ms, 5% phenyl polysilphenylenesiloxane (J&W Scientific, Folsom, CA, USA) primary column; 30 m × 0.25 mm id, 0.25 μm film thickness, coupled via a column connector (Agilent, Palo Alto, CA, USA) to a SUPELCOWAX 10, polyethylene glycol (Supelco, Bellefonte, PA, USA) second column of dimension 1.25 m × 0.1 mm id, 0.1 μm film thickness.

The mixture of *n*-alkanes (C_8–C_{20}) dissolved in *n*-hexane employed for retention index determinations was supplied by Supelco (Bellefonte, PA, USA). The calculation was done for components eluting between *n*-octane and *n*-eicosane.

3.3. Instrumentation

A Pegasus 4D system consisting of an Agilent 6890 N gas chromatograph equipped with a split/splitless injector (Agilent Technologies, Palo Alto, CA, USA), an MPS2 autosampler for automated SPME (Gerstel, Mülheim an der Ruhr, Germany), and a Pegasus III high-speed time-of-flight mass spectrometer (Leco Corp., St. Joseph, MI, USA) was used. Inside the GC oven a cryogenic modulator (N_2 jets–hot air jets technology) and a secondary oven (Leco Corp., St. Joseph, MI, USA) were mounted. Resistively heated air was used as a medium for hot jets, while cold jets were supplied by gaseous nitrogen cooled by liquid nitrogen.

3.4. Operating conditions

The operating conditions of the optimised HS-SPME–GC × GC–TOFMS method were as follows (Čajka et al., 2007):

- HS-SPME: incubation time: 5 min; incubation temperature: 40 °C; agitator speed: 500 rpm; extraction time: 20 min; desorption temperature: 250 °C; desorption time: 45 s (splitless). After 6 min exposure in the injector the fibre is automatically withdrawn and incubation and extraction of the next sample ensues.
- GC × GC: primary oven temperature program: 45 °C (0.75 min), 10 °C/min to 200 °C, 30 °C/min to 245 °C (1.25 min); secondary oven temperature: +20 °C above the primary oven temperature;

Table 1
Honey volatiles (markers) used for chemometric analysis.

Number of compound	Marker
1	Hexanal
2	Furan-2-carbaldehyde (furfural)
3	Hexan-1-ol
4	Heptanal
5	Heptan-1-ol
6	Benzaldehyde
7	Methylsulfanyldisulfanylmethane (dimethyl trisulfide)
8	Octanal
9	1-Methyl-4-propan-2-yl-benzene (*p*-cymene)
10	2-Phenylacetaldehyde
11	Octan-1-ol
12	1-Phenylethanone
13	Ethyl heptanoate
14	Nonanal
15	2-Phenylethanol
16	3,5,5-Trimethylcyclohex-2-en-1-one (isophorone)
17	Lilac aldehyde I
18	2,6,6-Trimethylcyclohex-2-ene-1,4-dione (4-oxoisophorone)
19	Lilac aldehyde II
20	Lilac aldehyde III
21	Nonan-1-ol
22	Ethyl octanoate
23	Decanal
24	Decan-1-ol
25	Ethyl nonanoate
26	Ethyl decanoate

modulator offset: +35 °C above the primary oven temperature; modulation period: 3 s (hot pulse 0.6 s); carrier gas: helium (purity 99.9999%); column flow: 1.3 mL/min.

- TOFMS: electron ionisation mode (70 eV); ion source temperature: 220 °C; mass range: *m/z* 25–300; acquisition rate: 300 spectra/s; detector voltage: −1750 V (first harvest), −1500 V (second harvest).

ChromaTOF (LECO Corp.) software (v. 2.31) was used for instrument control, data acquisition, and data processing. Identification of compounds was based on a NIST 2005 mass spectra library search and was further confirmed by comparing the linear retention indexes available in the same library.

In total, 26 aroma compounds (markers) were selected on the basis of a careful examination of the GC × GC chromatograms of honey samples by identifying the peaks that significantly varied in their intensities or those that described the quality of the honeys. The list of the markers selected is given in Table 1.

For the markers selected, the repeatability of SPME–GC × GC–TOFMS measurements (expressed as relative standard deviation, RSD) ranged between 2.1% and 12% ($n = 10$). Also, the quality of mass spectra was checked during the data processing of real samples and only those target volatiles with a mass-spectral match (i.e. similarity) higher than 700 were considered for chemometric analysis.

3.5. Data preprocessing prior to chemometric analysis

Before the construction of any classification/discriminant model, the raw data (374 × 26) presented in the form of absolute peak intensities were preprocessed using the row-closure operation (Vandeginste et al., 1998). This procedure involves the division of each row element by the corresponding row-sum. Row-closure enables an easier comparison of the sample profiles.

4. Results and discussion

LDA, DPLS, SIMCA and SVM were used to construct models for the discrimination of Corsican from non-Corsican honeys for each year of sampling and for the data of both years. This was done in order to investigate whether the year of sampling might have an influence on the discrimination. To estimate properly the predictive abilities of the built models, the data were divided into training and external test sets, respectively. Since LDA and DPLS are sensitive to the number of samples in the training set, special attention was paid when selecting this number. Table 2 shows the number of samples included in the training and test sets for each set of data.

One can see that the training set is balanced, i.e. it contains an equal number of Corsican and non-Corsican samples. This number was chosen as 75% of the total non-Corsican samples of each year and of both years because, in general, a smaller number of non-Corsican samples are available in comparison with the Corsican samples. The subsets were selected randomly and the selection was applied to each class separately. A leave-one object out cross-validation procedure was adopted to optimise all the models constructed and two figures of merit such as sensitivity and specificity were used to characterize the quality of the models' predictions. For a two-class problem, sensitivity is defined as the percentage of samples from the first class that are correctly predicted by the model, while specificity is the percentage of samples that are correctly classified as belonging to the other class. The ideal model would have sensitivity and specificity of 100%. Together with the sensitivity and specificity, one can define the so-called efficiency, also known in the literature as the non-error rate, which is the total percentage of correctly classified test samples.

4.1. Classification/discrimination of the Corsican honeys using data from 2006

The efficiency, sensitivity and specificity obtained from all models are listed in Tables 3 and 4.

The LDA model has an efficiency of 89.5% (see Table 3), which is related to eight incorrectly classified samples in the test set. An analysis of these misclassified samples showed that 6 out of 58 Corsican samples were recognized as non-Corsican and two non-Corsican samples fell within the domain of the Corsican samples. This resulted in a high sensitivity of 89.7% and a specificity of 88.9% of the LDA model (see Table 4). A slightly worse efficiency of 85.5% was obtained from the DPLS model with complexity four. RMSCV was equal to 0.59 and RMSEP was 0.63. Similar to the LDA model, six Corsican samples were incorrectly classified as non-Corsican resulting in the same sensitivity of 89.7%. However, compared to the LDA model, the DPLS model showed a decrease of 16.7% in specificity because of five non-Corsican honeys that were identified as Corsican. In contrast to the quality of the predictions obtained from LDA and DPLS, the eight-component SIMCA model showed a relatively poorer efficiency of 77.5%. Even though the sensitivity of the SIMCA model is the highest observed (93.1%), the specificity is quite low (64.8%). A good performance was achieved using the SVM method combined with the Pearson VII universal kernel (PUK) function. The optimal values for σ, w and C are presented in Table 5.

A total of ten samples were misclassified, when the SVM model was tested, which resulted in an efficiency of 86.8%. Again six Corsican samples were incorrectly rejected and the sensitivity is the same as the one found for the LDA and DPLS models. Because of the four incorrectly identified non-Corsican honeys, the SMV model has a specificity of 77.8%.

Comparing the prediction abilities of the models built for data of year 2006, all the models, except SIMCA, showed the same sensitivity, but they present different specificities. This means that the probability of recognizing a Corsican honey as a non-Corsican is the same with all methods, while the probability of identifying a non-Corsican honey as a Corsican increases in the order LDA, SVM, DPLS and SIMCA (see Table 4).

4.2. Classification/discrimination of the Corsican honeys using data from 2007

Again four different models were built for a randomly chosen training set. The number of samples considered in the training

Table 2
Number of Corsican and non-Corsican samples in the training and test sets for 2006, 2007 and for the two-year sampling.

	Training set		Test set	
	Corsica	Non-Corsica	Corsica	Non-Corsica
2006	53	53	58	18
2007	63	63	45	21
Two-year sampling	116	116	103	39

Table 3
Efficiency (in percentage) of the models.

	LDA	DPLS	SIMCA	SVM
Corsica vs. non-Corsica				
2006	89.5	85.5	77.5	86.8
2007	89.4	83.3	60.0	95.5
Two-year sampling	85.2	86.6	64.3	91.5

Table 4
Sensitivity (in percent) and specificity (in percent) of the models.

	LDA		DPLS		SIMCA		SVM	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Corsica vs. non-Corsica								
2006	89.7	88.9	89.7	72.2	93.1	64.8	89.7	77.8
2007	91.1	85.7	86.7	76.2	97.8	42.9	97.8	90.5
Two-year sampling	86.4	82.1	87.4	84.6	93.2	45.2	93.2	87.2

and test sets are presented in Table 2, while the efficiency, sensitivity and specificity of the models are shown in Tables 3 and 4. For honey data from 2007, all models, except SIMCA, presented comparable or improved efficiencies with respect to the corresponding models built for the data of the first year of sampling (see Table 3). The LDA model has an efficiency of 89.4%, when tested. After analyzing the misclassified test samples, it was found that four Corsican honeys were incorrectly recognized as non-Corsican and three non-Corsican honeys were found to have similar volatile compositions as the Corsican honeys. This analysis showed a fairly high sensitivity of 91.1% and a specificity of 85.7% for the LDA model.

The seven-component DPLS model with RMSCV equals 0.64 and a RMSEP of 0.72 lacks a correct assignment for six Corsican samples and therefore, a reduced sensitivity of 86.7% was observed in comparison with LDA. A decrease in specificity by 9.5% was also the result of the incorrect recognition of five non-Corsican honeys as Corsican. In contrast to both models built, the five-component SIMCA model had an efficiency of only 60.0% when tested. Although the model had a sensitivity of 97.8%, which is higher than the one found by LDA and DPLS, the majority of the non-Corsican honeys were assigned as Corsican and therefore, the specificity of the SIMCA model is considerably low (42.9%).

From the values presented in Table 3, it became clear that the SVM model has the best efficiency of 95.5% among the models built. Only three samples (one Corsican and two non-Corsican) were misclassified, when the SVM model was tested. Consequently, the SVM model offered the same fairly high sensitivity of 97.8% as the five-component SIMCA model and the best specificity (90.5%) among the all models constructed.

The difference in the classification results between the data of year 2006 and year 2007 might be explained by the possible differences in the volatile profiles of honeys from year to year.

4.3. Classification/discrimination of the Corsican honeys using the two-year data

One possibility is to construct a model for samples of one year and then to use this model to predict the belongingness of samples collected in the next year of sampling. However, due to the large variation in the sample profiles between the years possibly caused by different meteorological conditions, such a model would either be no longer valid or with poor predictive ability and therefore, a model updating would be required. A more reliable approach is to model the data of a two-year sampling together, i.e. for each class the training set contains representative samples of both years. Again, the training set was selected randomly and LDA, DPLS, SIMCA and SVM models were constructed.

Table 5
Optimal settings of σ, w and C of SVM models built using the PUK function.

	σ	w	C
Corsica vs. non-Corsica			
2006	2	10	10
2007	1	1	10^2
Two-year sampling	5	16	10^3

The efficiency of LDA (85.2%) is lower than that observed for the first and second year of sampling (see Table 3). There were 21 honey samples, which were incorrectly classified, when the model was tested. From the analysis of these samples, it followed that 14 Corsican and 7 non-Corsican honeys were predicted incorrectly, which led to a reduced sensitivity of 86.4% and a specificity of 82.1% in comparison with the predictions obtained from the LDA models constructed for the data from separate years of sampling (see Table 4).

DPLS showed an improved predictive ability compared to the LDA model. With DPLS, 13 out of 103 Corsican samples were identified as non-Corsican resulting in a sensitivity of 87.4%. Because 6 out of 39 non-Corsican honeys were recognized as Corsican, the specificity of the model was 84.6%. It should be noted that DPLS had a complexity of five factors, while RMSCV and RMSEP were both equal to 0.64.

Although, with the six-component SIMCA model 93.2% of the Corsican samples were predicted well, 54.8% of the non-Corsican samples were incorrectly recognized (see Table 4). Again, the worse predictive ability was observed for the SIMCA model.

The classification model constructed by SVM presented the best efficiency of 91.5% among all the methods. A total of 12 test samples (7 Corsican and 5 non-Corsican) were wrongly predicted and consequently the SVM model yielded a sensitivity of 93.2% and a specificity of 87.2%.

5. Conclusions

The main conclusion of this study is that the volatile profiles of Corsican honeys are specific enough and allow for their discrimination from honeys of different geographical origins. In general, all models, except SIMCA, showed good efficiencies, high sensitivities and specificities for data of separate sampling years and for data of both years. Moreover, all the models presented higher sensitivities than specificities. This demonstrates that the probability of recognizing Corsican honeys as non-Corsican on the basis of their volatile profiles is small. LDA, DPLS and especially SVM using the PUK function can be successfully used to detect the mislabeling of honeys.

Moreover, this study also demonstrated the flexibility of SVM using the PUK function in building models with better predictive abilities than those obtained from LDA and DPLS for the problem studied. The reason might be that the peak intensities of the 26 volatile compounds are not completely linearly related to the geographical origin of the honey samples.

It should be noted here that there is another important issue, namely which particular volatile compounds are the most descriptive in the classification/discrimination of the honey samples according to their origin. However, this matter will be the subject of a future study.

Disclaimer

The publication reflects only the authors' views and the European Commission is not liable for any use that may be made of the information contained therein.

Acknowledgements

The authors are grateful for the financial support concerning scientific activities within the Sixth Framework Programme of the European Union, project TRACE – "Tracing food commodities in Europe" (Project No. FOOD-CT-2005-006942). The implementation of TOFMS technology was funded by the project MSM 6046137305 supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Acquarone, C., Buera, P., & Elizalde, B. (2007). Pattern of pH and electrical conductivity upon honey dilution as a complementary tool for discriminating geographical origin of honeys. *Food Chemistry, 101*, 695–703.

Anklam, E. (1998). A review of the analytical methods to determine the geographical and botanical origin of honey. *Food Chemistry, 63*(4), 549–562.

Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A, 1158*, 196–214.

Brereton, R. (2006). Consequences of sample size, variable selection, and model validation and optimization, for predicting classification ability from analytical data. *TrAC – Trends in Analytical Chemistry, 25*, 1103–1111.

Čajka, T., Hajšlová, J., Cochran, J., Holadová, K., & Klimánková, E. (2007). Solid phase microexctraction–comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry for the analysis of honey volatiles. *Journal of Separation Science, 30*, 534–546.

Corbella, E., & Cozzolino, D. (2006). Classification of the floral origin of Uruguayan honeys by chemical and physical characteristics combined with chemometrics. *LWT – Food Science and Technology, 39*, 534–539.

Daszykowski, M., Kaczmarek, K., Stanimirova, I., Vander Heyden, Y., & Walczak, B. (2007). Robust SIMCA-bounding influence of outliers. *Chemometrics and Intelligent Laboratory Systems, 87*, 95–103.

Devillers, J., Morlot, M., Pham-Delegue, M. H., & Dore, J. C. (2004). Classification of monofloral honeys based on their quality control data. *Food Chemistry, 86*, 305–312.

Fernández-Torres, R., Pérez-Bernal, J., Bello-López, M., Callejón-Mochón, M., Jiménez-Sánchez, J., & Guiraúm-Pérez, A. (2005). Mineral content and botanical origin of Spanish honeys. *Talanta, 65*, 686–691.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188.

González, A. G. (2007). Use and misuse of supervised pattern recognition methods for interpreting compositional data. *Journal of Chromatography A, 1158*, 215–225.

Guyot, C., Scheirman, V., & Collin, S. (1999). Floral origin markers of heather honeys: *Calluna vulgaris* and *Erica arborea*. *Food Chemistry, 64*, 3–11.

Hernández, O. M., Fraga, J. M. G., Jiménez, A. I., Jiménez, F., & Arias, J. J. (2005). Characterization of honey from the Canary Islands: Determination of the mineral content by atomic absorption spectrophotometry. *Food Chemistry, 93*, 449–458.

Latorre, M. J., Peña, R., Pita, C., Botana, A., Garcia, S., & Herrero, C. (1999). Chemometric classification of honeys according to their type. II. Metal content data. *Food Chemistry, 66*, 263–268.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive analysis. Correspondence analysis and related techniques for large matrices*. USA: Wiley.

Marini, F., Magrì, A. L., Balestieri, F., Fabretti, F., & Marini, D. (2004). Supervised pattern recognition applied to the discrimination of the floral origin of six types of Italian honey samples. *Analytica Chimica Acta, 515*, 117–125.

Melssen, W. J., Üstün, B., & Buydens, L. M. C. (2006). SOMPLS: A supervised self-organising map partial least squares algorithm for multivariate regression problems. *Chemometrics and Intelligent Laboratory Systems, 86*, 299–309.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to multivariate calibration and classification*. Chichester: NIR Publications.

Radovic, B. S., Careri, M., Mangia, A., Musci, M., Gerboles, M., & Anklam, E. (2001). Contribution of dynamic headspace GC–MS analysis of aroma compounds to authenticity testing of honey. *Food Chemistry, 72*, 511–520.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge: MIT press.

Soria, A. C., González, M., de Lorenzo, C., Martínez-Castro, I., & Sanz, J. (2004). Characterization of artisanal honeys from Madrid (Central Spain) on the basis of their melissopaynological, physicochemical and volatile data. *Food Chemistry, 85*, 121–130.

Üstün, B., Melssen, W. J., & Buydens, L. M. C. (2006). Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems, 81*, 29–40.

Vandeginste, B. M. G., Massart, D. L., Buydens, L. M. C., de Jong, S., Lewi, P. J., & Smeyers-Verbeke, J. (1998). *Handbook of chemometrics and qualimetrics*. Part B. Amsterdam: Elsevier.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern Recognition, 8*, 127–139.

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S., et al. (1996). Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Analytica Chimica Acta, 329*, 257–265.