ANNALS of Faculty Engineering Hunedoara
– International Journal of Engineering
Tome XIII [2015] – Fascicule 4 [November]
ISSN: 1584-2673 [CD-Rom; online]
a free-access multidisciplinary publication of the Faculty of Engineering Hunedoara

[1.] Diana TSANKOVA, [2.] Svetla LEKOVA, [3.] Krastena NIKOLOVA, [4.] Georgi TERZIYSKI

# VIS SPECTROSCOPY-BASED CHEMOMETRIC ANALYSIS OF HONEY WITH RESPECT TO DISCRIMINATION OF ITS BOTANICAL ORIGIN

[1,3,4.] University of Food Technologies – Plovdiv, BULGARIA
[2.] University of Chemical Technology and Metallurgy – Sofia, BULGARIA

**Abstract**: The aim of the article is to investigate the potential of honey discrimination (on the base of its botanical origins) by Vis spectroscopy and chemometric analysis. Fifty-five samples from three types of honey (acacia, linden, and honeydew) are measured by a spectrophotometer "Helios Omega" with recorded wavelength range of 380~780 nm for calibration of honey classifier. Firstly, principal components analysis (PCA) is used for reducing the number of inputs and for a proper visualization of the experimental results. Next, the first two principal components (PCs) are combined separately with k-means clustering (KMC), Naïve Bayes classification (NBC), and linear and quadratic discriminant analysis (LDA, QDA) to develop PC-KMC, PC-NBC, PC-LDA and PC-QDA models, respectively. The comparative analysis of the four classifiers is based on leave-one-out-cross validation test carried out in MATLAB environment.
**Keywords:** Honey discrimination, Vis spectroscopy, k-means clustering, Naïve Bayes classification, discriminant analysis

## 1. INTRODUCTION

"Honey is the natural sweet substance, produced by honeybees from the nectar of flowers or from secretions of living parts of plants or excretions of plant sucking insects on the living parts of plants, which the bees collect, transform by combining with specific substances of their own, deposit, dehydrate, store and leave in honeycombs to ripen and mature" [2, 4]. Since honey is a natural product, nothing should be added to it. But for obtaining more profit, it is often subject to counterfeiting by adding sugar and other impurities. The botanical and geographical declaration of the origin seems to be one of the fundamental aspects of the honey quality that affects its commercial value [19, 22]. So in order to prevent fraud in the labeling, it should be developed a means of distinguishing between different types of honey. At the current stage of knowledge, a reliable authentication of floral origin of honey can be achieved by a global interpretation of sensory, pollen and physicochemical analyses carried out by an expert [16, 21, 20]. The content of different phenolic compounds is recognized to well reflect the type of honey and its quality, because phenolic acids and flavonoids are inherent chemical markers of the floral origin [17, 22]. Unfortunately, the most of these methods are generally too time-consuming, complex, and labour intensive for quality control application or require very specialized personnel to interpret the results.

In addition, most of the analytical techniques involve some kind of sample pre-treatment. The advantages of the technique of visible (Vis) and near infrared (NIR) spectroscopy with respect to other analytical methods are the non-invasive approach, the relatively easy and quick data acquisition. Some authors have used this technique with good accuracy for qualification of adulterants in honey [6, 25] or for distinguishing its floral origin [12, 24].

Among traditional classifiers, Discriminant Analysis (DA) is probably the most known method [15] and can be considered the first multivariate classification technique. Some authors [3, 1, 5, 12] have implemented linear discriminant analysis (LDA) for classification of the floral origin of honey, on the basis of its chemical and physical properties, including the mineral composition of honey. But due to the data correlation, the discriminant analysis (as well as other statistical classification methods) encounters some computational difficulties such as 'badly scaled or close to singular matrix'. Therefore, usually it is used in a combination with the principal components analysis (PCA) as a correlation reduction method.

The aim of the article is to investigate the possibility of distinguishing honey on the base of its botanical origin, using chemometric methods and spectral characteristics in the visible area. Spectroscopic data obtained undergo statistical processing including: principal components analysis (PCA); linear and quadratic discriminant analysis (LDA, QDA), Naïve Bayes classification (NBC) and k-means clustering (KMC) for cluster distinguishing. Technology of joint use of PCA and classification/clustering techniques, not only

contributes reduction of the area of the input data, but also overcomes some computational problems. The comparative analysis of the four classifiers is based on leave-one-out-cross validation test carried out in MATLAB environment.

## 2. MATERIALS AND METHODS

### ≡ Honey Spectrum Acquisition

Fifty-five samples of three different types of Bulgarian honey (acacia – 18 samples; linden – 25 samples; and honeydew – 12 samples) were purchased from supermarkets and from private producers. Before spectral measurement, the honey samples were placed in a water container at 50$^0$C until the soluble substances fully dissolved. Then the samples were annealed at room temperature (25-26$^0$C). The spectral characteristics of the honey were taken with a spectrophotometer Helios Omega ranging from 380 to 780 nm at 1 nm sampling space and using the software package VISIONlite ColorCalc. Spectral readings of the three types of honey were treated with the aid of the following methods, as follows. First, the method of the PCs has reduced dimensionality of the input data, then they were classified using the following four methods: the k-means clustering (KMC); Naïve Bayes classification (NBC); linear and quadratic discriminant analysis (LDA, QDA).

### ≡ Principal Components Analysis [7, 10]

The aim of the method is to reduce the dimensionality of multivariate data (e.g., wavelengths) whilst preserving as much of the relevant information as possible. PCA is a linear transformation that transforms the data (observations of possibly correlated variables) to a new coordinate system such that the new set of variables, the principal components, are linear functions of the original variables. PCs are uncorrelated, and the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. This is achieved by computing the covariance matrix for the full data set. Then, the eigenvectors and eigenvalues of the covariance matrix are computed, and sorted according to decreasing eigenvalue [7, 10]. All the principal components are orthogonal to each other. The full set of principal components is as large as the original set of variables. Usually the sum of the variances of the first few principal components exceeds 80% of the total variance of the original data [23].
In this study, the first two PCs are used by KMC, NBC, LDA and QDA calibration methods for developing the PC-KMC, PC-NBC, PC-LDA and PC-QDA discrimination models, respectively.

### ≡ K-means clustering [13, 8]

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Given an initial set of k means, the algorithm proceeds by alternating between two steps - assignment and update steps [13]. During the assignment step each observation is assigned to the cluster whose mean yields the least within-cluster sum of squares. Since the sum of squares is the squared Euclidean distance, this is the "nearest" mean. In the update step the new means are calculated and they have to be the centroids of the observations in the new clusters.
The algorithm has converged when the assignments no longer change. Since both steps optimize the within-cluster sum of squares, and there exists a finite number of such partitions, the algorithm must converge to a local optimum. There is no guarantee that the global optimum is found by k-means algorithm.)

### ≡ Naive Bayes classification [23, 9, 14, 18]

The Naive Bayes classifier is fast and easy to implement. It is designed for use when features are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid. In the training step, using the training samples, the method estimates the parameters of a probability distribution, assuming that features are conditionally independent given the class. During the prediction step, for any unseen test sample, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test sample according the largest posterior probability.
In this study the NBC algorithm uses normal (Gaussian) distribution. It is appropriate for features that have normal distributions in each class. The Naive Bayes classifier estimates a separate normal distribution for each class by computing the mean and standard deviation of the training data in that class.

### ≡ Linear and Quadratic Discriminant Analysis

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are two classic classifiers, with, as their names suggest, a linear and a quadratic decision surface, respectively. The basic idea of LDA is to find a linear transformation, such that the ratio of the between-class scatter and the within-class scatter is maximized. Samples are projected to a new space with smallest within-class distance and largest inter-class distance [11]. Although LDA usually gives a good discrimination performance, it suffers from some deficiencies if variables are highly correlated or class boundaries are complex or nonlinear [12]. To avoid such deficiencies, in the former case, variables are often transformed by correlation-reducing methods such as PCA, and in the latter case, LDA could be replaced by QDA. Unlike LDA, in QDA there is no assumption that the covariance of each of the classes is identical. To estimate the parameters required in quadratic discrimination more computation and data is required than in the case of linear discrimination.

Four methods mentioned above are used for classification of honey in 3 classes (acacia, lime, honeydew) in order to make a comparative analysis of their performance for this application. The results are confirmed by leave-one-cross-validation test in MATLAB environment.

## 3. RESULTS AND DISCUSSION

### ≡ Absorbance Spectra

Absorbance spectra of the three types of honey with wavelengths ranging from 380 to 780 nm are shown in Figure 1. The spectra present peaks at the band of 350 ~ 450 nm and low absorbance in the range above 760 nm. All the spectra are similar in spectral shape and absorbance. Therefore, it is necessary to apply appropriate methods of multivariate analysis to distinguish honey.

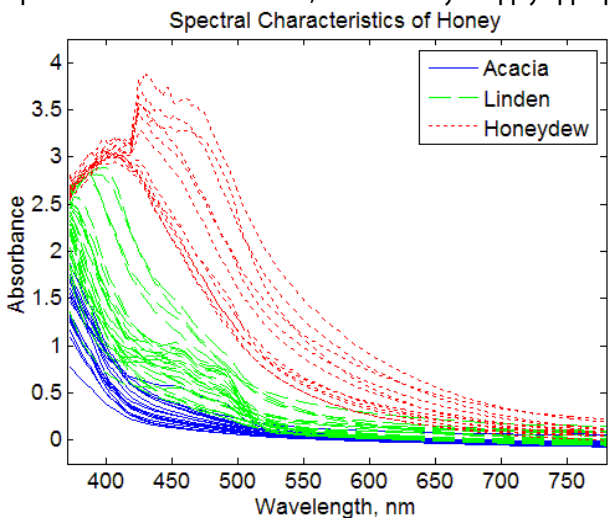


**Figure 1**. Absorbance spectra

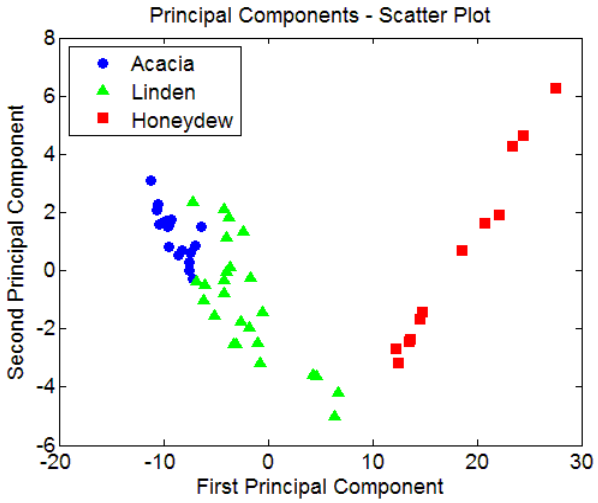


**Figure 2**. PCA of the three types of honey

### ≡ PC-KMC, PC-NBC, PC-LDA and PC-QDA Discrimination Models

The spectral dimensionality was reduced to a small number (two) of principal components using PCA. The scores scatter plot of the 1st and 2nd PCs is shown in Figure 2. Here, determining the type of honey is based solely on the inscription on the label by the manufacturer, i.e. trusting the manufacturer. The first two PCs explain as high as 98.63 % of variance of the spectra (94.29 % for PC-1 and 4.34 % for PC-2). The two PCs were chosen to develop PC-KMC, PC-NBC, PC-LDA and PC-QDA models. Leave-one-out-cross-validation test was used to check the performance of the classifiers. The prediction results of the honey's botanical origin made by the proposed classifiers (PC-KMC, PC-NBC, PC-LDA and PC-QDA) are shown in Figure 3 and Table 1, Table 2, and Table 3.

**Table 1**. Discrimination accuracy of PC-KMC based model

| Success: 81.82 % | | Predicted Class by PC-KMC | | | |
|---|---|---|---|---|---|
| | | Acacia | Linden | Honeydew | |
| Observed Class | Acacia | 18 | 0 | 0 | 18 |
| | Linden | 10 | 15 | 0 | 25 |
| | Honeydew | 0 | 0 | 12 | 12 |
| | | 28 | 15 | 12 | |

**Table 2**. Discrimination accuracy of PC-LDA based model

| Success: 92.73 % | | Predicted Class by PC-LDA | | | |
|---|---|---|---|---|---|
| | | Acacia | Linden | Honeydew | |
| Observed Class | Acacia | 18 | 0 | 0 | 18 |
| | Linden | 4 | 21 | 0 | 25 |
| | Honeydew | 0 | 0 | 12 | 12 |
| | | 22 | 21 | 12 | |

**Table 3**. Discrimination accuracy of PC-NBC based model and PC-QDA based one

| Success: 94.54 % | | Predicted Class by PC-NBC or PC-QDA | | | |
|---|---|---|---|---|---|
| | | Acacia | Linden | Honeydew | |
| Observed Class | Acacia | 17 | 1 | 0 | 18 |
| | Linden | 2 | 23 | 0 | 25 |
| | Honeydew | 0 | 0 | 12 | 12 |
| | | 19 | 24 | 12 | |

Figures and tables show the good performance of the models mentioned above. The minimum prediction accuracy (81.82 %) was obtained by the PC-KMC model. The PC-LDA performed significantly better (92.73 %) than PC-KMC one. The maximum prediction accuracy was obtained by the PC-NBC and PC-QDA calibration models. In this case (Table 3), 1 sample from observed class 'acacia' was predicted wrong as 'linden', while 2 samples from class 'linden' were predicted wrong as 'acacia'. The model predicted 52 out of 55 samples correctly. 94.54 % prediction accuracy (94.4 % class 'acacia', 92 % class 'linden', and 100 % class 'honeydew') was achieved.

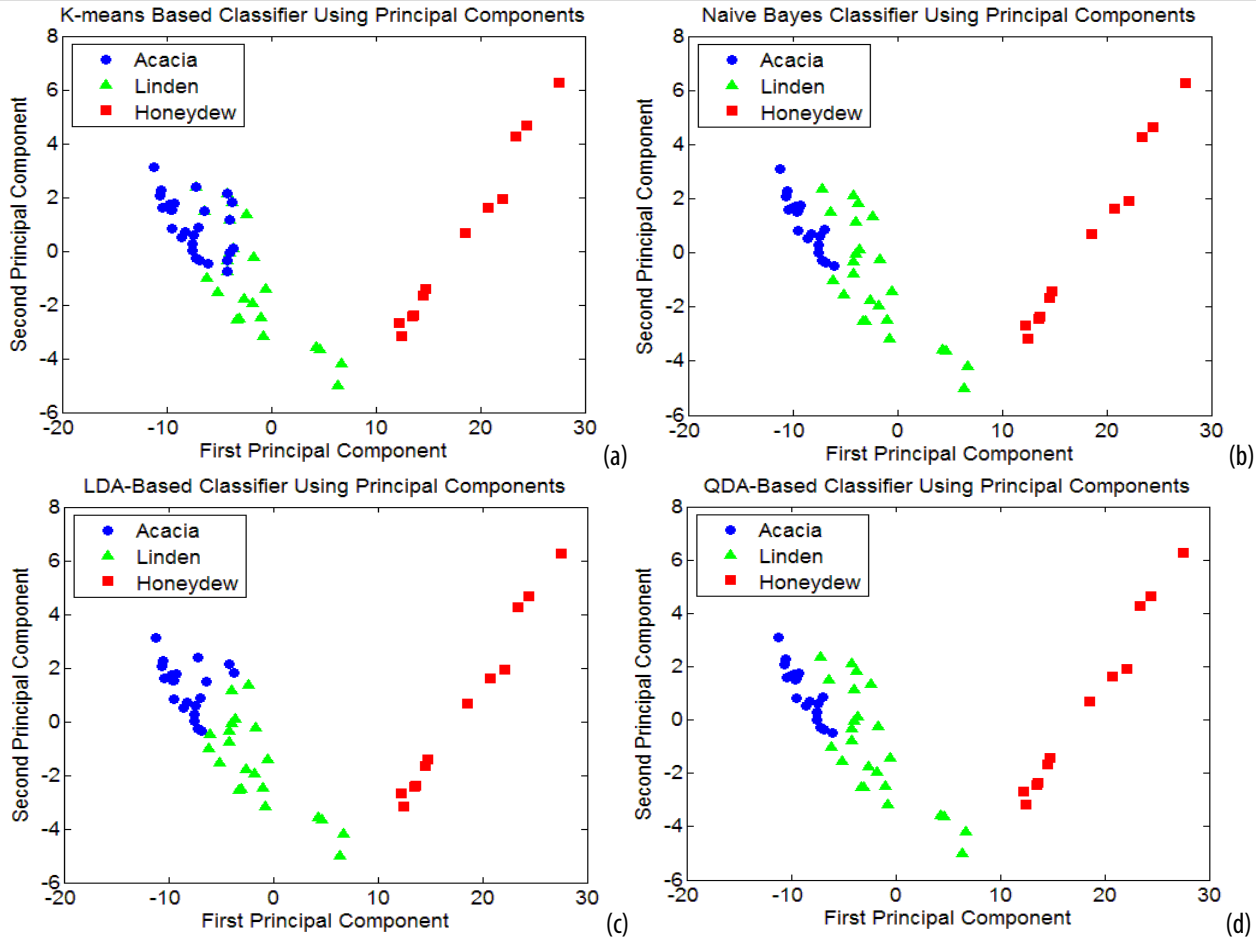


**Figure 3**. Classification models of honey: (a) PC-KMC, (b) PC-NBC, (c) PC-LDA, and (d) PC-QDA

## 4. CONCLUSIONS

In this article the possibility for distinguishing honey on the base of its botanical origin was investigated, using chemometric methods and spectral characteristics in the visible area. The proposed four calibration models are arranged in ascending order by their prediction accuracy as follows: PC-KMC (81.82 %), PC-LDA (92.73 %), PC-NBC and PC-QDA (both 94.54 %). The good performance of these models (models in the visible range) may allow developing a simpler and cheaper sensor for honey discrimination in practice. Future work will include adding new samples of other types of honey and using methods of artificial intelligence to increase the classifiers' accuracy.

**ACKNOWLEDGEMENTS**

The paper presents research and development, supported by Scientific Fund of Internal Competition of the University of Food Technologies – Plovdiv under the Research Project No.7/14-H.

**REFERENCES**

[1.] Chudzinska, M., Baralkiewicz, D., Application of ICP-MS method of determination of 15 elements in honey with chemometric approach for the verification of their authenticity, Food and Chemical Toxicology Vol.49, pp.2741–2749, 2011.

[2.] Codex Alimentarius: Draft revised standard for honey (at step 10 of the Codex procedure), Alinorm 01/25 19-26, 2001.

[3.] Corbella, E., Cozzolino, D., Classification of the floral origin of Uruguayan honeys by chemical and physical characteristics combined with chemometrics, LWT 39, pp.534–539, 2006, www.elsevier.com/locate/lwt

[4.] EU Council: Council directive 2001/110/EC of 20 December 2001 relating to honey, Official Journal of the European Communities L10:47–52, 2002.

[5.] Fernandez-Torres, R., Perez-Bernal, J.L., Bello-Lopez, M.-A., Callejon-Mochon, M., Jimenez-Sanchez, J.C., Guiraum-Perez, A., Mineral content and botanical origin of Spanish honeys, Talanta 65, pp.686–691, 2005.

[6.] Gallardo-Velazquez, T., Osorio-Revilla, G., Loa, M.Z.D., Rivera-Espinoza, Y., Application of FTIR-HATR spectroscopy and multivariate analysis to the quantification of adulterants in Mexican honeys, Food Research International, vol. 42, no.3, pp. 313–318, 2009.

[7.] Hotelling, H., Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24(6 & 7), 417–441 & 498–520, 1933.

[8.] http://en.wikipedia.org/wiki/K-means_clustering#Independent_component_analysis_.28ICA.29

[9.] http://en.wikipedia.org/wiki/Naive_Bayes_classifier#cite_note-rennie-3

[10.] Jolliffe, I. T., Principal Component Analysis. Second ed. Springer Series in Statistics. New York: Springer-Verlag New York, 2002.

[11.] Kim, H., Drake, B. L., Park, H., Multiclass classifiers based on dimension reduction with generalized LDA, Pattern Recognition, Vol. 40, No. 11, pp. 2939–2945, 2007.
[12.] Li, Y., Yang, H., Honey Discrimination Using Visible and Near-Infrared Spectroscopy, Hindawi, International Scholarly Research Network, ISRN Spectroscopy, Volume 2012, 2012, Article ID 487040, 4 pages, doi:10.5402/2012/487040
[13.] MacKay, David, An Example Inference Task: Clustering, Information Theory, Inference and Learning Algorithms, Cambridge University Press, Chapter 20, pp. 284–292. ISBN 0-521-64298-1. MR 2012999, 2003.
[14.] McCallum, A., Nigam, K., A comparison of event models for Naive Bayes text classification, AAAI-98 workshop on learning for text categorization, 752, 1998.
[15.] McLachlan, G., 1992, Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley
[16.] Persano Oddo, L., Bogdanov, S., Determination of Honey Botanical Origin: Problem and Issues. Apidologie, Vol.35 (special issue), pp.2-3, 2004.
[17.] Pyrzynska, K., Biesaga, M., Analysis of phenolic acids and flavonoids in honey, Trends in Analytical Chemistry, Vol.28, pp.893–902, 2009.
[18.] Rennie, J., Shih, L., Teevan, J., Karger, D., Tackling the poor assumptions of Naive Bayes classifiers. ICML, 2003.
[19.] Robbins, R.J., Phenolis acids in foods: An overview of analytical methodology, Journal of Agricultural and Food Chemistry, Vol.51, pp.2866–2887, 2003.
[20.] Ruoff, K., Authentication of the Botanical Origins of Honey, A Dissertation for the degree of Doctor of Sciences, University of Helsinky, p.32, 2006.
[21.] Ruoff, K., R. Karoui, E. Dufour, W. Luginbuhl, J.O. Bosset, S. Bogdanov, R. Amado, Authentication of the botanical origin of honey by front-face fluorescence spectroscopy. A preliminary study. J Agric Food Chem. Vol.53, No.5, pp.1343-1347, 2005.
[22.] Sergiel, I., P. Pohl, M. Biesaga, Mironczyk, A., Suitability of three-dimensional synchronous fluorescence spectroscopy for fingerprint analysis of honey samples with reference to their phenolic profiles, Food Chemistry, Vol.145, 319–326, 2014. http://dx.doi.org/10.1016/j.foodchem.2013.08.069
[23.] Statistics Toolbox™ User's Guide, R2014a, © COPYRIGHT 1993–2014 by The MathWorks, Inc.
[24.] Tsankova, D., Lekova, S., Botanical Origin-Based Honey Discrimination Using Vis-NIR Spectroscopy and Statistical Cluster Analysis, Journal of Chemical Technology and Metallurgy (JCTM), at press, 2015.
[25.] Zhu, X., S. Li, Y. Shan et al., Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics, Journal of Food Engineering, Vol.101, No. 1, pp. 92–97, 2010.

ANNALS of Faculty Engineering Hunedoara – International Journal of Engineering