Chemometric analysis of chemo-optical data for the assessment of olive oil blended with hazelnut oil

P. Kadiroğlu[1,2,*]
F. Korel[1,*]
M. Pardo[3,*]

[1] Food Engineering Department
Izmir Institute of Technology
Urla, Izmir, Turkey

[2] Food Engineering Department
Adana Science and Technology University
Sarıçam, Adana, Turkey

[3] Institute of Applied Mathematics
and Information Technology, CNR
Genova, Italy

(*) CORRESPONDING AUTHOR:
Prof. Dr. Figen Korel
Phone: 0090-232-7506228
Fax: 0090-232-7506196
E-mail: figenkorel@iyte.edu.tr

Matteo Pardo
Phone: 0049-30-25440157
Fax: 0049-30-25440130
E-mail: matteo.pardo@gmail.com

Author Information note:
Pınar Kadiroğlu
Adana, Turkey
E-mail: pınarkadiroglu@u.edu.tr

The main objective of this study was to determine different hazelnut oil concentrations in extra virgin olive oil (EVOO) belonging to different geographical regions inside Turkey using the combination of a SAW sensor based electronic nose (e-nose) and a machine vision system (MVS). We leveraged the oil characterisation given by the two easy-to-use and complementary experimental techniques through the adoption of conventional PCA for data exploration and random forests (RF) for supervised learning. The e-nose/MVS combination allows significantly better results both in adulteration detection independently of EVOO's geographical provenance and in EVOO geographical provenance determination, independently of the adulteration level, with respect to the single characterisation method. RF analysis also produces feature ranking, permitting to shed light on which oils' characteristics influence the learning result. We found that EVOO geographical provenance discrimination is mainly due to yellowness and guaiacol content, while (E)-2-hexenal chiefly determines the prediction of the hazelnut level.

Key words: Extra virgin olive oil, electronic nose, machine vision system, random forests, feature selection.

INTRODUCTION

Hybrid chemical sensing has been often used to improve electronic nose (e-nose) selectivity [1, 2]. The claimed advantage of such systems is the low(er) correlation between the responses of the different sensors types, which in turn is assured by the different sensors transduction principles. Even more diversification has been reached with joint e-nose and electronic tongue (e-tongue) experiments that measure samples in different phases [3-5]. Another experimental technique, which is complementary to the e-nose, is gas chromatography-mass spectrometry (GC-MS). GC-MS is the reference analytical technique for the characterisation of food headspace, yet its cost doesn't normally allow more than a double check of the e-nose results on a restricted sample subset. Only recently researchers have considered sample properties different from their chemical emission. In particular, adding spectral properties, as measured with a camera, colorimeter or electronic eye (e-eye), allows a more complete sample characterisation, due to the removal of feature correlation. Colour is the one of the most important quality criterions of virgin olive oil and highly affects consumer preferences. It is influenced by different factors such as environmental conditions, fruit variety, and degree of fruit ripeness, growing region, processing and storage techniques. The chlorophyll and carotenoid profiles of olive oils have been shown to correlate with several colour descriptors [6]. An e-nose, e-tongue and e-eye combination was used to discriminate red wines aged with different methods [7] and to discriminate Spanish olive

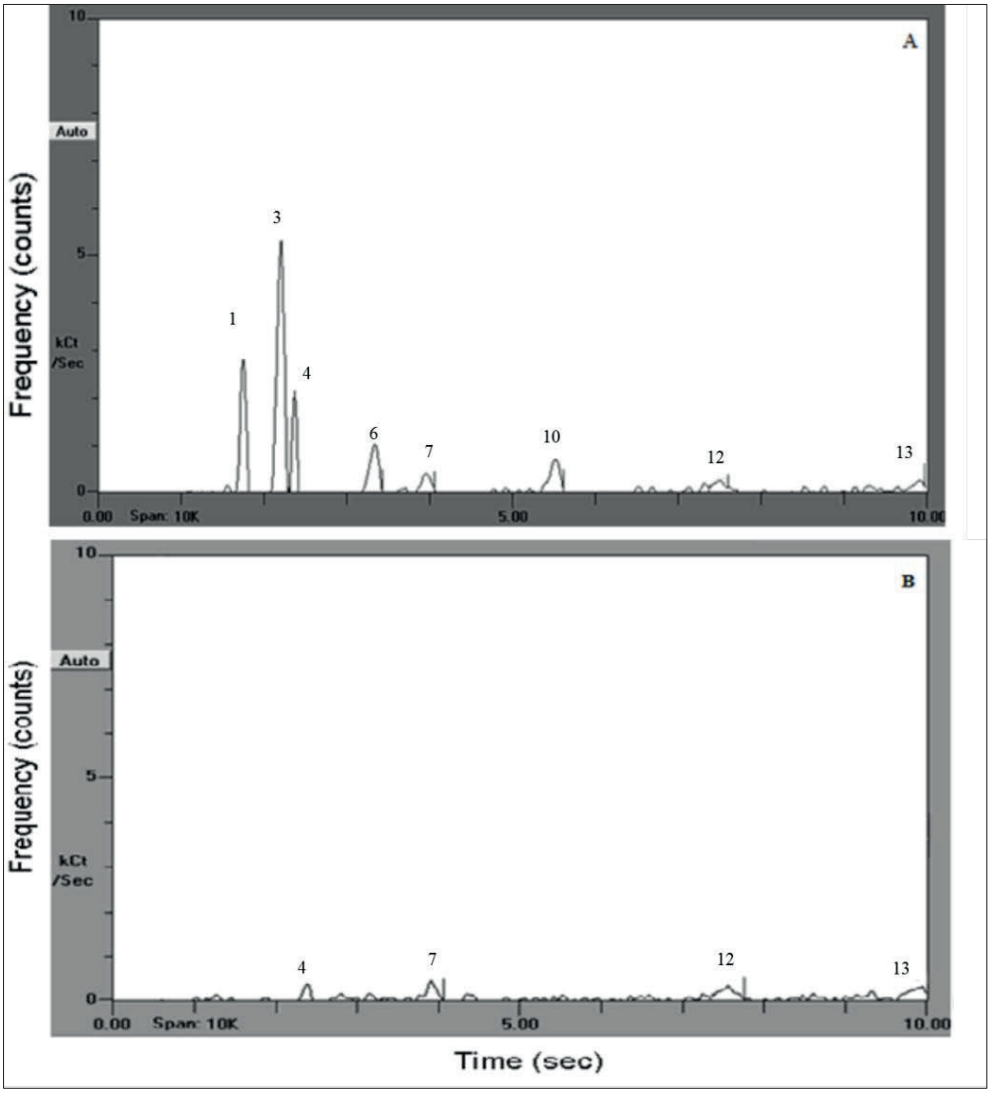

Figure 1 - The typical znose™ chromatogram of EVOO (a) and hazelnut oil (b).

varieties and predict their bitterness degree [8].

The authentication of virgin olive oil and adulteration is an important issue for the oil industry. Economic losses, health and safety problems arise from frauds with cheaper vegetable oils or low-quality olive oils [9]. Several studies focused on the detection of olive oil adulteration with traditional analytical techniques [10-14], and rapid methods, such as high-power pulsed-field gradient NMR [15], the combination of SPME/GC-FID, SPME/GC-MS and e-nose [16], and mid-infrared (IR) spectroscopy [17].

For the analysis of combined chemo-optical data, several standard chemometrics or pattern recognition algorithms can be considered. An important feature of any analysis method is the ability to perform feature selection that allows appreciating the features' relative contribution to the performance enhancement due to datasets fusion. We opted for Random Forests (RF), an 'ensemble learning method generating many classifiers and aggregating their results, which showed top learning performances in several comparative studies and easily produces a feature score [18]. Other well-known aggregation methods are boosting [19] and bagging [20]. In bagging, each classifier is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction. Starting from bagging, Breiman [18] proposed RF, which are ensembles of trees (classification or regression trees). In addition to constructing each tree using a different bootstrap sample of the data, RF change the way of the constructed trees. In standard trees, each node is split using the best split among all variables. In a RF, each node is split using the best split among a subset of predictors (i.e. features) randomly chosen at that node. This somewhat contrary strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support

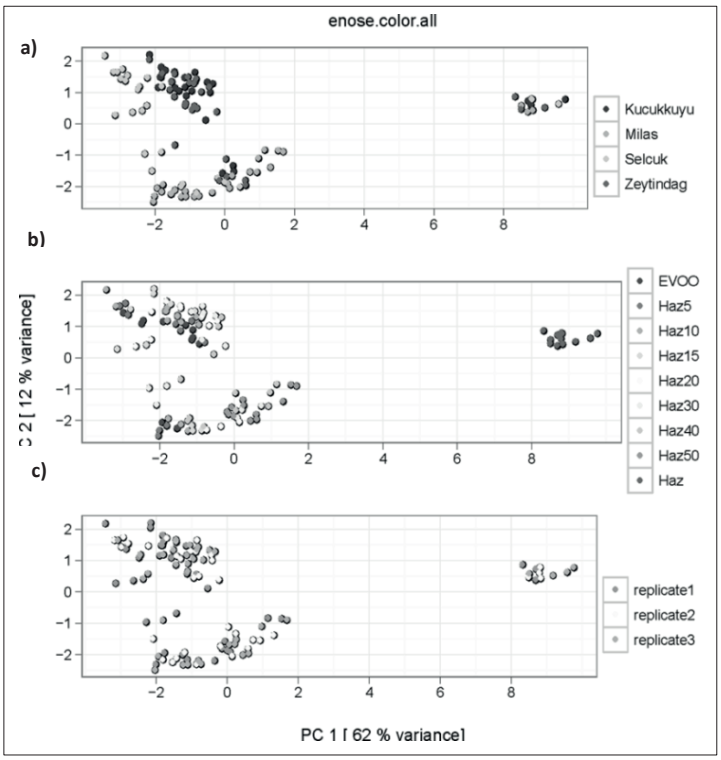

Figure 2 - PCA plots with different point labeling for the joined znose™ and color data

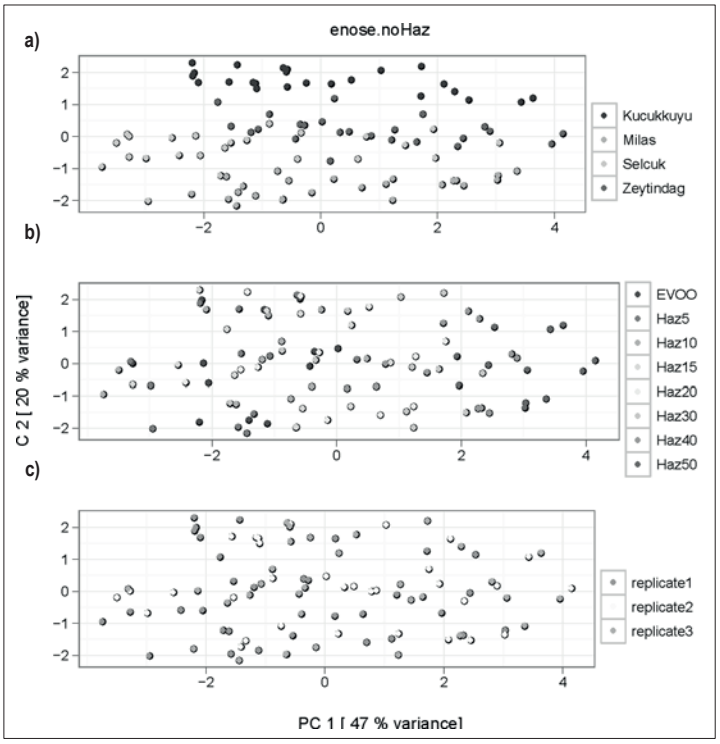

Figure 3 - PCA plots with different point labeling for the znose™ data, no pure hazelnut oil present.

vector machines and neural networks. It is rather robust against overfitting [18]. In addition, RF has only two hyper parameters which are the number of variables in the random subset at each node and the number of trees in the forest. They are usually not very sensitive to their values. Another advantage of RF is that it automatically outputs variable ranking. The RF algorithm estimates the importance of a variable by calculating how much the prediction error increases when data not in the bootstrap sample ('out-of-bag' data) for that variable is permuted (which amounts to making that variable useless), while all others are left unchanged. The rationale is that, if prediction deteriorates when a certain variable is made useless, that variable is important for prediction. The necessary calculations are carried out tree by tree as the RF is constructed.

In analytical chemistry there are a few papers on the advantage of classification performances and variable selection capabilities of RF. To our best knowledge, Hancock et al. were the first ones to make a performance comparison of RFs and other modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic quantitative structure-retention relationship (QSRR) studies [21]. Granitto *et al.* have a nice paper on the application of Random Forest-Recursive Feature Elimination (RF-RFE) algorithm to the identification of relevant features in the spectra produced by Proton Transfer Reaction-Mass Spectrometry (PTR-MS) analysis of agroindustrial products [22]. In previous papers of ours [23, 24] we found that RF and Support Vector Machines (SVM) have a similar classification performance [25, 26], while Nearest Shrunken Centroids (NSC) have worse performances [27]. It was also shown that RF and NSC, which both have a useful intrinsic feature selection mechanism, produce different feature rankings. In particular, NSC, scoring features independently, may lose some of the features found by RF.

The main objective of this study was to determine the ability of the combination of e-nose and machine vision system (MVS) to detect the adulteration level of hazelnut oil in EVOO, independently of their geographical provenance. We compare

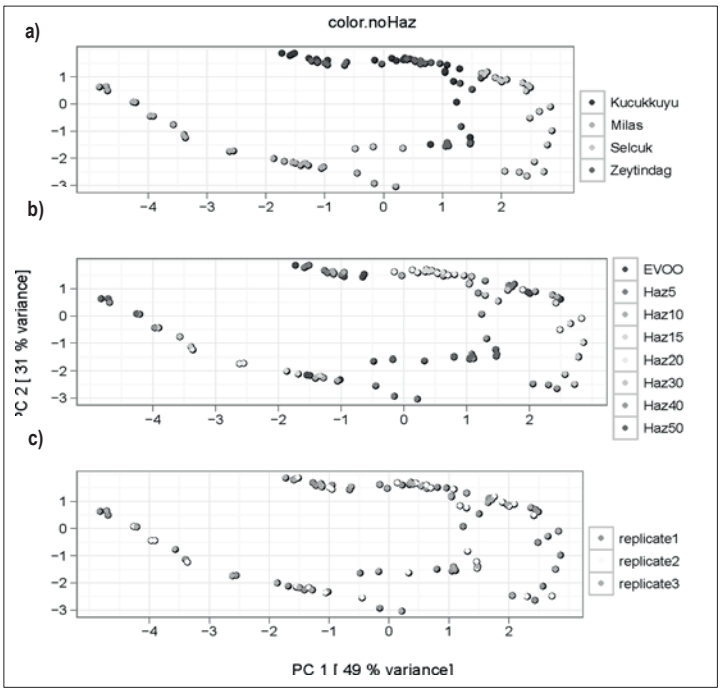

Figure 4 - PCA plots with different point labeling for the color data, no pure hazelnut oil present.

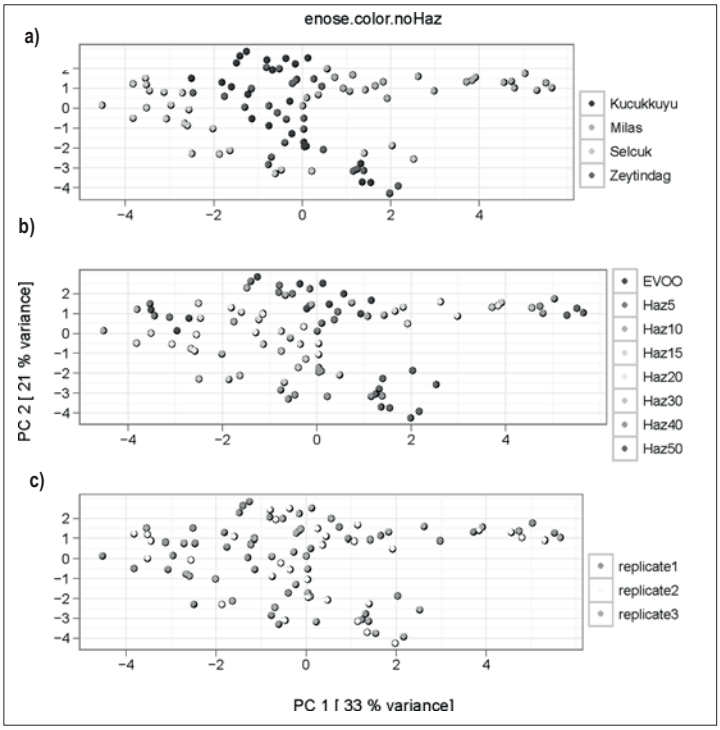

Figure 5 - PCA plots with different point labeling for the joined znose™ and color data, no pure hazelnut oil present.

this to the ability of the e-nose and MVS taken singularly. We also address the oil provenance discrimination problem. We make use of the visual data interpretation capabilities of PCA as much as possible; when PCA arrives to its limit we adopt state-of-the-art supervised analysis with RF.

EXPERIMENTAL PART

SAMPLE PREPARATION

Two North Aegean region (Zeytindag and Kucukkuyu) and two South Aegean regions (Milas and Selcuk) EVOOs, which were produced from olives harvested from one cultivar in a specific region, and hazelnut oil were purchased in İzmir (Turkey). Prior to the analyses, all samples were stored in the dark at 8°C. None of them were subjected to any treatment that might alter their composition. The study was carried out on 9 groups of samples: pure EVOO, pure hazelnut oil and seven groups of samples of EVOO blended with hazelnut oil (Haz) at different adulteration levels: 5, 10, 15, 20, 30, 40, and 50% (v/v). The experiment was performed in triplicate.

E-NOSE ANALYSIS

E-nose measurements were performed with the zNose™ 7100 vapor analysis system (Electronic Sensor Technology, Newbury Park, CA, USA) and the area of 8 chromatographic peaks were extracted, following the procedures stated in Kadiroğlu et al. [28]. Six e-nose readings were taken for each oil sample.

MACHINE VISION SYSTEM ANALYSIS

Oil samples (25 ml) were then transferred into glass Petri dishes (60 × 15 mm) and placed in MVS illuminated with two D65 fluorescent lamps (ECS Inc., Gainesville, FL, USA). An image was taken with a CCD digital camera (Sony DFK 21BF04, The Imaging Source Europe GmbH, Bremen, Germany). Due to the stability of the colour measurement, only one reading was taken. Five summary colour features were extracted with the ColorExpert software (ECS Inc., Gainesville, FL, USA): lightness (L), redness-greenness (a), yellowness-blueness (b), chroma and

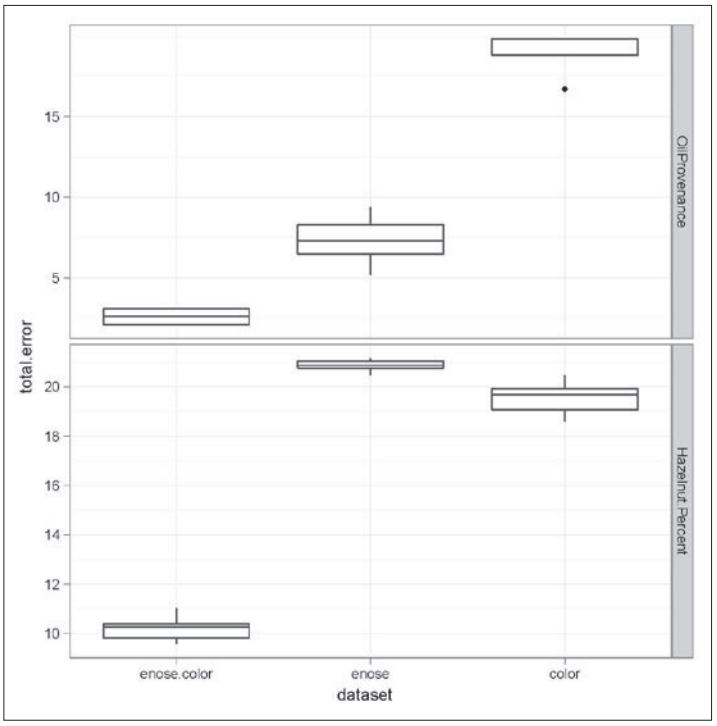

Figure 6 - Box plots summarizing test results over 10 runs of RF

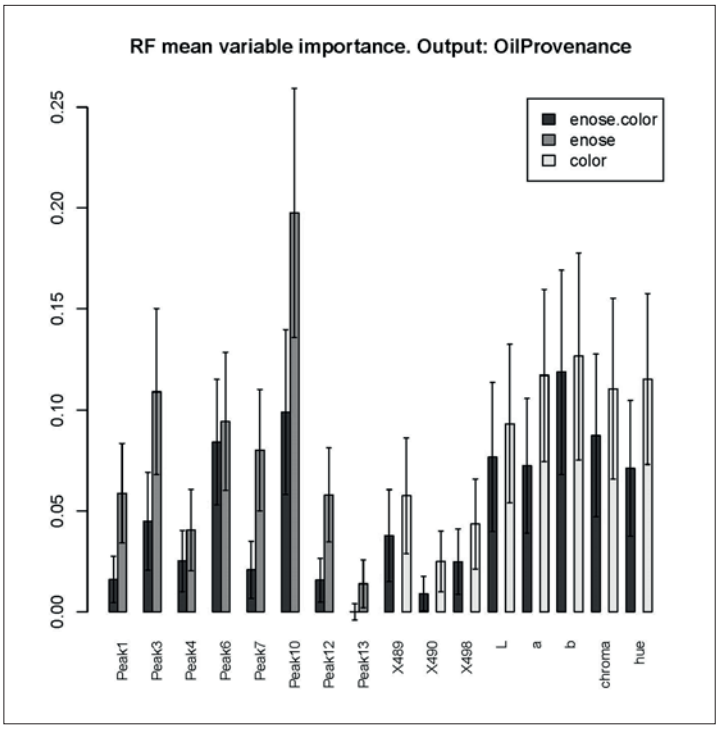

Figure 7 - Variable importance outputted by rf for the oil provenance discrimination problem.

hue. L values range from 0 (black) to 100 (perfect reflecting diffuser). Hue and chroma are derived from a and b: the hue angle, describes the sense of colour ($h = \tan^{-1}(b/a)$), and the saturation index or chroma ($C = (a^2 + b^2)^{0.5}$) is associated with brightness or vividness of a colour.

DATA ANALYSIS

The first data pre-processing step was to aggregate the six readings for every replicate of the sample measured with the e-nose (the mean value was taken). In order to visually understand how the output variables (geographical provenance and adulteration levels) affect sensors' response, Principal Component Analysis (PCA), the use of different labelling on the single and on the joint datasets was applied systematically. For the supervised analysis and feature selection with random forests, we adopted the *randomForest* package [29], an R interface to the original Fortran programs by Breiman and Cutler [30].

RESULTS AND DISCUSSION

Typical chromatograms for EVOO (A) and hazelnut oil (B), displayed in Figure 1, show that the differences between pure olive oil and pure hazelnut oil are significant.

In Figure 2 we show PCA plots -performed on the colour dataset- corresponding to three different data labelling. In the upper two plots data points are labelled according to quantities to be discriminated, i.e. geographical provenance in (a) and Haz content in (b). In (c) a quality check was performed: data are plotted according to replicate number. Replicates cluster together for any fixed 'provenance' – 'Haz content' pair, confirming measurements' reproducibility. In Figure 2b it is seen that 100% adulteration (samples 'Haz') is straightforward to discriminate (we show only the colour dataset, the same holds for the e-nose dataset). We therefore restrict further analysis to datasets without 'Haz' samples ('noHaz').

PCA plots of e-nose data alone, colour data alone and e-nose in combination with colour data are presented in Figure 3, Figure 4, Figure 5 respectively. Figure 3 is a fine example of different effects determining different principal

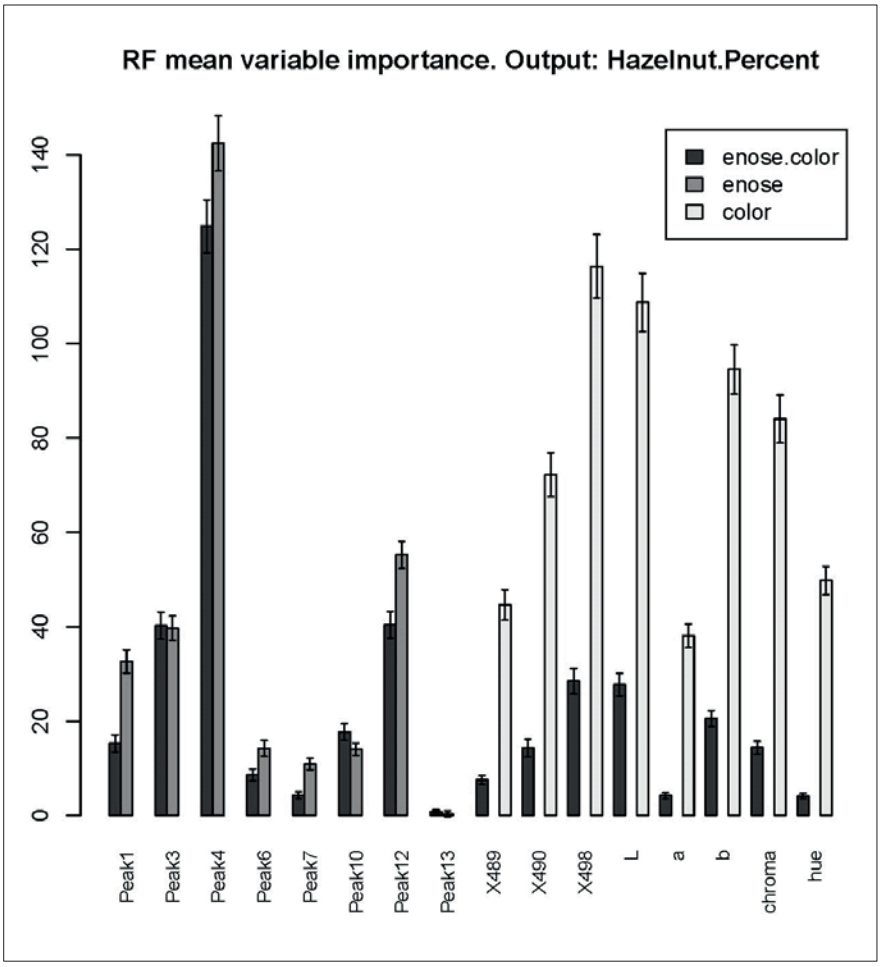

Figure 8 - Variable importance outputted by rf for the adulteration quantification problem.

components (PC). Figure 3a shows that differences in oil provenance determine the 2nd PC, where Selcuk and Zeytindag are partially superposed. In Figure 3b Haz concentrations clearly increase from left to right. Still, no satisfactory discrimination is possible; in particular, looking in detail, it can be noticed that Haz20 samples on the left-hand side, which are mixed with EVOO, Haz5 and Haz10, all have Selcuk origin.

The PCAs on colour data in Figure 4 tell a partially different story. Here provenance and Haz concentration influence both PCs. Moreover: 1) Kucukkuyu and Zeytindag are superposed and 2) three Selcuk points with high adulteration level (Haz50) are displaced at the centre of the PCA plot. Most importantly, from Figure 4b we learn that adulteration level is only clear inside a geographical homogeneous group.

Finally, Figure 5 does not show significant increase in discrimination of geographical provenance or adulteration levels. Based on the unsecular results Kucukkuyu and Zeytindag are superposed while Milas and Selcuk samples were discriminated clearly. Similarly, adulteration level discrimination could be observed within the geographical provenance. However, overall view demonstrates that Haz5 and Haz10 samples could not be differentiated from EVOO samples while the discrimination capability increased with the increment of Haz level. It may be noticed, though, that the 2D PCA projection captures far less variance: only 54%, versus 80% for the colour data and 67% for the e-nose. This means that e-nose and colour data are not very correlated. To check this hypothesis supervised learning needs to be applied.

RF test results displayed in Figure 6 confirm the hypothesis. The boxplots sum up the errors of 10 RF runs. Indeed e-nose and colour data taken together allow a significantly better prediction of both geographical provenance and adulteration level with respect to the two datasets taken singularly. The median geographical provenance misclassification error (upper plot) of e-nose+colour is circa 2.5% (with worse error of circa 3%), while the median error for the e-nose data is circa three times and that obtainable from the colour data more than seven times as much. The lower plot reports the percent of unexplained variance in hazelnut oil concentration prediction. Again, the joint datasets allow a much better prediction, the median error being circa one half of that obtainable by each component datasets.

Having established that joining the datasets increases prediction, we would like to know which features are more important for learning. We expect that, for the joint data, at least one feature from each dataset scores highly. Results are shown in Figure 7 and Figure 8 for oil geographical provenance and adulteration level, respectively. The mean variable importance, always on the 10 runs, is given by the coloured filled bars and the black lines display the standard deviation.

The relative importance of features, as gauged by the relative height of the dark grey bars, is quite different for the two prediction problems: while colour and e-nose feature sets in Figure 7 are of comparable size, in Figure 8 three e-nose features are higher than the strongest colour feature, with Peak4 having a dominant role. Nevertheless, also in the latter case, the inclusion of colour features is beneficial, as seen in Figure 6. This is a case in which complementary features – though having a small overall effect by themselves - can contribute to a performance enhancement when fused with strong features.

A related observation is that, according to Figure 6, colour features alone give a better prediction than e-nose features alone, if only by a small margin. Yet, we just saw that, when fused with e-nose features, colour features' importance is smaller. An explanation of this fact is that at least two-colour features have a similar role; therefore when one is taken out, performance does not drop that much (remember that it is this change of performance that the RF feature importance is measuring).

The two learning problems make use of two distinct set of features: for provenance discrimination several features have a similar importance when fused (in the order: b, Peak10, chroma, Peak6); for adulteration-level prediction Peak4 stands out, while it was of low importance in the first task. The b value indicates yellowness in olive oils: as Haz level increases, the b value increases. The EVOOs obtained from different provenances also have different yellowness values. So, a priori, the feature b could help both provenance discrimination and detection of Haz level. It is a result of the RF analysis that, in our experimental setting, b contributes mainly to geographical discrimination. Peak 10 was tentatively identified through the database of Kovats indices stored in the substance library of the Microsense software, using n-alkanes as standard. Peak 10 was identified as guaiacol (kovats index-1091), having burnt odour. South region EVOOs (Milas and Selcuk) have higher amounts of guaiacol whereas North region EVOOs (Kucukkuyu and Zeytindag) have lower amounts. Peak 4 was tentatively identified as (E)-2-hexenal (kovats index-854) and has a fatty odor. As the Haz level increases, the amount of (E)-2-hexenal decreases and this compound helps in the prediction of Haz level in EVOOs. Bozdoğan Konuşkan [31] stated that olive oils had guaiacol and (E)-2-hexenal as volatile compounds and had phenolic and burnt odour and green and apple odour, respectively.

It can be also noted that the uncertainty on the variable importance (extent of the black lines) is bigger for the oil provenance that for the adulteration level prediction. This could depend on the intrinsic discreetness of the classification problem, for which the change of a label results in a quantum difference in prediction. On average one third of the samples (i.e. 24 samples) are in the test set. One single differently classifies sample therefore changes the classification rate by $1/24 \approx 4\%$.

CONCLUSION

We demonstrated the advantage of joining e-nose and MVS for the rapid detection of hazelnut adulteration in extra virgin olive oils, independently of the geographical provenance of the oils (inside Turkey). PCA, if performed with different point labelling, already allows for smart data analysis when data variance is mainly explained by the first two principal components. For the combination of e-nose and MVS data, giving the complementarity of two experimental techniques, supervised learning with random forests was applied. Random forests permitted to clearly quantify the prediction differences between joint datasets and single datasets and, through feature ranking, gave an indication of the main chemical quantities responsible for the successful predictions.

BIBLIOGRAPHY

[1] H. Ulmer, J. Mitrovics, G. Noetzel, U. Weimar, W. Göpel, Odours and flavours identified with hybrid modular sensor systems. Sensor Actuat B-Chem *43* (1-3), 24-33 (1997)

[2] H. Ulmer, J. Mitrovics, U. Weimar, W. Göpel, Sensor arrays with only one or several transducer principles? The advantage of hybrid modular systems. Sensor Actuat B-Chem *65* (1-3), 79-81 (2000)

[3] F. Winquist, I. Lundström, P. Wide, The combination of an electronic tongue and an electronic nose. Sensor Actuat B-Chem *58* (1-3), 512-517 (1999)

[4] A.K. Deisingh, D.C. Stone, M. Thompson, Applications of electronic noses and tongues in food analysis. Int J Food Sci Tech *39* (6), 587-604 (2004)

[5] M.S. Cosio, D. Ballabio, S. Benedetti, C. Gigliotti, Evaluation of different storage conditions of extra virgin olive oils with an innovative recognition tool built by means of electronic nose and electronic tongue. Food Chem *101* (2), 485-491 (2007)

[6] B. Gandul-Rojas, MR-L. Cepero, M.I. Mínguez-Mosquera, Use of chlorophyll and carotenoid pigment composition to determine authenticity of virgin olive oil. J Am Oil Chem Soc *77*(8), 853-858 (2000)

[7] M.L. Rodriguez-Mendez, C. Apetrei, I. Apetrei, S. Villanueva, I.J.A . de Saja, I. Nevares, M. del Alamo, Combination of an electronic nose, an electronic tongue and an electronic eye for the analysis of red wines aged with alternative methods. ISIE 2007, 2007 IEEE International Symposium on Industrial Electronics 2782 - 2787 (2007)

[8] C. Apetrei, I.M. Apetrei, S. Villanueva, J.A. de Saja, F. Gutierrez-Rosales, M.L. Rodriguez-Mendez, Combination of an e-nose, an e-tongue and an e-eye for the characterisation of olive oils with different degree of bitterness. Anal Chim Acta *663* (1), 91-97 (2010)

[9] F.P. Capote, J.R. Jiménez, M.D.L. de Castro, Sequential (step-by-step) detection, identification and quantitation of extra virgin olive oil adulteration by chemometric treatment of chromatographic profiles. Anal Bioanal Chem *388* (8),1859-1865 (2007)

[10] L. Mannina, M. Patumi, P. Fiordiponti, M.C. Emanuele, A.L. Segre, Olive and hazelnut oils: A study by high-field 1H NMR and gas chromatography Ital. J. Food Sci. *2*, 139-149 (1999)

[11] G. Morchio, A. Pellegrino, C. Mariani, G. Bellan, Feasibility to check the presence of hazelnut oil in olive oil. Note 2 Riv. Ital. Sostanze Grasse (76), 115-127 (1999)

[12] W. Moreda, A. Cert, Algorithms for the detection of hazelnut oil in olive oil. Grasas y Aceites 51 (3), 143-149 (2000)

[13] S. Vichi, L. Pizzale, E. Toffano, R. Bortolomeazzi, L. Conte, Detection of hazelnut oil in virgin olive oil by assessment of free sterols and triacylglycerols. J AOAC Int. 84 (5), 1534-1541 (2001)

[14] C. Mariani, G. Bellan, E. Lestini, R. Aparició, The detection of the presence of hazelnut oil in olive oil by free and esterified sterols. European Journal Food Research and Technology, Eur Food Res Technol 223 (5), 655-661 (2006)

[15] D. Šmejkalová, A. Piccolo, High-power gradient diffusion NMR spectroscopy for the rapid assessment of extra-virgin olive oil adulteration. Food Chem *118* (1), 153-158 (2010)

[16] S. Mildner-Szkudlarz, H.H. Jeleń, The potential of different techniques for volatile compounds analysis coupled with PCA for the detection of the adulteration of olive oil with hazelnut oil. Food Chem *110* (3), 751-761 (2008)

[17] G. Gurdeniz, B. Ozen, Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. Food Chem *116* (2), 519-525 (2009)

[18] L. Breiman, Random forests. Mach Learn *45* (1), 5-32 (2001)

[19] R.E. Schapire, Y. Freund, P. Bartlett, W.S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods. Ann Stat *26* (5), 1651-1686 (1998)

[20] L. Breiman, Bagging predictors. Mach Learn *24* (2), 123-140 (1996)

[21] T. Hancock, R. Put, D. Coomans, Y.V. Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. Chemometr Intell Lab *76* (2), 185-196 (2005)

[22] P.M. Granitto, C. Furlanello, F. Biasioli, F. Gasperi, Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr Intell Lab *83* (2), 83-90 (2006)

[23] M. Pardo, G. Sberveglieri, Random forests, nearest shrunken centroids and support vector machines for the classification of diverse e-nose datasets. 2006 IEEE Sensors (1-3), 424-426 (2006)

[24] M. Pardo, G. Sberveglieri, Random forests and nearest shrunken centroids for the classification of sensor array data. Sensor Actuat B-Chem *131* (1), 93-99 (2008)

[25] V.N. Vapnik Statistical Learning Theory, Wiley, New York (1998)

[26] M. Pardo, G. Sberveglieri, Classification of electronic nose data with support vector machines. Sensor Actuat B-Chem *107* (2), 730-737 (2005)

[27] R. Tibshirani, T. Hastie, B. Narasimhan, . Chu, Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Stat Sci *18* (1), 104-117 (2003)

[28] P. Kadiroğlu, F. Korel, F. Tokatli, Classification of Turkish extra virgin olive oils by a SAW detector electronic nose. J Am Oil Chem Soc *88* (5), 639-645 (2011)

[29] A. Liaw, M. Wiener Classification and Regression by randomForest. The Newsletter of the R, Project 2 (2002).

[30] L. Breiman, A. Cutler, Random Forests. http://www.stat.berkeley.edu/~breiman/RandomForests / cc_home.htm. Accessed 12/11/2012 (2004)

[31] D. Bozdoğan Konuşkan, A. Karayiyen Volatile aroma compounds in virgin olive oil and their effects on sensory quality of oil. Gıda *36* (6), 357-364 (in Turkish) (2011)

Received: July 10, 2018
Accepted: December 12, 2018