Continuous statistical modelling for rapid detection of adulteration of extra virgin olive oil using mid infrared and Raman spectroscopic data

Georgouli, K., Martinez del Rincon, J., & Koidis, A. (2017). Continuous statistical modelling for rapid detection of adulteration of extra virgin olive oil using mid infrared and Raman spectroscopic data. Food Chemistry, 217, 735-742. DOI: 10.1016/j.foodchem.2016.09.011

Published in:
Food Chemistry

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Continuous statistical modelling for rapid detection of adulteration of extra virgin olive oil using mid infrared and Raman spectroscopic data

Konstantia Georgouli[a], Jesus Martinez Del Rincon[b], Anastasios Koidis[a,*]

[a] *Queens University Belfast, Institute for Global Food Security, Belfast, Northern Ireland, UK*

[b] *Queens University Belfast, Institute of Electronics, Communications and Information Technology, Belfast, Northern Ireland, UK*

Abstract

The main objective of this work was to develop a novel dimensionality reduction technique as a part of an integrated pattern recognition solution capable of identifying adulterants such as hazelnut oil in extra virgin olive oil at low percentages based on spectroscopic chemical fingerprints. A novel Continuous Locality Preserving Projections (CLPP) technique is proposed which allows the modelling of the continuous nature of the produced in-house admixtures as data series instead of discrete points. The maintenance of the continuous structure of the data manifold enables the better visualisation of this examined classification problem and facilitates the more accurate utilisation of the manifold for detecting the adulterants. The performance of the proposed technique is validated with two different spectroscopic techniques (Raman and Fourier transform infrared, FT-IR). In all cases studied, CLPP accompanied by k-Nearest Neighbors (kNN) algorithm was found to outperform any other state-of-the-art pattern recognition techniques.

Keywords: Continuous statistical modelling, dimensionality reduction, rapid detection, Adulteration, Extra virgin olive oil, FT-IR, RAMAN, spectroscopy

*Corresponding author

Email addresses: kgeorgouli01@qub.ac.uk (Konstantia Georgouli), j.martinez-del-rincon@qub.ac.uk (Jesus Martinez Del Rincon), t.koidis@qub.ac.uk (Anastasios Koidis)

2010 MSC: 00-01, 99-00

1. Introduction

The interdisciplinary collaborations between engineering, computer science and analytical science have led to the development of contemporary analytical instruments that allow the extraction of great amount of chemical information for a large number of samples relatively quickly and effortless. However, the produced analytical data (spectroscopic, chromatographic, isotopic, sensorial, etc.) are often multivariate data matrices which demand appropriate chemometric analysis. In chemometrics, mathematical and statistical methods are used for processing and capturing the most important and relevant content within the multivariate data. Despite the fact that a few multivariate methods are used in the area of food analysis either alone or in combination with other methods (Berrueta et al., 2007), there is an increasing demand for the introduction of novel and more intelligent pattern recognition methods for tackling more complex food analysis challenges such as food adulteration issues observed worldwide (Lohumi et al., 2015).

One of the most common adulterations occurring is mixing one commodity product or ingredient with another one in small percentages where the two ingredients are of a very similar chemical nature. In these cases, current chemometric techniques somehow fail to identify the fraudulent sample accurately (Ozen & Mauer, 2002; Šmejkalová & Piccolo, 2010) or use the same samples for both calibration and validation steps of the model (López-Díez et al., 2003; Christy et al., 2004), which biases the results. An indicative example of ongoing food fraud is the adulteration of extra virgin olive oil, a premium and high value commodity with renowned health properties (Zhang et al., 2011). Despite the establishment of a strict legislation framework, including specific analytical parameters defining the purity of the oil (International Olive Council, a; Agriculture and Rural Development, European Commission), the extra virgin olive oil adulteration with other lower value vegetable oils still remains

an important issue for the consumers and the olive oil sector alike (European Commission, 2013; Frankel, 2010).

One of these adulterants is hazelnut oil, which has very similar triacylglycerol, total sterol and fatty acid composition with extra virgin olive oil and has concerned numerous researchers (Pena et al., 2005; Parker et al., 2014; Koidis & Osorio Argüello, 2013). Extra virgin olive oil can be adulterated with hazelnut oil in two different ways: adulteration with crude hazelnut oil and adulteration with refined hazelnut oil. The identification of the adulteration with refined hazelnut oil is increasingly difficult due to the removal of markers like filberstone, a volatile compound unique to hazelnut oil, and other minor components through the refining process in addition to the similarity of the triacylglycerol profile of both oils (Flores et al., 2006).

Most research efforts aiming to address this adulteration problem have made use of chromatographic analytical methods. Despite providing satisfactory results by analysing the triacylglycerol content (International Olive Council, b), polar components (Zabaras & Gordon, 2004) and using sterol fractions, 4,4'-Dimethylsterols (Damirchi et al., 2005), n-alkanes (Webster et al., 2001) and filberstone (Flores et al., 2006) as possible markers, chromatographic methods involve complicated process steps, demand a large amount of time and financial resources and require access to laboratory facilities. Therefore, it is urgent to develop simple, inexpensive, rapid and accurate alternative methods to determine adulterants in extra virgin olive oil in environments that time and fast decisions are important (ports, control points, market surveys and other rapid testing environments).

Apart from chromatographic, several spectroscopic techniques in combination with chemometric methods have been proposed as rapid screening techniques for the authentication of extra virgin olive oil and the detection and quantification of its adulteration with hazelnut oil. Adulteration of olive oil with hazelnut oil at levels of 25% and higher was detected using Fourier transform infrared (FT-IR) coupled with partial least squares (PLS) analysis (Ozen & Mauer, 2002). Moreover, the same combination has been used for devel-

oping a method for the estimation of extra virgin olive oil adulteration with edible oils including hazelnut oil. The produced PLS models for the case of the hazelnut oil showed a relatively good performance (relative error of prediction, REP=20.8 and correlation factor R^2=0.9351) (Maggio et al., 2010). Multiple linear regression (MLR) models constructed using FT-IR data for extra virgin olive oil-hazelnut oil admixtures claim to be capable of detecting hazelnut oil content in olive oil with a 5% limit of detection (Lerma-García et al., 2010). In another study, high gradient diffusion NMR spectroscopy coupled with discriminant analysis (DA) was used for detecting rapidly the adulteration of extra virgin olive oils with seed and nut oils. The lower limit of detection for the case of hazelnut oil was 30% (Šmejkalová & Piccolo, 2010). The development of an artificial neural network in 600MHz ^{1}H-NMR and ^{13}C-NMR data achieved a limit of 8% (García-González et al., 2004). In a recent study, 60MHz ^{1}H NMR spectral data in combination with PLS regression achieved a limit of detection at the level of 11.2% w/w (Parker et al., 2014). However, it has to be highlighted that the aforementioned studies tackling this adulteration of extra virgin olive oil with little or great success do not claim explicitly if the hazelnut oil is refined or crude and they are not often validated adequately and correctly which might produce overestimated and /or overfitted results.

The detection of adulterants at low levels (5-20%) is still quite challenging even for high end methods such as chromatography (Zhang et al., 2011; Osorio et al., 2014a). There is a need for more research in the field of data analysis of complex chemical data, especially spectroscopic data which are by nature multivariate. More accurate statistical methods are required to be used on top of existing analytical methods that would not necessarily demand a large number of samples and are independent of statistical interpretations (Frankel, 2010).

The present work introduces a novel continuous statistical modelling technique which extends the Locality Preserving Projections (LPP) dimensionality reduction technique to the cases where data are considered as a continuous variable. Data are modelled as data series and the continuity is preserved during

the learning and dimensionality reduction by building two graphs incorporating neighbourhood information of the data set. In this way, the proposed technique has been designed, developed and tested coupled with k-Nearest Neighbors (kNN) classifier on the adulteration of extra virgin olive oil with hazelnut oil using spectra from two different spectroscopic techniques. Preliminary results obtained are compared with the performance of state-of-the-art supervised pattern recognition techniques.

2. Theory and algorithm

2.1. The proposed method: Continuous Locality Preserving Projections (CLPP)

Continuous Locality Preserving Projections technique is a semi-supervised linear method that enables the dimensionality reduction for learning manifolds characterised by continuous data. It extends the linear dimensionality reduction technique LPP (He & Niyogi, 2003) preserving continuity as in previous non-linear techniques such as Temporal Laplacian Eigenmaps (TLE) (Lewandowski et al., 2010). LPP was chosen as the base method due to its properties and advantages against other dimensionality reduction techniques such as principal component analysis (PCA) (Wold et al., 1987) or linear discriminant analysis (LDA) (Fisher, 1938), especially when the input data show linear properties (He & Niyogi, 2003). Given a set of $Y = y_1, y_2, \cdots, y_n$ data points in high dimensional space ($y_k \in R^D$) (see Fig. 1a), CLPP is able to transform this into its low dimensional space by mapping it to a set of points $Z = m_1, m_2, \cdots, m_n$ ($m_k \in R^d$) with $d \ll D$ (see Fig. 1b), while preserving the continuity of the data.

CLPP algorithm includes the construction of two different neighbourhood graphs preserving implicitly the continuous similarity in data points during the space transformation. These graphs express continuous dependencies and therefore local continuous neighbours in the high dimensional space are located nearby in the embedded space without enforcing any artificial embedded geometry. Two continuous neighbourhoods are produced for each data point m_k (see

Fig. 2):

- Continuous neighbourhood (C_k): the $2t$ nearest points in sequence of current data point:

$$C_k \in \{m_{k-t}, \ldots, m_k, \ldots, m_{k+t}\} \tag{1}$$

- Similarity neighbourhood (S_k): the r points parallel to m_k, acquired from the r repetitions of m_k in the r parallel trajectories $T_{(1..r)}$. Each trajectory is generated by the $2t$ continuous neighbours:

$$S_k \in \{T_{k,1}, \ldots, m_k, \ldots, T_{k,r}\} \tag{2}$$

Specifically, the steps for the dimensionality reduction comprise:

1. Assign weights to the edges of each graph using the LPP formulation:

$$Gc(k,j) = \begin{cases} e^{-||y^k - y^j||^2}, & \text{k,j} \in C_k. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

$$Gs(k,j) = \begin{cases} e^{-||y^k - y^j||^2}, & \text{k,j} \in S_k. \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

2. Compute the eigenvectors V of embedded space : The d eigenvectors V^* with the smallest nonzero eigenvalues make the embedded space. These eigenvectors and eigenvalues are calculated by solving the generalized eigenvalue problem:

$$\arg\min_{V^*}(V^T \cdot Y^T \cdot (L_C + \beta \cdot L_S) \cdot Y \cdot V) \tag{5}$$

subject to

$$V^T \cdot Y^T \cdot (D_C + \beta \cdot D_S) \cdot Y \cdot V = 1 \tag{6}$$

where $L_C = D_C - G_C$ and $L_S = D_S - G_S$ are the Laplacian matrices and D_C and D_S are diagonal matrices. β is a weighting factor for balancing the continuous and similarity variabilities.

CLPP applies the same principles than other continuous techniques that aim to preserve continuity (Lawrence, 2004; Lewandowski et al., 2010). Nevertheless, CLPP shows two main advantages regarding previous techniques: its simplicity and both directional mapping (from low to high and from high to low dimensional spaces) are provided automatically while reducing the space. This second advantage is crucial, since it has been proved that calculating those mappings from new data in non linear techniques is complex and inaccurate (Martinez-del Rincon et al., 2014). The linearity of the spectroscopic data as demonstrated by projecting them in a PCA space (Osorio et al., 2014b) proves the suitability of the CLPP to our application problem.

2.2. CLPP applied to oil adulteration

In order to apply CLPP framework to the extra virgin olive adulteration with hazelnut oil, it is important to understand how the raw data will be considered by the dimensionality reduction technique. Each adulterated olive oil sample will be considered as a data series T_r, where each data point m_k is the low dimensional representation of its corresponding spectra profile y_k at different percentage of adulteration from 0% to 100%, k=[0, 100]. M_{k+t} and m_{k-t}, composing the subset C_k, will be then the same oils admixture but at the immediate higher and lower levels of adulteration correspondingly. S_k will be the set of different adulterated oils samples (different olive oil samples or the same olive oil sample but adulterated with a different hazelnut oil) adulteration at the exact same level of adulteration k (see Fig. 2).

Following these indications, our new CLPP technique has potential to be applied to any food authenticity problem involving admixtures and/or adulteration. In this paper, the adulteration of vegetable oils is used as the test case.

2.3. Projection of new testing samples into CLPP space

Due to its linearity, CLPP provides a simple mapping function for projecting new testing samples between high and low dimensional space. Equation 7

provides the mapping mechanism for a new testing sample $Y_{test} \notin Y$, whose classification we want to estimate:

$$Z_{test} = V^{*T} * (Y_{test} - \overline{Y}) \tag{7}$$

where $\overline{Y}$ is the mean value of the Y, learned during the creation of the latent space.

3. Experimental results

3.1. Samples

Four extra virgin olive oil samples consisting of three Italian (var. *Toscano*, *Olivastra Seggianese* and *Tonda Iblea*) and one Greek (var. *Koroneiki*), two Turkish refined hazelnut oils and two crude hazelnut oils (Turkey and Italy) were collected directly from the producers. The olive oil samples were spiked accurately at percentages that vary from 1% to 90%.

A few adulteration levels are necessary for generating the desired continuity in the produced latent space, as it can be noticed in Fig. 3, which illustrates the space resulted by LDA and CLPP by using different number of adulteration levels for FT-IR data. Specifically, sixteen different concentration grades were selected, from 1% to 15% with an interval of 2, and from 20% to 90% with an interval of 10 (see Table 1). The higher resolution in the low concentrations of hazelnut oil was selected in order to cover the most challenging adulteration area (5-20%) to detect (Zhang et al., 2011). A total of 256 admixture samples were prepared for Raman and FT-IR spectroscopic analysis (n=264 samples including the pure extra virgin olive oils, refined hazelnut oils and crude hazelnut oils), belonging to 16 possible combinations between the 4 base extra virgin olive oils and the 4 hazelnut oil adulterants.

3.2. FT-IR/Raman spectral acquisition

For FT-IR spectroscopic analysis, the acquisition of all FT-IR spectra was performed using a Nicolet iS5 Thermo spectrometer (Thermo Fisher Scientific,

Dublin, Ireland) equipped with a DTGS KBr detector and a KBr beam splitter. Spectra were acquired from 4000 to 550 cm^{-1} co-addding 32 interferograms at 4 cm^{-1} resolution with a diamond attenuated total reflectance (iD5 ATR) accessory. Absorbance values were recorded at each spectrum point. Three replicates resulting in 7157 variables were measured for each sample and the average spectrum of these was used.

A benchtop Advantage 1064 Raman Spectrometer (DeltaNu Inc., Laramie, Wyoming, USA) with a scanning range from 200 to 2000 cm^{-1} and an excitation light of 1064 nm was used to collect the Raman spectra of the oil samples. The integration time for each Raman spectrum was 10 s. The final sample spectra was the average of two replicates with initial 1867 data points.

3.3. Data pre-treatment

The resulting FT-IR and Raman spectral profiles underwent some typical preprocessing techniques in order to reduce or remove any random or systematic variation in the data (Devos et al., 2014). This phase involves three steps. Specifically, Standard Normal Variate (SNV) (Barnes et al., 1989) and S-Golay filter (Savitzky & Golay, 1964) [polynomial order=2,frame size=9] were applied for removing the scatter and smoothing the data points respectively. At the end of this preprocessing procedure, the irrelevant spectra area was cut out. Regarding FT-IR, data fall between 690.39 and 1875.434 cm^{-1} and between 2750.476 and 3100.01 cm^{-1} which result in a spectrum of 3184 variables. In Raman dataset, 1038 variables between 800.314 and 1800.22 cm^{-1} were selected.

All chemometric data preprocessing was performed by means of in-house Matlab routines (The MathWorks Inc., USA).

3.4. Experimental setup

The performance of the proposed dimensionality reduction technique as part of a classification technique is evaluated by comparing it with the most used supervised pattern recognition techniques in the literature of food science (Berrueta et al., 2007), i.e. soft independent modelling of class analogy (SIMCA)

as the modelling method, partial least squares discriminant analysis (PLS-DA), kNN and nearest neighbour using Pearson's correlation for distance metric as discriminant methods, partial least squares (PLSR) (Wold et al., 1984) as the regression technique and unsupervised hierarchical clustering (UHC) (Di Girolamo et al., 2015) as an unsupervised learning technique. It is also compared against other pattern recognition techniques that we consider they have potential to tackle the adulteration problem. These were PCA + kNN, LDA + kNN and LDA + support vector machines (SVM) (Belousov et al., 2002) as discriminant methods. It has to be mentioned that the methodologies involving LDA also required PCA to be applied before LDA to reduce the dimensionality for solving LDA's limitation on a low sample-to-variable ratio (number of samples $\ll$ number of variables) (Szymańska et al., 2015). Parameter tuning was optimised empirically for every technique within the comparison in order to provide the highest classification rate in each of them. Details about the parameters values used in our measurements for gathering results are shown in the supplementary material. For CLPP, t=3 and r=5 were used in all experiments. It has to be noted that CLPP is a novel method that was conceived and developed by this research team and directly implemented in Matlab.

The main proposal of this work is the application of kNN on the CLPP space. CLPP has been also combined and tested with SVM, geodesic distance, clustering and Mahalanobis distance as classifiers for finding the best combination (data not shown). Furthermore, PLSR is applied in combination with the CLPP latent space for exploring the potential improvement regarding the conventional PLSR. The rationale of this experiment is that applying regression on a low dimensional space is simpler and computationally less expensive than on the raw data while preserving the advantages of regression outputs. For comparison purposes, the application of PLSR on PCA space was also examined.

As previously mentioned, two spectral datasets (Raman and FT-IR spectra) of 256 samples each were investigated for this work. It is accepted that to evaluate the classification ability of all the aforementioned multivariate techniques, the testing dataset must not be used in the building of the model

(Biancolillo et al., 2014). Therefore, experiments were conducted using leave-one-adulterated-oil-out cross validation in which two oils, one of the four extra virgin olive oils and one of the four hazelnut oils (crude or refined) and all their admixtures are taken for testing leaving the rest of them for the training of the model in each iteration. In total, sixteen iterations were performed for each experiment. Admixtures of the two testing oils with the remaining training oils are not used at all in the experiment iteration for producing unbiased, generalised and realistic results. This leads to training and testing sets consist of 168 samples and 18 samples respectively in each iteration.

The mean accuracy and the standard deviation over these iterations are the main evaluation metrics of this comparative analysis. Root mean square error (RMSE) of prediction was measured for the cases in continuous space (PLSR, PCA + PLSR and CLPP + PLSR) given the continuous nature of their output as an adulteration percentage in real numbers. For computing the classification rate for the PLSR experiments, if the PLSR output value of a testing sample is within the range of adulteration associated to a given class then this sample is classified to this specific class.

Two different classification scenarios on the adulteration of olive oil with hazelnut oil are considered with respect to the number of classes for establishing a clear idea of the behaviour of the compared techniques. Here the concept of the class is related to the expected level of resolution to be detected in the adulteration. The eighteen concentration grades (the 16 adulteration levels shown in Table 1 plus pure olive and pure hazelnut oil) of the in-house admixtures were grouped in 10 classes (1st class $\in$ [0,1), 2nd class $\in$ [1,5), 3rd class $\in$ [5,9), 4th class $\in$ [9,13), 5th class $\in$ [13,20), 6th class $\in$ [20,40), 7th class $\in$ [40,60), 8th class $\in$ [60,80), 9th class $\in$ [80,90), 10th class $\in$ [90,100]), where the numbers in the intervals represent the concentration of hazelnut oil within the mixture, in percentage. These classes were used for the calibration and validation of the model in a first scenario. Thereafter, the characterisation of a spectrum of an oil sample as pure extra virgin olive oil ($\in$ [0,1)), low adulterated extra virgin olive oil ($\in$ [1,12)), high adulterated extra virgin olive oil ($\in$ [12,90)) and mostly

pure hazelnut oil ($\in$ [90,100]) (4 classes) is addressed to the second scenario. This second scenario aims to evaluate the performance of our methodology in an adulteration screening system, where a simple decision is intended.

3.5. Discussion of the results

3.5.1. Qualitative analysis

An exploratory representation for FT-IR data is presented in Fig. 4 using PCA, LDA and CLPP with two latent dimensions. All three dimensionality reduction techniques were performed using the same values for the parameters for both scenarios (PCA: PCA_dims=2; LDA: LDA_dims=2; CLPP: CLPP_dims=2, β=0.50). The pattern of the mapped data of PCA and CLPP spaces remains similar in both scenarios. It appears that PCA, as an unsupervised dimensionality reduction technique, does not allow a clear separation of the admixtures for FT-IR data for all cases. Unlike PCA, admixtures are more discriminant in LDA space due to the pronounced supervised class membership. On the other hand, CLPP provides a better visualisation and dispersion of the continuous data. Specifically, it can be noticed that pure olive oils and hazelnut oils are plotted on the extremes of the produced CLPP arc respectively, whereas the different admixtures are lied across the arc that prove the data continuity. Similar conclusions can be drawn for Raman data (Figures not shown).

3.5.2. Quantitative analysis

The cross validation schema was applied as described in section 3.4 for two examined scenarios.

Classification problem with 10 classes. Table 2 presents the mean classification rate and the standard deviation of each pattern recognition technique. Only LDA and CLPP perform above the state-of-the-art techniques, i.e. SIMCA and PLS-DA in both Raman and FT-IR data. In spite of the difficulty and the complexity of this scenario, CLPP+kNN shows the best performance in both datasets regarding classification rate and standard deviation, obtaining around 40% of recognition rate of the adulteration level. In addition, the application

of CLPP on a PLSR framework performs better than the simple PLSR, which proves further the suitability of the CLPP reduced space to the adulteration problem. PLSR execution also exhibits a parallel reduction in the error of prediction (RMSE reducing from 0.19 to 0.18 for Raman spectral data and from 0.22 to 0.20 for FT-IR). PLSR on PCA space improves the classification ability of PLSR only using RAMAN spectra by retaining the same RMSE.

Classification problem with 4 classes. The decrease in the number of classes influences the classification considerably as it can be seen in Table 2. Using four different groups of classes, roughly 79% and 75% correct classification can be achieved with CLPP+kNN (see Table 2) in RAMAN and FT-IR respectively, being the best performing algorithm and with the smaller standard deviation (cross validation). Regardless of the number of classes in the problem, CLPP+PLSR enhances the performance of the simple PLSR in satisfied levels with simultaneous decrease in RMSE, from 0.23 to 0.18 for Raman and from 0.24 to 0.19 for FT-IR data. PCA+PLSR also improves the general PLSR performance and the RMSE (to 0.19 for Raman and to 0.20 for FT-IR), although in a smaller amount. Furthermore, an extra column has been included for indicating the classification ability of each technique in low percentages (1-12%) since this area is the most challenging for most analytical methods and particularly for rapid screening applications such as the current one. For the case of 10 classes, this area (1-12%) is not applicable since the number of classes provide already a more detailed partitioning. SIMCA exhibits a very low classification rate of 12.50% for Raman data because according to the literature it is very sensitive to handle unbalanced training datasets and classifies most testing samples to the class with the more representatives (12-90% hazelnut oil adulteration) (Alonso-Salces et al., 2010). CLPP+kNN exhibits again the highest performance in this measure for both datasets.

Referring to both scenarios, the option to model the adulteration of extra virgin olive oil with both crude hazelnut oil and refined hazelnut oil at the same time and the relatively small number of pure samples make the problem more

complicated and challenging but also demonstrate clearly the great potential of CLPP technique. Beyond the performance of CLPP+kNN, the classification ability of the application of PLSR on CLPP space is better compared with the simple PLSR and the qualitative analysis of the space is more continuous and coherent with the true nature of the data. In the first scenario, LDA+kNN and PCA+kNN produce comparable results with CLPP+kNN in some particular case. Although the difference between their performance is not statistical significant, since their error bars (see supplementary material) overlap i.e. P value > 0.05 (Cumming et al., 2007), CLPP+kNN is consistently more accurate and with smaller standard deviation in the most of the cases investigated. This can be justified from the systematic design of the training sample set that we designed and that allows the resulting latent space produced by LDA and PCA to become convergent to CLPP when the number of classes is large (see Fig. 4). Notably, the most widely applied and leading multivariate techniques like SIMCA, PLS-DA and PLSR, exhibit the weakest results in the condition of the first scenario where a ten classes classification problem is examined.

4. Conclusions

In this paper, a dimensionality reduction technique was developed to model the continuous nature of the admixtures as data series for addressing the adulteration of extra virgin olive oil with hazelnut oil. The food adulteration problem was modelled in two separate ways with a different number of classes. The results proved that CLPP coupled with kNN provides the best classification performance compared to state-of-the-art techniques (SIMCA, PLS-DA). This study confirms that the proposed solution could be very useful and effective for screening purposes. About 80% and 75% overall mean classification rate was obtained for the classification problem with four classes with more than 82% and 69% in low percentages (1%-12%) for Raman and FT-IR data respectively. Moreover, some interest remarks for the scientific chemometric community can be derived from this work. First, the adulteration problem is continuous by

nature and should be considered as such in the next generation chemometric analytic tools, as revealed by the low performance of current pattern recognition techniques and the improvement in performance when combining CLPP with PLSR in all investigated cases. Second, a detailed data with high number of samples and/or publicly available datasets for model training is crucial for developing new algorithms for tackling adulteration problems as evidenced by the good performance provided by LDA when samples were carefully prepared. Bearing in mind that this type of olive oil adulteration is a sophisticated and difficult analytical problem, this preliminary study demonstrates clearly that CLPP-based framework is able to preserve the continuous nature of the data that can be used for screening purposes on low adulteration olive oil mixtures.

Future work will look at the application of CLPP to other challenging food adulteration problems such as the authenticity of dairy powder and of herbs and spices, using FT-IR, Raman spectroscopic data, given CLPP's theoretical potential to be applied to any admixture problem, and higher number of samples.

Acknowledgements

This research was supported with funding from The Department Learning and Employment Northern Ireland (DELNI) and the Department of Environment, Food and Rural Affairs (DEFRA) of the UK.

References

Agriculture and Rural Development, European Commission (). Legislation on olive oil. `http://ec.europa.eu/agriculture/olive-oil/legislation/index_en.htm`. [Online; accessed 25-January-2016].

Alonso-Salces, R., Héberger, K., Holland, M., Moreno-Rojas, J., Mariani, C., Bellan, G., Reniero, F., & Guillou, C. (2010). Multivariate analysis of nmr fingerprint of the unsaponifiable fraction of virgin olive oils for authentication purposes. *Food Chemistry*, *118*, 956–965.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.*, *43*, 772–777. URL: `http://as.osa.org/abstract.cfm?URI=as-43-5-772`.

Belousov, A., Verzakov, S., & Von Frese, J. (2002). A flexible classification approach with optimal generalisation performance: support vector machines. *Chemometrics and Intelligent Laboratory Systems*, *64*, 15–25.

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

Biancolillo, A., Bucci, R., Magrì, A. L., Magrì, A. D., & Marini, F. (2014). Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication. *Analytica chimica acta*, *820*, 23–31.

Christy, A. A., Kasemsumran, S., Du, Y., & OZAKI, Y. (2004). The detection and quantification of adulteration in olive oil by near-infrared spectroscopy and chemometrics. *Analytical Sciences*, *20*, 935–940. doi:`10.2116/analsci.20.935`.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of cell biology*, *177*, 7–11.

Damirchi, S. A., Savage, G. P., & Dutta, P. C. (2005). Sterol fractions in hazelnut and virgin olive oils and 4, 4-dimethylsterols as possible markers for detection of adulteration of virgin olive oil. *Journal of the American Oil Chemists' Society*, *82*, 717–725.

Devos, O., Downey, G., & Duponchel, L. (2014). Simultaneous data preprocessing and svm classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food chemistry*, *148*, 124–130.

Di Girolamo, F., Masotti, A., Lante, I., Scapaticci, M., Calvano, C. D., Zambonin, C., Muraca, M., & Putignani, L. (2015). A simple and effective mass spectrometric approach to identify the adulteration of the mediterranean diet component extra-virgin olive oil with corn oil. *International journal of molecular sciences*, *16*, 20896–20912.

European Commission (2013). Workshop on olive oil authentication. `http://ec.europa.eu/agriculture/events/2013/olive-oil-workshop/newsletter_en.pdf`. [Online; accessed 25-January-2016].

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of eugenics*, *8*, 376–386.

Flores, G., Del Castillo, M. L. R., Blanch, G. P., & Herraiz, M. (2006). Detection of the adulteration of olive oils by solid phase microextraction and multidimensional gas chromatography. *Food chemistry*, *97*, 336–342.

Frankel, E. N. (2010). Chemistry of extra virgin olive oil: adulteration, oxidative stability, and antioxidants. *Journal of Agricultural and Food Chemistry*, *58*, 5991–6006.

García-González, D. L., Mannina, L., DImperio, M., Segre, A. L., & Aparicio, R. (2004). Using 1h and 13c nmr techniques and artificial neural networks to detect the adulteration of olive oil with hazelnut oil. *European Food Research and Technology*, *219*, 545–548.

He, X., & Niyogi, P. (2003). Locality preserving projections. In *Advances in Neural Information Processing Systems 16*.

International Olive Council (a). Standards. `http://www.internationaloliveoil.org/estaticos/view/222-standards`. [Online; accessed 25-January-2016].

International Olive Council (b). Testing methods: COI/T.20/DOC. NO 25 2013 Global Method for the detection of extraneous oils in

olive oils. http://www.internationaloliveoil.org/estaticos/view/224-testing-methods. [Online; accessed 25-January-2016].

Koidis, A., & Osorio Argüello, M. T. (2013). Identification of oil mixtures in extracted and refined vegetable oils. *Lipid Technology*, *25*, 247–250.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, *16*, 329–336.

Lerma-García, M., Ramis-Ramos, G., Herrero-Martínez, J., & Simó-Alfonso, E. (2010). Authentication of extra virgin olive oils by fourier-transform infrared spectroscopy. *Food Chemistry*, *118*, 78–83.

Lewandowski, M., Martinez-del Rincon, J., Makris, D., & Nebel, J.-C. (2010). Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 161–164). IEEE.

Lohumi, S., Lee, S., Lee, H., & Cho, B.-K. (2015). A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. *Trends in Food Science & Technology*, *46*, 85–98.

López-Díez, E. C., Bianchi, G., & Goodacre, R. (2003). Rapid quantitative assessment of the adulteration of virgin olive oils with hazelnut oils using raman spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry*, *51*, 6145–6150.

Maggio, R. M., Cerretani, L., Chiavaro, E., Kaufman, T. S., & Bendini, A. (2010). A novel chemometric strategy for the estimation of extra virgin olive oil adulteration with edible oils. *Food Control*, *21*, 890–895.

Osorio, M. T., Haughey, S. A., Elliott, C. T., & Koidis, A. (2014a). Evaluation of methodologies to determine vegetable oil species present in oil mixtures: Proposition of an approach to meet the eu legislation demands for correct vegetable oils labelling. *Food Research International*, *60*, 66–75.

Osorio, M. T., Haughey, S. A., Elliott, C. T., & Koidis, A. (2014b). Identification of vegetable oil botanical speciation in refined vegetable oil blends using an innovative combination of chromatographic and spectroscopic techniques. *Food Chemistry*, .

Ozen, B. F., & Mauer, L. J. (2002). Detection of hazelnut oil adulteration using ft-ir spectroscopy. *Journal of Agricultural and Food Chemistry*, *50*, 3898–3901.

Parker, T., Limer, E., Watson, A., Defernez, M., Williamson, D., & Kemsley, E. K. (2014). 60mhz 1 h nmr spectroscopy for the analysis of edible oils. *TrAC Trends in Analytical Chemistry*, *57*, 147–158.

Pena, F., Cárdenas, S., Gallego, M., & Valcárcel, M. (2005). Direct olive oil authentication: Detection of adulteration of olive oil with hazelnut oil by direct coupling of headspace and mass spectrometry, and multivariate regression techniques. *Journal of Chromatography A*, *1074*, 215–221.

Martinez-del Rincon, J., Lewandowski, M., Nebel, J.-C., & Makris, D. (2014). Generalized laplacian eigenmaps for modeling and tracking human motions. *Cybernetics, IEEE Transactions on*, *44*, 1646–1660.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, *36*, 1627–1639.

Šmejkalová, D., & Piccolo, A. (2010). High-power gradient diffusion nmr spectroscopy for the rapid assessment of extra-virgin olive oil adulteration. *Food Chemistry*, *118*, 153–158.

Szymańska, E., Gerretzen, J., Engel, J., Geurts, B., Blanchet, L., & Buydens, L. M. (2015). Chemometrics and qualitative analysis have a vibrant relationship. *TrAC Trends in Analytical Chemistry*, .

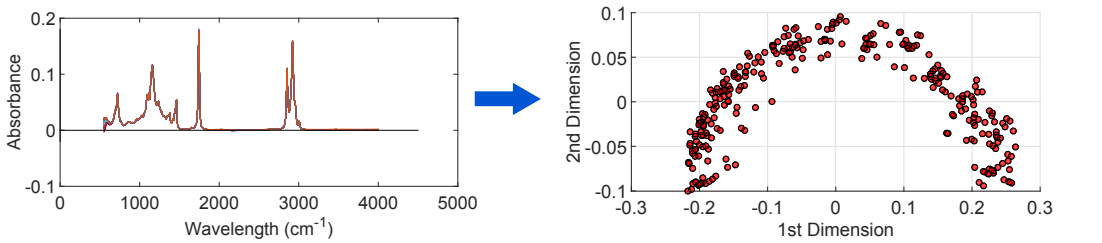
Webster, L., Simpson, P., & Shanks, A. (2001). Adulteration of olive oil with hazelnut oil: To enable detection of unrefined and refined hazelnut oil in virgin and refined olive oil, .

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*, 37–52.

Wold, S., Ruhe, A., Wold, H., & Dunn, W., III (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*, 735–743.

Zabaras, D., & Gordon, M. (2004). Detection of pressed hazelnut oil in virgin olive oil by analysis of polar components: improvement and validation of the method. *Food Chemistry*, *84*, 475–483.

Zhang, X., Qi, X., Zou, M., & Liu, F. (2011). Rapid authentication of olive oil by raman spectroscopy using principal component analysis. *Analytical Letters*, *44*, 2209–2220.



$Y = \{y_k\}_{(k=1..n)},\ y^k \in R^D$ (a)

$Z = \{m_k\}_{(k=1..n)},\ m^k \in R^d\ d \ll D$ (b)

Figure 1: Definition and application of CLPP: (a) Data points in high dimensional space; (b) Data points in low dimensional space.

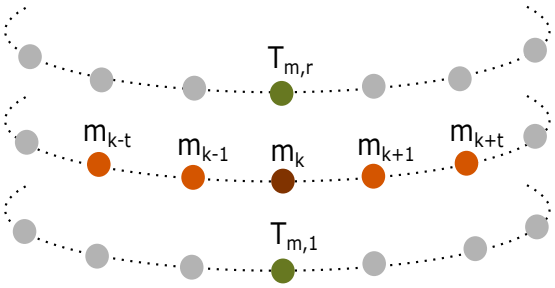


Figure 2: Continuous neighbours of a given sample, m_k: continuous (orange points) and similarity neighbours (green points).

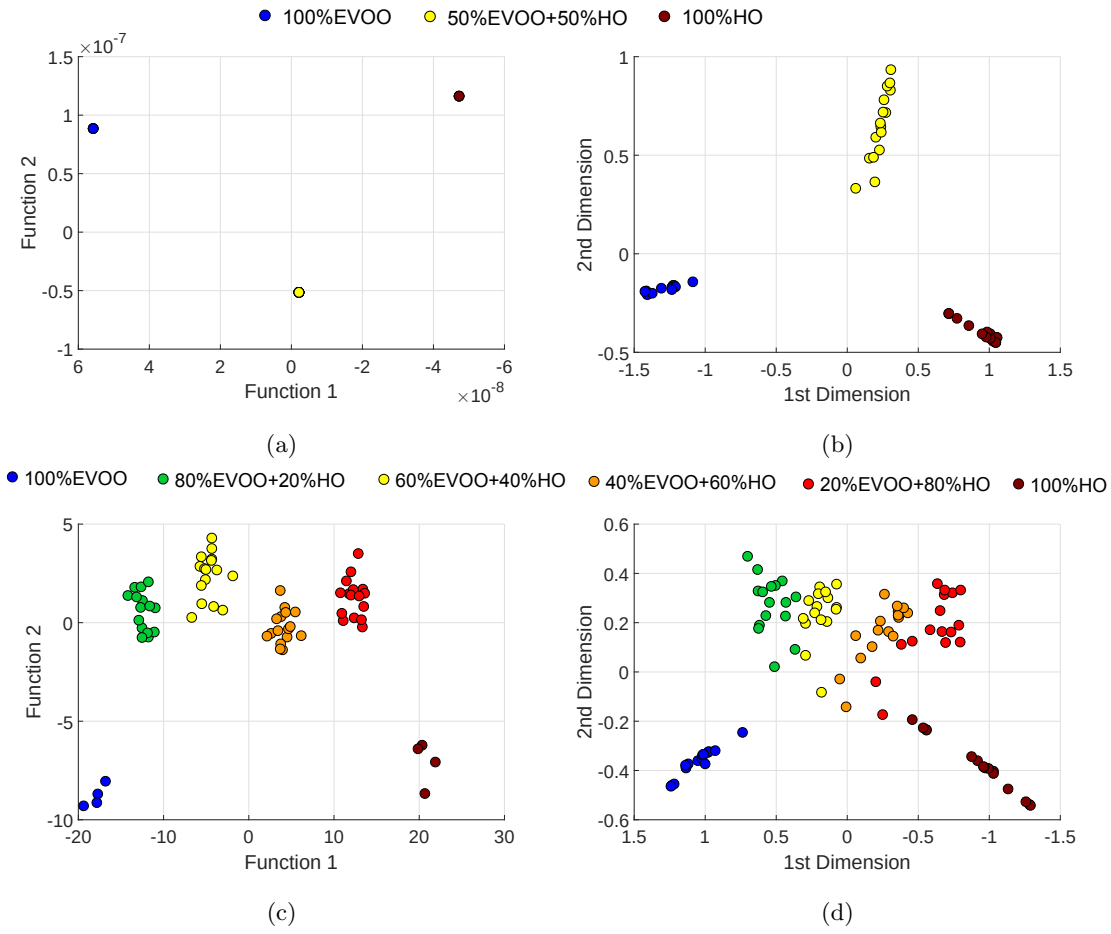
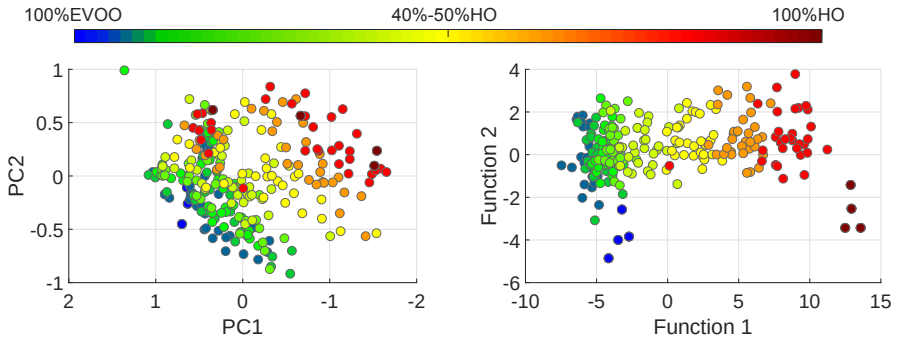
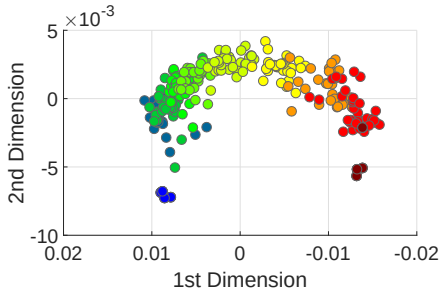


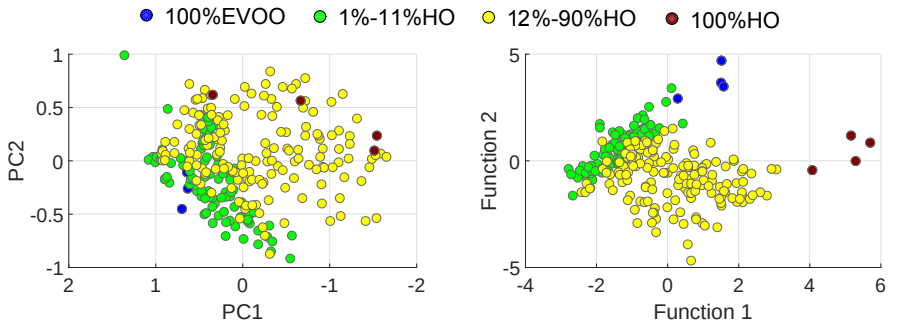
Figure 3: LDA space produced by FT-IR for: (a) three adulteration grades; (c) six adulteration grades. CLPP space for FT-IR data: (b) three adulteration grades; (d) six adulteration grades (see legend). EVOO, extra virgin olive oil; HO, hazelnut oil.



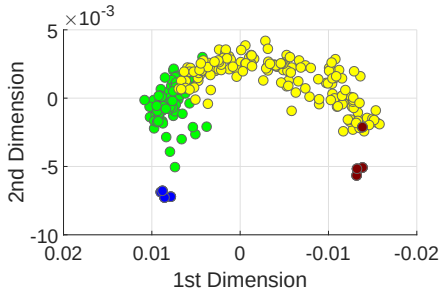
(a) (b)



(c)



(d) (e)



(f)

Figure 4: Exploratory analysis of FT-IR data for 10 classes: (a) PCA score plot; (b) LDA space; (c) CLPP space. For 4 classes: (d) PCA score plot; (e) LDA space; (f) CLPP space. EVOO, extra virgin olive oil; HO, hazelnut oil.

Table 1: Details of extra virgin olive and hazelnut oils for Raman and FT-IR analysis.

Admixtures		
Identity of the reference olive oil	Identity of the adulterant hazelnut oil	Concentration (%v/v) of hazelnut oil
EVOO1	RHO1	1.00
		3.00
		5.00
		7.00
		9.00
EVOO2	RHO2	11.00
		13.00
		15.00
		20.00
EVOO3	CHO1	30.00
		40.00
		50.00
		60.00
EVOO4	CHO2	70.00
		80.00
		90.00

'EVOO' indicates extra virgin olive oil; 'RHO' is refined hazelnut oil and 'CHO' is crude hazelnut oil.

Table 2: Mean classification rate (%) and standard deviations of the testing samples within each dataset for 10 different classes and for 4 percentage areas for the detection of olive oil adulteration using RAMAN and FT-IR.

CLASSIFICATION	RAMAN		FT-IR	
TECHNIQUE	**Overall (%)**	**For 1-12%**	**Overall(%)**	**For 1-12%**
For 10 different classes				
SIMCA	25.35±17.09	n/a	30.90±18.59	n/a
PLS-DA	26.39±8.24	n/a	25.69±10.12	n/a
PLSR	33.68±26.56	n/a	27.43±12.74	n/a
kNN	25.00±14.77	n/a	34.38±15.21	n/a
Pearson's correlation	26.04±15.01	n/a	30.90±15.45	n/a
UHC	23.96±11.06	n/a	21.18±10.78	n/a
PCA+kNN	25.00±14.77	n/a	35.07±16.45	n/a
LDA+kNN	40.63±25.15	n/a	32.29±19.05	n/a
LDA+SVM	33.33±19.25	n/a	26.61±13.98	n/a
PCA+PLSR	35.42±28.10	n/a	25.35±19.56	n/a
CLPP+PLSR	38.54±25.29	n/a	29.17±22.73	n/a
CLPP+kNN	**40.97±17.90**	**n/a**	**36.11±17.21**	**n/a**
For 4 different classes				
SIMCA	56.25±6.99	12.50	64.58±11.45	53.13
PLS-DA	66.32±14.41	65.63	64.93±12.94	58.33
PLSR	59.72±20.24	28.13	56.94±12.91	27.08
kNN	53.47±17.90	42.71	67.01±19.40	54.17
Pearson's correlation	54.17±18.31	43.75	68.75±15.57	58.33
UHC	58.68±11.47	57.79	56.60±13.02	56.25
PCA+kNN	53.82±16.94	41.67	68.06±16.67	58.33
LDA+kNN	74.31±13.59	72.92	69.44±15.45	61.46
LDA+SVM	63.19±14.47	57.29	60.07±28.13	64.58
PCA+PLSR	59.72±19.93	33.33	59.03±17.44	30.21
CLPP+PLSR	64.93±19.11	39.58	59.03±15.83	32.29
CLPP+kNN	**79.17±10.04**	**82.29**	**74.65±12.00**	**69.79**

SIMCA: Soft independent modelling of class analogy; PLS-DA: Partial least squares discriminant analysis; PLSR: Partial least squares regression; kNN: k-Nearest Neighbors; UHC: Unsupervised hierarchical clustering; PCA: Principal component analysis; LDA: Linear discriminant analysis; SVM: Support vector machines; CLPP: Continuous locality preserving projections; n/a: not applicable.