Geographical origin discrimination of lentils (*Lens culinaris* Medik.) using ^{1}H NMR fingerprinting and multivariate statistical analyses

Francesco Longobardi[a,]*, Valentina Innamorato[a], Annalisa Di Gioia[a], Andrea Ventrella[a], Vincenzo Lippolis[b], Antonio F. Logrieco[b] , Lucia Catucci[a,c], Angela Agostiano[a,c]

[a] Dipartimento di Chimica, Università di Bari “Aldo Moro”, Via Orabona 4, 70126 Bari, Italy.

[b] Consiglio Nazionale delle Ricerche (CNR), Istituto Scienze delle Produzioni Alimentari (ISPA),Via Amendola 122/O, 70126 Bari, Italy.

[c] Consiglio Nazionale delle Ricerche, Istituto per i Processi Chimico-Fisici (IPCF-CNR), sez. di Bari, Via Orabona 4, 70126 Bari, Italy.

*Corresponding author. Tel.: +39-080-5442042; fax: +39-080-5443607.

E-mail address: francesco.longobardi@uniba.it

Abstract

Lentil samples coming from two different countries, i.e. Italy and Canada, were analysed using untargeted ^{1}H NMR fingerprinting in combination with chemometrics in order to build models able to classify them according to their geographical origin. For such aim, Soft Independent Modelling of Class Analogy (SIMCA), k-Nearest Neighbor (k-NN), Principal Component Analysis followed by Linear Discriminant Analysis (PCA-LDA) and Partial Least Squares-Discriminant Analysis (PLS-DA) were applied to the NMR data and the results were compared. The best combination of average recognition (100%) and cross-validation prediction abilities (96.7%) was obtained for the PCA-LDA. All the statistical models were validated both by using a test set and by carrying out a Monte Carlo Cross Validation: the obtained performances were found to be satisfying for all the models, with prediction abilities higher than 95% demonstrating the suitability of the developed methods. Finally, the metabolites that mostly contributed to the lentil discrimination were indicated.

Keywords: ^{1}H NMR fingerprinting; Lentils; Geographical origin; Chemometrics

1. Introduction

Lentil (*Lens culinaris* Medik.) is the fourth most important pulse crop in the world after bean (*Phaseolus vulgaris* L.), pea (*Pisum sativum* L.), and chickpea (*Cicer arietinum* L.). Lentils are characterised by a high energy value and a high content of complex carbohydrates, proteins, dietary fibers, vitamins, minerals (de Almeida Costa, da Silva Queiroz-Monici, Pissini Machado Reis, & de Oliveira, 2006; Wang & Daun, 2006; Wang, Hatcher, Toews, & Gawalko, 2009) even if some anti-nutritional constituents are also present (Thavarajah, Thavarajah, See, & Vandenberg, 2010; Wang et al., 2009).

FAOSTAT reported that the world production of lentils was about 4.9 million of tons, primarily coming from Canada, India, Australia and Turkey; in particular, about a quarter of the production is from India but most of it is consumed in the domestic market, while Canada is the largest export producer of lentils in the world (FAOSTAT database 2014).

In Italy during the last years the lentil production declined from 14 k tons in the 60's to 1.9 k tons in 2014 due to several causes; therefore, as consequence, Italy annually imports about 29.6 million kg of lentils, mainly coming from Canada, USA, Turkey and China (Piergiovanni, 2000; Bacchi, Leone, Mercati, Preiti, Sunseri & Monti, 2010). However, Italian lentils, being cultivated mainly in specific localities, present unique and characteristic sensory and nutritional properties giving them a higher value; in fact, many Italian lentils gained international and national marks linked to their geographical origins, such as "protected geographical indication" (PGI), "traditional agricultural food products" (PAT) and Slow Food Presidium. Such labels allow to improve the commercial value of the food products, by guaranteeing a high quality level, and protect their typicality. Nevertheless, unscrupulous producers, driven by high illicit profits, often sell products that recall the "Italian Sounding" but are actually obtained blending or substituting the Italian products with foreign ones having low qualitative levels and commercial values.

Obviously, this kind of problems concerns not only the lentil production but all the traditional foods from raw materials to finished products. Therefore, it is clear why there is an increasing demand to have analytical methods able to certify the declared geographical origin of food products, in order to protect consumers and honest producers from fraud and unfair competition, respectively; consequently, during recent years, several food authentication techniques have been proposed (de la Guardia & Gonzalvez Illueca, 2013).

Among these techniques, the Nuclear Magnetic Resonance (NMR) has been considered a versatile and useful tool, due to its ability to provide a complete view of food metabolites, providing qualitative and quantitative information either on major and minor compounds (Mannina, Sobolev, & Viel, 2012). NMR has been regarded, in combination with multivariate statistical analysis, as a powerful tool for determining food quality and geographical origin, especially when used as untargeted method, where the whole spectra are used as fingerprints without assigning particular resonances to specific metabolites (Baiano, Terracone, Longobardi, Ventrella, Agostiano, & Del Nobile, 2012; Ferrara et al., 2013; Fiehn, 2001; Longobardi et al., 2012; Longobardi et al., 2013; Mannina, Patumi, Proietti, Bassi, & Segre, 2001; Vlahov, Del Re, & Simone, 2003).

As far as lentil authenticity is concerned, some studies are reported in literature. In particular, accessions of lentils from different countries were examined on the basis of some morphological characters by discriminant analysis and canonical analysis, showing regional grouping, even if misclassifications of individuals within groups were frequent (Erskine, Adham, & Holly, 1989). Moreover, the proteome of lentil seeds was used to identify specific markers and discriminate different plant landraces, through multivariate statistical analyses (Scippa et al., 2010).

In addition, DNA-based methods combined with high resolution melting analysis (Bosmali, Ganopoulos, Madesis, & Tsaftaris, 2012) were used to identify a particular lentil variety amongst other Greek varieties or admixtures, reaching a clear discrimination.

However, only few studies on geographical differentiation of lentil samples have been done; in particular, Diffuse Reflectance Fourier Transform Infrared Spectroscopy combined with discriminant analysis was proved to be convenient and fast, but the study, involving 27 samples grouped in two classes, i.e. “Greek” and “imported”, was carried out without performing a validation procedure, reducing the real applicability of the proposed method (Kouvoutsakis, Mitsi, Tarantilis, Polissiou, & Pappas, 2014). Other studies involved stable isotope ratios of $\delta^{13}C$, $\delta^{15}N$, whose values may depend on several factors, such as climatic parameters typical of the region (Zhang, Emeriau, & Martin, 1991); however, the $\delta^{2}H$, $\delta^{18}O$, $\delta^{34}S$ ratios are most linked to geographical origin (Rossmann, Reniero, Moussa, Schmidt, Versini, & Merle, 1999; Stöckigt, Schmidt, Rossmann, & Christoph, 2005; Ziegler, Osmond, Stichler, & Trimborn, 1976) and were analysed, in combination with chemometrics, to successfully discriminate geographical origin of lentils (Longobardi et al., 2015).

To the authors' knowledge, no study based on "NMR fingerprinting - multivariate statistical analysis" approach has been reported; thus, in this paper different statistical strategies, i.e. Principal Component Analysis followed by Linear Discriminant Analysis (PCA-LDA), k-Nearest Neighbor (k-NN), Partial Least Squares-Discriminant Analysis (PLS-DA), and Soft Independent Modeling of Class Analogy (SIMCA) were tested on ^{1}H NMR data of lentil samples aiming at discriminating them on the basis of their different geographical origin, i.e. Italy and Canada.

2. Materials and methods

2.1 Sample collection, sample preparation and NMR experiments

Lentil samples of the 2013 crop season were collected (as portions of about 500 g of seeds) from producers and supermarkets; the total number of samples was 85, subdivided into 43 Canadian (15 macrosperma and 27 microsperma subspecies) and 42 Italian (11 macrosperma and 31 microsperma) samples.

Herein, the sample preparation was carried out according to the procedure reported by Wu, Li, Li, & Tang (2014) with slight modifications, as reported in the following. After removing the foreign material, the lentil seeds were finely ground by using the Retsch ZM 200 (Retsch, Haan, Germany) laboratory mill equipped with 500-µm sieve and stored in sealed bags under vacuum until analysis.

About 400 mg of lentil flour were extracted with 4 mL of a mixture methanol/water (1:1, v/v) by mixing on vortex mixer for 10 s; the mixture was kept in an ice-water bath for 10 min. After centrifugation at 13000 rpm for 15 min, 3 mL of supernatant were transferred into a vial and dried under a nitrogen stream at 40°C, with the purpose to re-dissolve the dry extract in a smaller liquid volume constituted by 900 µL of buffer solution (phosphate buffer 50mM and NaN_3 1mM, pH 7.2) and 100 µL of 10 mM sodium salt of 3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid (TSP) in D_2O. This step allowed enhancing the signals related not only to major but also to minor lentil compounds. After 10 min of centrifugation at 9000 rpm, 600 µL of supernatant were transferred into NMR tubes (standard 5-mm tubes, Bruker BioSpin GmbH, Rheinstetten, Germany) for NMR measurements. All chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA).

One-dimensional ^1H NMR spectra were recorded on a Bruker Avance III 700 MHz NMR spectrometer (Bruker BioSpin GmbH, Rheinstetten, Germany) equipped with a cryogen cooled probe (cryoprobe QCI-^1H-^{19}F/^{13}C/^{15}N-^2H 5-mm with Z-gradient coils) using an autosampler (SampleXpress from Bruker BioSpin GmbH).

The spectra were acquired at 298 K under steady state conditions with non-spinning samples, using the Bruker 1D *noesygppr1d* pulse sequence. For each sample, 32 scans of 64 k data points with a receiver gain of 32 were recorded, applying a 90^0 pulse with an acquisition time of 2.28 s, a spectral width of about 20 ppm and a mixing time of 10 ms; during a relaxation delay of 10 s, a 25 Hz CW-based water peak suppression was performed. The offset for water suppression was previously optimised by applying a saturation power.

Each spectrum was recorded using TOPSPIN 3.1 software (Bruker BioSpin GmbH, Rheinstetten, Germany) in full automation mode in about 12 min. All NMR spectra were processed using the AU program apk0.noe, that automatically applied phase correction, baseline correction, and chemical shift correction referencing NMR spectra with respect to the TSP signal. NMR assignment of signal of metabolites was done through comparison with literature chemical shift data (Fan, 1996; Wu et al., 2014).

2.2 Bucketing and Chemometrics analysis

For spectra analyses, AMIX 3.8 software (Bruker BioSpin GmbH, Rheinstetten, Germany) was used. In particular, a "bucketing procedure" was applied to the NMR spectra after scaling all of them to the total intensity. In detail, the chemical shift axis of each spectrum, in the range 0.50-10.00 ppm (with the exclusion of the spectral region containing the suppressed water signal: 4.94-4.74 ppm), was divided into segments (buckets or bin) of a fixed width of 0.04 ppm converting each single spectrum into a row of values, i.e. the values assumed by the area subtended by the NMR intensity for each bucket considered. After that, each single spectrum was merged into a final matrix called "bucket table" composed by 233 columns (bin) and 85 rows (samples).

Statistical analyses were performed by using Statistica 8.0 (StatSoft Italia srl, Padova, Italy), V-Parvus 2010 (http://www.parvus.unige.it, Genova, Italy) and Classification Toolbox (Ballabio, & Consonni, 2013) in Matlab (Mathworks Inc., Natick, Massachusetts, USA). First of all, the Kennard and Stone Duplex algorithm (Casale et al., 2012) was applied in order to generate a subdivision of the whole dataset (bucket table) into a modeling (63 samples) and a test (22 samples) set; the modeling set was represented by 31 Italian and 32 Canadian samples, while the test set consited of 11 Italian and 11 Canadian samples. Then the modeling set, after removing outliers, was analysed by multivariate statistical techniques: in particular, the data were explored by means of the Principal Component Analysis (PCA) according to the NIPALS alghoritm (Jolliffe, 2002), while the samples were classified on the basis of their geographical origin carrying out discriminant statistical techniques, i.e. PCA-LDA, k-NN, PLS-DA (Barker & Rayens, 2003; Fisher, 1936; Oliveri & Downey, 2012), and also the class-modelling technique SIMCA (Wold & Sjöström, 1977). The suitability of a classification model coming from the discriminant techniques was evaluated by considering its recognition ability, i.e. its ability to correctly classify the samples used for the building of the model, and its cross-validation (CV) prediction ability, i.e. its ability to correctly classify samples of a test set generated in a V-fold cross validation (with V equal to 10). As regards the model obtained by SIMCA, its suitability was evaluated by considering its sensitivity (the percentage of samples correctly accepted by a class model) and its specificity (the percentage of samples correctly rejected by a class model). Finally, the models were validated by using the test set and a Monte Carlo Cross-Validation (MCCV) procedure. The MCCV procedure computes many models, each time creating a different evaluation set by random selection (each sample may fall many times, or even no times at all, in the evaluation set). In particular, a MCCV based on 1000 runs and involving a 20% of left-out samples in the evaluation sets was applied on the whole dataset (excluding outliers).

3. Results and discussion

In Figure 1 a typical ^{1}H NMR spectrum of a lentil extract is reported showing several signals, corresponding to many metabolites and in the following the main ones are commented. In particular, the triplet and the doublet observable at 0.93 ppm and 1.00 ppm can be assigned to the isoleucine methyl groups; the doublets at 0.96 and 0.94 ppm can be attributed to the methyl groups of leucine; at 0.98 and 1.01 ppm it is possible to notice the doublets attributed to the valine diastereotopic methyl groups; the lactate methyl group is responsible of the 1.33 ppm doublet, while the doublet at 1.48 ppm comes from the alanine methyl group; the multiplets at 1.56 ppm are due to γ-methylene protons of citrulline; the singlet at 1.92 ppm are due to the methyl group of acetate; the multiplets at 2.05, 2.12 and 2.34 ppm are due to protons of glutamate; the malate residue protons generate the 2.37 and 2.66 ppm double doublets and the 4.29 ppm double doublet; the doublets at 2.56 and 2.68 ppm are due to the methylene protons of citrate; the doublets observable at 2.68 and 2.81 ppm are attributable to the two aspartate diastereotopic methylene protons; the doublets of doublets at about 3.06 and 3.18 ppm, along with the doublets at about 6.90 and 7.18 ppm can be attributed to protons of tyrosine residue; the 3.2 and 3.22 ppm singlets can be assigned to choline and choline phosphate methyl groups, respectively; the singlet at 4.42 ppm, the triplet at 8.08 ppm, the multiplet at 8.82 ppm and the singlet at 9.11 ppm are attributed to protons of trigonelline; glucose is responsible for the intense signals in the 4.15–3.35 ppm range, for the 4.65 ppm doublet (ß anomer C1H) and for the 5.21 ppm doublet (α anomer C1H); the doublet at 4.59 ppm (C1H in ß anomer) and the doublet at 5.24 ppm (C1H in α anomer) are attributable to galactose; the doublet at 5.41 ppm are due to C1H of the glucose in sucrose; the doublet at 5.43 ppm are due to C1H of the glucose in raffinose family oligosaccharides (RFOs); the doublets at 5.92, 5.90 and 7.88 ppm are attributed to protons of uridine residue; the singlets at 8.25 and 8.35 ppm are due to C2H and C8H of inosine; the doublet at about 6.13 ppm, along with the singlets at 8.27 and 8.6 ppm are attributed to protons of inosine-5'-monophosphate residue; the singlet at 6.52 are due to protons of fumarate; the doublet at 7.72 is due to C4H of tryptophan; the doublet at about 7.33 ppm, along with the triplets at 7.37 and 7.43 ppm are attributed to protons of phenilalanine residue; the singlet at 8.46 ppm is due to protons of formate.

Subsequently, with the purpose to find if any anomalous sample was observable inside the space of a single class of origin, i.e. inside the single Italian or the single Canadian class, the data of the modelling set were processed by considering each class separately in a specific PCA model; the relevant influence plots were obtained and commented. In particular, Figure 2 represents the influence plots of the PCA models for the Italian (6 PCs explaining 78.0% of the total variance, Figure 2a) and the Canadian (6 PCs explaining 73.3% of the total variance, Figure 2b) classes, respectively. As observable, all the samples coming from a specific class fit in the relevant model (i.e. stay inside the space delimited by the two straight lines defining the model confidence limits at a level of 95%) with the exception of two outliers, one for Italy and one for Canada, that therefore were excluded from data in the further statistical treatments.

In order to get general indications about the capacity of the NMR variables to discriminate lentil samples on the basis of their different places of production, the new training set (30 Italian and 31

Canadian samples) was subjected to PCA. By plotting the sample scores in a PC1 vs. PC2 graph (Figure 3), overlapping regions were observed, obtaining only a modest visual clustering of the objects on the basis of the geographical origin (PC1 and PC2 explained respectively 32.6% and 21.7% of the total variance). No significant separation was evidenced even when observing the score plots of the remaining PCs. These findings highlighted the necessity to process the data by using supervised techniques, as commented in the following.

As first approach, SIMCA, a class-modelling technique, was used to classify the lentil samples coming from the two different geographical origins. From a general point of view, the class-modelling techniques aim at looking for similarities occurring among samples of the same class and model each category separately from the other ones, building them like defined space areas in the hyper-space of the model at a specified confidence level. Therefore, an object could be assigned to more than one class if it lies in an overlapping region, or it is even possible that a sample is assigned to none of the modeled classes, as long as it does not fit in any of the class spaces. This last feature could be particularly useful for the aims of this work, since it would be possible to know if a real sample comes from Italy, Canada or even from another different and not specified geographical origin.

The optimal complexity of the model, i.e. the number of PCs to be used to describe the class variability, was chosen on the basis of a CV procedure (V=10). In particular, the geometric average between sensitivity and specificity in CV was selected as an optimality criterion, so that the number of PCs was chosen as the one corresponding to the highest value of this figure of merit. In such a way, the optimal complexity of the model resulted to be in 5 PCs for each class of geographical origin at a confidence level of 95%. The SIMCA results are visualized by the Coomans plot in Figure 4: as showed, 5 Italian and 4 Canadian samples resulted to be out of the relevant SIMCA model boundaries, represented by the vertical and horizontal lines, respectively, thus demonstrating moderate sensitivities for both classes; in fact, the SIMCA model showed 85.7% mean sensitivity, since 25 Italian samples over a total of 30 were accepted by the relevant class model, with a specific sensitivity of 83.3%, while 27 Canadian samples over 31 were correctly accepted by the Canadian class model, with a specific sensitivity of 87.1%. Moreover, not satisfying results were obtained with regard to specificity; indeed, even if the mean specificity was 80.3% and all the Italian samples except one were not accepted by the Canadian model, resulting in a 96.7% model specificity, eleven Canadian objects over a total of 31 were incorrectly identified as Italians, resulting in a low Italian class specificity (64.5%). This latter result makes the SIMCA approach not suitable for the main aim of this work: indeed, it is not capable to satisfactorily indicate if Canadian samples are fraudulently sold as Italian ones, which is the most common fraud regarding Italian lentils.

Taking into account all reported above, it was considered advantageous to test other statistical analyses, such as discriminant techniques, which are more suitable for classification aims (Berrueta, Alonso-Salces, & Héberger, 2007). As a consequence, the classification techniques k-NN, LDA and PLS-DA were applied and their results were summarised in Table 1.

For k-NN, different k values were tested evaluating for each of them the prediction error rate in cross-validation (V=10); the smallest k value achieving the lowest error was 5 and therefore was selected as the optimal one. In detail, the recognition (classification) and the CV prediction ability were both 95.1%. This means that k-NN correctly classified and predicted 30 Canadian samples out of 31 and 28 Italian samples out of 30.

As a second discriminant technique, herein LDA was applied; preliminarily, a variable reduction was adopted in order to make the number of variables lower than the (n-g)/3 value (with n representing the number of samples, and g standing for the number of groups), so avoiding overfitting risks, as reported (Berrueta et al., 2007; Defernez & Kemsley, 1997).

In this work, the number of variables was reduced by applying PCA and selecting the first 20 PCs, so leading to a final PCA-LDA model. The value for the recognition (classification) ability was 100% and the value for the CV prediction ability was 96.7%, i.e. it correctly predicted 28/30 Italian samples and 31/31 Canadian samples.

As last supervised discriminant technique, here PLS-DA was applied: this particular technique has the advantage to process large data set, even when the sample number exceeds the number of variables. By implementing a 10-fold cross-validation, it was found that 5 latent variables guaranteed the optimal model complexity, leading to a 98.4% average recognition rate: more in detail, the totality of the Canadian samples were correctly classified, and only one over 30 Italian samples was not correctly assigned. The average CV prediction rate gained a value of 96.7%: the CV prediction abilities for the single Canadian and Italian categories were found to be 93.3% and 100%, respectively.

Nevertheless, it is well-established that the use of an external validation procedure is highly recommended to evaluate the reliability of a model in the prediction of unknown samples. Therefore, the models employed herein were validated and compared calculating the prediction abilities obtained both on the test set (soft validation) and by a Monte Carlo Cross-Validation (MCCV, hard validation). In soft validation, for PLS-DA the prediction abilities were found to be 100.0% for Canada and 90.9% for Italy, corresponding to an average prediction rate equal to 95.4%. Regarding the Italian category, only 1 over a total of 11 samples was misclassified. For k-NN and PCA-LDA, the resulting prediction abilities were 100.0% both for Italy and Canada.

The hard validation procedure evidenced a prediction ability of 95.3% for both PLS-DA and PCA-LDA, and of 95.2% for k-NN. These findings evidence that the topic of the discrimination of the lentil geographical origin is well addressed by the use of NMR data in combination with supervised statistical techniques.

With the purpose to get information about the metabolites responsible for the geographical discrimination, a combination of univariate and multivariate analysis was used (Wang et al., 2014; Cuevas, Moreno-Rojas, Arroyo, Daza, & Ruiz-Moreno, 2016). In particular, the potential discriminant metabolites were identified as the ones having both PLS-DA variable importance in the projection values (VIP) higher than 1 (supervised multivariate critrerion) and statistically different means on the basis of the geographical origin (t-test as univariate criterion, $p \leq 0.01$).

PLS-DA was used for the multivariate part of such criterion, since it was directly referable to the original variables and consequently to the metabolites; k-NN, in fact, does not provide any explicit classification rule based on the data patterns, and the PCA-LDA model was built by using PCs, and therefore resulted more difficult to directly relate to the original NMR variables for the relevant comments.

According to the adopted dual criterion, it was highlighted that the most contributing buckets were in the regions containing signals of isoleucine, alanine, citrulline, acetate, malate, citrate, aspartate, choline, choline phosphate, galactose, glucose in sucrose and in RFOs and other unidentified compounds; consequently, such metabolites can be considered important for the discrimination of the geographical origin of lentils. In particular, the mean values of isoleucine, alanine, citrulline, acetate, choline, choline phosphate, galactose were higher in Italian lentils than in Canadian ones; on the contrary the means of malate, citrate, aspartate, glucose in sucrose and in RFOs resulted to be higher in Canadian samples than in Italian lentils. However, for all indicated compounds, the data distributions around the mean values of the two origins overlapped, consequently no specific single markers were found confirming the need to employ supervised multivariate methods for origin discrimination.

By comparing the results here obtained with the results gained in a previous work regarding the use of IRMS for the same aim (Longobardi et al., 2015), it can be noticed that both techniques are valid and show a vocation for this kind of studies and applications. The choice of one of them should take into account a balance of advantages and drawbacks of each technique. In particular, although NMR is a more expensive technique (both considering the purchase and the maintenance of the instrumentation) and it needs highly specialized operators, it shows high repeatability and therefore it does not need replicates; moreover, NMR could provide qualitative and quantitative information about the metabolites contained in the analysed sample. On the other hand, even if IRMS cannot give an extensive description regarding the analytes but only bulk information, and even if it needs replicates, it is cheaper and easier to be performed.

4. Conclusions

This work contributed to highlight the advantages of applying ^{1}H NMR fingerprinting as instrumental technique, and k-NN, PCA-LDA and PLS-DA as statistical techniques, in the classification of the geographical origin of lentil samples.

In particular, the PCA-LDA model allowed obtaining the best performances with a recognition ability of 100%, a CV prediction ability of 96.7%, and external prediction rates of 100% and 95.3% on the test set and by a MCCV procedure, respectively. Moreover, very good results were obtained also with k-NN and PLS-DA discriminant models highlighting that the NMR data contained enough information to build adequate models. In addition, a pattern of metabolites which mostly contributed to the lentil discrimination based on their geographical origin was identified.

In conclusion, it can be stated that although the proposed NMR method could be considered expensive and it requires highly specialized operators, it is capable to give high prediction abilities and repeatability if used to solve geographical origin issues of lentils, offering in addition the possibility to obtain information about sample metabolites. This work open up possibilities to extend the results here obtained to different lentil crop seasons, even using a higher number of samples. A further improvement in the lentil authenticity topic could regard studying relationships occurring between lentil chemical composition and detailed pedoclimatic parameters by using NMR data.

Acknowledgements

This work has been carried out within "Apulian Food Fingerprint" project (Intervento Reti di Laboratori Pubblici di Ricerca PO Puglia FESR 2007-2013, Asse I, Linea 1.2 – PO Puglia FSE 2007–2013 Asse IV).

References

Bacchi, M., Leone, M., Mercati, F., Preiti, G., Sunseri, F., & Monti, M. (2010). Agronomic Evaluation and Genetic Characterization of Different Accessions in Lentil (Lens culinaris Medik.). *Italian Journal of Agronomy, 5,* 303–314.

Baiano, A., Terracone, C., Longobardi, F., Ventrella, A., Agostiano, A., & Del Nobile, M. A. (2012). Effects of different vinification technologies on physical and chemical characteristics of Sauvignon blanc wines. *Food Chemistry, 135,* 2694–2701.

Ballabio, D., Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods*, 5, 3790–3798.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17,* 166–173.

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

Bosmali, I., Ganopoulos, I., Madesis, P., & Tsaftaris, A. (2012). Microsatellite and DNA-barcode regions typing combined with High Resolution Melting (HRM) analysis for food forensic uses: A case study on lentils (Lens culinaris). *Food Research International, 46,* 141–147.

Casale, M., Oliveri, P., Casolino, C., Sinelli, N., Zunin, P., Armanino, C., Lanteri, S. (2012). Characterisation of PDO olive oil Chianti Classico by non-selective (UV-visible, NIR and MIR spectroscopy) and selective (fatty acid composition) analytical techniques. *Analytica Chimica Acta*, *712*, 56–63.

Cuevas, F. J., Moreno-Rojas, J. M., Arroyo, F., Daza, A., Ruiz-Moreno, M. J. (2016). Effect of management (organic vs conventional) on volatile profiles of six plum cultivars (Prunus salicina Lindl.). A chemometric approach for varietal classification and determination of potential markers. *Food Chemistry, 199,* 479–484.

de Almeida Costa, G. E., da Silva Queiroz-Monici, K., Pissini Machado Reis, S. M., & de Oliveira, A. C. (2006). Chemical composition, dietary fibre and resistant starch contents of raw and cooked pea, common bean, chickpea and lentil legumes. *Food Chemistry*, *94*, 327–330.

de la Guardia, M., & Gonzalvez Illueca, A. (2013). *Food Protected Designation of Origin: Methodologies and Applications (Comprehensive Analytical Chemistry)*. Elsevier.

Defernez, M., & Kemsley, E. K. (1997). The use and misuse of chemometrics for treating classification problems. *TrAC Trends in Analytical Chemistry*, *16*, 216–221.

Erskine, W., Adham, Y., & Holly, L. (1989). Geographic distribution of variation in quantitative traits in a world lentil collection. *Euphytica*, *43*, 97–103.

Fan, T. W. M. (1996). Metabolite profiling by 1D and 2D NMR analysis of complex mixtures. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *28*, 161–219.

FAOSTAT database 2014.

Ferrara, G., Mazzeo, A., Matarrese, A. M. S., Pacucci, C., Pacifico, A., Gambacorta, G., Mastrorilli, P. (2013). Application of Abscisic Acid (S-ABA) to “Crimson Seedless” Grape Berries in a Mediterranean Climate: Effects on Color, Chemical Characteristics, Metabolic Profile, and S-ABA Concentration. *Journal of Plant Growth Regulation*, *32*, 491–505.

Fiehn, O. (2001). Combining Genomics, Metabolome Analysis, and Biochemical Modelling to Understand Metabolic Networks. *Comparative and Functional Genomics*, *2*, 155–168.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*, 179–188.

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer.

Kouvoutsakis, G., Mitsi, C., Tarantilis, P. A., Polissiou, M. G., & Pappas, C. S. (2014). Geographical differentiation of dried lentil seed (Lens culinaris) samples using Diffuse Reflectance Fourier Transform Infrared Spectroscopy (DRIFTS) and discriminant analysis. *Food Chemistry*, *145*, 1011–1014.

Longobardi, F., Ventrella, A., Napoli, C., Humpfer, E., Schütz, B., Schäfer, H., Sacco, A. (2012). Classification of olive oils according to geographical origin by using ^{1}H NMR fingerprinting combined with multivariate analysis. *Food Chemistry*, *130*, 177–183.

Longobardi, F., Ventrella, A., Bianco, A., Catucci, L., Cafagna, I., Gallo, V., Mastrorilli, P., Agostiano, A. (2013). Non-targeted ^{1}H NMR fingerprinting and multivariate statistical analyses for the characterisation of the geographical origin of Italian sweet cherries. *Food Chemistry*, *141*, 3028–3033.

Longobardi, F., Casiello, G., Cortese, M., Perini, M., Camin, F., Catucci, L., & Agostiano, A. (2015). Discrimination of geographical origin of lentils (Lens culinaris Medik.) using isotope ratio mass spectrometry combined with chemometrics. *Food Chemistry*, *188*, 343–349.

Mannina, L., Patumi, M., Proietti, N., Bassi, D., & Segre, A. L. (2001). Geographical Characterization of Italian Extra Virgin Olive Oils Using High-Field ^{1}H NMR Spectroscopy. *Journal of Agricultural and Food Chemistry*, *49*, 2687–2696.

Mannina, L., Sobolev, A. P., & Viel, S. (2012). Liquid state ^{1}H high field NMR in food analysis. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *66*, 1–39.

Oliveri, P., & Downey, G. (2012). Multivariate class modeling for the verification of food-authenticity claims. *TrAC - Trends in Analytical Chemistry*, *35*, 74–86.

Piergiovanni, A. R. (2000). The evolution of lentil (Lens culinaris Medik.) cultivation in Italy and its effects on the survival of autochthonous populations. *Genetic Resources And Crop*

Evolution, *47*, 305–314.

Rossmann, A., Reniero, F., Moussa, I., Schmidt, H.-L., Versini, G., & Merle, M. H. (1999). Stable oxygen isotope content of water of EU data-bank wines from Italy, France and Germany. *Zeitschrift Für Lebensmitteluntersuchung Und - Forschung A*, *208*, 400–407.

Scippa, G. S., Rocco, M., Ialicicco, M., Trupiano, D., Viscosi, V., Di Michele, M., Scaloni, A. (2010). The proteome of lentil (Lens culinaris Medik.) seeds: Discriminating between landraces. *Electrophoresis*, *31*, 497–506.

Stöckigt, D., Schmidt, H.-L., Rossmann, A., & Christoph, N. (2005). Herkunft und Authentizität von Lebensmitteln: Stabilisotopenanalytik. *Chemie in Unserer Zeit*, *39*, 90–99.

Thavarajah, D., Thavarajah, P., See, C. T., & Vandenberg, A. (2010). Phytic acid and Fe and Zn concentration in lentil (Lens culinaris L.) seeds is influenced by temperature during seed filling period. *Food Chemistry*, *122*, 254–259.

Vlahov, G., Del Re, P., & Simone, N. (2003). Determination of Geographical Origin of Olive Oils Using ^{13}C Nuclear Magnetic Resonance Spectroscopy. I − Classification of Olive Oils of the Puglia Region with Denomination of Protected Origin. *Journal of Agricultural and Food Chemistry*, *51*, 5612–5615.

Wang, N., & Daun, J. K. (2006). Effects of variety and crude protein content on nutrients and anti-nutrients in lentils (Lens culinaris). *Food Chemistry*, *95*, 493–502.

Wang, N., Hatcher, D. W., Toews, R., & Gawalko, E. J. (2009). Influence of cooking and dehulling on nutritional composition of several varieties of lentils (Lens culinaris). *LWT - Food Science and Technology*, *42*, 842–848.

Wang, C., Dong, R., Wang, X., Lian, A., Chi, C., Ke, C., Guo, L., Liu, S., Zhao, W., Xu, G., & Li, E. (2014). Exhaled volatile organic compounds as lung cancer biomarkers during one-lung ventilation. *Scientific reports*, *4*, 1–8.

Wold, S., & Sjöström, M. (1977). SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy (pp. 243–282).

Wu, X., Li, N., Li, H., & Tang, H. (2014). An optimized method for NMR-based plant seed metabolomic analysis with maximized polar metabolite extraction efficiency, signal-to-noise ratio, and chemical shift consistency. *Analyst*, *139*, 1769–1778.

Zhang, B. L., Emeriau, L., & Martin, G. J. (1991). Comportements isotopiques comparés de constituants de légumineuses - Caractérisation de lentilles. *Sciences Des Aliments*, *11*, 291–304.

Ziegler, H., Osmond, C. B., Stichler, W., & Trimborn, P. (1976). Hydrogen isotope discrimination in higher plants: Correlations with photosynthetic pathway and environment. *Planta*, *128*, 85–92.

Figure captions

Figure 1. One-dimensional ^{1}H NMR spectrum of a lentil sample, obtained with selective suppression of the water signal.

Figure 2. Influence plots obtained for the Italian PCA model (a) and for the Canadian PCA model (b) at a confidence level of 95%. Geographical origins: Italy (□), Canada (●).

Figure 3. PC1 vs. PC2 scatter plot for lentil samples. Geographical origins: Italy (□), Canada (●).

Figure 4. Coomans plot for the Italian and Canadian SIMCA models with a confidence interval equal to 95%. Geographical origins: Italy (□), Canada (●).

Table 1

Recognition and prediction abilities for the k-NN, PCA-LDA and PLS-DA models classifying lentils according to their geographical origin.

Classification technique	**Model performance (%)**
***k*-NN**	
Recognition ability (modelling)	95.1
Prediction ability (CV 10)	95.1
Prediction ability (test set)	100
Prediction ability (MCCV)	95.2
PCA-LDA	
Recognition ability (modelling)	100
Prediction ability (CV 10)	96.7
Prediction ability (test set)	100
Prediction ability (MCCV)	95.3
PLS-DA	
Recognition ability (modelling)	98.3
Prediction ability (CV 10)	96.7
Prediction ability (test set)	95.4
Prediction ability (MCCV)	95.2

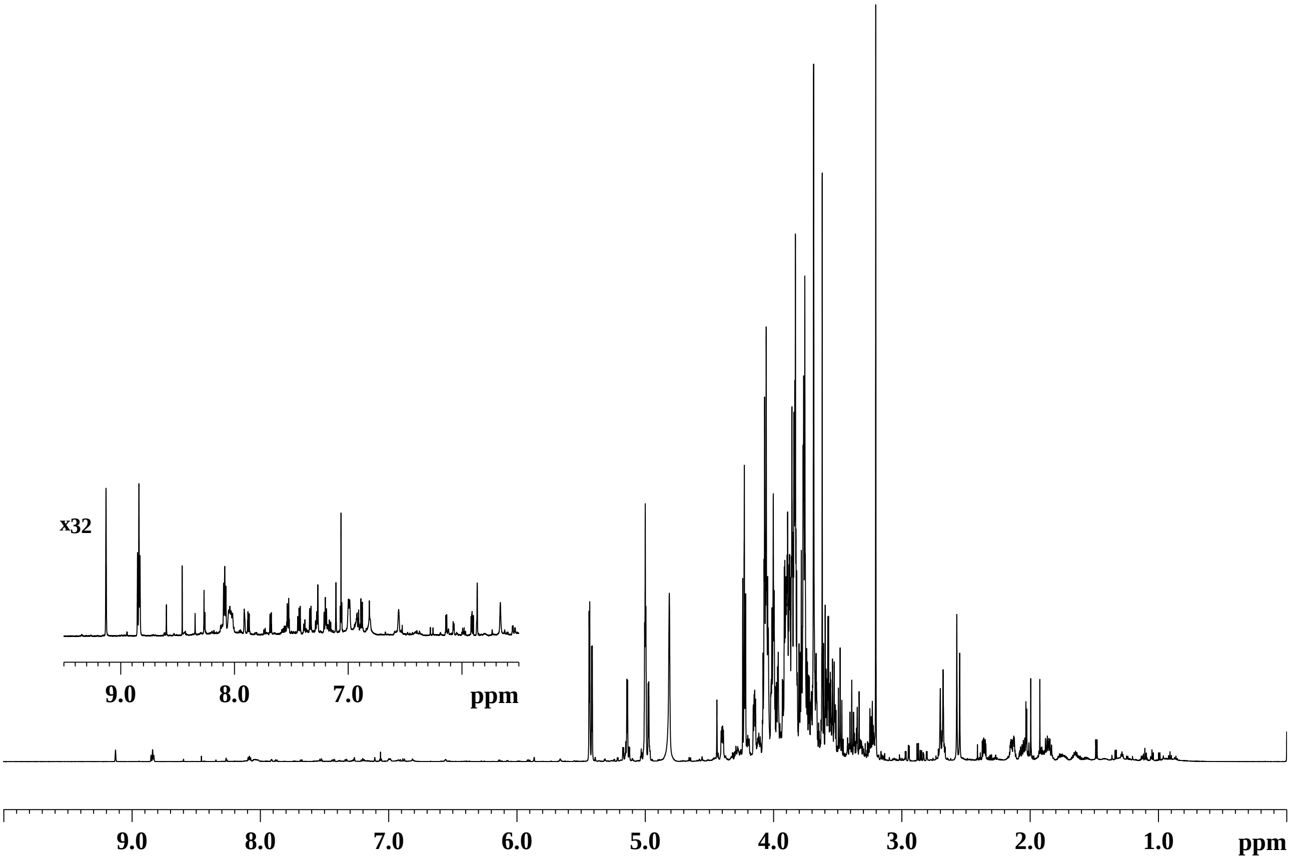

Figure 1

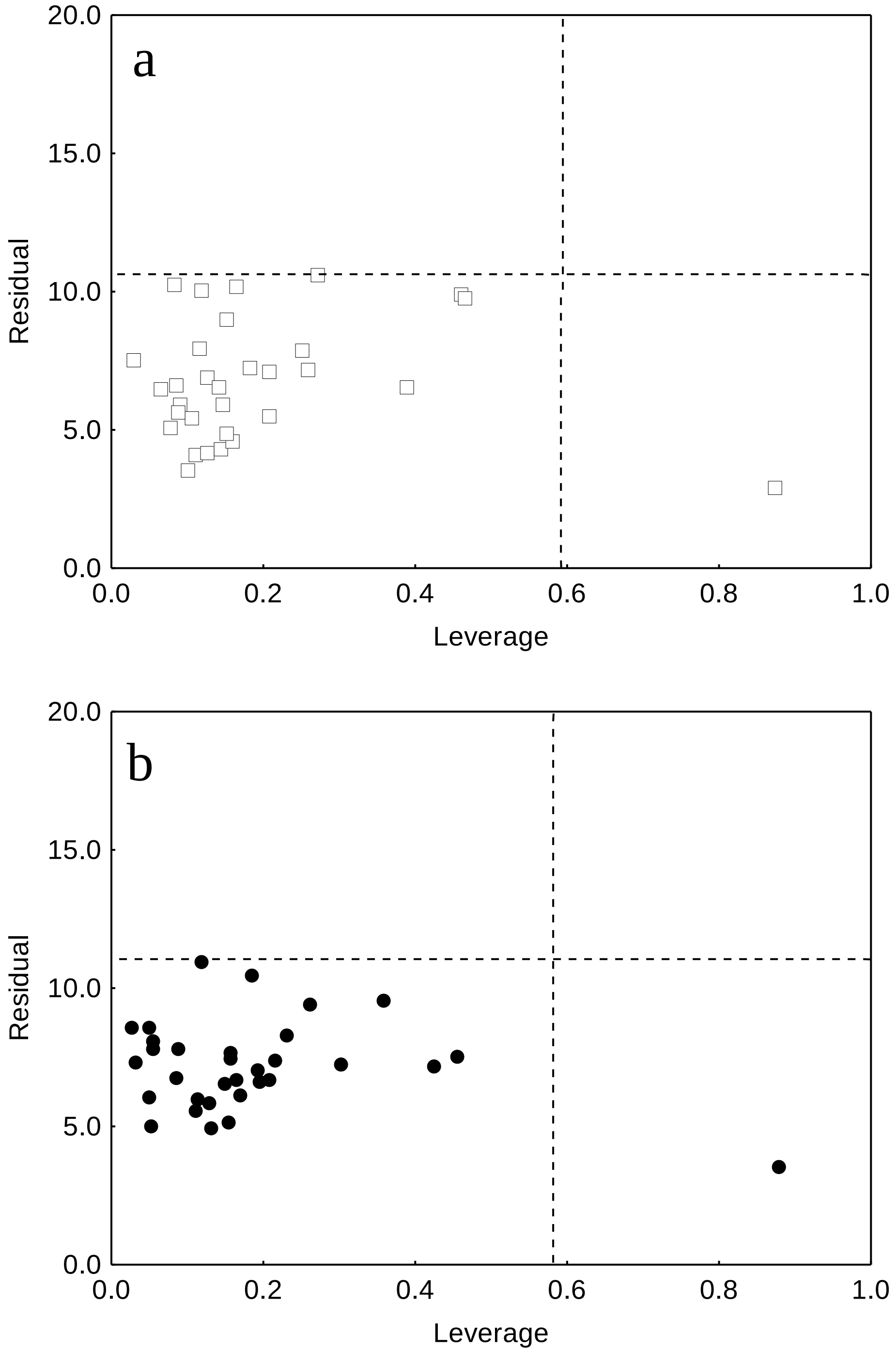
a
20.0
15.0
10.0
5.0
0.0
Residual
0.0
0.2
0.4
0.6
0.8
1.0
Leverage
b
20.0
15.0
10.0
5.0
0.0
Residual
0.0
0.2
0.4
0.6
0.8
1.0
Leverage

Figure 2

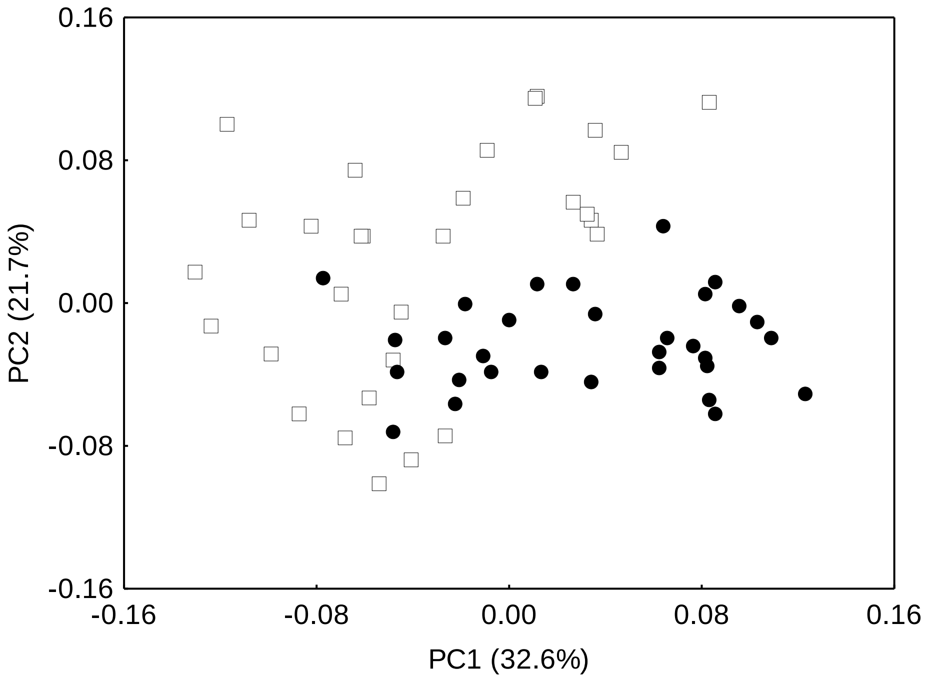

Figure 3

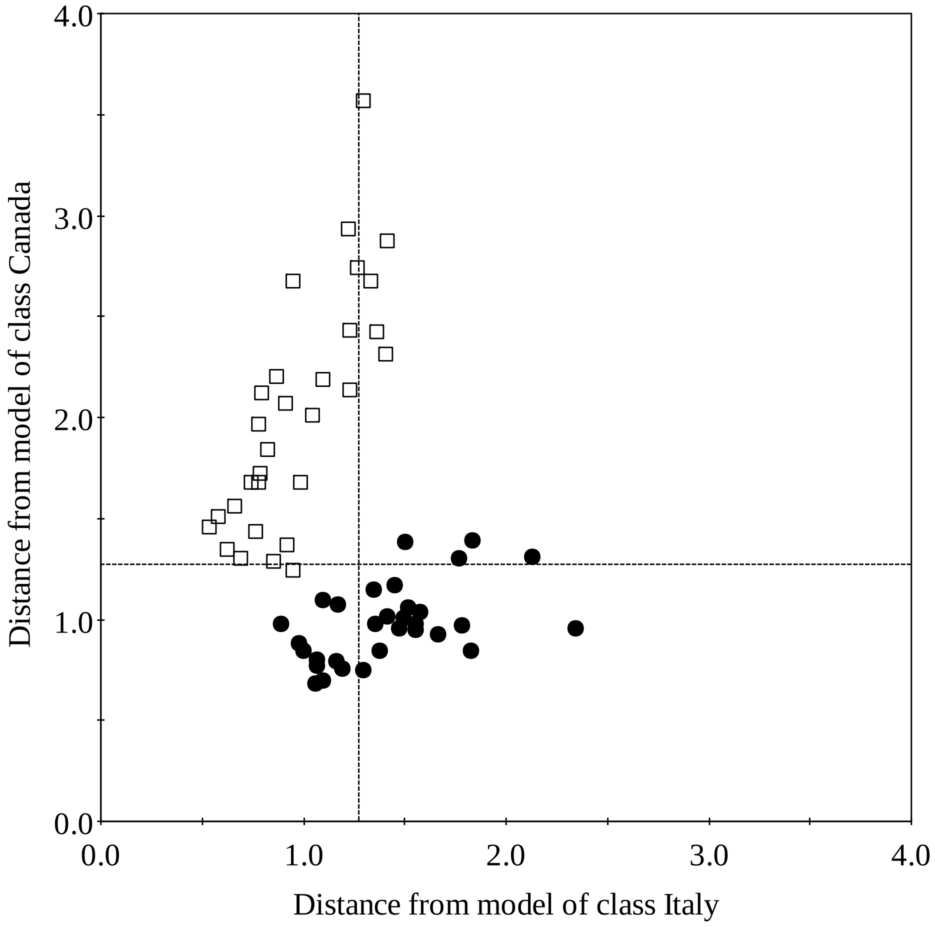
4.0
3.0
2.0
1.0
0.0
0.0
1.0
2.0
3.0
4.0
Distance from model of class Canada
Distance from model of class Italy

Figure 4

Highlights

Geographic origin of lentils was discriminated by ^{1}H NMR fingerprint and chemometrics

^{1}H NMR was used in an untargeted approach

Different supervised methods were tested

External validation procedures were applied on the supervised models

LDA gave 100% classification and test set prediction performances