Analytical Methods

A fast chemometric procedure based on NIR data for authentication of honey with protected geographical indication

C. Herrero Latorre *, R.M. Peña Crecente, S. García Martín, J. Barciela García

Universidad de Santiago de Compostela, Dpto. Química Analítica, Nutrición y Bromatología, Facultad de Ciencias, Alfonso X el Sabio s/n, 27002 Lugo, Spain

Article info
Article history:
Received 24 July 2012
Received in revised form 26 November 2012
Accepted 6 June 2013
Available online 15 June 2013

Keywords:
Honey
NIR
Chemometric procedures
Authentication system

Abstract
In this work, information contained in near infrared (NIR) spectra of honeys with protected geographical indication (PGI) “Mel de Galicia” was processed by means of different chemometric techniques to develop an authentication system for this high quality food product. Honey spectra were obtained in a fast and single way, and they were pretreated by means of standard normal variate transformation in order to remove the influence of particle size, scattering and other factors, and prior to their use as input data. As the first step in chemometric study, display techniques such as principal component analysis and cluster analysis were applied in order to demonstrate that the NIR data contained useful information to develop a pattern recognition classification system to authenticate honeys with PGI. The second step consisted in the application of different pattern recognition techniques (such as D-PLS: Discriminant partial least squares regression; SIMCA: Soft independent modelling of class analogy; KNN: K-nearest neighbours; and MLF-NN: Multilayer feedforward neural networks) to derive diverse models for PGI-honey class with the objective of detecting possible falsification of these high-quality honeys. Amongst all the classification chemometric procedures, SIMCA achieved to be the best PGI-model with 93.3% of sensitivity and 100% of specificity. Therefore, the combination of NIR information data with SIMCA developed a single and fast method in order to differentiate between genuine PGI-Galician honey samples and other commercial honey samples from other origins that, due to their lower price, could be used as substrates for falsification of genuine PGI ones.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

In the last decades, since authenticity is a relevant food quality criterion, fast and single methods to detect adulterations and to ensure the quality and the geographical origin of different commodities are widely demanded by consumers, producers, and regulatory bodies (Lees, 2003; Sun, 2008). For these reasons, European Union has recognised foodstuff produced using raw materials in a particular geographic location and following specific traditional procedures with the following labels: Protected Designation of Origin (PDO), Protected Geographical Indication (PGI) and Traditional Speciality Guaranteed (TSG), according to the case (Council Regulation (ECC) No. 2081/92, 1992; Council Regulation (ECC) No. 2082/92, 1992).

Honey is considered a valuable food product because its high monosaccharide (mainly glucose and fructose) composition provides immediate energy to the organism. It includes other beneficial substances such as proteins, amino acids, organic acids, aroma compounds, vitamins, enzymes, minerals, trace elements and polyphenols. Honey is defined according to the European legislation (Council Directive 2001/110/EC, 2002) as a natural sweet substance produced by Apis mellifera bees from the nectar and other secretions of plants. In 2007, the European Community recognised the PGI label “Mel de Galicia” for the honeys produced in the Galicia region (NW Spain) due to their high quality and specific characteristics (Commission Regulation EC 868/2007, 2007). The price achieved in the food markets for the honey labelled as PGI is higher than other commercial honeys without PGI, from other geographical origins. Different chemometric pattern recognition techniques have been applied for authentication of the geographical or botanical origin of honey based on several types of data. Reviews concerning these methods have been summarised in the works of Anklam (1998) and Wang and Li (2011). The first method employed for recognising honey origin used pollen analysis data (Terrab, Díez, & Heredia, 2003). However, melissopalynological work needs high qualified and trained analysts, it is an expensive and time consuming characterisation technique. Therefore, the utilisation of diverse kinds of chemical variables constitutes an appropriate alternative approach. Chemical markers and components used for honey characterisation included physicochemical parameters (Bogdanov, Ruoff, & Persano Oddo, 2004; Serrano, Villarejo, Espejo, & Jodral, 2004), flavonoids (Escriche,
with the resolution of 2 nm. Each sample has been characterised by its spectrum that was composed of a set of 700 variables. Therefore, the primary data set was a $X_{30 \times 700}$ matrix with two categories: PGI-Galician honey and other non-PGI honey. Diversity of the particle size in the sample analysed by NIR affects scattering and it constitutes an important source of variation in the spectra. This effect has an additive nature and varies from sample to sample. Therefore, raw spectral data need to be processed to avoid this drawback. Standard normal variate (SNV) transformation was the pretreatment used in the case at hand. SNV is often applied to every spectrum individually in NIR data for removing baseline shifts, slope changes and curvilinearity of spectra, i.e., they reduce the influence of particle size, scattering and other influencing factors. In SVN, the average and standard deviation of all the data points for that spectrum was calculated. The transformed spectra were obtained by subtracting the average value from the absorbance for every data point and dividing the result by the standard deviation.

2.3. Chemometric techniques, validation procedure and statistical software

In the present work, two types of chemometric procedures were employed. The first step (data visualisation) consisted of exploring if the NIR data contain useful information to build a classification model to identify PGI-Galician honey samples. Two multivariate display techniques such as principal component analysis (PCA) and cluster analysis (CA) were used for this task. In addition, PCA was also used as data reduction technique in order to reduce the number of variables to be considered in classification techniques. In the second step (classification and validation) other types of mathematical tools, called pattern recognition techniques (PRT), were used to develop a mathematical model to predict whether a given sample belongs to a given class, in this case, the PGI-Galician honey class. For this objective, different discriminant and modelling approaches were developed and compared: Discriminant partial least squares regression (D-PLS); Soft independent modelling of class analogy (SIMCA); K-nearest neighbours (KNN); and multi-layer feed-forward neural network (MLF-NN). Brief descriptions of these procedures with the appropriate references are given in Section 3.

For a model of prediction, it is important to extend the validation to further data, other data than the used for model construction. The ideal situation is when there are enough samples available to create separate (independent) training and test sets. When this is not possible because the number of available samples is low (as is the case at hand), the strategy commonly applied, named cross-validation, consists of dividing the available samples between training and test sets. The first one is used to develop the decision model while the second is employed for validation of the decision rule obtained. As the decision rule and the validation obtained depend on the objects in each set, this process must be repeated the appropriate number of times (with different constitutions of both sets) to guarantee that all samples, at least once, belong to test set. When only one sample at a time is selected as the training set, the validation strategy was named Leave one out cross-validation (LOOCV) method, and the procedure must be repeated a number of times equal to the number of samples (30 times in the present case). A single sample was considered as the validation set, and the remaining observations as the training data. This arrangement was repeated 30 times so that each sample was used once to validate the model (Vandeginste et al., 1998).

The statistical packages employed in this work were the following: PCA and CA were performed with the use of Statgraphics Centurion (Statistical Graphics Corporation, Rockville, MD, USA). SIMCA and KNN were carried out using V-Parvus (University of...
Genoa, Genoa, Italy), D-PLS was performed by Unscrambler (CAMO Software AS, Oslo, Norway), and the artificial neural network computations were done using MATLAB (Mathworks Inc., Natick, MA, USA).

2.4. Model evaluation criteria

The success in classification and the prediction for the PGI-honey classification models developed using the different pattern recognition techniques were evaluated according to their sensitivity and specificity. Sensitivity and specificity are adequate statistical measurements for the performance of binary classification systems. Sensitivity of a model for a class A is the proportion of genuine A-samples that the model recognises as belonging to this class. Specificity of a model for a class A is the proportion of non-A-samples that the model is refused as non-belonging to this model. It is clear that these two measures are related to type I and type II errors, and they can be denoted in terms of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Sensitivity is TP/(TP + FN) and specificity is calculated as TN/(TN + FP). A PGI-honey model with 100% sensitivity and 100% specificity will be a model capable to accept all the genuine PGI-samples and to refuse all the non-PGI-samples.

3. Results and discussion

The use of information contained in NIR data for developing authentication systems in combination with chemometric techniques present advantages compared to other approaches based on chemical information from different honey compounds: (i) minimal sample handling (minimising the risk of contamination and avoiding the problems of sample viscosity); (ii) fast and simple measurement eluding the use of high expensive equipment; and (iii) unnecessary quantification step avoiding preparation of calibration curves. In Fig. 1, the NIR spectra of honey samples obtained as described in Section 2.2 have been presented. Honey spectra present four intense absorption bands at 1480, 1580, 1935 and 2100 nm which are related with the first O–H overtone, the first overtone of the O–H stretching, the combination of O–H stretching and deformation, as well as the first overtone of O–H deformation and C–O stretching band, respectively. Other minor absorption bands at 1202 and 2321 nm are related with C–H bonds. It is evident that the spectrum profiles are very similar for the honey samples of the two categories: there is no wavelength or wavelength-group in the spectra that could differentiate between PGI and non-PGI classes. Therefore, a multivariate approach based on the use of all wavelengths simultaneously must be explored.

3.1. Data display techniques (CA and PCA)

In this work, the information contained in the honey spectra was revealed by means of two display chemometric techniques: CA and PCA, which are often used together to study the structure of a data set. CA is an unsupervised display procedure that allows the mapping of the 700-multidimensional variable space in a two-dimensional tree called dendrogram (Kaufman & Rousseeuw, 2008). Dendrogram was constructed from the similarities between samples measured as the squared Euclidean distance between them, whereas the Ward hierarchical agglomerative method was used to establish the clusters. In Fig. 2a, the dendrogram obtained from the NIR spectra of the 30 samples has been shown. As it can be seen, at a similarity level of 0.75 three clusters were identified. The majority of non-PGI-honey samples (coded as I) performed a cluster named as C1 that includes two sub-clusters with 6 and 7 samples, respectively. Two samples of this class appeared in a separate cluster. The samples of PGI-class (coded as A) are included in two clusters C2 and C3. Therefore, it is clear that PGI and non-PGI samples are located in separate areas in the multidimensional space of the variables. Principal component analysis is a chemometric technique that was used (complementary with CA) in order to achieve a primary data structure evaluation, as well as to produce a reduction of the data dimensionality with the minimum loss of useful information (Jolliffe, 2002). PCA extracts orthogonal axes according to the successive directions of maximum variance of data; the principal components (which are linear combinations of original variables, in this case the 700 wavelength data for each spectrum). PCA was applied to the X matrix of spectra, transformed according to SNV pretreatment, with the aim of studying the data structure in a reduced dimension retaining the maximum amount of variability. From the score-plot in Fig. 2b (which attains 97% of total variance), it can be observed as a natural separation between
PGI-samples (coded as A) and non-PGI-samples (coded as I) in the space defined by the principal components. As indicated in the previous CA, the two samples of non-PGI-class that are located in a different area of the multidimensional space are also revealed by PCA. As a conclusion from the two display techniques applied, it was proved that honey samples for both categories have different characteristics and they occupy different areas in the 700-multidimensional space. Thus, NIR spectra seem to contain useful information to perform classification models in order to assign an unknown honey sample to the class to which it belongs to.

Each of the NIR spectra used in this study is a set of 700 variables (wavelengths). However, it has been demonstrated that the use of such very-high dimensional data as input information for pattern recognition techniques is not a good strategy due to: (i) Certain PRT would be unstable solutions when the number of variables is extremely high compared to the number of samples; (ii) Another problem is the co-linearity of NIR data, correlation could cause problems and slowing in the training step; and (iii) The use of a set with a great amount of variables is extremely expensive in terms of storage memory and computing time. Therefore, it was necessary to reduce the dimension of the NIR data because the full spectra contain irrelevant variables for PRT. The estimation capability of the models derived will be improved by removing the spectral regions which do not contain useful information. In this case, the applied approach used to achieve this task was to use mathematical techniques for data compression such as PCA (Zhu et al., 2010). Once PCA was carried out, the scores of the samples in a selected number of component principals (retaining the maximum amount of variance) were used as input data for PRT instead of the complete spectra. As a consequence of this strategy, a drastic variable reduction has been achieved without loss of valuable information. From the evaluation of eigenvalues of the principal components (which represent the percentage of variance retained for the principal components), it was established that the first ten principal components represented 99.9% of the total data variance (See Fig. 3). Therefore, a new data matrix \( X'_{300-10} \) Containing the scores of the thirty honey samples on the first ten principal components was considered to be the input information for the PRT employed in the development of classification models to be performed in the next section.

### 3.2. Pattern recognition techniques

Different PRTs, such as D-PLS, SIMCA, KNN and ANN, were considered and compared to develop classification models in order to authenticate genuine PGI-Galician honey samples. In all cases the \( X'_{300-10} \) matrix described above was used as input data. The validation of the constructed models was performed, due to the limited number of samples available, by the use of training and test sets obtained using LOOCV procedure. The performance of each model derived using the diverse PRT was assessed taking into account its sensitivity and specificity for prediction in the validation samples.

#### 3.2.1. Discriminant partial least squares

D-PLS was utilised to express the relationship amongst a response binary class matrix \( Y_{301} \) (defined as “1” for PGI-samples and “0” for non-PGI-samples) and the set of predictor variables \( X'_{300-10} \) by means of a set of latent factors \( T_{30-1} \) which maximise the covariance between response and predictor variables (Stahle & Wold, 1987). Best models can be achieved when classes present similar or equal number of samples. The discrimination rule derived was validated using a leave one out cross-validation procedure as described in Section 2.3. The optimal number of factors \( f \) was selected on the basis of the minimum cross-validation residual validation variance. One sample was considered as correctly classified when the D-PLS predicted value was in the interval defined by the target ± 0.2. In Fig. 4, the D-PLS prediction values for honey samples in the validation set using a four-factor model have been presented. As it can be seen, even though the prediction values achieved were consistent with the sample class, one sample of non-PGI class and three samples of PGI class were misclassified. The performance for the model of the PGI-class was based on its sensitivity and specificity which are listed in Table 1. The PGI-class model performed by D-PLS achieves 80.0% of sensitivity and 93.3% of specificity. Therefore, this classification rule could be appropriate to detect falsification of genuine samples; because of its high specificity. In practice, a very high percentage of non-PGI samples could be detected and classified as non-PGI. However, there is a considerable risk of rejecting a genuine PGI-sample as false (20%).
3.2.2. Soft independent modelling of class analogy

SIMCA is a modelling pattern recognition technique that consists of the construction of a confidence box based on disjointed principal components around each class (Wold, 1976). The samples to be classified are projected in the principal component spaces defined in the previous step, and the unknown samples are placed in the class in which they match best. In the case at hand, the predictor variables $X^{30-10}$ were autoscaled separately for each class (the usual scaling procedure for SIMCA). The characteristics of the SIMCA model constructed have been summarised in Table 1. The PGI-class model obtained by SIMCA presents better results than those obtained in D-PLS. As it can be seen in Table 1, the sensitivity augmented up to 93.3% and the specificity reached 100%. This means that SIMCA developed a model for the PGI-class in which all samples from non-PGI origin were rejected by this model. In Fig. 5, the Coomans-plot for SIMCA was shown, the model for PGI-class at 95% (represented by the rectangle in the left side of the plot) do not include any non-PGI sample. However, there is a possibility of 6.7% of classifying a true PGI-sample as false.

3.2.3. K-nearest neighbours

KNN is a supervised classification procedure based on a local metric related to the similarity between objects. This PRT operates by classifying an unknown object in the category to which the $K$ nearest known objects of the training set belongs. Similarity between samples was calculated on the basis of the Euclidean distance between them and the importance of a given neighbour in the class assignation was weighted on the basis of the inverse squared distance (Coomans & Massart, 1982). The selection of the number of neighbours, $K$, was performed in a heuristic way, by assaying the correct classification rate achieved considering all samples as the training set. The best results were obtained for $K = 3$. This number of neighbours was employed for further calculations in which the validation of the model constructed was carried out in the 30-step LOOCV previously commented. As Table 1 shows, KNN achieved a PGI-class model with high sensitivity, but low specificity. This model is not appropriate for detection of PGI honey falsification because, in spite of all PGI-samples are classified as true; however there is an elevated risk (14.4%) of accepting a non-PGI sample as genuine.

| Table 1 |
|---|---|---|
| **PRT** | Characteristics of PRT used | Sensitivity(%) | Specificity(%) |
| D-PLS | Model built on the basis of four PCs | 80.0 | 93.3 |
| SIMCA | - SIMCA normal range model based on four PCs explaining (85.04% total variance) | 93.3 | 100.0 |
| | - Unweighted augmented SIMCA distance- Boundary for each class in the PC-space was determined according to 95% confidence limit | | |
| KNN | - $K = 3$ | 100.0 | 86.6 |
| | - Calculation of the distance between samples was carried out as the Euclidean distance | | |
| | - Contribution of each neighbor in the class assignation was weighted on the basis of the inverse squared Euclidean distance | | |
| MLF-NN | - Three layers (10–7–1) network was used | 100.0 | 93.3 |
| | - Sigmoidal transfer function used:$f(x)=1/(1+\exp(-x))$ | | |
| | - Initial values of the weights were selected randomly in the range from –3 to 3 | | |
| | - Maximum number of epochs was 2,000 | | |
| | - Initial $\eta$ (adaptive learning rate) was 0.2 | | |
| | - Momentum $i$ was 0.5 | | |
3.2.4. Multi-layer feedforward neural network

MLF-NN is a powerful pattern recognition technique that builds a model based on a set of input/output (score variables/category PGI and non-PGI) objects updating the weights of connections between neurons to obtain an appropriate output for each input (Zupan & Gasteiger, 1999). The weights contain information (which it is not chemically interpretable) about the relationship between the ensemble of inputs and the output. In this paper, a MLF-NN was used to predict the category of honeys on the basis of the $X_{30,10}^o$ data. For the selection of the adequate network structure, different network architectures with three and four layers were assayed using the complete data set. In all cases the input layer was composed of ten neurons (equal to input variables), and the output layer was formed by one neuron in which the target output was coded as “1” for the PGI-class and “0” for the non-PGI class, considering also a confidence interval (±0.2). For the hidden layers, different numbers of neurons between 4 and 8 were assayed. The lower root mean square error was achieved using a neural network with three layers (10–7–1) that was the one used for further calculations. The other characteristics of the MLF-NN used have been summarised in Table 1. The described MLF-NN was utilised for the modelisation of the proposed problem, using the thirty-step LOOCV procedure described earlier. The classification rule obtained (see Table 1) gave similar results than KNN. Complete sensitivity allowed the correct classification of every Galician sample as PGI, but also, from the specificity it is clear that there is certain possibility of assigning false honeys as genuine.

3.3. Comparison amongst different PRTs

According to the data shown in Table 1, the models created for PGI-class on the basis of KN and MLF-NN achieved a total sensitivity, but a lower specificity; while D-PLS and SIMCA produced inverse results, with complete (or very high) specificity and lower values for sensitivity. This fact could be explained taking into account the different nature of the diverse PRTs applied. KN and MLF-NN are discriminant techniques. Therefore, they perform a mathematical rule that define a complex boundary between classes, and this means that in these techniques the emphasis was placed on the difference between classes. D-PLS is a regression PRT and SIMCA is a modelling procedure. Thus, in these cases the classification rule was developed with an emphasis on the similarity within the class. The former techniques could be adequate for assuring the genuineness of true samples, while the last ones could be more useful for detecting falsifications of true samples.

4. Conclusion

As a conclusion, it can be indicated that honey NIR spectra contains useful information that can be processed in combination with PRTs to develop PGI-honey authentication procedures. A clear advantage in contrast to other approaches based on chemical data is the simplicity of obtaining the relevant NIR information by a non-destructive, single, fast and easy measurement without any pretreatment step of honey. NIR spectra can be obtained directly from honey samples and they were directly stored for their use in chemometric developments. Also, it was demonstrated that the selection of the PRT applied to the development of classification rule is critical. In the case at hand, SIMCA modelling technique was considered to be the most useful since it developed a PGI-model in which all non-PGI samples are identified as non-genuine ones. Thus, the combination of the NIR information data with the SIMCA constructed model developed a single and fast procedure for authenticate genuine PGI-Galician honeys. This approach supplies an easy-to-use tool to detect commercial honey samples from other origins that, due to their lower price, could be used as substrates for falsification of genuine PGI Galician honeys. Therefore, the proposed method avoids the use of organoleptic property measurements for PGI-honey authentication, and it has demonstrated to be a useful tool in order to guarantee the quality and to preserve the image of the product, to detect frauds (which is an important issue for producers) and to prevent the consumers from overpayment.

Acknowledgements

The authors are grateful to the Directive Council of PGI “Mel de Galicia” for providing Galician guaranteed honey samples. Financial support by Deputación Provincial de Lugo (Contract APMPL) is gratefully acknowledged.

References
