Comprehensive
REVIEWS
in Food Science and Food Safety

Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing

Daniel Granato, Predrag Putnik, Danijela Bursać Kovačević, Jânio Sousa Santos, Verônica Calado, Ramon Silva Rocha, Adriano Gomes Da Cruz, Basil Jarvis, Oxana Ye Rodionova, and Alexey Pomerantsev

Abstract: In the last decade, the use of multivariate statistical techniques developed for analytical chemistry has been adopted widely in food science and technology. Usually, chemometrics is applied when there is a large and complex dataset, in terms of sample numbers, types, and responses. The results are used for authentication of geographical origin, farming systems, or even to trace adulteration of high value-added commodities. In this article, we provide an extensive practical and pragmatic overview on the use of the main chemometrics tools in food science studies, focusing on the effects of process variables on chemical composition and on the authentication of foods based on chemical markers. Pattern recognition methods, such as principal component analysis and cluster analysis, have been used to associate the level of bioactive components with *in vitro* functional properties, although supervised multivariate statistical methods have been used for authentication purposes. Overall, chemometrics is a useful aid when extensive, multiple, and complex real-life problems need to be addressed in a multifactorial and holistic context. Undoubtedly, chemometrics should be used by governmental bodies and industries that need to monitor the quality of foods, raw materials, and processes when high-dimensional data are available. We have focused on practical examples and listed the pros and cons of the most used chemometric tools to help the user choose the most appropriate statistical approach for analysis of complex and multivariate data.

Keywords: classification, food authentication, multivariate statistical techniques, one-class classifiers, pattern recognition

Introduction

The term "chemometrics" describes the statistical and mathematical approaches used to optimize the design of experiments and extract useful information from large and complex datasets (Varmuza & Filzmoser, 2009). Chemical data commonly include values and properties of various compounds determined by laboratory experiments and having numerous sources of variance. Accordingly, statistical analysis of such data should employ one or more multivariate statistical tools (Varmuza & Filzmoser, 2009). Multivariate statistics encompasses the simultaneous analysis of one or more dependent (outcome) variables against two or more independent (input) variables (Hidalgo & Goodman, 2013); in many circumstances, the procedures can compare a large number of responses (dependent variables) against a plethora of independent variables (predictors). The most common types of multivariate tests include multivariate analysis of variance (MANOVA), various forms of factor analysis (such as principal components analysis, PCA), mathematical modeling approaches, artificial neural networks (ANN), discriminant analysis, and many others (Dziurkowska & Wesolowski, 2015). Simultaneous comparison of all independent variables in a single test requires multiple analyses of each independent variable against outcome measures that inflate type I errors (Dumancas, Ramasahayam, Bello, Hughes, & Kramer, 2015). Consequently, *P*-value estimates are affected by compound relationships, where probability is dependent on $[1 - (0.95)^n]$ where "*n*" is the number of single comparisons (Rutherford, 2011).

Experimental data in food science and other areas can be either qualitative or quantitative (Szymańska et al., 2015). Qualitative data are of three types: nominal (such as three types of foods, five types of process, and so on); dichotomous (such as male/female, authentic/adulterated); and ordinal (data ordered by criteria, for example, three levels of sensory evaluation, such as 1 = *unacceptable*, 2 = *marginally acceptable*, or 3 = *acceptable*). However, quantitative variables include continuous scales (temperature, pressure, time, concentration, mass, and so on), intervals, and ratios (Larson-Hall, 2010). Regardless of the type, all variables can be analyzed by chemometric approaches (Szymańska et al., 2015). The most

CRF3-2017-0235 Submitted 11/29/2017, Accepted 1/26/2018. Authors Granato and Santos are with Dept. of Food Engineering, State Univ. of Ponta Grossa, Av. Carlos Cavalcanti, 4748, 84030–900, Ponta Grossa, Brazil. Authors Putnik and Kovačević are with Faculty of Food Technology and Biotechnology, Univ. of Zagreb, Pierottijeva 6, 10000, Zagreb, Croatia. Author Calado is with School of Chemistry, Federal Univ. of Rio de Janeiro, Rio de Janeiro, Brazil. Authors Rocha and Cruz are with Dept. de Alimentos, Inst. Federal de Educação, Ciência e Tecnologia (IFRJ), 20270-021, Rio de Janeiro, Brazil. Author Jarvis is with Dept. of Food and Nutrition Sciences, School of Chemistry, Food and Pharmacy, The Univ. of Reading, Whiteknights, Reading, Berkshire RG6 6AP, U.K. Authors Rodionova and Pomerantsev are with Semenov Inst. of Chemical Physics RAS, Kosygin str. 4, 119991, Moscow, Russia. Direct inquiries to author Granato (E-mail: granatod@gmail.com; dgranato@uepg.br).

doi: 10.1111/1541-4337.12341

important statistical techniques for chemometrics are PCA and partial least squares (PLS) analysis (Varmuza & Filzmoser, 2009). These methods commonly require pretreatment of the data (preprocessing; Skov, Honoré, Jensen, Næs, & Engelsen, 2014), such as "normalization" and scaling, to remove systematic bias from the datasets, but with minimal influence on the quality of information (Nunes, Alvarenga, de Souza Sant'Ana, Santos, & Granato, 2015).

Foods are complex materials (Lucci, Saurina, & Núñez, 2017) that are commonly studied by engineers, technologists, chemists, physicists, microbiologists, and many other professionals (Munck, Nørgaard, Engelsen, Bro, & Andersson, 1998). Practical applications of chemometrics in food science include, but not limited to the authenticity, functionality, bioactivity, and food safety (Granato, Santos, Escher, Ferreira, & Maggio, 2018; Nascimento et al., 2018; Skov et al., 2014).

Many major problems currently faced by governmental agencies and industries are related to adulteration and food frauds. Concern is driven by public interest in food quality and safety (Danezis, Tsagkaris, Camin, Brusic, & Georgiou, 2016). The use of chemometrics with appropriate analytical techniques can identify adulteration of wines (Alañón, Pérez-Coello, & Marina, 2015), honey (Wu et al., 2017), essential oils (Do, Hadji-Minaglou, Antoniotti, & Fernandez, 2015), and many other high-value products. Aside from food adulterations, chemometrics can be used to analyze data on soil toxicity (Peng et al., 2016), influences of climate on the nutritional value of foods (Obranović et al., 2015), and changes in functional properties as a consequence of processing (Bursać Kovačević et al., 2016; Herceg et al., 2016), just to mention few practical examples that will be discussed later.

The objective of this updated review is to give a holistic insight of the advantages and disadvantages of multivariate statistical methods intended for analyzing complex datasets. There are numerous research studies and critical reviews covering various aspects of the application of chemometric tools to food chemistry problems (Danezis et al., 2016; Munck et al., 1998; Nunes et al., 2015; Skov et al., 2014). Due to the extent and multidisciplinary nature of available data, all aspects of chemometrics cannot be covered in a single paper. Therefore, the focus of this paper is the use of chemometric techniques for resolving issues of food authentication and food microbiology, and assessment of the effects of food processing. Moreover, the focus is on pattern recognition methods, providing exploratory, and classification procedures. Examples of practical applications have been selected to illustrate the benefits and shortcomings of these methods.

Chemometrics in Food-Related Disciplines

The concepts of chemometrics can be applied to data from chemical analyses that are used widely in the fundamental study of foods (Varmuza & Filzmoser, 2009). Chemometrics is useful also to bridge the gaps in multidisciplinary data needed for solid scientific conclusions, and to produce and add knowledge in food science (Munck et al., 1998). Many applications of chemometrics, coupled with both conventional and innovative analytical measurements, have been proposed and applied in the last decade to solve technological and legislative food control problems.

Main chemometric tools in food disciplines

Chemometric methods are used to separate useful information from noise, uncover hidden correlations, and provide a visual approach for multivariate data analysis. Overall, there are three general chemometric approaches: explorative analysis, classification, and calibration. These approaches are used for data analysis in food chemistry and other disciplines within food science and technology. The choice of the approach depends on the problem and on the type of experimental data.

Exploratory analysis. Principal component analysis helps to reveal the hidden structure of and to compress multivariate datasets (Wold, Esbensen, & Geladi, 1987). The Kaiser–Meyer–Olkin (KMO) and Bartlett's test of sphericity are used to test whether data are suited for PCA or other types of factor analysis (Tabachnick & Fidell, 2007). The KMO measure of sampling adequacy is a statistic that indicates the proportion of variance in your variables that might be caused by underlying factors. High values (close to 1.0) generally indicate that a factor analysis may be useful with your data. If the value is less than 0.50, the results of the factor analysis probably will not be very useful (Granato et al., 2018). Bartlett's test of sphericity tests the hypothesis that your correlation matrix is an identity matrix, which would indicate that your variables are unrelated and, therefore, unsuitable for structure detection. Small values of the significance level (less than 0.05) indicate that a factor analysis may be useful with your data.

Using PCA, a set of correlated variables is transformed into a set of uncorrelated principal components (PCs). The (I × J) data matrix $\mathbf{X}$ is decomposed by Equation (1):

$$\mathbf{X} = \mathbf{TP}^{t} + \mathbf{E} \tag{1}$$

where $\boldsymbol{T} = [t_{ia}]$ is the (I × A) scores matrix; $\boldsymbol{P} = [p_{ja}]$ is the (J × A) loadings matrix; $\boldsymbol{E} = [e_{ij}]$ is the (I × J) matrix of residuals; and A is the number of PCs.

The new coordinates are based on 18 analysis of the covariance matrix, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, which produces A pairs of eigenvalues and eigenvectors. PCs are extracted in such a way that the maximum amount of the data variance is associated with the 1st PC, and then progressively lesser variance are associated with each subsequent component. The important parameter A is the number of PCs. For meaningful results of PCA, it is crucial to report factor loadings, factor 18 values, and the amount of explained variance. The more components extracted, the better is the approximation of data matrix $\mathbf{X}$. For further analysis, $\mathbf{T}$ matrix represents objects in a new reduced space and the $\mathbf{P}$-matrix shows how well a variable is taken into account by model components.

The dominant patterns present within samples and variables are illustrated by plotting the columns of the score matrix and loading matrix, respectively. PCA is used in numerous applications as a 1st step, which helps to describe the data patterns. For instance PCA has been used to highlight differences between technological processes in production of olive oils (De Luca, Restuccia, Clodoveo, Puoci, & Ragno, 2016), to assess within-sample variation; and to compare processes with the variation observed among samples in the analysis of distillers' dried grains (Tena, Boix, & von Holst, 2015), and in many other studies.

Another chemometric tool that is used for explorative analysis is cluster analysis. Among the multiple clustering methods, the nonhierarchical methods, such as *k*-means and *k*-medians, and hierarchical cluster analysis (HCA) are the most frequently used methods. For instance, the HCA method was used for analysis of the ATR–FTIR spectra of gelatins of different origins (Cebi, Durak, Toker, Sagdic, & Arici, 2016), and for revealing clusters of the FTIR spectra of refrigerated and frozen/thawed chicken meat samples (Grunert, Stephan, Ehling-Schulz, & Johler, 2016). Use of clustering methods together with PCA have been widely described (Nunes et al., 2015; Zielinski et al., 2014). All these methods are often called the unsupervised multivariate methods

Table 1–Example of a confusion matrix that may be built using a supervised statistical technique.

Classified as → Class ↓	Brazil	China	Poland	USA	Mexico	Correct classification (%)
Brazil	**9**	2	0	1	1	69
China	0	**10**	0	0	4	71
Poland	0	0	**9**	0	0	100
USA	4	4	4	**10**	0	45
Mexico	0	0	0	2	**12**	86

(Pomerantsev, 2014) and are used in all fields of food science and technology. Other examples of these chemometrics tools are explained in the following sections.

Classification methods. Two types of classification techniques are distinguished regardless of the statistical background of a specific method. The first type of technique is used to assess to which of various predefined classes the object belongs. Such methods are referred to as discriminant analyses (DA) and are similar to logistic regression. The second type of method is referred to as "one class classifiers" - OCC (Tax & Duin, 1998), or class modeling methods (Derde & Massart, 1988). They confirm whether, or not, an object can be associated with a targeted class of interest. For example, using this method one can answer the questions: "does this olive oil come from Italy or Spain?" or "is this milk from organic or conventional sources?."

In these methods, a confusion matrix is generated from the classification model that allows the visualization of actual and predicted classification. In summary, confusion matrices are built when classification models are proposed in order to predict the number of correctly classified samples. For instance, if one aims to differentiate organic from conventional milk based on some analytical measurements, the confusion matrix will tell how many organic milks were correctly recognized as being from this class (Gondim, Junqueira, Souza, Ruisánchez, & Callao, 2017). A typical confusion matrix is shown in Table 1. Using the data shown in this table, one can easily see that all Polish samples were correctly identified, whereas only 45% of the USA samples were correctly classified.

Traditional statistical terms used for presenting the results of classification are the type I error (α), which is the rate of incorrect rejections of class membership; and the type II error (β), which is the rate of wrong acceptance of alien objects as members of a predefined class. The terms "sensitivity" and "specificity" are also used for presenting the results of classification. Class sensitivity is defined for each class as the percentage of samples of this class that are correctly recognized as the members of the class. It can also be defined as the rate of true positives, so it is complementary to the type I error (α). Class specificity is defined for each target class k as the percentage of samples from other classes (not k), which are correctly attributed as not belonging to the target class. This percentage value is complementary to the rate of false positives, and is given by 100 $(1 - \beta)$. It is important to note that any effort to reduce one type of error results in an increase in the other type of error. In order to reduce both simultaneously, a large number of samples must be used (Berrueta, Alonso-Salces, & Héberger, 2007).

When DA methods are applied, we must initially have a definitive list of classes; for example, discrimination between dry and wet processing methods for postharvest coffee (Hamdouche et al., 2016), discrimination of extra virgin olive oils from whole and stoned olive pastes (De Luca et al., 2016), discrimination between bovine, porcine, and fish gelatins (Cebi et al., 2016), rapid differentiation of fresh and frozen/thawed chicken (Grunert et al., 2016), evaluation of the provenance of honeys (Nascimento et al., 2018), among others. For discrimination purposes, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are usually employed because they are reliable, simple to understand, and the graphical outputs are appealing (Pomerantsev, 2014). Their main disadvantages are that (1) they do not work when the covariance matrix is singular; (2) they assume that data are normally distributed; and (3) they do not allow interchange of α and β errors.

LDA. In LDA, separations between classes are hyperplanes and the allocation of a given sample within one of the classes (that is, geographical origin, type of product, and cultivation system) is based on a maximum likelihood discriminant rule. Overall, the results of LDA classification can be graphically visualized by projecting the classes (preferably three or more) into the space of canonical variates, or discriminant functions (D'Archivio, Giannitto, Maggi, & Ruggieri, 2016). The discrimination model is calculated using all samples, which implies the model cannot be easily validated using external samples. However, some commercial packages (that is, MatLab) have the option to perform the model validation by using the leave-more-out cross-validation approach. When an LDA model is built and validated, the predictive capability of LDA discrimination data may be assessed. The method provides outputs that are easy to understand (confusion matrix and graphical representations) but it has the tendency to be over fitted if the number of samples in each class is restricted. Overfitting is characterized by high accuracy for a classifier when evaluated on a training set but low accuracy when evaluated on a separate test set in high-dimensional and low sample size settings (Subramanian & Simon, 2013). LDA should not be used if the design is not balanced (that is, if the number of samples in various classes is very different). In addition, the validation of the LDA classification output is not straightforward in many statistical packages, which causes the user not to undertake the validation procedure. The use of LDA, or other classification methods, requires that sample size within classes must be high enough to enable proper classification of samples. For classification purposes, if there is a limited number of samples available, multiple comparisons between groups using inferential analysis would be of interest.

A typical graphical output from LDA is shown in Figure 1. Imagine that the discrimination between GC-MS data of the lipids in various commercial milk samples produced in different locations is sought. From the projection of samples displayed in Figure 1, it is possible to observe a clear separation between samples from distinct origins and possible overlaps in a sample that was labeled as "from China" (included in the Brazilian cluster) and one sample labeled as "from Germany" (included in the Brazilian cluster). This bi-plot reflects the reciprocal location of milk samples in the canonical space such that possible adulterations (that is, food frauds) based on complex and large datasets can be easily assessed. Therefore, it is no wonder that LDA is the most widely used method of discrimination in food chemistry problems.

However, the main disadvantages of LDA are: it does not work when the covariance matrix is singular, for example, for large numbers of variables; it requires standardization, for example, by PCA; it is not suitable if the covariance matrices of the classes are different; it implicitly assumes normality of data; and it does not allow interchange of α and β errors.

A real example of the application of LDA in food chemistry is the geographical classification of 144 saffron (*Crocus sativus*) samples produced in some Italian regions (D'Archivio et al., 2016).

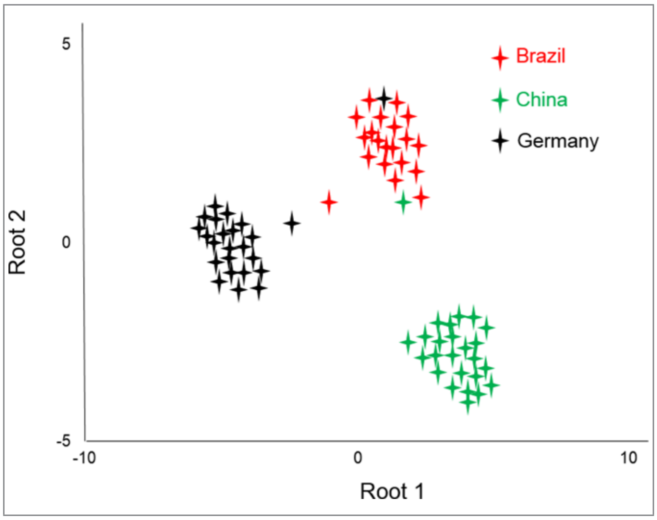

Figure 1–Linear discriminant analysis of commercial milk samples produced in different countries based on GC-MS data.

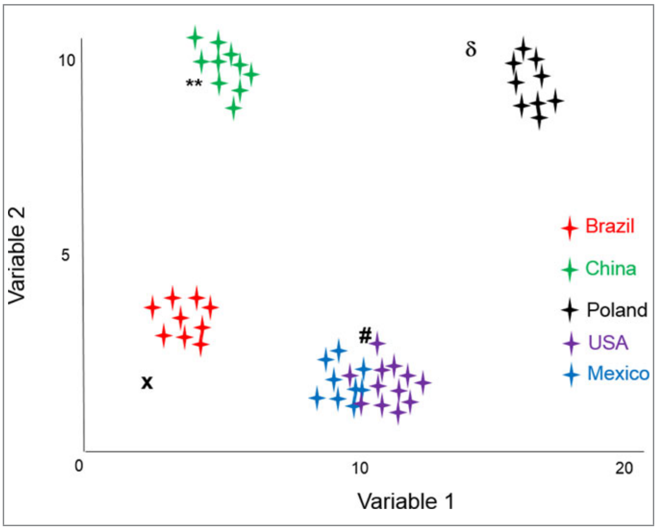

Figure 2–Classification of honey samples from different geographical origins using *k*-nearest neighbors (*k*-NN). Unknown honey samples are labeled as x, $^{\#}$, **, and. $^{\delta}$

Authors used LDA to classify the levels of crocin, safranal, picrocrocin, and their derivatives and flavonoids determined by HPLC. Use of the confusion matrix enabled correct classification of 88% of unknown saffron samples. The crocins trans-crocetin bis(b-D-glucosyl) ester, and *cis*-crocetin bis(b-D-glucosyl) ester, were the key elements that determine the geographical classification of Italian saffron. This work shows the benefit of LDA in identifying unique markers of high-value added herbs produced in different locations.

K-nearest neighbors. A simple method that does not use any type of distribution assumption and may be used for a small number of samples is the *k*-NN or *k*-nearest neighbor algorithm (Beebe, Pell, & Seasholtz, 1998). *k*-NN can be used for classification of categorical data variables and regression for continuous variables. Let X be a training set divided into class subsets, and let X be a new unknown object, which should be classified. First, calculate the distances (usually Euclidean) from x to all samples of the training set. Then select k nearest neighbors, which are located at minimal distances. The new object x belongs to a class which encompasses most of the k empirically chosen neighbors. An increase in the value of k reduces the impact of errors, whereas its decrease worsens classification (Pérez-Caballero et al., 2017; Reinholds, Bartkevics, Silvis, van Ruth, & Esslinger, 2015). This value is related to the number of neighbors (“votes”) to poll for future classifications, such that $k = 1$ provides a good classification rule. The more samples that agree within a particular classification, the more confidence can be placed in the classification data (Beebe et al., 1998). For example, *k*-NN has been applied for examining the trace and rare earth elemental fingerprint variations of “Fava Santorinis” over several harvesting years (Drivelos, Danezis, Haroutounian, & Georgiou, 2016).

An example of classification using the *k*-NN algorithm is shown in Figure 2 (example modified from Beebe et al., 1998), where two responses (moisture content and hydroxymethylfurfural level) were used to distinguish honey samples produced in different geographical origins. Several training samples are shown for each class and four unknown honey samples (x, #, **, and δ). Classification of the unknown samples using the one-nearest neighbors’ rule, shows that sample “**” would be from China, sample “δ” would be from Poland, and sample “x” would be from Brazil. However, the classification of sample “#” is less obvious as it is overlapped with honeys from the USA and Mexico. In this specific case, the nearest neighbor is from the USA, whereas the second nearest one is from Mexico. The analyst may use a different statistical method to try to differentiate honeys from these two countries or may use different analytical measurements to obtain a clearer differentiation. As *k*-NN is not a “soft” classification technique, that is, an object is always classified in one of the classes studied, most statistical software packages classify an unknown sample (for example, sample #) into the class containing most of the nearest neighbors (that is, USA with $n = 12$ objects). *k*-NN is not accurate when there are many features, as dimensionality eliminates the differences between samples; and results depend on the choice of metric and the number of neighbors.

k-NN was used to classify results of a study on sugar-based adulteration of honey (Soares, Amaral, Oliveira, & Mafra, 2017). They used a range of analytical techniques (TLC, C-isotope, HPAEC, GC, HPLC, IR, NMR, Raman spectroscopy, and Q-TOF-MS) to evaluate honey adulterants. A recent study used *k*-NN to investigate the influence of pesticides on the genome of honeybees, solitary bees, and bumblebees (Como et al., 2017). The authors developed their own software using two data sets to train and validate the analysis. They concluded that *k*-NN methodology enabled prediction of the toxicity of many different pesticides, which enabled their algorithm to be used to predict relevant patterns for other pesticides.

Partial least squares—discriminant analysis. Partial least squares—discriminant analysis (PLS-DA) is the most popular and efficient among the methods that can deal with a singular covariance matrix (Ståhle & Wold, 1987). It is a conventional PLS regression method where the fingerprints (I × J) matrix $\boldsymbol{X}$ is considered as a predictor matrix, and a specially constructed (I × K) response matrix $\boldsymbol{Y}$ comprises categorical (“dummy”) variables that describe class memberships for K classes (Ståhle & Wold, 1987). When PLS regression is used, the response value $\hat{\mathbf{Y}}$ is predicted for a new sample. The decision is based on the comparison of $\hat{\mathbf{Y}}$ with the given categorical variables in $\boldsymbol{Y}$. The sample is attributed to the class for which the “dummy” variable $\mathbf{Y}$ is closest to $\hat{\mathbf{Y}}$. In many of the reported applications (Cebi et al., 2016; De Luca et al., 2016;

Grunert et al., 2016; Hamdouche et al., 2016; Tena et al., 2015) PLS-DA was employed for making a final decision. One advantage of using PLS-DA is that one can understand which variables carry the class separating information, but a clear disadvantage of PLS-DA is that it is highly influenced by the number of classes and sample size distribution per class. This fact makes the use of PLS-DA appropriate only when there are few classes each with many samples to be differentiated. Other disadvantages of PLS-DA are: it requires preliminary regression analysis, which is sensitive to outliers. The result depends on the number of principal components in the PLS2 regression and PLS-DA does not work for small number training sample sets.

Soft independent modeling of class analogies. The procedure for OCC specifies the target class according to the properties of its representative members. These properties, "fingerprints," are multivariate analytical signals acquired by analysis using, for example, spectroscopic, chromatographic, electro-analytical, or other analytical techniques. The results of the "fingerprints" collection are presented in data matrices. The main matrix **X** is a set of data obtained using samples of the target class. The target class is always unique for a given OCC problem. Any other objects, or classes of objects, which are not the members of the target class, are considered as aliens.

Soft independent modeling of class analogies (SIMCA; Wold, 1977) is probably the best-known approach among different OCC. One of the main advantages of SIMCA relies on the fact that, as it is a "soft" classification technique, one object may be classified into one, several, or no classes (Gurbanov, Gozen, & Severcan, 2018). If an unknown object belongs to more than one class, this result may indicate that either the responses used to build the classification model lack sufficient power to discriminate between the classes or that there might be an error in the measurement. One main disadvantage of SIMCA is that it requires prior analysis by PCA; the result depends on the choice of PCs, which is facilitated by the minimum number for which the training set is correctly classified. SIMCA is sensitive to outliers, but they can be recognized by the method.

In food science, we observe many papers that improperly use SIMCA as a tool for discrimination between several predefined classes. SIMCA is an OCC that produces a description of a single target class of objects, and then detects whether, or not, a new object resembles this class. The rigorous version of SIMCA, as any other OCC does not utilize any information about nontarget (extraneous) classes, even when the data regarding such extraneous classes are available (Rodionova, Oliveri, & Pomerantsev, 2016).

SIMCA consists of two steps: At the first step, PCA (Eq. [(1)]) is applied to the training data extracted from the target class. At the second step, the PCA results are used for calculating two relevant distances for each object $i = 1, \ldots, I$ of the training set. As shown in Eq. (2), these are the score distance, h_i, and the orthogonal distance, v_i:

$$h_i = \mathbf{t}_i^t (\mathbf{T}^t \mathbf{T})^{-1} \mathbf{t}_i, \quad v_i = \sum_{j=1}^{J} e_{ij}^2 \qquad (2)$$

The score distance represents the position of a sample within the score space, and the orthogonal distance characterizes a sample distance to the score space.

A newly enhanced version of SIMCA is able to characterize classification results in a statistically sound way, that is, to calculate the errors of misclassification based on theoretical principles (Pomerantsev & Rodionova, 2014). Both the score distance and orthogonal distance follow the scaled chi-squared distribution and thus provides a possibility for developing tolerance areas for a given type I error. Regardless of the complexity of the initial data, the result of classification can be visualized in the two-dimensional acceptance plot (Figure 3). The solid line presents the border of the acceptance area developed for a predefined α-value. All samples located in this area are considered as the target class members. In Figure 4, all but one sample from class T3 are properly classified and one object is misclassified as alien (the type I error). All samples from class T1 are properly classified as aliens with respect to class T3.

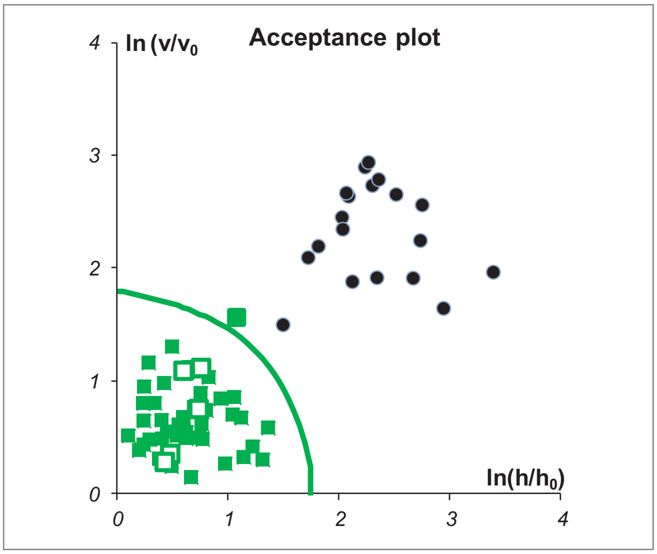

Figure 3–Result of newly enhanced version of soft independent modeling of class analogy (SIMCA) classification. Classification of olives in brine (Rodionova et al., 2016). Target class is class T3, PCA model with 6 PCs Filled green squares are training samples from class T3. Clear green squares are test objects from class T3. Black dots are the alternative objects from class T1.

The OCC methods are intensively applied in food chemistry for quality control and authentication of various foods, for example, for classification of olive oils (Paolo Oliveri et al., 2010), quality control of peanut oils (Xu, Cai, & Deng, 2011) and fruit juices (Fidelis et al., 2017), and many other cases. Sometimes, analysts try to compare the performance of SIMCA with PLS-DA or other discriminant methods. It is important to underline that these methods are not comparable, as they are aimed at solving different problems and, therefore, have different application areas (Paolo Oliveri & Downey, 2012; Rodionova, Titova, & Pomerantsev, 2016).

For instance, there is a clear advantage of using SIMCA over *k*-NN, which uses the closeness of samples in the space for classification, although SIMCA uses a defined boundary to classify unknown samples. However, SIMCA also has disadvantages over other classification methods: SIMCA is unable to produce probabilistic classifications and the number of responses and samples used in the classification affect the accuracy in a way that if the number of the independent variable and the sample size are bigger than 100, the model fails and results in overfitting (Kanik, Orekici Temel, Erdogan, & Ersoz Kaya, 2013). Another disadvantage of SIMCA is that no attempt is made to find directions that separate classes.

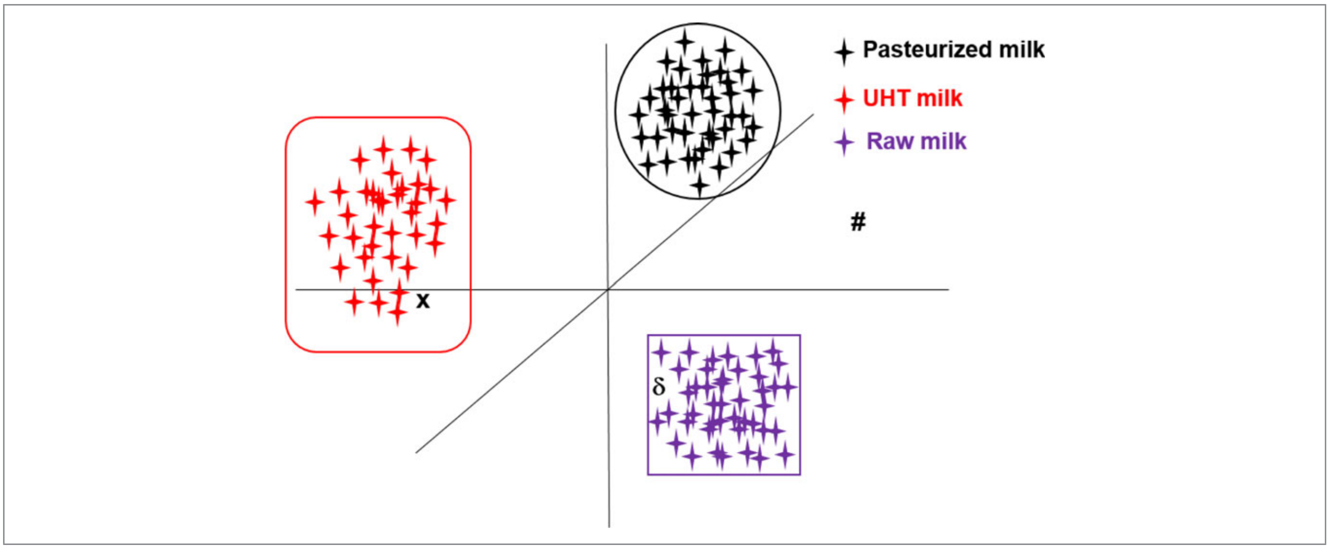

Figure 4–Example of SIMCA application on the classification of different milk types (3 classes). Samples x, #, and δ are unknowns.

Figure 4 illustrates another application of SIMCA to distinguish three milk types: sample "*X*" would be classified as "UHT milk," although sample "*δ*" would be classified as "raw milk." However, sample "#" would not be classified into either of these classes. As mentioned before, SIMCA may recognize one sample as fitting into one, two, multiple, or none of the classes. In authentication studies, the method sometimes does not provide useful information as it fails adequately to classify the samples. It was explained that an "inconclusive ratio" should also be calculated when SIMCA is used (Gondim et al., 2017). This parameter indicates the percentage of samples that were not assigned to any of the *k* classes and the samples that are assigned to more than one class. Authors have reported the use of SIMCA and mid-infrared spectroscopy to detect several adulterants in milk (for example, bicarbonate, NaOH, chloride, hypochlorite, water, sucrose, starch, H_2O_2, and carbonate) with over 80% correct classification. However, classification of a high proportion of samples (17% of samples) was inconclusive.

ANN. ANN is a technique that is used when the relationship between the dependent and the independent variables is not known a priori, representing complex, and nonlinear processes (Pacella & Semeraro, 2014). It is based on the structure of neurons (Boareto, De Souza, Valero, & Valdman, 2007) and, among the ANN types, the most common is the multilayer perceptron—MLP (Rasouli & Ghavami, 2016), where the neuron layers (input, hidden, and output) are connected by a feedforward connection (Figure 5). Some functions, with their parameters, are needed to train the network; usually, backpropagation (BP) is used as it is simple and versatile. The input values of these parameters are found after reaching an intended error level when comparing the output with the target values imposed. Broadly speaking, the idea is to have sets of experiment for use in different steps: the first set is needed to train the ANN; the second one is used to run a process simulation, and the last set is used to validate the experimental data. The node number must be adequate to train the network properly. According to several authors (dos Santos, Páscoa, & Lopes, 2017; Fan et al., 2013; Funes, Allouche, Beltrán, & Jiménez, 2015; Rasouli & Ghavami, 2016), one of the major advantages of ANN is that it may be applied when data are highly nonlinear correlated and there are uncertainties. It may be used also for process optimization.

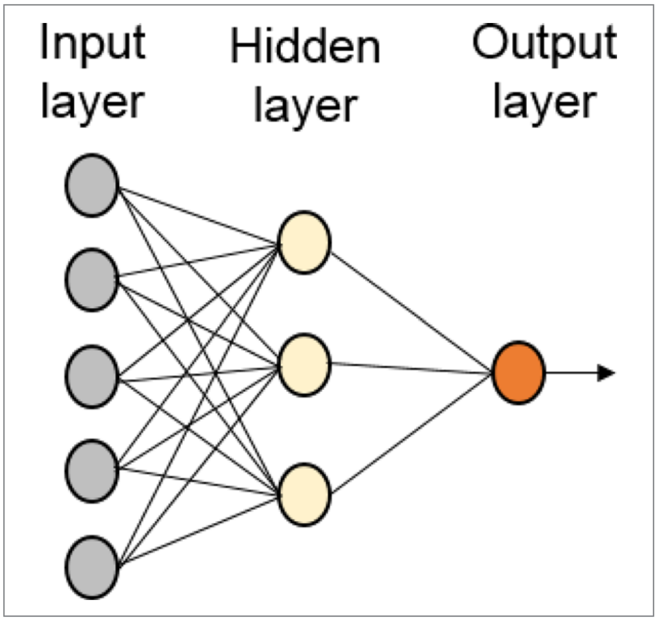

Figure 5–Schematic structure of a neural network with 5 input layers and one output layer.

It is important to note that when applying ANN, certain key issues need to be considered. Usually, an ANN has many parameters to be estimated. For example, for training an ANN, it is recommended that one uses an experimental data base at least three to five times greater than the number of parameters (some authors use 10 times), while paying attention to model overfitting. Once the parameters are estimated, another data set should be used to validate the proposed model (Souza Jr. & Trica, 2013). The ANN methodology can be applied in many food-related fields. Funes et al. (2015) presented well-known neural networks models, their limitations, listing many publications that have applied ANN methodology. Fan et al. (2013) used BP-ANN to obtain a relationship between instrumental color attributes and texture characteristics and to simulate and predict accurately the texture characteristics from the color values. Muñiz-Valencia, Jurado, Ceballos-Magaña, Alcázar, and Hernández-Díaz (2014) also used PCA and ANN to discriminate coffees from different regions of Mexico, by using estimates of mineral cations. They concluded that the concentrations of Ca, K, Mn, Mg, Na, and Zn provided the best descriptors and that differentiation of coffee sources was possible by using the ANN technique, with 93% of prediction ability and specificity of 98%.

Lohani and Muthukumarappan (2017) developed an interesting approach to compare two techniques in a study of the influence of process variables (fermentation time, flour to water ratio, flow rate, ultrasonication time, and ultrasonication intensity) on total phenolic content and antioxidant activity in sorghum flour. These techniques were response surface methodology (RSM; Box-Behnken design) and ANN. The RSM technique proposed a polynomial model (with quadratic and interaction terms) to describe a relationship involving all five process variables in 46 experiments that could predict a dependent variable value (Y_i). They adopted an input layer with five neurons (representing the five process variables) and an output layer with one neuron (for each answer variable). Both techniques presented models with high coefficients of determination (R^2).

Other studies have applied ANN and *k*-NN to classify food quality as a function of cooked food color (raw, light, optimal, dark, and burnt (O'Farrell, Lewis, Flanagan, Lyons, & Jackman, 2005a; O'Farrell, Lewis, Flanagan, Lyons, & Jackman, 2005b). They present a very good comparison of both methods using fresh minced beef burgers and chicken en croute for which the different color parameters were measured spectrometrically. They concluded that the ANN technique was better than *k*-NN, although the latter was satisfactory and needed simpler calculation.

Pérez-Caballero et al. (2017) applied other techniques to classify different types of tequilas (white, rested, aged, and extra-aged), using data from UV-Vis and CG-MS. They concluded that nonlinear models were best able to classify the tequilas samples. Among the techniques used, classification and regression trees (CART), random forest (RF), and support vectors machines (SVM) were the best, although *k*-NN, which is a simple method, also presented reliable results.

Effects of processing on food components

In food science, effects of processing on food components is most commonly discussed in a context of negative influences of thermal processing on food chemistry. Many reports show that thermal processing of foods adversely affects the nutritional content and sensory properties of foods, but thermal treatment is still necessary for inactivating microbial pathogens and enzymes, in order to prolong shelf-life (van Boekel et al., 2010). Raw fruits and vegetables (Poojary et al., 2017; Putnik et al., 2017; Putnik et al., 2017), their products (Bursać Kovačević et al., 2015), and byproducts (Putnik et al., 2017) are among the most sensitive materials for (thermal) processing, yet innovative applications of (nonthermal) technology can decrease or alleviate deterioration of important biologically active compounds. In these examples, use of multivariate statistics, especially multivariate regression, has enabled optimization of parameters.

Several examples of the use of PCA have reported on the further use of multivariate statistics for identifying changes in food chemistry during processing. One example focused on various polyphenolic compounds in chokeberry (*Aronia* spp.) juice, which was given by Bursać Kovačević et al. (2016) who studied the effects of cold atmospheric gas phase plasma on the levels of hydroxycinnamic acids, flavonols, and anthocyanins. For this purpose, authors used PCA, MANOVA, and multivariate regression. Results were evaluated using two types of controls (untreated and samples pasteurized for 2 min at 80 °C). PCA data showed that the pasteurization process strongly decreased the concentrations of hydroxycinnamic acids, although flavonols and anthocyanins increased slightly. In contrast, plasma-treated samples had increased levels of hydroxycinnamic acids and reduced anthocyanins compared to the untreated samples. The procedure was used to optimize the conditions for pasteurization of the juice.

For the purpose of meat processing, Shikha Ojha et al. (2018) used PCA to evaluate the multivariate effects of ultrasound frequency (25, 33, and 45 kHz), drying time, and addition of *Lactobacillus sakei* on proteolysis rate and levels of total protein, amino acids, organic acids, texture, and color of beef jerky. PCA was able to differentiate the control samples (without *L. sakei*) from the test samples using the responses (other than acidity and color) in a 2D projection. Although a high amount of the data variance was explained by two principal components (roughly 70%), no clear differences were observed between ultrasound frequencies. Overall, in technological applications, PCA enables clear differentiation process variables on selected responses.

The influence of harvesting, genotype, climate, and processing on levels of tocochromanols, carotenoids, and chlorophyll in flaxseed oils was evaluated using MANOVA (3-way ANOVA) and PCA (Obranović et al., 2015). MANOVA was used for the simultaneous comparison of three independent variables for each of the eight dependent variables, although PCA was used to construct a climate index, consisting of temperature (°C), sunshine hours, and levels of recorded rainfall. The oil from Genotype Biltstar had highest antioxidant activity, while that of Altess had the highest pigment contents. The fifth week after flowering was identified as the optimal maturation period to obtain the highest tocochromanol content. It was shown statistically that the content of γ-tocopherol and plastochromanol-8 increased with temperature and sunshine, and reduced with higher rainfall during the maturation period. In general, oils obtained by cold pressing yielded higher contents of tocochromanols and with less pigmentation. The study identified the most suitable flaxseed genotype for oil processing and shed some light on the ecological impact on biosynthesis of important food constituents.

Seafood, commonly packaged in a modified atmosphere to prolong the shelf-life, can be monitored instrumentally to detect spoilage odors composed of numerous volatile organic compounds. It is essential to be able to differentiate between the presence of large numbers of both odoriferous spoilage compounds and volatile compounds normally associated with the food. Multivariate tools can be used to analyze the large data sets in order to sort "the wheat from the chaff." Multivariate statistics, HCA, PCA, and PLS regression were used to identify relevant spoilage volatiles (Kuuliala et al., 2018). Amongst others, it was found that certain compounds (such as acetic acid, isobutyl alcohol, dimethyl sulfide, and trimethylamine) were associated with microbial growth and unacceptable sensory evaluation and identified a potential for development of smart packaging.

Food authentication based on chemical markers

In some countries, scandals involving adulteration of foods have become more frequent and most are economically-based (Hong et al., 2017; van Ruth, Huisman, & Luning, 2017). For instance, Brazil witnessed an extensive investigation into the largest food companies (especially meat-based ones) for bribing inspectors who allowed corruption in food production. This fraud enabled spoiled meals to be served in public schools and *Salmonella*-contaminated meats to be exported to Europe. The total nitrogen content of raw milk, powdered infant formula, and cereal-based formulations have been increased by addition of melamine ($C_3H_6N_6$) in many countries (Zhang & Xue, 2016). Addition of sucrose and water into honey, a high-value added food, also occurs worldwide (Soares et al., 2017; Spink, Ortega, Chen, & Wu, 2017). These are typical examples of economically motivated frauds of high-value added foods that lead consumers to have concerns about the authenticity of their foods. Such worldwide adulteration of food is illegal, and commercial products need to be monitored constantly by government agencies to detect and prevent adulteration (Spink et al., 2017).

Typical examples of questions that arise in food authenticity include: is the olive oil really from the Tuscany region? Are the meatballs from Wagyu beef (a Japanese cattle breed) really made from this breed? Is the Pinot Noir wine really made of this grape variety? Such questions are very hard to answer without the use of analytical fingerprints and authentic samples as references. Oliveri and Simonetti (2016) stated that to assess whether a food product is authentic is a complex task that involves assessment of multiple characteristics, including physical, chemical, microbiological, and biochemical properties. Hong et al. (2017) provided interesting and current information on the multiplicity of analytical methods that can be used to identify adulteration in foods. Such methods include vibrational spectroscopy (dos Santos et al., 2017), including near-infrared, NIRS (Chiesa et al., 2016), mid-infrared spectroscopy (Karoui, Downey, & Blecker, 2010), and Fourier-transform infrared (FTIR; Gao, Zhou, Han, Yang, & Liu, 2017), nuclear magnetic resonance, NMR (Gad & Bouzabata, 2017; Longobardi et al., 2017; Spiteri et al., 2017), mass spectrometry (Wu et al., 2017), proton transfer reaction mass spectrometry (Granato, Koot, & van Ruth, 2015), spectrophotometric, potentiometric, and chromatographic methods (Alonso-Salces, Serra, Reniero, & Heberger, 2009; Granato, Margraf, Brotzakis, Capuano, & van Ruth, 2015; Wu et al., 2017), and other methods (Azcarate, Gil, Smichowski, Savio, & Camiña, 2017; Bevilacqua et al., 2017; Dong, Zhao, Hu, Dong, & Tan, 2017). Such methods provide a robust fingerprint of the test samples and usually generate a large and complex data matrix that, if properly analyzed, can show even slight differences between factors (such as lots, manufacturers, geographical origin, and so on; Peng et al., 2016).

Because consumers do not have access to these sophisticated and highly sensitive analytical methods that are necessary to detect food fraud, they require assurance that purchased food is authentic, and/or compliant with legislation and with the statements on the food labels (Charlebois, Schwab, Henn, & Huck, 2016; Lucci et al., 2017; Walker, 2017). Therefore, strategies to increase and improve the monitoring of food authenticity are of major concern worldwide for entire food supply chains (Camin et al., 2017; Danezis et al., 2016; Manning, 2016; Sabir, Rafi, & Darusman, 2017).

In the period from 2000 to December 2017, more than 8,100 papers dealing with food authenticity have been recorded in the Science Direct database, and a crescent investigation and publication of articles is observed (total number of articles $= 49.238 \times \text{Year} - 98440$, $R^2 = 0.9399$). Studies conducted for government agencies show that the generation of large data sets is required to attest the authenticity of origin (that is, protected designation of origin or protected geographical indication), farming systems (that is, organic and biodynamic), safety, and overall expected quality of foods (Bajoub et al., 2017; Yang et al., 2016). Therefore, the use of multivariate statistical methods, for both exploratory and classification purposes provides an appropriate strategy for governmental, academic, and industrial stakeholders (Borràs et al., 2015; Pardo-Mates et al., 2017). The use of analytical methods with proper data analysis may support proper monitoring of food safety management systems in food and feed chains (Pustjens, Weesepoel, & van Ruth, 2016).

According to one study, authentication of a food requires procedures to determine whether a specific food conforms with its description (that is, organic, biodynamic, from a certain location, and so on; Rodionova et al., 2016). Authentication requires comparison with definitive reference materials (that is, authentic samples with known and stable physicochemical properties); this enables a fingerprint of such foods to be traced using OCC methods.

Here review some selected examples regarding the application of chemometrics for the assessment of food authentication and for the accurate spatiotemporal identification of food geographical origin. The origin of cocoa impacts the final product quality of chocolate. Traceability is a difficult task that requires refined analytical approaches to obtain accurate results. Marseglia et al. (2016) used a sophisticated high-resolution magic angle spinning ^{1}H nuclear magnetic resonance (HR-MAS ^{1}H NMR) method to authenticate the geographical origin of 60 fermented and dried Forasteiro cocoa beans from 23 different locations. Spectroscopic data were obtained, preprocessed (Fourier transform) and analyzed with the aid of PCA and OPLS-DA. Almost 65% of data variability was explained using 2D projection, and samples from Africa, America, and Asia/Oceania were well separated. PLS-DA results showed that American and African cocoa beans were efficiently classified using the analytical data. This example clearly shows that if a food company wishes to monitor the origin of food products with intrinsic characteristics (that is, foods or ingredients labeled as, for example, "protected geographical indication"), HR-MAS ^{1}H NMR coupled with chemometrics may provide a suitable analytical strategy for quality control purposes.

A group of authors studied the fatty acid profiles of 50 organic and 72 conventional chicken feed samples in the Netherlands (2009 to 2010) and used a classification method based on PLS-DA to authenticate the farming system (Alewijn, van der Voet, & van Ruth, 2016). More than 92% of samples were correctly differentiated in the validation (100% sensitivity and 95% specificity in the training set) and the measurement of fatty acids coupled with PLS-DA may represent a fast screening test to assess if organic laying-hen feed is produced according to the organic protocol. This study shows the importance of chemometrics for authentication purposes of organic feeds as such materials can be easily faked.

Salmon and herring from the Baltic Sea are constantly monitored because of contamination with polychlorinated biphenyls (PCBs), polychlorinated dibenzo-p-dioxins and dibenzofurans (PCDD/Fs). In this scenario, Sørensen, Lund, Cederberg, and Ballin (2016) quantified PCBs and PCDD/Fs in 79 salmon samples from Canada, Chile, China, Norway, USA, Vietnam, and the Baltic Sea near Denmark. PCA analysis of the PCB profiles showed substantial differences between salmon samples from different locations. The result was supported by the 2D projection of total variance. The compounds mainly responsible for the association with the geographical origins were identified by the factor loadings. The authors concluded that contamination of the Baltic Sea salmon is mostly influenced by the low-chlorinated congeners that are typical for industrial processes carried out in earlier times around the Baltic Sea. Hence, analysis of PCBs and PCDD/Fs compounds coupled with a simple PCA was an effective approach for monitoring both the quality and authentication of the origin of salmon samples. This paper provides a good example of the importance of PCA in data analysis.

Chiesa et al. (2016) authenticated 30 samples of a unique Italian "Valle d'Aosta Arnad" PDO lard using near-infrared spectroscopy (NIRS). For comparative purposes, 30 non-PDO lard samples were analyzed. Volatile organic compounds and fatty acid composition were used to characterize the lard samples and develop calibration models using PLS regression. PLS-DA of fatty acids and VOCs (100% sensitivity and 100% specificity) provided 100% correct differentiation of PDO and non-PDO lard samples. NIRS data also showed promising results with 100% sensitivity and 96.4% specificity. Because the Italian POD lard is a high-value meat product, the use of a nondestructive, solvent-free, and rapid technique (NIRS) coupled with chemometrics provides an ideal way to check for fraudulent products.

Bajoub et al. (2017) assessed the liquid chromatographic fingerprints of phenolic compounds from seven mono varieties of 140 extra-virgin olive oil samples processed during 2011 to 2014 in Morocco, and they used exploratory and classification methods to analyze the data. Three principal components explained about 80% of data variability and showed the clustering of Cornicabra and Frantoio varietals, whilst data for Picholine, Picual, and Languedoc were highly dissimilar. PLS-DA and SIMCA were used to classify the oils according to the varietals, and PLS-DA showed a high percentage of correct classification (that is, sensitivity and specificity) between varietals with external validation (between 90% and 100%). SIMCA data also showed correct classification rates (between 94% and 97%), and 100% correct classification in recognition and external validation was obtained for Arbequina and Cornicabra varietals.

Dong et al. (2017) sought to differentiate seven Robusta coffee cultivars (126 samples) using electronic nose and tongue by titratable acidity, pH, and soluble solids. For this purpose, *k*-NN and PLS-DA were the supervised chemometric tools used. 91.7% of the samples were differentiated by PLS-DA, although *k*-NN had an accuracy of 92.8% using only the e-nose results. Only two samples were misclassified by PLS-DA for the e-tongue data. These results showed that chemometrics coupled with the selected analytical methods is useful to authenticate coffees from different cultivars.

Consumers pay a premium price to acquire organic vegetables worldwide. Therefore, sound analytical methods are necessary to distinguish organically and conventionally grown crops and check for accuracy of the product labeling. Liquid chromatography-mass spectrometry (LC-MS) was used to characterize some chemical compounds from 140 carrot samples of Nerac and Namur varieties from Belgium (Cubero-Leon, De Rudder, & Maquet, 2018). PCA failed to separate organic from conventional carrots; the 2D projection explained only 22% of data variance. However, PCA was effective in showing clusters of carrots based on the production year. Orthogonal projections to latent structures-discriminant analysis (OPLS-DA) was used to differentiate the organic from conventional carrots. Results varied from 77.3% to 83.8%, with a total of 15 compounds being responsible for the classification. Nerac was differentiated from Namur variety with 90% accuracy.

One authentication strategy to assess the geographical origin of foods is to trace and validate specific stable markers. To authenticate the origin of 31 purple grape juices ($n = 6$ biodynamic, $n = 7$ organic, and $n = 18$ conventional) from different European countries, Granato et al. (2015) used a range of chemical, physicochemical, antioxidant activity, and instrumental taste parameters. No statistical differences were observed among farming systems by ANOVA, but PCA was effective only in showing that the total phenolic tongue, antioxidant activity, and taste of grape juices varied according to the production region. HCA showed no clear separation among farming systems and highlights the fact that this method is highly arbitrary and should not be used for "authentication" purposes. SIMCA, on the other hand, differentiated 12 of 13 organic/biodynamic juices and 17 of 18 conventional juices. PLS-DA classified 11 organic/biodynamic juices and all conventional juices. Instrumental richness, saltiness, and total soluble solids contributed significantly to the classification of grape juices.

Karabagias et al. (2017) characterized 37 Citrus spp. honeys from the Mediterranean region (Greece, Egypt, Morocco, and Spain) collected in 2013/2014 based on color and physicochemical parameters (moisture content, pH, acidity), individual minerals, and volatile organic compounds (VOCs). Analysis of variance was used to assess the effects of origin on the characteristics of the honeys, and LDA was used to trace the origin of the honeys. Results showed that Greek and Spanish honeys were very similar and formed a cluster that was statistically different from the Moroccan and Egyptian honeys. The honey samples from each country could be separated using LDA (100% correct classification) based on the mineral contents, of which Mo, Si, Se, Li, Ti, Ca, P, Sb, B, Sn, Sr, Ni, and Cu were the main predictors for portraying the origin of citrus honeys. LDA of 15 VOCs enabled 97% of correct differentiation of citrus honeys according to the geographical origin. LDA of the physicochemical and color parameters classified 100% of the honeys based on lightness (L^*), greenness (a^* coordinate), moisture, acidity, and pH as the most discriminatory variables. This is a good example where chemometrics using data from simple routine analyses can be used to classify high-value foods.

Evaluation of the adulteration of 50 samples of Greek saffron (*Crocus sativus* L.) with other less-expensive plants (such as safflower, turmeric, buddleia, calendula, and gardenia) using diffuse reflectance infrared spectroscopy (DRIS) in the range of 4,000 to 600 cm^{-1} was reported (Petrakis & Polissiou, 2017). Saffron samples were divided into organic ($n = 28$) and conventional ($n = 22$). PCA was used to assess the patterns for authentic saffron samples and to compare the purity of commercial samples, and PLS-DA was used to detect possible adulterations in the products. PCA accounted for more than 90% of the variability in the experimental results, but it was unable to differentiate between samples. PLS-DA analysis of DRIS data gave 95% correct classification of samples. Overall, use of a nondestructive analytical measurement of saffron samples may be a feasible tool to assess the authenticity of high-value added herbs and extracts and to enhance the potential of this approach for on-site analysis and fraud detection within the food chain.

The variety or cultivar plays an important role in determining key features of fruit and vegetable products. Margraf, Santos, de Andrade, van Ruth, and Granato (2016) studied the phenolic composition, physicochemical properties, and antioxidant activity of Brazilian grape juices from different geographical origins, varieties, and farming systems. PCA showed that the inhibition of lipid peroxidation of juices was highly associated with the total phenolics content, especially protocatechuic acid, flavonols, anthocyanins, and total flavonoids. However, when the farming system was analyzed, PCA could neither differentiate between data from organic and conventional juices nor among producing regions and grape varieties (Cubero-Leon et al., 2018). PLS-DA classified 80% of grape juices correctly assigned to their class (organic or conventional), with 72% sensitivity in the calibration step and 100% specificity in the external validation. Color intensity measured by a spectrophotometer, total flavonoids, p-coumaric, and syringic acids were the main discriminatory variables.

Application of chemometrics in microbiology

The definition of chemometrics implies that the technique is solely concerned with handling and analysis of chemical data, but increasingly in biological sciences many analyses are now based on "rapid" or alternative methods that determine chemical or physical criteria. There are four areas where chemometrics has been used extensively in microbiological studies:

(1) Taxonomic studies of food-associated organisms, especially species or strains of food-borne pathogens, such as *Salmonella;*
(2) Food spoilage studies, where chemical changes that occur during storage are evaluated in relation to microbial growth;
(3) Use of bioassays to determine food contaminants, such as veterinary antibiotic residues; and
(4) A range of other microbiologically related studies.

Microbial taxonomics. In tracing outbreaks of foodborne disease (or spoilage), detailed knowledge of the specific causative organism is often essential. Traditional diagnostic methods for the identification of microorganisms is adequate if you only need to know the genus, species, and strain of a specific organism involved in a food-borne disease episode (or even in a food spoilage incident). But there is often an epidemiological need for more definitive and rapid identification, in order to compare isolates from patients with isolates from suspect foods or environmental sources.

For example, the species *Salmonella enterica* contains many subspecies that can be identified by immunological and/or DNA typing, but this is very time-consuming and not always adequate for the purpose. One group of authors evaluated the use of FTIR spectroscopy and chemometrics to differentiate intact cells and crude lipopolysaccharide extracts from *S. enterica* serotypes and then applied canonical variance analysis (CVA) to appropriate regions of the FTIR spectra (Kim, Reuhs, & Mauer, 2005). Although the results from intact cells were inadequately differentiated (only 50% to 70% correct identifications), results from crude lipopolysaccharide extracts permitted 100% correct differentiation.

Kim, Kim, Reuhs, and Mauer (2006) extended this work by evaluating extracted cell outer membrane proteins (OMP) using FTIR/CVA and compared the results with SDS-PAGE analyses. For the serotypes studied, the FTIR/CVA method provided better differentiation than the electrophoretic analysis of the proteins, giving 100% correct identification of the serotypes studied (Kim et al., 2006).

The use of FTIR was evaluated to differentiate between isolates of *Escherichia coli* O157:H7 previously typed using multilocus variable number tandem repeat analysis (MLVA) and pulsed field gel electrophoresis (PFGE; Davis, Paoli, & Mauer, 2012). HCA and CVA of the FTIR spectra resulted in the clustering of the same or similar MLVA types and separation of different MLVA types of *E. coli* O157:H7. The developed FTIR method showed better discriminatory power than PFGE in sub-typing *E. coli* O157:H7, and it demonstrated that FTIR spectroscopy is suitable for rapid (≤16 hr) and economical subtyping of *E. coli* with comparable accuracy to MLVA typing. Strains were also classified (97% accuracy) based on the type of Shiga toxin present using CVA of the spectra.

Brandily et al. (2011) reported the use of fiber evanescent wave spectroscopy (FEWS) to classify diverse pathogenic organisms in samples of minced meat and sausage meat and in cheese. The technique requires the sample to be brought into contact with a special optical glass fiber, which is linked to an FTIR spectrometer. The output data were evaluated using PCA and logistic partial least squares (LPLS). PCA enabled differentiation of several pathogenic bacteria and the LPLS enabled further discrimination. However, the work has yet to be applied to realistic levels of pathogens in food samples.

Microbiological food spoilage. Although some purely chemical spoilage of food occurs, most food spoilage is associated with the growth of microorganisms. Several teams have used chemometric protocols to analyze results from food spoilage studies, not least because of the plethora of volatile and nonvolatile compounds often produced. Duflos and coworkers studied the spoilage of fish using headspace/mass spectrometry (HS/MS) and solid-phase micro-extraction gas chromatography/mass spectroscopy (SPME/GCMS) followed by PCA to analyze the outputs following storage of fish for 10 days at 4 °C (Duflos et al., 2010; Duflos, Coin, Cornu, Antinelli, & Malle, 2006). Of the 86 compounds identified, about 20 compounds typically produced by microbial activity could be used as indicators of spoilage. In a subsequent study on stored samples of Whiting, they used PCA to compare the results of subjective sensory assessments with the outputs from SPME/GCMS analyses to identify volatile compounds that were associated either with freshness or with spoilage.

Mikš-Krajnik, Yoon, Ukuku, and Yuk (2016) used similar analytical methods to study the spoilage of fresh salmon at various temperatures. HCA and Pearson's correlations were used to evaluate and to select compounds for use as chemical spoilage indices from the SPME/GCMS outputs. PLS regression was used to assess possible relationships between specific microorganisms and the production of volatile compounds in fish. Similar procedures were used also successfully by Vasconcelos, Saraiva, and de Almeida (2014) to study the relationship between microbial levels in chicken breast fillets stored at 3, 8, and 30 °C and changes in FTIR, pH and sensory assessments using PLS regression, discriminant analysis, and PCA procedures.

Bruckner, Albrecht, Petersen, and Kreyenschmidt (2012) studied the effects of variation in the chill storage conditions for raw pork chops and chicken breast fillets, packed aerobically in plastic film. A nondimensional sensory index score (SIS; nondimensional combination score based, as stated, on color, texture, and odor of the stored products) was used to provide an overall assessment of changes in color, texture, and odor of stored products; colony counts for TVC and pseudomonas were done throughout the storage period. The microbiological data were fitted to the Gompertz model using nonlinear regression and changes in time to the SIS were fitted by linear regression. Although the stated aims of the investigation were achieved, further analysis using PLS or other technique could have provided additional information from comparison of SIS results with the microbiological data.

Antimicrobial bioassays. Bioassays using selected strains of microorganism are used routinely to assess levels of antibiotic or other antimicrobial agents. Traditionally bioassays are used also to assess toxicity, levels of vitamins in food, and so on, Nagel et al. (2009) used a logistic regression model (LRM) and a concordance coefficient to assess the effects of chloramphenicol in the culture medium used for a "presence or absence" bioassay with

Geobacillus stearothermophilus subsp. *calidolactis* to detect and quantify tetracyclines in milk. Increasing concentrations of chloramphenicol significantly reduced the limit of detection for tetracyclines. In subsequent works, authors used the LRM to optimize bioassay conditions (Nagel, Molina, & Althaus, 2011, 2012).

In a search for new antimicrobials Suleman, van Vuuren, Sandasi, and Viljoen (2015) examined extracts from propolis, a sticky resin found in beehives, by LC-MS, chemical analyses, and bioassays. The data were evaluated using multivariate data analysis by orthogonal projections to latent structures (OPLS) and an S-plot function. Potential antimicrobials were identified as flavonoid compounds.

Other microbiological applications. Guo et al. (2011) used PLS followed by Monte Carlo PLS (MC-PLS) modeling to analyze the data from, and optimize the process for, at-line monitoring of the fermentation process used for nisin production. The nisin titer, reducing-sugar concentration, cell concentration, and pH values were compared with near-IR spectra obtained during fermentation. The optimal wavelengths for the NIR and the most efficacious methods for pre-processing the spectra were determined.

Lei and Jakobsen (2004) evaluated the diversity of lactic acid bacteria (LAB) in "koko" and "koko sour water," products of spontaneous fermentation of millet porridge, a traditional Ghanaian foodstuff. PCA and partial LSR analyses of data were used to link the LAB species to the different production stages and production sites. The isolates were found to have a great diversity at the intra-species level and were investigated for antimicrobial activity using agar diffusion assays, and acid and bile tolerance. Most isolates showed low levels of antimicrobial activity towards the test strain of *Listeria innocua*, and some were considered to have probiotic potential.

Boussard et al. (2012) used a battery of chemometric tools to examine the inter-relationships of yeast, glucose oxidase, horse bean, and soybean flours on the biochemical characteristics of white bread dough during the fermentation period. Free and bound polyunsaturated fatty acids (PUFAs), primary oxidation products of linoleic acid and other parameters were modeled using PCA, Pareto charts, and Mahalanobis distances.

Although not strictly a microbiological application, analyses of changes in the composition of bottled beers over a 12-month storage period were studied (Rendall et al., 2015) using PCA and the square prediction error (Q-statistic) to model GC-MS data. The authors observed a major change in composition after 7 months of storage. Importantly, this paper discusses in great detail the need to use a systematic approach to the chemometric analysis of data, especially in cases where many analytical data are to be assessed.

Trends and Final Comments

Chemometric tools have been studied for problem solving purposes in many scientific and technological fields. In food chemistry, multiple problems may be investigated using different analytical and statistical methods: adulteration, analysis of geographical/production origins, effects of processes, and unit operations on the quality attributes of foods. Although their use has grown, it is still very challenging to choose the best statistical method because all of them have both positive and negative attributes. This paper provides an overview of some of the most recent applications of *k*-NN, SIMCA, ANN, PLS-DA, and LDA in food science. The fundamental benefits and disadvantages have been illustrated using real world examples. Our main recommendation is not to apply any chemometric method at hand but to choose those that answer a specific need.

Acknowledgments

D. Granato acknowledges CNPq for a productivity grant (process 303188/2016-2). Jânio S. Santos thanks Fundação Araucária/CAPES for his PhD scholarship. O. Rodionova and A. Pomerantsev acknowledge a partial support within the Russian state assignment 0082-2014-0019. P. Putnik and D. B. Kovačević wish to acknowledge the Croatian Science Foundation project named: "High voltage discharges for green solvent extraction of bioactive compounds from Mediterranean herbs (IP-2016-06-1913)."

References

Alañón, M. E., Pérez-Coello, M. S., & Marina, M. L. (2015). Wine science in the metabolomics era. *TrAC Trends in Analytical Chemistry*, *74*, 1–20. https://doi.org/10.1016/j.trac.2015.05.006

Alewijn, M., van der Voet, H., & van Ruth, S. (2016). Validation of multivariate classification methods using analytical fingerprints – concept and case study on organic feed for laying hens. *Journal of Food Composition and Analysis*, *51*, 15–23. https://doi.org/10.1016/j.jfca.2016.06.003

Alonso-Salces, R. M., Serra, F., Reniero, F., & Heberger, K. (2009). Botanical and geographical characterization of green coffee (Coffea arabica and Coffea canephora): Chemometric evaluation of phenolic and methylxanthine contents. *Journal of Agricultural and Food Chemistry*, *57*(10), 4224–4235. https://doi.org/10.1021/jf8037117

Azcarate, S. M., Gil, R., Smichowski, P., Savio, M., & Camiña, J. M. (2017). Chemometric application in foodomics: Nutritional quality parameters evaluation in milk-based infant formula. *Microchemical Journal*, *130*, 1–6. https://doi.org/10.1016/j.microc.2016.07.016

Bajoub, A., Medina-Rodríguez, S., Gómez-Romero, M., Ajal, E. A., Bagur-González, M. G., Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2017). Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics. *Food Chemistry*, *215*, 245–255. https://doi.org/10.1016/j.foodchem.2016.07.140

Beebe, K. R., Pell, R. J., & Seasholtz, M. B. (1998). *Chemometrics: A practical guide* (1st ed). New York: Wiley & Sons.

Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A*, *1158*(1–2), 196–214. https://doi.org/10.1016/j.chroma.2007.05.024

Bevilacqua, M., Bro, R., Marini, F., Rinnan, Å., Rasmussen, M. A., & Skov, T. (2017). Recent chemometrics advances for foodomics. *TrAC Trends in Analytical Chemistry*, *96*, 42–51. https://doi.org/10.1016/j.trac.2017.08.011

Boareto, Á. J. M., De Souza, M. B., Valero, F., & Valdman, B. (2007). A hybrid neural model (HNM) for the on-line monitoring of lipase production by Candida rugosa. *Journal of Chemical Technology & Biotechnology*, *82*(3), 319–327. https://doi.org/10.1002/jctb.1678

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment – A review. *Analytica Chimica Acta*, *891*, 1–14. https://doi.org/10.1016/j.aca.2015.04.042

Boussard, A., Cordella, C. B. Y., Rakotozafy, L., Moulin, G., Buche, F., Potus, J., & Nicolas, J. (2012). Use of chemometric tools to estimate the effects of the addition of yeast, glucose-oxidase, soybean or horse bean flours to wheat flour on biochemical bread dough characteristics. *Chemometrics and Intelligent Laboratory Systems*, *113*, 68–77. https://doi.org/10.1016/j.chemolab.2012.01.006

Brandily, M. L., Monbet, V., Bureau, B., Boussard-Plédel, C., Loréal, O., Adam, J. L., & Sire, O. (2011). Identification of foodborne pathogens within food matrices by IR spectroscopy. *Sensors and Actuators B: Chemical*, *160*(1), 202–206. https://doi.org/10.1016/j.snb.2011.07.034

Bruckner, S., Albrecht, A., Petersen, B., & Kreyenschmidt, J. (2012). Influence of cold chain interruptions on the shelf life of fresh pork and poultry. *International Journal of Food Science & Technology*, *47*(8), 1639–1646. https://doi.org/10.1111/j.1365-2621.2012.03014.x

Bursać Kovačević, D., Gajdoš Kljusurić, J., Putnik, P., Vukušić, T., Herceg, Z., & Dragović-Uzelac, V. (2016). Stability of polyphenols in chokeberry juice treated with gas phase plasma. *Food Chemistry*, *212*, 323–331. https://doi.org/10.1016/j.foodchem.2016.05.192

Bursać Kovačević, D., Putnik, P., Dragović-Uzelac, V., Vahčić, N., Babojelić, M. S., & Levaj, B. (2015). Influences of organically and conventionally grown strawberry cultivars on anthocyanins content and color in purees and low-sugar jams. *Food Chemistry*, *181*, 94–100. https://doi.org/10.1016/j.foodchem.2015.02.063

Camin, F., Boner, M., Bontempo, L., Fauhl-Hassek, C., Kelly, S. D., Riedl, J., & Rossmann, A. (2017). Stable isotope techniques for verifying the declared geographical origin of food in legal cases. *Trends in Food Science & Technology*, *61*, 176–187. https://doi.org/10.1016/j.tifs.2016.12.007

Cebi, N., Durak, M. Z., Toker, O. S., Sagdic, O., & Arici, M. (2016). An evaluation of Fourier transforms infrared spectroscopy method for the classification and discrimination of bovine, porcine and fish gelatins. *Food Chemistry*, *190*, 1109–1115. https://doi.org/10.1016/j.foodchem.2015.06.065

Charlebois, S., Schwab, A., Henn, R., & Huck, C. W. (2016). Food fraud: An exploratory study for measuring consumer perception towards mislabeled food products and influence on self-authentication intentions. *Trends in Food Science & Technology*, *50*, 211–218. https://doi.org/10.1016/j.tifs.2016.02.003

Chiesa, L., Panseri, S., Bonacci, S., Procopio, A., Zecconi, A., Arioli, F., . . . Moreno-Rojas, J. M. (2016). Authentication of Italian PDO lard using NIR spectroscopy, volatile profile and fatty acid composition combined with chemometrics. *Food Chemistry*, *212*, 296–304. https://doi.org/10.1016/j.foodchem.2016.05.180

Como, F., Carnesecchi, E., Volani, S., Dorne, J. L., Richardson, J., Bassan, A., . . . Benfenati, E. (2017). Predicting acute contact toxicity of pesticides in honeybees (Apis mellifera) through a k-nearest neighbor model. *Chemosphere*, *166*, 438–444. https://doi.org/10.1016/j.chemosphere.2016.09.092

Cubero-Leon, E., De Rudder, O., & Maquet, A. (2018). Metabolomics for organic food authentication: Results from a long-term field study in carrots. *Food Chemistry*, *239*, 760–770. https://doi.org/10.1016/j.foodchem.2017.06.161

D'Archivio, A. A., Giannitto, A., Maggi, M. A., & Ruggieri, F. (2016). Geographical classification of Italian saffron (Crocus sativus L.) based on chemical constituents determined by high-performance liquid-chromatography and by using linear discriminant analysis. *Food Chemistry*, *212*, 110–116. https://doi.org/10.1016/j.foodchem.2016.05.149

Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *TrAC Trends in Analytical Chemistry*, *85*, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Davis, R., Paoli, G., & Mauer, L. J. (2012). Evaluation of Fourier transform infrared (FT-IR) spectroscopy and chemometrics as a rapid approach for sub-typing Escherichia coli O157:H7 isolates. *Food Microbiology*, *31*(2), 181–190. https://doi.org/10.1016/j.fm.2012.02.010

De Luca, M., Restuccia, D., Clodoveo, M. L., Puoci, F., & Ragno, G. (2016). Chemometric analysis for discrimination of extra virgin olive oils from whole and stoned olive pastes. *Food Chemistry*, *202*, 432–437. https://doi.org/10.1016/j.foodchem.2016.02.018

Derde, M. P., & Massart, D. L. (1988). Comparison of the performance of the class modelling techniques UNEQ, SIMCA, and PRIMA. *Chemometrics and Intelligent Laboratory Systems*, *4*(1), 65–93. https://doi.org/10.1016/0169-7439(88)80013-3

Do, T. K. T., Hadji-Minaglou, F., Antoniotti, S., & Fernandez, X. (2015). Authenticity of essential oils. *TrAC Trends in Analytical Chemistry*, *66*, 146–157. https://doi.org/10.1016/j.trac.2014.10.007

Dong, W., Zhao, J., Hu, R., Dong, Y., & Tan, L. (2017). Differentiation of Chinese robusta coffees according to species, using a combined electronic nose and tongue, with the aid of chemometrics. *Food Chemistry*, *229*, 743–751. https://doi.org/10.1016/j.foodchem.2017.02.149

dos Santos, C. A. T., Páscoa, R. N. M. J., & Lopes, J. A. (2017). A review on the application of vibrational spectroscopy in the wine industry: From soil to bottle. *TrAC Trends in Analytical Chemistry*, *88*, 100–118. https://doi.org/10.1016/j.trac.2016.12.012

Drivelos, S. A., Danezis, G. P., Haroutounian, S. A., & Georgiou, C. A. (2016). Rare earth elements minimal harvest year variation facilitates robust geographical origin discrimination: The case of PDO "Fava Santorinis". *Food Chemistry*, *213*, 238–245. https://doi.org/10.1016/j.foodchem.2016.06.088

Duflos, G., Coin, V. M., Cornu, M., Antinelli, J.-F., & Malle, P. (2006). Determination of volatile compounds to characterize fish spoilage using headspace/mass spectrometry and solid-phase microextraction/gas chromatography/mass spectrometry. *Journal of the Science of Food and Agriculture*, *86*(4), 600–611. https://doi.org/10.1002/jsfa.2386

Duflos, G., Leduc, F., N'Guessan, A., Krzewinski, F., Kol, O., & Malle, P. (2010). Freshness characterisation of whiting (Merlangius merlangus) using an SPME/GC/MS method and a statistical multivariate approach. *Journal of the Science of Food and Agriculture*, *90*(15), 2568–2575. https://doi.org/10.1002/jsfa.4122

Dumancas, G. G., Ramasahayam, S., Bello, G., Hughes, J., & Kramer, R. (2015). Chemometric regression techniques as emerging, powerful tools in genetic association studies. *TrAC Trends in Analytical Chemistry*, *74*, 79–88. https://doi.org/10.1016/j.trac.2015.05.007

Dziurkowska, E., & Wesolowski, M. (2015). Multivariate statistical analysis as a supplementary tool for interpretation of variations in salivary cortisol level in women with major depressive disorder. *The Scientific World Journal*, *2015*, 1–8. https://doi.org/10.1155/2015/987435

Fan, F. H., Ma, Q., Ge, J., Peng, Q. Y., Riley, W. W., & Tang, S. Z. (2013). Prediction of texture characteristics from extrusion food surface images using a computer vision system and artificial neural networks. *Journal of Food Engineering*, *118*(4), 426–433. https://doi.org/10.1016/j.jfoodeng.2013.04.015

Fidelis, M., Santos, J. S., Coelho, A. L. K., Rodionova, O. Y., Pomerantsev, A., & Granato, D. (2017). Authentication of juices from antioxidant and chemical perspectives: A feasibility quality control study using chemometrics. *Food Control*, *73*, 796–805. https://doi.org/10.1016/j.foodcont.2016.09.043

Funes, E., Allouche, Y., Beltrán, G., & Jiménez, A. (2015). A review: Artificial neural networks as tool for control food industry process. *Journal of Sensor Technology*, *05*(01), 28–43. https://doi.org/10.4236/jst.2015.51004

Gad, H. A., & Bouzabata, A. (2017). Application of chemometrics in quality control of Turmeric (Curcuma longa) based on ultra-violet, Fourier transform-infrared and ^{1}H NMR spectroscopy. *Food Chemistry*, *237*, 857–864. https://doi.org/10.1016/j.foodchem.2017.06.022

Gao, F., Zhou, S., Han, L., Yang, Z., & Liu, X. (2017). A novel FT-IR spectroscopic method based on lipid characteristics for qualitative and quantitative analysis of animal-derived feedstuff adulterated with ruminant ingredients. *Food Chemistry*, *237*, 342–349. https://doi.org/10.1016/j.foodchem.2017.05.011

Gondim, C. d. S., Junqueira, R. G., Souza, S. V. C. d., Ruisánchez, I., & Callao, M. P. (2017). Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies. *Food Chemistry*, *230*, 68–75. https://doi.org/10.1016/j.foodchem.2017.03.022

Granato, D., Koot, A., & van Ruth, S. M. (2015). Geographical provenancing of purple grape juices from different farming systems by proton transfer reaction mass spectrometry using supervised statistical techniques. *Journal of the Science of Food and Agriculture*, *95*(13), 2668–2677. https://doi.org/10.1002/jsfa.7001

Granato, D., Margraf, T., Brotzakis, I., Capuano, E., & van Ruth, S. M. (2015). Characterization of conventional, biodynamic, and organic purple grape juices by chemical markers, antioxidant capacity, and instrumental taste profile. *Journal of Food Science*, *80*(1), C55–C65. https://doi.org/10.1111/1750-3841.12722

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science and Technology*, *72*, 83–90.

Grunert, T., Stephan, R., Ehling-Schulz, M., & Johler, S. (2016). Fourier transform infrared spectroscopy enables rapid differentiation of fresh and frozen/thawed chicken. *Food Control*, *60*, 361–364. https://doi.org/10.1016/j.foodcont.2015.08.016

Guo, W.-L., Du, Y.-P., Zhou, Y.-C., Yang, S., Lu, J.-H., Zhao, H.-Y., . . . Teng, L.-R. (2011). At-line monitoring of key parameters of nisin fermentation by near infrared spectroscopy, chemometric modeling and model improvement. *World Journal of Microbiology and Biotechnology*, *28*(3), 993–1002. https://doi.org/10.1007/s11274-011-0897-x

Gurbanov, R., Gozen, A. G., & Severcan, F. (2018). Rapid classification of heavy metal-exposed freshwater bacteria by infrared spectroscopy coupled with chemometrics using supervised method. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *189*, 282–290. https://doi.org/10.1016/j.saa.2017.08.038

Hamdouche, Y., Meile, J. C., Nganou, D. N., Durand, N., Teyssier, C., & Montet, D. (2016). Discrimination of post-harvest coffee processing methods by microbial ecology analyses. *Food Control*, *65*, 112–120. https://doi.org/10.1016/j.foodcont.2016.01.022

Herceg, Z., Kovačević, D. B., Kljusurić, J. G., Jambrak, A. R., Zorić, Z., & Dragović-Uzelac, V. (2016). Gas phase plasma impact on phenolic compounds in pomegranate juice. *Food Chemistry*, *190*, 665–672. https://doi.org/10.1016/j.foodchem.2015.05.135

Hidalgo, B., & Goodman, M. (2013). Multivariate or multivariable regression? *American Journal of Public Health*, *103*(1), 39–40. https://doi.org/10.2105/ajph.2012.300897

Hong, E., Lee, S. Y., Jeong, J. Y., Park, J. M., Kim, B. H., Kwon, K., & Chun, H. S. (2017). Modern analytical methods for the detection of food fraud and adulteration by food category. *Journal of the Science of Food and Agriculture*, *97*(12), 3877–3896. https://doi.org/10.1002/jsfa.8364

Kanik, E. A., Orekici Temel, G., Erdogan, S., & Ersoz Kaya, I. (2013). Affected states soft independent modeling by class analogy from the relation between independent variables, number of independent variables and sample size. *Balkan Medical Journal*, *30*(1), 28–32. https://doi.org/10.5152/balkanmedj.2012.070

Karabagias, I. K., Louppis, A. P., Karabournioti, S., Kontakos, S., Papastephanou, C., & Kontominas, M. G. (2017). Characterization and geographical discrimination of commercial Citrus spp. honeys produced in different Mediterranean countries based on minerals, volatile compounds and physicochemical parameters, using chemometrics. *Food Chemistry*, *217*, 445–455. https://doi.org/10.1016/j.foodchem.2016.08.124

Karoui, R., Downey, G., & Blecker, C. (2010). Mid-infrared spectroscopy coupled with chemometrics: A tool for the analysis of intact food systems and the exploration of their molecular structure−quality relationships − A review. *Chemical Reviews*, *110*(10), 6144–6168. https://doi.org/10.1021/cr100090k

Kim, S., Kim, H., Reuhs, B. L., & Mauer, L. J. (2006). Differentiation of outer membrane proteins from Salmonellaenterica serotypes using Fourier transform infrared spectroscopy and chemometrics. *Letters in Applied Microbiology*, *42*(3), 229–234. https://doi.org/10.1111/j.1472-765X.2005.01828.x

Kim, S., Reuhs, B. L., & Mauer, L. J. (2005). Use of Fourier transform infrared spectra of crude bacterial lipopolysaccharides and chemometrics for differentiation of Salmonella enterica serotypes. *Journal of Applied Microbiology*, *99*(2), 411–417. https://doi.org/10.1111/j.1365-2672.2005.02621.x

Kuuliala, L., Abatih, E., Ioannidis, A. G., Vanderroost, M., De Meulenaer, B., Ragaert, P., & Devlieghere, F. (2018). Multivariate statistical analysis for the identification of potential seafood spoilage indicators. *Food Control*, *84*, 49–60. https://doi.org/10.1016/j.foodcont.2017.07.018

Larson-Hall, J. (2010). 2.1.1 Levels of measurement of variables in a guide to doing statistics. In *Second language research using SPSS* (pp. 33–35). New York: Taylor & Francis.

Lei, V., & Jakobsen, M. (2004). Microbiological characterization and probiotic potential of koko and koko sour water, African spontaneously fermented millet porridge and drink. *Journal of Applied Microbiology*, *96*(2), 384–397. https://doi.org/10.1046/j.1365-2672.2004.02162.x

Lohani, U. C., & Muthukumarappan, K. (2017). Modeling of continuous ultrasonication to improve total phenolic content and antioxidant activity in sorghum flour: A comparison between response surface methodology and artificial neural network. *International Journal of Food Engineering*, *13*(4). https://doi.org/10.1515/ijfe-2016-0086

Longobardi, F., Innamorato, V., Di Gioia, A., Ventrella, A., Lippolis, V., Logrieco, A. F., . . . Agostiano, A. (2017). Geographical origin discrimination of lentils (Lens culinaris Medik.) using ^{1}H NMR fingerprinting and multivariate statistical analyses. *Food Chemistry*, *237*, 743–748. https://doi.org/10.1016/j.foodchem.2017.05.159

Lucci, P., Saurina, J., & Núñez, O. (2017). Trends in LC-MS and LC-HRMS analysis and characterization of polyphenols in food. *TrAC Trends in Analytical Chemistry*, *88*, 1–24. https://doi.org/10.1016/j.trac.2016.12.006

Manning, L. (2016). Food fraud: Policy and food chain. *Current Opinion in Food Science*, *10*, 16–21. https://doi.org/10.1016/j.cofs.2016.07.001

Margraf, T., Santos, É. N. T., de Andrade, E. F., van Ruth, S. M., & Granato, D. (2016). Effects of geographical origin, variety and farming system on the chemical markers and in vitro antioxidant capacity of Brazilian purple grape juices. *Food Research International*, *82*, 145–155. https://doi.org/10.1016/j.foodres.2016.02.003

Marseglia, A., Acquotti, D., Consonni, R., Cagliani, L. R., Palla, G., & Caligiani, A. (2016). HR MAS ^{1}H NMR and chemometrics as useful tool to assess the geographical origin of cocoa beans – Comparison with HR ^{1}H NMR. *Food Research International*, *85*, 273–281. https://doi.org/10.1016/j.foodres.2016.05.001

Mikš-Krajnik, M., Yoon, Y.-J., Ukuku, D. O., & Yuk, H.-G. (2016). Volatile chemical spoilage indexes of raw Atlantic salmon (Salmo salar) stored under aerobic condition in relation to microbiological and sensory shelf lives. *Food Microbiology*, *53*, 182–191. https://doi.org/10.1016/j.fm.2015.10.001

Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., & Andersson, C. A. (1998). Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometrics and Intelligent Laboratory Systems*, *44*(1-2), 31–60. https://doi.org/10.1016/s0169-7439(98)00074-4

Muñiz-Valencia, R., Jurado, J. M., Ceballos-Magaña, S. G., Alcázar, Á., & Hernández-Díaz, J. (2014). Characterization of Mexican coffee according to mineral contents by means of multilayer perceptrons artificial neural networks. *Journal of Food Composition and Analysis*, *34*(1), 7–11. https://doi.org/10.1016/j.jfca.2014.02.003

Nagel, O. G., Molina, M. P., & Althaus, R. L. (2011). Optimization of bioassay for tetracycline detection in milk by means of chemometric techniques. *Letters in Applied Microbiology*, *52*(3), 245–252. https://doi.org/10.1111/j.1472-765X.2010.02990.x

Nagel, O. G., Molina, M. P., & Althaus, R. L. (2012). Use chemometric techniques in the optimization of a specific bioassay for betalactams in milk. *Letters in Applied Microbiology*, *54*(1), 32–38. https://doi.org/10.1111/j.1472-765X.2011.03169.x

Nagel, O. G., Zapata, M. D. L. L., Basílico, J., Bertero, J., Molina, M. P., & Althaus, R. L. (2009). Effect of chloramphenicol on a bioassay response for the detection of tetracycline residues in milk. *Journal of Food and Drug Analysis* 17(1), 36–42.

Nascimento, K., Sattler, J. A., Macedo, L. F. L., Gonzalez, C., Melo, I., Araujo, E., . . . Muradian, L. B. A. (2018). Phenolic compounds, antioxidant capacity and physicochemical properties of Brazilian Apis mellifera honeys. *LWT-Food Science and Technology*, *91*, 85–94.

Nunes, C. A., Alvarenga, V. O., de Souza Sant'Ana, A., Santos, J. S., & Granato, D. (2015). The use of statistical software in food science and technology: Advantages, limitations and misuses. *Food Research International*, *75*, 270–280. https://doi.org/10.1016/j.foodres.2015.06.011

O'Farrell, M., Lewis, E., Flanagan, C., Lyons, W., & Jackman, N. (2005a). Comparison of k-NN and neural network methods in the classification of spectral data from an optical fibre-based sensor system used for quality control in the food industry. *Sensors and Actuators B: Chemical*, *111–112*, 354–362. https://doi.org/10.1016/j.snb.2005.02.003

O'Farrell, M., Lewis, E., Flanagan, C., Lyons, W. B., & Jackman, N. (2005b). Combining principal component analysis with an artificial neural network to perform online quality assessment of food as it cooks in a large-scale industrial oven. *Sensors and Actuators B: Chemical*, *107*(1), 104–112. https://doi.org/10.1016/j.snb.2004.09.050

Obranović, M., Škevin, D., Kraljić, K., Pospišil, M., Neđeral, S., Blekić, M., & Putnik, P. (2015). Influence of climate, varieties and production process on tocopherols, plastochromanol-8 and pigments in flaxseed oil. *Food Technology and Biotechnology*, *53*, 496–504. https://10.17113/ftb.53.04.15.4252

Oliveri, P., Casale, M., Casolino, M. C., Baldo, M. A., Nizzi Grifi, F., & Forina, M. (2010). Comparison between classical and innovative class-modelling techniques for the characterisation of a PDO olive oil. *Analytical and Bioanalytical Chemistry*, *399*(6), 2105–2113. https://doi.org/10.1007/s00216-010-4377-1

Oliveri, P., & Downey, G. (2012). Multivariate class modeling for the verification of food-authenticity claims. *TrAC Trends in Analytical Chemistry*, *35*, 74–86. https://doi.org/10.1016/j.trac.2012.02.005

Oliveri, P., & Simonetti, R. (2016). Chemometrics for food authenticity applications. In G. Downey (Ed.), *Advances in food authenticity testing* (1st ed, pp. 701–728). Amsterdam: Elsevier.

Pacella, M., & Semeraro, Q. (2014). Application of neural-based algorithms as statistical tools for quality control of manufacturing processes. In D. Granato & G. Ares (Eds.), *Mathematical and statistical methods in food science and technology* (pp. 431–448). Los Angeles: Wiley & Sons.

Pardo-Mates, N., Vera, A., Barbosa, S., Hidalgo-Serrano, M., Núñez, O., Saurina, J., & . . . Puignou, L. (2017). Characterization, classification and authentication of fruit-based extracts by means of HPLC-UV chromatographic fingerprints, polyphenolic profiles and chemometric methods. *Food Chemistry*, *221*, 29–38. https://doi.org/10.1016/j.foodchem.2016.10.033

Peng, J., Liu, F., Zhou, F., Song, K., Zhang, C., Ye, L., & He, Y. (2016). Challenging applications for multi-element analysis by laser-induced breakdown spectroscopy in agriculture: A review. *TrAC Trends in Analytical Chemistry*, *85*, 260–272. https://doi.org/10.1016/j.trac.2016.08.015

Pérez-Caballero, G., Andrade, J. M., Olmos, P., Molina, Y., Jiménez, I., Durán, J. J., ... Miguel-Cruz, F. (2017). Authentication of tequilas using pattern recognition and supervised classification. *TrAC Trends in*

Analytical Chemistry, 94, 117–129. https://doi.org/10.1016/j.trac.2017.07.008

Petrakis, E. A., & Polissiou, M. G. (2017). Assessing saffron (Crocus sativus L.) adulteration with plant-derived adulterants by diffuse reflectance infrared Fourier transform spectroscopy coupled with chemometrics. *Talanta*, *162*, 558–566. https://doi.org/10.1016/j.talanta.2016.10.072

Pomerantsev, A. L. (2014). *Chemometrics in excel*. Hoboken NJ: John Wiley & Sons.

Pomerantsev, A. L., & Rodionova, O. Y. (2014). Concept and role of extreme objects in PCA/SIMCA. *Journal of Chemometrics*, *28*, 429–438.

Poojary, M. M., Putnik, P., Bursać Kovačević, D., Barba, F. J., Lorenzo, J. M., Dias, D. A., & Shpigelman, A. (2017). Stability and extraction of bioactive sulfur compounds from Allium genus processed by traditional and innovative technologies. *Journal of Food Composition and Analysis*, *61*, 28–39. https://doi.org/10.1016/j.jfca.2017.04.007

Pustjens, A. M., Weesepoel, Y., & van Ruth, S. M. (2016). Food fraud and authenticity: Emerging issues and future trends. In C. E. Leadley (Ed.), *Innovation and future trends in food manufacturing and supply chain technologies* (1st ed, pp. 3–20). Woodhead Publishing.

Putnik, P., Bursać Kovačević, D., Herceg, K., Roohinejad, S., Greiner, R., Bekhit, A. E.-D. A., & Levaj, B. (2017). Modelling the shelf-life of minimally-processed fresh-cut apples packaged in a modified atmosphere using food quality parameters. *Food Control*, *81*, 55–64. https://doi.org/10.1016/j.foodcont.2017.05.026

Putnik, P., Bursać Kovačević, D., Režek Jambrak, A., Barba, F., Cravotto, G., Binello, A., & . . . Shpigelman, A. (2017). Innovative "Green" and novel strategies for the extraction of bioactive added value compounds from citrus wastes—A review. *Molecules*, *22*(5), 680. https://doi.org/10.3390/molecules22050680

Putnik, P., Roohinejad, S., Greiner, R., Granato, D., Bekhit, A. E.-D. A., & Bursać Kovačević, D. (2017). Prediction and modeling of microbial growth in minimally processed fresh-cut apples packaged in a modified atmosphere: A review. *Food Control*, *80*, 411–419. https://doi.org/10.1016/j.foodcont.2017.05.018

Rasouli, Z., & Ghavami, R. (2016). Investigating the discrimination potential of linear and nonlinear spectral multivariate calibrations for analysis of phenolic compounds in their binary and ternary mixtures and calculation pKa values. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *165*, 191–200. https://doi.org/10.1016/j.saa.2016.04.044

Reinholds, I., Bartkevics, V., Silvis, I. C. J., van Ruth, S. M., & Esslinger, S. (2015). Analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments: A review. *Journal of Food Composition and Analysis*, *44*, 56–72. https://doi.org/10.1016/j.jfca.2015.05.004

Rendall, R., Reis, M. S., Pereira, A. C., Pestana, C., Pereira, V., & Marques, J. C. (2015). Chemometric analysis of the volatile fraction evolution of Portuguese beer under shelf storage conditions. *Chemometrics and Intelligent Laboratory Systems*, *142*, 131–142. https://doi.org/10.1016/j.chemolab.2015.01.015

Rodionova, O. Y., Oliveri, P., & Pomerantsev, A. L. (2016). Rigorous and compliant approaches to one-class classification. *Chemometrics and Intelligent Laboratory Systems*, *159*, 89–96. https://doi.org/10.1016/j.chemolab.2016.10.002

Rodionova, O. Y., Titova, A. V., & Pomerantsev, A. L. (2016). Discriminant analysis is an inappropriate method of authentication. *TrAC Trends in Analytical Chemistry*, *78*, 17–22. https://doi.org/10.1016/j.trac.2016.01.010

Rutherford, A. (2011). 3.6.3. Type 1 error rate control and analysis power. In *ANOVA and ANCOVA: A GLM Approach* (pp. 65–67). Hoboken, NJ: Wiley.

Sabir, A., Rafi, M., & Darusman, L. K. (2017). Discrimination of red and white rice bran from Indonesia using HPLC fingerprint analysis combined with chemometrics. *Food Chemistry*, *221*, 1717–1722. https://doi.org/10.1016/j.foodchem.2016.10.114

Shikha Ojha, K., Granato, D., Rajuria, G., Barba, F. J., Kerry, J. P., & Tiwari, B. K. (2018). Application of chemometrics to assess the influence of ultrasound frequency, Lactobacillus sakei culture and drying on beef jerky manufacture: Impact on amino acid profile, organic acids, texture and color. *Food Chemistry*, *239*, 544–550. https://doi.org/10.1016/j.foodchem.2017.06.124

Skov, T., Honoré, A. H., Jensen, H. M., Næs, T., & Engelsen, S. B. (2014). Chemometrics in foodomics: Handling data structures from multiple analytical platforms. *TrAC Trends in Analytical Chemistry*, *60*, 71–79. https://doi.org/10.1016/j.trac.2014.05.004

Soares, S., Amaral, J. S., Oliveira, M. B. P. P., & Mafra, I. (2017). A comprehensive review on the main honey authentication issues: Production and origin. *Comprehensive Reviews in Food Science and Food Safety*. *16*(5), 1072–1100. https://doi.org/10.1111/1541-4337.12278

Sørensen, S., Lund, K. H., Cederberg, T. L., & Ballin, N. Z. (2016). Identification of Baltic Sea salmon based on PCB and dioxin profiles. *Food Control*, *61*, 165–171. https://doi.org/10.1016/j.foodcont.2015.09.044

Souza, Jr., M. B., & Trica, D. (2013). *Introdução a Modelagem e Dinâmica para Controle de Processos* (1st. ed). Rio de Janeiro: Publit.

Spink, J., Ortega, D. L., Chen, C., & Wu, F. (2017). Food fraud prevention shifts the food risk focus to vulnerability. *Trends in Food Science & Technology*, *62*, 215–220. https://doi.org/10.1016/j.tifs.2017.02.012

Spiteri, M., Rogers, K. M., Jamin, E., Thomas, F., Guyader, S., Lees, M., & Rutledge, D. N. (2017). Combination of ^{1}H NMR and chemometrics to discriminate manuka honey from other floral honey types from Oceania. *Food Chemistry*, *217*, 766–772. https://doi.org/10.1016/j.foodchem.2016.09.027

Ståhle, L., & Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *Journal of Chemometrics*, *1*(3), 185–196. https://doi.org/10.1002/cem.1180010306

Subramanian, J., & Simon, R. (2013). Overfitting in prediction models – Is it a problem only in high dimensions? *Contemporary Clinical Trials*, *36*(2), 636–641. https://doi.org/10.1016/j.cct.2013.06.011

Suleman, T., van Vuuren, S., Sandasi, M., & Viljoen, A. M. (2015). Antimicrobial activity and chemometric modelling of South African propolis. *Journal of Applied Microbiology*, *119*(4), 981–990. https://doi.org/10.1111/jam.12906

Szymańska, E., Gerretzen, J., Engel, J., Geurts, B., Blanchet, L., & Buydens, L. M. C. (2015). Chemometrics and qualitative analysis have a vibrant relationship. *TrAC Trends in Analytical Chemistry*, *69*, 34–51. https://doi.org/10.1016/j.trac.2015.02.015

Tabachnick, B. G., & Fidell, L. S. (2007). Principal components and factor analysis. In *Using multivariate statistics* (5th ed, pp. 607–615). Needham Heights, MA: Allyn & Bacon.

Tax, D., & Duin, R. (1998). Outlier detection using classifier instability. *Lecture notes in Computer Science*, *1451*, 593–601.

Tena, N., Boix, A., & von Holst, C. (2015). Identification of botanical and geographical origin of distillers dried grains with solubles by near infrared microscopy. *Food Control*, *54*, 103–110. https://doi.org/10.1016/j.foodcont.2015.01.033

van Boekel, M., Fogliano, V., Pellegrini, N., Stanton, C., Scholz, G., Lalljie, S., & . . . Eisenbrand, G. (2010). A review on the beneficial aspects of food processing. *Molecular Nutrition & Food Research*, *54*(9), 1215–1247. https://doi.org/10.1002/mnfr.200900608

van Ruth, S. M., Huisman, W., & Luning, P. A. (2017). Food fraud vulnerability and its key factors. *Trends in Food Science & Technology*, *67*, 70–75. https://doi.org/10.1016/j.tifs.2017.06.017

Varmuza, K., & Filzmoser, P. (2009). Chemoinformatics-chemometrics-statistics. In *Introduction to multivariate statistical analysis in chemometrics* (pp. 1–26). Boca Raton, FL: CRC Press Taylor and Francis Group.

Vasconcelos, H., Saraiva, C., & de Almeida, J. M. M. M. (2014). Evaluation of the spoilage of raw chicken breast fillets using fourier transform infrared spectroscopy in tandem with chemometrics. *Food and Bioprocess Technology*, 7(8), 2330–2341. https://doi.org/10.1007/s11947-014-1277-y

Walker, G. S. (2017). Food authentication and traceability: An Asian and Australian perspective. *Food Control*, *72*, 168–172. https://doi.org/10.1016/j.foodcont.2016.01.028

Wold, M. S. (1977). SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In B. R. Kowalski (Ed.), *Chemometrics theory and application, American Chemical Society Symposium Series* (Vol. *52*, pp. 243–282). Wash., D.C.: American Chemical Society.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wu, L., Du, B., Vander Heyden, Y., Chen, L., Zhao, L., Wang, M., & Xue, X. (2017). Recent advancements in detecting sugar-based adulterants in honey – A challenge. *TrAC Trends in Analytical Chemistry*, *86*, 25–38. https://doi.org/10.1016/j.trac.2016.10.013

Xu, L., Cai, C.-B., & Deng, D.-H. (2011). Multivariate quality control solved by one-class partial least squares regression: Identification of adulterated peanut oils by mid-infrared spectroscopy. *Journal of Chemometrics*, *25*(10), 568–574. https://doi.org/10.1002/cem.1402

Yang, X.-T., Qian, J.-P., Li, J., Ji, Z.-T., Fan, B.-l., Xing, B., & Li, W.-Y. (2016). A real-time agro-food authentication and supervision system on a novel code for improving traceability credibility. *Food Control*, *66*, 17–26. https://doi.org/10.1016/j.foodcont.2016.01.032

Zhang, W., & Xue, J. (2016). Economically motivated food fraud and adulteration in China: An analysis based on 1553 media reports. *Food Control*, *67*, 192–198. https://doi.org/10.1016/j.foodcont.2016.03.004

Zielinski, A. A. F., Haminiuk, C. W. I., Nunes, C. A., Schnitzler, E., van Ruth, S. M., & Granato, D. (2014). Chemical composition, sensory properties, provenance, and bioactivity of fruit juices as assessed by chemometrics: A critical review and guideline. *Comprehensive Reviews in Food Science and Food Safety*, *13*(3), 300–316. https://doi.org/10.1111/1541-4337.12060