Talanta 83 (2010) 475–481

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Establishment of multivariate specifications for food commodities with discriminant partial least squares

Néstor F. Pérez, Ricard Boqué*, Joan Ferré

Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, C/Marcel·lí Domingo, s/n. 43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 29 April 2010
Received in revised form 1 September 2010
Accepted 25 September 2010
Available online 1 October 2010

Keywords:
Multivariate specifications
Food specifications
Discriminant partial least squares (DPLS)
Food authentication

ABSTRACT

A novel method for establishing multivariate specifications of food commodities is proposed. The specifications are established for discriminant partial least squares (DPLS) by setting limits on the predictions of the DPLS model together with Hotelling T^2 and square error of prediction (*SPE*). These limits can be tuned depending on whether type I error (i.e. a correct sample is declared out-of-specification) or type II error (i.e. an out-of-specification sample is declared within specifications) need to be minimized. The methodology is illustrated with a set of NIR spectra of Italian olive oils, corresponding to five regions and the class Liguria is the class of interest. The results demonstrate the possibility of establishing multivariate specification for olive oils from the Liguria region on the basis of spectral data obtaining type I and type II errors lower than 5%.

1. Introduction

According to the American Society for Testing and Materials (ASTM) [1], specification is “an explicit set of requirements to be satisfied by a material, product, system or service.” The document also states that “Examples of specifications include, but are not limited to requirements for: physical, mechanical, or chemical properties, and safety, quality, or performance criteria. A specification identifies the test methods for determining whether each of the requirements is satisfied”. Specifications have a large importance in engineering, manufacturing and trade, and the governments must ensure proper development or provision of services to establish the minimum requirements needed to ensure the quality and adequacy of the item or service provided. These quality requirements are of such importance that sometimes are regulated by laws or standards [2,3] and overseen by competent agencies [4].

Specifications can be derived in different ways. First, specifications can be a set of parameters or characteristics that the user defines that products must satisfy, such as the tolerances of materials. Specifications can also be derived from observations or researches, such as the minimum nutritional requirements in foods [3], or the maximum levels permitted of pesticides, the heavy metals and contaminants in general [4] and also the product specifications that refer to the constitution, origin and/or characteristics of the product (e.g. the specifications of a protected designation of origin of a food commodity).

Consumers feel product specifications as the way to evaluate whether a food has the optimal conditions for consumption (manufacturing, nutrition and health) [5]. Lately, it became necessary to include specifications that ensure the authenticity of food [6]; i.e. parameters that guarantee the origin and the production conditions (e.g. organic food) and that there are not counterfeits. This requirement is also demanded by the producers because it ensures that there is no unfair competition and because it adds value to the product. In response, the European Union has set up three mechanisms of protection: protected designation of origin (PDO), protected geographical indication (PGI) and traditional specific guaranteed [6,7]. Within this context, the primary objective of the European project “Tracing the Origin of Food (TRACE)” was to develop analytical methodologies to find a fingerprint for different food commodities and to identify counterfeit products [8].

A product specification can be either univariate or multivariate. Univariate specifications are the most commonly used and are defined by one or more individual variables (e.g. mass, length, or density). However, most specifications are multivariate by nature, that is, several variables must be measured. Also, many analytical methodologies provide the information of the product as a vector of measured variables (i.e. mass spectrum); thus, the specifications must be adapted to this multivariate context. The variables can be analyzed either separately, without taking into account the relationship between them, or using multivariate analysis, that takes into account the correlations between variables. Treating multivariate specifications as multi-univariate has been reported to lead to erroneous conclusions about the quality of the product [9]. A

* Corresponding author at: Universitat Rovira i Virgili, Analytical Chemistry and Organic Chemistry, Facultat de Química, Campus Sescelades, 43007 Tarragona, Spain. Tel.: +34 977 55 9564; fax: +34 977 55 8446.
E-mail address: ricard.boque@urv.cat (R. Boqué).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.09.037

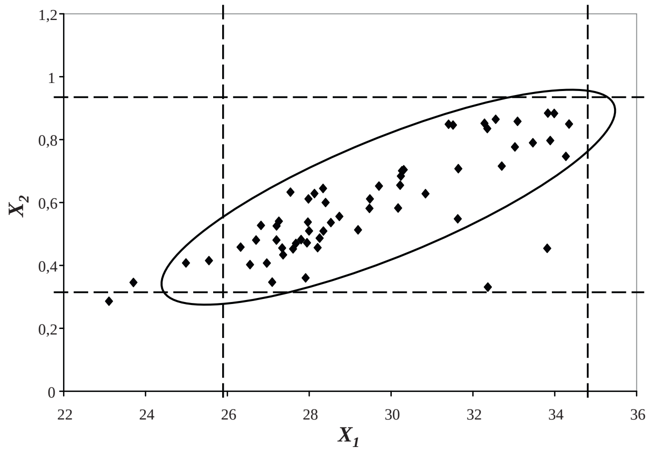

Fig. 1. Multivariate control charts for variables X_1 and X_2. Dotted lines represent the univariate specification limits, and the ellipse the bivariate specification limits.

clear example is shown in Fig. 1, where two variables, X_1 and X_2, and their corresponding univariate limits, are plotted. The limits create a rectangle that frames most of the points. However, it is more efficient if we frame the points within an ellipse, a bivariate limit. The ellipse fits better the distribution of the points, two "bad" objects are rejected, and two objects rejected by the variable X_1 are accepted. In this way, the joint analysis of the specifications can refine the specification limits and reduce the type I and type II errors. A Type I error is committed when a sample that comply with the specification is rejected, while a type II error is committed when a sample that does not meet the specification is accepted. In general, producers will require low type I errors, because they will not be satisfied if a complying product is said to be out of the specifications. On the contrary, consumers would like to be protected against out-of-specification products, and so want low type II errors. Multivariate analysis of specifications has become more important in recent years, also because of the vast amount of data generated by the analytical methods [9,10]. An example is the work by Novič and Grošelj [11], who established product specifications based on classification models with neural networks.

To verify one or more univariate specifications statistical quality control (SQC) tools are commonly used. Univariate SQC verifies if the variable is within limit values (e.g. Shewhart chart) [12]. For multivariate specifications, multivariate SQC (MSQC) tools are used, such as the Hotelling's T^2 statistics [13]. When the number of variables is large, such as those generated by spectral methods of analysis, principal component analysis (PCA) or partial least squares (PLS) regression are used to reduce the number of variables, so that the multivariate control limits or specification limits are defined using the significant PCA or PLS factors. PLS has the advantage over PCA that provides a control both on the input variables (e.g. the raw materials), and of the output variables (quality of the final product), and it has been applied in the control of chemical industrial processes [13].

Multivariate specifications will largely depend on the data analysis methods used. In this paper we present a procedure for establishing product specifications from discriminant PLS (DPLS) (PLS applied to classification) [14]. DPLS binary models are derived with the strategy "one against all" [15], thus obtaining as many models as classes are being modelled ($C = K$). For each of the models, the scores, the *x*-residuals and the predicted $\hat{y}$ of the samples of the class of interest are used to establish the boundaries of Hotelling T^2, Q (or squared prediction error (*SPE*)) statistics and also the PLS prediction $\hat{y}$. A product that meets the specification will be within the statistical limits. This procedure is illustrated with the olive oil data set, a dataset generated within the TRACE project.

2. Multivariate specifications

Unlike univariate specifications, which are used to define many of the regulations now in observance, multivariate specifications are scarcely mentioned in the literature. This contrasts with the extensive references to multivariate statistical process control (MSPC), which shares many concepts and analytical tools with multivariate specifications. Initially referred in econometrics [16], multivariate specifications were first studied and applied in the chemical industry by De Smet, Duchesne and MacGregor to ensure optimal raw materials in industrial processes [17].

To establish multivariate specifications De Smet, Duchesne and MacGregor proposed three steps: (1) acquire an adequate set of data, (2) develop the multivariate specifications, and (3) implement the multivariate specifications. When acquiring the data set, we must consider the item for which we want to establish specifications. Thus, for specifications of a final product only a single data matrix is needed. On the contrary, for specifications of raw materials to be input in a process, three data matrices may be required: properties of raw materials, process variables and properties of the finished product, because the characteristics of the finished product depend both on the variables of the process and on the raw materials. In step two, besides studying the possibility of reducing the dimensionality of the data by PCA or PLS, the possible correlations between matrices must be taken into account. Finally, the implementation requires an appropriate pretreatment of the problem item and to establish that the object meets the specification limits.

The most critical step is the second one where, in order to obtain robust specifications, it is necessary to detect and remove outliers [9,18]. Having defined the data set, we can define the region of multivariate specifications and monitor new objects with multivariate χ^2 or T^2 statistics. χ^2 or chi-square calculates the distance of a new object to the center of a data set when the covariance matrix is known. If the covariance matrix is not known and has to be estimated then the Hotelling T^2 statistic is applied [19]. The monitoring seeks that the samples are lower than the upper control limit (UCL), calculated usually at a 95% or 99% confidence level, since it is assumed that the data follow a normal distribution. Other possible methods of monitoring are the multivariate versions of EWMA and CUSUM charts [19]. However, when the number of variables is substantially increased, it is more difficult to use those monitoring methods [19]. Therefore, the dimensionality of the data set should be reduced with PCA or PLS [17]. PCA is used when only a single data matrix (e.g. quality properties of the product) is considered, so the T^2 statistic is applied to the scores of PCA, and is complemented by the Q (or *SPE*) statistic that takes into account the residuals, that is, the information not modelled by PCA [9]. When data are in different (correlated) matrices (e.g. composition of raw materials and properties of finished products) the dimensionality reduction is done with PLS, since it maximizes the covariance between the two data matrices [17]. Once the scores have been obtained we can apply the T^2 and the *SPE* statistics. Finally, the limits of the specifications are selected to produce the smallest type I and/or type II errors [17].

2.1. Multivariate specifications in DPLS

The way multivariate specifications are used depends on the multivariate analysis method used. It is therefore necessary to study the advantages and disadvantages of using DPLS [14] to define specifications. Take as an example a product with three classes ($C = 3$), e.g. production sites. In this case the set of data will consist of the data matrix **X**, with *I* objects (representing the products) and *J* measured variables, and a vector **y** that encodes the classes to which each object belongs. When developing the DPLS models the binarization strategy "one against all" (the class of interest (coded 1) is modelled against the rest (coded 0)) is used. However, to establish multivariate specifications only the scores of the I_c objects of class ω_c (class of interest) are used [15], i.e. for the model ω_1 *vs.* $\omega_2-\omega_3$, only the I_1 scores of the class ω_1 are used. Thus, the total number of models developed is equal to the number of classes ($K=C=3$).

Once the J variables have been reduced to an optimal number of significant factors A, using DPLS, we can monitor new objects using the Hotelling's T^2 [20] statistic:

$$T^2 = (\mathbf{t}-\bar{\mathbf{t}}_c)^{\mathrm{T}}\mathbf{S}_c^{-1}(\mathbf{t}-\bar{\mathbf{t}}_c) \tag{1}$$

where $\mathbf{t}$ is the vector of scores for a new object, $\bar{\mathbf{t}}_c$ and $\mathbf{S}_c$ are the mean vector and covariance matrix, respectively, for the scores of the training samples of class ω_c, these are calculated as:

$$\bar{\mathbf{t}}_c = \frac{1}{I_c}\sum_{i=1}^{I_c}\mathbf{t}_i \tag{2}$$

$$\mathbf{S}_c = \frac{1}{I_c-1}\sum_{i=1}^{I_c}(\mathbf{t}_i-\bar{\mathbf{t}}_c)(\mathbf{t}_i-\bar{\mathbf{t}}_c)^{\mathrm{T}} \tag{3}$$

where I_c is the number of objects in class ω_c, and $\mathbf{t}_i$ is the ith vector of scores for objects in class ω_c. Since Hotelling's T^2 monitoring assumes a normal distribution of the data, the upper control limit (UCL) is calculated as:

$$T^2_{\mathrm{UCL}} = \frac{(I_c-1)(I_c+1)A}{I_c(I_c-A)}F_\alpha(A, I_c-A) \tag{4}$$

where F_α (A, I_c-A) is the upper 100α% critical point of the F distribution with A and I_c-A degrees of freedom [13].

Another statistic used to monitor new objects is the squared prediction error (*SPE*), or *Q*. *SPE* provides, in the space of the original variables, the squared difference between the actual and predicted values:

$$SPE = \sum_{J=1}^{J}(x_j-\hat{x}_j)^2 \tag{5}$$

The upper control limits for *SPE* are based on a χ^2 distribution approximation:

$$SPE_{\mathrm{UCL}} = \left(\frac{v}{2m}\right)\chi^2_{1-\alpha}\left(\frac{2m^2}{v}\right) \tag{6}$$

where v and m are the variance and mean value, respectively, of the *SPE* values of the training objects and $\chi^2_{1-\alpha}$ is a weighted chi-square distribution ($g\chi^2_h$) with the weight g and h degrees of freedom [21].

In addition, different from other methods, in DPLS the predictions for the class of interest (ideally values around 1) can be used to complement the T^2 and *SPE* statistics. For the predictions of the class of interest it is necessary to establish a lower and an upper limit. Since the predictions are not necessarily normally distributed, the percentiles from the distribution of objects of the class of interest are used. Thus, the limits at a given confidence level are established from the percentage of objects that are within those limits (for example 95% of the data for a confidence level of 95%). Thus, the multivariate specification is defined by three limits: T^2, *SPE* and $\hat{y}$. The object is within specifications if it fulfils these three requirements simultaneously. That is:

$$\begin{gathered} T_i^2 < \lim T^2_{\mathrm{UCL},\alpha} \\ \text{and} \\ SPE_i < \lim SPE_{\mathrm{UCL},\alpha} \\ \text{and} \\ \lim_{\mathrm{low},\alpha}\hat{y} < \hat{y}_i < \lim_{\mathrm{up},\alpha}\hat{y} \end{gathered} \tag{7}$$

Although the commonly accepted limits in MSQC are built for α values of 5% or 1%; when defining multivariate specifications these limits must be based on the behaviour of the training data, i.e. we must find a limit that allows a balance between type I and type II errors. This makes it necessary to optimize the limits of T^2, *SPE* and $\hat{y}$.

To apply the specifications to new objects a series of steps must be followed. First, the object must be pretreated in the same way as the training objects (i.e. log transformed, mean centered, autoscaled, and so on). Second, the scores, the x-residuals, and the $\hat{y}$ predicted have to be calculated; and third, Hotelling T^2, *SPE* and $\hat{y}$ statistics have to be monitored to verify that the new objects are within the product specifications.

3. Experimental

3.1. Data set

The establishment of multivariate specifications is illustrated with the data set olive oil [8], which contains 166 Italian olive oils belonging to 5 different regions: Liguria (63), Sicilia (28), Lazio (29), Puglia (28) and Umbria (18). 700 variables were measured, corresponding to the values of absorbance in the near infrared region, measured between 1100 and 2498 nm, every 2 nm.

3.2. Procedure and software

The Kennard–Stone algorithm [22] applied to each class separately was used to split the data set into a training set (with 70% of the objects) and a test set (with 30% of the objects). The olive oil data were mean-centered and transformed into their first and second derivative before the DPLS models were calculated. The procedure is illustrated with the oils from the Liguria class, but it can be extrapolated to other classes. The DPLS models are developed as the class of interest against the other classes, that is, Liguria *vs.* Sicilia, Lazio, Puglia and Umbria. The optimal number of factors was selected by leave-one-out cross validation (LOOCV), and three criteria were tested: minimum type I error, minimum type II error and minimum overall error.

All calculations were done using in-house made Matlab (The Math Works, Inc) subroutines.

4. Results and discussion

4.1. Multivariate specifications. Italian olive oil data set

Fig. 2 shows the (mean centered) training data set. No important differences were observed among the objects of the class Liguria (solid line) and the rest (dotted line), except for the objects Liguria025 and Umbria184 that have the most extreme values.

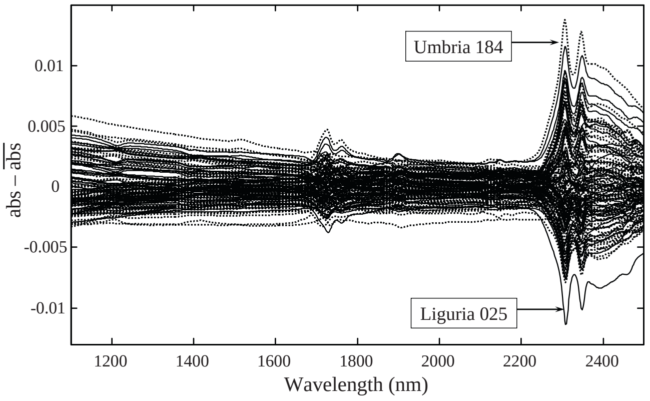

Fig. 2. Olive oil data: mean-centered training set. Liguria oils (solid lines) and non Liguria oils (dotted lines).

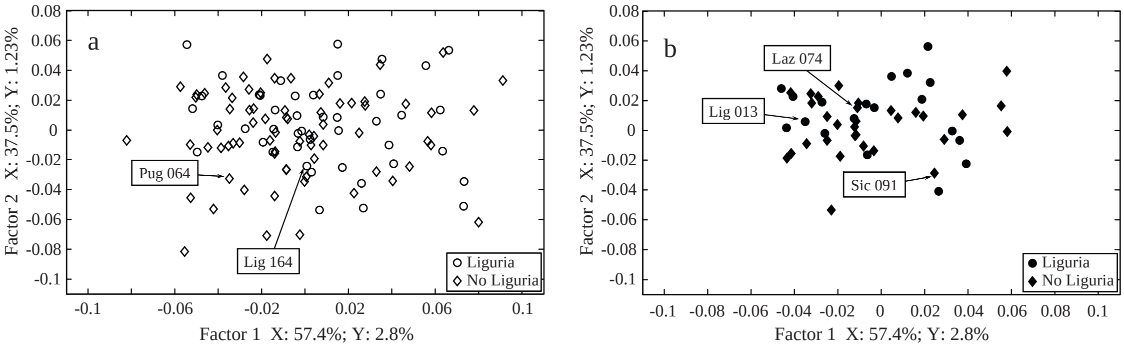

Fig. 3. DPLS scores plots for the first two factors. (a) Training set and (b) test set.

Fig. 3 shows the scores for the first two factors of the DPLS model Liguria *vs*. others (94.6% of cumulative variance explained in **X** and 4% in **y**). At a first glance there is no clear separation between Liguria class and non Liguria class groups and the objects are evenly spread in the factor space. As it is to be expected from Fig. 2, objects Liguria025 and Umbria184 have the most extreme scores. These objects, however, were not removed from the dataset since the distance to the other objects was not appreciably large.

After the DPLS had been calculated, the specifications for the class Liguria were established by defining limits for Hotelling T^2, *SPE* and $\hat{y}$ at a confidence level of 99%. Fig. 4 shows the values of Hotelling T^2 *vs*. $\hat{y}$ and Fig. 5 shows the values of *SPE vs*. $\hat{y}$, for the DPLS model with 15 factors and mean-centered data. By first considering $\hat{y}$, all the Liguria training objects are within the two $\hat{y}$ limits (0% of type I error). However, four non Liguria objects are also within the limits, equivalent to a 5.5% of type II error. This indicates that the specification of the Liguria class can be almost defined by the two limits of the $\hat{y}$ of the PLS model. By considering the limit value of the Hotelling T^2 statistic only, all Liguria oils fall below that specification limit (0% type I error) but many non Liguria oils are also below the limit (high type II error). The type II error is even higher, almost 100%, for *SPE* (Fig. 5), since most non Liguria oils are below the *SPE* limit. This is because DPLS, unlike PCA, models the two classes simultaneously and when an object is predicted, even if it is not Liguria, it has a low *SPE*. Note that $\hat{y}$ statistic enabled rejection of one non-Ligurian object that was inside the Hotelling T^2 limit. When the limits for Hotelling T^2, *SPE* and $\hat{y}$ are considered together, the type I error was 11.4% and the type II error was 9.7%.

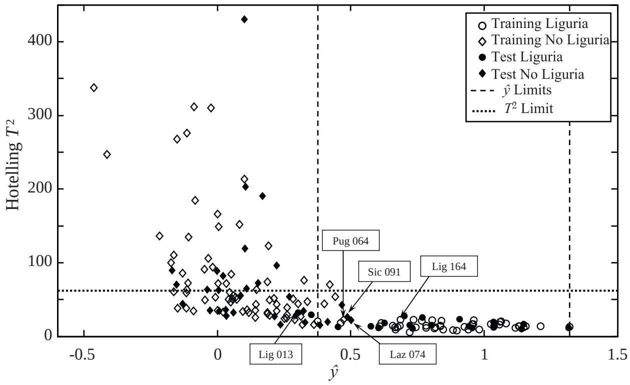

Fig. 4. Hotelling T^2 *vs*. $\hat{y}$ for training and test objects, showing the limits with 15 factors and confidence level of 99%. Some of the objects that have a particular behaviour are indicated.

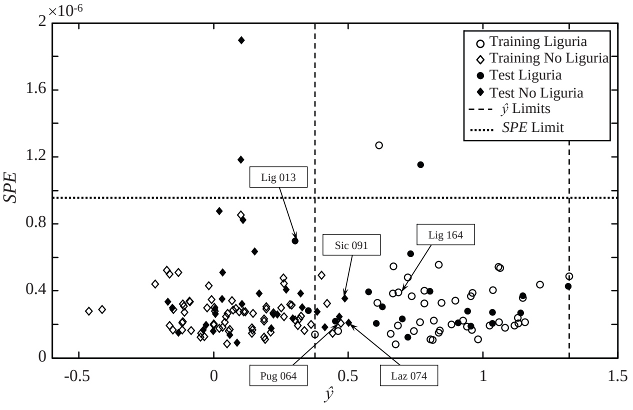

Fig. 5. *SPE vs*. $\hat{y}$ for the training and test objects, showing the limits with 15 factors and confidence level of 99%. Some of the objects that have a particular behaviour are indicated.

The type I and type II errors depend on the number of factors in the PLS model. For specifications that protect the producer, the optimal number of factors is the one that minimizes the type I error. To protect the consumer, the optimal number of factors is the one that minimizes the type II error. Fig. 6a shows the variation of type I error with respect to number of factors (data mean-centered) when the theoretical limits of Hotelling T^2, *SPE* and $\hat{y}$ are set at a confidence level of 95%. By considering Hotelling T^2 only, the calculated cross-validation type I error is almost constant at 5% for any number of factors, which agrees with the theoretical value. Even for the model with 17 factors the type I error for this statistic can be as low as 0% although this is likely an overfitted model (Fig. 6a). For the *SPE* statistic only, the theoretical 5% type I error is obtained only for the model with 2 factors, and the error increases when more factors are included in the model. The reason for this increase is that the residuals of the training samples decrease when so does the number of factors, thus making the limit of the *SPE* statistic lower. This makes that the *SPE* value for the cross-validated samples more easily exceeds the *SPE* limit, so more samples are rejected. For the predicted $\hat{y}$, the models from 1 to 5 factors maintain the type I error around the theoretical value of 5% and the error increases for models with more than 5 factors, for similar reasons than for the *SPE* described above (Fig. 6a). Since the objects that are declared out-of-specification by each statistic are not necessarily the same, the type I error of the combined use of the three statics is almost the sum of the type I errors of each statistic. Hence, a 13.6% type I error is obtained for the models with 1–4 factors, and then increases the more factors are added to the model. A global type I error of 5% can be obtained by increasing the confidence level of each statistic to 99%. Fig. 6b shows the variation of type I error with respect to number of factors, when the limits of the statistics are set to a theoretical confidence level of 99%. In general, the type I error for Hotelling T^2, *SPE* and $\hat{y}$ statistics decrease up to 0% in some cases. The combined type I error is 4.5% for the PLS model with 3 factors, close to the desired 5% for setting the specification.

To establish specifications that protect the consumer the type II error should be minimized. Fig. 7 shows the variation of type II error with the number of factors. In general, a large number of factors are required to obtain type II errors lower than 10%. For example, for a confidence limit of 95%, a model with 22 factors is required for the Hotelling T^2 statistic and a model with 15 factors for the $\hat{y}$. The *SPE* statistic required 36 factors, although this can be considered casual because the *SPE* of all objects (both Liguria and non Liguria) was so large that they all were considered out of specification. By combining the three statistics, 15 factors are required in the model (Fig. 7a). Moreover, when the confidence level is increased from 95% to 99%, more factors are required to have minimal type II error, 14 factors with 95% against 16 for 99% (Fig. 7). This behaviour is opposite to what it was observed with the type I error. Note that the limits on $\hat{y}$ have the largest effect in reducing the type II error. For example, for the model with 14 factors, the combined use of only T^2 and *SPE* statistics gives a type II error of 37.5%, while combining T^2, *SPE* and $\hat{y}$ the error decreases down to 2.8% (Fig. 7).

Given the behaviour observed with type I and II errors, the specification limits for individual statistics (Hotelling T^2, *SPE* and $\hat{y}$) should be established at a confidence level of 99% in order to obtain a combined type I error of 5%. The same procedure was applied to the data processed with the first and second derivative with similar results. It confirms the need for a confidence limit of 99% in the individual statistics, so that by combining the statistics the type I error is kept below 5%.

The specification limits for Liguria class with mean-centered data and first and second derivatives are then:

Data mean-centered

a) For specifications that protect the producer:

$$\text{Object } i \text{ is within specification if } T_i^2 < 13.8 \text{ and } SPE_i < 3.30 \times 10^{-4} \text{ and } -0.054 < \hat{y}_i < 1.02, \quad \text{for a DPLS model with 3 factors.} \tag{8}$$

b) For specifications that protect the consumer:

$$\text{Object } i \text{ is within specification if } T_i^2 < 68.2 \text{ and } SPE_i < 6.52 \times 10^{-7} \text{ and } 0.51 < \hat{y}_i < 1.31, \quad \text{for a PLS model with 16 factors.} \tag{9}$$

c) For specifications with balanced error:

$$\text{Object } i \text{ is within specification if } T_i^2 < 62.0 \text{ and } SPE_i < 9.57 \times 10^{-7} \text{ and } 0.38 < \hat{y}_i < 1.32, \quad \text{for a PLS model with 15 factors.} \tag{10}$$

Data with first derivative

a) For specifications that protect the producer:

$$\text{Object } i \text{ is within specification if } T_i^2 < 13.8 \text{ and } SPE_i < 3.71 \times 10^{-7} \text{ and } 0.018 < \hat{y}_i < 1.20, \quad \text{for a DPLS model with 3 factors.} \tag{11}$$

b) For specifications that protect the consumer:

$$\text{Object } i \text{ is within specification if } T_i^2 < 37.4 \text{ and } SPE_i < 4.51 \times 10^{-8} \text{ and } 0.61 < \hat{y}_i < 1.32, \quad \text{for a PLS model with 10 factors.} \tag{12}$$

c) For specifications with balanced error:

$$\text{Object } i \text{ is within specification if } T_i^2 < 33.5 \text{ and } SPE_i < 6.36 \times 10^{-8} \text{ and } 0.47 < \hat{y}_i < 1.27, \quad \text{for a PLS model with 9 factors.} \tag{13}$$

Data with second derivative

a) For specifications that protect the producer:

$$\text{Object } i \text{ is within specification if } T_i^2 < 16.8 \text{ and } SPE_i < 8.12 \times 10^{-8} \text{ and } 0.32 < \hat{y}_i < 1.14, \quad \text{for a DPLS model with 4 factors.} \tag{14}$$

b) For specifications that protect the consumer:

$$\text{Object } i \text{ is within specification if } T_i^2 < 29.8 \text{ and } SPE_i < 4.97 \times 10^{-8} \text{ and } 0.78 < \hat{y}_i < 1.22, \quad \text{for a PLS model with 8 factors.} \tag{15}$$

c) For specifications with balanced error:

$$\text{Object } i \text{ is within specification if } T_i^2 < 19.9 \text{ and } SPE_i < 7.47 \times 10^{-8} \text{ and } 0.36 < \hat{y}_i < 1.24, \quad \text{for a PLS model with 5 factors.} \tag{16}$$

Also note that the training objects Liguria164 and Puglia 064 were found out of specification and within specification, respectively, in the cross-validation step with data mean-centered. However, in the calibration step, which defines the specifications, both objects are within specifications. The object Liguria164 is near the limit of 99% in the three specifications types. The object has a $T^2 = 51.9$, $SPE = 6.71 \times 10^{-7}$ and $\hat{y} = 0.380$. For the error balanced specification, $\hat{y}$ is out of the limits (limit of 0.412), while the T^2 and *SPE* values are within the limits ($T^2{}_{\text{limit}} = 63.4$ and $SPE_{\text{limit}} = 8.73 \times 10^{-7}$). Since we require the object to simultaneously satisfy the three limits, the object is declared out of specifications. In contrast, the object Puglia064 is far from the three specification types. For example, for error balanced specifications the object has a T^2 of 18.6 (limit of 62.0), a *SPE* of 2.61×10^{-7} (limit of 9.14×10^{-7}) and a $\hat{y}$ of 0.631 against lower and upper limits of 0.370 and 1.28. Since the object is within the three boundaries it is considered to meet specifications.

The defined specifications were checked against a test set. Table 1 shows the percentages of classification for test objects when the limits are set with a theoretical confidence level of 99% with data mean-centered and transformed with the first and second derivative. For the specification that protects the producer and the specification that protects the consumer the results are good, with errors lower than for the training set. For example, for specifications that protect the producer, the type I error was 0%, against 4.5% (mean-centered) and 6.8% (first and second derivative) for the training set. For specifications that protect the consumer, the type II error was 0%, against 2.8% (mean-centered) and 1.4% (first and second derivative) for the training set. On the contrary, with the specifications for balanced error with mean-centered data, the test data produces larger errors than the training data (type I error of 11.4% and type II of 9.7%). In comparison, for first and second derivative data the type I error is smaller and the type II error is slightly larger. Thus we consider that the specifications defined for the class Liguria are appropriate, although the balanced-error specification has a higher type I and II error because the objects have to simultaneously comply with the three boundaries.

Three test objects (mean-centered data) require a particular analysis: the object Liguria013 that is out of specification, and the objects Sicilia091 and Lazio 074, that are within specifications. For the error-balanced specifications, Liguria013 is within the lim-

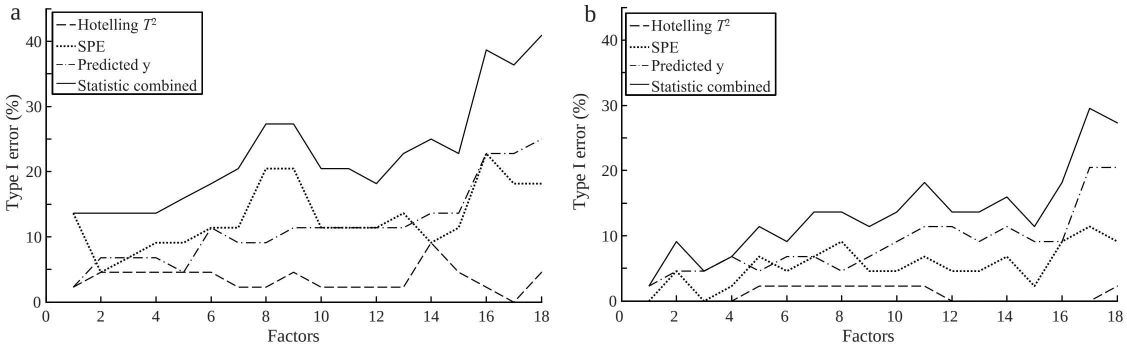

Fig. 6. Variation of the type I error of the hotelling T^2, *SPE* and $\hat{y}$ statistics and the three statistics combined, assuming a confidence level of 95% (a) and 99% (b).

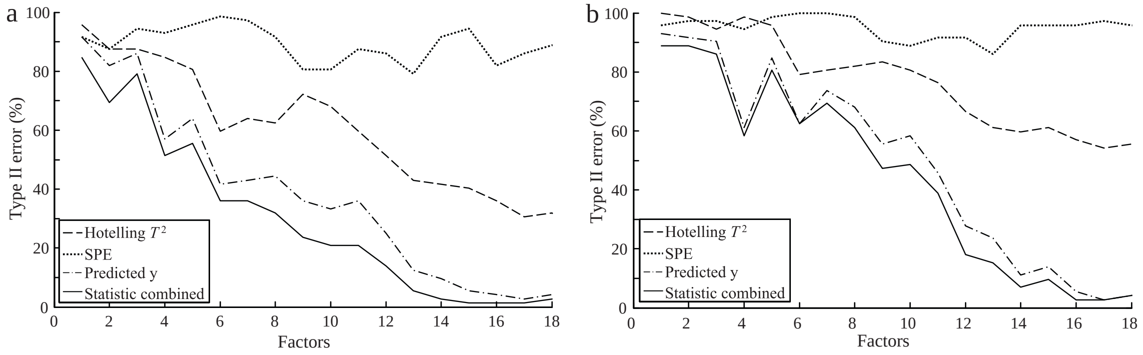

Fig. 7. Variation of the type II error of the Hotelling T^2, *SPE* and $\hat{y}$ statistics and the three statistics combined, assuming a confidence level of 95% (a) and 99% (b).

Table 1
Error performance from Liguria olive oil test data set, specification limits with confidence level of 99%.

Data preprocessing	Type of specification	Optimal factors	Type I error	Type II error
Mean-centered	Protect producer	3	0%	87.1%
	Protect consumer	16	26.3%	0%
	Balanced error	15	15.8%	16.1%
1st derivative	Protect producer	3	0%	83.9%
	Protect consumer	10	15.8%	0%
	Balanced error	9	0%	6.5%
2nd derivative	Protect producer	4	0%	19.4%
	Protect consumer	8	36.8%	0%
	Balanced error	5	0%	12.9%

its of T^2 (Fig. 5) and *SPE* (Fig. 6) but outside of (although close to) the limits of $\hat{y}$ (Fig. 6). This also occurs for the specification that protects the consumer, so the object is finally declared out of specification in the two cases. Objects Sicilia091 and Lazio074 are within the three limits both of the specification that protects the producer and the balanced-error specification, so they are within specifications.

5. Conclusions

Multivariate specifications based on T^2, *SPE* and predicted $\hat{y}$ have been established for the NIR spectra of olive oils from the Liguria region. Adding limits on $\hat{y}$, together with the commonly used T^2 and *SPE* statistics, improves the definition of the specification and reduces the number of factors needed in the DPLS model. This optimal number of factors depends on the type of specification (either specification that protect the producer, specification that protect the consumer or specification that provides a balance of type I and type II errors). Note that, in order to reach a general confidence level close to 95%, type I and II error of 5%, the individual confidence levels for T^2 and *SPE* and $\hat{y}$ had to be set to 99%.

Acknowledgments

The authors express their gratitude to Dr. Gerard Downey (Teagasc, Ireland) for providing the olive oil dataset. This work was supported by the project TRACE – "TRAcing food Commodities in Europe (EU IP 006942)" – from the Sixth Framework Programme of the European and by the Spanish Ministerio de Educación y Ciencia project CTQ2007-66918/BQU. This paper reflects only the authors' views and neither the Community nor the Spanish Ministerio are liable for any use that may be made of the information contained therein.

References

[1] http://www.astm.org/COMMIT/Regs.pdf (25/11/2009). ASTM International, "Regulations Governing ASTM Technical Committees", October 2009.

[2] D.L. Flumignan, G.C. Anaia, F.de O. Ferreira, A.G. Tininis, J.E. de Oliveira, Chromatographia 65 (9/10) (2007) 617–623.

[3] J.M. Betz, K.D. Fisher, L.G. Saldanha, P.M. Coates, Anal. Bioanal. Chem. 389 (2007) 19–25.

[4] M.L. Weiner, W.F. Salminen, P.R. Larson, R.A. Barter, J.L. Kranetz, G.S. Simon, Food Chem. Toxicol. 39 (2001) 759–786.

[5] M.R. Hubbard, Statistical Quality Control for the Food Industry, third edition, Kluwer Academic/Plenum Plublishers, USA, 2003, pp. 253–276.

[6] S. Kelly, K. Heaton, J. Hoogewerff, Trends Food Sci. Technol. 16 (2005) 555–567.

[7] http://ec.europa.eu/agriculture/foodqual/quali1_en.htm (last accessed in 21st of January 2010).

[8] Project TRACE – "TRAcing food Commodities in Europe" (project no. FOOD-CT-2005-006942). www.trace.eu.org.

[9] L.H. Chiang, L.F. Colegrove, Chemometr. Intell. Lab. Syst. 88 (2007) 143–153.

[10] D.M. Ennis, J. Bi, J. Food Qual. 23 (2000) 541–552.

[11] M. Novič, N. Grošelj, Anal. Chim. Acta 649 (2009) 68–74.

[12] D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, third edition, John Wiley & Sons, Inc., USA, 2003, pp. 595–648.

[13] T. Kourti, J.F. MacGregor, Chemometr. Intell. Lab. Syst. 28 (1995) 3–21.

[14] B.R. Kowalski, Chemometrics, Mathematics and Statistics in Chemistry, D. Reidel Publishing Company, Dordrecht, Holland, 1984, pp. 85–88.

[15] G. Ou, Y.L. Murphey, Pattern Recogn. 40 (2007) 4–18.

[16] M. Guidolin, A. Timmermann, J. Econometrics 131 (2006), 285-208.

[17] C. Duchesne, J.F. MacGregor, J. Qual. Technol. 36 (1) (2004) 78–94.

[18] M.-J. Bruwer, J.F. MacGregor, W.M. Bourg Jr., Food Qual. Prefer. 18 (2007) 890–900.

[19] C. Wikström, C. Albano, L. Eriksson, H. Fridén, E. Johansson, Å. Nordahl, S. Rännar, M. Sandberg, N. Kettaneh-Wold, S. Wold, Chemometr. Intell. Lab. Syst. 42 (1998) 221–231.

[20] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, fifth edition, Prentice Hall, USA, 2002, pp. 210–252.

[21] A. Nijhuis, S. de Jong, B.G.M. Vandeginste, Chemometr. Intell. Lab. Syst. 38 (1997) 51–62.

[22] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.