Determining performance parameters in qualitative multivariate methods using probability of detection (POD) curves. Case study: two common milk adulterants

Authors. Carina de Souza Gondim[a,b,c], Roberto Gonçalves Junqueira[a], Scheilla Vitorino Carvalho de Souza[a], M. Pilar Callao[c]*, Itziar Ruisánchez[c,]

Affiliations

[a] Department of Food Science, Faculty of Pharmacy (FAFAR), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG Brazil.

[b] CAPES Foundation, Ministry of Education of Brazil, 70040-020, Brasília, DF, Brazil.

[c] Chemometrics, Qualimetric and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007, Tarragona, Spain.

*Corresponding author: mariapilar.callao@urv.cat

Abstract

A strategy for determining performance parameters of two–class multivariate qualitative methods was proposed. As case study, multivariate classification methods based on mid-infrared (MIR) spectroscopy coupled with the soft independent modelling of class analogy (SIMCA) technique for detection of hydrogen peroxide and formaldehyde in milk were developed.

From the outputs (positive/negative/inconclusive) of the samples, which were unadulterated and adulterated at target value, the main performance parameters were obtained. Sensitivity and specificity values for the unadulterated and adulterated classes were satisfactory. Inconclusive ratios 12 and 21%, respectively, for hydrogen peroxide and formaldehyde were obtained.

To evaluate the performance parameters related to concentration, Probability of Detection (POD) curves were established, estimating the decision limit, the capacity of detection and the unreliability region. When inconclusive outputs were obtained, two additional concentration limits

were defined: the decision limit with inconclusive outputs and the detection capability with inconclusive outputs.

The POD curves showed that for concentrations below 3.7 g L^{-1} of hydrogen peroxide and close to zero of formaldehyde, the chance of giving a positive output (adulterated sample) was lower than 5%. For concentrations at or above 11.3 g L^{-1} of hydrogen peroxide and 10 mg L^{-1} of formaldehyde, the probability of giving a negative output was also lower than 5%.

Keywords: POD curves, multivariate qualitative analysis, milk adulteration, performance parameters, SIMCA, method validation

1 Introduction

The International Union of Pure and Applied Chemistry (IUPAC) stated that qualitative analysis is "the analysis in which substances are identified or classified on the basis of their chemical or physical properties" [1]. Nowadays, qualitative analysis is not only used to identify compounds but has been extended to classify and authenticate foodstuffs, products, specimens and materials [2,3]. This expansion is largely due to the introduction of chemometric techniques, which allowed multivariate data to be included, modeled and used for qualitative analysis [2].

Multivariate qualitative methods have been developed and applied in such fields as clinical, medicine, biology and chemistry [3-6]. In food field, analytical (mainly spectroscopy) and classification techniques have been successfully combined for food fraud detection, both adulteration [7-11] and authentication [11-15]. In this context, the analytical problems can require non-quantifiable or semi-quantifiable information: i.e. to authenticate a substance/product or to verify if a substance is present above or below a pre-established threshold concentration level (cut-off value). In these cases, the use of qualitative methods, which provide a binary response (positive/negative), might be suitable. These methods have been commonly used in systems that require immediate decisions and represent an appealing alternative to quantitative analysis,

which generally gives more but often unnecessary sample information and requires a greater investment of money and/or time.

The type of binary response required by qualitative analysis (yes/no, belongs/does not belong, etc.) can be obtained by applying a classification technique. The classification techniques can be divided into two main blocks: discriminant techniques that aims to divide data space up into separate regions, each of which corresponds to one class and class-modelling techniques which models each class independently. In discriminant techniques, new samples are always assigned to the closest class. So, the detection of outlier sample is difficult. In modelling techniques, samples can be classified into a just one class, in more than one or in no one, being able to obtain an ambiguous or inconclusive result [16].

Method validation is a fundamental process for quality assurance since it ensures that the method is reliable and fit for purpose [17-19]. For qualitative methods, a guide related to measures to monitor certain substances and residues in live animals and animal products is established in the Commission Decision CD/657/EC 2002 [20]. Nevertheless, this document has been interpreted ambiguously, which has led to some confusion in the terminology [21]. Today, there is considerable consensus about the definition of such performance parameters as sensitivity and specificity [22] but no agreement has been reached about other related indexes [23]. In 2013, AOAC International published a guideline for the validation of binary qualitative methods for detecting biological and chemical compounds [24]. It gives some guidance on the statistical treatment of the binary results so that additional performance parameters related to concentration (decision limit, detection capability and the unreliability region) can be obtained. This guideline have been adopted by some authors [17,18]. Macarthur and Holst [25] proposed the use of a plot of prediction intervals for the estimation of the probability of detection, providing some parameters related to concentration. Gondim et al. [26] developed a systematized procedure to validate qualitative methods, also including parameters related to concentration.

Probability of Detection (POD) curve is a tool that represents graphically the positive outputs of the established qualitative method with respect to concentration [25-28]. POD curve harmonizes

the statistical concepts and parameters between quantitative and qualitative method validation. In the literature, this curve is referred by various other names, such as Performance Characteristic Curves (PCC), mainly used in the quality characterization of univariate screening methodologies [29,30]. However, considering multivariate screening validation, PCC have seldom been used [17].

Milk adulteration is reported in recent literature [31-34], despite legislation and inspection by official agencies [35-38]. One of the most common types of milk fraud is the addition of compounds to reduce or maintain microbial counts, masking failures in good manufacturing practices. Formaldehyde and hydrogen peroxide are examples of such compounds. Aiming to detect this type of fraud, the development of rapid and non destructive methodologies based on spectroscopic techniques coupled with chemometrics has been reported [39-42].

The goal of the present study was to propose a strategy to assess the performance of a multivariate qualitative method and to generate information for decision-making in cases of milk adulteration, considering these two common adulterants - formaldehyde and hydrogen peroxide. The strategy was divided into two steps. In the first, a two-class SIMCA model was established. In the second, the main performance parameters from the positive/negative SIMCA outputs were estimated; samples adulterated at different concentration levels of the adulterant were analyzed and predicted in the SIMCA model; and the performance parameters related to concentration were estimated by fitting a POD curve with the positive SIMCA outputs.

2 Qualitative analysis for presence/absence of a target analyte

A qualitative binary method gives two possible (binary) outcomes: i.e. either a positive or negative output [24]. A binary response reflects a categorical property of the sample (compliant/non-compliant) or the concentration of a target analyte in a sample (presence/absence). In the latter case, positive and negative outputs are directly related to a

threshold concentration value generally imposed by regulation or clients. In absence of any other information, it is set by the analysts on the basis of their knowledge of the analytical problem.

In a multivariate qualitative method, the binary response is provided by multiple variables and treated mainly by multivariate classification techniques. Therefore, in a multivariate qualitative method for detecting a target analyte, two classes have to be modelled (two-class strategy): class 1 defined with samples without the analyte and class 2 defined with samples containing the analyte at the threshold concentration. The binary response is obtained for each pre-defined class. For instance, in the prediction step in the class 2 model, a positive output should be obtained for samples containing the target analyte at the threshold concentration while samples without the analyte should give a negative output. In the class 1 model, the opposite should occur.

Figure 1 shows the proposed two-class strategy scheme for assessing a multivariate qualitative method. In the first step, a two-class SIMCA model is constructed (sections 2.1 and 2.2). In the second step the main performance parameters are estimated from the positive/negative outputs obtained for each class of the SIMCA model (see section 2.3); samples at different concentration levels of the analyte are analyzed and predicted in the SIMCA model; and the performance parameters related to concentrations are estimated by fitting a POD curve (section 2.4).

Figure 1

2.1 Multivariate analysis and classification rules

SIMCA is a modelling technique based on Principal Component Analysis (PCA) where each class is modelled independently from all others [43]. For each sample, the Hotelling T^2 and Q statistics are calculated, which measure, respectively, the information of each sample within the SIMCA model and the amount of original information not included in the model [44].

Class frontiers (T^2_{lim} and Q_{lim}) are calculated for each pre-defined class (class model), at a specific significance level (α), usually set at 0.05. For the sake of simplicity, samples are assigned by means of the reduced statistical values (Hotelling T^2_r and Q_r) which are the ratio between the statistic for sample *i* and the class limit. A sample must have values lower than 1 for both reduced statistics to be considered "within the class model".

In addition to these criteria, another parameter of interest that can be used if a sample is assigned to more than one class is its distance from each class because if it is clearly closer to one class than another, it has a greater probability of belonging to it. The distance of a sample *i* from class *j* (d_{ij}) is a combination of its reduced statistic expressed as equation 1:

$$d_{ij} = (Q_{r,i}^2 + T^2_{r,i\ 2})^{1/2} \qquad \text{(eq. 1)}$$

The distance can be transformed into a probability value of belonging to each class.

2.2 Outputs of two-class models

When a sample is submitted to a qualitative method, there are four possible output scenarios,: true positive (TP), when the qualitative method gives a positive output for a sample that is positive; false positive (FP), when the qualitative method gives a positive output for a sample that is negative; true negative (TN), when the qualitative method gives a negative output for sample that is negative; false negative (FN), when the qualitative method gives a negative output for a sample that is positive [17,29].

As has been stated above, with a multi-class strategy, the four possible outputs are obtained for each pre-defined class. In addition to these well-known outputs, inconclusive output can be obtained. A sample is considered inconclusive when it cannot be undoubtedly assigned to one class (model): that is to say when it is either not assigned to a class or it is assigned to more than one class (both classes in a two-class strategy).

2.3 Main performance parameters

From the multivariate classification outputs, the main performance parameters are defined and calculated as follows:

Sensitivity is the ability of the classification model to recognize its own samples. The sensitivity of class *j* is estimated by considering only samples that belong to that class (equation 2):

$$Sensitivity_j = TP_j / n^oS_j \quad \text{(eq. 2)}$$

Where *j* indicates the class under study, TP_j means true positives (samples from class *j* that have been properly predicted by the model as belonging to class *j*), and n^oS_j the total number of samples that really belong to class *j*. Therefore, sensitivity indicates the likelihood of recognizing truly positive samples.

Specificity is the ability of the classification model to distinguish external samples and the specificity for class *j* is estimated by considering only samples that do not belong to that class (equation 3):

$$Specificity_j = TN_j / n^oS_{not\ j} \quad \text{(eq. 3)}$$

~~w~~Where *TN* means true negatives (samples that are not from class *j* and have been predicted as not belonging to class *j*), and $n^oS_{not\ j}$ means the total number of samples that really do not belong to class *j*. Therefore, specificity indicates the likelihood of recognizing samples that are truly different from the class.

The inconclusive ratio indicates the percentage of samples that cannot be undoubtedly assigned to one class, (equation 4):

$$Inconclusive\ ratio_j = (NA_j + MA) / n^oS_j \quad \text{(eq. 4)}$$

Where NA_j means unassigned samples (samples that are from class *j* that are not assigned to class *j* or to any other class); *MA* means multiple assignation samples (samples from class *j*

assigned to more than one class) and n^oS_j means the total number of samples that really belong to class *j*.

2.4 Performance parameters related to concentration: POD curves

Probability of detection (POD) is defined as “the proportion of positive outcomes (P(x)) for a qualitative method for a given matrix at a given analyte level or concentration” [24]. The POD curve graphically describes the relation between the probability of positive results and the analyte concentration . The probability values were estimated from independent measurements at each level of adulteration.

In an ideal qualitative binary method, the method is expected to give a very high percentage of positive outputs (P(x)=100%) when the analyte concentration *x* is equal to or higher than the threshold value. Similarly, the method is expected to give negative outputs (P(x)=0% or N(x)=100%) when the analyte concentration is lower than the threshold value [45~~36~~]. However, in real/experimental situations, behavior is not ideal and the P(x) or N(x) cannot be immediately changed from 0 to 1 at a single concentration level (at the threshold or cut-off). Figure 2 (solid line) shows an example of a POD curve presenting a sigmoidal shape. This curve is obtained by fitting the experimental P(x) values to a non-linear function using empirical models [2~~7~~9].

Figure 2

Due to the fact that there is no way to define in advance which is the best equation for a specific problem under study, a visual inspection of the experimental values plot can provide an indication of the equation type (e.g. sigmoid, exponential or polynomial) to be used to fit the data. This previous analysis is important since a nonlinear regression program cannot find the best mathematical model through a set of data points. It can only optimize the parameters once the equation model has been specified [46,47].

As in linear regression, the nonlinear least square procedure is used to determine the equation parameters that minimize the sum of the squares of the distances of the data points to the curve (residuals) [46,48-50]. However, contrary to linear regression, in nonlinear regression, an exact estimation of the parameters, in most cases, does not exist and cannot be calculated in one step. Equation parameters are determined iteratively from an initial estimate of the value of each parameter by applying numerical methods [46].

In the case of non-linear equations where many parameters (usually between two and four) have to be estimated, initial values need to be selected for each parameter so as not to go in a direction that the regression converges to a wrong solution (local minimum) or never converges. The initial values should be specified on the basis of previous experience, preliminary analyses, or just on a hunch [46].

The goodness parameters of the non-linear fit are evaluated by means of the root mean square of errors (RMSE) (equation 5), and the adjusted coefficient of determination (R^2_{adj}) which measures how well the calculated curve fits the original data (equation 6) [48]:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-m}} \qquad \text{(Eq. 5)}$$

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \cdot (n-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot (n-m)} \qquad \text{(Eq. 6)}$$

Where *i* is the number of the observation; y_i, is the value of *y* for a given *i*; $\hat{y}$, the predicted value of *y* for a given *i*; $\bar{y}$, the mean of values of *y*; *n*, the number of observations; and *m,* the number of equation parameters.

To obtain the concentration limits two horizontal lines have to be drawn representing the probabilities of committing false positive and negative errors. The upper line is usually set at P(x)=95% (or 0.95) and the lower line at P(x)=5% (or 0.05). The intersection of these horizontal lines with the POD curves (Figure 2, solid line) shows the performance parameters related to concentrations [17,18,30,45,51-53]:

- The CCα (decision limit) is obtained from the intersection between the lower horizontal line and the POD curve. CCα is the concentration limit at which the qualitative method detects the contaminant (it is present) with a 5% of error of stating that the contaminant is present when in fact it is not (false positive);

- The CCβ (detection capability) is obtained from the intersection between the upper horizontal line and the POD curve. It is the smallest amount of a substance that can be reliably detected, identified and/or quantified in a sample with a statistical certainty of 1-β (1-false negative);

- The unreliability region is between the two limits where there is the probability of false positive errors.

In the context of multivariate methods, when inconclusive outputs, I(x), are obtained, POD curves can also be fitted to take them into account. Therefore, another curve is fitted for the P(x)+I(x) probabilities versus the concentration of the target analyte (Figure 2, dashed line) and additional limits (CCα_I and CCβ_I) can be defined. CCβ_I is the concentration level at which there is a 5% chance of true negative outputs or a 95% chance of positive plus inconclusive outputs. When this limit is set, the unreliability region is divided into two sections. The section between CCα and CCβ_I is the concentration region in which negative, positive and inconclusive outputs are obtained. The section between CCβ_I and CCβ is the concentration region in which the probability of obtaining a negative output is equal to or lower than 5%. A practical implication is described in section 4 (Results and discussion). The interval between CCα_I and CCα has no practical implications.

3 Materials and methods

3.1 Sampling and sample preparation

A total of 30 raw milk samples were collected from the Professor Hélio Barbosa Experimental Farm of the Veterinary School of the Federal University of Minas Gerais State (EV/UFMG).

These samples were obtained from ten animals with different production characteristics by hand milking in three different weeks to ensure a representative sampling, covering a lot of the possible variability factors. Samples were kept in polypropylene bottles at refrigerated temperatures (4 to 7°C) before preparation and analysis.

Raw milk samples were used as unadulterated samples. To obtain the adulterated samples, a known quantity of aqueous standard solution of each adulterant was individually added to an aliquot of each raw milk sample in a polypropylene microtube. The adulterants solutions were freshly prepared at concentrations in which the amount of water added to unadulterated samples was ≤ 2% (w/v). Water is naturally present in milk at levels between 85 and 90% (w/v) and the detection of adulteration with water at levels between 0.5 and 10% (w/v) by mid-infrared spectroscopy was reported as not precise [39]. Thus, the addition of ≤ 2% (w/v) of water was not considered a critical factor for the samples classification.

Samples were adulterated with formaldehyde at six adulteration levels between 7.4 and 74 mg L^{-1} with 30 independent replicates for each concentration level (totaling 180 samples). Samples were adulterated with hydrogen peroxide at five concentration levels between 1 and 21 g L^{-1} also with 30 independent replicates for level (totaling 150 samples). The selected levels of adulteration corresponded to those reported as detectable by classical qualitative methods [54,55].

The microtubes were sealed, homogenized manually by inversion and kept refrigerated (4 to 7 °C) until MIR analysis.

3.2 Reagents, equipment and software

The reagents were of appropriate analytical grade. Formaldehyde and hydrogen peroxide were obtained from Synth, Diadema, SP, Brazil. Sodium carbonate and sodium citrate were supplied by Dinâmica Química Contemporânea Ltda., Diadema, SP, Brazil and Alphatec, São Bernardo

do Campo, SP, Brazil, respectively. The water used in adulterant solutions was obtained from the Direct-Q® 3 UV water purification system, Billerica, MA, USA.

The equipment (volumetric flasks, analytical balances, thermometers and spectrometer) used to prepare and analyze the samples were calibrated by external laboratories accredited according to ISO/IEC 17025 to ensure measurement traceability to the International System of Units (SI). The samples were analyzed using the spectrometer Spectrum One - FTIR Spectrometer, Perkin Elmer, Waltham, MA, USA, in attenuated total reflectance (ATR) mode, with a ZnSe crystal.

The spectra collected were the average of four scans, in the MIR region from 4000 to 650 cm^{-1}. The ATR accessory was cleaned and a background correction performed between each analysis. For all samples, the absorbance values for frequencies below 950 cm^{-1} were very high. Thus, the region between 950 and 650 cm^{-1} was excluded before the data analysis, being the final number of variables 3051.

The spectra were submitted to Multiplicative Scatter Correction (MSC) [56] pre-processing followed by mean centering. The spectra data were analyzed using MATLAB software, version R2010a (7.10.0.499), The MathWorks, Natick, MA, USA, with the PLS Toolbox, version 7.0.2, Eigenvector Technologies, Manson, WA, USA. MATLAB software version R2012b (The MathWorks, Natick, MA, USA) was used for curve fitting by nonlinear regression. Action software (Estatcamp, Campinas, Brazil) was employed in residual statistical analysis.

4 Results and discussion

A processed spectrum of unadulterated sample and the corresponding spectra of the samples adulterated with hydrogen peroxide and formaldehyde were randomly selected and illustrated in Figure 3. It was observed that the spectra were quite simple and similar. The main absorption peak was identified between 3400 and 3000 cm^{-1}, corresponding to the O-H stretching region, which is related to the presence of water in milk samples. A second informative region was observed between 1700 and 1500 cm^{-1}, representing the O-H bending also related to the water

content in milk samples. The absorption peak around 1700 cm^{-1} was attributed to amide I and II and related to protein content and lipid interactions [39,57]. Although slight differences were observed between the three spectra, there were not specific bands in the MIR spectra for the two investigated compounds, which it was expected due to the lack of selectivity of the technique. Then, it was decided to work with the entire spectrum.

Figure 3

Two multivariate classification methods (based on two-class SIMCA models) were established for each one of the adulterants considered. Due to the limited number of samples, the leave-on-out cross validation method was employed for the construction of the models. According Foca et al. [58], cross-validation can lead to highly over-optimistic estimates of the performance of the final model. However, when a small number of samples are available, segregate the whole data set into training and test sets could implies data sets with an insufficient number of data, which does not provide a sufficiently representative estimate of the predictive capability.

In both methods, class 1 was built with 29 corrected MIR spectra of unadulterated samples. This class model retained the first six PCs, which explained 98.3% of the total variance. The class 2 SIMCA model developed for hydrogen peroxide was built with 57 spectra of adulterated samples at the target value concentrations (15-21 g L^{-1}). This class model retained the first three PCs, which explained 90.8% of the total variance. The class 2 SIMCA model developed for formaldehyde was built with 60 spectra of adulterated samples at the target value concentrations (~~0.0~~56-~~0.0~~74 mg L^{-1}). This class model retained the first seven PCs, which explained 98.3% of the total variance.

Table 1 shows the main performance parameters for each multivariate method. All three main performance parameters were quite satisfactory for the hydrogen peroxide method. Only 12% of

the samples analyzed needed to be submitted to a confirmatory analysis: just one sample was classified in both classes, and the remaining samples did not fit any model (9 out of 86).

Table 1

The main performance parameters for the formaldehyde method were slightly poor for the unadulterated class but quite satisfactory for the adulterated class (Table 1). With this method, 21% of the analyzed samples needed to be submitted to a confirmatory analysis: 13 out of 89 samples were classified in both classes, and 6 were not assigned.

The fact that, in the worst case, just 21% of samples had to be submitted to a confirmatory analysis was quite satisfactory for a screening strategy. It should be borne in mind that the experimental cost of the proposed methodology (MIR+SIMCA) is lower than that of the classical determination of adulterants in milk.

The next step in the proposed assessment methodology consisted of estimating the performance parameters related to the concentration of adulterant by fitting the POD curve. Independent samples at several concentrations of adulterant (six levels for formaldehyde and five for hydrogen peroxide) were predicted by the SIMCA models and the percentages of true positive, TP or P(x) and inconclusive ratio were calculated.

Figure 4 shows the POD fitted curves. For each adulterant, two curves were drawn, one for the percentage of samples classified as adulterated, positive output, P(x), and the second for samples classified as positive plus inconclusive, P(x)+I(x). Users of this approach may decide by the use of POD curves with or without inconclusive data, depending on the fitness for purpose, if they want to be more stringent in type I or type II error control.

As indicated in section 2.4, the type of non-linear models for each case was established by the visual inspection of the plots, considering the goodness of fit parameters RMSE and R^2_{adj}. The

equations of the fitted curves with their respective statistics for each adulterant are presented in Table 2.

Figure 4

Table 2

For hydrogen peroxide, the P(x) POD curve was sigmoidal while the P(x)+I(x) was exponential. Although the shape was different, there was a gradual change in both POD curves with the concentration of adulterant. Fixing the probabilities of committing false positive and negative errors (α and β) at 0.05 (usual value), two out of the four concentration limits (section 2.5) were obtained from the equations presented in Table 2. $CC\alpha$ was set at 3.7 g L^{-1}, which means that when a sample was adulterated at that concentration level, there was a 5% probability of detecting the sample as adulterated. For concentrations below 3.7 g L^{-1}, the chance of giving a positive output (adulterated sample) was lower than 5%. Concentrations between $CC\alpha$ and $CC\beta_I$ (set as 11.3 g L^{-1}) made up the unreliability region where most of the errors or inconclusive outputs were obtained. For samples adulterated at or above $CC\beta_I$ (11.3 g L^{-1}), the classification method gave true positive outputs with a probability of at least 88% while the percentage of inconclusive outputs was equal to or lower than 12% (from 12% to 7%).

Because of the experimental P(x) and I(x) outputs, for the hydrogen peroxide method it was not possible to set the $CC\alpha_I$ and $CC\beta$ limits. This has no practical implications, as the unreliability region was set between 3.7 and 11.3 g L^{-1}, and their two corresponding concentration limits ($CC\alpha$ and $CC\beta_I$) were established. Thus, the classification method was characterized.

Both of the POD curves fitted for the formaldehyde method were exponential, but behaved differently as the concentration changed. The P(x) POD curve changed gradually with the concentration of adulterant, while the P(x)+I(x) POD curve changed abruptly from 0 to 7.4 mg L^{-1} and then remained practically constant. Fixing the probabilities of committing false positive and

negative errors (α and β) at 0.05 (usual value), two out of the four concentration limits (section 2.5) were obtained. CCα had a very low value (almost zero) which is characteristic of the P(x) POD curves that are exponential.

CCβ_I was also set at a very low concentration value of 0.01 g L^{-1}. From a practical point of view, samples adulterated around 10 mg L^{-1} will be submitted to a confirmatory analysis in most cases (60%) and in 35% of the cases it will be detected as adulterated (positive output). A narrow unreliability region was defined between zero (CCα) and 10 mg L^{-1} (CCβ_I), where the most probable outputs were inconclusive (60% or higher) and the probability of obtaining a negative output varied between 5 and 40% (for the limits of the unreliability region).

As observed for the hydrogen peroxide model, for the formaldehyde model CCβ could not be obtained by setting the β error at 0.05 since the sensitivity of the class 2 (adulterated class model) was lower than 95%. In this case, the analyst can establish additional concentration limits from the POD curves for the problem under study.

For instance, a very informative concentration limit was the concentration at which the probability of obtaining a negative output was zero. This point, CC_N, was obtained from the intersection of the P(x)+I(x) POD curve with the horizontal line set at one (or 100%) and was 17.5 g L^{-1} and 57 mg L^{-1}, respectively, for hydrogen peroxide and formaldehyde methods. This means that samples adulterated at this concentration gave either positive (85% of the cases) or inconclusive outputs (15% of the cases), but no negative outputs. The analyst might also consider another interesting limit: the concentration at which there was a 50% probability of having a positive output, called, by some authors, as the cut-off of the method. For the formaldehyde method the cut-off was set at 20 mg L^{-1}, which means that for samples adulterated at concentrations higher than 20 mg L^{-1}, the probability of having a positive output was higher than 50%.

Similarly, depending on the interest of the analyst and on the problem under study, additional limits can be set for the hydrogen peroxide method.

Conclusions

An strategy to estimate the performance parameters for a multivariate qualitative method based on MIR spectroscopy coupled with the SIMCA technique has been proposed. As a case study, detection of food fraud on raw milk with two common milk adulterants÷, hydrogen peroxide and formaldehyde, was evaluated.

The main performance parameters such as sensitivity, specificity and inconclusive ratio were estimated for each multivariate method. Both methods were implemented as a screening strategy where a small number of samples that had been inconclusively assigned needed to be subject to a confirmatory analysis: 12% of the samples adulterated with hydrogen peroxide and 21% for samples adulterated with formaldehyde, both at their respective threshold concentrations.

POD curves have proved to be a useful tool for setting the performance parameters related to the concentration of adulterant: the decision limit ($CC\alpha$) and the detection capability ($CC\beta$) of the method. Although there is no need to do so according to the regulations, when inconclusive outputs were obtained, two additional concentration limits can be defined: the decision limit with inconclusive ($CC\alpha_I$) and the detection capability with inconclusive ($CC\beta_I$). In the case study presented, having more concentration levels would have allowed to obtain more accurate adjusted curves.

The shape of the POD curve determines how many performance parameters (of the four mentioned) can be estimated as well as their practical significance. For POD curves with a sigmoidal shape, all four limits were estimated. For exponential POD curves, $CC\alpha$ values could not be estimated in all cases, and when possible it provided a very low concentration value (almost zero).

Once the POD curves were fitted, the analyst could establish additional concentration limits for the problem under study. For instance, the concentration limit at which the probability of

obtaining a negative output was ~~is~~ zero or the concentration at which there was 50% probability of obtaining a positive output.

The strategy proposed can be highly efficient, particularly if numerous samples need to be analyzed, because it considerably reduces experimentation time and there is a very low risk of error. It can also be applied to other signals, samples and/or adulterants.

Acknowledgments

The authors acknowledge the CAPES Foundation for fellowships (Proc. nº: 99999.006216/2015-03), the experimental farm 'Professor Hélio Barbosa' of the Veterinary School/ Federal University of Minas Gerais State (EV/UFMG) for providing the raw milk samples and the 'Laboratório de Medicamentos' of the Ezequiel Dias Foundation (FUNED) for allowing the use of their facilities and the analytical assistance.

References

[1] L.A. Currie, Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC recommendations 1995), Pure App. Chem., 67 (1995) 1699-1723.

[2] E. Szymanska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet, L.M.C. Buydens, Chemometrics and qualitative analysis have a vibrant relationship, Trends Anal. Chem., 69 (2015) 34-51.

[3] B.L. Milman, Chemical Qualitative Analysis II, in B.L. Milman (Ed.), Chemical Identification and Its Quality Assurance, Springer, 2011, pp. 235-253.

[4] N.N. Chidumayo, K. Yoshii, N. Saasa, M. Sakai, H. Kariwa, Development of a tick-borne encephalitis serodiagnostic ELISA using recombinant Fc-antigen fusion proteins, Diagn. Microbiol. Infect. Dis., 78 (2014) 373-378.

[5] S. Tomatsu, T. Fujii, M. Fukushi, T. Oguma, T. Shimada, M. Maeda, K. Kida,Y. Shibata, H. Futatsumori, A.M. Montaño, R.W. Mason, S. Yamaguchi, Y. Suzuki, T. Orii, Newborn screening and diagnosis of mucopolysaccharidoses., Mol. Genet. Metab., 110 (2013) 42-53.

[6] G.M. Reisfield, B.A. Goldberger, R.L. Bertholf, “False-positive” and “false-negative” test results in clinical urine drug testing, Bioanalysis, 1 (2009) 937-952.

[7] C.V. Di Anibal, M. Odena, I. Ruisánchez, M.P. Callao, Determining the adulteration of spices with Sudan I-II-II-IV dyes by UV-visible spectroscopy and multivariate classification techniques, Talanta, 79 (2009) 887-892.

[8] M. Maia, A.I.R.N.A. Barros, F.M. Nunes, A novel, direct, reagent-free method for the detection of beeswax adulteration by single-reflection attenuated total reflectance mid-infrared spectroscopy, Talanta, 107 (2013) 74-80.

[9] C.V. Di Anibal, I. Ruisánchez, M.P. Callao, High-resolution 1H nuclear Magnetic Resonance spectrometry combined with chemometric treatment to identify adulteration of culinary spices with Sudan dyes, Food Chem., 124 (2011) 1139-1145.

[10] D. Cozzolino, The role of vibrational spectroscopy as a tool to asses economically motived fraud and counterfeit issues in agricultural products and foods, Anal. Methods, 7 (2015) 9390-9400

[11] I. Reinholds, V. Bartkevics, I.C.J. Silvis, S. M. van Ruth, S. Esslinger, Analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments: A review, J. Food Comp. Anal., 44 (2015) 56-72

[12] Q. Zhang, C. Liu, Z. Sun, X. Hu, Q. Shen, J. Wu, Authentication of edible vegetable oils adulterated with used frying oil by Fourier Transform Infrared Spectroscopy, Food Chem., 132 (2012) 1607-1613.

[13] D. Cozzolino, Recent trends on the use of infrared spectroscopy to trace and authenticate natural and agricultural food products, Appl. Spectrosc. Rev., 47 (2012) 518-530.

[14] M. Bevilacqua, R. Bucci, A.D. Magri, A.L. Magri, F. Marini, Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: a case study, Anal. Chim. Acta, 717 (2012) 39-51.

[15] E. Ferrari, G. Foca, M. Vignali, L. Tassi, A. Ulrici, Adulteration of the anthocyanin content of red wines: Perspectives for authentication by Fourier Transform-Near InfraRed and 1H NMR spectroscopies, Anal. Chim. Acta, 701 (2011) 139-151.

[16] O.Y. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, Trends Anal. Chem, 78 (2016) 17-22.

[17] M.I. Lopéz, M.P. Callao, I. Ruisánchez, A tutorial on the validation of qualitative methods: from the univariate to the multivariate approach, Anal. Chim. Acta, 891 (2015) 62-72.

[18] L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, Trends Anal. Chem., 80 (2016) 612-624.

[19] M. Alewijn, H. van der Voet, S. van Ruth, validation of multivariate classification method using analytical fingerprints-concept and case study on organic feed for laying hens, J. Food Comp. Anal., 51 (2016) 15-23

[20] Comision Regulation (EC) No. 2002/657/ EC of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, Off. J. Eur. Communities, (2002) 8-36.

[21] R. Galarini, R. Buratti, L. Fioroni, L. Contiero, F. Lega, Development, validation and data quality assurance of screening methods: a case study, Anal. Chim. Acta, 700 (2011) 2-10.

[22] P. Feldsine, C. Abeyta, W.H. Andrews, Collaborative study for method validation, J. AOAC Int., 5 (2002) 1187–1200.

[23] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, Trends Anal. Chem., 35 (2012) 74–86.

[24] AOAC International, Guidelines for Validation of Qualitative Binary Chemistry Methods, Version 12.3 (2013).

[25] R. Macarthur, C. von Holst, A protocol for the validation of qualitative methods of detection, Anal. Methods, 4 (2012) 2744-2754.

[26] C.S. Gondim, O.A.M. Coelho, R.L. Alvarenga, R.G. Junqueira, S.V.C. Souza, An appropriate and systematized procedure for validating qualitative methods: Its application in the detection of sulfonamide residues in raw milk, Anal. Chim. Acta, 830 (2014) 11-22.

[27] H. Schneeweiß, P. Wilrich, A simple method of constructing a confidence interval for the mean probability of detection in collaborative studies of binary measurement methods, Accred. Qual. Assur., 19 (2014) 221-3.

[28] S. Uhlig, K. Frost, B. Colson, K. Simon, D. Mäde, R. Reiting, P. Gowik, L. Grohmann, Validation of qualitative PCR methods on the basis of mathematical–statistical modelling of the probability of detection, Accred. Qual. Assur., 20 (2015) 75-83.

[29] S.L.R. Ellison, T. Fearn, Characterising the performance of qualitative analytical methods: statistics and terminology, Trends Anal. Chem., 24 (2005) 468-476.

[30] E. Trullols, I. Ruisánchez, F.X. Rius, Validation of qualitative methods of analysis that use control samples, Trends Anal. Chem., 23 (2004) 137-145.

[31] Brasil - Portaria número 1 de 7 de outubro de 1981, http://www.agricultura.gov.br, (accessed on April 2015).

[32] Brasil - Instrução Normativa número 68 de 12 de dezembro de 2006, http://www.agricultura.gov.br, (accessed on April 2015)

[33] China - State Council of the People's Republic of China. Regulation on the Supervision e Administration of the Quality e Safety of Dairy Products - Order of the State Council of the People's Republic of China (No. 536), 9 October. 2008.

[34] India-Ministry of Health and Family Welfare of India, Manual of Methods of Analysis of Foods-milk & Milk Products , New Delhi, 2012.

[35] A. Afzal, M. S. Mahmood, I. Hussain, M. Akhtar, Adulteration and microbiological quality of milk (a review), Pak. J. Nutr., 10 (2011) 1195–1202.

[36] S.S. Souza, A G. Cruz, E H M. Walter, J.J.F. Faria, R.M.S. Celeghini, M.M.C. Ferreira, D.Granato, A.S. Santana, Monitoring the authenticity of Brazilian UHT milk: a chemometric approach, Food Chem., 124 (2011) 692-695.

[37] M. Kamal, R. Karouri, Analytical methods coupled with chemometric tools for determining the authenticity and detecting the adulteration of dairy products: a review, Trends Food Sci. Tech., 46 (2015) 27-48.

[38] C.E. Handford, K. Campbell, C.T. Elliot, Impacts of milk fraud on food safety and nutrition with special emphasis on developing countries, Compr. Rev. Food Sci. F., 15 (2016) 130-142.

[39] B.G. Botelho, N. Reis, L.R. Oliveira, M.M. Sena, Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA, Food Chem., 181 (2015) 31-37.

[40] C.S. Gondim, R.G. Junqueira, S.V.C. Souza, M.P. Callao, I. Ruisánchez, Detection of several common adulterants in raw milk by MID Infrared spectroscopy and one-class and multiclass multivariate strategies, Food Chem., in press.

[41] P.M. Ssantos, E.R. Pereira-Filho, L.E. Rodriguez Saona, Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis Food Chem. 138 (2013) 19-24.

[42] E. Capuano, R. Boerrigter-Eenling, A. Koot, S.M. van Ruth, targeted and untargeted detection of skim milk powder adulteration by near-infrared spectroscopy, Food Anal. Methods 8 (2015) 2125-34

[43] M. Bevilacqua, R. Bucci, A.D. Magri, A.L. Magri, R. Nescatelli, F. Marini, Classification and Class-Modelling, in F. Marini (Ed.), Data Handling in Science and Technology: Chemometrics in Food Chemistry 28, Elsevier B.V., Amsterdam, 2013, pp. 171-233.

[44] A. Rius, M.P. Callao, F.X. Rius, Multivariate statistical process-control applied to sulfate determination by sequential injection-analysis, Analyst, 122 (1997) 737-741.

[45] R. Song, P.C. Schlecht, K. Ashley, Field screening test methods: performance criteria and performance characteristics, J. Hazard. Mater., 83 (2001) 29-39.

[46] H.J. Motulsky, L.A. Ransnas, Fitting curves to data using nonlinear regression: a practical and nonmathematical review, FASEB Journal, 1 (1987) 365-374.

[47] J. Mazucheli, J.A. Achcar, Algumas considerações em regressão não linear, Acta Sci., 6 (2002) 1761-1770.

[48] N.R Draper, H Smith, Applied Regression Analysis, third ed., Wiley, New York, 1998.

[49] H.J. Motulsky, A. Christopoulus, Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting, GraphPad Software Inc., San Diego, California, 2003.

[50] S.V.C. Souza, R.G. Junqueira, A procedure to assess linearity by ordinary least squares method, Anal. Chim. Acta, 552 (2005) 25-35.

[51] A. Rios, D. Barceló, L. Buydens, S. Cárdenas, K. Heydorn, B. Karlberg, K. Klemm, B. Lendl, B. Milman, B. Neidhart, R.W. Stephany, A. Townshend, A. Zschunke, M. Valcárcel, Quality assurance of qualitative analysis in the framework of the European project ‘MEQUALAN’, Accred. Qual. Assur., 8 (2003) 68-77.

[52] B.M. Simonet, A. Rios, M. Valcárcel, Unreliability of screening methods, Anal. Chim. Acta, 516 (2004) 67-74.

[53] A. Rios, H. Téllez, Reliability of binary analytical responses, Trends Anal. Chem., 24 (2005) 509-515.

[54] L.C.C.D. Silva, R. Tamanini, J.R. Pereira, E.A. Rios, J.R. Ribeiro Junior, V. Beloti, Preservatives and neutralizing substances in milk: analytical sensitivity of official specific and nonspecific tests, microbial inhibition effect, and residue persistence in milk, Ciênc. Rural, 45 (2015) 1613-1618.

[55] C.S. Gondim, Métodos qualitativos para detecção de adultereantes em leite cru: validação intra e interlaboratorial. (PhD Thesis), Universidade Federal de Minas Gerais, 2016.

[56] M.R. Maleki, A.M. Mouazen, H. Ramon, J. De Baerdemaeker, Multiplicative scatter correction during on-line measurement with near infrared spectroscopy, Biosystems Engineering, 96 (2007) 427-433.

[57] A. Kohler, N.K. Afseth, K. Jørgensen, Å Randby, H. Martens, Quality analysis of milk by vibrational spectroscopy, in: Applications of Vibrational Spectroscopy in Food Science, John Wiley and Sons, Ltd., Chichester, 2010.

[58] G. Foca, M. Cochi, M.L. Vigni, R. Caramanico, M. Corbellini, A. Ulrici, Different feature selection strategies in the wavelet domain applied to NIR-based quality classification models of bread wheat flours, Chemometr. Intell. Lab., 99 (2009) 91-100.

Figure captions

Figure 1. Multivariate qualitative method: establishment of classification model and assessment process

Figure 2. Probability of detection (POD) curves with sigmoid shape and respective concentration limits

(----------) POD curve for positive outputs, (……) POD curve for positive plus inconclusive outputs.

Figure 3. Unprocessed raw milk spectra: (a) unadulterated; (b) adulterated with hydrogen peroxide at 21 g L^{-1}; (c) adulterated with formaldehyde at 74 mg L^{-1}.

Figure 4. Probability of detection (POD) curves and complementary performance parameters.

• positive outputs, P(x); ○ positive plus inconclusive outputs, P(x)+I(x); (---------) P(x) POD curve, (……) P(x)+I(x) POD curve.

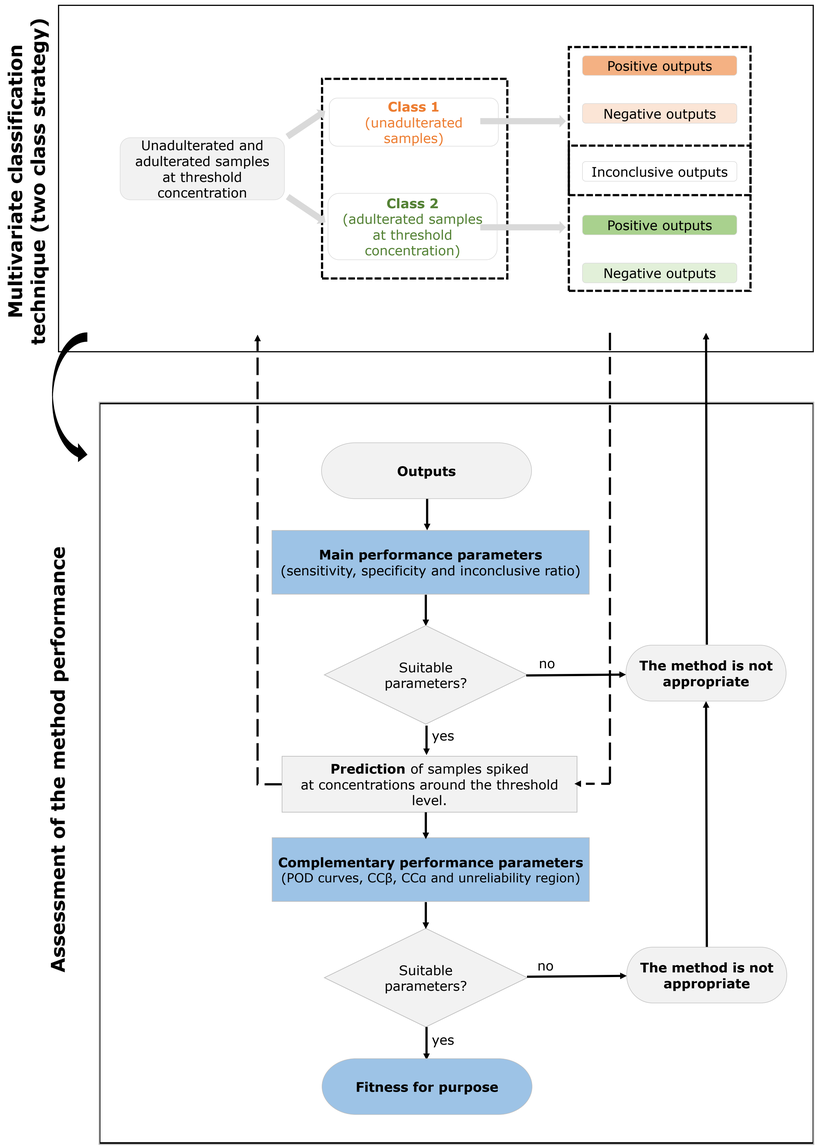
Table 1. Main performance parameters of the two-class SIMCA methods

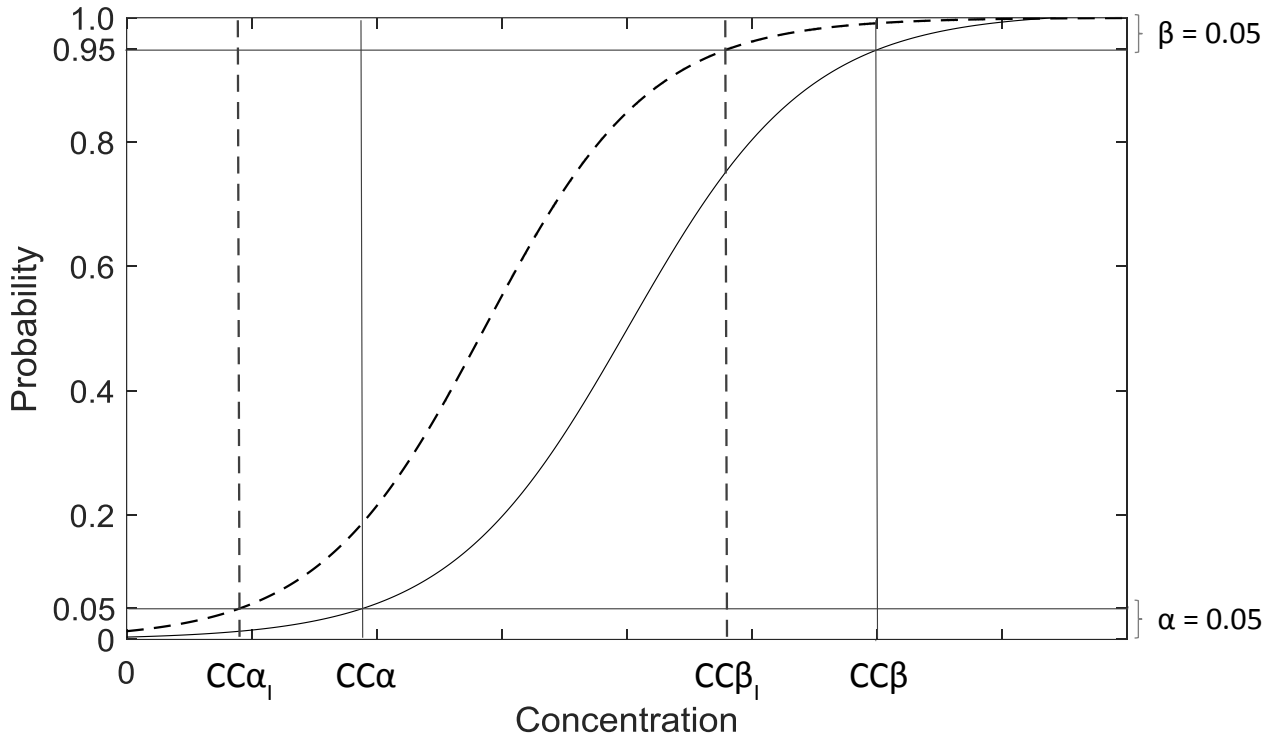
Parameter	Hydrogen peroxide method		Formaldehyde method	
	Unadulterated class	**Adulterated class**	**Unadulterated class**	**Adulterated class**
Sensitivity	0.90	0.88	0.62	0.85
Specificity	1.00	1.00	1.00	0.97
False negative rate	0.0	0.0	0.03	0.0
False positive rate	0.0	0.0	0.0	0.03
Inconclusive ratio	0.12		0.21	

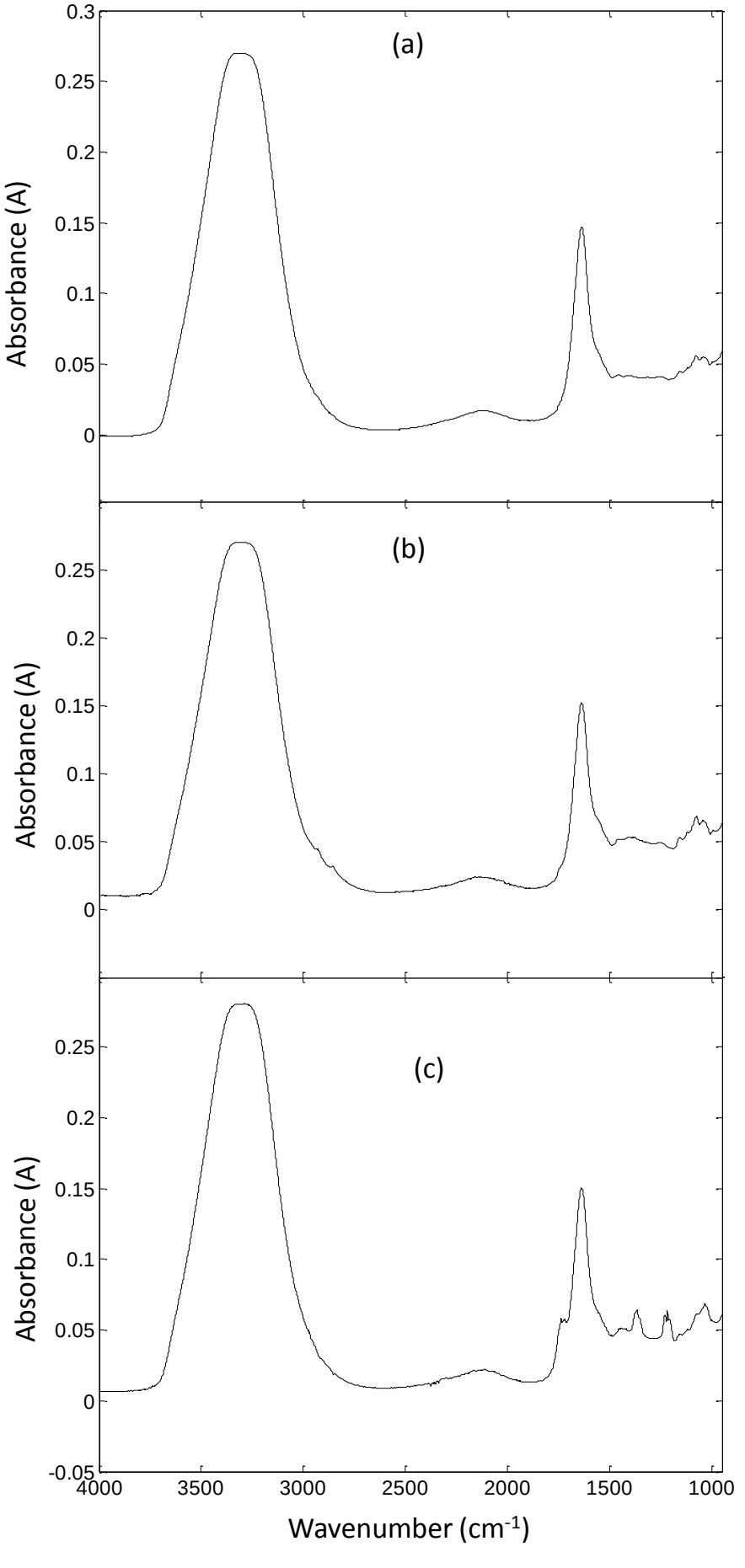
Table 2. Equations and goodness of fit statistics of the probability of detection (POD) curves

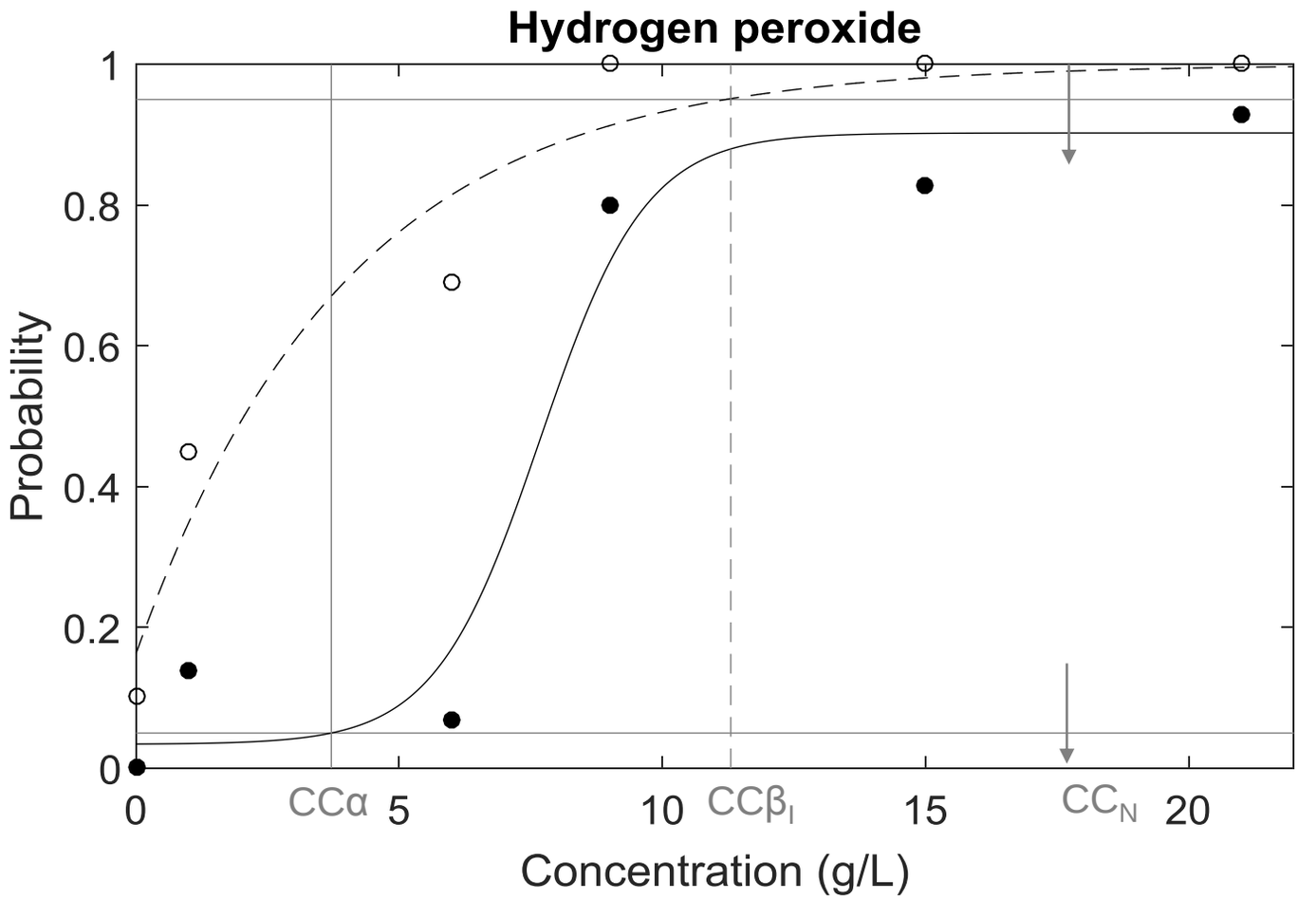
Adulterant	**POD curve**	**Equation**	**Goodness of fit statistics**
Hydrogen peroxide	P(x)	0.90 - 0.87 / [1 + exp(x - 7.68)]	R^2_{adj}: 0.9383; RMSE:0.1076
	P(x)+I(x)	-0.84·exp(-0.25·x) + 1.00	R^2_{adj}: 0.9116; RMSE:0.1104
Formaldehyde	P(x)	-0.92·exp(-431.5·x) + 0.93	R^2_{adj}: 0.8665; RMSE: 0.1083
	P(x)+I(x)	-0.60·exp(-5072·x) + 0.98	R^2_{adj}: 0.9763; RMSE:0.0258

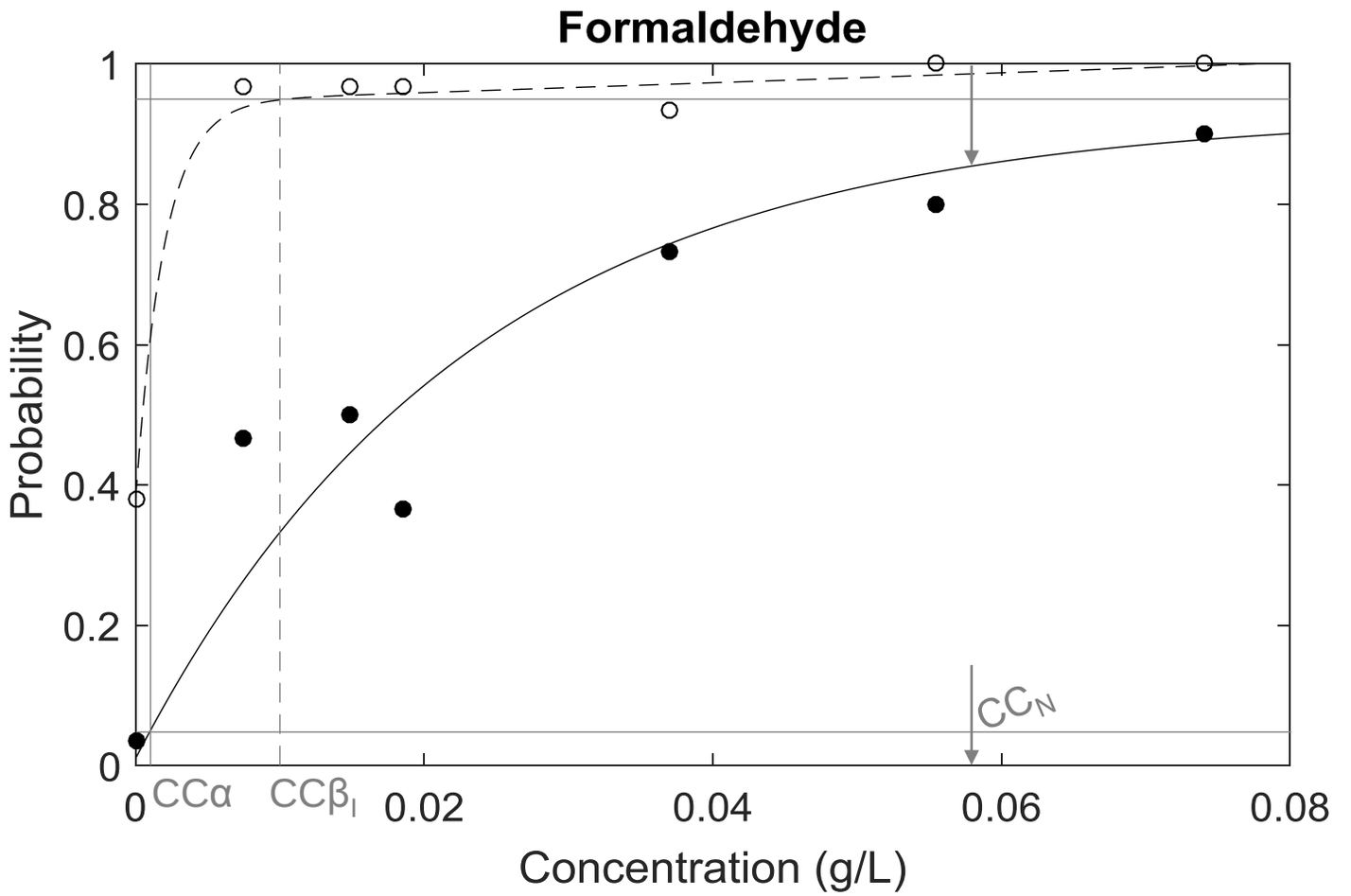
P(x) = POD curve for positive outputs; P(x)+I(x) = POD curve for positive plus inconclusive outputs; R^2_{adj} = adjusted coefficient of determination; RMSE = root mean square error.


Multivariate classification technique (two class strategy)
Unadulterated and adulterated samples at threshold concentration
Class 1 (unadulterated samples)
Class 2 (adulterated samples at threshold concentration)
Positive outputs
Negative outputs
Inconclusive outputs
Positive outputs
Negative outputs
Assessment of the method performance
Outputs
Main performance parameters (sensitivity, specificity and inconclusive ratio)
Suitable parameters?
no
The method is not appropriate
yes
Prediction of samples spiked at concentrations around the threshold level.
Complementary performance parameters (POD curves, CCβ, CCα and unreliability region)
Suitable parameters?
no
The method is not appropriate
yes
Fitness for purpose


1.0
0.95
0.8
0.6
0.4
0.2
0.05
0
Probability
β = 0.05
α = 0.05
0
CCα_l
CCα
CCβ_l
CCβ
Concentration


(a)
(b)
(c)
Absorbance (A)
Absorbance (A)
Absorbance (A)
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
4000
3500
3000
2500
2000
1500
1000
Wavenumber (cm-1)


Hydrogen peroxide
Probability
Concentration (g/L)
CCα
CCβI
CCN
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20


Formaldehyde
Probability
Concentration (g/L)
CCα
CCβI
CCN
0
0.2
0.4
0.6
0.8
1
0
0.02
0.04
0.06
0.08