Analytica Chimica Acta 827 (2014) 28–33

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Validation of multivariate screening methodology. Case study: Detection of food fraud

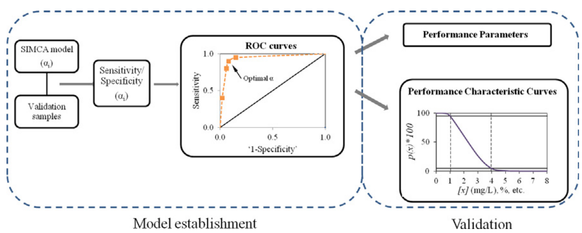
M. Isabel López, Núria Colomer, Itziar Ruisánchez *, M. Pilar Callao

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

HIGHLIGHTS

- New approaches are proposed in validation of multivariate screening.
- ROC curves were used to set the model boundaries at optimal significance level, α.
- Unreliability region, decision limit and detection capability were obtained by PCC curves.
- Suitable performance parameters were achieved in the study of hazelnut adulteration.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form 9 April 2014
Accepted 10 April 2014
Available online 15 April 2014

Keywords:
Multivariate screening validation
ROC curves
Performance characteristic curves
Food fraud
Performance parameters

ABSTRACT

Multivariate screening methods are increasingly being implemented but there is no worldwide harmonized criterion for their validation. This study contributes to establish protocols for validating these methodologies. We propose the following strategy: (1) Establish the multivariate classification model and use receiver operating characteristic (ROC) curves to optimize the significance level (α) for setting the model's boundaries. (2) Evaluate the performance parameter from the contingency table results and performance characteristic curves (PCC curves). The adulteration of hazelnut paste with almond paste and chickpea flour has been used as a case study. Samples were analyzed by infrared (IR) spectroscopy and the multivariate classification technique used was soft independent modeling of class analogies (SIMCA). The ROC study showed that the optimal α value for setting the SIMCA boundaries was 0.03 in both cases. The sensitivity value was 93%, specificity 100% for almond and 98% for chickpea, and efficiency 97% for almond and 93% for chickpea.

1. Introduction

Screening methods have been used with increasing success in routine analysis, thanks to their ability to identify the properties of samples at considerably reduced costs and times. They are characterized by their binary output – presence/absence, yes/no, etc. – according to a pre-set threshold. Screening methods were first used in univariate analysis, which usually require specific measurement (i.e., test kits) [1]. More recently, screening methods have been developed for multiple measurements (i.e., analysis of different properties) or nonspecific signals (i.e., spectroscopic data). In these cases, proper multivariate data treatment is required if the output is to be binary.

In the field of food, multivariate screening methodologies have increased in importance ever since it became important to detect anomalous samples as well as ensure quality and safety [2,3].

* Corresponding author. Tel.: +34977558490; fax: +34977558446.
E-mail address: itziar.ruisanchez@urv.cat (I. Ruisánchez).

http://dx.doi.org/10.1016/j.aca.2014.04.019

Table 1
2×2 contingence table.

Predictions	Actual	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

TP, true positive; TN, true positive; FP, false positive; FN, false negative.

According to the literature, analytical and classification techniques have been successfully combined in food fraud in both adulteration [4–8] and authentication [9–12]. We recently reported a multivariate screening methodology based on two classification approaches for a food adulteration problem [13].

As any analytical method, multivariate screening has to be validated to be implemented as routine methods in control laboratories. This involves establishing performance parameters. However, validation protocols for qualitative methods are not as developed as quantitative methods. For qualitative methods, the main guide is established in the Commission Decision CD/657/EC 2002 [14]. Nevertheless, this CD/657/EC 2002 has been interpreted ambiguously, which has led to a confusion in the terminology [15].

Multivariate validation is not as well established as univariate validation. There is considerable consensus about the definition of such performance parameters as sensitivity and specificity [16] but no agreement has been reached about other related indexes [2,17,18]. In addition, while a positive output in univariate analysis means the 'presence of the analyte under study', a positive output in multivariate analysis means that the 'sample belongs to the pre-established model'. Therefore, some performance parameters might have to be re-defined according to the compliance definition.

The goal of the present study is to establish a strategy for validating a multivariate screening methodology based on a parametric classification technique. In the proposed strategy, the steps were the following: (1) Establish the classification model. We propose using receiver operating characteristic (ROC) curves to optimize the significance level (α) for setting the model's boundaries. (2) Evaluate the performance parameters, some of which can be obtained directly from the output of the screening model. In this study, we propose using the performance characteristic curves (PCC curves) to obtain additional quality parameters such as the unreliability region and limits related to concentration (decision limit and detection capability). A hazelnut adulteration problem is considered as a case study. The price of hazelnuts depends on the market and can be reduced by adding such other ingredients as almond, because of its similarity, but other unexpected products might be added (for example, chickpea).

ROC curves have been extensively described by Fawcett in 2006 [19]. They have mainly been used in such fields as biomedicine, clinical analysis and biometrics [20,21] to set the cut-off value of a test or to compare the performance of different tests. However, their application in the field of food is less extensive, and they are usually used to select variables in multivariate classification techniques [22–24].

PCC curves have mainly been used to obtain performance parameters other than the ones obtained by contingency tables in the quality characterization of univariate screening methodologies [18,25–27]. No references have been found in which PCC curves are used in multivariate screening validation.

This paper takes a further step in several aspects. Firstly, we use ROC curves to optimize the significance level (α) used to establish the boundaries of the model. This α value is usually set to 0.05 by default. Secondly, we propose to adapt PCC curves to a multivariate methodology, for the first time, when dealing with modelling classification techniques.

2. Experimental

2.1. Samples

The unadulterated set is composed of 28 hazelnut pastes. Experience shows that the most common percentage of adulteration is around 7%. So, representative samples were selected and spiked with almond paste or chickpea flour at different levels (1–8%) so that the PCC curves could be established. In total, there were around 13 adulterated samples at each adulteration level and for each adulterant studied.

Additional details about hazelnuts, adulterants and sample preparation can be found in our previous study [13].

2.2. Instrumentation and software

The spectra data was acquired by an infrared (IR) spectrophotometer (FTIR 680 Plus JASCO) equipped with a diamond crystal in the spectrophotometer ATR cell, which was continuously purged with N_2. Spectra were the average of 32 scans, recorded in the spectral range of 4000–600 cm^{-1} every 2 cm^{-1}. The CO_2 and H_2O contributions were removed with the control software Spectra Manager before the spectra were exported to Matlab [28] and treated with PLS Toolbox [29].

3. Strategy

We propose a multivariate screening strategy based on one class approach, so only the unadulterated samples are modeled. Once the model is established, the performance parameters are evaluated. The algorithm used in the present work is soft independent modeling of class analogies (SIMCA), which has been widely applied for classification problems and it is of great interest whenever dealing with spectroscopic data since it is not influenced by working with collinear variables (highly correlated).

SIMCA was introduced by Wold in 1976 [30]. It is a classification technique based on principal component analysis (PCA) that characterizes each sample in relation to the build model by calculating two scalar statistics, Q and the Hotelling T^2. The Q-statistic is related to the amount of original information of each sample not included in the model whereas Hotelling T^2 measures the information of each sample considered by the model. Like any parametric classification technique, some limits, Q_{lim} and $T^2{}_{lim}$, must be set to delimit SIMCA boundaries, which depend on the significance level (α) [31].

Table 2
Description of the performance parameters.

Sensitivity	False negative (FN) rate (1 − sensitivity)	Specificity	False positive (FP) rate (1 − specificity)	Efficiency	Youden's index
$\frac{TP}{TP+FN}$	$\frac{FN}{TP+FN}$	$\frac{TN}{TN+FP}$	$\frac{FP}{TN+FP}$	$\frac{TN+TP}{TN+FP+TP+FN}$	(sensitivity + specificity − 1)