ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
19 (1999) 923–935

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA

A. Candolfi [a], R. De Maesschalck [a], D.L. Massart [a,*], P.A. Hailey [b], A.C.E. Harrington

[a] *ChemoAC, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium*
[b] *Analytical Research and Development, Pfizer Central Research, Ramsgate Road, Sandwich, Kent CT139NJ, UK*

Received 18 March 1998; received in revised form 26 June 1998; accepted 8 August 1998

Abstract

Soft independent modelling of class analogy (SIMCA) is applied to identify near-infrared (NIR) spectra of ten excipients used in the pharmaceutical industry. For each class at least 15 excipient samples were collected for the data base, considering different batches and occasionally various suppliers. Therefore the data of the classes are not always homogeneous. The performance of the original SIMCA method, which is usually described in the literature and also applied by the users, carried out at two confidence levels, 95 and 99%, on original data, SNV (standard normal variate transformation) and second derivative pre-processed data, is discussed. Reasons for the rejection rates are given. No objects were assigned to a wrong class using SIMCA. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Pharmaceutical industry; Excipients; NIR; Original SIMCA; Pre-processing

1. Introduction

The identification of excipients is an important task in the cGMP manufacture of drug products within the pharmaceutical industry. A positive test is required to release the excipient batches for use in the production of clinical trial or commercial batches. The identification of excipients is performed according to the various pharmacopoeias, for example the Pharm. Eur., which include tests for identification [1]. These tests are often wet chemical, with additional tests for providing evidence of functional groups such as aldehydes or ketones within the excipient structure. These mostly laborious compendial identification methods present an opportunity to improve our analyses for both time and quality and one technique that meets these requirements is near-infrared (NIR) spectroscopy. The application of NIR spectroscopy is growing very fast within the industry because it is rapid, requires no, or very little, sample preparation and is non-destructive [2]. In the latest European pharmaco-

* Corresponding author. Tel.: +32-2-4774734; fax: +32-3-4774735; e-mail: fabi@vub.vub.ac.be.

PII: S0731-7085(98)00234-9

poeia a specific monograph on NIR is implemented [1]. Moreover, the FDA (US Food and Drug Administration) is preparing a guideline for the applicability of NIR methods for pharmaceutical quality assessment.

NIR spectra can be analysed and interpreted using a variety of chemometric tools. The identification of the NIR spectra from excipient batches requires a suitable chemometric classification method which leads to the correct identification of unknown excipients. Several methods are reported which can be applied for this purpose [3–7], and in the present work SIMCA (Soft Independent Modelling of Class Analogy) is used. It was shown by Gemperline et al. [3], that raw material testing by SIMCA analysis of NIR spectra seems promising. In SIMCA, each class is modelled separately by PCA (Principal Component Analysis). Class borders, defining the quality of acceptable objects, are then constructed around the PC model. Therefore the performance of the method depends not only on the difference between classes, but also strongly on the training set for each class.

In the present work we focus on both the α-error, i.e. the rejection of correct samples from their class and the β-error, i.e. the acceptance of objects that do not belong to that class. Both α- and β-errors are essential for validating a database, and are vital for a successful classification. The performance of SIMCA is discussed on a small excipient data set, containing ten excipients. There are several variants of SIMCA concerned with the way class borders are defined (e.g. number of degrees of freedom, cross-validation or not). Within this publication the original SIMCA method is considered which is the most often described in the literature [8–11], and applied by the users [3,4]. Special attention is paid on the variability of the NIR spectra within certain excipient classes, and on what classification results can be expected in a real life situation, when samples are coming from different batches and suppliers. Additionally, we investigate whether and how pre-processing influences the performance of SIMCA.

2. Materials and methods

2.1. Excipients and instrumentation

An NIR database was constructed based upon a requirement of at least 15 batches per excipient class. Samples of the following ten excipients were collected:

Class	Excipient	Number of samples
1.	Anhydrous dicalcium phosphate	17
2.	Anhydrous lactose	16
3.	Explotab	19
4.	Lactose	22
5.	Magnesium stearate	15
6.	Methocel	18
7.	Povidone	15
8.	Sodium lauryl sulphate	17
9.	Starch	19
10.	Avicel	17

These substances are used in the manufacture of solid dosage forms as binders, diluents, disintegrates or lubricants. All the samples per class (excipient) were obtained from the warehouse of a pharmaceutical company. The materials were either already on stock or received by the company within a 9 month period. This time was spent collecting the data. The samples came from different excipient batches and occasionally from various suppliers. Multiple excipient grades within certain classes were considered, but the aim of the study is to identify the chemical product, not its specific grade.

The NIR spectra were collected in the reflectance mode with a NIRSystems 6500 spectrophotometer (NIRSystem, Silver Spring, MD, USA). Before the data acquisition, a successful system suitability test (wavelength scale, absorbance scale and noise) was performed. All spectra were ratioed versus a Spectralon standard (99% reflective, SRS-99-010, Labsphere, North Sutton, NH, USA). Each spectrum is the average spectrum of

32 scans. The spectral range used for the data analysis goes from 1100 to 2468 nm, the data were measured in 2 nm intervals, which results in 685 variables.

The standard sample cup (NIRSystems), was utilised for performing the measurements. For each excipient respectively, $7–15 \pm 0.1$ g of powder was filled into the cup in a standard procedure depending upon the bulk density of the material. The corresponding amount of powder was densely packed into the cup and compressed by closing it. Three analysts were involved in performing the NIR measurements.

The samples included in this database were also submitted to pharmacopoeial tests, which have all been passed.

2.2. Pre-processing of NIR spectra

Raw NIR-spectra often exhibit a baseline shift or drift due to variations in the sample presentation. Different parameters can affect the spectra, such as the stability of the instrument, temperature, humidity, the filling of the measurement cell or the particle size of powders. Instrument performance checks ensure compliance of the instrument to rigorous specifications whilst variances in the spectra due to different sample compaction, and hence pathlength can be reduced significantly using pre-processing techniques. Therefore data pre-treatment is an important step in the data analysis.

Two approaches to pre-process data are utilised in this work, SNV (standard normal variate transformation) [15], and second derivative [16]. SNV seems to be suitable to remove the multiplicative interferences of scatter and particle size. Scatter occurs on the surface of particles and depends on the physical nature of the material. The spectral pathlength depends on the particle size of the material. Using derivatives, a linear background is removed. In the case of first derivative it is converted to a constant level and in the second derivative to zero. Here, the second derivative transformation is utilised. In this application the modified Savitzky-Golay convolution method, proposed by Gorry [17], with a window width of 17 variables (i.e. 34 nm), is applied.

2.3. SIMCA

The theory of SIMCA was already extensively discussed by several authors [8–11]. Only a short introduction to the method is therefore presented. In this work the method is applied in its original form.

Each class is modelled separately based on the similarity of the objects within the class. The model is obtained by PCA with a certain number of significant PCs. This is described by the following equation for one class K,

$$\mathbf{X}_K = \bar{\mathbf{X}}_K + \mathbf{T}_K(nxr)\mathbf{V}_K^T(rxp) + \mathbf{E}_K(nxp) \quad (1)$$

where $\bar{\mathbf{X}}_K$ is the mean centred data matrix, $\mathbf{T}_K(nxr)$ the score matrix obtained for n objects and r selected PCs, $\mathbf{V}_K^T(rxp)$ the loading matrix obtained for r selected PCs and p variables and $\mathbf{E}_k(nxp)$ the residual matrix. The selection of the appropriate number of PCs, r, is a crucial point in SIMCA. Several methods are dedicated to this purpose, but in most applications cross-validation is performed. However, this procedure does not always seem to be the optimal one [12–14]. We wanted to apply the simplest possible method of selecting significant PCs to avoid possible errors when the method is used in a routine environment. In our study the number of PCs is therefore selected according to the percentile of the total variance which is expressed by each PC. The PCs containing more than 1% of the total variance are arbitrarily chosen for modelling. The class boundaries, or confidence limits, are then constructed around the PC model. They are based on the distribution of the distances (Euclidean distance) between the objects and the origin in the space of the residual PCs, i.e. based on $\mathbf{E}_K$, where e_{ki}^2 is the squared residual of the kth object for the ith (latent) variable.

$$s_0 = \sqrt{\sum_{k=1}^{n}\sum_{i=1}^{p} e_{ki}^2/[(p-r)(n-r-1)]}$$
$$= \sqrt{\sum_{k=1}^{n}\sum_{i=r+1}^{p} t_{ki}^2/[(p-r)(n-r-1)]} \quad (2)$$

s_0 is the mean distance between the objects belonging to class K and the class model. With the help of an F-test the critical distance can further be computed at a certain level of significance (α):

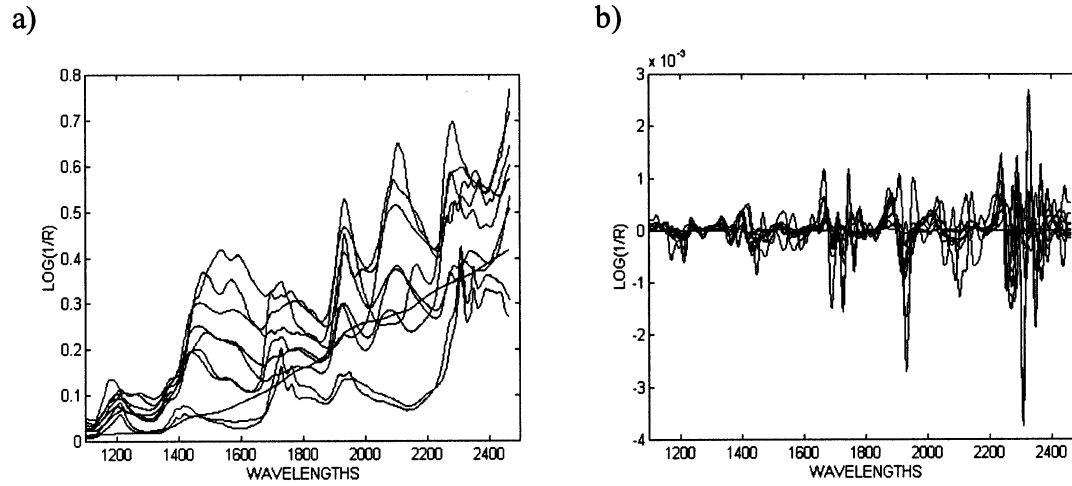

Fig. 1. Mean spectra for the ten excipients obtained from (a) original data and (b) second derivative data.

$$s_{crit} = \sqrt{F_{crit} s_0^2} \quad (3)$$

The choice of α has to be evaluated on the basis of the problem under investigation and is related to the percentile of wrongly rejected objects (α-error), which is considered acceptable. In this application two confidence-levels, 95 and 99%, are selected.

After the model has been developed on the training set, new objects can be classified. For the identification of such a new object it is projected into the PC space defined by the PCA model and its distance towards the class model (s_k) is compared to s_{crit}:

$$\tilde{\mathbf{x}}_{new}(1xp) = \bar{\mathbf{x}}_K + (\mathbf{x}_{new} - \bar{\mathbf{x}}_K)\mathbf{V}_K\mathbf{V}_K^T \quad (4)$$

$$\mathbf{e_{new}} = \mathbf{x}_{new} - \tilde{\mathbf{x}}_{new} \quad (5)$$

$$s_k = \sqrt{\sum_{i=1}^{p} e_{new,i}^2/(p-r)} \quad (6)$$

If $s_k < s_{crit}$, the object is considered part of the class for which the model was established, if the distance is larger it is considered to be an outlier.

The procedure described above is referred to as first part of SIMCA. Sometimes one applies a second step in which one closes the class boxes to detect outliers within the space of the modelled PCs. This does not appear relevant here. Spectra of other classes have basically a different shape and will therefore be outlying in the residual PCs, whereas outliers within the modelled PCs are mainly due to new spectra which are similar to the ones in the training class, but with higher or lower absorbances.

2.4. Software

All methods were programmed by ourselves in Matlab code (V.4.0) (Mathworks, Natick, USA). For the spectral acquisition NSAS (V.3.50) (NIRSystems) was used.

3. Results and discussion

Fig. 1(a) shows the mean spectra of all classes for the original data. The shapes of the spectra are quite different from each other. The spectrum obtained from anhydrous dicalcium phosphate, which is an inorganic substance, has very little feature but nonetheless is useful for the determination of the moisture content in what should be an anhydrous material. Although this excipient class is retained for the data analysis (rather as reference substance), it should be realised that NIR spectroscopy alone is not sufficient to identify inorganic substances. To show the effect of second derivative transformation, the mean spectra of the pre-processed data are presented in Fig. 1(b). The baseline shifts as described previously

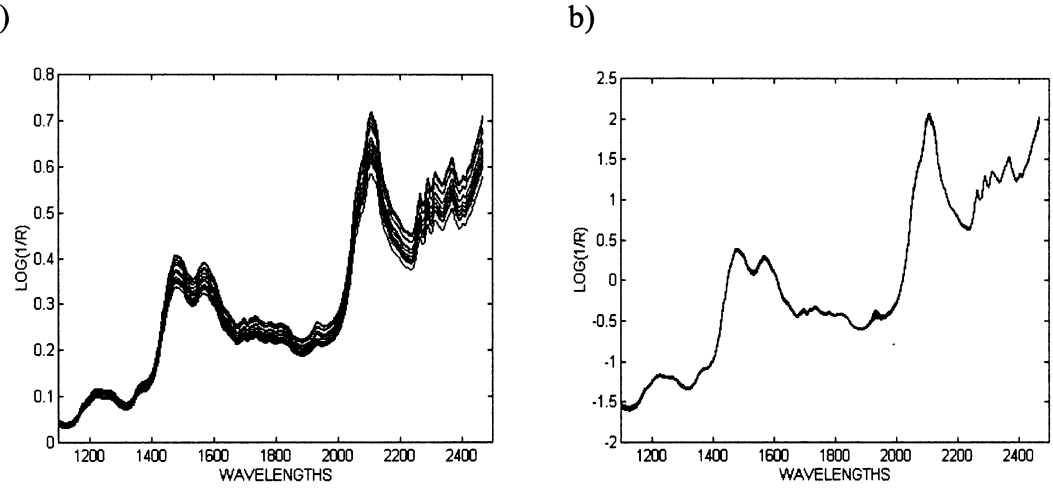

Fig. 2. Spectra for anhydrous lactose (class 2) obtained from (a) original data and (b) SNV data.

have been removed by this transformation. Derivatives emphasise small shoulders and peaks, so that the pre-processed spectra have more pronounced shapes. The spectra of anhydrous lactose without and with being pre-processed with SNV are shown in Fig. 2(a,b). In the original spectra a small offset can be observed at lower wavelengths. Towards higher wavelengths the spread of the spectra increases. These known phenomena, which typically occur for powdered materials due to multiplicative effects of scatter and particle size, are corrected by SNV producing spectra constant over the entire spectral range, except for 1938 nm, which corresponds to a water peak [16]. Anhydrous lactose contains 0.1–0.2% of absorbed water [18].

To study the structure of the data set and the effect of the transformations, PCA is carried out on the column centred, SNV and second derivative data. The PC1 versus PC2 score plots are shown in Fig. 3(a,b,c). On the score plot for the centred, but otherwise not transformed data (Fig. 3a), one can see that all classes are separated along the first two PCs. The spread of the objects within one class is rather small for some classes (classes, 1/2/3/5/8 and 10), while for other classes this is not the case (classes, 4/6/7 and 9). Especially the latter classes seem to be inhomogenous, containing sub-clusters. PC1 is determined both by the variation between and that within classes. On the PC1-PC2 score plot for the SNV data, displayed in Fig. 3(b), all classes are well separated on the plane spanned by the two PCs. This transformation clearly allows better discrimination between the groups. The within-class variance is decreased so that only the spread of the classes 3, 6 and 7, mainly on PC2, is still rather large. PC1 is now determined nearly exclusively by interclass differences, except for class 6, where two sub-clusters are separated. PC2 also describes differences within classes. These differences are particularly due to variations in the height of the log (1/R) values situated around wavelength 1938 nm. This is becoming evident when interrogating the loading plots (see Fig. 4). The highest absolute values for the loadings on the first PC (Fig. 4a), are located at the spectral regions where the mean spectra of the classes are the most different, while the highest absolute values for the loadings on the second PC (Fig. 4b), are obtained for the characteristic spectral band for water at 1940 nm.

The PC1-PC2 score plot for the second derivative data is shown in Fig. 3(c). Here, the first PC particularly separates two classes from the others, namely class 5 and 8. These excipients are the only ones containing fatty acids in their molecular structure. PC2 separates class 4 from the other classes. The spread of class 4 on PC2 is much larger compared to the one of the other classes. Class 1, anhydrous dicalcium phosphate, is lo-

cated around the zero value on PC1. Since it is an inorganic substance the spectra contain little information.

After the global PCA for all classes together each class is explored separately to reveal inhomogeneities. The phenomena, described in Fig. 2(a) (small offset and increasing curvature of the spectra towards higher variables), occur in all classes.

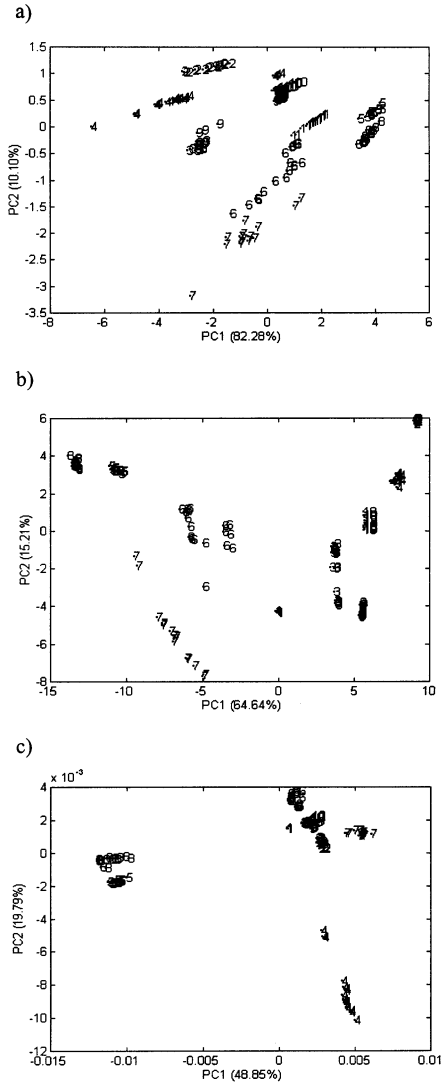

Fig. 3. PC1 versus PC2 score plot for the ten excipient classes obtained from (a) original, (b) SNV and (c) second derivative data. The classes are labelled according to their class label from 1–10 (see Section 2.4).

The plot of the original spectra of explotab (class 3) is shown in Fig. 5(a). The biggest variation in the spectra can be observed around wavelength 1934 nm, a characteristic wavelength for water. In fact, water influences the entire spectral range, but especially two spectral regions, 1450 and 1940 nm. On the PC1-PC2 score plot of the centred data of explotab, which is presented in Fig. 6(a), three to four sub-clusters are displayed. Three sub-clusters are separated along the first PC which contains almost 80% of the total variance. The highest absolute loading value on PC1 (Fig. 6b), is situated at wavelength 1934 nm, so that one can indeed conclude that the variability of the spectra in this class is mainly due to a different water content of the samples.

The original spectra of lactose (class 4) are presented in Fig. 5(b). Lactose is available in several particle size degrees [18], which are all summarised in one class here. The effect of multiplicative interferences due to the various particle sizes is clearly illustrated in this figure. The spectra with lower log (1/R) values over the full spectral range are obtained from powdered material, the ones with higher log (1/R) values from crystalline material. As a result this data set consists of one main group of spectra, two smaller sub-clusters and one single spectrum. This data structure can be observed on the PC1-PC2 score plot obtained for all classes together (Fig. 3a). SNV is able to remove multiplicative interferences due to particle size, and as a result mainly chemical information is retained in the spectra. The variance of the data for class 4 is much smaller after the transformation (Fig. 3b).

The original spectra of class 6, methocel, are displayed in Fig. 5(c). There is particularly one spectrum which is atypical. The main spectral differences are located around the wavelengths 1460 and 1920 nm, which correspond again to the two principal spectral regions for water. The grade of the sample is different compared to the others.

As can been seen in Fig. 5(d), which gives the plot of the spectra for the povidone objects (class 7), some of the spectra are quite unlike the major part of the spectra of this class. Two spectra show lower log (1/R) values over the full spectral range

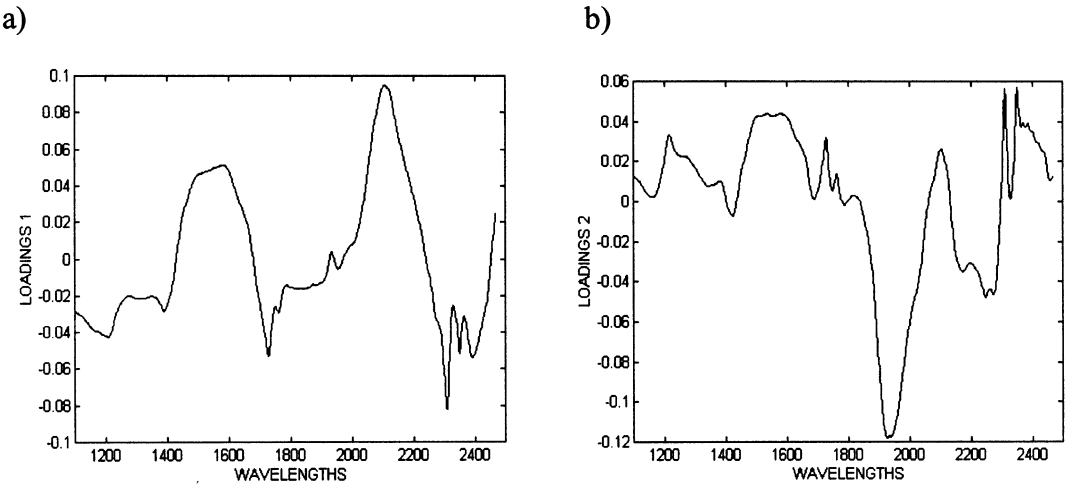

Fig. 4. Loading plots on (a) PC1 and (b) PC2 for the SNV data.

and also a somewhat different shape. A third spectrum is similar in shape to those two spectra, but has higher log (1/R) values. If several suppliers provide slightly different qualities of an excipient, it can be manifested in such variations of the spectra. The inhomogeneities in this class are also noticeable on the general PC1-PC2 score plot for all classes (Fig. 3a).

In class 9 (starch) three spectra are evidently distinct from the rest, since their overall absorbance is higher compared to the other spectra. The spectra are presented in Fig. 5(e). The samples may be corn starch, maize starch, wheat starch etc., but are globally all starch. These distinctions too can be perceived in the PC1-PC2 score plot for all classes (Fig. 3a).

The other classes, which are not discussed here in detail, contain some inhomogeneities, but of a lesser degree. In general one can state that some of the dissimilarities observed in the spectra within one class can be related to the moisture content of the samples and/or to the particle size and shape of the materials. The samples for each class are obtained from different batches and suppliers, so some natural heterogeneity in the individual classes can be expected. Whilst this natural variance is observed, each sample included in the database has passed all compendial tests and is therefore released for use in the production. As a result all spectra are kept for the data analysis with SIMCA, since they represent a real life situation one encounters in the pharmaceutical industry and hence cannot be considered as analytical outliers. However, checking for atypical objects must be carried out carefully, since the samples in the database define the quality of the classification models.

In SIMCA, a classification model (PC model) is constructed for each class individually. In this application we have arbitrarily decided to retain only the PCs containing more than 1% of the total variance for modelling. The remaining, residual PCs, are used to build the confidence interval around the model. The performance of the model is evaluated by doing leave-one-out cross-validation (LOOCV) within the corresponding class. This yields the α-error, which is the amount of correct samples, that are rejected, i.e. not considered part of the class. Two confidence levels are compared, 95 and 99%, which indicates that the theoretically expected amount of wrongly rejected samples is 5, and 1%, respectively. The number of PCs used for modelling, the correct classification rate (CCR) and the number of rejected objects (given between brackets), obtained by LOOCV for both confidence levels, are summarised in Table 1 for the original data.

For each class a PC model with maximum three PCs is built. However, in the cases, where three latent variables are retained, the third PC contains

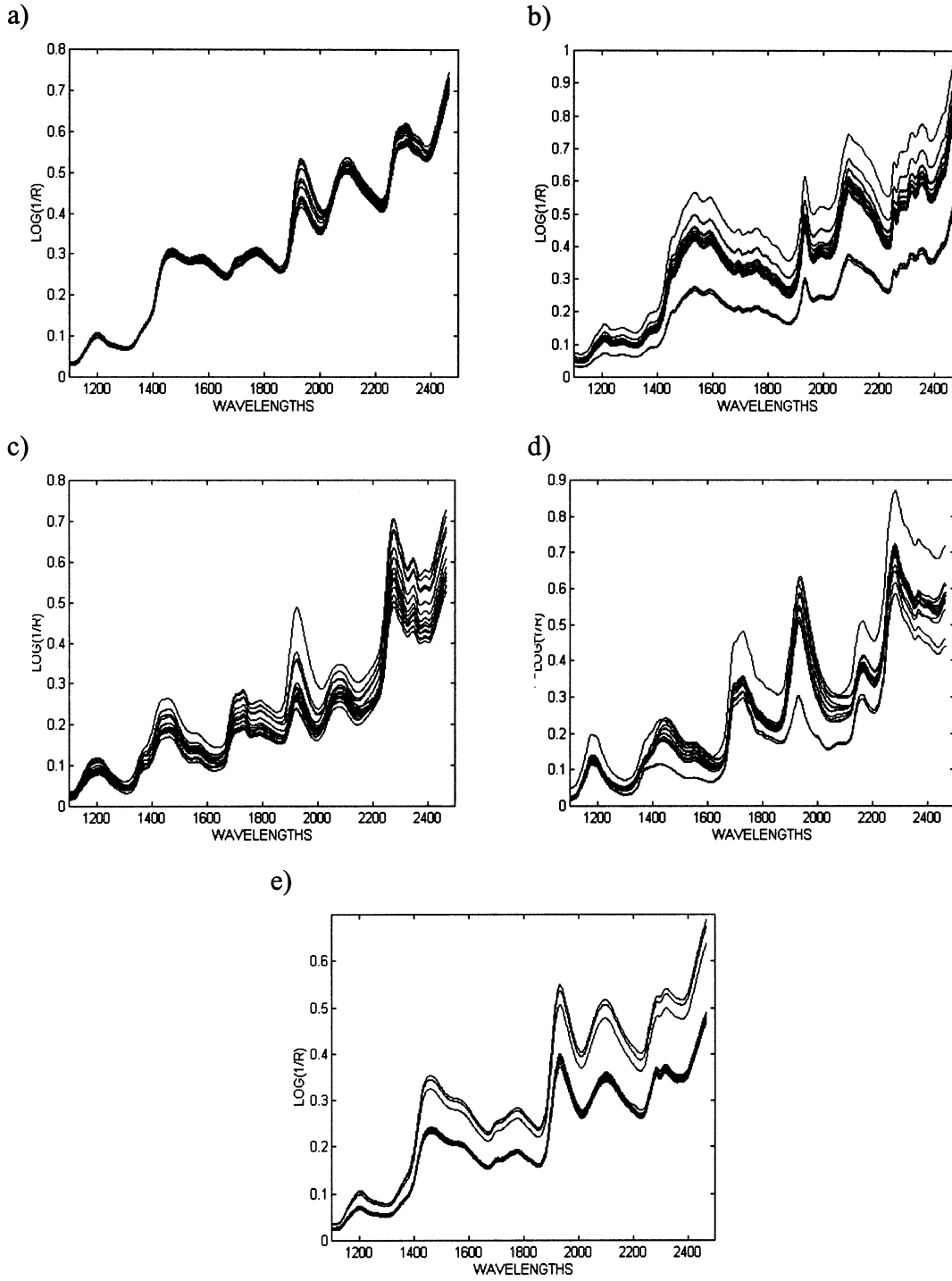

Fig. 5. Original spectra for (a) explotab (class 3), (b) lactose (class 4), (c) methocel (class 6), (d) povidone (class 7) and (e) starch (class 9).

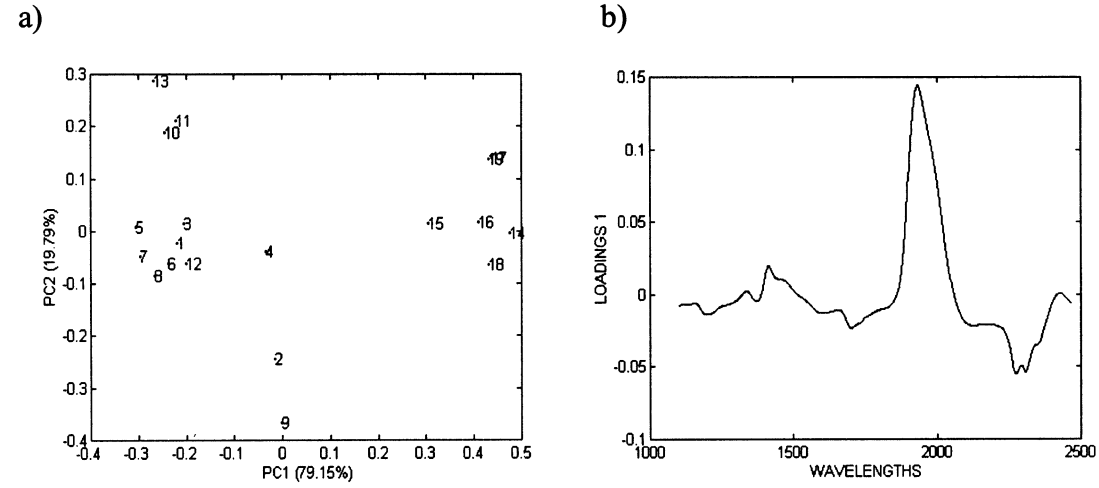

Fig. 6. (a) PC1 versus PC2 score plot and (b) loading plot on PC1 for explotab (class 3).

only around 1% of the total variance. To check, whether a latent variable explaining only such a small part of variance influences the results, the models are repeated with two PCs only. In the table the results obtained for both dimensionalities are separated by a '/'. The total variance explained by the modelled PCs amounts from 97.65 to 99.82% for the individual classes. For every model LOOCV is performed to determine the actual α-error, respectively the CCR. The CCR describes the ratio of the correctly classified samples/the total number of samples in one class. A CCR equal to 1 means that all samples are correctly classified. For the 95% confidence level a CCR between 0.67 and 0.86 is achieved, which signifies that 14–33% of the samples are misclassified, i.e. rejected from their class. This is much more than one expects, i.e. 5%. For the 99% confidence level a CCR between 0.67 and 0.94 is achieved, which means 6–33% rejected samples. In three situations two models are established, with either three or two PCs. Except for class 6, on the 95% confidence level, better results are obtained with the more parsimonious models, so they should be preferred.

Table 1
Classification results (CCR) of SIMCA based on LOOCV for the models obtained with different numbers of PCs (separated by '/') for the raw log (1/R) data[a]

	Number of PCs	CCR, 95% level	CCR, 99% level
Class 1	1	0.76 (4)	0.82 (3)
Class 2	1	0.81 (3)	0.88 (2)
Class 3	2	0.74 (5)	0.84 (3)
Class 4	1	0.86 (3)	0.95 (1)
Class 5	2/3	0.8/0.67 (3/5)	0.87/0.67 (2/5)
Class 6	2/3	0.67/0.72 (6/5)	0.78/0.72 (4/5)
Class 7	2	0.67 (5)	0.73 (4)
Class 8	2/3	0.82/0.71 (3/5)	0.82/0.71 (3/5)
Class 9	1	0.84 (3)	0.89 (2)
Class 10	2	0.76 (4)	0.94 (1)

[a] The number of rejected samples per class is indicated between brackets.

As already mentioned above too many objects are misclassified. We see different reasons for that, the first one has to do with the natural heterogeneity and the dimension of the data set. In LOOCV the system is perturbed by leaving out one object at the time to estimate the classification performance. As soon as an extreme sample is left out, the remaining objects will not span the same space anymore and as a result the object left out will not be classified in this class. This is a realistic way of evaluating the performance of the method, since in real life situations, one must expect that new samples with extreme characteristics will be submitted for prediction. The results show, that SIMCA is sensitive to dissimilarities between objects which is an advantage. Some authors pro-

Table 2
Classification results (CCR) of SIMCA based on LOOCV for the models obtained with different numbers of PCs (separated by '/') for the SNV transformed data[a]

	Number of PCs	CCR, 95% level	CCR, 99% level
Class 1	3	0.82 (3)	0.88 (2)
Class 2	3/4/5/6	0.68/0.63/0.44/0.63 (5/6/9/6)	0.88/0.81/0.63/0.69 (2/3/6/5)
Class 3	1	0.79 (4)	0.90 (2)
Class 4	3/4	0.68/0.68 (7/7)	0.68/0.68 (7/7)
Class 5	3/4/5	0.8/0.67/0.73 (3/5/4)	0.87/0.8/0.8 (2/3/3)
Class 6	2	0.78 (4)	0.78 (4)
Class 7	1/2	0.87/0.73 (2/4)	0.87/0.8 (2/3)
Class 8	3/4	0.65/0.65 (6/6)	0.88/0.77 (2/4)
Class 9	3/4	0.68/0.74 (6/5)	0.74/0.74 (5/5)
Class 10	2	0.94 (1)	0.94 (1)

[a] The number of rejected samples per class is indicated between brackets.

pose to develop stable models by deleting all outliers and by repeating the model for the remaining objects [19]. However, in this application no objects should be removed, since the data set represents real world variations. As more samples are included in the data set, better classification results will be obtained with SIMCA.

SIMCA's sensitivity towards dissimilarities can be related to the parametric character of the method, using the F-test as statistical tool for determining outliers. Parametric methods assume that the data population is normally distributed. This assumption is however not fulfilled for some of the excipient classes, where several excipient grades are summarised within one class. This might explain the too large α-errors.

As described in the theory section on SIMCA (Section 2.3), an important factor for the classification seems to be the number of PCs included in the SIMCA models. The classification results show, that models established with different numbers of PCs lead to different success rates. It still is a difficult task to determine the correct number of latent variables, and no general rules are given.

Another reason for the too high rejection rate is the way the confidence limits are constructed. In the original SIMCA, as applied here, they are built by using the fitted scores of the training set. For the identification of new objects the predicted scores are used to compute the distance to the class model in the space of the residual PCs. Due to the properties of least-square methods (here PCA) the fitted measures, i.e. residuals and also residual scores, are always smaller than the predicted ones. Consequently, they should not be directly compared. If one does so anyway, the α-error in prediction is too high, since the width of the class cylinder is too narrow. This problem has been discussed previously in the literature [12–14], but does not seem to have been considered in applications. Research on how to overcome this is now in progress.

The evaluation of the β-error in such a classification system is an important point. This is done by subjecting the objects belonging to any other class to the model of the class under investigation, to check whether some of these samples would be wrongly identified as belonging to it. None of these objects was wrongly identified, so that the β-error is equal to zero. This is an important result in the application of SIMCA for the identification of excipients. Indeed, β-errors have to be avoided, as this type of error would present significant concern. Since the β-error is equal to zero even at the 99% confidence level, this confidence level is preferred, in order to minimise the α-errors.

It was investigated whether pre-processing has an influence on the performance of SIMCA. This was of special interest, since the excipients should be classified based on their chemical structure. Pre-processing is a possible way to avoid problems occurring due to different particle sizes of powders. For this reason the data analysis is

Table 3
Classification results (CCR) of SIMCA based on LOOCV for the models obtained with different numbers of PCs (separated by '/') for the second derivative transformed data[a]

	Number of PCs	CCR, 95% level	CCR, 99% level
Class 1	3/4/5/6/7	0.82/1/1/1/1 (3/–/–/–/–)	0.88/1/1/1/1 (2/–/–/–/–)
Class 2	2/3/4	0.63/0.69/0.69 (6/5/5)	0.75/0.75/0.75 (4/4/4)
Class 3	1/2/3	0.79/0.74/0.68 (4/5/6)	0.84/0.79/0.74 (3/4/5)
Class 4	1/2/3	0.82/0.68/0.77 (4/7/5)	0.95/0.73/0.82 (1/6/4)
Class 5	3/4	0.67/0.67 (5/5)	0.87/0.67 (2/5)
Class 6	2/3/4	0.89/0.72/0.72 (2/5/5)	0.89/0.78/0.83 (2/4/3)
Class 7	2/3/4	0.8/0.73/0.67 (3/4/5)	0.87/0.8/0.73 (2/3/4)
Class 8	2/3	0.71/0.76 (5/4)	0.76/0.76 (4/4)
Class 9	1/2	0.79/0.68 (4/6)	0.79/0.74 (4/5)
Class 10	2/3	0.82/0.88 (3/2)	0.82/0.88 (3/2)

[a] The number of rejected samples per class is indicated between brackets.

repeated on SNV and second derivative pre-processed data. The results are summarised in Table 2 for the SNV data and in Table 3 for the second derivative data.

It is evident that after pre-processing the major variance explained in the PCs is often not only described by the first or first two latent variables, as it is the case for the original data, but that it is distributed over several PCs. As a result models with more PCs are constructed when the data are pre-processed. Since the dominant variance in the spectra of one class is removed by pre-processing (offset, baseline drift) the data are more similar. Therefore the individual PCs for one class explain less variance, than before pre-treatment. As was described for the original raw data, latent variables containing only a small part of the total variance, but more than 1% (1–2%), were removed and the models were repeated. In the table the results for the different dimensionalities are again separated by a'/'. In most cases the more parsimonious models with fewer PCs again lead to better results and are preferred. The α-errors are comparable to those achieved with the original data. In SIMCA the confidence interval is built based on the data-variance in the space of the residual PCs. If the data are more or less homogeneous (for instance after pre-processing) the variance is small and so is the confidence interval, so that even very small deviations lead to rejection. Pre-processing removes some general undesirable effects due to the measurement procedure and the sample itself, but cannot eliminate inhomogeneities in the data, as for instance, natural acceptable variations of the moisture content. In our data set there is still a natural heterogeneity within most classes, since the samples are coming from different batches and suppliers.

The main effect of pre-processing in the context of SIMCA is not to decrease the α-error, but to influence the β-error where necessary. However, the β-error with raw data was already excellent, since it was zero. To evaluate whether this remains so with the transformed data, all samples from the other classes were again subjected to the class model under investigation. None of the objects were wrongly classified, i.e. the β-error is again equal to zero. Pre-processing decreases the within-class variance and increases the between-class variance, so that the β-error could be reduced if needed. In order to demonstrate this relationship, the Fisher criterion (FC) is used [20]. The FC describes the ratio of between-class variance/within-class variance for each variable. Fig. 7(a,c) shows the superposed original and SNV spectra of class 2, and class 4, respectively. The spread of the original spectra of class 4 is much larger compared to that for class 2, after transformation the spreads are strongly reduced and almost comparable for both, and their between-class variance is increased (see Fig. 7b,d). In the case of the original data, there is only one

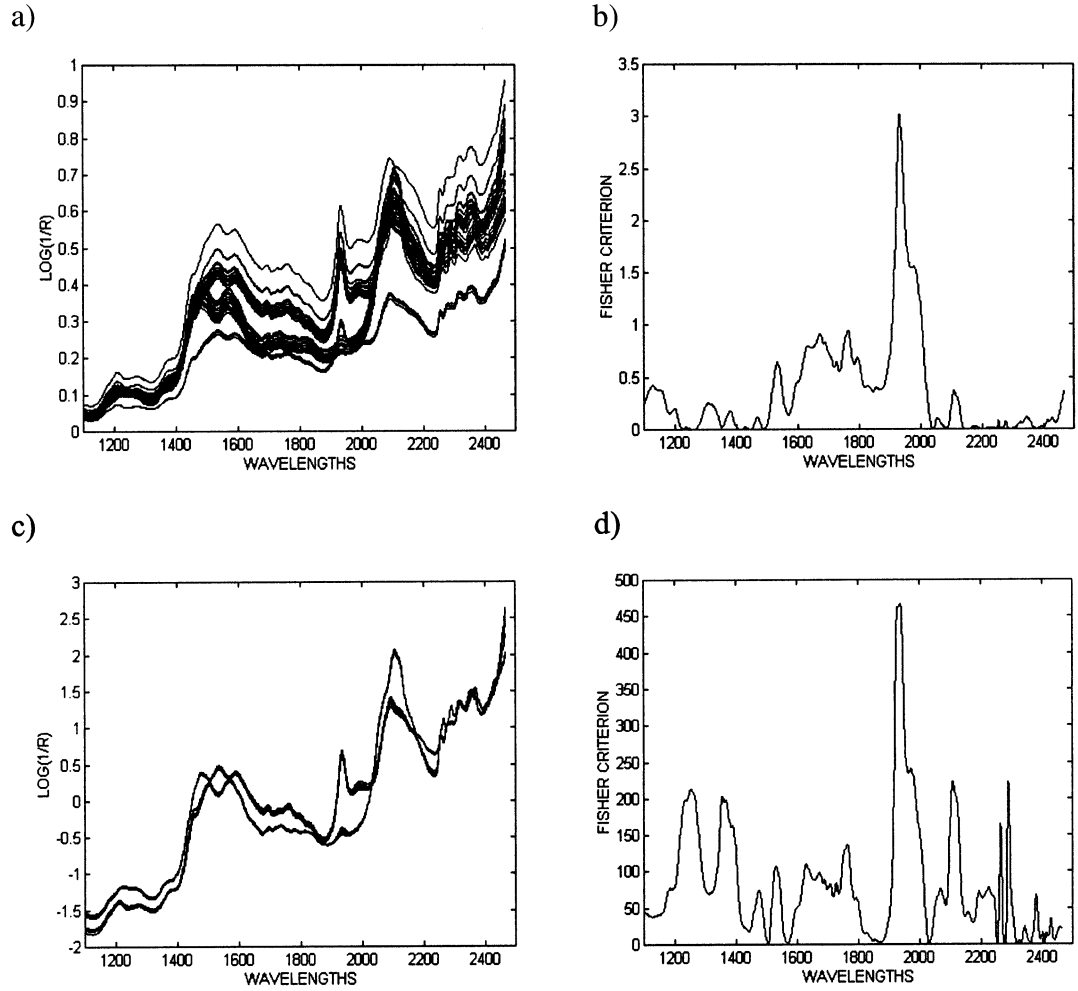

Fig. 7. Spectra for class 2 and class 4 obtained from (a) original data and (c) SNV data, corresponding FC obtained from (b) original data and (d) SNV data.

spectral region with discriminating power, situated around wavelength 1940 nm, one of the characteristic spectral regions where water absorbs. For the SNV data, the FC is much larger for almost all variables. The magnitude of the criterion, after applying SNV, is about 100 times increased. Although in this application the β-error remains unchanged after transforming the data, increasing the between-class variance by preprocessing is still useful, if in a later stage additional classes of excipients would be included in the classification system. Thereby the risk of overlapping classes would then be diminished.

4. Conclusions

NIR combined with SIMCA in its original form is used to identify the samples of an excipient data set. PCs containing more than 1% of the total variance are retained for building the model, while the residual PCs are used for the construction of the confidence limits. Two confidence levels, 95 and 99%, are selected to establish the confidence limits around the PC model. For both levels, the β-error, i.e. the wrong identification of samples, is equal to zero. Therefore it is proposed to work at a confidence level of 99%, since in this

situation the α-error is smaller. About 15% of the samples are rejected from their own classes (α-error), although they have all passed the compendial tests. Several reasons are given for explaining the high rejection rate. On the one hand they are mainly data related, i.e. the heterogeneous nature of the NIR spectra from different batches and suppliers and the limited number of objects available to build the model. On the other hand they are due to the pattern recognition method, i.e. its parametric character, the number of latent variables used for modelling and the way the confidence limits are constructed, namely with the use of the fitted scores.

To decrease the α-error several solutions can be proposed. Since one of the reasons for a high α-error is the natural heterogeneity of the data, one has to deal with this situation. It might be useful to measure several samples for each batch to increase the number of objects available when building the model. This could give a better estimate of the variability of the material and less extreme objects might be found. In time, also new incoming samples should be included in the database, in order to make the database as representative as possible. This too would probably lead to a smaller rejection rate. More elaborate algorithms of SIMCA, focusing on the way the confidence limits are constructed, may also solve part of the problem.

The β-error in the data analysis is zero, which is an important result, since it means that, if the substance is accepted to be a certain excipient, this conclusion is never wrong. However, the β-error was only tested by predicting samples from other present excipient classes by the models. This can be considered as a first stage of validating the database. For actually using the method in a quality control laboratory further tests are required, such as submitting excipient samples which are out of specification to the database or contaminated samples. It has also to be kept in mind, that NIR spectroscopy alone should not be used for the identification of inorganic materials. For these substances β-errors may occur due to a lack of characteristic spectral features.

Pre-processing of the spectra does not influence the results here, but seems nevertheless useful. It removes physical spectral information (due to particle size), so that the models are build based on mainly chemical spectral information. Moreover it increases the between-class variance. This is necessary to decrease a possible β-error if in a later step more classes would be included in the identification system or contaminated samples would be analysed.

References

[1] European Pharmacopoeia, 3rd Edition, 1997, Council of Europe, Strasbourg, 1996.
[2] W.F. McClure, Anal. Chem. 66 (1) (1994) 43–53.
[3] P.J. Gemperline, L.D. Webber, F.O. Cox, Anal. Chem. 61 (1989) 138–144.
[4] N.K. Shah, P.J. Gemperline, Anal. Chem. 62 (1990) 465–470.
[5] P.J. Gemperline, N.R. Boyer, Anal. Chem. 67 (1995) 160–166.
[6] P. Corti, L. Savini, E. Dreassi, G. Ceramelli, L. Montecchi, S. Lonardi, Pharm. Acta Helv. 67 (2) (1992) 57–61.
[7] M.A. Dempster, B.F. MacDonald, P.J. Gemperline, N.R. Boyer, Anal. Chim. Acta 310 (1995) 43–51.
[8] S. Wold, M. Sjöström, in: B.R. Kowalski (Ed.), Chemometrics: Theory and Application, American Chemical Society, Washington DC, 1977, pp. 243–281.
[9] M.P. Derde, D.L. Massart, Chemom. Intell. Lab. Syst. 4 (1988) 65–93.
[10] B. Mertens, M. Thompson, T. Fearn, Analyst 119 (1996) 2777–2784.
[11] R. De Maesschalck, A. Candolfi, S. Heuerding, D.L. Massart, Decision Criteria for SIMCA Applied to Near Infrared Data, Chemom. Intell. Lab. Syst. (1998) in print.
[12] J.B.M. Dröge, H.A. van't Klooster, J. Chemom. 1 (1987) 221–230.
[13] J.B.M. Dröge, W.J. Rinsma, H.A. van't Klooster, A.C. Tas, J. van der Greef, J. Chemom. 1 (1987) 231–241.
[14] S. Wold, M. Sjoestroem, J. Chemom. 1 (1987) 243–245.
[15] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[16] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR spectroscopy. 2nd ed., Longman Scientific and Technical, UK., 1993.
[17] P.A. Gorry, Anal. Chem. 62 (1990) 570–573.
[18] Hilfsstoffkatalog, Ciba-Geigy A.G., F. Hoffmann-La Roche, Aktiengesellschaft und Sandoz A.G., Basel (Switzerland), 1974.
[19] M.P. Derde, D. Coomans, D.L. Massart, Anal. Chim. Acta 141 (1982) 187–192.
[20] W. Wu, B. Walczak, D.L. Massart, K.A. Prebble, I.R. Last, Anal. Chim. Acta 315 (1995) 243–255.