Available online at www.sciencedirect.com
SCIENCE DIRECT®
Talanta 62 (2004) 25–35
Talanta
www.elsevier.com/locate/talanta

Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration

M.H. Zhang, J. Luypaert, J.A. Fernández Pierna, Q.S. Xu[1], D.L. Massart*

ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received 8 May 2003; accepted 4 June 2003

Abstract

A principal component regression (PCR) model is built for prediction of total antioxidant capacity in green tea using near-infrared (NIR) spectroscopy. The modelling procedures are systematically studied with the focus on outlier detection. Different outlier detection methods are used and compared. The root mean square error of prediction (RMSEP) of the final model is comparable to the precision of the reference method.

Keywords: Multivariate calibration; Robust; Leverage point; Residual; Total antioxidant capacity; NIR; PCR; Outlier detection; Green tea

1. Introduction

Green tea is of great interest due to its beneficial medicinal properties [1]. Many studies have suggested that these properties are related to the antioxidant activity coming from tea polyphenols [2,3]. Tea polyphenols account for 30–42% of the dry weight of green tea leaves [4]. The main polyphenols in green tea are epicatechin (EC), epicatechin-3-gallate (ECG), epigallocatechin (EGC) and epigallocatechin-3-gallate (EGCG), with the latter playing the most important role in the total antioxidant capacity of green tea. To quantitatively control the antioxidant capacity in green tea, there are three main classical methods. One is to determine individual polyphenols using high-performance liquid chromatography (HPLC) or capillary electrophoresis (CE) [5]. Another is to estimate the total phenolics content by a colorimetric method such as the Folin–Ciocalteu assay [6]. The third one is to analyze the total antioxidant capacity based on the reducing activity of polyphenols, such as the oxygen radical absorbance capacity (ORAC) assay and the trolox equivalent antioxidant capacity (TEAC) assay [7]. All these methods are time consuming and difficult to handle due to unstability of polyphenols and their unknown interactions.

Near-infrared (NIR) spectroscopy is a fast, accurate, easy and non-destructive technique that can be a candidate as a replacement of classical chemical analysis. The prerequisite of NIR application for quantitative purpose is building a reliable calibration model. Recently, Luypaert et al. investigated the feasibility for prediction of total antioxidant capacity in green tea using NIR [8] with a small calibration set. We collected a new data set comprising more varieties of green tea samples and used an improved method proposed by Re et al. [9] as the reference method.

We systematically studied the different steps that have to be gone through in multivariate calibration. The focus of this article is on outlier detection since it is the most important and difficult step for modelling. Outliers are observations that either are irrelevant, incorrect or abnormal in some other way when compared to the majority of the data. They often strongly influence the modelling. Outliers may occur due to avoidable and unavoidable mistakes in the analytical process or the presence of irrelevant sources of variance (bad outliers). Observations may behave as outliers because of their highest or lowest analyte levels. They then merely are extreme values and are in fact good "outliers". Bad outliers often reflect errors of some sort. If this is the case, they should be eliminated. In some cases it is due to the inclusion of a sample which behaves differently. In that case it means that the model should be extended by the inclusion of additional samples of that type. Good outliers

* Corresponding author. Tel.: +32-2-477-4734; fax: +32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).
[1] On leave from Hunan University, Changsha, PR China.

0039-9140/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0039-9140(03)00397-7

are often very valuable because they expand the calibration range and should be retained.

2. Experimental

2.1. Samples

One hundred twenty-three batches of tea were purchased in China, Belgium and Spain. Of the 99 batches bought in China, the origin is known. Thirty-five of them are from *Zhejiang*, 31 from *Hunan*, 7 from *Anhui*, 6 from *Shanxi*, 6 from *Fujian*, 5 from *Jiangsu*, 1 from *Jianxi* and 1 from *Sichuan*, respectively. Most of them are green tea of different grades and kinds, which include *Longjing*, *Maojian*, *Biluochun*, etc. Eight of them are yellow tea, which is slightly fermented green tea. Of the other 24 samples, it is only known that they are imported from China. Most of them are *Gunpowder* and some are *Chun Mee*.

2.2. Reagents and standards

2,2′-Azino-bis(3-ethylbenz-thiazoline-6-sulfonic acid (ABTS) diammonium salt) and potassium persulfate were purchased from Sigma–Aldrich Chemie GmbH (Steinheim, Germany). Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carbonsaeure; Aldrich, Gillingham, Dorset, UK) was used as the antioxidant standard.

2.3. Apparatus

NIR spectra of samples were recorded by BRAN + LUEBBE InfraAlyzer 500 (Bran + Luebbe GmbH, Norderstedt, Germany) using the diffuse reflectance mode. An UV-Vis scanning spectrophotometer (SHIMADZU, Japan) was used to monitor the time course of the antioxidative reaction.

2.4. Procedures

2.4.1. NIR spectra

NIR spectra were obtained by packing original sample leaves into a BRAN + LUEBBE standard closed cup and measuring the samples every 2 nm between 1100 and 2500 nm at room temperature. Each batch was measured three times in different days. Each day three samples from the same batch were measured three times with a rotation of 120°, respectively. As a result, each batch was scanned 27 times (9 times × 3 days). A replicate in our study refers to the mean of the nine spectra of the same batch measured the same day. Therefore each batch has three replicates and the mean of the replicates forms the matrix $\boldsymbol{X}$.

As a result, the matrix $\boldsymbol{X}$ contains 123 objects and 701 variables (Fig. 1). It is divided into two subsets using the DUPLEX method [10]: 100 objects in calibration set and 23 objects in test set. Objects are labelled according to the orders that they are randomly arranged in calibration set or test set. Each object has a fixed label. The test set is only used for external validation. Unless otherwise stated, only the calibration set is considered. Before calculation, the calibration set is column centred.

2.4.2. Total antioxidant capacity

Total antioxidant capacity was measured according to the method proposed by Re with slight modifications [9].

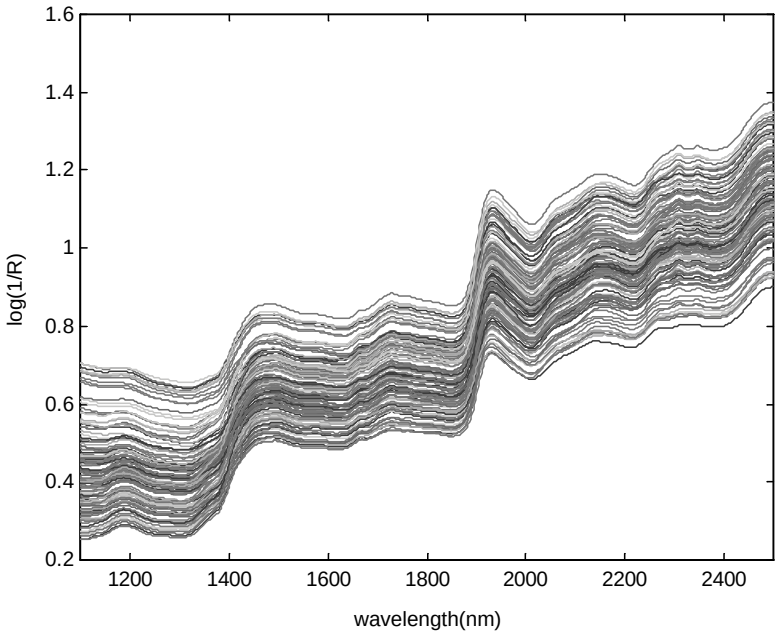


Fig. 1. NIR spectra of green tea leaves ($n = 123$).

First, the ABTS radical cation ($ABTS^{\bullet+}$) was generated by reacting 7 mmol ABTS water solution with 2.45 mmol potassium persulfate (final concentration). After standing the mixture for 12–16 h, the $ABTS^{\bullet+}$ was diluted with water to an absorbance of 0.70 (±0.02) at 734 nm. Then, 1 g of green tea sample was infused with 200 ml boiling water for 40 min in the dark. Finally 0.3 ml 200 times diluted infusion or Trolox standard were added into 1.0 ml diluted $ABTS^{\bullet+}$ and the decrease of the absorbance was read exactly 1 min after the mixing. The total antioxidant capacity value of the samples was obtained from the Trolox standard calibration curve. Each batch was measured three times on different days. The mean of the measurements for each batch forms the vector $\boldsymbol{y}$. The standard deviation of each batch i, determined as:

$$\mathrm{SD}_i = \sqrt{\frac{\sum_{u=1}^{u} \boldsymbol{y}_{iu} - \bar{\boldsymbol{y}}_i}{u-1}} \quad (1)$$

where u, the number of the measurement replicates, is an indication of measurement precision of that sample, including variance due to inhomogeneity of the sample. The mean of all SD (pooled SD), calculated as:

$$\mathrm{SD}_{\mathrm{pool}} = \sqrt{\frac{\sum_{i=1}^{n} \mathrm{SD}_i^2}{n}} \quad (2)$$

where n, the total number of samples, is used as the final value. The results showed that the total antioxidant capacity of 123 samples ranges from 14.53 to 35.79 (μmol Trolox/25 μg tea leaves) with an overall mean equalling 26.10 (Table 1). The pooled SD for 123 samples is 1.86 and the overall standard deviation, which describes the dispersion within 123 samples, is 3.93.

2.5. Data analysis

Principal component regression (PCR) was used to model the relation between the total antioxidant capacity and the NIR spectra of green tea. The performance of the final PCR model was evaluated in terms of bias (trueness), root mean square error of cross-validation (RMSECV) (precision) and square of multiple correlation coefficient R^2 (percentage of variation explained).

$$\mathrm{bias} = \frac{\sum_{i=1}^{n} (\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)}{n} \quad (3)$$

Table 1
The constitution of the calibration set and the test set by the reference method (unit: μmol Trolox/25 μg tea leaves)

Data set	n	Antioxidant capacity range	Mean	Pooled SD
Calibration set	100	19.61–35.73	26.44	1.86
Test set	23	14.53–35.79	24.64	1.83

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{\boldsymbol{y}}_{\setminus i} - \boldsymbol{y}_i)^2}{n-1}} \quad (4)$$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})^2}{\sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2} \quad (5)$$

where n is the number of the calibration samples, $\boldsymbol{y}_i$ the observed result for sample i, $\hat{\boldsymbol{y}}_i$ the estimated result for sample i, $\hat{\boldsymbol{y}}_{\setminus i}$ the estimated value for sample i when the model is constructed with sample i removed and $\bar{\boldsymbol{y}}$ the mean of the observed results for all calibration samples.

2.6. Software

Data analysis was performed in Matlab® for Windows, version 5.2 (The MathWorks Inc.) with the programs developed in our department.

3. Results and discussion

3.1. Data investigation

The data investigation includes three steps:

- Detection of outlying replicates.
- Investigation of clustering tendency. If clustering tendency is found, it should be decided whether to build separate models or not.
- Flagging of possible outliers. The PCR model is sensitive to the existence of outliers, it is necessary to detect possible outliers and check whether they can affect the model or not. Three kinds of outliers should be considered, namely outliers in $\boldsymbol{X}$, outliers in $\boldsymbol{y}$ and outliers towards model. Before modelling, outlier detection in $\boldsymbol{X}$ and $\boldsymbol{y}$ are performed to make sure the model will not be biased due to the existence of outliers. A true model, which is not influenced by the presence of bad outliers, is the base for correctly identifying outliers towards the model. Thus the outlier detection before modelling is important. Since at this stage the outlier detection is done separately in $\boldsymbol{X}$ or $\boldsymbol{y}$, it is quite possible that a sample with extreme value is detected as an outlier but it is later found to follow the model very well. Samples with extreme value always have larger influence on the model than other samples. Therefore it is necessary to be aware of them and discard them only when they are really of no use or detrimental to the model. Once the model is built, the presence of outliers towards the model is investigated. Such outliers should be eliminated.

3.2. Detection of outlying replicates

The Cochran test is used to detect outlying replicates [11]. No outlying replicate is detected.

Table 2
Hopkins statistic for the calibration data ($n = 100$)

Number of iterations	Population size (%)	$HS_{average}$	HS_{max}	HS_{min}	HS_{range}
5	20	0.41	0.46	0.36	0.09
5	100	0.40	0.43	0.37	0.06
10	10	0.39	0.58	0.23	0.35
20	5	0.37	0.65	0.11	0.54

3.3. Investigation of clustering tendency and inhomogeneity

3.3.1. Hopkins statistics

The Hopkins statistic (HS) [12], an index for clustering tendency, is used to examine whether objects in a data set are uniformly distributed in a multidimensional data space. It first calculates the Euclidean distances between randomly selected experimental objects and their nearest neighbours (Ed_{exp}). Then it generates some artificial objects. The Euclidean distances between artificially generated objects and their nearest experimental neighbours (Ed_{art}) are also calculated.

$$HS = \frac{Ed_{art}}{Ed_{art} + Ed_{exp}} \quad (6)$$

If HS > 0.75, the data set is considered to be significantly clustered.

In our study, Forina's modification is used to select suitable artificially generated objects [13]. Different combinations of the population size and the number of iterations are used to make sure each object can be chosen once (Table 2). No clustering tendency is found.

3.3.2. PCA score plot

Plotting PCA scores in two or three dimensions provides an easy way to observe the data distribution. In the PC1–PC2 plot, samples are unevenly distributed but no obvious cluster can be found (Fig. 2a). Although no obvious cluster can be observed visually, DBSCAN [14,15] shows there are three clusters (Fig. 2b). The second cluster includes six objects (objects 13, 92, 93, 94, 96 and 97), which are from the same kind called "Gunpowder". The third cluster consists of five objects, objects 7, 33, 44, 73, 83. No explanation can be found for this cluster. The five objects are from three kinds and three provinces. Different combinations of PCA score plots are investigated and no clear cluster is found. The possible outliers in each PC are listed in Table 3.

Table 3
The possible outliers on the more important PCs

PC	Describe variance (%)	Number of the object
1	91.8719	None
2	7.0726	4, 9
3	0.9004	None
4	0.0931	58, 83, 97
5	0.0394	3, 42, 46, 91, 99

3.3.3. Sammon's mapping

The PCA score plot requires pairwise plots of PC scores (e.g. PC1–PC2, PC1–PC3, etc.). Sammon's mapping supplies a way to visualize higher dimensional data in a single lower (normally 2) dimensional space while approximately preserving the inherent data structure [16]. To obtain stable results of the mapping, 200 iterations were used. The result shown in Fig. 3a indicates several unclear clusters. When the DBSCAN is applied to this result, three clusters are found (Fig. 3b). The formation of these clusters is similar to the result obtained by DBSCAN using the first two PCs (Fig. 2b). Three objects (objects 5, 9 and 98) are detected as outliers.

We conclude that careful observation shows some evidence of clustering. However, the clusters detected are not very clear and also too small to allow the building of separate models at this stage. If the method were to be applied on a very large scale and many more calibration samples were available, this might be reconsidered.

3.4. Detection of possible outliers in $\boldsymbol{y}$

Two kinds of Grubbs' tests [17] are used to detect the potential outliers in $\boldsymbol{y}$. In the single Grubbs' test

$$G = \frac{y_i - \bar{y}}{s} \quad (7)$$

where y_i is the suspected outlier that has the smallest or largest value, $\bar{y}$ is the mean of all samples including the suspected sample and s is the standard deviation of all samples. If the absolute G exceeds the critical value [17], the suspected sample is considered an outlier.

The double Grubbs' test is used for the detection of two outliers.

$$G = \frac{SS_{1,2}}{SS_0} \quad \text{or} \quad G = \frac{SS_{n-1,n}}{SS_0} \quad (8)$$

where SS_0 is the sum of squared deviations from the mean of the original samples and $SS_{1,2}$ and the $SS_{n-1,n}$ are the sum of squared deviations obtained after deletion of the two smallest or largest values. If G is smaller than the critical value [17], the two samples are considered to be outliers.

The prerequisite for using the Grubbs' test is that the data set should have a normal distribution. As shown in Fig. 4, the distribution of $\boldsymbol{y}$ for the 123 samples is close to normal. The Kolmogorov–Smirnov test shows that the antioxidant capacity values of the samples are normally distributed. The kurtosis value is 0.24 and the skewness value is 0.00. None of the outliers is detected by both Grubbs' tests.

Since the Grubbs' test concludes that there are no outliers in $\boldsymbol{y}$, all samples are considered for further analysis. It should be noted that in this case the DUPLEX method has selected five samples (objects 105, 108, 109, 119 and 122) for the test set that are not within the $\boldsymbol{y}$-range of the calibration set. The DUPLEX method makes a selection based on the spectral information and is chosen because it does

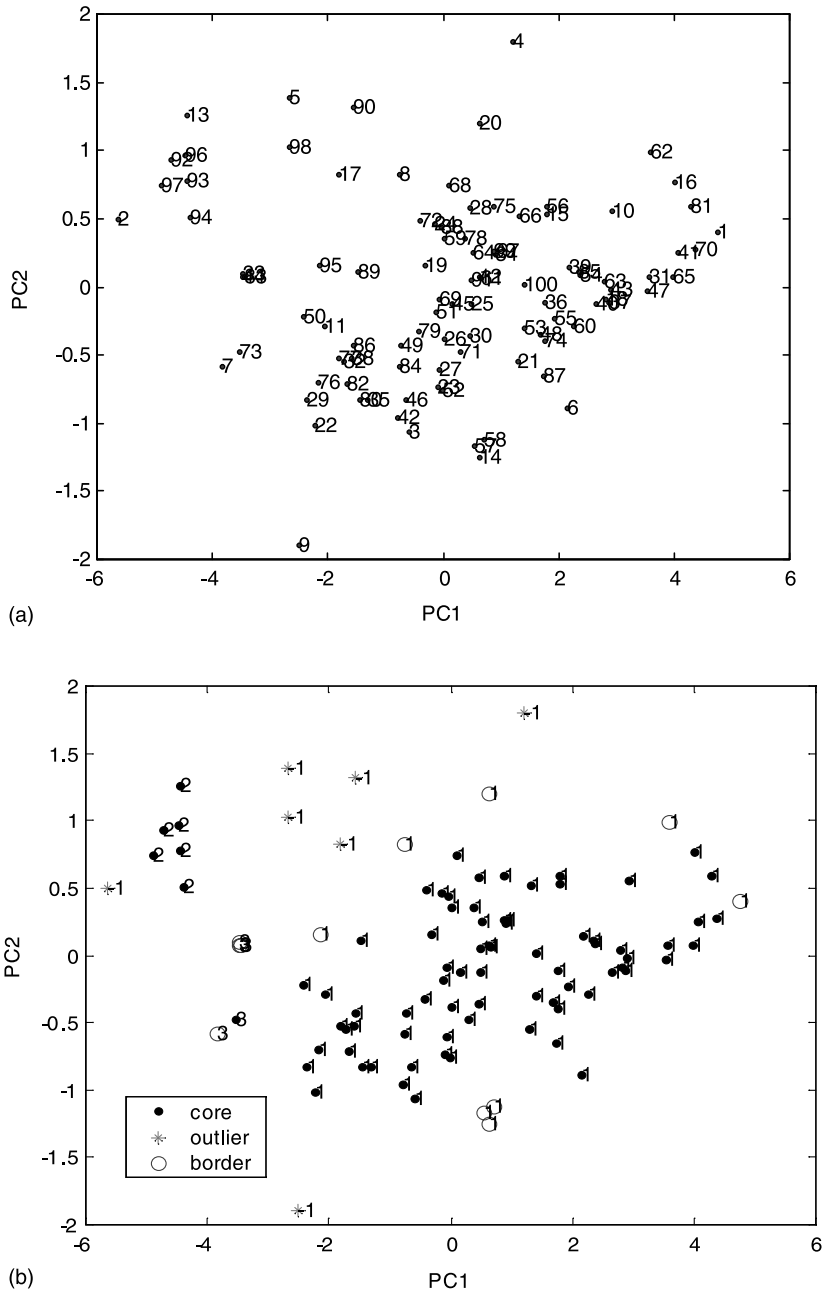


Fig. 2. Investigation of green tea data set using PCA scores: (a) PC1–PC2 (b) DBSCAN based on the first two PCs.

select some extreme samples for both sets, which is a desirable characteristic. However, it is not expected that the ***y***-values of the samples will be much outside the ***y***-range of the calibration samples and this is the case here for two prediction samples, objects 108 and 119, with ***y***-values of respectively, 14.53 and 16.47, nor is it expected that there would be as much as five samples out of the ***y***-range of the calibration set. It seems therefore possible that some of these objects are atypical (and should not be included) or that some ***y***-values are wrong. To test this, these five samples are individually put into the calibration set to see their influence on the model (Fig. 5).

It is found that object 108 and to a lesser extent object 122 have a very deleterious effect on the RMSECV. For that reason, these two objects are marked as suspect and the root mean square error of prediction (RMSEP) with and without objects 108 and 122 are computed. The result shows that objects 108 and 122 have the highest and the third highest residuals respectively in the test set when five selected PCs are used. Deleting them individually decreases the RMSEP. Sample re-examination finds that objects 108 and 122 indeed have special characteristics in appearance or composition, which might result in wrong measurement of ***X***. Another object in the test set, object 103, has the second highest

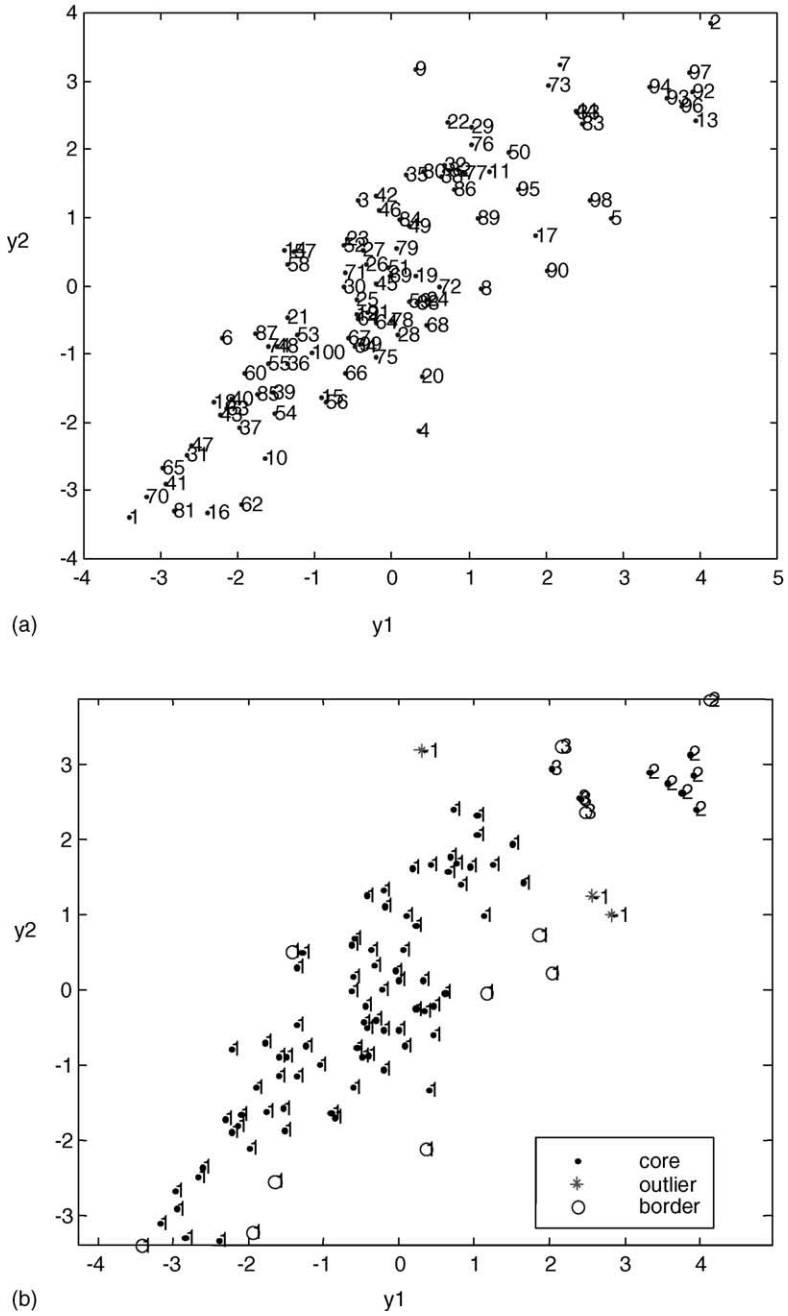


Fig. 3. Clustering investigation of green tea data set by Sammon's mapping: (a) Sammon's mapping using all PCs of the calibration set. (b) DBSCAN based on the result of Sammon's mapping.

residual. It is found that it also increases the RMSECV, when it is included in the calibration set, and the RMSEP. The high residual for this object may due to a wrong ***y***-value for an unknown reason. These three objects are removed from the test set.

3.5. Detection of possible outliers in *X*

3.5.1. Visual methods

The methods described in Section 3.3 allow finding some possible outliers. They are to be found in Table 4.

3.5.2. Mahalanobis distance (MD)

The squared MD is used as a distance measurement to detect outliers. The distance is calculated as [18,19]:

$$\mathrm{MD}_i^2 = (x_i - \bar{x})S^{-1}(x_i - \bar{x})' \quad (9)$$

where x_i is the ith value of x, $\bar{x}$ the mean of the x and S the variance-covariance matrix of the x. The distance is compared with a tabulated χ^2 value with A degrees of freedom (A is the total number of PC or PLS factors used) at the 5% significance level. The 5% objects with highest MD in the calibration set are listed in Table 4.

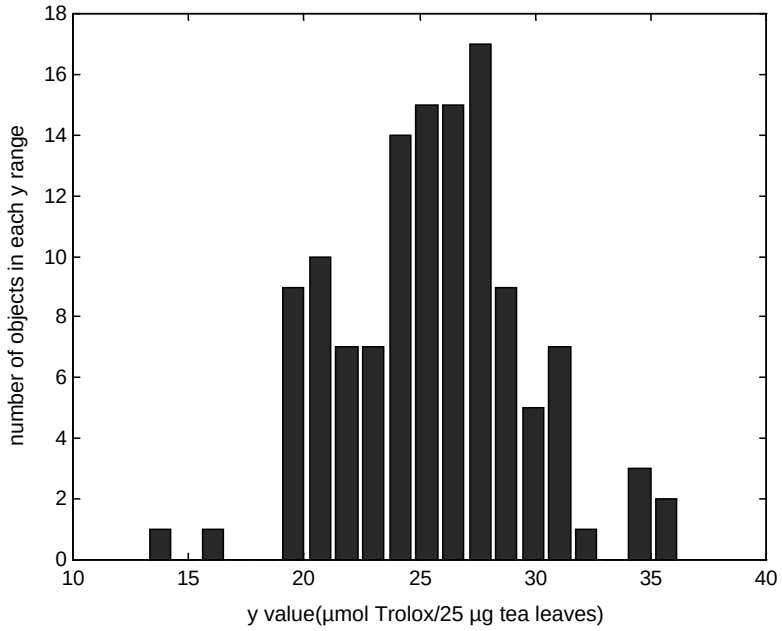


Fig. 4. Distribution of ***y***-values as obtained from the reference method.

*3.5.3. **X** residual standard deviation (XRSD)*

In this method, the total residual standard deviation of the matrix ***X*** (se) and the residual standard deviation of its individual object *i* (se_i) are calculated respectively. If se_i is three times larger than se as recommended by Martens [20], the object *i* is detected as an outlier (Table 4).

$$se^2 = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} e_{ik}^2}{df} \quad (10)$$

$$se_i^2 = \frac{\sum_{k=1}^{K} e_{ik}^2}{K - A} \quad (11)$$

where

$$e_{ik} = x_{ik} - \bar{x}_k - \sum_{a=1}^{A} t_{ia} p_{ka} \quad (12)$$

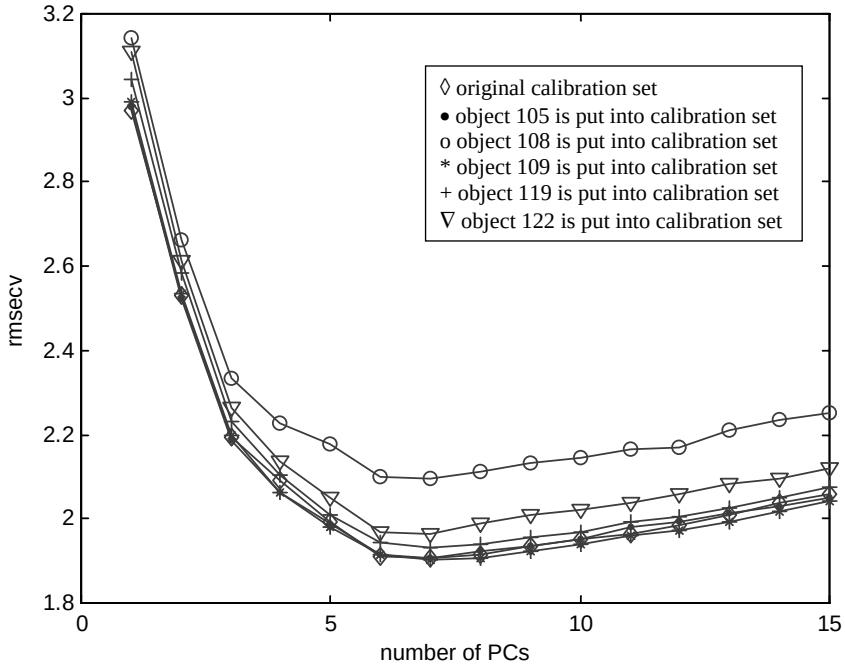


Fig. 5. The influence of five objects from the test set on the RMSECV. The model is based on the calibration set using PCs in the selected order. These five objects (objects 105, 108, 109, 119 and 122) have ***y***-values exceeding the ***y***-range of the calibration objects.

Table 4
Objects flagged as possible outlier by different methods

Method	Calibration set	Test set
PCA plot	4, 9 in PC2; 58, 83, 97 in PC4; 3, 42, 46, 91, 99 in PC5	109 in PC3; 117 in PC5
Sammon's mapping	5, 9, 98	Not applicable
MD[a]	4, 5, 42, 74, 91, 97	117
XRSD[a]	10, 26, 62, 85	109
RHM[a]	4, 5, 9, 42, 74, 91, 97	111, 114, 117
MVT[a]	4, 5, 13, 91, 97	None

[a] Five selected PCs (PC1, 2, 5, 6, 11) are used.

i is the number of objects of the matrix $\boldsymbol{X}$, K is the number of variables, A is the complexity of the PCR or PLS model, t are scores of $\boldsymbol{X}$, p are loadings and df is the number of degrees of freedom:

$$\mathrm{df} = IK - K - A(\max(I, K)) \quad (13)$$

3.5.4. Robust methods

Resampling by the half-means method (RHM) is a robust method to detect extreme objects [21]. In this test, the column mean and the standard deviation of 50% randomly selected samples from the whole matrix are calculated. The whole matrix is then autoscaled by this mean and standard deviation. According to this autoscaled matrix, a matrix of vector lengths for all objects is calculated. Objects are recorded if their vector lengths are larger than the fixed percentage of distribution (95% in our study). The whole procedure is repeated three times the number of objects. If an object has been recorded in many resampling experiments, it is considered to be a possible outlier (Table 4).

Multivariate trimming (MVT) [22] contains a loop that (1) calculates the squared MD of each object in the whole data set, (2) removes a fixed percentage (decided by the analyst beforehand, 70% in our study) of the objects with the highest MD, (3) calculates the covariance matrix and median of the 30% objects with the lowest MD and uses them in step (1) again. The loop is repeated until the covariance matrix and the median become stable. Based on the final median and the covariance matrix, the squared MDs of all objects are calculated. It is shown that objects 4, 5, 13, 91, 97 of the calibration set have the top 5% squared MD when five selected PCs are used (Table 4). None of the objects in test set has higher squared MD than those of these five objects.

In this study, different outlier detection methods are used. Visual methods, MD, XRSD, Grubbs' test are non-robust methods. They are effective for identifying a single outlier or influential observation in a data set. When there is more than one outlier or influential observation, the diagnosis may become difficult due to the masking or swamping effects. For this reason, robust methods, such as RHM and MVT are applied. In most of these methods, a cut-off value of 95% is used. This means that the listed objects in Table 4 are extreme values but might not be bad outliers.

The results from the MD, RHM and MVT are quite similar because they are based on mean (or median) and variance-covariance matrix. The MVT result is a little different because it uses median instead of mean. That the non-robust MD method performs as well as the robust methods might be due to the absence of bad $\boldsymbol{X}$-outliers in our data set.

It is noticeable that the objects detected with XRSD are entirely different from those by MD, RHM and MVT. The main reason may be that different distances are used. In XRSD the Euclidean distance is used and in the rest of the methods the Mahalanobis distance, which takes into account the correlation among variables in the data, is used.

All above methods only consider $\boldsymbol{X}$, not taking $\boldsymbol{y}$ into account. This is the limitation of these methods. Extreme objects detected by these methods mean that they are outside the $\boldsymbol{X}$ population, but they may not be outliers towards the model, i.e. be good outliers.

3.6. Model building

A principal component regression model is built and two measures of predictive ability are used. One is the RMSECV, which is obtained from the calibration set using a leave-one-out strategy. The other is the RMSEP. This is obtained by predicting an independent set, the test set. The modelling includes three steps. First, an initial model based on the calibration set without any refinement is constructed. Then by selecting a suitable pretreatment method, retaining meaningful PCs and eliminating bad outliers, the model is optimized. Finally, the model is validated through internal and external validation.

3.6.1. Selection of the initial model

Building a PCR model with PCs included according to the variance they represent shows that 11 PCs are needed to obtain a RMSECV (1.97) similar to the standard deviation of the reference method (Fig. 6a). A less complex model is preferred since, with increasing number of PCs, there is more chance to include noise and unrepresentative information from the calibration data. Using too many PCs may result in a phenomenon called overfitting, i.e. the calibration data are summarized well but the prediction will be bad. To decrease the complexity of the model, selected PCs are used. The PCs are arranged in decreasing order of correlation with the antioxidant capacity. Fig. 6b shows that the minimum RMSECV (1.91) is obtained using seven selected PCs but we preferred a less complex model with only four PCs (PC2, 6, 11, 1) which still achieves a relatively low RMSECV (2.09). The latter model has no significant difference with the former model when checked by

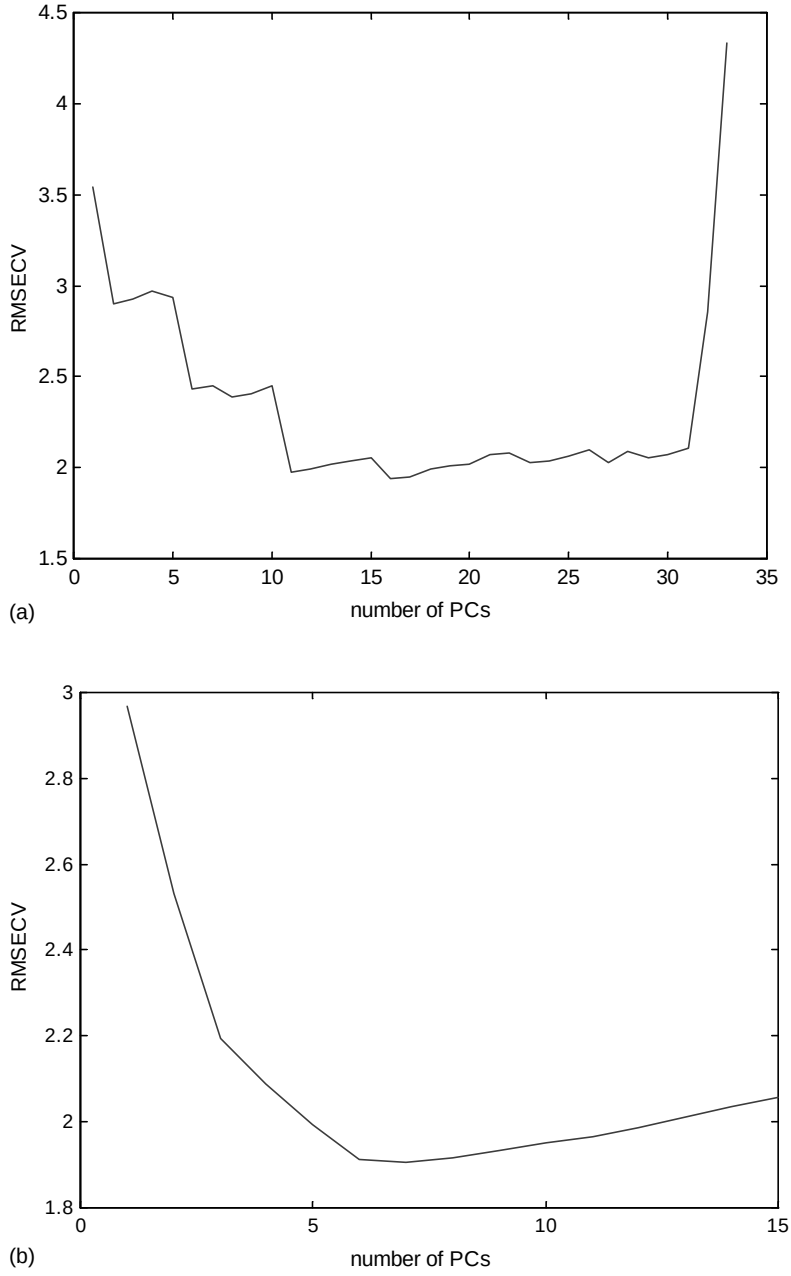


Fig. 6. (a) The RMSECV of the PCR model using top-down PCs. (b) The RMSECV of PCR model using selected PCs.

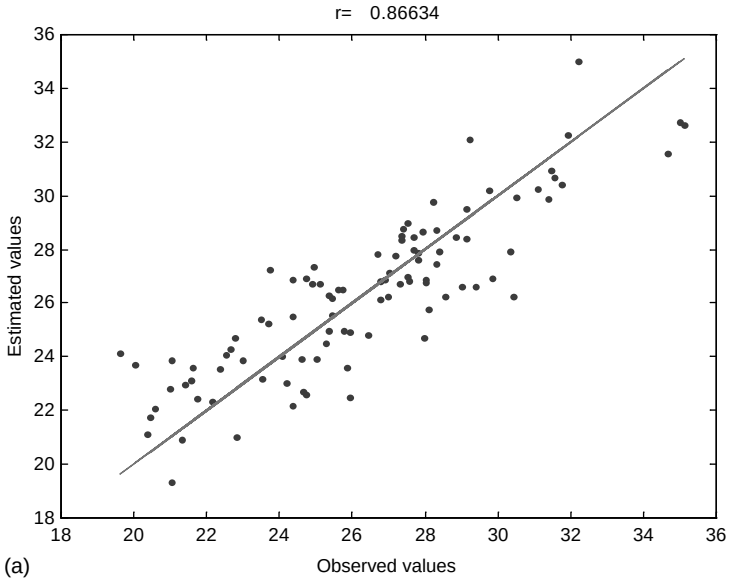
randomization *t*-test [23]. During the validation with the test set, though different samples are removed, the first minimum of RMSEP is always obtained using the five PCs (PC2, 6, 11, 1, 5) (data are not shown). Thus, PC5 is also retained. As a result, PC2, 6, 11, 1 and 5 are used in the further study. The preliminary model using these five PCs for the calibration set has RMSECV equalling 1.99 ($R^2 = 0.7279$) and test set has RMSEP equalling 1.91 ($R^2 = 0.6234$) with no sample deletion. The calibration model was suspected to be slightly non-linear. Therefore, a non-linearity test was performed on the model, but the result was negative.

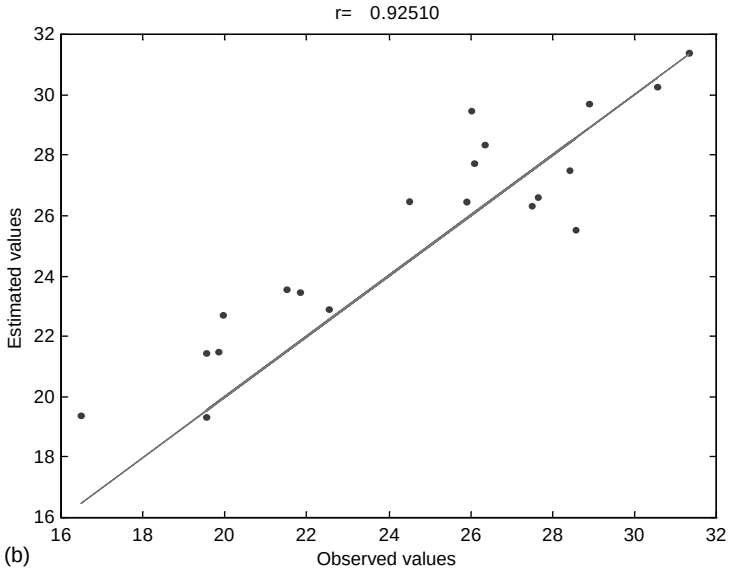
3.6.2. Selection of pretreatment method

Different spectral pretreatment methods, such as offset correction, detrend, multiple scatter correction (MSC), standard normal variate transformation (SNV), first and second derivative are applied to the calibration set. None of the methods improves the RMSECV, nor the RMSEP (Table 5).

3.6.3. Optimization and validation of the model

The possible outliers found in $\boldsymbol{X}$ (see Table 4) are removed from the model to see whether the model can be improved. The result shows that none of them affects the RMSECV





*r: correlation coefficient between the estimated values and the observed values

Fig. 7. The final model for prediction of the total antioxidant capacity in green tea using five selected PCs (PC2, 6, 11, 1, 5): (a) calibration set, (b) test set.

Table 5
The RMSECV and RMSEP of the preliminary model with different pretreatment

Pretreatment	RMSECV	RMSEP
None	1.99	1.91
Offset	2.00	1.94
Detrend	2.01	2.08
MSC	2.18	2.41
SNV	2.15	2.54
First derivative	2.27	2.13
Second derivative	2.02	2.21

No object is eliminated from the calibration set. From the test set, objects 103, 108 and 122 are removed.

or RMSEP much, indicating that they are not bad outliers (Table 6).

The studentized residual [24] for each object in the calibration set is calculated to study the outliers towards the model. Objects whose studentized residuals exceed 2.5 are identified as outliers at 99% confidence level [25]. It is found that objects 6 and 23 of the calibration set, which have the highest residual in the model, are outliers toward the model and they are eliminated. The RMSECV of this model is 1.86 ($R^2 = 0.7505$) when five selected PCs are used and the RMSEP is 1.81 ($R^2 = 0.7557$) (Fig. 7). The bias of the model is −0.08, which is not significantly different from 0.

Table 6
The RMSECV and RMSEP of the model with certain objects deleted (five selected PCs are used and the spectra processed no pretreatment)

Object deleted	RMSECV	RMSEP
–	1.99	1.91
4	1.98	1.86
5	2.00	1.90
6	1.94	1.87
9	1.98	1.90
23	1.94	1.86
42	2.02	2.39
74	2.00	2.25
91	1.96	2.01
97	2.07	1.93
6, 23	1.86	1.81

From the test set, objects 103, 108 and 122 are removed.

4. Conclusions

Our study shows that it is indeed possible to use NIR and chemometrics to estimate the total antioxidant capacity of green tea.

A first conclusion about the outlier study is that different methods give different results. The main conclusion is however that none of the outliers detected has an influence on the RMSEP. They are therefore "good" outliers that extend the range of calibration. Two outliers (objects 6 and 23 of the calibration set) were not detected with the outlier in $\boldsymbol{X}$ or $\boldsymbol{y}$ methods. They are outliers of a different type. The spectra are not outlying, but they do not follow the relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$, probably because of measurement errors either in the reference method or in the NIR study. More generally, it follows that simply identifying outliers by the tests described is not enough: their effect on RMSEP must always be tested. Moreover samples with a high residual towards the calibration model should also be considered potential outliers.

In this paper, a PCR model is built to predict the total antioxidant capacity contained in green tea samples. Different chemometric methods are applied to optimize the model. The best PCR model has a RMSEP of 1.81, comparable to the reference method, for which a pooled SD of 1.83 was found.

References

[1] C.S. Yang, P. Maliakal, X.F. Meng, Annu. Rev. Pharmacol. Toxicol. 42 (2002) 25–54.
[2] Y.D. Jung, L.M. Ellis, Int. J. Exp. Pathol. 82 (2001) 309–316.
[3] S.P. Pillai, C.A. Pillai, D.M. Shankel, L.A. Mitscher, Mutat. Res./Genet. Toxicol. Environ. Mutagen. 496 (2001) 61–73.
[4] C. Rice-Evans, Proc. Soc. Exp. Biol. Med. 220 (1999) 262–266.
[5] J.J. Dalluge, B.C. Nelson, J. Chromatogr. A 881 (2000) 411–424.
[6] ISO (International Standard Organization), ISO TC 34/SC 8 N444, 1994.
[7] G.R. Beecher, B.A. Warden, H. Merken, Proc. Soc. Exp. Biol. Med. 220 (1999) 267–270.
[8] J. Luypaert, M.H. Zhang, D.L. Massart, Anal. Chim. Acta 478 (2003) 303–312.
[9] R. Re, N. Pellegrini, A. Proteggente, A. Pannala, M. Yang, C. Rice-Evans, Free Radic. Biol. Med. 26 (1999) 1231–1237.
[10] R.D. Snee, Technometrics 19 (1977) 415–428.
[11] V. Centner, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 330 (1996) 1–17.
[12] B. Hopkins, Ann. Bot. 18 (1954) 213.
[13] J.A. Fernandez Pierna, D.L. Massart, Anal. Chim. Acta 408 (2000) 13–20.
[14] M. Daszykowski, B. Walczak, D.L. Massart, Chemom. Intell. Lab. Sys. 56 (2001) 83–92.
[15] M. Ester, H.P. Kriegel, J. Sander, X. Xu, in: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, August, 1996.
[16] J.W. Sammon, IEEE Trans. Comput. 18 (1969) 401–409.
[17] F.E. Grubbs, G. Beck, Technometrics 14 (1972) 847–854.
[18] P.C. Mahalanobis, On the generalised distance in statistics, Proc. Natl. Inst. Sci. India 12 (1936) 49–55.
[19] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemom. Intell. Lab. Sys. 50 (2000) 1–18.
[20] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[21] W.J. Egan, S.L. Morgan, Anal. Chem. 70 (1998) 2372–2379.
[22] R. Gnanadesikan, J.R. Kettenring, Biometrics 28 (1972) 81–124.
[23] H. van der Voet, Chemom. Intell. Lab. Sys. 25 (1994) 313–323.
[24] R. De Maesschalck, F. Estienne, J. Verdú-Andrés, A. Candolfi, V. Centner, F. Despagne, D. Jouan-Rimbaud, B. Walczak, D.L. Massart, S. de Jong, O.E. de Noord, C. Puel, B.M.G. Vandeginste, Int. J. Chem. 2 (1999) 19.
[25] International Organization for Standardization, ISO Standards Handbook 3; Statistical Methods, 3rd ed., ISO, Geneva, 1989.