ELSEVIER

Analytica Chimica Acta 408 (2000) 13–20

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Improved algorithm for clustering tendency

J.A. Fernández Pierna, D.L. Massart *

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

Received 20 August 1999; accepted 25 November 1999

Abstract

A modification of the Hopkins algorithm for clustering tendency is described in this study. Detection of clustering is an important problem in multivariate calibration. With this modified algorithm, the automatic detection of clusters, before any modelling, is carried out.

Keywords: Multivariate calibration; Chemometrics; Principal component analysis; Cluster

1. Introduction

An important subject to investigate in multivariate calibration is the homogeneity of the data. In order to guarantee the quality of the model, it is preferable that a data set be as homogeneous as possible [1]. If the data set is clustered, i.e. when the data set contains subgroups of similar objects inside the given population, one might prefer to use local modeling methods instead of global methods [2,3].

There are many possibilities used to detect clustering. The simplest way is the use of plots. A principal component plot can be used to give a good representation of the data because the highest amount of variance is explained by the first eigenvectors. In some cases, however, the clustering will become apparent only in plots of higher PCs so that one should look at several score plots. This is the case for the data studied here as an example. They do not show clustering in the PC1–PC2 plot, but do so in the PC1–PC3 plot. Another useful type of plot is the plot of the *y*-values (characteristic to be modelled) because clustering often becomes evident in the measured property (Figs. 1 and 2). Fig. 2 represents the moisture values for 100 NIR spectra of wheat samples.

An automatic warning using numerical criteria that clusters are present would be a useful diagnostic in multivariate calibration software. Hopkins statistic [4,5] has been proposed, but does not function well in certain cases. In this paper, a variant that should perform better in many cases is proposed.

2. Theory

Hopkins statistic is based on the null hypothesis, H_0, that the objects in a data set are uniformly distributed in multidimensional space and examines whether the observed distribution differs from this assumption.

In order to achieve this, the Euclidean distance from an experimental object to its nearest neighbour (W) is compared with the distance from an artificial object to the nearest real object (U). Artificial and experimental objects are randomly selected over the whole space. If there is a clustering tendency, W will tend to be smaller than U [6].

* Corresponding author. Tel.: +32-2-477-4737; fax: +32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0003-2670/00/$ – see front matter
PII: S0003-2670(99)00879-X

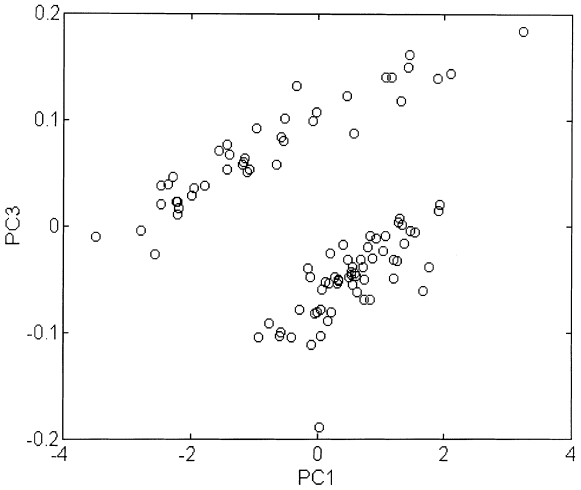

Fig. 1. PC plot (PC1 vs. PC3) for wheat data.

The equation for the Euclidean distance between two objects j and j' is

$$D_{jj'} = (\Sigma_k (x_{jk} - x_{j'k})^2)^{1/2} \quad (1)$$

where $k = 1,\ldots$ are the variables.

The Euclidean distances (U and W) are calculated and the minimal distances for both objects are found. These distances are summed for all objects and used for the determination of Hopkins statistic (H) [4,5]:

$$H = \frac{\Sigma_N U}{\Sigma_N U + \Sigma_N W} \quad (2)$$

where N is the number of selected objects.

This procedure is iterated several times with new random selections and H_{average} is calculated. H_{average} can range from 0.5 when distances U and W are equal (homogeneous set) to 1 when U is much larger than W (extreme clustering). If H is greater than 0.75, then there is more than 90% confidence that the null hypothesis can be rejected.

Forina [7] proposed a modification of this equation that should yield a more stable measure of the clustering tendency. In this study, a new H^* is calculated by using N_{real} (total number of objects) and an arbitrary N_{pseudo} (because the reliability of H depends on N) in order to obtain a clustering index that ranges from 0 to 1 with the critical value equal to 0.5. Eq. (2) is modified as follows:

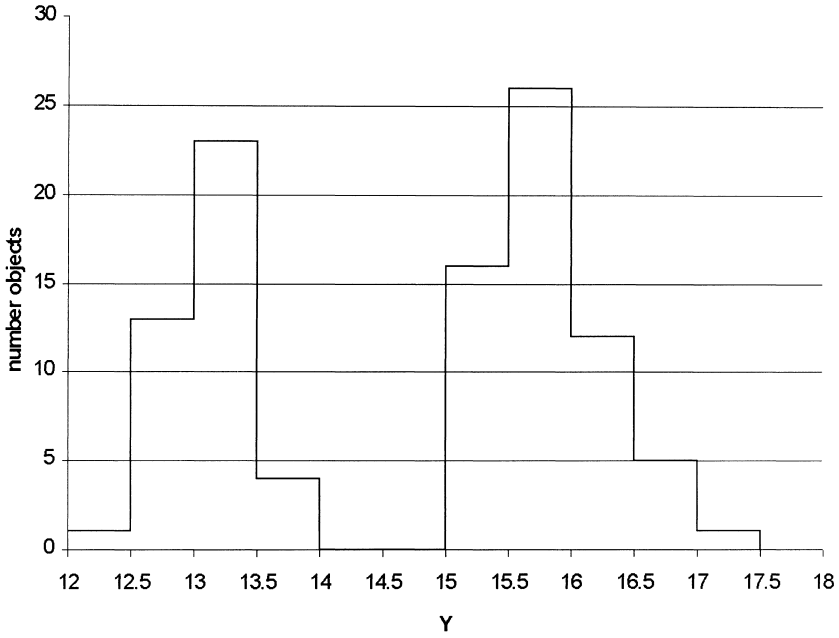

Fig. 2. Y plot for wheat data.

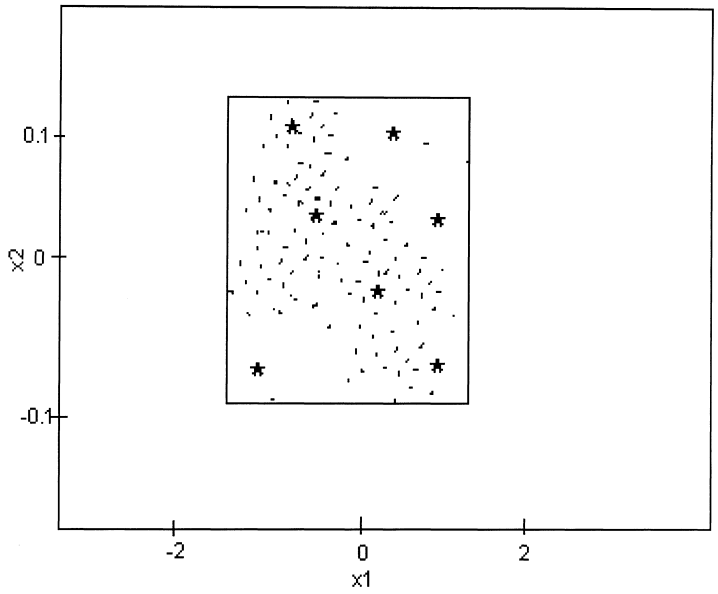

Fig. 3. Univariate space. Artificial (*) objects selected randomly.

$$H^* = \frac{N_{\text{real}}\Sigma_N U - N_{\text{pseudo}}\Sigma_N W}{N_{\text{real}}\Sigma_N U + N_{\text{pseudo}}\Sigma_N W} \quad (3)$$

One important step in this modification is to select correctly N_{pseudo}; it is recommended that a large value be taken for it (~80% of the objects).

Artefacts are possible in the original version and in the modification. Random selection within the domain bounded by univariate *x*-values (Fig. 3) leads to selection of points outside the multivariate limits of the experimental points. The Euclidean distance *U* becomes very large, *H* (Eq. (2)) would also be very large and clusters would be suspected when they do not exist. Even when one works in the PC space, for example PC1 versus PC2 (Fig. 4a), the same problem can be found although to a lesser extent. It is therefore necessary to assure that all artificial objects are selected only within the boundaries of the experimental objects (Fig. 4b).

To do this, a boundary is constructed through the extreme experimental objects around the whole data set [8] in the score plot. The first vertex of this figure is obtained as the point at the maximal distance of the centroid (mean of each of the two respective PCs), and a line b1 (Fig. 5a) is drawn between the centroid and this point. The minimum distances ($d1$) from each experimental point to b1 and the distances ($d2$) from these experimental points to the first vertex are calculated by using Eqs. (4) and (5). The next selected vertex is the point that forms a maximal angle between the line connecting the first chosen vertex and the candidate vertex and the line b1 according to $A = \text{arcsine}\,(d1/d2)$:

$$d1 = \frac{|Ax + By + C|}{(A^2 + B^2)^{1/2}} \quad (4)$$

where $Ax + By + C$ is the equation of the line.

$$d2 = ((x_2 - x_1)^2 + (y_2 - y_1)^2)^{1/2} \quad (5)$$

where $x_{1,2}$ and $y_{1,2}$ are the coordinates of each point.

A line (g1) between the two vertices is drawn. For the next selected vertex, the distances from each experimental point to the last selected vertex are calculated and the new selected vertex is the point for which the line forms a maximal angle with g1, a line g2 is drawn and so on.

Once the boundary is complete, random real objects and artificial objects are selected as before. Only artificial objects inside the boundary are included.

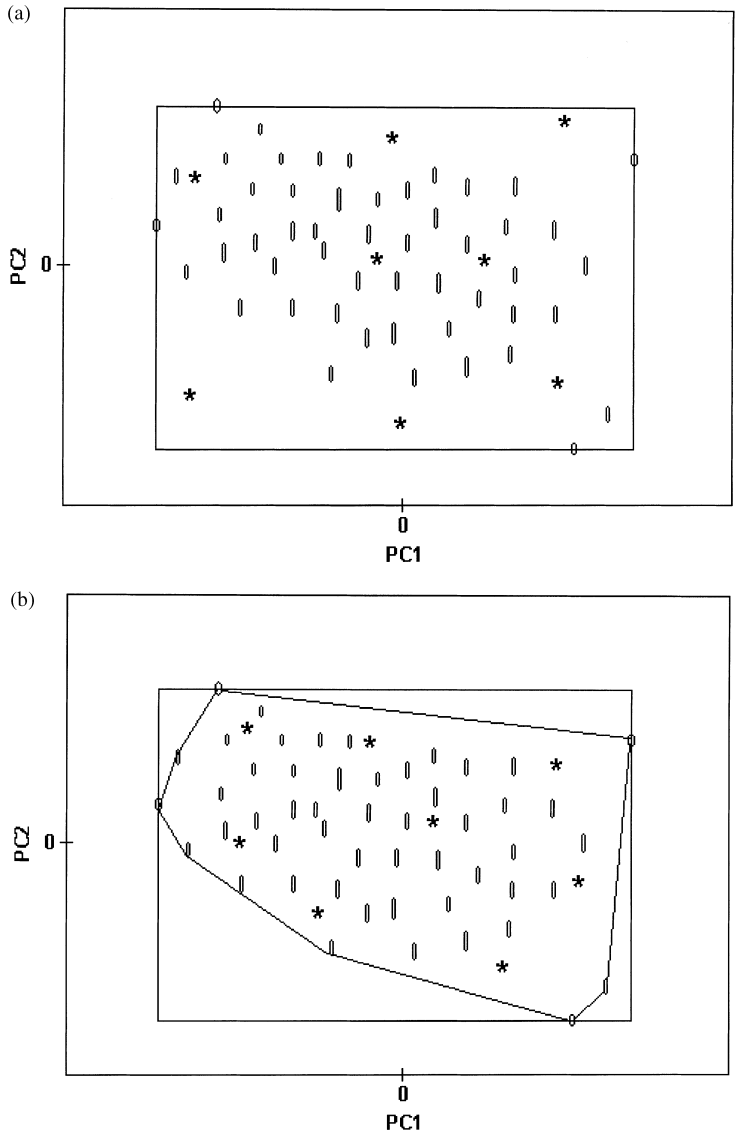

Fig. 4. PC1–PC2 plot. Artificial (*) objects selected randomly. (a) Artificial objects are selected over the univariately-determined space and (b) artificial objects are selected only in the range limited by the experimental objects.

3. Experimental

The data set consists of 100 NIR spectra of wheat samples (**X**-matrix) for the determination of moisture (**y**-matrix). The spectra were measured between 1100 and 2500 nm at each 2 nm interval in a Bran + Luebbe instrument. The data were corrected for offset, and are known to consist of two clusters.

This data set was obtained from the database of Chemometrics and Intelligent Laboratory Systems [9].

4. Computer programs

All procedures were programmed in Matlab for Windows, version 4.0.

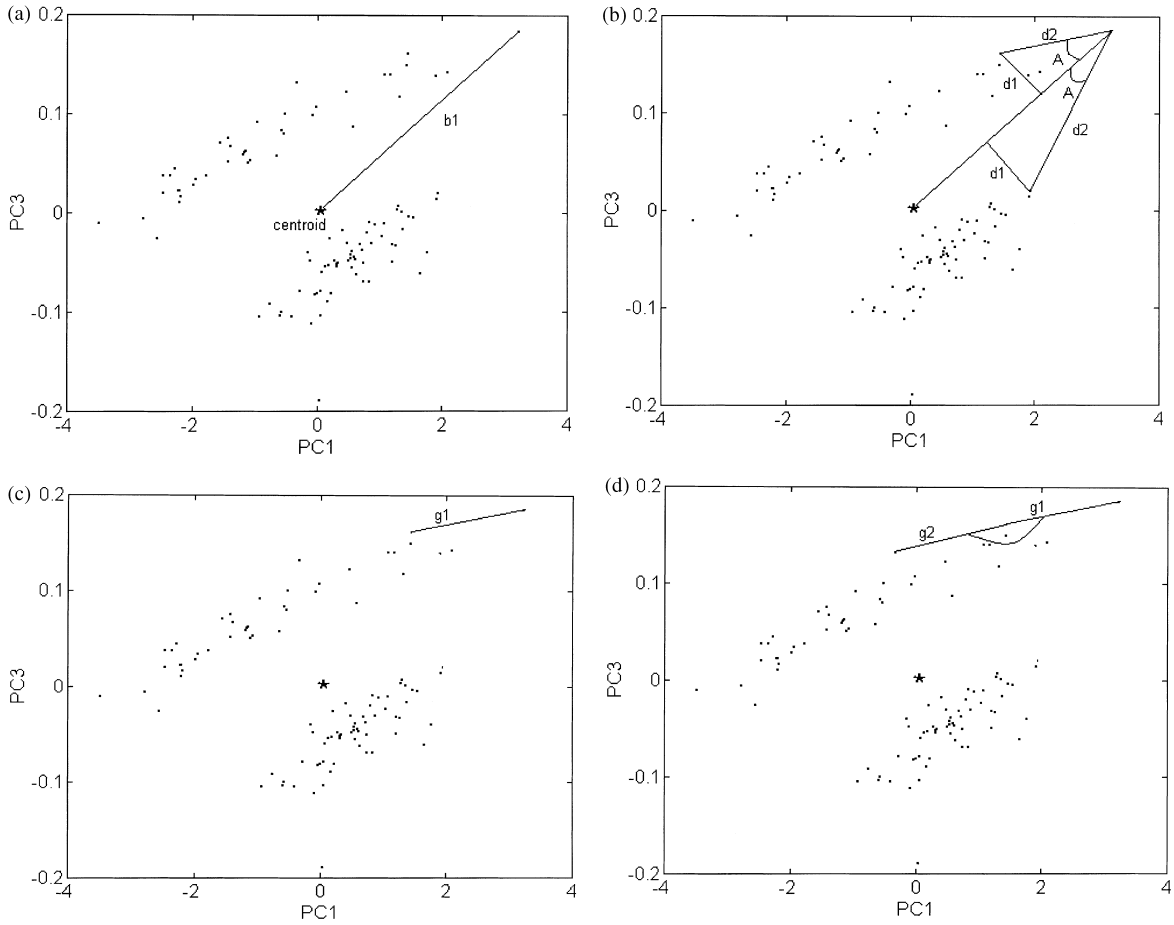

Fig. 5. Construction of the boundary.

5. Results and discussion

Hopkins algorithm is repeated several times. Centner et al. [10] recommend choosing the number of repetitions so that each of them uses more than 5% of the total population. In their article, they present a table with all combinations of population size and number of iterations. According to this table, it was decided that three repetitions and 10% of the population be used.

Three examples are proposed to demonstrate the modified algorithm. Before using the algorithm, a visual first study was performed to look for inhomogeneities, such as outliers and clusters. To do this, the PC plots are used. The PC plots for the centered data are given in Fig. 6. In the PC1–PC2 plot (Fig. 6a), no clusters were found, but when the PC1–PC3 plot (Fig. 6b) is studied, two clusters become evident. Outliers can also be found but the Hopkins algorithm will be applied before their elimination to study their possible influence.

In the space of PC1 versus PC3 where clusters are present, the original and the modified algorithm were applied and the results are presented in Table 1.

Both algorithms detect the presence of clusters, but the modified Hopkins algorithm gives clearer results than the original because the value for the Hopkins statistic (H_{average}) is larger. The original algorithm yields results that are only marginally higher than the critical 0.75 value. The modification introduced by Forina combined with the modification proposed here gives still better results for this case. Table 1 shows

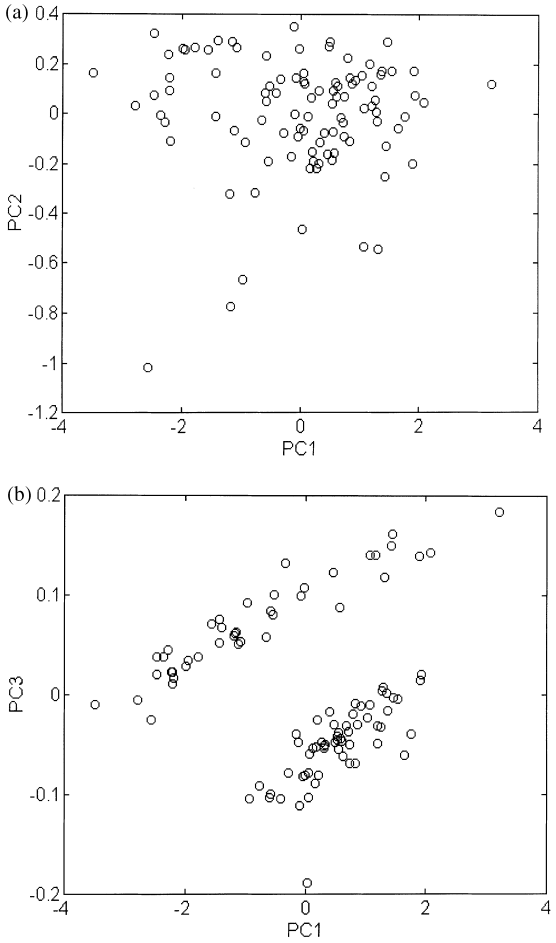

Fig. 6. (a) Plot of PC1 and PC2 after centering and (b) plot of PC1 and PC3.

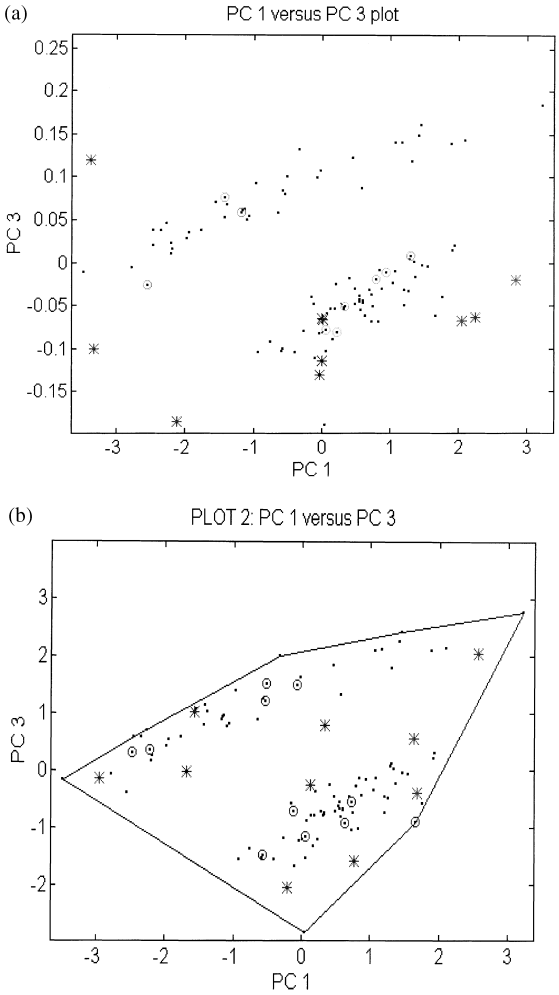

Fig. 7. Experimental (○) and artificial (*) objects selected randomly for one iteration. (a) Original Hopkins algorithm; (b) modified Hopkins algorithm.

the results and the best models are indicated by the boldfaced *H*-values.

Fig. 7 shows the selection of the experimental points by both algorithms.

The next example is a case where no clusters are present and both methods (original and modified algorithms) give good results. This example is useful to see the limits of clustering detection of both methods and shows that the modified algorithm also gives clearer answers even in the case where no clusters are present.

Table 1
Hopkins statistic for an example with clear clusters

	$H_{average}$	H_{min}	H_{max}	Range
Original algorithm	0.7767	0.7333	0.8106	0.0772
Modification Forina ($N_{ps} = 20$)	0.8312	0.7784	0.8987	0.1204
Modification Forina ($N_{ps} = 80$)	0.5691	0.5084	0.6299	0.1215
Modified algorithm	**0.8777**	**0.7957**	**0.9581**	**0.1624**
Modified algorithm + modification Forina (N_{pseudo}: 20)	**0.9195**	**0.8652**	**0.9718**	**0.1066**

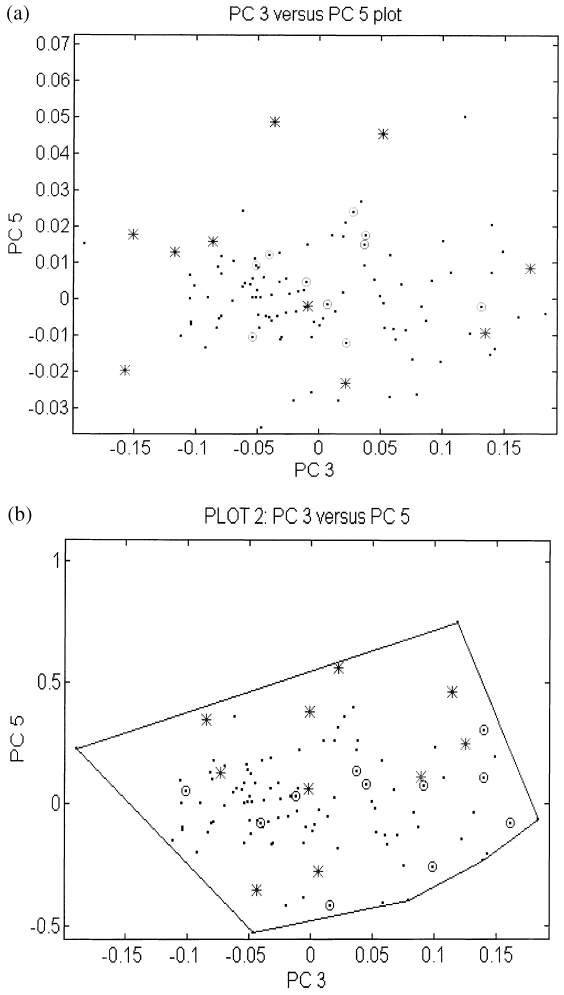

Fig. 8. Experimental (○) and artificial (*) objects selected randomly for one iteration. (a) Original Hopkins algorithm; (b) modified Hopkins algorithm.

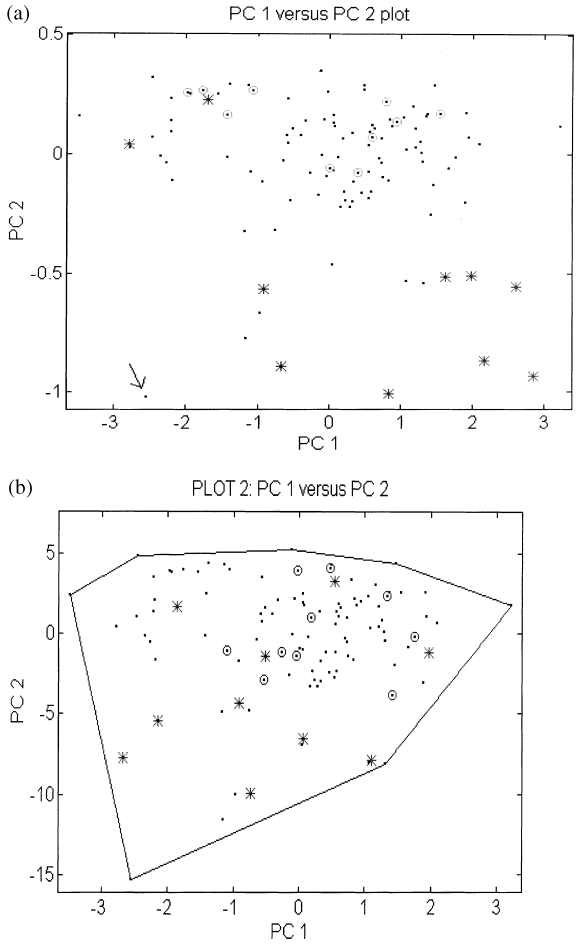

Fig. 9. Experimental (○) and artificial (*) objects selected randomly for one iteration. (a) Original Hopkins algorithm; (b) modified Hopkins algorithm; arrow: point added to the data set.

Table 2 and Fig. 8 show the results for this example.

In both cases, a value less than 0.75 is obtained but the H_{average} for the modified algorithm is better because, as no clusters exist, a smaller value for the Hopkins statistic is expected. The selection of N_{pseudo} in Forina's algorithm seems to be important as rather bad results are obtained when N_{pseudo} is 20 but good results are obtained when it is 80.

Table 2
Hopkins statistic for an example without clusters

	H_{average}	H_{min}	H_{max}	Range
Original algorithm	0.6635	0.6120	0.7317	0.1197
Modification Forina ($N_{\text{ps}} = 20$)	0.7995	0.7489	0.8731	0.1242
Modification Forina ($N_{\text{ps}} = 80$)	**0.4079**	**0.3888**	**0.4170**	**0.0181**
Modified algorithm	**0.6089**	**0.4383**	**0.6948**	**0.1909**
Modified algorithm + modification Forina (N_{pseudo}: 20)	0.6423	0.4742	0.7102	0.1237

Table 3
Hopkins statistic for an example without clusters

	$H_{average}$	H_{min}	H_{max}	Range
Original algorithm	0.7933	0.7731	0.8246	0.0515
Modification Forina ($N_{ps} = 20$)	0.8244	0.7790	0.8905	0.1114
Modification Forina ($N_{ps} = 80$)	**0.4700**	**0.4443**	**0.4956**	**0.0513**
Modified algorithm	**0.6707**	**0.6443**	**0.700**	**0.0557**
Modified algorithm + modification Forina (N_{pseudo}: 20)	0.7012	0.6742	0.7402	0.1237

For the next example, the data set was modified; the modification consists of the introduction of an extreme value.

In this example, the application of the two algorithms leads to the results presented in Table 3.

In the first and second cases, a value of H larger than 0.75 is obtained, so that it would be wrongly concluded that clusters are present. Concerning the algorithm by Forina, the same conclusions are reached as for the preceding example and the PC plots also give the same results as the previous example (Fig. 9).

The presence of the extreme value is responsible and we can conclude from the results in this data set that the new algorithm is less influenced by the presence of extreme values.

6. Conclusions

The modified Hopkins algorithm enables an automatic warning that clustering occurs. The results obtained agree to a great extent with the visual observation on PC plots. The boundary drawn through the multivariate limits of the experimental points allows to obtain better and more reliable results in order to detect and quantify clusters. In certain cases, good results are also obtained with Forina's method. These results can be better than those of the original Hopkins algorithm, but this depends on the choice of N_{pseudo}.

References

[1] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.
[2] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.
[3] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Vol. 2, Elsevier, Amsterdam, 1988.
[4] B. Hopkins, Ann. Bot. 18 (1954) 213.
[5] R.G. Lawson, P. Jurs, J. Chem. Inf. Comput. Sci. 30 (1990) 137.
[6] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.
[7] M. Forina, personal communication.
[8] A. Thielemans, H. de Brabander, D.L. Massart, J. Assoc. Off. Anal. Chem. 72 (1) (1989) 41.
[9] J. Kalivas, Chemom. Intell. Lab. Syst. 37 (1997) 255–259.
[10] V. Centner, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 330 (1996) 1.