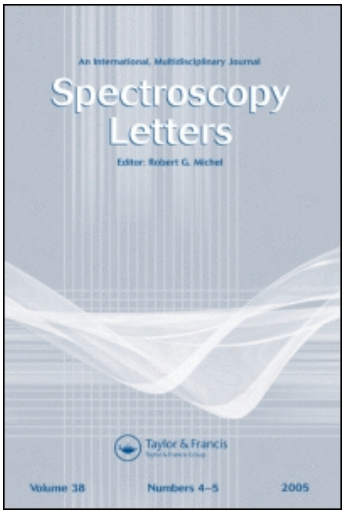

Spectroscopy Letters
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597299

Multivariate Calibration with the Delaunay Triangulation Method: Definition of the Calibration Domain

L. Jin[a]; Q. S. Xu[a]; D. L. Massart[a]
[a] ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Brussels, Belgium

To cite this Article Jin, L. , Xu, Q. S. and Massart, D. L.(2005) 'Multivariate Calibration with the Delaunay Triangulation Method: Definition of the Calibration Domain', Spectroscopy Letters, 38: 6, 787 — 807
To link to this Article: DOI: 10.1080/00387010500316148
URL: http://dx.doi.org/10.1080/00387010500316148

Spectroscopy Letters, 38: 787–807, 2005

ISSN 0038-7010 print/1532-2289 online
DOI: 10.1080/00387010500316148

Multivariate Calibration with the Delaunay Triangulation Method: Definition of the Calibration Domain

L. Jin, Q. S. Xu, and D. L. Massart
ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Brussels, Belgium

Abstract: The Delaunay triangulation (DT) method for multivariate calibration is a topological multivariate calibration method. In this paper, we present methods for the definition of the calibration domain. Outliers in the calibration set must be found and deleted and clusters detected. When clusters are found, it may be advantageous to make separate local models. Two methods are proposed. The first, called the DT calibration domain algorithm, is based on finding a kernel of samples that is then extended according to local rules. An alternative is to first eliminate gross outliers and then divide the data set in clusters, if such clusters exist, with Dbscan, a density-based clustering method. The cluster(s) is (are) then used as kernels(s) and extended with the same rules as the DT calibration domain algorithm to develop DT models for each cluster. The two methods and some of the difficulties that can be encountered with them are demonstrated with three simulated data sets and tested with three real NIR data sets (one agricultural, one food, and one industrial). It is shown that the methods perform well and are at least comparable in prediction performance to partial least squares (PLS).

Keywords: Dbscan, Delaunay triangulation, multivariate calibration, near-infrared spectroscopy, topological methods

Received 5 November 2004, Accepted 9 February 2005

This paper was by special invitation as a contribution to a special issue of the journal entitled "Quantitative Vibrational Spectrometry in the 21st Century." This special issue was organized by Professor Miguel de la Guardia, Professor of Analytical Chemistry at Valencia University, Spain.

Address correspondence to D. L. Massart, ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium. E-mail: fabi@vub.ac.be

INTRODUCTION

The Delaunay triangulation (DT) method for multivariate calibration[1] was recently proposed as an alternative to principal component regression (PCR) and partial least squares (PLS). It is a so-called topological method. Topological methods determine the property value, such as the concentration of a certain component or a physicochemical characteristic, of a new sample as a weighted mean of this value for the neighboring samples. For simplicity, we will call the property value concentration in what follows. Topological methods are local methods, meaning that the concentration of a new sample is determined by calibration samples that are close to it. Regression-based methods such as PLS and PCR are global methods, in the sense that the relationship between concentration and spectra is based on all samples of the calibration set. Compared to PLS and PCR, there has been little research published about topological methods for multivariate calibration.[2–4] However some methods such as Topnir[3] have been applied successfully in industry.

Delaunay triangulation was first developed in a very different context by Delaunay and is used for instance in crystallographic and geographical applications.[5–7] It consists in creating a mesh or network of points, such that each of them is part of a triangle (in two dimensions; see Fig. 1a), or the appropriate simplex with $K + 1$ points (in K dimensions). The network is optimized such that the circumsphere of any simplex contains no other points.

In the DT multivariate calibration method, a DT mesh is first obtained with the calibration samples. A new sample, the concentration of which has to be predicted, is then considered as a mixture of the calibration samples that constitute the triangle or simplex that contains it. If the number of variables is not too large, the DT mesh is constructed in the original data space. Otherwise, for instance when the variables are spectra as is the case in the applications we will describe, it is constructed in the PC-space. The number of dimensions is first reduced by obtaining PC scores for the calibration data and the scores are used as new variables.

The DT method for multivariate calibration was first presented in Ref. [1] and it was shown to have good prediction properties for new samples. Several expected advantages are discussed in the same reference. For instance, because it is a local method, there should be less problems with nonlinearity. However, as it is a new method, several aspects require further attention. One of them is the proper definition of the calibration set. As in other multivariate calibration methods, this consists of many samples, for which the concentration has been determined. These samples are collected in such a way that it can be hoped that all sources of variation, which could have an influence on the relationship between concentration and spectrum, are represented. These sources of variation are often not known, and therefore many samples are collected in a way that ensures diversity (e.g., by taking samples from different lots or origins).

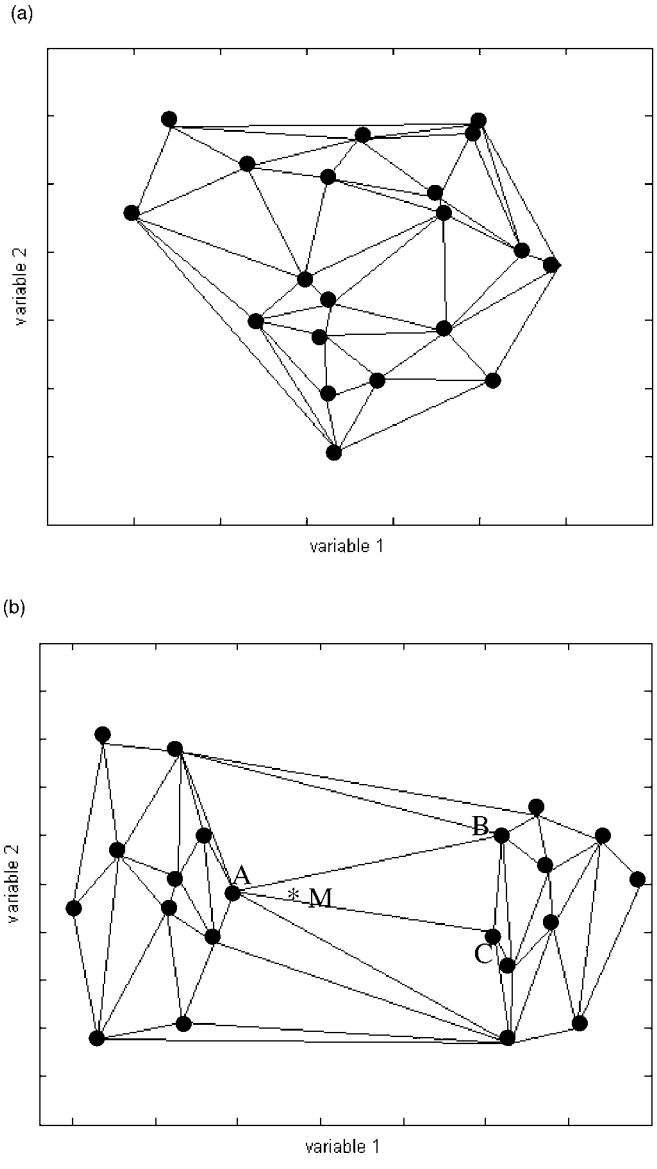

Figure 1. (a) A DT mesh in two dimensions; (b) a DT mesh when two clusters exist; M is a sample whose concentration must be predicted.

As in any other calibration method, the calibration set should not contain outliers. Also, if there are clusters, which are relatively far from each other, this can be a problem. For instance, the concentration of sample M in Fig. 1b would be computed as a weighted average of samples A, B, and C,

although two of them are far away from M. The problem does not arise when two clusters are close to each other. As we will explain later, it is possible to predict the concentration of samples that are close enough to a DT mesh and, when two clusters are close to each other, a sample that falls between the two clusters will also be close enough to at least one of the clusters.

In the DT method for multivariate calibration, separate DT meshes are made for each cluster. A local method is needed to detect for each cluster of the calibration set, which samples are part of it and to avoid inclusion of outliers and samples belonging to another cluster. We call this method the DT-calibration domain method. An alternative method based on defining clusters with a clustering method is also described.

THEORY

The DT Multivariate Calibration Method

The first step of the DT method is to construct a mesh of simplexes for the calibration data with the Delaunay triangulation method. The DT triangulation algorithm is described in Barber et al.[8]

Throughout the article, we will introduce the method in two dimensions, which means that the simplex is a triangle. However, the method can easily be generalized to higher dimensions. In Matlab 6.5, there is a function that allows construction of the DT mesh in K-dimensional space.

In the DT method for multivariate calibration in two dimensions, the following equations are used to obtain the coefficients of a new sample M, which should be predicted, with respect to the $k = 3$ neighbors (M_1, M_2, M_3) that surround it.

$$\alpha_{M_1} = \frac{(x_{2M} - x_{2M_2})(x_{1M_3} - x_{1M_2}) + (x_{1M} - x_{1M_2})(x_{2M_2} - x_{2M_3})}{(x_{2M_1} - x_{2M_2})(x_{1M_3} - x_{1M_2}) + (x_{1M_1} - x_{1M_2})(x_{2M_2} - x_{2M_3})} \quad (1)$$

$$\alpha_{M_2} = \frac{(x_{2M} - x_{2M_1})(x_{1M_3} - x_{1M_1}) + (x_{1M} - x_{1M_1})(x_{2M_1} - x_{2M_3})}{(x_{2M_2} - x_{2M_1})(x_{1M_3} - x_{1M_1}) + (x_{1M_2} - x_{1M_1})(x_{2M_1} - x_{2M_3})} \quad (2)$$

$$\alpha_{M_3} = \frac{(x_{2M} - x_{2M_2})(x_{1M_1} - x_{1M_2}) + (x_{1M} - x_{1M_2})(x_{2M_2} - x_{2M_1})}{(x_{2M_3} - x_{2M_2})(x_{1M_1} - x_{1M_2}) + (x_{1M_3} - x_{1M_2})(x_{2M_2} - x_{2M_1})} \quad (3)$$

where x_{1i} and x_{2i} are the values of the two variables in the original x-space or the scores of the objects in the PC-space depending on the situation, and α_{M1}, α_{M2}, and α_{M3} are the contribution of samples M_1, M_2, and M_3, respectively. The sum of the coefficients is always 1:

$$\alpha_{M_1} + \alpha_{M_2} + \alpha_{M_3} = 1$$

For the new samples, the concentration of which has to be predicted and which are found to be inside the mesh of the convex hull containing the calibration

data set, the coefficient limits are in the range [0, 1]. When a new sample is found to be outside the mesh, it is in principle still possible to predict its concentration. This can for instance be done by using the closest triangle(s). At least one coefficient will then be negative. Because the sum of the coefficients is 1, the coefficient limits must then be extended to values larger than 1. When the new sample is far from the calibration data, the negative coefficient is large and the limits will be extended to a large extent. An accurate prediction of the sample is then not possible. Earlier work showed that the prediction remains acceptable as long as the coefficients are in the range $[-1, 2]$. Samples that have at least one coefficient outside the range [0, 1], but for which the coefficients are within the range $[-1, 2]$ are outside the convex hull of the calibration samples, but close to it. They are called *marginal outliers* and can be predicted correctly. Objects that have at least one coefficient outside the range $[-1, 2]$ are considered to be too far away from the calibration data. We will call them *true outliers.*

The DT Calibration Domain Method

The DT method is a local method and a methodology for detection of samples not belonging to the calibration domain should be based on local characteristics. The algorithm we propose is based on two steps and is explained here in two dimensions. The steps are shown in Fig. 2.

When spectra are used without prior feature selection, the PC scores of the objects are used. They are obtained after column centering of the data.

Before starting with the algorithm, the data set is studied with a number of diagnostic methods (see further) with the aim of detecting clear-cut clusters. Although this is not necessary for the functioning of the method, gross outliers can be removed at the same time. If clear-cut clusters exist, a separate DT mesh will be obtained for each of the clusters that is considered large enough. This also means that the DT calibration domain method is applied to each cluster.

The first step of the algorithm consists of obtaining a *kernel calibration data set*. This set consists of the samples close to the center of the majority group in the calibration data. The "majority" group means that the group contains more than half the number of samples, and 51% samples is usually selected in the kernel set.

In the second step, the kernel calibration data set is extended with those samples that are not too far removed from it and can therefore be considered as compatible with the kernel set. The resulting data set is called the *cleaned calibration data set*. The samples in the kernel set are considered as seeds to obtain this data set. Calibration samples that are not part of the clean data set are set apart. It may happen that (a majority of) the remaining samples can form a second (or third,...) cluster that can be handled in the same way as

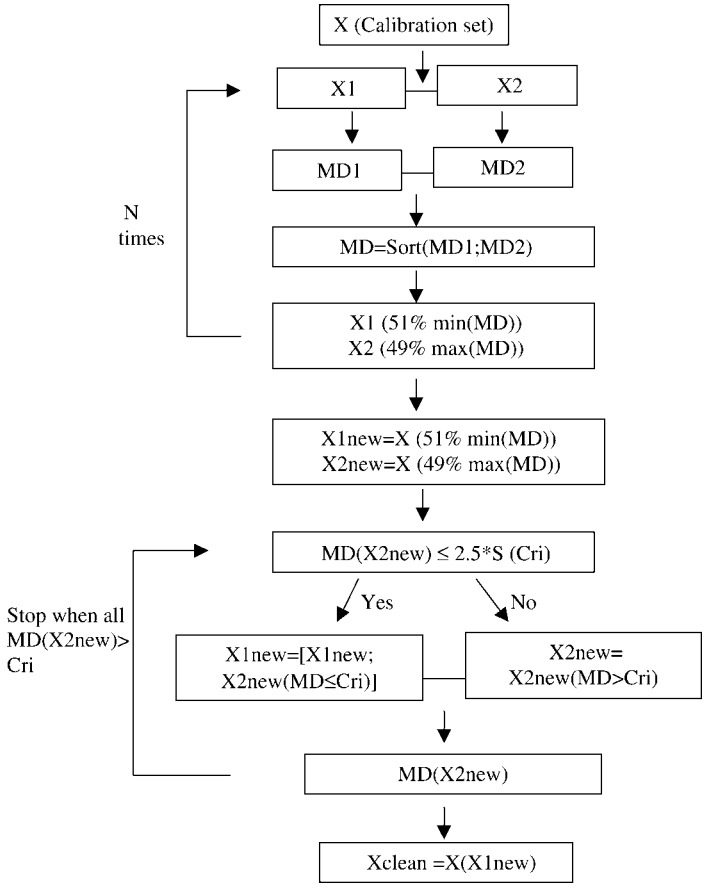

Figure 2. Steps of the DT calibration domain method. Xclean is the cleaned calibration set.

the first cluster. With 51% samples selected in the kernel set first, a stable subset for the *cleaned calibration data set* can be obtained in most cases.

The Kernel Calibration Data Set[9]

The kernel calibration set is defined as the set containing the 51% samples closest to the center of the majority group.

The algorithm is initiated by randomly splitting the calibration set into two subsets X1, the initial input of the kernel calibration data set, and X2, the other objects. The number of initial objects in X1 is $K + 1$ (K is the number of dimensions).

In a second step, the Mahalanobis distance (MD)[10,11] between each sample and the center of the objects in X1 is calculated. The MD between object M and the center is calculated by using the following equation:

$$MD = \sqrt{(\boldsymbol{t}_M - \bar{\boldsymbol{t}}_p)\boldsymbol{C}_{tp}^{-1}(\boldsymbol{t}_M - \bar{\boldsymbol{t}}_p)'} \quad (4)$$

where $\boldsymbol{C}_{tp} = [(\boldsymbol{T}_p - \bar{\boldsymbol{t}}_p)'\,(\boldsymbol{T}_p - \bar{\boldsymbol{t}}_p)]/(r-1)$, $\boldsymbol{t}_M$ (1 × K) is the score of sample M, $\boldsymbol{T}_p$ (r × K) are the scores of the samples selected in X1, $\bar{\boldsymbol{t}}_p$ (1 × K) is the mean column vector of $\boldsymbol{T}_p$, and **r** and **K** are the number of samples in X1 and the number of dimensions, respectively. When the number of variables of the original data is not too large, the variables in the original *x*-space are used instead of the scores for obtaining the MD values.

In a third step, the MD values of all samples are ranked. Fifty-one percent of the objects that have the smallest MD values are selected and put into X1 and the rest is kept in X2. They constitute the first kernel.

The second and third step are repeated N times to optimize the kernel. The 51% objects with smallest MD values at that stage constitute the final kernel calibration set (X1new). The other objects are put into X2new.

The structure of the algorithm is as follows:

```
K + 1 objects (X1) are selected randomly from the calibration set
for each iteration from 1 to N
   for each object i
      Calculate the MD of the object to the selected (new) X1
      end
      51% of the objects that have the smallest MD values are selected as
      new X1
 end
The 51% objects with smallest MD values are considered as the kernel
calibration set X1new. The others are put into X2new.
```

Extension of the Kernel Calibration Data Set

In order to extend the kernel set of the calibration data (X1new) with those samples that are sufficiently close to it, a criterion (Cri) is needed. A DT mesh is constructed with the data in X1new. To calculate the MD value of a sample in X2new, the MD of the samples in X2new to each triangle in the DT mesh is computed and the smallest is selected as the sample's MD value. The MD value between the vertices and the center of the triangle, which contains the vertices, is a constant S. In two dimensions, S = 1.1547, and in higher dimensions $S = \sqrt{K^2/(K+1)}$ (where K is the number of dimensions). Samples from X2new are considered close enough and are transferred to X1new if their MD ≤ Cri = 2.5* S, yielding an extended X1new.

When all samples from X2new have been considered, the MD for the objects remaining in X2new are considered again toward the extended X1new and included if they satisfy the criterion. This is repeated until no samples from X2new can be transferred X1new. In the process, the number of samples in X1new is extended step by step and the number of samples in X2new becomes smaller and smaller.

The remaining samples in X2new are considered not to belong to the cluster investigated and therefore also not to the calibration domain. The samples in X1new constitute the cleaned calibration data set (Xclean). It includes a majority group of the samples investigated, as at least 51% of the samples are included.

For the extension of the kernel calibration set, the structure of the algorithm is as follows:

```
for each object in X2new
   the MD of the object to the triangles constructed by the (extended)
   kernel set is calculated and the minimal one is selected
end
if MD ≤ Cri
   the object is added to extend the kernel set and deleted from X2new
else
   the object is remained in X2new
end
The previous steps are repeated until no object can be added to extend
the kernel set.
```


Prediction of the Concentration for New Samples

The cleaned calibration data Xclean is used as the calibration set to construct a DT mesh. The concentration of new samples can be predicted adequately if the sample is found to be within the convex hull of the data or a marginal outlier, that is, when all α-coefficients are within the range $[-1, 2]$.

New samples with at least one coefficient outside the range $[-1, 2]$ are considered true outliers toward the calibration set.

Diagnostic Methods

The proposed DT calibration domain method performs well in the presence of outliers and minority clusters. However, as with any calibration method, it is preferable not to work blindly and to investigate the data structure before starting with the calibration. At this stage, gross outliers can be eliminated and the clustering structure investigated. When the main cluster contains less than half of all samples, the DT calibration domain method will not

function properly. Although this can be corrected after the cleaned calibration set has been obtained, it is preferable to look at this stage for the existence of clear clusters. Therefore, a number of diagnostic methods are applied before starting the DT calibration domain algorithm. Visualization methods and a clustering method are used for that purpose. The first visualization method to be used is of course PCA. PCA is directed toward finding the largest variation in the data set. This is often caused by clustering, and in that case the PCA may allow to observe the clustering. In some cases, however, it is not apparent in the plots of the first PCs and may go undetected. Therefore, it is complemented with methods that find directions in multivariate space along which clusters are most evident and projects the samples on planes determined by these directions. For this purpose, we use projection pursuit (PP) with the Yenyukov projection index as criterion.[12,13]

Clustering is usually carried out with hierarchical or partitioning methods.[14,15] They have the disadvantage for our application that they always give a clustering, even when there is only one cluster, and that they tend to select clusters of a given (e.g., round) form, depending on the algorithm applied. Density-based approaches are much less known but do not suffer from these disadvantages. The density-based approach used here is the Dbscan method. It determines the number of clusters based on the characteristics of the data and can detect clusters of any form. It is also able to detect outliers, defined in this case as objects that are not close enough to enough other objects.

The method was originally proposed by Ester et al.[16] and first applied to chemical data by Daszykowski et al.[17] It is based on determining how many objects are situated in a given neighborhood of a certain object. If there are more than a certain minimum, that object is considered to be part of a cluster with the objects in the neighborhood. To apply the method, it is necessary to define two parameters: the minimal number of objects in the neighborhood (Minpt) and the radius of the neighborhood (ε). Minpt and ε were defined as proposed in Daszykowski et al.[17] and Ankerst et al.[18].

$$\text{Minpt} = \text{integer}(\text{m}/25) \tag{5}$$

$$\varepsilon = \sqrt[\text{K}]{\frac{\text{V} \times \text{Minpt} \times \Gamma(\text{K}/2+1)}{\text{m} \times \sqrt{\pi^{\text{K}}}}} \tag{6}$$

where m is the number of objects in the data set, K is the number of dimensions, and Γ is the gamma-function. V denotes the volume of the K-dimensional hypersphere, which is formed by the same number of objects as the experimental data set in the same range but uniformly distributed.

According to the determined Minpt and ε, objects situated in a region with relatively high density can be detected to form cluster(s).

An Alternative Approach for Determining the Calibration Domain

Based on the PC and projection pursuit plots, the analyst can try to define the clusters himself and use these, or large parts of these, as kernels that can then be extended to obtain the cleaned calibration set. However, as each plot is two-dimensional, it would not always be easy to decide exactly which samples to include and exclude. It is however possible to define clusters with Dbscan, which is done in multivariate space. The cluster or clusters can function as kernels, to which the extension step can then be applied.

EXPERIMENTAL

Three simulated and three real NIR data sets are used. The first real data set (the meat data) contains the spectra of 198 meat samples recorded by Foss-NIRSystems 5000 between 1300 and 2500 nm to determine the fat content.[19] The second (the alfalfa data) consists of 305 samples of forages[20] measured between 1108 and 2492 nm each 8 nm to determine the protein content. The third data set (the hydrogen data) contains 239 samples of gas oil measured between 4900 and 9000 cm^{-1} (each 2 nm) to determine the percentage of hydrogen.

The algorithms were programmed in Matlab version 6.5.

RESULTS AND DISCUSSION

Simulated Data Sets

To show how the DT-calibration domain algorithm works, three two-dimensional simulated data sets are presented. Because the aim of this section is a demonstration of the DT calibration domain algorithm, it is supposed that no preliminary investigation of the data set took place. In practice, at least some of the problems with the data would have been discovered and remedied before applying the algorithm. At the same time, we will also show how the alternative Dbscan-based methods works. In data set 1, the data consist of one group. In data set 2, a minor group of objects is somewhat removed from the majority group with different densities and it contains an outlier that is expected to cause difficulties with the alternative Dbscan-based approach. Data set 3 contains three equivalent clusters, and it is expected that the DT calibration domain method will not give good results.

Simulated Data Set 1

Data set 1 contains 78 samples that are chosen from one group of real NIR data: 60 for calibration and 18 for prediction (Fig. 3). Each sample is

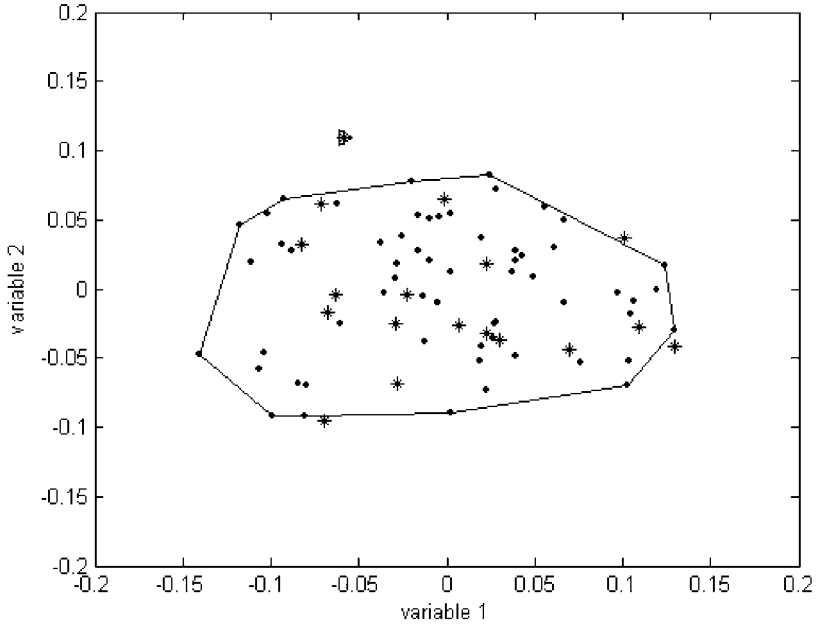

Figure 3. Simulated data set 1, where • represents the calibration samples, ∗ represents the samples in the test set, and △ is an outlier in the test set.

represented by a spectrum consisting of 174 wavelengths, and the first two PCs of the samples are used. The kernel calibration data set (31 samples) is obtained after 100 iterations. Then a Delaunay triangulation mesh is constructed by using the kernel calibration data set. For each of the remaining 29 samples, the MD values are calculated to the triangles of the DT, and the smallest one is selected as the MD value of the samples. If the MD value is smaller than Cri = 2.5*1.1547, the point is included into the kernel set. Twelve samples are included after the first iteration. This means that there are now 43 samples in the extended kernel set (X1new) and 17 objects in X2new. Then a new DT is performed on the 43 samples. The MD values of the objects in X2new are calculated again and the samples with MD $\leq$ 2.5*1.1547 are again included in X1new. This step is repeated till no MD of the samples in X2new is smaller than Cri. All 60 samples in the calibration set are part of the cleaned calibration set, and therefore no samples are detected as outliers in the calibration set.

One large cluster and a small cluster of 7 on the lower left of the figure are detected using the alternative Dbscan-based approach. The cluster is considered to be the kernel calibration data set. After extension of the cluster, the other 7 samples are included so that all 60 objects are contained in the cleaned calibration set. The DT and the Dbscan based alternative approach agree therefore completely in this simple case.

Of the new (prediction) samples, 14 are inside the DT mesh constructed with the cleaned calibration set, and 3 of them are inside the marginal outlier domain defined by the coefficient limits $[-1, 2]$, so that one outlier is detected in the test set as shown in Fig. 3.

Simulated Data Set 2

Simulated data set 2 contains 460 two-dimensional samples in the calibration set and consists of two clusters, a dense one of 400 samples that is drawn from the bivariate normal distribution $N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$ and a less dense one of 59 samples that is drawn from the bivariate normal distribution $N_2\left(\begin{bmatrix}0\\5.5\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$. Moreover, there is one large outlier (Fig. 4). In the DT calibration domain method, 235 samples are selected in the kernel calibration set with 500 iterations. After the extension step, the cleaned calibration set contains 394 samples. All 59 samples in the smaller cluster are recognized

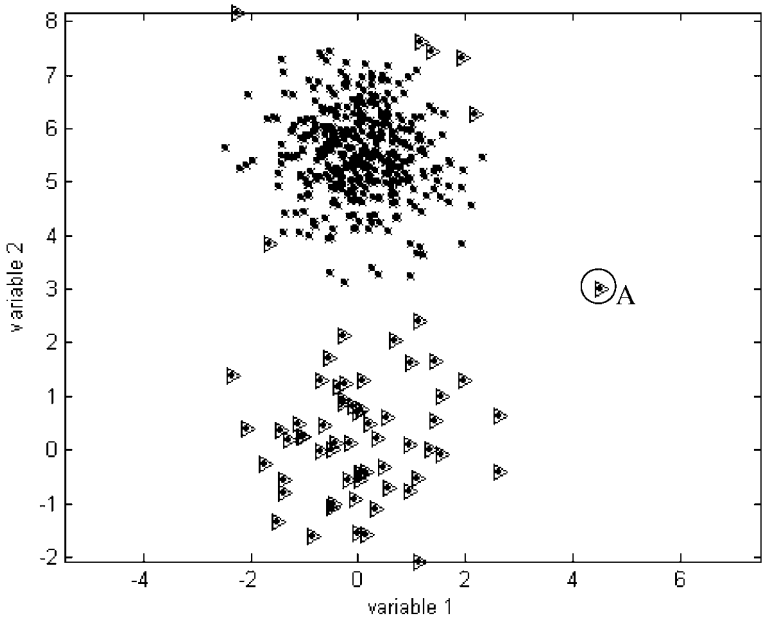

Figure 4. Simulated data set 2. Sample A is the outlier that is expected to cause difficulties for finding clusters in the data set; x represents samples in the cleaned calibration set, and △ represents samples that are not recognized as belonging to the calibration set.

as not being part of the calibration set. This is also the case for the large outlier and six samples on the border of the large cluster.

This simulated data set was constructed to show a situation in which Dbscan will not give the results wanted because Dbscan is sensitive to the parameters Minpt and ε. The value of the latter is determined based on the volume of the data space. When a large outlier is present, this increases the volume artificially. The clusters are then difficult to detect. With the Dbscan approach, only one cluster is detected, although there are two. When the outlier is removed, Dbscan, correctly finds two clusters, the larger one containing 399 samples and the smaller one 59 samples. There is also one outlier from the larger cluster.

After removal of the large outlier, Dbscan and the DT method find nearly the same calibration set. A few samples are considered outliers in the DT method and not in the Dbscan method. It should also be noted that in a case like this, with two clusters that are not far from each other, the analyst might decide to investigate with a test set if it is necessary to have two local models or whether the results with one model would not be equally acceptable.

In this case, it would be possible to create a second local model for the smaller cluster.

Simulated Data Set 3

Data set 3 contains 150 two-dimensional samples. There are three equivalent clusters that contain 60, 40, and 50 samples, respectively. The three clusters are simulated from the bivariate normal distribution $N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$, $N_2\left(\begin{bmatrix}-5\\7.5\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$, and $N_2\left(\begin{bmatrix}7\\-2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$, respectively. This data set is expected not to give the results wanted with the DT calibration domain method. In the DT calibration domain method, the first step is to obtain 51% of all samples as kernel calibration set. When clusters with equivalent size exist, the number of samples in each cluster is less than 51%, so that samples will come from at least two clusters. As shown in Fig. 5, the kernel set (77 samples) contains indeed samples from two clusters.

The Dbscan approach does work because three clusters are detected. The largest cluster contains 60 samples (cluster 1) and is considered to be the kernel calibration set. No additional samples are added in the extension step, and the cleaned calibration set contains 60 samples. A second local model is then made with the second largest cluster and a third one with the last cluster.

In this case, Dbscan and the DT method yield different results. The Dbscan method warns the analyst that the solution obtained with the DT method may not be (and in this case, is not) the right one.

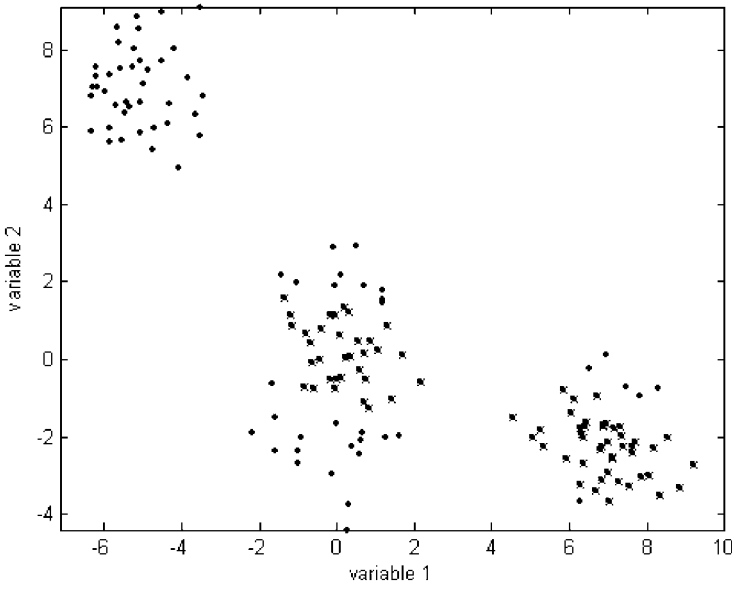

Figure 5. Simulated data set 3. Kernel calibration set for the DT-calibration domain method; x represents the kernel calibration data.

NIR Data Sets

The Meat Data

The set of 198 meat samples is split into two subsets by the Duplex method[21]: 150 in the calibration set and 48 in the test set. The PC1-PC2 plot is shown in Fig. 6. No clustering is apparent in the calibration set.

Monte Carlo cross validation was applied to determine the optimal number of PCs for the DT method. The root mean square error for cross-validation (RMSECV) was computed and the minimal RMSECV was obtained when five dimensions were used, so that the first five PCs are used as variables. Because there are five dimensions, S is 2.0412. With this value, a kernel set of 77 samples is selected with the DT calibration domain method and, after optimization, extended using the rules described in the "Theory" section. One sample is found to be an outlier; all other samples are incorporated into the cleaned calibration set.

When the alternative method based on Dbscan is used, it is found that there is one high-density cluster of 102 samples. The 48 other samples form minor clusters or are outliers. However, they are quite close to the central cluster. The 102 samples are considered to form the kernel, and after extension the same 149 samples are gathered in the cleaned calibration set as with the DT calibration domain method and the same outlier is found.

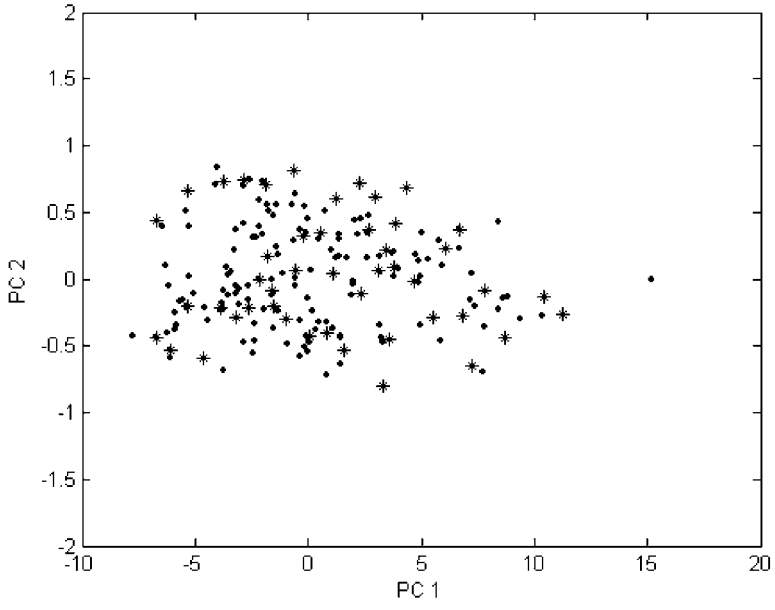

Figure 6. The meat data, PC1–PC2 plot; • represents the calibration samples, and * represents the samples in the test set.

Of the 48 samples in the test set, 24 are inside the DT mesh and 24 are marginal outliers. There is no true outlier in the test set. The RMSEP (root mean square error for prediction) value for the 48 samples in the test set is 2.196 with the DT method and 3.181 with PLS. The DT method performs markedly better than PLS.

The Alfalfa Data

This data set was split into two data sets by the providers of the data. It contains 205 samples for calibration and 100 samples for prediction.

In order to select the number of PC factors needed to construct the DT mesh, the leave-one-out cross-validation (LOOCV) method is used. The minimal RMSECV was obtained for five dimensions, so that the first five PCs are used as variables.

The PP plot with the entropy projection index as criterion[12,13] does not pinpoint a gross outlier. The PCA plot (Fig. 7a) and the PP plot with the Yenyukov projection index as criterion (Fig. 7b) show the existence of three clusters close to each other. This is confirmed with Dbscan, as 3 clusters are detected (Fig. 7a).

The DT calibration domain method was used with $S = 2.0412$. No outliers are detected in the calibration set, and all 205 calibration samples

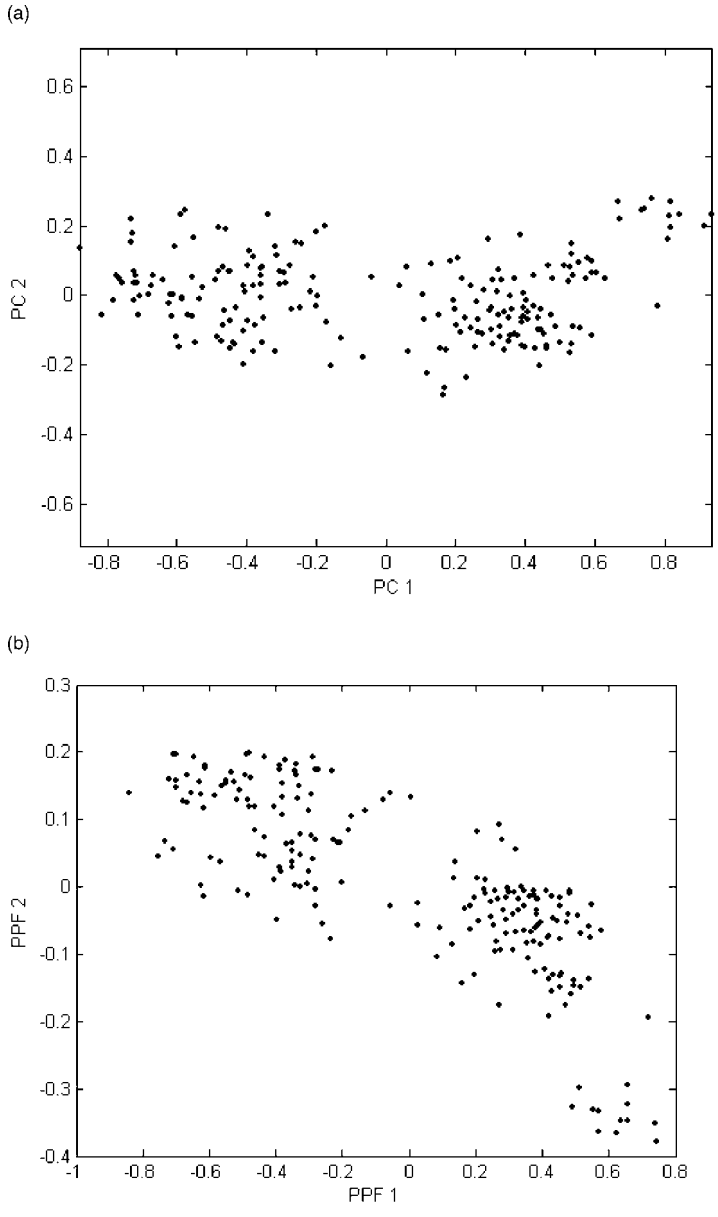

Figure 7. The alfalfa data: (a) PC1–PC2 plot; (b) projection pursuit plot.

are included in the cleaned calibration set. This means that no clusters are recognized. Because the largest cluster contains less than 51% of all data, it is expected that less than three clusters are found. When the alternative Dbscan method is used, the largest cluster is chosen as the kernel calibration

set, thereby confirming the DT result. After extension, all calibration samples are found to be contained in the cleaned calibration data set. This is not surprising, as the three clusters are only a little further away from each other than some of the samples on the borders of the clusters are from the other samples of the same cluster. As explained in the introduction, there is then no reason to make separate models for each cluster.

The calibration set is used to construct the DT domain, and 65 of the 100 samples of the test set are found to be inside it. The marginal outlier domain with coefficient limits $[-1, 2]$ contains 34 samples, so that 1 outlier is detected in the test set. The DT method yields an RMSEP of 1.18, whereas that of PLS is 1.33. When the outlier in prediction is deleted also for PLS, as it was for the DT method, the PLS value is 1.20 and therefore comparable to that for the DT method. The large influence of this one sample on the PLS RMSEP indicates that it is indeed a true outlier and shows that the DT method in prediction recognizes such outliers.

The Hydrogen Data

The data set is split into two subsets by using the Duplex method. There are 190 samples for calibration and 49 samples in the test set. Figures 8a and 8b show the PC1–PC2 and PC2–PC3 plot of this data set, and Fig. 8c shows the projection pursuit plot with the entropy projection index as criterion. The calibration data seem to contain a majority group, at least one smaller cluster and some large outliers.

Dbscan finds three clusters. Two of them are close to each other, a major one of 158 samples and a minor one of 9 samples. There is also a much further removed cluster of 18 samples and there are 5 outliers (see Fig. 8d).

Five dimensions are found to be the optimal complexity by using LOOCV in the DT method, even after the elimination of the three larger outliers. After applying the DT calibration domain method, 30 samples are found not to be included in the calibration set, so that 160 samples constitute the cleaned calibration set. It may be surprising to find some outliers in the majority group in Figs. 8a–8c. This is due to the fact that the figures are shown in only two dimensions, while in fact there are five.

The alternative Dbscan approach is also used for this data set. The majority group contains 167 objects after extension and consists of the large cluster and the minor one of 9 samples (Fig. 8d). It combines the two clusters that were very close to each other.

Because many objects were discarded from the cleaned calibration set and the dimensionality of a PC model is often increased by outliers or by the fact that there is more than one group of samples, it is necessary to reevaluate the dimensionality of the PC model for the cleaned calibration set. Therefore, LOOCV is carried out again, and the complexity is now found to be 4. There are 16 outliers in the test set when the calibration set of the DT

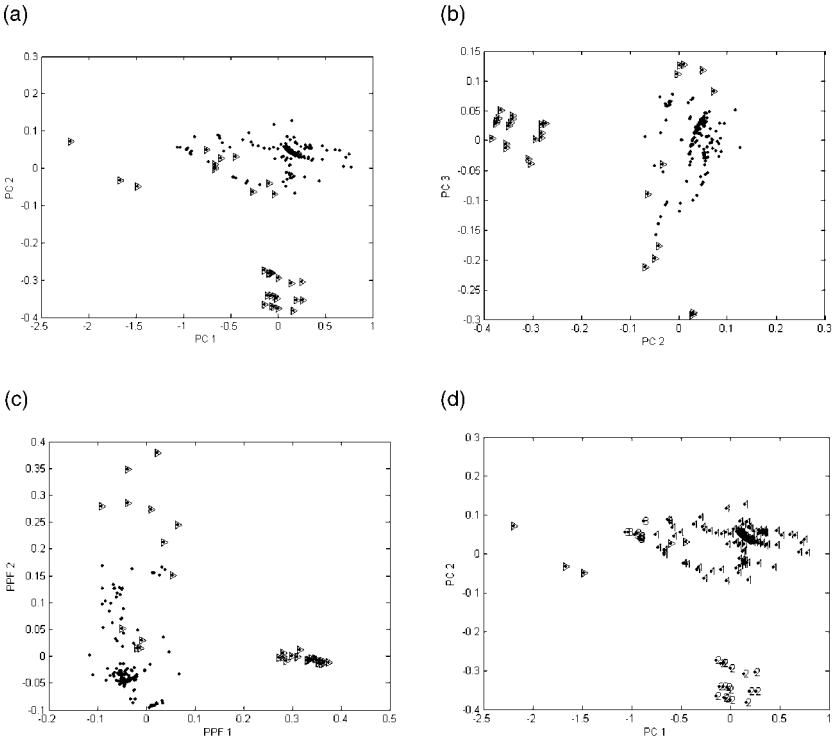

Figure 8. The hydrogen data: (a) PC1–PC2 plot; (b) PC2–PC3 plot; (c) projection pursuit plot. △ represents samples that are not recognized as belonging to the calibration set. (d) Clusters and outliers detected with the Dbscan approach where 1 represents cluster 1, 2 represents cluster 2, and 3 represents cluster 3. △ represents samples not belonging to any cluster.

calibration domain method is used and 11 when the one of the alternative methods based on Dbscan is applied.

Because there is at least one more cluster, a second model is made, still after deleting the three large outliers. Because the cleaned calibration set obtained with Dbscan seems more appropriate than that obtained with the DT calibration domain algorithm, the complete procedure is started all over again for the 23 samples not included in the first cleaned calibration set (i.e., starting with the construction of a DT mesh and a cross-validation to determine the number of dimensions). The second cleaned calibration set is found to consist of the 18 samples of the second cluster, and 3 dimensions are needed. When each sample of the test set is now predicted with the DT method, using the calibration set to which it is closest, 6 outliers are detected in the test set. PLS is applied with a global model including the two clusters. For both the DT and the PLS methods the three large outliers in the calibration set were omitted. With the DT method, the RMSEP is

0.0587, with PLS it is 0.0631, when the six outliers are not predicted. It might appear that DT is again better than PLS as for the alfalfa data, but when the 6 outliers are predicted, PLS outperforms in this case DT.

CONCLUSIONS

Two methods are proposed to determine the calibration domain for the Delaunay triangulation method, a topological multivariate calibration model. One of the methods is purely based on local characteristics and is called the DT calibration domain. The other first applies a clustering with Dbscan, a density-based clustering method. The two methods work well in most situations, but in some situations such as simulated data sets 2 and 3, one of the methods fails to find the cleaned calibration set, but the other does.

The Dbscan method is therefore complementary to the DT-outlier method, and each has some advantages and disadvantages. A difficulty with the DT calibration domain method is that it is not able to recognize whether the samples that are excluded from the calibration set are isolated outliers or belong to a cluster. In the latter case, one would like to know this to create an additional local model. Using the Dbscan-based method solves this problem. Dbscan has a tendency to find too small clusters in the initial stage, but the extension algorithm corrects this.

The DT method compared to other multivariate outlier methods such as PLS also has advantages and disadvantages. Apart from the advantages of local methods compared to global methods such as avoiding problems with nonlinearity, the study shows that the DT method has for instance a natural way of finding outliers in the prediction stage, as shown with the alfalfa data set. With PLS, additional diagnostics are needed to discover such outliers. On the other hand, the hydrogen data set shows that at least in some cases, PLS has a better chance of making an acceptable prediction of the concentration of such outliers.

Some of the difficulties in the DT method with the real NIR data sets are due to the fact that principal component scores are used instead of selected wavelengths. When the latter are used, the selection of the proper number of PCs is not needed, nor its reevaluation when a cleaned calibration set contains many samples less than the original set. Therefore, feature selection should be preferred when possible. A difficulty that was not studied here, but which might occur in some applications when using PC scores, is that the sample could be an outlier toward the PC model in the sense that it has a large residual. It is possible that such a sample would fall inside the calibration domain and that the large residual would go unnoticed, thereby leading to wrong predictions of the sample in question. How to cope with this will be the object of future research.

In general, we conclude that the potential of the DT method is confirmed but that further research is needed, for instance to find the best way of updating

the calibration set with new samples, to determine the effect of including more samples than the simplexes in the prediction, and to determine the uncertainty of the predicted concentration values. More applications and experience with the method would also be useful and would allow it to be fine-tuned.

ACKNOWLEDGMENT

We thank Dr. J. A. Fernández Pierna and Dr. P. Dardenne for their kindly provision of the meat data set. Thanks also to Dr. Michal Daszykowski for the helpful discussion on the Dbscan method.

REFERENCES

1. Jin, L.; Fernández Pierna, J. A.; Xu, Q.; Wahl, F.; De Noord, O. E.; Saby, C. A.; Massart, D. L. Delaunay triangulation method for multivariate calibration. *Anal. Chim. Acta* **2003**, *488*, 1–14.
2. Stone, C. J.; Bickel, P. J.; Breiman, L.; Brillinger, D. R.; Brunk, H. D.; Pierce, D. A.; Chernoff, H.; Cover, T. M.; Cox, D. R.; Eddy, W. F.; Hampel, F.; Olshen, R. A.; Parzen, E.; Rosenblatt, M.; Sacks, J.; Wahba, G. Consistent nonparametric regression. *Ann. Stat.* **1977**, *5*, 595–645.
3. Espinosa, A.; Sanchez, M.; Osta, S.; Boniface, C.; Gil, J.; Martens, A.; Descales, B.; Lambert, D.; Valleur, M. On-line NIR analysis and advanced control improve gasoline blending. *Oil Gas J.* **1994**, *17*, 49–56.
4. Danielsson, R.; Malmquist, G. Multi-dimensional simplex interpolation: an approach to local models for prediction. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 115–128.
5. Okabe, A.; Boots, B.; Sugihara, K. *Spatial Tessellation: Concepts and Application of Voronoi Diagrams*; Wiley: Chichester, 2000.
6. Gudmundsson, J.; Hammar, M.; Van Kreveld, M. Higher order Delaunay triangulation. *Comput. Geometry* **2002**, *23*, 85–98.
7. Available at http://www.personal.kent.edu/~rmuhamma/Compgeometry/MyCG/CG-Applets/DelaTessel/delacli.htm.
8. Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans. Mathematical Software* **1996**, *22* (4), 469–483.
9. Peña, D.; Yohai, V. A fast procedure for outlier diagnostics in large regression problems. *J. Am. Stat. Assoc.* **1999**, *94*, 434–445.
10. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18.
11. Available at: http://www.galactic.com/Algorithms/discrim_mahaldist.htm.
12. Croux, C.; Ruiz-Gazen, A. A fast algorithm for robust principal components based on projection pursuit. *COMPSTAT: Proceedings in Computational Statistics 1996*; Physica-Verlag: Heidelberg, 1996; pp. 211–217.
13. Daszykowski, M.; Walczak, B.; Massart, D. L. Projection methods in chemistry. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 97–112.
14. Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B*; Elsevier: Amsterdam, 1998.

15. Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990.
16. Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*; Simoudis, E., Han, J., Fayyad, U., Eds.; Poland, 1996; 226–231.
17. Daszykowski, M.; Walczak, B.; Massart, D. L. Looking for natural patterns in data: Part 1. Density-based approach. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 83–92.
18. Ankerst, M.; Breunig, M. M.; Kriegel, H.; Sander, J. OPTICS: ordering points to identify the clustering structure. *SIGMOD Conference*; Delis, A., Faloutsos, C., Ghandeharizadeh, S., Eds.; Philadelphia, 1999; 49–60.
19. Corbisier, S.; Sinnaeve, G; Baeten, V.; Sindic, M.; Dardenne, P.; Deroanne, C. Optimisation of the measurement of meat and meat products; valorisation of the databases. In *Proceedings of the 11th ICNIRS*, Córdoba, Spain, April 6–11, 2003; Davies, A., Garrido-Varo, A., Eds.; NIR Publications: Chichester, UK.
20. Ruisánchez, I.; Rius, F. X.; Maspoch, S.; Coello, J.; Azzouz, T.; Tauler, R.; Sarabia, L.; Ortiz, M. C.; Fernández, J. A.; Massart, D. L.; Puigdomènech, A.; García, C. Preliminary results of an interlaboratory study of chemometric software and methods on NIR data. Predicting the content of crude protein and water in forages. *Chemom. Intell. Lab. Syst.* 2002*63* (2), 93–105.
21. Snee, R. D. Validation of regression models: methods and examples. *Technometrics* 1977*19*, 415–428.