Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 488 (2003) 1–14
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Delaunay triangulation method for multivariate calibration

L. Jin [a], J.A. Fernández Pierna [a], Q. Xu [a], F. Wahl [b], O.E. de Noord [c], C.A. Saby [d], D.L. Massart [a,*]

[a] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Institut Français du Pétrole (IFP), BP3, 69390 Vernaison, France*
[c] *Shell Research and Technology Centre, Shell International Chemicals B.V., P.O. Box 38000, 1030 BN Amsterdam, The Netherlands*
[d] *Centre de Recherche TotalFinaElf, BP 22, F-69360 Solaize, France*

Received 13 May 2003; accepted 19 May 2003

Abstract

The Delaunay triangulation (DT) method is proposed as a new local multivariate calibration method. DT was developed within computational geometry, and it is shown that it has potential for applications in analytical chemistry, such as multivariate calibration. The study compares the performance of the DT method with the global methods principal component regression (PCR) and partial least squares (PLS) and the local methods locally weighted regression (LWR) and the law of mixtures (LM) method. For the datasets studied the DT method gives similar results when the root mean square error for prediction (RMSEP) value is compared. However, the DT method requires fewer components than PCR and PLS.

Keywords: Multivariate calibration; Local methods; Delaunay triangulation; Simplexes; Prediction

1. Introduction

In multivariate calibration, one can distinguish local and global methods. The latter use all calibration samples to construct one model, often using linear methods such as partial least squares (PLS), principal component regression (PCR) or multiple regression. The former use only the samples in the neighbourhood of the sample whose characteristics have to be predicted. It can eliminate the risk of using a wrong model, e.g. a linear model when, in fact, the relationship is non-linear.

There are essentially two types of local methods. The first consists of methods that apply PLS or PCR but use only calibration samples in the neighbourhood of the sample to be predicted. There are several such methods [1–4]. A second category consists of methods in which the property values, e.g. concentrations, of samples that are close to the unknown in the variable space (neighbouring samples) are averaged in some way. Such methods are sometimes called topological methods. The method we present here belongs to the second category.

The oldest and simplest topological method is the *k*-nearest neighbours (kNN) method, which is best known for classification, but is also used in quantitative prediction [5]. The patented TOPNIR method, which is used with near infrared spectroscopy in the oil industry, is based on this approach [6]. A crucial question is how to select the nearest neighbours. The most local method is obtained when the number of points selected is only one more than the number of *x*

* Corresponding author. Tel.: +32-2-4774737; fax: +32-2-4774735.
E-mail address: massart@vub.vub.ac.be (D.L. Massart).

0003-2670/03/$ – see front matter
doi:10.1016/S0003-2670(03)00629-9

variables, i.e. for k variables, $k+1$ nearest neighbours. Those should be selected in such a way that they surround the unknown sample. This means that calibration samples that form a so-called simplex should be selected in which the unknown sample is inscribed (a simplex is a geometrical figure in k dimensions with $k+1$ vertices, e.g. in two dimensions a triangle). The property of the prediction samples falling inside a certain simplex is then predicted using the property of the calibration samples forming that simplex. This was proposed in a method called the multi-dimensional simplex interpolation (MSI) [7]. In the MSI method, enclosing simplexes are generated for each unknown sample. It is shown that to obtain the best predictions, the distance between the unknown sample and the simplex points should be kept as small as possible. Several simplexes can be obtained and the MSI selects the simplex for which the distance is indeed smallest.

More recently we proposed a method that we call the law of mixtures (LM) method [8]. The basic principle of the method is to connect the calibration samples such that they form a predefined mesh of simplexes, while in the MSI method simplexes are selected each time an unknown sample must be predicted. In all cases studied, the LM method gave at least similar results as PCR or PLS, but the algorithm to form the mesh was very time consuming and the simplexes formed were not the best that could be imagined. A method to achieve simplexes, which is well known in some application areas such as in geometry, is the Delaunay triangulation (DT) [9–13]. Moreover Matlab version 6.1 [14] contains an efficient algorithm, making the method available to many users. DT has some important properties (see Section 2) that ensure that these simplexes are "good" simplexes.

The idea of the DT method originates from the study of structures in computational geometry. It is one of the most popular methods for generation of unstructured meshes. For a given set of points in two dimensions, it constructs a triangle's mesh using all the points as vertices. It is applied in several fields of science such as in metallurgy for examination of alloys, in cartography for town planning, in crystallography to simulate the growth for crystals, in mesh generation of finite elements methods, etc. [15,16]. As far as we know there are, however, no applications in chemometrics or even in analytical chemistry. The aim of this article is to introduce the DT method for multivariate calibration. The resulting method is extremely simple. It consists of four steps: (1) the number of variables is reduced by obtaining PC scores for the calibration dataset and using these as new variables; (2) the DT mesh is constructed; (3) the simplex in which the unknown sample is situated in the PC-space is determined; (4) the result for the unknown sample is the (weighted) average of the property values of the calibration samples that constitute the simplex. The same mesh is used for all unknown samples.

2. Theory

2.1. The Delaunay triangulation (DT) method [9–13]

The DT (in two dimensions) is defined by the empty circle condition; i.e. one triangle is a valid triangle only if its circumcircle encloses no other points of the dataset. The circumcircle of a Delaunay triangle (or Delaunay circle) is the circle that can be drawn through all the vertices of that triangle. Generally, the Delaunay triangulation of a set of points is unique [9,13]. The steps to construct the DT in two dimensions are as follows.

(1) Two points are selected randomly (e.g. points 1 and 2 in Fig. 1). They are connected by a line $\overline{1}\,\overline{2}$. The rest of the points are connected with these two points. A triangle is determined when there is no other point inside the Delaunay circle. Two triangles can be built in this case (triangle A and B in Fig. 1a). If on one side of line $\overline{1}\,\overline{2}$ there are no points, only one triangle on the other side can be constructed. Both triangles fulfill the empty circle condition and are retained.

(2) Triangle A is selected to continue building the mesh. One of the sides of the triangle that is not connected with B is selected, for instance, line $\overline{1}\,\overline{3}$ in Fig. 1b. This line now plays the same role as line $\overline{1}\,\overline{2}$ in the previous step. A new triangle is determined when there is no other point inside the Delaunay circle. In Fig. 1b, the Delaunay circle determined by the triangle C′ contains points 5 and 9 and therefore triangle C′ is not accepted. The points inside (points 5 and 9) are used in order

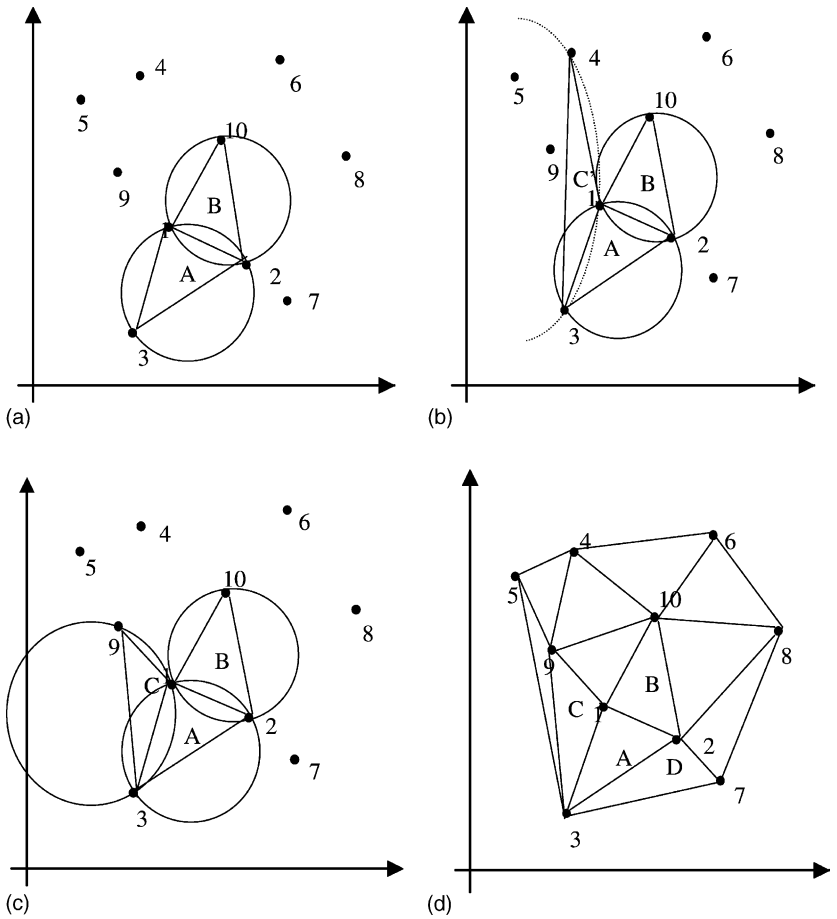


Fig. 1. (a–d) Steps to construct the Delaunay triangles.

to build new candidate triangles. Point 9 leads to triangle C (Fig. 1c) and it is accepted because no more points are inside its circumcircle.

(3) Step 2 is repeated for the other side of triangle A and triangle D is constructed in Fig. 1d.

(4) Steps 2 and 3 are now repeated for the retained triangles (B, C, D, ...) until all points are connected as shown in Fig. 1d.

The previous steps are performed in the PC-space. In this way the dimensionality of the DT space is reduced dramatically as compared to the original variable space. Two possibilities were considered, namely, to use the first PCs ranked according to the variance they express or to use the PCs selected according to their correlation to the property (y) in principal component regression (PCR). The second approach has the advantage that it reduces the dimensionality even more since PCs with large variance that are not related to y are not considered. On the other hand, it is not evident that using global correlation with y is a good criterion to select local models. Both approaches are compared in Section 3.

The LM method [8] is based on the same idea as DT. The main difference between these two methods is the way that the mesh is built. DT uses the natural neighbours of each calibration data point to construct the simplexes, while LM uses the nearest point to the centre of the calibration data to construct the initial simplexes with the boundary points, then the rest of

the points are connected depending on their situation with respect to the previous simplexes. The way that DT selects the simplexes is much more logical and elegant.

A question which arises is of course whether the simplexes generated by the DT method are "good", i.e. allow good prediction. Danielsson and Malmquist [7] showed for their MSI method that simplexes should be as small as possible. Another criterion is that the simplex points should surround the point to be predicted as well as possible. The DT has the property that the circumcircle of every triangle contains no other data points, thereby ensuring that simplexes are indeed small. The DT maximises the smallest interior angle of each triangle. This means that the minimum angle of any triangle is as large as possible, so that DT tends to triangles with more equidistant points than other triangulations. This property makes that the simplex surrounds the point to be predicted well and that it avoids co-linearity or ill-conditioning. It may be concluded that the DT generates simplexes with desirable properties.

2.2. Prediction

Once the mesh is constructed, it is used for the prediction of new samples. The new samples are projected into the PC-space where the mesh is built. If a new sample M falls within a simplex defined by k neighbours (k is equal to the number of PCs + 1), it is considered as a mixture of these samples in order to calculate the value of its associated property.

In two dimensions the following equations are used to obtain the position of the new sample with respect to its $k = 3$ neighbours (M_1, M_2, M_3) that are surrounding the new sample M [8]. They can easily be generalised to more dimensions.

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

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

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

where x_{1i} and x_{2i} are the scores of the objects in the PC-space, α_{M_1}, α_{M_2} and α_{M_3} are the contribution of samples M_1, M_2 and M_3, respectively, and the sum of the coefficients is 1:

$$\alpha_{M_1} + \alpha_{M_2} + \alpha_{M_3} = 1 \quad (4)$$

The property of an unknown sample M is obtained by using Eq. (5):

$$y_M = \alpha_{M_1} y_{M_1} + \alpha_{M_2} y_{M_2} + \alpha_{M_3} y_{M_3} \quad (5)$$

where y_M, y_{M_1}, y_{M_2} and y_{M_3} are the property for samples M, M_1, M_2 and M_3, respectively. Similar equations, but presented differently, were proposed in the MSI method [7].

For sample M, all the coefficients of each triangle are calculated. Only one triangle fulfils the condition: $0 \leq \alpha_{M_1}, \alpha_{M_2}, \alpha_{M_3} \leq 1$ and it is the triangle that contains M.

For the new samples falling into the mesh, there are two special situations, i.e. a new sample falls on a line (facet of the polyhedron in more than two dimensions) or it falls on a calibration point. In the first case, one of the coefficients is 0. In the second case, one of the coefficients is 1, i.e. the property of the new sample equals the property of the overlapped calibration point.

In most applications, e.g. in geography, the whole map can be triangulated. However, in our application some prediction objects are situated outside the mesh. Some of these are real outliers and some are not real outliers, but borderline objects. Real outliers in prediction are the samples that are inconsistent with the calibration data and they can be found by outlier detection methods [17–19]. The borderline objects are the objects that are outside the mesh constructed by the calibration set, but are not real outliers. The real outliers must be detected and eliminated, borderline objects should be predicted and the original DT method does not allow it. This is the main difficulty in applying DT to calibration problems. In order to obtain also results for the borderline objects, three different approaches are proposed.

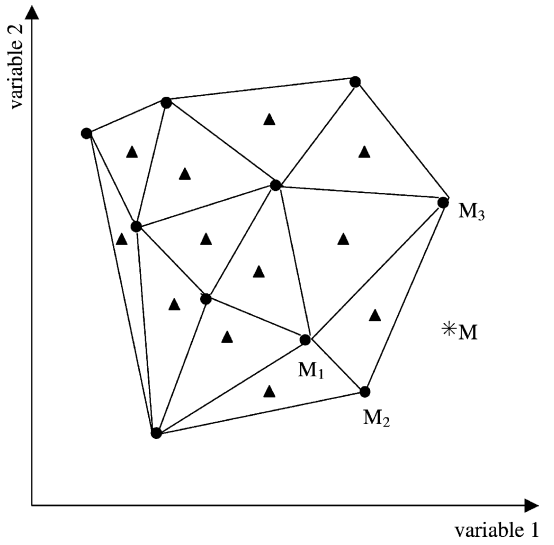


Fig. 2. Finding the centre of each triangle in approach 1, where (●) are the calibration objects; (▲) are the centres of the triangles; and (✳) (M) is the borderline object.

2.2.1. Approach 1

Due to the fact that the sum of the coefficients ($\sum \alpha$) is 1 and $\alpha > 0$ for the points inside the mesh (see Eq. (4)), the α values are limited to the range [0, 1]. To predict the borderline objects we should allow the coefficients of the mixture to be negative. The coefficients are determined for the triangle closest to the object. The following steps explain how to apply this approach in two dimensions, but it can easily be generalised.

(1) The centres of all triangles in the mesh are determined (Fig. 2). Then the Euclidean distances between the prediction object and each of the centres are calculated. The triangle with the smallest distance is selected. For instance, in Fig. 2, for the borderline object M, the triangle $M_1M_2M_3$ is selected.
(2) Using Eqs. (1)–(3), the coefficients α_1, α_2, α_3 are obtained. Because M is outside the triangle $M_1M_2M_3$, at least one of the coefficients is negative.
(3) The property (y_M) for the borderline object is obtained by using Eq. (5).

2.2.2. Approach 2

In approach 1 we allowed the coefficients to be negative without any further constraint. In approach 2, the range of α values is limited to, e.g. [−1, 2], [−2, 3], etc. When the point is outside the mesh, all triangles whose coefficients are within a given limit are considered. In this way a subspace is constructed around the triangle, defining the prediction limits of it. The limit [−1, 2] is preferred because it allows constructing the smallest subspace around the triangle, but in the case that the points cannot be predicted using this limit, the next one, e.g. [−2, 3] should be applied. Fig. 3 shows the prediction subspace of triangle $M_1M_2M_3$ when the coefficient's limit is [−1, 2], i.e. a new point A inside the subspace surrounded by the dashed line can be predicted using this triangle. In the example of Fig. 3, if one borderline object (B) is outside the subspace determined by the limit [−1, 2] for triangle $M_1M_2M_3$, it is verified if it belongs to the subspace determined by the rest of the triangles with the same limit. If this is not the case, the limits [−2, 3], [−3, 4], ... are used till object B belongs to at least one subspace. Because real outliers have been eliminated first, most of the borderline objects are close to the convex hull containing the calibration data and the subspace with the coefficient limits [−1, 2] includes most of the borderline objects to be predicted. One borderline object can also belong to the subspace of two or more triangles. In such cases the property of the borderline object is the average value of the prediction results obtained from all the possible triangles.

(1) For the borderline object M, the coefficients α_1, α_2, α_3 for each triangle are determined using Eqs. (1)–(3).
(2) The triangles for which the coefficients are between the given limits are retained. If, for instance, there are two triangles, $M_1M_2M_3$ and $M_4M_5M_6$, with coefficients between these limits, two prediction results for object M are obtained:

$$y_M(1) = \alpha_{M_1} y_{M_1} + \alpha_{M_2} y_{M_2} + \alpha_{M_3} y_{M_3} \quad (6)$$

$$y_M(2) = \alpha_{M_4} y_{M_4} + \alpha_{M_5} y_{M_5} + \alpha_{M_6} y_{M_6} \quad (7)$$

(3) The property of M ($\hat{y}_M$) is the average value of the prediction results from step 2.

2.2.3. Approach 3

Instead of choosing the simplex with the smallest distance between the centre of the triangles and M as in approach 1, the simplex can also be chosen with as criterion that $\max_j(|\alpha_{ij}|)$ should be smallest (in K

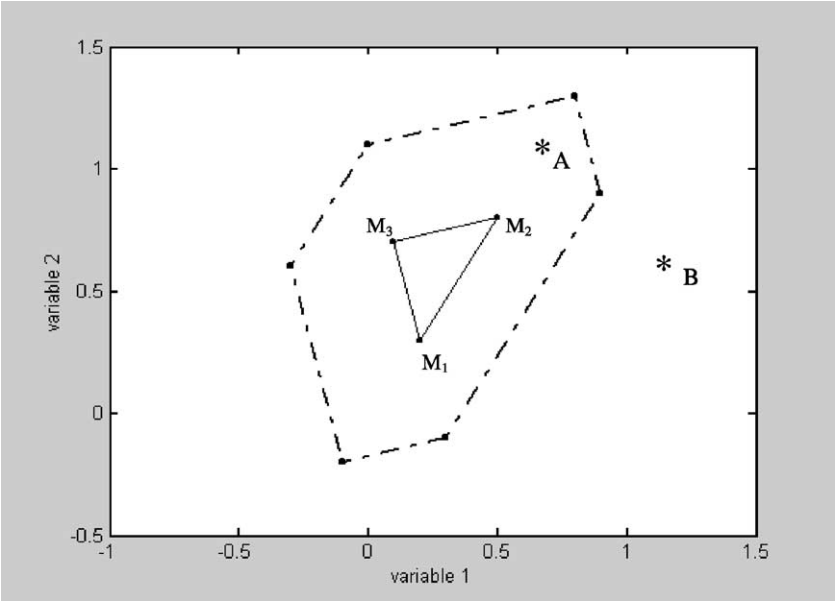


Fig. 3. The prediction subspace of triangle $M_1M_2M_3$ when the coefficient limits are $[-1, 2]$.

dimensions, $j = K + 1$). The steps for this approach (in two dimensions) are the following.

(1) The DT is constructed and the resulting mesh consists of N triangles.
(2) For each triangle, α_{i1}, α_{i2}, α_{i3} ($i = 1, 2, \ldots, N$) are obtained with Eqs. (1)–(3). The triangle with smallest $\max_j(|\alpha_{ij}|)$ to the borderline object M that has to be predicted is selected, i.e. for each triangle the maximum coefficient value $\max_j(|\alpha_{ij}|)$ is selected first, then these values are ranked and the smallest one is retained.
(3) Eq. (5) is used to obtain the property of M.

3. Experimental

Three NIR datasets are used in this study. They consist of the following.

3.1. Dataset 1

Two hundred thirty-nine samples of gasoil measured between 4900 and 9000 cm^{-1} (each 2 nm) to determine the percentage of hydrogen (y). The range of property (y) is from 10.61 to 14.40.

3.2. Dataset 2

Three hundred five samples of alfalfa (forage) [20] measured between 1108 and 2492 nm (each 8 nm) to determine the protein content (y). The range of property (y) is from 10.78 to 27.75.

3.3. Dataset 3

Eighty-seven samples of polyether polyols measured between 1100 and 2158 nm (each 2 nm) to determine the hydroxyl number (in mg KOH g^{-1}). The range of property is from 10.90 to 133.60.

4. Results and discussion

4.1. Dataset 1

The dataset was split into two subsets by using the Duplex algorithm [21]: 199 objects were used to build the model and 40 objects were selected for prediction.

In order to determine the optimal number of components in the model, leave-one-out cross-validation and the randomisation test [22] were used. In PLS, the

minimum root mean square error for cross-validation (RMSECV) value was obtained when eight components were used. After comparing the precision of the model using the randomisation test, we decided on a five components model. Another technique to determine the complexity of the model is the Monte-Carlo cross-validation (MCCV) method [23–25]. The optimal number of components is 5 when this method is used [8]. This was also the case for PCR with selection.

Leave-one-out cross-validation is also used in the DT method. When the left-out object is inside the mesh constructed by the rest of the data, the property of the object is obtained by using the DT method. If it is outside, approach 2 is used. The optimal RMSECV value was obtained for four dimensions when the randomisation test was used. In this dataset, the first four important PCs selected according to their correlation to y are first, third, fourth and sixth PC and DT is applied in the data space defined by these PCs.

Fig. 4a and b show the triangles obtained using the DT method and the LM method, respectively, in two dimensions. As was previously said, these methods are different in the way of constructing the mesh. From Fig. 4a and b one can see that DT gives, as could be expected from theory (see discussion at the end of Section 2) a better mesh than the mesh from LM.

A large number of prediction points is found to be outside the DT mesh. No real outliers are detected using the method from [19], which means that all the points outside the mesh are borderline objects. In two dimensions, 5 (12.5%) borderline objects are detected in prediction. In three dimensions, 16 (40%) borderline objects are detected. In four and five dimensions, 20 (50%) and 27 (67.5%) samples are detected as borderline objects, respectively.

There are more samples situated on the border when more dimensions are considered for the same calibration set. If the same population is considered in, e.g. one, two and three dimensions, the number of border samples in one dimension is two, in two dimensions all samples on the convex two dimensional hull are on the border. This is also true in three dimensions, but it is easy to understand that there will be more samples on the convex hull in three than in two dimensions. Consider now a population consisting of the calibration set and one new prediction object. As the number of dimensions grows the chance that this new object will be borderline object, as defined by us, increases. When it is borderline, it does belong to the population (is not a real outlier), but it will be outside the mesh of the calibration samples.

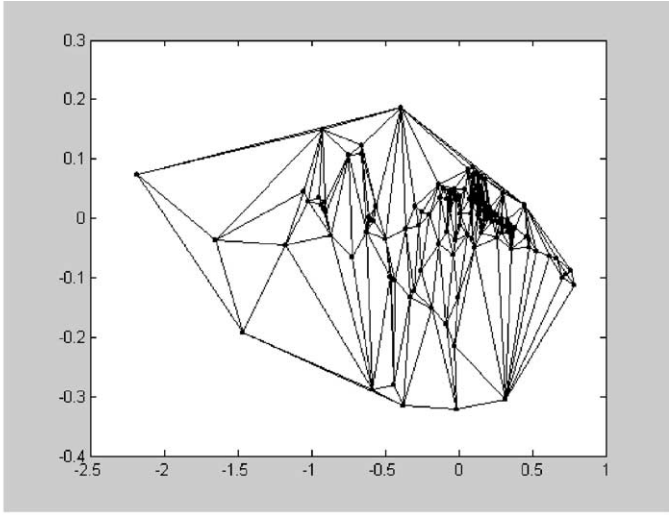
Table 1 shows the RMSECV values using leave-one-out cross-validation for all objects in the calibration dataset and the root mean square error for prediction (RMSEP) values after removing the borderline objects detected by DT. The RMSEP is determined according to the following equation:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)^2}{N}} \tag{8}$$

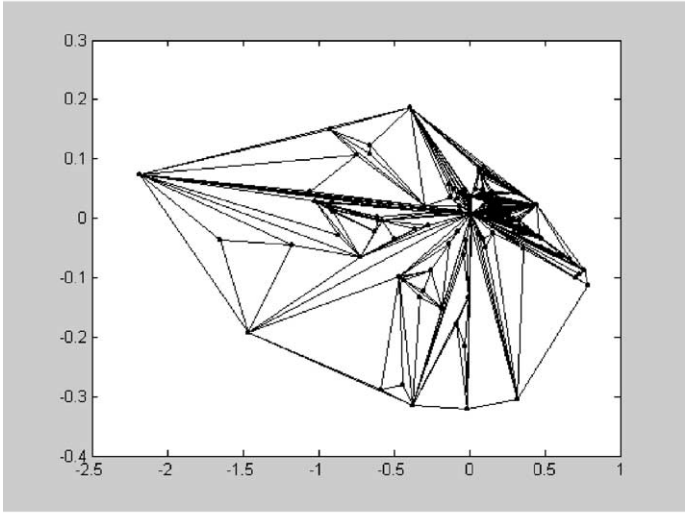
where N is the number of objects in the prediction set, $\hat{\boldsymbol{y}}_i$ and $\boldsymbol{y}_i$ are the predicted and the experimental property of the ith object, respectively.

Table 1
Dataset 1: RMSECV is obtained for all object in calibration set and RMSEP for the prediction of the prediction set from which the borderline objects have been removed

Dimensions	Number of borderline objects detected (%)		PCR	PLS	LWR	LM	DT
2	5 (12.5)	RMSECV	0.108	0.099	0.076	0.097	0.095
		RMSEP	0.148	0.134	0.141	0.112	0.138
3	16 (40.0)	RMSECV	0.071	0.089	0.065	0.084	0.068
		RMSEP	0.065	0.094	0.085	0.072	0.047
4	20 (50.0)	RMSECV	0.054	0.062	0.047	0.050	0.045
		RMSEP	0.066	0.062	0.048	0.050	0.044
5	27 (67.5)	RMSECV	0.052	0.049	0.043	0.049	0.042
		RMSEP	0.075	0.069	0.056	0.055	0.045



(a) PC1-PC3 plot



(b) PC1-PC3 plot

Fig. 4. The triangles obtained in two dimensions with (a) the DT method and (b) the LM method, respectively.

When more than three dimensions are used, LWR, LM and mainly DT become somewhat better than the global techniques for predicting objects within the calibration set.

For the determination of the property of the borderline objects, all three approaches are applied. In approach 2, the limits [−1, 2], [−2, 3], [−3, 4] and [−4, 5] are considered in order to compare them. For this dataset, all the borderline objects can be predicted when the limit [−1, 2] is used. Fig. 5 shows the RMSEP value of this dataset in different dimensions for those limits. It is clear that the limit [−1, 2] gives the minimum RMSEP in all cases except for two dimensions.

The RMSEP values for the borderline objects in prediction in different dimensions using the different

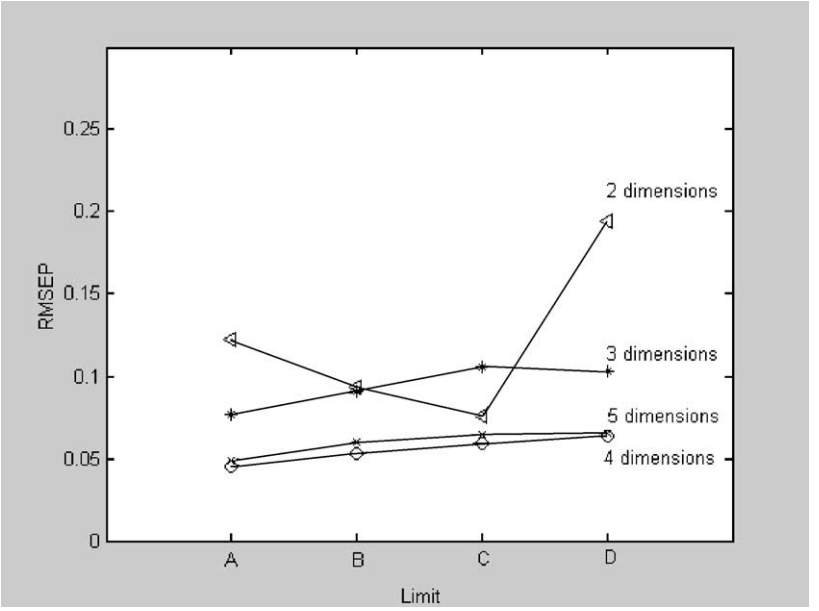


Fig. 5. The RMSEP values for dataset 1 in different dimensions for different limits of the coefficients using approach 2, where A represents the limit [−1, 2]; B represents the limit [−2, 3]; C represents the limit [−3, 4]; and D represents the limit [−4, 5].

Table 2
Dataset 1: RMSEP for the borderline objects using approaches 1–3

Dimensions	Approach 1 (borderline objects)	Approach 2 (borderline objects)	Approach 3 (borderline objects)
2	0.165	0.123	0.168
3	0.103	0.077	0.105
4	0.095	0.045	0.060
5	0.103	0.049	0.090

approaches are shown in Table 2. In four and five dimensions, the results using approach 2 are clearly better.

In order to compare the prediction ability of PLS, PCR, LWR, LM and DT (with approach 2 for borderline objects), the RMSEP values from these methods for all the objects, i.e. now including the borderline objects, are compared in Table 3. The DT was carried out both in the space of PCs ranked according to the variance (top-down PCs) and of PCs ranked according to correlation with y (selection PCs). The results with selected PCs appears to be better than the results from top-down PCs in DT with different dimensions. The optimal result (0.045) from DT with selection PCs when four dimensions are used is better than the results from PCR (0.068), PLS (0.064) and LWR (0.055) when five components are used and the result from LM (0.084) and DT with top-down PCs (0.057) with four dimensions. For this case, the DT (with selection PCs and approach 2 for borderline objects) requires

Table 3
Dataset 1: comparison of the RMSEP obtained from the DT method (approach 2 for borderline objects) with that obtained from PCR, PLS, LWR and LM for all objects (borderline objects + points within the calibration set space)

Dimensions	PCR	PLS	LWR	LM	DT (top-down PCs)	DT (selection PCs)
2	0.154	0.140	0.137	0.112	0.263	0.136
3	0.086	0.117	0.101	0.089	0.133	0.061
4	0.072	0.075	0.061	0.084	0.057	0.045
5	0.068	0.064	0.055	0.092	0.065	0.048

Table 4
Dataset 2: RMSECV is obtained for all object in calibration set and RMSEP for the prediction of the prediction set from which the borderline objects have been removed

Dimensions	Number of borderline objects detected (%)		PCR	PLS	LWR	LM	DT
2	3 (3.41)	RMSECV	1.18	2.20	1.52	1.22	1.23
		RMSEP	1.24	2.16	1.32	1.21	1.00
3	8 (9.09)	RMSECV	1.08	1.38	1.30	1.11	1.04
		RMSEP	1.05	1.40	1.24	0.98	0.81
4	18 (20.45)	RMSECV	1.01	1.16	1.10	0.98	0.90
		RMSEP	1.01	1.16	0.96	0.89	0.72
5	28 (31.82)	RMSECV	0.95	1.11	0.88	0.93	0.85
		RMSEP	0.90	0.94	0.70	0.92	0.73

fewer dimensions and yields a better RMSEP than the other methods. The randomisation test showed that the differences in RMSEP with PCR, PLS and LM were significant.

4.2. Dataset 2

Dataset 2 was split into two independent datasets by the providers of the data. It contains 205 samples for calibration and 100 samples for prediction.

With PCR the minimum RMSECV is obtained using 11 components and the randomisation test gives 10 components. For PLS regression the minimal RMSECV is obtained when 18 components are used, the randomisation test gives 16 and the MCCV method shows that one can consider five latent variables. In the DT method, five dimensions are considered. The first five important PCs are the 5th, 1st, 8th, 14th and 10th PC.

This dataset was well studied in [19] and it was concluded that 12 real prediction outliers are present. The real outliers were removed, leaving 88 samples in the test set. There are three borderline objects (i.e. not real outliers) for two dimensions; 8, 18 and 28 for three, four and five dimensions, respectively.

The results from PCR, PLS, LWR and LM are compared with the results from DT. The RMSECV values for all objects in the calibration dataset and the RMSEP values using PCR, PLS, LWR, LM and DT after removing the borderline objects are shown in Table 4. As in dataset 1 the results using DT are similar or a little better than the rest of the methods. Again, it appears that DT gives good results for samples within the calibration space.

The three approaches to obtain the property of the borderline objects are also investigated. The results are compared in Table 5. In approach 2, two objects are outside the subspaces, which are created with the limit $[-1, 2]$ in two dimensions. The limit $[-2, 3]$ is then used for these objects. The results from approach 2 are better than those from approaches 1 and 3.

Table 6 shows the results from PLS, PCR, LWR, LM and DT (with approach 2 for borderline object) for all samples in prediction. For the same number of dimensions the results from DT with selection of PCs are better than those for PLS, PCR, LWR, LM and DT with top-down PCs. When the DT method with selection PCs in five dimensions is used, the RMSEP is 0.82 which is similar to the result from LWR in five dimensions (0.87) and comparable with the minimal RMSEP values from PLS (0.72) and PCR (0.76) when 11 components are used and DT with top-down PCs (0.75) when eight dimensions are used.

Table 5
Dataset 2: RMSEP for the borderline objects using approaches 1–3

Dimensions	Approach 1 (borderline objects)	Approach 2 (borderline objects)	Approach 3 (borderline objects)
2	1.88	1.82	1.88
3	2.57	1.44	2.37
4	2.58	1.68	2.25
5	1.76	1.01	1.28

Table 6
Dataset 2: comparison of the RMSEP obtained from the DT method (approach 2 for borderline objects) with that obtained from PCR, PLS, LWR and LM for all objects (borderline objects + points within the calibration set space)

Dimensions	PCR	PLS	LWR	LM	DT (top-down PCs)	DT (selection PCs)
2	1.48	2.17	1.38	1.20	1.63	1.04
3	1.19	1.59	1.42	1.28	1.68	0.88
4	1.34	1.30	1.19	1.38	1.73	0.96
5	1.05	1.36	0.87	1.22	1.02	0.82

4.3. Dataset 3

This set was chosen because it was studied repeatedly using other methods [4,26,27] and because it has a relatively small number of objects and a high complexity. Then, the number of borderline objects should be very high and this dataset should present a worst case situation. Three real outliers were found by Centner and Massart [26], in that case 84 samples out of the 87 original ones were left. It was split into two subsets by using random selection: 60 samples for calibration and 24 samples for prediction.

For the PLS regression, a complexity of 8 is obtained according to the minimal RMSECV and 7 when the randomisation test and MCCV method are used. For PCR the minimum RMSECV is obtained using 10 components and the randomisation test and MCCV give eight and nine components, respectively. For the DT method, the RMSECV values for different dimensions are shown in Table 7. Seven dimensions are considered to be optimal. The first seven important PCs according to their correlation to $\boldsymbol{y}$ are the second, first, fifth, fourth, eighth, third and sixth PC.

In two dimensions, two borderline objects are detected in the test set (8.33% of the objects), in three to six dimensions, 7 (29.2%), 11 (45.8%), 14 (58.3%) and 19 (79.2%) borderline objects are detected in prediction, respectively. In seven dimensions, all prediction objects are outside the mesh constructed by the calibration data.

The RMSECV are shown in Table 7. It is clear that the minimal values are obtained when seven components are used. The RMSEP from PCR, PLS, LWR, LM and DT are also shown after removing the borderline objects. For seven dimensions, because no objects in prediction are inside the mesh that is constructed by the calibration data, a comparison is not possible and approaches 1–3 are applied.

Table 7
Dataset 3: RMSECV is obtained for all object in calibration set and RMSEP for the prediction of the prediction set from which the borderline objects have been removed

Dimensions	Number of borderline objects detected (%)		PCR	PLS	LWR	LM	DT
2	2 (8.33)	RMSECV	5.29	5.09	4.36	4.45	4.73
		RMSEP	4.94	4.71	4.17	4.15	4.56
3	7 (29.2)	RMSECV	4.34	4.21	3.79	4.20	4.19
		RMSEP	4.43	3.08	2.85	3.29	3.11
4	11 (45.8)	RMSECV	3.61	3.41	3.15	3.48	3.31
		RMSEP	2.75	2.41	2.16	2.35	2.09
5	14 (58.3)	RMSECV	2.88	2.92	2.18	2.04	2.52
		RMSEP	1.95	2.23	1.38	1.87	1.80
6	19 (79.2)	RMSECV	2.57	2.38	1.64	2.06	2.01
		RMSEP	2.82	1.33	1.48	2.82	2.75
7	24 (100)	RMSECV	2.10	1.61	1.44	1.77	1.79
		RMSEP	–	–	–	–	–

Table 8
Dataset 3: RMSEP for the borderline objects using approaches 1–3

Dimensions	Approach 1 (borderline objects)	Approach 2 (borderline objects)	Approach 3 (borderline objects)
2	9.31	14.8	11.9
3	3.37	2.01	2.50
4	3.71	2.12	3.84
5	2.89	1.71	2.04
6	3.68	1.28	1.44
7	2.65	1.47	1.52

The three proposed approaches for borderline objects are used as in the previous datasets in different dimensions. For this dataset, in three to seven dimensions, the best results are from approach 2 (Table 8). As in dataset 1, all the borderline objects are predicted using the limit [−1, 2] in approach 2. Normally the prediction for objects inside the mesh is better than for objects outside the mesh (borderline objects). However, it must be remembered that borderline objects are not real outliers and are therefore quite close to the calibration objects. They are therefore in general predicted quite well and in some cases better than one would expect. For instance, in six dimensions, the RMSEP value for the borderline objects is 1.28 (approach 2), which is better than the result for the objects inside the mesh (2.75), and similar to the RMSEP obtained for PLS and LWR. This shows that in this case it is advantageous to use more objects for averaging out errors in prediction.

In this dataset, the results (RMSEP) from different techniques for all objects are compared in Table 9. The results from DT in the space of selection PCs are better than in the space of top-down PCs (with approach 2 for borderline objects). For PCR, PLS, LWR, LM and DT with top-down PCs, the optimal RMSEP values are obtained when 9 (1.71), 7 (1.69), 7 (1.55), 7 (1.72) and 7 (1.78) components are considered, which are somewhat higher than the result from DT with approach 2 and selection of PCs in seven dimensions (1.47), but the randomisation test shows that the differences are not significant.

5. Conclusions

The results using DT in all datasets presented here are at least comparable with the results from PCR, PLS and LWR. Because we have shown in an earlier article [27] that PLS performs better than the simple kNN topological methods, it follows that DT is also superior to these kNN methods. Because the DT method uses the natural neighbourhood to construct the triangulation, the results from DT are better than the results from the LM method. Another advantage of the DT method is that fewer components are used for prediction as compared to PCR and PLS. In dataset 2, 11 components are necessary in PCR and PLS to achieve the minimal RMSEP, but DT requires only 5.

Borderline objects must be expected to occur very often. How often depends on the number of objects in the calibration set and the number of dimensions. Therefore, special attention was given to this problem. For the prediction of the borderline objects, we proposed three approaches. The results from approach 2, where all possible triangles with coefficients within a given limit are used, are always better than approaches 1 and 3. This is probably caused by the fact that more objects are used for prediction in approach 2 and therefore errors are averaged out to a larger extent. Approach 2 is therefore recommended to predict the

Table 9
Dataset 3: comparison of the RMSEP obtained from the DT method (approach 2 for borderline objects) with that obtained from PCR, PLS, LWR and LM for all objects (borderline objects + points within the calibration set space)

Dimensions	PCR	PLS	LWR	LM	DT (top-down PCs)	DT (selection PCs)
2	5.03	4.79	4.39	4.99	6.11	6.11
3	4.06	3.23	2.78	3.08	2.68	2.83
4	2.58	2.37	2.18	2.20	2.39	2.10
5	1.93	2.04	1.77	1.98	2.04	1.75
6	2.02	1.87	1.70	2.21	1.82	1.69
7	1.85	1.69	1.55	1.72	1.78	1.47

property of borderline objects. The DT method with approach 2 applied in all these datasets gives similar results to PCR, PLS and LWR and gives better results than the recently proposed LM method. In order to achieve even better results in approach 2, the limitation of the coefficients will be further investigated. As general conclusion one can say that the DT method not only performs well for objects within the calibration set, but also for borderline objects. It is noteworthy that even for dataset 3 with seven components, where all prediction objects are borderline objects (and thus near to, but not within the limits of the calibration set), the DT results are good. It should be noted here that the MSI method does not consider the borderline problem.

DT in the space of PCs ranked according to correlation with y gave somewhat better results than in the space with top-down PCs. However, more evidence is needed to conclude this definitively.

The most evident advantage of topological local methods compared to both global and local methods based on PLS is that the method is very simple to understand as it is simply based on taking averages, while PLS remains difficult to understand for the non initiated. Compared to global methods local methods do not require attention to the possible presence of non-linearity. One of the main difficulties in the continued use of regression based calibration methods is that they are not simple to update, i.e. to include new calibration samples. With topological methods this is much simpler. It has been shown that new points can be inserted in a DT mesh in a simple way without having to compute the mesh all over again [28].

Updating the calibration must be still simpler with the MSI method. However, using a predefined mesh as in the DT method is much simpler than having to rank simplexes as would be the case if MSI were used for solving the borderline samples problem. Using the latter method for borderline samples would mean considering simplexes for many possible combinations of points. Moreover, predictions with DT are continuous when the prediction point inside the mesh moves around. This is not the case for MSI, because the simplex on which the prediction is based may change.

This does not mean that the DT method is entirely free from problems. For instance, if there is a measurement error in the reference method of a certain calibration sample, this will affect to a larger extent prediction samples in that neighbourhood than it would in PLS. A possible disadvantage of all topological local methods is that they cannot be interpreted in terms of regression coefficients, so that at first sight it is less simple to decide which variables are important and which ones are not.

We can conclude that the DT method is shown to perform well for multivariate calibration. The DT could serve as a valuable tool in multivariate calibration, certainly in cases where the more common global techniques are less efficient. Local topological methods, such as the DT method and the MSI method, are not very widespread and merit a larger acceptance as they show a number of advantages. Further research into such methods would therefore be useful.

Acknowledgements

The authors thank an anonymous referee for giving useful suggestions to improve the manuscript.

References

[1] T. Naes, T. Isaksson, B. Kowalski, Anal. Chem. 62 (1990) 664–673.
[2] Z. Wang, T. Iasksson, B.R. Kowalski, Anal. Chem. 66 (1994) 249–260.
[3] Y. Xie, J.H. Kalivas, Anal. Chim. Acta 348 (1997) 29–38.
[4] V. Centner, D.L. Massart, Anal. Chem. 70 (1998) 4206–4211.
[5] C.J. Stone, Ann. Stat. 5 (1977) 595–645.
[6] A. Espinosa, M. Sanchez, S. Osta, C. Boniface, J. Gil, A. Martens, B. Descales, D. Lambert, M. Valleur, Oil Gas J. 17 (1994) 49–56.
[7] R. Danielsson, G. Malmquist, Chem. Intell. Lab. Syst. 14 (1992) 115–128.
[8] L. Jin, J.A. Fernández Pierna, F. Wahl, P. Dardenne, D.L. Massart, Anal. Chim. Acta 476 (2003) 73–84.
[9] A. Okabe, B. Boots, K. Sugihara, Spatial Tessellation: Concepts and Application of Voronoi Diagrams, Wiley, Chichester, 2000.
[10] J. Gudmundsson, M. Hammar, M. van Kreveld, Comput. Geometry 23 (2002) 85–98.
[11] http://www.sepwww.stanford.edu/oldsep/sergey/sepsergey/fmeiko/paper_html/node9.html.
[12] http://www.goanna.cs.rmit.edu.au/~gl/research/comp_geom/delaunay/delaunay.html.
[13] http://www.cage.rug.ac.be/~dc/alhtml/delaunay.html.
[14] http://www.mathworks.com/access/helpdesk/help/techdoc/math_anal/poly_i19.shtml.
[15] M.V. Anglada, N.P. Garcia, P.B. Crosa, Comput. Aided Geometric Design 16 (1999) 107–126.

[16] http://www.gris.uni-tuebingen.de/gris/proj/dt/dteng.html.
[17] 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.
[18] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.
[19] J.A. Fernández, L. Jin, M. Daszykowski, F. Wahl, D.L. Massart, in: Proceeding in the 2nd International Symposium on PLS and Related Methods, Capri, Italy, 2001, Chem. Intell. Lab. Syst.
[20] I. Ruisánchez, F.X. Rius, S. Maspoch, J. Coello, T. Azzouz, R. Tauler, L. Sarabia, M.C. Ortiz, J.A. Fernández, D.L. Massart, A. Puigdomènech, C. García, Chem. Intell. Lab. Syst. 63 (2) (2002) 93–105.
[21] R.D. Snee, Technometrics 19 (1977) 415–428.
[22] H. van der Voet, Chem. Intell. Lab. Syst. 25 (1994) 313–323; H. van der Voet, Chem. Intell. Lab. Syst. 28 (1995) 315.
[23] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486–494.
[24] R.R. Picard, R.D. Cook, J. Am. Stat. Assoc. 79 (1984) 575–583.
[25] Q. Xu, Y. Liang, Chem. Intell. Lab. Syst. 56 (2001) 1–11.
[26] V. Centner, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 330 (1996) 1–17.
[27] V. Centner, J. Verdú-Andrés, B. Walczak, D. Jouan-Rimbaud, F. Despagne, L. Pasti, R. Poppi, D.L. Massart, O.E. Noord, Appl. Spec. 54 (2000) 608–623.
[28] J. Boissonnat, M. Teillaud, Theor. Comput. Sci. 112 (1993) 339–354.