Proceedings of the 9th International Symposium on
Dynamics and Control of Process Systems (DYCOPS 2010),
Leuven, Belgium, July 5-7, 2010
Mayuresh Kothare, Moses Tade, Alain Vande Wouwer, Ilse Smets (Eds.)

A continuous regression function for the Delaunay calibration method

Francesco Corona * **Elia Liitiäinen** * **Amaury Lendasse** *
Roberto Baratti ** **Lorenzo Sassu** ***

* *Department of Computer and Information Science, Helsinki University of Technology, Helsinki, Finland (francesco.corona,elia.liitiainen,amaury.lendasse@hut.fi)*
** *Department of Chemical Engineering and Materials, University of Cagliari, Cagliari, Italy (baratti@dicm.unica.it)*
*** *Process Department, Sartec - Saras Ricerche e Tecnologie S.p.A., Assemini, Italy (lorenzo.sassu@sartec.it)*

Abstract: The Delaunay tessellation and topological regression is a local simplex method for multivariate calibration. The method, developed within computational geometry, has potential for applications in online analytical chemistry and process monitoring. This study proposes a novel approach to perform prediction and extrapolation using Delaunay calibration method. The main property of the proposed extension is the continuity of the estimated regression function also outside the calibration domain. To support the presentation, an application in estimating the aromatic composition in Light Cycle Oil by Near Infrared spectroscopy is discussed.

Keywords: Process monitoring, Multivariate calibration, Model maintenance, Spectroscopy, Delaunay tessellation, Non-parametric regression, Local estimation.

1. INTRODUCTION

Real-time monitoring is an essential component in modern process industry for optimizing production toward high-quality products while reducing operating and off-specification costs. The tools of process analytical chemistry like Infrared (IR) and Near Infrared (NIR) spectroscopy fulfill the necessary requirements for real-time analysis of important properties for a broad variety of materials, because based on inexpensive and continuously acquired spectral measurements (Workman, 1999).

The principle underlying process monitoring from spectra is the existence of a relationship between the spectrum of a given product and the property of interest. The relationship is rarely known *a priori* but it can be reconstructed from data by learning specifically tailored multivariate calibration models. Multivariate calibration methods are often divided into local and global approaches. The latter use all known (calibration) observations to learn the parameters of a single regression model. The former use only small subsets of the calibration data to build different calibration models located in the neighborhood of the observation whose properties have to be estimated. Widely used parametric models like Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) exist in both local and global variants (Gemperline, 2006). Among local methods, non-parametric approaches based on nearest neighbors or topological regression (Stone, 1977), have gained recent interest, mostly driven by industrial motivation (see Espinosa et al. (1994); Jin et al. (2003a,b) and references therein). This is because such methods are mostly non-parametric, possess an inherent ability to handle nonlinearities and, what is more important here, the possibility to minimize models' maintenance tasks while retaining the prediction accuracy. In fact, the number of spectroscopic models typically used in a production plant is rapidly increasing, and this implies money and time consuming trained personnel for design, calibration and maintenance of the estimation models.

With the scope to investigate alternative calibration methods that could reduce the maintenance costs associated to continuous recalibrations, the authors discussed an application of the Delaunay Tessellation and Topological Regression method (DTR) by Jin et al. (2003b) to calibrating the aromatic composition in Light Cycle Oil (LCO) by NIR spectroscopy (Corona et al., 2009). The DTR method was considered for its potentiality to achieve accuracies comparable with PCR and PLSR models while being much simpler to develop (a single model can be calibrated for all the properties to be estimated) and maintain/upgrade (Jin et al., 2005, 2006). In order to assess the potentiality of the method, a feasibility study with comparison to standard calibration methods was successfully performed. The study also highlighted a main limitation of the DTR method: a scarce extrapolation ability accompanied by a lack of stable methods for estimating a continuous regression function also outside the calibration domain.

This work proposes a consistent extension to the Delaunay Tessellation and Topological Regression method that permits to estimate a continuous regression function also for the observations situated outside the calibration domain. The paper is organized as follows: Section 2 overviews the DTR method and presents the new approach for the prediction of borderline objects and Section 3 discusses the results obtained for the estimation of aromatics in LCO.

2. THEORY

The Delaunay Tessellation and Topological Regression (DTR) proposed by Jin et al. (2003b) is a local multivariate calibration method developed from arguments in computational geometry . In its basic form, the DTR method consists of the following three main steps:

(1) a dimensionality reduction based on a set of known input observations (e.g., NIR absorbance spectra);
(2) the generation (in the low-dimensional space) of an unstructured mesh by Delaunay tessellation;
(3) a nearest neighbors (or topological) regression for estimating the outputs (e.g., aromatics in hydrocarbon mixtures) for a set of unknown observations.

This section overviews the steps in the DTR method for a set of observations $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $\mathbf{y}_i \in \mathbb{R}^P$ are the inputs (on-line spectrum) and output (off-line analysis) variables for the $i-$th observation, respectively.

2.1 Dimensionality reduction

Because of the high dimensionality of the input spectra $\mathbf{x}$ (usually hundreds, up to thousands) and the small number of samples N (usually tens), it is appropriate to operate in a reduced data space whose dimensionality is circumscribed by the intrinsic complexity of the observed system. The dimensionality reduction step thus aims at projecting the input observations onto a system of lower coordinates in such a way that certain properties of the original data points $\mathbf{x}_i$ are preserved as faithfully as possible by a new set of data points $\mathbf{x}'_i \in \mathbb{R}^S$, with $S \ll D$.

The mapping can be either driven only by the inputs (e.g., as in PCR) or by both the inputs and outputs (e.g., as in PLSR). In general, there is a wide range of methods for performing dimensionality reduction (Lee and Verleysen, 2007) that can be considered in this step. However, this study is confined to a projection based only on input data, because this representation can be common to all the output properties to be estimated; hence, capable to minimize problems and costs associated to models' recalibration and maintenance.

For the sake of simplicity, a Principal Component Analysis (PCA) is used to characterize the experiments; in that sense, the property of the data points that is preserved by the mapping is in the set of pair-wise distances between them (Jolliffe, 2002).

2.2 Delaunay tessellation

Once the input observations are projected onto a low dimensional system of coordinates (e.g., the principal components), the known part of this space is partitioned, by generating a mesh using all the available data points. The elements of the mesh are simplices delimited by known observations (i.e., projected input data points with known values for the output properties, the calibration set). Within each simplex, locality conditions are assumed because similar data should be mapped close to each other.

A well-known method for generating a mesh of simplices is the Delaunay tessellation (Gudmundsson et al., 2002). For a given set of point observations in two dimensions, the Delaunay tessellation constructs an unstructured mesh of triangular simplices (hence, the common name Delaunay triangularization) by using all the input data points as vertices; one triangle is a valid simplex if and only if its circumcircle does not enclose any other point in the set (the empty circle condition). The mesh is constructed in order to maximize the minimum angle and thus avoids the generation of spiky simplices. The Delaunay triangulation always exists and it is also unique, if no three points are on the same line and no four points are on the same circle.

In three dimensions the simplices are tetrahedrons and, for a reduction to S dimensions, the elements of the tessellation are polyhedrons defined by $K = S + 1$ points. In the general S-dimensional case, existence and uniqueness of the tessellation is also guaranteed, if no $K+1$ points are on the same hyper-plane and no $K + 2$ points are on the same hyper-sphere. Notice that the DTR method requires a mesh generation performed on as many dimensions as those obtained in the dimensionality reduction step.

2.3 Topological regression

Once the mesh is built, it is used for estimating the properties of new observations (i.e., data points for which only the input values are known). Topological regression is performed after projecting also the new observations onto the same low dimensional system obtained in the first step.

Inside objects The standard case for estimation is when a new observation $\mathbf{x}'_i$ (in the S-dimensional projection space) happens to fall within the convex hull that contains all the know data points. Since the union of all simplices in the tessellation is the convex hull of the points, the new data point also falls within one of the simplices.

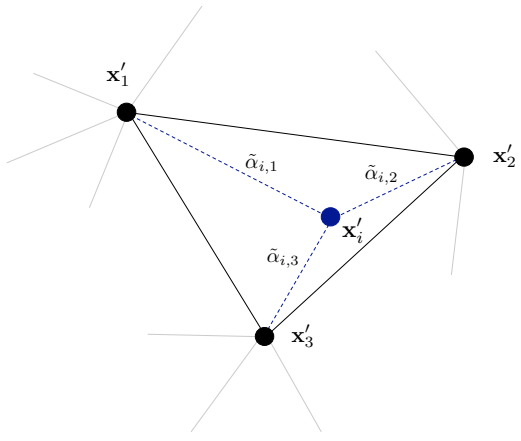


Fig. 1. Property estimation for inside points, on the plane.

In two dimensions, the enclosing simplex is a triangle with vertexes $\mathbf{x}'_1$, $\mathbf{x}'_2$ and $\mathbf{x}'_3$ (three known observations for which also the values of the properties $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$ have been measured), and the position of the new observation $\mathbf{x}'_i$ with respect to its three neighboring points (i.e., the vertexes) is expressed as a linear combination or weighted sum of their input coordinates subjected to the convexity constraints

(i.e., the weights $\boldsymbol{\alpha}_i$ are non-negative and sum up to one). For each such a new point, only the enclosing triangular simplex (Figure 1) fulfills the convexity constraints, with weights that will be also bounded to the unit interval.

The equation used for calculating the set of weights $\boldsymbol{\alpha}_i$ is

$$\begin{pmatrix}\alpha_{i,1}\\ \alpha_{i,2}\\ \alpha_{i,3}\end{pmatrix} = \begin{pmatrix}x'_{1,1} & x'_{2,1} & x'_{3,1}\\ x'_{1,2} & x'_{2,2} & x'_{3,2}\\ 1 & 1 & 1\end{pmatrix}^{-1} \begin{pmatrix}x'_{i,1}\\ x'_{i,2}\\ 1\end{pmatrix}, \qquad (1)$$

where $x'_{k,s}$ denotes the s-th coordinate of the k-th vertex of the enclosing triangle (here, $s \in \{1,2\}$ and $k \in \{1,2,3\}$).

The resulting weights $\alpha_{i,1}$, $\alpha_{i,2}$ and $\alpha_{i,3}$ are the barycentric coordinates of $\mathbf{x}'_i$ with respect to the vertices $\mathbf{x}'_1$, $\mathbf{x}'_2$ and $\mathbf{x}'_3$ of the triangle, with $0 \leq \alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3} \leq 1$ and $\alpha_{i,1}+\alpha_{i,2}+\alpha_{i,3} = 1$ being the enclosing simplex also convex. Thus, the weights can be understood as the contributions of these known observations to the new observation. Because of the local linearity assumption within simplices, any of the P properties $y_{i,p}$ of the new observation is then estimated from a linear combination (with convexity constraints) of the properties of the known observations:

$$\hat{y}_{i,p} = \alpha_{i,1}y_{1,p} + \alpha_{i,2}y_{2,p} + \alpha_{i,3}y_{3,p}, \qquad (2)$$

where $y_{1,p}$, $y_{2,p}$ and $y_{3,p}$ are the values of the p-th property at the vertexes (for any $p \in \{1,2,\ldots,P\}$). Once the weights are calculated, estimating any property is thus immediate and a complete and common map of the distribution of the output properties inside the calibration domain is easily constructed. Equation 1 and 2 easily generalize to any S-dimensional tessellation with K-hedrons.

For the general S-dimensional case, the consistency of the regression inside a simplex is shown in the noise-free case by taking the first-order Taylor expansion of the input-output relationship $y_p = f(\mathbf{x}')$ about the input point $\mathbf{x}'_i$,

$$f(\mathbf{x}') \approx f(\mathbf{x}'_i) + (\mathbf{x}' - \mathbf{x}'_i)^\top \nabla f(\mathbf{x}'_i).$$

Evaluating the truncated expansion at any of the vertexes $\mathbf{x}' = \mathbf{x}'_k$ with $f(\mathbf{x}') = f(\mathbf{x}'_k)$, gives the expression

$$f(\mathbf{x}'_k) \approx f(\mathbf{x}'_i) + (\mathbf{x}'_k - \mathbf{x}'_i)^\top \nabla f(\mathbf{x}'_i).$$

Because the general expression for the estimates is $\hat{y}_{i,p} = \sum_{k=1}^{K} \alpha_{i,k} y_{k,p}$ where $y_{k,p} = f(\mathbf{x}'_k)$, substituting the expansion into the estimation function and re-arranging yields

$$\sum_{k=1}^{K} \alpha_{i,k} y_{k,p} \approx \sum_{k=1}^{K} \alpha_{i,k} \Big(f(\mathbf{x}'_i) + (\mathbf{x}'_k - \mathbf{x}'_i)^\top \nabla f(\mathbf{x}'_i) \Big)$$

$$\hat{y}_{i,p} \approx \sum_{k=1}^{K} \alpha_{i,k} f(\mathbf{x}'_i) + \sum_{k=1}^{K} \alpha_{i,k} (\mathbf{x}'_k - \mathbf{x}'_i)^\top \nabla f(\mathbf{x}'_i).$$

Now, because $\mathbf{x}'_i = \sum_{k=1}^{K} \alpha_{i,k} \mathbf{x}'_k$, with $\sum_{k=1}^{K} \alpha_{i,k} = 1$, then

$$\begin{aligned}\hat{y}_{i,p} &\approx f(\mathbf{x}'_i) + (\sum_{k=1}^{K} \alpha_{i,k} \mathbf{x}'_k - \sum_{k=1}^{K} \alpha_{i,k} \mathbf{x}'_i)^\top \nabla f(\mathbf{x}'_i)\\ &= f(\mathbf{x}'_i) + (\mathbf{x}'_i - \mathbf{x}'_i)^\top \nabla f(\mathbf{x}'_i)\\ &= f(\mathbf{x}'_i),\end{aligned}$$

which demonstrates how for any point $\mathbf{x}'_i$ in the convex set, the estimation function is exact up to the second order (i.e., for the linear case assumed inside the simplices).

The underlying relationship is thus estimated by a piecewise linear regression function, which is continuous and continuosly differentiable inside the simplices (i.e., local $\mathcal{C}^1$ continuity) and continuous but not continuously differentiable at their junction (i.e., local $\mathcal{C}^0$ continuity).

Outside objects The special case is for the estimation of a new observation that does not fall inside the convex hull defined by the known data points; hence, not even inside any of the constructed simplices. In this situation, Equation 1 still holds but only the affine constraints are satisfied (i.e., the weights $\boldsymbol{\alpha}_i$ are still summing up to one but they are not bounded to the unit interval anymore). Some of these observations are outliers (in a strict sense) but they can also be borderline objects located in region of the input space that was unknown when the initial calibration set was defined. It is worthwhile noticing that the main limitation of the DTR method is, in this sense, its near-absolute lack of extrapolation ability. However, this limitation is not as dramatic as it may appear, because of the simplicity to update both projection and tessellation to account for outlying points (Jin et al., 2005, 2006).

For the estimation of the properties of such observations, several approaches are reported in the literature. All the approaches rely on Equation 1 and 2 for the calculation of the weights and property estimation, but they differ in the way they select the simplex over which they are resolved.

Jin et al. (2003b) proposed three different approaches:

(1) find the simplex whose centroid is the closest to the outside point and then allow for negative weights without any further constraint (Jin 1);
(2) find a simplex whose weights can be negative but limited within some interval (e.g., $[-1,2]$, $[-2,3]$, $[-3,4]$ and so on). If more that one simplex is found, the final estimate is found by averaging over all the simplices (Jin 2);
(3) find the simplex such that $\max(|\alpha_{i,k}|)$ is minimized and then allow for negative weights without any further constraint (Jin 3).

However, none of the approaches is necessarily capable of estimating a continuous regression function. Moreover, the first approach could be unstable as the weights may explode and the second and third solution rely on arbitrary intervals and criteria for searching the closest simplex.

Corona et al. (2009) contributed the centroid method (cent.), where an estimate for a new external observation is obtained by projecting it onto the closest simplex, as identified by its centroidal point (similar to Jin 1). In that sense, an artificial data point with a set of identical positive weights in the unit interval that also sum up to one is constructed and the property is then estimated as equal to what would be calculated for the centroid of such a simplex, again using Equation 2. Although the approach is stable, a noncontinuous regression function is estimated.

This paper proposes a consistent regression function also for the external data points by looking for their closest projection onto the convex hull (proj.). Since the projected

points are in the convex set used for calibration, their weights and any of their properties can be calculated and estimated with expressions similar to Equation 1 and 2.

Concretely, for an outside point $\mathbf{x}'_i$, its projection on the hull is computed by finding the closest point by going through all the facets (segments, on the plane) that bound the tessellation. For each facet determined by $K-1$ end-points $\mathbf{x}'_1, \ldots, \mathbf{x}'_{K-1}$, its closest point to $\mathbf{x}'_i$ is obtained from

$$\min_{\alpha_{i,1},\ldots,\alpha_{i,K-1}} \left\| \sum_{k=1}^{K-1} \alpha_{i,k}\mathbf{x}'_k - \mathbf{x}'_i \right\|^2 \quad \text{s.t.} \quad \begin{cases} \sum_{k=1}^{K-1} \alpha_{i,k} = 1 \\ 0 \leq \alpha_{i,k} \leq 1 \end{cases} ;$$

that is, from the optimal and unique set of convex weights $\alpha_{i,1}, \ldots, \alpha_{i,K-1}$ that characterize its position on the facet while minimizing its distance to $\mathbf{x}'_i$. The projected point $\tilde{\mathbf{x}}'_i$ that is closest to $\mathbf{x}'_i$ is then found through all the facets.

With a projected point $\tilde{\mathbf{x}}'_i$ and a set of optimal convex weights $\tilde{\alpha}_{i,1}, \ldots, \tilde{\alpha}_{i,K-1}$ on a facet determined by known end-points $\tilde{\mathbf{x}}'_1, \ldots, \tilde{\mathbf{x}}'_{K-1}$, the estimate of any property p of $\mathbf{x}'_i$ is again immediate and consistent because obtained as a linear combination of properties of known observations:

$$\hat{y}_{i,p} = \sum_{k=1}^{K-1} \tilde{\alpha}_{i,k} y_{k,p}, \quad \text{with} \quad p \in \{1, \ldots, P\}. \qquad (3)$$

If for simplicity we consider the planar case (Figure 2), we observe that the projected point $\tilde{\mathbf{x}}'_i$ can be located either 1) on a segment of the convex hull or 2) be one of its vertexes. Case 1 is characterized by two nonzero weights out of three, whereas for case 2 only one weight is nonzero and it also equals one. Case 2 occurs when point $\mathbf{x}'_i$ is in the portion of the space bounded by the external normals to two contiguous segments (i.e., sharing the same vertex).

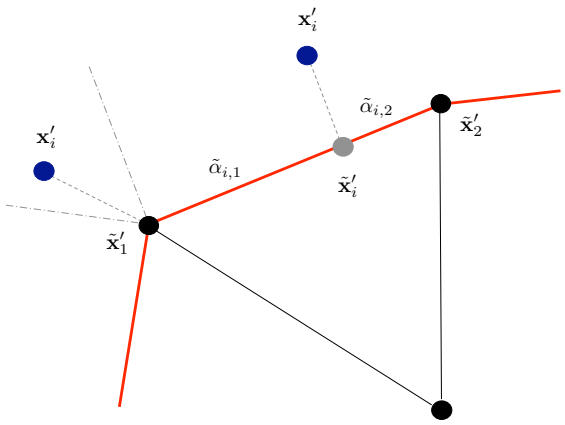


Fig. 2. Property estimation for outside points, on the plane. The red line denotes the bounding convex hull.

The two cases lead to two distinct modes of continuous variation of the estimates outside the calibration domain:

1) the solution remains constant when moving along the normal to the closest facet and varies linearly when moving orthogonally to the normal to the closest face;
2) the solution remains constant throughout the corresponding portion of the input space.

At the intersection between the two cases, the regression function is continuous but cannot be continuously differentiated because the estimates vary between piecewise linear and constant. In general, the solution is stable and does not rely on arbitrary criteria when compared to Jin et al. (2003b) and, it addresses the discontinuity issue associated with Corona et al. (2009) because it provides a continuous transition also at the boundary of the enclosing hull.

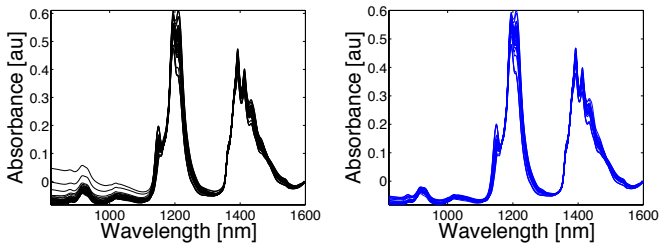
Again, a generalization of the proposed approach to any S-dimensional with K-hedrons is straightforward.

3. EXPERIMENTAL

The presented application is framed within the intense research activity that has characterized the recent trends in refining industry aimed at optimizing the use of low-value products. Light Cycle Oil (LCO) is a low-value stream in the diesel boiling range produced in Fluid Catalytic Cracking units. Due to its poor characteristics (e.g., a high total aromatics content, considerable percentages of compact structure poly-aromatics and a high sulfur content), LCO cannot be blended directly in the finished diesel fuel pool but it is preliminary upgraded to an higher value diesel in hydro-treatment units (where the poly-aromatics are hydrogenated). In order to satisfy the required process and environmental standards of hydro-treated products, rapid and cost effective (and possibly on-line) evaluation of the aromatic content is thus mandatory.

3.1 Materials

A total of 91 LCO and Hydro-treated LCO (HDT LCO) samples were acquired and used for the present study. The HDT LCO samples were obtained by Sartec S.p.A. in a bench-scale pilot unit (*Vinci Technologies*) operating at various temperatures and pressures, by processing from different LCO feeds provided by the Saras Refinery (Italy). The pilot unit mimics most typical industrial operations and ensures the range of variation in the total aromatic content and in the distribution the mono-, di- and tri-aromatics (AH) classes expected in the full-scale case.



(a) Learning data (b) Testing data

Fig. 3. The spectral measurements.

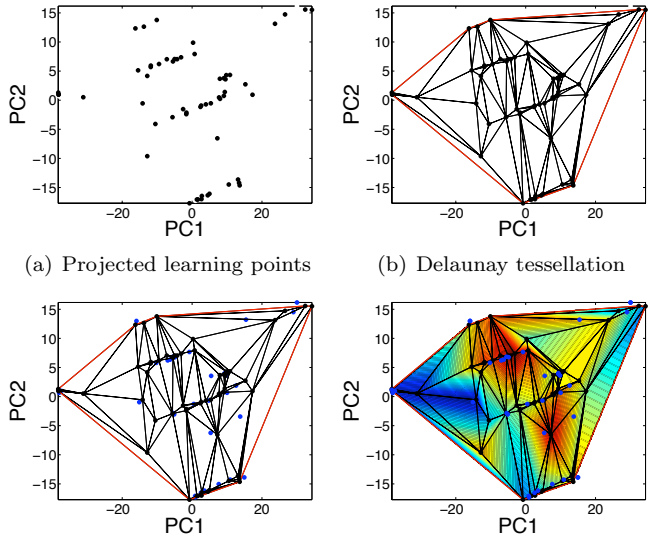
The NIR spectra of the samples (Figure 3) were recorded using a *Varian Cary 500 Scan* double-beam spectrometer in the wavelength range $1600-800nm$ with $1nm$ resolution ($\mathbf{x}_i \in \mathbb{R}^D$, with $D = 800$). The aromatic content ($w\%$), was determined with the HPLC method $EN-12916$ using an *Agilent 1100 Series* system with refraction index detection.

For the development of the multivariate models, the available data have been divided in calibration and testing sets ($N_c = 58$ and $N_t = 33$ observations). The two sets have been defined by Sartec S.p.A. in order to contain each some examples of all products' qualities and span the entire range of variation in the aromatics' concentration. As for the preprocessing of the spectral observations, the first derivative is used in the experiments.

3.2 Calibration

Based on the 58 samples in the calibration set, a dimensionality reduction with Principal Component Analysis has been performed, as a first step. As discussed in Section 2, the technique of choice uses only the input observations (the NIR spectra). After mean-centering the inputs, PCA is performed and the calibration (differentiated) spectra projected, Figure 4(a). The number of retained principal components is two ($\mathbf{x}'_i \in \mathbb{R}^S$, with $S = 2$). The selection is based on the inspection of the eigenvalues of the covariance matrix of the data; the two retained directions account for over 90% of the total variance observed in the input space.

Upon projecting the input observations in the calibration set onto the first two principal directions, a Delaunay tessellation has been performed. Each element of the mesh is a triangular simplex and the set of simplices is enclosed in a bi-dimensional convex hull, Figure 4(b). Subsequently, also the 33 testing (differentiated) spectra have been mean-centered (by removing the mean of the calibration set) and projected onto the same principal components space, Figure 4(c), where topological regression is performed.



(a) Projected learning points (b) Delaunay tessellation

(c) Projected testing points (d) Property distribution

Fig. 4. A graphical representation of the Delaunay method.

The regression model is found by resolving for the barycentric coordinates (Equation 1) of all the testing observations belonging to the convex set and, then, calculating the corresponding properties (Equation 2) from the known measurements. Again, it is worthwhile noticing that being the DTR model the same for all the properties to be estimated (the weights are calculated only once), only a single regression model is needed; thus, minimizing the

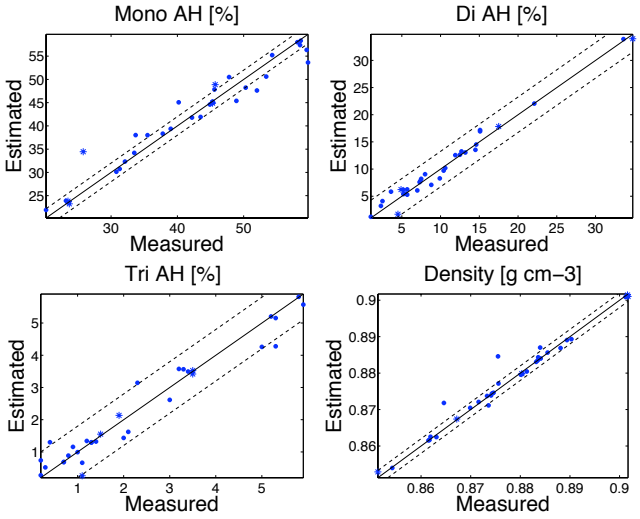


Fig. 5. Estimation results for the testing observations, including repeatability bands. Inside points are plotted as dots (·) and outside points as asterisks (∗).

calibration and maintenance tasks. Moreover, a common map (i.e., the estimated regression function) of the property distribution inside the calibration domain can be constructed, as depicted in Figure 4(d) where high values of the properties are dyed in red and low values in blue. As for the testing observations that do not fall inside the convex hull determined by the calibration points (only 5, and all rather close to the boundaries of the convex hull, Figure 4(c) and 4(d)), the projection method presented in Section 2 is used and a comparison with the other approaches available in the literature has been performed.

3.3 Results

The results for the independent set of 33 testing observations are depicted in Figure 5 and reported in Table 1 for the mono-, di- and tri-aromatics content. In addition, we are also reporting the results obtained when estimating the density of the samples: Density (gcm^{-3}) was measured according to the analytical method $ASTM - 4052$. It is worthwhile noticing that estimating such a property (as well as others not reported here) was straightforward because of the already calculated weights.

The accuracy of the estimation is reported in terms of Root Mean Squared Error for Prediction ($RMSEP$):

$$RMSEP_p = \sqrt{\frac{\sum_{i=1}^{N_t}(\hat{y}_{i,p} - y_{i,p})^2}{N_t}} \quad \text{with} \quad p \in \{1, \ldots, P\},$$

such a metric is preferred for it retains the original units of the measurements and thus also allows for a direct comparison with the repeatability of the analytical methods.

For all the properties, the accuracy of the estimates obtained with the DTR calibration and the projection method is found to be within the repeatability range of the analytical measurements. On this problem, the projection method has been always capable to outperform all the approaches proposed by Jin et al. (2003b). On the other hand, the centroid method proposed by Corona et al. (2009) remains a fairly accurate and simple alternative.

Since all the approaches perform equally on the data points that fall inside the convex hull bounding the calibration domain, all the differences in the estimates are only due to outside points. In that sense, the projection approach is not only theoretically more rigorous, but it has also demonstrated able to perform better on this practical case.

Table 1. Estimation results for the testing observations, as RMSEP. For the PLSR models, the cross-validated number of latent variables is also indicated.

	Mono AH [$w\%$]	**Di AH** [$w\%$]	**Tri AH** [$w\%$]	**Density** [gcm^{-3}]
DTR (proj.)	2.74	1.08	0.45	0.0022
DTR (cent.)	2.90	1.11	0.46	0.0023
DTR (Jin 1)	3.25	1.12	0.46	0.0022
DTR (Jin 2)	9.24	7.22	0.77	0.0145
DTR (Jin 3)	2.90	1.45	0.55	0.0021
PLSR	1.26(8)	0.67(7)	0.60(8)	0.0021(8)

For completeness, Table 1 also presents the results obtained with a set of PLSR models independently cross-validated by Leave One Out (Hastie et al., 2009) for the number of latent variables. Such models are presented because often more accurate but also over-parameterized (the number of latent variables is much higher than the two used by DTR) and thus clearly less robust and manageable, too. These limitations of the PLSR model are p-fold, when all the properties are considered.

4. CONCLUSION

Delaunay Tessellation and Topological Regression is a valid and accurate alternative for multivariate calibration in industrial process monitoring from spectral measurements. In the presence of model maintenance issues, the DTR method is capable to define a single regression model that can be used to estimate any set of properties. The model is easy to construct because non-parametric and it is also inherently able to handle nonlinearities, thus making the estimation accurate and computationally very efficient.

The major limitation of the Delaunay calibration method is, however, its near-absolute lack of extrapolation ability on samples that fall outside the calibration domain. Such samples are expected to occur rather often, depending on the number of available observations and the dimensionality of the problem. Therefore, this work devotes special attention to this problem and proposes a rigorous approach to estimate a continuous regression function also for the outside objects. The discussed approach projects borderline samples onto the calibration domain and uses known observations for defining the estimates.

When applied to the calibration of the aromatic content in Light Cycle Oils, the proposed DTR method with projection demonstrated capable to always outperform other DTR-based approaches available in the literature and often comparable in accuracy with standard PLSR models. When compared to PLSR, the main advantages of the DTR method are in the simplicity of the calibration and ease to upgrade but, also the fewer components, thus leading to more robust and manageable models.

REFERENCES

Corona, F., Liitiäinen, E., Lendasse, A., Baratti, R., and Sassu, L. (2009). Delaunay tessellation and topological regression: An application to estimating product properties from spectroscopic measurements. In *Proceedings of the International Symposium on Process Systems Engineering*, 1179–1184. Salvador Bahia, Brazil.

Espinosa, A., Sanchez, M., Osta, S., Boniface, C., Gil, J., Martens, A., Descales, B., Lambert, D., and Valleur, M. (1994). On-line NIR analysis and advanced control improve gasoline blending. *Oil & Gas Journal*, 92, 49–56.

Gemperline, P. (2006). *A Practical Guide to Chemometrics: Second Edition*. CRC Press - Taylor & Francis, Boca Raton.

Gudmundsson, J., Hammar, M., and van Kreveld, M. (2002). High order delaunay triangulations. *Computational Geometry*, 23, 85–98.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning: Second Edition*. Springer, New York.

Jin, L., Fernández Pierna, J.A., Wahl, F., Dardenne, P., and Massart, D.L. (2003a). The law of mixtures method for multivariate calibration. *Analytica Chimica Acta*, 476, 73–84.

Jin, L., Fernández Pierna, J.A., Xu, Q., Wahl, F., de Noord, O.E., Saby, C.A., and Massart, D.L. (2003b). Delaunay calibration method for multivariate calibration. *Analytica Chimica Acta*, 488, 1–14.

Jin, L., Xu, Q.S., Smeyers-Verbeke, J., and Massart, D.L. (2005). Updating multivariate calibrations with the delaunay triangulation method. *Applied Spectroscopy*, 59, 1125–1135.

Jin, L., Xu, Q.S., Smeyers-Verbeke, J., and Massart, D.L. (2006). Updating multivariate calibration with the delaunay triangulation method: The creation of a new local model. *Chemometrics and Intelligent Laboratory Systems*, 80, 87–98.

Jolliffe, I.T. (2002). *Principal Component Analysis: Second Edition*. Springer, New York.

Lee, J.A. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer, New York.

Stone, C.J. (1977). Consistent non-parametric regression. *The Annals of Statistics*, 80, 595–645.

Workman, J.J. (1999). Review of process and non-invasive near-infrared and infrared spectroscopy: 1993–1999. *Applied Spectroscopy Reviews*, 34, 1–89.