Special Issue

JOURNAL OF CHEMOMETRICS

Received: 8 October 2007, Revised: 20 June 2008, Accepted: 19 July 2008, Published online in Wiley Interscience: 29 September 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1188

# Wavelength selection using the measure of topological relevance on the self-organizing map

**Francesco Corona[a]*, Satu-Pia Reinikainen[b], Kari Aaljoki[c], Annikki Perkiö[d], Elia Liitiäinen[a], Roberto Baratti[e], Olli Simula[a] and Amaury Lendasse[a]**

**In this work, we investigated the possibility to perform wavelength selection by exploiting the metric structure of the spectrophotoscopic measurements. The topologically preserving representation of the data is performed using the self-organizing map (SOM) where the inputs' significance to the output is computed with the measure of topological relevance (MTR) on SOM. The MTR on SOM is a metric measuring the similarity between local distance matrices and we found that spectral inputs with a topology, which is, close to the output's are also associated to the wavelengths that chemically explain the influence of the spectra to the property of interest. As a result, we suggest a wavelength selection strategy based on the MTR on SOM, that is, interpretable to the domain experts and independent on the regression technique subsequently used for estimation. To support the presentation, a full-scale application from the oil refining industry is illustrated on the problem of estimating standard properties in a complex hydrocarbon product starting from spectrophotoscopic measurements. The method is further validated on the problem of octane number estimation in finished gasolines, under small sample conditions. The application led to accurate, parsimonious and understandable models. Copyright © 2008 John Wiley & Sons, Ltd.**

**Keywords:** variable selection; self-organizing map

## 1. INTRODUCTION

Spectrophotograms are recognized sources of information in a broad variety of fields ranging from analytical chemistry to process industry. Many applications reported in the research and industrial literature regard the estimation of important quality indexes in a material (typically, chemical and physical properties) starting from a collection of light absorbance spectra (e.g. see Reference [1]).

The information encoded in the spectra result from the interaction between light and matter and it is displayed as complex curves conditioned by the composition of the analyzed samples. In turns, the composition determines the properties of interest. However, without specific methods of analysis, such information is not easily accessible and, cannot be directly extracted and used for estimation purposes. In fact, one intrinsic characteristic of the measurements acquired by a high-resolution spectrophotometer is that the absorbance spectrum can be regarded as a regular function observed at discretized arguments in the instrument's operating range of wavelengths. Because of such a distinctive feature, the problem of estimating the output (the property of interest) is defined from very high-dimensional and collinear inputs (the observed spectra). Furthermore, it is not unusual to analyze datasets with a number of observations, which is, radically smaller than the number of input candidates.

To address the estimation problem, two regression approaches are commonly used in practice. One standard solution is to rely on full-spectrum methods for linear dimension reduction coupled with linear regression: the basic formulations of principal components regression (PCR) and partial least-squares regression (PLSR) are reference models. The natural refinement of such an approach benefits from a preliminary selection of relevant wavelength ranges [2] as performed by one of the many available techniques (e.g. see References [3–11]). Unfortunately, being based on derived variables, PCR and PLSR models are still not necessarily trivial to interpret and, the understandability of the models can be further reduced when nonlinear and *kernelized* extensions are considered. The alternative solution consists of selecting, among all the original candidates, individual inputs that truly contribute to a correct estimation of the output. Hence, wavelength selection is the limit extension of range selection where the origi-

* *Correspondence to: F. Corona, Department of Information and Computer Science, Helsinki University of Technology, Konemiehentie 2 (Room B313), Espoo, P.O. Box 5400, FI-02015 HUT, Finland.*
*E-mail: francesco.corona@hut.fi*

a *F. Corona, E. Liitiäinen, O. Simula, A. Lendasse*
*Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Finland*

b *S.-P. Reinikainen*
*Department of Chemical Technology, Lappeenranta University of Technology, P.O. Box 20, FI-53851 Lappeenranta, Finland*

c *K. Aaljoki*
*Neste Engineering. P.O. Box 310, FI-06101 Porvoo, Finland*

d *A. Perkiö*
*Neste Oil. P.O. Box 310, FI-06101 Porvoo, Finland*

e *R. Baratti*
*Department of Chemical Engineering and Materials, Università di Cagliari, Piazza d'Armi, I-09123 Cagliari, Italy*

nal interpretability of the inputs is explicitely retained. Typically, the selection is approached either from first-principle considerations known *a priori* or with data-derived methods based on model performances, stepwise strategies, dependence indexes and regularization (e.g. see References [12,13] and the references therein).

In this study, wavelength selection is approached by exploiting the metric structure of the data, leading to a method that identifies only the spectral inputs with a topology most similar to the output's. The topology preserving modeling of the data is carried out with the self-organizing map (SOM, [14]). The SOM is an adaptive algorithm to formulate the vector-quantization paradigm and perform mappings of high-dimensional data onto an ordered low-dimensional subspace. The SOM is widely employed in many fields including chemometrics (for reference, see the SOM References [15,16]). The SOM is mainly used to get a visual insight of the data and their structure, as well as for the investigation of potential relationships between variables [17,18]. Here, the SOM is used as a framework to investigate the topological similarities between the spectral inputs and the output according to a metric, the measure of topological relevance (MTR) on the SOM ([19]), which is derived from the assumed continuity of the unknown functionality existing between them.

The MTR on SOM measures such similarities from distances between the map nodes (the U-matrices, [20]), and we found that the inputs with a topology, which is, maximally similar to the output's are usually associated to the wavelengths that chemically explain the influence of the spectral inputs to the property of interest. This, suggests a simple strategy for wavelength selection leading to only few inputs still interpretable to the domain experts. Moreover, being the selection performed before building the estimation model, the approach is also model independent; in the sense that, once the inputs are selected, any regression technique can be used to reconstruct their relationship with the output. With simplicity in mind, the estimation techniques preferred in our experiments are classical linear models like ordinary least squares (OLS) and Ridge regression (RR). For completeness, also a *de facto* standard in nonlinear function estimation was considered; the least squares formulation of the support vector machine for regression (LS-SVM, [21]). With this respect, when the observations present a considerable curvature, the wavelengths selected by the MTR on SOM are expected to perform better using nonlinear regression techniques, because of the ability of the SOM to model nonlinearities in the data.

The study is organized as follows. In Section 2, we overview the rationale and algorithmic part of the investigation; Subsection 2.1 briefly illustrates the SOM paradigm and the MTR on SOM and Subsection 2.2 describes its application in the wavelength selection strategy. In Section 3, the direct application of the method is discussed on a set of full-scale problems from the oil refining industry and further validated on a small sample benchmark problem, from the same domain.

## 2. ALGORITHMS

### 2.1. The self-organizing map (SOM)

In its basic formulation, the SOM consists of a bi-dimensional regular array of nodes where a prototype vector $\mathbf{m}_k \in \mathbb{R}^p$ is associated with every node $k = 1, \ldots, K$. Each prototype acts as an adaptive model vector for the observations $\mathbf{v}_i \in \mathbb{R}^p$, with $i = 1, \ldots, N$. The nodes are arranged in a grid, that is, usually either hexagonal or rectangular.

During the computation of the map, the observations are projected onto the SOM's array and the model vectors adapted according to the learning rule:

$$\mathbf{m}_k(t+1) = \mathbf{m}_k(t) + \alpha(t) h_{k,c(\mathbf{v}_i)} \left( \mathbf{m}_k(t) - \mathbf{v}_i(t) \right)$$

where $t$ is the discrete-time coordinate of the mapping steps, and $\alpha(t) \in (0, 1)$ the monotonically decreasing learning rate. The scalar multiplier $h_{k,c(\mathbf{v}_i)}$ denotes a neighborhood kernel function centered at the best matching unit (BMU).

The BMU denotes the model vector $\mathbf{m}_c$ that best matches with the observation vector $\mathbf{v}_i$. The matching is determined according to a competitive criterion conventionally based on the Euclidean metric $\| \cdot \|$ and, at each step $t$, the BMU $\mathbf{m}_c(t)$ is hence the model vector $\mathbf{m}_k(t)$, that is, closest to the observation $\mathbf{v}_i(t)$; which is

$$\|\mathbf{m}_c(t) - \mathbf{v}_i(t)\| \leq \|\mathbf{m}_k(t) - \mathbf{v}_i(t)\|, \quad \forall k$$

The neighborhood kernel function $h_{k,c(\mathbf{v}_i)}$ is usually chosen in the Gaussian form:

$$h_{k,c(\mathbf{v}_i)} = \exp\left( - \frac{\|\mathbf{r}_k - \mathbf{r}_c\|^2}{2\sigma^2(t)} \right)$$

where the vectors $\mathbf{r}_k$ and $\mathbf{r}_c$ (in $\mathbb{R}^2$, for the 2D map) represent the geometric location of the nodes on the array, and $\sigma(t)$ denotes the monotonically decreasing width of the kernel. Over the SOM's array, the effect of the Gaussian kernel decreases smoothly with the distance from the BMU, thus allowing for a regular smoothing of the model vectors. The map is computed recursively for each observation and, as the term $\alpha(t) h_{k,c(\mathbf{v}_i)}$ tends to zero with $t$, the set of model vectors $\{\mathbf{m}_k\}_{k=1}^K$ is updated to represent, or prototype, similar observations in $\{\mathbf{v}_i\}_{i=1}^N$. By the end of the mapping, the models will converge toward their asymptotic limits [22,23].

As a result, the prototypes will form an ordered manifold in the original data space where the relevant topological and metric properties of the observations are preserved. In Figure 1, the training of a 2D SOM is depicted for a 2D set of synthetic data consisting of four clusters. In general, since the dimensionality of the data is higher than the SOM array's, the resulting map is to be understood as the organized image of the original data where the high-dimensional structures existing within the observations are represented on the low-dimensional grid with simple geometric relationships between the prototypes.

Because of such properties, the SOM is usually employed to getting a visual insight of the data structure by using the many available visualization techniques; for instance, the component planes [24] and the unified-distance matrices, or U-matrices. Figure 2 depicts such displays for the SOM trained onto the synthetic data. In the component planes representation, different gray shades indicate the distribution on the map of the values of the prototypes (along the original data directions), whereas the U-matrices display clusters (dark areas) and cluster separations (bright areas) by visualizing distances between neighboring prototypes.

#### 2.1.1. The measure of topological relevance (MTR on SOM)

The SOM is also used to starting a preliminary investigation of potential relationships between the component variables. From the map, dependencies can be searched either qualitatively (by looking for similar patterns in identical positions in the component

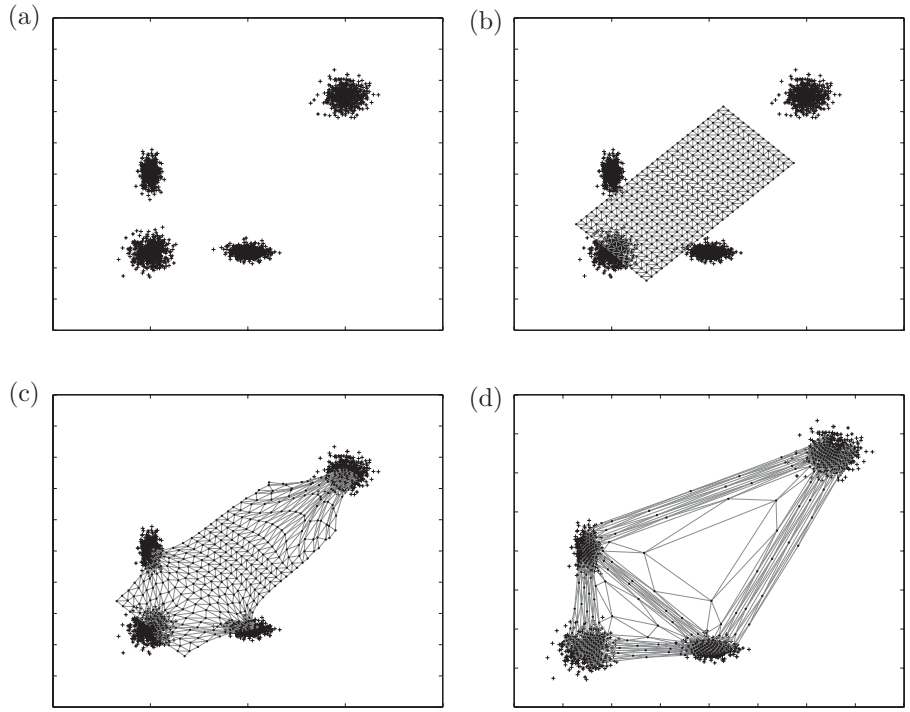


**Figure 1.** The training of the SOM: (a) the data, (b) the map initialized along the two largest principal components of the data, (c) the map after a few training steps and (d) the final result.

planes and distance-based representations) or calculated explicitely from their correlation [25].

Such metrics mostly exploit the quantization properties of the SOM. On the other hand, the MTR on the SOM calculates the output's relevance of a set of inputs from the topology preserving properties of the map. The metric is derived from the assumed continuity of the unknown functionality $y = f(\mathbf{x}) + r$ existing between the inputs and the output. As usual, the relationship is to be estimated from $N$ observations in the form $\{(\mathbf{x}_i; y_i)\}_{i=1}^{N}$ where, $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ are the inputs and output for the $i$th observation, respectively, and $r$ denotes the additive noise. Under the continuity hypothesis, if two points $\mathbf{x}_i$ and $\mathbf{x}_i'$ are close together in the input space, it is expectable that also $f(\mathbf{x}_i)$ and $f(\mathbf{x}_i')$ are close together in the output space. If the neighborhood continuity is not satisfied (i.e. $y_i$ and $y_i'$ are not close together), this can be either due to an high level of noise or because the inputs are actually not relevant for the output. For a SOM trained to represent input–output observations $\mathbf{v}_i = [\mathbf{x}_i; y_i]$, the MTR on SOM exploits this general principle directly from the model vectors $\mathbf{m}_k$ of the map and the U-matrices; that is, a relevant input is expected to have, on a map, a topology that is similar to the output's.

The global neighborhood topology of the data is extracted from a map as a matrix consisting of distances between each connected node (the full U-matrix **U** based on all the component variables, as in Figure 2(c)). For clarity, we recall that: letting $\mathbf{m}_k$ the prototype vector associated to node $k$ and $\mathcal{N}(k)$ its neighborhood of $L$ adjacent nodes $l$ (e.g. $L = 6$ almost everywhere, on the hexagonal grid), the entries of the unified-distance matrix **U** are essentially calculated as:

- local pairwise distances, for all $l$:

$$d(k, l) = \|\mathbf{m}_k - \mathbf{m}_l\|$$

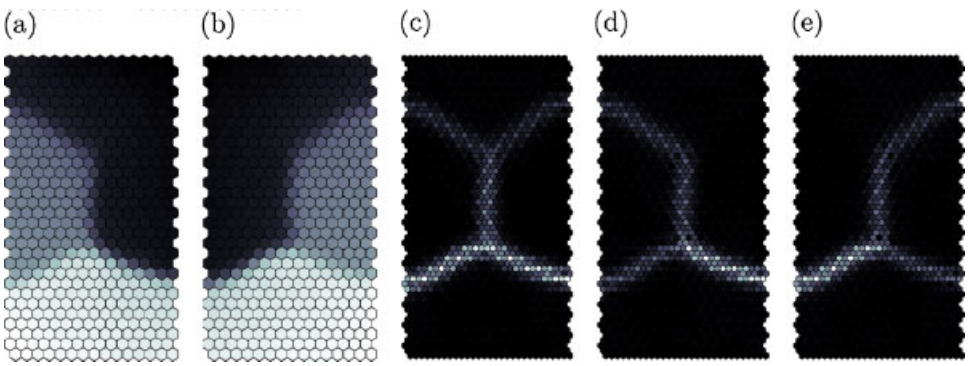


**Figure 2.** The visualization of the SOM: (a,b) the two component planes (one for each direction in the original data space), (c) the full U-matrix revealing the clustering structure of the data and (d,e) the two component U-matrices (again, one along each direction in the original data space). This figure is available in color online at www.interscience.wiley.com/journal/cem



  

- locally averaged distances in $\mathcal{N}(k)$:

$$d(k) = \frac{1}{L} \sum_{l \in \mathcal{N}(k)} \|\mathbf{m}_k - \mathbf{m}_l\|$$

Being based on a fixed nearest neighbor graph, such a matrix is hence structured and organized according to the local topology of the observations.

Analogous considerations apply when the U-matrix is calculated independently along each direction of the data space (the component U-matrices $\mathbf{U}_{x_j}$ for the input variables, with $j = 1, \ldots, d$, and $\mathbf{U}_y$ for the output, as in Figure 2(d) and (e)). By calculating the distance between each input–output pair of component U-matrices, the MTR on SOM is used to quantify the similarity between topologies and, thus, the significance of an input in reconstructing the neighborhood continuity with the output. The metric assessing such a similarity is formally expressed as:

$$\mathcal{T}(x_j, y) = \|\mathbf{U}_{x_j} - \mathbf{U}_y\|_F$$

where the Frobenius metric $\|\cdot\|_F$ is used to measure the Euclidean closeness between matrices; the closer to 0 is the measure, the more relevant is the input for the output. In order to clearly represent relevances the way they are commonly perceived, the measure $\mathcal{T}(\cdot,\cdot) \geq 0$ is preferably inverted and rescaled so that, larger values indicate stronger relevances (e.g. $\mathcal{T}(\cdot,\cdot) \geq 0 \rightarrow \mathcal{T}(\cdot,\cdot) \in [0, 1]$).

The MTR on SOM can be calculated using the SOM Toolbox. The Matlab version of the package is available from http://www.cis.hut.fi/projects/somtoolbox/.

### 2.2. The strategy for wavelength selection

In principle, given the values of the MTR on SOM for each input–output pair, variable selection could be simply performed by: (1) ranking the inputs according to their relevance to the output and (2) selecting a reduced but still representative subset $\check{\mathbf{x}} \in \mathbb{R}^s$ with $s \ll d$. However, this basic selection procedure when directly applied to spectrophotoscopic inputs is intrinsically limited by the continuous nature of the wavelengths' domain, since absorbances measured at neighboring wavelengths are characterized by a relevance to the output that is very similar. Therefore, the selection of an input $x_j$, that is, found to be relevant to predicting $y$ would be naturally accompanied by the selection of a broad range of contiguous inputs also characterized by high relevance, but redundant and collinear because embedding a near-identical informative content.

In such a context, it is possible to observe the existence of a regular function defined over the wavelengths' domain that describes the smooth changes in relevance as neighboring inputs are considered. The MTR on SOM also shows such a behavior, therefore, we propose a strategy for wavelength selection that retains only those inputs that manifest a topology, that is, maximally similar to the output's.

The procedure to perform wavelength selection summarizes as follows:

1. calculate the set $\mathcal{T} = \{\mathcal{T}(x_j, y)\}_{j=1}^{d}$ of pairwise relevances between each input–output pair;
2. select the subset of inputs, $\check{\mathbf{x}}$, with a topology that best matches the output's, that is

$$\check{\mathbf{x}} = \left\{ \check{x}_{j^*} \subset \mathbf{x} : j^* = \underset{j}{\operatorname{argmax}}\, \mathcal{T}(x_j, y) \right\}$$

The procedure identifies and selects only inputs whose wavelengths correspond to the local maxima of $\mathcal{T}$ (i.e. relevant to predicting the output). The selection is optimal with respect to the problem of predicting the output; in fact, among similar inputs, only the maximally relevant ones are retained and the neighboring redundancies are discarded. Being relevance to the output the only supervising criterion for selection, the procedure is still suboptimal with respect to the problem of selecting inputs that are also minimally redundant. Nevertheless, the selected variables $\check{\mathbf{x}}$ are implicitly also as much as possible dissimilar, because each $\check{x}_{j^*}$ prototypes the different wavelengths' ranges separated by the local minima of $\mathcal{T}$.

For the sake of clarity, the procedure to perform wavelength selection using the MTR of SOM is summarized on an illustrative example (Figure 3). The dimensionality of the spectra is 100 (i.e. $\mathbf{x} \in \mathbb{R}^d$, with $d = 100$) and a scalar property is to be estimated (i.e. $y \in \mathbb{R}$), 30 observations are available. In the first stage, both the spectra and property observations are mapped onto the SOM, here the topological relevance between each input and the output is measured. The metrics form a relevance function defined over the wavelength domain (i.e. $\mathcal{T}(x_j, y)$). In the second stage, the local maxima of $\mathcal{T}$ are identified and only the associated spectral variables are selected (i.e. $\check{\mathbf{x}} \in \mathbb{R}^s$, with $s = 4$). From the set of selected variables, $\check{\mathbf{x}}$, any model that estimates the functionality, $f$, can be learned and used to predict the output, $y$, as depicted in Figure 4.

The regression techniques preferred in our experiments are classical linear models like the OLS and the RR. Nevertheless, also nonlinear function estimators were used; specifically, we considered the LS-SVM ([21]. When needed and to avoid overfitting, the meta-parameters of the models (i.e. the penalty term in the Ridge model and the kernel width and the regularization term in the LS-SVM) were optimized by standard resampling methods to estimating the prediction accuracy. In our study cases, schemes like the leave one out (LOO-CV) and 10-times 5-fold cross-validation (CV) were adopted [26]. The same approach on the topological error of the SOM [27] is also used to optimize the size of the map and prevent it from overfitting the observations.

## 3. EXPERIMENTS

The American Society for Testing Materials standards (ASTM, [28]) for measuring the properties of an hydrocarbon mixture are often based on time-consuming procedures, expensive and maintenance-intensive laboratory equipment that require skilled labor. Nevertheless, real-time estimates of such properties are of fundamental significance for both the production and blending process of finished and semi-finished products. In particular, the availability of accurate and parsimonious models that estimate such properties in real-time from inexpensive and non-intrusive spectral measurements is beneficial when the predictions are used for on-line monitoring and control of the operating units, expecially if no sensible computational load is introduced to the plant's distributed control system.

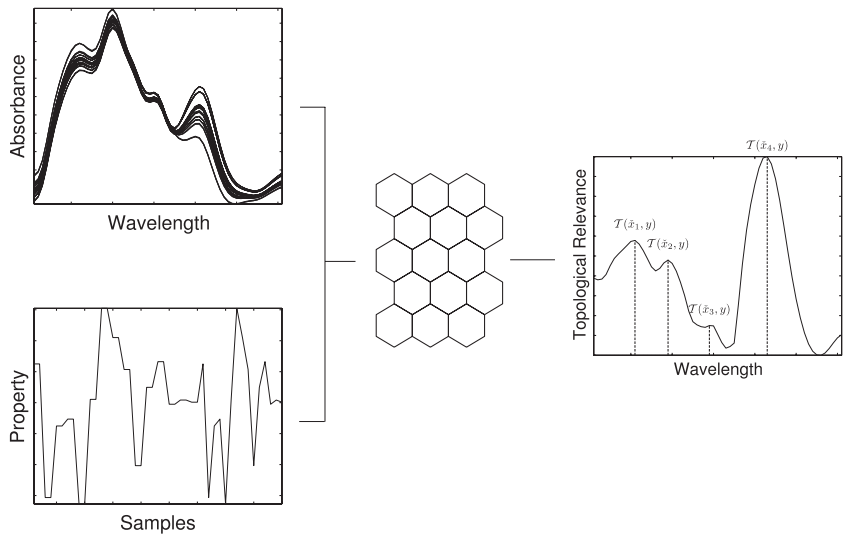


**Figure 3.** Wavelength selection using the MTR on SOM: a summary.

This section presents application of the investigated method on a set of monitoring tasks from the oil refining industry. The first study case is a full-scale application where the problem consists of estimating a set of different chemical and physical indexes in a hydrocarbon product processed by a separation unit. The second study case is a standard benchmark problem referring to the prediction of the octane number in finished gasolines from a set of only few measurements.

### 3.1. Study case I

The first application consists of estimating six different properties in a hydrocarbon mixture starting from the same set of spectral observations. The data are collected over an extended period of time (from February to August 2006, once per working shift) that spans most of the important variations in the processing unit; that is, major winter and summer assets, as well as minor changes in the production conducted according to operational necessities. Due to the commercial significance of the application and the secrecy agreement with our industrial partners, it is not possible to provide a more exhaustive description of the problem. Anyway, we believe that the reported information, along with the following results, will give the possibility to appreciate the method here proposed.

The absorbance spectra are acquired by means of a continuous-flow spectrophotometer operating in the 4408–9992 $cm^{-1}$ range (≈2270–1000 nm). The absorbance is measured on the basis of the NIR principle with a 4 $cm^{-1}$ resolution, in Figure 5(a). Each observation consists of the 1397-channel spectrum of absorbances

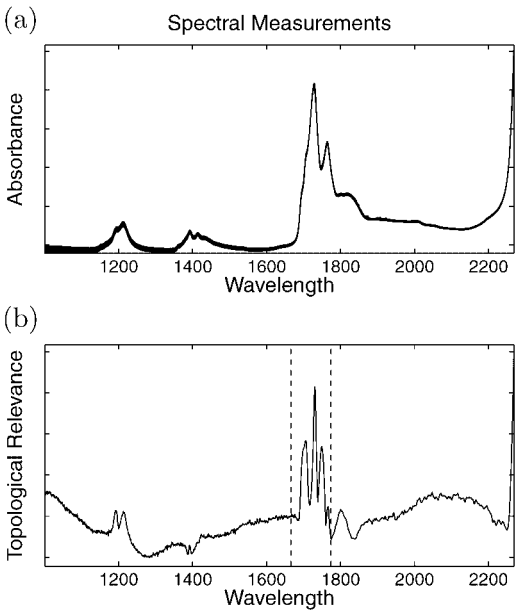


**Figure 5.** (a) The full spectral range and (b) the MTR on SOM between the inputs and the first output property. In (b), the vertical dashed lines delimit the spectral range initially selected, roughly corresponding to the first overtone.

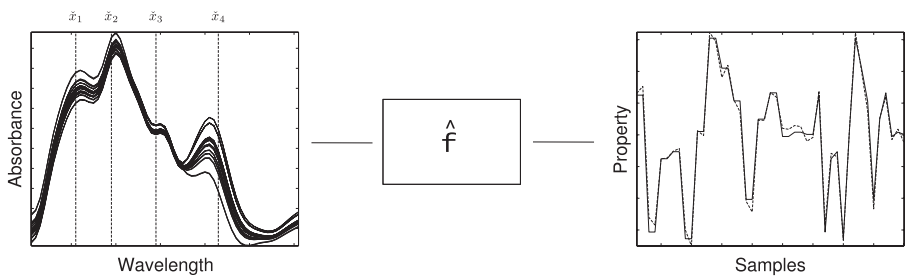


**Figure 4.** Estimating the property from the reduced set of spectral variables selected using the MTR on SOM: any estimator $\hat{f}$ can be used.



  

**Table I.** Data summary

| | Calibration samples (*n*) | Testing samples (*n*) | Missing samples (%) |
|---|---|---|---|
| Property 1 | 419 | 139 | ≈36 |
| Property 2 | 431 | 143 | ≈35 |
| Property 3 | 431 | 143 | ≈35 |
| Property 4 | 354 | 118 | ≈47 |
| Property 5 | 165 | 55 | ≈75 |
| Property 6 | 413 | 138 | ≈37 |

and the corresponding values of six properties. The measurements of the product's properties are obtained in laboratory by ASTM methods and the accuracy of the analysis are used to assess the quality of the estimates.

Because a different number of missing values is recorded for different properties, the avaliable dataset consists of a different number of observations to be used for learning and validating each estimation model (calibration). By the same token, a different number of observations is also available for testing the final models. Approximatively, 900 observations with a percentage of missing values ranging from 35 to 75% are available. In general, the first 2/3rds of the data are used for learning/validation and the last 1/3rd is used for testing the resulting models in extrapolation. Table I summarizes the modeling setup.

### 3.1.1. *Wavelength selection and chemical interpretability*

The analyzed spectra show the typical overlapped absorbance bands arising from different hydrocarbon functional groups and reflect the samples' composition. The major absorbance features in the experimental region are usually assigned to the second overtone (≈1100–1300 nm), the combination bands (≈1300–1550 nm) and the first overtone (≈1600–1800 nm) of the Carbon–Hydrogen vibrations [29,30]. The vibrations of the C–H bond on different functional groups lead to distinct absorption peaks, therefore, the chemical and physical properties of the hydrocarbon mixture can be reconstructed from spectra since phenomenological relationships between the chemical structure and the properties exist. In addition, being the relationships between the properties (6) and the spectra distributed among a large number of different inputs (1397), the application is interesting because wavelength selection cannot be easily performed only by an *a priori* interpretation of the spectra.

Starting from such a recognition, our first concern was to perform a preliminary selection of the most relevant spectral range. According to the method presented in Section 2, a first set of six 2D SOM on the full-spectrum inputs and each of the output properties was trained using the observations in the calibration set. Each map consisted of a hexagonal array of nodes firstly initialized along the subspace spanned by the eigenvectors corresponding to the two largest eigenvalues of the covariance matrix of the data. As usual, the ratio between these two eigenvalues was also used to cross-validate the size of the SOMs. Subsequently, the MTR on the SOM was calculated between each input-output pair, independently for each output. In Figure 5(b) the resulting relevance function based on the MTR on SOM is depicted for the first output. As for the other product properties, near-identical results were obtained; for the sake of compactness, the corresponding plots are not reported/discussed.

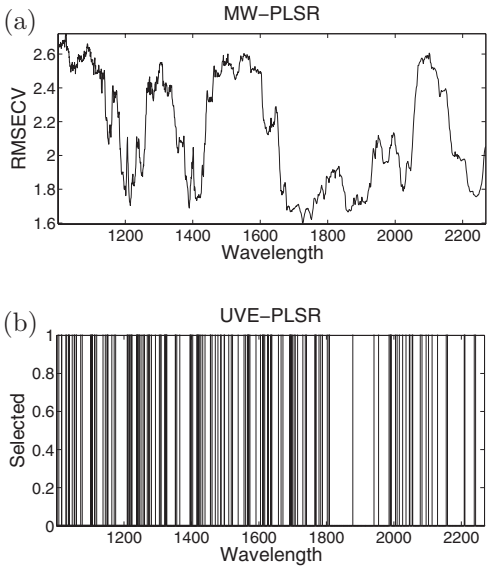


**Figure 6.** Comparison on the full range: (a) the accuracies obtained by MW-PLSR in cross-validation and (b) the set of variables selected by UVE-PLSR.

From Figure 5, it is possible to notice that the spectral region corresponding to the first overtone is characterized by values of topological relevance that are overall higher than those obtained in the other spectral ranges. Higher relevances were found for the combination bands arising from ≈2100 nm), but, because of the incompleteness of the peak, such a region was not considered. Based on such a result, we performed a preliminary selection of this range thus achieving a reduction of the dimensionality while retaining the most significant information. Moreover, because the behavior is shared between all the outputs, this also suggested that the problem could be globally modeled starting from the same spectral range. It is worthwhile noticing that the selection did not depend on the accuracy of any regression model and, only afterwards, its validity was experimentally confirmed by the models' accuracy.

The preliminary range selection was compared to several widely accepted methods for variable selection. In Figure 6, we reported the results obtained for the first output property by uniformative variable elimination (UVE-PLSR, [5]) and moving window (MW-PLSR, [10]), a variant of Interval PLS (iPLS, [9]). As expected, the qualitative results obtained by MW-PLSR confirmed the relevance of the selected interval (the first overtone), as associated with the best accuracies over the full range. As for UVE-PLSR, its application often led to an improvement in the overall performance of PLSR but its also associated to the selection of a still high number of spectral variables which is only partially simplifying the interpretability of the problem; the dimensionality reduction and the accuracy obtained by UVE-PLSR is reported in Table III.

By selecting the 1666–1744 nm range, in Figure 7(a), only 92 wavelengths were retained and used to perform wavelength selection. The range was selected after noticing that the peak of relevance found at approximatively 1800 nm could mostly be assigned to the presence of olefin in the mixture which should, however, be almost absent in the analyzed product (indeed, it was verified afterward that including such a input did not improve the accuracy of the models). For the purpose of selection, a second set of SOMs was trained, as above, and the sets of MTR on SOM

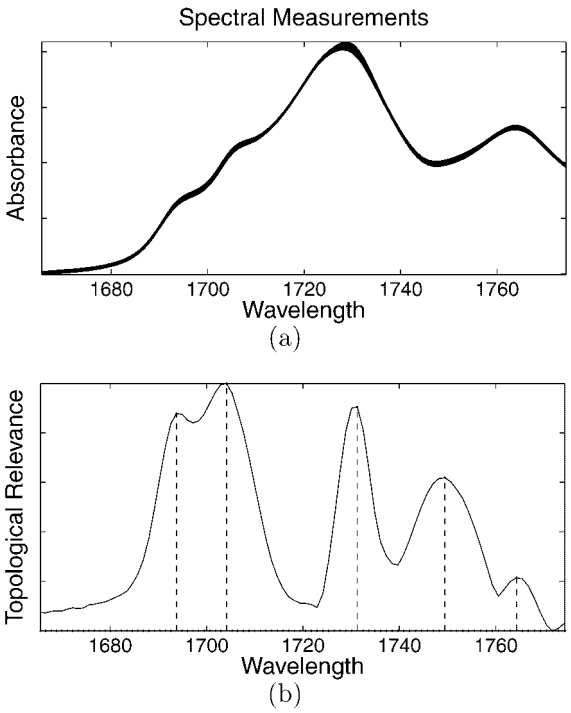


**Figure 7.** (a) The selected spectral range and (b) the MTR on SOM between the inputs and the first output property. In (b), the vertical dashed lines indicate the selected wavelengths.

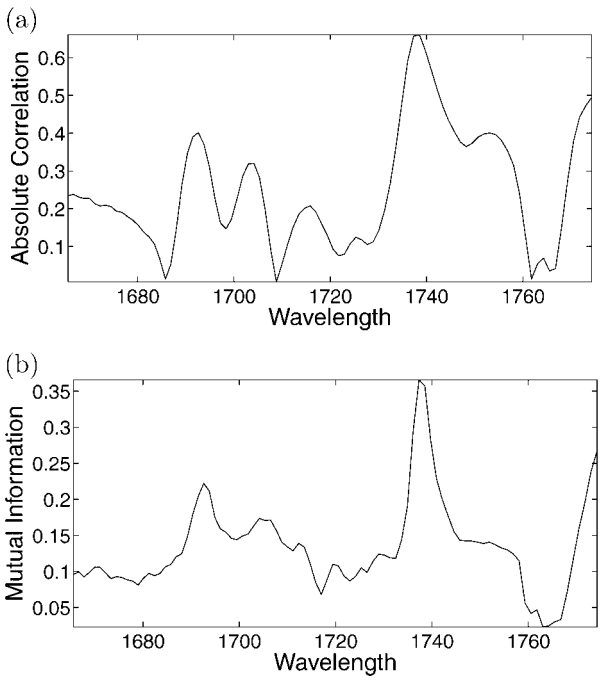


**Figure 8.** (a) The relevance function estimated from the absolute correlation coefficient and (b) the mutual information to the first output.

calculated for each output independently. The selected subset of maximally important inputs was then found in the correspondence of the local maxima of each relevance functions; only five wavelengths were identified as relevant to the outputs (Table II). In Figure 7(b), the resulting relevance function is depicted only for the first output, being the others again very similar.

The set of selected inputs is analogous for each output and, more importantly, it is in agreement with the chemical model explaining the influence for the chemical groups on the product properties (again, see References [29,30]). In detail:

- wavelengths $\check{x}_1$ and $\check{x}_2$ are related to the stretching vibrations of the methyl ($CH_3$) groups that manifest themselves as a doublet ($\approx$1695 and $\approx$1705 nm). Their presence indicates the presence of (possibly, large amounts of) branched hydrocarbons, although the absorbances can be also influenced by linear paraffins;
- wavelengths $\check{x}_3$ and $\check{x}_5$ are assigned to the stretching vibrations of the methylene ($CH_2$) groups also manifesting as a doublet ($\approx$1725 and $\approx$1765 nm). Their presence typically indicates the presence of linear hydrocarbons.

As for $\check{x}_4$, the selection of this wavelength could not be readily assigned to any neat spectral feature known to the authors and, therefore, remains unassigned.

In Figure 8, the results of the relevance function estimated with more conventional indexes of dependence like the absolute Pearson's correlation coefficient (CC) and the mutual information (MI) are presented for comparison. Based on the depicted results, both the measures present a behavior that resembles what obtained with the MTR on SOM. Only the CC is capable to represent the smooth nature of the observations but it is, however, restricted to detecting dependencies that manifest themselves in the covariance. On the other hand, the more general MI presents a very scattered behavior (due to difficulty of estimating such measure from real-world measurements) that prevents the application of an automatic wavelength selection strategy.

**Table II.** The subset of selected wavelengths

| | $\check{x}_1$ | $\check{x}_2$ | $\check{x}_3$ | $\check{x}_4$ | $\check{x}_5$ |
|---|---|---|---|---|---|
| [nm] | 1694 | 1704 | 1731 | 1749 | 1764 |

For comparison, Figure 9 depicts what is obtained by using established variable selection methods (MW-PLSR, UVE-PLSR and genetic algorithms (GA-PLSR, [8]) now a suitable technique because of the reduced number of spectral variables). Although characterized by an almost flat relevance over the spectral range, again the MW-PLSR qualitatively confirmed the significance of the variables selected by the MTR on SOM. As for UVE-PLSR, the high number of selected variables prevents from a direct assignment of the absorbtion bands (nevertheless, this result is mostly true only for the first output). On the other hand, GA-PLSR performed a parsimonious selection of wavelengths that, though, is again not necessarily interpretable. Anyway, both methods are often capable to improve the performance of PLSR. A detailed comparison of the results is presented in Table III.

### *3.1.2. Regression models and estimation results*

From the set of five selected wavelengths, both linear (OLS and RR) and nonlinear (LS-SVM) models were calibrated to reconstruct the functionality to the set of six output properties (independently). The results are summarized in Table III where the accuracy and the complexity of the models are also compared with standard PLSR, UVE-PLSR and GA-PLSR, when suitable. The prediction accuracy of the models is reported in extrapolation (i.e. only for the independent set of testing data), in terms of the root mean square



  

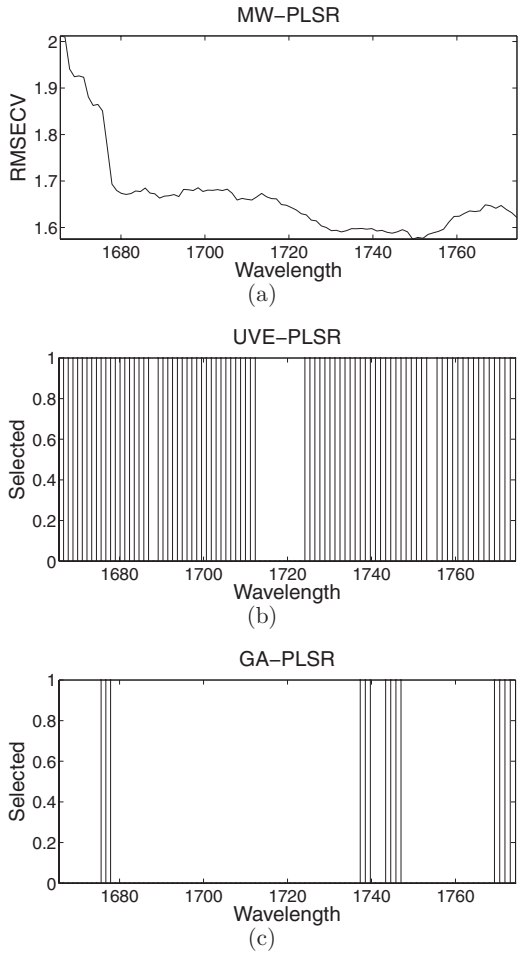


**Figure 9.** Comparison on the selected range: (a) the accuracies obtained by MW-PLSR in cross-validation, (b) the set of variables selected by UVE-PLSR and (c) GA-PLSR.

error in prediction (RMSEP) and the determination coefficient $\mathcal{R}^2$. All the models were preliminary cross-validated using only the calibration/validation set of observations.

From Table III, it is possible to notice that the preliminary selection of the spectral range corresponding to the first overtone suggested by the MTR on SOM has proven to be effective. In fact, the prediction results are always benefiting when the basic PLSR models are applied on such a range and, moreover, the complexity (i.e. the number of latent variables actually cross-validated) is often reduced. This result is often further confirmed by the improved accuracy of the UVE- and GA-PLSR models.

In addition, the prediction accuracies are also improved after performing wavelength selection with the MTR on SOM and, then, building the OLS and RR models. The result accredited that the MTR on SOM was capable to select only a parsimonious set of inputs still carrying an important informative content. This is also demonstrated by almost negligible values of penalty term cross-validated for most of the RR models: this indicates a near-absolute absence of shrinkage for the OLS estimated regression coefficients, as well as that the models did not overfit the observations. The performances could be further ameliorated when the selected variables were used to calibrate the more powerful LS-SVM model. In particular, this has shown to be valid when the relationship between the spectra and the property of interest cannot be fully recovered by the linear methods (no matter whether PLSR, OLS or RR models are used). On the other hand, when a linear functionality is experienced using the LS-SVM did not lead to any sensible improvement, as expected.

The results obtained with the best models calibrated on the selected variables (linear for the first three properties and nonlinear for the last three) are depicted in Figure 10. In the diagrams, the estimated properties are plotted against the laboratory measurements for the entire testing period (over 2 months of continuous functioning) and compared to the accuracy of the corresponding ASTM test. The representation provides an important summary of the properties that we sought during the development of the models for such a critical application: (1) accuracy (always comparable with the prediction accuracy of reference methods); (2) simplicity (a reduced set of variables and the simplest methods can be used to produce valuable results); (3) interpretability (the

**Table III.** A comparison between the prediction results obtained by the regression models on the independent set of testing observations

| | Full range [1000–2270 nm] | | 2nd Overtone [1100–1300 nm] | Combination [1300–1550 nm] | 1st Overtone [1666–1744 nm] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PLSR | UVE | PLSR | PLSR | PLSR | UVE | GA | OLS | RR | LS-SVM |
| Property 1 | 1.33 (0.76) 1397/6 | 3.30 (0.60) 129/9 | 1.31 (0.77) 300/5 | 1.42 (0.72) 300/7 | 1.24 (0.80) 92/4 | 1.20 (0.81) 82/3 | 1.32 (0.77) 14/6 | 1.16 (0.83) | 1.16 (0.82) | 1.18 (0.83) |
| Property 2 | 2.15 (0.79) 1397/5 | 2.04 (0.80) 128/5 | 2.01 (0.80) 300/5 | 2.38 (0.73) 300/5 | 1.80 (0.85) 92/5 | 2.02 (0.82) 29/9 | 1,75 (0.86) 31/6 | 1.65 (0.90) | 1.59 (0.90) | 1.60 (0.90) |
| Property 3 | 0.25 (0.85) 1397/5 | 0.23 (0.88) 162/5 | 0.25 (0.85) 300/5 | 0.30 (0.78) 300/5 | 0.23 (0.88) 92/5 | 0.22 (0.82) 23/7 | 0.23 (0.88) 20/6 | 0.21 (0.92) | 0.20 (0.91) | 0.21 (0.90) |
| Property 4 | 285 (0.60) 1397/5 | 313 (0.47) 103/5 | 460 (0.65) 300/7 | 405 (0.62) 300/5 | 450 (0.66) 92/8 | 332 (0.35) 42/9 | 300 (0.53) 39/8 | 225 (0.76) | 220 (0.75) | 220 (0.82) |
| Property 5 | 1.61 (0.82) 1397/5 | 3.08 (0.43) 77/8 | 2.13 (0.65) 300/8 | 1.88 (0.74) 300/2 | 1.52 (0.84) 92/4 | 1.63 (0.82) 38/3 | 1.85 (0.76) 10/6 | 2.00 (0.80) | 1.83 (0.81) | 1.39 (0.88) |
| Property 6 | 3.51 (0.36) 1397/9 | 3.68 (0.30) 145/5 | 2.61 (0.67) 300/3 | 2.77 (0.62) 300/4 | 2.55 (0.69) 92/6 | 2.41 (0.73) 29/5 | 2.65 (0.66) 13/3 | 2.62 (0.67) | 2.62 (0.67) | 2.23 (0.83) |

The accuracy is evaluated with the root mean square error of prediction (RMSEP) and, in between brackets, the determination coefficient $\mathcal{R}^2$. Information about the model complexity is indicated by the number of variables and latent factors cross-validated for the PLSR models. As for OLS, RR and LS-SVM, the models are based only on the five variables selected by the MTR on SOM.

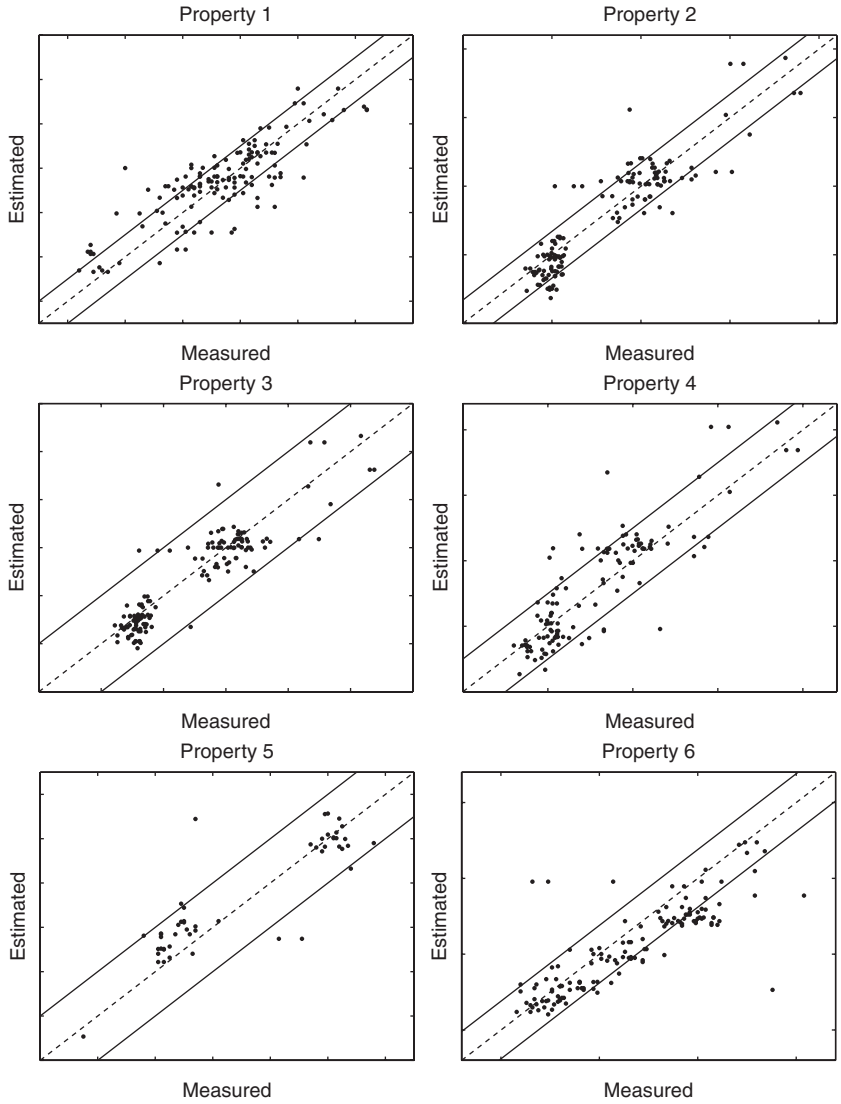


**Figure 10.** A comparison between measured and estimated product properties in simulating the real testing scenario (over 2 months of continuous operation). The bands represent the accuracy of the ASTM methods.

selected wavelengths are easily accessible and can be used to analyze the models and their performances) and (4) the robustness required to effective implementations in a full-scale production environment (different operating assets and corrupted measurements).

### 3.2. Study case II

The second application consists of estimating the octane number in gasolines. The application of the methodology is discussed on a dataset of spectral measurements and associated evaluations of the octane number provided by Camo A/S (Trondheim, Norway), which is gratefully acknowledged.

The absorbance spectra are acquired by means of a spectrophotometer operating in the 1100–1550 nm wavelengths' range with a 2 nm resolution, in Figure 11. The measurements of the octane number (in the 86–92 range) are evaluated in laboratory by reference ASTM motor tests. Therefore, each sample consists of the 226-channel input spectrum of absorbances and the corresponding output, the octane number. The application of the MTR on SOM on this benchmark is interesting because of the small sample size; only 24 observations for model calibration/validation and 9 observations for testing the final model.

According to the methodology, the input and output observations in the calibration set were mapped onto a 2D SOM. On the map, the set of topological relevances between each input–output pair was calculated and only a subset of six relevant inputs selected (Table IV).

The set of selected inputs is again in agreement with the chemical model explaining the influence for the chemical groups on the octane number [31]. The major absorbance features in the experimental region are usually assigned to the second overtone and to the combination bands of the C–H vibrations. In detail:

**Table IV.** The subset of selected wavelengths

| | $\check{x}_1$ | $\check{x}_2$ | $\check{x}_3$ | $\check{x}_4$ | $\check{x}_5$ | $\check{x}_6$ |
|---|---|---|---|---|---|---|
| [nm] | 1146 | 1215 | 1336 | 1394 | 1416 | 1518 |



  

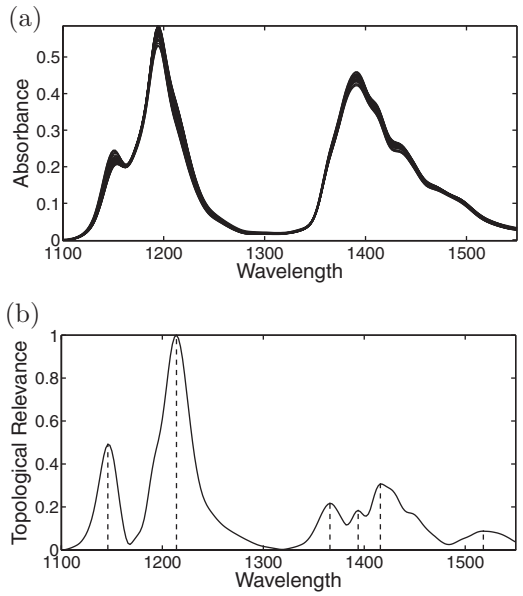


**Figure 11.** (a) The full spectral range and (b) the MTR on SOM between the spectra and the octane number. In (b), the vertical dashed lines are drawn in the correspondence of the selected variables.

- the aromatic bonds at 1150 nm ($\check{x}_1$) are related to an increase in octane number. Conversely, the methylene bonds at 1220 nm ($\check{x}_2$) indicate the presence of linear hydrocarbons which are responsible for a reduction in the gasoline quality. The methyl bonds at 1200 nm indicate a larger amount of branched hydrocarbons although the absorbance is also influenced by the amount of linear paraffin: in fact, its effect on octane is not readily explained and the contribution, usually, varies with the gasoline type. Actually, this occurs with the present spectra in which, even if the relevance $\mathcal{T}$ shows an inflection at 1200 nm, the absorbance does not correspond to a local maximum and, thus, the associated input is not selected;
- by the same token, the effect of the combination bands for methylene (1395/1416 nm) and methyl (1360/1345 nm) on octane mimics what observed in the short-wavelength range. With this respect, the methylene absorbance wavelengths are correctly identified ($\check{x}_4$ and $\check{x}_5$), while $\check{x}_3$ accounts for the first methyl band. As already noticed above, again the second methyl band is only partially recovered by an inflection in $\mathcal{T}$.

As for variable $\check{x}_6$, no spectral features are readily assignable. Its selection can be ascribed to baseline effects.

Starting from the six selected spectral variables, the regression models were calibrated to estimate the octane number and the prediction accuracy evaluated on the independent set of testing data (Table V and Figure 12). When needed, the meta-parameters of the models (the penalty term in RR and the kernel width and regularization term in LS-SVM) were optimized by LOO-CV. In this specific case, the use of leave one out cross-validation was adopted because of the reduced sample size. The same scheme was applied for cross-validating the optimal size of the SOM used for wavelength selection, as well as the number of latent factors for the PLSR model.

Based on the experimental results, it is possible to notice that again the MTR on SOM was capable to select an informative and interpretable subset of spectral inputs to be used in parsimonious and yet accurate regression models.

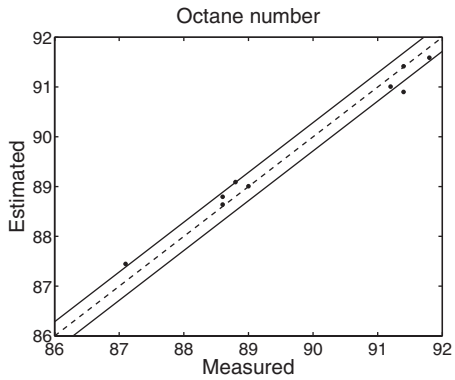


**Figure 12.** A comparison between measured and estimated octane number. The band represents the accuracy of the ASTM methods (0.25 RON).

**Table V.** A comparison between prediction results

| | Number of variables | RMSEP |
|---|---|---|
| PLSR | 4 (Latent) | 0.28 |
| OLS | 6 (Original) | 0.34 |
| RR | 6 (Original) | 0.31 |
| LS-SVM | 6 (Original) | 0.24 |

## 4. CONCLUSIONS

In this work, we investigated the possibility to exploit the metric structure of spectrophotoscopic data in order to perform wavelength selection. For the purpose, we analyzed the results obtained by using the MTR on the SOM.

The application of the method on a full-scale set of monitoring problems from process industry showed that the MTR on SOM was successfull, firstly, in the preliminary selection of the most relevant spectral range and, secondly, in identifying a significant subset of wavelengths. The selected inputs are in agreement with the chemical model explaining the composition of the analyzed samples and, therefore, understandable to the domain experts. Moreover, the selected variables are also characterized by an important informative content that can be exploited to develop parsimonious regression models. This led to develop prediction models that always achieved the accuracy required for an effective use in real-time monitoring. Given the independence from the regression method, several methods could be investigated and developed. The predictions obtained by building the simplest linear models were always comparable to what is achieved with the standard methods. On the other hand, when more accuracy was required, the possibility to use more powerful nonlinear methods directly on the selected variables allowed to improve the results in a sensible fashion.

### Acknowledgements

This investigation was supported by the Finnish Funding Agency for Technology and Innovation (TEKES) under the project CHESS (Developing Chemometrics with the tools of Information Science), which is gratefully acknowledged.

## REFERENCES


1. Workman JJ, Jr. Review of process and non-invasive near-infrared and infrared spectroscopy: 1993–1999. *Appl. Spectrosc. Rev.* 1999; **34**: 1–89.
2. Nadler B, Coifman RR. The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration. *J. Chemom.* 2005; **19**: 107–118.
3. Rossi DT, Pardue HL. Effects of wavelength range on the simultaneous quantitation of polynuclear aromatic hydrocarbons with absorption spectra. *Anal. Chem.* 1985; **175**: 153–161.
4. Lindgren F, Geladi P, Rännar S, Wold S. Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms. *J. Chemom.* 1994; **8**: 349–363.
5. Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* 1996; **68**: 3851–3858.
6. Osborne SD, Jordan RB, Kunnemeyer R. Method of wavelength selection for partial least squares. *Analyst* 1997; **122**: 1531–1537.
7. Forina M, Casolino C, Pizzarro Millan C. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *J. Chemom.* 1999; **13**: 165–184.
8. Leardi R. Application of genetic algorithm PLS for feature selection in spectral data sets. *J. Chemom.* 2000; **14**: 643–655.
9. Nørgaard L, Saudland A, Wagner J, Nielsen JP. Munck L, Engelsen SB. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* 2000; **54**: 413–419.
10. Jiang JH, Berry RJ, Siesler HW, Ozaki Y. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. *Anal. Chem.* 2002; **74**: 3555–3565.
11. Abrahamsson C, Johansson J, Sparén A, Lindgren F. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. *Chemom. Intell. Lab. Syst.* 2003; **69**: 3–12.
12. Rossi F, Lendasse A, François D, Wertz V, Verleysen M. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemom. Intell. Lab. Syst.* 2006; **80**: 215–226.
13. Stout F, Kalivas JH, Hebergér K. Wavelength selection for multivariate calibration using Tikhonov regularization. *Appl. Spectrosc.* 2007; **61**: 85–95.
14. Kohonen T. *Self Organizing Maps*. Springer: Berlin, 2001.
15. Pöllä M, Honkela T, Kohonen T. Bibliography of self-organizing map (SOM) papers: 2002–2005. Unpublished manuscript 2006. Laboratory of Computer and Information Science, Helsinki University of Technology, Finland.
16. Oja M, Kaski S, Kohonen T. Bibliography of self-organizing map (SOM) papers: 1998–2001 addendum. *Neural Computing Surveys* 2003; **3**: 1–156.
17. Kaski S. Data exploration using self-organizing maps. *Doctoral Dissertation* 1997. Laboratory of Computer and Information Science. Helsinki University of Technology, Finland.
18. Vesanto J. Data exploration process based on the self-organizing map. *Doctoral Dissertation* 2002. Laboratory of Computer and Information Science. Helsinki University of Technology, Finland.
19. Corona F. Development and applications of data-derived models for process industry. *Doctoral Dissertation* 2007. Department of Chemical Engineering and Materials, University of Cagliari, Italy.
20. Ultsch A. Self-organizing neural networks for visualization and classification. *Information and Classification*. Opitz O, Lausen B and Klar R. Springer: Berlin, 1993; 307–313.
21. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vanderwalle J. *Least-squares Support-vector Machines* 2002. World Scientific: Singapore, 2003.
22. Ritter H, Schulten K. Convergence properties of Kohonen's topology conserving maps: fluctuations, stability and dimension selection. *Biol. Cybern.* 1988; **60**: 59–71.
23. Erwin E, Obermayer K, Schulten K. Self-organizing maps: stationary states, metastability and convergence rate. *Biol. Cybern.* 1992; **97**: 35–45.
24. Vesanto J. SOM-based data visualization methods. *Intelligent Data Analysis: An International Journal* 1999; **3**: 111–126.
25. Vesanto J, Ahola J. Hunting for correlations in data using the self-organizing map. *Proceedings of CIMA'99—International Conference on Computational Intelligence Methods and Applications* 1999; 279–285.
26. Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer: New York, 2001.
27. Kiviluoto K. Topology preservation in self-organizing maps. *Proceedings ICNN'96—International Conference on Neural Networks* 1996; 294–299.
28. A.S.T.M. Committee. Petroleum products and lubricants. *Annual Book of ASTM Standards*, Vol. 4. ASTM International: West Conshohoken. 2007.
29. Wheeler OH. Near infrared spectra of organic compounds. *Chem. Rev.* 1959; **59**: 629–666.
30. Weyer LG. Near-infrared spectroscopy of organic substances. *Appl. Spectrosc. Rev.* 1985; **21**: 1–43.
31. Kelly JJ, Callis B. Nondestructive procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. *Anal. Chem.* 1996; **62**: 1444–1451.




  