Neurocomputing 73 (2009) 71–79

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A SOM-based approach to estimating product properties from spectroscopic measurements

Francesco Corona [a,*], Elia Liitiäinen [a], Amaury Lendasse [a], Lorenzo Sassu [b], Stefano Melis [b], Roberto Baratti [c]

[a] Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 HUT, Finland
[b] SARTEC Saras Ricerche e Tecnologie S.p.A., Process Department, Strada 5 Traversa C, I-09032 Assemini, Italy
[c] University of Cagliari, Department of Chemical Engineering and Materials, Piazza d'Armi 1, I-09123 Cagliari, Italy

ARTICLE INFO

Available online 3 August 2009

Keywords:
Self-organizing maps
Variable selection
Chemometrics

ABSTRACT

In this work, the problem of real-time monitoring of products' properties from spectrophotoscopic measurements is presented. Light absorbance spectra are used as inputs to software sensors that estimate outputs otherwise difficult to measure on-line. We approached the problems associated to calibrating the estimation models from very high-dimensional inputs and a reduced number of observations by selecting only a subset of relevant inputs emerging from the topological structure of the data. The topologically preserving representation is performed using the self-organizing map (SOM) where the input significance to the output is computed with the measure of topological relevance (MTR on SOM). As a result, we found that spectral inputs with a topology that 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. Being based on a selection of original spectral variables, the resulting models retain the chemical interpretability of the underlying system. Moreover, the selection approach is independent on the regression model to be embedded in the soft sensors. To support the presentation, the utility of the MTR on SOM is discussed on full-scale problems from pharmaceutical and refining industry. Based on our results, the approach leads to accurate and parsimonious models that can be efficiently implemented in industrial settings.

1. Introduction

Real-time monitoring has become an essential component of modern process industry for optimizing the production toward high-quality products while reducing operating costs. The tools of on-line analytical chemistry and chemometrics fulfill the necessary requirements for real-time analysis of key chemical and physical properties for a broad variety of materials. This paper focuses on monitoring products' properties from non-invasive and non-destructive measurements obtained by light spectroscopy analysis.

The principle underlying process monitoring from infrared (IR), near- and medium-infrared (NIR and MIR) spectroscopic measurements is the existence of a relationship between the light absorbance spectrum of a given product and the property of interest. In fact, the spectrum is conditioned by the composition of the product and, in turn, the composition determines the property of interest. This relationship is rarely known *a priori* and it is usually reconstructed by calibrating specific data-derived models, without an explicit regard to first-principle criteria. The resulting spectrophotoscopic models are used to generate interesting insights on the underlying chemistry. Moreover, the wide availability of continuous-flow spectrophotometers makes the modeling approach suitable for the design of soft sensing devices that monitor the key properties of the products starting from the measured spectra [1].

However, the problem of estimating the property (the output) is defined from very high-dimensional and intrinsically redundant inputs (the spectrum). Redundancy is observed as the inherent collinearity existing between the spectral inputs. Furthermore, it is not unusual to calibrate models on a number of observations (the product's samples) that is radically smaller than the number of input candidates. To address these problems, two approaches are commonly used. One standard solution is to rely on full-spectrum methods for dimension reduction coupled with regression: principal components regression (PCR) and partial least-squares regression (PLSR) are reference models [2]. The natural refinement of such an approach is to perform a preliminary selection of relevant spectral ranges [3]. However, PCR and PLSR models are intrinsically limited by their linear

[*] Corresponding author.
E-mail address: francesco.corona@hut.fi (F. Corona).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.10.026

structure and, because based on combinations of the original variables, are not trivial to interpret. When "kernelized" [4] or other nonlinear [5,6] generalizations of methods are considered, the insight can be further reduced [7]. Analogous considerations apply to the functional extensions of the methods [8,9]. The alternative solution consists of selecting, among all spectral candidates, only those inputs that truly contribute to a correct estimation of the output and, that are as much as possible not collinear. Thus, variable selection is understood as the limit extension of range selection where the chemical interpretability of the system is explicitly retained. Some recent advances in spectroscopic modeling are based on such an idea. In the absence of a chemical model, the approach is either based on model properties [10] or on relevance indexes [11]. In both cases, however, the computational burden associated to variable selection can be demanding and the approach unpractical because of the large number of input candidates.

In this study, variable selection is approached by exploiting the metric structure of the spectral data, leading to a method that identifies only the spectral inputs with a topology that best matches the output's. The topology preserving modeling of the data is carried out with the self-organizing map (SOM, [12]) over which the measures of topological relevance (MTR on SOM, [13,14]) between the inputs and the output are estimated from Unified-distance matrices (U-matrices, [15]). We found that the inputs with a topology that 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. Thus, suggesting a simple strategy for wavelength selection that leads 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 model can be used to reconstruct their relationship with the output. The regression technique preferred in our applications is the least squares formulation of the support-vector machine (LS-SVM, [16]). For completeness and with simplicity in mind, we also considered classical linear models for ordinary least squares (OLS) and ridge regression. When appropriate, the meta-parameters of the models were validated with standard resampling methods to estimating the prediction accuracy; the leave-one-out cross-validation (LOO-CV) is here adopted [17].

The presentation is organized as follows. Section 2 introduces the monitoring problem and briefly overviews the adopted approach to variable selection using the MTR over the SOM. In Section 3, the applications to real-world problems in process monitoring from the pharmaceutical and oil refining industry are presented and discussed.

2. Methodology

The problem of monitoring product properties from light absorbance spectra can be reformulated within the context of variable selection and associated function estimation. That is, given observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$—where $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]^\top$ and y_i are the inputs (on-line spectrum) and output (off-line analysis) variables for the i-th observation, respectively—the task consists of modeling the underlying functionality $y = f(\mathbf{x})$ that is assumed to exist between the observations. Because of the very high dimensionality d of $\mathbf{x}$ (several hundreds, up to thousands) and the small number N of observations (several tens, up to few hundreds), it is appropriate to operate in a reduced data space whose dimensionality is circumscribed by the intrinsic complexity of the system. Formally, being $\mathbf{x} \in \mathbb{R}^d$ the given set of candidate input variables, it is necessary to select a subset $\check{\mathbf{x}} \in \mathbb{R}^s$, with $s \ll d$, that builds the best model for f, according to some predefined criterion [18].

Here, a three-stage methodology stemming from [19,13] is adopted. The methodology summarizes as follows:

(1) the first stage models the input and output observations onto a self-organizing map where the topological structure of the data is preserved;
(2) the second stage investigates, from the SOM, how the output's topology is related to the topology of the input;
(3) only the inputs with a topology that best matches the topology of the output are selected as relevant.

Once the subset $\check{\mathbf{x}}$ of inputs is selected, any regression model can be used to reconstruct f and predict the output y.

2.1. Topology preserving mappings with the SOM

The self-organizing map, SOM [12], is an adaptive algorithm to formulate the vector-quantization paradigm [20]. In the following, the basic formulation and essential properties of the SOM algorithm are briefly reported.

The SOM consists of a low-dimensional (typically, 2D) regular array of K nodes where a prototype vector $\mathbf{m}_k \in \mathbb{R}^p$ is associated with every node k. Each prototype acts as an adaptive model vector for the observations $\mathbf{v}_i \in \mathbb{R}^p$. In the addressed context of spectroscopy, both the inputs and the output are considered; i.e., $\mathbf{v}_i = [\mathbf{x}_i; y_i]$ and $p = d + 1$. During the computation of the map, the observations are mapped onto the SOM's array and the prototyping model vectors adapted according to the learning rule:

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

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

$$\|\mathbf{m}_c(t) - \mathbf{v}_i(t)\| \le \|\mathbf{m}_k(t) - \mathbf{v}_i(t)\|, \quad \forall k = 1, \ldots, K. \tag{2}$$

The kernel $h_{k,c(\mathbf{v}_i)}$ centered at $\mathbf{m}_c(t)$ is usually chosen in Gaussian form:

$$h_{k,c(\mathbf{v_i})} = \exp\left(-\frac{\|\mathbf{r}_k - \mathbf{r}_c\|^2}{2\sigma^2(t)}\right), \tag{3}$$

where the vectors $\mathbf{r}_k$ and $\mathbf{r}_c$ (in $\mathbb{R}^2$, for a 2D map) represent the geometric location of the nodes on the array, and $\sigma(t)$ denotes the monotonically decreasing width of the kernel that allows for a regular smoothing of the prototypes. On the array, the effect of the kernel decreases with the distance between the BMU and the other prototypes.

The map is computed recursively for each observation. As $\alpha(t) h_{k,c(\mathbf{v}_i)}$ tends to zero with t, the set of prototype model vectors $\{\mathbf{m}_k\}_{k=1}^K$ is updated to represent similar observations in $\{\mathbf{v}_i\}_{i=1}^N$ and the prototypes converge toward their asymptotic limits [21,22]. The resulting model vectors form a submanifold in the original data space where the relevant topological and metric properties of the observations are preserved. Thus, the SOM is to be understood as an ordered image of the original high-dimensional data manifold modeled with a low-dimensional array of prototypes. On the SOM's array, the complex structures existing in the data are represented with simple geometric relationships.

2.1.1. The MTR based on the U-matrix of the SOM

The self-organizing map is widely employed to getting a visual insight of the data and to starting a preliminary investigation of potential relationships between the component variables. From the SOM, dependencies can be either searched by looking for similar patterns in identical positions in component plane and distance-based representations of the map [23] or estimating the correlation coefficients between such displays, as proposed in [24].

We identify the relevant inputs by exploiting the topology preserving properties of the SOM of the input and output data according to a relevance measure derived from the assumed continuity of the unknown functionality $y = f(\mathbf{x})$. Under this hypothesis, if two points $\mathbf{x}_i$ and $\mathbf{x}'_i$ are close together in the input space, it is expectable that $f(\mathbf{x}_i)$ and $f(\mathbf{x}'_i)$ are also close together in the output space. Therefore, the continuity of f is also represented in the local topology of the data and, thus, recoverable from nearest neighbors graphs. If the neighborhood continuity is not satisfied (i.e., the points y_i and y'_i are not close together in the output space) it can be either due to the presence of noise or because the inputs are not related to the output. In order to benefit from the noise-filtering properties of the SOM, this general principle can be directly explored from the set of model vectors $\{\mathbf{m}_l\}_{l=1}^{M}$ of the map.

The standard approach to recover the topological structure of the data from the SOM is to compute the Unified-distance matrix, or U-matrix [15]. The U-matrix $\mathbf{U}$ is built from local distances for each SOM node and, thus, defines a neighboring graph based on the model vectors of the map. To represent the local topology of the component variables, the corresponding U-matrices are calculated independently along each direction of the data space; that is, $\mathbf{U}_{x_j}$ (with $j = 1, \ldots, d$) for the input variables, and $\mathbf{U}_y$ for the output. The measure of topological relevance on the self-organizing map (MTR on SOM, [13,14]) assesses the significance of the input x_j to the output y by calculating the distance between the respective topologies, that is:

$$\mathscr{T}(x_j, y) = \|\mathbf{U}_{x_j} - \mathbf{U}_y\|_F, \tag{4}$$

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

2.2. An input selection strategy for spectroscopy

In principles, variable selection using the MTR on SOM can be simply performed by ranking the inputs according to their relevance to the output, and selecting a reduced but still representative subset $\check{\mathbf{x}} \in \mathbb{R}^s$. However, this basic procedure applied to spectroscopy data is intrinsically limited by the continuous nature of the light's wavelengths domain, regardless the employed relevance index as long as it is continuous. In fact, it is intuitive that 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 because embedding a near-identical informative content.

In such context, the selection scheme proposed in [19] and adapted to the MTR on SOM in [13] can be adopted. The procedure was originally defined for a standard measure of dependence, Pearson's correlation coefficient (CC):

$$\mathscr{R}(x_j, y) = \frac{E[x_j y] - E[x_j]E[y]}{\sqrt{E[x_j^2] - E[x_j]^2}} \sqrt{E[y^2] - E[y]^2}, \tag{5}$$

where, in practice, the expectations are approximated based on a finite number of observations. However, the CC is only able to capture dependencies that manifest themselves in the covariance. This motivated the use of alternative measures of relevance.

In the case of MTR over the SOM, the selection procedure summarizes as:

(1) calculate the full set $\mathscr{T} = \{\mathscr{T}(x_j, y)\}_{j=1}^{d}$ 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: i.e.,

$$\check{\mathbf{x}} = \{\check{x}_{j^*} \subset \mathbf{x} : j^* = \operatorname*{argmax}_{j} \mathscr{T}(x_j, y)\}_{j^*=1}^{s}.$$

The procedure identifies only the inputs that are associated to the local maxima of $\mathscr{T}$, thus, relevant to predict the output. In such a condition, 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 discarded. Being relevance to the output the only supervising criterion for selection, the procedure is still sub-optimal with respect to problem of selecting inputs that are also minimally redundant. Nevertheless, the selected variables are implicitly as much as possible dissimilar, because each prototypes different subsets of inputs separated by the local minima of $\mathscr{T}$.

Because the selection scheme is general and valid for any measure of relevance, as long as it is defines a continuous function in the operating domain of wavelengths of the spectrophotometer, in this study we also considered other measures: namely, (i) mutual information (MI, [25]), and; (ii) noise variance estimates (NVE, [26]), as well as the forementioned CC. For completeness, a brief overview on such measures is reported in the following:

- Mutual information measures the distance between the joint density $p(x_j, y)$ and the product density $p(x_j)p(y)$ in the sense of Kullback–Leibler divergence. The analytic form of the MI is given by

$$\mathscr{I}(x_j, y) = \int p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)} \, dx_j \, dy. \tag{6}$$

  It can be shown that $I(x_j, y) \geq 0$ and $I(x_j, y) = 0$ if and only if the variables x_j and y are independent. The integral can be viewed as a measure of distance between the actual joint distribution and the joint distribution under the assumption of independence of the variables. To estimate MI we used the estimator introduced in [27];
- Noise variance estimation is a technique that, under the assumption that there is a functional relationship between x_j and y, estimates the part of the output that cannot be modeled with the given inputs (i.e., the noise). As such noise variance estimates can be also understood as the best possible mean squared error (MSE) obtainable by any model. The task can be done in various ways of which we chose the well-known estimator proposed in [26].

3. Experimental

The development and application of the studied soft-sensors is illustrated on a selection of actual monitoring tasks from pharmaceutical and refining industry. The selected applications are referenced and industrial full-scale problems for variable selection and interpretation, as well as prediction purposes.

3.1. *Application to pharmaceutical industry: predicting the composition of active substance in tablets*

The first application consists of estimating the content of active substance in pharmaceutical tablets. The problem is discussed in detail for Escitalopram® tablets produced by H. Lundbeck A/S (Valby, Denmark) using the measurements provided by the Spectroscopy and Chemometrics Group at the Faculty of Life Science, University of Copenhagen (Denmark), which is kindly acknowledged for sharing the data.

The case is interesting because the identification of the inputs associated to the active substance can be prevented by the superposition of interfering artifacts due to the presence of the excipients and the production processes. Moreover, the good manufacturing practice (GMP) requires pharmaceutical industries to perform frequent content uniformity (CU) controls on the finished products; a requirement that is usually fulfilled by time and solvent consuming chromatographic analysis operated by specifically trained laboratory personnel. Therefore, the production of such a drug would greatly benefit from the availability of fast and reliable methods alternative to conventional tests.

3.1.1. *Problem and data description*

Four different dosages (5, 10, 15 and 20 mg) of the drug are used (see Table 1). The 10, 15 and 20 mg tablets have the same concentration of active substance (i.e., they are dose-proportional with a nominal content equal to 8.0% w/w) and have a slot and a print on one side, whereas the 5 mg tablets have a nominal content of active substance equal to 5.6% w/w.

The tablets have different total weight and also different shape and size. Seven full-scale production batches and 12 batches from pilot plant production are available. Furthermore, three specially prepared batches were produced to extend the calibration range to 85–115% of the nominal content for each dosage form, giving 12 additional laboratory-scale batches. In total 31 batches are used, each batch consisting of 10 tablets that were individually analyzed by the spectroscopic method as well as the reference method. The pilot plant batches are film-coated, while full- and lab-scale batches are not. The tablets contain several excipients, the dominating one being microcrystalline cellulose and, for the coated tablets, the coating material contains titanium dioxide. In addition, it is worthwhile noticing that all the laboratory-scale tablets were stamped with a press using only one punch, whereas the pilot- and full-scale tablets are produced after a total of 40 different punches. Dyrby et al. in [28] provide a detailed description of the experimental setting.

The spectra were acquired in the $4000-14\,000\,\mathrm{cm}^{-1}$ wavenumbers' range (corresponding to the 700–2500 nm wavelengths' range) with a resolution of $16\,\mathrm{cm}^{-1}$. The measurements were recorded with an ABB Bomem FT-NIR (Fourier transform NIR) model MB-160 performing 128 transmittance scans per sample. The main advantage of the transmission mode, when compared to the reflectance mode, is that the resulting spectra contain also information on the inside of the tablets; thus, making the method less sensitive to samples heterogeneity and use of coating materials. The transmittance mode is, however, more sensitive to the pressing process used in producing the tablets.

The absorbances are available only for the $7400-10\,500\,\mathrm{cm}^{-1}$ interval (Fig. 1) because the $4000-7400\,\mathrm{cm}^{-1}$ range was very noisy, while in $10\,500-14\,000\,\mathrm{cm}^{-1}$ very little information is present. The content of active substance in each tablet was evaluated by the reference high performance liquid chromatography (HPLC) method performed in laboratory. Thus, each observation consists of a 404-channel spectrum (i.e., $\mathbf{x} \in \mathbb{R}^d$ with $d = 404$) and the content of active substance (i.e., $y \in \mathbb{R}$). The available measurements summarize to 120 laboratory-scale observations, 120 pilot-scale observations and 70 full-scale observations.

In this study, our objective consists of developing an estimation model that, although calibrated using only the laboratory-scale and the pilot-scale measurements, is directly usable in monitoring the full-scale production.

3.1.2. *Laboratory- and pilot-scale modeling and full-scale predictions*

For the purpose, a preliminary analysis was performed considering the three datasets independently and analyzing, for each, the inputs relevances to the corresponding output.

According to the method discussed in Section 2, three 2D SOMs of the input and output observations in each calibration set were computed. Because no discrimination in learning and testing set is provided with the data, we used the first $\frac{2}{3}$ of each dataset for

Table 1
Study case 1 (tablets): nominal specifications.

Active substance Weight (mg)	Tablet Weight (mg)	Active substance Content (% w/w)	Batches Number
5.0	90	5.6	1 full- + 3 pilot-scale
10.0	125	8.0	2 full- + 3 pilot-scale
15.0	188	8.0	2 full- + 3 pilot-scale
20.0	250	8.0	2 full- + 3 pilot-scale
4.3–5.7	90	4.8–6.3	3 laboratory-scale
8.3–11.4	125	6.9–9.1	3 laboratory-scale
12.9–17.1	188	6.9–9.1	3 laboratory-scale
17.3–22.8	250	6.9–9.1	3 laboratory-scale

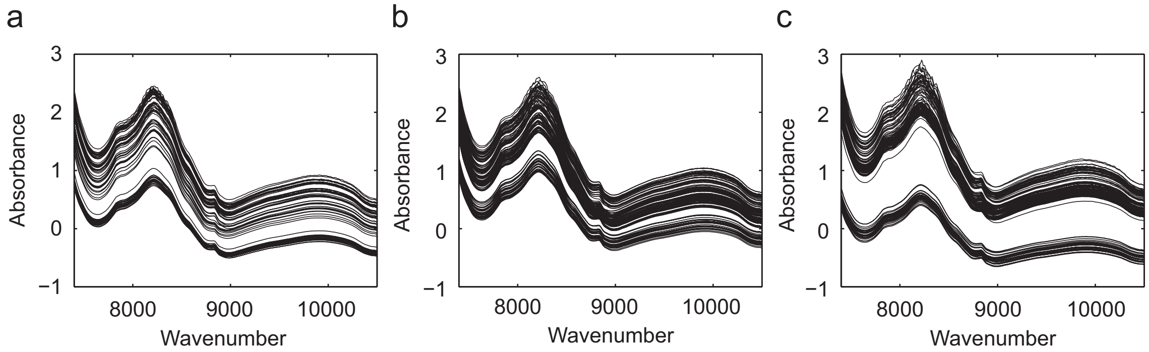

Fig. 1. Study case 1 (tablets): a selection of spectral observations from laboratory (a), pilot-scale (b) and full-scale (c) production.

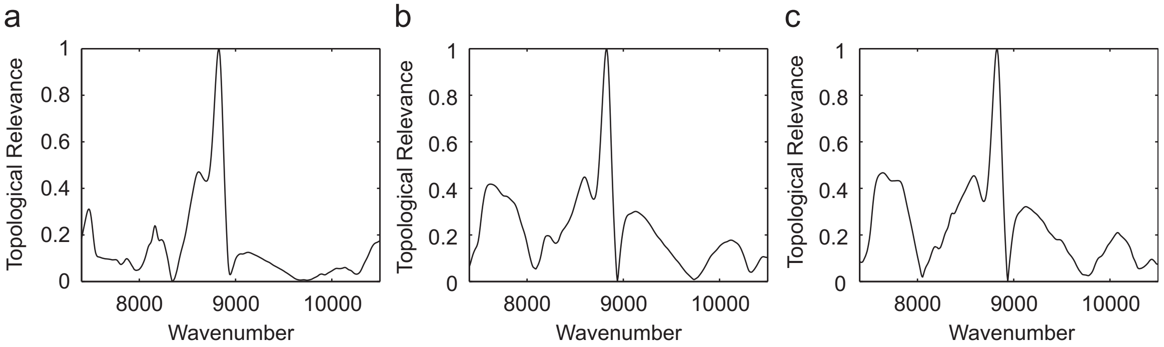

Fig. 2. Study case 1 (tablets): the measures of topological relevance on the self-organizing map for laboratory (a), pilot- (b) and full-scale (c) production.

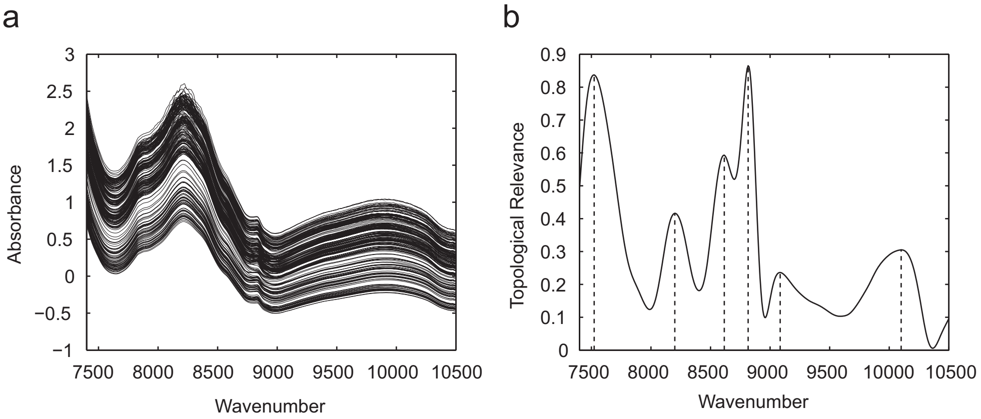

Fig. 3. Study case 1 (tablets): the laboratory and pilot-scale spectral observations (a) and the corresponding MTR on SOM between the inputs and the output (b). The vertical dashed lines are drawn in the correspondence of the selected variables.

calibration. Each map consists of a hexagonal array of nodes initialized in the space spanned by the eigenvectors corresponding to the two largest eigenvalues of the covariance matrix of the data. As usual, the ratios between these eigenvalues were also used to calculate the size of the maps. For each map, the set of topological relevances $\mathscr{T} = \{\mathscr{T}(x_j, y)\}_{j=1}^{d}$ between each input–output pair was calculated. The results are depicted in Fig. 2, notice that no variable selection is performed in this phase.

The NIR spectrum of the active substance is highly overlapped with the excipients' in the tablets, leaving just a single working region (around $8800\,\text{cm}^{-1}$) relatively free of interference, see Fig. 1. In this region, the peak corresponding to the active substance (assigned to the C–H aromatic bond at $\sim 8830\,\text{cm}^{-1}$) is visible as the shoulder of the broadband of the primary excipient ($\sim 8200\,\text{cm}^{-1}$). As expected, the application of the MTR on SOM correctly identifies the matching input as the global maximum of $\mathscr{T}$ for all the production scales, in Fig. 2. Other significant inputs, whose detailed assignment to specific vibrational bands is beyond the scope of this work, are also identified correctly in correspondence to the local maxima; for example, the approach is able to find a match with specific features in the active substance's spectrum (e.g., ~ 7500 and $\sim 8600\,\text{cm}^{-1}$) while assigning a reduced relevance to secondary inputs that are known to be less informative.

Given the analogy between the results obtained with the different production scales, we then considered only the laboratory-scale and pilot-scale measurements and re-applied the methodology to perform variable selection. On the resulting SOM, the set of topological relevances $\mathscr{T} = \{\mathscr{T}(x_j, y)\}_{j=1}^{d}$ between each input–output pair was calculated and the subset $\check{\mathbf{x}} = \{\check{x}_{j^*}\}_{j^*=1}^{s}$ of relevant inputs was selected, $s = 6$. Being the six inputs maximally relevant, they are identified by the local maxima of $\mathscr{T}$, in Fig. 3 and Table 2.

Table 2
Study case 1 (tablets): the set of selected inputs and associated wavenumbers.

	$\check{x}_1$	$\check{x}_2$	$\check{x}_3$	$\check{x}_4$	$\check{x}_5$	$\check{x}_6$
(cm^{-1})	7539	8200	8631	8831	9101	10 116

In Fig. 4, the results obtained with the absolute Pearson's correlation coefficients (Fig. 4(a)), mutual information (Fig. 4(b)) and Gasser's noise variance estimates are presented (Fig. 4(c)). Notice that, in the case of NVE, the local minima reflect the highest relevance. Based on the depicted results, all the measures are capable of identifying either part or all the relevant variables and their behavior resembles, qualitatively, the relevance estimated by the MTR on SOM. Nevertheless, none of the measures is able to represent the smooth nature of the observations and, thus, allow a direct selection of the local maxima in the relevance function. The impossibility to recover such property of the data prevents from an automatic variable selection procedure.

Finally, both linear (OLS and ridge regression) and nonlinear (LS-SVM) models were calibrated to represent f from the six selected inputs $\check{\mathbf{x}}$. When needed, the meta-parameters of the models (the penalty term in ridge regression and the kernel width and regularization term in LS-SVM) were validated by LOO-CV. The prediction accuracy of the models was evaluated in terms of root mean squared error ($RMSE_T$) on an independent set of testing data. The prediction accuracy of the regression models used to

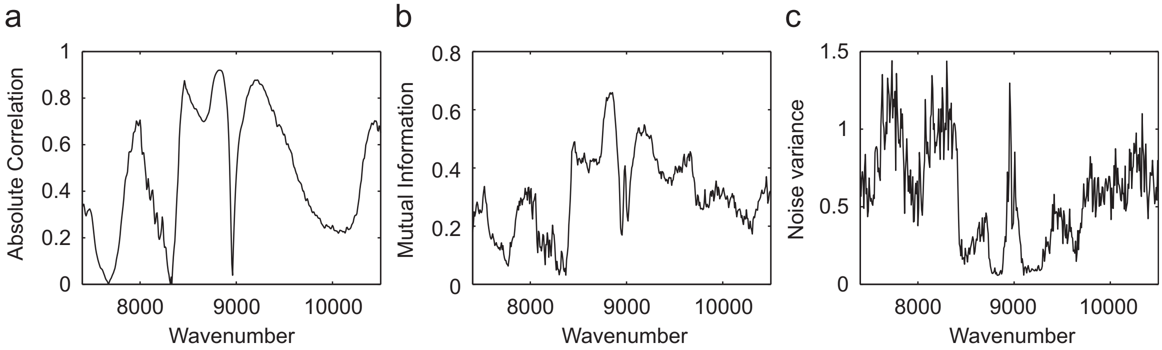

Fig. 4. Study case 1 (tablets): other input–output measures of dependence—correlation coefficient (a), mutual information (b) and Gasser's noise variance (c).

reconstruct f from the six selected inputs $\check{\mathbf{x}}$ is reported in Table 3. The results refer to a direct application of the regression models on the entire set of full-scale measurements.

In Table 3, the prediction results are compared to the two standard calibration methods used in spectroscopy, the full-spectrum PLSR and PCR. The number of latent variables in the PLSR and PCR models was also selected by LOO-CV. From the table, it is possible to notice that the adopted method is not only capable to select the relevant inputs but shows that the associated LS-SVM model gives a prediction accuracy that outperforms the standard methods. Interestingly, also the linear models produce accurate results confirming the quality of the selected variables; this is also demonstrated by an almost negligible value of the penalty term selected for the ridge regression, indicating a near-absolute absence of shrinkage for the regression coefficients.

Based on the experimental results, we can conclude that the method proved capable to select only those inputs carrying important information, thus, leading to parsimonious models based on only six original variables with a clear chemical understandability. Together with the high accuracy, such properties suggest the possibility for an efficient port of the models to the on-line rating of the content of active substance in the full-scale production. In fact, the models could be successfully embedded in a soft sensing device capable of obtaining very accurate results also robust to the different properties of the tablets deriving from interfering artifacts and different production operations.

Table 3
Study case 1 (tablets): a comparison between the results in full-scale production.

	Number of variables	$RMSE_T$
PCR	6 (latent)	0.44
PLSR	5 (latent)	0.42
OLS	6 (original)	0.38
Ridge	6 (original)	0.38
LS-SVM	6 (original)	0.22

3.2. *Application to oil refining industry: predicting quality properties of finished gasolines and diesels*

3.2.1. *Octane number of gasolines*

The application consists of estimating the octane number in gasolines. The American Society for Testing and Materials (ASTM) standard for obtaining such a property is based on an internal combustion engine in which the octane number is measured [29]. The procedure is time consuming, involves expensive and maintenance-intensive equipment and requires skilled labor and, therefore, is not well suited for on-line monitoring. Nevertheless, real-time measurements of such a property are of fundamental importance for both the production and blending processes of the finished fuel. The application of the methodology is discussed on a set of spectral data and associated ratings of octane 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, in Fig. 5. The absorbance is measured on the basis of the NIR transmission principle with a 2 nm resolution. The measurements of the octane number (in the 86–92 range) are evaluated in laboratory by the reference ASTM motor tests. Therefore, each sample consists of the 226-channel spectrum of absorbances and the corresponding octane number; that is, $\mathbf{x} \in \mathbb{R}^d$ with $d = 226$, and $y \in \mathbb{R}$. The dataset consists of 24 observations for model calibration and validation and nine observations for testing the final model. The data were preliminary preprocessed by removing the outliers and mean-centering. Although in reduced amount, the data are collected over a sufficient period of time considered to span all the important variations in the production of the finished product. Being the relationship between the octane and the spectrum distributed among different inputs, the application is also interesting because variable selection cannot be easily performed through first-principle interpretation of the spectra [30,31].

According to the methodology, the 2D SOM of the input and output observations in the calibration set was computed. On the map, the set of topological relevances $\mathscr{T} = \{\mathscr{T}(x_j, y)\}_{j=1}^d$ between each input–output pair was calculated and the subset $\check{\mathbf{x}} = \{\check{x}_{j^*}\}_{j^*=1}^s$ of relevant inputs selected, $s = 6$. Being the six inputs maximally relevant, they are identified by the local maxima of $\mathscr{T}$.

The set of selected inputs (see Fig. 5(b) and Table 4) is in agreement with the chemical model explaining the influence for the chemical groups on the octane number [32]. 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) and to the combination bands (1300–1550 nm) of the C–H vibrations. In detail:

- 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 hydrocarbon although the absorbance is also influenced by the amount of linear paraffin: in fact, its effect

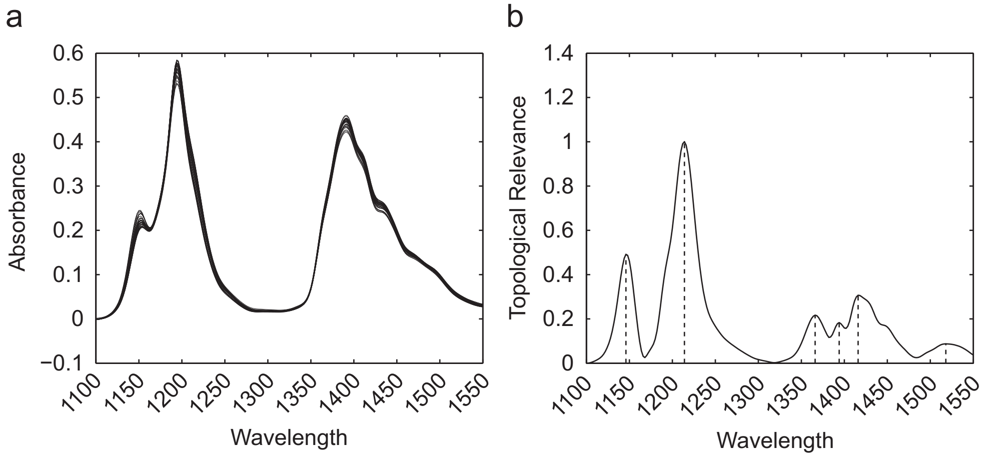

Fig. 5. Study case 2 (gasolines): the spectral observations (a) and the MTR on SOM between the inputs and the output (b). The vertical dashed lines are drawn in the correspondence of the selected variables.

Table 4
Study case 2 (gasolines): the set of selected inputs and associated wavelengths.

	$\check{x}_1$	$\check{x}_2$	$\check{x}_3$	$\check{x}_4$	$\check{x}_5$	$\check{x}_6$
(nm)	1146	1214	1366	1394	1416	1518

Table 5
Study case 2 (gasolines): a comparison between prediction results.

	Number of variables	$RMSE_T$
PCR	3 (latent)	0.21
PLSR	4 (latent)	0.28
OLS	6 (original)	0.34
Ridge	6 (original)	0.31
LS-SVM	6 (original)	0.24

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.

Subsequently, the regression models were calibrated to represent f from the six selected inputs $\check{\mathbf{x}}$ and the prediction accuracy evaluated on the independent set of testing data (Table 5). From the table, it is possible to notice that all the regression models achieve accuracies that are comparable to the ASTM standard of reference. In detail, the LS-SVM gives prediction results that are analogous to the standard PLSR model, whereas the PCR model slightly outperforms all the other methods.

3.2.2. Density of diesels

The last application that we discuss refers to estimating the density of finished diesel fuels from FT-IR (Fourier transform IR) spectra. The issue here is tied to the need of real time collection of process data for quality control during refinery operation. For the purpose, the measurements were acquired from the SARAS Refinery in Sarroch (Italy), which is acknowledged for the support.

The on-line estimation of diesel properties for either quality or process control during refinery operations is usually accomplished inferentially through spectroscopy-based analyzers and chemometric modeling as they usually guarantee high stability (precision) and accuracy (reproducibility) of the measurements The latter being evaluated through comparison with the accuracy laboratory method according to ASTM [33]. Among the many physical and chemical diesel properties routinely measured online, density is one of the most important but also one of the most challenging to be determined inferentially within specs. This is due to the fact that density is a bulk property resulting from the contribution of all components in the hydrocarbon mixture (i.e., it is not associable at any particular spectral feature). In such cases, the full spectral range, containing all the input variables, is usually adopted. Nevertheless, considering the collinearity and the high-dimensionality of the inputs, we investigated the possibility to obtain equivalent or better results selecting fewer variables, although encompassing the complete spectral data range.

The available measurements consist of diesel fuel spectra and associated density measurements, summarizing to 264 observations for model learning and validation and 108 observations for testing. The data were acquired over a year round period during refinery operations yielding to a collection of fuel samples that were as diverse as possible and included the broadest range of values for the property of interest. The spectra were measured with an Analect Diamond 20 FTIR spectrometer (Applied Instrument Technologies, Pomona, CA) over the 6000−1000 cm^{-1} region (approx., 1667–10 000 nm) by using a flow cell, with optical pathlength of 0.5 mm and conditioned at 25 °C. The instrument resolution is 8 cm^{-1}, so that 1297 spectral variables are measured ($d = 1297$). Each sample was scanned 64 times and ratioed with a background spectrum of the empty cell flushed with nitrogen recorded over 128 scans. Density measurements spanning from 845 to 822 kg/m^3 were obtained by using the conventional ASTM method D1298 [29].

Given the experimental setting, the application also offers an interesting extension of the methodology to a broader spectral range. In fact, MIR spectral bands associate to fundamental vibrations giving peaks that are specific and sensitive, in contrast

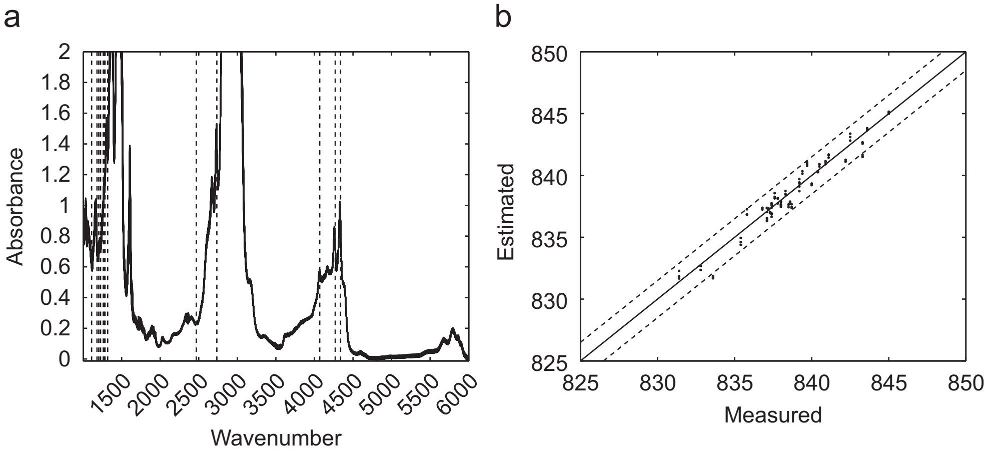

Fig. 6. Study case 3 (diesels): the spectra with vertical dashed lines corresponding to the selected variables (a) and the predictions with 95% confidence bands (b).

to NIRs' that associate to overtone and combination bands originating from the fundamental vibrations in the MIR region (i.e., characterized by low molar absorptivity, or low sensitivity). Although, each technique has distinctive advantages and disadvantages, NIR has historically been the most used of the two. Hence, this application also investigates the potentialities of the methodology on a spectral region which includes portions of both the NIR ($4000-1000\,\mathrm{cm}^{-1}$) and MIR ($6000-4000\,\mathrm{cm}^{-1}$), trying to take advantage of the information carried in each range.

The application led to the selection of 13 input variables, in Fig. 6(a). The selection can be mostly assigned to the spectral features corresponding to the aliphatic C–H bending vibrations (approx., $1503-1296\,\mathrm{cm}^{-1}$), asymmetrical C–O–C stretching vibrations of aliphatic ethers (approx., 1290–1133 and $1133-1000\,\mathrm{cm}^{-1}$), as well as the combination bands of the C–H vibrations (approx., $4000-4500\,\mathrm{cm}^{-1}$). When used to calibrate the regression models, the selected subset of inputs achieved accuracies that are equivalent to the reference values reported in literature [34] and comparable to the analytical methods of measurement. In particular, already a simple linear model like OLS is capable to predict the density of the samples in the independent testing set with a 95% confidence and a $RMSE_T = 0.7\,\mathrm{kg/m}^3$. Such a result, when compared to the full spectrum PCR and PLSR models with 11 latent variables developed with the Saras refinery dataset ($RMSE_T = 0.93$), and to the results from other examples reported in literature (for instance, $RMSE_T = 0.9$ in [34]), shows a significant improvement and suggests direct use of the model in on-line monitoring such a property of the finished diesel fuels.

4. Conclusions

In this paper, the application of a methodology for variable selection based on the measures of topological relevance the self-organizing map was presented and discussed within the context of spectroscopic modeling. The selection methods were investigated on a set of different monitoring problems in industry.

From the obtained results a major consideration can be drawn. The interpretability of the selected variables and sparsity of the obtained models is, indeed, an advantage because of the easy understandability for the domain experts. Moreover, the selected variables are also characterized by an important informative content that can be exploited to develop simple and robust estimation models that always demonstrated capable to achieve the accuracy required for an effective use in real-time monitoring properties of the materials otherwise difficult to measure on-line.

References

[1] J.J.J. Workman, Review of process and non-invasive near-infrared and infrared spectra, Applied Spectroscopy Reviews 34 (1999) 1–89.
[2] P. Geladi, Recent trends in calibration literature, Chemometrics and Intelligent Laboratory Systems 60 (2002) 211–224.
[3] B. Nadler, R.R. Coifman, Prediction error in CLS and PLS: the importance of feature selection prior multivariate calibration, Journal of Chemometrics 19 (2005) 107–118.
[4] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.
[5] E. Oja, Neural networks, principal components, and subspaces, International Journal of Neural Systems 1 (1989) 61–68.
[6] A. Cichocki, R. Unbehauen, Neural Networks for Optimization and Signal Processing, Wiley, New York, 1993.
[7] R. Bolton, D. Hand, A. Webb, Projection techniques for nonlinear principal component analysis, Statistics and Computing (2003) 267–276.
[8] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, in: second ed, Springer, New York, 2005.
[9] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer, New York, 2006.
[10] N. Benoudjit, E. Cools, M. Meurens, M. Verleysen, Chemometric calibration of infrared spectrometers: selection and validation of variables by non-linear model, Chemometrics and Intelligent Laboratory Systems 70 (2004) 47–53.
[11] F. Rossi, A. Lendasse, D. François, W. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometrics and Intelligent Laboratory Systems 80 (2006) 215–226.
[12] T. Kohonen, Self-Organizing Maps, in: third ed, Springer, Berlin, 2001.
[13] F. Corona, L. Sassu, S. Melis, R. Baratti, Measures of topological relevance for soft sensing product properties, IFAC International Symposium on Dynamics and Control of Process Systems, 2007, pp. 175–180.
[14] F. Corona, S.-P. Reinikainen, K. Aaljoki, A. Perkiö, E. Liitiäinen, R. Baratti, O. Simula, A. Lendasse, Wavelength selection using the measure of topological relevance on the self-organizing map, Journal of Chemometrics 22 (2008) 610–620.
[15] A. Ultsch, Self-organizing neural networks for visualization and classification, in: Information and Classification, Springer, Berlin, 1993, pp. 307–313.
[16] J.A.K. Suykens, T.V. Gestel, J. de Brabanter, B. de Moor, J. Vanderwalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002.
[17] T. Hastie, R. Tibshirani, J. Friedman, Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, New York, 2001.
[18] A.J. Miller, Subset Selection in Regression, Chapman & Hall, London, 1990.
[19] F. Corona, A. Lendasse, Input selection and function approximation using the som: an application to spectrometric modeling, In: Workshop on Self-Organizing Maps, 2005, pp. 653–660.
[20] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer, Boston, 1992.
[21] H. Ritter, K. Schulten, Convergence properties of Kohonen's topology conserving maps: fluctuations, stability and dimension selection, Biological Cybernetics 60 (1988) 59–71.
[22] E. Erwin, K. Obermayer, K. Schulten, Self-organizing maps: stationary states, metastability and convergence rate, Biological Cybernetics 67 (1992) 35–45.
[23] J. Vesanto, SOM-based data visualization methods, Intelligent Data Analysis 3 (1999) 111–126.

[24] J. Vesanto, J. Ahola, Hunting for correlations in data using the self-organizing map, in: Computational Intelligence Methods and Applications, 1999, pp. 279–285.

[25] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[26] P.A. Gasser, H.G. Müller, M. Köhler, L. Molinari, A. Prader, Nonparametric regression analysis of growth curves, Annals of Statistics 12 (1984) 210–229.

[27] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Physical Review, Series E 69 (2004) 066138.

[28] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric quantification of the active substance (containing $C\equiv N$) in a pharmaceutical near-infrared (NIR) transmittance tablet using NIR FT-Raman spectra, Applied Spectroscopy 56 (2002) 579–585.

[29] A.S.T.M. International, Annual Book of ASTM Standards—Petroleum Products and Lubricants, vol. 05, 2007.

[30] O.H. Wheeler, Near-infrared spectra of organic compounds, Chemical Reviews 59 (1959) 629–666.

[31] L.G. Weyer, Near infrared spectroscopy of organic compounds, Applied Spectroscopy Reviews 21 (1985) 1–43.

[32] J.J. Kelly, B. Callis, Nondestructive procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines, Analytical Chemistry 62 (1990) 1444–1451.

[33] A.S.T.M. International, Annual Book of ASTM Standards—Molecular Spectroscopy and Surface Analysis, vol. 03, 2007.

[34] G.E. Fodor, R.L. Mason, S.A. Hutzler, Estimation of middle distillate properties, Applied Spectroscopy 53 (1999) 1292–1298.

Francesco Corona received the Laurea degree (M.Sc.) in Chemical Engineering and the Dottorato di Ricerca (Ph.D.) in Industrial Engineering from the University of Cagliari (Italy). He is a researcher in the Laboratory of Computer and Information Science at the Helsinki University of Technology (Finland) where his activity concentrates on the development of data-derived methods for process modeling and their industrial application.

Elia Liitiäinen received the M.Sc. in Automation with a major in Information Technology in 2005 from the Helsinki University of Technology (Finland). He is pursuing the Ph.D. in Computer Science at the same university where his research activities include local learning models and statistical estimation of relevance.

Amaury Lendasse received the M.Sc. in Mechanical Engineering from the Université Catholique de Louvain (Belgium) in 1996, the M.Sc. in Control and Ph.D. in 2003 from the same university. In 2003, he was a researcher in the Computational Neurodynamics Laboratory at the University of Memphis (USA). Since 2004, he is a senior researcher in the Laboratory of Computer and Information Science at the Helsinki University of Technology where he leads the Time series Prediction and Chemominformatics group. Since 2007, he holds a Docentship in the same university. His interests include nonlinear modeling for time series prediction and function approximation, chemometrics, variable selection and noise variance estimation.

Lorenzo Sassu is a senior researcher with Saras Ricerche e Tecnologie S.p.A., Assemini (Italy). He has over 10 years experience in Industrial Chemistry in academic and industry research. Current research interests include process analytical technologies and chemometrics. He holds a Laurea degree (M.Sc.) in Industrial Chemistry from the University of Bologna (Italy) and a Ph.D. in Chemical Engineering from the University of Cagliari (Italy).

Stefano Melis received a Laurea degree (M.Sc.) in Chemical Engineering from University of Cagliari (Italy) in 1992 and a Dottorato di Ricerca (Ph.D.) also in Chemical Engineering, from University of Bologna (Italy) in 1996. Following, he joined ETH Zurich (Switzerland) where he served for 3 years as research assistant. In 1999, he started to work for Saras Ricerche e Tecnologie S.p.A. (Assemini, Italy), where he worked till 2007 as person in charge for the Process Department. Currently, he is working for Albemarle Catalyst Corporation as HPC specialist.

Roberto Baratti received the Laurea degree (M.Sc.) from the University of Cagliari (Italy) in 1982 and the Dottorato di Ricerca (Ph.D.) from the University of Pisa (Italy) in 1986. He is full Professor of Process Dynamics and Control Design at the University of Cagliari in the Department of Chemical Engineering and Materials. His research interests include the development of process models for control, monitoring and optimization purposes and development of control strategies using neural network models. He is member of IFAC, AIDIC and AIChE.