Analytica Chimica Acta 1316 (2024) 342851

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Implementing multiblock techniques in a full-scale plant scenario: On-line prediction of quality parameters in a continuous process for different acrylonitrile butadiene styrene (ABS) products

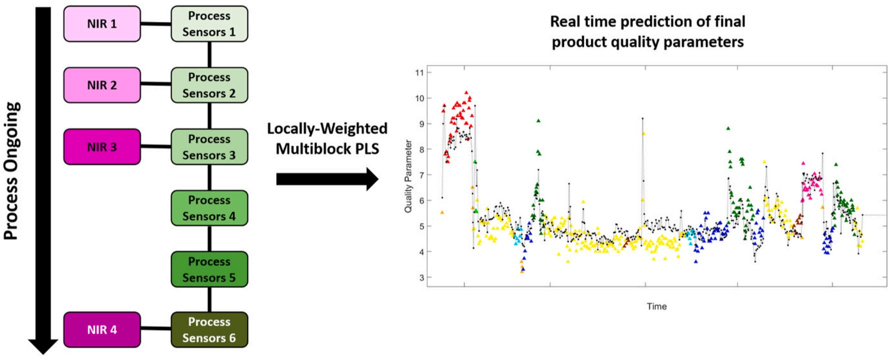
Daniele Tanzilli[a,b], Lorenzo Strani[a,*], Francesco Bonacini[c], Angelo Ferrando[c], Marina Cocchi[a], Caterina Durante[a]

[a] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via 4 Campi 103, 41125, Modena, Italy
[b] Centre National de la Recherche Scientifique (CNRS), Laboratoire de Spectroscopie pour les Interactions, la Réactivitè et l'Environnement (LASIRE), Cité Scientifique, University Lille, F-59000, Lille, France
[c] Research Center, Versalis (ENI) S.p.A., Via Taliercio 14, 46100, Mantova, Italy

HIGHLIGHTS

- Explores issues in handling multiblock data in a highly complex industrial scenario.
- Integration of multivariate local regression with a multiblock approach is proposed.
- Good on-line quality prediction of different grade products obtained by ROSA and LW-MB-PLS.
- LW-MB-PLS effectively reduces systematic prediction errors for specific products.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Prof. L. Buydens

Keywords:
Process monitoring
Real-time predictions
Industrial scale plant
ROSA
Locally weighted multiblock partial least squares
On-line NIR

ABSTRACT

Background: The study explores the challenges of handling multiblock data of different natures (process and NIR sensors) for on-line quality prediction in a full-scale plant scenario, namely a plant operating in continuous on an industrial scale and producing different grade Acrylonitrile Butadiene Styrene (ABS) products. This environment is an ideal scenario to evaluate the use of multiblock data analysis methods, which can enhance data interpretation, visualization, and predictive performances. In particular, a novel multiblock extension of Locally Weighted PLS has been proposed by the authors, namely Locally Weighted Multiblock Partial Least Squares (LW-MB-PLS). Response-Oriented Sequential Alternation (ROSA) has also been employed to evaluate the diverse block relevance for the prediction of two quality parameters associated with the polymer. Data are split in blocks both according to sensor type and different plant sections, and different models have been built by incremental addition of data blocks to evaluate if early estimation of product quality is feasible.

Results: ROSA method showed promising predictive performance for both quality parameters, highlighting the most influential plant sections through the selection of data blocks. The results suggested that both early and late-

[*] Corresponding author. Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125, Modena, Italy.
E-mail address: lostrani@unimore.it (L. Strani).

https://doi.org/10.1016/j.aca.2024.342851
Received 22 December 2023; Received in revised form 5 May 2024; Accepted 7 June 2024
Available online 8 June 2024
0003-2670/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

stage sensors play crucial roles in predicting product quality. A reasonable estimation of quality parameters before production completion has been achieved. On the other hand, the proposed LW-MB-PLS, while comparable in predictive performances, allowed reducing systematic prediction errors for specific products.

Significance: This study contributes valuable insights for continuous production processes, aiding plant operators and paving the way for advancements in online quality prediction and control. Furthermore, it is implemented as a locally weighted extension of MB-PLS.

1. Introduction

In a production process monitoring context, dealing with data from multiple sources is a quite common scenario [1–5], such as when analyzing the same sample with different instruments, to gain more comprehensive information about its features (e.g. in raw material characterization), or when sensors of different nature, typically measuring pressure, temperature, flows etc., are installed throughout a production line, aiming at analyzing the evolution of the product/process in time [1]. Therefore, in these scenarios, data is not merely multivariate, but is also multi-source [5]. For instance, considering data acquired by two different techniques, such as Near Infrared (NIR) spectroscopy and Ultraviolet–Visible (UV–Vis) spectroscopy, the spectral profiles are multivariate, as responses are captured at different wavelengths, and the sources are delineated by the two distinct spectroscopic methods [6]. Moreover, multi-source data may also be acquired when operating under diverse conditions, such as when various batches of an industrial process yield data under distinct processing parameters [5,7,8]. In addition, the quality of the intermediate product can be monitored on-line trough spectroscopic techniques and one of the most used techniques is certainly NIR spectroscopy, due to its non-destructive nature, rapidity, and suitability to be implemented on-line, examples can be found in food process monitoring [9–12], pharmaceutical [13–16] and chemical [17–19] industry. The combined information of these diverse sensors can be employed both to monitor the process ongoing and to predict in real time the quality parameters of the products normally assessed by off-line laboratory analyses [20].

Dealing with multivariate and multi-source data without using the proper chemometric tools can lead to inappropriate interpretation of the results [21]. In this respect, multiblock data analysis methods might be highly valuable for harnessing complementary information from data generated through different sources [21,22]. These methods enable a deeper comprehension of information within this kind of data, improving data visualization, predictive performances and identification of critical variables that significantly influence the models [21–26]. In the predictive context, Multiblock Partial Least Squares (MB-PLS) [27,28] was the first proposed and it is one of the most employed. This prevalence is largely attributed to its simplicity and integration into numerous instrument and statistical software platforms. However, several other methods have been developed, which are more focused on the interpretation of the role of the different blocks [22], such as highlighting the common [29,30] and/or specific information carried by each data block [31–33]. Sequential methods such as Sequential-Orthogonalized Partial Least Squares (SO-PLS) [23] or Response-Oriented Sequential Alternation (ROSA) [34] extract non-redundant information, most salient for prediction, from each different data block analyzed.

In a preliminary study involving a continuous styrenic polymer production plant, the authors evaluated the predictive performances of MB-PLS and ROSA methods, and ROSA gave reliable prediction models, exhibiting solid predictive performance and offering a transparent understanding of the impact of each block on the results [35].

However, continuous processes carried out in industrial scale plants can be extremely complex not only because of their numerous sensors of different nature, but also because different products can be manufactured in the same production line at different times, by changing operational conditions and formulations without interrupting production. In such instances, the plant requires time to adapt to the new conditions, often resulting in the production of non-compliant products. As well as the distinct features of each product introduce additional sources of variance which may lower the prediction performance of the model, as well as because a consistently different range of the parameters to be predicted can take place. On the other hand, computing a separate prediction model for each product type would not be efficient. For instance, attempting to predict the quality of a product that has not been produced for a significant period might lead to inaccuracies due to the lack of process evolution information over time. In this scenario, local regression methods can help in improving the model robustness, as they focus on creating models that adapt to the local characteristics of the data rather than assuming a global relationship, considering information regarding the process evolution at the same time. This allows for a more flexible and nuanced representation of complex patterns [36]. Notwithstanding, to the authors' knowledge, a method that integrates multivariate local regression with a multiblock approach has not been proposed yet.

Hence, in the present work, we afford to build a single real-time predictive model, for a new campaign, from the same styrenic production plant, encompassing two production years, and data collected on several different products produced within the same production line/campaign without interruptions. To this aim, we developed a novel multiblock extension of the local regression method Locally-Weighted-PartialLeast Squares (LW-PLS) [36], namely Locally-Weighted Multiblock Partial Least Squares (LW-MB-PLS). The prediction capability of this method has been evaluated and compared to the one obtained with ROSA and Multiblock Partial Least Squares (MB-PLS). This process constitutes an ideal benchmark for developing real-time predictions at plant scale, showing the features highlighted above, i.e. a high number of diverse process sensors together with four NIR probes, so that the resulting data, also split according to the different sections of the plant, led to diverse data blocks, together with smooth formulation transition to make several different products.

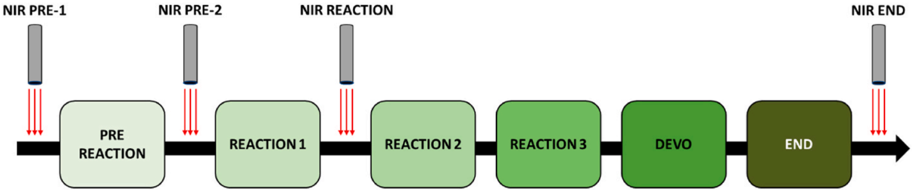


Fig. 1. Schematic representation of the ABS production line. The green blocks represent the six different sections into which the PS have been divided, whereas the gray bars and the red arrows represent the positions where the four on-line NIR probes were placed. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

2. Materials and methods

2.1. Process and sampling

The data collection was conducted in an Acrylonitrile Butadiene Styrene-(ABS) full scale industrial production plant, which operates in continuous, owned by Versalis company (ENI group). The process involves the production of nine different ABS types, which slightly differ in formulation and/or operative conditions. These products will be referred to as "product 1–9". The process can be described by considering six different sections, as shown in Fig. 1. In the first one, called "PRE REACTION", the monomers, namely styrene, butadiene and acrylonitrile, are mixed together. In the following three called "REACTION 1", "REACTION 2" and "REACTION 3", the monomers react, starting to form ABS polymer. In the last two sections, indicated as "DEVO" and "END", take place the removal of the residual monomers and the cut of the final product, respectively. In each section between 5 and 30 Process Sensors (PS) that measure temperatures, flow rates, pressures, and motor speeds are installed, for a total of 118 sensors. Furthermore, along the process line, four NIR probes are also installed. The first two, referred to as "NIR PRE-1" and "NIR PRE-2", are placed at the beginning and at the end of PRE REACTION section, before the occurrence of the reaction, to monitor the reagents before and after their mixture. The third NIR probe is located between REACTION 1 and REACTION 2 sections, to inspect the state of the reaction. Finally, a fourth NIR probe is placed in the END section, at the very end of the process, just before the product is cut. A schematic representation of the production line is displayed in Fig. 1.

In the current work were considered and analyzed process and NIR sensor data acquired on-line from this plant in the period January 2020 to April 2022, as well as quality data collected off-line and analyzed by the company laboratory in the same period.

2.2. ABS quality parameters

Due to confidentiality agreements with the company, the specific names of the two distinct ABS quality parameters considered in this study will remain undisclosed, and they will be denoted as "Quality Parameter 1" (QP1) and "Quality Parameter 2" (QP2). QP1 and QP2 are evaluated through offline analyses of ABS samples, specifically, the final product. This is done three times a day for QP1 and two times a day for QP2. QP1 and QP2 provide insights into the physical attributes of the product. The first one provides information about the fluid dynamic behaviour of the polymer, with the corresponding reference values expressed in grams, whereas QP2 determines the resistance of the product to impacts, and it is expressed in Joule. The company established upper and lower threshold values for both parameters for every ABS product. If either of these values falls outside the specified limits, the end product is deemed to be of lower quality and will be sold at a reduced price. Throughout the duration of this study, a total of 2184 tests were conducted, evenly distributed over time, to assess QP1, while 1349 tests were carried out for QP2. The values for QP1 ranged from 1.6 to 11.1 g (the values have been transformed with logarithm as a preprocessing during model calculation), while QP2 values spanned from 4.1 to 38.9 J.

2.3. NIR measurements

Spectra were collected on-line from the four distinct acquisition points using a Matrix FT-NIR spectrometer (Bruker Optics, Milan, Italy). The instrument was equipped with optical fibers (length of 100 m and a diameter of 600 μm). These fibers were linked directly to the acquisition sites on the process pipe through HT immersion probes (Drawing-no. 661.2350_1, Hellma GmbH and Co. KG, Müllheim, Germany). The acquisition was conducted in transmission mode, spanning the spectral range of 12,500 to 4000 cm^{-1}, with a nominal resolution of 4 cm^{-1} (64 scans per sample).

2.4. Data analysis

The collected data was organized into ten different data blocks, categorized based on data type and the acquisition area in the process. Specifically, PS measurements were arranged into five data blocks, each corresponding to a specific area of the plant. On the other hand, NIR spectra were divided into four blocks, each associated with a single optical probe. Fig. 1 provides the names and abbreviations (which will be used henceforth) of all the blocks, along with their respective positions within the plant. This also serves as an indication of their temporal sequence, given the continuous nature of the process.

2.4.1. Data synchronization

For each applied multiblock technique, the data blocks used for the analysis were constructed following the chronological progression of the ABS production process, considering the placement of the various sensors throughout the production line. In simpler terms, each data point within the datasets corresponds to information gathered at distinct time points, yet it is accurately associated with the same processed material, ensuring data synchronization. The time delay between the various plant sections, indicating the duration for the same material to transfer from one section to another, has been determined using the flow rate values derived from the pumps installed throughout the plant. These specific PS provide information on the material flow (in kg h^{-1}) passing through a reactor or tank. With knowledge of their volumes and the assurance that they are consistently full, it becomes feasible to approximate the time required for the material to traverse from one section to another.

2.4.2. Single block data preprocessing

Each data block underwent distinct preprocessing. Specifically, autoscaling was applied to each PS data block in order to make all the variables to have unit variance, considering their different nature and scales. While, in each NIR data block, spectra were cut in order to consider only the spectral range from 6500 to 5000 cm^{-1}, which displays spectral bands attributable to either reactants or products, and then treated with Standard Normal Variate (SNV) for the analysis of QP1 and with Savitzky-Golay First Derivative (1D) using a 15 points window for the analysis of QP2.

2.4.3. Multiblock methods

To create predictive models for the two parameters under consideration in this study and to evaluate which data blocks are most crucial for their estimation, two multiblock methods were examined, i.e. Response-Oriented Sequential Alternation (ROSA) and a newly developed multiblock implementation of Locally Weighted Partial Least Squares regression (LW-MB-PLS), which will be described in the following sections. The results of the latter were also compared with MB-PLS.

2.4.3.1. Response-Oriented Sequential Alternation. Response-Oriented Sequential Alternation (ROSA) is a multiblock regression approach introduced by Liland et al. [34], based on Partial Least Squares (PLS) regression. ROSA operates as a sequential algorithm, computing a PLS component at time from a single block, in this way the method is invariant to block-scaling (blocks are just mean-centered) distinguishing it from multiblock PLS (MB-PLS), and also to block ordering, distinguishing it from other sequential multiblock methods such as Sequential Orthogonal-PLS [23]. These characteristics enable ROSA to handle numerous blocks of varying dimensions. Additionally, ROSA boasts high computational efficiency. In fact, it bypasses the need for iterative convergence in optimizing criteria, and it only deflates the response variable rather than all the blocks.

Specifically, each PLS component is selected from a single block, choosing the block which gives a single PLS component with the smallest prediction residuals with respect to the other candidate blocks. Subsequent components are constrained to be orthogonal to the subspace spanned by the previously selected components, ensuring orthogonality in scores and loadings.

The main steps of the ROSA algorithm are described by the following equations:

$$\mathbf{w}_b = \mathbf{X}_b^{T}{}^{*}\ \mathbf{y} \tag{1}$$

$$\mathbf{t}_b = \mathbf{X}_b\ {*}\ \mathbf{w}_b/\mathbf{norm}(\mathbf{X}_b\ {*}\ \mathbf{w}_b) \tag{2}$$

$$\mathbf{r}_b = \mathbf{y}\ \text{-}\ \mathbf{t}_b\ (\mathbf{t}_b^T\mathbf{y}) \tag{3}$$

where $\mathbf{X}_b$ represents a single data block, and $\mathbf{w}_b$, $\mathbf{t}_b$, and $\mathbf{r}_b$ denote block weights, scores, and residuals, respectively. The first component, or Latent Variable (LV), is chosen from the b_{th}-block, resulting in the smallest residuals ($\mathbf{r}_b$). The scores ($\mathbf{t}_1$) are set equal to the $\mathbf{t}_b$ of the victorious block. The corresponding weights and scores are subsequently normalized and orthogonalized with respect to the preceding LVs, beginning from the second LV onwards. The **y**-loadings (**q**) are then calculated according to Equation (4):

$$q_a = \mathbf{y}^T\ \mathbf{t}_a \tag{4}$$

$\mathbf{t}_a$ are the previously selected scores for the a_{th} LV. For the calculation of subsequent LVs, steps 1 to 4 are repeated updating y with y-residual relative to the winning block ($r_{b\_winning}$).

The **X**-loadings (**P**) and PLS regression coefficients (**b**) (potentially including a constant term $\mathbf{b}_0$) can be computed using equations (5)–(7), once the optimal number of LVs has been determined, and the corresponding scores, weights and **y**-loadings are gathered in matrices **T**, **W**, and **q**.

$$\mathbf{P} = \mathbf{X}^T\ \mathbf{T} \tag{5}$$

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{q} \tag{6}$$

$$\mathbf{b}_0 = \mathbf{y}_m\text{-}\mathbf{x}_m{}^{*}\mathbf{b} \tag{7}$$

Here, $\mathbf{x}_m$ is a vector containing the mean of every variable of **X**, whereas $\mathbf{y}_m$ is the mean of **y**. In ROSA, every chosen LV carries information exclusively from the winning $\mathbf{b}_{th}$-block (the one with the smallest residuals as per equation (3)), and all LVs are orthogonal. It is crucial to emphasize that all blocks are always considered as candidates at every step of the algorithm. Consequently, successive LVs may contain information from the same previously chosen block or from a different one.

2.4.3.1.1. Selection of model dimensionality. To determine the model's complexity, i.e. the number of PLS components, venetian blinds cross-validation with ten cancellation groups was employed. Cross-validation has been implemented by applying the same samples splitting to each block, prior to single block preprocessing.

2.4.3.2. Locally Weighted Multiblock Partial Least Square regression (LW-MB-PLS). The Locally Weighted Partial Least Squares (LW-PLS) method [36,37] is an extension of PLS designed to provide accurate predictions even in the presence of complex data structures, such as clusters and non-linear relationships [37–39] between independent variables (**X**) and dependent variables (**Y**). In this study, we employed a K-Nearest Neighbors Locally Weighted (KNN-LW) [36] strategy. For a single data set (only one block) this involves selecting from the calibration set the *k* nearest neighbors to each new observation to be predicted. These neighbors are then weighted based on a function [36] that considers a dissimilarity (d_i), measure, e.g. using metrics like the Euclidean distance or Mahalanobis distance, between the selected *k* neighbors and the observation to be predicted. The weight function $f(d_i)$ is defined as:

$$f(d_i) = \exp(\text{-}d_i{}^{*}/(h{}^{*}\sigma(\mathbf{d}^{*})) \tag{8}$$

Table 1
Parameters considered for optimization in Cross-Validation with their respective tested values.

Parameter	Values
Number of LVs (a)	1, 2, 3, 4, 5
Number of nearest neighbors (k)	100, 200, 300, 400, 500, 600
Shape factor (h)	0.1, 0.2, 0.5, 1, 2, 4

where d_i * represents the normalized dissimilarity of the i_{th} neighbor (among the *k* nearest), $\sigma(\mathbf{d}^*)$ is the standard deviation of the vector $\mathbf{d}^*$ (holding the dissimilarity values of all the *k* nearest neighbor) and *h* is a parameter influencing the shape of the weighting function **f.** A higher value of *h* reduces the impact of dissimilarity on the weights. Once the weights are determined, a local PLS model is then calculated, where to each neighboring calibration sample is assigned a different weight according to equation (8). Both **X** and **Y** are mean-centered, and, similarly to standard PLS, the XY covariance between X-scores and Y-scores is maximized, as well the scores of different components are constrained to be orthogonal. The locally weighted PLS model can be expressed by the equations:

$$\text{Cov}(\mathbf{t}_a,\mathbf{u}_a) = \mathbf{t}_a^T\mathbf{D}\mathbf{u}_a \tag{9}$$

$$\mathbf{t}_a^T\mathbf{D}\mathbf{t}_k{=}\mathbf{u}_a\mathbf{D}\mathbf{u}_k = 0\ \text{for}\ a{\neq}k \tag{10}$$

Where $\mathbf{t}_a$ and $\mathbf{u}_a$ are the X-scores and Y-scores vectors for the a^{th} LV, respectively, whereas the **D** matrix holds the local weights for each sample. In terms of regression modeling, the predictions for new samples ($\hat{\hat{\boldsymbol{Y}}}$) are obtained using the equation $\hat{\hat{\boldsymbol{Y}}} = \mathbf{XB}$, where **B** holds the regression coefficients.

We propose in this work a straightforward implementation to extend this locally weighted approach to the multiblock case, i.e. the development of Locally Weighted Multiblock Partial Least Square regression (LW-MB-PLS), maintaining the core of both methods and computational efficiency. The proposed algorithm first performs a low-level data fusion of all blocks by concatenation and applying block scaling. This ensures that a single block of data does not dominate the others solely due to a larger number of variables. Then, the locally weighting scheme is applied to the fused data set. This ensures a unique set of neighbors for each new sample to be predicted, and a single set of weights to be optimized by tuning the *h* parameter. A possible counter side of this unique selection could be that, hypothetically, if the neighbors, for a given sample, were calculated independently for each block of data, they could not necessarily be the same and thus this might result in a suboptimal local model. However, we are convinced that the optimization of both the weights and the number of neighbors can compensate for that providing good predictive performance while maintaining a simpler model.

2.4.3.2.1. Tuning of model parameters. The various parameters, such as the number of latent variables (a), the number of nearest neighbors (k), and the shape factor (h), were optimized through Cross-Validation. The set of values explored for each parameter are reported in Table 1, and all possible combinations were tested. The optimal values were then established by inspection of the corresponding Mage plot [32]. This plot is employed to identify the optimal combination of factors (i.e., those yielding the lowest prediction error) for the input blocks [40].

2.4.3.2.2. Evaluating the block salience. To assess the contribution of each block to the LW-MB-PLS model, analogously to what proposed by Westerhuis et al. [28] we calculated the explained variance for each block. In addition, VIP values for each block (VIP_b) were obtained by summing the VIP values of the variables belonging to the block. In this case for VIP_b a significance threshold equal to the number of variables in a block was used, considering the threshold of one usually set for each single variable. However, since a specific model is used for each sample to be predicted, both parameters attain a different value per block and sample, allowing studying if and how the local models vary when different grade products are considered.

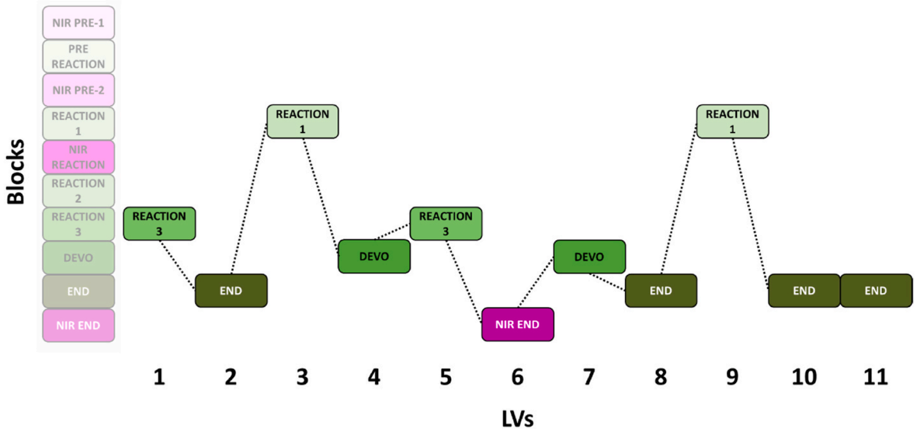


Fig. 2. ROSA model (using all the available blocks) for QP2 prediction. The winning block selected for each LVs is shown in correspondence of the component number. The left bar reports the time order of the blocks along the process.

2.4.3.3. Model building. The data at hand was first split into calibration and validation sets for both QP1 and QP2. To evaluate the models under conditions simulating real-time application, the validation set was constituted of observation pertaining to production period successive to the calibration one. However, as explained in the following, two distinct time windows were considered to take into account that instrumentation maintenance occurred soon after the 2021 summer stop. Hence, the calibration sets consisted of data gathered from January 11th' 2020 to January 23rd' 2021 and from September 22nd to February 6th' 2022 (approximately 70 % of the total data), whereas the validation sets encompassed data from January 24th' 2021 to June 8th' 2021 and from February 7th' 2022 to April 30th' 2022. Furthermore, it should be noted that the plant was not operational from June 9th to July 24th, 2020, and from June 9th to September 22nd. Consequently, no data was recorded during these periods. The data partitioning into calibration and validation sets was performed in this way because after the summer 2021 production stop, the source of the NIR spectrometer was changed.

The preprocessing applied to the different NIR data blocks have been described in section 2.3, whereas to each PS block was applied autoscaling, as explained in section 2.4.

In the case of QP1 models, the PS blocks DEVO and END were treated as a combined block (DEVO-END). This decision stemmed from the fact that the plant experts were primarily concerned with understanding how PS affects QP1 values in the final stages of the process. Determining the individual significance of DEVO or END areas for predicting this parameter was neither useful nor meaningful. Consequently, QP1 models only incorporated nine blocks.

However, for QP2, all ten original blocks were retained, as in this scenario, maintaining the final PS blocks as distinct entities can offer valuable insights. Moreover, in the case of QP2 models, data pertaining to product type 9 was excluded. The choice was motivated by the lower production entity of this specific product and the significantly higher QP2 values observed in comparison to all others, making the resulting models less effective.

The reliability of the predictive models was assessed using the root mean square error in prediction (RMSEP) as well as compared with the root mean square error in cross-validation (RMSECV). The CV-ANOVA [41] approach was employed to assess which are the models that give significantly different RMSECV and RMSEP. This was carried out by two approaches: i) comparing models obtained using the same technique but computed with different blocks used for model building, and ii) comparing models obtained using different techniques but computed with the same blocks used for model building. This approach allowed for the investigation of the significance of both the prediction method utilized and the different starting data blocks employed.

For ROSA method, the importance of each variable within a block was evaluated by inspecting the PLS regression coefficients and the Variable Importance in Prediction (VIP) values [42,43]. For what concerns LW-MB-PLS block explained variance and VIP block values were employed to assess the influence of each block in the ultimate predictive model. Although PLS weights were also examined, the associated figures are omitted for brevity, as the insights gleaned from them were comparable to those obtained from the regression coefficients.

2.5. Software

The chemometric analyses were conducted utilizing routines and toolboxes integrated into the MATLAB environment (the Mathworks Inc., Natick, MA, USA).

The ROSA method, including options for venetian blind cross-validation, VIP calculation, and validation sample response prediction, was implemented by the authors in MATLAB based on the code outlined in Ref. [34].

The LW-MB-PLS algorithm was developed and implemented in MATLAB by the authors starting from the code provided in Ref. [36].

3. Results and discussion

The following sections present the outcomes of the prediction models generated using both ROSA and LW-MB-PLS methods, utilizing different number of blocks, following the process timeline. Initially, ROSA results will be examined, followed by the LW-MB-PLS results. The concluding section will offer a comparative analysis of the two methods.

3.1. ROSA results

The first ROSA prediction models were built involving all the available blocks (9 for QP1 and 10 for QP2). After inspecting the RMSECV values (the maximum number of explored LVs was 20), 13 LV for QP1 model and 11 LVs for QP2 were selected. As described in section 2.4.3.1, ROSA algorithm selects a winning block for each LV, in this case providing information on which are the most influent sections of the plant for the prediction of the parameters under study.

A weakness of ROSA is that it uses the global minimum of residuals to select a "winner block" for each component, while there may be other blocks with residuals that are not statistically significantly different. To investigate this issue, we performed a trial on the ROSA model with all blocks by forcing a different block selection for the first component, selecting in turn each of the blocks with equivalent residuals. As shown in Fig. S1 of Supplementary Material, none of these alternative models was significantly better in term of RMSEP, while some were worse. However, the choice of the first block influences which blocks enter the

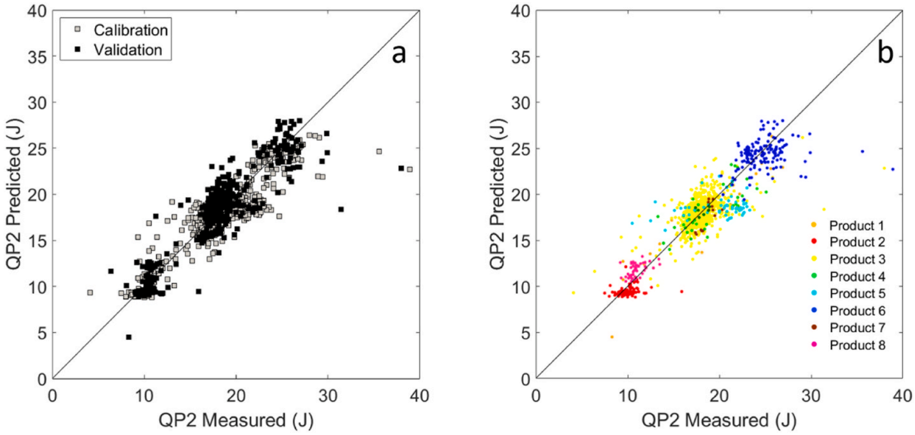


Fig. 3. Plots of predicted vs measured values of QP2 obtained by the ROSA model using all the available blocks. In (a) Samples are colored according to calibration (gray) and validation (black) and in (b) according to ABS product type.

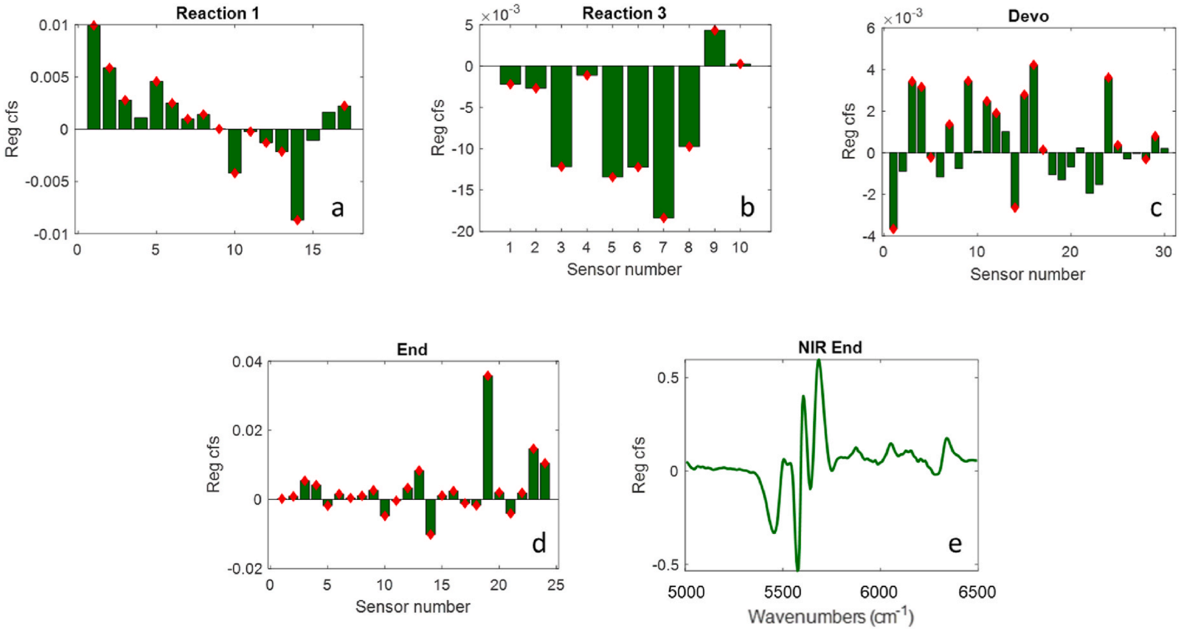


Fig. 4. Regression coefficients for REACTION 1 (a), REACTION 3 (b), DEVO (c), END (d) and NIR END (e). The red diamonds indicate variables with VIP scores exceeding one. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

model at later components and the number of components giving the lowest RMSECV. Finding an algorithmic solution to improve this aspect will be the objective of a dedicated work.

In Fig. 2 is reported the number of times and the order in which the blocks have been selected by the algorithm for each consecutive LV in the case of QP2 model building. It can be observed how the most picked block is END, a predictable outcome since in that area the product can be considered complete. Furthermore, DEVO has been chosen two times, hence the PS blocks ascribable to the final part of the plant have been selected six times out of eleven components. Nonetheless, the first selected block is REACTION 3, selected again for the fifth LV, proving that information retained by sensors in that specific area of the process are important to predict QP2 values. A similar observation applies to the REACTION 1 sensors, which analyze a product that is far from being completed, highlighting that even sensors operating in the early stages offer valuable insights. It can be also easily observed how there is only one NIR among the winning block, that is NIR END, indicating that the NIR spectra collected at the very end of the process, referring to an almost finished product, carry information on the specific QP2 quality parameter, which cannot be gathered by NIR spectra acquired at earlier production phases.

This model yielded a RMSEP of 2.04 J, and Fig. 3a shows a uniform distribution of predictions for the objects in the validation set, all falling within the expected range of QP2. In Fig. 3b the same plot is colored according to the eight different ABS products, highlighting how each product covers a specific part of QP2 range values. For instance, products 2 and 8 presents QP2 values around 10 J, whereas product 6 shows only QP2 values higher than 25. On the other hand, product 3 is the most produced one and its values fall between 15 and 25 J. Product 1 presents few samples scattered all around the QP2 range. Actually, this product serves as intermediate between the productions of other two ABS products that requires a change in formulation and/or in plant settings. Consequently, during a specific time frame, the ABS produced may be labeled as a different product, but it is anticipated that results for this product will be somewhat unstable.

Section 3.3 is devoted to a thorough comparison of ROSA and LW-MB-PLS results, anyhow it is anticipated that the corresponding LW-MB-PLS model using all available data blocks show equal performance

Table 2
Results obtained by applying ROSA.

Blocks used for model building[a]	LVs	RMSECV (g)	RMSEP (g)
QP1			
NP1,PR,NP2,R1,NR,**R2**,**R3**,**DE**,NE[b]	13	0.55^{a}	0.74^{a}
NP1,**PR**,NP2,**R1**,NR,**R2**,**R3**	6	0.72^{b}	0.81^{b}
NP1,**PR**,**NP2**,**R1**,NR,**R2**	10	0.75^{bc}	0.83^{bc}
NP1,**PR**,NP2,**R1**,NR	8	0.83^{c}	0.86^{cd}
PR,R1,**R2**,**R3**,**DE**	13	0.55^{a}	0.74^{a}
PR,**R1**,**R2**,**R3**	6	0.72^{b}	0.81^{b}
PR,**R1**,**R2**	10	0.74^{bc}	0.85^{cd}
PR,**R1**	8	0.83^{c}	0.86^{cd}
NP1,**NP2**,NR,**NE**	11	0.84^{c}	0.89^{d}
NP1,NP2,**NR**	7	1^{d}	1.19^{e}
NP1,**NP2**	6	1.12^{e}	1.27^{f}
Blocks used for model building[a]	LVs	RMSECV (J)	RMSEP (J)
QP2			
NP1,PR,NP2,**R1**,NR,R2,**R3**,**D**,**E**,**NE**	11	1.62^{a}	2.04^{a}
NP1,**PR**,NP2,**R1**,**NR**,R2,**R3**,**D**	9	1.82^{b}	2.46^{b}
NP1,**PR**,NP2,**R1**,**NR**,**R2**,**R3**	7	1.92^{bc}	2.62^{b}
NP1,**PR**,NP2,**R1**,NR,**R2**	8	2.04^{c}	3.52^{d}
NP1,**PR**,NP2,**R1**,**NR**	11	2.11^{c}	3.93^{e}
PR,**R1**,**R2**,**R3**,**D**,**E**	13	1.67^{ab}	2.06^{a}
PR,**R1**,**R2**,**R3**,**D**	12	1.75^{ab}	2.12^{a}
PR,**R1**,**R2**,**R3**	12	1.85^{b}	2.67^{b}
PR,**R1**,**R2**	13	2.25^{d}	2.69^{b}
PR,**R1**	13	2.33^{d}	3.25^{c}
NP1,NP2,**NR**,**NE**	12	1.59^{a}	2.57^{b}
NP1,**NP2**,**NR**	10	2.18^{cd}	3.28^{c}
NP1,**NP2**	11	2.45^{e}	3.4^{cd}

In a column, values with the same letter are not statistically different between each other ($p > 0.05$).

[a] Block names in bold indicate which blocks have been selected by ROSA.

[b] D = DEVO, DE = DEVO-END, E = END, NE=NIR END, NP1=NIR PRE 1, NP2=NIR PRE 2, NR=NIR REACTION, PR=PRE REACTION, R1=REACTION 1, R2=REACTION 2, R3=REACTION 3

in terms of RMSEP (Table 3 and Table S1 of Supplementary Material) while showing less systematic deviations for product 2 and 5 (Fig. S2 of Supplementary Material).

In Fig. 4a to e are displayed the PLS regression coefficients linked to the five blocks selected by ROSA. The red diamonds indicate variables with VIP scores exceeding one, and it's noticeable that nearly all the PS present in each block reach it, except for a few sensors in the Reaction 1 block (Fig. 4a, variables number 4, 15,16) and in the DEVO block (Fig. 4c, variables number 2, 6, 8, 10, 13, 18–23, 26, 27, 30). On the other hand, no wavelengths show VIP scores higher than one (Fig. 4e), suggesting that the NIR contribution for the prediction of QP2 is lower than the one provided by PS blocks. The specific names of the PS must remain undisclosed in compliance with the confidentiality agreement with the company. However, the type of sensor can be disclosed, it is evident that, for the two PS blocks, namely REACTION 1 and REACTION 3 (Fig. 4a and b, respectively), temperature sensors (number 1, 14 in Fig. 4a, 3 and 5, 6, 7, 8 in Fig. 4b) exhibited notably higher regression coefficients compared to others. Similarly, in DEVO and END blocks (Fig. 4c and d, respectively), pressure sensors (number 1, 4, 14 in Fig. 4c and 19 in Fig. 4d) and temperature sensors (number 3, 9, 16, 24 in Fig. 4c, 23 and 24 in Fig. 4d) displayed elevated regression coefficient values. In general, variables that show VIP scores higher than one, but low regression coefficient absolute values are influent just for few LVs, meaning that overall their influence is not highly significant. This information may allow the plant operators to understand which are the specific critical sensors of the plant to keep monitored in order to obtain a final product inside its threshold limits for QP2. In fact, an uncontrolled variation of one of these sensors could heavily influence the quality of the final product.

For the sake of brevity, results obtained by ROSA using all blocks on QP1 are not displayed, but similar results have been obtained. In this case, 13 LVs, according to minimum RMSECV, were used to build the model, obtaining an RMSEP of 0.74 g, and the algorithm selected no NIR blocks. On the other hand, the DEVO-END block resulted winner 10 times out of 13, meaning that the estimation of QP1 strongly relies on the PS data at the end of the process. The other selected blocks were REACTION 2 (1 time, fifth LV) and REACTION 3 (2 times, first and tenth LVs).

While the current results are already promising in terms of prediction performance, two further aspects warrant exploration: firstly, the potential to achieve reasonably accurate QP1 and QP2 predictions before the product is complete. In particular, company was interested in testing prediction models without relying on late-stage sensors, namely REACTION 2, REACTION 3, DEVO, END and NIR END. Secondly, whether relying solely on spectral or process sensors could suffice for a reliable estimation of this quality parameter. In pursuit of this, in addition to the comprehensive dataset encompassing all blocks, different ROSA models were constructed using different datasets assembled as follow: comprising only the blocks preceding the END zone; exclusively PS data; and exclusively NIR data (both with and without the spectra contained in the NIR-END block). The models built in this manner and their respective outcomes are presented in Table 2.

The models that presented the best performance prediction in terms of RMSEP are the ones which starts with all the available blocks, and lowering the number of blocks generally increase the prediction error significantly ($p < 0.05$). This is an expected result, as more information is available and especially that related to the almost finished product, it is possible to observe that for QP1 the data blocks related to NIR are always systematically discarded, whereas for QP2 the NIR REACTION and the NIR END blocks are selected at least one time. To confirm that, the models built starting only with NIR blocks lead to the worst results. These findings can be understood in the context that QP1 and QP2 are not strictly correlated with the chemical composition of ABS. Instead, it assesses the performance of the end-product through mechanical and physical tests. As a result, these product quality parameters are more susceptible to variations occurring during processing, which may introduce substantial changes even if the chemical composition remains constant. In general, RMSEP values for models that excluded blocks associated with the final stages of the process were found to be higher, although still deemed acceptable by process operators. This clearly demonstrates the feasibility of obtaining reasonable estimates for both QP1 and QP2 values before the ABS production process reaches completion. Consequently, two potential approaches emerge for the real-time prediction and control of QP1 and QP2: 1) leveraging both types of data to gain a more precise understanding of crucial process areas and sensors throughout the production plant; or 2) exclusively utilizing PS data for more streamlined data management and to mitigate the impact of noise in the data. Both approaches are extremely relevant for the company. On one hand, it is crucial to obtain accurate predictions of the quality parameters in order to significantly reduce the off-line analyses, saving workforce and reducing wastes. On the other hand, simplifying the data management is equally important in order to make the interpretation of the results more accessible to all the plant operators.

3.2. LW-MB-PLS results

The data analysis using the LW-MB-PLS method followed the same workflow as that employed with the ROSA method. The first inspected model was the one built with all the available blocks and, in this case, results obtained using QP1 as Y are described. Fig. 5a shows the Måge plot used to assess which is the combination of h and k parameters that provided the lowest RMSECV for a specific LV. Combinations that provided very high RMSECV values were not included in order to improve the figure clarity. It emerges that on the Pareto front are present only combinations with h spanning the higher values tested (1–4) while almost all k values are present (except the smallest value of 100) and there is not an interaction between h and k (similar low RMSECV values

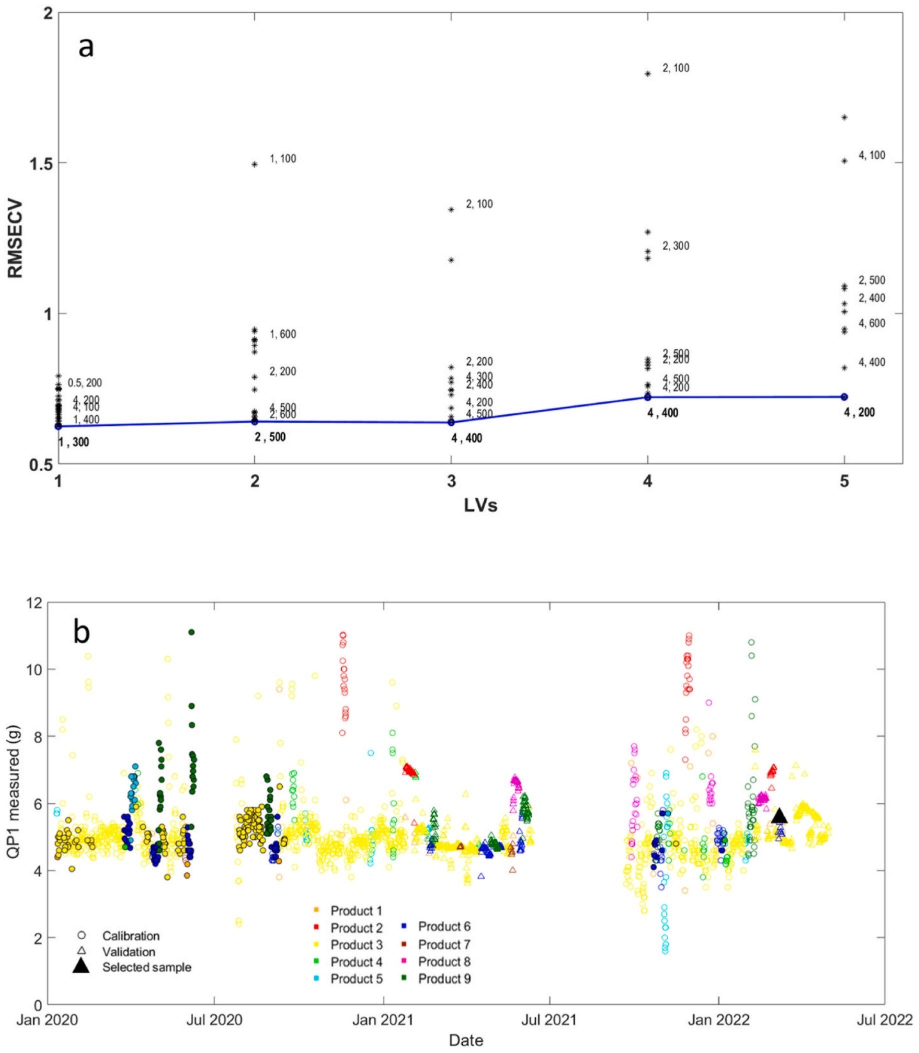


Fig. 5. (A) Måge plot for the LW-MB-PLS QP1 model. The point label report first the value of *h*, then that of *k*. The points on the Pareto front have labels in bold; (b) QP1 values vs time, colored by ABS product. Circles refer to calibration samples, whereas triangles refer to validation samples. The samples represented by the filled circles denote the selected neighbors to build the predictive model for the sample depicted by the black triangle (which belong to Product 6 type). Non-filled symbols represent samples that have not been selected by the model as neighbors.

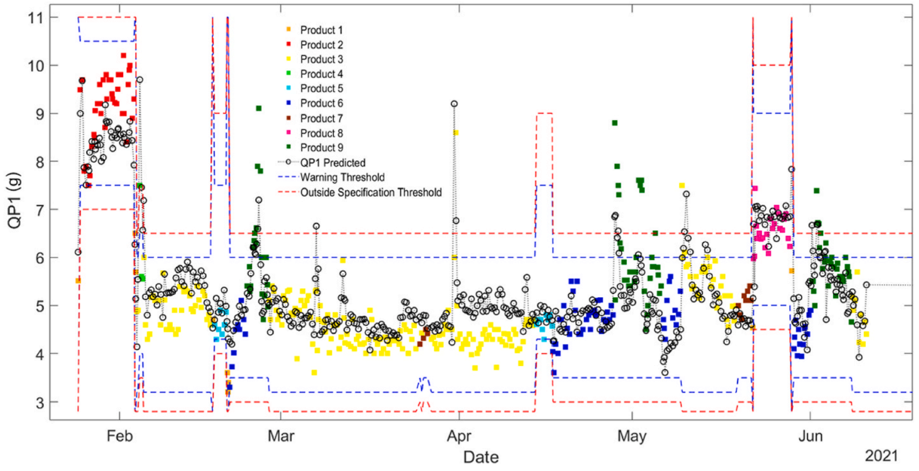


Fig. 6. Time evolution of the measured (colored filled squares) and predicted values (black non-filled circles) of QP1 for the January–June 2021 validation period. The predictions were obtained by means of the LW-MB-PLS model that employed all the available data blocks. Blue and red dashed lines represent the warning thresholds and the actual low-quality threshold, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

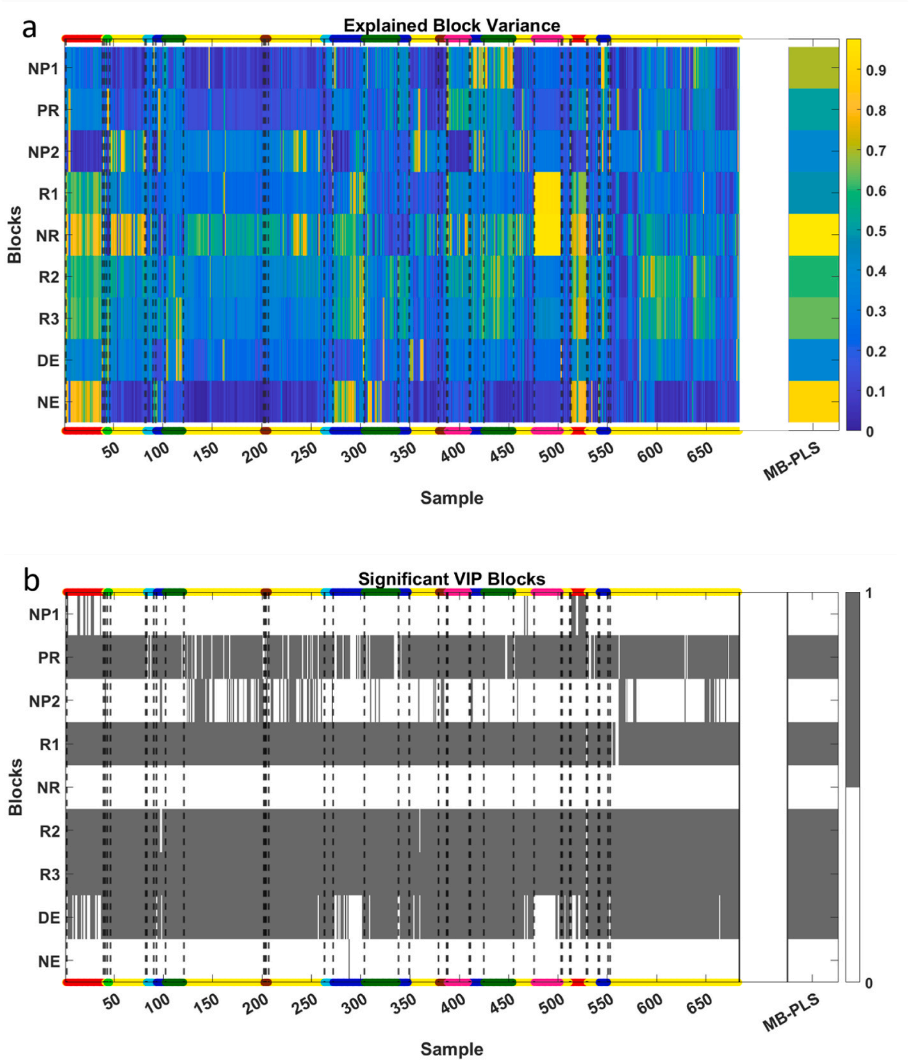


Fig. 7. The results shown refer to validation and prediction sets, i.e. covering the whole production time, for QP1. Explained variance for each block (a) and block VIPs (b) related to the LW-MB-PLS model built with all the available data blocks; values are shown in coded color according to the color bar. Colored lines at the top and the bottom of the figure indicate the product grade, whereas the dashed black lines indicate a product change. On the right of the figures, for comparison, are shown the results of the MB-PLS model computed with the same blocks. In (b), dark gray areas indicate a significant VIP value for the specific block. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

are attained either by small or high *k* notwithstanding which value of *h*). The most influential parameter is the number of LVs, in fact there is a clear error increase after 3 LV. Among the combinations attaining a similar RMSECV value, the more parsimonious one, in term of both number of LVs and neighbors, was selected. The minimum RMSECV corresponds to the following settings: one LV, 300 neighbors (*k*) and an *h* value of 1.

Moreover, the neighbors selected by the algorithm using the Euclidean distance in PCA space for a given sample (exploring several different samples) were inspected to assess if the neighborhood included samples that shares the same type of product or products with similar QP1 values. As an example, in Fig. 5b is displayed a plot of all the QP1 values obtained in the lab (reference method) versus production time, where the 300 neighbors of validation sample number 544 (black filled triangle of March 7, 2022) are represented by the filled circles. It is noticeable that a significant portion of the samples chosen by the algorithm as neighbors belongs to the same ABS product category as the sample to be predicted. In the same way, the majority of QP1 values align closely with the QP1 value exhibited by the sample to be predicted. This means that the chosen sample is predicted using almost only samples similar to it, without considering very different samples that could negatively influence the prediction performances.

The model obtained this way presented a RMSEP of 0.75 g. The prediction trend can be observed Fig. 6. Here, the predicted values of QP1 obtained by the LW-MB-PLS model using all available blocks are represented by the black non-filled circles squares, whereas the filled squares colored according to the different product types represent the QP1 reference values obtained from the off-line laboratory analysis. The Figure specifically displays data from January to June 2021, which corresponds to data included in the validation set. The model's predicted values cover a range very close to that of the validation set, following the production changes. Indeed, even in the event of a formulation alteration or a shift in plant operational parameters, the model nicely tracks the trend in predictions. Blue and red dashed lines represent the two thresholds set by the company to assess if a product is under specification or not. Specifically, blue line represents a warning value, where the product is still considered of high quality but close to the out of specific threshold represented by the red line. Obviously, these thresholds vary according to the different ABS products. The predictions follow the trend of the reference analysis when they fall above the thresholds, even if sometimes it seems that the model underestimates some of these values.

For comparative purposes, analogously to Fig. 6, Fig. S3

Table 3
Results obtained through LW-MB-PLS and MB-PLS.

Blocks for model building	LVs	RMSECV (g)	RMSEP (g)	MB-PLS RMSEP (g)
QP1				
NP1,PR,NP2,R1,NR,R2,R3, DE,NE[a]	1	0.62^{a}	0.75^{a}	0.82^{a}
NP1,PR,NP2,R1,NR,R2,R3	1	0.64^{ab}	0.91^{bc}	0.99^{b}
NP1,PR,NP2,R1,NR,R2	1	0.67^{ab}	0.97^{c}	0.97^{b}
NP1,PR,NP2,R1,NR	1	0.73^{bc}	1.28^{d}	0.97^{b}
PR,R1,R2,R3,DE	2	0.57^{a}	0.78^{a}	0.80^{a}
PR,R1,R2,R3	1	0.59^{a}	0.77^{a}	0.87^{a}
PR,R1,R2	2	0.63^{ab}	0.85^{b}	0.84^{a}
PR,R1	1	0.67^{ab}	0.85^{b}	1.05^{b}
NP1,NP2,NR,NE	2	0.74^{c}	1.34^{d}	2.15^{d}
NP1,NP2,NR	3	0.81^{d}	1.67^{e}	2.64^{e}
NP1,NP2	3	0.98^{e}	1.31^{d}	1.26^{c}
Blocks for model building	LVs	RMSECV (J)	RMSEP (J)	MB-PLS RMSEP (J)
QP2				
NP1,PR,NP2,R1,NR,R2,R3, D,E,NE	3	1.5^{a}	2.13^{a}	2.37^{a}
NP1,PR,NP2,R1,NR,R2,R3, D	2	1.67^{ab}	3.12^{b}	3.66^{c}
NP1,PR,NP2,R1,NR,R2,R3	3	1.7^{bc}	3.11^{b}	3.17^{bc}
NP1,PR,NP2,R1,NR,R2	4	1.72^{bc}	3.45^{b}	4.07^{d}
NP1,PR,NP2,R1,NR	4	1.86^{c}	4.35^{c}	4.64^{fg}
PR,R1,R2,R3,D,E	2	1.61^{ab}	2.1^{a}	2.74^{b}
PR,R1,R2,R3,D	2	1.57^{ab}	2.3^{a}	7.14
PR,R1,R2,R3	1	1.68^{ab}	2^{a}	3.92^{cd}
PR,R1,R2	1	1.67^{ab}	2^{a}	2.06^{a}
PR,R1	1	2.57^{e}	4.06^{c}	4.25^{de}
NP1,NP2,NR,NE	4	1.57^{ab}	3.8^{bc}	4.14^{d}
NP1,NP2,NR	4	2.32^{d}	4.36^{c}	4.51^{f}
NP1,NP2	4	2.5^{e}	4.98^{d}	4.78^{g}

In a column, values with the same letter are not statistically different between each other ($p > 0.05$).

[a] D = DEVO, DE = DEVO-END, E = END, NE=NIR END, NP1=NIR PRE 1, NP2=NIR PRE 2, NR=NIR REACTION, PR=PRE REACTION, R1=REACTION 1, R2=REACTION 2, R3=REACTION 3

(Supplementary Material) show the prediction versus time for QP1 obtained by the ROSA model when all available blocks are considered for model building. The general trend is similar, however for some products and time periods there is evidence of systematic errors, even if the entity is inside the warning thresholds, hence acceptable. The only exception is in April where the product 3 sample which is far above the thresholds is well predicted by LW-MB-PLS (Fig. 6) and not by ROSA (Fig. S3).

3.2.1. *Role of the single block in the local models used for predictions*

Fig. 7a and b represent the explained variance for each block and block VIPs of the inspected model (i.e. QP1, all available block), respectively. At the top and bottom edge of the figures there is a line colored according to each product, while the dashed black lines indicate a product change. The samples shown are order per production time and comprise both validation and prediction sets (before and after the production stop). The right part of the figures also shows, for comparison, the explained variance per block from the MB-PLS model computed with the same blocks. In general, comparing the figures, it can be noticed that in the LW-MB-PLS model there is a certain consistency between the VIP values, or the explained block variance, for the same type of product. Therefore, depending on the product, the blocks relevance changes (e.g., for products 2, red, block NE is contributing to the model, explained variance above 70 %, while for product 3, yellow, it is not), and sometimes also for the same product with time (i.e. block R1 and NR are much contributing for product 8 in the time period June 2022 (about 500 as sample number in Fig. 7a) while not in late May 2021 (about 400 as sample number in Fig. 7a). Noteworthy, between the two time periods the production stop took place. In addition, considering the VIP values, although the NIR blocks contribute a lot to the model for some products/periods, for the same products/periods the VIPs are below the significance threshold, which means that these blocks contribute to components that explain a small percentage of QP1. This is consistent with the fact that ROSA does not select them.

Also in this case, different models were calculated considering different combinations of data blocks. However, the results will be summarized in the next section for a comparison with the ones obtained by ROSA.

3.3. *Comparison between the multiblock methods*

The results obtained with LW-MB-PLS are summarized in Table 3. As in the case of ROSA method, the models that exhibited the most accurate predictive performance in terms of RMSEP are those computed with all available blocks, and reducing the number of blocks tends to increase the prediction error significantly ($p < 0.05$) for both QP1 and QP2, except for QP1 when the excluded blocks are the NIR ones (i.e. the models holding all process sensors blocks and all but the last DE, have same performance). Thus confirming that these are not relevant for predicting QP1. In general, the same consideration done in section 3.1 can be confirmed here. Table 3 also shows the results obtained with standard MB-PLS, which in most cases show higher RMSEP values than those obtained with LW-MB-PLS.

The differences among predictive performance for the three methods, ROSA, LW-MB-PLS and MB-PLS, were evaluated according to ANOVA conducted by considering the error in prediction, as detailed in section 2.4.3.3. The results are shown in Table S1 of Supplementary Materials. Generally, MB-PLS shows significantly worse prediction performances than the other two methods in almost any case (i.e. blocks used for model building), with few exceptions where it performs equally to LW-MB-PLS. In the case of QP1 it is observable how ROSA and LW-MB-PLS demonstrate similar performance mostly when NIR blocks are not present in the considered blocks for model building. On the other hand, when NIR data are present together with process sensor data, LW-MB-PLS provides significantly higher RMSEP values than ROSA, which does not select the NIR blocks (or select just one of them). Thus, confirming that NIR blocks are not important for predicting QP1, and could add noise in the model, the only exception is the LW-MB-PLS model built with all available blocks whose performance does not differ from ROSA. When only NIR blocks are given to build the model again ROSA performs better when it selects only few of them, while performs equally when it selects all of them. Concerning QP2, for which NIR data are generally useful for improving the predictions, ROSA performs better when NIR blocks are involved as model building blocks (the only exception also in this case being when all blocks are available). LW-MB-PLS gives equal or better performance when only process sensors are involved (the only exception is when only the first two, PR and R1, are considered). In general, ROSA performs better when noisy blocks are present because it can select only few of the blocks and only non-redundant information. However, authors observed that for some ABS products LW-MB-PLS helped in decreasing a systematic error in prediction that in ROSA was quite evident.

This can be appreciated by looking at Fig. 8a and b, where the final portion of the validation period corresponding to February–April 2022 for QP1 is reported. Here, the adoption of LW-MB-PLS reduced the model bias, making the prediction trend more accurate. The mean prediction error is equal or slightly lower for ROSA, meaning that LW-MB-PLS outperforms ROSA for the prediction of certain products, such as product 3 in the figure, but for other ABS products the performances are worse.

In conclusion, a first general remark is that ROSA and LW-MB-PLS are based on different methodology. LW-MBPLS, being based on MBPLS, does not provide a clear extraction of the common and distinctive information retrievable from each blocks, since block importance is evaluated only in term of block weights in the final model. On the other hand, ROSA aims at retrieving unique complementary

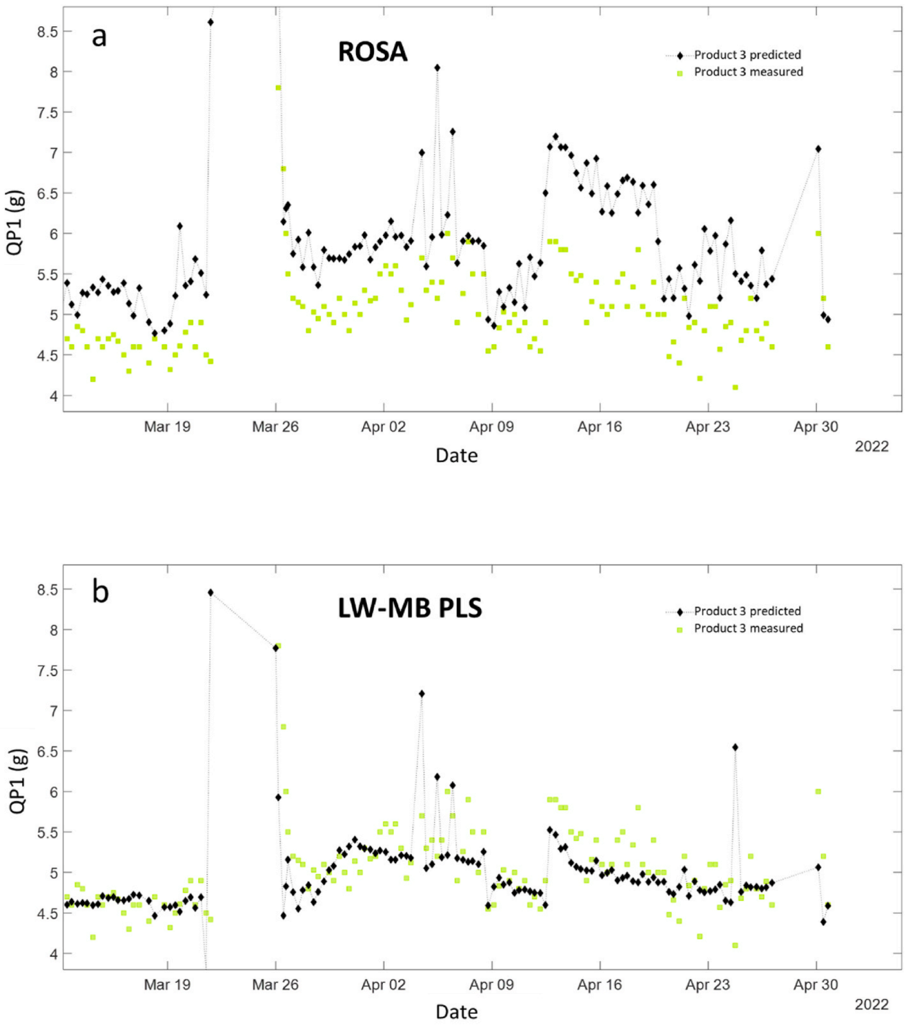


Fig. 8. Time evolution of the measured (green squares) and predicted values (black diamonds) of QP1 for the final portion of February–April 2022 validation period by ROSA (a) and LW-MB-PLS (b). The predictions were obtained by means of the models that employed all the available data blocks. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

information by applying block orthogonalization w.r.t. to the previous extracted component before going for next component, whereas MBPLS does not remove already used information in a block. From an applicative point of view, which is the one concerned in this paper, we may observe that both methods provide models with good predictive capability. ROSA has the advantage of furnishing a single model, hence being very easy to implement in real-time scenario (only the b coefficients need to be stored and used for prediction). LW-MBPLS requires the calculation of the distances between the sample to be predicted and all the calibration samples (slow step) and the fit of a PLS model with the selected neighbors before the prediction step (fast step). Another appealing feature of ROSA is to filter out redundant information among blocks that may guarantee models that are more robust. However, the method is dependent on the choice of the winner block and often several block share similar error, therefore this aspect needs further investigation. However, in cases such as the process studied with multiple product grades, or in presence of non-linearities, a local approach is required to reduce systematic errors.

4. Conclusions

This paper investigated the application of two multiblock regression methods, namely Response Oriented Sequential Alternation (ROSA) and Locally-Weighted Multiblock Partial Least Squares (LW-MB-PLS), a novel extension of Locally-Weighted-Partial Least Squares, for online prediction of quality parameters (QP1 and QP2) in a full-scale styrenic polymer production plant. The study expanded on previous research by incorporating a larger dataset covering all products manufactured by the plant and introduced a new multiblock approach (LW-MB-PLS). The analysis of the results revealed valuable insights into the predictive capabilities of these methods.

The ROSA method demonstrated promising predictive performance for both QP1 and QP2, with the selection of influential blocks providing information about critical sections of the plant. The importance of sensors in early and late stages of the process was highlighted, and the impact of specific sensors on the final product quality was elucidated. The results indicated the feasibility of obtaining reasonable estimates for QP1 and QP2 values before the completion of the production process, offering potential approaches for real-time prediction and control. On the other hand, the LW-MB-PLS method, while generally exhibiting comparable predictive accuracy, demonstrated effectiveness in reducing systematic errors for certain products. The computational efficiency of ROSA was acknowledged, although LW-MB-PLS presented advantages in mitigating bias in predictions for specific ABS products.

From an applicative point of view, both methods are implementable for real time predictions. LW-MBPLS can be recommended when nonlinearity is observed, or as in the present case when different grade of products must be handled. ROSA is especially fast and can be used to sequentially assess the relevance of each block, in addition it may bring to more robust model by filtering redundant information among blocks. In perspective, ROSA can be used in the process-understanding phase to exploit the possible scenarios and then if a prediction bias is observed it can be resorted to local modelling using only the most salient block. However, a drawback of ROSA, which require further investigation, is how to deal with blocks showing similar error in the selection phase.

Overall, this study contributes to the understanding of multiblock regression techniques in the context of continuous production processes, providing valuable insights for plant operators and paving the way for further advancements in online quality prediction and control.

Declaration of generative AI in scientific writing

During the preparation of this work, the authors used ChatGPT in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Daniele Tanzilli: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Lorenzo Strani:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis. **Francesco Bonacini:** Writing – review & editing, Validation, Resources, Methodology, Conceptualization. **Angelo Ferrando:** Writing – review & editing, Supervision, Resources, Conceptualization. **Marina Cocchi:** Writing – review & editing, Validation, Supervision, Software, Project administration, Methodology, Investigation, Conceptualization. **Caterina Durante:** Writing – review & editing, Validation, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

Emilia Romagna Region supported PhD grant of one of the author (D. Tanzilli) fund: PA 2023–20467/RER CUP E83C23002540002.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2024.342851.

References

[1] L. Strani, E. Mantovani, F. Bonacini, F. Marini, M. Cocchi, Fusing NIR and process sensors data for polymer production monitoring, Front. Chem. 9 (2021) 748723.

[2] Z. Ge, Review on data-driven modeling and monitoring for plant-wide industrial processes, Chemometr. Intell. Lab. Syst. 171 (2017) 16–25.

[3] B.S. Silva, M.J. Colbert, M. Santangelo, J.A. Bartlett, P.P. Lapointe-Garant, J. S. Simard, R. Gosselin, Monitoring microsphere coating processes using PAT tools in a bench scale fluid bed, Eur. J. Pharmaceut. Sci. 135 (2019) 12–21.

[4] S.G. Wubshet, J.P. Wold, N.K. Afseth, U. Böcker, D. Lindberg, F.N. Ihunegbo, I. Måge, Feed-forward prediction of product Qualities in enzymatic protein hydrolysis of poultry by-products: a spectroscopic approach, Food Bioprocess Technol. 11 (2018) 2032–2043.

[5] E. Strelet, Y. Peng, I. Castillo, R. Rendall, Z. Wang, M. Joswiak, B. Braun, L. Chiang, M.S. Reis, Multi-source and multimodal data fusion for improved management of a wastewater treatment plant, J. Environ. Chem. Eng. 11 (2023) 111530.

[6] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31.

[7] R. Vitale, O.E. de Noord, J.A. Westerhuis, A.K. Smilde, A. Ferrer, How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding, J. Chemom. 35 (2) (2020) e3266.

[8] M.P. Campos, M.S. Reis, Data preprocessing for multiblock modelling – a systematization with new methods, Chemometr. Intell. Lab. Syst. 199 (2020) 103959.

[9] D. Tanzilli, A. D'Alessandro, S. Tamelli, C. Durante, M. Cocchi, L. Strani, A feasibility study towards the on-line quality assessment of pesto sauce production by NIR and chemometrics, Foods 12 (8) (2023) 1679.

[10] S. Grassi, A. Giraudo, C. Novara, N. Cavallini, F. Geobaldo, E. Casiraghi, F. Savorani, Monitoring chemical changes of coffee beans during roasting using real-time NIR spectroscopy and chemometrics, Food Anal. Methods 16 (5) (2023) 947–960.

[11] G. Gorla, A. Ferrer, B. Giussani, Process understanding and monitoring: a glimpse into data strategies for miniaturized NIR spectrometers, Anal. Chim. Acta 1281 (2023) 341902.

[12] S. Grassi, L. Strani, C. Alamprese, N. Pricca, E. Casiraghi, G. Cabassi, A FT-NIR process analytical technology approach for milk renneting control, Foods 11 (1) (2021) 33.

[13] C.V. Möltgen, T. Puchert, J.C. Menezes, J.C.D. Lochmann, G. Reich, A novel in-line NIR spectroscopy application for the monitoring of tablet film coating in an industrial scale process, Talanta 92 (2012) 26–37.

[14] M. Verstraeten, D. Van Hauwermeiren, M. Hellings, E. Hermans, J. Geens, C. Vervaet, I. Nopens, T. De Beer, Model-based NIR spectroscopy implementation for in-line assay monitoring during a pharmaceutical suspension manufacturing process, Int. J. Pharm. 546 (1–2) (2018) 247–254.

[15] A.Q. Vo, H. He, J. Zhang, S. Martin, R. Chen, M.A. Repka, Application of FT-NIR analysis for in-line and real-time monitoring of pharmaceutical hot melt extrusion: a technical note, AAPS PharmSciTech 19 (2018) 3425–3429.

[16] N.L. Velez, J.K. Drennen, C.A. Anderson, Challenges, opportunities and recent advances in near infrared spectroscopy applications for monitoring blend uniformity in the continuous manufacturing of solid oral dosage forms, Int. J. Pharm. 615 (2022) 121462.

[17] R.R. de Oliveira, R.H. Pedroza, A.O. Sousa, K.M. Lima, A. de Juan, Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy, Anal. Chim. Acta 985 (2017) 41–53.

[18] K. He, M. Zhong, Z. Li, J. Liu, Near-infrared spectroscopy for the concurrent quality prediction and status monitoring of gasoline blending, Control Eng. Pract. 101 (2020) 104478.

[19] L. Strani, F. Bonacini, A. Ferrando, A. Perolo, D. Tanzilli, R. Vitale, M. Cocchi, Real time quality assessment of general purpose polystyrene (GPPS) by means of multiblock-PLS applied on on-line sensors data, Chem. Eng. Trans. 100 (2023) 175–180.

[20] A. Diez-Olivan, J. Del Ser, D. Galar, B. Sierra, Data fusion and machine learning for industrial prognosis: trends and perspectives towards Industry 4.0, Inf. Fusion 50 (2019) 92–111.

[21] A.K. Smilde, I. Måge, T. Naes, T. Hankemeier, M.A. Lips, H.A. Kiers, E. Acars, R. Bro, Common and distinct components in data fusion, J. Chemom. 31 (7) (2017) e2900.

[22] P. Mishra, J.M. Roger, D. Jouan-Rimbaud-Bouveresse, A. Biancolillo, F. Marini, A. Nordon, D.N. Rutledge, Recent trends in multi-block data analysis in chemometrics for multi-source data integration, TrAC, Trends Anal. Chem. 137 (2021) 116206.

[23] A. Biancolillo, T. Naes, The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions, in: M. Cocchi, Data Handling in Science and Technology, Elsevier, Amsterdam, pp. 157-177.

[24] M. Alinaghi, H.C. Bertram, A. Brunse, A.K. Smilde, J.A. Westerhuis, Common and distinct variation in data fusion of designed experimental data, Metabolomics 16 (2019) 2.

[25] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemom. 33 (2019) e3085.

[26] Y. Song, J.A. Westerhuis, A.K. Smilde, Separating common (global and local) and distinct variation in multiple mixed types data sets, J. Chemom. 34 (2020) e3197.

[27] J.A. Westerhuis, P.M. Coenegracht, Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares, J. Chemom. 11 (1997) 379–392.

[28] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemom. 12 (1998) 301–321.

[29] A. El Ghaziri, V. Cariou, D.N. Rutledge, E.M. Qannari, Analysis of multiblock datasets using ComDim: overview and extension to the analysis of (K+ 1) datasets, J. Chemom. 30 (8) (2016) 420–429.

[30] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemom. 10 (5-6) (1996) 463–482.

[31] T. Næs, O. Tomic, B.H. Mevik, H. Martens, Path modelling by sequential PLS regression, J. Chemom. 25 (2011) 28–40.

[32] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, Food Qual 24 (1) (2012) 8–16.

[33] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemom. 25 (8) (2011) 441–455.

[34] K.H. Liland, T. Næs, U.G. Indahl, ROSA—a fast extension of partial least squares regression for multiblock data analysis, J. Chemom. 30 (2016) 651–662.

[35] L. Strani, R. Vitale, D. Tanzilli, F. Bonacini, A. Perolo, E. Mantovani, A. Ferrando, M. Cocchi, A multiblock approach to fuse process and near-infrared sensors for on-line prediction of polymer properties, Sensors 22 (4) (2022) 1436.

[36] M. Lesnoff, M. Metz, J.M. Roger, Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data, J. Chemom. 34 (5) (2020) e3209.

[37] S. Kim, M. Kano, H. Nakagawa, S. Hasebe, Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection, Int. J. Pharm. 421 (2) (2011) 269–274.

[38] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Non-linear regression methods in NIRS quantitative analysis, Talanta 72 (1) (2007) 28–42.

[39] K. Hazama, M. Kano, M. Covariance-based locally weighted partial least squares for high-performance adaptive modeling, Chemometr. Intell. Lab. Syst. 146 (2015) 55–62.

[40] E. Menichelli, T. Almøy, O. Tomic, N.V. Olsen, T. Naes, SO-PLS as an exploratory tool for path modelling, Food Qual. Prefer. 36 (2014) 122–134.

[41] U.G. Indahl, T. Naes, Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling, J. Chemom. 12 (4) (1998) 261–278.

[42] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design. Theory, Methods and Applications, ESCOM Science Publishers, Leiden, 1993, pp. 523–550.

[43] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemometr. Intell. Lab. Syst. 129 (2013) 76–86.