Article:
de Oliveira, RR, Avila, C, Bourne, R orcid.org/0000-0001-7107-6297 et al. (2 more authors) (2020) Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control. Analytical and Bioanalytical Chemistry, 412. pp. 2151-2163. ISSN 1618-2642

https://doi.org/10.1007/s00216-020-02404-2

Data fusion strategies to combine sensor and multivariate model outputs for Multivariate Statistical Process Control

Rodrigo R. de Oliveira[1*], Claudio Avila[2], Richard Bourne[2], Frans Muller[2], Anna de Juan[1]

[1]Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Diagonal 645, 08028, Barcelona, Spain

[2]School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, UK

[*]Corresponding author. Tel.: +34 934034445

e-mail address: rodrigo.rocha@ub.edu; RR de Oliveira ORCID: 0000-0002-4309-5236

Acknowledgements

Funding from the European Community's Framework program for Research and Innovation Horizon 2020 (2014-2020) under grant agreement number 637232, related to the ProPAT project is acknowledged. Funding from Spanish government under the project CTQ2015-66254-C2-2-P is also acknowledged.

Abstract

Process analytical Technologies (PAT) applied to process monitoring and control generally provide multiple outputs that can come from different sensors or from different model outputs generated from a single multivariate sensor. This paper provides a contribution to current data fusion strategies for the combination of sensor and/or model outputs in the development of multivariate statistical process control (MSPC) models. Data fusion is explored through three real process examples combining output from multivariate models coming from the same sensor uniquely (in the near infrared (NIR)-based end-point detection of a two-stage polyester production process) or the combination of these outputs with other process variable sensors (using NIR-based model outputs and temperature values in the end-point detection of a fluidised bed drying process and in the on-line control of a distillation process). The three examples studied show clearly the flexibility in the choice of model outputs (e.g., key properties prediction by multivariate calibration, process profiles issued from a multivariate resolution method, etc.) and the benefit of using MSPC models based on fused information

including model outputs towards those based on raw single sensor outputs for both process control and diagnostic and interpretation of abnormal process situations. The data fusion strategy proposed is of general applicability for any analytical or bioanalytical process that produces several sensor and/or model outputs.

Keywords: Data fusion, Multivariate statistical process control, Near-infrared, spectroscopic sensors, Chemometrics.

1. Introduction

Recent process analytical technology (PAT) applications in analytical and bioanalytical processes generally use data from process analysers, mostly based on spectroscopic measurements, to provide single or several outputs related to process quality indicators [1–5]. The outputs based on spectroscopic measurements come from the use of different multivariate analysis tools, e.g. multivariate calibration models provide prediction of product key properties [6], multivariate curve resolution (MCR) deliver concentration profiles associated with the evolution of compounds in a process [7] and multivariate statistical process control (MSPC) gives indicators that may tell whether the process is on- or off-specifications [8]. In addition to the spectroscopic sensors, most processes are also monitored with simpler devices providing other univariate measurements, such as temperature, pressure, pH or flow rates.

To handle and interpret the measurements of the sensors above in a process monitoring context, MSPC is a well-established methodology for statistical process control and fault diagnosis and identification [9, 10]. However, MSPC tends to be used either on the original multivariate sensor information, e.g., spectroscopic information [1, 11–13] or on sets of univariate process sensors, e.g., temperature, flow, etc. [14–17], but the combination of both kinds of sensors is seldom found.

Indeed, few works are found in the literature combining information from spectroscopic sensors with other process variables to build MSPC models. Gabrielsson et al. have shown that an MSPC model combining UV spectroscopy and process data provides better performance than models built separately for each kind of data set [2]. In this case, no compression of the spectroscopic information was used in the data fusion. Independent MSPC models were developed using data from an electronic nose based on an array of sensors, NIR spectroscopy, mass spectrometry, on-line HPLC and standard on-line bioreactor sensors in a tryptophan fermentation process, but neither the sensor measurements nor the output of the individual MSPC models were combined afterwards [18]. Another work used forward variable selection and cascade artificial neural network procedures to monitor a yoghurt fermentation process [19]. To do so, variables were selected from an electronic nose for prediction of key properties in a primary net followed by a secondary net that used the predicted key properties from the previous net combined with selected NIR wavelength channels and temperature measurements for estimation of a discrete process state variable. However, such a fusion was not meant to perform process control. Another data fusion strategy used for classification problems was

conceived under a mid-level data fusion framework where the multivariate information was merely compressed into scores (typically from Principal Component Analysis, PCA) and fused with other sensor outputs for further analysis; however, no examples of this strategy are found for process control [20].

Data fusion in MSPC offers two main assets: a) the model or sensor outputs joined probe different physicochemical aspects of the processes under study and offer a more accurate description of the system of interest, and b) the fact of considering the different sensor and model outputs together allows testing not only the behavior of each of the process parameters fused, but also that the natural relationship among them be the correct one, something absolutely impossible out of a fusion scenario.

In this study, the concept of data fusion for process control is widened to enclose both the combination of several model outputs from a single multivariate sensor and/or of several sensor outputs in a single data structure following a mid-level fusion strategy [21]. In this way, both measurement and modeling tasks for the same process are interconnected. Indeed, model outputs derived from multivariate sensors, such as predictions of key properties, process concentration profiles, etc., are compressed information much more specific, diverse and interpretable than mere scores and help better to find out the cause of process malfunctions or off-specification situations in a data fusion context.

The data fusion MSPC strategies presented in this work are applied to three real scenarios described below that show diverse combinations of multivariate model outputs and process sensor information.

a) **Pharmaceutical drying process**. This process is monitored with NIR spectroscopy and with temperature sensors placed at different points of a fluidised bed dryer reactor. Process end-point detection is carried out via a data fusion MSPC model combining NIR-based multivariate model outputs, such as moisture prediction and NIR-based MSPC indicators, with temperature measurements (see Fig. 1(a)).

b) **Polyester production process**. This process is monitored only with NIR spectroscopy. Process end-point detection is carried out using a data fusion MSPC model combining different NIR-based model outputs coming from predictions of key properties and NIR-based MSPC information (see Fig. 1(b)).

c) **Distillation process.** This process is monitored by NIR spectroscopy and vapour temperature measurements [7]. Here, the evolution of the distillation process was controlled via data fusion on-line MSPC models based on the combination of compressed NIR-based information, expressed by the concentration profiles derived from multivariate resolution analysis (MCR) of the process spectra and vapour temperature measurements (see Fig. 1(c)).

More detailed comments on the way to build the data structures displayed in Fig. 1 and on the interpretation of the related data fusion MSPC models will be described throughout the text.

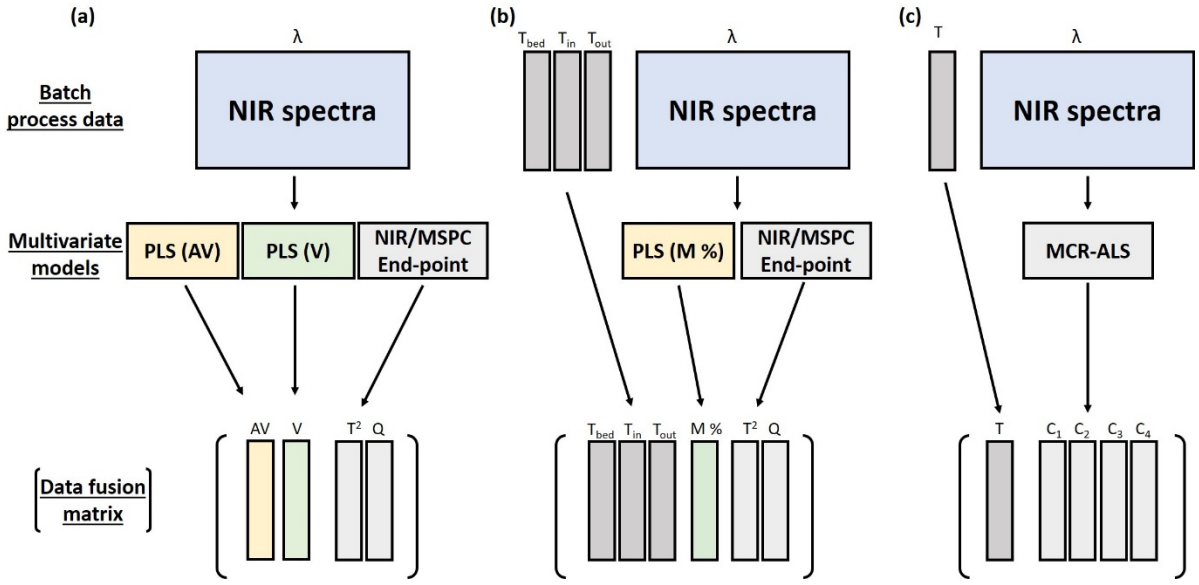

Fig. 1 Data fusion strategies used to combine the several sensor and/or model outputs for batches from (a) *Process 2*, (b) *Process 1;* and (c) *Process 3*

The three processes studied are very different in nature, models and sensors combined in order to show the general applicability of this methodology in any analytical or bioanalytical process context. In all cases, the performance of MSPC models built with the proposed data fusion strategies (hereafter DF-MSPC models) is compared with that of MSPC models built with the sole NIR information (hereafter $MSPC_{NIR}$) through control charts obtained from validation batches. The results obtained clearly show that the use of information coming from different models and/or sensor outputs in data fusion process control models overcomes the performance of the control procedures based on single sensor information and provides a more useful way to identify the causes related to process faults and off-specification situations.

2. Experimental

Three case studies illustrate the different data fusion strategies employed to build multivariate statistical process control (MSPC) models. The experimental monitoring of these processes where NIR spectroscopy and other process variables are monitored is described below.

2.1. Process 1: Fluidized bed drying of pharmaceutical granules

Fourteen batches of 500 g (1-L equivalent) of pharmaceutical wet granules (dry mass fraction of Mannitol >50 %, Avicel PH-101 <30 %, Hypromellose 2910 < 10 %, and other excipients <10 %) were dried in a 4-L fluidized bed (4M8-Trix Formatrix, ProCepT, Belgium). The fluidized bed air inlet flow was controlled at 0.6 or 0.85 m^3/min and a temperature of 22 to 30 ºC. Temperature sensor readings of the fluidized material (T_{bed}), inlet (T_{in}) and outlet air (T_{out}) were recorded simultaneously for each in-situ NIR spectrum. The spectra cover a wavelength range of 1750 to 2150 nm at 1 nm intervals using a spectrophotometer with a novel MEMS Fabry-Perot interferometer (N-Series 2.2, Spectral Engines, Finland) coupled to a diffuse reflectance immersion probe (OFS-6S-100HO/080704/1, Solvias, Switzerland). In-line measurements were collected approximately every second. Off-line reference moisture content analysis was carried out using a thermogravimetric moisture analyzer (MB120, Ohaus, Germany) from samples retrieved at six-minute interval. These moisture reference values were used afterwards to build NIR-based models for moisture predictions. On-specification moisture content was set to be below 2 %. More information can be found in reference [22].

2.2. Process 2: Polyester production process

The production of saturated polyester resins following a commercial recipe was used in this process example and is described in reference [23]. Thirteen batches were carried out with an average batch run time of 22 h. Process monitoring was carried out by in-line NIR absorbance spectra collected inside the two-liter round flask reactor using a NIR immersion probe (Excalibur 20, Hellma Analytics, Germany) with an optical pathlength of 5 mm working in the transmission mode and connected through 2-m fiber optic cables to a spectrophotometer with MEMS Fabry-Perot interferometer (N-Series 1.7, Spectral Engines, Finland) in the (1350 to 1650) nm wavelength range. Spectra were collected every 4 seconds. The key properties selected to follow the progress of the reaction were the acid value (AV) and the high shear

viscosity (V). Off-line determination of AV was carried out by manual acid-base titration following the ASTM D1639-03 method. Off-line values of V were obtained using a cone/plate viscometer (CAP 2000, Brookfield, USA) operating at 200 °C following the procedure described in the ASTM D4287-00 method. These reference values were used to build NIR-based calibration models for in-line prediction of AV and V values. This polyester production process involves two steps and end-point detection models had to be built for each one of them. The targeted ranges to indicate the end-point for the first step are 8 to 12 mg KOH g^{-1} for AV and 10 to 14 P for V, whereas the end-point of the second d stage requires 45 to 63 mg KOH g^{-1} for AV and 25 to 45 P for V. More information can be found in reference [23].

2.3. Process 3: Automated benchtop batch gasoline distillation

An automated batch distillation process with synchronized temperature readings, percentage of distilled mass fraction of initial sample weight and in-line FT-NIR absorption spectra (900 nm to 2600 nm, Rocket, ARCoptix ANIR, Switzerland) was designed and used to monitor the distillation of 100-ml volume of synthetic gasolines [7]. The gasoline batches were prepared by mixing ethanol AR (99 % Sigma-Aldrich) and pure gasoline (type A, from Petrobras refinery) at different ratios. A set of 23 blends was performed: 11 samples containing a volume fraction of 27 % ethanol (on-specification gasolines) and 12 with 10 % to 25 % and 30 % to 40 % ethanol (off-specification gasolines, according to Brazilian legislation). More detailed information can be found elsewhere [7]. NIR spectra in the 1103 – 2228 nm wavelength range (573 channels) and vapor temperature were recorded every unit of percent of distilled mass fraction in the 5 % to 90 % range.

3. Data treatment

3.1. NIR data preprocessing

For processes 1 and 2, similar preprocessing steps were employed to filter out noise and baseline fluctuations on NIR spectra observations. A certain number of consecutive raw spectra measurements, N_{RAW}, were averaged into a single spectrum, then a moving average smoothing with window size, N_{MA}, was employed using the previously averaged spectra. Finally, to remove any unwanted baseline spectral variation in the moving averaged spectra, standard normal variate (SNV) normalization [24] was applied in *Process 1* and Savitzky-Golay derivative [25] (1^{st} order derivative, 2^{nd} order polynomial function and 15 points window) in *Process 2* data.

Temperature measurements were averaged as NIR spectra in *Process 1*, covering the same time window as the number of spectra N_{RAW} averaged to obtain a single one.

Fig. 2(a) shows the raw (left plot) and preprocessed (center plot) NIR observations using N_{RAW} = 10 and N_{MA} = 75 for one typical drying process batch (*Process1*). The right plot shows the related batch temperature profiles. Fig. 2(b) shows the raw (left plot) and (right plot) preprocessed NIR observations using N_{RAW} = 13 and N_{MA} = 30 for the polyester production process batch (*Process 2*).

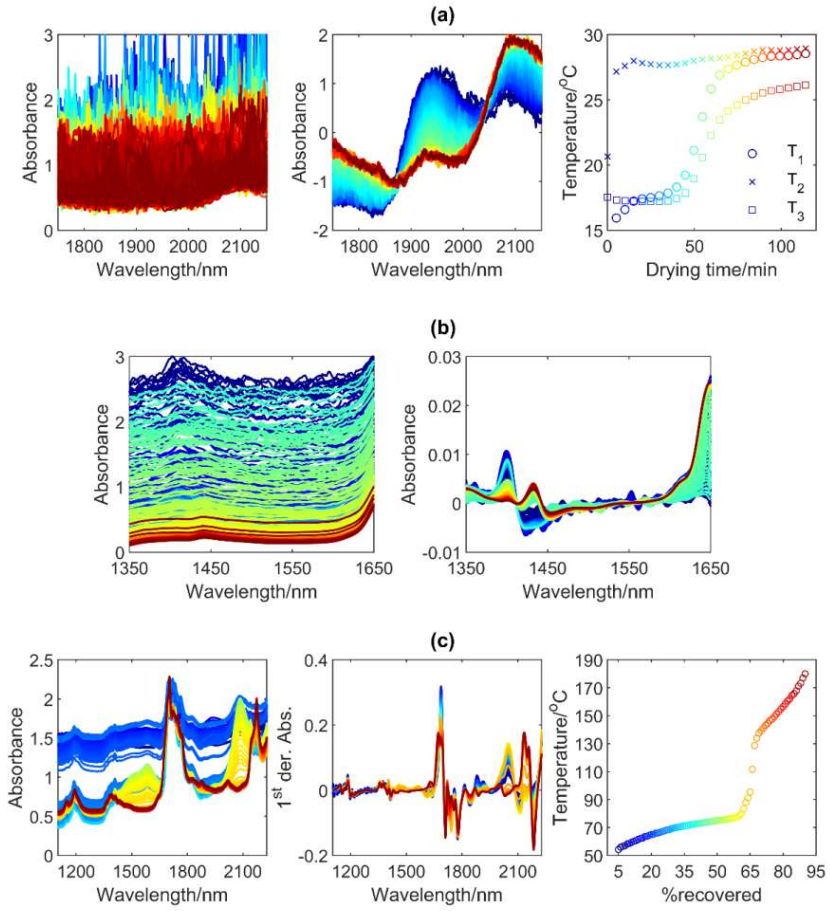

Fig. 2 Data related to a batch from (a) *Process 1:* raw (left) preprocessed (center) NIR spectra and (right) temperature profiles, for better visualization the interval was set to 5 minutes. (b) *Process 2*: raw (left) and preprocessed (right) NIR spectra. (c) *Process 3*: raw (left) and preprocessed (center) NIR spectra and distillation curve of vapor temperature (right). Color scale indicates the temporal variation of batch observations, from the beginning (blue) to the end (red)

In *Process 3*, raw NIR spectra were preprocessed for baseline correction by Savitzky-Golay derivative (1st order derivative, 2nd order polynomial function and 9 points window) followed by spectral normalization to mitigate signal intensity fluctuations. Fig. 2(c) shows the raw (left plot) and preprocessed (center plot) NIR spectra, respectively, and the related

distillation curve (right plot) with recorded boiling temperatures during the 5 to 90 % distillation fractions of an on-specification batch.

3.2. NIR-based model outputs used in data fusion MSPC models

As could be seen in Fig. 1, different kinds of information, issued from the application of different multivariate analysis methods, were used to build MSPC data fusion models. Below, a brief description of the multivariate methods used and the kind of outputs provided is presented.

Partial least squares regression (PLS) was used to build models able to predict key properties of processes that can be estimated from NIR measurements. PLS was used to build models able to predict moisture in *Process 1* and acid number or viscosity in *Process 2*. PLS is the most often used multivariate calibration method in chemometrics [26, 27]. This method relates the **X** matrix (formed by NIR spectra in these examples) to the matrix of parameters to be predicted **Y** (e.g. formed by moisture content, acid number or viscosity values) to build a calibration model with predictive ability that expresses the maximum covariance between **X** and **Y.** More details and description of PLS algorithm can be found elsewhere [28–30]. **Y** predicted values by PLS models are afterwards used in the design of data fusion MPSC models for both *Processes 1* and *2*, as seen in Figures 1 (a) and 1 (b), respectively.

Multivariate curve resolution - alternating least squares (MCR-ALS) is a method that can provide concentration profiles and related spectral signatures for the compounds involved in a process using only the spectroscopic information recorded during process monitoring. MCR-ALS was used to model the NIR data of the distillation process, *Process 3*. MCR-ALS assumes a bilinear model, $\mathbf{D = CS^T + E}$, which is the multiwavelength extension of the Lambert-Beer's law [31–34]. In this context, **D** is a data table with the NIR spectra from several on-specification distillation batches. $\mathbf{S^T}$ contains the pure spectra profiles of the components needed to describe the distillation process and **C** the related concentration (distillation) profiles. Thus, MCR-ALS provides the concentration and spectral profiles of the different distillation fractions of the system. To ensure obtaining meaningful process and spectra profiles, MCR-ALS was applied using non-negativity and unimodality constraints to model **C** profiles, whereas $\mathbf{S^T}$ profiles were not constrained. Details on the implementation of MCR-ALS for *Process 3* modelling can be found in reference [7]. As shown in Fig. 1 (c), **C** profiles are afterwards used as input information for on-line batch MSPC data fusion models described in the next section.

Multivariate Statistical Process Control (MSPC) models aim at providing statistical boundaries that allow building control charts that help to know whether a process is on- or off-specification based on the measurement of NIR spectra. MSPC models based uniquely on NIR multivariate observations were used to provide an additional compressed indication of process evolution (see more detail afterwards in this same section about the MSPC model construction). Summarizing, MSPC indicators T^2 and Q serve as parameters to enclose information related to the unspecific NIR variation linked to the expected process variation and to the acceptable residual variation, respectively. These NIR-derived indicators are afterwards used in data fusion strategies linked to *Process 1* and *Process 2*, as seen in Fig. 1(a) and Fig. 1(b).

3.3. Construction of Multivariate statistical process control (MSPC) models

This section covers the steps required to build an MSPC model, either based on the raw output of an NIR sensor or on combined information leading to a data fusion scenario, as previously described. PCA-based MSPC models are always built using multivariate observations from normal operating condition (NOC) batches to set the statistical boundaries of normal operation. Afterwards, observations of new batches are submitted to the MSPC model to check whether they are within the normal operation boundaries or not.

MSPC models can have different goals, such as end-point detection or checking the process evolution. For *Processes 1 and 2*, MSPC models were designed for batch end-point detection; meanwhile, for *Process 3*, local on-line batch MSPC models were built to check the process evolution using the strategy described in reference [7].

To use MSPC either for end-point detection or for on-line batch evolution monitoring, MSPC models should be built using datasets formed by NOC observations, $\mathbf{X}_{NOC}$, which can be full NIR spectra or the combination of different NIR-based model and/or sensor outputs, as shown in Fig. 1. The dataset $\mathbf{X}_{NOC}$, is modeled by PCA in order to set the statistical boundaries of the experimental domain (space) of NOC observations according to the equation below [35],

$$\mathbf{X_{NOC}} = \mathbf{T_{NOC}}\mathbf{P_{NOC}^{T}} + \mathbf{E_{NOC}} \quad \textbf{(1)}$$

where $\mathbf{T_{NOC}}$ is the scores matrix of the NOC observations used to build the model and $\mathbf{P_{NOC}^{T}}$, the loadings matrix (which is the link between scores and original variables in $\mathbf{X}_{NOC}$). $\mathbf{E}_{NOC}$

describes the residual variation unexplained by the PCA model and is used to define the Q-statistic control chart limit, $Q_{lim.}$, according to Jackson and Mudholkar equation [36].

For any new observation (NIR spectrum or the combined information) acquired in real-time, $\mathbf{x}_{\text{new}}$, the PCA model is used to obtain its related score value, $\mathbf{t}_{\text{new}}$, as follows:

$$\mathbf{t}_{\text{new}} = \mathbf{x}_{\text{new}}\mathbf{P}_{\mathbf{NOC}} \quad \textbf{(2)}$$

Then, the residuals for the new observation are obtained as:

$$\mathbf{e}_{\mathbf{new}} = \mathbf{x}_{\text{new}} - \mathbf{t}_{\text{new}}\mathbf{P}_{\text{NOC}}^{\text{T}} \quad \textbf{(3)}$$

And the related Q-statistic value as:

$$Q = \mathbf{e}_{\mathbf{new}}^{\mathbf{T}}\mathbf{e}_{\mathbf{new}} \quad \textbf{(4)}$$

When the new observation follows the NOC described by the MSPC models, the residual $\mathbf{e}_{\mathbf{i,new}}$ will be small and the related Q value will appear below the chart control limit. Conversely, when the observation does not follow the NOC, the related Q value will appear above the control chart indicating that the process is deviating from the normal process trajectory or that the batch is far from the end-point, depending on the type of MSPC model used. The contribution plot associated with a high Q value can be assessed by plotting the related $\mathbf{e}_{\mathbf{new}}$ vector for the sought observation. High absolute values related to elements in $\mathbf{e}_{\mathbf{new}}$ will identify variables showing abnormal behavior.

From the PCA model another statistical parameter can be obtained, the Hotelling's T^2, which represents the estimated Mahalanobis distance to the compressed subspace represented by the PCA model built with NOC observations.

The T^2 is calculated for any new observation using the predicted $\mathbf{t}_{\text{new}}$ and the following equation:

$$T^2 = \mathbf{t}_{\mathbf{new}}^{\text{T}}\boldsymbol{\Theta}^{-1}\mathbf{t}_{\mathbf{new}} \quad \textbf{(5)}$$

where $\boldsymbol{\Theta}$ is the PCA scores covariance matrix [37, 38].

T^2 can also be used to build MSPC control charts [7]. However, in this work, T^2 together with Q statistics were also used as means to represent compressed process information from purely NIR-based MSPC models in *Processes 1* and *2*, as mentioned in section 3.2.

Often, for an easier interpretation of MSPC control charts and indicators, reduced NIR-MSPC statistics ($Q_{red.}$ and $T^2_{red.}$) are calculated by dividing the obtained Q and T^2 values by their related 95 % confidence interval (CI) control limit; therefore, the control limit of derived charts becomes equal to one.

Software

Data handling and chemometric model building were carried out using own routines programmed in Matlab R2017a (Mathworks, USA) and PLS_Toolbox 8.2.1 (Eigenvector Research, USA) running under Matlab.

4. Results and discussion

The specific details of the data fusion strategies and the related DF-MSPC results are presented for each process application studied in this work. The aim of this section is showing the high performance of DF-MSPC models and how these models clearly improve the performance of models based on the sole use of NIR spectra ($MSPC_{NIR}$) both in terms of detecting process faults and identifying the causes of the abnormal process behaviour.

4.1. Process1: Fluidized Bed Drying Process

In this process, the DF-MSPC model combines temperature sensor readings of the fluidized material (T_{bed}), inlet (T_{in}) and outlet air (T_{out}) and information coming from two NIR-based multivariate models, a PLS regression model for prediction of moisture content and a $MSPC_{NIR}$ model for end-point detection, which provided $T^2_{red,NIR}$ and $Q_{red,NIR}$ process indicators. The PLS model was built using the off-line moisture content values as measured with the reference method and the related NIR spectra from the off-line pharmaceutical granules sampled. Meanwhile, an $MSPC_{NIR}$ end-point model was built with NIR spectra related to process observations obeying the moisture content specification at the end-point (below 2 %). Details of the drying process and the results describing the quality of moisture determination PLS models and $MSPC_{NIR}$ end-point detection model were discussed in a previous work [22].

To combine the temperature and the NIR sensor information, a data fusion strategy was implemented as shown in Fig. 1(b), including M_{PLS} %, $T^2_{red,NIR}$ and $Q_{red,NIR}$ and the three

temperature values T_{bed}, T_{in}, and T_{out}. After variable autoscaling, a DF-MSPC model for end-point detection using NOC batches was built.

The performance of the DF-MSPC model for end-point detection on new batches is shown using two validation batches, an on-specification batch (labelled batch 5) and an off-specification batch (batch 12). Fig. 3 shows the information to be submitted to the DF-MSPC model, i.e., T_{bed}, T_{in}, and T_{out} and M %, $T^2_{red,NIR}$ and $Q_{red,NIR}$ for validation batches 5 and 12 during the drying process.

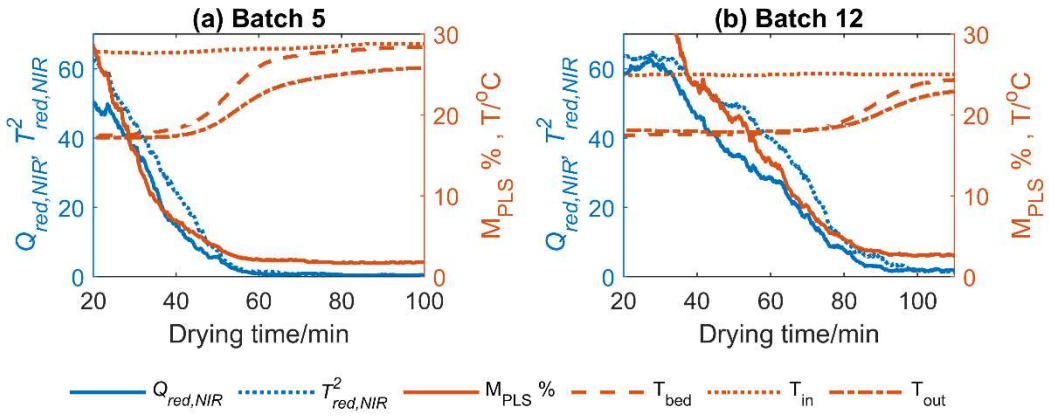

Fig. 3 Information to be analyzed by the DF-MSPC model for two validation batches a) Batch 5 (on-specification) and b) Batch 12 (off-specification). T_{bed}, T_{in} and T_{out} are the temperature sensors readings placed at granules, inlet and outlet air, respectively M_{PLS} % is the predicted moisture by PLS, $T^2_{red,NIR}$ and $Q_{red,NIR}$, are the process indicators obtained from $MSPC_{NIR}$ model . First 20 minutes are not shown because of unstable measurements at the beginning of the process

Fig. 3(a) shows the information submitted to DF-MSPC for validation batch 5, considered to be on-specification. At the end of the drying process, the predicted moisture level (M_{PLS}%) was found to be 1.8 %, the temperature readings were T_{bed}, 28.4 ºC, T_{in}, 28.8 ºC and T_{out}, 25.8 ºC, and the MSPC indicators $Q_{red,NIR.}$, 0.5, and $T^2_{red,NIR}$, 0.2, both below the control limit set equal to one, indicating that the end-point was detected and the batch was considered correct. Batch 12 in Fig. 3(b) is an example of off-specification batch. At the end of the drying process, the predicted moisture level was 2.7 %, the temperature readings T_{bed}, 24.4 ºC, T_{in}, 25.0 ºC and T_{out}, 22.9 ºC, respectively, and the MSPC indicators $Q_{red,NIR.}$, 1.9, and $T^2_{red,NIR.}$, 1.3, were both above the control limit set equal to one indicating that the batch did not reach the end-point and could be considered as off-specification. The main reason why the batch 12 did not reach the specification moisture level during the drying process time was because of the low inlet air temperature, about 4 ºC lower than in batch 5, which could be due to changes in the process environment or other uncontrolled causes.

The data fusion information from each validation batch, shown in Fig. 3, was submitted to the DF-MSPC model for end-point detection. To understand the better performance of the DF-MSPC model when compared with the $MSPC_{NIR}$ model, Figure 4 shows overlapped Q_{red} charts for both approaches. The blue line shows the evolution of $Q_{red,DF}$, derived from the DF-MSPC model, and the dashed orange line the evolution of $Q_{red,NIR}$, derived from the $MSPC_{NIR}$ model,

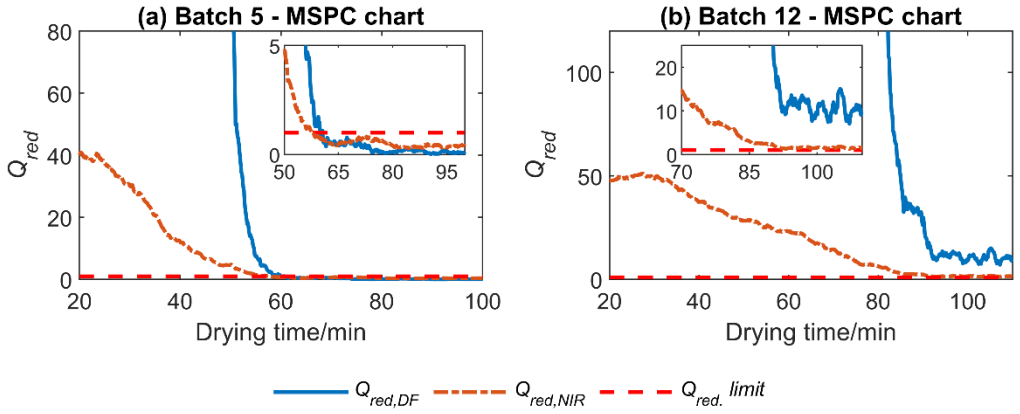

Fig. 4 Reduced Q (Q_{red}) MSPC charts for drying end-point detection using DF-MSPC model, $Q_{red,DF}$ (solid blue curve) and MSPC model based on the sole NIR information, $Q_{red,NIR}$ (dash-dotted orange curve) for validation batches a) Batch 5 (on-specification) and b) Batch 12 (off-specification). 95 % CI reduced Q control limit is represented by the dashed red flat line equal to one. Inset plots show a zoom of the last minutes at the end of each batch drying process

For batch 5, where moisture content reached the desired 2 % level, both data fusion and NIR-based control charts detected the end-point at approximately 60 min of drying time, as shown in Fig. 4(a). On the other hand, batch 12 did not reach the specified moisture level and both control charts did not detect the end-point during the entire batch duration, see Fig. 4(b). However, $Q_{red,DF.}$ values, coming from the DF-MSPC model, diagnose significantly better off-specification observations than $Q_{red,NIR.}$, obtained using only NIR spectral information. This is clearly noticed during the last minutes of batch 12, where $Q_{red,DF.}$ values are significantly higher and clearly further from the control limit than $Q_{red,NIR.}$ values, as a consequence of including the information from process temperature in the DF-MSPC models. Moreover, for both batch 5 and 12, the use of DF-MSPC models also provides a much clearer difference between the $Q_{red,DF.}$ values before and after moisture stabilization than the $Q_{red,NIR.}$ values, being the decrease of the $Q_{red,DF}$ curve always much steeper than that of $Q_{red,NIR.}$

In off-specification situations, it is also important to observe the contribution plots associated with the abnormal observations to understand the causes of the process malfunction. The contribution plot of observation at 100 min of drying batch 12 related to the $Q_{red,DF}$ DF-MSPC chart is shown in Fig. 5 (bar plot in orange). To compare with the residual level of an on-specification observation, the contribution plot related to the observation at 95 min of batch 5 is shown as well (bar plot in blue). It was observed that the main contributions to the high $Q_{red,DF.}$ values of batch 12 are the high $Q_{red,NIR.}$ and high M_{PLS} % prediction for moisture content. Indeed, the moisture content is higher than expected in on-specification values and, as a consequence, the residual related to the spectral shape expected at the end point (represented by $Q_{red,NIR.}$ from $MSPC_{NIR}$ model) is also higher. Fig. 5 also indicates that the temperature readings gave low contribution to the residuals, but the absolute value was higher than their contribution to batch 5 observation. The negative temperature contribution of batch 12 is in line of the fact that T_{in} at 100 min was lower than the expected temperature at the end-point stage of a NOC dry batch and, such a fact, caused the lower values for the related T_{bed} and T_{out}.

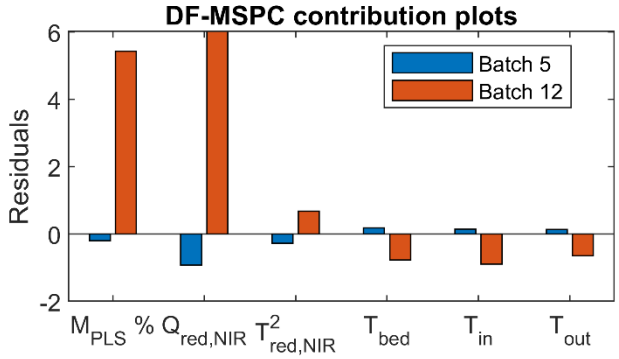

Fig. 5 Residual contribution plots related to the observations at 95 min for batch 5 and 100 min for batch 12 evaluated with the DF- MSPC model for drying end-point detection

In this example, both types of MSPC models have shown satisfactory performance for the detection of on- and off-specification situations. However, end-point control charts based on the DF-MSPC model provide a much clearer diagnostic of on- and off-specification situations and include all available process information, i.e. sensor and model outputs.

4.2. Process 2: Polyester production process

The singularity of this example is that, in this case, the data fusion concept is understood as the fusion of several model outputs coming from a single NIR sensor. In this process, the task of end-point detection should be applied to two different reaction stages and, therefore, two separate DF-MSPC models are built. The information for each DF-MSPC model comes

from two PLS models for prediction of acid number (AV_{PLS}) and viscosity (V_{PLS}) and one $MSPC_{NIR}$ model for end-point detection, providing the $T^2_{red,NIR}$ and $Q_{red,NIR}$ process indicators (see Fig. 1(a)). Acting in this way, key process parameters were used together with general unspecific NIR information linked to other physicochemical aspects of the end-point process stage. The two PLS models were built using the in-line NIR spectra related to off-line AV and V reference measurements from samples taken during the production process of calibration batches. Meanwhile, the NIR-based MSPC models were built with NIR spectra from the last 15 minutes measurements at the end-point from each process stage to define properly the process stage to be controlled. Details of the polymerization process and the results describing the quality of PLS models for the determination of AN and V and the use of $MSPC_{NIR}$ models for end-point detection for the two stages of the process were discussed in a previous work [23].

The DF-MSPC models linked to the end-point of the two reaction stages were built using information of NOC batches, as shown Fig. 1(a). Thus, outputs from PLS models (AV_{PLS} and V_{PLS}) and from the end-point detection $MSPC_{NIR}$ model, $T^2_{red,NIR}$ and $Q_{red,NIR}$, were combined into a data fusion matrix that was further autoscaled before model building.

The two DF-MSPC models related to end-point detection of each of the reaction stages were validated using real-time observations compressed in the same way as shown in Fig. 1 (a) for new batches. Fig. 6 shows the information from PLS and $MSPC_{NIR}$ models for three validation batches of the polymerization process (labelled batch 10, 11 and 13), used to show three different situations encountered in the batch polyester production process. Because of the presence of solids in the reaction, which caused spectra saturation, the information in Figure 6 is omitted for the first 12 to 15 hours of the first reaction stage and for approximately one hour between stage transitions.

Batch 10, shown in Fig. 6(a), represents a batch in which only the first stage of the process reached the end-point specification. The first stage ended at approximately 21 h of process time; at this point, AV_{PLS} and V_{PLS} were 9 mg KOH/g and 12 P, respectively, and $Q_{red,NIR}$, 0.4, and $T^{2,}_{red,NIR}$, 0.2, both below the control limit equal to one, where the batch was considered on-specification. On the other hand, the second stage was terminated at approximately 23.5 h, but did not meet the desired specifications. At this time AV_{PLS} and V_{PLS} were 62 mg KOH/g and 32 P, respectively, and $Q_{red,NIR}$ and $T^2_{red,NIR.}$, 10.2 and 0.6, respectively. Although the predictions of process key properties were within the targeted ranges and $T^2_{red.}$ below the control limit, the high $Q_{red.}$ value indicates the off-specification situation of the NIR

observations at the end of the second stage of batch 10, which was afterwards confirmed through off-line determinations of the end-product.

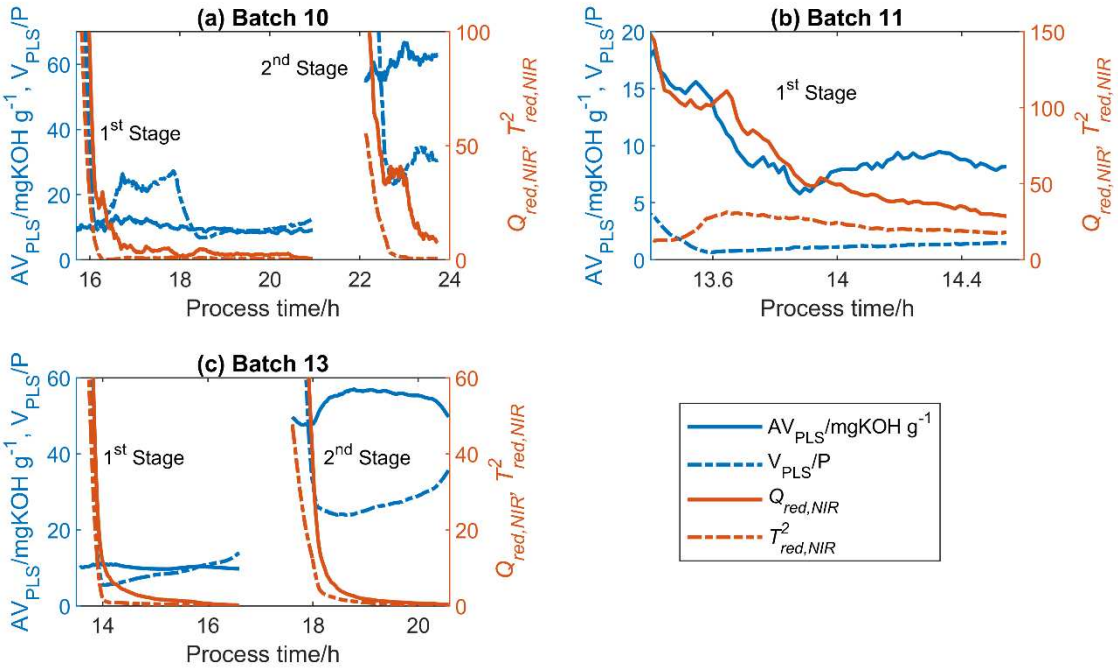

Fig. 6 Information to be analyzed by the DF-MSPC model from the polyester production process validation batches (a) Batch 10 (1st stage on-specification, 2nd stage off-specification), (b) Batch 11 (1st stage off-specification) and (c) Batch 13 (1st and 2nd stages on-specification). AV_{PLS} /mg KOH g^{-1} and V_{PLS} /P are represented by blue solid and dashed curves (left axis), respectively. $Q_{red,NIR}$ and $T^2_{red,NIR}$, are represented by orange solid and dashed curves, respectively (right axis)

Batch 11 was terminated before completing the first stage because of gel formation inside the reactor. The related data fusion information is shown in Fig. 6(b). At approximately 14.5 h of process time, AV_{PLS} 8 mg KOH/g and V_{PLS} 1 P, and $Q_{red,NIR.}$, 28.6 and $T^2_{red,NIR}$, 18.1, both above the control limit indicating that indeed batch 11 did not reach the end-point specification.

Batch 13 reached the end-point specifications for both process stages and the related information is shown in Fig. 6(c). The first stage ended at approximately 16.5 h of the process time; at this point AV_{PLS} was 10 mg KOH/g and V_{PLS}, 12 P. $Q_{red,NIR.}$, 0.2, and $T^2_{red,NIR.}$, 0.3, were both below the control limit indicating that the $MSPC_{NIR}$ model detected the process stage end-point and the batch was considered on-specification. The second stage was completed at approximately 20.5 h with AV_{PLS} and V_{PLS} of 50 mg KOH/g and 35 P, respectively, and $Q_{red,NIR.}$ and $T^2_{red,NIR}$, 0.3 and 0.1, respectively, which indicate that batch 13 was considered on-specification for both process stages.

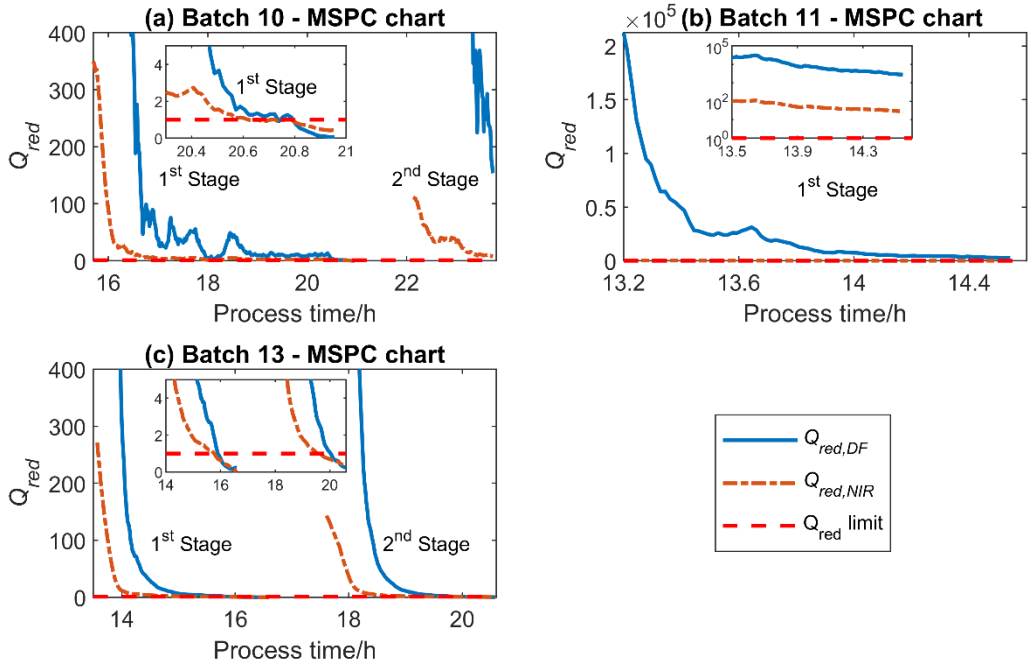

Fig. 7 Q_{red} MSPC charts for polymerization process stage end-point detection using DF-MSPC model (solid blue curve) and MSPC$_{NIR}$ model (dashed orange curve) for validation batches a) Batch 10 (1st stage on-specification, 2nd stage off-specification), b) Batch 11 (off-specification) and c) Batch 13 (1st and 2nd stages on-specification). 95 % CI Q_{red} control limit is represented by the dashed red flat line equal to one. Inset plots show a close view of the last minutes at the end of the 1st stage from batches 10 and 11 and the end of both stages for batch 13

The information shown in Fig. 6 for each validation batch was submitted to the related stage end-point detection DF-MSPC model. Fig. 7 shows overlapped Q_{red} charts for the DF-MSPC model (blue line) and the MSPC$_{NIR}$ model (orange dotted line).

Fig. 7(a) shows the MSPC control charts for batch 10, which met the end-point specifications only for the first stage, as indicated by the low $Q_{red} < 1$ values at approximately 21 h of process time in both DF-MSPC and MSPC$_{NIR}$ control charts. For the second stage, after 22 h of process time, Q_{red} values were above the control limit in both control charts, but clearly higher when using the DF-MSPC model as a consequence of including the explicit predictions of product quality parameters, i.e. AV and V, in the model. Batch 11 represents a faulty batch in the first process stage and no end-point was detected in batch 11 control charts shown in Fig. 7(b). $Q_{red,DF.}$ values obtained from DF-MSPC models were extremely higher and further from the control limit that $Q_{red,NIR.}$ values issued from MSPC$_{NIR}$ models, which can only be seen when the control charts are represented in a log-scale, as shown in the inset graph of Fig. 7(b). Conversely to batches 10 and 11, batch 13 was found to be a NOC batch and, therefore, both process stages reached the end-point specifications, see Fig. 7(c). Both DF-MPSC and MSPC$_{NIR}$

control charts detected the batch end-points at approximately 16 h and 20 h of process time for the first and second stages, respectively, see inset of Fig. 7(c).

Like in *Process 1*, end-point control charts based on the DF-MSPC model provide a much clearer diagnostic of on- and off-specification situations for both polymerization process stages. This conclusion seems to point out that using compressed and interpretable NIR information, such as key properties and process evolution indicators, seems to be more efficient than the mere use of direct NIR spectral information for process control.

4.3. Process 3: Gasoline Distillation

This example shows a different combination of temperature and NIR sensor model outputs for process evolution control and is the clearest example of combination of integral process modeling (as provided by the **C** profiles from MCR-ALS, see section 3.2.) and process control. In this case, NIR observations from NOC distillation batches were modelled with MCR-ALS, which provided distillation profiles (**C**) and NIR spectral signatures ($\mathbf{S}^T$) for the four different fractions distilled in the gasoline system studied. The four distillation profiles from the MCR-ALS **C** matrix were combined with the related boiling temperature measurements from each batch as shown in Fig. 1(c). The data fusion matrix from NOC batches was hereafter used to build local on-line batch DF-MSPC models based on the fixed sized moving window (FSMW) strategy described in [7].

Local on-line batch MSPC models were built as described in reference [7]. In this process, the information per batch observation is formed by five values, the four concentration values of the distilled fractions and the related distillation temperature. To control the process evolution, as many local DF-MSPC models as process observations are built, each one representing the variability allowed per each process observation by taking as initial information the observation at a particular distillation stage (% distilled mass fraction) and a window of 15 neighbouring observations.

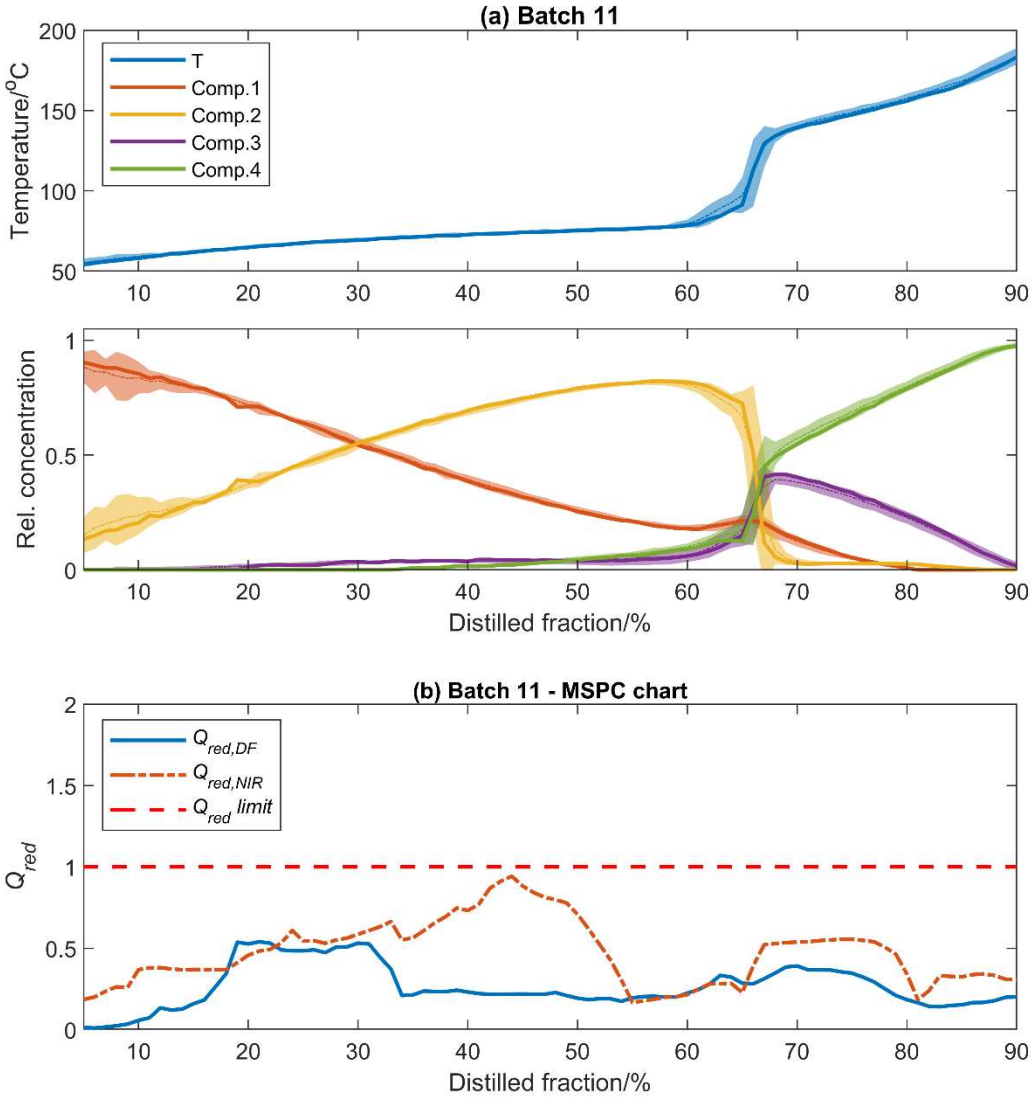

Fig. 8 Information related to gasoline distillation batch 11 (on-specification gasoline). a) Top plot: distillation temperature; bottom plot: concentration profiles of distilled fractions; batch 11 information (solid line curves), NOC batches information (thin dashed line curves are the average and the related ± 2 standard deviation bounds are the shaded area). b) $Q_{red,NIR}$ control charts from DF-MSPC models (solid blue curve) and from MSPC$_{NIR}$ model (orange dashed curve) models using the FSMW on-line MSPC strategy

Fig. 8(a) and Fig. 9(a) show the information (distillation curve and NIR-based MCR concentration profiles of distilled fractions) for the NOC batches used to build the local DF-MSPC models and for two validation batches submitted to the local DF-MSPC models for testing. In both Figures 8 (a) (showing an on-specification batch) and 9 (a) (showing an off-specification batch), the information from the validation batches is represented in solid lines. Instead, the behaviour of NOC distillation batches used for DF-MSPC model building is shown through a thin dashed line and a colour band surrounding it that represents the NOC average ± 2 standard deviation bounds.

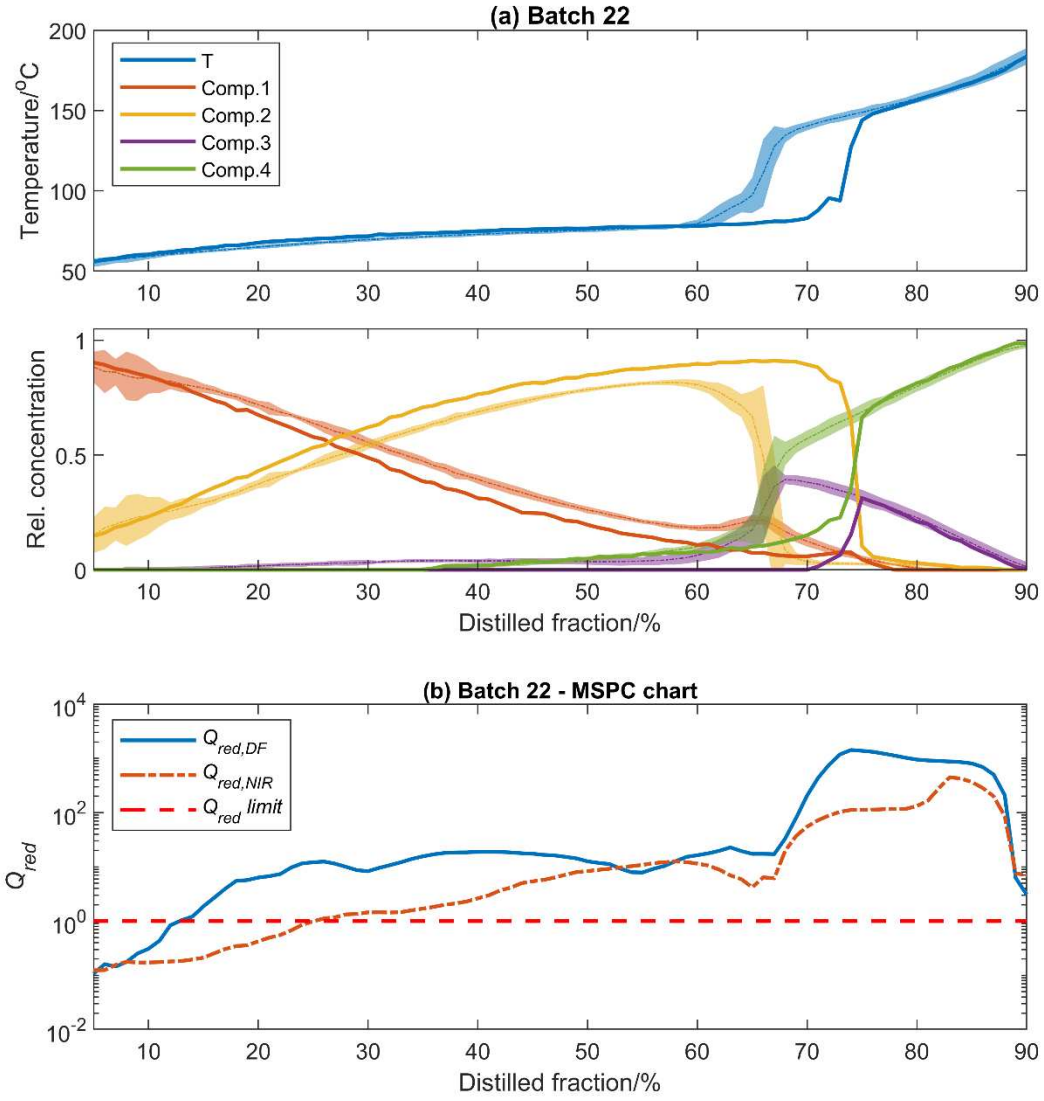

Fig. 9 Information related to gasoline distillation batch 22 (off-specification gasoline). a) Top plot: distillation temperature; bottom plot: concentration profiles of distilled fractions; batch 22 information (solid line curves), NOC batches information (thin dashed line curves are the average and the related ± 2 standard deviation bounds are the shaded area). b) $Q_{red,NIR}$ control charts from DF-MSPC models (solid blue curve) and from MSPC$_{NIR}$ model (orange dashed curve) models using the FSMW on-line MSPC strategy. y-axis is represented with log$_{10}$ scale

Batch 11 is a distillation batch from an on-specification gasoline, i.e. contains 27 % of ethanol. As depicted in Fig. 8(a), the output from the boiling temperature sensor and the distillation **C** profiles obtained from NIR spectra by MCR-ALS indicate that it follows the expected NOC behaviour. On the other hand, Figure 9 (a) shows the information from the off-specification batch 22, with 35 % of ethanol, which clearly shows the deviation from the NOC behaviour after 15 % distilled mass fraction.

The related data fusion information from each validation batch, shown in Fig. 8(a) and Fig. 9(a), was submitted to the related local DF- MSPC models. For comparison, Fig. 8(b) and Fig. 9(b) show the control charts obtained for each validation batch when using the local DF-MSPC models (blue line) or the $MSPC_{NIR}$ models (orange dotted line). Please note that the $Q_{red,DF}$ and $Q_{red,NIR}$ value associated with every observation in Figures 8(b) and 9(b) comes from a different local DF-MSPC or $MSPC_{NIR}$ model, respectively, as described in detail in [7]. In this example, the x-axis in the control charts refer to % distilled mass fraction and Q_{red} values need to be always below one to indicate that the process evolves correctly. One or more values above one indicate that, in those particular process stages, the new batch does not proceed as NOC batches.

Fig. 8(b) shows that all Q_{red} values related to process observations in batch 11 (on-specification gasoline), issued from DF-MSPC and $MSPC_{NIR}$ models, are below the control limit, which agrees with the results shown in Fig. 8(a). Moreover, as observed in the previous examples, the overall $Q_{red,DF}$ values are lower than $Q_{red,NIR}$ values, indicating more clearly the on-specification scenario. On the other hand, control charts for validation batch 22, which contains 35 % of ethanol, showed the deviation from the NOC batch distillation behaviour using $Q_{red,DF}$ or $Q_{red,NIR}$ values in Fig. 9(b). However, faulty observations were detected earlier when using the DF-MSPC models, which coincides with the deviation observed after 13 % from the NOC MCR-ALS distillation profiles in Fig. 9(a). Furthermore, $Q_{red,DF}$ values obtained when data fusion information was used were, in general, higher than $Q_{red,NIR}$ confirming the more efficient diagnostic for faulty observations of data fusion models over those using pure NIR information.

Although the detection of on- or off-specification distillation batch is clear in Figures 8 (a) and 9 (a) when looking at the distillation temperature plot and the concentration profile plot, control charts based on the DF-MSPC approach provide a much clearer diagnostic of on- and off-specification situations based on a multivariate statistical approach considering not only each individual variable but their interactions as well. Ref. [7] shows that off-specification situations in batches where the nominal concentration of faulty batches was much closer to the NOC composition were equally identified as non-acceptable. Like in the previous examples, using compressed and interpretable NIR information obtained from MCR-ALS decomposition combined with temperature sensor information is more efficient than the mere use of direct NIR spectral information for on-line batch process control.

In this particular scenario, the use of MCR results could have also been done selecting only some of the profiles, related to key fractions in the distillation process, to be submitted to the DF-MSPC model. This possibility is generalizable to any reaction process modelled by MCR, where not all concentration profiles would be necessarily included in the MSPC model, but only those related to critical components in the process evolution. The option of selecting a particular part of the NIR-based information would be completely impossible if non-compressed NIR spectra or PCA-based scores were taken as original information for MSPC models.

5. Conclusions

This work provides a relevant contribution to the current data fusion PAT strategies for the development of MSPC models combining all available process related information. In this sense, data fusion strategies extend to incorporate the combination of outputs from multivariate models coming from the same sensor or the combination of these model outputs with other process variable sensors. In this way, modelling and measurement tasks linked to the same process are considered altogether for the process control diagnostic. This concept has been illustrated with data fusion-based MSPC models for three different PAT applications.

Multivariate spectral information was compressed by means of multivariate models into process meaningful information such as key process quality parameters from PLS, concentration profiles from MCR-ALS or statistical process parameters according to each application. DF-MSPC models based on the different strategies were successfully validated for batch process end-point detection and on-line batch statistical process control. In all process examples, the data fusion methodologies have shown a high performance at detecting on- and off-specification batch situations and the model outputs used, much more interpretable than compressed abstract scores, were clearly helpful to identify the sources of process abnormalities.

The presented strategies could be extended to other industrial and bioprocess applications where dealing with several process outputs derived from multivariate and univariate sensors for building statistical process control are envisioned. The combination of modelling outputs used in the DF-MSPC models is very flexible and can be tailored according to the relevant information, e.g., key properties, evolution of one or more process compounds of the process under study. In this way, only the relevant information related to the multivariate

sensor measurement is used for process control and a more efficient and interpretable information on process performance is provided.

6. Acknowledgements

Funding from the European Community's Framework program for Research and Innovation Horizon 2020 (2014-2020) under grant agreement number 637232, related to the ProPAT project is acknowledged. Funding from Spanish government under the project CTQ2015-66254-C2-2-P is also acknowledged.

Conflict of Interest: The authors declare that they have no conflict of interest.

7. References

1. Gurden SP, Westerhuis JA, Smilde AK (2002) Monitoring of batch processes using spectroscopy. AIChE J 48:2283–2297
2. Gabrielsson J, Jonsson H, Trygg J, Airiau C, Schmidt B, Escott R (2006) Combining process and spectroscopic data to improve batch modeling. AIChE J 52:3164–3172
3. Huang J, Kaul G, Utz J, Hernandez P, Wong V, Bradley D, Nagi A, O'Grady D (2010) A PAT Approach to Improve Process Understanding of High Shear Wet Granulation Through In-Line Particle Measurement Using FBRM C35. J Pharm Sci 99:3205–3212
4. Jin Y, Wu Z, Liu X, Wu Y (2013) Near infrared spectroscopy in combination with chemometrics as a process analytical technology (PAT) tool for on-line quantitative monitoring of alcohol precipitation. J Pharm Biomed Anal 77:32–39
5. Lourenço ND, Lopes JA, Almeida CF, Sarraguça MC, Pinheiro HM (2012) Bioreactor monitoring with spectroscopy and chemometrics: A review. Anal Bioanal Chem 404:1211–1237
6. Zhao C, Gao F, Wang F (2010) Phase-based joint modeling and spectroscopy analysis for batch processes monitoring. Ind Eng Chem Res 49:669–681
7. de Oliveira RR, Pedroza RHP, Sousa AO, Lima KMG, de Juan A (2017) Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy. Anal Chim Acta 985:41–53
8. Catelani TA, Santos JR, Páscoa RNMJ, Pezza L, Pezza HR, Lopes JA (2018) Real-time monitoring of a coffee roasting process with near infrared spectroscopy using multivariate statistical analysis: A feasibility study. Talanta 179:292–299
9. Nomikos P, MacGregor JF (1994) Monitoring batch processes using multiway principal component analysis. AIChE J 40:1361–1375
10. Wold S, Kettaneh N, Friden H, Holmberg A (1998) Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments. Chemom Intell Lab Syst 44:331–340
11. Huang H, Qu H (2011) In-line monitoring of alcohol precipitation by near-infrared spectroscopy in conjunction with multivariate batch modeling. Anal Chim Acta 707:47–56

12. Huang J, Goolcharran C, Utz J, Hernandez-Abad P, Ghosh K, Nagi A (2010) A PAT approach to enhance process understanding of fluid bed granulation using in-line particle size characterization and multivariate analysis. J Pharm Innov 5:58–68
13. Mattila M, Saloheimo K, Koskinen K (2007) Improving the Robustness of Particle Size Analysis by Multivariate Statistical Process Control. Part Part Syst Charact 24:173–183
14. Faggian A, Facco P, Doplicher F, Bezzo F, Barolo M (2009) Multivariate statistical real-time monitoring of an industrial fed-batch process for the production of specialty chemicals. Chem Eng Res Des 87:325–334
15. Marjanovic O, Lennox B, Sandoz D, Smith K, Crofts M (2006) Real-time monitoring of an industrial batch process. Comput Chem Eng 30:1476–1481
16. Aguado D, Ferrer A, Ferrer J, Seco A (2007) Multivariate SPC of a sequencing batch reactor for wastewater treatment. Chemom Intell Lab Syst 85:82–93
17. González-Martínez JM, Ferrer A, Westerhuis JA (2011) Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping. Chemom Intell Lab Syst 105:195–206
18. Cimander C, Mandenius CF (2002) Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modelling. J Chem Technol Biotechnol 77:1157–1168
19. Cimander C, Carlsson M, Mandenius CF (2002) Sensor fusion for on-line monitoring of yoghurt fermentation. J Biotechnol 99:237–248
20. Jiang H, Chen Q (2014) Development of electronic nose and near infrared spectroscopy analysis techniques to monitor the critical time in SSF process of feed protein. Sensors (Switzerland) 14:19441–19456
21. Cocchi M (ed) (2019) Data Fusion Methodology and Applications. In: Data Handl. Sci. Technol. Elsevier Ltd, pp 1–370
22. Avila C, Ferré J, de Oliveira, Rodrigo Rocha de Juan A, Sinclair W, Mahdi F, Hassanpour A, Hunter TN, Bourne RA, Muller FL (2019) Process monitoring of moisture content and mass transfer rate in a fluidised bed with a low cost inline MEMS NIR sensor. Submitted
23. Avila C, Mantzaridis C, Ferré J, et al (2019) Monitoring the production of saturated polyester resins using novel MEMS FPI near infrared spectral sensor. Submitted
24. Zeaiter M, Rutledge D (2010) Preprocessing Methods. In: Compr. Chemom. pp 121–231
25. Savitzky A, Golay MJE (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem 36:1627–1639
26. Brereton RG (2000) Introduction to multivariate calibration in analytical chemistry. Analyst 125:2125–2154
27. Booksh KS, Kowalski BR (1994) Theory of Analytical Chemistry. Anal Chem 66:782A-791A
28. Martens H, Næs T (1991) Multivariate Calibration. John Wiley & Sons, New York
29. Thomas EV (1994) A Primer on Multivariate Calibration. Anal Chem 66:795A-804A
30. Haaland DM, Thomas EV (1988) Partial Least-Squares Methods for Spectral Analyses . 1 . Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. Anal Chem 60:1193–1202
31. de Juan A, Jaumot J, Tauler R (2014) Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. Anal Methods 6:4964–4976
32. de Juan A, Tauler R (2003) Chemometrics applied to unravel multicomponent processes and mixtures. Anal Chim Acta 500:195–210
33. Tauler R, Kowalski BR, Fleming S (1993) Multivariate Curve Resolution Applied to Spectral Data from Multiple Runs of an Industrial Process. Anal Chem 65:2040–2047

34. Tauler R, Maeder M, de Juan A (2009) Multiset Data Analysis: Extended Multivariate Curve Resolution. In: Compr. Chemom. Chem. Biochem. data Anal. four-volume set. Vol. 2, Chapter 2.24, S.D. Brown, R. Tauler, B. Walcz. Elsevier, pp 473–505
35. Kourti T (2009) Multivariate Statistical Process Control and Process Control, Using Latent Variables. In: Compr. Chemom. Elsevier, pp 21–54
36. Jackson JE, Mudholkar GS (1979) Control Procedures for Residuals Associated with Principal Component Analysis. Technometrics 21:341–349
37. MacGregor JF, Kourti T (1995) Statistical process control of multivariate processes. Control Eng Pract 3:403–414
38. Kourti T (2002) Process analysis and abnormal situation detection: From theory to practice. IEEE Control Syst Mag 22:10–25