



UNIVERSITAT DE  
BARCELONA

## Development and implementation of strategies for process data fusion, modelling and control

Rodrigo Rocha de Oliveira

**ADVERTIMENT.** La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX ([www.tdx.cat](http://www.tdx.cat)) i a través del Dipòsit Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

**ADVERTENCIA.** La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR ([www.tdx.cat](http://www.tdx.cat)) y a través del Repositorio Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING.** On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX ([www.tdx.cat](http://www.tdx.cat)) service and by the UB Digital Repository ([diposit.ub.edu](http://diposit.ub.edu)) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

# **DEVELOPMENT AND IMPLEMENTATION OF STRATEGIES FOR PROCESS DATA FUSION, MODELLING AND CONTROL**

**RODRIGO ROCHA DE OLIVEIRA**



**UNIVERSITAT DE  
BARCELONA**









UNIVERSITAT DE  
BARCELONA

FACULTAT DE QUÍMICA  
DEPARTAMENT D'ENGINYERIA QUÍMICA I QUÍMICA ANALÍTICA

Programa de doctorat: QUÍMICA ANALÍTICA i MEDI AMBIENT

**DEVELOPMENT  
AND IMPLEMENTATION OF STRATEGIES FOR PROCESS  
DATA FUSION, MODELLING AND CONTROL**

Memòria presentada per

**Rodrigo Rocha de Oliveira**

Per optar al grau de Doctor per la Universitat de Barcelona

**Directora**

**Dra. Anna de Juan Capdevila**

Departament d'Enginyeria Química i Química Analítica

Universitat de Barcelona



A Dalva, Euca e Anna,

“Todos nós sabemos alguma coisa. Todos nós ignoramos  
alguma coisa. Por isso, aprendemos sempre.”

— Paulo Freire

“ – bem feito >>> Perfeito”





## ACKNOWLEDGEMENTS

It arrives at the end of a very important cycle of my life. An unforgettable cycle, not only because of what I have accomplished, but also because of the people who have been part of it. That is why I will try to express my gratitude on this piece of paper. Hard task, sure it will be incomplete, in all aspects.

Primeiro quero agradecer a minha família. Aos meus pais, Dalva e Raimundo, que sempre demonstraram fé, carinho e o apoio para voar bem longe do ninho que jamais esquecerei. Às minhas irmãs, Rosiene e Rosineide, e irmão, Romildo, que também me deram todo suporte e confiança. A Mateus e Maria também quero agradecer o carinho dado ao tio mesmo a distância. Especialmente a minha esposa. Eucástila, meu bem, minha *pareia*, obrigado pela paciência, companheirismo, colaboração, carinho e muito amor. Admiro muito tua coragem e *espirituosidade* que me deram muita força nesse caminho.

Agradecer aos meus amigos Potiguares, Ricardo e Thiago, que assim como eu seguem se aventurando longe de suas casas em busca do futuro que sonham. Também ao velho amigo Shell e seus eventos “organizados”. À equipe Hempense por entender minha ausência nesse empreendimento que irá decolar.

Gracias por vuestra amistad, Neus y Javier, por los buenos momentos de risas y quedadas gastronómicas.

A mis *hermanes científicos*: Sara, Silvia M, Sanae y Victor. Gracias por acogerme en el grupo de quimiometría desde cuando era *peque*. Gracias también a Raimundo, por siempre mantener vivo mi espíritu de ingeniero.

A Carmen, Nuria, Iñigo, Andreu de la secretaria del departament i tot el personal de la facultat de química, FBiG i de la UB en general que de distintes formes van contribuir amb la evolució d’aquesta tesi.

No podría dejar olvidado el soporte se los TFGs y TFMs que han contribuido de manera importante a este trabajo. Gracias, Carla, Julián, Laia y Miriam. – También aprendí mucho con vosotros.

A Adri, Ale, Dario, David i Iker por todos los momentos, por las bravas, la pasta amatriciana con aceite de coco y los brócolis *basilicosos*, – hummmm, también por mantener mis plantas vivas, por los jueves estrambóticos y los nuevos conceptos de liquidez. Por supuesto a Alberto y sus birras gratis, a Juan y Ana y sus caramelos contrabandeados, tampoco serán olvidados.

---

To all international students for the moments of knowledge, culture and friendship exchange. Special thanks to Andrés, Betta, Kudi, Lorenzo, Inal, Petra, Sara Z, Silvia C and Siewert.

Gracias Nerea, José, Maria por los buenos momentos de charla y comida.

My sincere thanks also go to Marina for hosting me at UNIMORE with open arms and all exchanged knowledge and support during those first pandemic days. To my colleagues that I have met in Modena and shared the first Covid days: Alec, Lena and Mohamad, thanks for the good moments spent together. Grazie Mille!

I have also to acknowledge the EU for funding the ProPAT project. Thanks to all ProPAT partners for the collaboration.

Finally, I have to admit that no words can describe how thankful I am for having Anna as my PhD supervisor. Intentaré descriure amb el meu propi català, – sense *translator*. Gràcies Anna, per sempre confiar en el meu potencial i estimular les meves capacitats científiques. El teu pragmatisme em va ajudar a avançar contínuament. T'admiro molt per ser qui ets. He après molt amb tu, no només al àmbit científic. Gràcies per tots els moments que hem passat junts, congressos, excursions, dinars, calçotada, experiments nocturns al synchrotron ... També per confirmar els senyals de que estic fent més gran i ja no soc el petit Rodrigo que va arribar fa casi deu anys. Gràcies per ajudar-me a créixer com persona i com científic. No hauria arribat fins aquí sense tu.

Thank you! This accomplishment would not have been possible without all of you.

Barcelona, November 2021

Rodrigo Rocha de Oliveira

# CONTENTS

CONTENTS.....	i
ABSTRACT.....	iii
RESUM.....	v
CHAPTER 1. OBJECTIVES AND STRUCTURE OF THE THESIS.....	1
1.1 Objectives .....	3
1.2 Structure of the thesis .....	4
1.3 List of scientific publications presented in this work.....	5
CHAPTER 2. GENERAL INTRODUCTION .....	7
2.1 Industry 4.0. Process Analytical Technology .....	9
2.2 Analytical tools for process monitoring. Sensor typology and sensor data .....	12
2.3 Chemometric tools for process monitoring, modelling and control .....	15
2.3.1 Principal Component Analysis (PCA).....	15
2.3.2 Partial Least Squares Regression (PLS).....	17
2.3.3 Multivariate Statistical Process Control (MSPC) .....	19
2.3.4 Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) .....	21
CHAPTER 3. PROCESSES STUDIED. EXPERIMENTAL AND DATA PREPROCESSING.....	27
3.1 Processes monitored with spectroscopic probes and process sensors .....	29
3.1.1 Multistep polyester production.....	29
3.1.2 Fluidized bed drying of pharmaceutical granules .....	31
3.1.3 Benchtop batch distillation of gasoline and ethanol blends .....	33
3.2 Process monitoring using hyperspectral imaging (HSI) .....	35
3.2.1 Blending process monitoring with atline NIR-HSI.....	37
3.2.2 Blending process monitoring with inline NIR-HSI .....	38
CHAPTER 4. RESULTS AND DISCUSSION .....	41
SECTION I – Process monitoring, modeling and control using spectroscopic probes and process sensors.....	43
4.1 Data configurations. Synchronized and non-synchronized multibatch data.....	45
4.2 Process modelling and control for synchronized batch processes .....	47
4.3 Process modeling and control for non-synchronized batch processes .....	73
4.3.1 MSPC models for endpoint detection. Data fusion strategies.....	73
4.3.2 Online synchronization-free MSPC for batch process evolution assessment .....	129
SECTION II – Process monitoring using hyperspectral imaging.....	161
4.4 Blending process monitoring and heterogeneity assessment .....	163
4.4.1 Methodology to assess heterogeneity from HSI.....	191
4.4.2 Use of heterogeneity indices for blending process understanding .....	202
CONCLUSIONS .....	209
REFERENCES .....	215



# ABSTRACT

With the emergence of Industry 4.0 and the increasing availability of sensors and data acquisition systems, modern manufacturing processes are now generating large amounts of process data on a scale as never seen before. During the past few decades, the intense development of powerful data-driven methodologies for process analytics has demonstrated the importance of multivariate data analysis for this field. Still, new strategies inspired by current methodologies and yet to be developed will continuously be required to tackle new challenges posed by the digital revolution in process analytics.

This thesis has been focused on the development and application of chemometric tools for process analytical technology (PAT) and includes approaches for process monitoring, modeling and control of batch processes. All the methodology proposed has been tested on real batch processes of diverse nature monitored with sensor of different typology. The chemometric tools developed in this thesis are meant to be used in two different contexts: a) process monitoring, modeling, and control using spectroscopic probes and process sensors, and b) process monitoring using hyperspectral images.

In the context of process monitoring using spectroscopic probes and process sensors, different methodologies have been designed to handle information coming from synchronized and non-synchronized batch process data. For synchronized batch process data, new strategies for offline and online Multivariate Statistical Process Control (MSPC) have been designed. Offline MSPC models, meant to control complete batches, were built based on information coming from original sensor variables or from compressed spectral information, issued from multivariate exploratory and resolution analysis outputs. Online process control methodologies were based on the use of local MSPC models built exploring the effect of different designs of process time windows onto the capacity to discriminate between observations following normal operation conditions (NOC) and showing an abnormal behavior. For non-synchronized batch data, a novel batch synchronization-free online MSPC methodology for tracking process evolution and control was proposed based on the idea of a global batch process trajectory and the use of local MSPC models.

A clear improvement of the results linked to all MSPC scenarios is linked to the use of new mid-level data fusion strategies. The novel contribution in this thesis is the extension of the idea of data fusion to incorporate both diverse sensor outputs and diverse model outputs issued from the same sensor, but related to different modeling tasks. These model outputs, which are much more specific than mere compressed scores, help significantly to tune the information introduced in the MSPC models and to a better interpretation of the sources of abnormal process behavior.

The chemometric solutions proposed for process monitoring using hyperspectral images (HSI) were mainly oriented to take advantage of the spatial information of the measurement for the qualitative and quantitative heterogeneity assessment in

blending processes. The qualitative description of heterogeneity is linked to HSI unmixing analysis, which provides pure component distribution maps that offer a good visual representation of the evenness in the spatial distribution of the different materials in the blending formulation. The quantitative characterization of heterogeneity is obtained from the variographic analysis of the distribution maps and results in two indices: the Global Heterogeneity Index (GHI), related to the scatter of the individual pixel concentration values, and the Distributional Uniformity Index (DUI), describing the distributional heterogeneity, usually overlooked in traditional approaches, that expresses the evenness in the spatial distribution of the different materials forming a blend. These indices have been proven to be a powerful process analytical tool to characterize the heterogeneity in blending processes monitored atline and inline with NIR-HSI. For image-based inline process monitoring, an extension of this methodology, called SWiVIA (Sliding Window Variographic Image Analysis), has been adapted for the continuous assessment of heterogeneity in real-time blending process monitoring. The versatility of the SWiVIA methodology enables heterogeneity assessment at the time resolution and spatial scale of scrutiny required for the blending application of interest.

# RESUM

Amb l'arribada de la Indústria 4.0 i la creixent disponibilitat de sensors i sistemes d'adquisició de dades, els processos de fabricació moderns generen quantitats ingents de dades de procés a una escala mai vista. Durant les últimes dècades, el desenvolupament continuat de metodologies d'anàlisi de processos basades en la interpretació directa de la mesura ha confirmat la importància de l'anàlisi multivariant de dades en aquest camp. Tot i així, caldrà desenvolupar noves aproximacions inspirades en metodologies existents o encara per descobrir per afrontar els nous reptes que planteja la revolució digital en l'anàlisi de processos.

Aquesta tesi s'ha centrat en el desenvolupament i aplicació d'eines quimiomètriques lligades a la tecnologia analítica de processos (PAT) per al seguiment, modelització i control de processos per lots. Tota la metodologia proposada ha estat provada en processos reals de diversa naturalesa monitorats amb sensors de diferents tipologies. Les eines quimiomètriques desenvolupades en aquesta tesi estan pensades per ser utilitzades en dos contextos diferents: a) el seguiment, modelització i control de processos mitjançant sondes espectroscòpiques i sensors de procés, i b) el seguiment de processos mitjançant imatges hiperespectrals.

En el context del monitoratge de processos mitjançant sondes espectroscòpiques i sensors de procés, s'han dissenyat diferents metodologies per gestionar la informació procedent de dades de procés per lots sincronitzats i no sincronitzats. Per a dades de lots sincronitzats, s'han dissenyat noves estratègies per al control estadístic multivariant de processos (MSPC, Multivariate Statistical Process Control) *offline* i *online*. Els models MSPC *offline*, destinats a controlar lots complets, es van construir a partir d'informació associada a variables originals de sensors o d'informació espectral comprimida, procedent de resultats de models d'anàlisi exploratòria i de resolució multivariant. Les metodologies de control de processos *online* es van basar en l'ús de models locals de MSPC construïts explorant l'efecte de diferents dissenys de finestres de temps de procés sobre la capacitat de discriminar observacions seguint condicions normals d'operació (NOC, *Normal Operation Conditions*) d'observacions amb un comportament anòmal. Per a les dades de lots no sincronitzats, es va proposar una nova metodologia MSPC *online* exempta de l'etapa de sincronització per fer un seguiment de l'evolució i el control del procés basada en l'ús d'una trajectòria global del procés per lots, que serveix per a la construcció de models locals de MSPC.

Una millora clara dels resultats associada a tots els escenaris de models MSPC està vinculada a l'ús de noves estratègies de fusió de dades de nivell intermedi (*mid-level data fusion*). La nova contribució d'aquesta tesi és l'extensió de la idea de fusió de dades a la incorporació tant de respostes de sensors diversos com de resultats de models multivariants obtinguts de respostes d'un mateix sensor, però relacionats amb



diferents tasques de modelització. Aquests resultats de models multivariants, que aporten informació molt més específica que els scores de PCA, per exemple, permeten una tria més acurada de la informació que s'introdueix en els models MSPC i faciliten una millor interpretació de les causes de comportaments anòmals en el procés.

Les solucions quimiomètriques proposades per al seguiment de processos mitjançant imatges hiperespectrals (*HSI, Hyperspectral Images*) es van orientar principalment a aprofitar la informació espacial de la mesura per a l'avaluació qualitativa i quantitativa de l'heterogeneïtat en els processos de mescla. La descripció qualitativa de l'heterogeneïtat està vinculada al resultat de l'anàlisi de resolució multivariant de les dades HSI, que proporciona mapes de distribució de components purs que ofereixen una bona representació visual de la uniformitat en la distribució espacial dels diferents materials en la mescla estudiada. La caracterització quantitativa de l'heterogeneïtat s'obté de l'anàlisi variogràfica dels mapes de distribució i està basada en dos índexs: l'índex d'heterogeneïtat global (*GHI, Global Heterogeneity Index*), relacionat amb la dispersió dels valors de concentració dels píxels individuals, i l'índex d'uniformitat distribucional (*DUI, Distributional Uniformity Index*), que descriu l'heterogeneïtat distribucional, normalment ignorada en plantejaments tradicionals, que expressa el grau d'uniformitat en la distribució espacial dels diferents materials que formen una mescla. S'ha demostrat que aquests índexs són una eina PAT potent per caracteritzar l'heterogeneïtat dels processos de mescla seguits amb mesures discretes o en temps real mitjançant imatgeria hiperespectral d'infraroig proper (*NIR-HSI*). Per al seguiment de processos en temps real basat en imatges, s'ha adaptat una extensió d'aquesta metodologia, anomenada *SWiVIA (Sliding Window Variographic Image Analysis – Anàlisi variogràfica d'imatges basada en finestres mòbils)*, per a l'avaluació en temps real de l'heterogeneïtat en el seguiment continu de processos. La versatilitat de la metodologia *SWiVIA* permet l'avaluació de l'heterogeneïtat amb la resolució temporal i l'escala espacial d'escrutini desitjada segons les característiques del procés de mescla estudiat.

# **CHAPTER 1. OBJECTIVES AND STRUCTURE OF THE THESIS**



## 1.1 Objectives

The interest of the industry sector towards the implementation of process analytical technologies (PAT) in their manufacturing processes has been continuously increasing during the last decades. Although great progress has been made in the development and integration of many PAT tools into production processes, there are still many challenges to handle adequately the increasing volume of process data so that the final products can meet the quality requirements from consumers and regulatory agencies.

The central goal of this thesis is proposing advanced PAT tools to improve industrial process understanding, monitoring and control. The main focus will be on the development of new multivariate analysis methodologies for process monitoring and Multivariate Statistical Process Control (MSPC). The methodologies proposed will be addressed to improve: a) process monitoring, modeling and control using spectroscopic probes and process sensors; and b) process monitoring using hyperspectral imaging.

### **Process monitoring, modeling and control using spectroscopic probes and process sensors**

The specific objectives linked to this research area are:

- The use of process modeling tools, such as Principal Component Analysis (PCA) and Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) for process understanding and to obtain compressed and interpretable seeding inputs for future process control models.
- The design or improvement of methodologies for offline and online MSPC for synchronized batch data. Improvement of offline process control methodologies will be based on the adequate selection of the input information used to build the MSPC models. New and flexible online MSPC methodologies based on the use of local models employing different designs of process time windows will be proposed.
- The design a novel synchronization-free process control methodology for online MSPC of non-synchronized batch processes. This approach will allow tracking process evolution by using local MSPC models covering a global NOC process trajectory, obtained from the overlap of desynchronized individual NOC batch trajectories.
- The proposal of new mid-level data fusion strategies incorporating diverse sensor outputs and/or several multivariate model outputs issued from the same sensor, but related to different modeling tasks. This approach will improve the process

control performance and will help in the interpretation of the causes of process upsets.

### **Process monitoring using hyperspectral imaging**

The specific objectives linked to this research area are:

- The use of the spatial information from hyperspectral images (HSI) to obtain a qualitative and quantitative characterization of the heterogeneity in blending formulations. HSI unmixing analysis, specifically based on the Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) algorithm, will be used to obtain distribution maps of the compounds in the blend. These maps will provide a visual representation of the spatial distribution of the different materials in a blending formulation and, hence, a qualitative heterogeneity description. The maps will be subsequently used for the quantitative characterization of the heterogeneity based on variographic analysis.
- The design of quantitative heterogeneity indices issued from the variographic analysis of distribution maps. A Global Heterogeneity Index (GHI), related to the scatter of the individual pixel concentration values, and a Distributional Uniformity Index (DUI), linked to the distributional heterogeneity expressed as the evenness in the spatial distribution of the different materials forming a blend, will be proposed.
- The use of heterogeneity indices to characterize the heterogeneity in blending processes monitored atline and inline. For real-time continuous assessment of heterogeneity in blending processes monitored with pushbroom HSI systems, an adapted methodology, called SWiVIA (Sliding Window Variographic Image Analysis) will be proposed. The methodology will be designed so that the blending time resolution and the spatial scale of scrutiny of the blend will be tuned according to the characteristics of the process of interest.

## **1.2 Structure of the thesis**

This thesis describes the research developed and collected in seven publications around the topic of development and application of multivariate approaches for process monitoring, modeling and control. The manuscript is structured in six chapters that include an introductory part and the results and discussion of the research carried out.

In the first chapter, the objectives and structure of this thesis are presented. Additionally, the scientific publications derived from this work are listed. In the second chapter, a general introduction about the current context of the work related to process analytical technologies in modern process manufacturing is given. A brief definition of analytical tools, sensors and data typologies for process monitoring is presented. The last part is devoted to describing the basics of the chemometric tools used to develop

the new strategies for process monitoring, modeling, and control. The third chapter is divided into two parts related to the two main categories of processes studied in this work: a) processes monitored with spectroscopic probes and b) processes monitored with hyperspectral images. For each process case study, a description of the experimental setup, materials, sensors and data generated is provided. The raw data preprocessing steps are also included.

In the fourth chapter, a detailed presentation of the results obtained from the publications in this Thesis is provided. This chapter is divided into two main sections related to the two groups of processes presented in Chapter 3. In Section I, the different data configurations used to organize synchronized and non-synchronized batch data are presented together with the related discussion of the results for process modeling, monitoring, and control. Special attention is provided to the results related to the implementation of new methodologies for process data fusion and batch synchronization-free online process control. In Section II, the results related to processes monitored with hyperspectral images (HSI) are presented. First, the new methodology to assess qualitative and quantitative information related to different aspects of heterogeneity in chemical images is introduced. To conclude, the results related to the application of this approach to the atline and inline monitoring of real blending processes are discussed.

In the fifth chapter, the main conclusions resulting from the present work are presented. Finally, in the sixth and final chapter, the list of references to the literature cited in this Thesis is provided.

### **1.3 List of scientific publications presented in this work**

The work performed in this thesis resulted in the seven scientific publications below, grouped by topics and following the sequence in the thesis manuscript.

#### **Publication I. Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy.**

Authors: de Oliveira, R. R., Pedroza, R. H. P., Sousa, A. O., Lima, K. M. G., and de Juan, A.

Citation reference: *Analytica Chimica Acta* (2017), 985: 41–53.

DOI: [10.1016/j.aca.2017.07.038](https://doi.org/10.1016/j.aca.2017.07.038)

#### **Publication II. Process Monitoring of Moisture Content and Mass Transfer Rate in a Fluidised Bed with a Low Cost Inline MEMS NIR Sensor.**

Authors: Avila, C. R., Ferré, J., de Oliveira, R. R., de Juan, A., Sinclair, W. E., Mahdi, F. M., Hassanpour, A., Hunter, T. N., Bourne, and R. A., Muller, F. L.

Citation reference: *Pharmaceutical Research* (2020), 37: 84.

DOI: [10.1007/s11095-020-02787-y](https://doi.org/10.1007/s11095-020-02787-y)

**Publication III. Acid number, viscosity and end-point detection in a multiphase high temperature polymerisation process using an online miniaturised MEMS Fabry-Pérot interferometer.**

Authors: Avila, C., Mantzaridis, C., Ferré, J., Rocha de Oliveira, R., Kantojärvi, U., Rissanen, A., Krassa, P., de Juan, A., Muller, F. L., Hunter, T. N., Bourne, R. A.

Citation reference: *Talanta* (2021), 224: 121735.

DOI: [10.1016/j.talanta.2020.121735](https://doi.org/10.1016/j.talanta.2020.121735)

**Publication IV. Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control.**

Authors: de Oliveira, R. R., Avila, C., Bourne, R., Muller, F., and de Juan, A.

Citation reference: *Analytical and Bioanalytical Chemistry* (2020), 412:2151–2163.

DOI: [10.1007/s00216-020-02404-2](https://doi.org/10.1007/s00216-020-02404-2)

**Publication V. Synchronization-Free Multivariate Statistical Process Control for Online Monitoring of Batch Process Evolution.**

Authors: Rocha de Oliveira, R. and de Juan, A.

Citation reference: *Submitted to Frontiers in Analytical Science; Specialty section: Chemometrics; Research Topic: Novel Applications of Chemometrics in Analytical Chemistry and Chemical Process Industry.*

**Publication VI. Design of Heterogeneity Indices for Blending Quality Assessment Based on Hyperspectral Images and Variographic Analysis.**

Authors: Rocha de Oliveira, R. and de Juan, A.

Citation reference: *Analytical Chemistry* (2020), 92: 15880–15889.

DOI: [10.1021/acs.analchem.0c03241](https://doi.org/10.1021/acs.analchem.0c03241)

**Publication VII. SWiVIA – Sliding window variographic image analysis for real-time assessment of heterogeneity indices in blending processes monitored with hyperspectral imaging.**

Authors: Rocha de Oliveira, R. and de Juan, A.

Citation reference: *Analytica Chimica Acta* (2021), 1180: 338852.

DOI: [10.1016/j.aca.2021.338852](https://doi.org/10.1016/j.aca.2021.338852)

## **CHAPTER 2. GENERAL INTRODUCTION**





## 2.1 Industry 4.0. Process Analytical Technology

The lifestyle and development of mankind have been significantly marked by the industrial revolution. Nowadays, four different periods are often differentiated that relate to crucial milestones related to the industrial progress over more than 200 years. Thus, the first industrial revolution happened during the second half of the 18<sup>th</sup> century and the advent of the use of water- and steam-powered mechanical manufacturing systems was a breakthrough in the production carried out in the first large factories. About 100 years later, the second industrial revolution began when mass production assembly lines and electrical energy arrived, resulting in much higher production efficiency. The third industrial revolution, in the 1970s, brought advanced electronics, which enabled digital programming for the automation of production processes and important development of communication technologies, e.g. internet and wireless communications. The fourth industrial revolution, known as Industry 4.0, will integrate current and yet to come technologies such as the Internet of Things, artificial intelligence, machine learning, robotics and advanced computing to dramatically change the landscape of manufacturing.

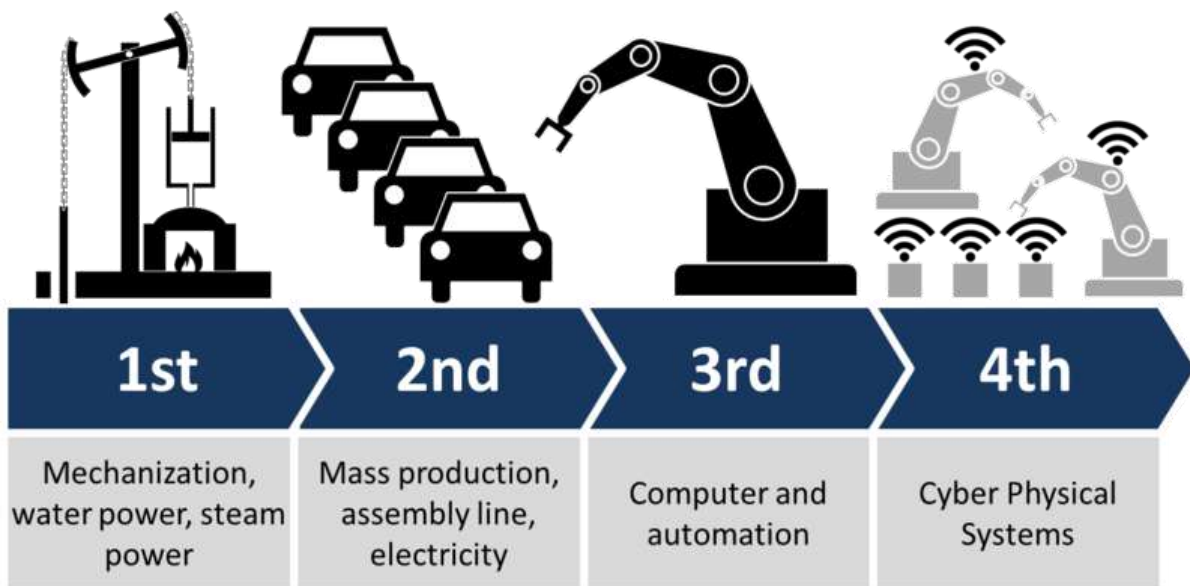


Figure 1 Industrial revolutions in the history of humankind. Reproduced from (Roser, 2015) at AllAboutLean.com.

The term “Industry 4.0” was coined in 2011 by a German initiative named *Industrie 4.0* (Drath and Horch, 2014; Hermann et al., 2016). This term was rapidly adopted by the federal government, which announced that Industry 4.0 would be one of the key initiatives of its “High-Tech Strategy 2020 for Germany”. Later, in 2013, the “*Industrie 4.0 Working Group*” published a report naming three key components for its implementation: the internet of things (IoT), Cyberphysical Systems (CPS), and Smart Factories (*Final report of the Industrie 4.0 Working Group*, 2013). Although it is advocated that the third industrial revolution, “the Digital Revolution”, has not yet reached its full potential (Rifkin, 2016), the term Industry 4.0 was born referring to the

next industrial revolution, the fourth industrial revolution (Figure 1), which is about to take place right now (Drath and Horch, 2014; Hermann et al., 2016).

It is claimed that the Industry 4.0 will integrate rapidly evolving technologies, such as the IoT, artificial intelligence (AI), robotics, and advanced computing to dramatically change the landscape of manufacturing. This will improve current industrial workflows and will create new advanced manufacturing technologies enabling autonomous, and self-organizing manufacturing systems that operate independently of human intervention (Arden et al., 2021). In other words, Industry 4.0 is the automation of conventional manufacturing and industrial processes using modern (current and yet to be developed) smart technologies. Such technologies will be integrated for increased automation, improved communication and self-monitoring due to the ability to analyze and detect issues without the need for human involvement. Other related concepts bring the same idea under the name of “Industrial Internet”, “Integrated” or “Smart Industry” as well as “Smart or Advanced Manufacturing” (Hermann et al., 2016).

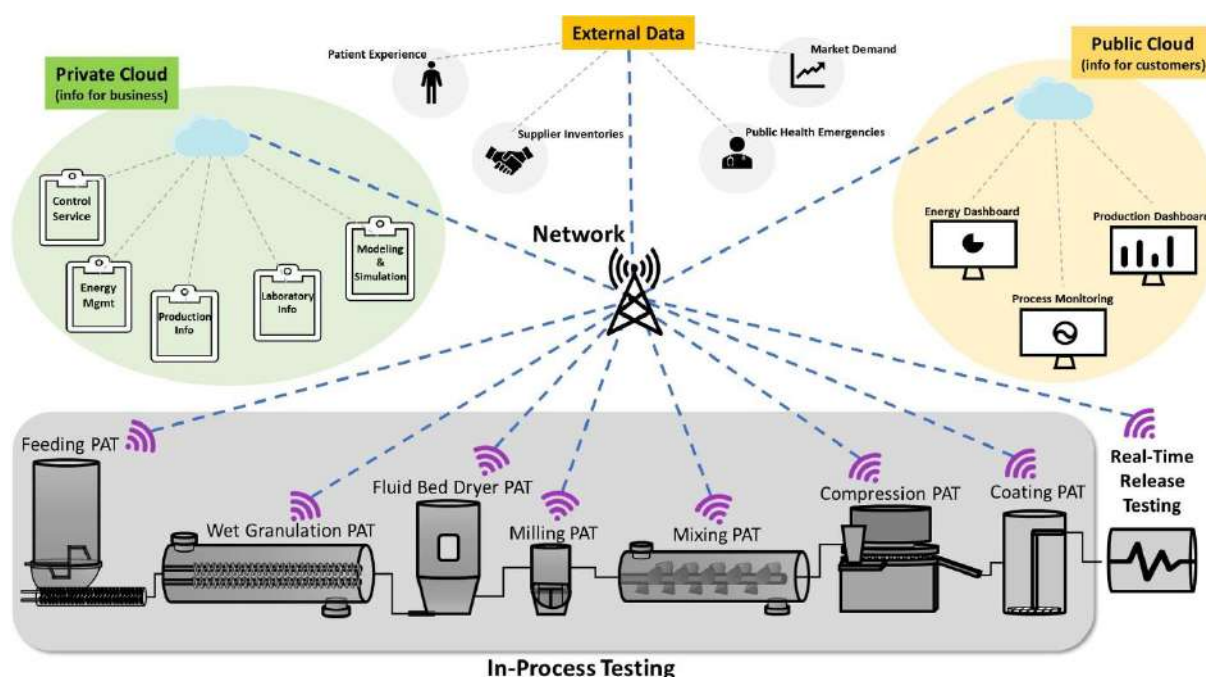


Figure 2 Representation of how a cyber-physical system (CPS) for pharmaceutical manufacturing will look like in Industry 4.0. The key parts of a CPS include the public-cloud, private-cloud, and manufacturing floor (in gray). Reproduced from (Arden et al., 2021) under the terms of (CC BY NC ND) license.

One of the most promising elements in Industry 4.0 are the cyber-physical systems (CPS). They are defined as “*the integration of computation and physical processes. Embedded computers and networks will serve to monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa*” (Lee, 2008). The idea is to digitally connect machines and also process units, e.g. process reactors. As a consequence, information collected about the current process status, e.g. sensor measurements and process metadata, should be available

much more readily, and reliable process understanding will be derived from the large quantities of data provided for analysis. This fusion of the physical and the virtual world has been illustrated in a recent work (Arden et al., 2021) for pharmaceutical manufacturing in Industry 4.0, as shown in Figure 2.

In the pharmaceutical context, the scheme shows in the top part the use of information, from external inputs coming from the clinical experience and definition of the needed product to economic production and market facts. The right and left cloud design output information, always related to the actual process taking place, and having a bidirectional communication. The gray rectangle at the bottom in Figure 2 highlights the in-process testing of the pharmaceutical manufacturing floor in Industry 4.0. This includes a series of continuous operation processes, e.g., feeding, wet granulation, fluid bed drying, milling, blending, compression, and tablet coating. Moreover, note that together with the operation name, the “PAT” acronym is recurrent.

PAT stands for Process Analytical Technology and it is a technology already introduced during the Industry 3.0 to enhance understanding and to control the manufacturing process (Arden et al., 2021; FDA, 2004). In 2004, the United States Food and Drug Administration (FDA) published a guidance document defining PAT as *“a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality”* (FDA, 2004). Although it may look as if this approach was mainly promoted by the pharmaceutical industry, PAT had been adopted by many industries including the chemical, petrochemical and food industries even before its formalization by the FDA in 2004. The adoption of PAT by industries implies a transition from releasing final products based on traditional post-process quality control by offline analysis, to release products meeting the specifications using in-process testing and analysis of process data and measurements. This paradigm shift recommended by the FDA guidance is grounded on the concept that *“quality cannot be tested into products; it should be built-in or should be by design”* (FDA, 2004).

This quality concept is also a fundamental part of the CPS in Industry 4.0 and many PAT tools are used to achieve this goal. These tools, when used within a system, can provide effective and efficient means for acquiring information to facilitate process understanding, continuous improvement, and the development of risk-mitigation strategies. The FDA has categorized PAT tools into four categories:

- a) Multivariate tools for process design, data acquisition and analysis
- b) Process analyzers
- c) Process control tools
- d) Continuous improvement and knowledge management tools

When some or all of these tools are appropriately combined, they may be applied to a single-unit operation or to an entire manufacturing process to ensure quality assurance (FDA, 2004).

Recent PAT applications involve process monitoring by several advanced process analyzers, such as multivariate spectroscopic sensors (e.g. near-infrared and Raman spectroscopic probes) combined with univariate process sensors (e.g. devices to measure temperature or pressure). These analyzers can generate a large volume of data requiring the use of chemometrics tools based on multivariate analysis for process monitoring, modeling, and control. The combination of these tools allows the extraction of relevant process information for continuous process improvement. This information can be related to critical quality attributes to be assessed during different steps of the production process, be used for process understanding or consist of statistical parameters applied to control the process evolution and endpoint.

In this thesis, appropriate combinations of PAT tools have been proposed and applied to improve process understanding, monitoring and control of different processes. Although the main focus of this work has been the development of new multivariate analysis approaches for process monitoring and control, a brief description of the main instrumental process analyzers used in industry, with specific attention to those applied to the real processes in this thesis, is provided in the next subsection.

## **2.2 Analytical tools for process monitoring. Sensor typology and sensor data**

It is said that in Industry 4.0 "*nothing goes without sensor systems*" (Arnold, 2014). Indeed, as mentioned before, one of the four PAT tool categories defined by the FDA guideline includes the process analyzers (FDA, 2004). This subsection provides a basic introduction to the typology of process analyzers and their related data, generated for real-time process monitoring in this thesis.

Before describing the main typologies of analytical sensors, it is relevant mentioning that the location of a sensor in a process line determines whether the measurements obtained are *atline*, *online* or *inline*. *Atline* defines measurements coming from a sample manually extracted from the process line and analyzed in the immediate environment of the manufacturing process. *Online* applies to measurements acquired from a sample diverted, e.g. via bypass, from the process stream that can be returned to the process line. Finally, *inline* measurements take place when the sample is not removed from the process stream (FDA, 2004). Inline and online process analyzers can provide measurements at high frequency and are thus preferable for real-time process monitoring. Atline process analyzers perform measurements at lower frequencies but are more accurate.

The measurement of different physicochemical parameters during a process can provide valuable information that can be related to the physical and chemical state of

a process. The most conventional process sensors aim at measuring single parameters, also known as engineering variables, such as temperature, pressure, tank level and flow rate. Most of these measurements are currently carried out inline. Figure 3A shows the data related to a process sensor measuring vapor temperature during a batch distillation process (de Oliveira et al., 2017).

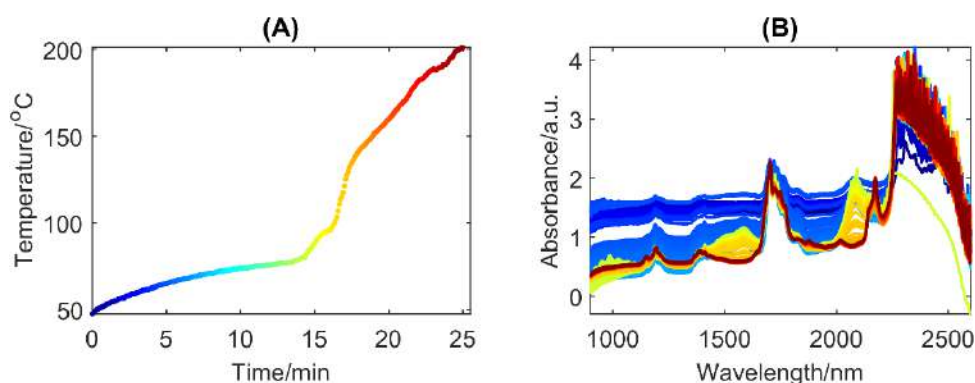


Figure 3 Inline measurements of a batch distillation process. (A) Vapor temperature and (B) raw NIR absorption spectra of distilled product. Observations are colored according to batch time, from dark blue (beginning of the process) to dark red (end of the process).

Thanks to the development of advanced analytical tools for process monitoring, more complex information can be obtained in real-time, which can be used for process understanding, monitoring and control. In this sense, spectroscopic probes are the main option and provide a multivariate response, i.e., a full spectrum, that informs about physical and chemical properties of the process point tested. Spectroscopic techniques, especially near-infrared (NIR) spectroscopy, stand out as attractive analytical tools for inline process monitoring. NIR spectroscopy is a type of vibrational spectroscopy ranging from 780 to 2500 nm in the electromagnetic spectrum and allows probing overtones and combination bands of the fundamental mid-infrared vibrations of -CH, -NH, and -OH groups (Burns and Ciurczak, 2009; Ozaki and Morisawa, 2021). Due to the ability to measure solid, liquid and gaseous samples, the fast acquisition of spectral data, the minimal or no sample preparation, and the low cost, NIR has become one of the most advantageous technique for process monitoring (Pasquini, 2018). The fiber optic probes coupled to NIR spectrometers can be located directly in the process streams, allowing continuous real-time in-process measurements, even in harsh environments (Avila et al., 2020; Grassi et al., 2019; Pasquini and Scafi, 2003). A NIR process analyzer provides chemical and physical information about the process, enclosed in a number of spectroscopic variables that can reach up to hundreds or thousands spectral channels, which is a much richer way to monitor a process than using only a few parameters issued from conventional process sensors. To exemplify this, Figure 3B shows several raw inline NIR absorption spectra collected during the batch distillation process. The data were collected synchronously to the temperature measurements in Figure 3A and represent the evolution of a single batch run. It is important to mention that other spectroscopic techniques such as UV-VIS, FT-IR, Raman, molecular fluorescence and even NMR are also used for inline process monitoring (Bowler et al., 2020; Dalitz et al., 2012;

Takahashi et al., 2015). Moreover, gas and liquid chromatography have also been adapted as online process analyzers. Even when there has been a big progress in the adaptation in terms of speed and accuracy of other spectroscopic sensors, these techniques still lack many of the advantages mentioned for the NIR probes and this is the reason why NIR analyzers are still the most current option in the industrial environment.

The versatility of NIR spectroscopy for process monitoring is not limited to the collection of a single spectrum per process point. Several NIR spectra can be collected in a spatially resolved manner using near-infrared hyperspectral image (NIR-HSI) process analyzers. Hyperspectral images (HSI), also called chemical images (CI) are a special type of spectroscopic measurement that gives both spectral and spatial information from a sample (Amigo, 2020). In this way, HSI connects chemical and spatial information and provide excellent information to study the composition of the chemical constituents of the sample and their spatial distribution. The spatial information is of utmost importance to study the heterogeneity evolution in mixing processes (Rocha de Oliveira and de Juan, 2021a, 2020). An HSI analyzer works by attributing a spectrum to every individual pixel in the image. As a consequence, the information of a HSI can be displayed as a 3D data cube, where two dimensions of the cube are the spatial coordinates ( $x$  and  $y$ ) of pixels and the third is the spectral dimension ( $\lambda$ ). Figure 4A illustrates this data cube for an atline NIR-HSI collected from a pharmaceutical formulation sample before a blending process (Rocha de Oliveira and de Juan, 2020). In Figure 4B, the NIR spectra related to three HSI pixels indicated in Figure 4A are shown.

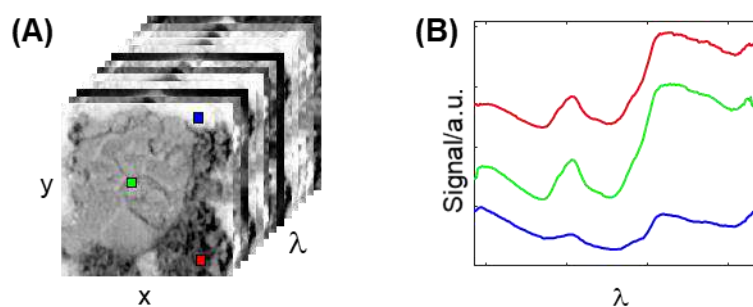


Figure 4 (A) Data cube representation of a NIR-HSI collected from a pharmaceutical formulation sample. (B) Representation of three pixel spectra from the NIR-HSI data.

HSI data can also be collected using other types of spectroscopic techniques such as Raman, FT-IR and fluorescence spectroscopies (Amigo, 2020; Gómez-Sánchez et al., 2021). Due to the inherent advantages related to NIR spectroscopy, NIR is one of the most used spectroscopic techniques to obtain HSI data for process monitoring (Boldrini et al., 2012). Another advantage of NIR-HSI compared to other imaging systems is the acquisition mode, very suited for real-time process monitoring. Thus, NIR-HSI acquisition for real-time process monitoring is usually carried out employing a line scanning or push-broom configuration. In a push-broom configuration, a detector is continuously scanning lines of pixel spectra over a sample and the HSI is built up

by moving the sample longitudinally. This configuration allows fast collection of HSI turning into a convenient industrial solution for the inline continuous and non-invasive monitoring of material on conveyor belt systems, for instance. This time saving comes at the expense of lower spectral and spatial resolution when compared to other configurations such as point and plane scanning. Description of these different image acquisition configurations can be found elsewhere (Grassi, 2020).

Process analyzers such as temperature, NIR probes and NIR-HSI were used in this thesis for the inline and atline monitoring of different types of batch processes. Atline measurements of some critical quality attributes were carried out using reference analytical methods. The batch process data generated by these process analyzers allowed the development of data-driven models for batch process understanding, monitoring and control using new approaches, often based on evolutions or combinations of the chemometrics tools described in the next subsection. It is important to note that, although NIR has been the spectroscopy used throughout the thesis, the chemometric solutions proposed can be applied to any other spectroscopic probe meant to be used in a process analysis context.

## **2.3 Chemometric tools for process monitoring, modelling and control**

Many chemometric tools have been designed for monitoring, modelling and control of industrial processes through multivariate analysis of batch process data. Exploratory tools are often used to model batch process data and visualize the variability among batches in a low-dimensionality space representing their process trajectories. Multivariate Curve Resolution offers deeper insight for modeling tasks and provides process profiles connected with spectroscopic fingerprints for all compounds involved in a process. Multivariate regression models allow the continuous monitoring of processes by predicting key properties from sensor measurements during the process evolution. The models and results provided by these tools can be combined to develop multivariate statistical process control charts, which are useful either for batch endpoint detection or to control the complete batch process evolution. The basics of the tools used to develop the new strategies for process monitoring and control developed in this thesis are described in the following subsections.

### **2.3.1 Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) is one of the most common exploratory methods for modeling multivariate batch process data (Wold et al., 2009, 1998). It is also the ground for the construction of most of the multivariate statistical process control (MSPC) charts, also called latent variable-based MSPC charts (Ferrer-Riquelme, 2010; Kourti and MacGregor, 1995). As many other chemometric tools, PCA is a data-driven method and provides information based only on the measurements acquired



during the process, without using any prior knowledge of the process mechanism or about the identity of the compounds involved.

Batch process data can be formed by many observations defined by many variables, e.g., hundreds or thousands of spectra collected during the process evolution, organized into a data matrix. The objective of PCA is to compress the information offered by the high-dimensionality batch process data matrix into a low-dimensional subspace defined by a small number of principal components (PC's). By doing so, the maximum non-random variance of the process data is explained by the PC's, which are orthogonal linear combinations of the original variables (Jolliffe, 2002; Wold et al., 1987a). The PCA model for a process data matrix  $\mathbf{X}(N \times J)$  is expressed as:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (1)$$

where  $\mathbf{T}(N \times A)$  is the scores matrix, related to the  $N$  observations of batch process data,  $\mathbf{P}^T(A \times J)$  is the loadings matrix, related to the importance of the  $J$  variables in the description of the  $A$  PC's, and  $\mathbf{E}(N \times J)$  is the residual matrix after modeling. The dimension of the PCA model, i.e. the number of PC's required to describe the relevant variation in the original data set, can be found using a suitable cross-validation method (Eastment and Krzanowski, 1982).

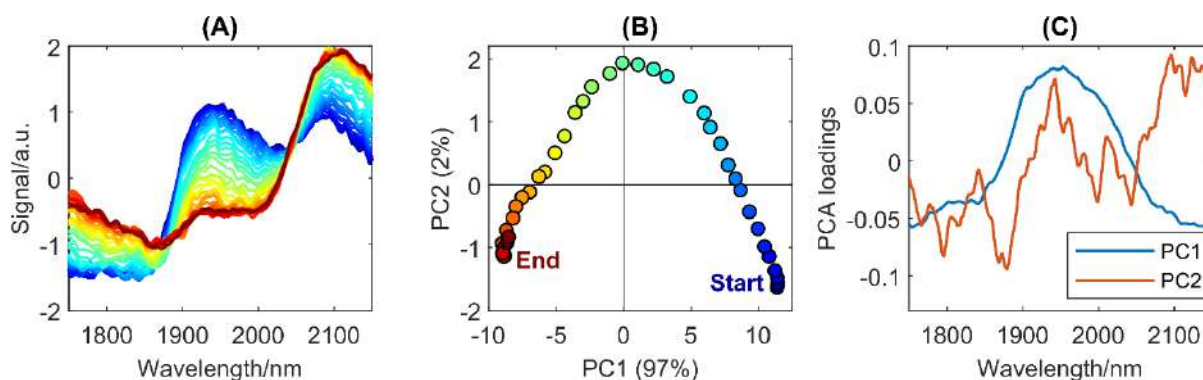


Figure 5 (A) Forty NIR spectra observations from a batch run of a fluidized bed drying of pharmaceutical granules process. (B) Related PCA scores scatter plot; colors are related to the spectral observations. (C) Loadings plot of the PCA model. The inline NIR spectral observations and related scores are colored according to batch time, from dark blue (first) to dark red (last). PCA results are obtained from centered data.

In the context of a process monitored spectroscopically,  $\mathbf{X}$  would be the matrix of spectra acquired,  $\mathbf{T}$  a matrix describing the evolution of the observations of the process in the small space of PC's and  $\mathbf{P}$  the loading matrix that would express the importance of the different spectroscopic variables to define the PC's. To illustrate this, Figure 5A represents 40 NIR spectra spanning 401 spectral channels from 1750 to 2150 nm, related to observations collected during a fluidized bed drying batch process (Avila et al., 2020; Rocha de Oliveira and de Juan, 2021b). The PCA decomposition of this spectroscopic data set is represented by the scatter score plot of the two first PC's in Figure 5B and the related loadings in Figure 5C. These two PC's account for 99% of

the total variance. The scatter plot of the spectral observations in the PCA space, Figure 5B, allows the visualization of the complete drying process trajectory. Note that the variation expressed by the 401 spectroscopic variables in Figure 5A is well captured using only two PC's. When more batches are included in this PCA model, the individual batch trajectories can be overlaid in the same score plot giving a global trajectory of this process. This trajectory also permits the visualization of batch deviations when a process disturbance occurs. High loadings (in absolute values) in Figure 5C point out at the most relevant spectroscopic variables to explain the spectral evolution along the process. As can be seen in the loading of PC1 (expressing 97% of the variance) the three bands of maximum spectral variation, with bigger signal changes around 1750, 1950 and 2150 nm, are clearly recognized.

### 2.3.2 Partial Least Squares Regression (PLS)

Batch processes were traditionally monitored by atline or offline determinations of critical parameters related to physical and chemical properties that defined the process evolution and the quality attributes of the end-products. Now, with the increasing importance of process digitalization and the continuous effort towards the *Industry 4.0* principles, many industrial processes are monitored with inline real-time process sensors. These sensors, often spectroscopic probes, provide the seeding information to set multivariate calibration models to relate the inline spectral observations collected to the related critical parameters of interest. The first step is establishing the model between the spectra measured (**X** matrix) and the critical parameters of interest (**Y** matrix) using a set of calibration samples, for which both kinds of information are available. Afterwards, the developed multivariate calibration model can be used for continuous monitoring of new batch processes through the real-time prediction of the critical process parameters from the inline spectroscopic measurements acquired.

Partial least squares (PLS) regression is the multivariate calibration method most frequently used to provide suitable models for real-time process monitoring (Wold et al., 2001). The objective of PLS is to build a calibration model that expresses the maximum covariance between the data matrix **X** of sensor measurements (e.g., spectra) and the matrix of parameters to be predicted **Y** (e.g., % moisture) (Martens and Næs, 1991; Næs, 2004). This covariance information is expressed by a few successive abstract factors, called latent variables, which compress the relevant information in **X** and **Y** matrices. The PLS algorithm decomposes the matrices **X** and **Y** in factor scores **T** and **U** related to samples in **X** and **Y**, respectively, and factor loadings **P** and **Q** related to variables in **X** and **Y**, respectively. The factor decomposition can be expressed by the equations below (Wold et al., 1987b).

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (2)$$

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F} \quad (3)$$

where  $\mathbf{E}$  and  $\mathbf{F}$  are the residuals in  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively, not explained by the latent variables in the models.

The regression model is obtained by the eq.(4) using  $\mathbf{T}$  and  $\mathbf{U}$ .

$$\mathbf{U} = \mathbf{T}\mathbf{B}_{PLS} + \mathbf{H} \quad (4)$$

where  $\mathbf{B}_{PLS}$  is a matrix that contains the PLS coefficients relating the information of  $\mathbf{X}$  and  $\mathbf{Y}$ , expressed by their respective scores. More details on how the regression coefficients are obtained can be found elsewhere (Haaland and Thomas, 1988; Martens and Næs, 1991; Næs, 2004; Wold et al., 2001, 1987b).

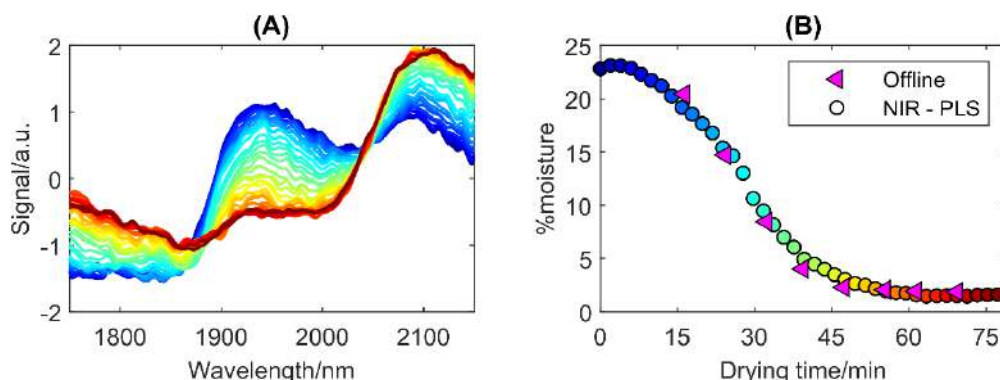


Figure 6 (A) NIR spectra observations from a fluidized bed drying of pharmaceutical granules. (B) PLS predictions of % moisture content based on the NIR observations. The inline NIR spectral observations and related PLS predictions are colored according to batch time, from dark blue (first) to dark red (last). Once the PLS regression model is built, it can be used for the prediction of critical parameters related to a set of new spectral observations,  $\mathbf{X}_{new}$ . First, the scores related to  $\mathbf{X}_{new}$  are obtained using the loading matrix  $\mathbf{P}$ ,

$$\mathbf{T}_{new} = \mathbf{X}_{new}\mathbf{P} \quad (5)$$

Then, using the PLS model coefficients,  $\mathbf{Y}$  scores ( $\mathbf{U}_{NEW}$ ) are calculated,

$$\mathbf{U}_{new} = \mathbf{T}_{new}\mathbf{B}_{PLS} \quad (6)$$

Finally, using the loading matrix  $\mathbf{Q}$ , the matrix of the parameters of interest,  $\mathbf{Y}_{new}$ , is predicted, as:

$$\mathbf{Y}_{new} = \mathbf{U}_{new}\mathbf{Q}^T \quad (7)$$

To illustrate the use of PLS for process monitoring, the drying spectroscopic data shown in the PCA example will be used. Thus, Figure 6A shows again the inline collected NIR spectra during the drying process and Figure 6B the related PLS-based predictions of moisture content as a function of the drying time (colored circles). The magenta triangles in Figure 6B represent the moisture content from samples retrieved for offline analysis, which indicates the good quality of the PLS predictions. The

establishment of the PLS calibration model was carried out using multivariate data collected from previous batch runs.

In this thesis, PLS regression is used to build multivariate calibration models for the prediction of one or more critical parameters in different batch process applications. The PLS predictions from the models developed will be also useful in the context of multivariate statistical process control to be combined with outputs from process sensors and other multivariate models.

### 2.3.3 Multivariate Statistical Process Control (MSPC)

Multivariate statistical process control (MSPC) models aim at providing statistical boundaries that allow building control charts to assess whether a process is on- or off-specification based on the measurement of multivariate process variables, e.g. NIR spectra (Kourti, 2009). MSPC models can have different goals, such as batch endpoint detection or checking process evolution. To use MSPC in any context, MSPC models should be built using multivariate process observations from batches that are representative of normal operating conditions (NOC). Afterwards, observations of new batches are submitted to the MSPC model to check whether they are within the NOC boundaries or not. In this section, only the basics on how to build and use MSPC models for endpoint detection are described. Later, in Section I of Chapter 4, new proposals to use MSPC for process evolution assessment will be described in detail.

Based on the same process example used to illustrate the PCA and PLS methods, an MSPC model for drying endpoint detection would be built using NIR spectra observations from the endpoint of several NOC drying batches, arranged in a data matrix,  $\mathbf{X}_{\text{NOC}}$ . The statistical boundaries for the drying endpoint detection would be set based on the PCA decomposition of  $\mathbf{X}_{\text{NOC}}$ .

$$\mathbf{X}_{\text{NOC}} = \mathbf{T}_{\text{NOC}}\mathbf{P}_{\text{NOC}}^T + \mathbf{E}_{\text{NOC}} \quad (8)$$

Here  $\mathbf{T}_{\text{NOC}}$  is the scores matrix of the endpoint NOC observations used to build the model and  $\mathbf{P}_{\text{NOC}}^T$ , the loadings matrix (which is the link between scores and the original matrix  $\mathbf{X}_{\text{NOC}}$ ).  $\mathbf{E}_{\text{NOC}}$  describes the residual variation unexplained by the PCA model.

Using this PCA model, the scores ( $\mathbf{t}_{\text{new}}$ ) for any new inline NIR spectrum acquired in real-time,  $\mathbf{x}_{\text{new}}$ , are obtained as follows,

$$\mathbf{t}_{\text{new}} = \mathbf{x}_{\text{new}}\mathbf{P}_{\text{NOC}} \quad (9)$$

And the related vector of residuals is obtained as:

$$\mathbf{e}_{\text{new}} = \mathbf{x}_{\text{new}} - \mathbf{t}_{\text{new}}\mathbf{P}_{\text{NOC}}^T \quad (10)$$

From the PCA-based MSPC model built with endpoint observations, two MSPC control charts can be built, in which observations of new batches are represented: a) a Hotelling's  $T^2$  chart, also referred as D-statistic ( $D_{stat.}$ ), where the Mahalanobis distance from the center of the latent subspace, representing the average NOC observation, to the location of the new observation in the score plot is displayed, and b) the Q-statistic ( $Q$ ) chart, where the residual of the new observation, related to the variation not explained by the PCA model of NOC observations, can be seen.

The Hotelling's  $T^2$  is calculated for any new observation using the predicted  $\mathbf{t}_{new}$  and the following equation,

$$T^2 = \mathbf{t}_{new}^T \mathbf{\Theta}^{-1} \mathbf{t}_{new} \quad (11)$$

where  $\mathbf{\Theta}$  is the PCA scores covariance matrix (Kourti, 2002; MacGregor and Kourti, 1995). Under the assumption that the scores follow a multivariate normal distribution, the control limit for this chart is calculated according to the equation estimated by (Jackson, 1991),

$$T_{lim}^2 = \frac{A(I-1)}{I-A} F(A, I-A, \alpha) \quad (12)$$

where  $I$  is the number of observations in  $\mathbf{X}_{NOC}$  used to build the PCA model with  $A$  PC's and  $F(A, I-A, \alpha)$  is the  $100(1-\alpha)$  percentile of the corresponding  $F$ -distribution.

The  $Q$ -statistic is calculated for any new observation using the vector of residuals  $\mathbf{e}_{new}$  as,

$$Q = \mathbf{e}_{new}^T \mathbf{e}_{new} \quad (13)$$

Regarding the control limit for the  $Q$ -statistic control chart limit,  $Q_{lim}$ , (Jackson and Mudholkar, 1979) showed that an approximate critical value for  $Q$  at significance level  $\alpha$  is given by,

$$Q_{lim} = \theta_1 \left[ \frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \quad (14)$$

where,  $\theta_1 = \sum_{j=A+1}^{rank(\mathbf{X}_{NOC})} \lambda_k^k$  and  $h_0 = 1 - \left( \frac{2\theta_1 \theta_3}{3\theta_2^2} \right)$ ,  $\lambda_j$  are the eigenvalues of the PCA residual covariance matrix and  $z_\alpha$  is the  $100(1-\alpha)$  standardized normal percentile.

Often, for an easier interpretation of the MSPC charts, reduced MSPC statistics ( $Qr$  and  $Tr^2$ ) are calculated by dividing the obtained  $Q$  and  $T^2$  values by their related

control limits; therefore, the control limit of charts based on reduced values becomes equal to one.

For the drying endpoint detection associated with the NIR spectra shown in Figure 7A, the related reduced  $Q$ -statistical control chart,  $Q_r$ , is depicted in Figure 7B. When the drying batch run started, the new spectral observations were far from the endpoint of previous batches (NOC observations); thus, their related residual vector,  $\mathbf{e}_{\text{new}}$ , was large and  $Q_r$  values appear above the chart control limit, shown as a discontinuous red line in Figure 7B. Conversely, as the process progresses towards the endpoint, new observations get gradually more similar to the NOC observations and the  $Q_r$  values decrease until the endpoint is reached, where  $Q_r < 1$ .

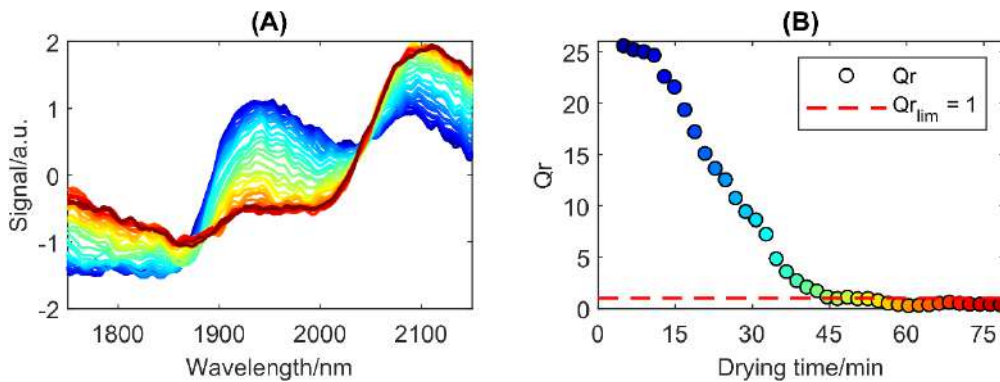


Figure 7 (A) NIR spectra observations from a fluidized bed drying of pharmaceutical granules, used for endpoint detection using an MSPC model. (B) MSPC  $Q$ -statistical control chart for endpoint detection related to the inline NIR observations, where the control limit ( $Q_{r\text{lim}} = 1$ ) is represented by the discontinuous red line. The inline NIR spectral observations and related  $Q_r$  values are colored according to batch time, from dark blue (first) to dark red (last).

In this thesis, PCA-based MSPC models were built for endpoint detection of single and multiphase batch process applications. The multivariate data used to build the MSPC models were based on NIR spectral measurements or on the combination of this information with other process sensor measurements in data fusion scenarios (See section 4.3.1). Although the developed MSPC for online control of batch process evolution is not described in this section, the basic principles to build control charts described here will be used.

### 2.3.4 Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS)

Another powerful tool to extract qualitative and quantitative information from multivariate process data is Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) (de Juan and Tauler, 2021; Tauler et al., 1993). For a spectroscopic-monitored process, MCR-ALS can provide the concentration profiles and pure spectral fingerprints for all compounds involved by using only the spectra acquired during the process evolution and without assuming any postulated process mechanism, i.e., it is a data-driven soft-modeling method. An MCR-ALS model tries to explain the maximum possible variance of the initial measurements through a bilinear model, as PCA does.

However, in contrast to PCA, MCR-ALS gives physically and chemically meaningful concentration and spectral profiles of the pure components of the system.

MCR-ALS assumes that the original set of process observations behaves following a bilinear model, which is the multiwavelength extension of Lambert-Beer's law and is described by the following expression (de Juan et al., 2014; de Juan and Tauler, 2003; Tauler et al., 2009, 1993):

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \quad (15)$$

In this context,  $\mathbf{X}$  is a data table with spectroscopic process observations from a single batch.  $\mathbf{S}^T$  contains the pure spectra signatures of the components needed to describe the process and  $\mathbf{C}$  the related concentration profiles.  $\mathbf{E}$  is the matrix with the residual part not explained by the model related to the experimental error. Figure 8A shows the graphical representation of the MCR-ALS decomposition of the data matrix  $\mathbf{X}$  into  $\mathbf{C}$  and  $\mathbf{S}^T$  profiles.

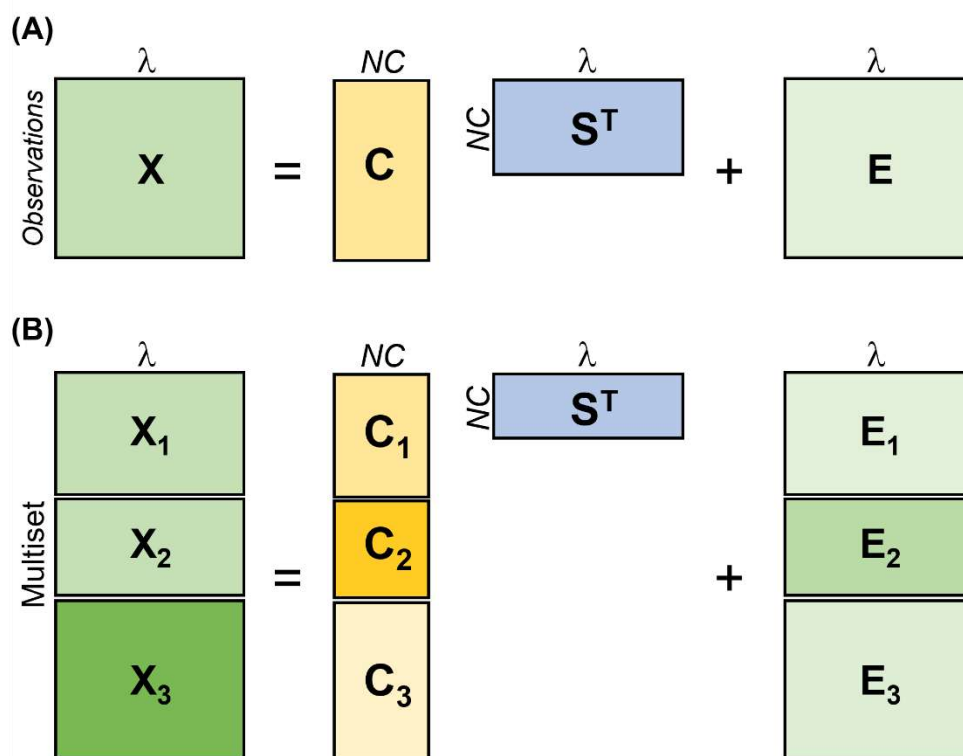


Figure 8 Graphical representation of MCR-ALS decomposition of a single dataset (A), and a multiset structure (B).

MCR-ALS obtains  $\mathbf{C}$  and  $\mathbf{S}^T$  matrices using an alternating iterative optimization method. First, an initial estimate of  $\mathbf{C}$  or  $\mathbf{S}^T$  should be used to start the iterative procedure. Then, in each iterative cycle, the  $\mathbf{C}$  and  $\mathbf{S}^T$  matrices are calculated under the action of some selected constraints, applied to reduce the ambiguity of the final solutions and to give physicochemical meaning to the profiles retrieved (Tauler et al.,

1995). The optimization continues until an optimal solution is obtained that fulfills the constraints assumed and the preset convergence criterion. Methods such as Evolving Factor Analysis (EFA) (Maeder, 1987) or SIMPLS-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) (Windig and Guilment, 1991) can provide good initial estimates to start the MCR-ALS iterative optimization. Typical MCR-ALS constraints commonly used are non-negativity and unimodality, i.e., presence of a single maximum per concentration profile. Additionally, local rank constraints, i.e., setting the absence of certain compounds in observations of the concentration profiles, can be useful to improve the quality of the resolved spectral signatures. In the context of process analysis, hard-modeling constraints, focused on fitting some concentration profiles according to a predefined physicochemical model can also be applied. More details about the MCR-ALS optimization procedure and available constraints can be found elsewhere (de Juan and Tauler, 2021; Jaumot et al., 2014; Tauler et al., 1993).

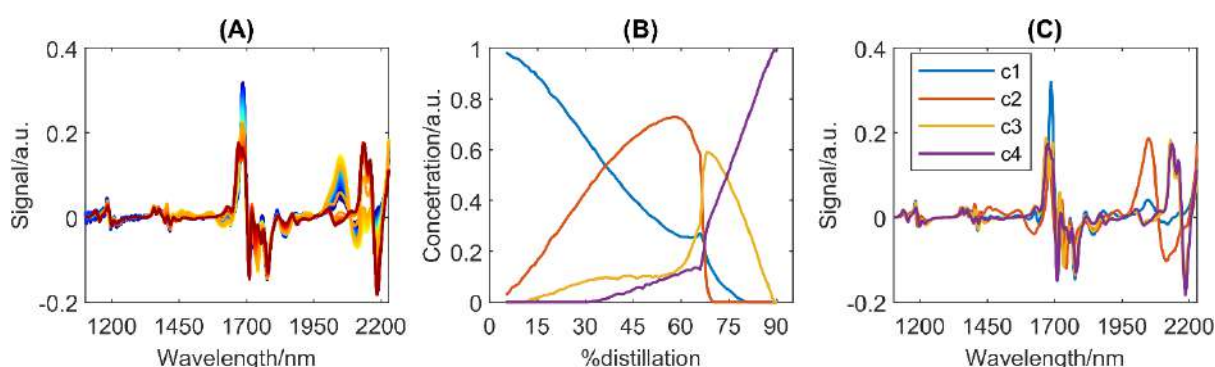


Figure 9 MCR-ALS process modeling of a distillation process monitored spectroscopically. (A) preprocessed inline NIR spectra colored according to % of the batch distillation from dark blue (5%) to dark red (90%). (B) The distillation (concentration) profiles for and (C) the related pure spectral profiles for the four components (c1-4)

To illustrate how MCR-ALS can be used to model a single batch process, Figure 9 shows the data of a gasoline/ethanol blend distillation batch monitored with NIR spectroscopy and the related MCR-ALS results (de Oliveira et al., 2017). Figure 9A shows the preprocessed spectral observations (86 NIR spectra with 573 spectral channels from 1100 to 2230 nm) collected inline from the distilled product. The MCR-ALS decomposition of this spectroscopic data set is represented by the concentration (distillation) profiles of four components in Figure 9B and the related spectral fingerprint profiles in Figure 9C. From the spectral profiles, the components in Figure 9C can be assigned to the main distilled fractions of gasoline/ethanol blends: light hydrocarbons (c1), ethanol (c2), and mid to high molecular weight (MW) hydrocarbons and aromatic compounds (c3 and c4). The identity of these compounds is confirmed when looking at the distillation profiles in Figure 9B. Note that the low MW hydrocarbons fraction is mainly distilled together with ethanol as azeotropes at the beginning of the distillation, as observed in the concentration profiles of components c1 and c2. After 70 wt% of the distillation process, almost all ethanol, component c2, was boiled-off remaining most of the mid to high MW fractions of gasoline, rich in aromatic compounds, components (3) and (4). See Publication I (de Oliveira et al., 2017) for a more detailed interpretation.



The same MCR-ALS bilinear model can also be used to perform image unmixing from processes monitored with hyperspectral imaging (HSI) systems (de Juan et al., 2019; de Juan, 2020; Piqueras et al., 2011; Rocha de Oliveira and de Juan, 2020). Although the dataset of a single HSI can be visualized as a 3D data cube (Figure 10A), where two dimensions ( $x$  and  $y$ ) are the pixel coordinates and the third is the spectral dimension ( $\lambda$ ), the data cube can be unfolded into a 2D matrix  $\mathbf{X}$  with all pixel spectra one under the other, with rows ( $x \times y$  pixels) and columns ( $\lambda$ ). The resolution of  $\mathbf{X}$  using MCR-ALS provides the  $\mathbf{C}$  and  $\mathbf{S}^T$  matrices related to the concentration and pure spectral fingerprint of the components present in the HSI. The stretched concentration profiles in matrix  $\mathbf{C}$  can be appropriately folded back to recover the unmixed distribution maps of each component, see Figure 10A.

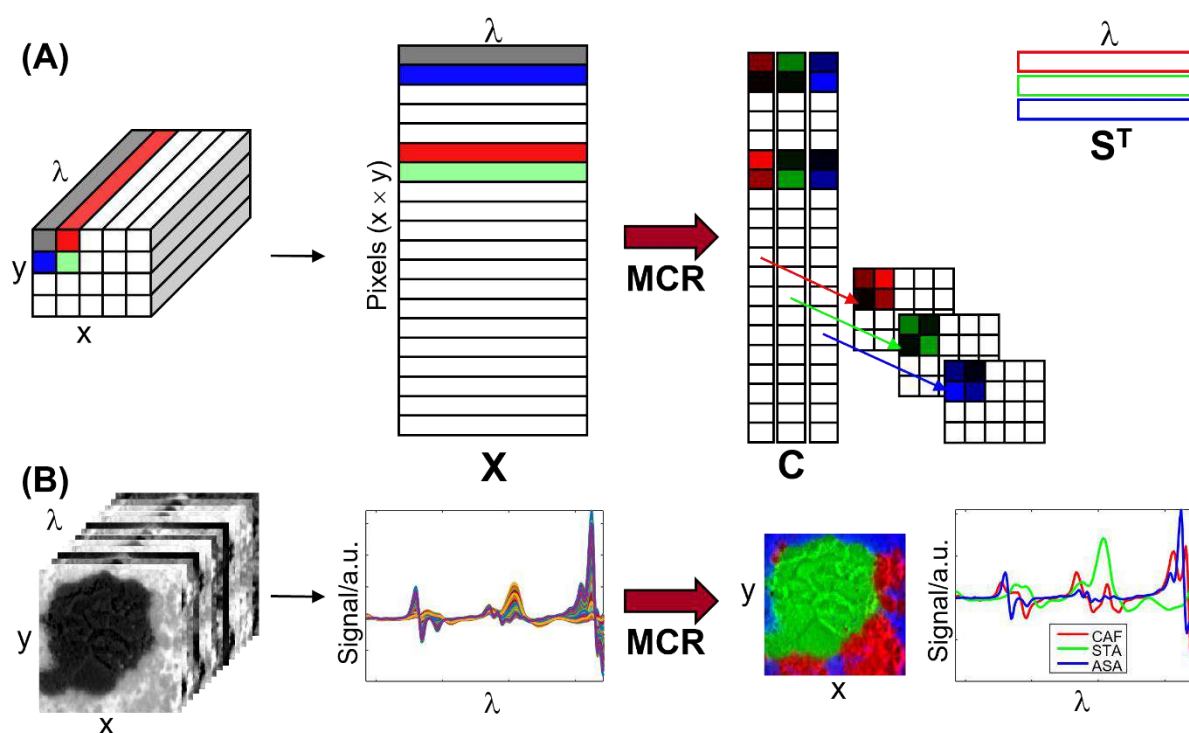


Figure 10 (A) Graphical representation of the application of Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) for unmixing of an HSI dataset. (B) Representation of the MCR-ALS unmixing of a real dataset, NIR-HSI (150 × 150 pixels × 224 spectral channels) of a pharmaceutical mixture (CAF – caffeine, STA – starch, and ASA – acetylsalicylic acid) into distribution maps and related pure spectral fingerprints.  $x$  and  $y$  are spatial pixel coordinates and  $\lambda$  represents the spectra wavelengths. Note that both distribution maps and spectral fingerprints of the three components are overlaid in a single plot.

Figure 10B illustrates how MCR-ALS can be used to unmix a real HSI dataset collected from a pharmaceutical mixture (CAF – caffeine, STA – starch, and ASA – acetylsalicylic acid) during a blending process (Rocha de Oliveira and de Juan, 2020). The data cube in Figure 10B (left) represents a NIR-HSI with 150 × 150 pixels and 224 spectral channels scanned, from 930 to 1750 nm. The preprocessed NIR pixel spectra, representing the unfolded HSI datacube (matrix  $\mathbf{X}$  in Figure 10A), are depicted next to the datacube in Figure 10B. The MCR-ALS unmixed distribution maps

(refolded  $\mathbf{C}$  matrix) and related spectral profiles (matrix  $\mathbf{S}^T$ ) are depicted on the right of Figure 10B. The distribution maps of the three pure components are overlaid in a combined RGB map (Red for CAF, green for STA, and blue for ASA). Note that after the MCR-ALS unmixing of the HSI dataset, the distribution of the three ingredients can be easily visualized. From these distribution maps, quantitative quality parameters related to different aspects of heterogeneity can be extracted allowing better blending process monitoring. This aspect will be discussed later in Section II of Chapter 4.

Either for process modeling or image unmixing, multiple data sets can be analyzed simultaneously using the same MCR-ALS bilinear model (Tauler et al., 2020). In this case, the individual matrices,  $\mathbf{X}_i$ , for each batch or HSI unfolded matrix can be arranged in a column-wise augmented multiset and the eq. (15) extended as,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \end{bmatrix} \quad (16)$$

In this data configuration, the components in the different matrices share the same spectral profiles  $\mathbf{S}^T$ , but may have different concentration profiles in each subset of the augmented matrix.  $\mathbf{C}_i$  submatrices can also have different numbers of rows. See Figure 8B for the graphical representation of eq. (16). The multiset configuration presented relates to multisets of  $\mathbf{X}_i$  matrices monitored using the same spectroscopic technique, as will be the case in the examples of this thesis, but other row-wise and/or column-wise augmented multiset arrangements, coupling several experiments monitored with different techniques are also possible. (Tauler, 1995; Tauler et al., 2020)

In this thesis, MCR-ALS was used for process modeling of a distillation process (Publication I) and image unmixing for a blending process monitored with NIR-HSI (Publication VI). MCR-ALS was also employed to compress spectroscopic information to design strategies for MSPC based on the sole NIR information or the combination with temperature profiles as discussed in Publications I and IV.



**CHAPTER 3. PROCESSES STUDIED.  
EXPERIMENTAL AND DATA  
PREPROCESSING**



### **3.1 Processes monitored with spectroscopic probes and process sensors**

The PAT tools presented in Section I of Chapter 4 were designed to tackle different situations related to the application of process modeling methodologies and the development of multivariate statistical process control approaches for batch processes. In this thesis, process control models have been applied for endpoint detection or to control the evolution from the start to the end of the batch process. Whereas endpoint detection is a simpler task and the main problems arise from the nature of the process and the sensors used, controlling the batch process evolution has required the development of new approaches to ensure that synchronized or non-synchronized batch process data could be handled adequately. Finally, all PAT tools developed were designed to handle data from batch processes monitored with single or multiple process sensors. To demonstrate these situations and the PAT tools developed, three batch processes with different characteristics and monitored by different sensors were used in this thesis. A description of each process and the purposes for which it has been used is presented in the next subsections.

#### **3.1.1 Multistep polyester production**

Saturated polyester resins are used in diverse applications, especially for powder coating in automotive and construction segments. The growing global demand for this product requires the optimization of the production process to ensure consistent final product quality. The industrial production of saturated polyester resins consists of the polycondensation reaction of polycarboxylic acids or their anhydrides and polyalcohols, producing water as a by-product. This is a batch process carried out in a two-stage esterification steps. The first stage involves the preparation of a precondensate by reaction of the acids with an excess of polyalcohols and the second stage takes place when the reaction of the remaining polyalcohols with additional acids is carried out. The reaction takes place at high temperatures and sensors need to be adapted to work in this harsh environment. To monitor each process stage completion and control the final product quality specification, some key analytical indicators such as acid number (AN) and viscosity (V) were usually monitored atline. These atline methods satisfied the needs for quality control tests, but they were labor and time-consuming and prone to sampling errors, making difficult the quick correction of process upsets and the fast and reliable detection of process stage endpoints.

For real-time process monitoring and implementation of statistical process control strategies, in situ NIR spectroscopy was proposed as a better approach both to do real-time prediction of the key parameters of interest and to detect the endpoint of each of the two process stages. In this thesis, the NIR data collected during process monitoring were used for the development of PAT tools for process monitoring and control.

Several batch runs for the production process of saturated polyester resins were carried out at Megara Resins industrial facility in Megara, Greece, following their commercial recipe. Figure 11A shows the complete process workflow. Liquid and solid reactants were added at the beginning of the two process stages, which took place under a continuous stream of nitrogen gas inside the reactor. Once the second stage reached the endpoint, additives were added to obtain the final polyester resin. The whole process was carried out at temperatures between 200 and 240 °C. Inline NIR spectroscopic process monitoring was carried out using an immersion transmittance probe coupled with optical fibers inside a small-scale reactor. Figure 11B shows the schematic experimental setup and a picture of the actual system where all batches were carried out. Key properties AN and V were determined atline by manual acid-base titration and using a cone/plate viscometer, respectively. These reference values were used to build NIR-based calibration models for inline prediction of AN and V values.

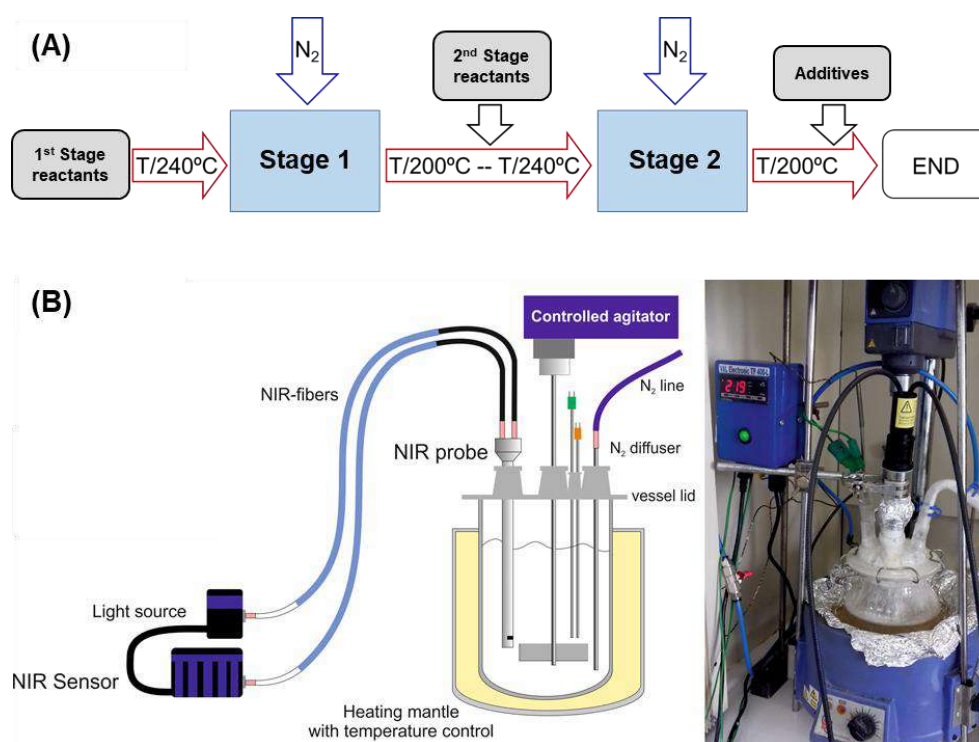


Figure 11 (A) Complete polymerization process workflow. (B) Schematic and actual picture of the experimental setup to produce the polyester resins, reproduced from (Avila et al., 2021).

The presence of a multiphasic and complex mixture of different materials, mainly at the beginning of each process stage, affected severely the NIR measurements, which showed a high noise level and strong baseline fluctuations due to the light scattering caused from both solid particles and gas bubbles. Spectral preprocessing consisted of the application of a moving average of consecutive NIR observations followed by Savitzky-Golay derivative (Savitzky and Golay, 1964) (1<sup>st</sup>-order derivative, 2<sup>nd</sup>-order polynomial function and 15-point window). Figure 12 shows the NIR infrared spectra before (Figure 12A) and after (Figure 12B) spectral preprocessing.

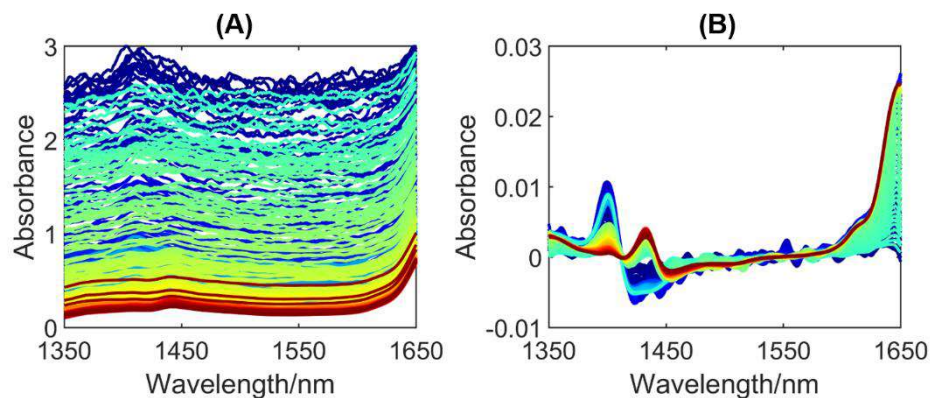


Figure 12 NIR spectra from a complete polyester production process before (A) and after (B) spectral preprocessing. Color scale indicates the temporal variation of batch observations, from the beginning (blue) to the end (red), reproduced from (de Oliveira et al., 2020).

Due to differences in process evolution among the different batch runs, this is an example of non-synchronized batch process data. With the NIR data, predictive multivariate models (PLS) for AN and V and PCA-based MSPC for endpoint detection of the two process stages have been developed. Also, a mid-level data fusion strategy to combine the several NIR information outputs issued from the PLS and MSPC multivariate models was designed to improve endpoint detection.

### 3.1.2 Fluidized bed drying of pharmaceutical granules

Fluidized bed (FB) drying is a common unit operation in the pharmaceutical industry. It is used to remove water or other solvents added to the dry powder pharmaceutical mixture during the wet granulation process. This process is performed before further processing operations, such as tablet compression. Therefore, controlling the endpoint moisture content is key to guarantee downstream processability and final product quality. Traditional methods for the monitoring of the moisture during a drying process include the loss on drying (LOD) and Karl Fischer titration methods; however, these methods can only be used for atline samples and are time-consuming (Green et al., 2005). Consequently, many pharmaceutical industries are introducing fast inline analytical techniques, such as NIR spectroscopy for real-time drying process monitoring (Green et al., 2005; Nieuwmeyer et al., 2007; Peinado et al., 2011). NIR spectra can be affected by both chemical and physical changes of pharmaceutical granules during the drying process. When coupled to chemometric tools, this information can be used to provide quantitative information about moisture content and to develop MSPC charts. Even though temperature sensors are also used for the inline monitoring of inlet and outlet air temperature during FB drying operation, this information is often ignored when developing PAT tools based on NIR spectroscopy. In this thesis, the FB drying process of pharmaceutical granules is monitored acquiring simultaneously NIR spectra and temperature readings for the development of PAT tools based on the complete process data information.



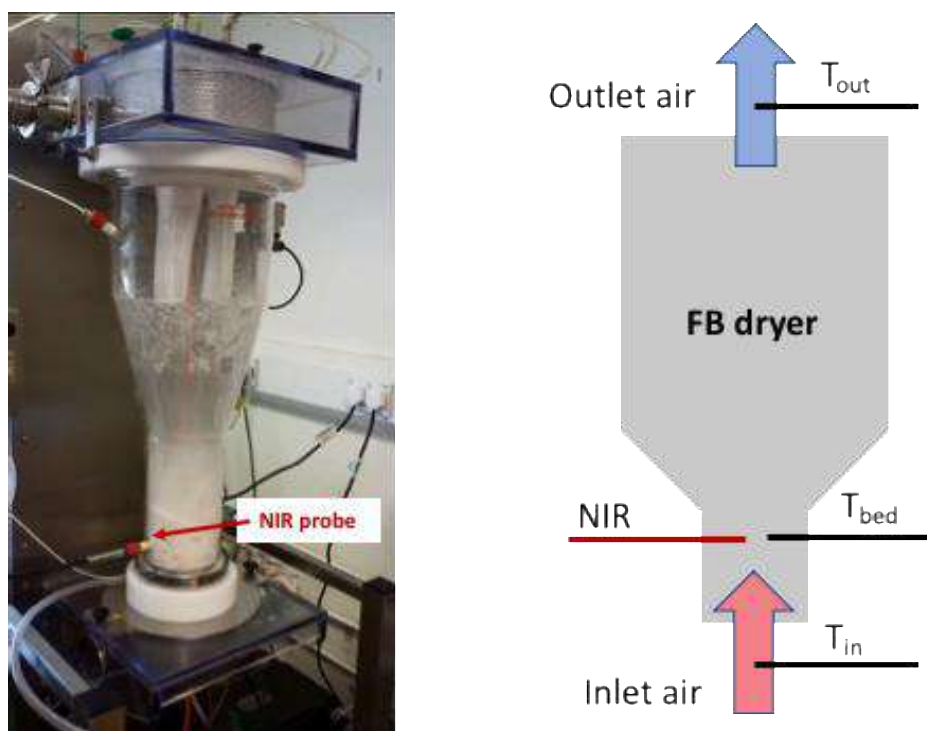


Figure 13 Left, Picture of the fluidized bed drying reactor with pharmaceutical granules. Right, Drawing of the FB drying reactor indicating the positions of the NIR and the three thermocouple sensors.

To achieve this purpose, several batches of pharmaceutical wet granules (mannitol > 50% and excipients) were dried in a pilot-scale fluidized bed reactor. The drying process was monitored with inline measurements of NIR (1750-2150 nm) reflectance spectra of the fluidized material. Simultaneously, three thermocouple sensors recorded the temperatures of the fluidized material ( $T_{bed}$ ), inlet air ( $T_{in}$ ) and outlet air ( $T_{out}$ ). Figure 13A shows a real picture of the FB drying reactor and Figure 13B a drawing with the different inline sensor locations. For each batch run, atline reference moisture content analysis was carried out using a thermogravimetric LOD moisture analyzer to build predictive multivariate models based on the NIR measurements. Due to the continuous pharmaceutical granules flow, the *in situ* NIR measurements suffered from a high noise level, which required suitable spectral preprocessing. To filter out the noise and correct the baseline fluctuations of the raw NIR spectra (Figure 14A) the preprocessing steps employed were the application of a moving average of consecutive NIR observations followed by standard normal variate (SNV) normalization (Figure 14B). The preprocessing applied to the temperature profiles was only the moving average, Figure 14C.

Due to differences in process conditions such as ambient moisture, inlet air temperature and airflow, the initial and final moisture content of the solid material among the drying runs is not the same. This variability causes batches with different drying durations and differences in the process progress of the different batch runs, which are not synchronized. The drying studied is also an example of multisensory

process monitoring since several temperature probes and an NIR probe are used simultaneously for process monitoring.

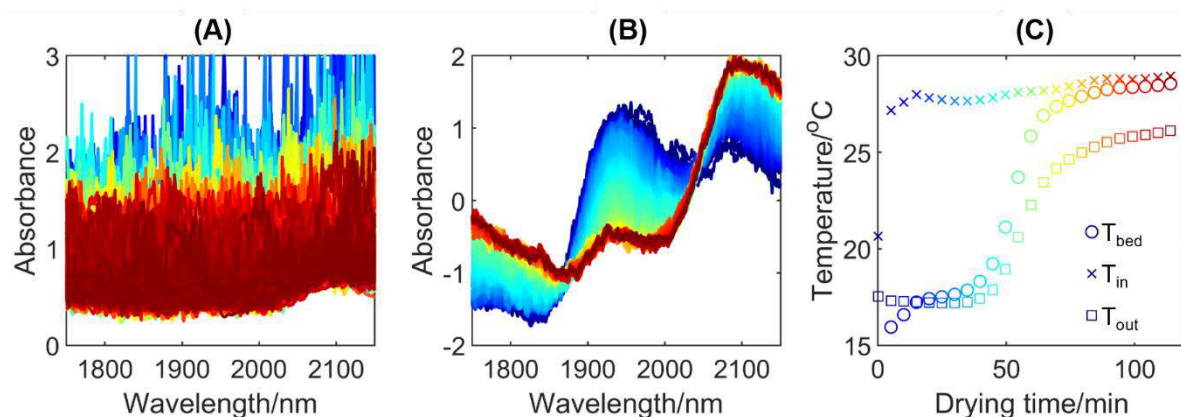


Figure 14 NIR spectra from a complete drying batch before (A) and after (B) spectral preprocessing. (C) FB temperature profiles. Color scale indicates the temporal variation of batch observations, from the beginning (blue) to the end (red), reproduced from (de Oliveira et al., 2020).

The data from this batch process have been used to test exploratory and process modeling tools for a better understanding of the process evolution. Predictive multivariate models (PLS) for moisture content and MSPC for drying endpoint detection using the NIR data have been also developed. Besides, a mid-level data fusion strategy to combine NIR information and temperature profiles has been designed to build MSPC for drying endpoint detection.

Additionally, new MSPC approaches designed to monitor the process evolution when batch process data are not synchronized could also be explored.

### 3.1.3 Benchtop batch distillation of gasoline and ethanol blends

Gasoline distillation is a usual standard procedure to assess the composition and quality of this commercial product. In some countries, such as Brazil, there are regulations related to the composition of the accepted commercial blends of gasoline and ethanol for particular kinds of fuel. Gasoline 'type C' is defined as the blend of pure gasoline and  $(27 \pm 1)\%$  ethanol (v/v). The standard control procedure relies only on the measurement of temperature during distillation to check whether the boiling point of the product studied matches the specifications of the regulated blend. However, in this procedure, no check of the chemical composition of the blend is carried out. In this thesis, the gasoline distillation process is monitored acquiring simultaneously temperature readings and NIR spectra for a more complete description of the process evolution and chemical characterization of the product studied.

To do so, 23 batches of synthetic gasoline samples were distilled in an automated batch distillation device designed for the inline monitoring of vapor temperature and distilled product with NIR spectroscopy. Synthetic gasoline samples were prepared by mixing ethanol and pure gasoline at different ratios. Distillation of on-specification

gasoline blends with 27% (v/v) of ethanol and off-specification blends with lower and higher ethanol content was carried out.

Figure 15 shows the distillation and inline data acquisition setup. For every distillation batch, vapor temperature readings using a thermocouple and inline FT-NIR absorption spectra in the range of 900 to 2600 nm were recorded in a synchronized way. The fraction of distilled mass was continuously monitored by an analytical balance. The temperature and NIR spectra stored are averages of all measurements recorded during every 1% distillation interval, in the 5 to 90% distilled mass range. Consequently, the information from every distillation batch contains the same number of NIR and temperature observations (86 NIR spectra and temperature measurements) related to the same distillation process stages, as defined by the percentage % (w/w) of distilled sample mass.

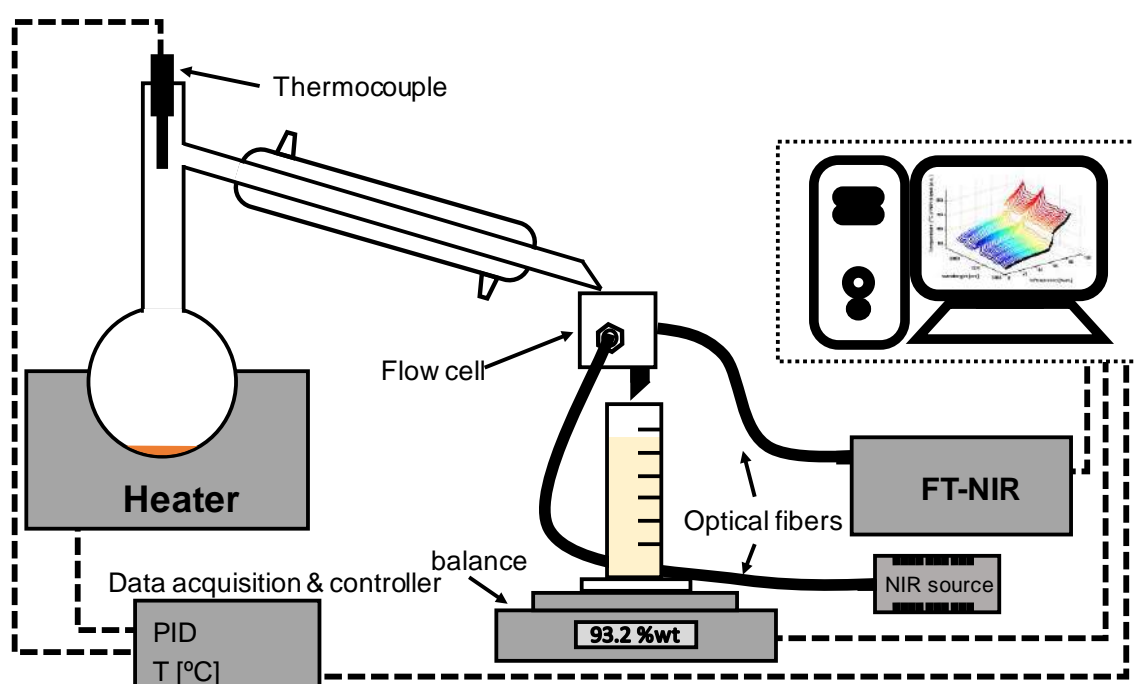


Figure 15 Experimental setup of the automatic distillation device with inline NIR and temperature monitoring, reproduced from (de Oliveira et al., 2017).

The spectral preprocessing steps used for the NIR spectra collected during the distillation process were the Savitzky-Golay (Savitzky and Golay, 1964) first-order derivative for baseline correction followed by spectral normalization to mitigate signal intensity fluctuations of the NIR spectra. Figure 16 shows the NIR spectra before and after preprocessing for a particular batch. Temperature readings were used as such.

Due to the procedure employed to store the information, i.e., averaged temperature and NIR spectra associated with every 1% distilled mass fraction, this process is a very good example of a situation where batches are naturally synchronized because the percentage of distillation weight gives a direct reference for batch progress

evolution. It is also an example of multisensory process monitoring since temperatures and NIR spectra are simultaneously acquired during the distillation progress.

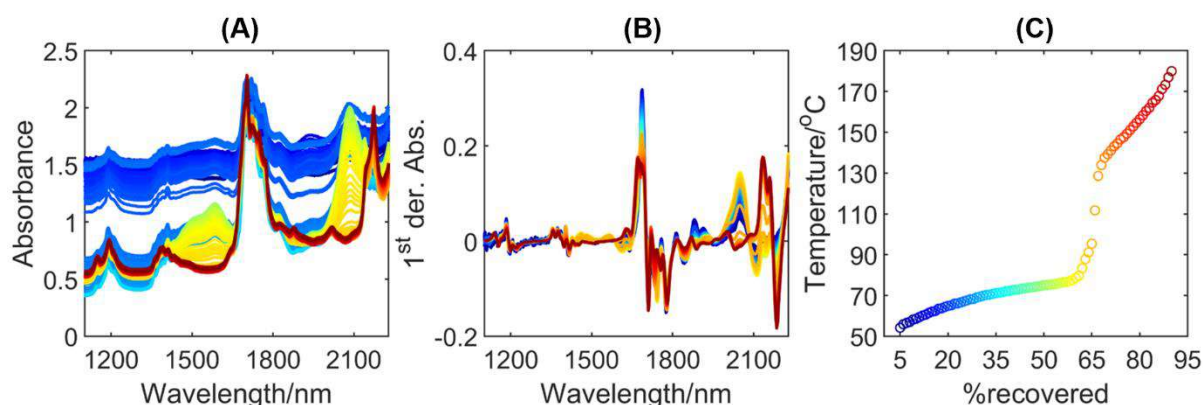


Figure 16 Data obtained from the distillation of a gasoline blend with 27% ethanol. (A) Raw and (B) preprocessed NIR spectra. (C) Vapor temperature profile. Color scale indicates the temporal variation of batch observations, from the beginning (blue) to the end (red), reproduced from (de Oliveira et al., 2020).

The batch process data from this example have been used to test several exploratory and process modeling tools for a better understanding of the process evolution.



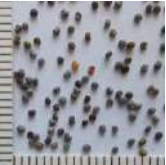
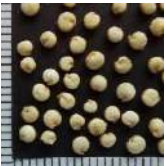




Besides, new MSPC approaches designed to monitor the process evolution for synchronized and non-synchronized batch data could also be explored. Finally, work using only NIR measurements and checking different NIR/T data fusion strategies has been carried out.

### 3.2 Process monitoring using hyperspectral imaging (HSI)

Process monitoring by hyperspectral imaging provides the added value of having spatial and spectral information about the samples or process stages studied. The use of this kind of measurement is particularly interesting when some relevant aspects of the process evolution, e.g., heterogeneity, require necessarily a spatial description. However, in general, scanning a sample area using hyperspectral imaging will always provide more representative information about samples and processes than relying on the measurement of a sensor probe, capable to scan a single sample point or a limited field of view.

Blending processes are the clearest example where a spatial description of the system studied will help to check whether the mixing action is evolving adequately and whether the endpoint of the process has been achieved. Indeed, when a good blending is reached, the spatial distribution of the compounds in the blend is even and no agglomerations are present. A proper combination of hyperspectral imaging and suitable chemometric analysis can help to monitor the heterogeneity of the blend and detect when it is sufficiently low to consider the blending process ended.

Table 1 Properties and picture with 1 mm reference scale of the material used to prepare the blending runs.

Category	Material	Bulk density ( $g\ mL^{-1}$ )	Particle size (mm)	Picture (1 mm scale)
Food	Ground coffee (GC)	0.350(0.003) <sup>a</sup>	<0.5 <sup>c</sup>	
	Rice grits (RG)	0.763(0.007)	1 <sup>b</sup>	
	Poppy seeds (PS)	0.616(0.008)	1 <sup>b</sup>	
	Quinoa seeds (QS)	0.75(0.02)	2.3 <sup>b</sup>	
Pharma	Acetylsalicylic acid (ASA)	0.78(0.01)	1 <sup>b</sup>	
	Caffeine (CAF)	0.77 (0.02)	<0.5 <sup>c</sup>	
	Citric acid (CA)	0.87(0.04)	1 <sup>b</sup>	
	Sodium starch glycolate (SSG)	0.817(0.006)	<0.106 <sup>d</sup>	

<sup>a</sup>Standard deviation from triplicate measurements of bulk density in parenthesis.

<sup>b</sup>Aproximate average particle size as determined by image-based particle size analysis.

<sup>c</sup>Particle size was too small to be determined by the image-based particle size analysis.

<sup>d</sup>Particle size through 140 mesh (min. 99%). Provided by JRS Pharma certificate of analysis.

Two types of blending process monitoring by Near Infrared-Hyperspectral Imaging (NIR-HSI) were carried out in this work. The first was an atline process monitoring, based on acquiring NIR-HSI measurements from static material related to different blending times. The second blending monitoring consisted of an in-situ NIR-HSI continuous monitoring of blending material. Specificities of each blending monitoring are described in the subsections below.

In both cases, several blending batches of different solid materials were carried out. The materials present in the blending mixtures consisted of pharmaceutical compounds and food products with diverse physical properties in terms of particle size, particle shape and density for a better understanding of the influence of these characteristics in the blending evolution and final blend quality. A description of relevant physical characteristics of the materials used is shown in Table 1.

### 3.2.1 Blending process monitoring with atline NIR-HSI

This blending monitoring was carried out using ternary mixtures of ASA, CAF and SSG at different ratios. Atline NIR-HSI data were collected from the blending material after stopping the blending process at different blending times providing 11 NIR images per batch run.

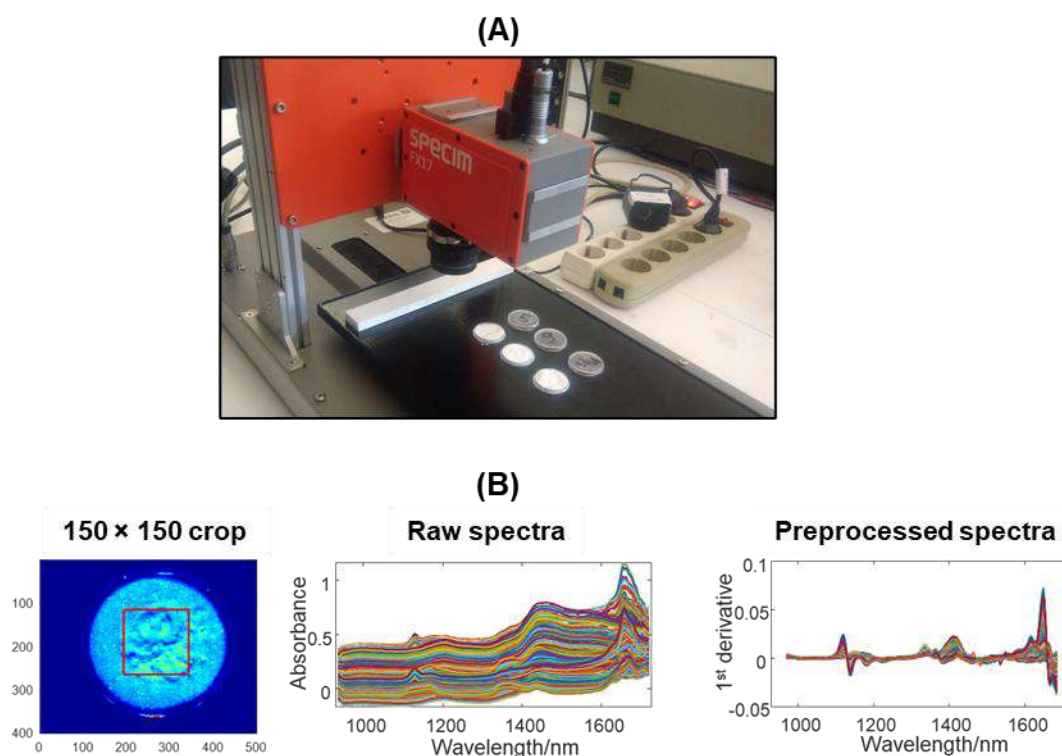


Figure 17 (A) NIR-HSI acquisition setup for the atline monitoring of the blending processes. (B) NIR-HSI data preprocessing steps with the selected analyzed area (150 × 150 pixels square area), and the raw and preprocessed NIR spectra from the selected area.

The hyperspectral images at each blending time of this process were acquired with a pushbroom NIR camera (935-1720 nm, Specim FX17 by Spectral Imaging Ltd., Oulu,

Finland), which allowed the sequential collection of lines of 640 pixels spectra with a frame rate of 35 Hz and sample scanning rate of 3.2 mm/s. The system setup allowed the collection of hyperspectral data in diffuse reflectance mode. Pixel size in all images was approximately  $0.1 \times 0.1 \text{ mm}^2$ . For more detail on the instrumentation, see reference (Rocha de Oliveira and de Juan, 2020). Figure 17A shows the NIR-HSI data acquisition system and the atline monitoring of the blending material.

Before data analysis of the NIR-HSI data, a squared region of interest (ROI) from the center of each raw image ( $150 \times 150$  pixels) was cropped for further analysis, which represented a sample area of ca.  $15 \times 15 \text{ mm}$ . The raw NIR absorbance spectra of the cropped image were preprocessed using Savitzky–Golay first derivative for baseline correction. Figure 17B shows the preprocessing steps carried out.

### **3.2.2 Blending process monitoring with inline NIR-HSI**

Continuous blending monitoring was performed on ten blending batches using binary and ternary mixtures made of different combinations of the food or pharmaceutical materials described in Table 1. The blending runs were carried out in a lab-scale horizontal rotary blender, which consisted of a 40-mL glass vial attached directly to a stepper motor that enabled rotation of the vial around its longitudinal axis with a constant rotational speed, see Figure 18A. The same NIR camera from the previously described process was used to record the hyperspectral data in this setup. However, in this case, the camera was positioned below the blending vial allowing continuous acquisition of hyperspectral data, i.e., lines of pixel spectra covering all the width of the vial during the blending runs.

Typical raw NIR spectra from a scanned line of pixels of the push broom camera are shown in Figure 18A. Spectral baseline variation was corrected using Savitzky-Golay first derivative followed by spectral normalization using the Euclidean norm, see Figure 18B.

The data collected in the atline and inline blending monitoring by NIR-HSI measurements were used to propose quantitative heterogeneity indicators used to describe the blending progress and the quality of the final blend. These indicators could be used to define the heterogeneity at a compound-specific and global sample level.

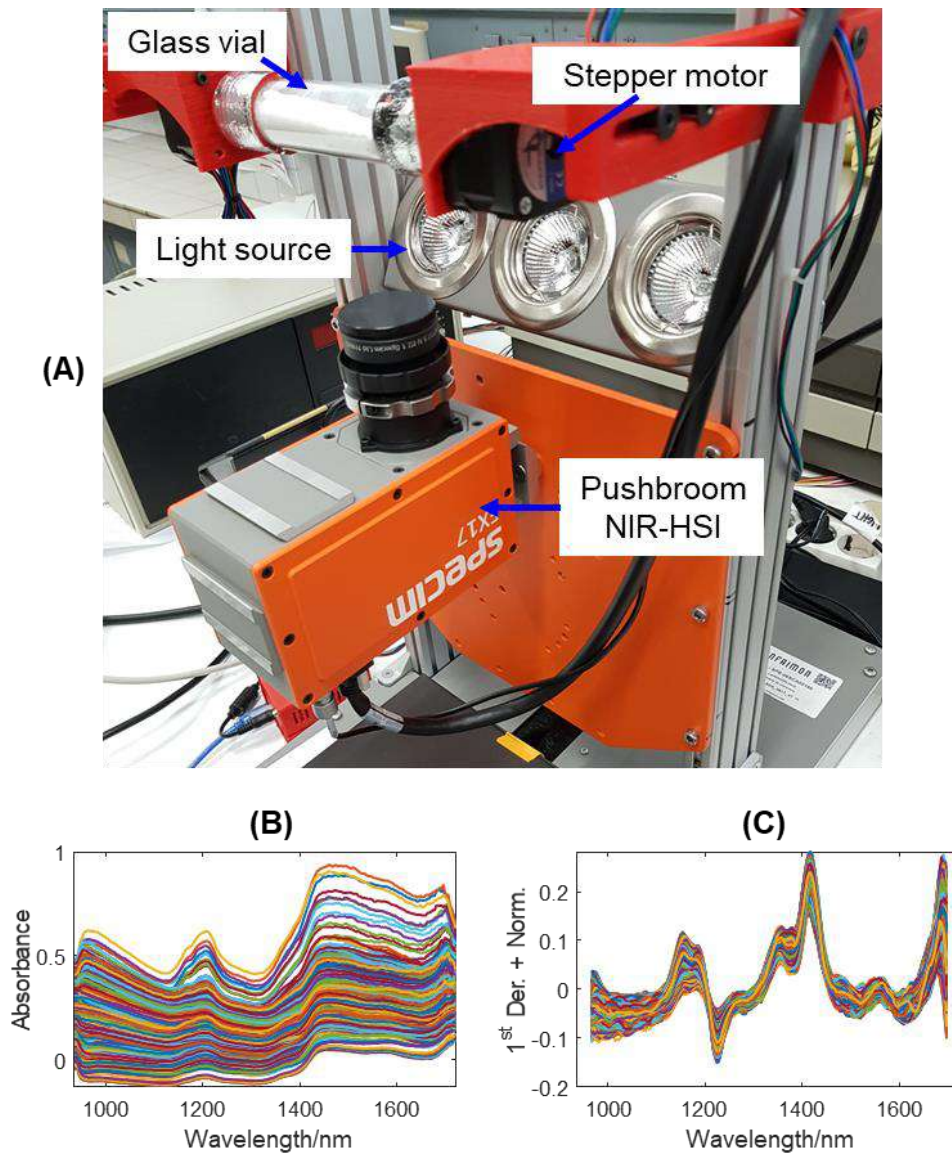


Figure 18 (A) Picture of the experimental setup for the continuous monitoring of the blending process using the inline NIR-HSI. (B) Typical raw NIR spectra from a scanned line of pixel spectral and (C) the same spectra after preprocessing using Savitzky-Golay first derivative and spectral normalization (1<sup>st</sup> Der. + Norm.).





## **CHAPTER 4. RESULTS AND DISCUSSION**



## **SECTION I – Process monitoring, modeling and control using spectroscopic probes and process sensors**

This section gathers several scientific publications focused on the application and development of new PAT tools for batch process monitoring, modeling and control of diverse processes. New MSPC approaches designed to deal with synchronized and non-synchronized batch data as well as the combination of information from several sensors or model outputs were explored in this thesis.



#### 4.1 Data configurations. Synchronized and non-synchronized multibatch data

As mentioned in section 2.3, process data measurements from a single batch consist of the collection of several variables,  $J$ , (process data and/or spectroscopic measurements) at different process observations,  $K$ , throughout the batch. These measurements are usually organized in a data matrix,  $\mathbf{X}(K \times J)$ , to be used for process monitoring, modeling or control purposes. When several NOC batches of the same process are monitored, the total number of batch process data matrices  $\mathbf{X}$  will be as many as  $I$  batches monitored, see Figure 19A. Because of the inherent batch process complexity and non-stationary behavior, key process events may not occur at the same time point when comparing different NOC batch runs of the same process. This usually leads to NOC batches with different durations, i.e. different number of process observations  $K$ , and, most important, to process evolutions that do not follow the same and synchronized process pattern. When these differences occur, we talk about non-synchronized batch data.

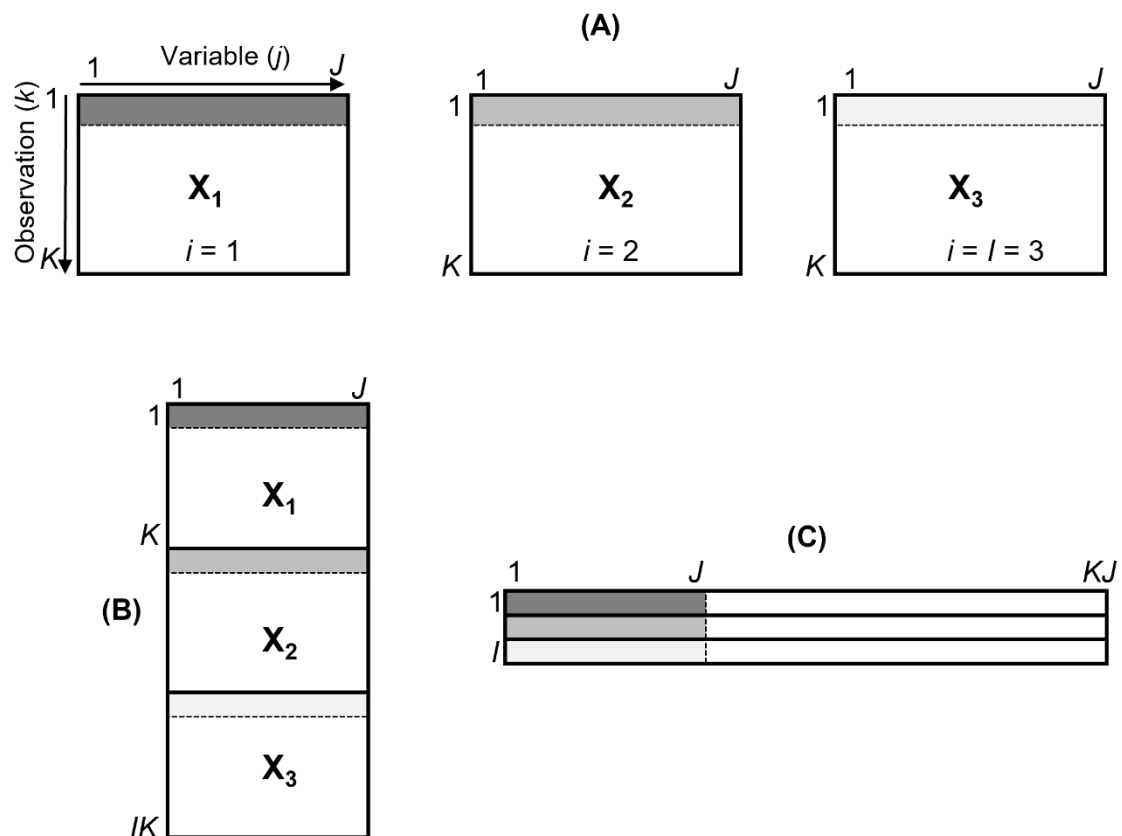


Figure 19 (A) Multibatch data. (B) Variable-wise and (C) batch-wise multibatch structures.  $I$  is the number of batches;  $K$ , the number of observations or rows of  $\mathbf{X}$ ; and  $J$ , the number of variables or columns of  $\mathbf{X}$ .

Although non-synchronized batches are the most common situation in practice, sometimes the possibility to monitor a variable that is related to batch evolution, sometimes called generically maturity variable, can be used to directly align batch data

and obtain synchronized batches. An example of this scenario are the distillation processes that will be presented in detail in chapter 3.1.3. Batch alignment can also be performed using PLS (using local batch time or a maturity variable as the  $y$  variable) (Wold et al., 2009) or using more advanced algorithms, such as correlation optimized warping (COW) or dynamic time warping (DTW) (José M. González-Martínez et al., 2014; Kassidas et al., 1998; Liu et al., 2017; Ramaker et al., 2004; Zhao et al., 2020).

For batch process modeling using PCA or MCR-ALS as introduced in sections 2.3.1 and 2.3.4, respectively, data from multiple batches are usually organized into a variable-wise augmented matrix sized  $(IK \times J)$ , as shown in Figure 19B. This augmented matrix is obtained by appending each batch matrix  $\mathbf{X}_i$  one on top of each other and the only requirement is that the variable dimension,  $J$ , i.e., the spectroscopic and/or process variables, be common among all batches. This is an easy condition since batch monitoring is carried out with the same process sensors and covering the same spectral range. This data arrangement adapts to both synchronized and non-synchronized batch process data since only the variable dimension is common, but the augmented observation dimension allows for differences in size and process dynamics among batches.

Another way to organize multiple batch data is using the so-called batch-wise augmented matrix sized  $(I \times KJ)$ , as shown in Figure 19C. It is done by first performing a batch vectorization, i.e. unfolding the batch matrix into a row vector with  $KJ$  elements, then placing the vectorized batch data one on top of each other in a batch-wise augmented matrix. This type of multibatch structure can only be used with synchronized batch data since it requires not only the same number of observations per batch,  $K$ , but, for a meaningful bilinear model, it assumes that NOC batches share the same batch dynamics in the new column direction formed by the vectorized batch data. This type of multibatch structure used for process control was first named after their authors as the NM approach (Nomikos and MacGregor, 1995) in the literature. MSPC approaches most often use the variable-wise structure, which was formerly named as the WKFH approach (Wold et al., 1998).

The next subsections will present different strategies to perform multibatch process modeling and control using the multibatch structures above. The choice of the multibatch data arrangement will depend on the goal of the data analysis task and on the synchronized or non-synchronized nature of the batches to be coupled.

## 4.2 Process modelling and control for synchronized batch processes

This subsection shows the discussion of the results related to the work published in Publication I. In this article, different strategies for process modelling and control of gasoline batch distillation processes are used. Chemometric tools described in section 2.3, such as PCA and MCR-ALS, were used for process understanding. New MSPC strategies were specifically designed for real-time control of the process evolution of synchronized batch process data.

**Publication I.** de Oliveira, R. R., Pedroza, R. H. P., Sousa, A. O., Lima, K. M. G., and de Juan, A. **Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy.** *Analytica Chimica Acta* (2017), 985: 41–53.

DOI: [10.1016/j.aca.2017.07.038](https://doi.org/10.1016/j.aca.2017.07.038)







Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: [www.elsevier.com/locate/aca](http://www.elsevier.com/locate/aca)

## Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy



Rodrigo R. de Oliveira <sup>a, b</sup>, Ricardo H.P. Pedroza <sup>b</sup>, A.O. Sousa <sup>d</sup>, Kássio M.G. Lima <sup>b, c</sup>, Anna de Juan <sup>a, \*</sup>

<sup>a</sup> Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Diagonal 645, 08028, Barcelona, Spain

<sup>b</sup> LabPVT, Federal University of Rio Grande do Norte, Av. Senador Salgado Filho, 3000, Natal 59078-970, RN, Brazil

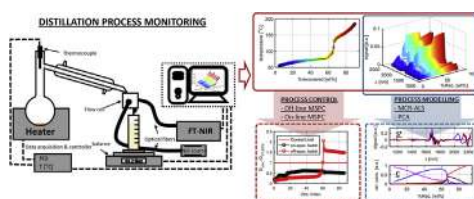
<sup>c</sup> Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Av. Senador Salgado Filho, 3000, Natal 59078-970, RN, Brazil

<sup>d</sup> Departamento de Física – CCET – UFRN Campus Universitário, Lagoa Nova 59072-970, Natal, RN, Brazil

### HIGHLIGHTS

- Design of a PAT oriented batch distillation device recording synchronously NIR, T and distilled mass.
- Process understanding achieved by using global (PCA) and component (MCR-ALS) trajectories.
- MCR-ALS distillation profiles are used as starting information to build MSPC models.
- Evolving batch MSPC strategies are proposed to control on-line gasoline distillation processes.

### GRAPHICAL ABSTRACT



### ARTICLE INFO

#### Article history:

Received 20 December 2016

Received in revised form

12 July 2017

Accepted 13 July 2017

Available online 21 July 2017

#### Keywords:

Near-infrared spectroscopy

On-line multivariate statistical process control - MSPC

Process modeling

Distillation process

Petroleum

### ABSTRACT

A distillation device that acquires continuous and synchronized measurements of temperature, percentage of distilled fraction and NIR spectra has been designed for real-time monitoring of distillation processes. As a process model, synthetic commercial gasoline batches produced in Brazil, which contain mixtures of pure gasoline blended with ethanol have been analyzed. The information provided by this device, i.e., distillation curves and NIR spectra, has served as initial information for the proposal of new strategies of process modeling and multivariate statistical process control (MSPC). Process modeling based on PCA batch analysis provided global distillation trajectories, whereas multiset MCR-ALS analysis is proposed to obtain a component-wise characterization of the distillation evolution and distilled fractions. Distillation curves, NIR spectra or compressed NIR information under the form of PCA scores and MCR-ALS concentration profiles were tested as the seed information to build MSPC models. New on-line PCA-based MSPC approaches, some inspired on local rank exploratory methods for process analysis, are proposed and work as follows: a) MSPC based on individual process observation models, where multiple local PCA models are built considering the sole information in each observation point; b) Fixed Size Moving Window – MSPC, in which local PCA models are built considering a moving window of the current and few past observation points; and c) Evolving MSPC, where local PCA models are built with an increasing window of observations covering all points since the beginning of the process until the current observation. Performance of different approaches has been assessed in terms of sensitivity to fault detection and number of false alarms. The outcome of this work will be of general use to define

\* Corresponding author.

E-mail address: [anna.dejuan@ub.edu](mailto:anna.dejuan@ub.edu) (A. de Juan).

strategies for on-line process monitoring and control and, in a more specific way, to improve quality control of petroleum derived fuels and other substances submitted to automatic distillation processes monitored by NIRS.

© 2017 Elsevier B.V. All rights reserved.

## 1. Introduction

Distillation curves are frequently used for quality control of petroleum products. The evolution and shape of these curves is directly related to the composition and chemical characteristics of these products and, hence, a temperature deviation from normal distillation behavior may be an indicator of adulteration. ASTM D86 [1] is the standard test method required to obtain distillation curves and classical process control is made by comparing the temperature at specific distillation points with standard specification limits.

However, distillation curves, based only on boiling temperature monitoring, are not conclusive to identify adulterations in product composition. Adulterants can nowadays be chosen so that the modified petroleum products show normal distillation curve behavior. Near-infrared spectroscopy (NIRS) may help to overcome such scenario because of the rich physicochemical information associated with this spectroscopic technique and the existence of many NIR sensors designed for on-line process monitoring. Along this line, distillation devices that incorporate NIR sensors and collect synchronized distillation temperatures and related NIR absorption spectra measurements, as proposed by Pasquini and Scafi, are a suitable solution [2]. Thus, the fiber optic probes coupled to NIR spectrometers can be located directly in the distillation process stream, allowing continuous real-time in-process measurements [2–4]. Therefore, information representing both physical and chemical properties of the distilled sample can be derived from each distillation batch.

In Brazil, commercial gasoline is blended with ethanol. Thus, gasoline derived directly from refineries without ethanol addition is denominated “type A”. Gasoline “type C” is the commercial mixture of gasoline “type A” and  $(27 \pm 1)\%$  of ethanol (% v/v) [5]. As a process model for this work, a study of quality control of Brazilian gasolines regarding ethanol content specification is proposed. To do the experimental process monitoring, an improved version of the automatic distillation device monitored by NIRS proposed by Pasquini and Scafi [2], which allows continuous and synchronized data acquisition and storage of distillation temperatures, distilled mass and related NIR spectra, is proposed. Detailed description of the experimental setup is found in section 2 below.

The distillation curves and NIR spectra collected from distillation batch processes can be modeled with principal component analysis (PCA) [6] and multivariate curve resolution – alternating least squares (MCR-ALS) [7] for better process understanding and use of this information in further process control. PCA batch analysis provides global distillation trajectories, whereas MCR-ALS offers the additional value of describing the temperature-dependent evolution and characterization of the different distilled fractions during the process.

MSPC has been used to control processes related to very diverse fields, such as pharmacy [8–11], petrochemistry [12–14] and biotechnology [15,16]. Batch MSPC using NIRS has been described in recent works [3,4,8–11,16,17]; however, no MSPC using NIRS to monitor batch distillation process has been reported in the literature.

In this study, different off-line process control models are studied using complete batch information collected during the

distillation process monitored by NIRS. Distillation curves, original NIR spectra, as well as the compressed spectral information contained in PCA scores or MCR-ALS concentration profiles are used to build off-line PCA-based multivariate statistical process control (MSPC) models. To our knowledge, there is no report in the literature about using the concentration profiles from MCR-ALS analysis as starting information to build PCA-based MSPC models. In this framework, the description of the separate components of the process provided by MCR-ALS would allow for using all concentration profiles on the MSPC model or profiles of selected compounds that could be envisioned as more specific indicators of process evolution.

NIR measurements obtained from distillation processes are also used to build on-line batch MSPC models. On-line batch MSPC approaches commonly used are based on the methods proposed by Nomikos and MacGregor [18–20] and Wold et al. [21]. Other approaches are proposed by Rännar et al. for adaptive batch monitoring using hierarchical PCA [22], by Zhao et al. using multiple PCA models for local model building at each observation point [23] and using moving window [24]. In this work, chemometric tools typically used to perform local exploratory analysis of the evolution of processes, such as evolving factor analysis (EFA) or fixed size moving window - EFA (FSMW-EFA) [25,26], have been adopted to propose new on-line batch MSPC strategies. The performance of these on-line MSPC approaches has been studied in terms of sensitivity to fault detection and number of false alarms.

The outcomes of this study will be of general applicability, as guidelines for process modeling and control based on spectroscopic measurements, and suppose a significant improvement on the specific field of quality control based on distillation processes, both from the instrumental point of view and from the way to handle the derived information from coupled temperature-NIR distillation curves.

## 2. Experimental

### 2.1. Automatic distillation device setup

The automatic distillation device designed is shown in Fig. 1. It is formed by a distillation glassware setup (125 ml), a transmittance flow cell connected through optical fibers to a FT-NIR spectrophotometer (Rocket, ARCOptix ANIR, Switzerland), an analytical balance (XS204, Mettler-Toledo, Switzerland), a thermocouple and heater controlled by a data acquisition device and a personal computer with a data acquisition software that connects and controls the distillation setup. Heating mantle power applied is automatically controlled based on a feedback controller to keep distillation rate constant rather than keeping constant power as in Ref. [2].

### 2.2. Batch distillation process

For every distillation batch, 100 mL from the suitable sample, previously weighed, are introduced in the distillation flask. The heater is started and once the initial boiling point (IBP) is automatically detected, distillation process starts and synchronized

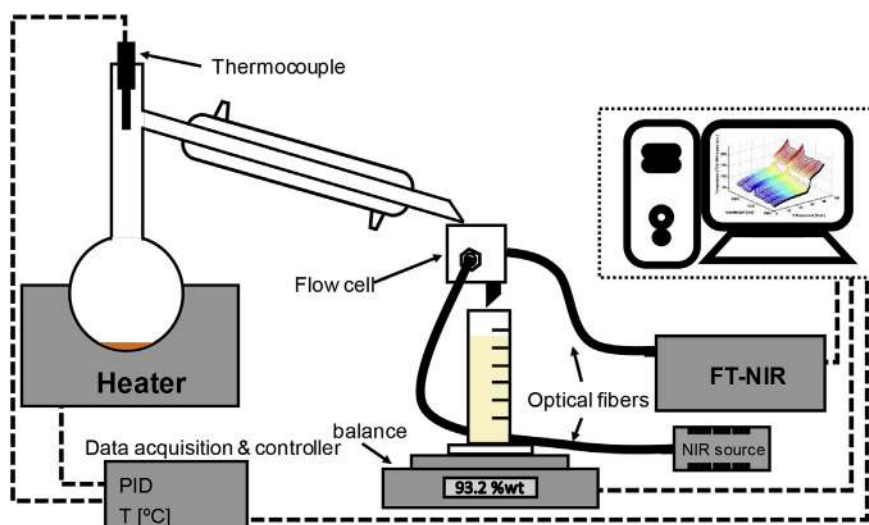


Fig. 1. Experimental setup of the automatic distillation device with on-line NIRS monitoring.

measurements of temperature, distillation recovered percentage (wt%) of initial sample weight and NIR absorption spectra (900–2600 nm) are taken every five seconds until the end point (EP) is reached. Data are stored in MATLAB format in such a way that values every 1 wt% are saved. Temperature and NIR spectra are averages of all measurements recorded during every 1 wt% distillation interval.

Synthetic gasoline (type C) batches were distilled using the designed automatic distillation device. 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 27%(v/v) ethanol (on-specification gasolines) and 12 with 10–25%(v/v) and 30–40%(v/v) ethanol (off-specification gasolines). Table 1 describes the gasoline batches prepared with their related composition. These batch ID labels will be used to identify the batches throughout the manuscript.

### 3. Data treatment

#### 3.1. Raw data and preprocessing

Temperature, distilled weight and NIR spectra were obtained synchronously every 5 s and averaged measures were stored every 1 wt% of distilled weight increment from IBP until EP. The final process range considered was from 5 to 90 wt% distilled weight, which corresponded to  $K = 86$  observation points. Observations at the beginning (<5 wt%) and end (>90 wt%) of the distillation process were unstable and, therefore, not used for process control. NIR spectra working wavelength range was 1103–2228 nm due to high

noise observed in measurements out of these wavelength boundaries. This range contained  $J = 573$  spectral channels. For each distillation batch, a column-vector sized ( $K \times 1$ ) with the temperatures associated with the distillation curve and a matrix sized ( $K \times J$ ) with the related NIR infrared spectra were obtained.

Data obtained from on-specification batch B07 are used to illustrate the typical data obtained at the end of a batch distillation run. Fig. 2(a) shows the distillation curve with the recorded boiling temperatures, Fig. 2(b) the related raw NIR spectra and Fig. 2(c) the raw NIR spectra at the four observation points indicated in Fig. 2(a).

NIR spectra were preprocessed for baseline correction by Savitzky-Golay derivative [27] (1st order derivative, 2nd order polynomial function and 9 points window) followed by signal intensity fluctuation corrected by spectral normalization, see Fig. 3(b).

#### 3.2. Data analysis

##### 3.2.1. Process modeling. Principal Component Analysis (PCA) and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

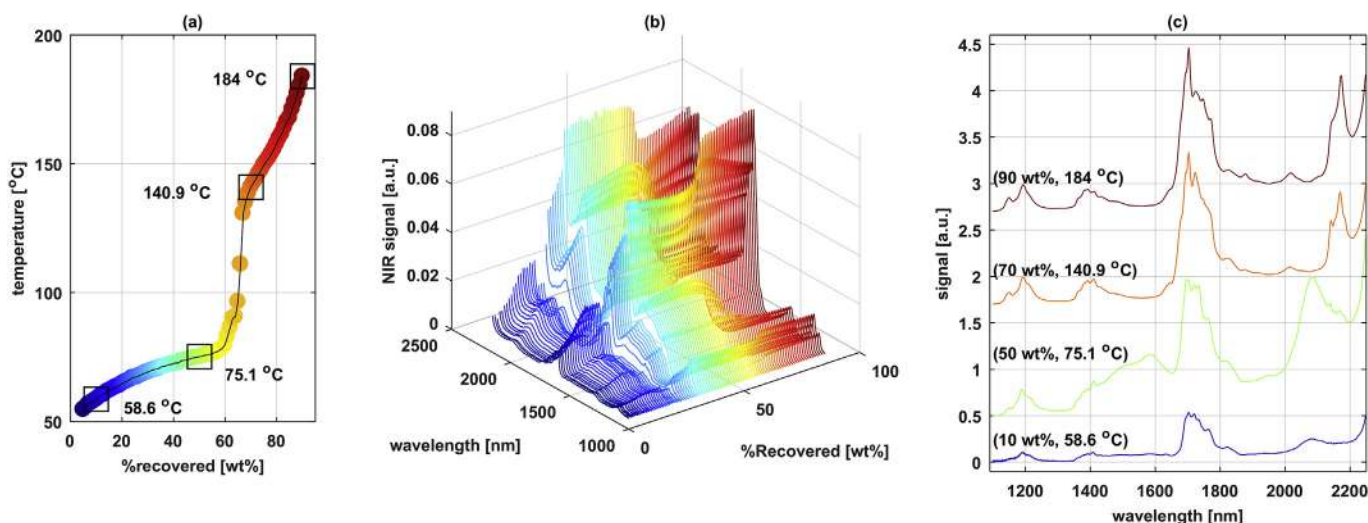
The matrices with the NIR data from each on-specification batch were arranged one on top of each other into a column-wise augmented multiset structure,  $\mathbf{D}$ , and modeled using principal component analysis (PCA) and multivariate curve resolution-alternating least squares (MCR-ALS). PCA provided a global model of trajectories explaining the overall process evolution, whereas MCR-ALS provided a model describing the evolution and chemical identity of each component (distinct distilled fraction) in the distillation batches analyzed.

PCA was used to reduce the dimensionality of the spectral data from the distillation processes by compressing the high-dimensional mean-centered original NIR data matrix into a low-dimensional subspace of principal components. These components explain most of the data variability and are orthogonal linear combinations of the original spectroscopic variables [6]. The PCA model of column-wise augmented matrix  $\mathbf{D}$  is expressed as:  $\mathbf{D} = \mathbf{TP}^T$ , where  $\mathbf{T}$  are the scores, related to the observations of the distillation process and  $\mathbf{P}^T$  are the loadings, related to the importance of the NIR wavelengths in the description of the principal components. The scatter plot of scores provides the global trajectories of the processes analyzed.

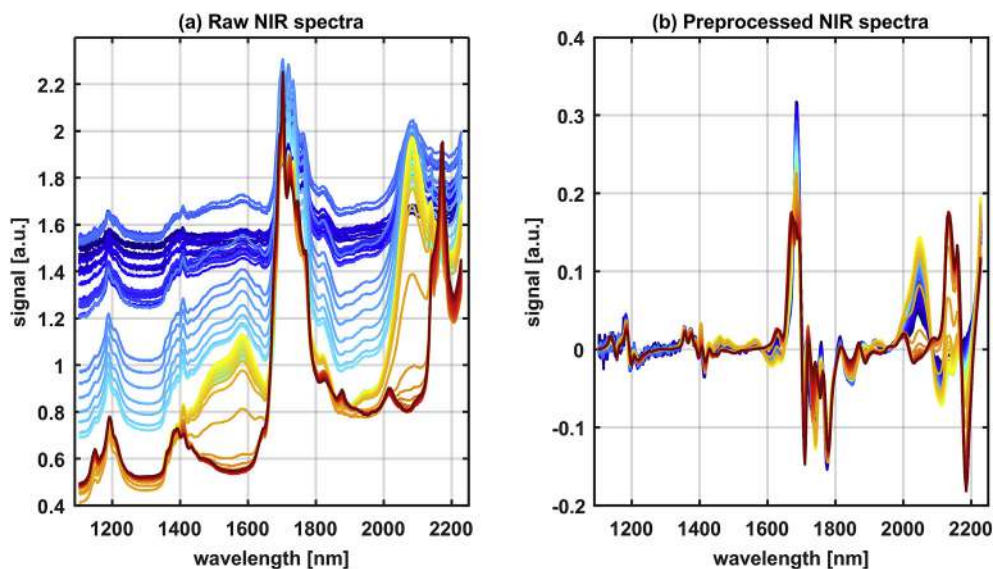
The same multiset structure was modeled using multivariate

Table 1  
Description of Batch ID and their related composition, as used in this work.

Batch ID	%(v/v) Gasoline	%(v/v) Ethanol	Class
B01-B11	73	27	On-specification
B12	90	10	Off-specification
B13	85	15	
B14-B16	80	20	
B17-B18	75	25	
B19-B20	70	30	
B21-B22	65	35	
B23	60	40	



**Fig. 2.** Process data from distillation batch B07. (a) Distillation curve, (b) On-line raw NIR spectra vs. percentage of the distilled fraction [wt%] and (c) raw NIR spectra at distilled fraction at 10, 50, 70 and 90 wt% indicated in (a), the spectra were vertically offset for clear comparison.



**Fig. 3.** Plot of the (a) raw and (b) preprocessed NIR spectra obtained from the distillation batch B07 between 5 and 90 wt% with 1 wt% interval, 5 wt% (blue) → 90 wt% (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

curve resolution - alternating least squares (MCR-ALS). MCR-ALS assumes a bilinear model,  $\mathbf{D} = \mathbf{C}\mathbf{S}^T$ , which is the multiwavelength extension of the Lambert-Beer's law [7,28–30].  $\mathbf{S}^T$  contains the pure spectra of the components needed to describe the distillation process and  $\mathbf{C}$  the concentration (distillation profiles). In contrast to PCA, MCR-ALS gives real meaningful concentration and spectral profiles of pure components of the system. MCR-ALS works by alternatingly optimizing  $\mathbf{C}$  and  $\mathbf{S}^T$  under constraints. Initial estimates of  $\mathbf{S}^T$  were performed by using a pure variable selection method based on SIMPLISMA [31]. Constraints applied in this work were non-negativity and unimodality, i.e., presence of a single maximum per profile, for the concentration ( $\mathbf{C}$ ) profiles. Local rank constraints, i.e., setting the absence of certain compounds in observations of the concentration profiles, were used to improve the quality of the resolved spectral signatures [32]. This was done by appending pure ethanol NIR spectra to the column-wise multibatch structure (in this case, only the ethanol was set to be present in the

concentration elements linked to the appended pure ethanol spectra).

MCR-ALS provides a much more detailed description of the process than PCA in terms of characterization of process profiles and spectral signatures, related to distillation fractions in this case. However, the single process trajectory provided by the scatter score plot of PCA is a global description of process evolution and a quick visual way to observe when a batch process evolves as NOC batches or does differently. Being complementary views about the evolution of a process, we found relevant to include both in this study. Both PCA scores and MCR  $\mathbf{C}$  profiles are afterwards used as starting information for off-line batch MSPC models described in the next section.

### 3.2.2. Process control

From the batches analyzed, nine on-specifications or NOC (Normal Operation Conditions) batches (batches B01-09), were

selected to build PCA-based MSPC models (see Table 1). These models were afterwards used to detect whether a new batch (or observations within it) is in or out of control [33]. Two on-specification batches (B10-11) and twelve off-specification batches (B12-23) were used to test the MSPC models.

The PCA-based MSPC model is built using the preprocessed and mean-centered data matrix of NOC batches,  $\mathbf{X}_{\text{NOC}}$ , sized (nr. of NOC batches  $\times$  observed measurements per batch) according to the equation below,

$$\mathbf{X}_{\text{NOC}} = \mathbf{T}_{\text{NOC}}\mathbf{P}_{\text{NOC}}^T + \mathbf{E}_{\text{NOC}} \quad (1)$$

where  $\mathbf{T}_{\text{NOC}}$  is the scores matrix of all NOC batches and  $\mathbf{P}_{\text{NOC}}^T$  is the loadings matrix. The number of components used in an MSPC model is a critical parameter and has been established by cross-validation [34].

The scores for new batches are obtained multiplying the measured preprocessed batch information,  $\mathbf{X}_{\text{NEW}}$ , with the loadings matrix  $\mathbf{P}_{\text{NOC}}^T$  from the model built with the NOC batches, using the following equation:

$$\mathbf{T}_{\text{NEW}} = \mathbf{X}_{\text{NEW}}\mathbf{P}_{\text{NOC}} \quad (2)$$

Then, the residuals are obtained using the new batch scores, as:

$$\mathbf{E}_{\text{NEW}} = \mathbf{X}_{\text{NEW}} - \mathbf{T}_{\text{NEW}}\mathbf{P}_{\text{NOC}}^T \quad (3)$$

From the PCA model built with NOC batches, two MSPC control charts can be built, in which observations of new batches are represented: a) Hotelling's  $T^2$  chart, usually referred as  $D$ -statistic ( $D_{\text{stat.}}$ ), represents the estimated Mahalanobis distance from the center of the latent subspace, representing the average in control conditions of a batch, to the projection of a new batch (or observation) onto this subspace and the b)  $Q$ -statistic chart ( $Q_{\text{stat.}}$ ) accounts for the residual part of the process variation not explained by the PCA model.

The Hotelling statistic,  $D_{\text{stat.}}$ , was calculated using the following equation:

$$D_{\text{stat.}} = \mathbf{t}^T\Theta^{-1}\mathbf{t} \quad (4)$$

Where  $\mathbf{t}$  is the vector containing the scores of a new given batch with the  $A$  retained principal components (PC's), and  $\Theta$  is the scores covariance matrix with  $(A \times A)$  size. The control limit for this chart is calculated according to the equation proposed by Jackson [35].

$$D_{\text{CL}} = \frac{A(I-1)}{I-A}F(A, I-A, \alpha) \quad (5)$$

where  $I$  is the number of in control batches used to build the model with  $A$  PC's and  $F(A, I-A, \alpha)$  is the  $100(1-\alpha)$  percentile of the corresponding  $F$  distribution.

The  $Q_{\text{stat.}}$  for the  $i$ th new batch  $\mathbf{x}_i$  is given by

$$Q_{\text{stat.}} = \mathbf{e}_i^T \mathbf{e}_i \quad (6)$$

where  $\mathbf{e}_i$  is the residual vector of the  $i$ th new batch from the PCA model. Regarding the control limit for the  $Q_{\text{stat.}}$  chart, Jackson and Mudholkar [36] showed that an approximate  $Q_{\text{stat.}}$  critical value at significance level  $\alpha$  is given by

$$Q_{\text{CL}} = \theta_1 \left[ \frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \quad (7)$$

where,  $\theta_k = \sum_{j=A+1}^{\text{rank}(X)} \lambda_j^k$  and  $h_0 = 1 - (2\theta_1\theta_3/3\theta_2^2)$ ,  $\lambda_j$  are the

eigenvalues of the PCA residual covariance matrix and  $z_\alpha$  is the  $100(1-\alpha)\%$  standardized normal percentile.

Two MSPC approaches were applied in this work, devoted to off-line and on-line control, respectively. Both approaches and related control charts are explained below.

### 3.2.3. Off-line batch MSPC

Off-line batch MSPC charts were built using data provided from completed distillation processes. Different models were built according to the starting information used, either temperatures from distillation curves or information derived from NIR spectra, Fig. 4.

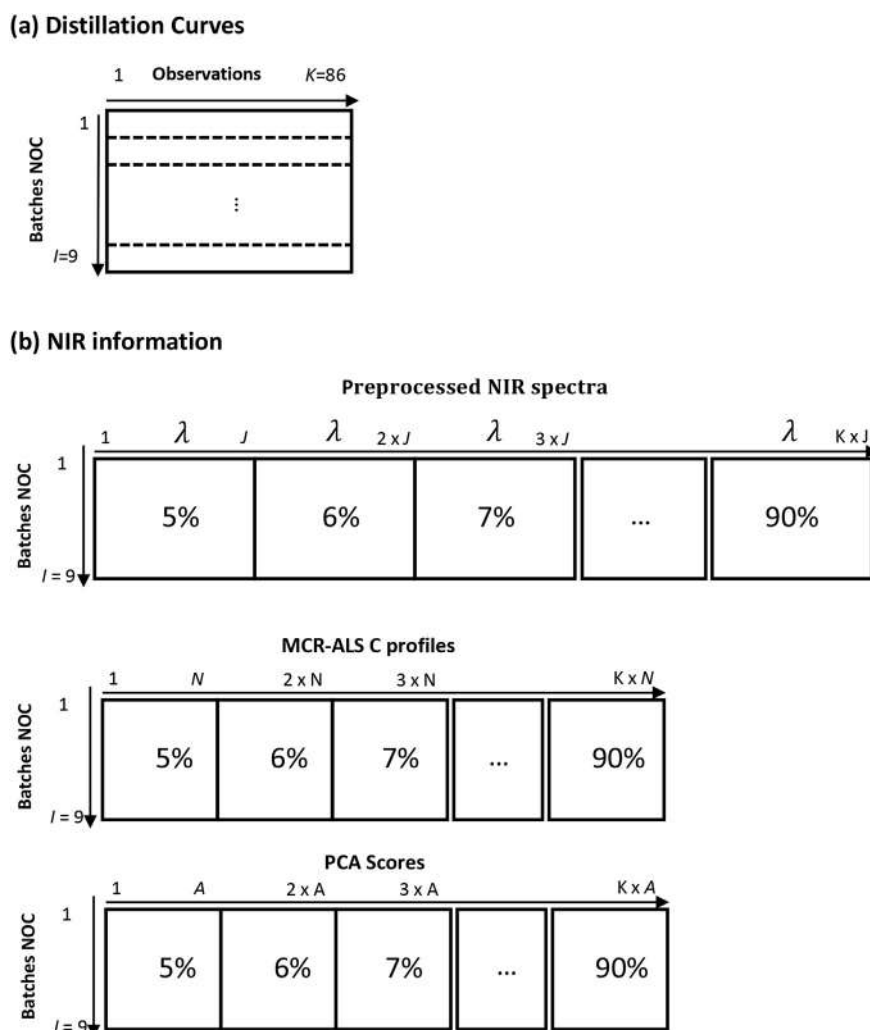
#### a) Off-line batch MSPC models using distillation curves

The distillation curves from the 9 NOC distillation batches were arranged in a matrix  $(I \times K)$ , with  $I = 9$  rows and  $K = 86$  observation points of the distillation curve. This matrix was mean-centered and decomposed by PCA to obtain the model loadings and MSPC limits, see Fig. 4(a). New batch data were projected into the model to obtain the related statistical parameters ( $D_{\text{stat.}}$  and  $Q_{\text{stat.}}$ ).

#### b) Off-line batch MSPC models using NIRS data

Different off-line MSPC models were built with the NIRS-derived information. All models were built on data sets with  $I = 9$  rows and a variable number of columns depending on the kind of NIRS-derived information, see Fig. 4(b). This gave rise to three different MSPC models:

- i. **Models based on the original preprocessed NIR data matrix.** This model is done using a matrix containing the NIR readings from each individual NOC batch row-wise unfolded into a vector, i.e. the matrix of a batch with dimensions  $(K \times J)$ , where  $K = 86$  are batch observation points and  $J = 573$  wavelengths, is arranged in a row vector with dimension  $(1 \times KJ)$ , with  $K = 86$  and  $J = 573$ . Then, the information of  $I = 9$  NOC batches was arranged in a matrix sized  $(I \times KJ)$ , on which the MSPC model was built.
- ii. **Models based on the batch scores from PCA decomposition of the NOC multiset structure.** The information of a NOC batch are the scores obtained in the PCA model of the related NIR spectra, row-wise unfolded into a vector sized  $(1 \times KA)$  with  $A$  being the number of retained principal components. Then, the information of  $I = 9$  NOC batches was arranged in a matrix sized  $(I \times KA)$ , on which the MSPC model was built.
- iii. **Models based on the resolved concentration profiles from MCR-ALS decomposition of the NOC multiset structure.** The information of a NOC batch are the concentration profiles obtained in the MCR-ALS model of the related NIR spectra, row-wise unfolded into a vector sized  $(1 \times KN)$  with  $N$  being now the number of MCR contributions needed to describe the process. Then, the information of  $I = 9$  NOC batches was arranged in a matrix sized  $(I \times KN)$ , on which the MSPC model was built. Please note that, generally speaking, the use of only some of the concentration profiles modeled in a batch could be an option to build the MSPC model, provided that the selected profiles were proven to be very specific indicators of the process evolution or that the discarded profiles belonged to spurious process contributions, e.g., modeled background contributions if existing. Please note that even if the use of C-profiles implies a noise-filtered compression of the original information, the size of the unfolded profiles, sized  $(1 \times KN)$  per each NOC batch, requires a PCA-based MSPC model for easier interpretability.



**Fig. 4.** Different starting information used to build off-line PCA-based batch MSPC models (a) Distillation curves, (b) NIR information, from top to bottom: Original preprocessed NIR variables, concentration profiles from MCR-ALS and scores from PCA extracted from the multibatch structure for process modeling.

The MSPC PCA models built with the different kinds of starting information were used to extract the related  $D_{stat.}$  and  $Q_{stat.}$  charts limits. Suitable data from new batches, not used to build the model, were projected onto the MSPC PCA model to test the performance of the models built.

### 3.2.4. On-line batch MSPC

Different on-line MSPC monitoring charts were developed using the data provided from NIRS measurements. As in the off-line approach, the same unfolded NOC matrix with the original NIR variables was used in the on-line approach. However, three on-line MSPC approaches were proposed using multiple PCA models based on different intervals of observation points, as described below:

#### a) On-line MSPC based on individual process observation models

This approach is the most straightforward method. An individual model is built per each observation point using historical data from on-specification completed batches as illustrated in Fig. 5(a). Thus, during a new batch, the new on-line data obtained (NIR spectrum of current observation) is projected into the respective observation point model and the statistical parameters compared with the control chart limits.

#### b) On-line MSPC based on evolving MSPC models

MSPC models with increasing number of observation points are built adding the new current distillation point in every new model until all distillation process is covered. As illustrated in Fig. 5(b), the first MSPC model is built using only the NIRS data matrix of the NOC historical data batches at the first recovered point (5 wt%), the second model using two observation points (5 and 6 wt%) and so on. For new batch monitoring, the data up to the current observation point are projected into the model for the related observations points and statistically tested.

#### c) On-line MSPC based on fixed size moving window, FSMW-MSPC, models

Several MSPC models built with a fixed size window (FSMW) including the current observation and several consecutive past observation points are built using the NOC historical data. The window slides one observation ahead in each new model until all observation points are covered. For instance, in Fig. 5(c) the window moves from  $k$  to  $k + 1$  and so on until  $k = K$ . For new batch monitoring, the data from the observation points covered by the moving window are projected into the model for the respective observations points and statistically tested.

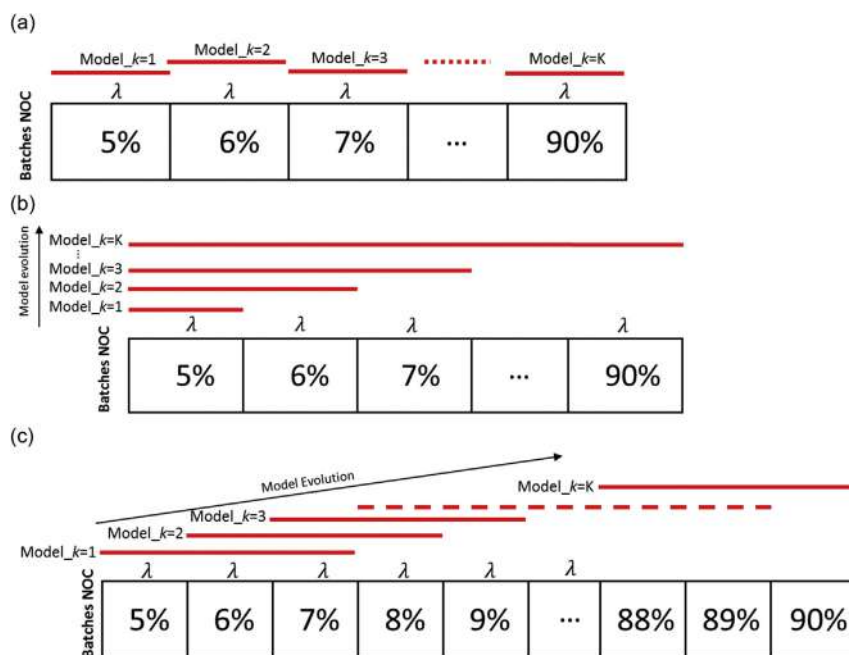


Fig. 5. Different evolving on-line MSPC models approaches. a) Individual process observation models, b) Evolving MSPC models and c) FSMW-MSPC models.

The three approaches aim at on-line process control, but there are important differences due to the use of the different information in the models. Thus, the modality looking at individual process points does not take into account the neighbouring past observations and, hence, the evolution of the process. In the modalities FSMW-MSPC and evolving MSPC, the process evolution is taken into account and not only the new process observation of interest. In the case of the FSMW-MSPC model, only the recent past observation points (those within the window) are taken into account and the window size established is related to the number of relevant neighbouring process observations. Instead, the evolving-MSPC takes into account all process evolution until the present observation, giving potentially the same importance to all the past observations analyzed.

## 4. Results and discussion

### 4.1. Visual interpretation of distillation curve and NIR process data

Prior to chemometric analysis, the distillation curve and raw NIR spectra obtained during the distillation process were visually interpreted. Fig. 2(a) illustrates the distillation curve of an on-specification batch (B07), the related process raw NIR spectra, Fig. 2(b), and NIR spectra selected at four specific distillation points, Fig. 2(c).

A sudden change in temperature can be observed through a simple visual inspection of the distillation curve between 60 and 70 wt%. This behavior is observed in gasoline-ethanol blends due to the formation of azeotropes of ethanol and hydrocarbons [37–40]. Distillation curve for gasoline-ethanol blend show three distinct regions: a plateau or azeotropic region (ethanol-hydrocarbon azeotropes are boiled) in the beginning of the distillation process, a transition region (sudden change in temperature) and a dilution-only region at the end of the distillation (after all added ethanol is boiled-off), as observed by French and Malone [38].

Four observation points (10, 50, 70 and 90 wt%) at the start and end of each distillation region were chosen to visualize the changes in NIR spectra with the evolution of distillation process. Fig. 2(c)

shows the complexity of the many superimposed absorption bands of the NIR spectra acquired during the distillation process. The bands around 1180 nm correspond to the second overtone, around 1400 nm to the 1st overtone combination and around 1700 nm to the first overtone region of carbon-hydrogen (C-H) bonds present in all points observed. The band around 2080 nm observed in the fractions at 10 and 50 wt% is related to the absorption of a combination of oxygen-hydrogen (O-H) stretching and bending from ethanol added to the gasoline. An absorbance increment in the band around 2080 nm was observed as the distillation was evolving from 10 to 50 wt%, mainly related to the increase of the ethanol relative concentration in the distilled fractions. A new band around 2170 nm appears in the spectra of the fraction at 70 and 90 wt%. This new band is related to absorption of aromatic compounds in the heavy fractions of the gasoline [41–43].

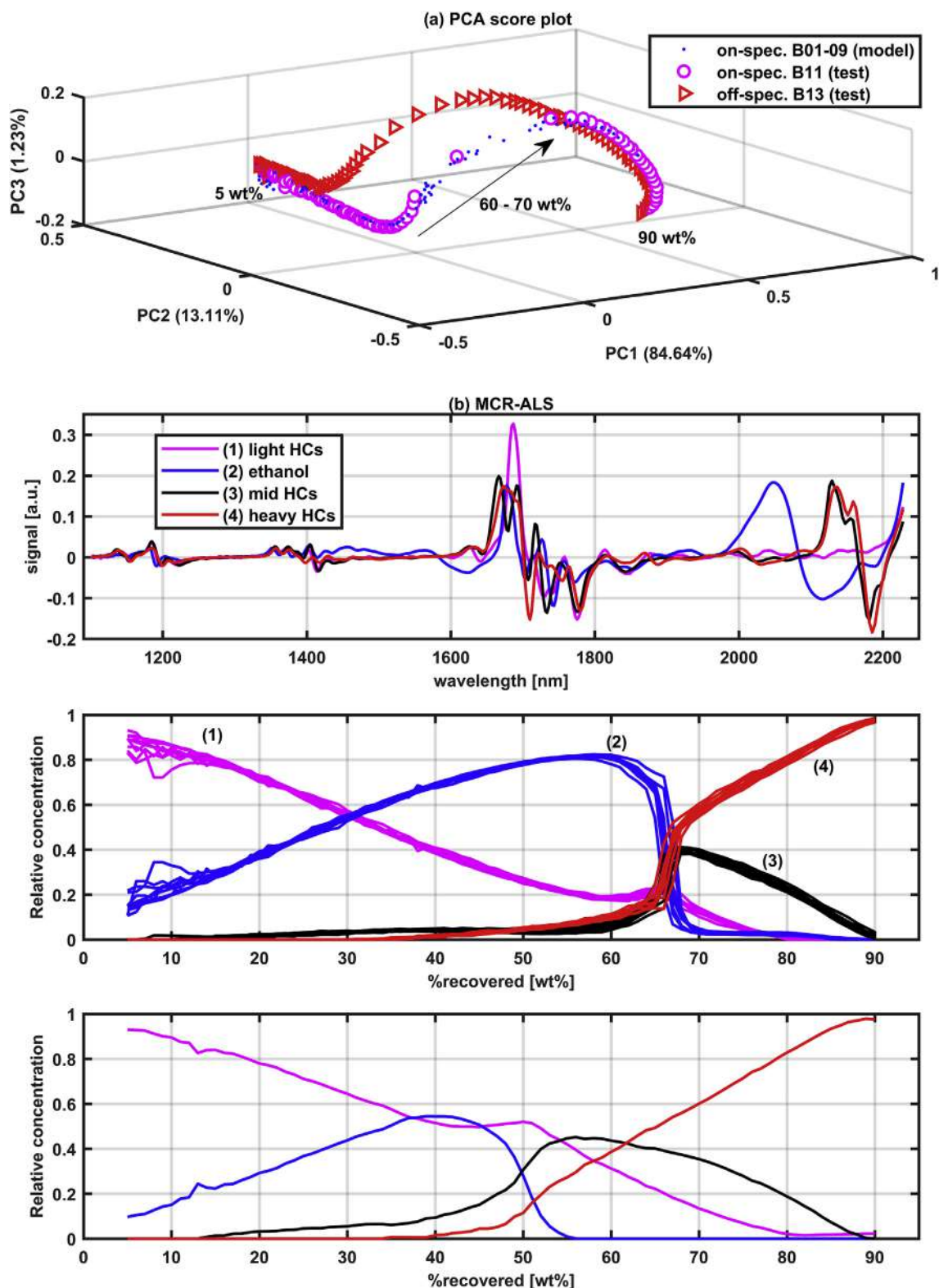
### 4.2. Process modeling of NIR data

#### 4.2.1. Global process description (PCA model)

The NIR data were mean-centered and decomposed by PCA. Venetian blinds cross-validation method was used to find the number of principal components. 3 principal components explained 98.98% (PC1 84.64%, PC2 13.11% and PC3 1.23%) representing a good summary of batch variability.

Fig. 6(a) shows the principal components score plot distribution for PC1 and PC2 extracted from NIRS data of batches B01–09 used to build the PCA model (blue dots). The distribution of scores illustrates the process trajectory of on-specification batches and its variability. Because of the unstable distillation rate at the start of the distillation process, more variation was observed in these observation points as compared to the rest of the process. In addition, the NIRS data collected from the distillations of on-specification gasoline batch B11, not used in the PCA process modeling, and off-specification batch B13, which had only 15%(v/v) of ethanol added, were projected in the PCA model. The scores obtained from PCA projection allowed the observation of the process trajectory of the new projected batches. Batch B11 (in magenta circles) was observed to follow the same on-specification process





**Fig. 6.** Process modeling. (a) PCA map of the 3 PC's scores from on-specification batches B01-09 used to build PCA model, and after projection in the PCA model batches B11 (on-specification) and B13 (off-specification). Start (5 wt%), transition region (60–70 wt%) and end (90 wt%) of distillation process are indicated; (b) Multibatch MCR-ALS showing from top to bottom the pure spectral profiles, the superimposed concentration profiles for on-specification batches B01-09 and concentration profiles for off-specification batch B13 (components (1) to (4)).

trajectory, while batch B13 (in red triangles) deviated from NOC trajectory, as illustrated in Fig. 6(a). On-specification batch B10 when projected to the PCA model showed the same behavior as B11. Off-specification batches B12, B14–23 also deviate from NOC trajectory as batch B13, (data not shown for clarity). The deviation becomes larger when the ethanol content is further from the ethanol specification level of NOC batches.

#### 4.2.2. MCR-ALS

The dataset decomposition through MCR-ALS provides a model of process components easy to interpret and complementary to the global process description provided by PCA. The multibatch structure with the preprocessed (not mean-centered) data obtained from the distillation batches was decomposed by MCR-ALS. Four components were found through singular value decomposition, which agrees with the three contributions found in PCA of mean-centered data, since the rank decreases in one when mean centering is performed.

The four components concentration (distillation) and spectral profiles obtained after MCR-ALS decomposition of the multiset structure are shown in Fig. 6(b). The components resolved from the distillation process are related to the main distilled fractions of gasolines “type C”: First, light hydrocarbons; second, ethanol; third and fourth, mid to high molecular weight (MW) hydrocarbons and aromatic compounds, as reported elsewhere [44]. The identity of these compounds is confirmed when looking at the spectral features found in the related pure spectra and at the temperature distillation range.

The low MW hydrocarbons fraction is mainly distilled together with ethanol as azeotropes at the beginning of the distillation, i.e., at lower temperatures, as observed in the concentration profiles of components (1) and (2), see Fig. 6(b). After 70 wt% of the distillation process, almost all ethanol, component (2), was boiled-off remaining most of the mid to high MW fractions of gasoline, rich in aromatic compounds, components (3) and (4). This region was observed in the distillation curves and is characterized by an increase in the slope of the distillation curve, as observed in Fig. 2(a).

For comparison, Fig. 6(b) shows the distillation profiles of B13 (with only 15%(v/v) ethanol). Although the component spectra are the same, all distillation profiles are shifted to lower wt% of distillate, as expected for a batch with lower ethanol content.

### 4.3. Process control

#### 4.3.1. Off-line batch process control

Off-line batch MSPC charts were built working with data coming from completed distillation batches. Specificity and sensitivity were adopted as quality parameters to assess the performance of MSPC charts for off-line batch process control. Specificity stands for the ratio of NOC batches (on-specification) correctly identified over the total NOC batches used to test the MSPC charts. Sensitivity is derived as the ratio of out of NOC (off-specification) batches correctly identified as out of NOC over the total out of NOC tested.

##### a) Process control starting information

The starting information used to build off-line batch MSPC models came either from distillation curves or NIR process data. The different starting information is depicted in section 3.2. Full distillation curves or observations within a selected temperature range were used to build off-line PCA-based MSPC models. Derivative form of the distillation curves was also used to improve the models. As for NIR information, full original preprocessed NIR spectra or selected spectral ranges were used to build the models. MSPC models were also built with the PCA scores, extracted from

the process modeling by PCA, with all the distillation concentration profiles or only with the component related to ethanol, extracted from MCR-ALS decomposition, as described in section 3.2.

##### b) Off-line batch MSPC results

Table 2 shows the summary of the results using the different starting information to build and test off-line batch MSPC models.

An MSPC PCA model with mean-centered full distillation curve data (5–90 wt%) from NOC batches was built with 2 PC's and explained 90.91% of data variance. MSPC chart based on  $Q_{stat.}$  parameter correctly identified NOC and off-spec batches used to test the control charts as observed in Table 2 (row #1 has 100% specificity and sensitivity of  $Q_{stat.}$ ). However, despite  $D_{stat.}$  chart correctly identifies NOC batches, some off-spec batches are below the  $D_{stat.}$  limit, see Fig. 7(b), the sensitivity observed was 73.33%, Table 2 row #1. This may have happened because distillation curves of off-specification batches with ethanol concentration near to the on-specification level, 27%(v/v), have extensive distillation ranges with similar behavior (except for the points in the steepest zone of the curve) and, when considered the full curve, stay within the accepted variability of the NOC batches.

Another PCA model was built using the same data used previously, but this time preprocessed by Savitzky-Golay derivative and mean-centered. Results showed an improvement on the sensitivity, but still some batches were misidentified in the  $D_{stat.}$  chart, as reported in Table 2 row #2. All off-specification batches could be correctly identified using the derivative curve data only in the distillation range between 25 and 75 wt%, (Table 2 row #3). This range showed most of the variation in the distillation curves due to different ethanol content and avoided the instability and, hence, undesired and non-composition related variability in the beginning of the distillation.

As observed in the MSPC charts built with the distillation curve data, the specificity and sensitivity of the  $Q_{stat.}$  charts for all models built with NIRS data were 100%. However, different strategies were necessary to improve the sensitivity of  $D_{stat.}$  MSPC charts.

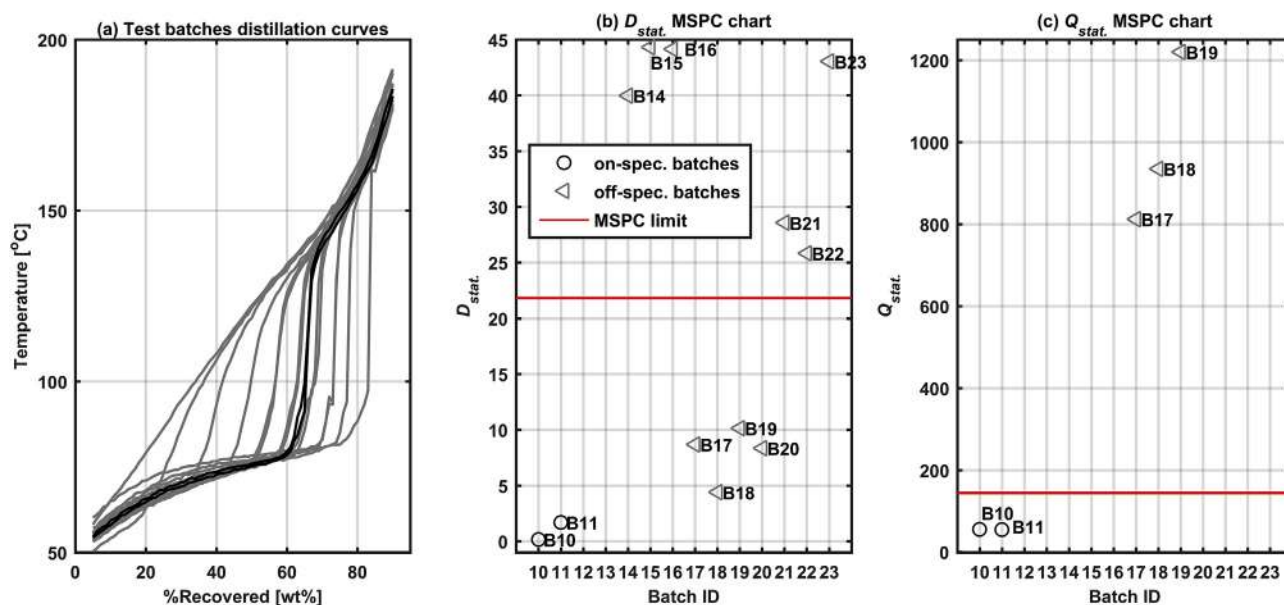
The off-line PCA-based batch MSPC charts built using information from NIR spectra are explained below. Table 2, row #4, shows the results from a model built using the full preprocessed spectra (1103–2228 nm) and distillation (5–90 wt%) range. Despite of the 100% specificity in  $Q_{stat.}$  chart, none of the off-specification batches was detected as faulty by the  $D_{stat.}$  chart. The  $D_{stat.}$  MSPC chart sensitivity was significantly improved to 75% when the NIRS data were reduced taking only the NIR observations within the distillation range from 60 wt% to 70 wt%, see Table 2, row #5. NIRS data were also reduced by selecting the most expressive spectral bands related mainly to hydrocarbons (1600–1800 nm) and ethanol absorption regions (2000–2200 nm). The MSPC chart built with this reduced spectral and distillation range improved the  $D_{stat.}$  sensitivity to 83% (row #6), but still some samples with composition similar to the on-specification batches were missed by the control chart.

Off-line MSPC models were built with the NIR information compressed by PCA and MCR-ALS. Similar results were observed. The sensitivity for  $D_{stat.}$  MSPC charts built with concatenated PCA scores or MCR-ALS concentration profiles (row #7 and #10) improved when compared with full spectral and distillation range data without data compression (row #4), see Table 2. MSPC models built with the compressed information extracted from the NIR observations within the 60–70 wt% distillation range showed an expressive improvement of the  $D_{stat.}$  sensitivity to 91.67% (row #8 and #11).  $D_{stat.}$  charts (row #9 and #12) showed 100% sensitivity when MSPC models were built using the Savitzky-Golay derivative of the PCA scores or the MCR-ALS concentration profiles within the

**Table 2**  
Off-line batch MSPC results.

#	Description	NC	%EV	%Specificity		%Sensitivity	
				$D_{stat.}$	$Q_{stat.}$	$D_{stat.}$	$Q_{stat.}$
<b>Distillation Curves</b>							
1	DistRange (5–90 wt%)	2	90.91	100	100	66.67	100
2	DistRange(5–90 wt%)_1 <sup>st</sup> diff.SG	2	94.83	100	100	91.67	100
3	DistRange(25–75 wt%)_1 <sup>st</sup> diff.SG	2	98.36	100	100	100	100
<b>NIR information</b>							
<b>Unfolded NIRS data</b>							
4	FullSpec_DistRange(5–90%)	3	68.91	100	100	0.00	100
5	FullSpec_DistRange(60–70%)	2	92.56	100	100	75.00	100
6	SelSpec_DistRange(60–70%)	2	94.54	100	100	83.33	100
<b>scores from PCA modeling (3 PC's)</b>							
7	DistRange(5–90%)	2	87.53	100	100	66.67	100
8	DistRange(60–70%)	2	96.52	100	100	91.67	100
9	DistRange(60–70%)_1 <sup>st</sup> diff.SG	2	98.67	100	100	100	100
<b>C profiles from MCR-ALS modeling (4comp)</b>							
10	DistRange(5–90%)	2	81.07	100	100	33.3	100
11	DistRange(60–70%)	2	94.33	100	100	91.67	100
12	DistRange(60–70%)_1 <sup>st</sup> diff.SG	2	96.99	100	100	100	100
13	DistRange(5–90%)_EtOHcomp	2	90.60	100	100	58.33	100
14	DistRange(60–70%)_1 <sup>st</sup> diff.SG_EtOHcomp	2	99.65	100	100	100	100

# Model number, **NC** number of PCA principal components, **%EV** cumulative explained variance by NC principal components, **diff.SG** Savitzky-Golay derivative, **DistRange** Distillation Range used to build the model, **FullSpec** Complete NIRS measurement range, **SelSpec** Small more selective to ethanol signal, **EtOHcomp** component 2 related to ethanol.



**Fig. 7.** (a) Full distillation curves used to test PCA-based off-line batch MSPC model, on-specification in black and off-specification in gray, (b)  $D_{stat.}$  and (c)  $Q_{stat.}$  MSPC charts. Some batches show higher  $D_{stat.}$  and  $Q_{stat.}$  values and are not shown for better visualization of the control limits.

same distillation range (60–70%). MSPC models were built also using only the ethanol distillation profile. Results are shown in Table 2, rows #13 and #14. The  $D_{stat.}$  sensitivity was higher than in models built with all four components for models built with the full distillation range. Moreover, when the derivative ethanol profile in the 60–70 wt% distillation range, 100% specificity in  $D_{stat.}$  chart was achieved. The improvement of results when using only the ethanol concentration profile might be related to the better definition of this compound in the MCR-ALS results.

#### 4.3.2. On-line batch MSPC on the NIR data

On-line batch MSPC control charts were built following the strategies described in section 3.2. For the distillation batches

studied, PCA models were calculated for each observation point (86 models) following each one of the strategies described using the mean-centered data collected from NOC batches. Individual observation models (see Fig. 5(a)) and evolving models (see Fig. 5(b)) were calculated as described. For FSMW evolving models (see Fig. 5(c)), the window selected enclosed 15 neighbouring observations. Thus, for observations nr. 1 to 14, PCA models were calculated as in the evolving strategy (see Fig. 5(b)), whereas from observation nr. 15 and on, the full sliding window of 15 points was applied, as seen in Fig. 5(c). PCA models for individual observation and FSMW evolving strategies were built with one principal component for all observation points, while in evolving models, one PC was used in evolving models from 5 to 20 wt% and three PC's

in the remaining observation points. A confidence interval of 99% was considered to calculate the MSPC charts limits,  $D_{CL99\%}$  and  $Q_{CL99\%}$ , for each model, as described earlier in section 3.2.

The NIR measurements for a new batch observation were mean-centered according to the mean of NOC batches and each observation (or set of observations) projected into the PCA model built for each strategy to extract the MSPC statistics,  $D_{stat.}$  and  $Q_{stat.}$ .

At this point, it is important to stress the difference between off-line and on-line MSPC control charts.

Off-line MSPC control charts are based on a single PCA model built on the completed NOC batches. The final  $D_{stat.}$  and  $Q_{stat.}$  charts represent the values of these statistics vs. the batch index of each analyzed new batch. Every new batch is represented by a point.

On-line MSPC control charts display simultaneously the information of many PCA models, as many as observations in each batch, see Fig. 5. Therefore, each new observation (NIR spectrum) acquired in a new batch is tested to see whether it is in- or out of control on a different PCA model. The process control is done at an observation level and not at a full batch level, as in the off-line approach. As a consequence, every new batch has a full  $D_{stat.}$  and a full  $Q_{stat.}$  plot, where the x-axis refers now to the different observations studied along the process evolution.

Control limits in  $D_{stat.}$  and  $Q_{stat.}$  would change per each new observation analyzed, since a different PCA model is used for projection every time. To facilitate visualization, the y-axis represents scaled values of  $D_{stat.}$  and  $Q_{stat.}$ , defined as  $D_{stat.}/D_{CL}$  and  $Q_{stat.}/Q_{CL}$ . In this way, a flat line at value 1 represents the control limits for all models used in  $D_{stat.}$  and  $Q_{stat.}$  for all observations. On-line  $D_{stat.}$  and  $Q_{stat.}$  charts, which represent the evolution of the related scaled statistics as a function of the observation (% distillate) analyzed, allow not only identifying on- and off-specification batches, but to know when the anomaly in an abnormal batch starts.

The results after monitoring new batches through the three different on-line MSPC strategies (individual observation model, FSMW MSPC evolving models and evolving MSPC models) are summarized in Table 3. Table 3 shows whether a new batch was diagnosed as on-specification or not and which MSPC chart ( $D_{stat.}$ ,  $Q_{stat.}$  or both) detected the fault. (Please note that the behavior of the full distillation batch is analyzed in this section for a better comparison of the three approaches. In a real on-line control

context, the distillation would be stopped as soon as found to be out of specification).

Observing the information summarized in Table 3,  $Q_{stat.}$  on-line MSPC charts detected correctly a fault in all off-specification batches by using any of the three on-line strategies.  $D_{stat.}$  charts worked generally well, except when using evolving models, which were not able to detect fault in off-specification batches with ethanol content very approximate to the accepted specification and above, i.e. batches B17 to B23, see Tables 1 and 3. The on-specification batch B11 was wrongly detected as faulty by the  $Q_{stat.}$  on-line MSPC chart using the individual observation model MSPC strategy.

Fig. 8 shows the  $D$  and  $Q$  statistics on-line control charts from distillation batches B11 (on-specification, with 27% ethanol added) and B18 (off-specification, with 25% ethanol added) for the three different on-line batch MSPC strategies.

Some comments need to be done for each on-line strategy according to the observed results.

#### a) Individual process observation models

$Q_{stat.}$  MSPC charts were observed to be very sensitive to fault detection. Besides, they show clearly the point where the batch starts to be anomalous. However,  $Q_{stat.}$  charts were more prone to show false alarms, Fig. 8(a). This happens because each model was built using a single observation point and a slight variation in an individual observation for a new batch process leads to a fault detection.

#### b) Evolving MSPC models

The evolving MSPC strategy considered the evolution of the process since the start, building models with increasing number of observations. This caused less sensitive  $D_{stat.}$  charts, since past NOC observations may have a lot of weight in the models and batches can be detected easily as faulty only when the fault observations occurred at the beginning of the distillation process (batches B11–16). Batches with ethanol concentration near to the specification value and above were not detected by  $D_{stat.}$  charts since the fault occurred too late and was not large enough to compensate the weight of the large number of initial NOC observations. Despite of this fact, the evolution of  $D_{stat.}$  values for undetected off-specification batches show a different trend (a clear increase when the abnormal behavior starts) as compared to on-specification batches (presenting a flat constant tendency), as observed in  $D_{stat.}$  charts Fig. 8(b). This may suggest that the  $D_{stat.}$  chart could still be used in these instances if the control limits were set empirically.  $Q_{stat.}$  charts performed more satisfactorily in fault detection. However, faults were detected later than in individual observation models due to the excessive weight of past NOC observations as well.

#### c) Fixed size moving window, FSMW-MSPC models

The FSMW strategy considered only a few past observation points, set according to the window size (15 observations were used in this study). This feature produced more sensitive  $D_{stat.}$  charts because past NOC observations had less weight in models, Fig. 8(c). FSMW strategy was observed to be less prone to false alarms on  $Q_{stat.}$  charts than individual observation charts, Fig. 8(a), since individual point fluctuations have less impact in the window-based PCA models. This strategy has been found to be the most flexible of the three, showing efficient and easy detection of faults and avoiding false alarms. Obviously the performance of this approach may depend on the width of the window: if too small,

**Table 3**  
On-line batch MSPC results on test batches B10–23.

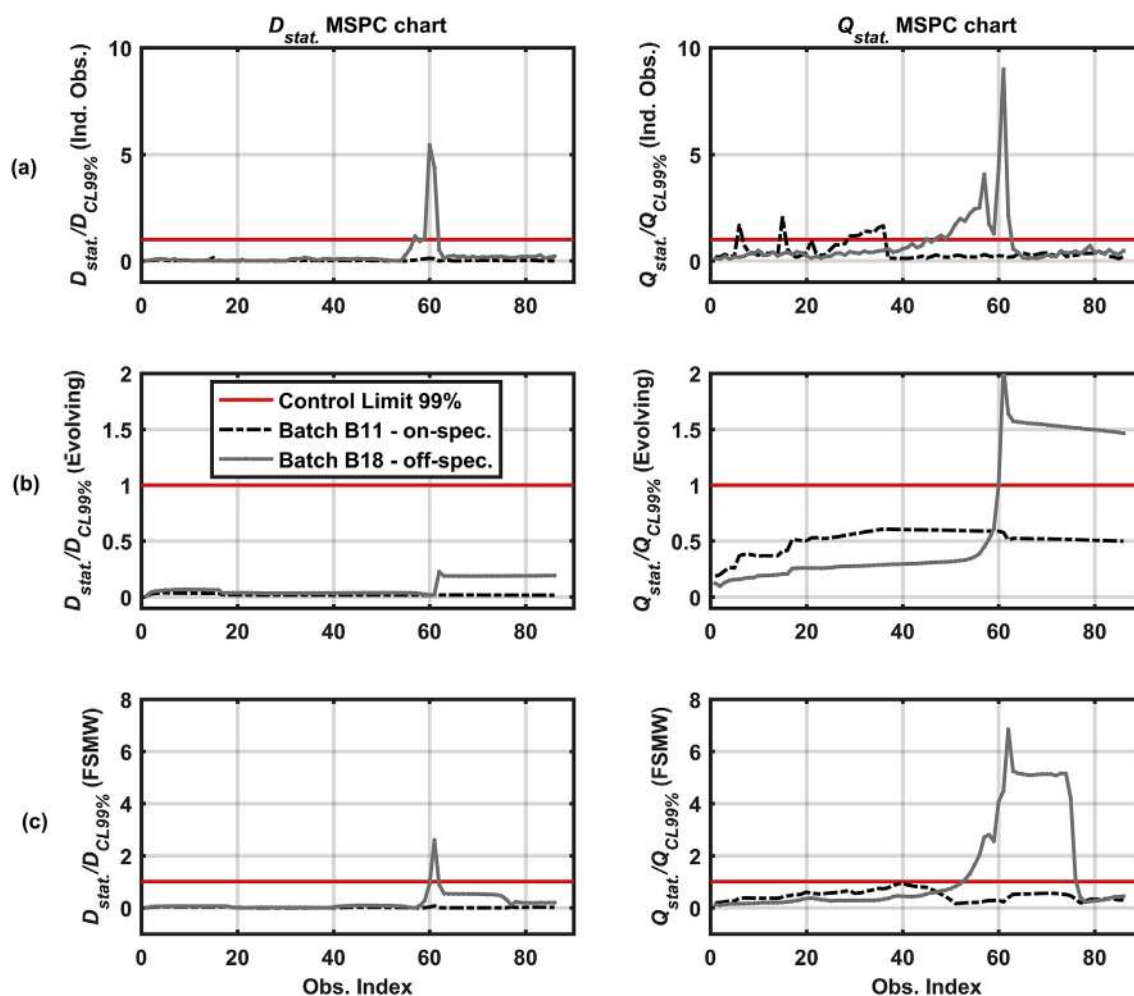
Test Batch	Method		
	Ind. Obs. <sup>a</sup>	FSMW <sup>b</sup>	Evolving <sup>c</sup>
<b>on-spec</b>			
B10	on-spec. <sup>d</sup>	on-spec.	on-spec.
B11	$Q_{stat.}$	on-spec.	on-spec.
<b>off-spec</b>			
B12	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$
B13	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$
B14	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$
B15	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$
B16	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$
B17	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}$
B18	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}$
B19	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}$
B20	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}$
B21	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}$
B22	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}$
B23	$Q_{stat.}, D_{stat.}$	$Q_{stat.}, D_{stat.}$	$Q_{stat.}$

<sup>a</sup> Ind. Obs., Individual observation MSPC models.

<sup>b</sup> FSMW, fixed size moving window MSPC models.

<sup>c</sup> Evolving MSPC models.

<sup>d</sup> On-spec means the batch is on-specification according to both  $D_{stat.}$  and  $Q_{stat.}$ .  $D_{stat.}$  means the batch is off-spec according to  $D_{stat.}$  chart.  $Q_{stat.}$  means the batch is off-spec. according to  $Q_{stat.}$  chart.



**Fig. 8.** On-line MSPC charts for batches B11 (on-specification) and B18 (off-specification) for the a) Individual process observation models, b) Evolving MSPC strategies and c). FSMW evolving MSPC.

false alarms may show up in analogy to what happens in individual observation models; if too big, sensitivity in  $D_{stat.}$  chart may decrease because of the weight of too many past NOC observations in the chart. This inconvenience is clearly surmountable if the window width is set by using representative off-spec batches that may allow setting the correct window width to avoid the malfunctions described in the other two on-line MSPC approaches.

## 5. Conclusions

The present work provides an improvement of PAT technologies for distillation-based quality control procedures through the design of an automatic distillation device that allowed synchronized measurements of the distilled mass percentage, distillation temperature and NIR spectra during the distillation process.

Process modeling on NIR spectra by PCA and MCR-ALS allowed understanding the process evolution from a global (scores plot) and component-wise (distillation profiles) point of view, respectively. In this sense, MCR-ALS provides a good thermal and physico-chemical characterization of distilled fractions, even if coming from a simple distillation process.

MSPC strategies based on the different kinds of data obtained from the designed device are proposed. Off-line models using distillation curves were able to detect off-specification batches when suitable preprocessing and distillation curve range were

used. Successful off-line MSPC models were built with the NIR spectra information compressed into PCA scores or MCR-ALS concentration profiles. The possibility to perform a sensible selection of some of the MCR-ALS concentration profiles, linked to particularly relevant process contributions has proven to improve MSPC results.

On-line batch MSPC strategies were proposed for fault detection during the distillation process using the collected NIRS data. Individual process observations MSPC models showed  $D_{stat.}$  charts very sensitive to fault detection; however, false alarms were observed in the  $Q_{stat.}$  charts. Evolving MSPC models were able to solve the false alarms observed with the individual observation strategy, but failed to detect some off-specification batches with similar composition to NOC batches when using  $D_{stat.}$  charts. The FSMW-MSPC approached used a flexible combination of the other two strategies and succeeded to detect all off-specification batches and correctly identify on-specification batches during the test with both  $D_{stat.}$  and  $Q_{stat.}$  control charts, avoiding false alarms.

## Acknowledgements

R.R. de Oliveira acknowledges the EMQAL Grant. EMQAL is a Joint European Master Programme selected under Erasmus Mundus coordinated by University of Barcelona. Funding support from the European Community's Framework programme for Research and Innovation Horizon 2020 SPIRE - Integrated Process Control

(ProPAT), grant Agreement no. 637232 and Spanish government through project CTQ2015-66254-C2-2-P are also acknowledged. K.M.G. Lima acknowledges the CNPq (Grant 305962/2014–4) for financial support. A.O. Sousa acknowledges the Petrobras (Grant 2012/00217-7) for financial support.

## References

- [1] ASTM D86-15, Standard Test Method for Distillation of Petroleum Products and Liquid Fuels at Atmospheric Pressure, ASTM B. Stand., 2015.
- [2] C. Pasquini, S.H.F. Scafi, Real-time monitoring of distillations by near-infrared spectroscopy, *Anal. Chem.* 75 (2003) 2270–2275.
- [3] V.A. Corro-Herrera, J. Gómez-Rodríguez, P.M. Hayward-Jones, D.M. Barradas-Dermitz, M.G. Aguilar-Uscanga, A.C. Gschaedler-Mathis, *In-situ* monitoring of *Saccharomyces cerevisiae* ITV01 bioethanol process using near-infrared spectroscopy NIRS and chemometrics, *Biotechnol. Prog.* 32 (2016) 510–517.
- [4] H. Xiong, X. Gong, H. Qu, Monitoring batch-to-batch reproducibility of liquid-liquid extraction process using in-line near-infrared spectroscopy combined with multivariate analysis, *J. Pharm. Biomed. Anal.* 70 (2012) 178–187.
- [5] RESOLUÇÃO ANP No 40, DE 25.10.2013 DOU 28.10.2013, [http://nxt.anp.gov.br/NXT/gateway.dll/leg/resolucoes\\_anp/2013/outubro/ranp40-2013.xml?fn=templates\\$fn=document-frame.htm\\$3.0\\$Q=\\$x=\\$nc=2230](http://nxt.anp.gov.br/NXT/gateway.dll/leg/resolucoes_anp/2013/outubro/ranp40-2013.xml?fn=templates$fn=document-frame.htm$3.0$Q=$x=$nc=2230). Accessed August 1, 2016.
- [6] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52.
- [7] R. Tauler, M. Maeder, A. de Juan, Multiset data analysis: extended multivariate curve resolution, in: *Compr. Chemom. Chem. Biochem. Data Anal. Four-Volume Set, vol. 2*, Elsevier, 2009, pp. 473–505. Chapter 2.24, S.D. Brown, R. Tauler, B. Walcz.
- [8] Y. Jin, Z. Wu, X. Liu, Y. Wu, 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 (2013) 32–39.
- [9] M.C. Sarragaça, P.R.S. Ribeiro, A.O. dos Santos, J.A. Lopes, Batch statistical process monitoring approach to a cocrystallization process, *J. Pharm. Sci.* 104 (2015) 4099–4108.
- [10] H. Howland, S.W. Hoag, Analysis of curing of a sustained release coating formulation by application of NIR spectroscopy to monitor changes physical-mechanical properties, *Int. J. Pharm.* 452 (2013) 82–91.
- [11] R. Kona, H. Qu, R. Mattes, B. Jancsik, R.M. Fahmy, S.W. Hoag, Application of in-line near infrared spectroscopy and multivariate batch modeling for process monitoring in fluid bed granulation, *Int. J. Pharm.* 452 (2013) 63–72.
- [12] T. Kourti, J.F.J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, *Chemom. Intell. Lab. Syst.* 28 (1995) 3–21.
- [13] A. AlGhazzawi, B. Lennox, Monitoring a complex refining process using multivariate statistics, *Control Eng. Pract.* 16 (2008) 294–307.
- [14] A. AlGhazzawi, B. Lennox, Model predictive control monitoring using multivariate statistics, *J. Process Control* 19 (2009) 314–327.
- [15] T.C. Ávila, R.J. Poppi, I. Lunardi, P.A.G. Tizei, G.A.G. Pereira, Raman spectroscopy and chemometrics for on-line control of glucose fermentation by *Saccharomyces cerevisiae*, *Biotechnol. Prog.* 28 (2012) 1598–1604.
- [16] J. Alves-Rausch, R. Bienert, C. Grimm, D. Bergmaier, Real time in-line monitoring of large scale *Bacillus* fermentations with near-infrared spectroscopy, *J. Biotechnol.* 189 (2014) 120–128.
- [17] H. Huang, H. Qu, In-line monitoring of alcohol precipitation by near-infrared spectroscopy in conjunction with multivariate batch modeling, *Anal. Chim. Acta* 707 (2011) 47–56.
- [18] P. Nomikos, J.F. MacGregor, Multivariate statistical process control charts for monitoring batch processes, *Technometrics* 37 (1995) 41–59.
- [19] P. Nomikos, J.F. MacGregor, Monitoring batch processes using multiway principal component analysis, *AIChE J.* 40 (1994) 1361–1375.
- [20] P. Nomikos, J.F. MacGregor, Multi-way partial least squares in monitoring batch processes, *Chemom. Intell. Lab. Syst.* 30 (1995) 97–108.
- [21] S. Wold, N. Kettaneh, H. Friden, A. Holmberg, Modelling and diagnostics of batch processes and analogous kinetic experiments, *Chemom. Intell. Lab. Syst.* 44 (1998) 331–340.
- [22] S. Rännar, J.F. MacGregor, S. Wold, Adaptive batch monitoring using hierarchical PCA, *Chemom. Intell. Lab. Syst.* 41 (1998) 73–81.
- [23] L. Zhao, T.-Y. Chai, G. Wang, A Nonlinear Modeling and Online Monitoring Method for the Batch Process Using Multiple Local PCA, 2003, pp. 2–5.
- [24] L. Zhao, T. Chai, Adaptive moving window MPCA for online batch monitoring, in: 2004 5th Asian Control Conf, vol. 2, 2004, pp. 1290–1295.
- [25] M. Maeder, Evolving factor analysis for the resolution of overlapping chromatographic peaks, *Anal. Chem.* 59 (1987) 527–530.
- [26] M. Maeder, A. de Juan, Two-way data analysis: evolving factor analysis, *Compr. Chemom.* 2 (2010) 261–274.
- [27] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Anal. Chem.* 36 (1964) 1627–1639.
- [28] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, *Anal. Methods* 6 (2014) 4964.
- [29] A. de Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures, *Anal. Chim. Acta* 500 (2003) 195–210.
- [30] R. Tauler, B.R. Kowalski, S. Fleming, Multivariate curve resolution applied to spectral data from multiple runs of an industrial process, *Anal. Chem.* 65 (1993) 2040–2047.
- [31] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, *Anal. Chem.* 63 (1991) 1425–1432.
- [32] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, *Chemom. Intell. Lab. Syst.* 140 (2014) 1–12.
- [33] T. Kourti, 4.02-Multivariate statistical process control and process control, using latent variables, in: *Compr. Chemom.* 2009, pp. 21–54.
- [34] S. Wold, Cross -validatory estimation of the number of components in factor and principal components models, *Technometrics* 20 (1978) 397–405.
- [35] J.E. Jackson, *A User's Guide to Principal Components*, Wiley, New York, 1991.
- [36] J.E. Jackson, G.S. Mudholkar, Control procedures for residuals associated with principal component analysis, *Technometrics* 21 (1979) 341–349.
- [37] E.V. Takeshita, R.V.P. Rezende, S.M.A.G.U. de Souza, A.A.U. de Souza, Influence of solvent addition on the physicochemical properties of Brazilian gasoline, *Fuel* 87 (2008) 2168–2177.
- [38] R. French, P. Malone, Phase Equilibria Ethanol Fuel Blends 229 (2005) 27–40.
- [39] R.M. Balabin, R.Z. Syunyaev, S.A. Karpov, Quantitative measurement of ethanol distribution over fractions of ethanol-gasoline fuel, *Energy Fuels* 21 (2007) 2460–2465.
- [40] V.F. Andersen, J.E. Anderson, T.J. Wallington, S.A. Mueller, O.J. Nielsen, Distillation curves for alcohol-gasoline blends, *Energy Fuels* 24 (2010) 2683–2691.
- [41] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, *Anal. Chim. Acta* 667 (2010) 14–32.
- [42] C. Pasquini, Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications, *J. Braz. Chem. Soc.* 14 (2003) 198–219.
- [43] D.A. Burns, E.W. Ciurczak, *Handbook of near-infrared analysis*, third ed. *Anal. Bioanal. Chem.* 393 (2009) 1387–1389.
- [44] J.L. Burger, N. Schneider, T.J. Bruno, Application of the advanced distillation curve method to fuels for advanced combustion engine gasolines, *Energy Fuels* 29 (2015) 4227–4235.



Every distillation batch monitored in this work came from a mixture of gasoline and ethanol and provided synchronized vapor temperature measurements and NIR spectra associated with equispaced %mass fraction distilled. The additional reference of the %mass fraction distilled allowed organizing the stored information per batch in such a way that all batch data matrices had the same length and, most important, the process trajectory for NOC batches was also synchronized, as described in detail in section 3.1.3. Thus, both types of multibatch data structures (batch- and variable-wise augmented multisets) could be used for process modeling and control, as described below.

### **Process modeling using PCA and MCR-ALS**

For batch process modeling, variable-wise augmented multisets were used, as shown in Figure 19B, and only the NIR data collected during the distillation process were analyzed. Thus, the data matrices with the preprocessed NIR data from nine on-specification (27% ethanol) gasoline NOC distillation batches were arranged into a variable-wise augmented multiset and modeled using principal component analysis (PCA) and multivariate curve resolution-alternating least squares (MCR-ALS). PCA provided a global model of trajectories explaining the overall process evolution, whereas MCR-ALS provided a model describing the evolution and chemical identity of each component (distinct distilled fraction) in the distillation batches analyzed. The variable-wise arrangement of the NOC multibatch data allows obtaining specific process profiles per each batch and thus observing the natural variability among NOC batches. The results for the process modeling of the gasoline distillation batches using PCA and MCR-ALS are shown in Figure 20.

Figure 20a displays the scatter score plot issued from the PCA model of a variable-wise augmented multiset formed by the nine NOC batch data. The model was done on mean-centered data and three PC's explained 99% variance. The distribution of scores illustrates the similar process trajectories of NOC batches, depicted with little blue dots. This PCA model was used to obtain the scores from new batches not used in the model building step, as shown in equation (9) of section 2.3.3. Thus, NIR data from on-specification NOC batch B11, with 27%(v/v) ethanol, and off-specification batch B13, which had only 15 %(v/v) of ethanol added, were studied with this PCA model. On-specification batch B11 (magenta circles) was observed to follow the same process trajectory of the NOC batches used to build the PCA model, while off-specification batch B13 (red triangles) clearly deviated from the NOC trajectory, as illustrated in Figure 20a.



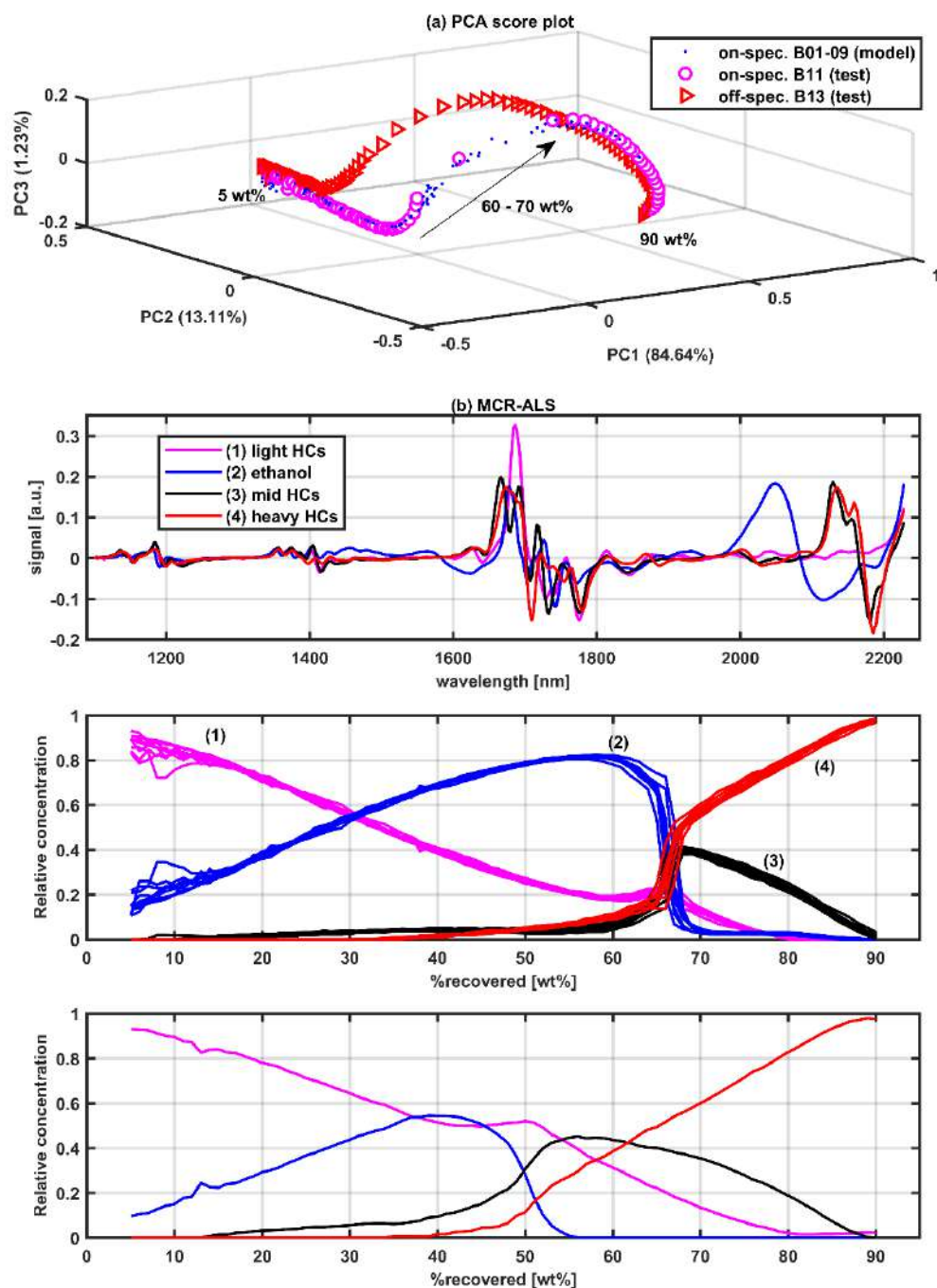


Figure 20 Multibatch distillation process modeling. (a) Scatter score plot issued from PCA of on-specification NOC batches B01-09 (displayed as little blue dots). Overlaid process trajectory of on-specification NOC batch B11 (pink circles) and off-specification batch B13 (red triangles). Start (5 wt%), transition region (60-70 wt%) and end (90 wt%) of distillation process are indicated; (b) Multibatch MCR-ALS showing from top to bottom the pure spectral profiles, the superimposed concentration profiles for on-specification NOC batches B01-09 and concentration profiles for off-specification batch B13 (components (1) to (4)).

The MCR-ALS analysis of multibatch NOC distillation data provided four components. The concentration (distillation) and spectral profiles obtained after the MCR-ALS decomposition are shown in Figure 20b (top panel with overlaid concentration profiles of the different batches and mid panel with the related spectral signatures). The components resolved from the distillation process are related to the four main distilled fractions of gasolines blended with ethanol. From low to high distillation temperatures, the first fraction relates to light hydrocarbons, the second to ethanol and the third and fourth fractions to mid and high molecular weight (MW) hydrocarbons and aromatic compounds, as reported elsewhere (Burger et al., 2015). The identity of these compounds is confirmed when looking at the spectral features found in the related pure spectra and at the temperature distillation range. The low MW hydrocarbons fraction is mainly distilled together with ethanol as azeotropes at the beginning of the distillation, i.e., at lower temperatures, as observed in the concentration profiles of components (1) and (2), see Figure 20b. After 70 wt% of the distillation process, almost all ethanol, component (2), was boiled-off, remaining most of the mid to high MW fractions of gasoline, rich in aromatic compounds, components (3) and (4). In the context of MCR-ALS, concentration profiles of new batch data from the same process can be used for visual inspection of its evolution. Figure 20b (bottom panel) shows the distillation profiles of batch B13 (with only 15 % (v/v) ethanol). Although the component spectra are the same as for the on-specification batches, all distillation profiles are shifted to lower wt% of distillate, as expected for a batch with lower ethanol content.

Besides their process modeling value, both PCA scores and MCR-ALS concentration profiles provide excellent NIR-related compressed information about the evolution of the distillation process that can be further used for process control purposes.

### **Process control of synchronized batches**

Since NOC distillation batches are synchronized, the process control models proposed were always developed using batch-wise augmented data sets. Two types of MSPC models were built, *offline* and *online* MSPC models. Offline MSPC models were used to detect faulty batches using the complete batch information collected during the distillation process, whereas online MSPC models could detect batch upsets related to new batch observations during process monitoring.

It is important to stress the difference between offline and online MSPC control charts. Offline MSPC control charts are based on a single PCA model built on the data from completed NOC batches. The final  $Q_{stat.}$  and  $D_{stat.}$  charts represent the values of these statistics vs. the batch index of each new batch. Every new batch is represented by a point in these charts. Instead, online MSPC control charts display simultaneously the information of many PCA models, as many as observations in each batch. Therefore, each new observation acquired in a new batch is tested to see whether it is in- or out of control on a different PCA model. The process control is done at an observation level and not at a full batch level, as in the offline approach. As a consequence, every new batch has a full  $D_{stat.}$  and a full  $Q_{stat.}$  plot, where the x-axis refers now to the

different observations studied along the process evolution. The description on how to build and use both types of MSPC models for the studied process are provided in the next subsections. In all models, data from nine NOC batches were used to develop the different PCA-based multivariate statistical process control (MSPC) models. These models were afterwards used to detect whether full new validation batches (in offline MSPC models) or individual observations of validation batches (in online MSPC models) were in or out of control.

### Offline batch MSPC models

Offline batch process control was carried out using the classical MSPC approach presented in section 2.3.3. The focus in this case was studying the influence of the starting information used, i.e., temperature profiles from complete distillation curves or information derived from NIR spectra, on the efficiency of the MSPC models to differentiate among NOC and abnormal batches. Thus, the data from the complete distillation of NOC batches were arranged in a batch-wise augmented matrix before model building. When using temperature information, the  $\mathbf{X}_{\text{NOC}}$  matrix included the full distillation curve per every batch. When using NIR-derived information, every row in the  $\mathbf{X}_{\text{NOC}}$  matrix contained the complete vectorized information from one batch, formed by all concatenated original preprocessed spectra, or by concatenated NIR-compressed information, such as selected spectral ranges, PCA scores or MCR-ALS distillation profiles. Due to the nature of the distillation processes, in which the main variation takes place in a narrow % mass distilled fraction range, the effect of taking information around this critical range instead of using the full distillation batch and the use of derivative preprocessing to enhance the changes among batches was also considered. Figure 21 illustrates this last point, showing the information of distillation curves covering the full % mass distilled fraction, their related derivative curves and the derivative curves covering the critical range of variation. From left to right, it is visible that the difference of behavior between NOC batches (in black) and off-specification batches (colored lines) becomes clearer.

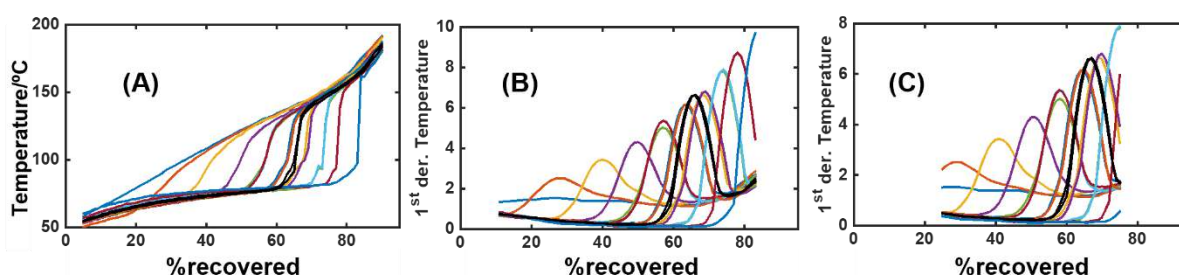


Figure 21 (A) Full distillation curves used to test PCA-based offline batch MSPC model, (B) Distillation curves after Savitzky-Golay 1<sup>st</sup> derivative. (C) Distillation curves after Savitzky-Golay 1<sup>st</sup> derivative between 25 and 75% of the distillation range. On-specification batches are shown as black curves.

To assess the performance of the MSPC charts for offline batch process control using different kinds of starting information, specificity and sensitivity were adopted as quality parameters. Specificity stands for the ratio of NOC batches (on-specification) correctly identified over the total number of NOC batches used to test the MSPC

charts. Sensitivity is derived as the ratio of out of NOC (off-specification) batches correctly identified as out of NOC over the total out of NOC batches tested. Table 2 contains a summary of the model characteristics and quality parameters of the MSPC models built with different kinds of starting information.

Table 2 Offline batch MSPC results, reproduced from (de Oliveira et al., 2017).

#	Description	NC	%EV	%Specificity		%Sensitivity	
				<i>D</i> <sub>stat.</sub>	<i>Q</i> <sub>stat.</sub>	<i>D</i> <sub>stat.</sub>	<i>Q</i> <sub>stat.</sub>
<b>Distillation Curves</b>							
1	DistRange (5-90 wt%)	2	90.91	100	100	66.67	100
2	DistRange(5-90 wt%)_1 <sup>st</sup> diff.SG	2	94.83	100	100	91.67	100
3	DistRange(25-75 wt%)_1 <sup>st</sup> diff.SG	2	98.36	100	100	100	100
<b>NIR information</b>							
<b>Unfolded NIRS data</b>							
4	FullSpec_DistRange(5-90%)	3	68.91	100	100	0.00	100
5	FullSpec_DistRange(60-70%)	2	92.56	100	100	75.00	100
6	SelSpec_DistRange(60-70%)	2	94.54	100	100	83.33	100
<b>scores from PCA modelling (3 PC's)</b>							
7	DistRange(5-90%)	2	87.53	100	100	66.67	100
8	DistRange(60-70%)	2	96.52	100	100	91.67	100
9	DistRange(60-70%)_1 <sup>st</sup> diff.SG	2	98.67	100	100	100	100
<b>C profiles from MCR-ALS modelling (4comp)</b>							
10	DistRange(5-90%)	2	81.07	100	100	33.3	100
11	DistRange(60-70%)	2	94.33	100	100	91.67	100
12	DistRange(60-70%)_1 <sup>st</sup> diff.SG	2	96.99	100	100	100	100
13	DistRange(5-90%)_EtOHcomp	2	90.60	100	100	58.33	100
14	DistRange(60-70%)_1 <sup>st</sup> diff.SG_EtOHcomp	2	99.65	100	100	100	100

# Model number, **NC** number of PCA principal components, **%EV** cumulative explained variance by NC principal components, **diff.SG** Savitzky-Golay derivative, **DistRange** Distillation Range, in % distilled mass, used to build the model, **FullSpec** Complete NIRS measurement range, **SelSpec** Small spectral range more selective for ethanol signal, **EtOHcomp** component 2 related to ethanol.

As a general conclusion, the *Q*<sub>stat</sub> control chart always perform correctly, irrespective of the kind of initial information used, i.e., NOC batches are recognized as such and abnormal batches are also clearly identified. When using the *D*<sub>stat</sub> chart, no problems are encountered to identify correctly NOC batches; however, some off-specification batches are mistaken as NOC batches depending on the starting information used to build the MSPC models. A general way to mitigate the problem is working with initial information covering only the process range showing more variation among batches (in the distillation context, the % distilled mass fractions around the steepest zone of the distillation curve for NOC batches, i.e., between 60-70 % distilled mass fractions). An additional way to enhance differences between NOC batches and off-specification batches with very similar characteristics to them is using the derivative version of the initial information used. Using NIR starting information, the worse results are obtained when the full NIR spectrum is used due to the presence of big spectral regions with very low and noisy signal, similar in all batches compared. In line with the selection of information along the process direction, better results are obtained when only selected

spectral regions with more significant variation are taken into account. It is worth commenting that the use of compressed expressions of the NIR information, such as PCA scores or MCR-ALS concentration profiles, provide always better results than the NIR spectral information because variations of composition are better captured and the compressed profiles are additionally noise-filtered. Finally, when all PCA scores and all MCR-ALS concentration profiles are used, the results are comparable. However, the use of MCR-ALS profiles offers a more flexible framework since, when needed, only the compound profiles showing the clearest variation of batch behavior or having the most critical information, e.g., ethanol in this context, can be adopted to build the necessary MSPC models.

### Online batch MSPC models

Control charts from offline MSPC models are ideal to detect faults using complete batch data. However, most of the times, fast detection of process upsets in real-time is required. In this work, new PCA-based online MSPC strategies were proposed using the NIR spectra collected during the distillation process. As in the offline approach, batch-wise augmented data matrices were used to build the MSPC models. In this case, only the batch-wise augmented matrix containing the original preprocessed NIR readings from all nine individual NOC batches was used to test the different online MSPC strategies. The augmented matrix had the NIR spectra of all NOC batches for each observation point, i.e. 5 to 90% mass distilled fraction, arranged side by side as shown in Figure 22A. Based on this augmented matrix, three online MSPC approaches were proposed using multiple PCA models based on different intervals of observation points, as described below:

- a) *Online MSPC based on individual process observation models.* This approach is the most straightforward method. An individual MSPC model  $(\mathbf{T}_k, \mathbf{P}_k)$ , where  $\mathbf{T}_k$  are the scores and  $\mathbf{P}_k$  are the loadings for the  $k^{th}$  observation is built per each observation point, i.e. from  $k = 1$  to  $k = K$ , using historical data from that particular batch observation in on-specification completed batches, as illustrated in Figure 22B. Thus, during a new batch, the new online data obtained (the NIR spectrum of the current  $k^{th}$  observation) is projected into the respective observation point model and the statistical parameters compared with the control chart limits.
- b) *Online MSPC based on Evolving MSPC models.* Local MSPC models with an increasing number of observation points are built adding the new current distillation point in every new model until all distillation process is covered. As illustrated in Figure 22C, the first MSPC model is built using only the NIRS data matrix of the NOC historical data batches at the first recovered point (5 wt%), the second model using two observation points (5 and 6 wt%) and so on until the last model uses all observation points. For new batch monitoring, the data up to the current observation point are projected into the MSPC model with the same related observations points and statistically tested.

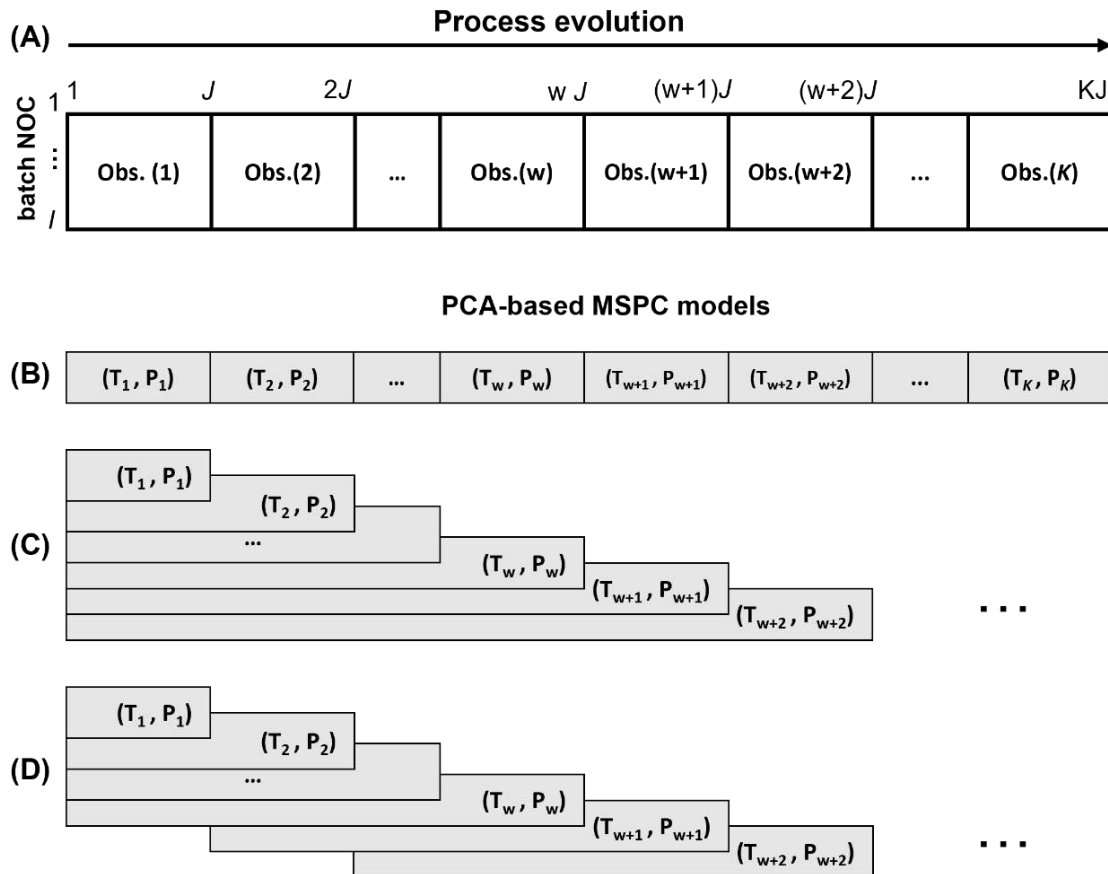


Figure 22 Different online MSPC model approaches. (A) Individual process observation models, (B) FSMW-MSPC models. (C) Evolving MSPC models.  $(\mathbf{T}_k, \mathbf{P}_k)$  represents the scores  $\mathbf{T}_k$  and loadings  $\mathbf{P}_k$  of the local MSPC model related to the evolving strategy used for the  $k^{\text{th}}$  batch observation.

- c) *Online MSPC based on Fixed Size Moving Window, FSMW-MSPC, models.* Several MSPC models built with a moving window including the current observation and several consecutive past observation points are built using the NOC historical data. The window size,  $w$ , defines the total number of observations inside the moving window. However, for the first observations with  $k < w$ , the MSPC models are calculated as in the evolving strategy, adding every new observation until  $k = w$ , when the sliding window is full. Then, the window slides one observation ahead from  $k$  to  $k = k + 1$ , to build each new model until all observation points are covered,  $k = K$ , see Figure 22D. In every new model, the current observation is included and the furthest one in time from the previous model is not considered anymore. For new batch monitoring, the data from the observation points covered by the moving window are projected into the model for the respective observations points and statistically tested.

The three approaches aim at online process control, but there are important differences due to the information included in each of the model typologies. Thus, the modality looking at individual process points (a) does not take into account the neighboring past observations and, hence, the evolution of the process. In the modalities FSMW-MSPC (c) and evolving MSPC (b), the process evolution is taken into account together with the new process observation of interest. In the case of the

FSMW-MSPC model, only the recent past observation points (those within the window) are taken into account and the window size established is related to the number of relevant neighboring process observations. Instead, the evolving-MSPC takes into account all process evolution until the present observation, giving potentially the same importance to all the past observations analyzed.

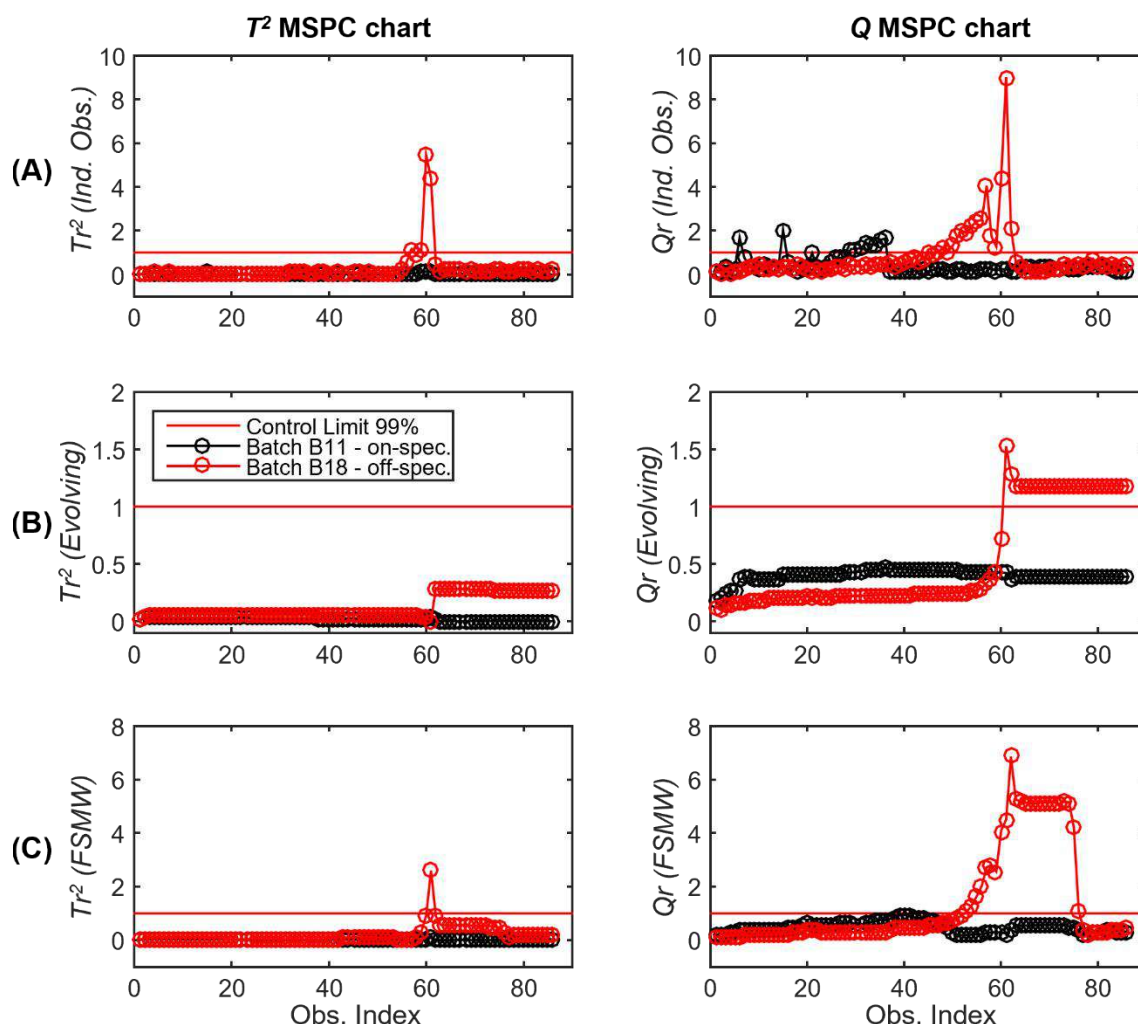


Figure 23 Online MSPC charts for batches B11 (on-specification) and B18 (off-specification) for the (A) Individual process observation models, (B) Evolving MSPC strategies and (C) FSMW evolving MSPC, reproduced from (de Oliveira et al., 2017).

For the distillation batches studied, PCA-based online MSPC models were calculated for each observation point (86 models) following each of the strategies described using the mean-centered data collected from NOC batches. For FSMW evolving models, the window selected enclosed  $w = 15$  neighboring observations. PCA models for individual observation and FSMW evolving strategies were built with one principal component for all observation points, while in evolving models, one PC was used from 5 to 20 wt% range and three PC's in the models referring to the remaining observation points. A confidence interval of 99% was considered to calculate the MSPC chart limits for each model, as described in section 2.3.3. The NIR measurements for a new batch

observation were mean-centered according to the mean of NOC batches. Each new observation (or set of observations) was projected into the suitable PCA model for each strategy to obtain the reduced MSPC statistics,  $Tr^2$  and  $Qr$ .

The online  $T^2$  and  $Q$  charts in Figure 23, which represent the evolution of the related reduced statistics as a function of the observation (%distillate mass fraction) analyzed, allow identifying on- and off-specification batches and when the anomaly in an abnormal batch starts. The three typologies of online MSPC charts related to two validation batches, on-specification batch B11 (with 27% ethanol added, displayed in black) and off-specification batch B18 (with 25% ethanol added, displayed in red), are shown in Figure 23. The  $Q$  online MSPC charts detected the fault in the off-specification batch by using any of the three online strategies, see Figure 23 (right plots).  $T^2$  charts worked generally well, except when using evolving models, which were not able to detect faults in off-specification batches with ethanol content very approximate to the accepted specification, as happened with batch B18, see Figure 23B (left plot). The on-specification batch B11 was wrongly detected as faulty by the  $Q$  charts using the individual observation model MSPC strategy, Figure 23A (right plot).

Based on the results shown in Figure 23, some conclusions can be extracted for each online strategy. For individual process observation models, MSPC charts are very sensitive to fault detection, in particular  $Q$ -residuals charts. Although this helps to accurately detect the point where the batch starts to be anomalous in off-specification batches, a slight variation in an individual observation of a new on-specification batch process leads to show false alarms, as in Figure 23A (right plot). This is due to the restricted amount of information used in the models, just related to a single observation, that may lead to interpret any accidental small fluctuation (probably absent in neighbouring observations) as an abnormal behavior. For evolving MSPC models, when the local models increase in number of observations, the  $T^2$  charts become less sensitive to fault detection. This is because the evolving MSPC strategy considers the evolution of the process since the start and, therefore, the large number of past NOC observations has a lot of weight in the evolving models and may hinder to detect abnormal observations that happen at long batch times unless the related anomaly is very clear. In evolving MSPC models, the detection of faulty batches happens easily only when the fault occurred at the beginning of the batch, such as for gasoline batches with ethanol content lower than the specification.  $Q$  charts performed more satisfactorily in fault detection. However, faults were detected later than in individual observation models due to the excessive weight of past NOC observations as well. The FSMW strategy considered only a few past observation points, set according to the window size (15 observations in this study). This feature produced more sensitive  $T^2$  charts because past NOC observations had less weight in models than in the evolving MSPC strategy. It was also observed to be less prone to false alarms on  $Q$  charts than individual observation charts since the higher number of observations in the window helps to distinguish more clearly accidental fluctuations of



individual observations from abnormal process trends. This strategy is the most flexible of the three, showing efficient and easy detection of faults and avoiding false alarms. The performance of FSMW-MSPC models depends on the width of the window: if too small, false alarms may show up in analogy to what happens in individual observation models; if too big, sensitivity in  $T^2$  chart may decrease because of the weight of too many past NOC observations in the chart. The window width should be set by using representative validation batches (on- and off-specification) that may allow setting the correct window width avoiding the problems of low sensitivity and false alarms as described in the other two online MSPC approaches.

## 4.3 Process modeling and control for non-synchronized batch processes

### 4.3.1 MSPC models for endpoint detection. Data fusion strategies.

The detection of a batch endpoint is crucial to secure final product quality consistency, save time, energy and reduce waste products. The endpoint detection of a batch process can be based on the use of a single MSPC model taking as input information sensor measurements, e.g. NIR spectra, collected at the end of several NOC batches. Multiple endpoints can also be detected in batch processes that take place following several steps with related MSPC models. Endpoint MSPC models can be built using only NIR information or combining it with information from other process sensors, like temperature. An additional idea around the data fusion concept is combining process sensors and/or PLS, PCA and MCR-ALS model outputs issued from the use of the information from spectroscopic probes. In this way, data fusion applies to both sensor combination and model output combination.

This subsection gathers three scientific publications related to the topics described above. Publication II describes the use of PLS for real-time monitoring of moisture content and the application of an MSPC model for endpoint detection of an industrial drying process using NIR spectroscopy. Publication III presents the use of PLS to predict in real-time several critical quality attributes and the application of several PCA-based MSPC models to detect multiple endpoints in an industrial polyester production process using NIR spectroscopy. Finally, Publication IV proposes new data fusion strategies to build endpoint MSPC models combining sensor and multivariate model outputs for the processes described in Publications I, II and III.

**Publication II.** Avila, C. R., Ferré, J., de Oliveira, R. R., de Juan, A., Sinclair, W. E., Mahdi, F. M., Hassanpour, A., Hunter, T. N., Bourne, and R. A., Muller, F. L., **Process Monitoring of Moisture Content and Mass Transfer Rate in a Fluidised Bed with a Low Cost Inline MEMS NIR Sensor**, *Pharmaceutical Research* (2020), 37: 84. DOI: [10.1007/s11095-020-02787-y](https://doi.org/10.1007/s11095-020-02787-y)

**Publication III.** Avila, C., Mantzaridis, C., Ferré, J., Rocha de Oliveira, R., Kantojärvi, U., Rissanen, A., Krassa, P., de Juan, A., Muller, F. L., Hunter, T. N. and Bourne, R. A., **Acid number, viscosity and end-point detection in a multiphase high temperature polymerisation process using an online miniaturised MEMS Fabry-Pérot interferometer.** *Talanta* (2021), 224: 121735. DOI: [10.1016/j.talanta.2020.121735](https://doi.org/10.1016/j.talanta.2020.121735)

**Publication IV.** de Oliveira, R. R., Avila, C., Bourne, R., Muller, F., and de Juan, A., **Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control**, *Analytical and Bioanalytical Chemistry* (2020), 412:2151–2163. DOI: [10.1007/s00216-020-02404-2](https://doi.org/10.1007/s00216-020-02404-2)





# Process Monitoring of Moisture Content and Mass Transfer Rate in a Fluidised Bed with a Low Cost Inline MEMS NIR Sensor

Claudio R. Avila<sup>1</sup> · Joan Ferré<sup>2</sup> · Rodrigo Rocha de Oliveira<sup>3</sup> · Anna de Juan<sup>3</sup> · Wayne E. Sinclair<sup>4</sup> · Faiz M. Mahdi<sup>1</sup> · Ali Hassanpour<sup>1</sup> · Timothy N. Hunter<sup>1</sup> · Richard A. Bourne<sup>1</sup> · Frans L. Muller<sup>1</sup> Received: 20 December 2019 / Accepted: 18 February 2020 / Published online: 21 April 2020  
© The Author(s) 2020

## ABSTRACT

**Purpose** The current trend for continuous drug product manufacturing requires new, affordable process analytical techniques (PAT) to ensure control of processing. This work evaluates whether property models based on spectral data from recent Fabry–Pérot Interferometer based NIR sensors can generate a high-resolution moisture signal suitable for process control.

**Methods** Spectral data and offline moisture content were recorded for 14 fluid bed dryer batches of pharmaceutical granules. A PLS moisture model was constructed resulting in a high resolution moisture signal, used to demonstrate (i) endpoint determination and (ii) evaluation of mass transfer performance.

**Results** The sensors appear robust with respect to vibration and ambient temperature changes, and the accuracy of water content predictions ( $\pm 13\%$ ) is similar to those reported for

high specification NIR sensors. Fusion of temperature and moisture content signal allowed monitoring of water transport rates in the fluidised bed and highlighted the importance water transport within the solid phase at low moisture levels. The NIR data was also successfully used with PCA-based MSPC models for endpoint detection.

**Conclusions** The spectral quality of the small form factor NIR sensor and its robustness is clearly sufficient for the construction and application of PLS models as well as PCA-based MSPC moisture models. The resulting high resolution moisture content signal was successfully used for endpoint detection and monitoring the mass transfer rate.

**KEY WORDS** fluidised bed drying · mass transfer resistance · MEMS Fabry–Pérot interferometer sensor · near infrared spectroscopy · online process monitoring

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11095-020-02787-y>) contains supplementary material, which is available to authorized users.

✉ Frans L. Muller  
F.L.Muller@leeds.ac.uk

Claudio R. Avila  
Clavila@udec.cl

Joan Ferré  
joan.ferre@urv.cat

Rodrigo Rocha de Oliveira  
rrochade10@alumnos.ub.edu

Anna de Juan  
anna.dejuan@ub.edu

Wayne E. Sinclair  
wayne.e.sinclair@gsk.com

Faiz M. Mahdi  
F.M.Mahdi@leeds.ac.uk

Ali Hassanpour  
A.Hassanpour@leeds.ac.uk

Timothy N. Hunter  
T.N.Hunter@leeds.ac.uk

Richard A. Bourne  
R.A.Bourne@leeds.ac.uk

<sup>1</sup> School of Chemical and Process Engineering, University of Leeds Leeds LS2 9JT, UK

<sup>2</sup> Department of Analytical Chemistry and Organic Chemistry Universitat Rovira i Virgili, 43007 Tarragona, Spain

<sup>3</sup> Department of Chemical Engineering and Analytical Chemistry Universitat de Barcelona, 08028 Barcelona, Spain

<sup>4</sup> Analytical Sciences, GlaxoSmithKline, Stevenage, UK

## NOMENCLATURE

$a_w$	activity of water in the solid phase (-)
$b, b_0$	regression coefficients
$C_g$	Water conc. in the gas phase ( $mol/m^3$ )
$C_L^*$	Gas phase conc. in equilibrium with the liquid in the granules ( $mol/m^3$ )
$C_S^*$	Gas phase conc. in equilibrium with the water in the solid phase ( $mol/m^3$ )
<b>E</b>	Matrix with residual errors of the PCA model (-)
$e_y$	Error between measured and predicted values of $y$ (-)
$e_{rel}$	Relative error (-)
$f_s$	Volume fraction solids in the fluidised bed (-)
$f_{MTR}$	Extent to which mass transfer is limiting (-)
$f_w$	Moisture fraction (wt% of total mass)
$F_w$	Moisture fraction (wt% of dry solids)
$m_w, m_s$	Granule's mass of water and solids respectively (kg)
$m_{adsorb}$	Mass of water absorbing components in the granule (kg)
$MTR_w$	Mass transfer rate ( $mol/s.m^3$ )
$N_w$	Number of moles of water (mol)
$\dot{N}_w$	Maximum drying rate ( $mol/s$ )
$\dot{N}_w$	Molar drying rate ( $mol/s$ )
<b>P</b>	Loadings matrix (-)
$P_w^*$	Vapour pressure of water (Pa)
$Q_{stat}$	Q-statistic (-)
$S_{in}$	Degree of saturation of the inlet gas (-)
<b>T</b>	Scores matrix (-)
$T_{bed}$	Bed temperature (K, or °C)
$V_{bed}$	Volume of the fluidised bed ( $m^3$ )
<b>y</b>	Response vector (-)
<b>X</b>	Matrix containing one spectrum on each row (-)
$x_{new}$	Row vector containing 1 spectrum (-)

## GREEK LETTERS

$\varphi_g$	Gas flowrate ( $m^3/s$ )
$\rho_s$	Fluid bed density ( $kg/m^3$ )
$\Omega$	Mass transfer resistance (1/s)
$\Omega_{ext}, \Omega_{max}$	Fitting parameters (1/s)
$\Omega_{tot}$	Overall or total mass transfer resistance (1/s)

## INTRODUCTION

Near infrared spectroscopy (NIR) has been established as a key tool for process analysis technology (PAT) (1). For pharmaceutical applications, it has proven to be an effective technique for gathering relevant chemical data to build up process understanding (2,3), contributing to new process development (4), and for process monitoring and control during the drug manufacturing processes (5). Recent advances in instrumentation and chemometrics have been identified as the main pillars advancing NIR spectroscopy towards application in manufacturing, but cost and measurement

robustness remain barriers to widespread implementation in the pharmaceutical industry.

The recently developed Micro-electro-mechanical system Fabry-Pérot Interferometers (MEMS FPI) for near infrared wavelengths are miniaturised tuneable optical filters formed by two facing reflectors separated by an air gap. The distance between the two reflectors is controlled by the voltage applied (6). Light with a wavelength of the gap size will interfere and pass the filter and be collected by a single-point detector positioned below. The range of distances the device can set will thus determine the range of wavelengths the device can measure (7). MEMS-FPI based devices with different reflector gap widths target specific regions of the NIR spectral ranges (e.g. 1.7  $\mu m$ , 2  $\mu m$ ). For a specific process application, the appropriate spectral range can be matched with the sensor range (8–10).

The resolution of the MEMS FPI sensors is lower than in Fourier Transform NIR spectrometers or traditional diffraction gratings spectrometers, but the compact form factor and cost-effective pricing enable new applications that are not possible with aforementioned traditional spectral sensing technologies. To measure spectra, the voltage is changed gradually and the detector signal is recorded. Spectral data can be recorded up to speeds of 1000 spectral points per second, providing the user a high rate of data at a few spectral positions, or a full spectrum over the range available at a lower rate.

MEMS-FPI are made from a single wafer without assembly steps, creating a single solid structure with no wearing parts which makes the devices position and vibration insensitive; staying very stable over time. Developed low-cost MEMS-FPI and detector modules have been successfully miniaturised, with systems weighing as little as 60 g (11), making them suitable for widespread application in process sensors in the manufacturing industry. This would be a major step forward to continuous, inline composition measurement.

For the pharmaceutical industry, accessing reliable spectral information at a low cost could bring immediate benefits. For instance, complementary compositional and physical information could be obtained for several drug manufacturing stages before attempting to replace conventional quality control or research analytical systems. A specific example is the monitoring of moisture content, a control parameter used in several stages of the solid-dose form production process (5), normally required for milling and blending (12,13), granulation (14,15), tablet coating (16), and drying, particularly fluidised bed granule drying, which has been extensively studied (17–20).

In examples reported, real-time moisture determination using NIR spectroscopy relies on correlating online NIR spectra to offline analytical moisture measurements (typically Karl Fisher titrations or loss on drying (LOD) (14,21)). Partial Least Squares regression (PLS) is the algorithm of choice to model

the measured moisture content (vector  $\mathbf{y}$ ,  $n$  data points) as a linear combination of  $n$  measured spectra (matrix  $\mathbf{X}$ , one spectra on each row<sup>1</sup>):

$$\mathbf{y} = \mathbf{X} \mathbf{b} + b_0 + \mathbf{e}_j \quad (1)$$

where  $\mathbf{b}$  the vector with relative contributions of each spectral point,  $b_0$  is a constant (22) and  $\mathbf{e}_j$  is the vector of errors. With  $\mathbf{b}$  and  $b_0$  obtained from a calibration data ( $\mathbf{y}$ ,  $\mathbf{X}$ ), a real-time moisture content signal can be inferred from a new NIR spectrum by application of Eq. 1 (23).

The second common application of NIR data is for drying endpoint detection. Here, Multivariate Statistical Process Control (MSPC) models based on Principal Component Analysis (PCA) are built from sets of spectra of different Normal Operating Condition (NOC) batches that represent the process end-point well. In-line spectra from new batches are tested in real time to determine whether they behave, or not, as the NOC spectra used to build the model (24,25).

To build MSPC models for process end-point detection, a data set formed by NIR spectra of on-specification batches where the end-point has been reached are used in a matrix  $\mathbf{X}$  containing of  $n$  end-point spectra. A PCA model is built in order to set the statistical boundaries of the experimental domain (space) of end-point NIR spectra (2,3):

$$\mathbf{X} = \mathbf{T} \mathbf{P} + \mathbf{E} \quad (2)$$

where  $\mathbf{P}$  is the loadings matrix (nr of principal components, nr of wavelengths which are the link between scores and original NIR spectra) and  $\mathbf{T}$  is the scores matrix of all end-point spectra (nr of spectra, nr of principle components).  $\mathbf{T}$  spans the valid experimental domain for on-specification measurements in the space of principal components. The matrix  $\mathbf{E}$  describes the residual errors of the PCA model. For any new (pre-processed) spectrum  $\mathbf{x}_{\text{new}}$  acquired in the current on-line monitored batch, the difference between the spectrum  $\mathbf{x}_{\text{new}}$  and its description by the PCA model  $\mathbf{x}_{\text{new}} \mathbf{P}^T \mathbf{P}$  is:

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

$\mathbf{e}_{\text{new}}$  is a row vector containing the residual error for each wavelength.  $Q_{\text{stat}}$  control charts are developed on this basis of the sum of squares of this error:

$$Q_{\text{stat}} = \mathbf{e}_{\text{new}} \mathbf{e}_{\text{new}}^T \quad (4)$$

When  $Q_{\text{stat}}$  falls below a minimum error, the chart control limit, the new spectra resembles the typical endpoint spectrum shape as defined by the NOC batches.

Acquiring online spectra from multiphase processing equipment (air-solid) (26) such as fluidised bed dryers, wet granulators and excipient blenders, presents several challenges such as noise and probe window fouling. For instance, within a fluidised bed, drying granules will flow past the NIR probe window, and particles will interact with the probe's NIR light

at a wide range of distances and orientations. The significant changes in material density and air gaps with variable distance between the solids and the field of view of the probe results in an intermittent signal (27). These interactions lead to significant levels of noise that distort the coefficients in Eqs. 1 and 2–4 resulting in large residual errors.

Under these conditions, it is essential to reduce the noise introduced during spectral measurement by suitable pre-processing of the NIR spectra (4,28), for instance by averaging a number of them.

This study aims to evaluate the accuracy, robustness and reliability of the spectral response obtained from the MEMS-FPI NIR sensor by monitoring the moisture content during the drying of pharmaceutical granules in a pilot-scale fluidised bed dryer. PLS models resulting from a validation data set transform online spectral measurements to predictions of the moisture content and MSPC tools use spectra for detecting the process end-point. In the final part, we convert the NIR derived moisture content signal to the water mass transfer rate and demonstrate how this can be applied so as to provide an insight in to the underlying processing phenomena.

## MATERIALS AND METHODS

### Granulation and Drying

Pharmaceutical granules were produced using a standard recipe supplied by GlaxoSmithKline (GSK, United Kingdom) employing Mannitol (64 wt%, solid wt% are with respect to the dry granule weight), Microcrystalline Cellulose Avicel PH-101 (29 wt%), Hypromellose 2910 (5 wt%, binder), AC-DI-SOL (1.5 wt% disintegrant) and colloidal  $\text{SiO}_2$  (0.5 wt%). A top driven high shear mixer wet-granulator model MiPro from ProCepT (Belgium) was used to prepare fourteen 1 kg batches of granules in a 2 l vessel with a three-bladed impeller rotating at 800 RPM. During granulation, Water (0.5 L/kg<sub>solid</sub>) was added at a constant feeding rate of 5 mL/min. Due to intense mixing the temperature rises from room temperature to approximately 45°C. On completion of the granulation, granules are transferred into a container and sealed up before cooling down and storing in a fridge at 5°C to minimise changes in moisture content before drying. Batch-to-batch repeatability in terms of moisture content and particle size was achieved by using the granulator torque profile as an online indicator, in parallel to a correlation between the water addition with the particle size (using an image analysis method) obtained from a control batch. This methodology mirrors a granulation procedure previously reported (29) and further

<sup>1</sup> Note we adhere to the standard notation for chemometrics, using row vectors  $\mathbf{r} = (1, 2, 3)$  for data over a series of wavelengths. For row vectors the order of multiplication is reversed. So for  $\mathbf{r} = \mathbf{v}^T$  if  $\mathbf{v}_2 = \mathbf{A} \mathbf{v}$  then  $\mathbf{v}_2^T = \mathbf{r}_2 = \mathbf{r} \mathbf{A}^T$ . Bold small letters refer to vectors, bold capitals to matrices.

information related to the granulation steps is provided in the [Supplementary Material](#).

Drying was performed in a 0.129 m ID, 4 l fluidised bed model 4 M8-Trix Formatrix from ProCepT (Belgium). A picture of the equipment used is shown in Fig. 1, and a summary with the experimental drying conditions of each batch is presented in Table I. For each batch, 0.5 kg of wet granules (1 l equivalent) were manually transferred into the fluidised bed, and continuously dried by passing a constant flow of air at  $25 \pm 3^\circ\text{C}$  to fluidise the granule bed. The air flow was set to 850 L/min for batches 1 to 10 and to 600 L/min for batches 11 to 14. Visual observation indicates that at both the air flowrates the fluidised bed operates in the bubbling regime, though the NIR signal readings were smoother at the lower gas flowrate.

Every 6–7 min a sample of approximately 5 g was retrieved from the fluidised bed using a vacuum suction for offline moisture analysis. Sampling required the NIR probe to be disconnected from the vessel for approximately 10 s (to make sampling port available). For batches 10 and 14 the fluidisation air was off for 15 s after each manual sampling, allowing collection of NIR spectra with reduced fluctuations in particle density and orientation.

### Analytical Characterisation

Reference moisture content (loss on drying, LOD) was measured using a thermogravimetric moisture analyser model MB120 from Ohaus (Germany), operating at a constant temperature of  $105^\circ\text{C}$ . Granule samples obtained from the fluidised bed were directly deposited in the MB120 sample

chamber and dried in a 5 to 10 min period (depending on the moisture content).

The LOD moisture fraction ( $f_w$ ) reported is defined by the weight of water ( $m_w$ ) and dry weight of the solid ( $m_s$ ) as follows:

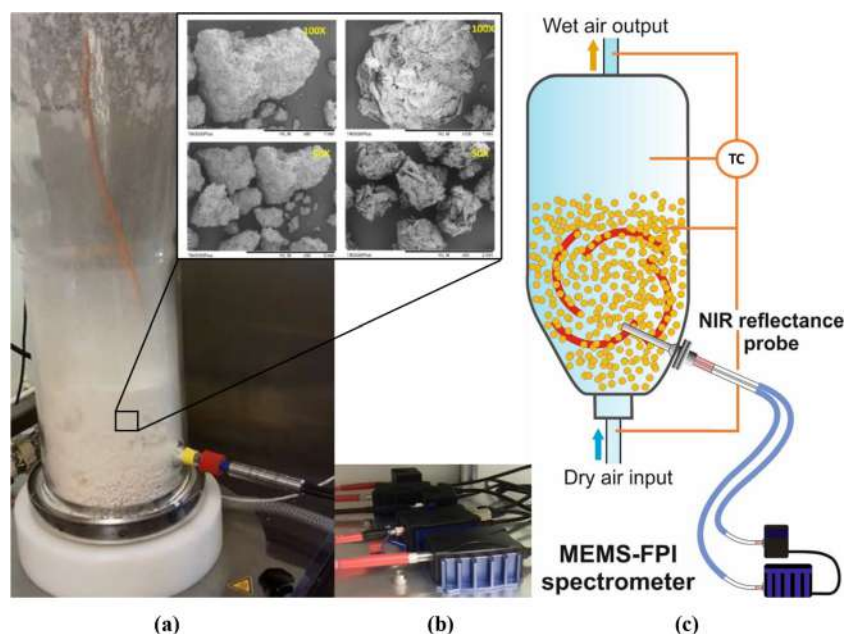
$$f_w = \frac{m_w}{m_w + m_s} \quad (5)$$

In and outlet temperatures were recorded simultaneously to the NIR spectral measurements using K type immersion thermocouples from Omega (United Kingdom) connected to a TC-08 AD converter from Pico technologies (USA). The gas flowrate is measured by the ProCepT control system.

### MEMS-FPI NIR Sensor and Data Acquisition

A sensor from Spectral Engines (Finland) model N-Series 2.2, operating from 1750 nm to 2150 nm wavelength range was used for the acquisition of NIR spectra (ca.  $4650\text{--}5714\text{ cm}^{-1}$ ). It has a tuneable MEMS Fabry–Pérot Interferometer acting as the spectral element and a single element extended InGaAs detector (Fig. 2; additional information about the scanning principle of the device can be found in the [Supplementary Material](#)). The sensor has an integrated light source model LS-PRO that utilises a miniature tungsten vacuum lamp as the illumination source. The energy output of the lamp was set to 50% of the maximum level. For all drying batches, the integration time of the sensor was set to 0.1 ms and the wavelength step to 1 nm (10 ms to set a step). This results in 401 wavelength points which including data transfer time results in an acquisition time of approximately 1 full NIR spectrum per second (single scan, no averaging). Control and communication of the NIR sensor, data logging of the NIR spectra and recording of temperature readings, were performed using a

**Fig. 1** (a) The fluidised bed dryer with NIR immersion probe attached (Top) SEM pictures of granules after the drying process, (b) the spectral sensor and light source connected to light guides. (c) Schematic view of the system including the position of the MEMS-FPI sensor used, NIR probe location, and temperature (TC) measuring points (right).



**Table 1** Summary of Experimental Conditions for Fluidised Bed Drying of Granules Spectra where Recorded at 1 Scan per Second

Sample name	PLS Model	Moisture drying range (from/to)	Flow rate L/min	Recording time (min)	time to reach 5% (min)**
Batch 1	Calibration	35.03% to 1.99%	850	137.3	58
Batch 2	Calibration	34.11% to 1.77%	850	117.1	46
Batch 3	Calibration	34.45% to 1.86%	850	123.3	54
Batch 4	Calibration	34.24% to 1.85%	850	108.8	42
Batch 5	Calibration	37.60% to 1.86%	850	112.5	43
Batch 6	Calibration	33.94% to 1.51%	850	114.5	39
Batch 7	Validation	33.55% to 1.32%	850	78.0	38
Batch 8	Validation	33.89% to 1.46%	850	85.1	39
Batch 9	Validation	33.83% to 1.62%	850	86.8	41
Batch 10*	Validation	33.92% to 1.63%	850	79.5	40
Batch 11	Validation	34.09% to 1.88%	600	88.0	49
Batch 12	Validation	34.30% to 2.55%	600	117.7	78
Batch 13	Validation	33.88% to 2.18%	600	100.1	64
Batch 14*	Validation	33.42% to 1.86%	600	116.3	66

\*After each manual sampling fluidisation was switched off for 10 s

\*\*Estimated time since the fluidisation started up to reaching 5% of moisture content

bespoke application developed in LabVIEW 2015 by the University of Leeds (ChemView version 3.4 (30)).

### NIR Reflectance Probe and Calibration

A 6 mm diameter immersion NIR diffuse reflectance probe model OFS-6S-100HO/080704/1 from Solvias (Switzerland) was inserted horizontally with the probe's tip located 45 mm above the bottom and in the centre of the bed (Fig. 1). The probe has a stainless steel body with a sapphire window and contains two fibre optic cables, one connecting to the light source, the other to the NIR sensor using standard SMA adaptors. A bright reference spectrum was obtained before starting each batch. A calibration block was used to position the tip of the NIR probe in a 90-degree angle relative to a

diffuse reflectance standard (model Spectralon USRS-99-010 from Labsphere, USA).

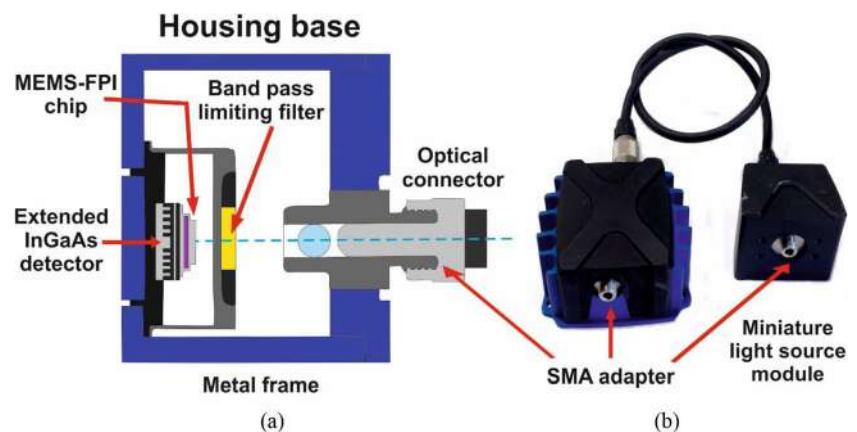
### Chemometric Modelling

Chemometric modelling and validation were carried out with in-house routines programmed in Matlab R2017a (Mathworks, USA) and with PLS\_Toolbox 8.5.1 (Eigenvector Research, USA) running under Matlab.

### Spectral Preprocessing

Intensive preprocessing was required to remove the spectral artefacts generated as result of inhomogeneity in the fluidised bed system. First, 10 intensity raw spectra were averaged into a single raw spectrum. The resulting signal was transformed

**Fig. 2** The tuneable MEMS Fabry–Pérot Interferometer: (a) schematic of the internals showing the MEMS FPI chip mounted on top of the InGaAs NIR detector and (b) a photo of the miniaturised NIR sensor with the light source. The footprint of the assembled sensor chassis is approximately 58 mm length by 57 mm width by 27 mm high, with a weight of 1.25 g.





into absorbance using the bright and dark reference spectra. As artefacts are still present in the data, a moving average filter was applied using the current and 74 prior spectra. Finally, spectra were mean-centred before being submitted to the PLS algorithm. The Standard Normal Variate method was evaluated, but in this case did not improve over the pre-processing method described.

#### PLS Regression for Moisture Prediction

The PLS regression model was built relating the NIR pre-processed spectra ( $\mathbf{X}$ ) to the mean-centred logarithm of LOD moisture content of samples collected during the drying of the 6 initial batches ( $y = \log_{10}f_w - \overline{\log_{10}f_w}$ ; Batches 1 to 6, see Table I). The  $\log_{10}$  transformation was used to minimise the relative error; the error is large at high values of  $f_w$  due to sticking etc. and reduces as the system dries. This improved the predictions over the wide range of moisture content (1.51% to 37.60%). The PLS regression model was calculated using the NIPALS algorithm (31). Finally, PLS outputs were back-transformed to provide moisture values in the original units. Note that when applying Eq. (1) to calculate the log moisture content  $y$  for a new spectrum, the error in Eq. (1) represents a relative error in the  $f_w$  domain; i.e. the expected moisture content  $f_w(X) = e^{y + \overline{\log_{10}f_w}} \times e^{\pm \epsilon_y} \approx f_w(X)|_{est} \times (1 \pm \epsilon_y)$ .

#### MSPC Charts for Process End-Point Detection

PCA-based MSPC model charts were built using NIR spectra corresponding to an offline determined moisture content below 2% chosen as the process end-point criterion. The MSPC model was built using NIR spectra from batches 2, 3, 4, 6, 7, 9 and 10, that each reached the desired process end-point i.e. moisture content below 2%, using the same pre-processing as for the PLS model. All the spectra were gathered in matrix  $\mathbf{X}$  (nr. of spectra, each with 400 wavelengths). After this step, a standard normal variate (SNV) normalisation was applied to remove any unwanted baseline spectral variation followed by mean-centring of the  $\mathbf{X}$  matrix before building the PCA-based MSPC model. The number of components in the PCA model was estimated by cross-validation. Finally, a  $Q$ -statistic control chart ( $Q_{stat}$ ) was built from the resulting MSPC model and control limits at 95 and 99% confidence interval were estimated according to Jackson and Mudholkar equation (32). For external validation,  $Q_{stat}$  control charts were obtained for batches 1, 5, 8, 11, 12, 13 and 14 (not used in the PCA model building step). When the shape of a new spectrum is similar to the end-point spectra used to build the PCA model, the residuals are small and the related  $Q_{stat}$  value appears below the chart control limit. Conversely, when a spectrum is far from the end-point, the spectral shape is clearly different and the

resulting  $Q_{stat}$  value appears well above the chart control limit. The point in time where the  $Q_{stat}$  value goes below the control chart limit at a 95% confidence interval for 10 consecutive observations was used as criterion to indicate the process end-point. Batches that do not reach the end-point should consistently show  $Q_{stat}$  values above the chart control limit. It is important to remind that the end-point detection by MSPC uses the sole information provided by the NIR spectra and does not require any reference moisture content.

## RESULTS AND DISCUSSION

The first aim of the work is to assess the robustness of the new reduced cost and small form factor MEMS FPI NIR sensor over a 9 month period during which 14 batches of placebo granules were manufactured and dried in a fluidised bed (Table I). The objective was to repeat essentially identical batches in order to evaluate the robustness and consistency of the sensor, and the predictions based on models derived from the NIR spectra. The system was however subject to changes uncontrolled variables: such as batch to batch variations in granulate (e.g. size, moisture content, storage time) and the ambient and air inlet temperature and humidity (experiments started with bx1 in August to bx 14 in mid December). To probe the sensitivity of the PLS moisture predictions to a change in flowrate we reduced the air flowrate from 850 to 600 L/min in the last 4 batches.

Overall, the MEMS-FPI sensor performance showed a very satisfactory stability and reproducibility. The spectral signal remained stable and repeatable under all ambient conditions. The bright reference intensity levels and spectra shape measured at the beginning of each experiment remained similar. The device proved to be free of interferences generated by mechanical vibrations; e.g. spectral readings did not alter when the device was installed directly next to the fluidised bed vessel. Ambient temperature variation did lead to minor variations. However, using a second spectrometer it was demonstrated that this was in fact due to effects of temperature on the transmission through the fibre optic cables, rather than on the light source or the MEMS-FPI sensor.

In the fluidised bed process granules are dried using ambient air of uncontrolled humidity. As a result the drying rate varies from batch to batch, with the final drying end-point determined by relative humidity conditions found on the day when a batch is dried. Table I summarises the observations for all batches. A typical data set for a drying experiment is shown in Fig. 3: Granules are charged cold and fluidised by a gas stream of 600 L/min at  $\sim 25^\circ\text{C}$ . As water evaporates, heat is removed from the gas stream resulting in a  $\sim 10^\circ\text{C}$  temperature difference between the bed and the fluidising gas. The moisture content measurement is based on a PLS model

(Eq. 1) based on 6 reference batches and validated using the NIR spectra from a further 8 validation batches. Moisture content is monitored continuously by NIR using the PLS model, and measured once every  $\sim 10$  min by sampling and performing offline analysis by LOD. When the granule water content is reduced by  $\sim 85\%$  the evaporation rate drops and the bed temperature rises. Equilibrium between the inlet gas (ambient RH) and the granules ( $f_w \approx 1.5\text{--}2$  wt%) is reached after  $\sim 90$  min at 600 L/min.

To test the robustness of the NIR measurement and the associated PLS model, drying was performed using two different flow rates (600 and 850 L/min) at constant air inlet temperature ( $25 \pm 3^\circ\text{C}$ ). As expected, the flowrate strongly affects the drying rate: thus, the drying time reduces to  $\sim 60$  min when the flowrate is set at 850 L/min. Even though the PLS model is based on data from runs 1 to 6, all operated at 850 L/min, it still correctly predicts the moisture content of runs at 600 L/min without further chemometric analysis, since only the drying rate, but not the sample composition, is changed. Thus, the fitted parameters for Eq. (1) successfully correlate  $f_w$  to the NIR spectra for the reference batches, and the same parameters allow prediction of moisture content for all validation batches irrespective of the flowrate. This demonstrates (i) the robustness of the data strategy applied (e.g. data treatment, data treatment prior to PLS analysis) and (ii) the excellent stability and robustness of the MEMS-FPI NIR sensor.

When drying from  $\sim 33$  to 10 wt% moisture contents it was observed that the granules shrank due to the loss of 26% of their original mass. At this stage no significant fines were observed. At lower moisture levels (5–1%), granules appear dried at the surface. Granule size continue to reduce, but now by attrition which generated a significant quantity of fines that were observed to build up on the fluidised bed filters,

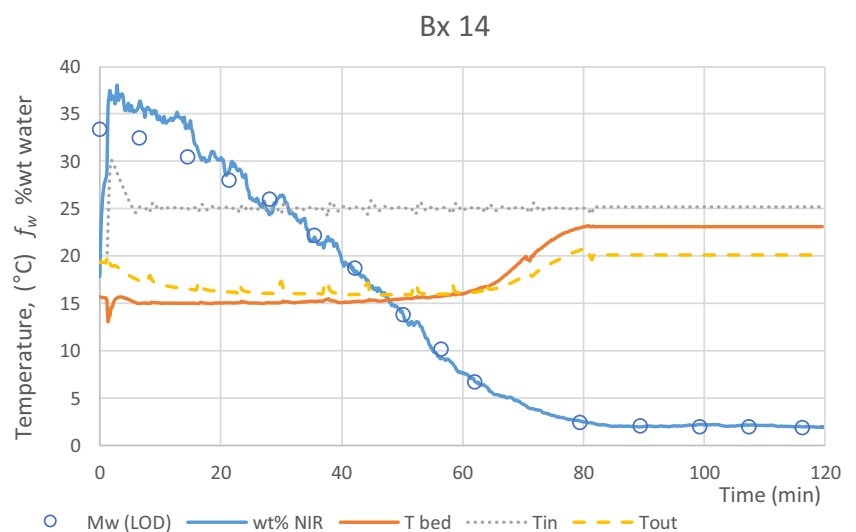
resulting in increasing pressure drop across the outlet filters. The size reduction, coupled to the reduction in moisture content did appear to increase the intensity of the reflected signal as the drying process progressed, but this did not seem to have a significant impact on the moisture content predictions.

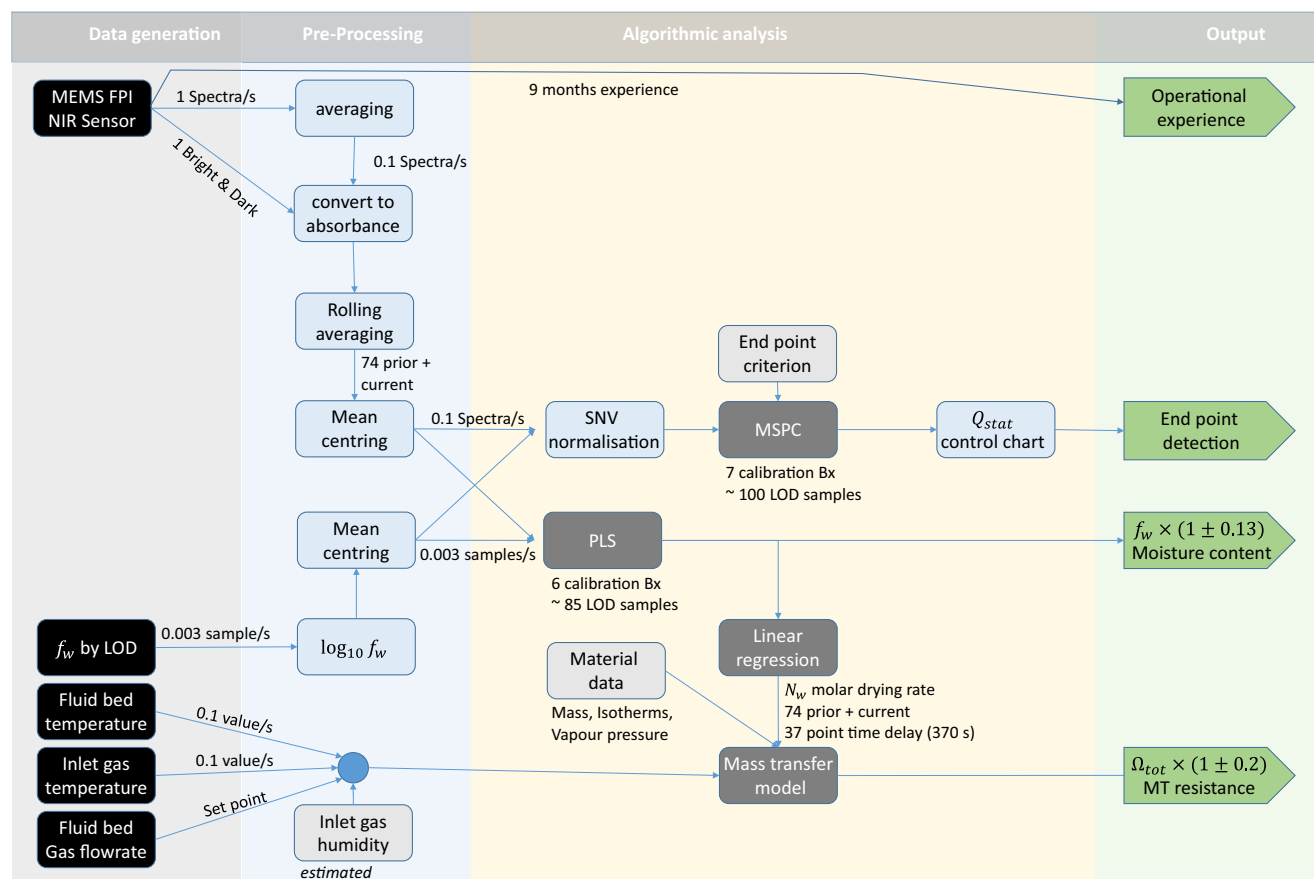
A data analysis strategy was designed (Fig. 4) to demonstrate that the low cost NIR sensors delivers high quality data robustly during fluid bed drying on placebo pharmaceutical granules; a typical processing scenario with a very low signal to noise ratio. We tested three applications: Moisture monitoring, end-point detection and process analysis (mass transfer monitoring) that will be discussed in more detail below

### NIR Signal Acquisition during Fluidised Bed Drying

The presence of large fluctuations in individual NIR scans when sensing *in situ* from a fluidised bed dryer has been previously reported and necessitates intensive data pre-processing steps and/or modification of the way a spectrum is measured (e.g. scoop devices to hold the sample in place while measuring) (28). The contact between the fluidised material and the tip of the NIR probe is variable and characterised by air gaps appearing in front of the sensing area at random intervals. This produces abrupt changes in the spectra collected from scan to scan. Figure 5a, b, c shows 10 consecutive single spectral readings and their corresponding average (darkened line) for three different time periods of the drying process (a,  $t = 0$  min; b,  $t = 41.2$  min; and c,  $t = 81.2$  min). In this work 10 single spectral scans were averaged to give a single NIR spectrum every 10 s. This procedure is exemplified in Fig. 5, where the average of the 10 single scans was converted to the absorbance spectra. The resulting absorbance signal still shows a relatively high level of noise, which is further reduced with a moving averaging filter

**Fig. 3** A typical data set for a drying experiment (Bx 14). Granules are charged cold and fluidised by a gas stream of 600 L/min at  $\sim 25^\circ\text{C}$ . As water evaporates, heat is removed from the gas stream resulting in a  $\sim 10^\circ\text{C}$  temperature difference between the bed and the fluidising gas. Moisture content is monitored by NIR using the PLS model (Eq. 1), and measured offline by LOD. When the granule water content is reduced by 85% the evaporation rate drops and the bed temperature rises. Equilibrium between the inlet gas (ambient RH) and the granules ( $\sim 1.5$  wt% water) is reached after  $\sim 90$  min.





**Fig. 4** An overview of the sensor data analysis strategy used to demonstrate the versatility of the new MEMS FPI NIR sensor. Data generated by sensors is pre-processed and delivered to algorithms to construct models from calibration data. With these models in place, new NIR spectra recorded can be interpreted immediately.

that averages the data of the currently acquired spectrum with the previous 74. The large filter window (750 s) provides a smoother variation of the spectral observations over time and only slightly reduces the response time of the moisture content predictions.

After the pre-processing steps, the drying evolution can be clearly observed in the spectra (Fig. 6): the absorbance of the -OH absorption band (1940 nm), associated with the presence of water in the material, strongly reduces as drying progresses.

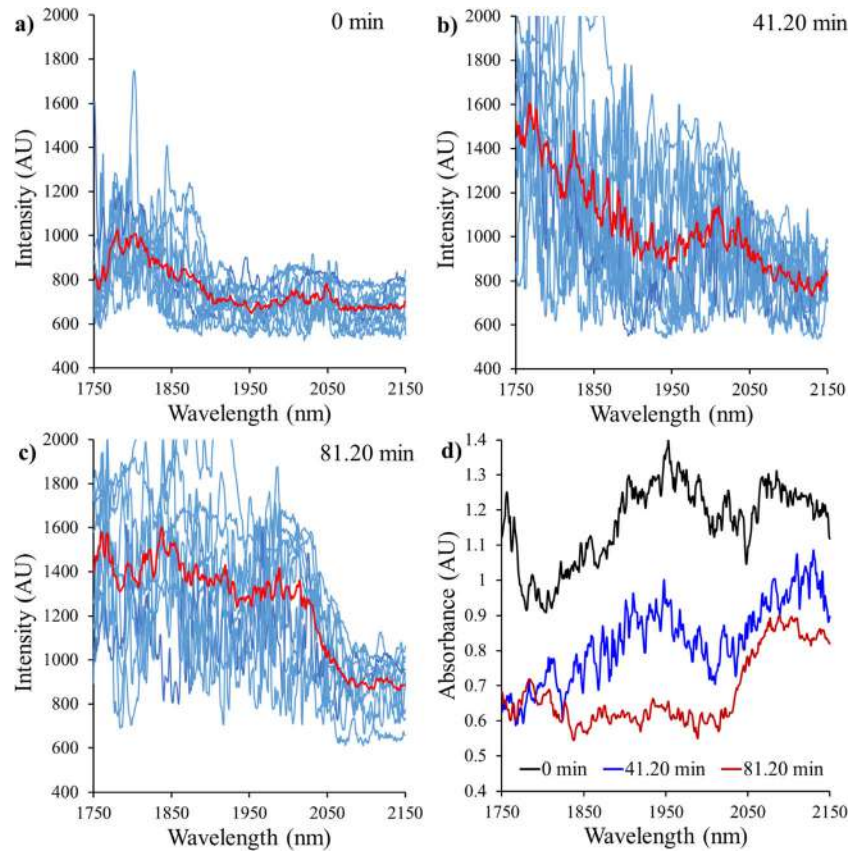
Fouling is another issue observed when acquiring *in situ* NIR spectra in a fluidised bed system. This problem often occurred at the beginning of the drying process, since granules with a moisture content above 20 wt% are very wet, and have a strong tendency to stick to the probe's window. This causes more reflection, and thus higher intensity NIR spectra, which shifts the signal to greater values compared to normal operation with a clean window. Fortunately, continuous collision of granules on the probe window causes a degree of self-cleaning. Hence, if granules stick for periods significantly smaller than 750 s, the pre-processing steps will reduce the impact of fouling on moisture content measurements.

### Moisture Content Prediction with the NIR Sensor

A PLS model was developed by relating the pre-processed NIR spectra from the initial six batches (Table I), to the offline measured LOD moisture content. For the six calibration batches (Bx1 to 6), different, narrow spectral regions including the spectral range corresponding to the strong moisture band (1900–2000 nm) were tested separately but the smallest overall moisture prediction error was found using the full spectral range available from the sensor (1750–2150 nm).

Figure 7 shows the results obtained for six selected batches comparing the LOD analytical moisture content (circles) with the predicted moisture profiles obtained from the online NIR spectra and the PLS regression model (continuous line). The figure shows (i) two calibration batches used for developing the PLS model (batches 2 and 4, using 850 L/min), (ii) two validation batches with the same flow rate (batches 7 and 9, using 850 L/min), and (iii) two validation batches using a slower flow rate (batches 13 and 14 using 600 L/min). Similar plots showing the results obtained for all fourteen batches can be found in the [Supplementary Material](#). Generally, the moisture

**Fig. 5** Construction of absorbance signal from the average (red line) of 10 consecutive single scans (blue thin background lines) for three time periods of the drying process: **(a)**  $t = 0$ ; **(b)**  $t = 41.2$  min; and **(c)**  $t = 81.2$  min, **(d)** Absorbance spectra obtained from the resulting average intensity. Dark signal level of the detector was approximately 500 intensity units.



content resulting from the NIR spectra provides a good estimate of the data measured offline by LOD when the moisture content  $f_w < 20wt\%$ .

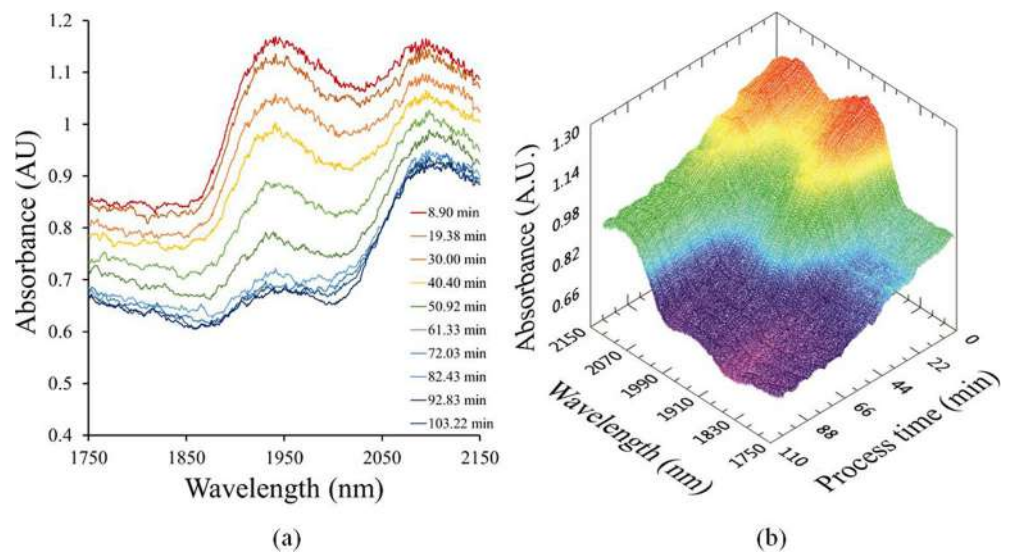
A direct comparison of the moisture content measured using LOD with the predicted results from online NIR spectra is given in Fig. 8. As the PLS model fitted  $\log_{10}f_w$ , we expect the relative error in the prediction to be constant for different levels of water content.

The average residual for all  $N$  LOD measurements ( $\sim 200$  LOD values) may be defined as:

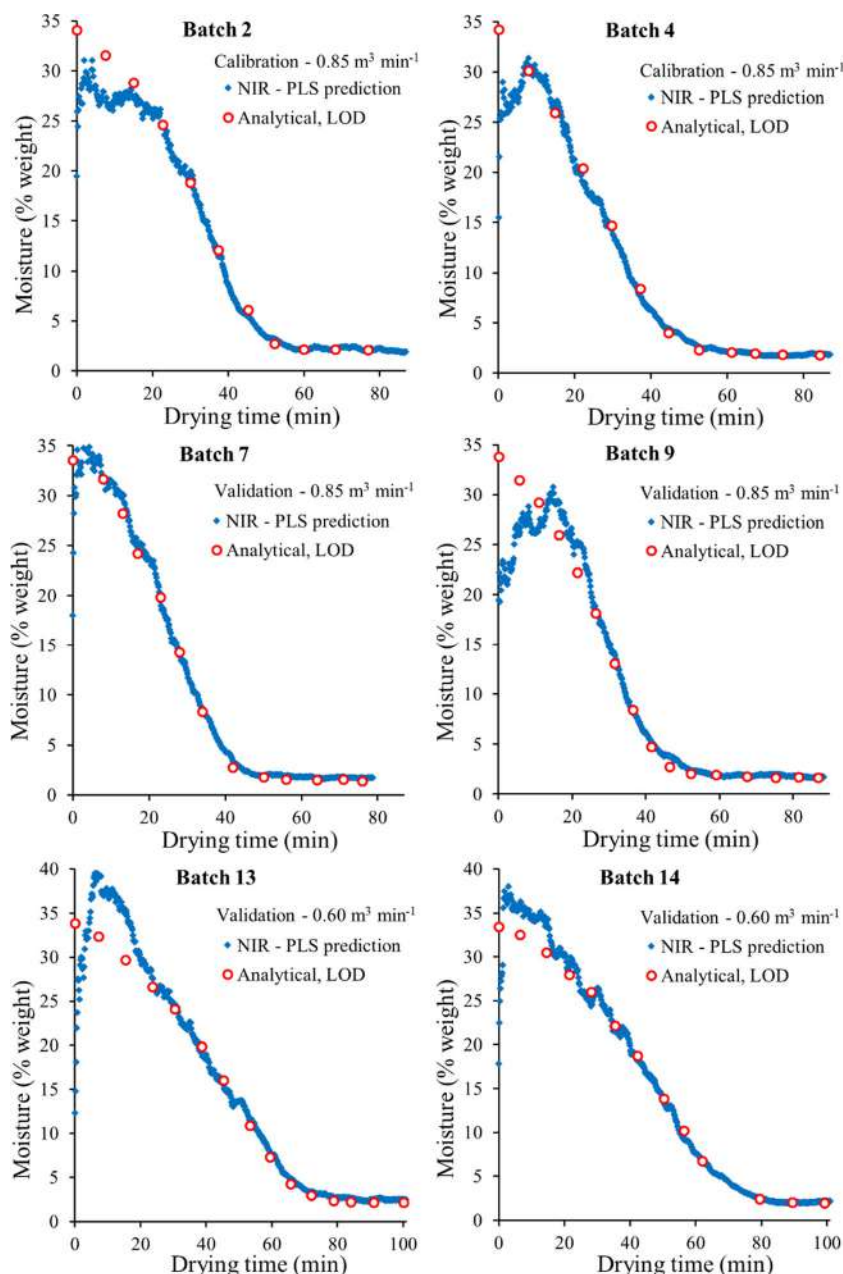
$$e_{rel} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left( \frac{f_{w,NIR} - f_{w,LOD}}{f_{w,LOD}} \right)^2} = 13\% \quad (6)$$

From Fig. 8 it is clear that the error in the NIR based water content is independent of the absolute extent of the moisture

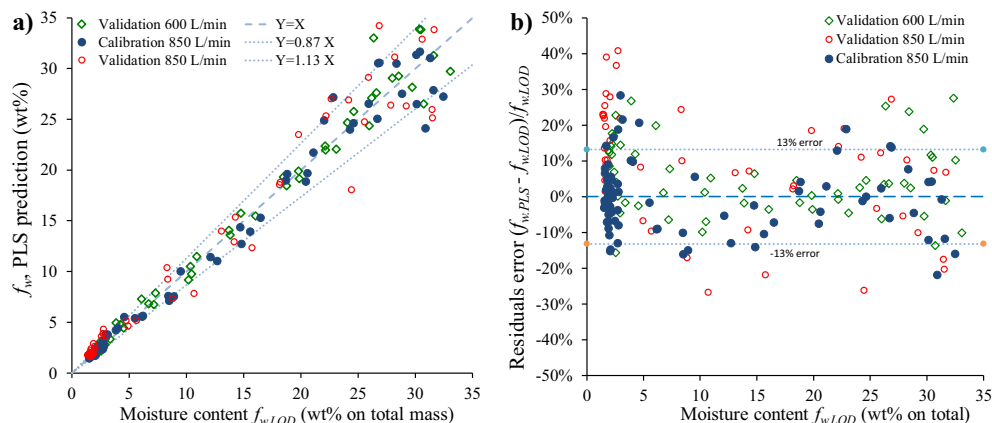
**Fig. 6** Change in water content observed from variations in the absorbance spectra **(a)**: for 10 specific periods of the drying process (obtained after applying the pre-processing steps). **(b)**: NIR profile evolution observed for the complete drying process.



**Fig. 7** Analytical moisture content determined using LOD (discrete circles) compared to the prediction profiles obtained from the PLS regression model using online NIR spectra (semi continuous line), including two calibration batches (batches 2 and 4 using 850 L/min), and four validation batches using two flow rates (batches 7 and 9 using 850 L/min; batches 13 and 14 using 600 L/min).



**Fig. 8** Comparison of the moisture content  $f_{w, PLS}$ , as measured by the NIR sensor with the PLS statistical model, and the analytical LOD data  $f_{w, LOD}$ . **(a)** direct comparison and **(b)** relative residuals of the prediction,  $\{(f_{w, PLS} - f_{w, LOD})/f_{w, LOD}\}$ .



content. The error for the calibration batches is  $\pm 10\%$ . However, all models over predict the LOD concentration below 5 wt%: 1% higher for the calibration batch, 18% for the 850 L/min and 8% for the 600 L/min validation batches. The high deviation at low water levels for the 850 L/min batches may be due to an uncontrolled parameter such as the humidity.

Comparing results to previous studies is difficult as the water concentration ranges used to build principal component models vary significantly. The pre-processing strategy was similar in all cited cases, averaging a large number of scans (from 32 to 300 vs 750 in this study) to compensate for the spectral noise. Peinado *et al.* (23) reported data for water contents  $f_w$  between 0.6 wt% to 2.8 wt% with a relative error  $\varepsilon_{rel} \approx 15\%$  using an ABB Fourier Transform Process Analyser Near Infrared spectrometer, with thermo electrically cooled InGaAs detectors (ABB-FTPA2000260). Fonteyne *et al.* (19) reported data for  $f_w$  in the range of 3.5 to 7 wt% with a relative error  $\varepsilon_{rel} = 5\%$  using a Matrix™ –F Duplex, Bruker Optics Ltd., FT-NIR spectrometer (32 single scans averaged; using 1000–2220 nm for analysis). The residuals obtained using a commercial dispersive spectrometer varied  $\pm 4$  wt% over 4–20 wt% water content (32 single scans averaged; using 1100–2500 nm for analysis) (28). The results obtained for the calibration and the validation batches with the novel MEMS-FPI sensor ( $\varepsilon_{rel}=13\%$ ) were similar to those reported with conventional spectrometers. We did however observe significantly larger relative errors below 5 wt% water in the validation batches. The absolute errors remained low ( $\sim 0.4$  wt%). Based on the operation over a significant period, we feel this is more likely to be due to changes in uncontrolled parameters

### Process End-Point Detection from MSPC Charts

A different application to evaluate the performance of the MEMS-FPI is the detection of the process end-point from the NIR signal. We use the MSPC model (Eq. 2) based on the end-point spectra from the calibration batches. This model required two principal components, PC1 marks the inverted water band and PC2, with a less interpretable shape, is required for the description of batch-to-batch variability (the PC loadings and related description are provided in the supplementary information S4).  $Q_{stat}$  control charts were calculated from Eqs. 2–4 for validation batches by projecting NIR spectral observations (using same pre-processing procedure as before) onto the developed model. Figure 9 shows the  $Q_{stat}$  MSPC charts obtained for six validation batches. Detected end-points are indicated with a yellow diamond marker in the  $Q_{stat}$  control charts. Batches 1, 5, 8 and 14 reached a final moisture content below 2% (on-specification), batches 12 and 13 did not (off-specification). For comparison, moisture content levels from the NIR spectra were compared in the MSPC control charts for batches 5 and 13 (moisture axis is at the right of the plot). These plots in Fig. 9 show that

based on moisture levels the endpoint would have been delayed for batch 5; both  $Q_{stat}$  and the moisture level agree that batch 13 is off specification.

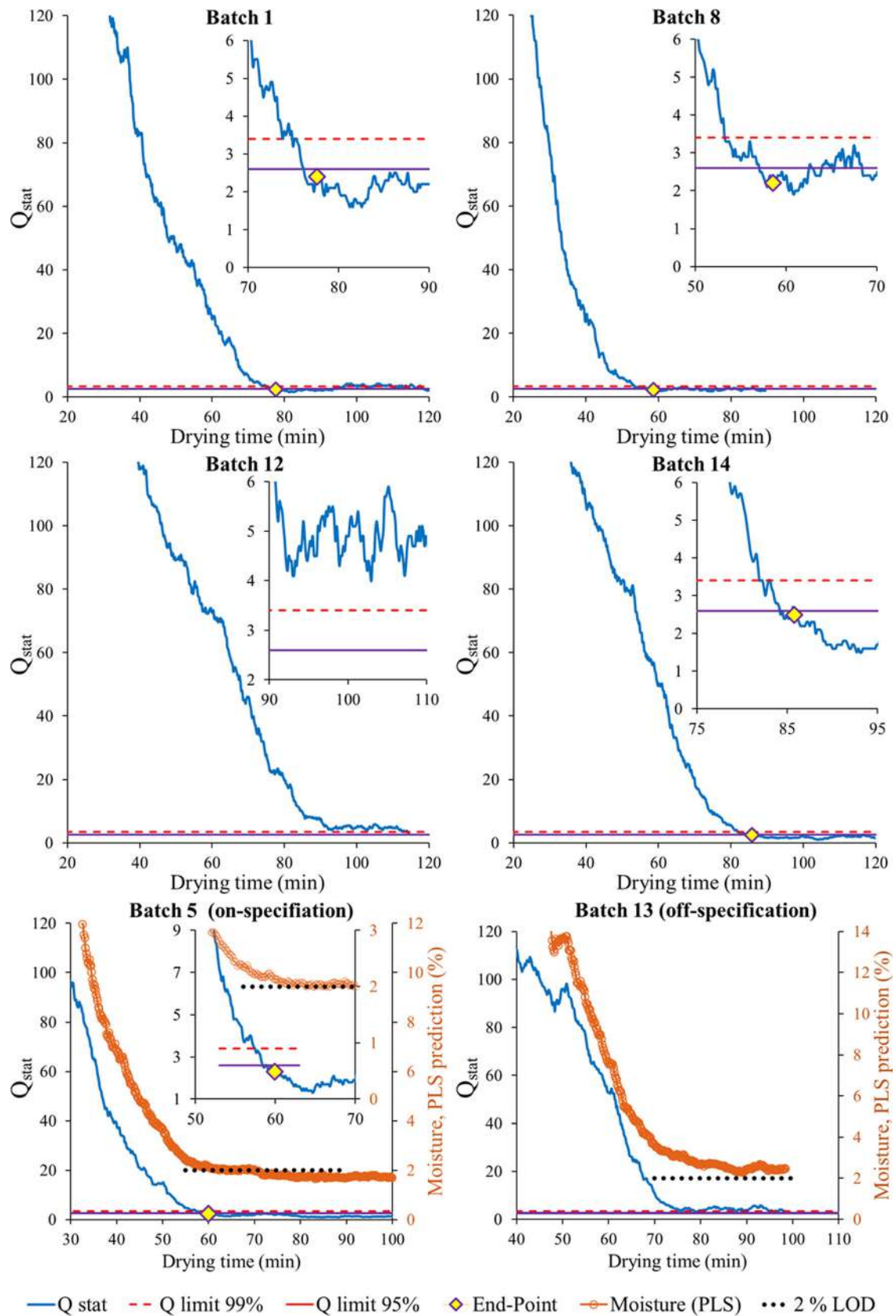
The results in Fig. 9 also confirm that a lower gas flow rate (batches 12–14, see Table I) significantly increases the endpoint process time (compared with on-specification batches, 1, 5 and 8). Similar results were observed for the other validation batches (see Supplementary Material). The spectral quality of the NIR sensor and its robustness is clearly sufficient for the construction and application of PCA-based MSPC models. The device has clear potential in end-point detection applications, promising a significant reduction in offline moisture measurements.

### Process Monitoring (Mass Transfer Resistance)

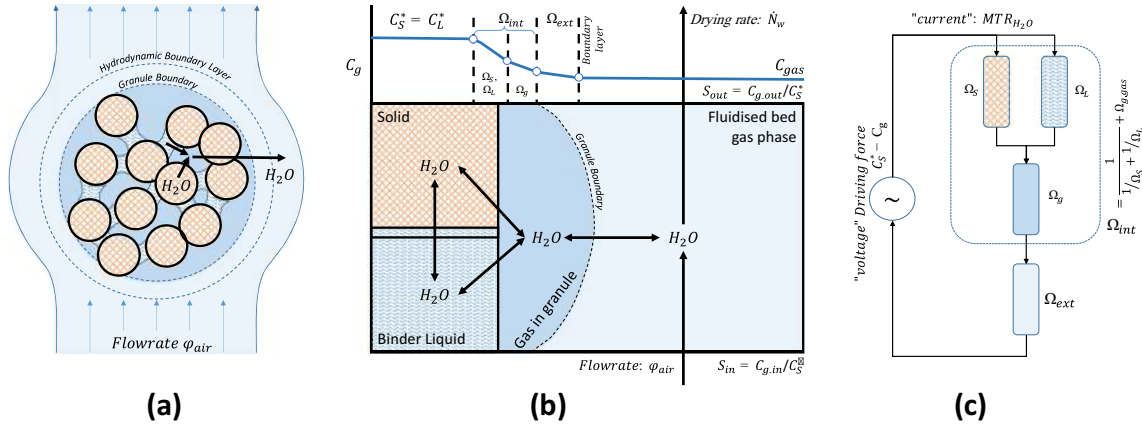
The promise of low cost NIR devices lies in the common availability of online compositional analysis. Our statistical analysis (PLS and MSPC) demonstrates the low cost MEMS-FPI NIR sensor to be a device that is suitable for composition measurement, with sufficient performance compared to conventional sensors when applied to fluid bed systems. To date NIR data has not been used in conjunction with mechanistic drying models to provide scale up data. As can be seen in Fig. 4, such analysis is complex, and requires the fusion of data from multiple sensor, as well as an understanding of material properties such as water adsorption isotherms and the vapour pressure of water. Methods to determine the bed moisture content and drying rate from temperature and humidity data show a significant deviation from samples analysed by LOD (33). To demonstrate the utility and value of continuous composition data in process monitoring we developed a methodology to monitor the process by evaluating the mass transfer resistance(s) from the available data. These resistances underpin fluidised bed scale up calculations. An overview of the mass transfer model is given in Fig. 10.

#### Step 1: Obtain the Molar Drying Rate

To evaluate the mass transfer we first must convert the Drying curve  $f_w(t)$ , to the drying rate in mol/s by differentiation of the water content  $N_w = \frac{1}{0.018} \frac{m_w f_w(t)}{1-f_w(t)}$  in the bed. The slope of the drying curve results from linear regression of a line to 36 data points either side of time  $t$  (a total of 72 points over 720 s). The standard error in the slope obtained is between 3%–5%. An example of  $f_w$ , and its derivative  $df_w/dt$  are shown in Fig. 11 row 1. The first graph shows the conventional drying curve consisting of the initial transient phase followed by the constant drying rate period (20–60 min,  $\frac{df_w}{dt} \approx 0.8$  wt%/min) and the falling rate period and finally equilibrium ( $f_w \approx 1.6$  wt%). Row 2 of the same figure shows the molar drying rate at approximately 0.15 mol/s in the constant drying rate regime.



**Fig. 9** MSPC control charts for batches 1, 5, 8 and 14 (on-specification), 12 and 13 (off-specification). Inserted figures show in detail the final time range of the drying process and the process end-point, identified when 10 consecutive observations of  $Q_{stat}$  values were below the 95% control limit. Batches 5 and 13 include the moisture predictions from PLS model for reference (secondary axis).



**Fig. 10** (a) a granule consists of solid held together by liquid bridges formed by the binder fluid. Water is also absorbed by solids ( $\sim 0.3 \text{ g}_{\text{water}}/\text{g}_{\text{solid}}$ ). (b) The process scheme shows the location of water in different environments (“phases”) with arrows representing mass transfer between the environments. The gas phase concentration over (in equilibrium with) the solid ( $C_S^*$ ) and liquid ( $C_L^*$ ) are assumed to be similar. (c) The mass transfer may be represented as a resistance model with a “current” of water ( $MTR_{H_2O}, \frac{\text{mol}}{\text{s} \cdot m_{\text{bed}}^2}$ ) flowing from high to low concentration. The transport through each environment requires a fall in concentration that is proportional to the “current”:  $\Delta C = MTR_{H_2O} \times \Omega$ , where  $\Omega$  is the so called mass transfer resistance. The driving force (the “voltage”, the sum of all  $\Delta C$ ) equals the concentration difference between the source of the water and the final sink, the fluidising gas:  $C_S^* - C_g$ .

**Step 2 Work out the Driving Force**

Granules consist of solids bound together by a binder fluid (Fig. 10a). Water is present in liquid bridges, but also adsorbed to some of the solid materials. The process scheme (Fig. 10b) shows the location of water in different environments (e.g. different phases and segregated domains) with arrows representing mass transfer between the environments. The mass transfer process is a diffusion process driven by the concentration gradient between water at the source in the granule and the water in the fluidising gas (the sink). As the liquid and solid phases in the granule are in intimate contact we assume equilibrium hence the gas phase concentration over (in equilibrium with) the solid ( $C_S^*$ ) and liquid ( $C_L^*$ ) are the same. As at some point the liquid phase will disappear, we focus our attention on obtaining  $C_S^*$  as function of temperature and water content. Thermodynamically, the vapour pressure over a solid or liquid phase is represented by the product of the activity of water in that phase  $a_w$  and the vapour pressure of water at the bed temperature,  $P_w^*(T_{bed})$ :

$$C_S^* = \frac{a_w P_w^*(T_{bed})}{RT_{bed}} \frac{\text{mol}}{\text{m}^3} \tag{7}$$

The equilibrium vapour pressure of pure water is given by the Antoine equation (34), supplementary material, here expressed relative to the vapour pressure at a reference temperature of 20 °C:

$$P_w^*(T(^{\circ}\text{C})) = 2339.1 * e^{-4078.8 * (\frac{1}{236.63+T(^{\circ}\text{C})} - \frac{1}{236.63})} \pm 1 \text{ Pa} \tag{8}$$

The water activity ( $a_w$ ) follows from the water adsorption isotherms of the materials present in the placebo.<sup>2</sup> Such isotherms relate  $a_w$  to the water content on a *dry weight* bases (Fig. 12a). The GAB correlation, an extended BET equation developed by Guggenheim, Andersen and de Boer (35,36), expresses water content of solid  $i$  ( $F_{wi}$ , dry basis) as function of the water activity (Table II, and supplementary material):

$$F_{wi}(a_w) = \frac{m_{wi}}{m_{si}} = \frac{k_{wi} a_w C_{GAB_i} m_{o_i}}{(1 - k_{wi} a_w)(1 + (C_{GAB_i} - 1)k_{wi} a_w)} \tag{9}$$

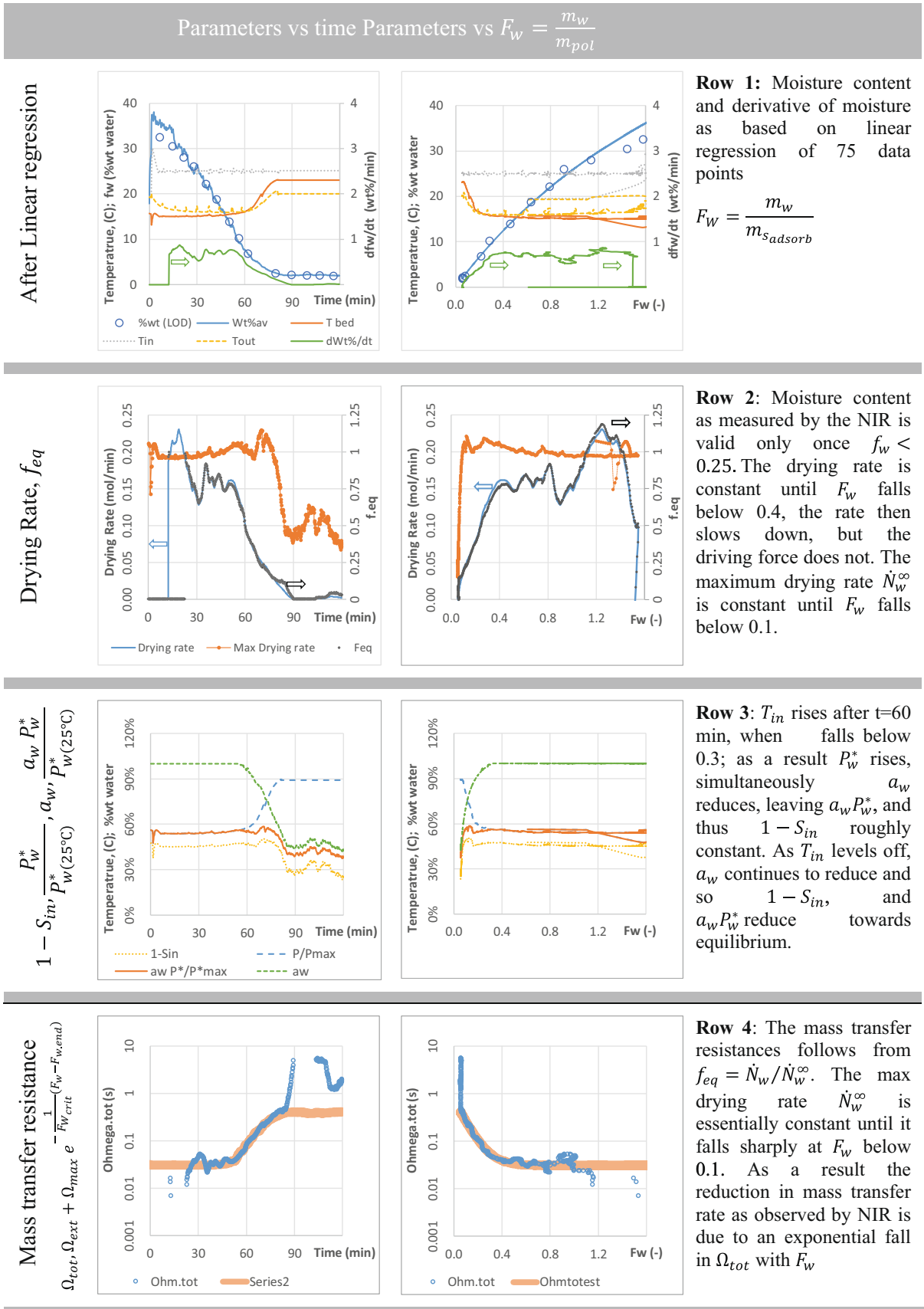
At the beginning of the drying the Hypromellose 2910 (5 wt% dry basis) and Ac-Di-Sol (1.5 wt%) contain up to 50% of the adsorbed water with the remainder adsorbed onto the Avicel (29 wt%). As the granule dries, this reduces to ~10% and most of the remaining water is associated with the Avicel. Assuming the materials do not interact, an aggregate isotherm may be constructed by combining the water content adsorbed by the various materials at the same water activity (Fig. 12b). The aggregate isotherm correlates the granule’s water activity to  $F_w = m_w/m_{adsorb}$ , where  $m_{adsorb}$  is the combined mass of Avicel, Hypromellose and Ac-Di-Sol (Fig. 12c, fitted GAB parameters in Table 22).

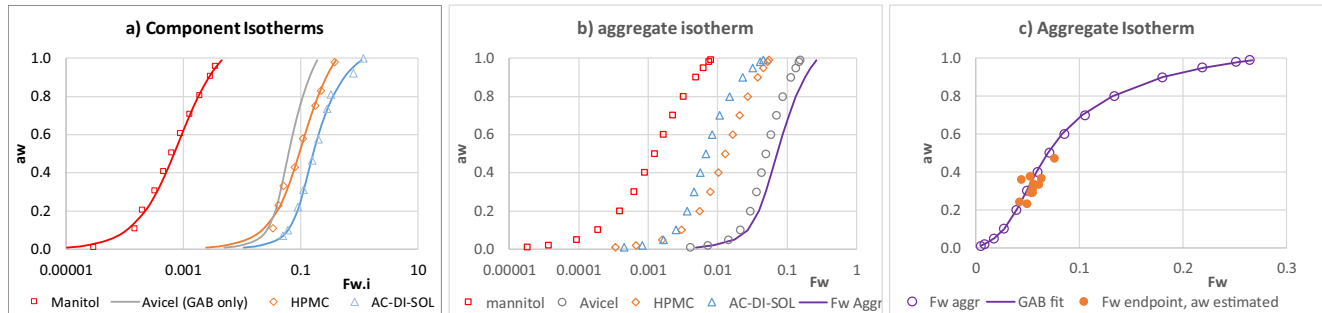
**Step 3: Link the Molar Drying Rate and the Driving Force**

The mass transfer process can be represented with a resistance model ((41), Fig. 10c) that sees a “current” of water ( $MTR_w, \frac{\text{mol}}{\text{s} \cdot m_{\text{bed}}^2}$ ) flowing from the high concentration at the

<sup>2</sup> We ignore the reduction in water activity due to the presence of solutes, predominantly Mannitol at ~ 1 mol/L, which reduce  $a_w$  from 1 to 0.98 by Raoult’s law.







**Fig. 12** The construction of the aggregate isotherms: **(a)** Isotherms of individual components data is taken from the references listed table X, lines are the GAB equation fitted to the data using the parameters in Table X. **(b)** Contribution of the individual materials to the total water content  $F_w$  based on the mass of the polymers (mannitol weight not included) **(c)** The GAB fit to the aggregate isotherm. The estimated relative humidity plotted versus the moisture content of the final product measured by LOD corresponds well to the aggregate isotherm.

source (liquid bridges, solids) to low concentration in the gas used to fluidised the bed. The transport through each environment requires a fall in concentration that is proportional to the “current”,  $\Delta C_i = MTR_{iw} \times \Omega_i$ , where  $\Omega_i$  is the a called a mass transfer resistance which has the units of seconds. The overall driving force (equivalent to “voltage”) is the sum of all these  $\Delta C_i$  and equals the concentration difference between the source of the water and its final sink, the fluidising gas:

$$C_S^* - C_g = \sum \Delta C_i = MTR_w \times \sum \Omega_i \tag{10}$$

The above equation shows the total mass transfer resistance  $\Omega_{tot} = \sum \Omega_i$  to be the sum of the individual resistances, in similarity with Ohm’s law. As the residence time of the gas is short ( $< 100$  ms) it is common in most fluidised bed models to assume that the mass transfer resistance  $\Omega_{tot}$  and the bed’s temperature and moisture content are constant on the time scale required for the gas to flow from the bottom to the top of the bed. A mass balance over a horizontal slice of the bed with volume  $dV_{bed}$  requires the gain of water in the gas flow ( $\phi_g dC_g$ ) to be equal to the mass transferred from the granules to the air ( $MTR_w dV_{bed}$ ):

$$\phi_g dC_g = MTR_w dV_{bed} = MTR_w \frac{1}{f_s \rho_s} dm_s \tag{11}$$

Here  $f_s$  is the volume fraction solids in the fluidised bed (estimated at 40%), and the  $\rho_s$  solids skeletal density (averaged at  $1500 \text{ kg/m}^3$ ). Substitution of Eq. 7 in 11 and integration yields (see [supplementary material](#)):

$$C_{g,out} - C_{g,in} = f_{MTR} (1 - S_{in}) C_S^* \text{ with } f_{MTR} = \left( 1 - e^{-\frac{m_s}{\Omega_{tot} f_s \rho_s \phi_g}} \right) \tag{12}$$

Here  $S_{in} = C_{g,in} / C_S^*$  is the degree of saturation of the inlet gas which varies between 0 (no water) to 1 for an inlet gas in equilibrium with the water in the granules. It is important to realise that the saturation of the inlet gas ( $S_{in}$ ) may change during processing, as  $C_S^*$  varies with both bed temperature and water content. We estimated  $C_{g,in}$  such that the outlet air is saturated at the beginning of the constant drying rate period.

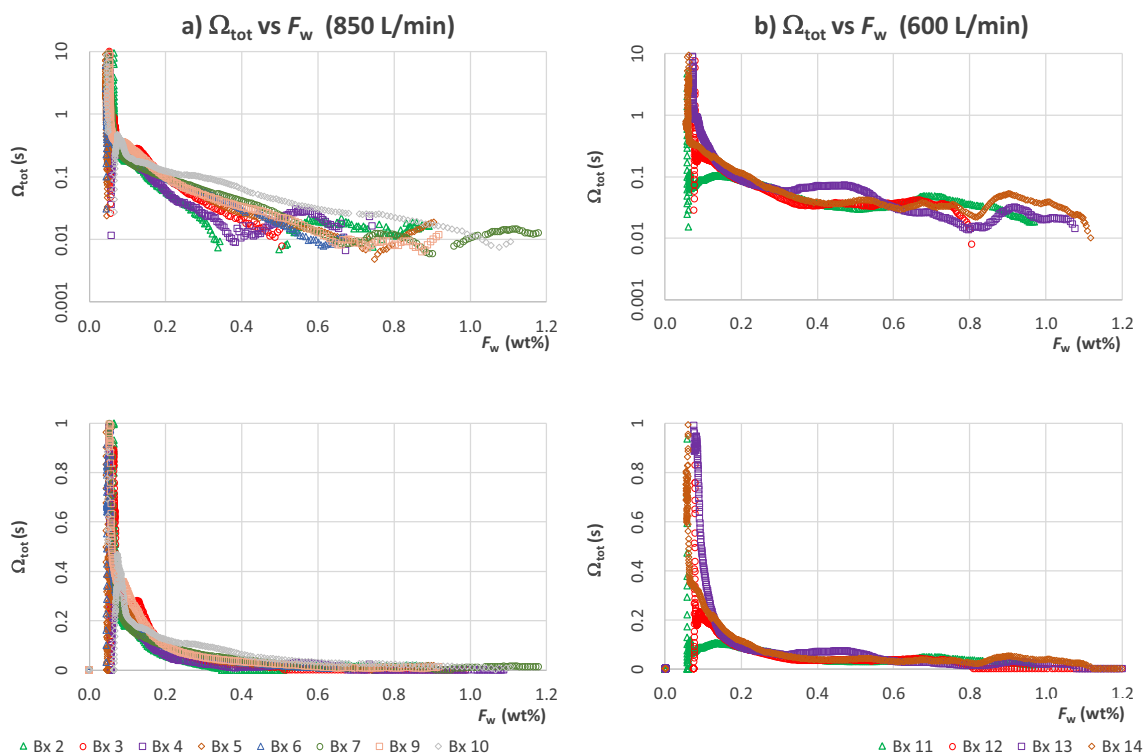
The molar drying rate  $\dot{N}_w$  in the fluid bed dryer now follows from the air flowrate  $\phi_g$  and the concentration change calculated from Eq. 12:

$$\begin{aligned} \dot{N}_w &= (C_{g,out} - C_{g,in}) \phi_g = f_{MTR} \dot{N}_w^\infty \text{ with } \dot{N}_w^\infty \\ &= (1 - S_{in}) C_S^* \phi_g \end{aligned} \tag{13}$$

$f_{MTR}$  is the extent to which mass transfer is limiting: when  $f_{MTR} = 0$  the mass transfer resistances are high and no significant transfer occurs. If on the other hand  $f_{MTR} = 1$  then mass

**Table II** Fitting Parameters for the GAB Equation of the Materials in the Placebo Formulation

Material	mass (gr)	$m_{0i}$	$C_{GAB_i}$	$k_{wi}$	Reference
Mannitol	213	0.00068	1.73	0.87	Data (37)
Avicel PH-101	96	0.040	17.4	0.80	GAB param (38)
Hypromellose 2910	17	0.018	18.9	0.99	Data (39)
AC-Di-Sol	5	0.095	13.4	0.92	Data (40)
Aggregate	118	0.0441	14.30	0.846	Mass weighted average



**Fig. 13** Mass transfer resistance curves for repeat batches; top) logarithmic Y axis, bottom) linear Y axis. The data demonstrates that at air flowrate of 850 L/min (a) the initial and final mass transfer resistance are relatively constant, but the internal resistance starts to dominate at widely different moisture content  $F_w = m_w/m_{adsorb}$ . At 600 L/min (b) we observe more consistent mass transfer resistance trajectories.

transfer is instantaneous, and the gas phase leaves saturated resulting in the maximum drying rate  $\dot{N}_w^\infty$ . The 2rd row in Fig. 11 shows the molar and maximum drying rates. The drying rate is about 80% of the maximum drying rate in the constant rate period which ends at  $F_w \approx 0.35$ , after which the rate reduces in a manner that appears proportional with  $F_w$ . Conversely, the maximum drying rate remains stable at  $F_w < 0.35$ , as  $T_{bed}$  and  $P_W^*(T_{bed})$  increase balanced by a reduction in the water activity  $a_w$  as water is removed. The reduction of  $a_w$  becomes dominate when  $F_w < 0.1$  the driving force and drying rates reduce then sharply.

**Step 4 Calculate the Overall Mass Transfer Resistance**

The mass transfer resistance  $\Omega_{tot}$  follows from the ratio of the molar drying rate  $\dot{N}_w$  measured by NIR, and the maximum

drying rate  $\dot{N}_w^\infty$  that follows from the bed temperature and the gas flowrate:

$$f_{MTR} = \frac{\dot{N}_w}{\dot{N}_w^\infty} = \frac{\dot{N}_w}{(1 - S_{in})C_S^*\phi_g} = 1 - e^{-\frac{m_s}{\Omega_{tot}f_s\rho_s\phi_g}} \quad (14)$$

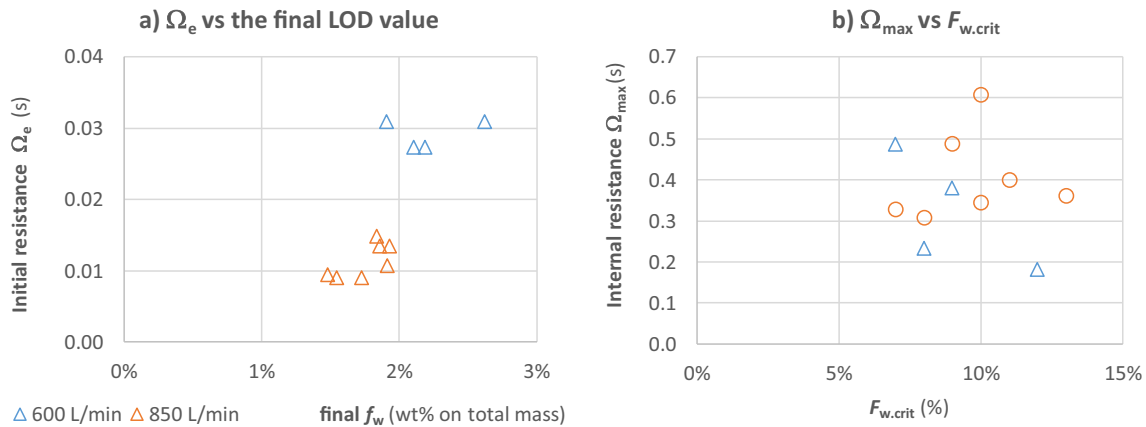
The measured temperature and water content data combined with the aggregate isotherm and the vapour pressure of water allows calculation of  $f_{MTR}$ . The overall mass transfer resistance then follows by rearranging

$$\Omega_{tot} = -\frac{m_s}{Ln(1 - f_{MTR})f_s\rho_s\phi_g} \quad (15)$$

This is shown in row 4 of Fig. 11. After the steady state is reached,  $\Omega_{tot} \approx 0.03s$ , but once  $F_w$  drops below 0.4, the mass transfer resistance starts to increase, eventually it is an order of magnitude higher. This behaviour is observed in all batches

**Table III** Average Mass Transfer Resistances for the Placebo Granules

Flowrate	600 L/min	850 L/min	All
Data points	4	8	12
$\Omega_{ext}$ (s)	$0.028 \pm 15\%$	$0.013 \pm 27\%$	
$\Omega_{max}$ (s)	$0.58 \pm 33\%$	$0.68 \pm 34\%$	$0.64 \pm 33\%$
$F_{W_{crit}}$ (-)	$0.10 \pm 20\%$	$0.10 \pm 31\%$	$0.10 \pm 27\%$



**Fig. 14** Mass transfer resistance parameters (a) the external resistance  $\Omega_e$  reduces with increasing flowrate, (b) the internal resistance increases exponentially with water content  $F_w = m_w/m_{absorb}$ . The resistance at the end of drying,  $\Omega_{max}$ , is an order of magnitude higher than the initial external resistance  $\Omega_e$ .

(Fig. 13). At 600 L/min, the curves of the different repeat batches are consistent, even though the 20% relative error in moisture content level does result in significant fluctuations around the mean. The increase in the internal resistance by a factor 30 to 70 is clearly visible in all batches displayed bar Bx 11.

#### Step 5 Evaluation of the Observed Mass Transfer Resistance

To parametrise, the observed mass transfer resistances we based on a constant resistance external to the granule, and an internal granule resistance that falls exponentially with moisture content:

$$\Omega_{tot} = \Omega_{ext} + \Omega_{max} e^{-\frac{1}{F_{w,crit}}(F_w - F_{w,end})} \quad (16)$$

Here  $F_{w,end}$  is the final moisture content relative to  $m_{adsorb}$ . Table III and Fig. 14 shows these parameters for the experiments conducted. As expected, the external resistance varies with airflow reducing from  $0.028 \text{ s} \pm 15\%$  at 600 L/min, to  $0.013 \text{ s} \pm 27\%$  for 850 L/min. The maximum resistance ( $0.64 \pm 33\%$ ) and the critical moisture content  $F_{w,crit}$  ( $0.10 \pm 27\%$ ) appears to be independent of the flowrate. The external resistance will dominate at moisture contents above  $F_w = 0.5$  as only  $e^{-(0.5-0.1)/0.1} \approx 2\%$  of the internal resistance remains ( $\sim 0.014 \text{ s}$ ).

The obtained mass transfer parameters are difficult to reconcile with literature, as generally only the drying rate is reported (33,42). It is worth noting that the variation in the estimated parameters related to mass transfer is double the error in the NIR based moisture content ( $\pm 13\%$ ). Even so, the mass transfer analysis using low cost NIR sensors is able to detect the change from externally controlled mass transfer (the so called constant rate period) to mass transfer limited by the internal resistance of the granule. The rate of change of the internal resistance is exponential, which is inconsistent with a shrinking core model in which the volume of the granule that contains

water shrinks towards the core of the granule (43). Further work with controlled humidity will be required to see if the analysis is robust.

A detailed phenomenological interpretation of the presented results is beyond the scope of this paper, but it appears that the desorption kinetics dominate mass transfer once the binder liquid droplets have evaporated. It follows that the ingredient selection and adsorption characteristics can have a profound effect on the drying time required.

## CONCLUSIONS

A new reduced cost and small form factor MEMS FPI NIR sensor has been tested over a 9 month period during which 14 batches of placebo granules have been manufactured and dried in a fluidised bed. Overall, the MEMS-FPI sensor performance gave a very satisfactory stability and reproducibility and delivered high quality, continuous data robustly during fluid bed drying of placebo pharmaceutical granules; a typical processing scenario in which acquired NIR spectra have a very low signal to noise ratio. We tested the sensors performance with three applications: moisture monitoring, end-point detection and process analysis (mass transfer monitoring).

In fluidised beds abrupt changes in the spectra collected from scan to scan occur because of the random motion of the placebo granules. Using spectra averaged over 12 min, a satisfactory and robust PLS regression model was developed to predict the moisture content from NIR. The accuracy of the moisture content prediction over a significant number of batches and an experimental period of 3 months remained constant at a relative error of 13%. This is of a similar magnitude as reported for fluidised beds dryers using high specification commercial NIR spectrometers. The consistent performance demonstrates the NIR sensor potential for use as process sensor.

In the second application, NIR spectra collected were used to develop a MSPC model with a 2% endpoint moisture

content target. This allowed successful endpoint detection, and correct identification of off-specification batches using only the NIR sensor's data.

To demonstrate the utility and potential benefits of having cheap online sensors available for process monitoring, we fused temperature and NIR generated moisture data to determine the mass transfer resistance. Our analysis indicates that for the placebo granules the overall mass transfer resistance is the sum of a gasflow dependent external resistance (0.01–0.03 s) and a moisture content dependent internal resistance that increases exponentially as moisture content reduces (0.0 to 0.7 s). We demonstrated that this low cost NIR sensor allows the detection of changes in the drying mechanisms, which may give an early warning if unspecified physico-chemical properties of input materials (such as water adsorption isotherms) have changed.

In summary, the small form factor MEMS-FPI sensors has been shown to be a robust alternative for process monitoring. It is robust to vibration and temperature changes and straightforward to install. Its NIR spectra are of a sufficient quality to deliver composition related predictions with the same accuracy as commercial spectrometers in a system with an extremely low signal to noise ratio. The MEMS chip spectrometer can be mass produced and has a small enough form factor to be integrated in the next generation of plant sensors. Besides, it is cheap enough to allow multi point composition sensing in the way that is done to day for temperature, pressure and flow rate measurement.

## ACKNOWLEDGMENTS AND DISCLOSURES

The presented work has received funding from the ProPAT project (European Union Horizon 2020 research and innovation programme under grant agreement No 637232). The authors also thank Spectral Engines (Finland) for facilitating the MEMS-FPI sensors used in this study, and particularly Uula Kantojärvi and Matti Tammi and for their support.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## REFERENCES

1. Pasquini C. Near infrared spectroscopy: a mature analytical technique with new perspectives - a review. *Anal Chim Acta*. 2018;1026:8–36.
2. Calvo NL, Maggio RM, Kaufman TS. Characterization of pharmaceutically relevant materials at the solid state employing chemometrics methods. *J Pharm Biomed Anal*. 2018;147:538–64.
3. Rantanen J, Wikstrom H, Turner R, Taylor LS. Use of in-line near-infrared spectroscopy in combination with chemometrics for improved understanding of pharmaceutical processes. *Anal Chem*. 2005;77(2):556–63.
4. Roggo Y, Chalou P, Maurer L, Lema-Martinez C, Edmond A, Jent N. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J Pharm Biomed Anal*. 2007;44(3):683–700.
5. De Beer T, Burggraef A, Fonteyne M, Saerens L, Remon JP, Vervaeke C. Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *Int J Pharm*. 2011;417(1–2):32–47.
6. Blomberg M, Torkkeli A, Lehto A, Helenelund C, Viitasalo M. Electrically tuneable micromachined Fabry-Perot interferometer in gas analysis. *Phys Scr*. 1997;T69:119–21.
7. Vaughan J. The Fabry-Perot-interferometer - history, theory, practice and applications. New York: Taylor & Francis; 1989.
8. Akujarvi A, Guo B, Mannila R, Rissanen A. MOEMS FPI sensors for NIR - MIR microspectrometer applications. *Moems and Miniaturized Systems Xv* 2016;9760.
9. Rissanen A, Guo B, Saari H, Nasila A, Mannila R, Akujarvi A, *et al*. VTT's Fabry-Perot interferometer technologies for hyperspectral imaging and mobile sensing applications. *Moems and Miniaturized Systems Xvi* 2017;10116.
10. Vakili H, Wickstrom H, Desai D, Preis M, Sandler N. Application of a handheld NIR spectrometer in prediction of drug content in inkjet printed orodispersible formulations containing prednisolone and levothyroxine. *Int J Pharm*. 2017;524(1–2):414–23.
11. Huck CW. Selected latest applications of molecular spectroscopy in natural product analysis. *Phytochem Lett*. 2017;20:491–8.
12. Rosas JG, Blanco M, Gonzalez JM, Alcalá M. Real-time determination of critical quality attributes using near-infrared spectroscopy: a contribution for Process Analytical Technology (PAT). *Talanta*. 2012;97:163–70.
13. Scheibelhofer O, Balak N, Wahl PR, Koller DM, Glasser BJ, Khinast JG. Monitoring blending of pharmaceutical powders with multipoint NIR spectroscopy. *AAPS PharmSciTech*. 2013;14(1):234–44.
14. Chablani L, Taylor MK, Mehrotra A, Rameas P, Stagner WC. Inline real-time near-infrared granule moisture measurements of a continuous granulation-drying-milling process. *AAPS PharmSciTech*. 2011;12(4):1050–5.
15. Otsuka M, Koyama A, Hattori Y. Real-time release monitoring for water content and mean particle size of granules in lab-sized fluid-bed granulator by near-infrared spectroscopy. *RSC Adv*. 2014;4(34):17461–8.
16. Lee MJ, Park CR, Kim AY, Kwon BS, Bang KH, Cho YS, *et al*. Dynamic calibration for the in-line NIR monitoring of film thickness of pharmaceutical tablets processed in a fluid-bed coater. *J Pharm Sci*. 2010;99(1):325–35.
17. Mark J, Karner M, Andre M, Rueland J, Huck CW. Online process control of a pharmaceutical intermediate in a fluidized-bed drier environment using near-infrared spectroscopy. *Anal Chem*. 2010;82(10):4209–15.
18. Nieuwmeijer FJ, Damen M, Gerich A, Rusmini F, van der Voort MK, Vromans H. Granule characterization during fluid bed drying by development of a near infrared method to determine water

- content and median granule size. *Pharm Res.* 2007;24(10):1854–61.
19. Fonteyne M, Arruabarrena J, de Beer J, Hellings M, Van Den Kerkhof T, Burggraeve A, *et al.* NIR spectroscopic method for the in-line moisture assessment during drying in a six-segmented fluid bed dryer of a continuous tablet production line: validation of quantifying abilities and uncertainty assessment. *J Pharm Biomed Anal.* 2014;100:21–7.
  20. Burgbacher J, Wiss J. Industrial applications of online monitoring of drying processes of drug substances using NIR. *Org Process Res Dev.* 2008;12(2):235–42.
  21. Heigl N, Koller DM, Glasser BJ, Muzzio FJ, Khinast JG. Quantitative on-line vs. off-line NIR analysis of fluidized bed drying with consideration of the spectral background. *Eur J Pharm Biopharm.* 2013;85(3):1064–74.
  22. Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear-regression - the Partial Least-Squares (PLS) approach to generalized inverses. *SIAM J Sci Stat Comput.* 1984;5(3):735–43.
  23. Peinado A, Hammond J, Scott A. Development, validation and transfer of a near infrared method to determine in-line the end point of a fluidised drying process for commercial production batches of an approved oral solid dose pharmaceutical product. *J Pharm Biomed Anal.* 2011;54(1):13–20.
  24. Kona R, Qu HB, Mattes R, Jancsik B, Fahmy RM, Hoag SW. Application of in-line near infrared spectroscopy and multivariate batch modeling for process monitoring in fluid bed granulation. *Int J Pharm.* 2013;452(1–2):63–72.
  25. Macgregor JF, Kourti T. Statistical process-control of multivariate processes. *Control Eng Pract.* 1995;3(3):403–14.
  26. Chen ZP, Lovett D, Morris J. Process analytical technologies and real time process control a review of some spectroscopic issues and challenges. *J Process Control.* 2011;21(10):1467–82.
  27. Andersson M, Svensson O, Folestad S, Josefson M, Wahlund KG. NIR spectroscopy on moving solids using a scanning grating spectrometer - impact on multivariate process analysis. *Chemom Intell Lab Syst.* 2005;75(1):1–11.
  28. Green RL, Thurau G, Pixley NC, Mateos A, Reed RA, Higgins JP. In-line monitoring of moisture content in fluid bed dryers using near-IR spectroscopy with consideration of sampling effects on method accuracy. *Anal Chem.* 2005;77(14):4515–22.
  29. Mahdi F, Hassanpour A, Muller F. An investigation on the evolution of granule formation by in-process sampling of a high shear granulator. *Chemical Engineering Research & Design.* 2018;129:403–11.
  30. Avila C. ChemiView V3.4; Available from <http://chemiview.leeds.ac.uk/> Accessed 03 Mar 2020.
  31. Martens H, Naes T. *Multivariate calibration.* Chichester: Wiley; 1989. p. 419.
  32. Jackson JE, Mudholkar GS. Control procedures for residuals associated with principal component analysis. *Technometrics.* 1979;21(3):341–9.
  33. Kemp IC, Sohet Q. Scale-up, optimization, and control of industrial batch fluidized bed dryers using multilevel theoretical models. *Dry Technol.* 2010;28(5):710–22.
  34. Buck AL. New equations for computing vapor-pressure and enhancement factor. *J Appl Meteorol.* 1981;20(12):1527–32.
  35. Boer JH. *The dynamical character of adsorption.* London: Oxford University Press; 1968. p. 240.
  36. Zografi G. States of water associated with solids. *Drug Dev Ind Pharm.* 1988;14(14):1905–26.
  37. Lin L, Quan G, Peng T, Huang Z, Singh V, Lu M, *et al.* Development of fine solid-crystal suspension with enhanced solubility, stability, and aerosolization performance for dry powder inhalation. *Int J Pharm.* 2017;533(1):84–92.
  38. Roja J, Moren S, Lopez A. Assessment of the water sorption properties of several microcrystalline celluloses. *Journal of Pharmaceutical Sciences and Research.* 2011;3(7):1302–9.
  39. Perfetti G, Alphazan T, Wildeboer WJ, Meesters GMH. Thermo-physical characterization of Pharmacoat((R)) 603, Pharmacoat((R)) 615 and Mowiol((R)) 4-98. *J Therm Anal Calorim.* 2012;109(1):203–15.
  40. Faroongsarng D, Peck GE. The Swelling and Water-Uptake of Tablets .3. Moisture Sorption Behavior of Tablet Disintegrants. *Drug Dev Ind Pharm* 1994;20(5):779–798.
  41. Kunii D, Levenspiel O. *Fluidization engineering*, vol. xxvii. 2nd ed. Boston; London: Butterworth-Heinemann; 1991. p. 491.
  42. Chen H, Liu X, Bishop C, Glasser BJ. Fluidized bed drying of a pharmaceutical powder: a parametric investigation of drying of dibasic calcium phosphate. *Dry Technol.* 2017;35(13):1602–18.
  43. Manganaro JL. A quasi-steady state shell and shrinking core approach to the drying of porous particles and an example of parameter identification. *Can J Chem Eng.* 2007;85(3):313–25.
- Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





Contents lists available at ScienceDirect

Talanta

journal homepage: [www.elsevier.com/locate/talanta](http://www.elsevier.com/locate/talanta)

## Acid number, viscosity and end-point detection in a multiphase high temperature polymerisation process using an online miniaturised MEMS Fabry-Pérot interferometer

Claudio Avila<sup>a, \*\*</sup>, Christos Mantzaridis<sup>b</sup>, Joan Ferré<sup>c</sup>, Rodrigo Rocha de Oliveira<sup>d</sup>, Uula Kantojärvi<sup>e</sup>, Anna Rissanen<sup>f</sup>, Poppy Krassa<sup>b</sup>, Anna de Juan<sup>d</sup>, Frans L. Muller<sup>a</sup>, Timothy N. Hunter<sup>a</sup>, Richard A. Bourne<sup>a, \*</sup>

<sup>a</sup> School of Chemical and Process Engineering, University of Leeds, Leeds, LS2 9JT, United Kingdom

<sup>b</sup> Megara Resins, 38th Km New National Rd. Athens-Corinth, Megara, 191 00, Greece

<sup>c</sup> Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira I Virgili, Tarragona, 43007, Spain

<sup>d</sup> Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, Barcelona, 08028, Spain

<sup>e</sup> Spectral Engines Oy, Kutomotie 16, Helsinki, Finland

<sup>f</sup> VTT Technical Research Centre of Finland, Tietotie 3, Espoo, Finland

### ARTICLE INFO

#### Keywords:

Near infrared spectroscopy  
MEMS Fabry-Pérot interferometer  
Online process monitoring  
High temperature polymerisation  
Saturated polyester resin  
Chemometrics

### ABSTRACT

Recent advances in the latest generation of MEMS (micro-electro-mechanical system) Fabry-Pérot interferometers (FPI) for near infrared (NIR) wavelengths has led to the development of ultra-fast and low cost NIR sensors with potential to be used by the process industry. One of these miniaturised sensors operating from 1350 to 1650 nm, was integrated into a software platform to monitor a multiphase solid-gas-liquid process, for the production of saturated polyester resins. Twelve batches were run in a 2 L reactor mimicking industrial conditions (24 h process, with temperatures ranging from 220 to 240 °C), using an immersion NIR transmission probe. Because of the multiphase nature of the reaction, strong interference produced by process disturbances such as temperature variations and the presence of solid particles and bubbles in the online spectra required robust pre-processing algorithms and a good long-term stability of the probe. These allowed partial least squares (PLS) regression models to be built for the key analytical parameters acid number and viscosity. In parallel, spectra were also used to build an end-point detection model based on principal component analysis (PCA) for multivariate statistical process control (MSPC). The novel MEMS-FPI sensor combined with robust chemometric analysis proved to be a suitable and affordable alternative for online process monitoring, contributing to sustainability in the process industry.

### 1. Introduction

The production of saturated polyester resins is a process of global relevance, with large production volumes and a considerable environmental footprint [1]. These are condensation polymers, normally formed in a polycondensation reaction between polycarboxylic acids or their anhydrides and polyalcohols, producing water as a by-product. This is a reversible equilibrium reaction, industrially performed

between 220 and 240 °C, where the formation of polyester is promoted when water and low boiling point products are distilled out [2]. The composition of the polyester resin is critically important in achieving the balance of glass transition temperature, acid number, hydroxyl number and viscosity of the resin that characterize the quality of the product [3]. Commercial saturated polyester resins are manufactured predominantly from a combination of polycarboxylic compounds including isophthalic acid, terephthalic acid, adipic acid, trimellithic acid anhydride and the

\* Corresponding author.

\*\* Corresponding author.

E-mail addresses: [C.R.Avila@leeds.ac.uk](mailto:C.R.Avila@leeds.ac.uk) (C. Avila), [cmantzaridis@gmail.com](mailto:cmantzaridis@gmail.com) (C. Mantzaridis), [joan.ferre@urv.cat](mailto:joan.ferre@urv.cat) (J. Ferré), [rodrigo.rocha@ub.edu](mailto:rodrigo.rocha@ub.edu) (R. Rocha de Oliveira), [uula@spectralengines.com](mailto:uula@spectralengines.com) (U. Kantojärvi), [Anna.Rissanen@vtt.fi](mailto:Anna.Rissanen@vtt.fi) (A. Rissanen), [p.krassa@megararesins.com](mailto:p.krassa@megararesins.com) (P. Krassa), [anna.dejuan@ub.edu](mailto:anna.dejuan@ub.edu) (A. de Juan), [F.L.Muller@leeds.ac.uk](mailto:F.L.Muller@leeds.ac.uk) (F.L. Muller), [T.N.Hunter@leeds.ac.uk](mailto:T.N.Hunter@leeds.ac.uk) (T.N. Hunter), [R.A.Bourne@leeds.ac.uk](mailto:R.A.Bourne@leeds.ac.uk) (R.A. Bourne).

<https://doi.org/10.1016/j.talanta.2020.121735>

Received 24 May 2020; Received in revised form 29 September 2020; Accepted 2 October 2020

Available online 4 November 2020

0039-9140/© 2020 Elsevier B.V. All rights reserved.



polyalcohols ethylene glycol, neopentylglycol, trimethylolpropane and glycerol. The production process required to achieve high molecular weight carboxyl-functional saturated polyester resins is a two stage esterification, in which the first stage involves the preparation of a precondensate by reacting the acids with excess of diols, and a second stage by reacting the remaining diols with additional acids.

For polyester production, chemometric modelling has been used to correlate analytical properties such as acid number [4–8] and hydroxyl number [4,6–8] with offline NIR spectra. Offline analysis satisfies the needs for quality control tests, but it is time and labour intensive. Hence, it is not efficient enough to implement feedback control in an industrial production process. For continuous process monitoring, in situ NIR methods could offer a better approach. However, in situ NIR spectra are greatly affected by the physical and chemical variations found in large-scale reaction systems [9]. For instance, the variation of process variables such as temperature [10], the presence of two-phase interfaces between liquid and solids [11,12], immiscible liquids and gas bubbles [13], the change in optical properties of the material during reaction [8], as well as changes in the NIR instrumentation (e.g. temporal variation of illumination, changes in light transmission due to fiber optics related issues [13]), need to be addressed on a case-by-case basis. As a result, transferring the advantages of offline NIR spectroscopy to real time process monitoring remains a challenge for the polyester industry and for similar applications.

To generate process understanding through online NIR monitoring, chemometric models including partial least squares (PLS) regression are typically used to correlate the key analytical properties with the online measured spectra [8]. Likewise, end-point detection models based on principal component analysis (PCA) for multivariate statistical process control (MSPC) have been used to control the process evolution through sole spectral variations [14–16]. Requirements that must be followed in building these models include the need for calibration data sets to be representative of future process data [17], and that pre-processing steps need to be applied to prevent the negative effects from process disturbances in the quality of the spectral signal [18,19]. When these issues are not addressed, accuracy and robustness of the chemometric models is compromised [20]. Any action directed to improve the quality of the spectra acquired, minimising the effect of disturbing factors on the signal and the models, is highly beneficial [21].

In this context, the quality of the online NIR spectra depends on two main factors: the interactions of the process disturbances with the process interface, and the method or acquisition strategy implemented by the spectrometer selected for the application [22]. Additionally, conventional spectrometers are often installed in safe areas distant from the process vessels, limited by their size, high cost and mechanical stability to obtain the demanded performance. These requirements impact in both the instrumentation installation cost and the quality of the online NIR signal used.

A recent alternative to the use of conventional spectrometers are spectral sensors using miniaturised and low cost MEMS-FPI chips (micro-electro mechanical system – Fabry-Pérot interferometer) developed for NIR wavelengths. MEMS-FPI are miniaturised tuneable optical filters that limit the pass of light in a narrow frequency range by using a set of two facing reflectors separated by an adjustable gap modified with a change in voltage [23]. These micro devices allow the scanning of specific regions of the spectra relevant to the process application, without incorporating moving parts such as those found in conventional FTIR spectrometers; and without diffraction gratings such as those found in dispersive spectrometers. These devices have additional advantages over conventional systems [24]: the size of the MEMS-FPI chip and the detector are considerably reduced, the system is position and vibration insensitive, and the spectral resolution does not suffer from tilting effects. Also, the device is very stable over time since the fabrication from a single wafer, without any additional assembly steps, creates a single solid structure with no wearing parts. Finally, thermal stabilization of the detector is straightforward because only a single-point detector is

used, compared to conventional technologies that normally require linear array detectors [25]. MEMS-FPI sensors have been used for mid infrared (MIR) [26] and lately for NIR [27] applications, with a wide industrial application potential [28–30], although they still require further validation under a variety of laboratory and industrial conditions to understand their limitations and develop their potential further.

This paper investigates the use of a novel MEMS-FPI spectral sensor to monitor the high temperature production of saturated polyester resins. The performance of the NIR device was evaluated under the complex multiphase reaction conditions by using the online spectral information combined with PLS regression models to predict acid number and viscosity, and to identify the process end-point by using MSPC tools. The potential benefits to the process industry in terms of miniaturisation and low cost offered by these sensors were also explored.

## 2. Materials and methods

### 2.1. Reaction system

Twelve experimental batches of the saturated polyester resin were synthesised following a commercial process description at Megara Resins industrial facility in Greece. For the reaction, industrial grade terephthalic acid, isophthalic acid and adipic acid were the dicarboxylic acids used; ethylene glycol, diethylene glycol, neopentyl glycol, trimethylolpropane and glycerol were the polyols used. Butylstannic acid was used as the esterification catalyst. The reagent ratios are kept undisclosed for confidentiality.

A 2 L round flask with external heating and temperature control was used as the reaction vessel, keeping a continuous stirring rate of 200 rev per min. In order to prevent the discoloration due to the oxidation reaction, the reactor was continuously purged with nitrogen. In the first step, the reactant mixture was prepared by adding the fraction rich in diols into the vessel at approximately 80 °C. Once the diols were melted, the fraction rich in acids was added to the vessel under constant agitation. The temperature was then ramped up to 180 °C, where it was held for a 3 h period, then increasing 20 °C every 3 h up to reaching 240 °C, where the first reaction stage proceeds.

A hydroxyl-terminated polyester was formed by reacting the dibasic acids, polyols and optional branching agents like trimethylolpropane at a temperature in the range of 160–240 °C in the presence of esterification catalyst and colour stabilizer to form a hydroxyl-terminated prepolymer. At this stage, the water of esterification was collected. When the acid number of the resin fell below the value determined by the specifications, the first stage of the reaction was completed, providing a hydroxyl terminal polyester. In the second stage, the hydroxyl groups were end-capped with carboxylic acids or their anhydrides to form a carboxylated polyester. The amount of end-capping agent used was determined by the hydroxyl number of the polyester. The end-capping agent was added to the prepolymer and the esterification was continued until the desired acid number was obtained. Vacuum was applied towards the end of the reaction in order to eliminate volatile products and thus shift the equilibrium towards the formation of the polymer. Finally, after a period determined by the analytical indicators, the temperature was lowered to 200 °C to add product enhancing additives and finish the production process.

### 2.2. Key analytical indicators

The analytical indicators selected to follow the progress of the reaction were acid number (AN) and viscosity ( $\mu$ ). Acid number was measured by manual acid-base titration following the ASTM (American Society for Testing and Materials) method D 1613-03 and it was reported as milligrams of potassium hydroxide (KOH) per gram of sample. Viscosity (high shear viscosity) was measured using a cone/plate viscometer model CAP 2000 from Brookfield (USA), operating at 200 °C

following the procedure described in the ASTM method D-4287-00.

The targeted ranges for the first reaction stage were AN 8–12 (mg KOH g<sup>-1</sup>) and  $\mu$  10–14 (P or g cm<sup>-1</sup> s<sup>-1</sup>); and for the second reaction stage AN 45–63 (mg KOH g<sup>-1</sup>) and  $\mu$  25–45 P. In case the measurements were out of specifications during any of the stages, additional reactants were added to reach the desired conditions. During the analytical sampling, online NIR spectra were collected simultaneously from the reaction vessel.

### 2.3. MEMS-FPI NIR sensor and data acquisition

A novel spectral sensor model N-Series 1.7 by Spectral Engines (Finland) was used for the acquisition of the NIR spectra from 1350 nm to 1650 nm. A diagram of the sensor is shown in Fig. 1. The sensor has a single element extended InGaAs detector, with a tuneable MEMS-FPI filter acting as the spectral element. The sensor had an integrated light source model LS-PRO equipped with a miniature tungsten vacuum lamp as the illumination source. Additional details about the scanning mechanism used by the sensor can be found in the Appendix section.

The spectral sensor was connected to a stainless steel NIR immersion probe (transmission mode, 5 mm optical pathlength) model Excalibur 20 by Hellma Analytics (Germany). The probe has two 2 m fibre optic cables, connecting one end to the light source and the other to the spectral sensor. The probe was designed to operate from ambient temperature up to 260 °C, and it was immersed with the transmission gap positioned perpendicular to the centre of the vessel (facing the flow created by the agitator) during the entire reaction time, without fouling or blocking of the sample gap for any of the batches performed.

For all experiments, the energy output for the lamp was set to 25% of the maximum level. This value was selected for the specific polymerisation system investigated, since higher values saturated the maximum input of the sensor and lower values were attenuated by the sample. The sensor integration time was set to 0.1 ms and the wavelength step set to 1 nm (301 points obtained from the operational sensor range).

The software used to operate and record NIR data from the spectral sensor was an in-house application developed by the University of Leeds using LabVIEW 2015 (ChemiView V 3.4 [31]). Process temperature readings were acquired using a TC-08 temperature reader from Pico Technologies (USA), using K-type immersion temperature probes from Omega (UK). For batches 1 to 10, a single NIR spectrum was obtained every 5 s as the average of 50 sensor readings (internal FPI scanning sequence implemented by the sensor, delivering 1 output spectra every 5 s). For batches 11 and 12, each NIR spectrum was obtained every 0.83 s from a single FPI scanning sequence (minimum possible). The information for all batches is included in Table 1, with batches labelled

according to the sequence of acquisition.

### 2.4. Process data treatment

Multivariate calibration models using PLS regression to determine AN and  $\mu$  parameters and PCA-based MSPC models for end-point detection were created from the online NIR data. In both cases, modelling and validation were carried out with in-house routines programmed in Matlab R2017a (Mathworks, USA) and PLS\_Toolbox 8.2.1 (Eigenvector Research, USA) running under Matlab.

For each batch, the influence of process disturbances in the quality of the NIR signal was considerable (discussed within results). In order to attenuate these effects, a pre-processing step was introduced. In this, 13 output spectra (as delivered by the sensor) were averaged into a single spectra, emulating the averaging that can be instrumentally obtained by increasing the number of FPI scans. This action reduced the number of spectra and the noise in the signal, at the expense of introducing a small delay time of 65 s per useable spectrum. Afterwards, the resulting averaged signal was transformed to absorbance. Since artifacts could not be completely removed, a moving average filter was applied to the absorbance spectra in the time dimension. Each spectrum was replaced by the average of itself and the  $N = 30$  previous spectra, where  $N$  was chosen as a compromise between small prediction delay and good quality. This means 30 absorbance spectra are required to build-up the moving average before the data can be used for monitoring purposes, which occurs at the beginning of the process stages where predictions are not required (latent phase, discussed in results). Finally, a 1st order Savitzky-Golay derivative [32] followed by column mean-centring was applied to correct baseline variations before submitting the resulting dataset to the PLS algorithm or to the end-point detection model. Under these conditions, the models deliver 1 prediction every 65 s.

- (a) **PLS regression models:** The polymerisation process has two very distinct reaction stages, the first stage to form the hydroxyl-terminated prepolymer, and the second reaction stage to form the final carboxylated polyester. Therefore it was not possible to develop a single PLS for each property (AN and  $\mu$ ) that could provide predictions accurate enough for the entire process. The solution was to develop a PLS model for each property and for each stage, resulting in four multivariate calibration models relating the calibration spectra to AN and  $\mu$  using PLS regression [33]. The averaged absorbance spectra corresponding to the times when samples were collected during the reaction (known acid number and viscosity) were placed as the rows of data matrix  $X$  (samples  $\times$  wavelengths). The reference values of acid number

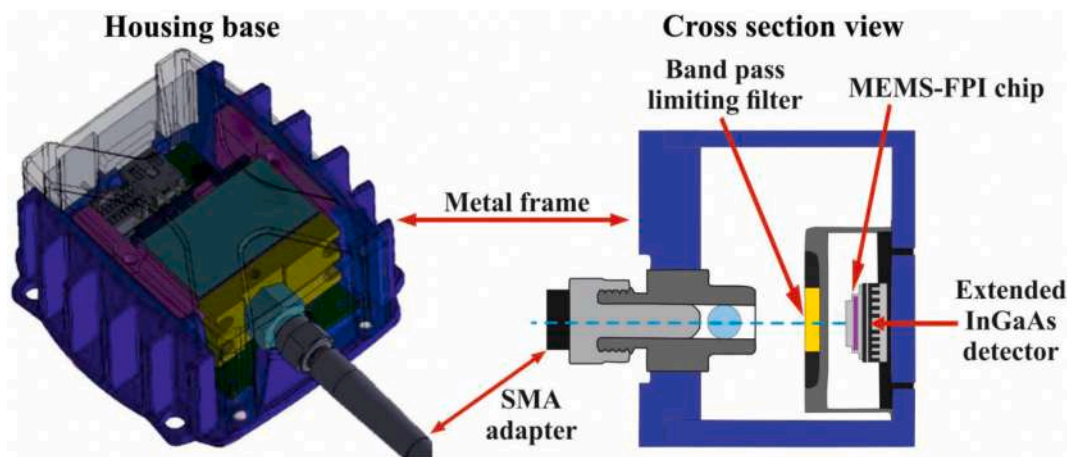


Fig. 1. Diagram of the NIR spectral sensor base (a) with the MEMS-FPI tuneable filter (b). The assembled sensor weight 125 g, with the metal chassis measuring 58 mm length by 57 mm width by 27 mm high.

**Table 1**  
Summary of analytical parameters measured, total reaction time and NIR spectra acquired.

Batch	AN <sup>a,b,c</sup>	$\mu$ <sup>a,d</sup>	AN <sup>b,c</sup>	$\mu$ <sup>b,d</sup>	Final process outcome	Reaction time	NIR scans per single spectrum	Number of spectra
1 <sup>e</sup>	8.4	7.0	–	–	Out of specification	21 h, 40 min	50 averaged in 5s	16,427
2	6.0	16.0	49.9	36.4	Within specification	21 h, 10 min	50 averaged in 5s	16,753
3	8.4	13.6	48.0	50.3	Within specification	26 h, 25 min	50 averaged in 5s	20,052
4	7.9	14.5	55.6	36.7	Within specification	21 h, 25 min	50 averaged in 5s	16,456
5	7.8	14.1	54.0	38.4	Within specification	22 h, 30 min	50 averaged in 5s	17,230
6	8.4	9.6	51.0	29.9	Within specification	22 h, 20 min	50 averaged in 5s	16,199
7	8.9	9.5	54.0	42.5	Within specification	22 h, 55 min	50 averaged in 5s	16,621
8	9.5	12.3	54.0	39.1	Within specification	22 h, 5 min	50 averaged in 5s	15,940
9	9.0	12.1	53.3	41.0	Within specification	24 h, 50 min	50 averaged in 5s	17,970
10 <sup>f</sup>	8.7	10.1	56.0	53.7	Out of specification	24 h, 15 min	50 averaged in 5s	17,149
11	8.3	10.3	55.0	31.7	Within specification	24 h, 45 min	1 scan in 0.83s	104,987
12	7.6	10.8	51.0	36.2	Within specification	21 h, 05 min	1 scan in 0.83s	92,204

<sup>a</sup> At the end of the first reaction stage.

<sup>b</sup> At the end of the second reaction stage.

<sup>c</sup> AN in mg KOH g<sup>-1</sup>.

<sup>d</sup>  $\mu$  in Poise.

<sup>e</sup> Batch 1 ended out of specification after first reaction stage.

<sup>f</sup> Batch 10 ended out of specification after the second reaction stage.

and viscosity made up column vectors  $\mathbf{y}_{av}$  (samples  $\times$  1) and  $\mathbf{y}_{vi}$  (samples  $\times$  1), respectively, and a separate model was completed to relate each of these properties to the NIR information. Because there are two clear different stages in the process,  $\mathbf{X}$ ,  $\mathbf{y}_{av}$  and  $\mathbf{y}_{vi}$  were split in two sets, one for the first stage of the reaction ( $\mathbf{X}_1$ ,  $\mathbf{y}_{av,1}$  and  $\mathbf{y}_{vi,1}$ ), and another for the second stage of the reaction ( $\mathbf{X}_2$ ,  $\mathbf{y}_{av,2}$  and  $\mathbf{y}_{vi,2}$ ). Pre-processed NIR spectra from batches 1 to 5 were used to generate the training set for the PLS models, with 7 additional batches used as external validation set.

(b) **MSPC models:** To build MSPC models, a data set formed by NIR spectra collected at the end of each stage from normal operating condition (NOC) batches were used. All the end-point spectra were organized in a data matrix  $\mathbf{X}_{NOC}$  (number of end-point NIR spectra  $\times$  wavelengths). A PCA model was built with these data to set the statistical boundaries of the experimental domain (space) of end-point NIR spectra [34,35]:

$$\mathbf{X}_{NOC} = \mathbf{T}_{NOC}\mathbf{P}_{NOC}^T + \mathbf{E}_{NOC}$$

where  $\mathbf{T}_{NOC}$  is the scores matrix of all end-point spectra (spanning the valid experimental domain for on-specification measurements in the space of principal components) and  $\mathbf{P}_{NOC}^T$  is the loadings matrix (which is the link between scores and original NIR spectra).  $\mathbf{E}_{NOC}$  describes the residual variation unexplained by the PCA model. The number of components used in the PCA model was established by cross-validation [36]. From the PCA model, a Q-statistic control chart  $Q_{stat}$  was built, the boundary of which was based on the residual part of the process variation not explained by the PCA model. The control limit for the  $Q_{stat}$  chart,  $Q_{lim}$ , was set according to the Jackson and Mudholkar equation [37], in which  $Q_{lim}$  represents an approximate  $Q_{stat}$  critical value with significant level  $\alpha = 0.05$ . For any new (pre-processed) spectrum acquired in an online monitored batch,  $\mathbf{x}_{i,new}$ , the PCA model obtained above was used as follows:

$$\mathbf{t}_{i,new} = \mathbf{x}_{i,new}\mathbf{P}_{NOC}$$

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

$$\mathbf{e}_{i,new} = \mathbf{x}_{i,new} - \mathbf{t}_{i,new}\mathbf{P}_{NOC}^T$$

and the related  $Q_{stat}$  value as:

$$Q_{stat} = \mathbf{e}_{i,new}^T \mathbf{e}_{i,new}$$

When the shape of the new spectrum is sufficiently similar to the on-specification end-point spectra used to build the PCA model, the residual  $\mathbf{e}_{i,new}$  will be small and the related  $Q_{stat}$  value will appear below the chart control limit,  $Q_{lim}$ . When a spectrum is far from the end-point, the

spectral shape will be clearly different from that of end-point spectra and will show up above the chart control limit. The point in time where the  $Q_{stat}$  value of the on-line monitored NIR spectra goes below the control chart limit will indicate the end-point of the process. Off-specification batches will consistently show  $Q_{stat}$  values above the chart control limit.

For this study, data from on-specification batches 2 to 5 were used to extract NIR data to build an initial end-point detection model for each reaction stage. Subsequently, data from on-specification batches 2 to 9 were used to build an updated version of the same models. The remaining batches out of their modelling sets were used for external validation. The pre-processed NIR spectra used were collected during the last 15 min before the end of each reaction stage for each batch. Two matrices with 60 spectra (4 batches  $\times$  15 spectra) were generated with the selected end-point NIR spectra to build two separate end-point MSPC models for stages 1 and 2 of the process. Spectral pre-processing was performed as explained above.

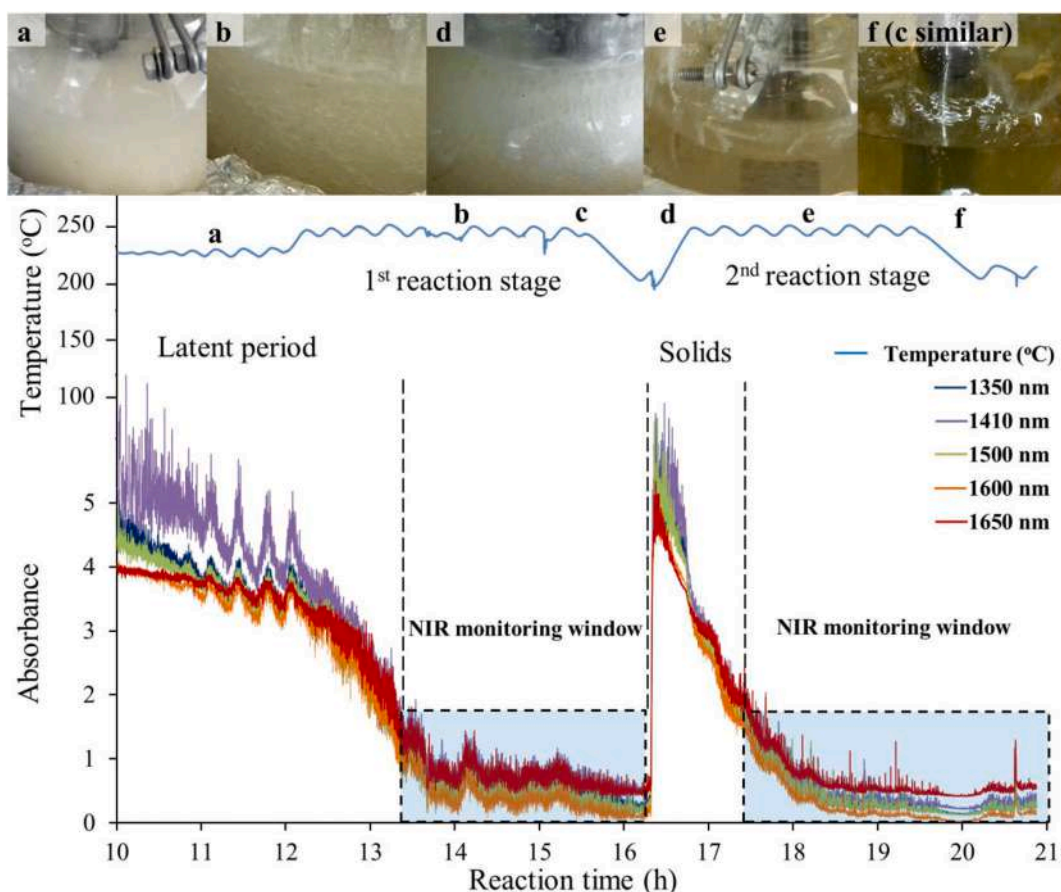
### 3. Results and discussion

#### 3.1. Saturated polyester resin production process and online NIR sampling

The production of saturated polyester resins progressed as a multi-phase reaction, in which gas bubbles and suspension solids considerably affected the spectral measurements during all the process stages. Fig. 2 (top) shows images of the different reactions periods relating the presence of bubbles, solid particles, while Fig. 2 (middle and bottom) presents related fluctuations in temperature and the NIR signal by these process disturbances in the time domain.

For instance, at the beginning of the process, the carboxylic acids were solids in suspension forming the liquid polymer as the reaction progresses. The solids totally attenuated the NIR signal over the initial 10–12 h of the process (also known as latent phase; Fig. 2, a), which gradually changed as the carboxylic acids reacted and the solution became transparent to NIR light at the beginning of the first reaction stage (Fig. 2, from a to b), clearing further towards the end of this (Fig. 2, c). During the latent phase, light absorption and scattering produced by the particles were the predominant effect. This phenomenon occurred again when the chemicals for the second reaction stage were added (a large fraction of carboxylic acids in solid form), and also when performing small corrections (adding small quantities of the same solids) required to drive the analytical properties towards the desired values (Fig. 2, d).

Simultaneously, as the reaction progressed, gas bubbles were generated due to the formation of water and low boiling point products resulting from the transesterification reactions, and also due to the



**Fig. 2.** Images showing typical process conditions (top): a) bubbles and solids in suspension during the latent phase; b) bubbles in suspension in the middle of first stage; c) homogeneous solution at the end of the first stage; d) bubbles and solids in suspension after adding second stage chemicals; e) bubbles in suspension in the middle of second stage; f) final product. The influence of temperature fluctuations (middle, blue line) mirrored by the absorbance NIR spectra for 5 selected wavelengths (bottom; for batch 5, similar to all batches) as a function of time, for the final 11 h of the process. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

nitrogen stream passing through the reaction mixture. These bubbles tended to remain in the system for extended periods of time due to the high viscosity of the mixture, which dissipated slowly when reaching the surface of the vessel or forced to leave when a vacuum was applied to the system. The last action also contributed to drive the key analytical properties towards the desired values. Bubbles scattered the NIR light, but still allowed a useable signal to reach the detector. Bubbles appeared at the intermediate phases of each reaction stage, when the solids had completely reacted. Towards the end of each reaction stage, bubbles also gradually disappeared, with the sample becoming fully transparent (Fig. 2, e).

Finally, the temperature of the reactor also fluctuated around the set point of the heating control as shown by the temperature readings and mirrored by the NIR spectra, particularly noticeable during the latent phase (Fig. 2, blue line). Fluctuations were due to the limitations of the heating element control. When none of these phenomena disturbed NIR acquisition, the signal had a stable amplitude and was very repeatable between scans, especially at the end of the reaction process (Fig. 2, f).

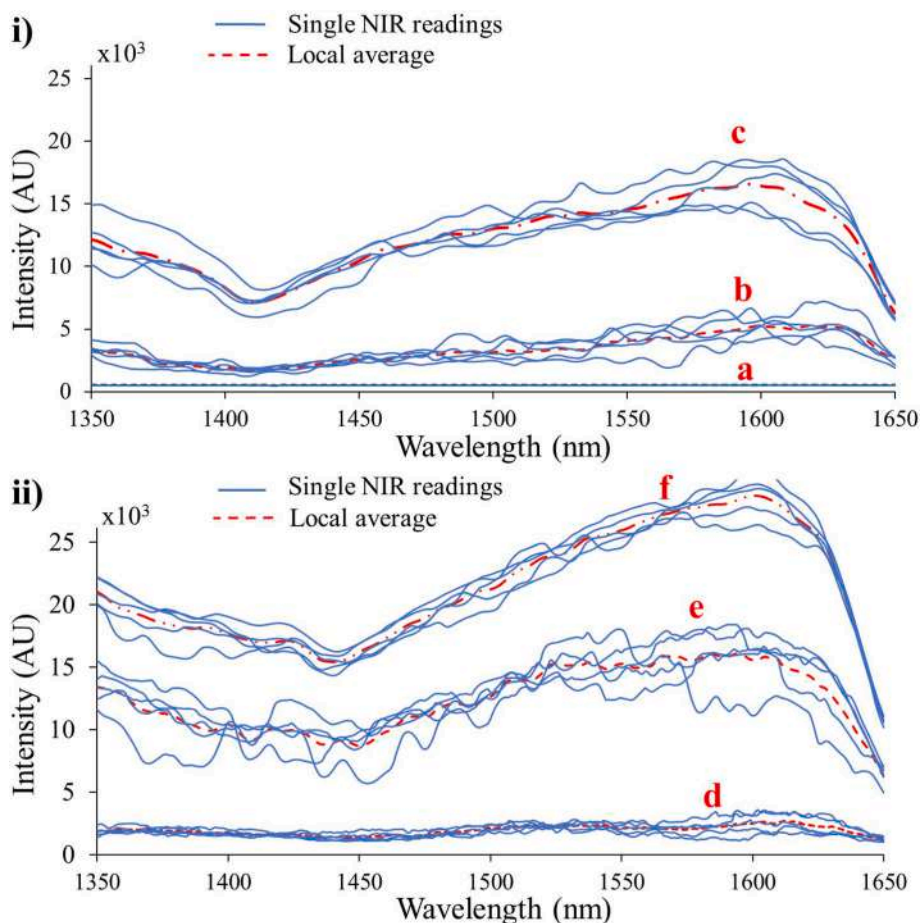
Compared to previous reports using offline NIR spectra to correlate key analytical properties [6], the fluctuations produced by process disturbances in the NIR spectra were the main obstacle to perform online monitoring. The attenuation effect produced by solid particles was the main restricting factor that limited the time window to obtain useful NIR measurements in transmission mode. On average, the complete reaction process takes approximately 25 h per batch, from which the first 12–14 h corresponded to the latent phase (non-transparent), with periods of approximately 5 h for each reaction stage (transparent). Under these

conditions, the time frame for measuring useful NIR spectra that could be correlated to the key analytical properties was 3–6 h for each stage. Fig. 2 illustrates the NIR monitoring window observed for batch number 5.

Fig. 3 shows groups of five consecutive NIR scans (raw intensity spectra, thin blue lines) and their corresponding average (red dashed lines), obtained for specific periods of the first (i) and second (ii) reaction stages during the NIR monitoring window. These groups correspond to similar time periods for the specific process conditions shown in Fig. 2. As observed from Fig. 3, the intensity of the signal tends to increase as the reaction progress, with the exception of the transition period between stages one and two, when a large fraction of solids was added causing the signal to drop. Regarding the active NIR groups for the polyester system relative to the spectral range of the NIR sensor used, the wavelength range 1400–1500 nm relating to first overtone of –OH vibration was the most important for prediction. It also allowed differentiating clearly between reaction stage 1 and stage 2 of the process. Although wavelengths longer than 1500 nm are less important for prediction, they allowed a better outlier detection and, therefore, the full wavelength range covered by the NIR spectral sensor was found useful for modelling purposes (an absorbance plot for the same spectra and time periods as shown in Fig. 3 is available in Appendix).

### 3.2. Prediction of key analytical properties using PLS and MSPC models

For the 12 batches performed, the analytical indicators measured at the end of each stage and the final process outcome are listed in Table 1.



**Fig. 3.** Example NIR spectra shown for six different process periods, displaying the disturbances generated by bubbles and solids particles in suspension. i) First reaction stage: a) bubbles and solids in suspension during the latent phase; b) bubbles in suspension in the middle of first stage; c) homogeneous solution at the end of the first stage. ii) Second reaction stage: d) bubbles and solids in suspension after adding second stage chemicals; e) bubbles in suspension in the middle of second stage; f) final product. Data from batch number 5, and similar to all other batches. Groups of five consecutive raw spectra (thin blue lines), and red dashed lines corresponding to the average spectrum obtained for each group. Absorbance plot for the same spectra is available in Appendix. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Two out of twelve batches ended up out of specification in relation to the commercial product, after a reasonable number of attempts to correct the direction of the process towards the desired analytical control parameters. The time difference observed between batches was due to the number of chemical adjustments carried out for each case. After each chemical correction, it was necessary to wait for thermal stabilization of the system and the reaction of the solids in suspension before obtaining the next analytical measurement.

Fig. 4 compares the acid number determined offline (circles) and the continuous prediction generated from the online NIR spectra after applying the PLS models for six batches (batches 3 to 5 used for calibration, and batches 7 to 9 used for validation were included in this figure. Similar plots for all twelve batches can be found in Appendix). The analytical measurements and the predictions shown in these figures were obtained during the NIR monitoring window, in which the time gap between reaction stages corresponded to the addition of the second stage chemicals (solids).

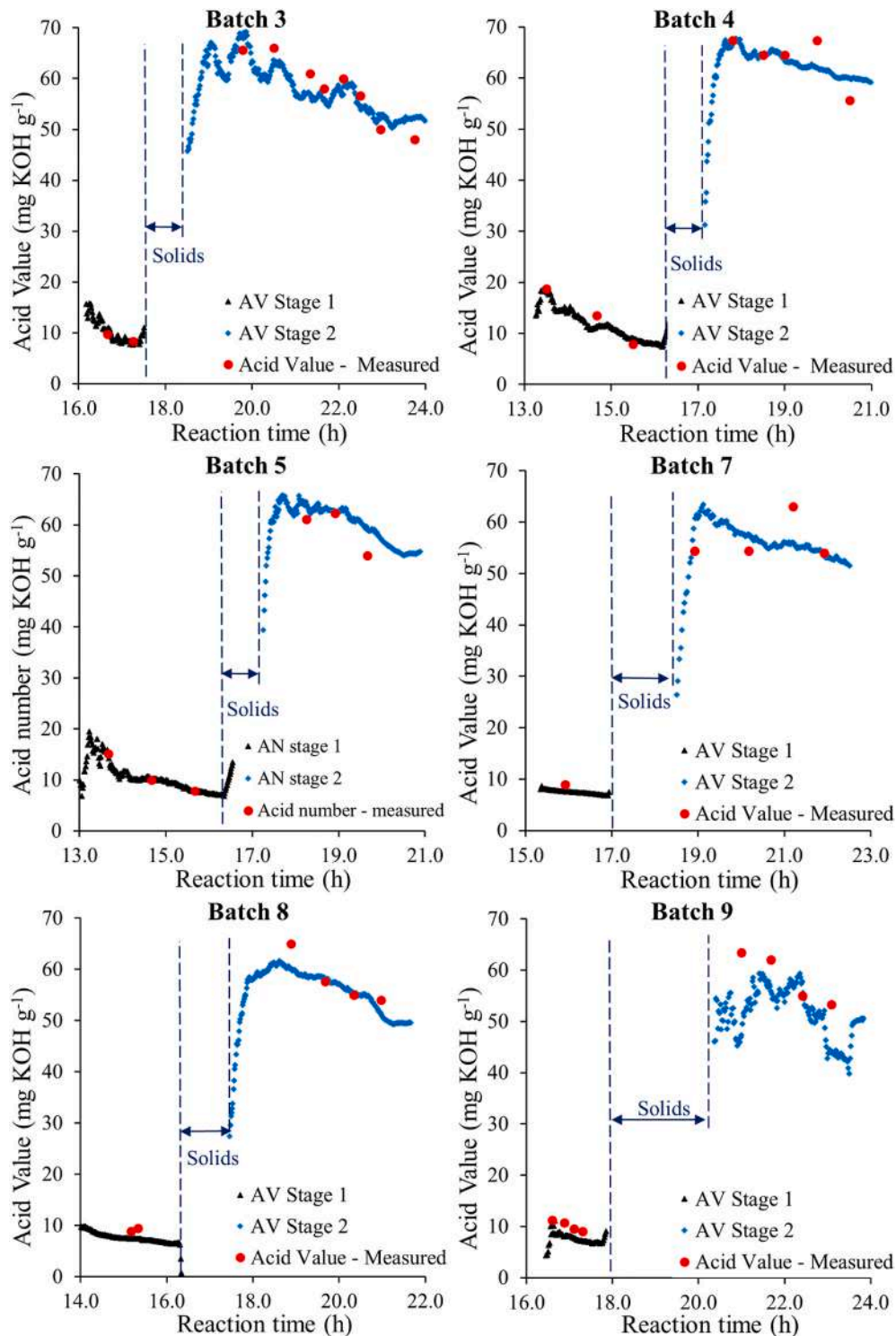
Continuous predictions obtained from the PLS models against the offline viscosity measurements are shown in Fig. 5, illustrating the same batches used for Fig. 4. For both process stages, viscosity values always increase due to the increasing length of the polymer branches formed and the PLS model predictions followed this trend.

For both acid number and viscosity predictions, sharp variations between consecutive spectra due to bubbles and solids in suspension were the most important data issue to be solved when building and implementing PLS models. These variations affected the transmission of light both in the wavelength dimension and in the time dimension randomly e.g. one spectrum may suffer artifacts at certain wavelengths, while the next was affected at different wavelengths (as shown in Fig. 3). Normally, the referential analytical properties vary slowly during the

reaction, except when adding chemical corrections to the system or when changing operational parameters such as the flow of inert gas. Under normal conditions, it is expected that the model predictions should also evolve slowly, although in this case process interferences still created fluctuations that could not be completely attenuated. Spectrum averaging compensated these undesired effects to a large extent, but could not completely remove them. The spectral moving average over time improved the stability of the predictions, and the remaining fluctuations were considered to be acceptable, taking into account the complexity of the data, and followed the evolution of the process satisfactorily.

Regarding to the accuracy of the predictions obtained, Fig. 6 compares the acid number and viscosity measured for the 12 batches against the predictions obtained from the PLS models. From these figures, it is evident that both key parameters differed considerably for the second reaction stage relative to those predicted for the first stage. Although the process fluctuations observed in both cases were very similar, changes in the first reaction stage were slower and observed at the end of longer time period (latent phase + first reaction stage). Conversely, for the second reaction stage, changes were more vigorous and produced in a shorter time interval, which led to slight increases to the variations on the NIR spectra and resulting predictions.

Results generally indicated that acid number predictions were more precise for the first stage than those obtained for the second stage. This difference can be explained by the chemistry of the system, which has smaller changes during the first stage, as it reacts under an excess of diols, with acid number ranging from 5 to 20 mg KOH g<sup>-1</sup>. For the second reaction stage, the end groups contributing to the acid number were targeted, with a variation fluctuating between 50 and 70 mg KOH g<sup>-1</sup>, almost an order of magnitude higher compared to the first stage

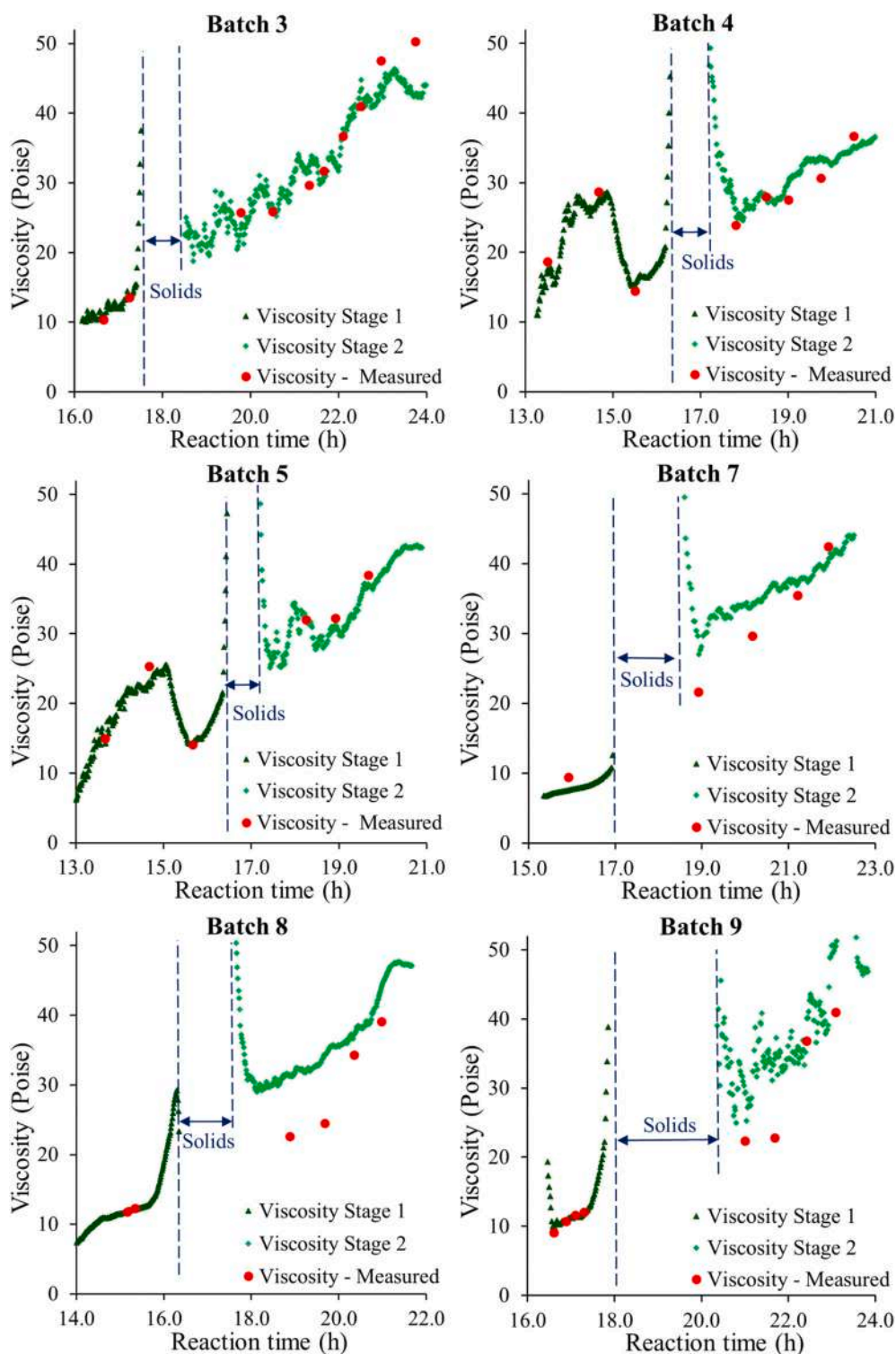


**Fig. 4.** Experimental acid number obtained (red dots) compared to continuous PLS model predictions based on NIR measurements, for the first and second reaction stages. Batches 3, 4 and 5 used for model making; batches 7, 8 and 9 used as external validation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

range. Finally, it is noted that viscosity predictions were more precise and accurate relative to those obtained for the acid number. This difference may simply be due to the higher repeatability of the analytical measurements obtained using the cone viscometer, compared to reference acid number obtained by manual titration that had greater higher variability.

Models for acid number and viscosity were developed with the data obtained from batches 1 to 5 (from February 2017), and predictions for

batches 6 to 12 considered new data (collected in September 2017). As a consequence, there was some increase in prediction variability for batches 6 to 12. Some of the slight reduction in predictive performance may have been due to some introduced systematic bias, because the system had to be reinstalled in Megara after a six month period. Even though the optical components e.g. fibre optic cables and sensors were the same, the system setup was not absolutely be identical e.g. fibre bending radius and ambient temperature was not exactly the same.



**Fig. 5.** Experimental viscosity values obtained (red dots) compared to continuous PLS model predictions based on NIR measurements, for the first and second reaction stages. Batches 3, 4 and 5 used for model making; batches 7, 8 and 9 used as external validation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

However, even accounting these differences, the model prediction was within an acceptable range e.g. within the intrinsic error of the wet chemistry analysis, and highlighted the real potential to use the NIR system for process monitoring.

Additionally, information from the PCA models obtained directly from the sole NIR spectra (without using calibration samples) provided another perspective to evaluate the use of the MEMS-FPI sensors. Fig. 7

shows the end-point detection MSPC model predictions obtained for all the batches during the NIR monitoring window, using an initial model created with batches 2 to 5 (black symbols). For a better visualization of the control chart and the related limit, reduced Q-statistics ( $Q_{red}$ ), expressed as,  $Q_{red} = Q_{stat}/Q_{lim}$ , were used. In this way, the limit in all  $Q_{red}$  charts is equal to 1. An initial qualitative analysis from the profiles suggests a clear decreasing pattern of the  $Q_{stat}$  values as the process

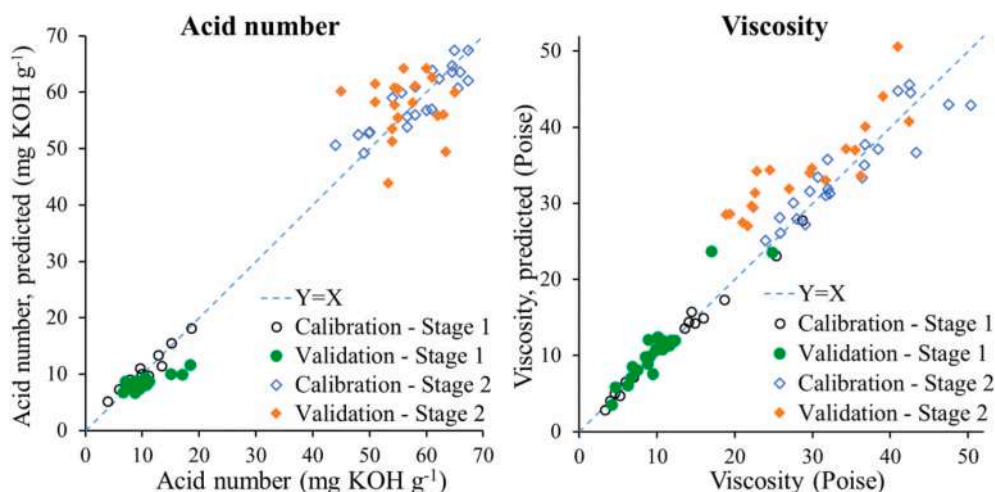


Fig. 6. Acid number and viscosity compared against NIR model predictions using calibration and validation batches for the two reaction stages.

progresses towards completion. Although, the overall final end-point values obtained could be more precise, given the complete experimental set, initial model performance was acceptable, considering the small number of available batches used to build the PCA model.

However, in order to improve the definition of the process end-point, the PCA-based MSPC models were updated to include a larger number of batches (2–9). Batches 10 to 12 were not included in the updated model and used for external model validation. Predictions using the updated model are also shown in Fig. 7 (red dots). Analysing the validation batch 10, we can observe that its second stage did not reach the end-point control limit, which agrees with experimental observation reported in Table 1, where this batch was considered as out of product specification. Although on specification batches 11 and 12 did not cross the end-point control limit for long time, they showed trend towards it, which indicates that these batches could be accepted according to these observations. The results suggest that a larger number of batches will improve the repeatability and robustness of the control models implemented, and that the sole online NIR information obtained from the sensor was sensitive enough to detect the process end-point.

The effect of the process disturbances was also observed for the PCA-based models, although its influence in the identification of the end-point reached for each stage was limited. This is explained by the quality of the selected NIR spectra used to build the end-point detection model, which correspond to the last 15 min of each stage. This particular time interval of the process had two key distinctive differences; firstly, the NIR spectra collected have a higher optical transmittance since the presence of bubbles and solids present was minimum at the end of each stage (Fig. 2, c and f). Secondly, there was a clear difference in the shape of the absorbance spectra at the ending period compared to the initial reaction interval, which produced more intense NIR profiles with stronger peak association (Fig. 3 illustrated this).

The use of a large number of averages to minimise the influence of process disturbances in the NIR spectra had a small impact on the response time of the MSPC model predictions. However, it was not great enough to hide the fluctuations produced by adding corrective chemicals to the reaction vessel (emphasized in Fig. 7 for batches 1 and 10, although this action was performed for most of the batches) to drive the key analytical parameters to their control values. Since the anticipated outcome for this model was a single parameter to identify the process end-point, the implementation was simpler than predicting the analytical properties over short time intervals and required only NIR spectra for generating the training set, without any additional experimental calibration.

Finally, the results obtained from the PLS prediction of viscosity and acid number can be used together with the MSPC control chart to

provide additional supporting information to the end-user. Although using the PLS models as an alternative to the traditional offline analytical analysis still need to be further demonstrated, the results obtained show clearly the NIR sensor performance, even when challenged by severe process fluctuations encountered in the pilot scale process. Under these conditions, predicted viscosity and acid number were within the acceptable limits required for monitoring the synthesis of saturated polyester resins. A reduction in the variability observed (Fig. 6) between calibration and validation batches could be achieved by increasing the number of calibration batches under a permanent installation of the NIR monitoring system.

In addition, the miniaturised size was a distinctive characteristic of the MEMS-FPI sensor, which enabled its installation attached to the reaction vessel, minimising the use of fibre optics cables for transmitting the NIR light. Instead, a standard electrical signal was transmitted from the sensor to the computer, reducing the installation and maintenance costs considerably. Another factor to consider was the stability observed for the MEMS-FPI sensor during the experimental trials, allowing to maintain the calibration for the PLS-based and end-point models. In this study, the whole system was dismantled and reassembled between the two experimental campaigns and, although updated models offered better results, predictions based on the models initially built were still acceptable for the 7 new batches.

Access to affordable process monitoring and control technologies for small and medium enterprises (SMEs) has been identified as a contributing factor to improve process sustainability [38]. For the synthesis of polyester resins (or similar challenging reactions), real-time access to the key process indicators can minimise the number of manual sampling points collected from the high temperature reaction vessels, helping to reduce the risks associated to a minimum. Also, this low-cost information can help to improve batch-to-batch consistency (e.g. observing the development of detrimental disturbances in real-time, and implementing control actions faster than using time delayed off-line data), which reduces the loss of materials and equipment due to batch failure. Finally, access to online monitoring tools can help SMEs to implement more advanced process optimisation strategies, saving cycle time by reducing the number of off-line controls, and bringing further reductions in material consumption and energy savings.

The use of this new generation of MEMS-FPI NIR sensors appears to be a suitable alternative to traditional spectroscopy systems, and particularly adapted to harsh industrial environments such as the production of saturated polyester resins.



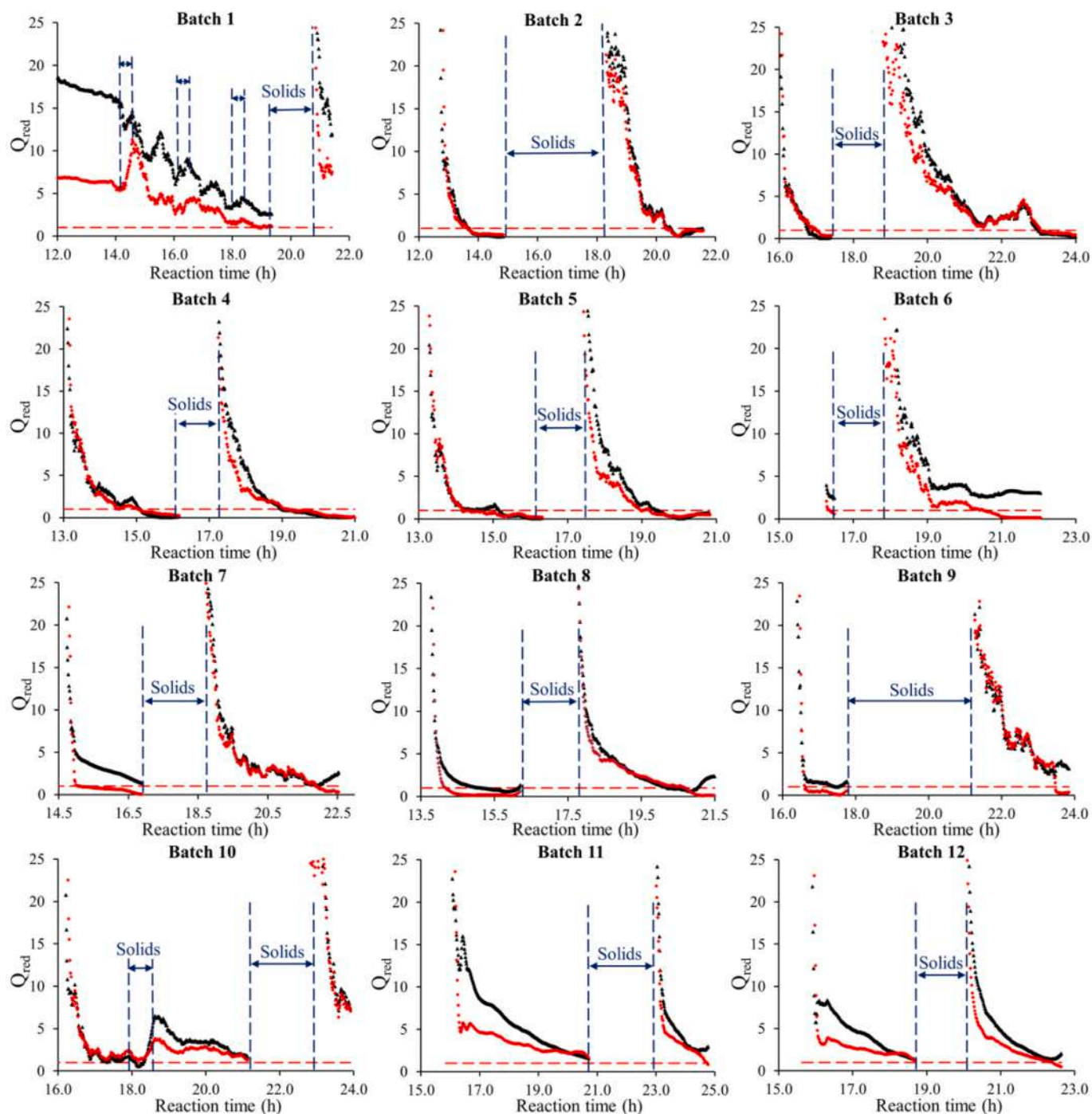


Fig. 7. PCA-based end-point detection MSPC  $Q_{red}$  charts predictions for all batches. Black dots indicate  $Q_{red}$  predictions from model developed using on-specification batches 2 to 5; red dots indicate  $Q_{red}$  predictions from updated model developed using on-specification batches 2 to 9; discontinuous red line indicates the end-point control limit. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

#### 4. Conclusion

A new MEMS-FPI NIR sensing technology combined with suitable chemometric data processing was used for effective monitoring of multiphase production of saturated polyester resins. This process presented several challenges, which are often encountered in similar industrial applications, including variations between spectra, due to the presence of bubbles and solids particles in suspension, and temperature fluctuations. These process disturbances affected the transmission of light both in the wavelength and in the time domains, and also limited the time window to observe the reaction in NIR transmission mode. These issues

where addressed by extensive pre-processing and allowed satisfactory implementations of PLS and PCA-based end-point detection models. In addition, the stability of the optical system over a long time period, achieved by the single frame MEMS-FPI chip architecture and integrated light source, helped to generate a high quality and robust NIR signal. Hence, the combination of the notable optical properties of the sensor combined with chemometric tools to address process-related signal distortions, provided excellent results for monitoring of the key analytical properties (acid number and viscosity) as well as end-point control. This new generation of NIR sensors presented a number of advantages over traditional spectral systems, such as miniaturisation, low

cost and stability, providing an affordable alternative to improve process performance, reduce costs and contribute to sustainability in the process industry.

### Credit author statement

**Claudio Avila:** Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Software. **Christos Mantzaridis:** Investigation, Resources, Writing - Review & Editing. **Joan Ferré:** Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Software, Writing - Review & Editing; Funding acquisition. **Rodrigo Rocha de Oliveira:** Methodology, Formal analysis, Conceptualization, Investigation, Visualization, Software, Writing - Review & Editing. **Uula Kantojärvi:** Resources. **Anna Rissanen:** Resources. **Poppy Krassa:** Resources, Writing - Review & Editing. **Anna de Juan:** Conceptualization, Writing - Review & Editing, Supervision; Funding acquisition. **Frans L. Muller:** Conceptualization, Writing - Review & Editing, Supervision; Funding acquisition. **Timothy N. Hunter:** Conceptualization, Writing - Review & Editing, Supervision; Funding acquisition. **Richard A. Bourne:** Conceptualization, Writing - Review & Editing, Supervision; Funding acquisition.

### 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.

### Acknowledgements

This work was supported by the ProPAT Project (European Union, Horizon 2020 Research and Innovation Programme. Grant agreement 637232). The authors thank Megara Resins for facilitating their facilities for this study.

### References

- [1] Saturated polyester resin market worth \$4,436 million by 2019, Focus Powder Coating. 2015 (7) (2015), [https://doi.org/10.1016/s1364-5439\(15\)30025-3](https://doi.org/10.1016/s1364-5439(15)30025-3).
- [2] M.P. Stevens, *Polymer Chemistry: an Introduction*, second ed., Oxford University Press, New York ; Oxford, 1990.
- [3] B. Parkyn, *Chemistry of polyester resins*, Composites 3 (1972) 29–33, [https://doi.org/10.1016/0010-4361\(72\)90468-5](https://doi.org/10.1016/0010-4361(72)90468-5).
- [4] E. Marengo, M. Bobba, E. Robotti, M. Lenti, Hydroxyl and acid number prediction in polyester resins by near infrared spectroscopy and artificial neural networks, *Anal. Chim. Acta* 511 (2004) 313–322, <https://doi.org/10.1016/j.aca.2004.01.053>.
- [5] S.R.K. Chalasani, S. Dewasthale, E. Hablot, X.K. Shi, D. Graiver, R. Narayan, A spectroscopic method for hydroxyl value determination of polyols, *J. Am. Oil Chem. Soc.* 90 (2013) 1787–1793, <https://doi.org/10.1007/s11746-013-2334-9>.
- [6] M. Blanco, J. Cruz, M. Armengol, Control production of polyester resins by NIR spectroscopy, *Microchem. J.* 90 (2008) 118–123, <https://doi.org/10.1016/j.microc.2008.04.004>.
- [7] M. Blanco, V. Villaescusa, Use of NIR spectroscopy in the production of modified industrial resins, *Talanta* 71 (2007) 1333–1338, <https://doi.org/10.1016/j.talanta.2006.07.028>.
- [8] R.A. Heikka, K.T. Immonen, P.O. Minkinen, E.Y.O. Paatero, T.O. Salmi, Determination of acid value, hydroxyl value and water content in reactions between dicarboxylic acids and diols using near-infrared spectroscopy and non-linear partial least squares regression, *Anal. Chim. Acta* 349 (1997) 287–294, [https://doi.org/10.1016/S0003-2670\(97\)00215-8](https://doi.org/10.1016/S0003-2670(97)00215-8).
- [9] N. Heigl, C.H. Petter, M. Rainer, M. Najam-ul-Haq, R.M. Vallant, R. Bakry, G. K. Bonn, C.W. Huck, Near infrared spectroscopy for polymer research, quality control and reaction monitoring, *J. Near Infrared Spectrosc.* 15 (2017) 269–282, <https://doi.org/10.1255/jnirs.747>.
- [10] O.R. Ghita, D.C. Baker, K.E. Evans, An in-line near-infrared process control tool for monitoring the effects of speed, temperature, and polymer colour in injection moulding, *Polym. Test.* 27 (2008) 459–469, <https://doi.org/10.1016/j.polymertesting.2008.01.010>.
- [11] M.M. Reis, P.H. Araujo, C. Sayer, R. Giudici, Spectroscopic on-line monitoring of reactions in dispersed medium: chemometric challenges, *Anal. Chim. Acta* 595 (2007) 257–265, <https://doi.org/10.1016/j.aca.2007.04.048>.
- [12] M.M. Reis, P.H.H. Araujo, C. Sayer, R. Giudici, In situ near-infrared spectroscopy for simultaneous monitoring of multiple process variables in emulsion copolymerization, *Ind. Eng. Chem. Res.* 43 (2004) 7243–7250, <https://doi.org/10.1021/ie034277u>.
- [13] Y.J. Wu, Y. Jin, Y.R. Li, D. Sun, X.S. Liu, Y. Chen, NIR spectroscopy as a process analytical technology (PAT) tool for on-line and real-time monitoring of an extraction process, *Vib. Spectrosc.* 58 (2012) 109–118, <https://doi.org/10.1016/j.vibspec.2011.10.006>.
- [14] T.A. Catelani, J.R. Santos, R. Pascoa, L. Pezza, H.R. Pezza, J.A. Lopes, Real-time monitoring of a coffee roasting process with near infrared spectroscopy using multivariate statistical analysis: a feasibility study, *Talanta* 179 (2018) 292–299, <https://doi.org/10.1016/j.talanta.2017.11.010>.
- [15] H. Huang, H. Qu, In-line monitoring of alcohol precipitation by near-infrared spectroscopy in conjunction with multivariate batch modeling, *Anal. Chim. Acta* 707 (2011) 47–56, <https://doi.org/10.1016/j.aca.2011.09.031>.
- [16] R.R. de Oliveira, C. Avila, R. Bourne, F. Muller, A. de Juan, Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control, *Anal. Bioanal. Chem.* 412 (2020) 2151–2163, <https://doi.org/10.1007/s00216-020-02404-2>.
- [17] D.S. Bu, B.Y. Wan, G. McGeorge, A discussion on the use of prediction uncertainty estimation of NIR data in partial least squares for quantitative pharmaceutical tablet assay methods, *Chemometr. Intell. Lab. Syst.* 120 (2013) 84–91, <https://doi.org/10.1016/j.chemolab.2012.11.005>.
- [18] Z.P. Chen, J. Morris, Pat: the extraction of maximum information from messy spectral data, *IFAC Proceedings Volumes* 40 (2007) 7–12, <https://doi.org/10.3182/20070604-3-mx-2914.00003>.
- [19] A. Cherfi, G. Fevotte, C. Novat, Robust on-line measurement of conversion and molecular weight using NIR spectroscopy during solution polymerization, *J. Appl. Polym. Sci.* 85 (2002) 2510–2520, <https://doi.org/10.1002/app.10727>.
- [20] T. Chen, E. Martin, The impact of temperature variations on spectroscopic calibration modelling: a comparative study, *J. Chemometr.* 21 (2007) 198–207, <https://doi.org/10.1002/cem.1041>.
- [21] A.E. Cervera, N. Petersen, A.E. Lantz, A. Larsen, K.V. Gernaey, Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation, *Biotechnol. Prog.* 25 (2009) 1561–1581, <https://doi.org/10.1002/btpr.280>.
- [22] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives - a review, *Anal. Chim. Acta* 1026 (2018) 8–36, <https://doi.org/10.1016/j.aca.2018.04.004>.
- [23] O.S. Heavens, The fabry-perot-interferometer - history, theory, practice and applications - vaughan,Jm, *Nature* 341 (1989), <https://doi.org/10.1038/341194a0>, 194-194.
- [24] J. Antila, M. Tuohiniemi, A. Rissanen, U. Kantojärvi, M. Lahti, K. Viherkanto, M. Kaarre, J. Malinen, MEMS- and MOEMS-based near-infrared spectrometers, *encyclopedia of analytical chemistry* (2014) 1–36, <https://doi.org/10.1002/9780470027318.a9376>.
- [25] A. Rogalski, Progress in focal plane array technologies, *Prog. Quant. Electron.* 36 (2012) 342–473, <https://doi.org/10.1016/j.pquantelec.2012.07.001>.
- [26] M. Blomberg, A. Torkkeli, A. Lehto, C. Helenelund, M. Viitasalo, Electrically tuneable micromachined Fabry-Perot interferometer in gas analysis, *Phys. Scripta* T69 (1997) 119–121, <https://doi.org/10.1088/0031-8949/1997/T69/018>.
- [27] A. Ajujarvi, B. Guo, R. Mannila, A. Rissanen, MOEMS FPI sensors for NIR - MIR microspectrometer applications, *Proc. SPIE* (2016) 9760, <https://doi.org/10.1117/12.2214710>.
- [28] H. Vakili, H. Wickstrom, D. Desai, M. Preis, N. Sandler, Application of a handheld NIR spectrometer in prediction of drug content in inkjet printed orodispersible formulations containing prednisolone and levothyroxine, *Int. J. Pharm.* 524 (2017) 414–423, <https://doi.org/10.1016/j.ijpharm.2017.04.014>.
- [29] A. Rissanen, B. Guo, H. Saari, A. Nasila, R. Mannila, A. Ajujarvi, H. Ojanen, VTT's Fabry-Perot interferometer technologies for hyperspectral imaging and mobile sensing applications, *Moems and Miniaturized Systems Xvi* (2017) 10116, <https://doi.org/10.1117/12.2255950>.
- [30] C.R. Avila, J. Ferre, R.R. de Oliveira, A. de Juan, W.E. Sinclair, F.M. Mahdi, A. Hassanpour, T.N. Hunter, R.A. Bourne, F.L. Muller, Process monitoring of moisture content and mass transfer rate in a fluidised bed with a low cost inline MEMS NIR sensor, *Pharm. Res. (N. Y.)* 37 (2020) 84, <https://doi.org/10.1007/s11095-020-02787-y>.
- [31] C. Avila, *ChemView*, <https://chemview.leeds.ac.uk/>, 2015. (Accessed 18 September 2020).
- [32] P.A. Gorry, General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method, *Anal. Chem.* 62 (1990) 570–573, <https://doi.org/10.1021/ac00205a007>.
- [33] H. Martens, T. Naes, *Multivariate Calibration*, Wiley, Chichester, 1989.
- [34] T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry, *Int. J. Adapt. Contr.* 19 (2005) 213–246, <https://doi.org/10.1002/acs.859>.
- [35] J.F. Macgregor, T. Kourti, Statistical process-control of multivariate processes, *Contr. Eng. Pract.* 3 (1995) 403–414, [https://doi.org/10.1016/0967-0661\(95\)00014-L](https://doi.org/10.1016/0967-0661(95)00014-L).
- [36] S. Wold, Cross-validatory estimation of number of components in factor and principal components models, *Technometrics* 20 (1978) 397–405, <https://doi.org/10.2307/1267639>.
- [37] J.E. Jackson, G.S. Mudholkar, Control procedures for residuals associated with principal component analysis, *Technometrics* 21 (1979) 341–349, <https://doi.org/10.2307/1267757>.
- [38] H.A. Kishawy, H. Hegab, E. Saad, Design for sustainable manufacturing: approach, implementation, and assessment, *Sustainability* (2018) 10, <https://doi.org/10.3390/su10103604>.





# Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control

Rodrigo R. de Oliveira<sup>1</sup> · Claudio Avila<sup>2</sup> · Richard Bourne<sup>2</sup> · Frans Muller<sup>2</sup> · Anna de Juan<sup>1</sup>

Received: 4 September 2019 / Revised: 8 January 2020 / Accepted: 10 January 2020 / Published online: 21 January 2020  
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

## 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 fluidized 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) 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

## 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 and flow [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

Published in the topical collection *Advances in Process Analytics and Control Technology* with guest editor Christoph Herwig.

✉ Rodrigo R. de Oliveira  
rodrigo.rocha@ub.edu

<sup>1</sup> Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

<sup>2</sup> School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, UK

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 was 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 behaviour 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 modelling tasks for the same process are interconnected. Indeed, model outputs derived from multivariate sensors, such as predictions of key properties and process concentration profiles, 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 fluidized 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.

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<sub>NIR</sub>) 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.

## Experimental

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

### 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<sup>3</sup>/min and a temperature of 22 to 30 °C. Temperature sensor readings of the fluidized material ( $T_{\text{bed}}$ ), inlet air ( $T_{\text{in}}$ ) and outlet air ( $T_{\text{out}}$ ) 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 analyser (MB120, Ohaus, Germany) from samples retrieved at 6-min 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].

### 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 2-L round flask reactor using a NIR immersion probe (Excalibur 20, Hellma Analytics, Germany) with an optical path length of 5 mm working in the transmission mode and connected through 2-m fibre 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 s. 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<sup>-1</sup> for AV and 10 to 14 P for V, whereas the end point of the second stage requires 45 to 63 mg KOH g<sup>-1</sup> for AV and 25 to 45 P for V. More information can be found in reference [23].

### 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 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].

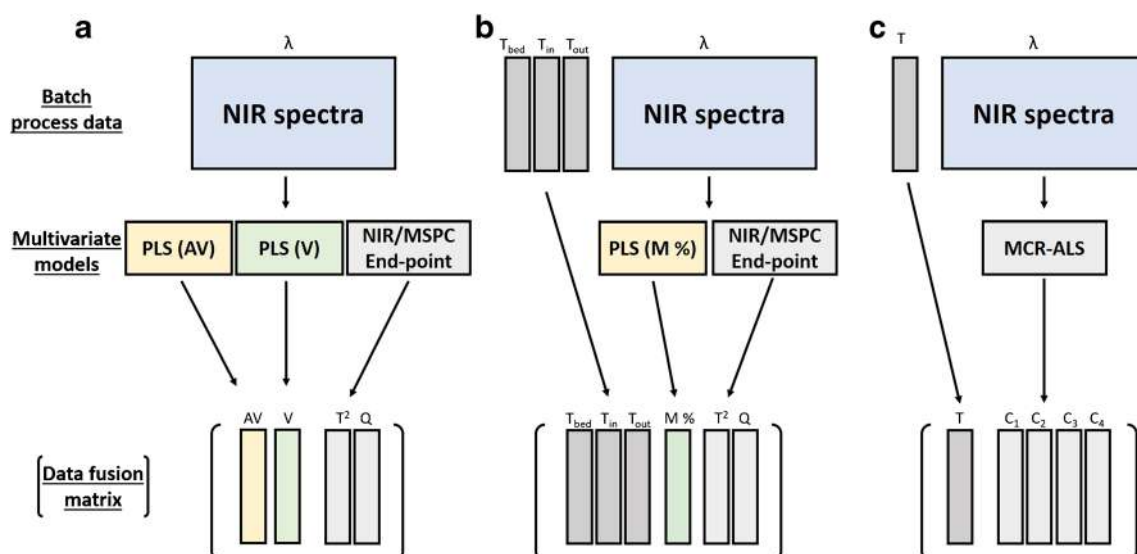


Fig. 1 Data fusion strategies used to combine the several sensor and/or model outputs for batches from (a) process 2; (b) process 1; and (c) process 3

NIR spectra in the 1103–2228-nm wavelength range (573 channels) and vapour temperature were recorded every unit of percent of distilled mass fraction in the 5 to 90% range.

## Data treatment

### NIR data preprocessing

For processes 1 and 2, similar preprocessing steps were employed to filter out noise and baseline fluctuations on NIR spectrum observations. A certain number of consecutive raw spectra measurements,  $N_{\text{RAW}}$ , were averaged into a single spectrum; then, a moving average smoothing with window size,  $N_{\text{MA}}$ , 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] (1st-order derivative, 2nd-order polynomial function and 15-point 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  $N_{\text{RAW}}$  averaged to obtain a single one.

Figure 2 (a) shows the raw (left plot) and preprocessed (centre plot) NIR observations using  $N_{\text{RAW}} = 10$  and  $N_{\text{MA}} = 75$  for one typical drying process batch (*process 1*). The right plot shows the related batch temperature profiles. Figure 2 (b) shows the raw (left plot) and (right plot) preprocessed NIR observations using  $N_{\text{RAW}} = 13$  and  $N_{\text{MA}} = 30$  for the polyester production process batch (*process 2*).

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

### 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  $\mathbf{X}$  matrix (formed by NIR spectra in these examples) to the matrix of parameters to be predicted  $\mathbf{Y}$  (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  $\mathbf{X}$  and  $\mathbf{Y}$ . More details and description of PLS algorithm can be found elsewhere [28–30].  $\mathbf{Y}$  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 Fig. 1 (a) and (b), respectively.

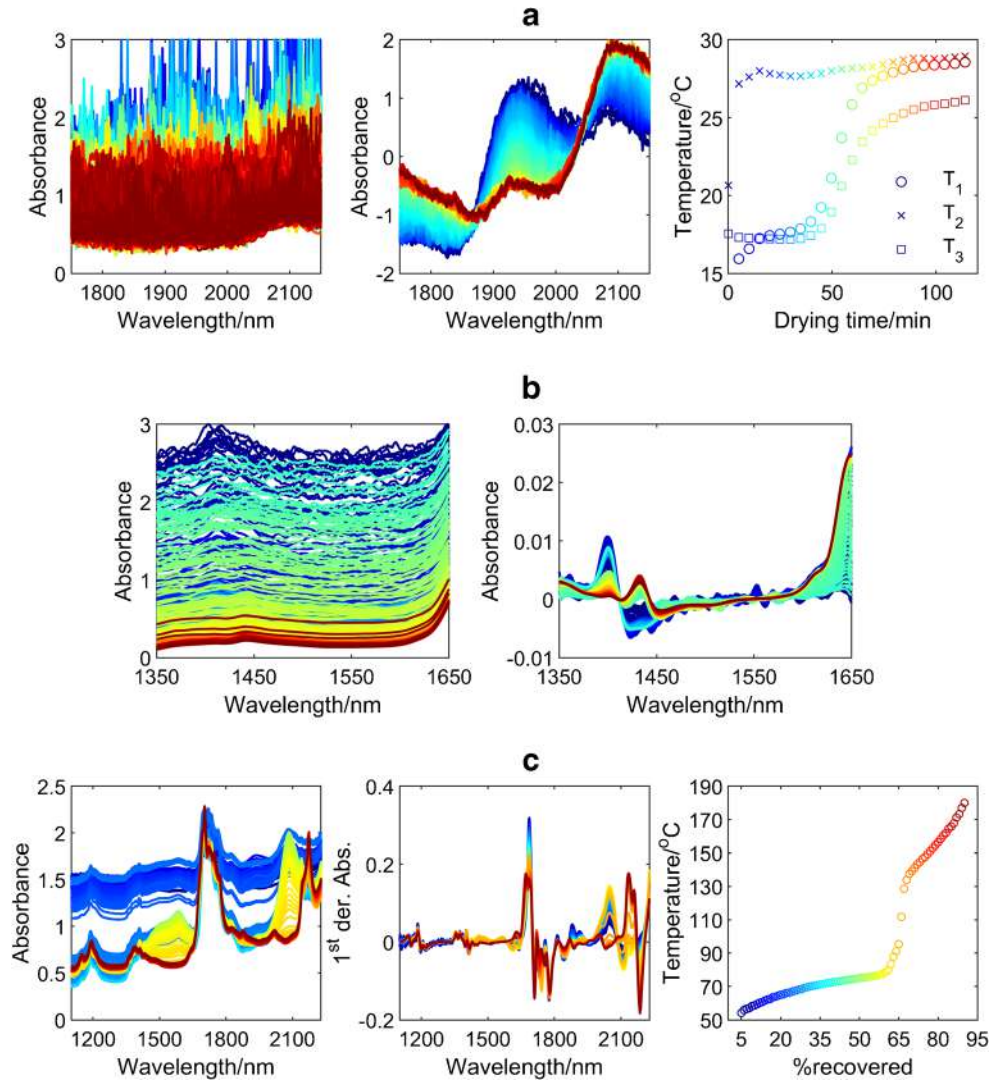
Multivariate curve resolution-alternating least squares (MCR-ALS) is a method that can provide concentration profiles and related spectral signatures for the compounds involved in a process using only the spectroscopic information recorded during process monitoring. MCR-ALS was used to model the NIR data of the distillation process, *process 3*. MCR-ALS assumes a bilinear model,  $\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E}$ , which is the multiwavelength extension of Lambert-Beer's law [31–34]. In this context,  $\mathbf{D}$  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  $\mathbf{C}$  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  $\mathbf{C}$  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),  $\mathbf{C}$  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 (b).

### Construction of multivariate statistical process control 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

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



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,  $\mathbf{X}_{\text{NOC}}$ , 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  $\mathbf{X}_{\text{NOC}}$  is modelled by PCA in order to set the statistical

boundaries of the experimental domain (space) of NOC observations according to the equation below [35],

$$\mathbf{X}_{\text{NOC}} = \mathbf{T}_{\text{NOC}} \mathbf{P}_{\text{NOC}}^T + \mathbf{E}_{\text{NOC}} \quad (1)$$

where  $\mathbf{T}_{\text{NOC}}$  is the scores matrix of the NOC observations used to build the model and  $\mathbf{P}_{\text{NOC}}^T$ , the loadings matrix (which is the link between scores and original variables in  $\mathbf{X}_{\text{NOC}}$ ).  $\mathbf{E}_{\text{NOC}}$  describes the residual variation unexplained by the PCA model and is used to define the  $Q$ -statistic control chart limit,  $Q_{\text{lim}}$ , according to the 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:

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

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

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



And the related  $Q$ -statistic value as:

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

When the new observation follows the NOC described by the MSPC models, the residual  $\mathbf{e}_{\text{i,new}}$  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  $\mathbf{e}_{\text{new}}$  vector for the sought observation. High absolute values related to elements in  $\mathbf{e}_{\text{new}}$  will identify variables showing abnormal behaviour.

From the PCA model, another statistical parameter can be obtained, 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:

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

where  $\Theta$  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 was also used as means to represent compressed process information from purely NIR-based MSPC models in *processes 1* and *2*, as mentioned in "NIR-based model outputs used in data fusion MSPC models" section.

Often, for an easier interpretation of MSPC control charts and indicators, reduced NIR-MSPC statistics ( $Q_{\text{red}}$  and  $T_{\text{red}}^2$ ) 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 1.

## 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.

## 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<sub>NIR</sub>) both in terms of detecting process faults and identifying the causes of the abnormal process behaviour.

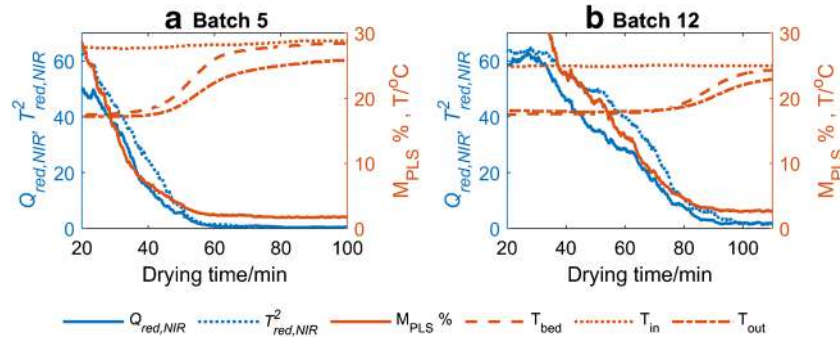
### Process1: Fluidized bed drying process

In this process, the DF-MSPC model combines temperature sensor readings of the fluidized material ( $T_{\text{bed}}$ ), inlet air ( $T_{\text{in}}$ ) and outlet air ( $T_{\text{out}}$ ) and information coming from two NIR-based multivariate models, a PLS regression model for prediction of moisture content and a MSPC<sub>NIR</sub> model for end point detection, which provided  $T_{\text{red,NIR}}^2$  and  $Q_{\text{red,NIR}}$  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<sub>NIR</sub> 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<sub>NIR</sub> 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  $M_{\text{PLS}}$  %,  $T_{\text{red,NIR}}^2$  and  $Q_{\text{red,NIR}}$  and the three temperature values  $T_{\text{bed}}$ ,  $T_{\text{in}}$ , and  $T_{\text{out}}$ . 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). Figure 3 shows the information to be submitted to the DF-MSPC model, i.e.  $T_{\text{bed}}$ ,  $T_{\text{in}}$  and  $T_{\text{out}}$  and  $M$  %,  $T_{\text{red,NIR}}^2$  and  $Q_{\text{red,NIR}}$  for validation batches 5 and 12 during the drying process.

Figure 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 ( $M_{\text{PLS}}$  %) was found to be 1.8%, the temperature readings were  $T_{\text{bed}}$ , 28.4 °C;  $T_{\text{in}}$ , 28.8 °C; and  $T_{\text{out}}$ , 25.8 °C, and the MSPC indicators  $Q_{\text{red,NIR}}$ , 0.5, and  $T_{\text{red,NIR}}^2$ , 0.2, both below the control limit set equal to 1, 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 were  $T_{\text{bed}}$ , 24.4 °C,  $T_{\text{in}}$ , 25.0 °C, and  $T_{\text{out}}$ , 22.9 °C, respectively; and the MSPC indicators  $Q_{\text{red,NIR}}$ , 1.9, and  $T_{\text{red,NIR}}^2$ , 1.3, were both above the control limit set equal to 1 indicating that the batch did not reach the end point and could be considered off-specification. The main reason why batch 12 did not reach the specification moisture level



**Fig. 3** Information to be analysed 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 sensor readings placed at granules, inlet air 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<sub>NIR</sub> model. First 20 min is not shown because of unstable measurements at the beginning of the process

during the drying process time was because of the low inlet air temperature, about 4 °C lower than that in batch 5, which could be due to changes in the process environment or other uncontrolled causes.

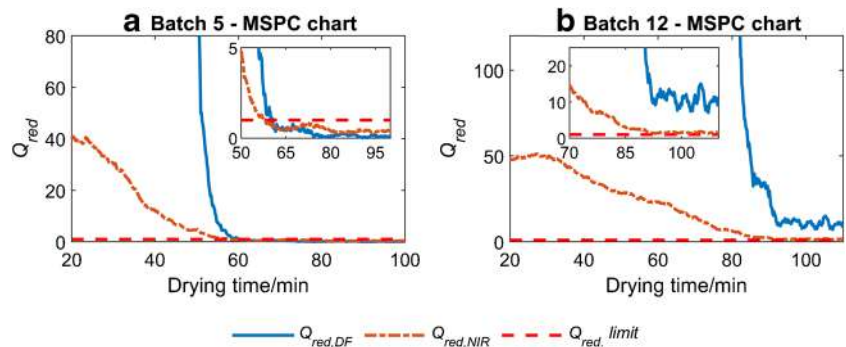
The data fusion information from each validation batch, shown in Fig. 3, was submitted to the DF-MSPC model for end point detection. To understand the better performance of the DF-MSPC model when compared with the MSPC<sub>NIR</sub> model, Fig. 4 shows overlapped  $Q_{red}$  charts for both approaches. 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<sub>NIR</sub> model.

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

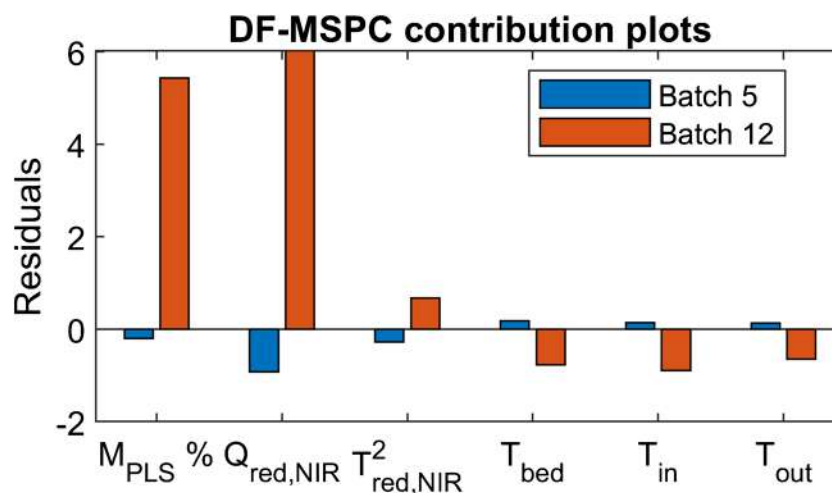
temperature in the DF-MSPC models. Moreover, for both batches 5 and 12, the use of DF-MSPC models also provides a much clearer difference between the  $Q_{red,DF}$  values before and after moisture stabilization than the  $Q_{red,NIR}$  values, being the decrease of the  $Q_{red,DF}$  curve always much 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<sub>NIR</sub> model) is also higher. Figure 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

**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



**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



dry batch and, such a fact, caused the lower values for the related  $T_{bed}$  and  $T_{out}$ .

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.

## 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-min 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].

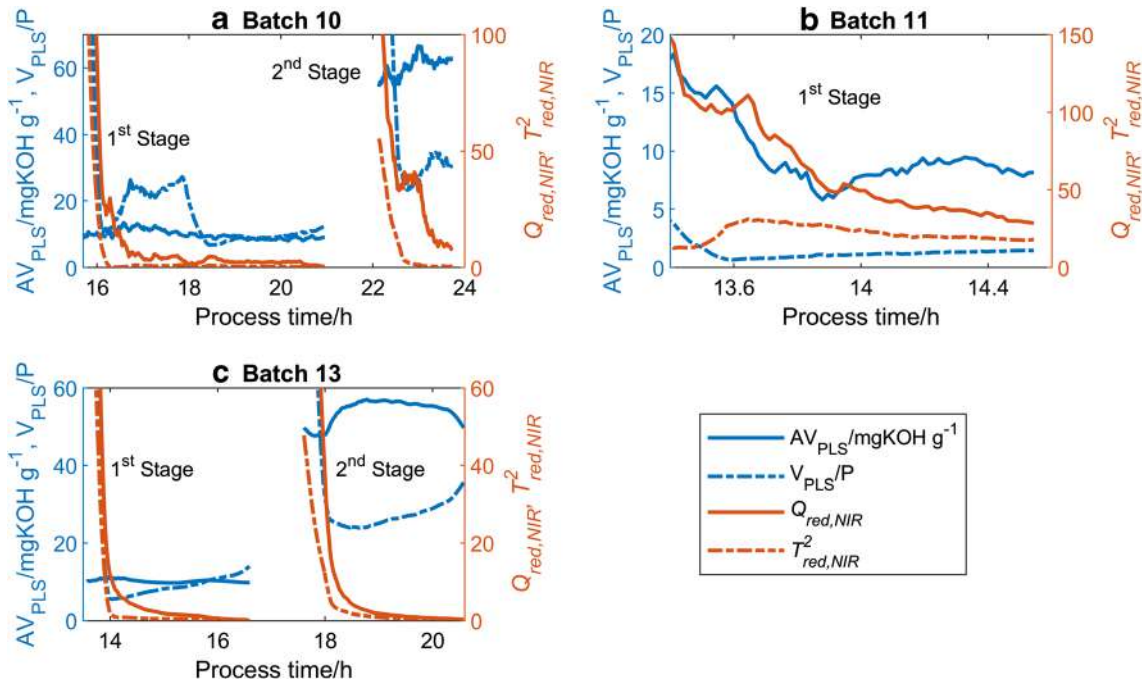
The DF-MSPC models linked to the end point of the two reaction stages were built using information of NOC batches,

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. Figure 6 shows the information from PLS and  $MSPC_{NIR}$  models for three validation batches of the polymerization process (labelled batches 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 spectrum saturation, the information in Fig. 6 is omitted for the first 12 to 15 h of the first reaction stage and for approximately 1 h 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 1, 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.

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



**Fig. 6** Information to be analysed 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).

14.5 h of process time,  $AV_{PLS}$  was 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 indicates that batch 13 was considered on-specification for both process stages.

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

Figure 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

$AV_{PLS}$  (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)

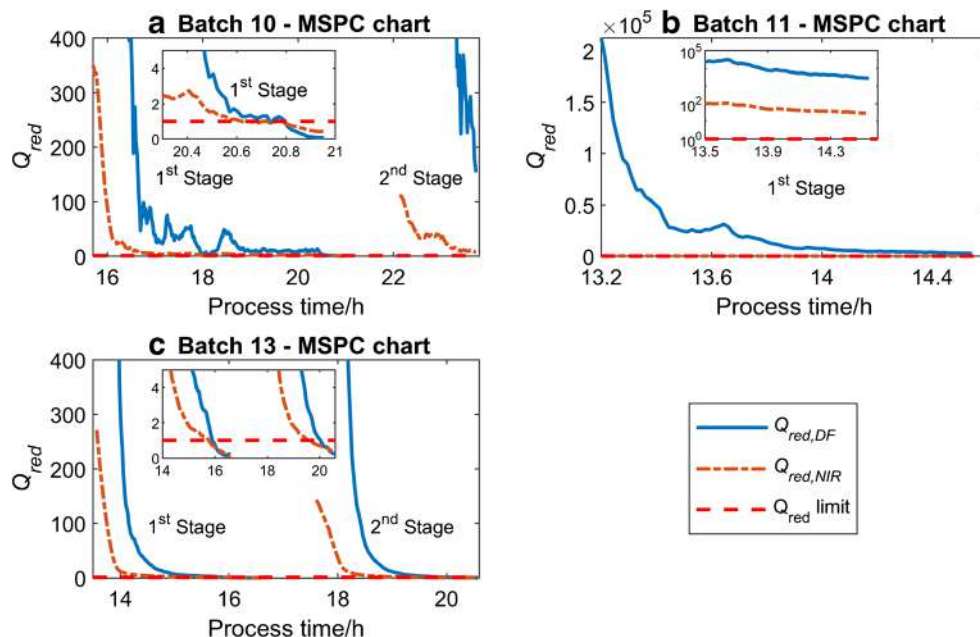
consequence of including the explicit predictions of product quality parameters, i.e. AV and V, in the model. Batch 11 represents a faulty batch in the first process stage and no end point was detected in batch 11 control charts shown in Fig. 7(b).  $Q_{red, DF}$  values obtained from DF-MSPC models 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-MSPC 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.

### Process 3: Gasoline distillation

This example shows a different combination of temperature and NIR sensor model outputs for process evolution control

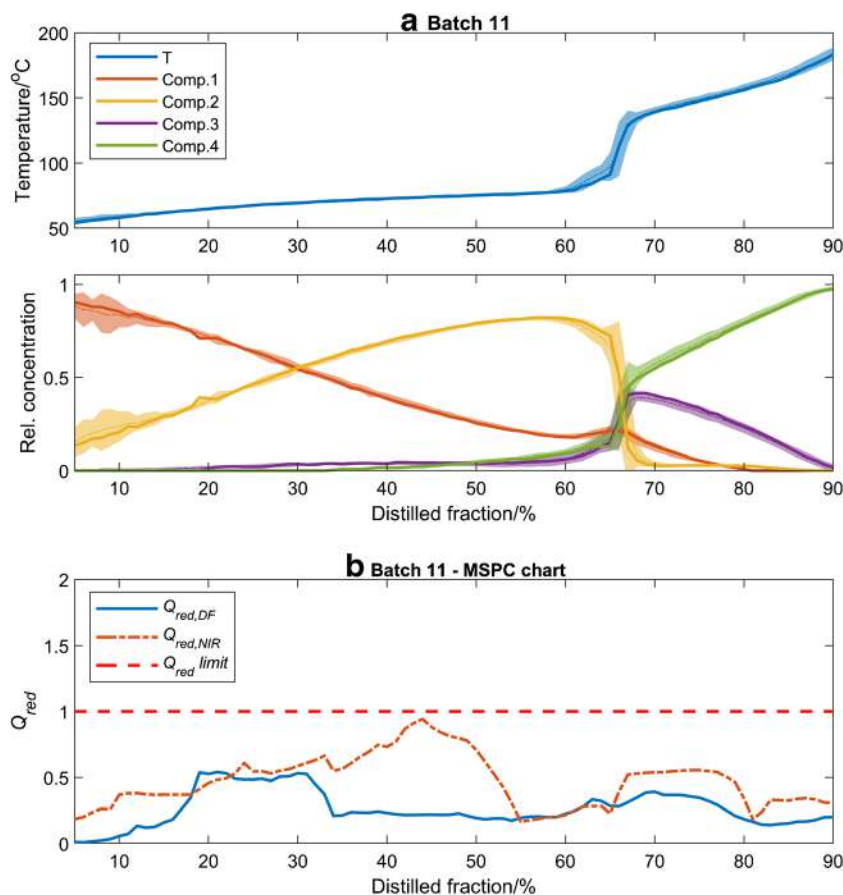
**Fig. 7**  $Q_{red}$  MSPC charts for polymerization process stage end point detection using DF-MSPC model (solid blue curve) and MSPCNIR model (dashed orange curve) for validation batches: (a) batch 10 (1st stage on-specification, 2nd stage off-specification), (b) batch 11 (off-specification) and (c) batch 13 (1st and 2nd stages on-specification). 95% CI  $Q_{red}$  control limit is represented by the dashed red flat line equal to 1. 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



and is the clearest example of combination of integral process modelling (as provided by the C profiles from MCR-ALS, see the “NIR-based model outputs used in data fusion MSPC models” section) 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 ( $S^T$ ) for the four different fractions distilled in the gasoline system studied. The four distillation

**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 batch information (thin dashed line curves are the average and the related  $\pm 2$  standard deviation bounds are the shaded area). (b)  $Q_{red,NIR}$  control charts from DF-MSPC models (solid blue curve) and from MSPCNIR model (orange dashed curve) using the FSMW on-line MSPC strategy



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-size 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.

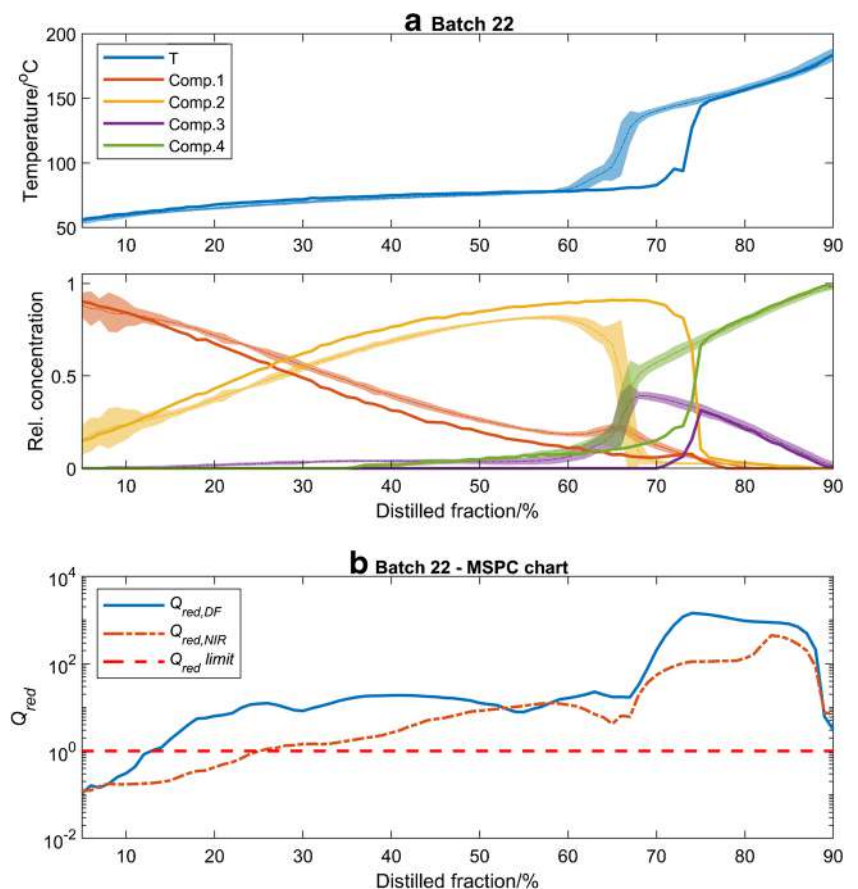
Figures 8(a) and 9(a) show the information (distillation curve and NIR-based MCR concentration profiles of distilled fractions) for the NOC batches used to build the local DF-MSPC models and for two validation batches submitted to the local DF-MSPC models for testing. In both Fig. 8(a) (showing an on-specification batch) and Fig. 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  $\pm 2$  standard deviation bounds.

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, Fig. 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 Figs. 8(a) and 9(a), was submitted to the related local DF-MSPC models. For comparison, Figs. 8(b) and 9(b) show the control charts obtained for each validation batch when using the local DF-MSPC models (blue line) or the MSPC<sub>NIR</sub> models (orange dotted line). Please note that the  $Q_{red, DF}$  and  $Q_{red, NIR}$  values associated with every observation in Figs. 8(b) and 9(b) come from a different local DF-MSPC or MSPC<sub>NIR</sub> model, respectively, as described in detail in [7]. In this example, the  $x$ -axis in the control charts refers to % distilled mass fraction and  $Q_{red}$  values need to be always below 1 to indicate that the process evolves correctly. One or

**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 batch information (thin dashed line curves are the average and the related  $\pm 2$  standard deviation bounds are the shaded area). (b)  $Q_{red, NIR}$  control charts from DF-MSPC models (solid blue curve) and from MSPC<sub>NIR</sub> model (orange dashed curve) using the FSMW on-line MSPC strategy.  $y$ -axis is represented with  $\log_{10}$  scale



more values above 1 indicate that, in those particular process stages, the new batch does not proceed as NOC batches.

Figure 8 (b) shows that all  $Q_{\text{red}}$  values related to process observations in batch 11 (on-specification gasoline), issued from DF-MSPC and MSPC<sub>NIR</sub> 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  $Q_{\text{red, DF}}$  values are lower than  $Q_{\text{red, NIR}}$  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  $Q_{\text{red, DF}}$  or  $Q_{\text{red, NIR}}$  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,  $Q_{\text{red, DF}}$  values obtained when data fusion information was used were, in general, higher than  $Q_{\text{red, NIR}}$  values, 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 Figs. 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 also their interactions as well. Reference [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.

## 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 is 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 and 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.

**Funding information** This study received 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. This study also received funding from the Spanish government under the project CTQ2015-66254-C2-2-P.

## Compliance with ethical standards

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

## References

1. Gurden SP, Westerhuis JA, Smilde AK. Monitoring of batch processes using spectroscopy. *AIChE J.* 2002;48:2283–97.
2. Gabrielsson J, Jonsson H, Trygg J, Airiau C, Schmidt B, Escott R. Combining process and spectroscopic data to improve batch modeling. *AIChE J.* 2006;52:3164–72.
3. Huang J, Kaul G, Utz J, Hernandez P, Wong V, Bradley D, et al. A PAT approach to improve process understanding of high shear wet

- granulation through in-line particle measurement using FBRM C35. *J Pharm Sci.* 2010;99:3205–12.
4. Jin Y, Wu Z, Liu X, Wu Y. 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.* 2013;77:32–9.
  5. Lourenço ND, Lopes JA, Almeida CF, Sarraguça MC, Pinheiro HM. Bioreactor monitoring with spectroscopy and chemometrics: a review. *Anal Bioanal Chem.* 2012;404:1211–37.
  6. Zhao C, Gao F, Wang F. Phase-based joint modeling and spectroscopy analysis for batch processes monitoring. *Ind Eng Chem Res.* 2010;49:669–81.
  7. de Oliveira RR, Pedroza RHP, Sousa AO, Lima KMG, de Juan A. Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy. *Anal Chim Acta.* 2017;985:41–53.
  8. Catelani TA, Santos JR, Páscoa RNMJ, Pezza L, Pezza HR, Lopes JA. Real-time monitoring of a coffee roasting process with near infrared spectroscopy using multivariate statistical analysis: a feasibility study. *Talanta.* 2018;179:292–9.
  9. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal component analysis. *AIChE J.* 1994;40:1361–75.
  10. Wold S, Kettaneh N, Friden H, Holmberg A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom Intell Lab Syst.* 1998;44:331–40.
  11. Huang H, Qu H. In-line monitoring of alcohol precipitation by near-infrared spectroscopy in conjunction with multivariate batch modeling. *Anal Chim Acta.* 2011;707:47–56.
  12. Huang J, Goolcharran C, Utz J, Hernandez-Abad P, Ghosh K, Nagi A. A PAT approach to enhance process understanding of fluid bed granulation using in-line particle size characterization and multivariate analysis. *J Pharm Innov.* 2010;5:58–68.
  13. Mattila M, Saloheimo K, Koskinen K. Improving the robustness of particle size analysis by multivariate statistical process control. *Part Part Syst Charact.* 2007;24:173–83.
  14. Faggian A, Facco P, Doplicher F, Bezzo F, Barolo M. Multivariate statistical real-time monitoring of an industrial fed-batch process for the production of specialty chemicals. *Chem Eng Res Des.* 2009;87:325–34.
  15. Marjanovic O, Lennox B, Sandoz D, Smith K, Crofts M. Real-time monitoring of an industrial batch process. *Comput Chem Eng.* 2006;30:1476–81.
  16. Aguado D, Ferrer A, Ferrer J, Seco A. Multivariate SPC of a sequencing batch reactor for wastewater treatment. *Chemom Intell Lab Syst.* 2007;85:82–93.
  17. González-Martínez JM, Ferrer A, Westerhuis JA. Real-time synchronization of batch trajectories for on-line multivariate statistical process control using dynamic time warping. *Chemom Intell Lab Syst.* 2011;105:195–206.
  18. Cimander C, Mandenius CF. Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modelling. *J Chem Technol Biotechnol.* 2002;77:1157–68.
  19. Cimander C, Carlsson M, Mandenius CF. Sensor fusion for on-line monitoring of yoghurt fermentation. *J Biotechnol.* 2002;99:237–48.
  20. Jiang H, Chen Q. Development of electronic nose and near infrared spectroscopy analysis techniques to monitor the critical time in SSF process of feed protein. *Sensors (Switzerland).* 2014;14:19441–56.
  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. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem.* 1964;36:1627–39.
  26. Brereton RG. Introduction to multivariate calibration in analytical chemistry. *Analyst.* 2000;125:2125–54.
  27. Booksh KS, Kowalski BR. Theory of analytical chemistry. *Anal Chem.* 1994;66:782A–91A.
  28. Martens H, Næs T. *Multivariate calibration.* New York: John Wiley & Sons; 1991.
  29. Thomas EV. A primer on multivariate calibration. *Anal Chem.* 1994;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. Multivariate curve resolution (MCR). Solving the mixture analysis problem. *Anal Methods.* 2014;6:4964–76.
  32. de Juan A, Tauler R. Chemometrics applied to unravel multicomponent processes and mixtures. *Anal Chim Acta.* 2003;500:195–210.
  33. Tauler R, Kowalski BR, Fleming S. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Anal Chem.* 1993;65:2040–7.
  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. Control procedures for residuals associated with principal component analysis. *Technometrics.* 1979;21:341–9.
  37. MacGregor JF, Kourti T. Statistical process control of multivariate processes. *Control Eng Pract.* 1995;3:403–14.
  38. Kourti T. Process analysis and abnormal situation detection: from theory to practice. *IEEE Control Syst Mag.* 2002;22:10–25.
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





### **Process modeling and control based on the sole use of NIR spectroscopic information**

The results presented in Publications II and III demonstrate the use of multivariate calibration models (PLS) for real-time monitoring of critical quality attributes and the application of process control (MSPC) models for endpoint detection in two inline NIR-monitored batch processes: a fluidized bed drying of pharmaceutical granules (Publication II) and a high-temperature multi-step polyester production process (Publication III). Both processes were carried out in pilot reactors designed to study the PAT application for a real industrial environment.

The constant flow of multiphasic materials, i.e., fluidized granules in the drying process and N<sub>2</sub> bubbles, solid and liquid reactants and products in the polyester reaction, presented an important challenge to extract useful information from the raw NIR spectra acquired during inline process monitoring. Intense preprocessing of the raw NIR data, as shown in sections 3.1.1 and 3.1.2, was required to obtain sensor measurements that could allow setting reliable multivariate calibration and MSPC models. Thus, using the preprocessed NIR data, useful models for real-time process monitoring of critical quality attributes and process control for endpoint detection were successfully implemented for both applications.

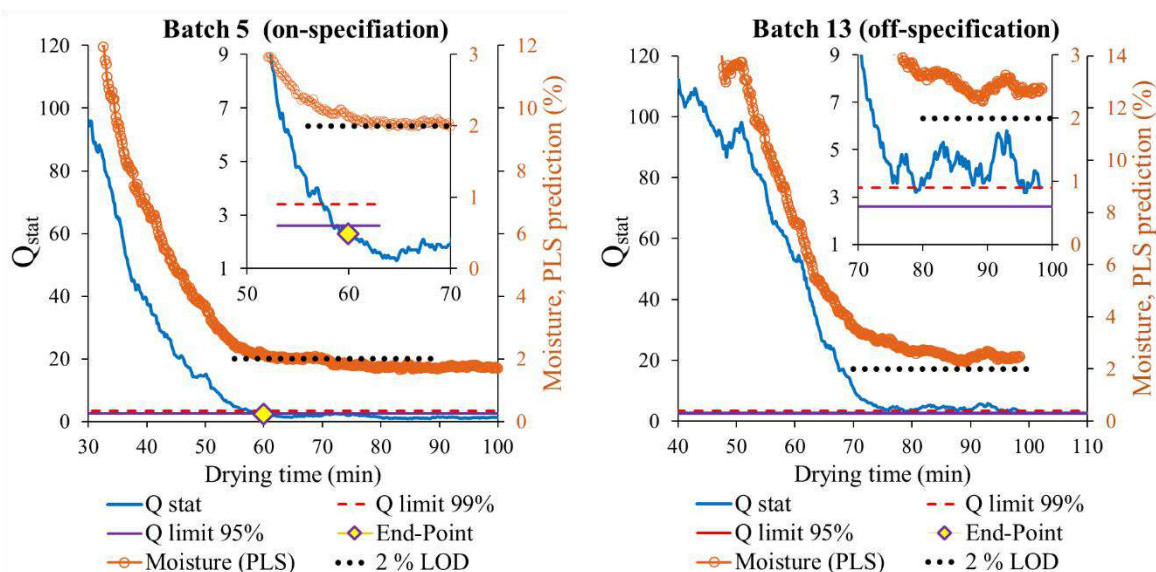


Figure 24 Results issued from the NIR inline monitoring of a fluid bed drying process. Moisture predictions from a PLS model (red lines in plots, right y axis) and Q control charts from an MSPC model for batch endpoint detection (blue lines and left y axis) for on-specification batch 5 (left) and off-specification batch 13 (right). Inserted figures show in detail the final time range of the drying process and the process endpoint, identified when 10 consecutive observations of  $Q_{stat}$  values were below the 95% control limit. Reproduced from (Avila et al., 2020).

Figure 24 shows the results for the process monitoring and control during the inline NIR monitoring of two drying batches. Figure 24A (left panel) shows the results of a NOC batch that correctly dried the pharmaceutical granules, seen because the moisture content of the end product was below the 2% (w/w) specification, as shown

by the PLS predictions. The batch endpoint was also successfully detected using the  $Q$ -residual MSPC chart using the sole NIR information. On the other hand, an example of the results from a batch that did not reach the moisture specification level and so the desirable endpoint can be visualized in Figure 24A (right panel). In this case, at the end of this batch run, both PLS predictions and  $Q$  residuals were above the expected on-specification characteristics of dry granules, see the inset plot in Figure 24A (right panel). These results demonstrate the value of using inline NIR spectroscopy combined with multivariate models for process monitoring and control of FB drying processes. More examples of on- and off-specification batches can be found in Paper II leading to similar conclusions. The design of the MEMS sensor and the physicochemical description of the drying process are contributions not linked to the scope of this thesis and performed by other coauthors and, hence, are not commented in this section.

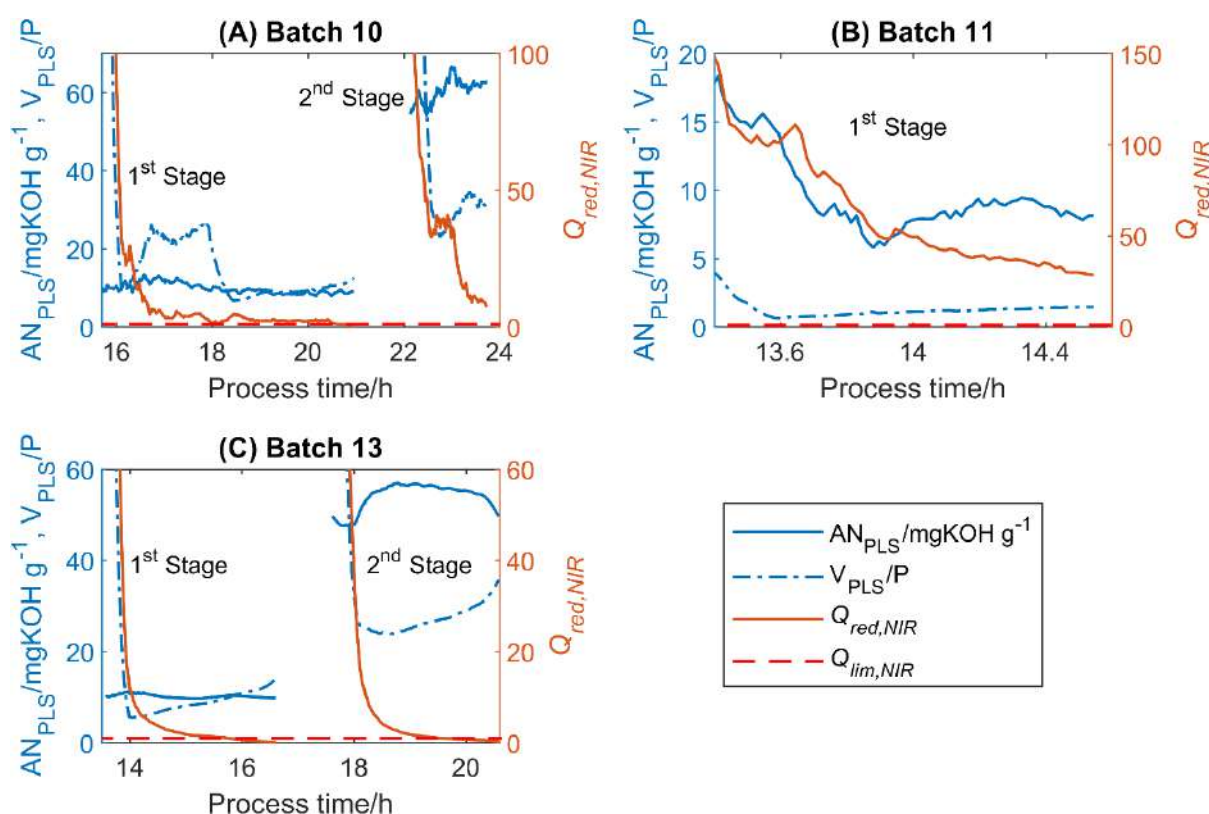


Figure 25 Results issued from the inline NIR monitoring of a polyester production process. PLS predictions of acid number ( $AN_{PLS}$ ) and viscosity ( $V_{PLS}$ ) and  $Q_{red}$  process control MSPC charts for the two-step MSPC endpoint detection for three validation batches. (A) batch 10 (1<sup>st</sup> stage on-specification, 2<sup>nd</sup> stage off-specification), (B) batch 11 (1<sup>st</sup> stage off-specification) and (C) batch 13 (1<sup>st</sup> and 2<sup>nd</sup> stages on-specification).  $AN_{PLS}$  and  $V_{PLS}$  are represented by blue solid and dashed curves, respectively (y left axis).  $Q_{red}$  is represented by the orange solid curve (right axis) and its control limit  $Q_{lim} = 1$ , is represented by the dashed red line. Reproduced from (de Oliveira et al., 2020).

Figure 25 shows the results for process monitoring and control using the inline NIR measurements for three polyester production batches. As mentioned in section 3.1.1., the reactions linked to this process were taking place in two main steps. For each of these steps, it was relevant monitoring some key parameters and setting accurately the endpoint. Hence, the inline collected NIR information was used to set multiple PLS

models to predict acid number (AN, expressed as mg KOH g<sup>-1</sup>) and viscosity (V, expressed as poises, P). It is important to mention that due to the two-step batch reaction (1<sup>st</sup> Stage and 2<sup>nd</sup> Stage) and the important differences in chemical composition between the two steps, different PLS models were built for each stage. For the first stage, accepted values were AN 8–12 (mg KOH g<sup>-1</sup>) and V 10–14 (P or g cm<sup>-1</sup> s<sup>-1</sup>) and for the second reaction stage AN 45–63 (mg KOH g<sup>-1</sup>) and V 25–45 P. Analogously to PLS models, two MSPC models for endpoint detection were developed to detect the endpoint of each process stage. Figure 25 shows selected validation batches that display the different situations encountered during the study of these polyester production processes. First, the monitoring of the on-specification Batch 13, in Figure 25C, shows that it reached the endpoint for both process stages. This can be seen because the  $Q_{red}$  parameter goes down in both stages and AN and V values are on-specification. The second situation can be visualized in Figure 25A, where Batch 10 performed correctly during the first stage but suffered from process upset and did not reach the second stage endpoint. The abnormal behavior in the second phase is seen both observing the high  $Q_{red}$  values and the inappropriate AN and V values, clearly off-specification. Finally, Batch 11 had to be terminated already in the first stage, see Figure 25B, since the abnormal behavior at this stage did not allow pursuing the continuation of the process. It is important to note that this process was formerly monitored atline and, therefore, any abnormal behavior was detected late and derived in the production of high amounts of waste material. Even when the process was developed in a suitable manner, the endpoint of the different stages was detected much later than with the proposed inline monitoring. The new inline procedure for process monitoring and control has implied a decrease in waste and a significant time and energy saving in the normal production tasks.

### **Data fusion strategies for MSPC**

Even though the good performance of MSPC models based on the sole original NIR spectral information was successfully demonstrated using the two process applications described above, MSPC models can work with input information obtained by data fusion. The data fusion-based MSPC models offer two main advantages: a) the joint model of sensor outputs gives a more accurate description of the system of interest since they probe different physicochemical aspects of the processes under study and b) by considering together different outputs, these MSPC models allow 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 impossible out of a fusion scenario.

The need of strategies combining information from spectroscopic sensors with other process variables to build MSPC models is linked to mid-level fusion strategies (Cocchi, 2019), where the balance between the different sources of sensor information is achieved by concatenating the univariate process sensor measurements with compressed versions of the information contained in multivariate spectroscopic

sensors, often expressed as abstract scores. Under the umbrella of mid-level fusion, the novel contribution of this thesis is extending the idea of data fusion for MSPC models to enclose either the sole combination of several model outputs from a single multivariate sensor or the combination of model and sensor outputs of univariate and multivariate sensors in a single data fusion structure. By doing so, different process measurements and modeling tasks for the same process are interconnected. Indeed, model outputs derived from multivariate sensors, such as predictions of quality attributes, process concentration profiles, etc., provide much more specific, diverse and interpretable compressed information than mere abstract scores. As a consequence, they serve more efficiently to diagnose and understand the cause of process malfunctions or off-specification situations in a data fusion context.

The data fusion strategies proposed to build related MSPC models address the different scenarios provided by the three batch processes presented in Chapter 3. To do that, diverse sensor outputs and NIR-based model outputs are properly combined. Each of the data fusion strategies proposed is briefly described below, showing the diversity of possible combinations of multivariate model outputs and process sensor information for the construction of MSPC models.

- a) **Pharmaceutical drying process.** Process endpoint detection is carried out via a data fusion-based MSPC model combining NIR-based multivariate model outputs, such as moisture prediction (%M) and NIR-based MSPC indicators ( $T^2$  and  $Q$ ), with temperature sensors placed at different points of a fluidized bed dryer reactor (see Figure 26A).
- b) **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. The two-step batch process endpoint detection is carried out using data fusion-based MSPC models combining different NIR-based model outputs coming from predictions of key properties (AN and V) and NIR-based MSPC information ( $T^2$  and  $Q$  values), see Figure 26B.
- c) **Distillation process.** In this process, data fusion is carried out based on the combination of compressed NIR-based information, expressed by the concentration profiles linked to the different distilled fractions derived from MCR-ALS, and vapor temperature measurements (see Figure 26C).

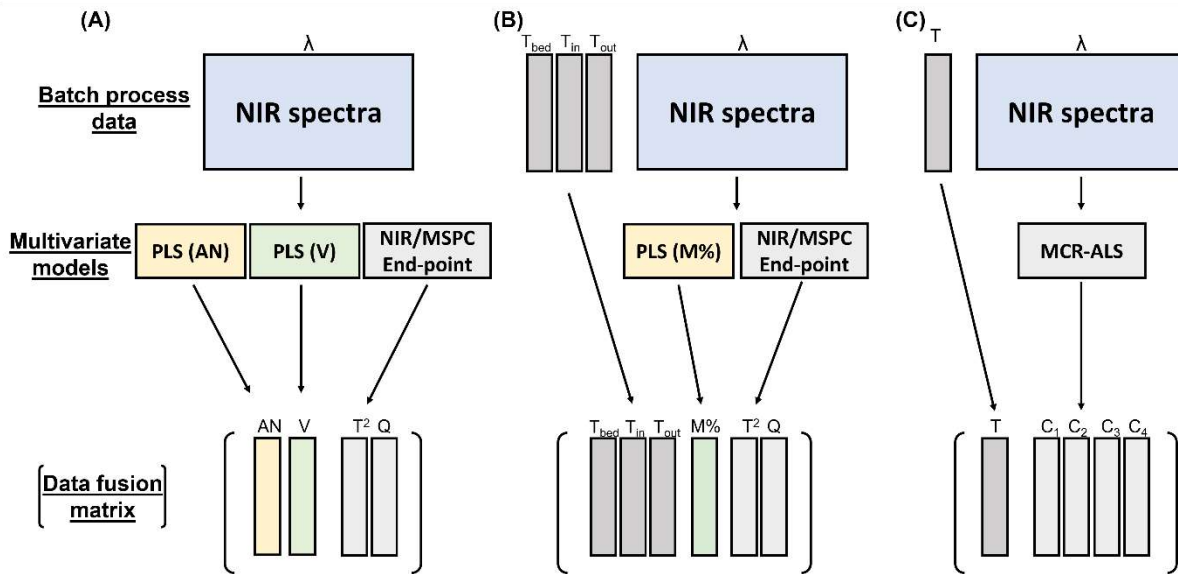


Figure 26 Data fusion strategies used to combine the several sensors and/or model outputs for batches from (A) polyester production process, (B) fluidized bed drying process; and (C) batch distillation process. Reproduced from (de Oliveira et al., 2020).

Figure 26 shows that different kinds of information, issued from the application of different multivariate analysis methods to NIR spectra, can be used as seeding information to build MSPC data fusion models. The output of PLS regression models provides straightforwardly values of key properties of processes. For the FB drying process, the PLS predictions of moisture (M%) were used, whereas for the polyester production process, the predictions of acid number (AN) and viscosity (V) were adopted. In the context of endpoint detection, the MSPC models based on the sole NIR information provide additional parameters linked to the acceptable variation at this process stage. Thus, the  $T^2$  values derived from NOC observations enclose information related to the global acceptable unspecific NIR variation in endpoint observations, whereas the related  $Q$  values express the acceptable residual variation, respectively. These NIR-derived MSPC indicators are afterwards used in data fusion strategies linked to the endpoint detection of FB drying and polyester production processes, as seen in Figures 26A and 26B. Another option to compress the NIR spectra is by using MCR-ALS. For the distillation process described in section 4.2, MCR-ALS provides concentration (distillation) profiles of the different distillation fractions involved in the process that afterwards can be used as input information for online batch MSPC data fusion models. It is important to note that, when MCR-ALS outputs are used, all concentration profiles or only some of them can be introduced in the data fusion MSPC model.

MSPC models based on the proposed data fusion strategies (hereafter DF-MSPC models) are built in the same way as MSPC models built with the sole NIR information (hereafter MSPC<sub>NIR</sub>). However, a DF-MSPC model uses the combined information instead of the raw output from the NIR sensor. For endpoint detection, DF-MSPC models are built as introduced in section 2.3.3, whereas online DF-MSPC models use

the method based on the evolving windows presented in section 4.2. In general, all DF-MSPC models are built using the data fusion 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 DF-MSPC model to check whether they are within the normal operation boundaries or not. It is important to mention that when building DF-MSPC models, column autoscaling should be applied to compensate for the differences in scale of the variables coming from the diverse multivariate model outputs and process sensors.

The three processes presented are very different in nature and so are the data fusion strategies used in the related MSPC models, showing the general applicability of this methodology for diverse PAT applications. However, for brevity, only results related to the application of DF-MSPC models for batch endpoint detection of the FB drying and polyester production processes presented in this section are presented. For the results of the distillation process, please refer to Publication IV (de Oliveira et al., 2020).

Figure 27 compares the Q-residuals control charts of DF-MSPC models and  $MSPC_{NIR}$  models for the FB drying (Figure 27A) and polyester production processes (Figure 27B) using different validation batches. For a better comparison, Figure 27 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. In general, the results obtained clearly show that the use of information coming from different models (polyester process) and sensor outputs (FB drying) in DF-MSPC models overcomes the performance of the control procedures based on single sensor information,  $MSPC_{NIR}$ .

For both processes,  $Q_{red,DF}$  values, coming from the DF-MSPC model, generally diagnose significantly better off-specification observations than  $Q_{red,NIR}$ , obtained using only NIR spectral information. This is observed during the last observations of faulty batches, inset plots in Figure 27A (right) and Figure 27B (top right), where  $Q_{red,DF}$  values are significantly higher and clearly further from the control limit than  $Q_{red,NIR}$  values. In the case of the polyester production process, the explicit inclusion of the predictions of product quality parameters AN and V instead of using only general unspecific NIR information helps in this purpose. Besides, the remaining unspecific NIR information is enclosed in the  $T^2$  and Q parameters coming from the  $MSPC_{NIR}$  model. In the case of the FB drying process, the temperature profiles and the explicit use of the moisture PLS prediction explain the improvement. Moreover, for on-specification batches, the use of DF-MSPC models also provide a much clearer difference between the  $Q_{red,DF}$  values before and after crossing the control limit than the  $Q_{red,NIR}$  values, being the decrease of the  $Q_{red,DF}$  curve always much steeper than that of  $Q_{red,NIR}$ . See inset plots for the endpoint from drying Batch 5 in Figure 27A (left) and the two-step endpoints related to polyester batch 10 in Figure 27B (bottom left).

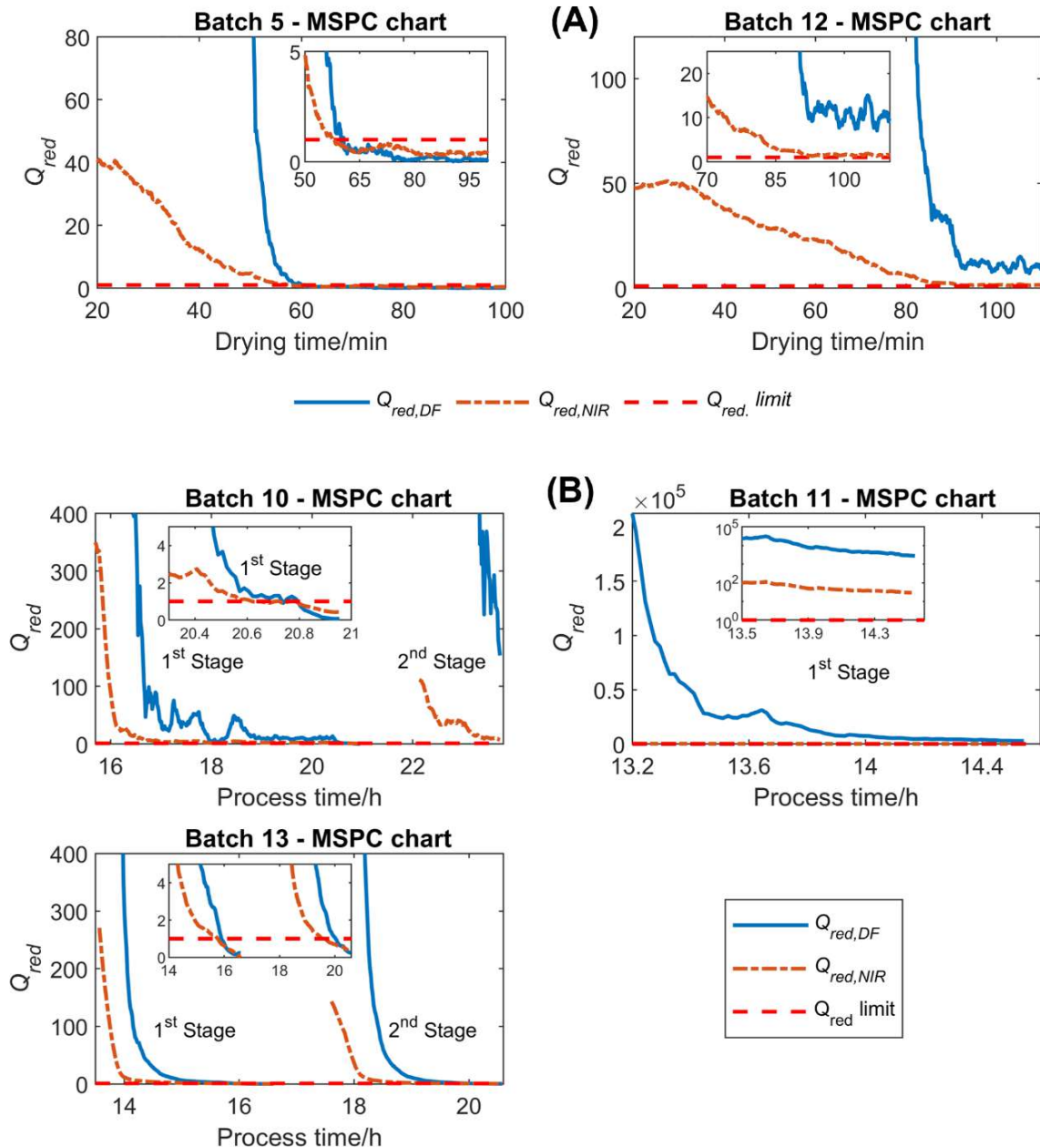


Figure 27 Reduced  $Q$  ( $Q_{red}$ ) MSPC charts for endpoint detection using DF-MSPC model,  $Q_{red,DF}$  (solid blue curve) and MSPC model based on the sole NIR information,  $Q_{red,NIR}$  (dash-dotted orange curve) for validation batches from (A) FB drying process, on-specification Batch 5 (left) and off-specification Batch 12 (right); and (B) polyester production process, Batch 10 (top left), Batch 11 (top right) and Batch 13 (bottom left). 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 process. Reproduced from (de Oliveira et al., 2020).

Another added value of using information coming from different models and/or sensor outputs in DF-MSPC models is the easier interpretation of the contribution plots linked to abnormal observations, helpful to understand the causes related to process faults and off-specification situations. Figure 28 shows the residual contribution plot (bar plot in orange) related to an abnormal observation at the end of the off-specification drying batch 12. To compare with the residual level of an on-specification observation, the contribution plot related to a NOC observation, i.e. after reaching the endpoint, of batch



5 is shown as well (bar plot in blue). As can be observed, the main contributions to the high  $Q_{red,DF}$  values ( $Q_{red,NIR}$  and high %MPLS), of batch 12 are meaningful process information easier to interpret than looking at residuals related to spectral wavelengths or PCA scores. Indeed, it is clear that the moisture in the off-specification of batch 12 is much higher than it should, seen by the high positive residual of the MPLS %. Besides, the overall abnormal spectrum obtained is also seen through the large value of the residual  $Q_{red,NIR}$  for this observation.

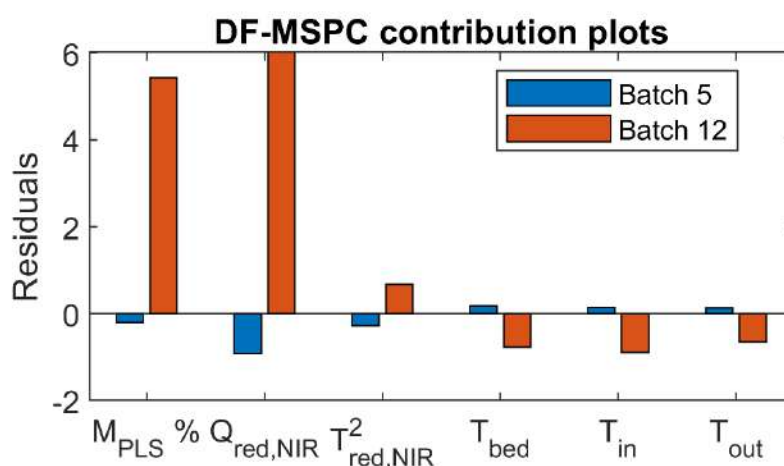


Figure 28 Residual contribution plots related to the FB drying observations at 95 min for on-specification batch 5 and 100 min for off-specification batch 12 evaluated with the DF- MSPC model for drying endpoint detection. Reproduced from (de Oliveira et al., 2020).

In both process examples, the two types of MSPC models (DF or sole NIR spectra) have shown satisfactory performance for the detection of on- and off-specification situations. However, endpoint 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 by using compressed and interpretable NIR information, such as key properties and process evolution indicators. The fact of using compressed interpretable NIR information is of great help also to understand the underlying causes of process malfunctions. In difference with the use of mere PCA scores that compress all the relevant information in the NIR acquired spectra, the use of more specific model outputs can also help to develop tailored DF-MSPC models where only particular key parameters or concentration profiles are taken into account, i.e., where only the part of NIR information linked to the target of interest is considered.

### **4.3.2 Online synchronization-free MSPC for batch process evolution assessment**

This section introduces a new batch synchronization-free methodology proposed to build online MSPC for tracking batch process evolution. The methodology is based on a first step related to the process modeling of non-synchronized NOC batch trajectories using PCA followed by the construction of local MSPC models covering the global NOC batch trajectory. The last step is the use of local MSPC charts to test in real-time whether the new batch observations are following or not the NOC trajectory in real-time. In this section, the methodology is demonstrated using the FB drying process presented in Chapter 3. This methodology is presented in Publication V where it is tested in two batch processes showing the flexibility of the methodology to deal with non and synchronized batch data.

**Publication V.** Rocha de Oliveira, R. and de Juan, A. **Synchronization-Free Multivariate Statistical Process Control for Online Monitoring of Batch Process Evolution**, *Submitted to Frontiers in Analytical Science; Specialty section: Chemometrics; Research Topic: Novel Applications of Chemometrics in Analytical Chemistry and Chemical Process Industry.*



# SYNCHRONIZATION-FREE MULTIVARIATE STATISTICAL PROCESS CONTROL FOR ONLINE MONITORING OF BATCH PROCESS EVOLUTION

Rodrigo Rocha de Oliveira\*, Anna de Juan\*

Chemometrics Group, Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, Diagonal 645, 08028, Barcelona, Spain

**\* Correspondence:**

R. Rocha de Oliveira and A. de Juan

\*rodrigo.rocha@ub.edu; anna.dejuan@ub.edu

**Keywords:** Batch process; online process monitoring; statistical process control; synchronization-free MSPC; local MSPC modeling.

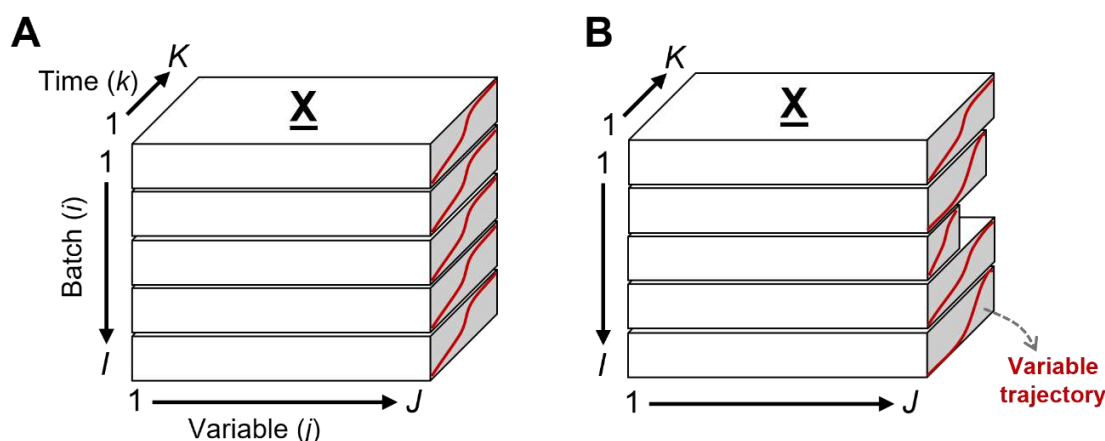
**Abstract**

Synchronization of variable trajectories from batch process data is a delicate operation that can induce artifacts in the definition of multivariate statistical process control (MSPC) models for real-time monitoring of batch processes. The current paper introduces a new synchronization-free approach for online batch MSPC. This approach is based on the use of local MSPC models that cover a NOC trajectory defined from principal component analysis (PCA) modeling of non-synchronized historical batches. The rationale behind is that, although non-synchronized NOC batches are used, an overall NOC trajectory with a consistent evolution pattern can be described, even if batch-to-batch natural delays and differences between process starting and end points exist. Afterwards, the local MSPC models are used to monitor the evolution of new batches and derive the related MSPC chart. During the real-time monitoring of a new batch, this strategy allows testing whether every new observation is following or not the NOC trajectory. For a NOC observation, an additional indication of the batch process progress is provided based on the identification of the local MSPC model that provides the lowest residuals. When an observation deviates from the NOC behavior, contribution plots based on the projection of the observation to the best local MSPC model identified in the last NOC observation are used to diagnose the variables related to the fault. This methodology is illustrated using two real examples of NIR-monitored batch processes: a fluidized bed drying process and a batch distillation of gasoline blends with ethanol.

## 1 Introduction

Industrial sectors often rely on batch processes to produce their intermediate or final products. Batch processes consist of cyclic repetitions of an established recipe aiming at the production of products meeting specific quality specifications. They are also characterized by complex, dynamic and nonstationary behavior. Thus, monitoring a batch evolution in real-time is a challenging, but essential action to obtain end products with desired quality, reducing costs and increasing process understanding. (Rato and Reis, 2020; Rendall et al., 2019; van Sprang et al., 2002).

Nowadays, with the emergence of Industry 4.0, batch processes are monitored not only with typical process sensors, e.g. temperature, pressure, flow, etc, but also with advanced sensors probes based on spectroscopic techniques such as near-infrared (NIR), mid-infrared, and Raman (Avila et al., 2021; Avila et al., 2012; Besenhard et al., 2018; Cimander and Mandenius, 2004; Grassi et al., 2019; Pöllänen et al., 2006). The collection and use of process sensor measurements from historical batches that followed the normal operating conditions (NOC) and reached the targeted product specifications is the basis for the development of multivariate statistical process control (MSPC) models and related charts, ready to be used to test the evolution of new batches (Ferrer-Riquelme, 2010; Kourti, 2005; Nomikos and MacGregor, 1995; Wold et al., 2009). Off-line MSPC charts can be used to diagnose the root cause of a disturbance from a finished faulty batch. However, it is even more important the on-line use of MSPC charts for real-time monitoring of batch evolution to enable taking quick action in case of detection of process disturbances.



**Figure 1** Three-dimensional data array,  $\underline{X}$ , with aligned NOC batch data (A) and uneven and not synchronized batch data (B).

Process data measurements from a single batch consist of the collection of several variables,  $J$ , (process data and/or spectroscopic measurements) at different process points throughout the batch,  $K_i$ . These measurements are usually organized in a data matrix,  $\mathbf{X}_i$ , with dimension  $(K_i \times J)$  to be used for process monitoring and/or control purposes. Most data-driven modeling strategies aiming at building online MSPC charts to monitor process evolution require that data from several NOC batches,  $I$ , that have the same batch length, i.e. batch data matrices with the same numbers of rows  $K$ , and follow the same and synchronized process dynamics. When this happens, the data can be arranged in a three-dimensional data array,  $\underline{X}$ , with dimensions  $I \times K \times J$ , (Figure 1A). Most of the MSPC models are built based on data-driven multivariate analysis methods, such as principal component analysis (PCA) and partial least squares (PLS); for this purpose, different unfolding strategies of the  $\underline{X}$  array can be used according

---

**SYNCHRONIZATION-FREE ONLINE BATCH MSPC**

to the modeling approach used (Nomikos and MacGregor, 1995; Wold et al., 1998). However, because of the inherent batch process complexity and nonstationary behavior, the batch duration,  $K_i$ , is not always the same and, equally relevant, key process events do not occur at the same time point when comparing different NOC batch runs of the same process. This uneven and not synchronized batch data (Figure 1B) cannot be represented in this perfect three-dimensional data array,  $\underline{X}$ , unless adjusted using different batch synchronization tools to cope with this problem (González-Martínez et al., 2014b).

Great progress has been made to develop strategies for batch alignment based on a maturity index coming directly from a process variable or estimated by PLS models or using more advanced algorithms, such as correlation optimized warping or dynamic time warping (González-Martínez et al., 2014a; Kassidas et al., 1998; Liu et al., 2017; Ramaker et al., 2004; Zhao et al., 2020). Most of these methods were designed for the monitoring of finished batches using offline MSPC models and only an attempt proposed by (González-Martínez et al., 2011) described a method based on time warping that allows batch alignment for online MSPC.

Despite the methodologies mentioned above, having naturally non-synchronized batches is the most common situation in practice and batch alignment is a delicate operation that can induce artifacts in the definition of MSPC models when scarce information is available or when is not properly applied. Hence, the need for MSPC approaches that can circumvent the synchronization step for online process monitoring and control. Very few attempts have been carried out in this direction. (Rato et al., 2017) used the translation-invariant wavelet decomposition and PCA for the monitoring of the semiconductor manufacturing process. Another method based on a search grid capturing the batch trajectory in the PCA score space was proposed by (Westad et al., 2015) and was used for the monitoring of two industrial processes.

In this paper, a new synchronization-free approach of multivariate statistical process control (MSPC) for online monitoring and diagnostics of batch processes is introduced. It is based on the modeling of an overall NOC historical batch trajectory, defined by individual non-synchronized NOC batches, and the subsequent construction of derived PCA-based local MSPC models covering the complete process, i.e., the complete overall NOC batch trajectory. These local models are used to identify whether new batch observations are inside the NOC trajectory and, when this is the case, to provide an estimate of the process progress. The approach is illustrated using two real examples of NIR-monitored batch processes but is readily applicable for the online monitoring of batch processes of different typologies monitored by one or more diverse sensors.

## 2 Process case studies and data sets

Two case studies from previous works are used to illustrate and test the online batch MSPC models for tracking process trajectories. A brief experimental description of these NIR-monitored processes with the related spectral preprocessing implemented is presented below.

### Process 1: Fluidized bed drying of pharmaceutical granules

Batches of 500-g pharmaceutical wet granules (dry mass fraction of mannitol > 50% and excipients) were dried in a 4-L fluidized bed (FB) (4M8-Trix Formatrix, ProCepT, Belgium). The FB air inlet flow was controlled at 0.6 or 0.85 m<sup>3</sup>/min and a temperature range from 22 to 30 °C. In-line NIR measurements were collected approximately every second using a spectrophotometer with a 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). The spectra covered a wavelength

range from 1750 to 2150 nm at 1-nm intervals. For each batch, off-line reference moisture content analysis was carried out using a thermogravimetric moisture analyzer (MB120, Ohaus, Germany) from samples retrieved at 6-min intervals to detect drying endpoint (moisture < 2%). Because of different process conditions at the beginning and during each batch run, such as inlet air temperature and flow, different batch durations were required for each trial to reach the defined <2% moisture level, therefore, providing data matrices with uneven lengths. Faulty batches used in the testing of the proposed approach did not reach this moisture level. Suitable preprocessing was employed to filter out noise and baseline fluctuations on the NIR raw data observations before data analysis. The preprocessing steps included the application of a moving average of consecutive NIR observations followed by standard normal variate (SNV) normalization. For a detailed description of the experimental procedure and the visualization of the spectral data, the reader is referred to (Avila et al., 2020; Rocha de Oliveira et al., 2020). Some batches were selected from the previous work and additional faulty batches were used for model validation. Ten NOC batches, NOC1 to NOC10, were used for MSPC model building, and three for validation (one NOC, BN1, and two faulty batches, BF1 and BF2). This is an example of a batch process where the evolution of drying in time is not synchronized among batches since the initial and final material in every batch does not necessarily have the same moisture level.

### Process 2: Automated benchtop batch gasoline distillation

Batches of 100-mL gasoline blends (mixture of pure gasoline and ethanol) were distilled in an automated batch distillation device designed for the in-line monitoring of distilled product with NIR spectroscopy. For every batch, vapor temperature readings and in-line FT-NIR absorption spectra (900 to 2600 nm with 4 cm<sup>-1</sup> resolution; Rocket, ARCoTix ANIR, Switzerland) were recorded for every unit of percentage distilled mass fraction of initial sample weight, in the 5 to 90% range. Therefore, the data matrices obtained had the same number of NIR observations per batch (86 NIR spectra) and every observation was related to the same distillation process stage, as defined by the percentage (w/w) of distilled sample mass. The gasoline batches were prepared by mixing ethanol AR (99% Sigma-Aldrich) and pure gasoline (from Petrobras refinery, Brazil) at different volume ratios from 10 to 40 %. Distillation batches of gasoline blends with 27% ethanol were defined as NOC batches and all batches with a different ratio as faulty, or out of specification according to Brazilian legislation. The preprocessing steps used in this data set were Savitzky-Golay derivative (1<sup>st</sup>-order derivative, 2<sup>nd</sup>-order polynomial function and 9-point window) for baseline correction followed by spectral normalization to mitigate signal intensity fluctuations of the NIR spectra. More detailed information related to the experiments and spectra preprocessing can be found elsewhere (de Oliveira et al., 2017). In this work, nine NOC distillation batches were used to build the MSPC control charts for tracking process trajectory (B1 to B3, B5 to B9 and B11), and three for validation, where one was NOC (B4) and two were faulty batches (B13, B19). In this case, batch process trajectories were synchronized because the percentage of distillation weight gives a direct reference for batch progress evolution.

### 3 Data treatment

The online batch MSPC model building procedure for tracking process evolution in synchronized or non-synchronized batch processes is described below. The complete methodology involves the following steps:

- a) *Modeling of NOC batch process trajectories.*
- b) *Construction of local MSPC models based on NOC batch process trajectories.*
- c) *Use of an MSPC chart based on local MSPC models to track the evolution of new batches.*

## SYNCHRONIZATION-FREE ONLINE BATCH MSPC

The first two steps are involved in the generation of the MSPC models, whereas the last step involves the use of the local MSPC models on new batches to test whether they follow the NOC trajectory or to detect faults. A detailed description of each step is presented below together with a visual description of the approach in Figure 2.

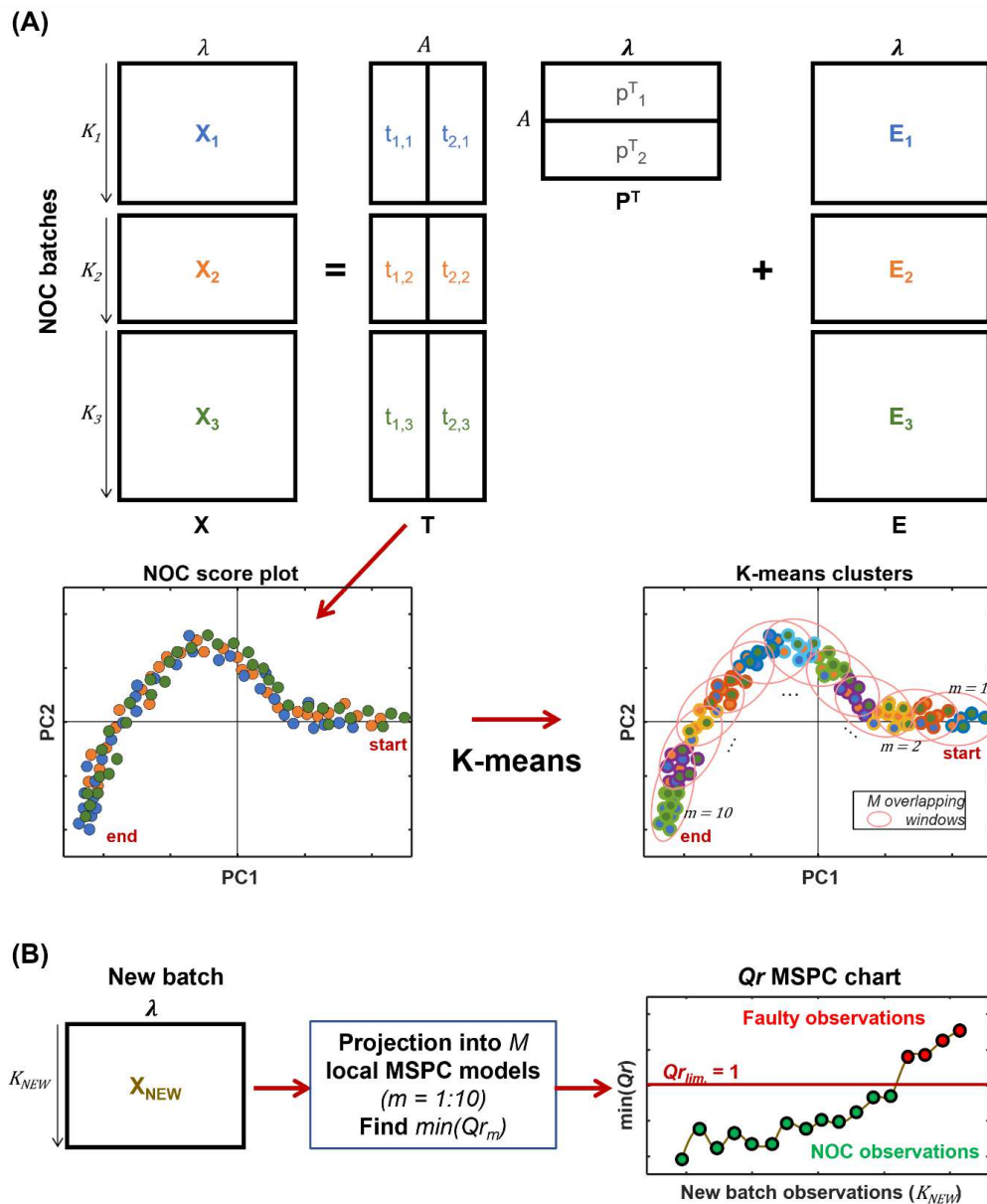


Figure 2 Illustrative description of the different steps involved in the implementation of the local MSPC models for online monitoring of batch evolution. (A) PCA modeling of original batch process data for several NOC batches, visualization of process trajectories in the scatter score plot and definition of local regions in the NOC trajectory using k-means. (B) Monitoring of the evolution of a new batch using the projection of each observation onto the local MSPC models. The related reduced  $Q$ -statistics control chart is obtained plotting the minimum  $Q_r$  value obtained in all  $M$  model projections per each observation.



*a) Modeling of NOC batch process trajectories*

The evolution of NOC batches, a.k.a “golden batches”, can be defined using different multivariate analysis modeling strategies, such as PCA, ICA, MCR, PARAFAC, etc. (Bogomolov, 2011; de Oliveira et al., 2017; Gomes et al., 2019; Haack et al., 2004; Mortensen and Bro, 2006; Skibsted et al., 2006). In this work, we use PCA as the basis to define the general NOC batch process trajectory.

The NIR spectra obtained in a NOC batch  $i$  are structured in a data matrix  $\mathbf{X}_i(K_i \times J)$ , where  $K_i$  are the number of spectra collected (related to time points for *Process 1* and to % of distillation for *Process 2*) and  $J$  are the NIR channels per spectrum.

When several NOC batches are used to define the general process trajectory, the data matrices from the different NOC batches,  $\mathbf{X}_i(K_i \times J)$ , are placed one on top of each other to build an augmented multiset structure  $\mathbf{X}(N \times J)$ , where  $N$  is the number of rows related to the total number of observations from the  $I$  NOC batches, that is,  $N = \sum_i^I K_i$ . Note that this strategy does not require resizing or synchronization of uneven batch lengths, since the only requirement is that all batches share a common spectral dimension,  $J$  (Wold et al., 1998). The next step is to column mean-center this multi-batch structure and analyze with PCA. Note that this centering operation does not remove the mean trajectory of the batches in time.

Principal component analysis (PCA) is used to obtain a global model of batch trajectories explaining the overall NOC process evolution. PCA is used to reduce the dimensionality of the preprocessed spectral data into a low-dimensional subspace of principal components (PC's), orthogonal among them, that preserve the relevant information of the original data and explain the maximum non-random variance (Jolliffe, 2002). The PCA model for the augmented process data matrix  $\mathbf{X}(N \times J)$  is expressed as in eq. (1),

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (1)$$

where  $\mathbf{T}(N \times A)$  is formed by the scores matrix, related to the observations of the batch process data,  $\mathbf{P}^T(A \times J)$  is the loadings matrix, related to the importance of the NIR variables in the description of the  $A$  PC's and  $\mathbf{E}(N \times J)$  is the residual matrix after modeling. The number of principal components of the model,  $A$ , can be found using a suitable cross-validation method. The loading matrix,  $\mathbf{P}^T$ , is common to all batches and the augmented score matrix,  $\mathbf{T}$ , accommodates  $\mathbf{T}_i$  blocks, related to every batch, that can be formed by a different number of observations,  $K_i$ . The multiset structure for three NOC batches and the related PCA model is illustrated in Figure 2A (top left), where  $\lambda$  represents the  $J$  spectral channels of the NIR spectra.

*b) Construction of local MSPC models based on NOC batch process trajectories.*

From the augmented score matrix of all NOC batches, individual batch score trajectories can be overlapped on a scatter score plot, as shown in Figure 2A (bottom left). The dots represent the scores for each observation and are colored according to the NOC batches used in the PCA model. Note that the overall trajectory evolution is the same for all NOC batches, but in a general non-synchronized case, the starting and endpoint of every batch do not need to coincide. The overlapped individual batch process trajectories define a global description of the variability of the NOC process evolution, helpful to observe whether a new batch process evolves as NOC batches or not, independently from the batch length and dynamics. The evolution described by the overlapped NOC trajectories can be divided into

---

**SYNCHRONIZATION-FREE ONLINE BATCH MSPC**

a sufficient number of  $C$  local regions using a cluster analysis methodology, such as K-means. Figure 2A (bottom right) illustrates these local regions for  $C = 11$ , as indicated by the outer circle color of the neighbor observations inside each cluster. The seeding information for the local MSPC models is formed by the observations in two consecutive clusters. Therefore, the first local MSPC model contains the observations in the first two clusters of the process trajectory, the second local MSPC model uses the observations in clusters two and three and so forth until all the NOC process trajectory is covered. The observations used in consecutive local MSPC models overlap with each other so that all process trajectory regions are covered. As can be seen in Figure 2A, for a k-means analysis providing 11 clusters, 10 local MSPC models with overlapping information as defined by the red ellipses can be built.

The local MSPC models are built based on PCA and control chart limits are defined using the suitable local model statistics. The operational procedure to build each local MSPC model can be described as follows. First, the original observations, i.e. NIR spectra, for each local model are placed into a data matrix  $\mathbf{X}_m(K_m \times J)$ , where  $m$  indicates the index of the local model (from 1 to  $M$ ) and  $K_m$  is the number of observations used to build the model. Then, this matrix is mean-centered and modeled with PCA, as in eq. (1), generating the matrices of scores  $\mathbf{T}_m(K_m \times A_m)$ , loadings  $\mathbf{P}_m^T(A_m \times J)$ , and residuals  $\mathbf{E}_m(K_m \times J)$ . Note that the mean-centering step is performed using the mean of the matrix  $\mathbf{X}_m$  and not the global mean of the multibatch structure. Enough PC's,  $A_m$ , are included in each local model to provide the best fit using cross-validation (Wold, 1978). Finally, the control limits of the local control charts can be derived using the residuals and the scores from the local PCA model (Aguado et al., 2007; Rännar et al., 1998; Wold et al., 1998). In this work, the controls charts are based only on the residual matrix,  $\mathbf{E}_m$ , deriving the Q-statistic control chart limit,  $Q_{lim}$ ; however, other statistical parameters can readily be used to track the process evolution. The  $Q_{lim}$  is calculated according to the equation proposed by (Jackson and Mudholkar, 1979). Thus, once the local MSPC models and their related multivariate control charts limits are set, the online process evolution of new batches can be tracked based on the local models defined.

*c) Use of an MSPC chart based on local MSPC models to track new batch evolution*

**Calculation of squared residuals statistics ( $Q$ ).** For online batch monitoring of new batch observations ( $\mathbf{X}_{NEW}$  in Figure 2B), every new observation is projected onto all local MSPC models and a set of related sum of squared residuals statistics,  $Q_{k,m}$ , are obtained as shown in Figure 2B. Thus, for every new online observation,  $\mathbf{x}_k$  (a NIR spectrum in  $\mathbf{X}_{NEW}$ ), its scores values,  $\mathbf{t}_{k,m}$ , are obtained for each local MSPC model using its related PCA loadings,  $\mathbf{P}_m$ , as follows,

$$\mathbf{t}_{k,m} = \mathbf{x}_k \mathbf{P}_m \quad (2)$$

Then, the residuals for the new observation in each local model are obtained as,

$$\mathbf{e}_{k,m} = \mathbf{x}_k - \mathbf{t}_{k,m} \mathbf{P}_m^T \quad (3)$$

And the related  $Q_{k,m}$  as:

$$Q_{k,m} = \mathbf{e}_{k,m} \mathbf{e}_{k,m}^T \quad (4)$$

For an easier interpretation of the global multivariate control chart obtained from the outputs of the local MSPC models, reduced  $Q$ -statistics,  $Qr_{k,m}$ , are calculated by dividing the obtained  $Q_{k,m}$  values by the related local model  $Q_{lim}$ . Thus, the control limits for all local MSPC models become equal to one,  $Qr_{lim} = 1$ . The reduced  $Q$  values for every new observation,  $Qr_{k,m}$ , are checked to see whether they are above or below the  $Qr_{lim}$ . If all  $Qr_{k,m}$  values for the observation  $k$  are large and above one, this observation is diagnosed as faulty, and it is an indicator that the process is deviating from the NOC trajectory. Conversely, if one or more  $Qr_{k,m}$  values are below the control limit, the observation follows the NOC trajectory. An easy way to visualize the diagnostic of every new observation by using a single  $Q$  chart is shown in Figure 2B (bottom right), where only the minimum  $Qr$  parameter after the projection in all local models is displayed for every new observation. Observations that follow the NOC trajectory are depicted by the green dots below the  $Qr_{lim} = 1$ , and the eventual deviations from it, with  $\min(Qr_{k,m}) > 1$ , in red. To assess the spectral variables making the greatest contributions to the deviation in  $Q$  we can display the  $Q$ -statistics contribution plots for the sought observation by plotting the elements of the residual vector,  $\mathbf{e}_{k,m}$ . The residuals used for the contribution plots are calculated using the best local MSPC model related to the last NOC observation.

For NOC observations, it is also possible to estimate the process stage of every observation by identifying the local MSPC model providing the lowest  $Qr_{k,m}$  value. This visualization approach will be provided for the real process applications studied in this work in the next section.

## 4 Results and discussion

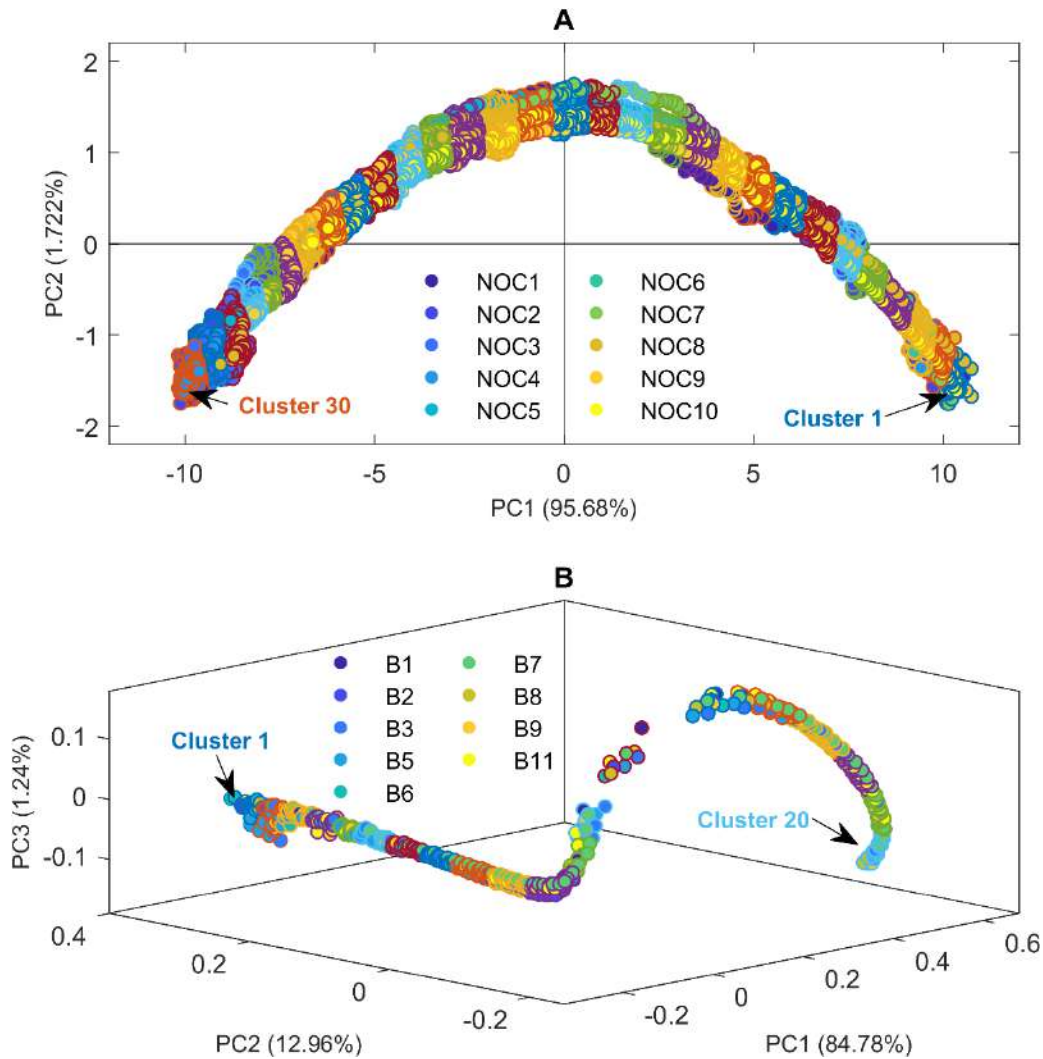
In this section, the results related to the construction of NOC trajectories and local MSPC models for each process case study are shown. Afterwards, the resulting MSPC charts for the online monitoring of new NOC and faulty batches are shown for each process. Complementary visualization of MSPC charts and fault diagnostics based on contribution plots are also presented.

### 4.1 Construction of NOC trajectories and local MSPC models

The construction of PCA-based NOC trajectories for each process was calculated as explained in *step a* of the Data Treatment section using the training dataset, i.e. all NIR observations from selected complete NOC batches. This step was followed by k-means analysis on the overlapped individual NOC batch trajectories to define the clusters used to build the local MSPC models covering the overall NOC process trajectory (Data Treatment section *step b*). Figure 3 shows the PCA score scatter plot and the k-means clusters used to build the local MSPC models describing the overall NOC batch process trajectories for the drying (*Process 1*) and the distillation processes (*Process 2*).

Principal Component Analysis of the NOC batches from *Process 1* (Fluidized bed drying) allowed description of the process evolution using only two PC's explaining a total of 97.61% of the data variance, as shown in the score plot of Figure 3A. The score plot described mostly the variation of the moisture content with the drying evolution from beginning to end of every NOC batch. Note that, because each batch had different initial and final moisture conditions, they started and finished at different points of the overall NOC trajectory; however, all individual batch trajectories followed the same evolution pattern, as shown in the PCA score plot. Once the overall NOC trajectory was defined, the k-means analysis allowed the identification of 30 clusters along this trajectory, as displayed by the different outer circle colors associated with the observations inside each cluster in Figure 3A. After that, a number indicating the process stage evolution was automatically assigned to each cluster according to the position in the overall NOC trajectory.

## SYNCHRONIZATION-FREE ONLINE BATCH MSPC



**Figure 3** PCA score plot for the online NIR observations showing the NOC batch process trajectories and local clusters found by k-means for (A) *Process 1*, fluidized bed drying, and (B) *Process 2*, gasoline blend distillation. The inner part of the circles is colored according to the related NOC batch, whereas the outer part reflects the observations included in every cluster and, hence, in the related local MSPC model.

For *Process 2* (Distillation), three components were required by PCA to explain 98.99% of NOC batches variance because of the complex gasoline sample and the continuous variation of the distilled material composition. The complex overall NOC trajectory associated with the distillation process is shown in the 3 PC score scatter plot in Figure 3B. Despite the higher complexity of the overall NOC trajectory linked to the distillation process, all individual batches trajectories followed the same evolution pattern with good reproducibility. In contrast to the drying process, the NIR observations of the distillation process were acquired at specific percentages of distillation weight; therefore, the observations were naturally synchronized according to the process evolution. Note that all batches started and finished at the same point of the overall NOC batch trajectory in the score plot. The k-means algorithm applied on the PCA scores of Figure 3B identified 20 clusters along the overall NOC batch trajectory, displayed in Figure 3B. The number of clusters is lower than in the previous example because of the limited number of available observations per batch run (only 86) and the need to avoid having clusters with a very low number of observations to build the local MSPC models.

Once the overall NOC batch process trajectories were defined for each process case, the original NIR observations inside the suitable two consecutive k-means clusters were used as seeding information to build local MSPC models for each step of the batch trajectory, as described in the Data Treatment section (*step b*). Thus, a total of 29 and 19 local PCA-based MSPC models were built for *Processes 1* and *2*, respectively. Local MSPC control chart limits based on the  $Q$ -statistics with a 99% confidence interval were calculated for each local MSPC model to be used for the online tracking of new batches evolution, as shown in the next subsection.

## 4.2 Online tracking of new batch evolution with local MSPC models

The results of the use of local MSPC models for the online tracking of new batch evolution are described separately for each process case, as shown below. The new batches used were identified in previous studies as NOC or faulty; therefore, they will be useful to demonstrate and validate the proposed methodology.

### 4.2.1 Application to Process 1 (FB Drying)

The tracking of every observation in new fluidized bed drying batches was performed as described in the Data treatment section (*step c*), using the 29 local MSPC models built as explained above (Figure S1 and a related animation S2 in the supplementary material help to display how the  $Q_r$  values issued from every MSPC local model are obtained for every observation in a batch).

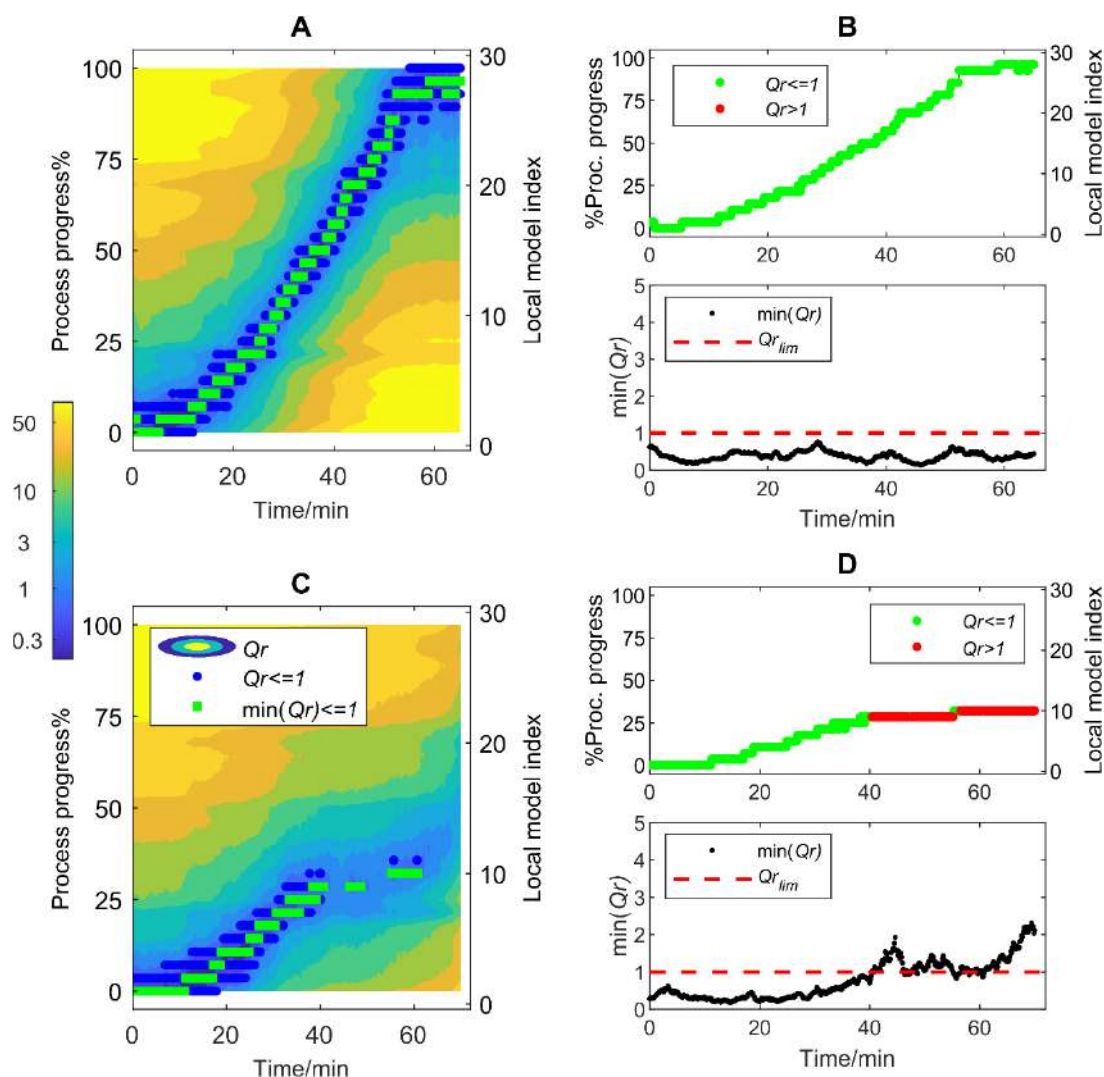
The  $Q_r$ -based MSPC control charts for the online tracking of observations in two drying batches are shown in Figure 4. Figures 4A and 4C are contour plots related to NOC batch BN1 and faulty batch BF1, respectively, that show all the  $Q_r$  values calculated after the projection of each online NIR observation of the batch onto all local MSPC models. A log-scale colormap has been used to highlight the differences at low  $Q_r$  values. The horizontal axis of the contour plot represents the batch time at which every observation was collected and the right vertical axis the indices related to the local MSPC model used to describe the *Process 1* NOC batch trajectory, i.e. from 1 to 29. Additionally, in the left vertical axis, each local MSPC model index is associated with a percentage of the process progress from 0-100%, defined making a linear scaling that links the initial local model to 0% process progress and the final local model to 100% process progress. The process progress in this approach plays the same role as the process maturity concept proposed by other authors (Westad et al., 2015; Wold et al., 1998).

Thus, to track the behavior of an observation of a new batch, their related  $Q_r$  values (associated with a specific process time) are examined. In the contour plots in Figures 4A and 4C, the  $Q_r$  values below the control limit, i.e.  $Q_r < 1$ , are depicted as blue dots and the  $\min(Q_r < 1)$  for every observation in green. If an observation shows a NOC behavior (as all do in Figure 4A related to batch BN1), there will always be one or more  $Q_r$  values below 1; i.e., all observations will show one or more blue dots and a green dot. Instead, when an observation deviates from the NOC trajectory, as in batch BF1 (Figure 4C), all  $Q_r$  values related to that observation are above the control limit of 1 and neither blue nor green dots are observed.

To facilitate the interpretation and summarize the relevant information of the results in the contour plots, graphics displaying the  $\min(Q_r)$  value and the related process progress for every batch observation are proposed (see Figures 4B and 4D for batches BN1 and BF1, respectively). Figure 4B shows that all observations for batch BN1 followed the NOC batch trajectory, seen because all  $\min(Q_r)$  values were below the control limit of 1 (bottom panel), and that the process progress covered the complete range (0-100%) (top panel). Figure 4D shows that batch BF1 deviated from the NOC

## SYNCHRONIZATION-FREE ONLINE BATCH MSPC

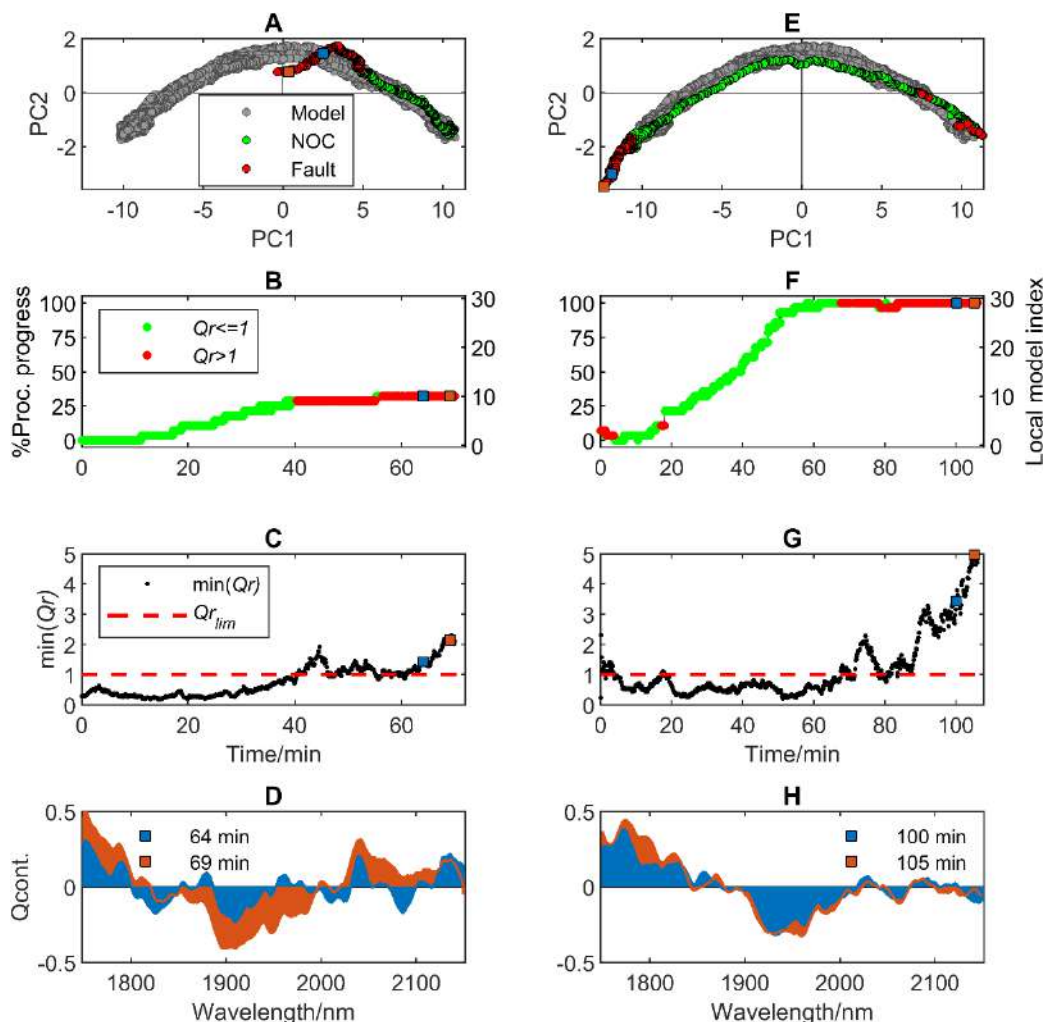
trajectory after approximately 40 min of batch time as flagged by the  $Q_r$  above the local MSPC control limits ( $\min(Q_r) > 1$ ) (bottom plot). When a fault happens, the related observations are displayed in red in the process progress plot to indicate that the evolution of the process is abnormal (top plot).



**Figure 4**  $Q_r$ -based MSPC charts for FB drying NOC batch BN1 (A and B) and faulty batch BF1 (C and D). (A and C) Contour plots of the  $Q_r$  values calculated after the projection of each NIR observation onto the local MSPC models. Blue dots show values of  $Q_r < 1$  (control limit), green squares the  $\min(Q_r) < 1$ . (B and D) Charts show the  $\min(Q_r)$  value (bottom panel) and the related process progress associated with it (top panel) for every batch observation. In the process progress plot, NOC observations are displayed in green and faulty observations in red.

Detailed results and interpretation of the abnormal behavior for the online tracking of two faulty batches, BF1 and BF2, are shown in Figure 5 (left and right plots, respectively). Figures 5A and 5E show the deviations of the two batches by displaying the score plot projections of NIR observations of these new batches onto the global PCA model used to describe the NOC batch trajectory. The score plot shows all training NOC batch trajectories as gray dots whereas the NOC observations from the new batches are overlaid as green dots when identified as NOC and as red dots when faulty. Figures 5B to 5G show the batch process progress and  $\min(Q_r)$  MSPC chart for the tracking of the online observations, where the abnormal observations are associated with  $\min(Q_r)$  values higher than 1 and

flagged in red color in the process progress plot. Moreover,  $Q$  contribution plots from two faulty observations selected from each batch are shown in Figures 5D and 5H. The contribution plots were used to understand the reasons for the deviations from the NOC batch trajectory, as described below for each batch.



**Figure 5** Results for the online tracking of new batch evolution using the local MSPC for faulty batches BF1 (A to D) and BF2 (E to H). (A and E) PCA score plot showing the NOC trajectory (gray dots) and new batch trajectory in green (NOC observations) and red dots (faulty observations). (B and F) MSPC chart showing the process progress. (C and G)  $Qr$ -based MSPC charts. (D and H) are the  $Q$  contribution ( $Qcont.$ ) plots for two faulty observations selected for each batch and represented by the blue and orange squares in the MSPC control charts.

The deviation of batch BF1 from the NOC trajectory was detected after approximately 40 min of batch time, see Figures 5B and 5C. Although in Figure 5A the faulty observations (red dots) right after 40 min were still close to the NOC trajectory, the related  $\min(Qr)$  after projection onto local MSPC models was above the control limit indicating a deviation, which became even larger after ca. 65 min of batch time, see Figure 5C. To help to diagnose this deviation, contribution plots are shown in Figure 5D for two faulty observations selected at 64 and 69 min of batch time. These observations are marked in blue and orange squares in the score plot and MSPC charts. The  $Q$  contribution plots show that the absorption bands that gave higher contributions to  $Q$  were around 1750 and 1900 nm related to the 1<sup>st</sup> overtone of CH and OH bonds. No clear trend was observed when comparing the contribution plots of

---

**SYNCHRONIZATION-FREE ONLINE BATCH MSPC**

the two observations suggesting that this deviation may have been caused by changes of heterogeneity or particle comminution of the pharmaceutical granules.

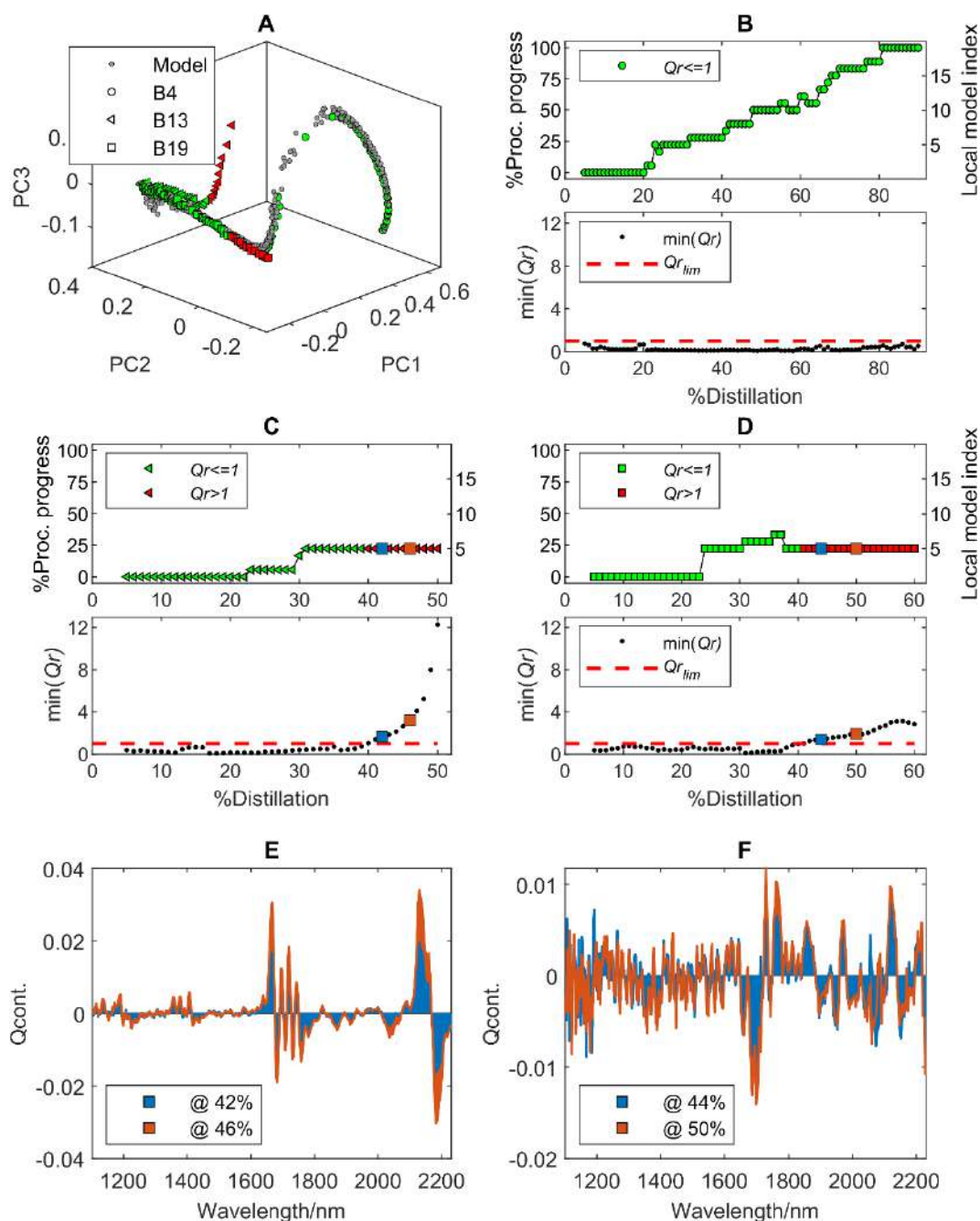
During the tracking of the additional batch BF2, three clusters of faulty observations were detected, see Figures 5F and 5G. The first faulty observations were detected during the first few minutes of the batch process. This deviation was related to the initial moisture content higher than the common starting point for the NOC batches used to build the MSPC models at the beginning of the process trajectory. However, after few minutes of drying, the online observations fell inside the confidence interval. The second faulty situation occurred after ca. 18 min of batch time during just four consecutive observations, but it quickly returned inside the control limit. This probably was related to a fast change of moisture content sensed by the NIR probe due to granule heterogeneity. This can be noticed by the fast change in process progress just before minute 20 in Figure 5F. From this point until approximately 60 min of batch time, the batch followed the NOC trajectory reaching 100% of batch progress, that is, reaching the minimum moisture level of the NOC batches used to train the local MSPC models at the end of the process trajectory, see Figure 5F. However, this batch was left to overdry reaching moisture levels lower than the endpoint of the historical NOC batches used for model training. The consequence of this action was successfully detected after approximately 70 min of the batch time by the MSPC chart (Figure 5G), where almost all consecutive observations were above the control limit. Looking at the bottom left of Figure 5E it can be observed how the PCA projections of these faulty observations were outside the NOC trajectory, but still following the drying process trend. Finally, two faulty observations at the end of this validation batch (at 100 and 105 min) were selected to check the contribution plots. These observations are marked in blue and orange squares in the score plot and MSPC charts. The  $Q$  contribution plots (Figure 5H) show that the absorption bands that contributed more to  $Q$  were around 1750 and 1950 nm related to the 1<sup>st</sup> overtone of CH and OH bonds, respectively, being the band at 1950 nm identified generally as the most dominant water band. The  $Q$  contribution positive and negative sign for the bands at 1750 and 1950 nm, respectively, indicates that the moisture level for these two observations was lower than the endpoint of the historical batches used in the model building. Also, when comparing the two faulty contribution plots, the systematic growth of the  $Q$  contributions at 1750 and 1950 nm bands, indicates the continuing moisture content decrease. It is important to note that this overdrying batch was used in this work to demonstrate the ability of the local MSPC models to detect such situations. In real-time monitoring, this batch would have been terminated once reached 100% of process progress, thus, avoiding energy waste and possible detrimental effects due to the excessive granules processing time.

#### 4.2.2 Application to Process 2 (gasoline distillation)

The local MSPC models built to track the batch gasoline distillation were tested. Three validation batches were used: one batch of on-specification gasoline blend with 27% of ethanol (batch B4) and two off-specification gasoline distillation batches, B13 and B19, with 15% and 30% ethanol blends, respectively. The results for all testing batches are shown in Figure 6.

The scatter score plot projections of the NIR observations for all three validation batches in the global PCA model used to build the *Process 2* NOC batch trajectory are represented in Figure 6A. In the score plot, gray dots identify the observations from the training batches describing the NOC batch trajectory, while the circles, triangles and squares are the projected observations from testing batches B4, B13 and B19, respectively. For the testing batches, the symbol face color indicates whether the observation was detected by the MSPC charts as faulty (red) or not (green). Process progress and  $\min(Qr)$  MSPC charts for the testing batches are shown in Figure 6B to Figure 6D for batches B4, B13 and B19, respectively. Additionally,  $Q$  contribution plots for two selected faulty observations are shown in Figure 6E to Figure 6F for batches B13 and B19, respectively.





**Figure 6** Global PCA score plot with the NOC batches (gray dots) used to define Process 2 batch trajectory, and the projection of the three validation batches (B4 circles, B13 triangles and B19 squares) (A). Process progress and  $Q_r$ -based MSPC charts for validation batch B4 (B), B13 (C) and B19 (D), green or red marker face color in process progress chart indicate that the observation is inside or outside the confidence limits, respectively.  $Q$  contribution ( $Q_{cont.}$ ) plots for selected faulty observations (indicated as blue and orange squares) for test batch B13 (E) and B19 (F).

The projections of the validation batch B4 in the global PCA model (Figure 6A) followed the NOC batch trajectory described by the cloud of gray dots. Indeed, when looking at the MSPC charts in Figure 6B, all observations are below the  $Q_r$  control limit and the batch process progressed accordingly to the on-specification gasoline batches. On the other hand, when looking at the projections of batch B13 observations to the global PCA model, an obvious deviation of the NOC batch trajectory was observed,

**SYNCHRONIZATION-FREE ONLINE BATCH MSPC**

see the red triangles in Figure 6A. This deviation was detected by the  $\min(Q_r)$  local MSPC charts (Figure 6C) after 40% of the initial batch weight was distilled. Note the interruption of the process progress after this point and several consecutive observations on until after ca. 70% of the distillation some observations fell back into the confidence interval defined by the local MSPC models at the end of the NOC batch trajectory. The off-specification batch B19 deviation from the NOC batch trajectory was lightly noticed by the PCA score plot projections in Figure 6A (red squares). However, this batch deviation was still detected by the local MSPC charts in Figure 6D. Note that this sensitivity is important since batch B19 contains 30% alcohol (v/v), only a 3% more than the NOC batches. Similarly, the fault was first detected after ca. 40% of the distillation batch and all consecutive observations since then were detected outside the confidence interval for all local MSPC models.

The contribution plots (Figure 6E) for the selected fault observations at 42% (in blue) and 46% (in orange) fraction of distilled material of the B13 batch show that the two bands covering the 1650-1700 nm and 2100-2200 nm NIR contributed the most to the  $Q$ . The absolute increase of  $Q$  contributions at 1665, 2130 and 2180 nm indicated a possible increment of mid and high-density hydrocarbon fractions at these distillation points. Additionally, the negative contribution at 1685 nm indicated a lower content of ethanol and light hydrocarbon compounds. This agrees with the expected distillation behavior for off-specification gasoline blends with low ethanol content. This is confirmed when looking at the distillation profiles obtained by Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) for these compounds presented in our previous work for this specific batch (de Oliveira et al., 2017). For batch B19, Figure 6F shows the contribution plots for the faulty observation at 44% (in blue) and 50% (in orange) of the batch distillation. The high negative contribution between 1680 and 1700 nm suggested the presence of a lower content of mid and heavy hydrocarbons fraction than expected for NOC batches at this point of distillation. These ethanol-rich fractions were related to the fact that this off-specification gasoline batch had a slightly higher ethanol content (30%) than NOC gasolines (27%).

## 5 Conclusions

The present work introduces a new approach for online monitoring of batch process evolution through the design of local MSPC models covering an overall NOC batch process trajectory, defined from the PCA modeling of non-synchronized NOC batches. The tracking of the evolution of new batches does not require synchronization either. The methodology has been demonstrated with the building and validation of online MSPC charts for the monitoring of two real batch process data of different nature using in-situ NIR measurements. In both process examples, the implementation of local MSPC charts has been successfully validated for the tracking of well-known new batches that followed or deviated from the overall NOC batch trajectory. The use of  $Q$  contribution plots was helpful to identify the sources of process abnormalities based on the chemical information provided by the NIR signal.

The fact that the proposed methodology does not require batch synchronization makes the data analysis pipeline simpler and flexible and offers many advantages for real-time process monitoring, from the building of the reference MSPC models to the test of new batches. Thus, the designed methodology allows the model building with historical NOC process data acquired with different online sampling rates and spanning evolution in different time (or process variable) ranges. The monitoring of new batches is also independent of the sampling rate used in the model building, which allows for changes in the sampling interval if required. Furthermore, the fact that the exam of the quality of new batch observations provides additionally a good indication of the process progress enables the potential use of this online tracking methodology for end-point detection, providing a single tool to control both the evolution and the end of the process. The presented methodology has been applied to NIR monitored processes but could be readily adapted to deal simultaneously with the output from several sensor

outputs in a sensor fusion scenario. That would allow an integral control of the process evolution by combining the output from advanced sensors with other process data (temperature, flow, pressure, ...)

## 6 Conflict of Interest

*The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.*

## 7 Author Contributions

R. Rocha de Oliveira: Conceptualization, Methodology, Investigation, Data Curation, Visualization, Software, Formal analysis, Writing – original draft, Writing – review & editing.

A. de Juan: Conceptualization, Methodology, Supervision, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition.

## 8 Funding

Spanish government. PID 2019-1071586B-IOO

Catalan Government: Excellence research group (2017 SGR 753).

## 9 References

- Aguado, D., Ferrer, A., Ferrer, J., and Seco, A. (2007). Multivariate SPC of a sequencing batch reactor for wastewater treatment. *Chemom. Intell. Lab. Syst.* 85, 82–93. doi:10.1016/j.chemolab.2006.05.003.
- Avila, C., Mantzaridis, C., Ferré, J., Rocha de Oliveira, R., Kantojärvi, U., Rissanen, A., et al. (2021). Acid number, viscosity and end-point detection in a multiphase high temperature polymerisation process using an online miniaturised MEMS Fabry-Pérot interferometer. *Talanta* 224, 121735. doi:10.1016/j.talanta.2020.121735.
- Avila, C. R., Ferré, J., de Oliveira, R. R., de Juan, A., Sinclair, W. E., Mahdi, F. M., et al. (2020). Process Monitoring of Moisture Content and Mass Transfer Rate in a Fluidised Bed with a Low Cost Inline MEMS NIR Sensor. *Pharm. Res.* 37, 84. doi:10.1007/s11095-020-02787-y.
- Ávila, T. C., Poppi, R. J., Lunardi, I., Tizei, P. A. G., and Pereira, G. A. G. (2012). Raman spectroscopy and chemometrics for on-line control of glucose fermentation by *Saccharomyces cerevisiae*. *Biotechnol. Prog.* 28, 1598–1604. doi:10.1002/btpr.1615.
- Besenhard, M. O., Scheibelhofer, O., François, K., Joksich, M., and Kavsek, B. (2018). A multivariate process monitoring strategy and control concept for a small-scale fermenter in a PAT environment. *J. Intell. Manuf.* 29, 1501–1514. doi:10.1007/s10845-015-1192-8.
- Bogomolov, A. (2011). Multivariate process trajectories: Capture, resolution and analysis. *Chemom. Intell. Lab. Syst.* 108, 49–63. doi:10.1016/j.chemolab.2011.02.005.
- Cimander, C., and Mandenius, C. F. (2004). Bioprocess control from a multivariate process trajectory. *Bioprocess Biosyst. Eng.* 26, 401–411. doi:10.1007/s00449-003-0327-z.

## SYNCHRONIZATION-FREE ONLINE BATCH MSPC

- de Oliveira, R. R., Pedroza, R. H. P., Sousa, A. O., Lima, K. M. G., and 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. doi:10.1016/j.aca.2017.07.038.
- Ferrer-Riquelme, A. J. (2010). “Statistical Control of Measures and Processes,” in *Comprehensive Chemometrics*, 97–126. doi:10.1016/B978-044452701-1.00096-X.
- Gomes, F. P. C., Garg, A., Mhaskar, P., and Thompson, M. R. (2019). Data-Driven Advances in Manufacturing for Batch Polymer Processing Using Multivariate Nondestructive Monitoring. *Ind. Eng. Chem. Res.* 58, 9940–9951. doi:10.1021/acs.iecr.8b05675.
- González-Martínez, J. M., de Noord, O. E., and Ferrer, A. (2014a). Multisynchro: A novel approach for batch synchronization in scenarios of multiple asynchronisms. *J. Chemom.* 28, 462–475. doi:10.1002/cem.2620.
- González-Martínez, J. M., Ferrer, A., and Westerhuis, J. A. (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. doi:10.1016/j.chemolab.2011.01.003.
- González-Martínez, J. M., Vitale, R., De Noord, O. E., and Ferrer, A. (2014b). Effect of synchronization on bilinear batch process modeling. *Ind. Eng. Chem. Res.* 53, 4339–4351. doi:10.1021/ie402052v.
- Grassi, S., Strani, L., Casiraghi, E., and Alamprese, C. (2019). Control and monitoring of milk renneting using FT-NIR spectroscopy as a process analytical technology tool. *Foods* 8. doi:10.3390/foods8090405.
- Haack, M. B., Eliasson, A., and Olsson, L. (2004). On-line cell mass monitoring of *Saccharomyces cerevisiae* cultivations by multi-wavelength fluorescence. *J. Biotechnol.* 114, 199–208. doi:10.1016/j.jbiotec.2004.05.009.
- Jackson, J. E., and Mudholkar, G. S. (1979). Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics* 21, 341–349. doi:10.1080/00401706.1979.10489779.
- Jolliffe, I. T. (2002). *Principal components analysis*. 2nd ed. New York: Springer.
- Kassidas, A., Macgregor, J. F., and Taylor, P. A. (1998). Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J.* 44, 864–875. doi:10.1002/aic.690440412.
- Kourtí, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *Int. J. Adapt. Control Signal Process.* 19, 213–246. doi:10.1002/acs.859.
- Liu, Y. J., André, S., Saint Cristau, L., Lagresle, S., Hannas, Z., Calvosa, É., et al. (2017). Multivariate statistical process control (MSPC) using Raman spectroscopy for in-line culture cell monitoring considering time-varying batches synchronized with correlation optimized warping (COW). *Anal. Chim. Acta* 952, 9–17. doi:10.1016/j.aca.2016.11.064.
- Mortensen, P. P., and Bro, R. (2006). Real-time monitoring and chemical profiling of a cultivation

- process. *Chemom. Intell. Lab. Syst.* 84, 106–113. doi:10.1016/j.chemolab.2006.04.022.
- Nomikos, P., and MacGregor, J. F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* 37, 41–59. doi:10.1080/00401706.1995.10485888.
- Pöllänen, K., Häkkinen, A., Reinikainen, S.-P., Rantanen, J., and Minkkinen, P. (2006). Dynamic PCA-based MSPC charts for nucleation prediction in batch cooling crystallization processes. *Chemom. Intell. Lab. Syst.* 84, 126–133. doi:10.1016/j.chemolab.2006.04.016.
- Ramaker, H. J., Van Sprang, E. N. M., Westerhuis, J. A., and Smilde, A. K. (2004). The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring. *Chemom. Intell. Lab. Syst.* 73, 181–187. doi:10.1016/j.chemolab.2003.12.015.
- Rännar, S., MacGregor, J. F., and Wold, S. (1998). Adaptive batch monitoring using hierarchical PCA. *Chemom. Intell. Lab. Syst.* 41, 73–81. doi:10.1016/S0169-7439(98)00024-0.
- Rato, T. J., Blue, J., Pinaton, J., and Reis, M. S. (2017). Translation-Invariant Multiscale Energy-Based PCA for Monitoring Batch Processes in Semiconductor Manufacturing. *IEEE Trans. Autom. Sci. Eng.* 14, 894–904. doi:10.1109/TASE.2016.2545744.
- Rato, T. J., and Reis, M. S. (2020). An integrated multiresolution framework for quality prediction and process monitoring in batch processes. *J. Manuf. Syst.* 57, 198–216. doi:10.1016/j.jmsy.2020.09.007.
- Rendall, R., Chiang, L. H., and Reis, M. S. (2019). Data-driven methods for batch data analysis – A critical overview and mapping on the complexity scale. *Comput. Chem. Eng.* 124, 1–13. doi:10.1016/j.compchemeng.2019.01.014.
- Rocha de Oliveira, R., Avila, C., Bourne, R., Muller, F., and de Juan, A. (2020). Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control. *Anal. Bioanal. Chem.* 412, 2151–2163. doi:10.1007/s00216-020-02404-2.
- Skibsted, E. T. S., Boelens, H. F. M., Westerhuis, J. A., Witte, D. T., and Smilde, A. K. (2006). Simple assessment of homogeneity in pharmaceutical mixing processes using a near-infrared reflectance probe and control charts. *J. Pharm. Biomed. Anal.* 41, 26–35. doi:10.1016/j.jpba.2005.10.009.
- van Sprang, E. N. ., Ramaker, H.-J., Westerhuis, J. a, Gurden, S. P., and Smilde, A. K. (2002). Critical evaluation of approaches for on-line batch process monitoring. *Chem. Eng. Sci.* 57, 3979–3991. doi:10.1016/S0009-2509(02)00338-X.
- Westad, F., Gidskehaug, L., Swarbrick, B., and Flåten, G. R. (2015). Assumption free modeling and monitoring of batch processes. *Chemom. Intell. Lab. Syst.* 149, 66–72. doi:10.1016/j.chemolab.2015.08.022.
- Wold, S. (1978). Cross -Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* 20, 397–405.
- Wold, S., Kettaneh-Wold, N., MacGregor, J. F., and Dunn, K. G. (2009). “2.10 - Batch Process Modeling and MSPC,” in *Comprehensive Chemometrics*, 163–197. doi:10.1016/B978-

**SYNCHRONIZATION-FREE ONLINE BATCH MSPC**

044452701-1.00108-3.

Wold, S., Kettaneh, N., Fridén, H., and Holmberg, A. (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* 44, 331–340. doi:10.1016/S0169-7439(98)00162-2.

Zhao, J., Li, W., Qu, H., Tian, G., and Wei, Y. (2020). Real-time monitoring and fault detection of pulsed-spray fluid-bed granulation using near-infrared spectroscopy and multivariate process trajectories. *Particuology* 53, 112–123. doi:10.1016/j.partic.2020.02.003.



The online MSPC modeling strategy introduced in chapter 4.2 cannot be used for process control of non-synchronized batch data unless batch alignment tools are used to cope with this problem (J. M. González-Martínez et al., 2014). However, batch alignment is a delicate operation that can induce artifacts in the definition of MSPC models when scarce information is available or when it is not properly applied. Hence, the need for MSPC approaches that can skip the synchronization step for online process monitoring and control. In this work, a new synchronization-free online MSPC approach for tracking process evolution in synchronized or non-synchronized batch processes is introduced. This methodology is based on three main steps: a) modeling of NOC batch process trajectories; b) construction of local MSPC models based on the information provided by NOC batch process trajectories; and c) use of an MSPC chart based on local MSPC models to track the evolution of new batches.

The first two steps involve the generation of the MSPC models, whereas the last step involves the use of the local MSPC models on new batches to test whether they follow the NOC trajectory or present deviations from the expected NOC behavior. The steps of the methodology proposed and the results obtained after the application to a fluidized bed drying process are discussed below.

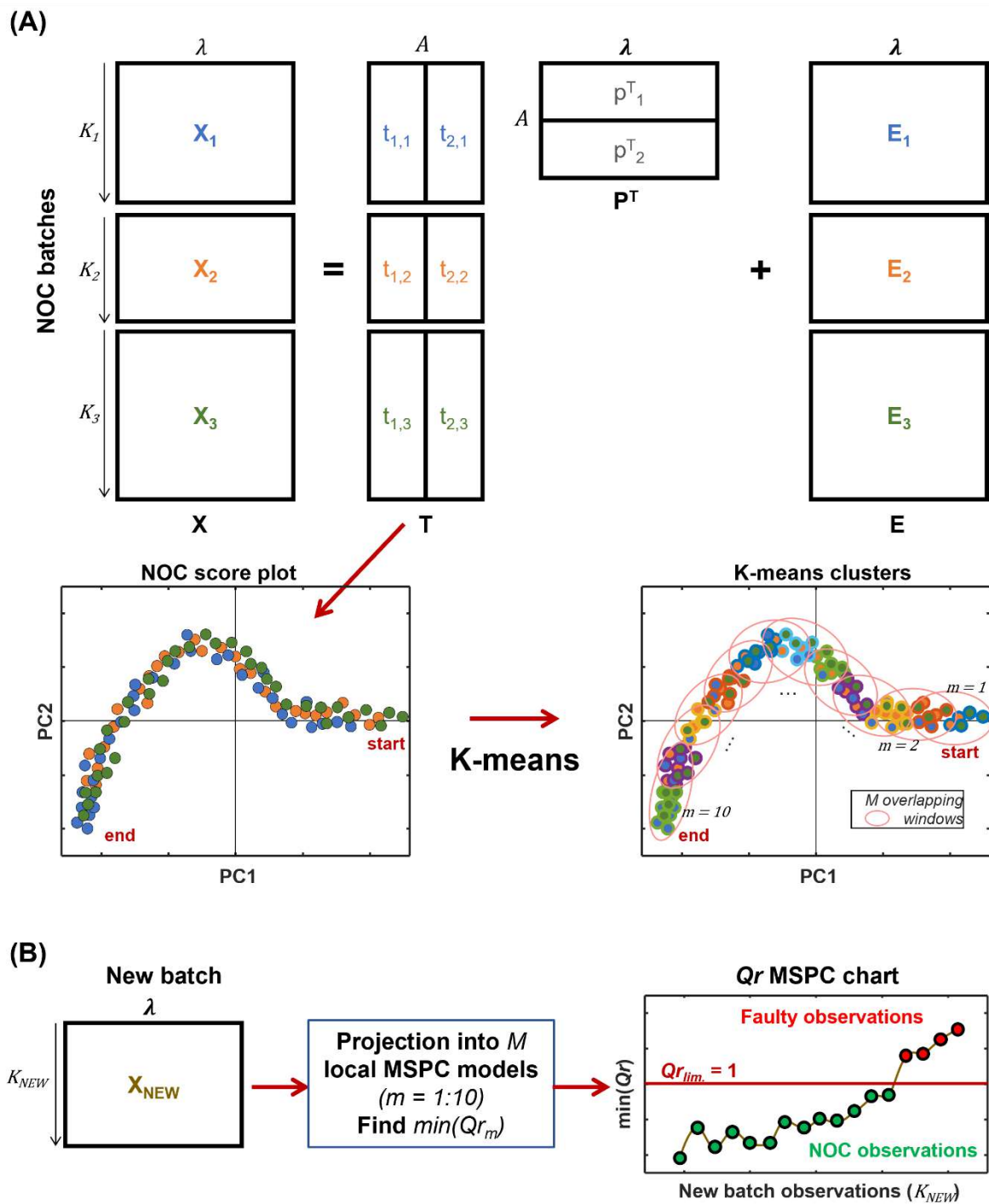
### **Online MSPC methodology**

#### **Step 1: modeling of NOC batch process trajectories**

The data from a complete NOC batch  $i$  are usually structured in a data matrix  $\mathbf{X}_i (K_i \times J)$ , where  $K_i$  are the number of spectra related to the process observations and  $J$  is the number of measured variables, e.g. NIR channels per spectrum. When several data matrices  $\mathbf{X}_i$  from non-synchronized NOC batches are used to define the general process trajectory, they are organized in a variable-wise multibatch structure  $\mathbf{X} (N \times J)$ , where  $N$  is the number of rows related to the total number of observations from the  $I$  NOC batches, that is,  $N = \sum_i^I K_i$ . As mentioned in section 4.1, this strategy does not require resizing or synchronization of uneven batch lengths, since the only requirement is that all batches share a common variable dimension  $J$ . See Figure 29A (top left) for a multibatch structure formed by three non-synchronized and differently sized batches.

To model the overall NOC batch trajectory, this multibatch structure is column mean-centered and analyzed with PCA, as explained in section 2.3.1 and illustrated in Figure 29A. Note that this centering operation does not remove the mean trajectory of the batches in time. The PCA modeling of this variable-wise augmented structure provides a loading matrix,  $\mathbf{P}^T (A \times J)$ , common to all batches and an augmented score matrix,  $\mathbf{T} (N \times A)$ , that accommodates the  $\mathbf{T}_i$  scores related to every batch.  $A$  is the number of PC's required by the PCA model.





Step 2: construction of local MSPC models based on NOC batch process trajectories

The augmented score matrix of all NOC batches can be used to represent the overlapped NOC batch process trajectories in the reduced PCA space, as illustrated in Figure 29A (bottom left). In this context, the dots represent the scores for each observation and are colored according to the NOC batches in the PCA model. Note that the overall trajectory evolution is the same for all NOC batches but, in a general non-synchronized case, the starting and endpoint of every NOC batch do not need to coincide.

The global description of the variability in the NOC process evolution through the overlapped NOC trajectories can be helpful to observe whether a new batch process evolves as NOC batches do or not, independently from the batch length and dynamics. To track locally the evolution of new batches, the overlapped NOC trajectories can be divided into a sufficient number of  $C$  local regions using a cluster analysis methodology, such as k-means. Figure 29A (bottom right) illustrates these local regions for  $C = 11$ , as indicated by the outer circle color of the neighbor observations inside each cluster.

To model the NOC process variability and set local MSPC control charts, the original NIR spectra from the observations in two consecutive clusters are used as seeding information to construct local MSPC models. Therefore, the first local MSPC model contains the observations in the first two clusters of the process trajectory, the second local MSPC model uses the observations in clusters two and three and so forth until all the NOC global process trajectory is covered. The observations used in consecutive local MSPC models overlap with each other so that all process trajectory regions are covered. As can be observed in Figure 29A (bottom right), for a k-means analysis providing 11 clusters, 10 local MSPC models with overlapping information as defined by the red ellipses can be built.

The PCA-based local MSPC models are built and the related control chart limits set using the procedure explained in section 2.3.3. For the online MSPC approach proposed, the specific operational procedure to build each local MSPC model can be described as follows. First, the original observations, e.g. NIR spectra, inside the consecutive clusters for each local model are placed into a data matrix  $\mathbf{X}_m (K_m \times J)$ , where  $m$  indicates the index of the local model (from 1 to  $M$ ) and  $K_m$  is the number of observations used to build the model. Then, this matrix is mean-centered and modeled with PCA generating the matrices of scores  $\mathbf{T}_m (K_m \times A_m)$ , loadings  $\mathbf{P}_m^T (A_m \times J)$ , and residuals  $\mathbf{E}_m (K_m \times J)$ . Note that the mean-centering step is performed using the mean of the matrix  $\mathbf{X}_m$  and not the global mean of the multibatch structure. In this work, the controls charts are based only on the residual matrix,  $\mathbf{E}_m$ , deriving the Q-statistic and control chart limit,  $Q_{lim}$  calculated as in section 2.3.3. Now that all local MSPC models and related control chart limits are set, the next step is to track the evolution of new batches based on the local models constructed.

Step 3: use of an MSPC chart based on local MSPC models to track new batch evolution.

For the online batch monitoring of new batch observations ( $\mathbf{X}_{\text{NEW}}$  in Figure 29B), every new observation, e.g. NIR spectrum, is projected onto all local MSPC models, for  $m = 1:M$ , and the related reduced Q-residuals,  $Qr_{k,m}$ , are obtained for each new observation,  $\mathbf{x}_k$ , as shown in Figure 29B.

Then, the reduced Q values for every new observation,  $Qr_{k,m}$ , are checked against the reduced control limit to see whether they are above or below the  $Qr_{lim} = 1$ . If all  $Qr_{k,m}$  values for the observation  $k$  are large and above one, this observation is diagnosed as faulty, and it is an indicator that the process is deviating from the NOC trajectory. Conversely, if one or more  $Qr_{k,m}$  values are below the control limit, the observation follows the NOC trajectory. An easy way to visualize the diagnostic of every new observation is by using a single  $Q$  chart, as in Figure 29B (right), where only the minimum  $Qr$  parameter after the projection onto all local models for every new observation is displayed for every new observation. Observations that follow the NOC trajectory are depicted by the green dots below the  $Qr_k < 1$ , and the eventual deviations from it, with  $\min(Qr_{k,m}) > 1$ , in red. To identify the spectral variables making the greatest contributions to the deviation in  $Q$ , the Q-statistics contribution plots for the sought observation can be displayed. The residuals used for the contribution plots are calculated using the best local MSPC model related to the last NOC observation. For observations following the NOC trajectory, it is also possible to estimate the process stage of every observation by identifying the local MSPC model providing the lowest  $Qr_{k,m}$  value. The indices of the local MSPC models can be used to set an approximate maturity index, as will be explained afterwards. Different visualization approaches to interpret the results issued from the application of the online MSPC for tracking the drying process will be presented.

**Online MSPC applied to monitor the fluidized bed drying process**

Figure 30 shows the score plot related to the PCA model of the non-synchronized NOC batches from the FB drying process. The model needed two components and had an explained variance of 97.61%. The score plot described mostly the variation of the moisture content with the drying evolution from the beginning to the end of every NOC batch. Because all batches had different initial and final moisture conditions, they show different process dynamics with time; however, when overlaying all individual batch trajectories in the same score plot, the same evolution pattern can be observed. Once the overall NOC trajectory has been defined, k-means analysis allowed the identification of 30 clusters along this trajectory, as displayed by the different outer circle colors in Figure 30. After that, a number indicating the process stage evolution was automatically assigned to each cluster according to the position in the overall NOC trajectory.

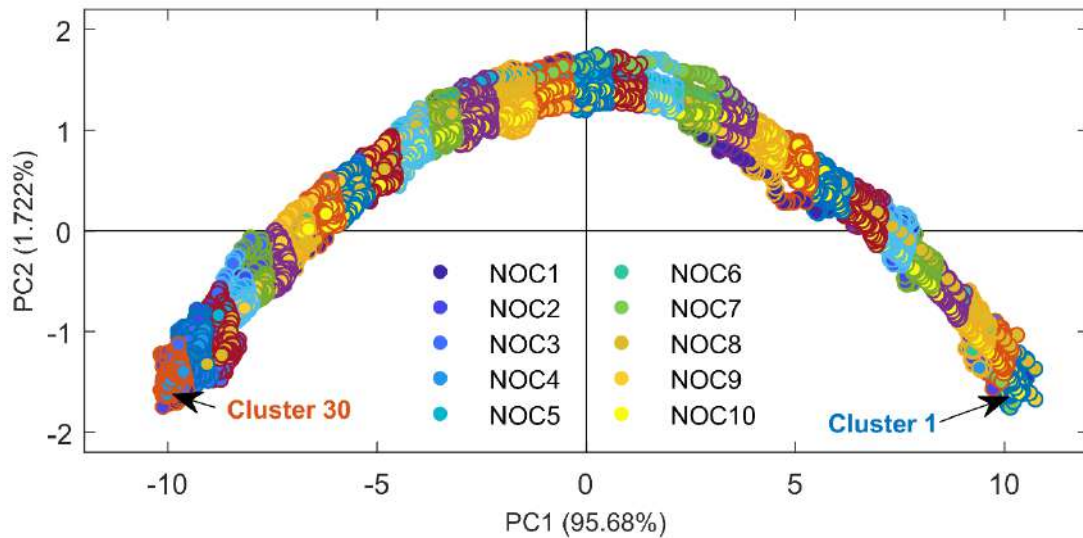


Figure 30 PCA score plot for of the observations of the online NIR monitored FB drying process showing the training NOC batch process trajectories and local clusters found by k-means. The inner part of the circles is colored according to the related NOC batch, whereas the outer part reflects the observations included in every cluster and, hence, in the related local MSPC model (reproduced from Publication V).

At this point, the original NIR observations inside the suitable two consecutive k-means clusters were used as seeding information to build local MSPC models for each step of the batch trajectory. Thus, for the 30 clusters formed, 29 local PCA-based MSPC models with overlapping information were built. Control chart limits based on the  $Q$ -statistics with a 99% confidence interval were calculated for each local MSPC model and used for the online tracking of new batches evolution.

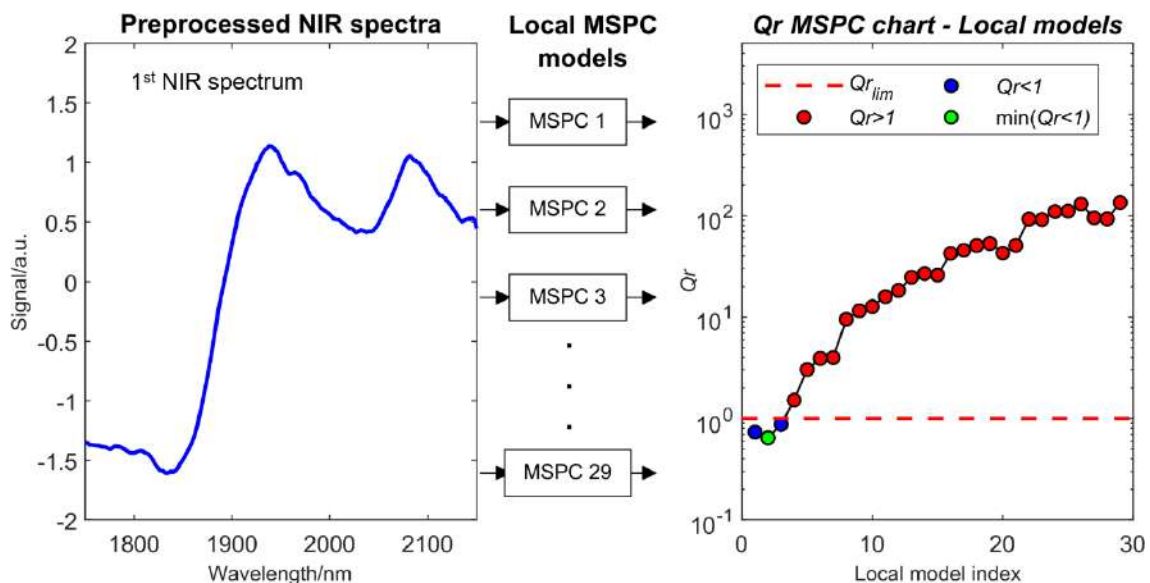


Figure 31 Projection of the first NIR observation of test batch BN1 (left plot) into all 29 local MSPC models and related  $Q_r$  values obtained (right plot). Red circles indicate  $Q_r$  values above the control limit; blue, below the limit; and green, the minimum  $Q_r$  below the control limit. The vertical axis is log-scaled for a better visualization.

The tracking of a new drying batch is performed by projecting every new observation, i.e., preprocessed NIR spectrum, onto all 29 local MSPC models. Figure 31 illustrates

the projection of the first preprocessed NIR spectrum (left plot) onto all local MSPC models and the related (right plot)  $Q_r$  values issued from every MSPC local model. From the  $Q_r$  control chart, it can be observed that for this spectral observation, the  $Q_r$  values related to the first three local MSPC models are below the control limit and that the minimum is related to the second local model in the NOC trajectory. This behavior is logical, since the first spectrum of a batch is expected to be more similar to the observations acquired at the beginning of the NOC batches, represented by the first local MSPC models.

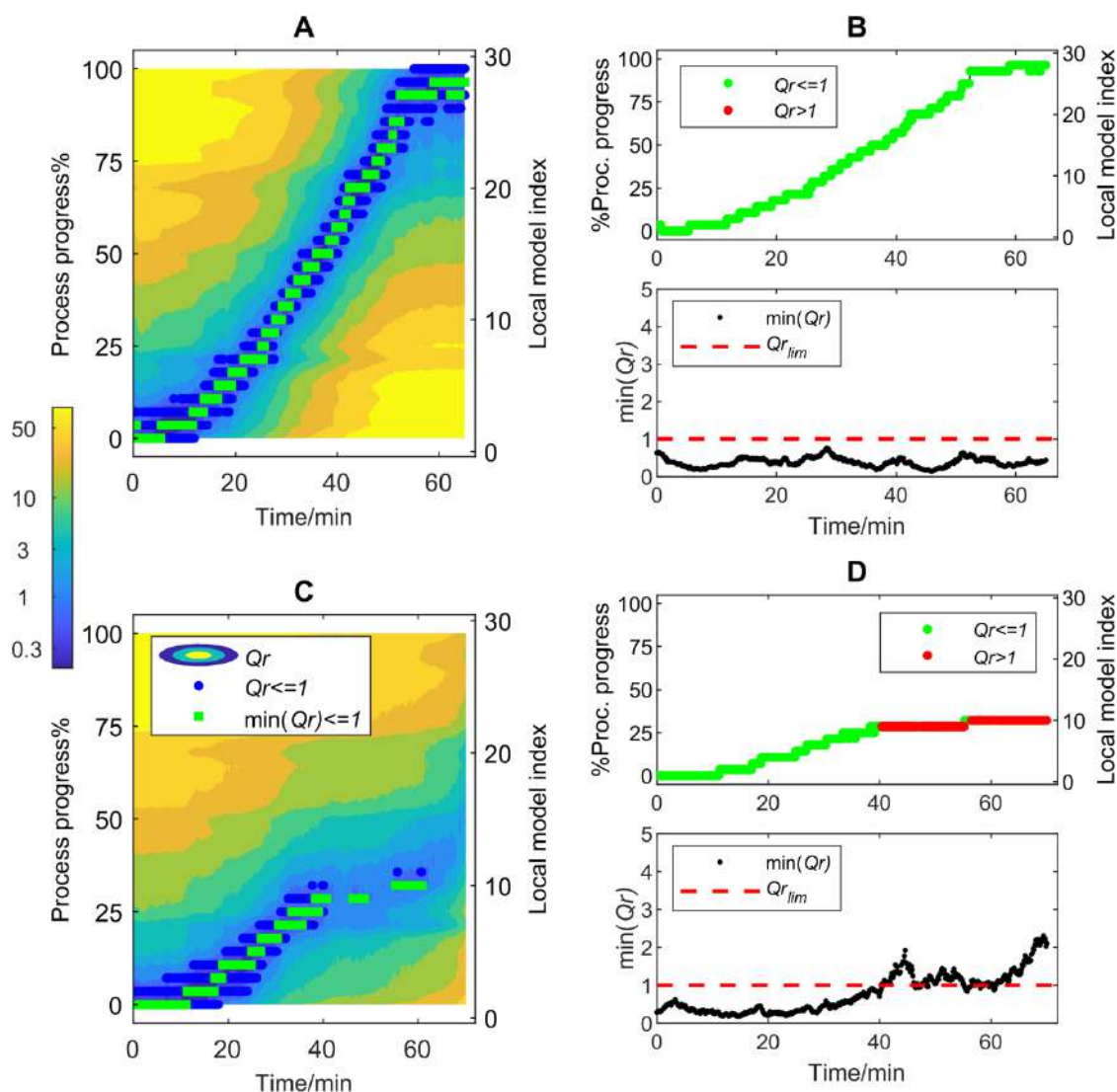


Figure 32  $Q_r$ -based MSPC charts for FB drying NOC batch BN1 (A and B) and faulty batch BF1 (C and D). (A and C) Contour plots of the  $Q_r$  values calculated after the projection of each NIR observation onto the local MSPC models. Blue dots show values of  $Q_r < 1$  (control limit), green squares the  $\min(Q_r < 1)$ . (B and D) Charts show the  $\min(Q_r)$  value (bottom panel) and the related process progress associated with it (top panel) for every batch observation. In the process progress plot, NOC observations are displayed in green and faulty observations in red (reproduced from Publication V).

Thus, for every new spectral observation, a vector of  $Q_r$  values with as many elements as local MSPC models,  $M$ , can be stored into a matrix sized  $(M \times K_{NEW})$ , where  $K_{NEW}$  is the total number of observations points of the new batch. This data matrix with all

$Qr$  values can be easily visualized through a contour plot, as shown in Figure 32 (left plots).

The  $Qr$ -based MSPC control charts for the online tracking of observations from two drying batches are shown in Figure 32. Figure 32A and Figure 32C are contour plots related to a NOC batch, BN1, and a faulty batch, BF1, respectively, that show all the  $Qr$  values obtained after the projection of each NIR observation of the batch onto all local MSPC models. A log-scale colormap has been used to highlight the differences at low  $Qr$  values. The horizontal axis of the contour plot represents the batch time at which every observation was collected and the right vertical axis the indices related to the different local MSPC models calculated, i.e. from 1 to 29. Additionally, in the left vertical axis, each local MSPC model index is associated with a percentage of the process progress from 0-100%, defined making a linear scaling that links the initial local model to 0% process progress and the final local model to 100% process progress. The process progress in this approach plays the same role as the process maturity index proposed by other authors (Westad et al., 2015; Wold et al., 1998). Thus, to track the behavior of an observation of a new batch, their related  $Qr$  values (associated with a specific process time) are examined. In the contour plots in Figure 32A and Figure 32C, the  $Qr$  values below the control limit, i.e.  $Qr < 1$ , are depicted as blue dots and the  $\min(Qr < 1)$  for every observation in green. If an observation shows a NOC behavior (as all do in Figure 32A related to NOC batch BN1), there will always be one or more  $Qr$  values below 1; i.e., all observations will show one or more blue dots and a green dot. Instead, when an observation deviates from the NOC trajectory, as in batch BF1 (in Figure 32C), all  $Qr$  values related to that observation are above the control limit of 1 and neither blue nor green dots are observed.

To summarize and facilitate the interpretation of the relevant information of the contour plots, graphics displaying the  $\min(Qr)$  value and the related process progress for every batch observation are provided (see Figure 32B and Figure 32D for batches BN1 and BF1, respectively). Figure 32B shows that all observations for batch BN1 followed the NOC batch trajectory, seen because all  $\min(Qr)$  values were below the control limit of 1 (bottom panel), and that the process progress covered the complete range (0-100%) (top panel). Figure 32D shows that batch BF1 deviated from the NOC trajectory after approximately 40 min of batch time as flagged by the  $Qr$  above the local MSPC control limits ( $\min(Qr) > 1$ ) (bottom plot). When a fault happens, the related observations are displayed in red in the process progress plot to indicate that the evolution of the process is abnormal (top plot).

Additional results and interpretation of the abnormal behaviour for the online tracking of two faulty batches, BF1 and BF2, are shown in Figure 33 (left and right plots, respectively). Figure 33A and Figure 33E show the deviations of the two batches as seen by the score plot projections of their observations onto the global PCA model used to describe the NOC batch trajectory. The score plot shows all training NOC batch trajectories as gray dots, whereas the new observations are represented in

green (NOC) or in red (fault) according to the MSPC detection. Figure 33B and Figure 33F show the batch process progress plot and Figure 33C and Figure 33G the  $\min(Q_r)$  MSPC chart for the tracking of the online observations. Abnormal observations are associated with  $\min(Q_r)$  values higher than 1 and flagged in red color in the process progress plot. To assess the spectral variables making the greatest contributions to the deviation from the NOC batch trajectory, Figure 33D and Figure 33H show the  $Q$  contribution plots from two faulty observations selected for each batch.

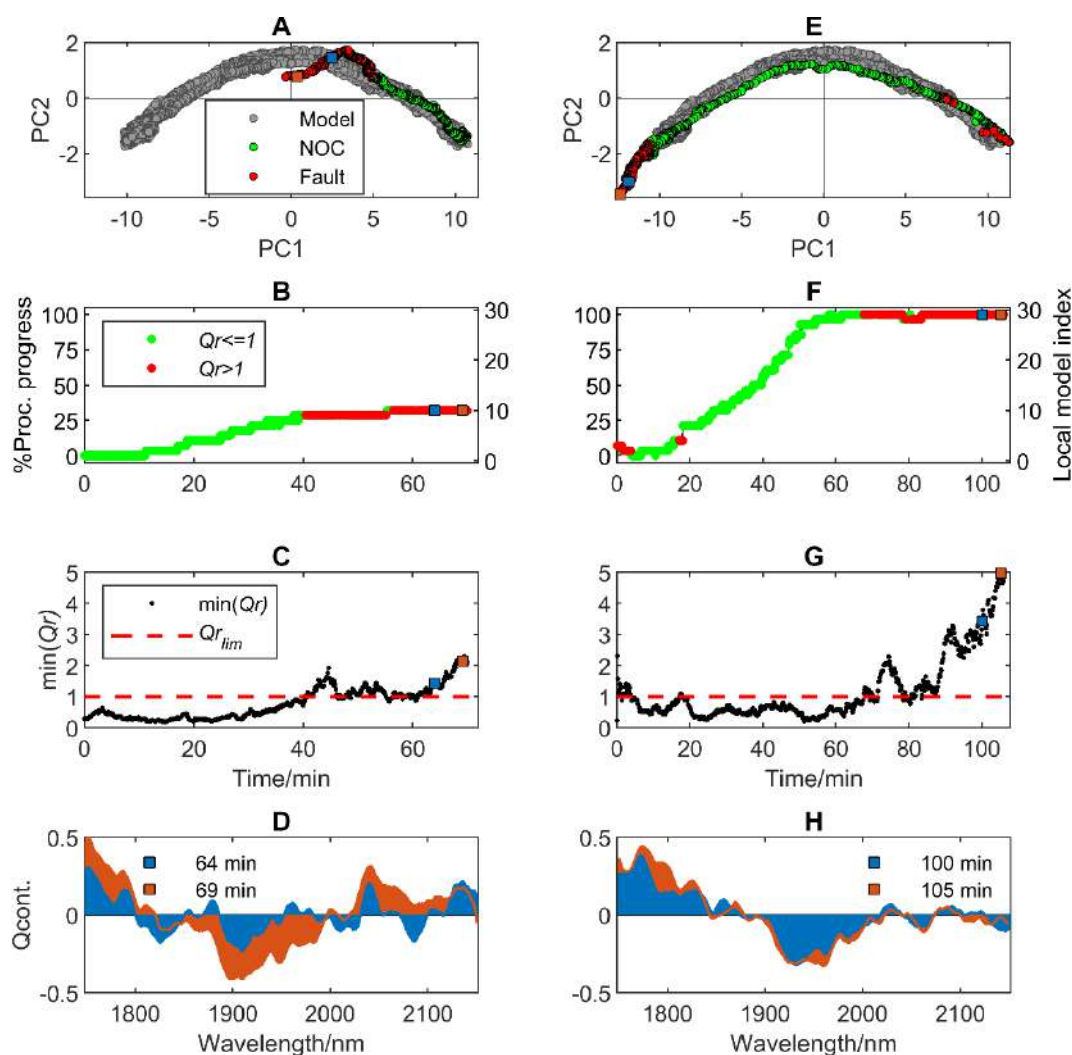


Figure 33 Results for the online tracking of new batch evolution using the local MSPC for faulty batches BF1 (A to D) and BF2 (E to H). (A and E) PCA score plot showing the NOC trajectory (gray dots) and new batch trajectory in green (NOC observations) and red dots (faulty observations). (B and F) MSPC chart showing the process progress. (C and G)  $Q_r$ -based MSPC charts. (D and H) are the  $Q$  contribution ( $Q_{cont.}$ ) plots for two faulty observations selected for each batch and represented by the blue and orange squares in the MSPC control charts (reproduced from Publication V).

Looking at Figure 33B and Figure 33C, we can observe that this methodology is capable to detect batch BF1 deviation from the NOC trajectory after approximately 40 min of batch time. Although in Figure 33A the faulty observations (red dots) right after 40 min were still close to the NOC trajectory, the related  $\min(Q_r)$  after projection onto local MSPC models was above the control limit indicating a deviation, which became

even larger towards the end of the batch run, see Figure 33C. The contribution plots shown in Figure 33D are related to the selected faulty observations as marked in blue and orange squares in the score plot and MSPC charts. The contribution plots show that the absorption bands that gave higher contributions to  $Q$  were around 1750 and 1900 nm related to the 1<sup>st</sup> overtone of CH and OH bonds. No clear trend was observed when comparing the contribution plots of the two observations for this batch suggesting that this deviation may have been caused by changes of heterogeneity or particle comminution of the pharmaceutical granules.

During the tracking of the additional faulty batch BF2, several faulty observations were detected, see Figure 33F and Figure 33G. The faulty observations during the first 20 minutes of batch time can be assigned to a few anomalous observations that might be related to the unusual higher initial moisture content and fast changes of granule heterogeneity sensed by the NIR probe. Except for these isolated situations, the batch followed satisfactorily the complete NOC trajectory reaching 100% of batch progress after approximately 60 min of batch time. This means that the batch reached the minimum moisture level of the NOC batches used to train the local MSPC models at the end of the process trajectory, see Figure 33F. However, this batch was left to overdry reaching moisture levels lower than the endpoint of the historical NOC batches used for model training. The consequence of this action is reflected by the continuously increasing  $\min(Qr) > 1$  values in the MSPC chart (Figure 33F and 33G) towards the end of the process. The same can be observed looking at the bottom left of the score plot projections in Figure 33E. The  $Q$  contribution plots from the selected observations at the end of the batch run, shown in Figure 33H, indicated that the most dominant contribution is related to the water band at 1950 nm. When comparing both contributions plots, the systematic growth of the  $Q$  contributions indicates the continuing moisture content decrease. It is important to note that in a real industrial application, this batch would have been terminated once reached 100% of process progress. Nevertheless, this application has demonstrated the ability of the local MSPC models to detect such situations and that the real-time monitoring would have resulted in a correct batch endpoint detection, which would avoid energy waste and possible detrimental effects due to the excessive granule processing time.

In summary, these results demonstrate the potential of this online MSPC methodology for batch process evolution without the requirement of batch synchronization. The tracking of the evolution of new batches does not require synchronization either. This methodology also works with naturally synchronized batch data, such as the distillation process described in section 3.1.3. The use of  $Q$  contribution plots is also helpful to identify the sources of process abnormalities based on the chemical information provided by the NIR signal. The proposed batch synchronization-free methodology makes the data analysis pipeline simpler and flexible and offers many advantages for real-time process monitoring, from the building of the reference MSPC models to the test of new batches. Thus, the designed methodology allows the model building with historical NOC process data acquired with different online sampling rates and



spanning evolution in different time (or process variable) ranges. The monitoring of new batches is also independent of the sampling rate used in the model building, which allows for changes in the sampling interval if required. Furthermore, the fact that the exam of the quality of new batch observations provides additionally a good indication of the process progress enables the potential use of this online tracking methodology for endpoint detection, providing a single tool to control both the evolution and the endpoint of the process. Although this methodology has been tested with NIR monitored processes, it can be adapted to deal simultaneously with the output from several sensor outputs in a sensor fusion scenario, such as presented in (de Oliveira et al., 2020). That would allow an integral control of the process evolution by combining the output from advanced sensors with other process data (temperature, flow, pressure, ...).

## **SECTION II – Process monitoring using hyperspectral imaging.**

This section is focused on the development of a novel PAT tool for the assessment of heterogeneity during the monitoring of blending processes using hyperspectral images (HSI).



#### 4.4 Blending process monitoring and heterogeneity assessment

A new methodology for qualitative and quantitative assessment of heterogeneity during atline and inline monitoring of blending processes using hyperspectral images (HSI) is presented. The methodology is based on a first step of HSI unmixing that provides the pure distribution maps of the blending constituents as a function of blending time. These maps allow visualizing qualitatively the heterogeneity variation during the blending process. In the second step, these maps are used as seeding information for a subsequent variographic analysis to extract quantitative heterogeneity indices for blending quality assessment. This study resulted in two scientific publications. The first article (Publication VI) introduces this novel methodology and defines the heterogeneity indices using atline NIR-HSI data collected at different blending times of a pharmaceutical formulation. The second publication (Publication VII) extends the use of the variogram-based heterogeneity indices for the real-time monitoring of food and pharmaceutical blending processes using inline NIR-HSI.

**Publication VI.** Rocha de Oliveira, R., de Juan, A. **Design of Heterogeneity Indices for Blending Quality Assessment Based on Hyperspectral Images and Variographic Analysis.** *Analytical Chemistry* (2020), 92: 15880–15889.  
DOI: [10.1021/acs.analchem.0c03241](https://doi.org/10.1021/acs.analchem.0c03241)

\*Awarded with the 7<sup>th</sup> SIEMENS PAT Award for Young Scientists at EuroPACT 2021

**Publication VII.** Rocha de Oliveira, R., de Juan, A. **SWiVIA – Sliding window variographic image analysis for real-time assessment of heterogeneity indices in blending processes monitored with hyperspectral imaging.** *Analytica Chimica Acta* (2021), 1180: 338852.  
DOI: [10.1016/j.aca.2021.338852](https://doi.org/10.1016/j.aca.2021.338852)



# Design of Heterogeneity Indices for Blending Quality Assessment Based on Hyperspectral Images and Variographic Analysis

Rodrigo Rocha de Oliveira\* and Anna de Juan\*

Cite This: *Anal. Chem.* 2020, 92, 15880–15889

Read Online

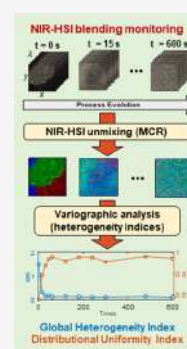
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** Heterogeneity characterization is crucial to define the quality of end products and to describe the evolution of processes that involve blending of compounds. The heterogeneity concept describes both the diversity of physicochemical characteristics of sample fragments (constitutional heterogeneity) and the diversity of spatial distribution of the materials/compounds in the sample (distributional heterogeneity, DH). Hyperspectral images (HSIs) are unique analytical measurements that provide physicochemical and spatial information on samples and, hence, are ideal to perform heterogeneity studies. This work proposes a new methodology combining HSI and variographic analysis to obtain a good qualitative and quantitative description of global heterogeneity (GH) and DH for samples and blending processes. An initial step of image unmixing provides a set of pure distribution maps of the blending constituents as a function of time that allows a qualitative visualization of the heterogeneity variation along the blending process. These maps are used as seeding information for a subsequent variographic analysis that furnishes the newly designed quantitative global heterogeneity index (GHI) and distributional uniformity index (DUI), related to GH and DH indices, respectively. GHI and DUI indices can be described at a sample level and per component within the sample. GHI and DUI curves of blending processes are easily interpretable and adaptable for blending monitoring and control and provide invaluable information to understand the sources of the abnormal blending behavior.



Blending process monitoring and control is an essential operation in many industrial processes. Indeed, a good blend is the necessary ground to ensure many other quality attributes linked to physical and compositional properties of manufactured products. Understanding blending means understanding heterogeneity, with all the complex aspects encompassed by this concept. The theory of sampling (TOS) by Gy provided an excellent and renewed definition of heterogeneity.<sup>1,2</sup> TOS distinguishes between constitutional heterogeneity (CH) and distributional heterogeneity (DH). Whereas CH focuses on the diversity of physical and chemical properties that present individual fragments of the materials in a sample, DH is focused on the quality of spatial distribution of the different materials/compounds in the sample, that is, on how far they are of presenting an even distribution. Because the DH concept is very linked to spatial correlation, studying this heterogeneity side requires looking at the properties of neighboring fragments (increments).

Traditionally, blending was controlled by off-line analysis of material increments taken every certain time or, more recently, by on-line spectroscopic monitoring using diverse sensor typologies that provide a single spectrum (or few spectra) per sample.<sup>3–6</sup> In most of these studies, a good blend implies that a reference composition is achieved and gets stabilized in time. Bulk sample properties are thus controlled, but the spatial distribution side linked to a good blend is overlooked. Nowadays, hyperspectral imaging (HSI) techniques work attributing a spectrum to every individual pixel in the image and, thus, connect chemical and spatial information of samples.

Hence, HSI are excellent measurements for a deeper study of heterogeneity.<sup>7,8</sup>

The heterogeneity concept in TOS can be used to interpret this kind of information in HSI. Indeed, different heterogeneity aspects can be addressed focusing on the study of properties of individual pixels or drawing the attention to properties of neighboring pixels or neighboring pixel areas. It is very tempting associating the first approach with the concept of CH and the latter to the definition of DH. However, whereas looking at pixel areas or neighboring pixels will provide a good indication of DH, CH cannot be derived from the study of properties of individual pixels because every pixel in an HSI may offer information on one or more fragments of the material scanned. From now on and to be accurate, we will use the term global heterogeneity (GH) to design the heterogeneity information issued from the independent exam of individual pixel properties, which reflects both CH and DH, and the term DH to express the information coming from the analysis of neighboring pixels or pixel areas.

Even ignoring the TOS formulation, some attempts to use images to define the different heterogeneity aspects mentioned can be found. Thus, GH has often been defined using histograms

Received: July 30, 2020

Accepted: November 12, 2020

Published: November 25, 2020



derived from pixel image intensities or pixel concentration values issued from multivariate calibration models.<sup>9,10</sup> Or, approaches such as macropixel analysis, were connected to the definition of DH by studying properties of pixel neighborhood areas of different increasing sizes covering all scanned images.<sup>11–14</sup>

Within the TOS context, variographic analysis has been proposed to statistically study the influence of spatial correlation in heterogeneity.<sup>1,2,15</sup> A recent work has been published and monitors blending by using variographic analysis based on large field of view single spectroscopic measurements acquired as a function of time on the material circulating on a conveyor belt.<sup>16</sup> An attempt of using variographic analysis on HSI can be encountered but is limited to extract and interpret variogram parameters obtained after fitting the experimental variogram with models inspired in geostatistical theory.<sup>17</sup>

In our work, we have designed quantitative indices of GH and DH directly derived from the experimental HSI variograms and easy to be interpreted. The full data analysis pipeline incorporates the use of the multivariate curve resolution–alternating least squares (MCR–ALS) method on the raw image to compress HSI information and obtain the distribution maps of the pure compounds in the sample analyzed.<sup>18–20</sup> This step allows defining GH and DH per sample and also individually per compound, thus completing the heterogeneity description. The distribution maps are the seeding information to obtain the so-called GHI (global heterogeneity index) and the DUI (distributional uniformity index), related to GH and DH, respectively. When images are collected as a function of blending time, GHI and DUI curves provide a very good tool to understand the evolution of GH and DH along the blending process and can be potentially used for end-point blending detection or for blending control of end-products.

The indices designed are tested on simulated data and real in-house blending runs of pharmaceutical products monitored by NIR imaging. GHI and DUI curves have provided insight on the quality of the blending evolution and on the detection and characterization of blending faults at a sample and at a compound level. Although the blending runs mimic a batch process, the same methodology would apply to blending control of continuous processes.

## ■ EXPERIMENTAL SECTION

A process mimicking the blending of a solid pharmaceutical formulation was carried out using caffeine (CAF) and acetylsalicylic acid (ASA) as active pharmaceutical ingredients (APIs), both purchased at Sigma-Aldrich (a.r.), and sodium starch glycolate, Explotab (EXP) as an excipient, donated by JRS Pharma. Three batches were performed with API mass proportions of 10:1, 1:1, and 1:10 (ASA/CAF), named B1, B2, and B3, respectively. The mass fraction of EXP was kept at 15% in all batches. An approximate total mass of 0.8 g of the formulation was weighed in a 2-halves cylindrical capsule (23 mm diameter × 5 mm height). Before starting the blending process, an initial NIR HSI was collected at time = 0 s ( $t_0$ ) from the capsule containing the three segregated ingredients. The closed capsule was placed in a rotating device for mixing and a total of 11 NIR images at cumulative blending times of 15, 30, 45, 60, 120, 180, 240, 300, 480, and 600 s, were recorded per batch.

The images from the pharmaceutical mixture at each blending time have been acquired with a pushbroom NIR image acquisition system Specim FX17 by Spectral Imaging Ltd.,

Oulu, Finland, for industrial and laboratory use. The imaging system consists of a hyperspectral camera and a 20 cm × 40 cm scanning bed. From the raw signal provided by the camera, reflectance and related absorbance spectra were calculated as explained in Section 1 of the [Supporting Information](#).

The camera frame rate was set to 35 Hz and the scanning bed speed to 3.2 mm/s to keep an adequate aspect ratio of the image. The FX17 sensor exposure time was set to 2 ms according to the signal provided by the “white” reference to avoid saturated signals. Spectra were recorded in the 900–1700 nm NIR spectral range with a spectral resolution of 3.5 nm. The pixel size in all images is approximately 0.1 × 0.1 mm<sup>2</sup>.

To study the reproducibility of the proposed heterogeneity indices, several images were collected from the same sample with different sensor exposure times in different days, see Section 4 of the [Supporting Information](#) for detailed experimental description.

## ■ DATA TREATMENT

The data treatment is oriented to monitor the evolution of the sample heterogeneity during a blending process. Below, the step related to MCR–ALS analysis of blending images to obtain the distribution maps of the pure ingredients of the formulation and the subsequent use of these maps to obtain heterogeneity indices based on variographic analysis per component and per sample is described.

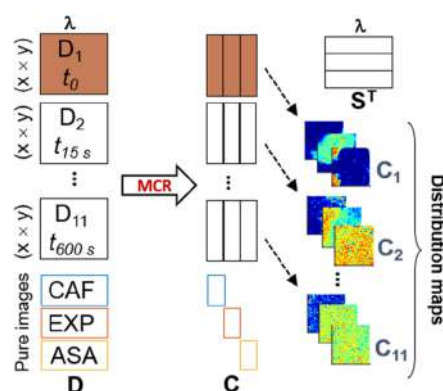
**HSI Unmixing. Image Preprocessing.** Before data analysis, a squared area (150 × 150 pixels) from the center of each image was cropped for further analysis, which represents a sample area of ca. 15 × 15 mm. The NIR spectra of the image were preprocessed using Savitzky–Golay first derivative (second order polynomial and window size of five points)<sup>21</sup> for baseline correction. See Figure S1 in the [Supporting Information](#).

**Multivariate Curve Resolution–Alternating Least Squares.** Image unmixing was performed with MCR–ALS, which provides the iterative decomposition of the preprocessed hyperspectral data (**D**) into concentration profiles, from which distribution maps can be derived (matrix **C**), and pure spectra (**S**<sup>T</sup>) of the sample constituents. Although a HSI dataset can be visualized as a three-dimensional (3D) data cube, where two dimensions ( $x$  and  $y$ ) are the pixel coordinates and the third is the spectral dimension ( $\lambda$ ), the data cube is unfolded into a two-dimensional (2D) matrix **D** with rows ( $x \times y$  pixels) and columns ( $\lambda$ ) that is decomposed according to the bilinear model in eq 1<sup>18,20,22</sup>

$$\mathbf{D} = \mathbf{CS}^T + \mathbf{E} \quad (1)$$

where **D** is the data matrix containing the preprocessed NIR pixel spectra and **C** and **S**<sup>T</sup> are the matrices with the concentration and spectral profiles of the pure components in the samples, respectively. **E** contains the variance not explained by the bilinear model, related to the experimental error. After the MCR–ALS resolution of the HSI dataset, the pure distribution maps of the image constituents can be obtained by folding back the stretched concentration profiles in **C** to recover the original 2D spatial structure of the image (see [Figure 1](#)).

The same bilinear model of MCR–ALS holds for multiset analysis, which consists of the simultaneous analysis of multiple images.<sup>19,20,23,24</sup> In this case, multiset structures **D** are built appending the submatrices **D**<sub>*i*</sub> linked to the pixel spectra of the images collected in the different blending steps and three additional matrices with spectra coming from images of the pure ingredients to help in the unmixing analysis, as shown in [Figure](#)



**Figure 1.** MCR–ALS analysis of an image multiset, where  $x$  and  $y$  are spatial pixels and  $\lambda$  represents the spectra wavelengths.

1. The decomposition of the multiset structure using eq 1 provides a single matrix  $S^T$  of pure spectra, valid for all the images analyzed, and a matrix  $C$ , formed by as many  $C_i$  submatrices as images in the data set. The profiles in each of these  $C_i$  submatrices can be appropriately folded back to recover the related distribution maps of the images recorded at the different blending times, see Figure 1.

The MCR–ALS algorithm requires an initial estimate of either  $C$  or  $S^T$  matrices to start the iterative optimization. In this work,  $S^T$  was estimated based on the selection of the purest pixel spectra<sup>25</sup> from the matrix  $D$ . The constrained ALS calculation of  $C$  and  $S^T$  was performed until convergence was reached.<sup>18–20</sup> The constraints used were normalization of pure spectra in  $S^T$  and non-negativity in the concentration profiles in  $C$ .

The correspondence among species constraint, which sets presence/absence of components in the different images, was applied to the pure component images in order to decrease ambiguity in the MCR solutions and provide more accurate results<sup>19</sup> (see Figure 1). The MCR analysis was carried out using an in-house GUI developed under MATLAB and related routines.<sup>26</sup>

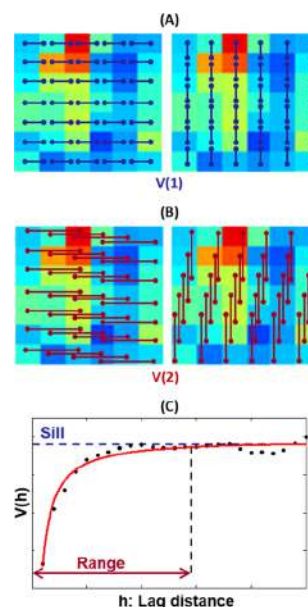
**Design of Heterogeneity Indices Based on Variographic Analysis of Images.** Heterogeneity can be very well-studied with variograms. A variogram displays the evolution of the variance as a function of a lag (expressed in time or distance units). In a variogram, the variance values are estimated by comparing pairs of observations separated at different lags.<sup>1,2,15</sup> Variograms can be easily adapted to explore correlation phenomena in 2D images (or 2D derived maps from 3D HSI) and, if needed, in 3D images formed by three spatial coordinates. For 2D image maps, experimental variograms are calculated comparing properties of pixel pairs separated a certain lag using the following equation

$$V(h) = \frac{1}{2} \cdot \frac{1}{N(h)} \sum_{i=1}^{N(h)} [c(x_i + h) - c(x_i)]^2 \quad (2)$$

where  $V(h)$  is the variance associated with the lag ( $h$ ), which is found as half of the average of the squared differences of all  $N(h)$  pairs of measured pixel values  $c(x_i + h)$  and  $c(x_i)$  separated by a lag distance ( $h$ ). Note that eq 2 expresses variance in absolute units. If the results need to be expressed in relative terms, the expression must be divided by the square of the average  $c$  value for all pixels in the image analyzed.

The variogram represents the variance estimated by the comparison of pixel pairs along the image as a function of the lag

distance among the pixels compared. In this work, the  $V(h)$  values of image variograms were calculated using concentration values ( $c$ ) extracted from distribution maps obtained from MCR–ALS and taking the lag distance in both vertical and horizontal directions of the square image as depicted in Figure 2.



**Figure 2.** Representation of the pixel pairs used for the variance calculation in lag distance (a)  $h = 1$  and (b)  $h = 2$ . (c) Representative variogram showing the extension of the correlation part, range, and the sill, linked conceptually to the variance in the absence of correlation.

Figure 2A,B displays the pairs of pixels compared to calculate  $V(1)$  and  $V(2)$ , variances associated with a lag  $h = 1$  and  $h = 2$ , respectively. Figure 2C shows the complete variogram obtained once  $V(h)$  values are calculated for all lags from 1 until the maximum lag distance, which is set to half the number of pixels of the squared image side, that is, 75 pixels in this study.

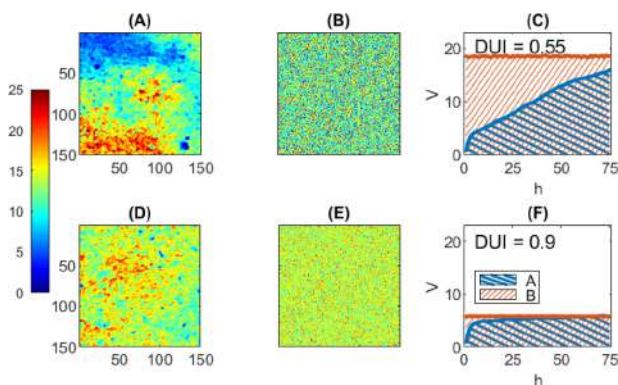
A representative shape for a variogram obtained from a 2D image is shown in Figure 2C. When a moderate level of GH exists, typical from pharmaceutical or alimentary mixtures, neighboring pixel pairs, with small lag  $h$ , are expected to present more similar properties than pixel pairs far away from each other; therefore, variance values will be smaller for low lag distances and will increase as the lag does, until a stabilization is reached, which indicates that correlation among pixel pairs does not exist anymore. The extension of the increasing part of the variogram is called the range and represents the lag distance in which there is correlation between the pixel pairs compared. Beyond that distance, there is no correlation anymore and the variance values get very similar to each other. The range defines the extension of the spatial correlation within the image and, therefore, relates to the DH. The sill is the maximum variance in a variogram, although technically in increasing variograms is often computed as the average of variance values.<sup>2</sup> The sill can be used as an estimate of the GH of the sample material.

In this work, the 2D distribution maps from MCR–ALS have been chosen for variogram calculations because heterogeneity can be estimated at individual component and at a sample level. However, the same approach could use as initial information predicted pixel ( $c$ ) values obtained from multivariate calibration models, pixel intensities from a specific spectral band, global



pixel intensities of all spectral range scanned, or any other type of measurement that represents a 2D map of the specific material or property to be analyzed.

To illustrate how variographic analysis can be used to study heterogeneity, Figure 3 shows the procedure followed to extract



**Figure 3.** Distribution map of EXP at the beginning (A) and in the middle (D) of a blending process (the colorbar refers to the MCR-derived pixel concentration values in the maps). (B,E) are the randomized maps from pixels in (A,D), respectively. (C) Overlapped variograms from maps in (A), blue curve, and (B), red curve. (F) Overlapped variograms from maps in (D), blue curve, and (E), red curve. Striped blue and striped red areas are the areas under the real map variogram curves and the randomized map variogram curves, respectively.

heterogeneity indices from two distribution maps related to real images at different blending times obtained in this work. Thus, Figure 3A,D show the distribution maps of the excipient, Explotab, at the beginning and in the middle of a blending batch, respectively. In Figure 3A, the excipient was not yet uniformly mixed, as seen by large clumps with high (in red) or very low concentration (in blue) of the substance. In Figure 3D, the excipient was found to be better mixed with the other ingredients, as shown by the small clumps present in the distribution map and the narrower range in the pixel concentration values. Figure 3C and 3F show the variograms (in blue) related to maps in Figure 3A and 3D, respectively. In Figure 3C a continuously increasing variogram is obtained, with a range beyond the maximum lag distance used. On the other hand, the variogram in Figure 3F shows a shorter range, around 35 pixel distance. There is also a clear difference in the sill of both variograms, with the largest values found for the variogram related to the map in Figure 3A, at the beginning of the blending, where the variance among pixel concentration values is larger.

From a qualitative point of view, it can be concluded that both GH, linked to the sill of the variogram, and DH, linked to the range, are higher for the map in Figure 3A than for that in Figure 3D. However, there is a need for a quantitative reference indicating how far from perfect mixing, that is, minimum DH, the material of each map is. To set this ideally mixed reference, the pixels of each map were randomized, as seen in Figure 3B and 3E for the maps in Figure 3A and 3D, respectively. These randomized maps have a double advantage: (a) the GH of the real material is preserved, that is, the pixel concentration values are the same as for the real map, and (b) there is a complete lack of correlation among pixel concentration values, that is, the situation that would happen when perfect mixing is achieved and no DH is present. Figure 3C and 3F show the variograms of the

randomized maps (in red) in Figure 3B and 3E, respectively. As expected, flat variograms with steady variance values are obtained for all lag distances  $h$  because the lack of correlation makes that neighboring pixels show concentration values as similar as those shown by pairs of pixels very distant from each other.

Looking at the variograms of the randomized maps, the sill of the flat variogram obtained from the randomized map in Figure 3B is higher than the map in Figure 3E because the variation in pixel concentration values is higher at the beginning of the blending process, that is, many pixels have very high or very low concentration values, whereas when the ingredients are better mixed, the pixel concentration values get more similar. This fact connects in a straightforward way with the expected decrease of GH during blending.

In addition, comparing the variograms of the original maps (in blue) with the related variograms of the randomized maps (in red), it can be observed that the shape of the real variogram gets closer to the shape of the flat variogram as blending progresses, that is, the two variograms are more similar at the middle of the process (Figure 3F), when the range for the real variogram gets shorter and variance stabilizes at earlier lags, than at the beginning of the blending (Figure 3C). This observation connects with the expected decrease of DH during blending.

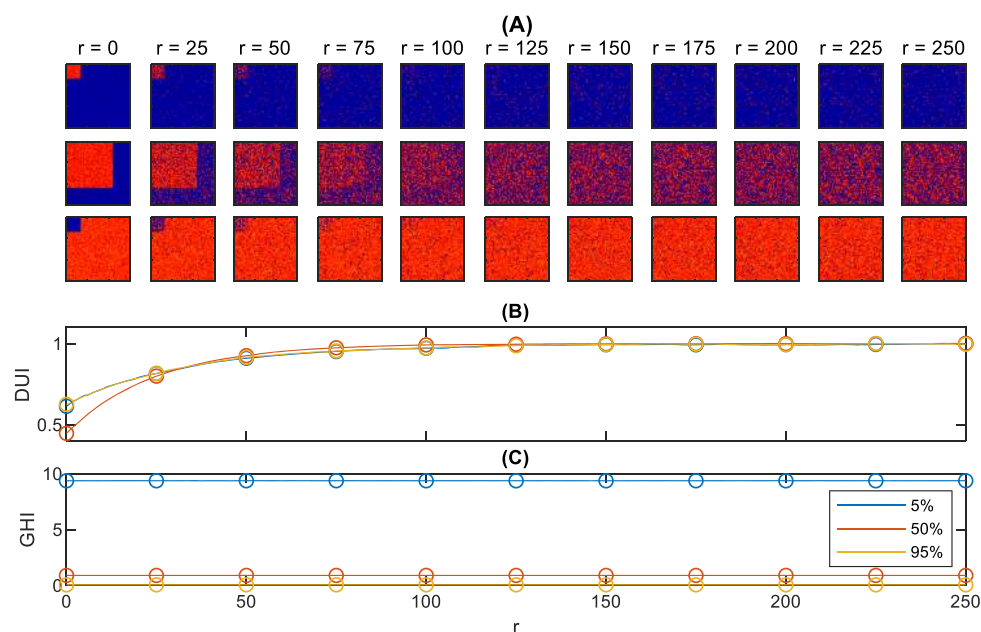
Based on the previous observations of the variographic analysis of images with different degree of mixing, two heterogeneity indices are proposed using information that can be extracted from the real and randomized variograms obtained from a component distribution map, namely:

- The GHI: estimated from the sill, that is, the average of the variance for all lags of the flat variogram from randomized maps. Actually, the sill of the flat variogram is an approximate estimation of the global variance of all pixel concentration values in the image.<sup>27</sup> Hence, GHI can be easily interpreted as the variance (absolute or relative) of pixel concentration values in the image.
- The DUI: estimated by calculating the ratio of the area of the variogram obtained from the real distribution map to the area of the flat variogram derived from the related randomized map

$$\text{DUI} = \frac{A}{B} \quad (3)$$

where  $A$  is the blue striped area under the variogram for the real map and  $B$  is the red striped area under the variogram for the randomized map (see Figure 3C,F).

The DUI can vary between 0 and 1 and allows quantifying the variation of DH based on variographic analysis. Experimental variograms far from their related flat horizontal variogram, as in Figure 3C, will give low DUI values, indicating high DH. Instead, experimental variograms close to its randomized map variogram provide DUI values close to 1 indicating that the mixture has low DH. In this case, the DUI values for the maps in Figure 3A and 3D are 0.55 and 0.9, respectively, meaning that 55 and 90% of ideal mixing is reached, respectively. It is relevant to note that the DUI value changes depending on the extent of the lag scale, that is, looking at Figure 3D,F, it is easy to see that if the lag scale had a limit lower than 75, the DUI values would be lower and if the images and related variograms had extended until a longer lag scale limit, the DUI values would be smaller. This means that the lag scale should adjust to the spatial scale level of heterogeneity that needs to be studied. Far from being a disadvantage, this



**Figure 4.** Simulated system I. (a) Simulated maps before mixing,  $r = 0$  and at different blending steps until  $r = 250$  for a bulk abundance of compound of interest equal to 5% (top), 50% (middle), and 95% (bottom); (b) DUI and (c) GHI curves calculated using the proposed variographic analysis of the generated maps for all steps. Circles represent the heterogeneity indices related to the maps shown above.

means that, if needed, the DH index can be studied at different spatial scale levels.

Both GHI and DUI indices can be obtained from variograms of individual components, but also total indices can be calculated for several selected components or all components of the formulation together. In this case, the indices are estimated from the total variogram obtained by averaging the variograms of the individual components using the following equation

$$V_T(h) = \frac{1}{nc} \sum_{j=1}^{nc} V_j(h) \quad (4)$$

where  $V_T(h)$  is the variance at lag  $h$  for the total variogram,  $V_j(h)$  the variance at lag  $h$  for the variogram of component  $j$  and  $nc$  is the total number of components considered in the calculation. In this work, the proposed heterogeneity indices have been calculated using the distribution maps obtained at different times of each blending process per component and per total formulation.

The evolution of these indices has been used to follow the progress of the blending process and understand faults related to it.

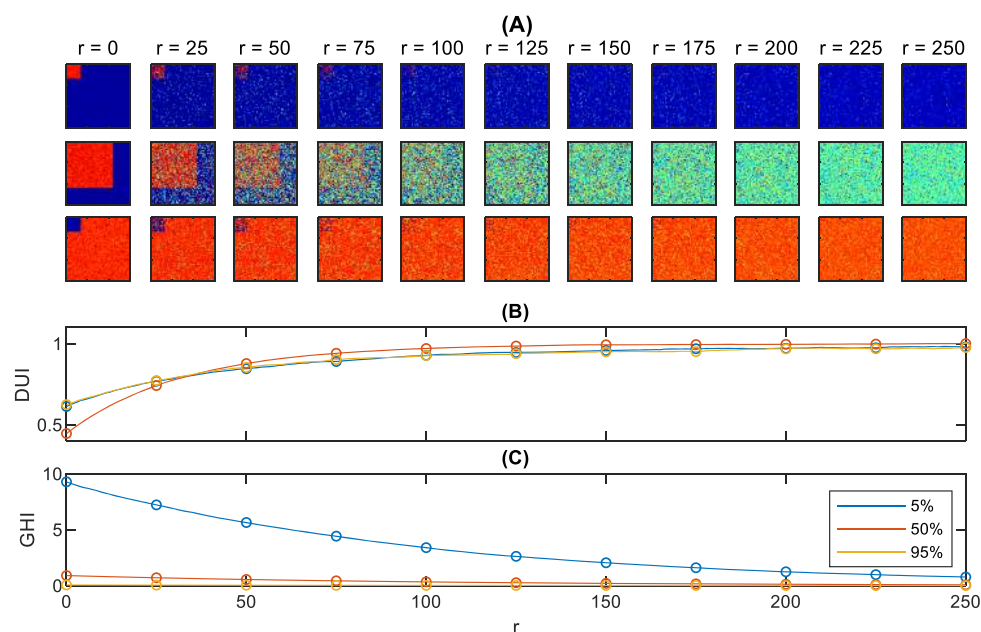
## RESULTS AND DISCUSSION

**Simulated Systems.** For a better understanding of the heterogeneity indices proposed, the approach has been tested in two systems that mimic the evolution of the distribution map of an individual compound from a binary mixture during blending. In system I (see maps in Figure 4A), it is assumed that the pixel size is equal to the fragment size of the compound. The blending simulation is carried out so that the fragments of the compound of interest only change position, but the pixel concentration keeps invariant (meaning that every pixel contains the compound or not). Top, middle, and bottom lines in Figure 4A indicate situations where the compound of interest have a

bulk abundance of 5, 50, or 95% in the sample. The color code in the maps is red for pixels with 100% abundance of the compound and dark blue when the abundance is 0%. A certain amount of noise has been added to the concentration values (see Section 2 of the Supporting Information for more details on the simulation).

Figure 4B shows the evolution of the DUI curves for the maps in Figure 4A. Note that the increase of the DUI index matches perfectly well the change in the spatial distribution of the compound of interest (in red). The closer the maps to a uniform distribution, the highest the DUI index. In all three systems, DUI values very close to one (ideal mixing) are found when the map is shown at  $r = 125$  or higher. It is also interesting to note that DUI curves when the compound of interest is at a 5% of abundance or 95% of abundance are almost identical. This clearly proves that the DUI index, as mentioned in the data treatment section, only relates to variations in the distributional pattern of compounds, not to their concentration level (it is a concentration-independent index). Bearing this in mind, the initial maps at 5% of abundance and 95% of abundance show an identical spatial pattern, with a big 95% area with similar concentrations (either low at the 5% abundance map or high at the 95% abundance map) and a small 5% zone different from the rest. From a spatial point of view, the initial situation when the abundance of the compound of interest is 50% is worse than the previous ones because two big different regions in the map are present; hence, the lower initial DUI value.

Figure 4C shows GHI indices calculated in relative scale. As mentioned in the data treatment section, this index is calculated from the sill of the variogram of the randomized map of the image. The reason why the GHI index remains invariant during all blending processes in Figure 4A is due to the nature of the blending simulation. Remember that in this case, mixing was simulated by changing the pixel positions in the map, but not their concentration values. Hence, the variance of the pixel concentration values in the randomized maps at all blending



**Figure 5.** Simulated system II. (A) Simulated maps before mixing,  $r = 0$  and after every 25 mixing steps interval until  $r = 250$  for a bulk abundance of compound of interest equal to 5% (top), 50% (middle), and 95% (bottom); (B) DUI and (C) GHI curves calculated using the proposed variographic analysis of the generated maps for all steps. Circles represent the heterogeneity indices related to the maps shown above.

stages, represented by the sill of the variogram, is identical because the pixel concentration values are the same (see the invariance of the histograms of the distribution maps in Figure S4 of the Supporting Information). Note that the GHI value derived from the variogram sill is practically identical to the variance of the pixel concentration values of the distribution map studied, as shown in Figure S4 of the Supporting Information. Because GHI is given here in relative terms, the GHI is higher when the compound of interest is minor and decreases as its abundance increases.

System II shows the scenario where the pixel size can enclose several fragments of the material. As a consequence, when blending progresses, the fragments of the compound of interest change position, but a single pixel can contain fragments of the two different compounds in the binary mixture and, hence, the pixel concentration of the compound of interest can acquire different values from 0 to 100% depending on the proportion of fragments present in the pixel. As in system I, top, middle, and bottom lines in Figure 5A indicate situations where the compound of interest have a bulk abundance of 5, 50, or 95% in the sample (see Section 2 of the Supporting Information for more details in the simulation).

It is important to note in maps of Figure 5A that not only the distributional pattern gets more uniform as blending progresses, but also the pixel concentration range is reduced (see the evolution of the histograms of the distribution maps in Figure S5 from the Supporting Information). These two phenomena reflect in the DUI and GHI curves, respectively.

Figure 5B shows the DUI curves for the blendings in 5A and they are very similar to those shown in system I (Figure 4B) because the modification in distributional pattern of the blending has been done in the same manner. Again, curves for 5 and 95% compound abundance are very similar and differ from the 50% compound abundance blending.

GHI curves in Figure 5C show a clear change with respect to those in Figure 4C. Because blending causes that the pixel

concentration range narrows, the variance associated with pixel concentration values, reflected by the sill of the variograms of the randomized maps, decreases and so does the GHI index. Note again that the GHI values derived from the variogram sill agree with the variances of the pixel concentration values of the distribution maps studied, as shown in Figure S5 of the Supporting Information. The decrease in these relative GHI indices happens for the three blendings studied, but the decay can be more clearly seen when the compound of interest is in a minor proportion.

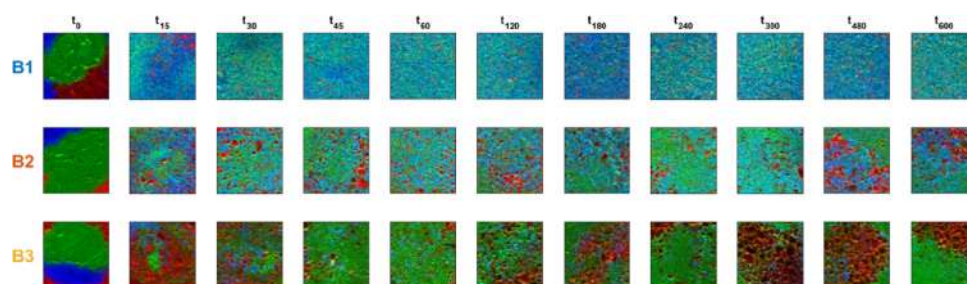
The increase of the DUI curve and the decrease of the GHI curve is the behavior expected for real blending processes monitored by imaging when mixing proceeds in a correct way. Deviations from this behavior are indications of blending problems in the formulation studied or in individual compounds.

**Real Blending Processes.** The real blending processes studied correspond to the scenario simulated in system II, where the pixel size is clearly bigger than the fragment size of the different materials in the formulation.

Note that each of the images recorded provides a number of pixel spectra large enough to derive reliable statistical indicators and, besides, covers a sample area slightly higher than a pill size ( $15 \times 15$  mm<sup>2</sup>). This means that the lag scale in the variograms will adjust to the spatial level of heterogeneity that needs to be studied.

As mentioned in the Data Treatment section, the study of the real blending processes first requires an unmixing step to obtain the pure distribution maps of the compounds in the formulation, followed by the computation of the GHI and DUI curves associated with the formulation and with each of their individual compounds.

**Unmixing of NIR-HSI Data and Qualitative Evaluation of Blending Evolution.** NIR HSI unmixing by MCR-ALS was carried out on a multiset structure containing a total of 36 preprocessed images, structured as a column-wise augmented



**Figure 6.** Combined RGB maps with overlaid pure component distribution maps obtained with MCR–ALS for batches B1 (top row), B2 (middle row), and B3 (bottom row). Red—CAF, green—EXP, and blue—ASA.

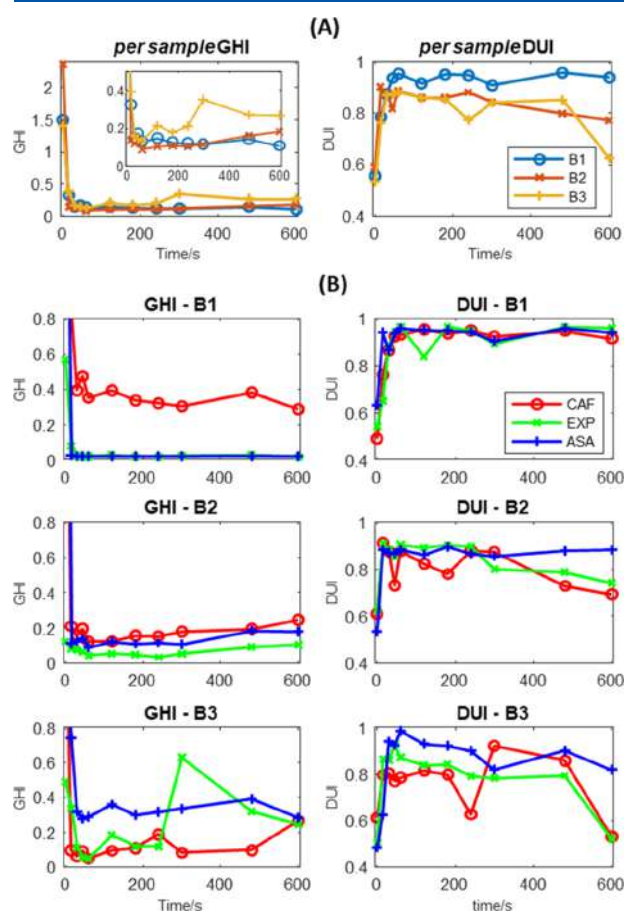
matrix. The multiset structure analyzed was formed by the preprocessed images of the three batches B1, B2, and B3, at 11 blending times each, and the three images of the pure ingredients (ASA, CAF, and starch, EXP). The pure resolved spectral profiles of the mixture formulation ingredients are shown in Figure S6. The evolution of the blending process can be qualitatively assessed by observing the MCR–ALS pure component distribution maps from the NIR HSI. Figures S7–S9 in the Supporting Information show the evolution of the distribution maps of the three ingredients, CAF, ASA, and EXP, at the 11 blending times for batch B1 (10:1—ASA/CAF), B2, and B3, respectively.

Figure 6 shows combined RGB maps overlaying the information of the three pure component maps (red for CAF, green for EXP, and blue for ASA) for batches B1, B2, and B3 in the first, second, and third row of the figure, respectively. In the blending evolution of batch B1 (10:1—ASA/CAF), the RGB map at  $t_0$  shows the segregated ingredients before blending started. The succeeding maps,  $t_{15}$  and  $t_{30}$ , show the decrease of the segregation level, but still some clumps of pure ingredients are visible. After consecutive blending steps, from  $t_{45}$  onward, all three components were visually more evenly distributed in the imaged area. For batch B2,  $t_0$  shows the segregated ingredients and a decrease of the segregation level is observed in distribution maps from consecutive blending times. However, at long blending times, the ingredients start to segregate, probably due to overmixing, as can be observed by the visible large clumps of different ingredients in the last two maps, at  $t_{480}$  and  $t_{600}$ . Last, batch B3 (1:10—ASA/CAF) starts with the segregated situation at  $t_0$  and a certain blending improvement in the immediate blending times. However, from blending time  $t_{120}$  and beyond, an increase of segregation was observed mainly because of the formation of large granules of pure CAF, the major ingredient of this formulation. It is also interesting to note that the segregation behavior is different in every component. Thus, when visualizing maps from batches B2 and B3, clumps are generally associated with CAF (in red) and starch (in green), whereas ASA (in blue) seem to show a more even distribution (see maps in Figures S3–S5 for more clarity).

From a qualitative point of view, it can be observed that the blending quality decreases from batch B1 to B2, being B3 the worst blended batch. It is also seen that blending quality is compound-dependent. These different situations will be quantitatively confirmed using the heterogeneity indices proposed in this work.

**Blending Process Monitoring with Image Variogram-Derived Heterogeneity Indices.** In this section, the assessment of blending quality using the proposed indices related to GHI and DUI is presented.

First, a description of the blending quality of the formulation for the three batches is provided. Thus, Figure 7A shows the per



**Figure 7.** (A) Per sample GHI and DUI curves for blending of batches B1, B2, and B3. Inset plot zooms per sample GHI values after  $t_0$ . (B) GHI and DUI curves per component for blending batches B1, B2, and B3. Left plots, GHI curves. Right plots, DUI curves. Note that some per component GHI values are outside y-axis scale at the beginning of the process.

sample heterogeneity indices GHI (left plot) and DUI (right plot) obtained from the total variograms (see eq 4) taking into account all formulation ingredients in the blendings. The evolution of the quantitative heterogeneity indices confirms the qualitative interpretation pointing out that batch B1 had a good blending evolution, whereas abnormal blending behaviors were detected in batches B2 and B3.

Thus, observing the per sample GHI curves in Figure 7A (left plot), a rapid drop of GHI is observed right after the blending of all batches started. That was an expected behavior considering that the formulation ingredients were completely segregated in the capsule before the start of the blending process. Per sample GHI kept decreasing for all batches until 60 s of blending time, when all batches reached a GHI, expressed in relative variance, below 0.13. After  $t_{60}$ , GHI stabilized for batch B1, but batches B2 and B3 showed an increasing trend, more visible and erratic for B3, see inset plot in Figure 7A (left plot). Per sample GHI at the end of each blending were 0.1, 0.18, and 0.26 for batches B1, B2, and B3, respectively.

Per sample DUI curves, Figure 7A (right plot), confirmed the visual interpretation of the spatial component distribution seen in maps of Figure 6B. Whereas the DUI curve had stabilized for batch B1 after the first minute of blending, a decreasing trend was observed for batches B2 and B3, with the most erratic behavior linked again to batch B3. The increase of segregation in the formulation ingredients for blends B2 and B3 shown in Figure 6B is reflected by the low DUI values obtained at the end of the blending. Indeed, the DUI value for batch B3 at  $t_{600}$  is almost as low as before the blending started at  $t_0$ .

The per sample DUI values obtained at the end of each blending at  $t_{600}$  were 0.94, 0.77, and 0.62 for batches B1, B2, and B3, respectively, meaning that the batch material reached 94, 77, and 62% of ideal mixing in the three batches.

To complement the heterogeneity description per formulation for all three batches in Figure 7A,B shows the per component heterogeneity indices for all batches studied, estimated as described in the data analysis section. In Figure 7B, the GHI curves for all compounds of batch B1 showed the same behavior as the formulation GHI curve, defined by a decrease and stabilization of GHI values. The higher GHI values obtained for CAF are related to its low concentration level in the B1 formulation, which resulted in a higher relative variance. The DUI curves of the different components of batch B1 showed that as the blending proceeded, the DH decreased and, consequently, DUI values increased for all components. Indeed, after 200 s of blending time, the DUI curves stabilized with a value circa 0.95 for all components of the formulation in B1. Both GHI and DUI curves show the expected evolution for a good blending behavior (as happened in Figure 5 for the simulated system II).

Figure 7B (middle row) shows GHI and DUI curves of batch B2 at the left and right plots, respectively. In this case, although the increasing trend of GHI after  $t_{60}$  was observed for all ingredients, slightly higher changes were associated with CAF and ASA. Regarding the DUI curves obtained per component in batch B2, Figure 7B shows that the decreasing trend of the DUI curve for the total formulation seen in Figure 7A was clearly associated only with CAF and EXP (the more even spatial distribution of ASA can be clearly seen in the individual distribution maps of this compound in Figure S8). Thus, while the DH of ASA kept stable and showed steady DUI values around 0.88 during most of the blending process, CAF and EXP decreased from a DUI value roughly equal to 0.9 at  $t_{60}$  to a value lower than 0.75 at  $t_{600}$ . This increase of DH, quantitatively represented by the decrease of the DUI value, matches the visual qualitative interpretation of CAF and EXP maps of batch B2 in Figure 6B.

Finally, for batch B3, Figure 7B (bottom left) shows that the GHI curve stabilized for the ASA component, although the value of the index remained high because of its low concentration level. Thus, the irregular behavior of the sample GHI curve for

batch B3 seen in Figure 7A is mainly due to CAF, with a steady increasing tendency, and EXP, with an erratic evolution reaching a maximum GHI value at  $t_{300}$ . Figure 7B (bottom right) shows the DUI curves for the three components in batch B3. Although an even spatial distribution is not fully achieved by any component, the segregation tendency is much more clearly associated with CAF and EXP than with ASA (see separate distribution maps of this compound in Figure S9 for further clarification). Indeed, the lowest DUI values at the end of batch B3 are obtained for CAF and EXP, even though these are the two major ingredients of this formulation. The clear irregular and decreasing tendency of the DUI curves of CAF and EXP matches the emergence of large clumps of these two compounds in their distribution maps at long blending times, particularly visible at  $t_{600}$  in red for CAF and in green for EXP, as seen in Figure 6 (bottom).

There are some interesting additional remarks linked to the plots observed. GHI values can be expressed in absolute or relative variance units. When linked to pixel concentration variation, GHI values in absolute variance scale would need maps issued from a calibration-based model, for example, PLS, to obtain a useful interpretation. GHI values in a relative scale allow working with maps derived from calibration-free methodologies, such as MCR. When working with GHI values in a relative scale, it should be reminded that high relative variance values may just appear because a minor compound is studied. This effect is clearly seen in GHI curves of ASA (in blue), the compound that tends to have the best blending in all batches, where the magnitude of GHI values increases from B1 to B2 to B3, matching the decreasing content of this compound in the three batch formulations. Therefore, interpretation should not be focused only on the GHI value, which is concentration scale-dependent, but on the evolution tendency of GHI, that is, whether it gets stabilized during blending or presents an increasing or irregular tendency.

This is not the case for DUI values, which only refer to the spatial distribution pattern of compounds. In this case, compounds present in different concentration levels may reach very similar and equally good DUI values when blending is correct (see the case of the DUI curves of ASA, CAF, and EXP in batch B1, where the ratio ASA/CAF is 10:1). DUI curves, as mentioned before and proven in the simulated blendings, do not suffer from scale-dependency and can be interpreted looking both at the DUI values obtained and at the shape of the curve.

As seen throughout this work, the good performance and easy interpretability of the heterogeneity indices proposed has been proven in simulated and real blending systems. From an analytical point of view, the robustness of the quantitative values of the indices proposed has also been tested and is described in detail in Section 4 of the Supporting Information. To do so, images from the same sample showing a mixture of the same composition as B1, collected at different exposure times and in different days have been acquired. The per sample and per component values of the GHI and DUI indices show a very good reproducibility, as seen in Figures S10 and S11 of Section 4 in the Supporting Information. It is interesting to note that DUI indices are particularly stable because they are obtained from area ratios between variograms of real and randomized maps and all variability contributions other than the distributional pattern of the material are cancelled out. In the case of GHI, satisfactory values are obtained with slightly bigger fluctuations in minor compounds than in major compounds, as expectable in any analytical parameter.

To conclude, both GHI and DUI indices are needed to describe properly the blending behavior because they focus on global and DH, respectively. When blending evolution is good for all sample constituents, DUI curves reaching high and stable values tend to go with GHI curves that also remain low and stable, as happens in batch B1. However, many other situations can be encountered where there is no synchronicity between the evolution of GHI and DUI curves and strong variations in DUI values do not lead to clear changes in GHI values or vice versa, as seen in batches B2 and B3. Likewise, a complete heterogeneity description should join a per sample and a per component description because the different sample constituents do not necessarily show the same heterogeneity pattern.

## CONCLUSIONS

HSI followed by image unmixing and variographic analysis provides an excellent combination to describe the global heterogeneity and DH in samples and the dynamic evolution of these attributes in blending processes. Indeed, a first visual qualitative description of heterogeneity can be extracted from the distribution maps retrieved by MCR–ALS, whereas the quantitative estimation of GH and DH is achieved through the proposed GHI and DUI, respectively.

The design of the GHI and DUI indices allows heterogeneity descriptions at a sample and component level. The assessment of blending evolution using per sample heterogeneity indices is appropriate to see the overall process evolution and to detect possible abnormal behaviors. Per sample GHI and DUI values can also be adopted as quantitative criteria to define blending quality or blending end-point by setting threshold values that need to be reached to stop a blending process, for example, a desired relative variance level for GHI and/or a preset percentage of ideal mixing for DUI. Although these indices have been tested in batch blending processes, their use can be directly transferred to monitoring and control of continuous blending operations.

The use of unmixing methods on the collected images provides maps that allow a per component description of heterogeneity through GHI and DUI indices that reflect appropriately the individual behavior of the sample or blending constituents. This individual description of heterogeneity offers additional advantages, such as a higher flexibility in blending monitoring and control protocols, for example, if there is only a single or some critical components in a blending process that need to be controlled, and contributes to a better understanding of the sources of abnormal global blending behaviors.

In general, the methodology proposed provides a good qualitative and quantitative description of heterogeneity for any kind of sample and for monitoring and control of processes that involve heterogeneity variations, such as blending operations.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.0c03241>.

NIR imaging preprocessing steps and complete visual output of MCR–ALS analysis of the blending batches studied; thorough description of the simulated systems; and measurements and results associated with the study of the robustness of the quantitative values of GHI and DUI indices (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

**Rodrigo Rocha de Oliveira** – Chemometrics Group, Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, 08028 Barcelona, Spain; [orcid.org/0000-0002-4309-5236](https://orcid.org/0000-0002-4309-5236); Email: [rodrigo.rocha@ub.edu](mailto:rodrigo.rocha@ub.edu)

**Anna de Juan** – Chemometrics Group, Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, 08028 Barcelona, Spain; [orcid.org/0000-0002-6662-2019](https://orcid.org/0000-0002-6662-2019); Email: [anna.dejuan@ub.edu](mailto:anna.dejuan@ub.edu)

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.analchem.0c03241>

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

R.R.d.O. acknowledges the research contract linked to the EU Horizon 2020 funds from the ProPAT project, grant no. 637232. Funding from Spanish government under the project PID 2019-1071586B-IOO is also acknowledged. The authors belong to the Catalan excellence research group (2017 SGR 753).

## REFERENCES

- (1) Gy, P. *Sampling for Analytical Purposes*; Wiley: West Sussex, U.K., 1998.
- (2) Esbensen, K. H. *Introduction to the Theory and Practice of Sampling*; IM Publications Open, 2020.
- (3) Razuc, M.; Grafia, A.; Gallo, L.; Ramirez-Rigo, M. V.; Romañach, R. J. *Drug Dev. Ind. Pharm.* **2019**, *45*, 1565–1589.
- (4) Li, Y.; Anderson, C. A.; Drennen, J. K.; Airiau, C.; Igne, B. *Anal. Chem.* **2018**, *90*, 8436–8444.
- (5) Durão, P.; Fauteux-Lefebvre, C.; Guay, J.-M.; Abatzoglou, N.; Gosselin, R. *Talanta* **2017**, *164*, 7–15.
- (6) Igne, B.; de Juan, A.; Jaumot, J.; Lallemand, J.; Preys, S.; Drennen, J. K.; Anderson, C. A. *Int. J. Pharm.* **2014**, *473*, 219–231.
- (7) Hyperspectral Imaging. In *Data Handling in Science and Technology*; Amigo, J. M., Ed.; Elsevier, 2020; Vol. 32.
- (8) Ma, H.; Anderson, C. A. *J. Pharm. Sci.* **2008**, *97*, 3305–3320.
- (9) Piqueras, S.; Burger, J.; Tauler, R.; de Juan, A. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 169–182.
- (10) Burger, J.; Geladi, P. *Analyst* **2006**, *131*, 1152–1160.
- (11) Hamad, M. L.; Ellison, C. D.; Khan, M. A.; Lyon, R. C. *J. Pharm. Sci.* **2007**, *96*, 3390–3401.
- (12) Rosas, J. G.; Blanco, M. J. *Pharm. Biomed. Anal.* **2012**, *70*, 680–690.
- (13) Sacré, P.-Y.; Lebrun, P.; Chavez, P.-F.; Bleye, C. D.; Netchacovitch, L.; Rozet, E.; Klinkenberg, R.; Streel, B.; Hubert, P.; Ziemons, E. *Anal. Chim. Acta* **2014**, *818*, 7–14.
- (14) de Moura França, L.; Amigo, J. M.; Cairós, C.; Bautista, M.; Pimentel, M. F. *Chemometr. Intell. Lab. Syst.* **2017**, *171*, 26–39.
- (15) Esbensen, K. H.; Friis-Petersen, H. H.; Petersen, L.; Holm-Nielsen, J. B.; Mortensen, P. P. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 41–59.
- (16) Sánchez-Paternina, A.; Sierra-Vega, N. O.; Cárdenas, V.; Méndez, R.; Esbensen, K. H.; Romañach, R. J. *Comput. Chem. Eng.* **2019**, *124*, 109–123.
- (17) Herrero-Langreo, A.; Gorretta, N.; Tisseyre, B.; Gowen, A.; Xu, J.-L.; Chaix, G.; Roger, J.-M. *Anal. Chim. Acta* **2019**, *1077*, 116–128.
- (18) de Juan, A.; Rutan, S. C.; Tauler, R. Two-Way Data Analysis: Multivariate Curve Resolution, Iterative Methods. In *Comprehensive Chemometrics*; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier, 2019; pp 153–171.
- (19) Tauler, R.; Maeder, M.; de Juan, A. Multiset Data Analysis: Extended Multivariate Curve Resolution. In *Comprehensive Chemo-*

*metrics*; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier, 2020; Vol. 2, pp 305–336.

(20) de Juan, A.; Tauler, R. *Anal. Chim. Acta* **2020**, DOI: 10.1016/j.aca.2020.10.051.

(21) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627–1639.

(22) de Juan, A. Multivariate Curve Resolution for Hyperspectral Image Analysis. In *Hyperspectral Imaging; Data Handling in Science and Technology*; Amigo, J. M., Ed., 2020; Vol. 32, pp 115–150.

(23) Piqueras, S.; Duponchel, L.; Tauler, R.; de Juan, A. *Anal. Chim. Acta* **2014**, *819*, 15–25.

(24) de Juan, A.; Gowen, A.; Duponchel, L.; Ruckebusch, C. Image Fusion. In *Data Handling in Science and Technology*; Elsevier, 2019; Vol. 31, pp 311–344.

(25) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425–1432.

(26) Jaumot, J.; de Juan, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 1–12.

(27) Gy, P. *Chemom. Intell. Lab. Syst.* **2004**, *74*, 39–47.



Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: [www.elsevier.com/locate/aca](http://www.elsevier.com/locate/aca)

# SWiVIA – Sliding window variographic image analysis for real-time assessment of heterogeneity indices in blending processes monitored with hyperspectral imaging

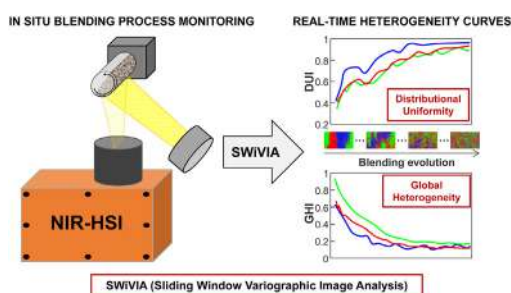
Rodrigo Rocha de Oliveira<sup>\*</sup>, Anna de Juan<sup>\*\*</sup>

Chemometrics Group, Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, Diagonal 645, 08028, Barcelona, Spain

## HIGHLIGHTS

- Novel PAT tool for real-time image-based blending quality assessment.
- Heterogeneity indices are continuously obtained from image variographic analysis.
- Indices are linked to distributional (DUI) and global (GHI) heterogeneity.
- Heterogeneity DUI and GHI curves allow blending process understanding and control.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

### Article history:

Received 5 May 2021

Received in revised form

8 July 2021

Accepted 12 July 2021

Available online 16 July 2021

### Keywords:

Heterogeneity indices

Variographic analysis

Blending

Hyperspectral imaging

## ABSTRACT

Controlling blending processes of solid material using advanced real-time sensing technologies tools is crucial to guarantee the quality attributes of manufactured products from diverse industries. The use of process analytical technology (PAT) tools based on chemical imaging systems are useful to assess heterogeneity information during mixing processes. Recently, a powerful procedure for heterogeneity assessment based on the combination of off-line acquired chemical images and variographic analysis has been proposed to provide specific heterogeneity indices related to global and distributional heterogeneity. This work proposes a novel PAT tool combining in situ chemical imaging and variogram-derived quantitative heterogeneity indices for the real-time monitoring of blending processes. The proposed method, so called sliding window variographic image analysis (SWiVIA), derives heterogeneity indices in real-time associated with a sliding image window that moves continuously until the full blending time interval is covered. The SWiVIA method is thoroughly assessed paying attention at the effect of relevant factors for continuous blending monitoring and heterogeneity description, such as the scale of scrutiny needed for heterogeneity definition or the blending period defined to set the sliding image window. SWiVIA is tested on blending runs of pharmaceutical and food products monitored with an in situ near-infrared chemical imaging system. The results obtained help to detect abnormal mixing phenomena and can be the basis to establish blending process control indicators in the future. SWiVIA is adapted to study blending behaviors of the bulk product or compound-specific blending evolutions.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

<sup>\*</sup> Corresponding author.

<sup>\*\*</sup> Corresponding author.

E-mail addresses: [rodrigo.rocha@ub.edu](mailto:rodrigo.rocha@ub.edu) (R. Rocha de Oliveira), [anna.dejuan@ub.edu](mailto:anna.dejuan@ub.edu) (A. de Juan).

<https://doi.org/10.1016/j.aca.2021.338852>

0003-2670/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

The blend quality of solid materials is key to guarantee the



quality attributes of manufactured products by industries such as pharmaceutical, food, plastic, and ceramic [1–3]. Lack of blend quality affects not only the uniformity of the product composition, but other physical and chemical properties of the final product. In parallel to the intense development of computational modeling methods devoted to understand the mechanisms involved during mixing [4–6], the blending research field needs the development of process analytical technology (PAT) including the use of advanced real-time sensing technologies and multivariate analysis tools [7,8] to properly interpret the blending evolution. Different advanced real-time sensors combined with PAT tools have been developed recently for the efficient optimization, monitoring and control of blending processes [9–13]. However, in most of these studies, the sensor provides a single-point spectrum per measurement ignoring the spatial distribution of the blending material. Some developments have been proposed to tackle this limitation of local probe measurements by using several spectroscopic probes attached at different positions of the blender vessel [14] or working with imaging sensors [15–17]. Imaging sensors offer the advantages of increasing the mass of material probed and, most importantly, are able to provide information on the spatial distribution of the blending material.

Among the different PAT imaging options for blending monitoring, there is a growing interest in the use of hyperspectral imaging (HSI) techniques. Specially, push-broom near-infrared hyperspectral image (NIR-HSI) systems, offering a fast image acquisition, are deemed suitable for on-line monitoring in industrial environments [18]. Hyperspectral images, also called chemical images, are formed by a large number of spectra, each of them associated with an individual sample pixel. In this way, they connect chemical and spatial information of the measured sample and provide excellent information to study the heterogeneity evolution in mixing processes.

Two types of heterogeneity information can be extracted from chemical images, the global heterogeneity (GH), related to the independent exam of individual pixel properties, and the distributional heterogeneity (DH), that expresses the information coming from the analysis of neighboring pixels or pixel areas, linked to the evenness in the spatial distribution of the different materials forming a blend. Classical assessment of GH information has often been defined using histogram-derived parameters from pixels of the chemical image [19,20]. Different strategies have been proposed to assess the DH from chemical images, based on the study of the domain size of clumps of pure material in the mixture [21,22], texture analysis [23], macropixel analysis [24] or variographic analysis [25]. Recent works have defined quantitative DH indices based on some of the strategies mentioned above for off-line blending monitoring or for the analysis of end products [25–27] that could be potentially extended to real-time monitoring of blending processes.

In this study, we propose a new PAT tool for real-time monitoring of blending processes combining an in situ push-broom NIR-HSI system with a suitable data analysis pipeline to continuously provide variogram-derived GH and DH heterogeneity indices [25] that depict adequately the blending evolution. A first step, based on the use of the non-negativity-constrained least squares (FNNLS) algorithm [28], is used to compress and transform in real-time the initial HSI information into distribution (concentration) maps related to each of the compounds in the blending process. These distribution maps are the input information for the proposed SWiVIA (Sliding Window Variographic Image Analysis) method. SWiVIA works using a fixed size sliding window that moves across the increasing maps obtained during blending. The map area covered by the window is used to derive the variogram-related GH and DH indices related to a particular blending time. Every time the

sliding window moves one point ahead in blending time, a new set of heterogeneity indices is obtained until the full blending time interval is covered. The SWiVIA method is used to continuously generate the variogram-derived heterogeneity indices GHI (Global Heterogeneity Index) and DUI (Distributional Uniformity Index), related to GH and DH, respectively [25]. The compound maps, obtained from the in-line push broom NIR-HSI system and the related SWiVIA-derived heterogeneity indices, organized in GHI and DUI curves, provide a very good tool to understand the evolution of GH and DH along the blending process. As described in a previous work, GHI and DUI indices can describe the blending evolution *per component* or *per total mixture*.

The proposed method is tested on several blending runs of pharmaceutical and food products performed in an in-house rotating blending device monitored with an in situ NIR chemical imaging system. First, a description of the effect of the scale of scrutiny and the blending time covered by the sliding window on the evolution of GHI and DUI curves is provided. Afterward, the SWiVIA-derived GHI and DUI curves obtained from several blending runs are described. From them, differences in blending behavior among runs and among compounds can be clearly seen and detection of de-mixing, an undesired blending-related phenomenon that can be induced by excessive blending [29,30], is also shown.

Although the blending runs were carried out using a small-scale blending device, the same methodology can be applied to in-line monitoring of batch or continuous industrial-scale blenders and can be easily adapted to the seeding information provided by other machine vision systems. The PAT tool proposed can be potentially used for the detection of blending completion in batch blending processes, feedback control of continuous blending processes, the characterization of new blending devices, or the study of the blending behavior of new formulations.

## 2. Experimental

### 2.1. Materials

To demonstrate the applicability of the proposed methodology for continuous monitoring of blending processes, several blending batches of solid material were studied in this work. The materials present in the mixtures to be blended consisted of pharmaceutical compounds, food products and plant seeds with diverse physical properties, such as particle size, particle shape or density. Considering that the seeds used are edible, the materials were organized into two main categories:

- A) Food materials: Rice grits (RG), ground coffee (GC), quinoa seeds *Chenopodium quinoa* (QS) and common poppy seeds *Papaver rhoeas* (PS), all commercial products.
- B) Pharmaceutical compounds: Caffeine (CAF) and acetylsalicylic acid (ASA) from Sigma-Aldrich, a.r., citric acid (CA) from Merck, and sodium starch glycolate (SSG) from JRS Pharma.

Information about bulk density and particle size of the materials are shown in Table 1 below.

### 2.2. Blending batches

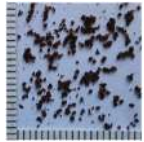

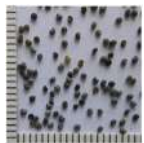


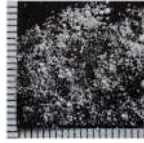

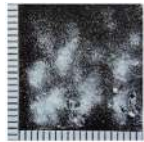
Ten blending batches were carried out using binary and ternary mixtures of the materials described above, trying to reflect combinations of materials with different physical properties. Table 2 summarizes the information about the composition and experimental settings used in the blending runs carried out in this work.

Each blending run was carried out in a lab-scale horizontal

rotary blender, which consisted of a 40-mL glass vial (h = 95 mm and O.D. = 27 mm) attached directly to a stepper motor that enabled rotation of the vial around its longitudinal axis with a constant speed. The stepper motor was controlled with the aid of a microcontroller and stepper motor driver connected to a PC through an in-house developed control interface. This system was mounted on an in-house-built 3D printed stand attached to the NIR camera system. The position of the vial was adjusted so that the bottom part was 7 cm above the NIR camera objective lens, see Fig. 1.

The mixtures for the blending batches were prepared by introducing separate horizontal layers of each material inside the vial. The volume of the initial mixture of all segregated solid ingredients in a batch amounted to approximately 20 mL, leaving sufficient space for the material to freely move inside the vial during the blending run. Once all ingredients of the batch were introduced, the vial was capped and attached to the stepper motor. The blending experiment started right after the NIR acquisition system was initialized until the total blending time set, see Table 2. All blending runs were performed at a rotational speed of 25 rpm, i.e. one

**Table 1**  
Properties and picture with 1 mm reference scale of the material used to prepare the blending runs.

Category	Material	Bulk density ( $\text{g mL}^{-1}$ )	Particle size (mm)	Picture (1 mm scale)
Food	Ground coffee (GC)	0.350(0.003) <sup>a</sup>	<0.5 <sup>c</sup>	
	Rice grits (RG)	0.763(0.007)	1 <sup>b</sup>	
	Poppy seeds (PS)	0.616(0.008)	1 <sup>b</sup>	
	Quinoa seeds (QS)	0.75(0.02)	2.3 <sup>b</sup>	
Pharma	Acetylsalicylic acid (ASA)	0.78(0.01)	1 <sup>b</sup>	
	Caffeine (CAF)	0.77 (0.02)	<0.5 <sup>c</sup>	
	Citric acid (CA)	0.87(0.04)	1 <sup>b</sup>	
	Sodium starch glycolate (SSG)	0.817(0.006)	<0.106 <sup>d</sup>	

<sup>a</sup> Standard deviation from triplicate measurements of bulk density in parenthesis.

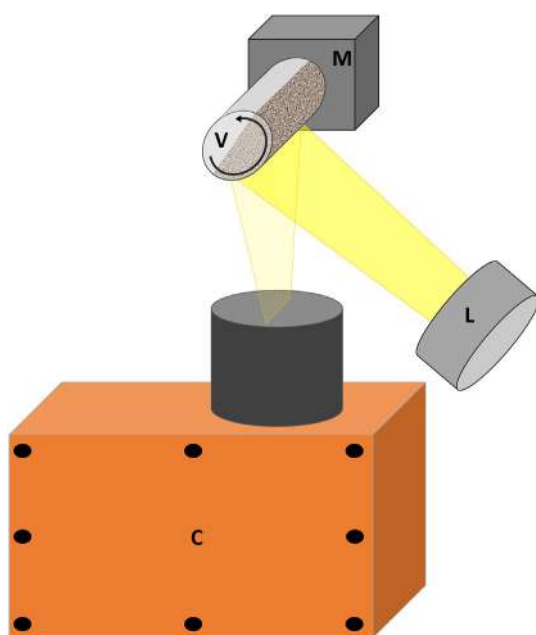
<sup>b</sup> Approximate average particle size as determined by image based particle size analysis.

<sup>c</sup> Particle size was too small to be determined by the image based particle size analysis.

<sup>d</sup> Particle size through 140 mesh (min. 99%). Provided by JRS Pharma certificate of analysis.

**Table 2**  
Description and experimental parameters (blending time and NIR image acquisition) of the blending runs analyzed.

Category	Batch ID	Relative composition	Total blending time (min)	Frame rate (Hz)	Integration time (ms)
Food	BF1	Coffee:Rice grits (1:1)	5	10	3
	TF1R1	Poppy:Coffee:Quinoa (1:1:1)	5	303	2
	TF1R2	Poppy:Coffee:Quinoa (1:1:1)	5	303	2
	TF2R1	Poppy:Coffee:Rice grits (1:1:1)	5	303	2
	TF2R2	Poppy:Coffee:Rice grits (1:1:1)	5	303	2
Pharma	BP1	ASA:Caffeine (1:1)	5	10	3
	BP2	ASA:CitricAcid (1:1)	15	10	2.9
	TP1R1	ASA:Citric acid:Starch (1:2:1)	5	303	1.2
	TP1R2	ASA:Citric acid:Starch (1:2:1)	5	303	1.2
	TP2	ASA:Citric acid:Starch (1:10:1)	15	10	2.9



**Fig. 1.** Experimental setup. Rotary blender device formed by the vial (V) with material mixture and stepper motor (M). The NIR hyperspectral image acquisition system was formed by the NIR camera (C) and the light source (L).

rotations takes 2.4 s.

### 2.3. In-line NIR hyperspectral image acquisition system

An NIR hyperspectral image acquisition system was used to continuously collect hyperspectral data during the blending experiment. The system consisted of a pushbroom NIR camera (Specim FX17 by Spectral Imaging Ltd., Oulu, Finland) that collects hyperspectral images in the NIR spectral region. Three halogen light bulbs were used as NIR irradiation source. The system setup allowed the collection of hyperspectral data in the diffuse reflectance mode, as shown in Fig. 1. In every frame, the pushbroom camera acquired a line of 640 NIR pixel spectra, formed by 224 spectral channels covering the 935–1720 nm spectral range.

Before each blending experiment, “dark” and “white” reference spectra were recorded in order to convert raw sensor data into reflectance and consecutively to absorbance spectra. “Dark” reference refers to the signal provided by the sensor background noise and was obtained by averaging the signal of 100 frames with the camera shutter closed. The white reference was obtained by scanning a vial filled with barium sulfate powder, a highly reflective material used to set the 100% reflectance reference. Vials filled with

the pure materials of the blending mixture were also scanned to obtain pure average reference spectra of the blending components in the same conditions as spectra from the blending run.

The integration time of the NIR measurements was set in the range of 1–3 ms to avoid saturated sensor signals. The camera frame rate or the number of line scans per second was set to 10 Hz or 303 Hz depending on the blending run, see Table 2. The images acquired from blending runs collected with a frame rate of 10 Hz had a pixel size of 2.5 mm and for runs collected with 303 Hz, 0.1 mm. The complete data acquisition settings for each blending run are shown in Table 2.

### 3. Data treatment

The data treatment section is focused on the description of the steps to obtain the real-time evolution of heterogeneity indices during blending batches. The full flowchart associated with the data treatment is displayed in Fig. 2. As can be seen in Fig. 2, the first step describes the preprocessing of raw image data and the construction of distribution maps for each component of the mixture during the blending process. After that, the continuous extraction of variogram-derived heterogeneity indices from the distribution maps is explained.

#### 3.1. Hyperspectral data preprocessing and generation of distribution maps

**Data description.** During the real-time monitoring of the blending process with the NIR-HSI system, every HSI data frame provided a line of 640-pixel spectra, where each pixel is formed by absorbance values of the 224 spectral channels in the NIR range. The final image size depended on the total number of frames collected, associated with the frame rate and total blending time.

**Data preprocessing.** The raw HSI data should be preprocessed for further analysis. Typical push broom NIR cameras present few *dead* sensor signals, also called *bad* sensor pixels [31]. The positions of the camera bad pixels were found using a “white” reference image scanned previously to run the blending experiments. In the system used in this work, the scanned line of pixels is sized (640 × 224), i.e., 640 spectra with 224 spectral channels = 143360 sensor elements in the FPA detector. From those, only 35 dead sensor elements (“bad sensor pixels”) distributed in five clumps in the detector array were found. The signal at each bad pixel detector position was discarded and replaced by an interpolated value from neighboring sensor pixels in the spectral direction in all subsequent scanned hyperspectral data. A shape-preserving piecewise cubic interpolation was the method used for this purpose. Subsequently, pixels from the edges of the scanned line were discarded because of some image artifacts, such as out of focus pixels and presence of the 3D printed plastic support used to attach the vial to the stepper

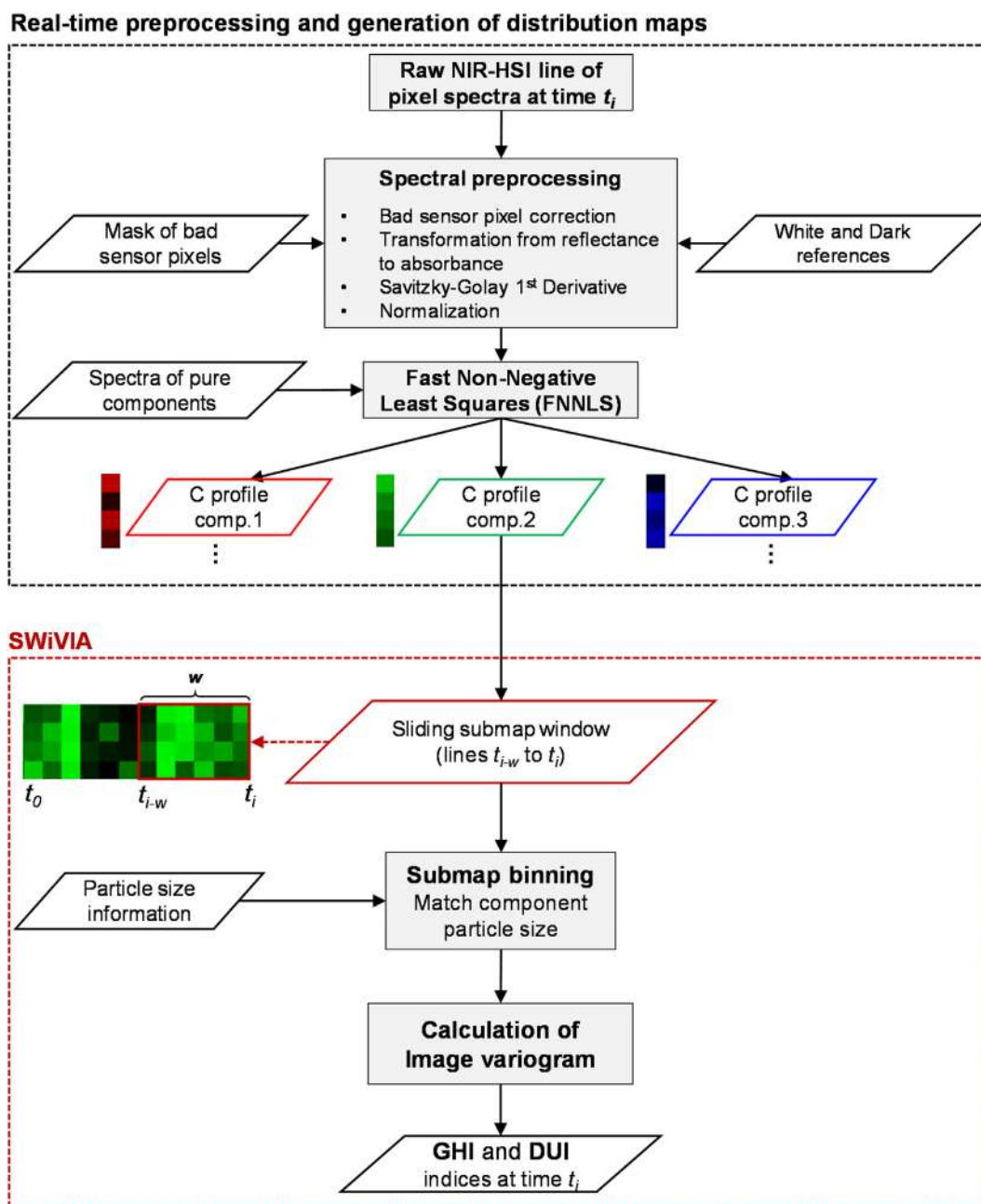


Fig. 2. Flowchart associated with the steps described in the data treatment section.

motor. From the 640 pixels/line provided by the camera sensor, at least 500 pixels were kept for further analysis. Typical raw NIR spectra baseline variation, as shown in Fig. S1A of the Supporting Information (SI), was corrected using Savitzky-Golay first derivative (second-order polynomial and window size of nine points) [32] followed by spectral normalization using the Euclidean norm. Fig. S1B of the SI shows the spectra after preprocessing.

**Generation of distribution maps with FNNLS.** The fast non-negativity least squares (FNNLS) algorithm was used to generate distribution maps for each component of the mixture using the preprocessed hyperspectral data collected during the blending experiments and the pure spectral signatures of the compounds involved in the blending. These signatures were obtained as the average of spectra of pure compound NIR images. FNNLS is a fast algorithm implementation of the non-negativity-constrained

linear least squares regression [28]. This algorithm is applied on every line of preprocessed pixel spectra taking as a basis the bilinear model for spectroscopic data presented in eq. (1),

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \quad (1)$$

where  $\mathbf{D}(x \times \lambda)$  contains the  $x$  spectra in a pixel line,  $\mathbf{C}(x \times k)$  is the matrix of concentration profiles for the  $k$  components of the blending run and  $\mathbf{S}^T(k \times \lambda)$  is the matrix of the known pure spectra signatures of the blending components.

Fig. 3A illustrates the FNNLS generation of concentration profiles,  $\mathbf{C}_{t1}$ , using the first scanned line of preprocessed pixel spectra,  $\mathbf{D}_{t1}$ , and the matrix of pure spectral signatures,  $\mathbf{S}^T$ , for a three-component system. Each column in matrix  $\mathbf{C}_{t1}$ , related to one of the species of the blending will be part of the related distribution

map. This step is repeated for every new frame of hyperspectral data obtained during the blending monitoring and the concentration profiles in the **C** matrix are arranged to form pure distribution maps for each component of the mixture. Fig. 3B shows the distribution maps obtained with FNNLS for each species after scanning  $f$  frames from  $t_1$  to  $t_f$ .

The distribution maps generated by FNNLS can only provide real-time qualitative visual information about the heterogeneity of the mixture during the blending process. In order to find quantitative heterogeneity information during the process, a continuous extraction of heterogeneity indices from the distribution maps using variographic analysis is proposed in this work. In the calculation of the heterogeneity indices, the distribution maps are used as such when the pixel size is bigger than the particle size of all compounds in the mixture (images collected at frame rate 10 Hz). Instead, when the pixel size is smaller than the particle size of a particular compound (images collected at frame rate 303 Hz), a binned version of the distribution map, where the pixel size is approximately equal to the particle size, is used. The binning factor can be easily deduced looking at the particle size of materials in Table 1.

### 3.2. Continuous extraction of heterogeneity indices using a sliding window variographic image analysis (SWiVIA)

Variographic analysis has been recently proposed to extract heterogeneity indices from distribution maps coming from NIR hyperspectral images [25]. A variogram shows the evolution of the variance as a function of a lag distance. Variograms can be easily adapted to analyze 2D images or distribution maps such as those generated with the FNNLS algorithm (see Fig. 3B). The variance

values from image variograms are estimated by comparing properties of pixel pairs separated by a certain lag, in both horizontal and vertical directions, using the following equation:

$$V(h) = \frac{1}{2} \cdot \frac{1}{N(h)} \sum_{i=1}^{N(h)} [c(x_i + h) - c(x_i)]^2 \quad (2)$$

where  $V(h)$  is the variance found as half of the average of the squared differences of all  $N(h)$  pairs of measured pixel values  $c(x_i + h)$  and  $c(x_i)$ , separated by a lag distance ( $h$ ). In this work, the variance values  $V(h)$  were calculated using the concentration values ( $c$ ) extracted from the distribution maps obtained by FNNLS. Note that the image generated during the blending experiment has one spatial dimension linked to the pixel line and another one related to the blending time dimension. This relationship between the spatial and blending time variables will be used to understand the evolution of the blending process.

The suitable combined use of the variogram that comes from the real distribution map (showing the real mixing situation) together with the flat variogram that is obtained when the concentration values of the real map are shuffled to provide a randomized map (displaying the map for the randomized mixture) is the basis to derive two heterogeneity indices: the distributional uniformity index (DUI), related specifically to the distributional heterogeneity of the material, i.e., the evenness of the spatial distribution of the material, and a global heterogeneity index, designed GHI, linked to constant variance value (sill) of the flat variogram of the randomized map (more detail on how these indices are calculated is given later on in this section).

In this work, these variogram-derived heterogeneity indices are used to continuously follow the blending process. For this purpose, a method based on sliding window variographic image analysis (SWiVIA) is proposed. The SWiVIA method extracts the variogram-derived heterogeneity indices from a submap defined by a sliding window that moves every time one pixel line ahead until the full map collected during the blending process is covered, as shown in Fig. 4.

Fig. 4A shows the distribution map in grayscale where dark and white pixels represent low and high concentration of the material, respectively. The blending process evolution of the material is shown from left to right in the map, starting at time  $t_1$  and finishing after collecting  $f$  lines of pixels at time  $t_f$ . Bands of segregated material are observed at the beginning of the process fading away as new lines of pixels are obtained mimicking the kind of distribution map obtained in the blending processes performed with a rotary blender in this work.

The sliding window sized ( $x \times w$ ) is delimited by the number of pixels ( $x$ ) of the scanned line and the width ( $w$ ) of the sliding window. Therefore, the first full sliding window is obtained once  $w$  lines of  $x$  pixels are collected as delimited by the red rectangle in Fig. 4A, covering from  $t_1$  until  $t_w$ . Then, the related heterogeneity indices can be calculated from the image inside this first window. Fig. 4B shows the two indices, GHI and DUI, from the image inside the window at  $t_w$ . After the next line of pixels is recorded, at  $t_{w+1}$ , the window slides to the right including this new line and discarding the oldest pixel line inside the window at  $t_1$ , as delimited by the green rectangle in Fig. 4A. Hence, new variogram-derived indices are calculated for this new window and plotted next to the previous point in Fig. 4B. This procedure is repeated for every new line of pixel recorded during the blending process allowing the real-time representation of the heterogeneity evolution until the last line is scanned at  $t_f$ . The last window and its related heterogeneity indices are shown in magenta in Fig. 4A and B, respectively.

At this point, all necessary steps displayed in the workflow

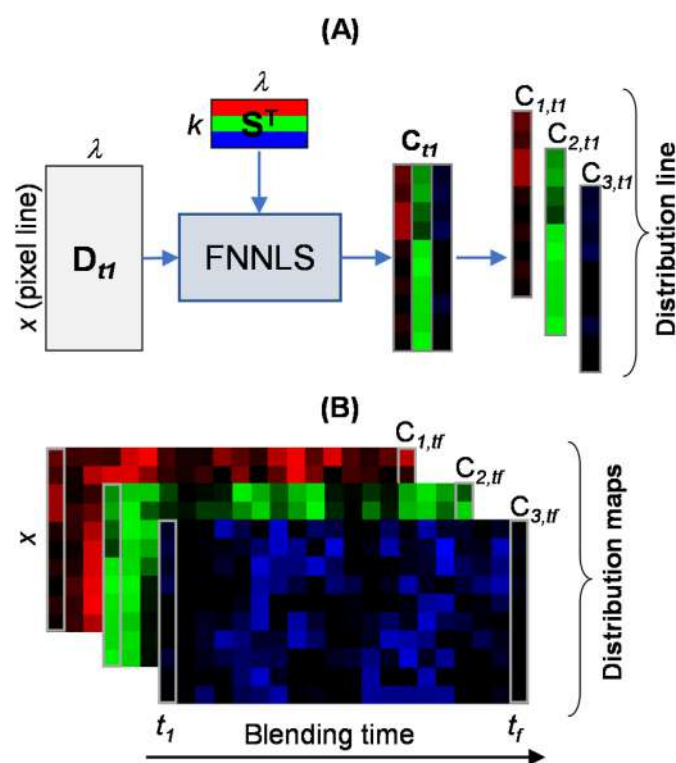
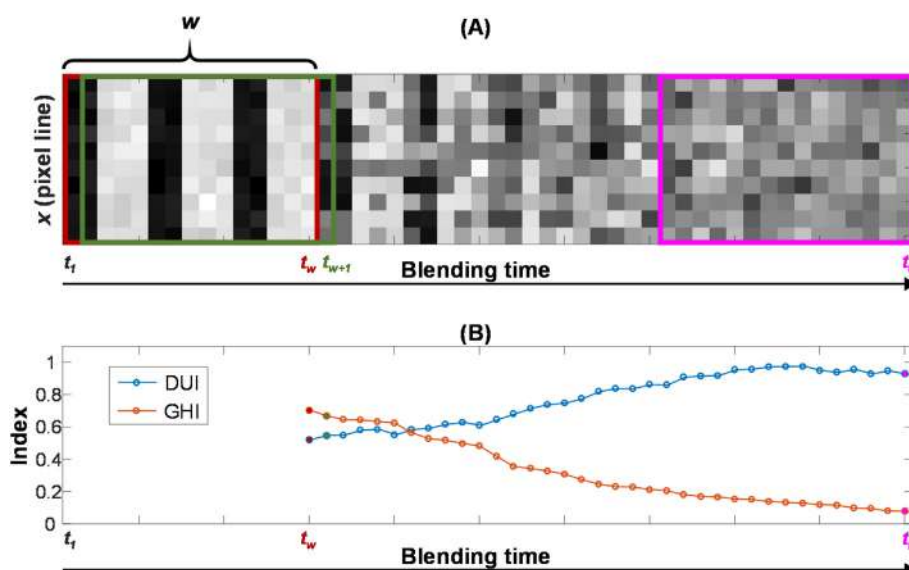


Fig. 3. (A) FNNLS generation of concentration profiles from the first scanned line of hyperspectral data after preprocessing. (B) Distribution maps generated by FNNLS for each component of the mixture during the blending process after scanning  $f$  frames (pixel lines) from  $t_1$  to  $t_f$ .



**Fig. 4.** Representation of the continuous extraction of variogram-derived heterogeneity indices from a blending distribution map using the sliding window variographic image analysis (SWiVIA) method. (A) distribution map and the sliding window with width,  $w$ , at time  $t_w$  (first window),  $t_{w+1}$  and  $t_f$ . (B) heterogeneity indices (DUI and GHI) calculated for each image inside the sliding window as a function of the blending time.

presented in Fig. 2 have been described. Note that the computation time needed to handle every new line of pixel spectra, from pre-processing until providing a new set of GHI and DUI values may go from 0.06 s (if no binning is required) to 0.01 s if a binning of  $5 \times 5$  pixels is used. This means that the approach is suitable to provide indices in real-time even if a high frame rate is used to monitor the process.

The SWiVIA method described above can be used to follow the evolution of a blending process based on the heterogeneity curves presented in Fig. 4B for each individual component of the mixture, *per component* indices. However, *per total mixture* indices can also be calculated for several selected components or all components of the mixture together as described elsewhere [25]. A weighted *per total mixture* index can also be used taking into account the nominal concentration of each blending component.

For a deeper understanding of the effect of the choice of some parameters on the SWiVIA results, it is important to describe in detail how the two variogram-derived heterogeneity indices are calculated from a sliding window. Fig. 5A displays a real distribution map generated by FNNLS from poppy seeds at the beginning of the TF2R2 blending run. The dashed green rectangle delimits the sliding window from which the variogram showed in Fig. 5B (blue line) was calculated using eq. (2). The GHI value displayed in Fig. 5B by the flat red line is a very good approximation of the flat variogram obtained from the randomized map, as designed in Ref. [25]. To save computational cost for online blending monitoring, the GHI is here estimated from the variance of the pixel concentration values inside the window, conceptually identical to the sill of the flat variogram. The DUI was estimated by calculating the ratio of the area of the variogram obtained from the window map (striped blue area) to the area of the flat variogram (striped red area), see Fig. 5B. The more evenly distributed the material is, the more similar the real variogram to the variogram of the randomized map (or to the line defined by the GHI index) will be. For a perfect mixture, the DUI value will be equal to 1.

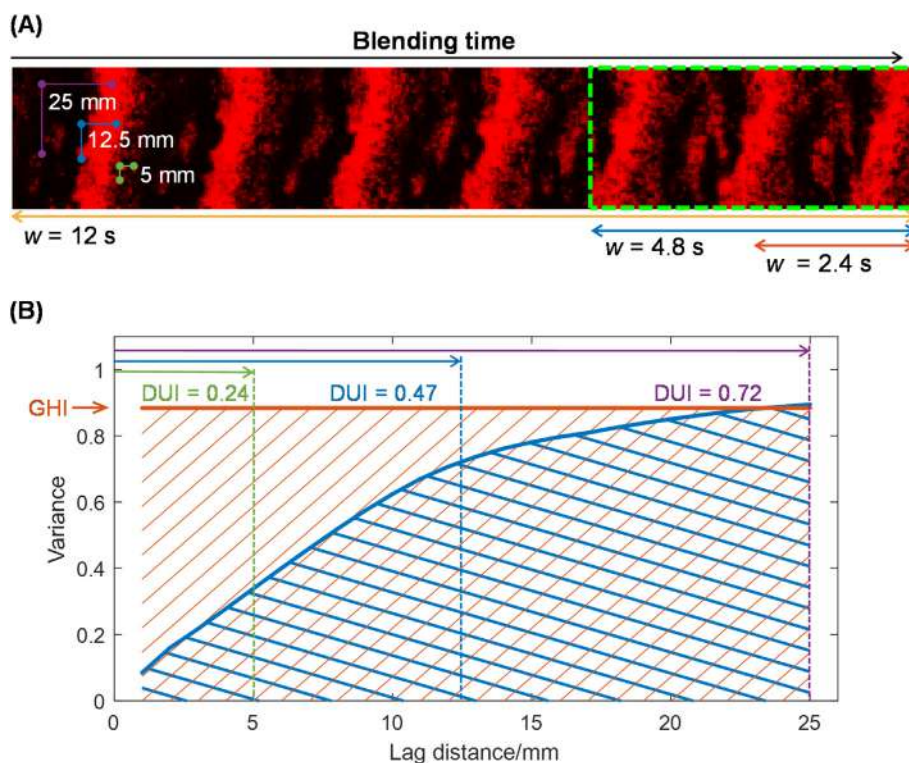
The value of the heterogeneity indices calculated from the submap delimited by the sliding window may vary according to the following SWiVIA parameters:

- Window size:** The size of the sliding window influences the value of both heterogeneity indices, since the related map studied can cover more or less blending time range. The window size needs to be sufficiently long to provide a good appreciation of the blending state at a certain blending time, but not excessively long to avoid that the indices derived are affected by too old observations that may cause some delays on the perception of the real evolution of the blending progress. Variation in window size may affect the evolution of both GHI and DUI indices. Fig. 5A shows the three window sizes (related to 2.4, 4.8 and 12 s of blending time) that will be studied in this work to address the effect of this parameter.
- Maximum variogram lag ( $h_{max}$ ):** The  $h_{max}$  parameter limits the distance in which pairs of pixels are compared and, therefore, defines the scale of spatial scrutiny defined to study the distributional heterogeneity. Reference scale for the  $h_{max}$  values (5, 12.5 and 25 mm) can be seen inside the distribution map in Fig. 5A. The variogram plot in Fig. 5B clearly shows the effect of changing this parameter in the DUI index obtained. Indeed, limiting the lag scale of the variogram at the different  $h_{max}$  values, the related DUI value changes significantly. Looking at Fig. 5B, it is easy to see that for small  $h_{max}$  the DUI value is smaller than for large  $h_{max}$  because of the different ratios between the two areas used to derive the index. This means that the maximum lag scale should be adjusted to the spatial scale of scrutiny sought for the problem of interest to obtain useful results. Changes in maximum variogram lag do not affect the GHI index since it is a constant parameter estimated as the variance of all pixel concentration values inside the full window studied.

The choice of these two parameters is process-dependent and provides a flexible framework to adapt the blending monitoring to different spatial scales of scrutiny and higher or lower blending time resolution.

### 3.3. Software

The NIR-HSI data measurements were recorded using data



**Fig. 5.** Reference scale for the different parameter settings for the SWiVIA method. (A) Poppy seeds distribution map representing 12 s of blending time at the beginning of batch TF2R1 and the reference scale for moving window size (2.4, 4.8 and 12 s) and maximum lag (5, 12.5 and 25 mm). (B) Map variogram in blue and global variance reference line in red from the sliding window with size equal to 4.8 s represented by the dashed green rectangle in (A) and the DUI values calculated for each maximum lag reference. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

acquisition software (Specim Lumo, Spectral Imaging Ltd., Finland). Data analysis was performed basically with in-house scripts for FNNLS and the heterogeneity index derivation and PLS\_Toolbox 8.7 (Eigenvector Research, USA) for data preprocessing running on MATLAB R2020b (Mathworks, USA).

## 4. Results and discussion

### 4.1. Study of SWiVIA parameters effect on heterogeneity curves

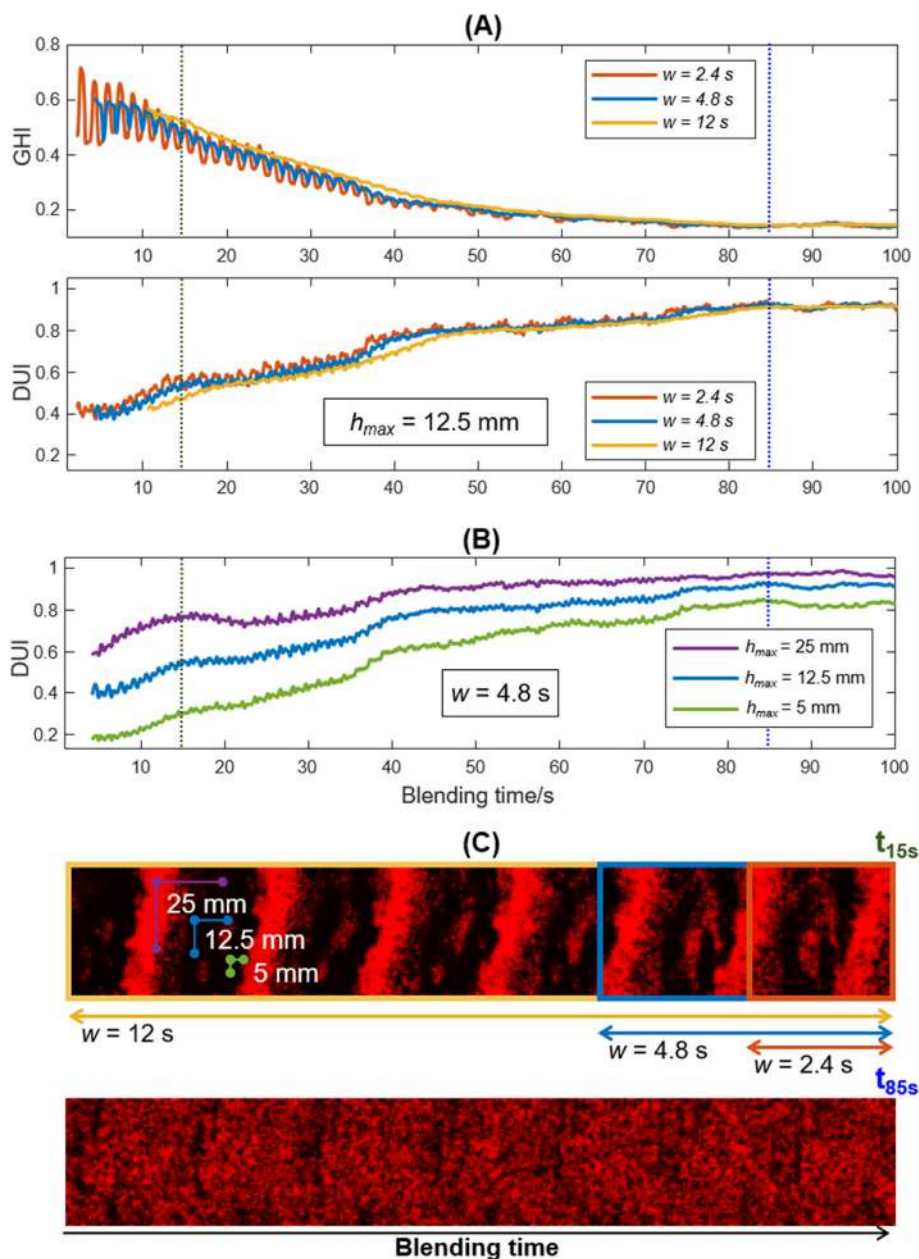
For demonstration of the influence of the two SWiVIA parameters, i.e. window size ( $w$ ) and maximum lag distance ( $h_{max}$ ), on the variogram-derived heterogeneity indices, different combinations of these parameters were tested for the analysis of the same distribution map.

Fig. 6 shows the DUI curves generated with the different combinations of the SWiVIA parameters for the poppy seeds (PS) distribution map from batch TF2. For a better interpretation, the results are shown only for the first 100 s of blending run. Fig. 6A shows the heterogeneity curves obtained when varying the moving window size ( $w = 2.4, 4.8$  and  $12$ ) with a fixed  $h_{max} = 12.5$  mm. Fig. 6B shows the DUI curves obtained varying the maximum lag distance ( $h_{max} = 5, 12.5$  and  $25$  mm) with fixed  $w = 4.8$  s. Additionally, Fig. 6C shows two PS submaps covering blending times from 13 s to 15 s and from 73 s to 85 s of batch TF2, which are related to the dotted vertical lines placed in Fig. 6A and B. Also, note that the reference scales for both  $w$  and  $h_{max}$  parameters are shown in Fig. 6C.

In general, the heterogeneity curves generated with the different SWiVIA parameters show a good blending evolution of PS for the first 100 s of this batch run (see Figs. 6A and 6B). The low DUI

values at the beginning of the process reflect the initial segregated material as shown in Fig. 6C, top distribution map. As blending progresses, a steady increase of the DUI value is observed until its stabilization and maximum distributional uniformity is reached after approximately 85 s of the blending run. The bottom map of Fig. 6C shows how PS is more uniformly distributed when the DUI curve stabilized at approximately 85 s. The effect of each individual parameter on the DUI curves is described below.

Fig. 6A shows that the three DUI curves obtained using different window sizes start and end at similar DUI values, approximately at 0.4 and 0.9, respectively. The DUI values are associated with the last blending time in the map window studied. Therefore, the longer the window size, the higher the number of past scanned lines included in the map window used to calculate the related heterogeneity indices. Since the indices shown in Fig. 6A all come from full map windows, different starting times are observed in the DUI curves of Fig. 6A. When using too long window sizes, the global evolution of the DUI curve suffers from a time delay, as shown in Fig. 6A for  $w = 12$  s. Such a delay comes from the higher weight of past observations (less mixed) in the index derivation. If the window size is too long, the situation of good mixing or the detection of blending faults may be slower than required. However, although a short window may detect faster variations on heterogeneity, if too short, it may be very sensitive to small variations resulting in a high amplitude oscillation of the DUI curve. As a conclusion, the window selected should have an adequate size to avoid the time delays in the description of the blending evolution associated with too long windows, but needs to include sufficient information to avoid the presence of too local blending phenomena that can hinder the visualization of the global blending trend when too short windows are selected. The same effect of the window size in the GHI curves



**Fig. 6.** GHI and DUI curves after the application of SWiVIA in the poppy seeds distribution map from batch TF2R2. (A) GHI (top) and DUI (bottom) curves using different sliding window sizes ( $w = 2.4, 4.8$  and  $12$  s) with fixed maximum lag ( $h_{max} = 12.5$  mm) for DUI calculation. (B) DUI curves using different maximum lag ( $h_{max} = 5, 12.5$  and  $25$  mm) with fixed  $w = 4.8$  s (C) Distribution maps from time 3 s to time 15 s (top) and from time 73 s–85 s (bottom), reference scale for  $h_{max}$  and  $w$  SWiVIA parameters are shown.

was observed as seen in Fig. 6A top panel. In this work, SWiVIA-derived heterogeneity indices built with  $w = 4.8$  s were found to be suitable for the application studied, since smooth DUI curves with no time delays were obtained. Thus, this value was fixed to study the effect of the  $h_{max}$  on the generation of DUI curves.

Fig. 6B shows the DUI curves generated with the three different  $h_{max}$  studied. Increasing DUI curves with no time delay among them were obtained when using different  $h_{max}$ . However, as described in the data treatment section, the larger the  $h_{max}$  value, the larger the DUI value obtained. In this case, DUI values for the reference time  $t_{15s}$  are 0.29, 0.54 and 0.75 for  $h_{max}$  of 5, 12.5 and 25 mm, respectively, and at  $t_{85s}$ , DUI values of 0.85, 0.93 and 0.97 relate to  $h_{max}$  of 5, 12.5 and 25 mm, respectively. This confirms that different DUI values are obtained according to the scale of spatial scrutiny selected, consequently changing the interpretation of the results

during the monitoring of a blending process. Indeed, when the scale of scrutiny is set to cover large spatial areas, the blending process would be considered complete earlier. For instance, if the blending endpoint for the DUI curves shown in Fig. 6B is set to DUI limit of 0.9, that is, 90% of the ideal mixture, different endpoints are achieved according to the  $h_{max}$  used. In this case, for the PS blending evolution in batch TF2, the endpoint is reached after approximately 45 s for a  $h_{max} = 25$  mm, after 80 s for  $h_{max} = 12.5$  mm, and did not reach the endpoint for  $h_{max} = 5$  mm. The selection of the  $h_{max}$  parameter will depend on the particle size of the materials mixed (very small  $h_{max}$  parameters are not suitable for materials with big particle size) and on the degree of spatial precision required for heterogeneity estimation, which may be higher in products where an insufficient blending has more critical effects. As mentioned before changes in  $h_{max}$  do not affect the GHI index.



Because of the direct influence of the SWiVIA parameters on the generated heterogeneity indices, the selection of both parameters is of great importance and must be adjusted according to the quality requirements defined by the application of interest. In this work, the SWiVIA parameters were set to  $w = 4.8$  s and  $h_{max} = 12.5$  mm to generate the heterogeneity indices curves used for monitoring all blending runs.

#### 4.2. Real-time monitoring of blending processes with SWiVIA-derived heterogeneity indices

In this section, the general results for the real-time monitoring of the blending runs studied in this work are presented based on the application of the SWiVIA approach and its derived heterogeneity indices. First, the general use of SWiVIA is described to study the blending evolution *per component* and/or *per total mixture* using the heterogeneity indices proposed. An additional specific section is also used to show how the proposed approach detects de-mixing, an undesired blending-related phenomenon that was observed in this study and has been reported in practice.

#### 4.3. General application of SWiVIA for the real-time monitoring of blending processes

The results of the continuous monitoring of two replicate food batches TF1R2 and TF1R1 using the SWiVIA method are shown in Fig. 7A and Fig. 7B, respectively. Mid and bottom plots show the evolution of the DUI and GHI indices, respectively, for every component in the mixture. The top plot shows combined RGB submaps overlaying the pure component distribution maps (PS in red, GC in green, and QS in blue) generated by FNNLS at selected time ranges.

The evolution of blending run TF1R2 and submaps used to calculate the indices at 5 s, 20 s and 290 s are shown in Fig. 7A. The top left plot shows the mixture distribution submap window (4.8 s) immediately before reaching 5 s of the blending time. This submap clearly shows the initial high segregation level as seen by the different layers of segregated materials at the beginning of the process. The next submap in Fig. 7A shows that the segregation level decreased after 20 s of blending time, but still some diffuse layers of clumped material were visible, mainly for GC and QS. Finally, all three components were visually more evenly distributed at the end of the blending run, as shown in the last submap, after 290 s. The quantitative heterogeneity indices generated by the SWiVIA method reflected the visual qualitative observation described above. The *component* DUI curves for batch TF1R2 show that the PS reached high and stable DUI values faster than the other two compounds. This can be visualized by the high slope of PS DUI curve (red curve in Fig. 7A mid panel) at the beginning of the blending run and the more even spatial distribution of the red component as seen in the submap at 20 s (top panel in Fig. 7A). Even though unique *component* DUI curves evolution were observed at the beginning of TF1R2 blending run, all three components reached the same level of spatial distribution after approximately 75 s with an average DUI value of 0.9. The *component* GHI curves for batch TF1R2, bottom panel of Fig. 7A, shows that all three curves started with high GHI values. This high variance at the beginning of the blending run is related to the large number of pixels with very high and very low concentration values in the areas with the presence and absence of the segregated material, respectively. Like the DUI curves, the GHI values changed significantly during the first minute of the TF1R2 blending run. However, because this index represents only the relative variance of the concentration values within the submap, independently to the spatial distribution, the GHI curves reached to a stabilization at

different times than for DUI curves. In this case, the GHI curves for PS and GC reached a stable and similar GHI value after approximately 40 s of blending time with a slightly variation around  $GHI = 0.2$  during the rest of the blending run. On the other hand, the GHI of the QS reached a steady  $GHI = 0.4$  after the first minute of the process.

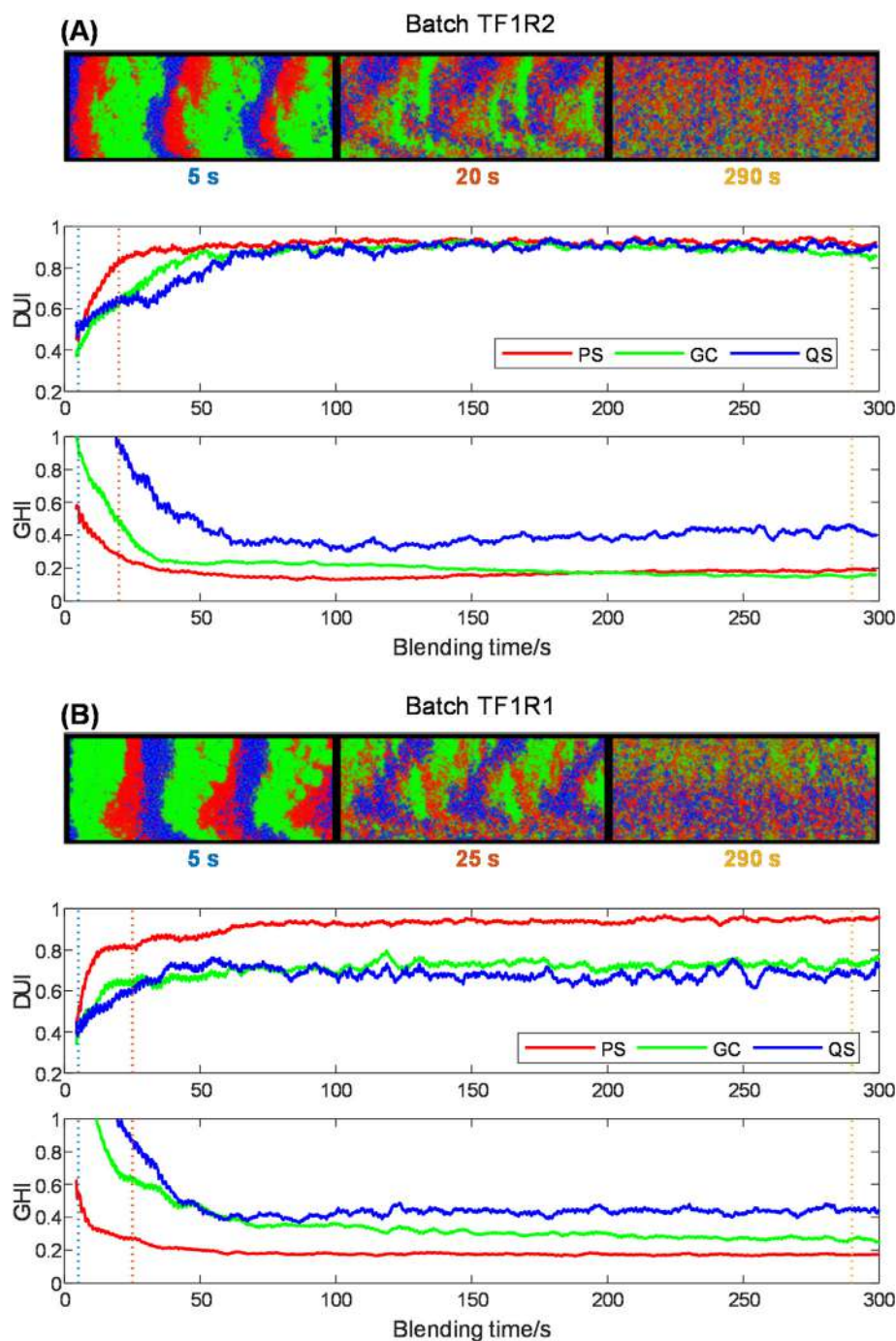
The *per component* heterogeneity curves for blending run TF1R1 are shown in Fig. 7B together with the combined RGB submaps at 5 s, 25 s and 290 s. Similar behavior at the beginning of this batch was observed when compared to batch TF1R2; however, after the first minute of the blending run, only PS reached the DUI level of 0.9 as shown in Fig. 7B (red curve in the mid panel). The other two components, GC and QS, reached a stabilization of the DUI curve, but with a lower DUI level around 0.7. This can also be visualized in the submap at 290 s, while PS (in red) was evenly distributed, the green (GC) and blue (QS) horizontal bands indicates their poor blend because of the accumulation to each side of the rotating vial. The *per component* GHI curves for batch TF1R1, Fig. 7B bottom panel, show the low relative variance for PS with GHI values below 0.2 after 60 s of blending time, with higher GHI values for GC and QS.

The results of the continuous monitoring with the heterogeneity curves for both replicate batches show the expected natural evolution of the blending process, that is, increasing DUI curves and decreasing GHI curve with blending evolution. Although blending runs TF1R1 and R2 had the same mixture composition, clear differences in the evolution of the heterogeneity curves were observed. This same behavior was observed for the other replicated batches studied in this work, as shown in Figs. S2 and S3 for food batches TF2R1 and TF2R2, respectively, and in Figs. S4 and S5 for pharmaceutical batches TP1R1 and TP1R2, respectively (see SI). This unique heterogeneity evolution for each batch indicates that specific continuous monitoring of each blending run is required to ensure blending quality.

The *per component* GHI and DUI heterogeneity indices obtained for all the compounds in a blending process can also be combined to obtain single *per total mixture* GHI and DUI indices, respectively. Two batches studied in this work were chosen to show how *per total mixture* indices can be used to see in a global manner the continuous monitoring of a blending run.

The first situation corresponds to the blending monitoring of a binary mixture. In this case, it is usual that the blending distribution pattern for one component is complementary to the other. Indeed, when looking at the distribution maps of a binary mixture, such as for batch BP1 shown in Fig. 8A, we can see this behavior for the combined submaps of CAF (red) and ASA (green) at the different blending times (5 s, 50 s and 290 s). This results in DUI curves evolving with very similar shape and scale for both compounds, see Fig. 8B (top panel); the same situation happens for GHI curves, but with different variance scale, see Fig. 8B (bottom panel). Thus, the mixing evolution of batch BP1 using the SWiVIA-derived *per total mixture* heterogeneity indices (GHI and DUI) is shown in Fig. 8C. Both curves indicated that the pharmaceutical binary blending run, BP1, mixed properly during the total blending time. The batch followed the regular blending pattern, starting with  $DUI < 0.4$  and  $GHI > 0.8$  for the initial individual layers (submap at 5 s) and ending with stable  $DUI > 0.8$  and  $GHI < 0.1$ , as can be seen in the even distribution of CAF and ASA at 290 s in the top right panel of Fig. 8A.

Often, the mixture to be blended can include major and minor compounds that contribute differently to the bulk heterogeneity. Here, the ternary pharmaceutical batch, TP2, is used to exemplify the use of weighted *total mixture* heterogeneity indices. The batch TP2 is formed by a mixture of the pharmaceutical compounds ASA, CA and SSG, where the content of CA was tenfold of the ASA and SSG, see Table 2. Therefore, this nominal concentration values were

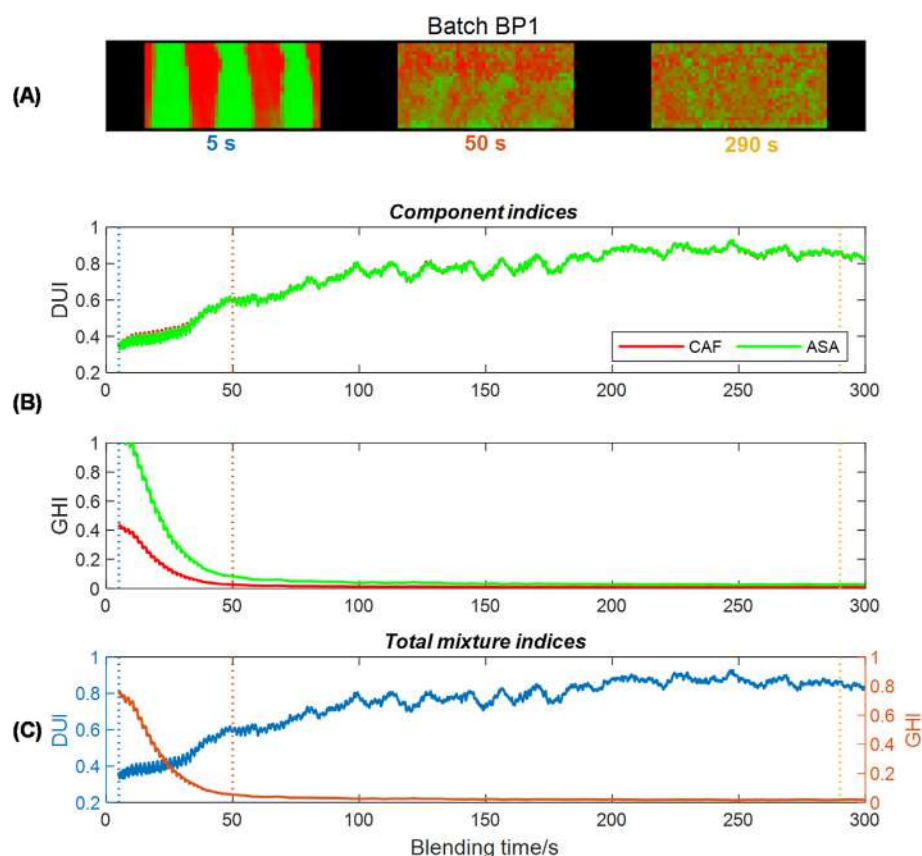


**Fig. 7.** SWiVIA-derived heterogeneity curves for the continuous monitoring of replicate batches TF1R2 (A) and TF1R1 (B). Top panel shows the combined RGB submaps overlaying the pure component distribution submaps (poppy seeds (PS) in red, ground coffee (GC) in green and quinoa seeds (QS) in blue) for selected reference times. Mid and bottom panels show the DUI and GHI curves, respectively. Selected reference times are indicated by the vertical dotted lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

used for the calculation of the concentration-weighted *total mixture* heterogeneity indices. Such a strategy avoids that very minor compounds, which can show an expected high heterogeneity, increase artificially the heterogeneity associated with the global formulation. Fig. 9A shows the combined RGB distribution submaps at 5 s, 100 s and 700 s of the TP2 blending run with ASA in red, CA in green and SSG in blue. Fig. 9B shows the *component* SWiVIA-derived heterogeneity indices and Fig. 9C the concentration-weighted *total mixture* heterogeneity indices for the same

blending run.

At the beginning of batch TP2, the initial layers of segregated ingredients are shown in the submap at 5 s in Fig. 9A, where the relative larger extension of the major compound CA, in green, can be seen. The blending evolution of this batch reached the highest and steady distributional homogeneity after approximately 100 s of the blending time with weighted *per total mixture* DUI values above 0.95 (see the submap at 100 s in Fig. 9A and the DUI curve in Fig. 9C for a better illustration). The weighted *total mixture* GHI, however,



**Fig. 8.** Results for the continuous monitoring of batch BP1. (A) Combined red and green submaps overlaying the pure component distribution submaps acetyl salicylic acid (ASA) in red and caffeine (CAF) in green. (B) *Component* SWiVIA-derived heterogeneity curves, GHI (top) and DUI (bottom) and (C) *Total mixture* heterogeneity curves, DUI (blue) and GHI (orange). Selected reference times are indicated by the vertical dotted lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

stabilized faster reaching to variance values below 0.1 after only 20 s of the blending time, see Fig. 9C. From time 300 s and forward the decreasing *total mixture* DUI curve show that a de-mixing process occurred. This phenomenon is described with more detail in the following section.

The use of the SWiVIA-derived *per total mixture* indices are helpful to see the overall evolution of the blending process based on the average or concentration-weighted average based on the concentration of all components of the mixture. The control of the blending end-point can be established using the combination of *total mixture* GHI and DUI indices, using predefined blending completion threshold values associated with heterogeneity quality specifications or using thresholds derived from the study of historical batches having achieved a satisfactory blending. Such a use of the *total mixture* GHI and DUI indices would save blending time and would avoid problems associated with excessive blending, as described below.

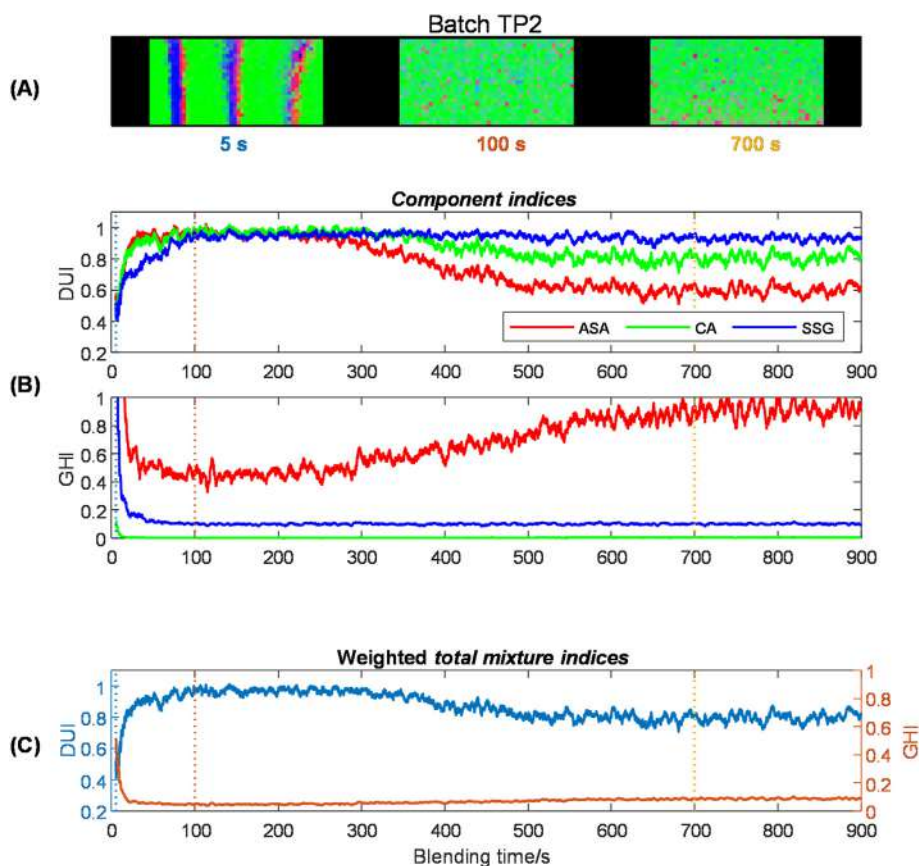
#### 4.4. De-mixing phenomenon

Attempting to ensure a perfect mixed product, industrial blending operations may “overblend” their mixtures. This practice increases energy and labour costs, and may induce a de-mixing or segregation of the blended material [2]. In this work, the SWiVIA approach allowed the visualization of the de-mixing phenomenon based on the GHI, and most clearly, the DUI indices during the continuous monitoring of some blending runs. Because of the small-scale system, the main factors that induced the de-mixing process in this work were the differences in particle size, density

among the materials and the circular blender design used.

Fig. 10A shows the heterogeneity curves and distributions maps at 5 s, 134 s and 290 s of BF1 blending run formed by GC (red) and RG (green). The results indicated that the two ingredients followed a good blending evolution during the first half of the blending run. The initial layers of material shown in the submap at 5 s started with a DUI = 0.3 and gradually were mixed reaching a maximum DUI levels above 0.8 after approximately 2 min after the start of the blending run. This moderately even distribution can be visualized in the submap at 134 s of Fig. 10A (top panel). Despite the continuous decreasing GHI curve, the DUI curves show that the de-mixing process has taken place during the second part of this blending run. After reaching the maximum distribution at 134 s, the two compounds started to agglomerate reaching lower DUI values around 0.6 at the end of BF1 blending run. The result of this de-mixing process can be clearly visualized in the submap at 290 s, Fig. 10A, where large clumps of GC (red) and RG (green) can be seen at the top part and bottom part of the blending map, respectively. Note that top and bottom part of the map correspond to the left and right side of the mixing vial. In this context, this kind of segregation could happen if the axis of the rotary blender is slightly tilted. GC and RG, showing clear differences of density and particle size, could be prone to show this de-mixing pattern.

Another example of de-mixing phenomenon was the pharmaceutical batch BP2 as shown in Fig. 10B, formed by ASA (red) and CA (green). In this case, the DUI curves show that the initial layers of the two ingredients (see submap at 20 s) fade away during the first minute of the blending run and, after that, the blending improves at a slow pace reaching maximum values of DUI = 0.95 after 2 min, as



**Fig. 9.** Results for the continuous monitoring of batch TP2. (A) Combined RGB submaps overlaying the pure component distribution submaps with acetyl salicylic acid (ASA) in red, citric acid (CA) in green and sodium starch glycolate (SSG) in blue. (B) *Component* SWiVIA-derived heterogeneity curves, GHI (top) and DUI (bottom). (C) weighted *total mixture* heterogeneity curves, DUI (blue) and GHI (orange). Selected reference times are indicated by the vertical dotted lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

shown by the even distribution in the submap at 128 s. Afterwards, the excessive blending slowly induces the de-mixing of the two ingredients, the red component accumulating on top of the map and the green component at the bottom. This phenomenon can be visualized through the steady decreasing DUI curves reaching minimum values below 0.6 at the end of the run and the horizontal bands of segregated material present in submap at 887 s in Fig. 10B. The GHI curves also reflected the de-mixing behaviour of this batch, as shown by the gradually increasing relative variance after 2 min of blending run.

Looking back at the *component* heterogeneity curves of the blending run TP2 shown in Fig. 9B (top panel), an interesting de-mixing process can be visualized. During this run, as mentioned before, the blend reached the maximum distributional homogeneity after approximately 100 s of the blending time. After this, despite of the steady DUI curves for about 2 min, the overblending caused the de-mixing of ASA and CA reaching at the end of the run DUI levels of approximately 0.8 and 0.6, respectively. The SSG, however, kept with the same elevated DUI level during the rest of the blending run. In this example, where more than two compounds participate in the blending, it is clearly seen that the blending pattern of every compound is not necessarily the same and that the de-mixing process does not need to involve all blending components in a mixture. A slight de-mixing also happened at the end of batches TF2R2 and TP1R1, as shown in the SWiVIA derived heterogeneity curves in Figs. S4 and S5 of the SI, respectively.

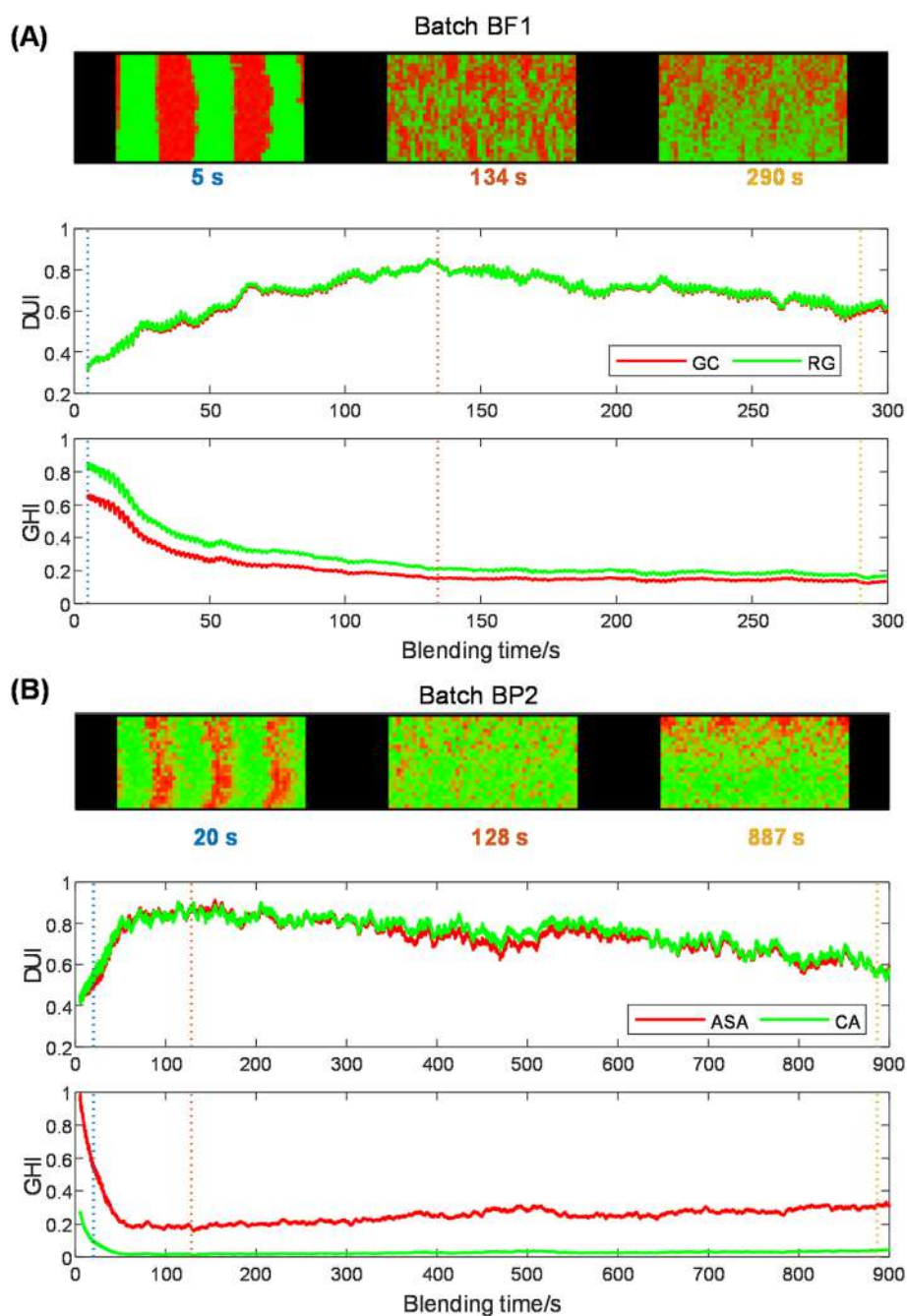
Regarding the blending performance, it seems that differences

among physical properties of the particles of the materials to be blended, including shape, size or density may derive in a blending worsening. In this work, the limited number of batch runs and blending materials was not sufficient to extract solid conclusions regarding the influence of materials geometry in blending. However, it seems that materials with spherical shape, such as PS, tend to provide a better mixing and fine particles with easy tendency to agglomerate, such as CAF and SSG, offer worse blending and are more prone to de-mixing in general.

It is important to note that the de-mixing process is always competing against the blending process. Due to the complex mechanisms involved in the mixing of particulate material and to the unavoidable differences linked to slight variations in the feeding and nature of the initial materials and in the blender operation, the required time to reach the endpoint of the blending process is not reproducible and specific control per each individual blending run should be carried out.

## 5. Conclusions

The combination of in situ acquired chemical images and the SWiVIA method allow a real time continuous heterogeneity assessment in blending processes. In fact, distribution maps generated from the imaging system provide visual qualitative heterogeneity information, while quantitative heterogeneity information is achieved via the proposed SWiVIA-derived global heterogeneity index (GHI) and distributional uniformity index (DUI) curves.



**Fig. 10.** SWiVIA-derived heterogeneity curves for the continuous monitoring of batches and combined red and green submaps overlaying the pure component distribution submaps for (A) batch BF1 with ground coffee (GC) in red and rice grits (RG) in green and (B) batch BP2 with acetyl salicylic acid (ASA) in red and citric acid (CA) in green. Vertical dotted lines show time reference of the submaps. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The methodology proposed can be tailored to process-specific conditions related to the degree of spatial scrutiny sought to study distributional heterogeneity and to the resolution time required to describe the blending evolution. The SWiVIA-derived GHI and DUI curves allow the description of heterogeneity at *component level* or *total mixture level*. The *total mixture* indices are suitable to see the overall evolution of the blending process and, if an abnormal behavior is detected, *component* indices can be used to understand the component or components responsible for the detected problem. The approach proposed provides a good qualitative and quantitative description of heterogeneity for any kind of blending run and allows detection of abnormal blending behavior,

such as the de-mixing process, usually related to overblending.

There are many potential applications of the developed methodology, such as the characterization of the mixing behavior of new blending devices or mixture formulations, the use of GHI and DUI indices to define quality control limits for end-point blending detection (in batch blending processes) or for continuous quality control (in continuous operations that involve heterogeneity variations, such as blending, granulation, tableting, etc.). Finally, it should be added that the input information required by this methodology is not limited to that provided by hyperspectral image platforms, but can be extended to other kinds of machine vision systems and, in general, to any kind of instrumental technique that

can provide spatially resolved responses.

### CRediT authorship contribution statement

**Rodrigo Rocha de Oliveira:** Conceptualization, Methodology, Investigation, Data curation, Visualization, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Anna de Juan:** Conceptualization, Methodology, Supervision, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition.

### 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.

### Acknowledgements

Funding from Spanish government under the project PID 2019-1071586B-I00 is acknowledged. The authors belong to the Catalan excellence research group (2017 SGR 753).

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.aca.2021.338852>.

### References

- [1] F.J. Muzzio, A. Alexander, C. Goodridge, E. Shen, T. Shinbrot, Chapter 15 Part A: fundamentals of solids mixing, in: E.L. Paul, V.A. Atiemo-Obeng, S.M. Kresta (Eds.), *Handb. Ind. Mix.*, John Wiley & Sons, Hoboken, New Jersey, 2004, p. 887.
- [2] C.D. Cullen, P.J. Romañach, Rodolfo J. Abatzoglou, Nicolas Rielly, *Pharmaceutical Blending and Mixing*, John Wiley & Sons, Ltd, Chichester, UK, 2015, <https://doi.org/10.1002/9781118682692>.
- [3] B. Cuq, H. Berthiaux, C. Gatamel, Powder mixing in the production of food powders, in: *Handb. Food Powders*, Elsevier, 2013, pp. 200–229, <https://doi.org/10.1533/9780857098672.1.200>.
- [4] A. Hassanpour, M. Ghadiri, Discrete element method (DEM) simulation of powder mixing process, in: *Pharm. Blending Mix*, John Wiley & Sons, Ltd, Chichester, UK, 2015, pp. 459–477, <https://doi.org/10.1002/9781118682692.ch17>.
- [5] N. Govender, D.N. Wilke, C.-Y. Wu, R. Rajamani, J. Khinast, B.J. Glasser, Large-scale GPU based DEM modeling of mixing using irregularly shaped particles, *Adv. Powder Technol.* 29 (2018) 2476–2490, <https://doi.org/10.1016/j.apt.2018.06.028>.
- [6] F. Yu, Z. Yao, G. Chen, Y. Zhang, Y. Zheng, DEM simulations of tote blenders for enhanced axial mixing efficiency, *Particology* 55 (2021) 199–208, <https://doi.org/10.1016/j.partic.2020.08.006>.
- [7] A.L. Bowler, S. Bakalis, N.J. Watson, A review of in-line and on-line measurement techniques to monitor industrial mixing processes, *Chem. Eng. Res. Des.* 153 (2020) 463–495, <https://doi.org/10.1016/j.cherd.2019.10.045>.
- [8] G. Shi, L. Lin, Y. Liu, G. Chen, Y. Luo, Y. Wu, H. Li, Pharmaceutical application of multivariate modelling techniques: a review on the manufacturing of tablets, *RSC Adv.* 11 (2021) 8323–8345, <https://doi.org/10.1039/D0RA08030F>.
- [9] A. Crouter, L. Briens, Methods to assess mixing of pharmaceutical powders, *AAPS PharmSciTech* 20 (2019) 84, <https://doi.org/10.1208/s12249-018-1286-7>.
- [10] B. Igne, A. De Juan, J. Jaumot, J. Lallemand, S. Preys, J.K. Drennen, C.A. Anderson, Modeling strategies for pharmaceutical blend monitoring and end-point determination by near-infrared spectroscopy, *Int. J. Pharm.* 473 (2014) 219–231, <https://doi.org/10.1016/j.ijpharm.2014.06.061>.
- [11] Y. Li, C.A. Anderson, J.K. Drennen, C. Airiau, B. Igne, Method development and validation of an inline process analytical technology method for blend monitoring in the tablet feed frame using Raman spectroscopy, *Anal. Chem.* 90 (2018) 8436–8444, <https://doi.org/10.1021/acs.analchem.8b01009>.
- [12] J.M. Guay, P.P. Lapointe-Garant, R. Gosselin, J.S. Simard, N. Abatzoglou, Development of a multivariate light-induced fluorescence (LIF) PAT tool for in-line quantitative analysis of pharmaceutical granules in a V-blender, *Eur. J. Pharm. Biopharm.* 86 (2014) 524–531, <https://doi.org/10.1016/j.ejpb.2013.12.013>.
- [13] T. Casian, A. Gavan, S. Iurian, A. Porfire, V. Toma, R. Stiufluic, I. Tomuta, Testing the limits of a portable NIR spectrometer: content uniformity of complex powder mixtures followed by calibration transfer for in-line blend monitoring, *Molecules* 26 (2021) 1129, <https://doi.org/10.3390/molecules26041129>.
- [14] O. Scheibelhofer, N. Balak, D.M. Koller, J.G. Khinast, Spatially resolved monitoring of powder mixing processes via multiple NIR-probes, *Powder Technol.* 243 (2013) 161–170, <https://doi.org/10.1016/j.powtec.2013.03.035>.
- [15] A.S. El-Hagrasy, H.R. Morris, F. D'Amico, R.A. Lodder, J.K. Drennen, Near-infrared spectroscopy and imaging for the monitoring of powder blend homogeneity, *J. Pharmaceut. Sci.* 90 (2001) 1298–1307, <https://doi.org/10.1002/jps.1082>.
- [16] J.G. Osorio, G. Stuessy, G.J. Kemeny, F.J. Muzzio, Characterization of pharmaceutical powder blends using in situ near-infrared chemical imaging, *Chem. Eng. Sci.* 108 (2014) 244–257, <https://doi.org/10.1016/j.ces.2013.12.027>.
- [17] D.L. Galata, L.A. Mészáros, M. Ficzer, P. Vass, B. Nagy, E. Szabó, A. Domokos, A. Farkas, I. Csontos, G. Marosi, Z.K. Nagy, Continuous blending monitored and feedback controlled by machine vision-based PAT tool, *J. Pharmaceut. Biomed. Anal.* 196 (2021) 113902, <https://doi.org/10.1016/j.jpba.2021.113902>.
- [18] Hyperspectral imaging, in: J.M. Amigo (Ed.), *Data Handl. Sci. Technol.*, Elsevier, 2020. <https://www.sciencedirect.com/bookseries/data-handling-in-science-and-technology/vol/32/suppl/C>.
- [19] S. Piqueras, J. Burger, R. Tauler, A. de Juan, Relevant aspects of quantification and sample heterogeneity in hyperspectral image resolution, *Chemometr. Intell. Lab. Syst.* 117 (2012) 169–182, <https://doi.org/10.1016/j.chemolab.2011.12.004>.
- [20] J. Burger, P. Geladi, Hyperspectral NIR imaging for calibration and prediction: a comparison between image and spectrometer data for studying organic and biological samples, *Analyst* 131 (2006) 1152–1160, <https://doi.org/10.1039/b605386f>.
- [21] H. Ma, C.A. Anderson, Characterization of pharmaceutical powder blends by NIR chemical imaging, *J. Pharmaceut. Sci.* 97 (2008) 3305–3320, <https://doi.org/10.1002/jps.21230>.
- [22] W. Li, A. Woldu, R. Kelly, J. McCool, R. Bruce, H. Rasmussen, J. Cunningham, D. Winstead, Measurement of drug agglomerates in powder blending simulation samples by near infrared chemical imaging, *Int. J. Pharm.* 350 (2008) 369–373, <https://doi.org/10.1016/j.ijpharm.2007.08.055>.
- [23] M.H. Bharati, J.J. Liu, J.F. MacGregor, Image texture analysis: methods and comparisons, *Chemometr. Intell. Lab. Syst.* 72 (2004) 57–71, <https://doi.org/10.1016/j.chemolab.2004.02.005>.
- [24] M.L. Hamad, C.D. Ellison, M.A. Khan, R.C. Lyon, Drug product characterization by Macropixel Analysis of chemical images, *J. Pharmacol. Sci.* 96 (2007) 3390–3401, <https://doi.org/10.1002/jps.20971>.
- [25] R. Rocha de Oliveira, A. de Juan, Design of heterogeneity indices for blending quality assessment based on hyperspectral images and variographic analysis, *Anal. Chem.* 92 (2020) 15880–15889, <https://doi.org/10.1021/acs.analchem.0c03241>.
- [26] P.Y. Sacré, P. Lebrun, P.F. Chavez, C. De Bleye, L. Netchacovitch, E. Rozet, R. Klinkenberg, B. Streef, P. Hubert, E. Ziemons, A new criterion to assess distributional homogeneity in hyperspectral images of solid pharmaceutical dosage forms, *Anal. Chim. Acta* 818 (2014) 7–14, <https://doi.org/10.1016/j.aca.2014.02.014>.
- [27] N.C. da Silva, L. de Moura França, J.M. Amigo, M. Bautista, M.F. Pimentel, Evaluation and assessment of homogeneity in images. Part 2: homogeneity assessment on single channel non-binary images. Blending end-point detection as example, *Chemometr. Intell. Lab. Syst.* 180 (2018) 15–25, <https://doi.org/10.1016/j.chemolab.2018.06.011>.
- [28] R. Bro, S. Jong, A fast non-negativity-constrained least squares algorithm, *J. Chemom.* 11 (1997) 393–401.
- [29] A.-Z.M. Abouzeid, D.W. Fuerstenau, Mixing–demixing of particulate solids in rotating drums, *Int. J. Miner. Process.* 95 (2010) 40–46, <https://doi.org/10.1016/j.minpro.2010.03.006>.
- [30] I. Jones, C. Smalley, Equipment qualification, process and cleaning validation, in: *Pharm. Blending Mix*, John Wiley & Sons, Ltd, Chichester, UK, 2015, pp. 369–399, <https://doi.org/10.1002/9781118682692.ch14>.
- [31] R. Dorrepaal, C. Malegori, A. Gowen, Tutorial: time series hyperspectral image analysis, *J. Near Infrared Spectrosc.* 24 (2016) 89–107, <https://doi.org/10.1255/jnirs.1208>.
- [32] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Anal. Chem.* 36 (1964) 1627–1639, <https://doi.org/10.1021/ac60214a047>.



#### **4.4.1 Methodology to assess heterogeneity from HSI**

The characterization of heterogeneity during a blending process is crucial to ensure the quality consistency of end products in diverse industrial sectors. HSI's are unique analytical measurements that provide physicochemical and spatial information on samples and, hence, are ideal to perform heterogeneity studies. Two types of heterogeneity information can be extracted from chemical images, the global heterogeneity (GH), representing the scatter associated with individual pixel properties, and the distributional heterogeneity (DH), linked to the evenness in the spatial distribution of the different materials forming a blend. This thesis proposes a new methodology combining HSI and variographic analysis to obtain a good qualitative and quantitative description of both heterogeneity aspects from samples and blending processes. This methodology consists of two steps, namely:

*Step 1. Extraction of distribution maps from HSI.* Hyperspectral image unmixing provides a set of pure distribution maps of the sample constituents. This allows a qualitative visualization of the heterogeneity variation during a blending process for each component of the mixture. These maps are used as seeding information for the variographic derivation of heterogeneity indices.

*Step 2. Derivation of heterogeneity indices from distribution maps.* Variographic analysis of these distribution maps gives quantitative heterogeneity indices to study the variation of both GH and DH during a blending process.

Below, each of these steps is described in detail. Specificities applied to adapt the methodology to discontinuous atline HSI monitoring and to continuous inline blending monitoring are also described.

##### **Step 1. Extraction of distribution maps from HSI**

The extraction of the pure component distribution maps during blending process monitoring was carried out using two strategies for hyperspectral image unmixing, selected according to the type of HSI process monitoring, i.e. atline (Publication VI) or inline (Publication VII).

For the atline monitored blending process described in section 3.2.1, different images at specific blending times were acquired and the extraction of the distribution maps from the NIR-HSI data was carried out using the multiset extension of the MCR-ALS algorithm. The multiset structure contained the multiple images recorded at different blending times for a blending run plus the images of the pure ingredients of the pharmaceutical formulation, as shown in Figure 34. As a result of the MCR analysis, a single set of pure spectral signatures of the compounds in the blending are obtained and their related set of distribution maps as a function of the blending time. Figure 35 shows the maps obtained by MCR-ALS for two blending batch processes where caffeine, CAF, starch, EXP, and acetylsalicylic acid, ASA, are mixed. The sequence



of maps as a function of blending time is displayed using combined RGB maps overlaying the information of the three MCR-ALS resolved compounds (red for CAF, green for EXP, and blue for ASA).

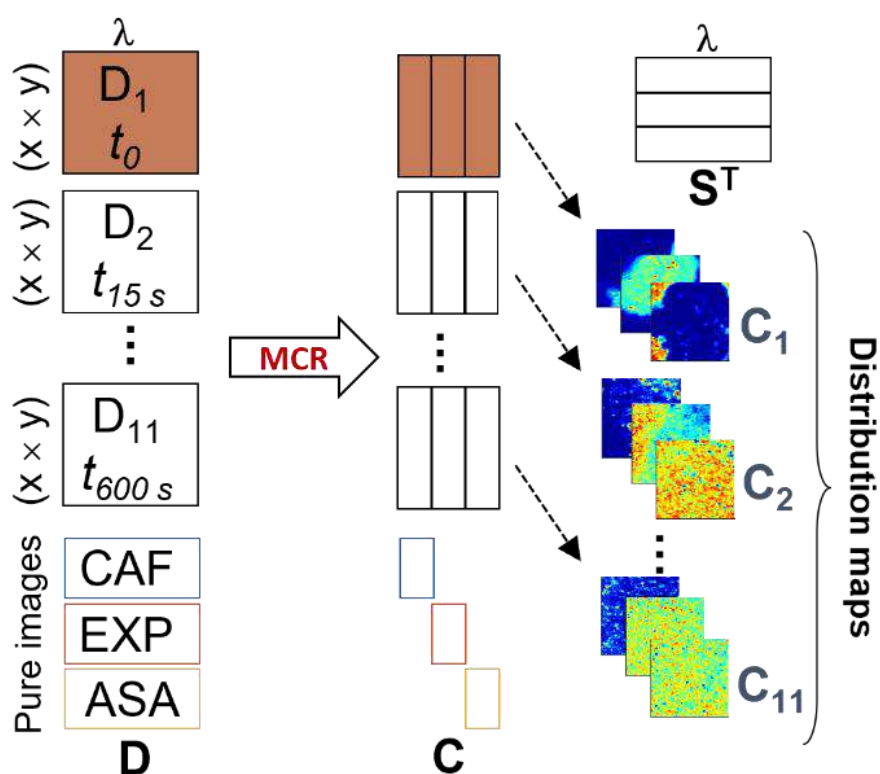


Figure 34 MCR-ALS analysis of an image multiset, where  $x$  and  $y$  are spatial pixels and  $\lambda$  represents the spectra wavelengths, reproduced from (Rocha de Oliveira and de Juan, 2020).

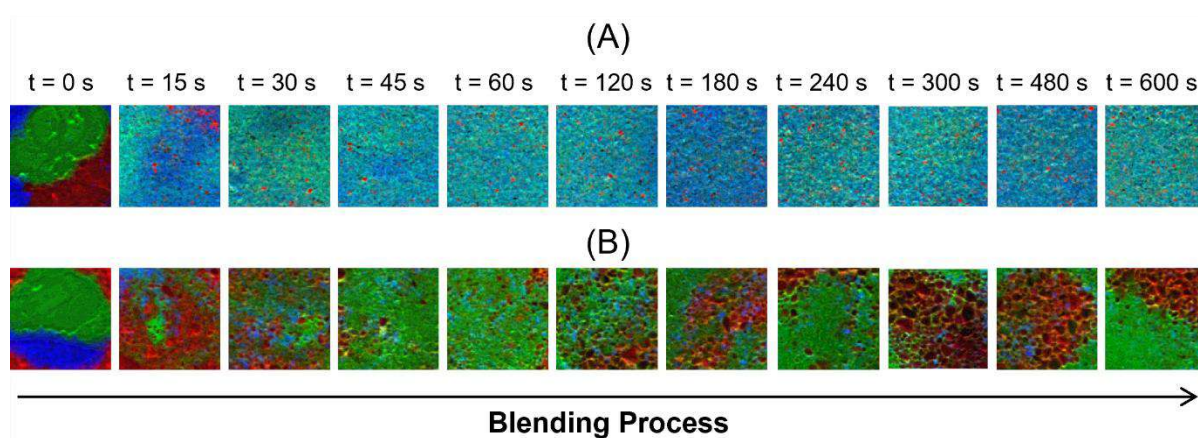


Figure 35 Combined RGB maps with overlaid pure component distribution maps obtained with MCR-ALS for two atline monitored pharmaceutical blending runs with different API mass proportions of 10:1 and 1:10 (ASA/CAF) for the blending runs in (A) and (B) respectively. Red – CAF, green – EXP, and blue – ASA, reproduced from (Rocha de Oliveira and de Juan, 2020).

The maps of the blending batch in Figure 35A demonstrate that the evolution of this blending run can be qualitatively assessed through the visualization of the combined distribution maps at each blending time. At  $t_0$ , before blending started, the segregated ingredients can be visualized in the RGB map. Although a significant reduction in segregation can be observed when looking at the succeeding maps at  $t_{15}$  and  $t_{30}$ , some clumps of pure ingredients are still visible. Then, from  $t_{45}$  onwards, all three components were visually more evenly distributed in the imaged area. In contrast to the good blending represented in Figure 35A, the blending batch shown in Figure 35B presents a deficient blending behavior, clearly visible from blending time  $t_{120}$  and beyond, where the initial blending worsens due to an increase of segregation associated with the formation of large granules of pure CAF (in red).

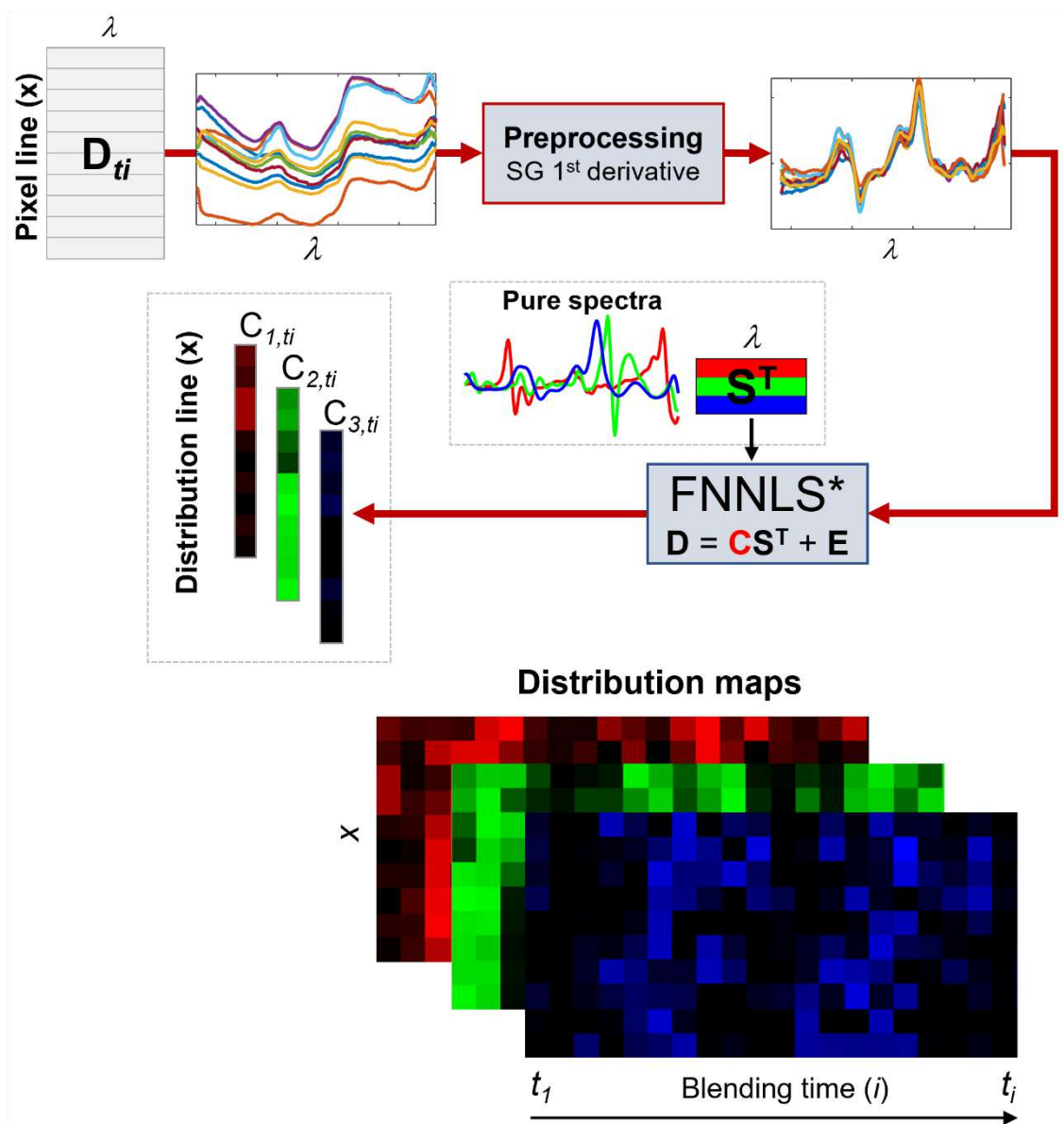


Figure 36 FNNLS generation of concentration maps from scanned lines of pixel spectra. The first step involves preprocessing, which is followed by the FNNLS generation of distribution lines used to build the distribution maps (bottom) by concatenating the distribution lines one after the other for each component.

In the context of a blending process monitored inline with an HSI system, as described in section 3.2.2, the fast non-negativity least squares (FNNLS) algorithm was used to provide the pure component distribution maps. In this context, the iterative *modus operandi* of MCR-ALS was not suitable to derive distribution maps in real-time. FNNLS is a fast algorithm implementation of the non-negativity-constrained linear least squares regression (Bro and Jong, 1997). In the context of a blending process monitored inline with a pushbroom NIR-HSI system, FNNLS works by taking every line of pixel spectra,  $\mathbf{D}$ , and the pure spectral signatures of the blending compounds,  $\mathbf{S}^T$ , obtained as the average of spectra of pure compound NIR images, to obtain a distribution line,  $\mathbf{C}$ , of the components involved in the blending process. The calculation takes as a basis the bilinear model for spectroscopic data ( $\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E}$ ) and the  $\mathbf{C}$  line is calculated under non-negativity constraints. This step is repeated for every new line of pixel spectra obtained during the blending monitoring and the concentration lines in the  $\mathbf{C}$  matrix are arranged to form pure distribution maps for each component of the mixture during the complete blending process, as shown in Figure 36.

Similar to the atline monitored process, the pure component distribution maps can be used for qualitative visualization of the blending process. In this case, FNNLS allows real-time and continuous qualitative heterogeneity assessment during the process. Figure 37 shows some RGB snapshots overlaying the pure component distribution submaps generated by FNNLS at different stages of a three-component blending run of food material. Every submap covers a blending time window of five seconds. The distribution submaps show how the initial highly segregated material, covering the first 5 s of blending, gets gradually mixed until the three components are visually more evenly distributed at the end of the blending run, after almost three minutes of blending time, as shown in the last snapshot.

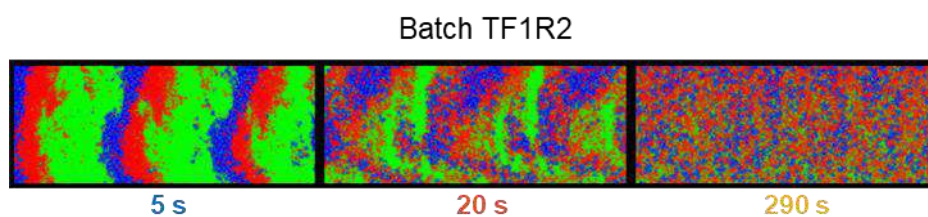


Figure 37 Combined RGB submaps overlaying the pure component distribution submaps (poppy seeds (PS) in red, ground coffee (GC) in green and quinoa seeds (QS) in blue) obtained with FNNLS during a blending run. Submaps show three different 5 s time windows of the blending process. The time displayed corresponds to the end time of the related submap, reproduced from (Rocha de Oliveira and de Juan, 2021a).

The distribution maps generated by MCR-ALS or FNNLS only provide qualitative visual information about the heterogeneity of the mixture during the blending process. To find quantitative heterogeneity information from these maps, a new methodology based on the use of variographic analysis is described in the next step.

### **Step 2. Derivation of heterogeneity indices from distribution maps**

Either for atline or inline monitored processes, the assessment of quantitative heterogeneity information linked to global heterogeneity (GH) or distributional heterogeneity (DH) from the distribution maps is carried out using variographic analysis. A variogram is the graphical representation of the evolution of variance as a function of a lag (expressed in time or distance units).

In the variographic analysis of distribution maps, the variance values are estimated by comparing the concentrations of pixel pairs separated at different lags, in both the horizontal and the vertical direction of the image, using the following equation:

$$V(h) = \frac{1}{2} \cdot \frac{1}{N(h)} \sum_{i=1}^{N(h)} [c(x_i + h) - c(x_i)]^2 \quad (17)$$

where  $V(h)$  is the variance associated with lag ( $h$ ), found as half of the average of the squared differences of all  $N(h)$  pairs of pixel concentration values,  $c(x_i + h)$  and  $c(x_i)$ , separated by a lag distance ( $h$ ). In this thesis, the variance values  $V(h)$  were calculated using the concentration values ( $c$ ) extracted from the distribution maps, but any other pixel property could be used for a variographic analysis.

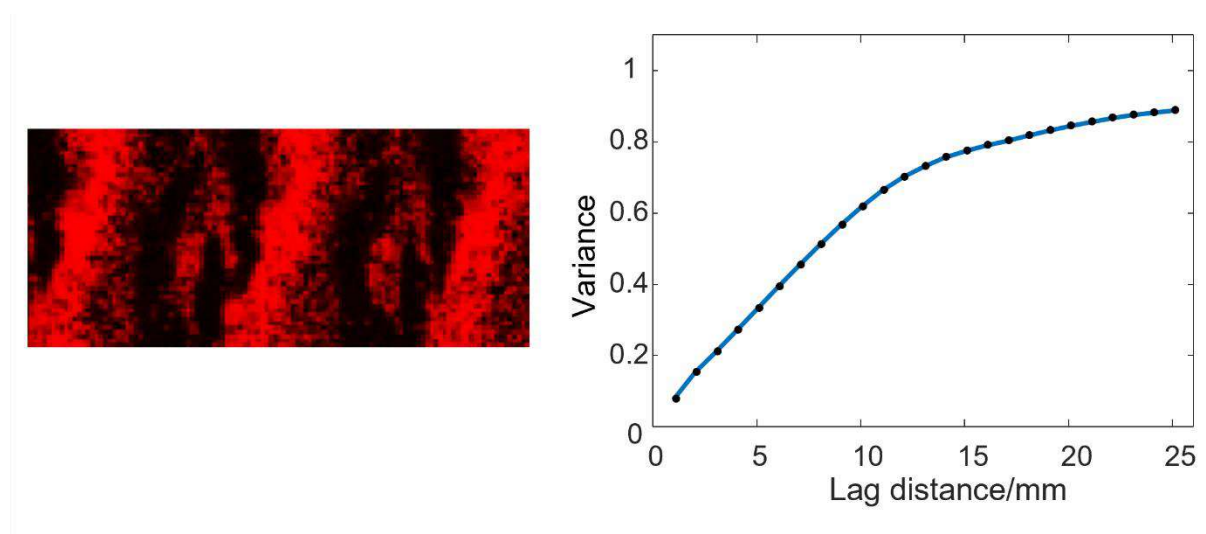


Figure 38 Variogram (right) issued from a distribution map (left) where complete mixing has not been achieved.

Figure 38 shows a typical variogram shape issued from a distribution map where complete mixing has not been achieved. In this instance, pairs of close pixels have more similar concentrations than pixel pairs far away from each other, which is seen because the variance at low lags is smaller than when the lag increases. This phenomenon is linked to the presence of distributional heterogeneity because the material is not evenly distributed. From a certain distance (*range*) and on, the variance stabilizes and there is no spatial correlation among pixel properties anymore. The value of the stabilized variance (*sill*) is related to the scatter among the pixel

concentration values of the full map, linked to the global heterogeneity of the material. These ideas will be used to derive quantitative GH and DH indices.

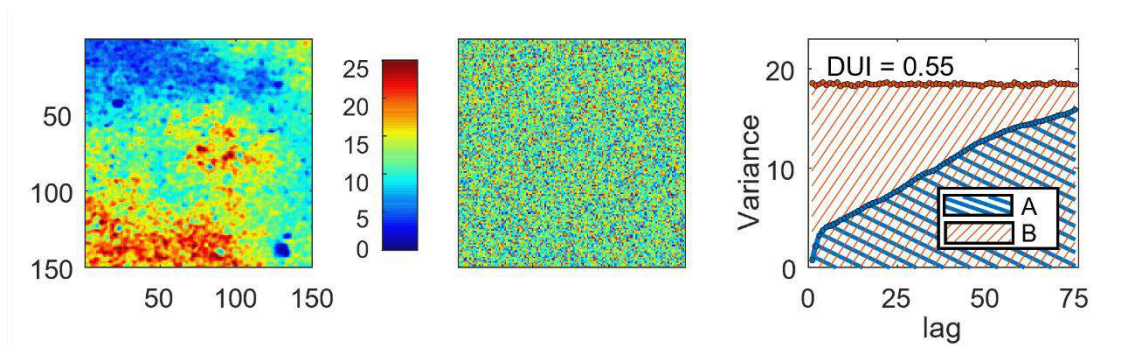


Figure 39. Left, distribution map of starch obtained with MCR-ALS from a blending time of the atline monitored blending process. Middle, randomized map from pixels in the left plot. Right, overlapped variograms from the real map, blue curve, and randomized map, red curve. Blue and red striped areas are related to the areas below the real and randomized maps variogram curves, respectively, reproduced from (Rocha de Oliveira and de Juan, 2020).

To understand how a variographic analysis can be used to extract global heterogeneity (GH) and distributional heterogeneity (DH) information from a distribution map, Figure 39 shows on the left a distribution map from starch obtained using MCR-ALS from an atline monitored blending process. The variogram related to this map is depicted as a blue curve in Figure 39 (right plot). Since a high level of DH still exists, i.e., a big clump of material is seen at the bottom of the map, variance values are smaller for low lag distances and increase as the lag does, until variance stabilization is reached, see Figure 39 (right plot). Figure 39 (middle plot) shows a randomized map obtained shuffling the pixel concentration values of the real starch map (in the left). This randomized map has the same GH as the real map because the pixel concentration values are identical, but it is an excellent reference for the ideal mixing, with no DH, since the distribution of the material is uniform. Because there is a complete lack of spatial correlation among pixel concentration values in the randomized map, a flat variogram with steady variance values for all lag distances is obtained, as depicted by the red flat curve in Figure 39 (right plot). The suitable use of the variogram from the real distribution map (showing the real mixing situation) and the flat variogram of the related randomized map (displaying the reference for the ideal mixture) helps to obtain the two heterogeneity indices related to GH and DH described below.

- i. The Global Heterogeneity Index (GHI). Related to GH, it is estimated as the average of the variance for all lags (*sill*) of the flat variogram from the randomized map. The GHI is mathematically identical to the variance of all pixel concentration values in the distribution map and can be interpreted in absolute or relative terms.
- ii. The Distributional Uniformity Index (DUI). Related to DH, it is estimated as the ratio of the area of the variogram from the real distribution map to the area of the flat variogram from the related randomized map,  $DUI = A/B$ . In Figure 39 (right plot), A is the blue striped area under the variogram for the real map and B is the red

striped area under the variogram for the randomized map. DUI values can vary between 0 and 1. When variograms from real distribution maps are far from their related flat horizontal variogram, i.e., when mixing is deficient and the spatial correlation among pixel properties extends a long distance, low DUI values are obtained, indicating high DH. On the other hand, maps that have real variograms very similar to their randomized map variograms provide DUI values close to 1 indicating that the mixture has low DH and the distribution of the material is almost uniform. In the case of the map shown in Figure 39, the DUI value is 0.55, meaning that 55% of the ideal mixing has been reached.

Both GHI and DUI indices can be obtained from variograms of individual components (*per component*), but also pooled indices can be calculated for several selected components or all components of the blending formulation together (*per sample* or *total mixture*). In this case, the total indices are estimated by averaging the variograms of the individual components. Weighted averages based on concentrations can also be used. Figure 40 shows both GHI (left) and DUI (right) indices calculated from the distribution maps for the atline monitored blending batch shown in Figure 35A. The plots show the evolution of GHI and DUI indices during the blending run *per component* (top plots) and *per total mixture* (bottom plot), taking into account all formulation ingredients in the blending run.

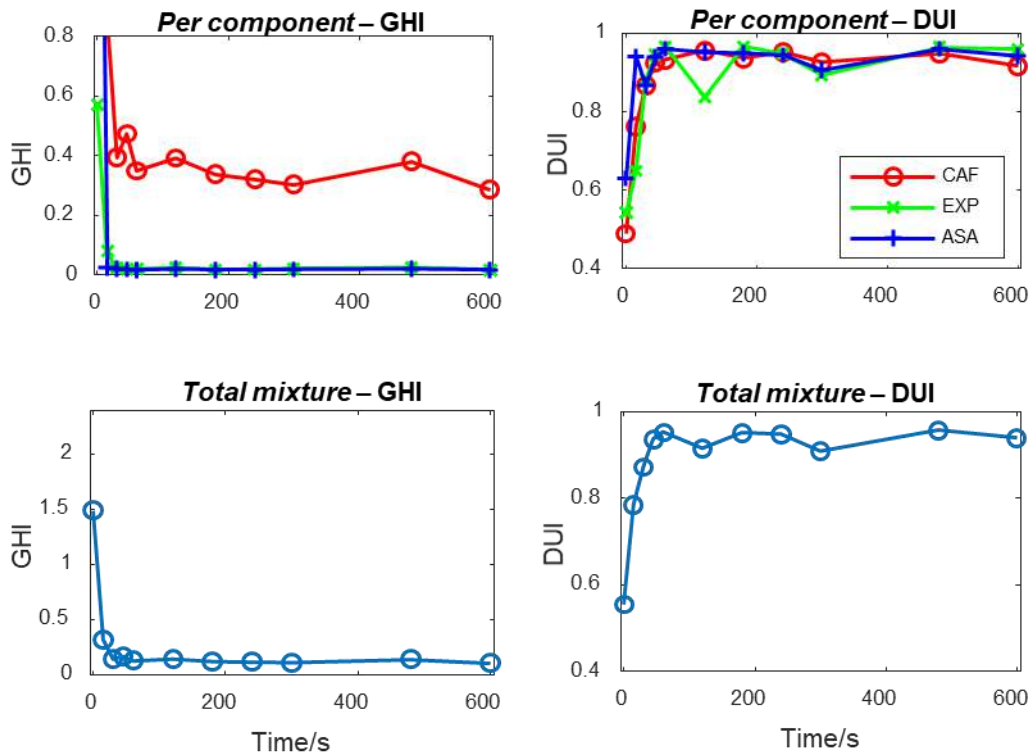


Figure 40 *Per component* GHI (top left) and DUI (top right) indices calculated from the distribution maps for each component of the blend formulation for the atline monitored blending run shown in Figure 35A. Total mixture indices, GHI (bottom left) and DUI (bottom right). Note that some *per component* GHI values are outside the y-axis scale at the beginning of the process, reproduced from (Rocha de Oliveira and de Juan, 2020).

The evolution of the quantitative heterogeneity indices in Figure 40 reflects the qualitative interpretation from Figure 35A, as discussed before. The rapid drop of GHI observed right after the start of the blending run indicates the fast reduction of scattering among pixel concentration values when changing from a completely segregated state at  $t_0$  to a point where the pixel concentration values are more similar and stable, about one minute of blending time ( $t_{60}$ ). Note that relative variances were used to calculate the GHI values; therefore, high GHI values are obtained for caffeine (CAF) since it has a low concentration in this blending run formulation (see Figure 40 top left). When studying the DH looking at the *per component* DUI curves (Figure 40 top right), it can be observed that as blending progressed DH decreased and, consequently, DUI values increased for all components. Indeed, after 200 s of blending time, the DUI curves stabilized for all components of this blending run at values higher than 0.9. Both GHI and DUI curves for the *total mixture* show the expected evolution for a good global blending behavior (see Figure 40, bottom plots).

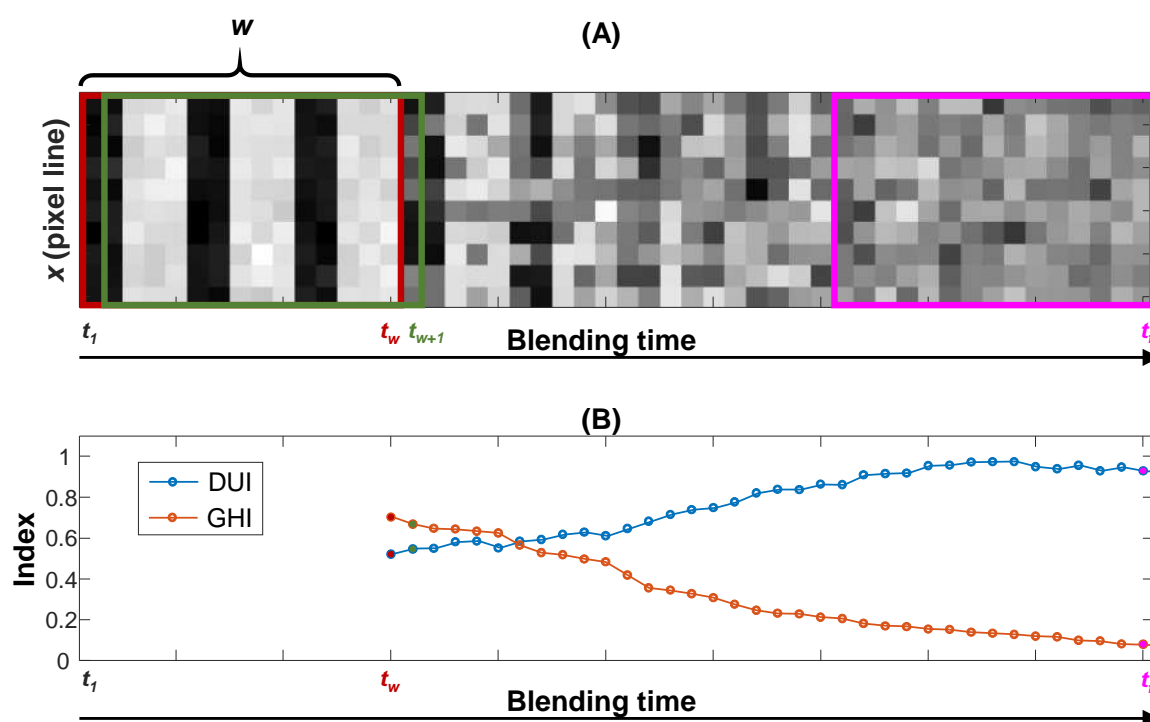


Figure 41 Representation of the continuous extraction of variogram-derived heterogeneity indices from a blending distribution map using the sliding window variographic image analysis (SWiVIA) method. (A) distribution map and the sliding window with width,  $w$ , at time  $t_w$  (first window),  $t_{w+1}$  and  $t_f$ . (B) heterogeneity indices (DUI and GHI) calculated for each map inside the sliding window as a function of the blending time, reproduced from (Rocha de Oliveira and de Juan, 2021a).

These variogram-derived heterogeneity indices can also be used for continuous blending processes monitored inline with an HSI system. For this purpose, and to adapt to the continuously increasing distribution maps, a Sliding Window Variographic Image Analysis (SWiVIA) method has been proposed in this thesis. The SWiVIA method extracts the variogram-derived heterogeneity indices from a submap defined

by a sliding window that moves every time a pixel line ahead until the full map collected during the blending process is covered, as shown in Figure 41. To mimic the kind of distribution map obtained during an inline monitored blending process, Figure 41A shows a simulated distribution map in grayscale where dark and white pixels represent low and high concentrations of the material, respectively. The blending process evolution of the material is shown from left to right in the map, starting at time  $t_1$  and finishing after collecting  $f$  lines of pixels at time  $t_f$ . Bands of segregated material at the beginning of the process fade away as mixing progresses, as reflected by the pixel lines acquired at longer blending times.

The sliding window, sized  $(x \times w)$ , is delimited by the number of pixels ( $x$ ) of the scanned line and the width ( $w$ ) of the sliding window. Therefore, the first full sliding window is obtained once  $w$  lines of  $x$  pixels are collected, as delimited by the red rectangle in Figure 41A, covering from  $t_1$  until  $t_w$ . Then, the related heterogeneity indices (GHI and DUI) can be calculated from the submap inside this first window. Figure 41B shows the two indices, GHI and DUI, from the map inside the window finishing at  $t_w$ . After the next line of pixels is recorded, at  $t_{w+1}$ , the window slides to the right including this new line and discarding the oldest pixel line inside the window at  $t_1$ , as delimited by the green rectangle in Figure 41A. Hence, new variogram-derived indices are calculated for this new window and plotted next to the previous point in Figure 41B. This procedure is repeated for every new line of pixel recorded during the blending process allowing the real-time representation of the heterogeneity evolution until the last line is scanned at  $t_f$ . The last window and its related heterogeneity indices are shown in magenta in Figure 41A and Figure 41B, respectively. This method can be used to generate heterogeneity curves based on indices related to each component (*per component*) or for several selected components or all components of the mixture together (*per total mixture*) as previously described.

The values of the heterogeneity indices calculated from the submap delimited by the sliding window may vary according to the following SWiVIA parameters:

- **Window size ( $w$ ):** The size of the sliding window influences the value of both GHI and DUI heterogeneity indices since the related submap studied can cover a shorter or a longer blending time range. The window size needs to be sufficiently long to provide a good appreciation of the blending state at a certain blending time, but not excessively long to avoid that the indices derived are affected by too old observations that may cause some delays on the perception of the real evolution of the blending progress. Variation in window size may affect the evolution of both GHI and DUI indices.
- **Maximum variogram lag ( $h_{max}$ ):** The  $h_{max}$  parameter limits the distance in which pairs of pixels are compared and, therefore, defines the scale of spatial scrutiny defined to study the distributional heterogeneity. The selection of this parameter will depend on the particle size of the materials mixed (very small  $h_{max}$  parameters are not suitable for materials with big particle size) and on the degree of spatial



precision required for heterogeneity estimation, which may be higher in products where an insufficient blending has more critical effects, such as pharmaceutical blends. Changes in  $h_{max}$  affect only the DUI index.

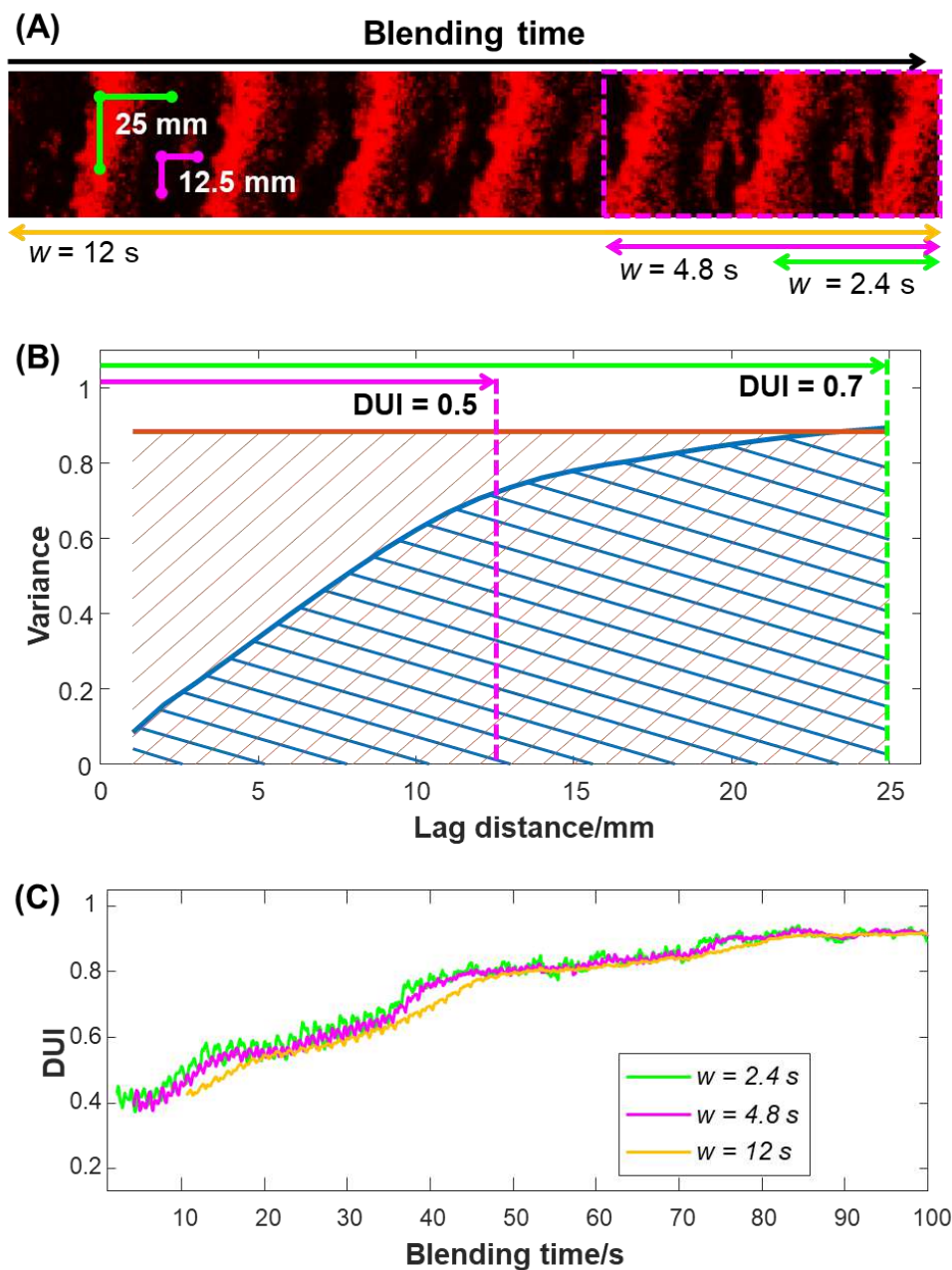


Figure 42 Reference scale for the different parameter settings for the SWIVIA method. (A) Poppy seeds distribution map representing 12 s of blending time at the beginning of a food blending batch and the reference scales for moving window size (2.4, 4.8 and 12 s) and maximum lag (12.5 and 25 mm). (B) Map variogram in blue and global variance reference line in red from the sliding window with size equal to 4.8 s represented by the dashed magenta rectangle in (A) and the DUI values calculated for each maximum lag reference. (C) DUI curve obtained with the SWIVIA method for the first 100 s of a food blending batch using different window sizes and fixed  $h_{max} = 12.5$  mm, reproduced from (Rocha de Oliveira and de Juan, 2021a).

The choice of these two parameters is process-dependent and provides a flexible framework to adapt the blending monitoring to different spatial scales of scrutiny and

higher or lower blending time resolution. Thus, the effect of each parameter on the interpretation of heterogeneity curves should be studied beforehand.

To understand the effect of these two parameters on the SWiVIA results, Figure 42A displays a real distribution map generated by FNNLS from poppy seeds (PS) at the beginning of a food blending run. Reference scales for window size and maximum variogram lag,  $h_{max}$ , are also displayed. The reference scale related to the  $h_{max}$  parameter is shown inside the distribution map in Figure 42A for  $h_{max}$  values of 12.5 and 25 mm. The variogram plot in Figure 42B clearly shows the effect of changing this parameter in the DUI index obtained. Indeed, limiting the lag scale of the variogram at 25 mm or 12.5 mm, the related DUI value changes significantly from 0.7 to 0.5, respectively. For small  $h_{max}$ , the DUI value is smaller than for large  $h_{max}$  because of the different ratios between the two areas used to derive the index. This means that the maximum lag scale should be adjusted to the spatial scale of scrutiny sought for the problem of interest to obtain useful results. The GHI index is not affected by this parameter since it is a constant parameter estimated as the variance of all pixel concentration values inside the full window studied.

The reference scale for different window sizes (related to 2.4, 4.8 and 12 s of blending time) can be seen in Figure 42A. The DUI curves shown in Figure 42C were calculated using the different windows sizes and a fixed  $h_{max} = 12.5$  mm and represent the evolution of the DUI values for the PS component during the first 100 s of a food blending run. The DUI values are associated with the last blending time in the map window studied. Therefore, since the indices shown in Figure 42C all come from full map windows, different starting times are observed in the DUI curves depending on the window size. The three DUI curves obtained using different window sizes start and end at similar DUI values, approximately at 0.4 and 0.9, respectively, and all show a good blending evolution. It is important to note that the longer the window size, the higher the number of past scanned lines included in the map window used to calculate the related heterogeneity indices. Therefore, when using too long window sizes, the global evolution of the DUI curve suffers from a time delay, as shown in Figure 42C for  $w = 12$  s. Such a delay comes from the higher weight of past observations (less mixed) in the index derivation. If the window size is too long, the situation of good mixing or the detection of blending faults may be slower than required. However, although a short window may detect faster variations on heterogeneity, if too short, it may be very sensitive to small and very local variations resulting in a high amplitude fluctuation of the DUI curve. In general, the window selected should have an adequate size to avoid the time delays in the description of the blending evolution associated with too long windows and the presence of local blending phenomena that can hinder the visualization of the global blending trend when too short windows are selected. Although not shown, the same effect of the window size can also be observed on the GHI curves. Since smooth heterogeneity curves with no major time delays were obtained, the SWiVIA-derived heterogeneity indices built with  $w = 4.8$  s were found to be suitable for the applications studied.

Because of the direct influence of the SWiVIA parameters on the generated heterogeneity indices, the selection of both parameters is of great importance and must be adjusted according to the quality requirements defined by the application of interest. In this thesis, the SWiVIA parameters were set to  $w = 4.8$  s and  $h_{max} = 12.5$  mm to generate the heterogeneity indices curves used for the inline monitoring of all blending runs.

### **4.4.2 Use of heterogeneity indices for blending process understanding**

In this subsection, general comments on the results obtained for the real-time monitoring of the blending runs studied in this thesis are presented based on the application of the SWiVIA approach and its derived heterogeneity indices. Special attention is paid to the effect of the physical properties of the materials in their blending behavior and to the description of demixing phenomena.

The SWiVIA results for the continuous monitoring of two replicate food batches of a blending of poppy seeds (PS in red), ground coffee (GC in green) and quinoa seeds (QS in blue), TF1R2 and TF1R1, are shown in Figure 43A and Figure 43B, respectively. Mid and bottom plots show the evolution of the DUI and GHI indices, respectively, for every component in the mixture. The top plot shows some snapshots with the combined RGB submaps overlaying the pure component distribution maps generated by FNNLS at selected time ranges.

The evolution of blending run TF1R2 and submaps used to calculate the indices at 5 s, 20 s and 290 s are shown in Figure 43A. The top left plot shows the mixture distribution submap window (4.8 s) immediately before reaching five seconds of the blending time. This submap clearly shows the initial high segregation level as seen by the separate layers of blending materials at the beginning of the process. The next submap in Figure 43A shows that the segregation level decreased after 20 s of blending time, but still some diffuse layers of clumped material were visible, mainly for GC and QS. Finally, all three components were visually more evenly distributed at the end of the blending run, as shown in the last submap, after 290 s. The quantitative heterogeneity indices generated by the SWiVIA method reflect the visual qualitative information observed in the selected submaps. The *per component* DUI curves for batch TF1R2 show that PS reached high and stable DUI values faster than the other two compounds (red curve in Figure 43A mid panel). This matches the more even spatial distribution of the red component, already seen in the submap at 20 s (top panel in Figure 43A). However, all three components reached the same level of even spatial distribution after approximately 75 s with an average DUI value of 0.9.

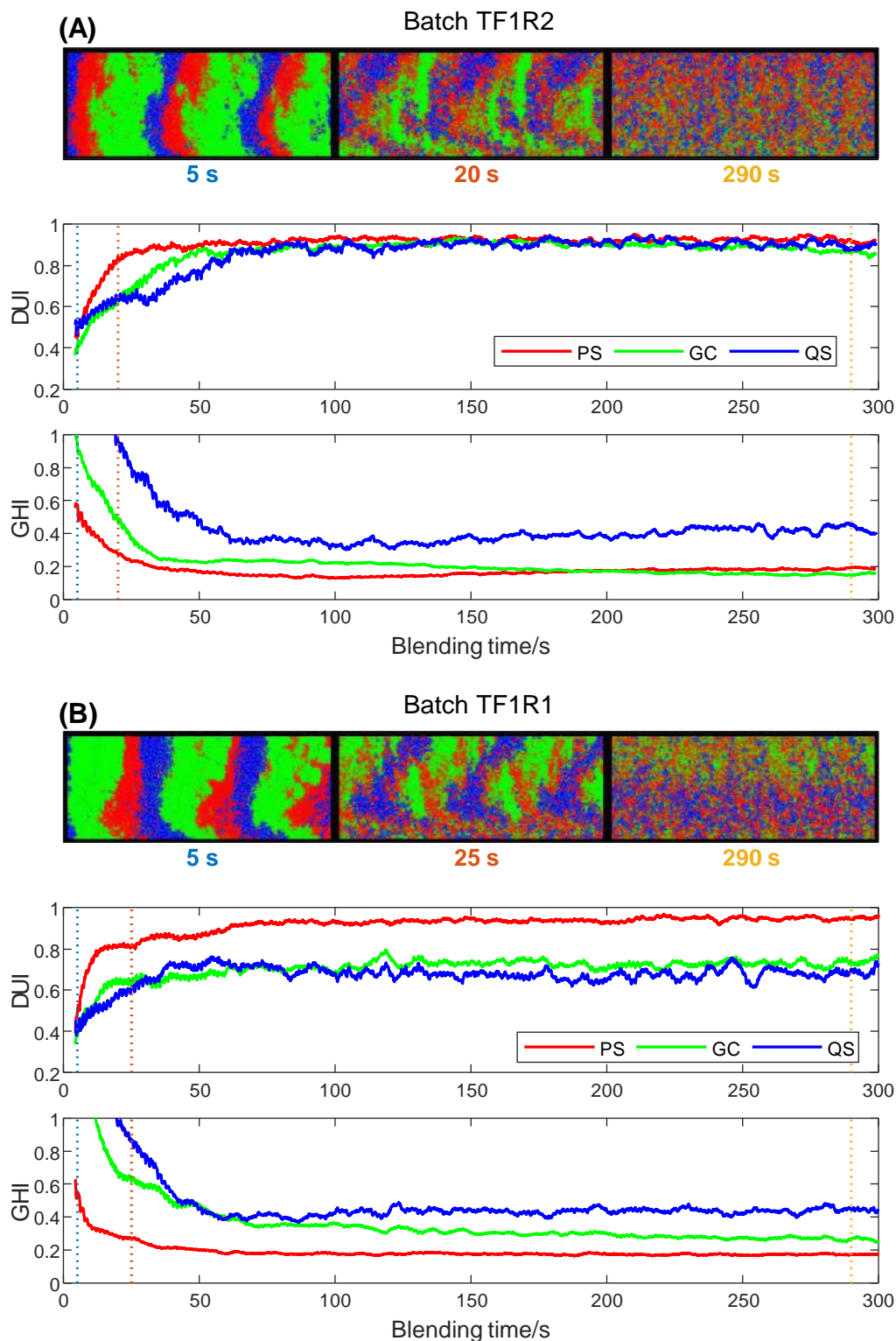


Figure 43 SWiVIA-derived heterogeneity curves for the continuous monitoring of replicate batches TF1R2 (A) and TF1R1 (B). Top panel shows the combined RGB submaps overlaying the pure component distribution submaps (poppy seeds (PS) in red, ground coffee (GC) in green and quinoa seeds (QS) in blue) for selected reference times. Mid and bottom panels show the DUI and GHI curves, respectively. Selected reference times are indicated by the vertical dotted lines, reproduced from (Rocha de Oliveira and de Juan, 2021a).

The *per component* GHI curves for batch TF1R2, in the bottom panel of Figure 43A, show that all components started with high GHI values. This high variance at the beginning of the blending run is related to the presence of a large number of pixels with very high and very low concentration values in the areas with the presence and absence of the segregated material, respectively. Like the DUI curves, the GHI values changed significantly during the first minute of the TF1R2 blending run. However, because this index represents only the relative variance of the concentration values within the submap, independently of the spatial distribution, the GHI curves reached a stabilization at different times than for DUI curves. In this case, the GHI curves for PS and GC reached a stable and similar GHI value after approximately 40 s of blending time with a slight variation around  $GHI = 0.2$  during the rest of the blending run. On the other hand, the GHI of the QS reached a steady  $GHI = 0.4$  after the first minute of the process.

Similar behavior at the beginning of the replicate TF1R1 blending run was observed as shown in Figure 43B. However, after the first minute of the blending run, only PS reached the DUI level of 0.9 as shown in Figure 43B (red curve in the mid panel). The other two components, GC and QS, reached a stabilization of the DUI curve, but with a lower DUI level around 0.7. This can also be visualized in the submap at 290 s, where PS (in red) is evenly distributed, while the green (GC) and blue (QS) components show a poor blend, seen through the accumulation to each side of the submap. The *per component* GHI curves for batch TF1R1, Figure 43B bottom panel, show the low relative variance for PS with GHI values below 0.2 after 60 s of blending time, with higher GHI values for GC and QS. This poor blending behavior can be related to the fact that the rotary blender (glass vial) used to carry out the blending process could be slightly tilted in this run. This probable deficient blending setup can be aggravated by the large differences in physical properties of QS and GC such as shape, density and particle size that might have further contributed to this improper blending.

The overall results for the continuous monitoring of both replicate blending batches using the heterogeneity curves show the expected natural evolution, that is, increasing DUI curves and decreasing GHI curve with blending evolution. Although blending runs TF1R1 and R2 had the same mixture composition, clear differences in the evolution of the heterogeneity curves were observed. This reveals that any blending run should be monitored inline to guarantee final product quality.

As a general trend linked to the blending performance, it seems that differences among physical properties of the particles of the materials to be blended, including shape, size or density may derive in a blending worsening. In this thesis, the limited number of batch runs and blending materials was not sufficient to extract solid conclusions regarding the influence of materials geometry in blending. However, it seems that materials with a spherical shape, such as PS, tend to provide a better mixing and materials formed by fine particles with an easy tendency to agglomerate

offer worse blending and are more prone to de-mixing in general. It is important to note that the de-mixing process is always competing against the blending process. Due to the complex mechanisms involved in the mixing of particulate material and to the unavoidable differences linked to slight variations in the feeding and nature of the initial materials and the blender operation, the required time to reach the endpoint of the blending process is not reproducible and specific control per each blending run should be carried out.

Additionally to the variations in blending behaviour that can be encountered among batch replicates and among components of the same batch, another problem that arises when performing blending processes is the de-mixing phenomenon. De-mixing or segregation can be induced when the blended materials are “overblended”, i.e., mixed for an excessive blending time. In this thesis, the SWiVIA approach allowed the visualization of this phenomenon based on the GHI, and most clearly, the DUI indices during the continuous monitoring of some blending runs. The main factors that induced the de-mixing process in this work were the differences in particle size, density among the materials and the *modus operandi* of the single-axis rotatory blender used.

As an example of the de-mixing phenomenon, Figure 44 shows the results of the pharmaceutical blending run, TP2. Three combined RGB submaps at 5 s, 100 s and 700 s are shown at the top of Figure 44 related to the mixture formed by acetylsalicylic acid (ASA) in red, citric acid (CA) in green and sodium starch glycolate (SSG) in blue. As expected, the initial layers of segregated ingredients are present at the beginning of the process, as shown in the submap at 5 s. The blending evolution of this batch reached the highest distributional homogeneity for all ingredients after approximately 100 s of the blending time, as can be observed by the DUI curves and the related distribution map in Figure 44. Despite the steady DUI curves for the following couple of minutes, the “overblending” caused the de-mixing of ASA and CA reaching at the end of the run DUI levels of approximately 0.8 and 0.6, respectively. Indeed, in the submap at 700 s, more accumulation of the red compound is seen at the bottom whereas the green compound dominates at the top. The SSG, however, kept the same elevated DUI level during the rest of the blending run. In this example, where more than two compounds participate in the blending, it is seen that the blending pattern of every compound is not necessarily the same and that the de-mixing process does not need to involve all blending components in a mixture. In this context, this kind of segregation could have happened if the axis of the rotary blender was slightly tilted. Also, ASA and CA showing clear differences in particle shape and size when compared to SSG, were possibly more prone to show this de-mixing pattern (see Table 1 in section 3.2).

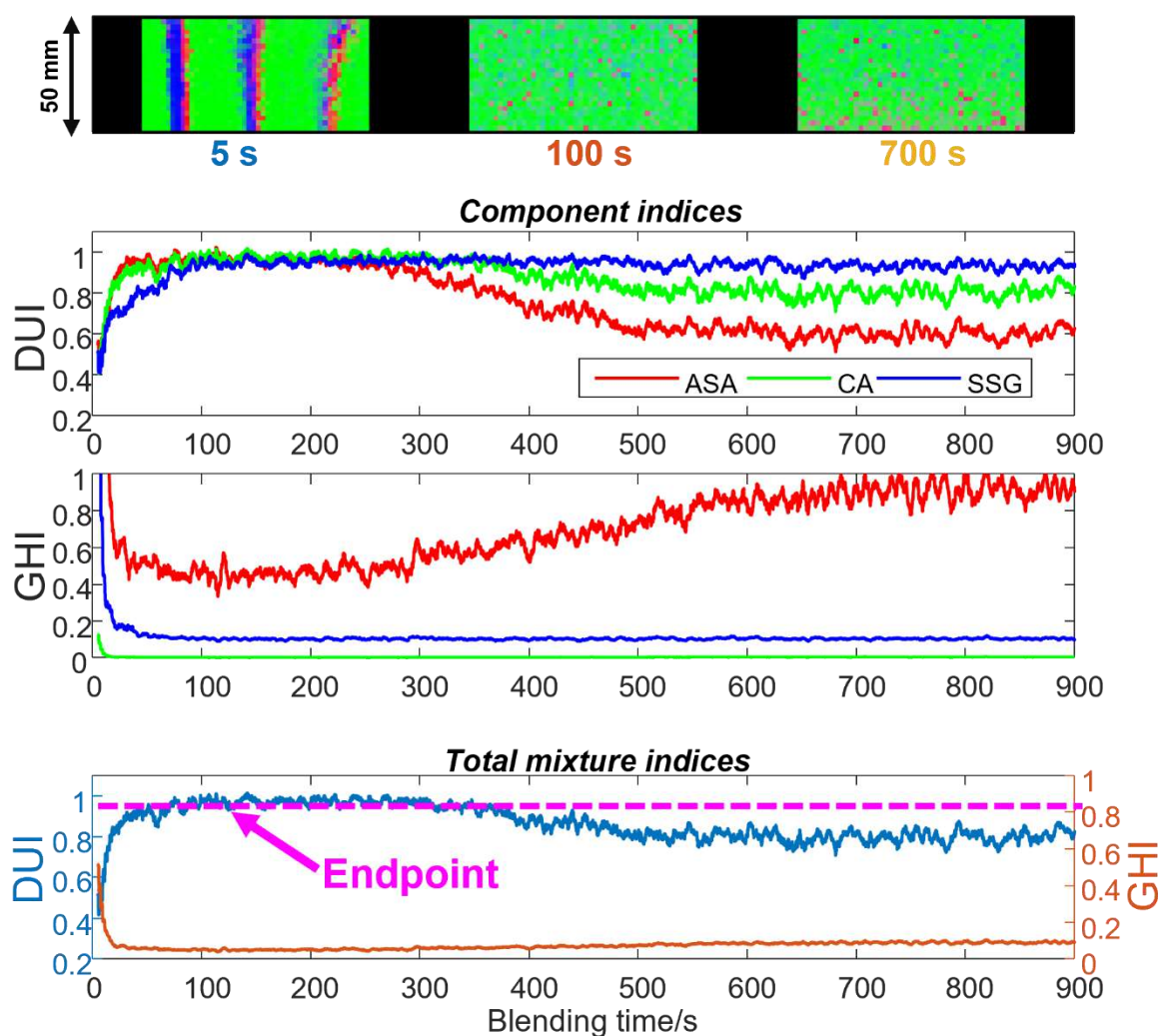


Figure 44 SWiVIA-derived heterogeneity curves for the continuous monitoring of pharmaceutical material blending batch TP2 (second and third middle plot) and combined RGB submaps overlaying the pure component distribution submaps for **acetyl salicylic acid (ASA) in red**, **citric acid (CA) in green** and **sodium starch glycolate (SSG) in blue** (top plot). Per component SWiVIA-derived heterogeneity curves and weighted total mixture heterogeneity curves, DUI (blue) and GHI (orange) (bottom plot), reproduced from (Rocha de Oliveira and de Juan, 2021a).

A good measure to prevent the demixing phenomenon seen in Figure 44 would have been setting a *total mixture* DUI threshold value (displayed in magenta in the bottom plot) to set the blending endpoint. In this way, blending would have been terminated after approximately two minutes with all ingredients evenly distributed, and the demixing phenomenon would have not taken place. In general, prevention of demixing phenomena can be achieved adopting process control models based on the monitoring of the presented GHI and DUI heterogeneity indices. Many possible modalities of endpoint control can be envisaged. Thus, the control of the blending endpoint can be established using the combination of *total mixture* GHI and DUI indices, using predefined blending completion threshold values associated with heterogeneity quality specifications, or using thresholds derived from the study of historical batches having achieved a satisfactory blending. The possibility to obtain DUI and GHI indices associated with every blending component can also point out to

select a critical component, e.g., an active principle of interest in a pharmaceutical formulation, as the key factor to set the blending endpoint. GHI and DUI indices can be used separately or be joined in desirability functions. In any of the forms described above, the use of heterogeneity indices for endpoint detection is an excellent potential approach to save blending time and avoid demixing problems associated with excessive blending.





## **CONCLUSIONS**



---

The general conclusions of this thesis are divided into two blocks, related to Sections I and II of Chapter 4.

**Process monitoring, modeling and control using spectroscopic probes and process sensors**

1. Process modeling of batch process data has been shown to be an essential tool for process understanding and an excellent way to compress complex spectral information into a small number of interpretable profiles. The different strategies used for process modeling allowed process understanding using global abstract PCA components, useful to define process trajectories, or using physicochemical meaningful MCR-ALS components for a more complete process description. PLS models have also helped to describe in real-time the evolution of key parameters along the processes investigated. Compressed MCR-ALS, PLS and other multivariate model outputs have been used alone or in combination with other process variables as input information for the development of data fusion-MSPC models. The batch process trajectories obtained by PCA have been the seeding information for the development of online MSPC models for tracking process evolution in non-synchronized batch processes.
2. PCA-based multivariate statistical process control (MSPC) models are adaptable to handle diverse batch process data for different purposes. In this thesis, we have used different batch data configurations for MSPC model construction. Batch-wise augmented multisets have been used to build MSPC models for synchronized batch data, whereas variable-wise augmented multisets have been shown to be adaptable to handle both non-synchronized and synchronized batch data for different purposes.
3. When using synchronized batch data, multiple batches can be organized in a batch-wise augmented structure. Based on this kind of structure, offline MSPC models have been proposed using the complete batch information to test whether new complete batches followed the normal operating conditions (NOC) or not. In this context, the influence of the input information in the performance of the MSPC models has been tested. In general, information derived from full NIR spectra has been seen to be more useful for process control than univariate sensors, e.g., NIR spectra collected in a distillation towards a temperature distillation profile. MSPC models were also better when using compressed spectral information, as selected spectral ranges, PCA scores or MCR-ALS concentration profiles, than when using the raw full spectra. Finally, narrowing the process region used to the range where more process evolution was seen, e.g., the steep zone in a distillation curve vs. the use of the full curve where flat regions are also incorporated, helped for a better discrimination between on- and off-specification batches.
4. For synchronized batches and batch-wise augmented multisets, new online batch MSPC approaches for real-time tracking of process evolution have been proposed.

The different strategies tested were based on: a) the construction of a local MSPC model for each individual process observation point; b) building local MSPC models considering a fixed size moving window (FSMW-MSPC) covering the current and few past process observation points; and c) building evolving MSPC models, where local PCA models were built with an increasing window of observations covering all points since the beginning of the process until the current observation. The study carried out has shown that online Q control charts were performing correctly for any of the strategies tested in the example of the distillation batches. When using  $D_{stat}$  control charts, local models made on individual observations were shown to be very sensitive to detect process disturbances, but were prone to false alarms. Evolving MSPC models avoided false alarms, but failed to detect small faults or detected them with a certain delay because of the high weight of correct past observations in the model. Finally, the FSMW-MSPC model surmounted the limitations of the two previous approaches, i.e., it was less local than the individual observation model, but also less affected by past observations because a limited window of observations was used in every model. As a consequence, FSMW-MSPC models could successfully detect faulty observations and correctly identify NOC observations without incurring false alarms.

5. MSPC strategies apt to deal with synchronized and non-synchronized batch data have been designed to control process evolution and to detect batch completion (endpoint). For both purposes, variable-wise augmented data configurations have been used. In addition, information coming from different sensor and model outputs have been integrated using mid-level data fusion strategies.
6. Endpoint detection MSPC models were applied to two NIR-monitored real industrial pilot-scale batch processes. In the fluidized bed drying process, a single NIR-based endpoint MSPC model was developed. In a high-temperature multi-step polyester production process, two endpoint MSPC models were designed to flag when each of the two steps of the polyester reaction was complete. The implementation of this PAT tool benefited both process applications by ensuring final product quality consistency, saving time and energy and avoiding waste generation during production.
7. New data fusion strategies to develop MSPC models using the combination of outputs from multivariate models, e.g., PLS predictions or MCR-ALS profiles, issued from the same sensor measurement or the combination of these model outputs with other process sensor outputs, e.g. temperature, have been proposed. These strategies were demonstrated to improve the ability of MSPC charts to detect batch endpoint by increasing the contrast among NOC and faulty observations compared to MSPC models based only on the preprocessed spectra. Similar conclusions were obtained when these DF-MSPC models were used for online MSPC. By using these data fusion strategies, the diagnostic of process upsets gets also more interpretable because the model outputs used provide more

specific information about the process than compressed abstract scores. Besides, it has also been shown that the possible combinations of sensor and model outputs are very diverse and can easily adapt to tailor the DF-MSPC model to control specific process information of interest.

8. A novel synchronization-free methodology for online batch MSPC has been introduced in this thesis. This methodology is based on a first step related to process modeling of NOC batch trajectories using PCA on a variable-wise augmented multibatch data matrix. NOC process trajectories from different batches overlap with each other in the reduced PCA score space, even if they start and end in different points because of the lack of synchronization. This idea is used to build local MSPC models using information from clusters of score observations covering the complete global process trajectory. To test new batch observations, they are projected in all local MSPC models. If a new observation shows a  $Q$  value below the control limit in one or more local MSPC models, the process evolves correctly; if none of the models provides acceptable  $Q$  values, the process starts deviating from the normal evolution expected. For observations in control, additional information about the batch progress along the NOC trajectory can also be achieved. This new methodology circumvents the problem of batch alignment in non-synchronized batch processes, which is a delicate operation that can easily induce artifacts in MSPC models.

### ***Process monitoring using hyperspectral imaging***

9. A new methodology to characterize heterogeneity in material blends from hyperspectral image (HSI) information has been proposed. Inspired by the heterogeneity concept defined in the theory of sampling (TOS) by P. Gy, two heterogeneity aspects are described using chemical images: the global heterogeneity (GH), related to the scatter among the properties of individual pixels, and the distributional heterogeneity (DH), related to the evenness in the spatial distribution of the different materials forming a blend, extracted from the analysis of neighboring pixels or pixel areas. This methodology has proven valuable to study heterogeneity from samples in blending processes monitored atline or inline with a pushbroom NIR-HIS system, but it is easily extendable to handle information collected using other kinds of hyperspectral images or machine vision devices.
10. A first qualitative description of the heterogeneity evolution in blending processes has been carried out by the extraction of pure component distribution maps using image unmixing methods on the image information obtained during the full blending run. MCR-ALS and FNNLS have been the unmixing methods of choice for processes monitored with atline and inline NIR-HSI, respectively. The evolution of the distribution maps with the blending time provides rich information about GH and DH and enables a visual interpretation of the blending progress associated with each of the compounds in the blend formulation. Besides, these distribution maps are the seeding information to derive quantitative heterogeneity indices.

11. Two quantitative and easily interpretable heterogeneity indices have been proposed in this thesis: a global heterogeneity index (GHI) related to GH and a distributional uniformity index (DUI) related to DH. Both GHI and DUI are extracted from the variographic analysis of the compound distribution maps. The designed indices can be used to study the blending behavior of each compound in the mixture or be combined in a pooled index for heterogeneity assessment of the total blend formulation. The versatility and efficiency of this methodology have been demonstrated in atline and inline blending process monitoring using the pushbroom NIR-HSI system.
12. For continuous inline blending monitoring, a methodology called SWiVIA (Sliding Window Variographic Image Analysis) has been designed to estimate GHI and DUI values in real-time, an essential asset to use these indices in real industrial environments. SWiVIA has shown to be tunable to adapt the blending monitoring at different blending time resolutions and different scales of spatial scrutiny, according to the nature and specifications of the blend formulation investigated.
13. SWiVIA has been used to follow several real blending processes using mixtures of materials with diverse physical properties, mainly related to food and pharmaceutical formulations. The GHI and DUI curves derived as a function of the blending time have allowed perceiving easily blending variations linked to the nature of the compounds blended and also differences in behavior among replicate blending batches. Although fully conclusive results could not be obtained, physical parameters such as particle size, shape or density have a clear influence on the blending behavior. Another relevant information easily detected using the SWiVIA approach is the presence of de-mixing phenomena during a blending process, seen with clear drops in the DUI curve after reaching a maximum, that may affect some or all compounds.
14. The risk to incur in demixing phenomena and the possibility to save blending time suggests the use of DUI and GHI indices, per component or per total blend formulation, to set univariate or multivariate statistical process control charts to detect blending endpoint. This would safeguard the final product quality, reduce costs related to unnecessary long blending times and avoid material “overblend” which can lead to the de-mixing phenomenon.

## **REFERENCES**





- Amigo, J.M. (Ed.), 2020. Hyperspectral imaging, in: *Data Handling in Science and Technology*. Elsevier.
- Arden, N.S., Fisher, A.C., Tyner, K., Yu, L.X., Lee, S.L., Kopcha, M., 2021. Industry 4.0 for pharmaceutical manufacturing: Preparing for the smart factories of the future. *Int. J. Pharm.* 602, 120554. <https://doi.org/10.1016/j.ijpharm.2021.120554>
- Arnold, H., 2014. Kommentar: Industrie 4.0: Ohne Sensorsysteme geht nichts [WWW Document]. URL <https://www.elektroniknet.de/messen-testen/industrie-4-0-ohne-sensorsysteme-geht-nichts.110776.html> (accessed 11.5.21).
- Avila, C., Mantzaridis, C., Ferré, J., Rocha de Oliveira, R., Kantojärvi, U., Rissanen, A., Krassa, P., de Juan, A., Muller, F.L., Hunter, T.N., Bourne, R.A., 2021. Acid number, viscosity and end-point detection in a multiphase high temperature polymerisation process using an online miniaturised MEMS Fabry-Pérot interferometer. *Talanta* 224, 121735. <https://doi.org/10.1016/j.talanta.2020.121735>
- Avila, C.R., Ferré, J., de Oliveira, R.R., de Juan, A., Sinclair, W.E., Mahdi, F.M., Hassanpour, A., Hunter, T.N., Bourne, R.A., Muller, F.L., 2020. Process Monitoring of Moisture Content and Mass Transfer Rate in a Fluidised Bed with a Low Cost Inline MEMS NIR Sensor. *Pharm. Res.* 37, 84. <https://doi.org/10.1007/s11095-020-02787-y>
- Boldrini, B., Kessler, W., Rebner, K., Kessler, R.W., 2012. Hyperspectral Imaging: A Review of Best Practice, Performance and Pitfalls for in-line and on-line Applications. *J. Near Infrared Spectrosc.* 20, 483–508. <https://doi.org/10.1255/jnirs.1003>
- Bowler, A.L., Bakalis, S., Watson, N.J., 2020. A review of in-line and on-line measurement techniques to monitor industrial mixing processes. *Chem. Eng. Res. Des.* 153, 463–495. <https://doi.org/10.1016/j.cherd.2019.10.045>
- Bro, R., Jong, S., 1997. A fast non-negativity-constrained least squares algorithm. *J. Chemom.* 11, 393–401.
- Burger, J.L., Schneider, N., Bruno, T.J., 2015. Application of the advanced distillation curve method to fuels for advanced combustion engine gasolines. *Energy and Fuels* 29, 4227–4235. <https://doi.org/10.1021/acs.energyfuels.5b00749>
- Burns, D.A., Ciurczak, E.W., 2009. *Handbook of near-infrared analysis*, 3rd ed. Anal. Bioanal. Chem. <https://doi.org/10.1021/ja015320c>
- Cocchi, M. (Ed.), 2019. *Data Fusion Methodology and Applications*, in: *Data Handling in Science and Technology*. Elsevier Ltd, pp. 1–370. <https://doi.org/10.1016/B978-0-444-63984-4.00001-6>
- Dalitz, F., Cudaj, M., Maiwald, M., Guthausen, G., 2012. Process and reaction monitoring by low-field NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* 60, 52–70. <https://doi.org/10.1016/j.pnmrs.2011.11.003>
- de Juan, A., 2020. Multivariate curve resolution for hyperspectral image analysis. In *Hyperspectral imaging*, in: Amigo, J.M. (Ed.), *Data Handling in Science and Technology*. pp. 115–150. <https://doi.org/10.1016/B978-0-444-63977-6.00007-9>
- de Juan, A., Jaumot, J., Tauler, R., 2014. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods* 6, 4964–4976. <https://doi.org/10.1039/c4ay00571f>

## References

---

- de Juan, A., Rutan, S.C., Tauler, Romà, 2019. Two-Way Data Analysis: Multivariate Curve Resolution, Iterative Methods, in: Brown, S., Tauler, R., Walczak, B. (Eds.), *Comprehensive Chemometrics*. Elsevier, pp. 153–171. <https://doi.org/10.1016/B978-0-12-409547-2.14752-3>
- de Juan, A., Tauler, R., 2021. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review. *Anal. Chim. Acta* 1145, 59–78. <https://doi.org/10.1016/j.aca.2020.10.051>
- de Juan, A., Tauler, R., 2003. Chemometrics applied to unravel multicomponent processes and mixtures. *Anal. Chim. Acta* 500, 195–210. [https://doi.org/10.1016/S0003-2670\(03\)00724-4](https://doi.org/10.1016/S0003-2670(03)00724-4)
- de Oliveira, R.R., Avila, C., Bourne, R., Muller, F., de Juan, A., 2020. Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control. *Anal. Bioanal. Chem.* 412, 2151–2163. <https://doi.org/10.1007/s00216-020-02404-2>
- de Oliveira, R.R., Pedroza, R.H.P., Sousa, A.O., Lima, K.M.G., 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. <https://doi.org/10.1016/j.aca.2017.07.038>
- Drath, R., Horch, A., 2014. Industrie 4.0: Hit or Hype? *IEEE Ind. Electron. Mag.* 8, 56–58. <https://doi.org/10.1109/MIE.2014.2312079>
- Eastment, H.T., Krzanowski, W.J., 1982. Cross-Validatory Choice of the Number of Components From a Principal Component Analysis. *Technometrics* 24, 73–77. <https://doi.org/10.1080/00401706.1982.10487712>
- FDA, 2004. United States Food and Drug Administration, Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance, U.S. Department of Health and Human Services. Rockville, 2004.
- Ferrer-Riquelme, A.J., 2010. Statistical Control of Measures and Processes, in: *Comprehensive Chemometrics*. pp. 97–126. <https://doi.org/10.1016/B978-044452701-1.00096-X>
- Final report of the Industrie 4.0 Working Group, 2013.
- Gómez-Sánchez, A., Marro, M., Marsal, M., Zacchetti, S., Rocha de Oliveira, R., Loza-Alvarez, P., de Juan, A., 2021. Linear unmixing protocol for hyperspectral image fusion analysis applied to a case study of vegetal tissues. *Sci. Rep.* 11, 18665. <https://doi.org/10.1038/s41598-021-98000-0>
- González-Martínez, José M., de Noord, O.E., Ferrer, A., 2014. Multisynchro: A novel approach for batch synchronization in scenarios of multiple asynchronisms. *J. Chemom.* 28, 462–475. <https://doi.org/10.1002/cem.2620>
- González-Martínez, J. M., Vitale, R., De Noord, O.E., Ferrer, A., 2014. Effect of synchronization on bilinear batch process modeling. *Ind. Eng. Chem. Res.* 53, 4339–4351. <https://doi.org/10.1021/ie402052v>
- Grassi, S., 2020. Configuration of hyperspectral and multispectral imaging systems. *Data Handl. Sci. Technol.* 32, 17–34. <https://doi.org/10.1016/B978-0-444-63977-6.00002-X>

- Grassi, S., Strani, L., Casiraghi, E., Alamprese, C., 2019. Control and monitoring of milk renneting using FT-NIR spectroscopy as a process analytical technology tool. *Foods* 8. <https://doi.org/10.3390/foods8090405>
- Green, R.L., Thureau, G., Pixley, N.C., Mateos, A., Reed, R.A., Higgins, J.P., 2005. In-Line Monitoring of Moisture Content in Fluid Bed Dryers Using Near-IR Spectroscopy with Consideration of Sampling Effects on Method Accuracy. *Anal. Chem.* 77, 4515–4522. <https://doi.org/10.1021/ac050272q>
- Haaland, D.M., Thomas, E.V., 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.
- Hermann, M., Pentek, T., Otto, B., 2016. Design Principles for Industrie 4.0 Scenarios, in: 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, pp. 3928–3937. <https://doi.org/10.1109/HICSS.2016.488>
- Jackson, J.E., 1991. *A User's Guide to Principal Components*, Wiley Series in Probability and Statistics. Wiley, New York.
- Jackson, J.E., Mudholkar, G.S., 1979. Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics* 21, 341–349. <https://doi.org/10.1080/00401706.1979.10489779>
- Jaumot, J., de Juan, A., Tauler, R., 2014. MCR-ALS GUI 2.0: New features and applications. *Chemom. Intell. Lab. Syst.* 140, 1–12. <https://doi.org/10.1016/j.chemolab.2014.10.003>
- Jolliffe, I.T., 2002. *Principal components analysis*, 2nd ed. Springer, New York.
- Kassidas, A., Macgregor, J.F., Taylor, P.A., 1998. Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J.* 44, 864–875. <https://doi.org/10.1002/aic.690440412>
- Kourti, T., 2009. Multivariate Statistical Process Control and Process Control, Using Latent Variables, in: *Comprehensive Chemometrics*. Elsevier, pp. 21–54. <https://doi.org/10.1016/B978-044452701-1.00013-2>
- Kourti, T., 2002. Process analysis and abnormal situation detection: From theory to practice. *IEEE Control Syst. Mag.* 22, 10–25. <https://doi.org/10.1109/MCS.2002.1035214>
- Kourti, T., MacGregor, J.F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemom. Intell. Lab. Syst.* 28, 3–21. [https://doi.org/10.1016/0169-7439\(95\)80036-9](https://doi.org/10.1016/0169-7439(95)80036-9)
- Lee, E.A., 2008. Cyber Physical Systems: Design Challenges, in: 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC). IEEE, pp. 363–369. <https://doi.org/10.1109/ISORC.2008.25>
- Liu, Y.J., André, S., Saint Cristau, L., Lagresle, S., Hannas, Z., Calvosa, É., Devos, O., Duponchel, L., 2017. Multivariate statistical process control (MSPC) using Raman spectroscopy for in-line culture cell monitoring considering time-varying batches synchronized with correlation optimized warping (COW). *Anal. Chim. Acta* 952, 9–17. <https://doi.org/10.1016/j.aca.2016.11.064>
- MacGregor, J.F., Kourti, T., 1995. Statistical process control of multivariate processes. *Control Eng. Pract.* 3, 403–414. [https://doi.org/10.1016/0967-0661\(95\)00014-L](https://doi.org/10.1016/0967-0661(95)00014-L)

## References

---

- Maeder, M., 1987. Evolving Factor Analysis for the Resolution of Overlapping Chromatographic Peaks. *Anal. Chem.* 59, 527–530.
- Martens, H., Næs, T., 1991. *Multivariate Calibration*. John Wiley & Sons, New York.
- Næs, T., 2004. *A user-friendly guide to multivariate calibration and classification*. NIR Publications, Chichester, UK.
- Nieuwmeyer, F.J.S., Damen, M., Gerich, A., Rusmini, F., Van Der Voort Maarschalk, K., Vromans, H., 2007. Granule characterization during fluid bed drying by development of a near infrared method to determine water content and median granule size. *Pharm. Res.* 24, 1854–1861. <https://doi.org/10.1007/s11095-007-9305-5>
- Nomikos, P., MacGregor, J.F., 1995. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* 37, 41–59. <https://doi.org/10.1080/00401706.1995.10485888>
- Ozaki, Y., Morisawa, Y., 2021. Principles and Characteristics of NIR Spectroscopy, in: *Near-Infrared Spectroscopy*. Springer Singapore, Singapore, pp. 11–35. [https://doi.org/10.1007/978-981-15-8648-4\\_2](https://doi.org/10.1007/978-981-15-8648-4_2)
- Pasquini, C., 2018. *Analytica Chimica Acta* Near infrared spectroscopy : A mature analytical technique with new perspectives e A review. *Anal. Chim. Acta* 1026, 8–36. <https://doi.org/10.1016/j.aca.2018.04.004>
- Pasquini, C., Scafi, S.H.F., 2003. Real-Time Monitoring of Distillations by Near-Infrared Spectroscopy. *Anal. Chem.* 75, 2270–2275. <https://doi.org/10.1021/ac034054d>
- Peinado, A., Hammond, J., Scott, A., 2011. Development, validation and transfer of a Near Infrared method to determine in-line the end point of a fluidised drying process for commercial production batches of an approved oral solid dose pharmaceutical product. *J. Pharm. Biomed. Anal.* 54, 13–20. <https://doi.org/10.1016/j.jpba.2010.07.036>
- Piqueras, S., Duponchel, L., Tauler, R., De Juan, A., 2011. Resolution and segmentation of hyperspectral biomedical images by Multivariate Curve Resolution-Alternating Least Squares. *Anal. Chim. Acta* 705, 182–192. <https://doi.org/10.1016/j.aca.2011.05.020>
- Ramaker, H.J., Van Sprang, E.N.M., Westerhuis, J.A., Smilde, A.K., 2004. The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring. *Chemom. Intell. Lab. Syst.* 73, 181–187. <https://doi.org/10.1016/j.chemolab.2003.12.015>
- Rifkin, J., 2016. The 2016 World Economic Forum Misfires with its Fourth Industrial Revolution Theme [WWW Document]. *IndustryWeek*. URL <https://www.industryweek.com/technology-and-iiot/information-technology/article/21967057/the-2016-world-economic-forum-misfires-with-its-fourth-industrial-revolution-theme> (accessed 11.4.21).
- Rocha de Oliveira, R., de Juan, A., 2021a. SWIVIA – Sliding window variographic image analysis for real-time assessment of heterogeneity indices in blending processes monitored with hyperspectral imaging. *Anal. Chim. Acta* 1180, 338852. <https://doi.org/10.1016/j.aca.2021.338852>
- Rocha de Oliveira, R., de Juan, A., 2021b. Synchronization-Free Multivariate Statistical Process Control for Online Monitoring of Batch Process Evolution. *Front. Anal. Sci.* Submitted.

- Rocha de Oliveira, R., de Juan, A., 2020. Design of Heterogeneity Indices for Blending Quality Assessment Based on Hyperspectral Images and Variographic Analysis. *Anal. Chem.* 92, 15880–15889. <https://doi.org/10.1021/acs.analchem.0c03241>
- Roser, C., 2015. A Critical Look at Industry 4.0 [WWW Document]. URL <https://www.allaboutlean.com/industry-4-0/> (accessed 11.5.21).
- Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* 36, 1627–1639. <https://doi.org/10.1021/ac60214a047>
- Takahashi, M.B., Leme, J., Caricati, C.P., Tonso, A., Fernández Núñez, E.G., Rocha, J.C., 2015. Artificial neural network associated to UV/Vis spectroscopy for monitoring bioreactions in biopharmaceutical processes. *Bioprocess Biosyst. Eng.* 38, 1045–1054. <https://doi.org/10.1007/s00449-014-1346-7>
- Tauler, R., 1995. Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst.* 30, 133–146. [https://doi.org/10.1016/0169-7439\(95\)00047-X](https://doi.org/10.1016/0169-7439(95)00047-X)
- Tauler, R., Kowalski, B.R., Fleming, S., 1993. Multivariate Curve Resolution Applied to Spectral Data from Multiple Runs of an Industrial Process. *Anal. Chem.* 65, 2040–2047.
- Tauler, R., Maeder, M., de Juan, A., 2009. Multiset Data Analysis: Extended Multivariate Curve Resolution, in: *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis Four-Volume Set. Vol. 2, Chapter 2.24*, S.D. Brown, R. Tauler, B. Walczak. Elsevier, pp. 473–505.
- Tauler, R., Smilde, A., Kowalski, B.R., 1995. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemom.* 9, 31–58.
- Tauler, Romà, Maeder, M., de Juan, A., 2020. Multiset Data Analysis: Extended Multivariate Curve Resolution, in: Brown, S., Tauler, R., Walczak, B. (Eds.), *Comprehensive Chemometrics*. Elsevier, pp. 305–336. <https://doi.org/10.1016/B978-0-12-409547-2.14702-X>
- Westad, F., Gidskehaug, L., Swarbrick, B., Flåten, G.R., 2015. Assumption free modeling and monitoring of batch processes. *Chemom. Intell. Lab. Syst.* 149, 66–72. <https://doi.org/10.1016/j.chemolab.2015.08.022>
- Windig, W., Guilment, J., 1991. Interactive Self-Modeling Mixture Analysis. *Anal. Chem.* 63, 1425–1432.
- Wold, S., Esbensen, K., Geladi, P., 1987a. Principal component analysis. *Chemom. Intell. Lab. Syst.* 2, 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9)
- Wold, S., Geladi, P., Esbensen, K., Öhman, J., 1987b. Multi-way principal components-and PLS-analysis. *J. Chemom.* 1, 41–56. <https://doi.org/10.1002/cem.1180010107>
- Wold, S., Kettaneh-Wold, N., MacGregor, J.F., Dunn, K.G., 2009. 2.10 - Batch Process Modeling and MSPC, in: *Comprehensive Chemometrics*. pp. 163–197. <https://doi.org/10.1016/B978-044452701-1.00108-3>
- Wold, S., Kettaneh, N., Fridén, H., Holmberg, A., 1998. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* 44, 331–340. [https://doi.org/10.1016/S0169-7439\(98\)00162-2](https://doi.org/10.1016/S0169-7439(98)00162-2)

## References

---

- Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58, 109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
- Zhao, J., Li, W., Qu, H., Tian, G., Wei, Y., 2020. Real-time monitoring and fault detection of pulsed-spray fluid-bed granulation using near-infrared spectroscopy and multivariate process trajectories. *Particuology* 53, 112–123. <https://doi.org/10.1016/j.partic.2020.02.003>





