Università degli Studi di Modena e Reggio Emilia

Dottorato di ricerca in
"Models and Methods for Material and Environmental Sciences"

Ciclo XXXVII

Designing new tools for real-time product quality control towards sustainable manufacturing

in co-tutela con l'Université de Lille

Candidato : **Tanzilli Daniele**

Relatore italiano (Tutor) : Prof.ssa Marina Cocchi

Relatore francese (Tutor) : Prof. Cyril Ruckebusch

Correlatore francese (Cotutor) : Prof. Raffaele Vitale

Correlatore italiano (Cotutor) : Dr. Lorenzo Strani

Coordinatore del Corso di Dottorato : Prof. Stefano Lugli

Université de Lille

École doctorale Science de la Matière, du Rayonnement et de l'Environnement

Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement

---

Designing new tools for real-time product quality control towards sustainable manufacturing

Conception de nouveaux outils pour le contrôle en temps réel de la qualité des produits en vue d'une fabrication durable

---

Thèse préparée et soutenue publiquement par Daniele Tanzilli le 06/05/2025, pour obtenir le grade de Docteur en Chimie théorique, physique, analytique

Thèse dirigée par :

Prof. Cyril Ruckebusch – Université de Lille

Prof. Raffaele Vitale – Université de Lille

Prof. Marina Cocchi – Università di Modena e Reggio Emilia

Dr. Lorenzo Strani – Università di Modena e Reggio Emilia

Composition du jury :

Prof. José Camacho Páez – *Rapporteur*, Université de Granada

Prof. Barbara Giussani – *Rapporteuse*, Université de l'Insubria

Prof. Federico Marini – *Examinateur*, Université de Rome "La Sapienza"

Prof. Pierantonio Facco – *Examinateur*, Université de Padova

Prof. Cyril Ruckebusch – *Directeur de thèse*, Université de Lille

Prof. Marina Cocchi – *Directrice de thèse*, Università di Modena e Reggio Emilia

Università degli Studi di Modena e Reggio Emilia

Models and Methods for Material and Environmental Sciences

Ciclo XXXVII

Université de Lille

Ecole doctorale Science de la Matière, du Rayonnement et de l'Environnement

---

Designing new tools for real-time product quality control towards sustainable manufacturing

Conception de nouveaux outils pour le contrôle en temps réel de la qualité des produits en vue d'une fabrication durable

---

Candidato : **Tanzilli Daniele**

Relatore italiano (Tutor) : Prof.ssa Marina Cocchi
Relatore francese (Tutor) : Prof. Cyril Ruckebusch
Correlatore francese (Cotutor) : Prof. Raffaele Vitale
Correlatore italiano (Cotutor) : Dr. Lorenzo Strani

A Francesca

Contents

ABSTRACT

In the era of Industry 4.0, real-time process monitoring and quality control are crucial for industries aiming at optimising production, minimising waste and improving sustainability. Chemometrics is increasingly being considered in this context, providing efficient tools to deal with large amounts of data and extract useful information for enabling data-driven decision making. Among these tools, Partial Least Squares regression (PLS) is widely used for real-time quality prediction, a key step to monitor the behaviour of industrial processes. However, when dealing with data collected during different stages of a process or that encode information not accurately reflecting the relevant variability of the process itself, PLS can face significant challenges due to the presence of strong non-linearities and severe outliers. These issues can negatively affect predictive performance. Furthermore, the common need of processing multiple blocks of data of diverse nature at the same time may sometimes render such scenarios even more challenging.

This thesis addresses these issues by introducing novel PLS-based algorithms.

Two of the proposed methods, Locally-Weighted-Multiblock-PLS (LW-MB-PLS) and Local- Weighted-Response Oriented Sequential Alternation (LW-ROSA) were designed to handle non-linearities in situations where multiple data blocks are handled. In particular, for LW-MB-PLS, which fits a separate linear calibration model for each new incoming sample, a particular focus was also made on the interpretability of the local sub-models.

To deal with both non-linearities and outliers simultaneously, this thesis introduces another novel method called Locally-Weighted-RoBoost-PLS (LW-RoBoost-PLS). This method combines the strengths of local modelling strategies to effectively handle non-linearities with those of robust regression techniques to mitigate the impact of outliers. It has also been extended to the multiblock framework.

These methods were specifically developed and evaluated in two industrial continuous process scenarios.

The first concerns a real Industry 4.0 setting, i.e. a chemical plant owned by Versalis S.p.A. where different in-line sensors monitor the nine-stage production of Acrylonitrile Butadiene Styrene (ABS). In this context, LW-MB-PLS, LW-ROSA and LW-RoBoost-PLS were found to yield significant improvements in terms of predictive performance compared to traditional approaches. In addition, in order to determine whether satisfactory estimates of ABS quality could be obtained before the end of the production process, the ensemble of data blocks resulting from all the process phases was modelled according to an incremental scheme. This way, it was possible to assess the evolution of the predictive ability of the aforementioned regression methods as the amount of information available increased.

The second scenario relates instead to the production process of pesto alla Genovese sauce at Barilla G. e R. F.lli S.p.A. Here, typical Industry 4.0 challenges such as data synchronization as well as the construction of real-time control charts and predictive models were tackled. The analysis of images delivered by cameras installed along the manufacturing line was also performed to evaluate the quality of the raw materials exploited for the pesto sauce preparation.

The proposed algorithms and methodologies constitute a significant contribution to the field of complex process data analysis, permitting to achieve improved predictive performance while overcoming several of the main limitations traditional multivariate regression approaches such as standard PLS suffer from. The results reported in this thesis could potentially boost the advancement of the domain of Industry 4.0, enhancing process control and sustainability.

RIASSUNTO

Nell'era dell'Industria 4.0, il monitoraggio in tempo reale dei processi e il controllo della qualità rappresentano strumenti cruciali per le aziende che puntano a ottimizzare la produzione, ridurre gli sprechi e incrementare la sostenibilità. In questo scenario, la chemiometria assume un ruolo sempre più rilevante, offrendo metodologie efficaci per gestire grandi quantità di dati e per estrarre informazioni utili a supporto di decisioni basate sui dati. Tra gli strumenti più diffusi in questo ambito, la regressione Partial Least Squares (PLS) è ampiamente utilizzata per la previsione della qualità in tempo reale, aspetto centrale per il monitoraggio del comportamento dei processi industriali. Tuttavia, quando i dati riflettono fasi differenti di un processo o includono informazioni che non catturano adeguatamente la variabilità del sistema, PLS può incontrare difficoltà, a causa della presenza di non linearità e outliers. Inoltre, la necessità di trattare simultaneamente più blocchi di dati eterogenei rende questi scenari ancora più complessi.

Questa tesi affronta tali problematiche proponendo nuovi algoritmi basati su PLS. Due di essi, Locally-Weighted-Multiblock-PLS (LW-MB-PLS) e Local-Weighted-Response Oriented Sequential Alternation (LW-ROSA), sono stati ideati per gestire le non linearità nei casi in cui è necessario trattare blocchi di dati multipli. Il metodo LW-MB-PLS adatta un modello di calibrazione lineare locale per ogni nuovo campione, mantenendo al contempo un'attenzione particolare all'interpretabilità dei modelli locali.

Per affrontare simultaneamente le non linearità e la presenza di outliers, è stato sviluppato un terzo metodo, Locally-Weighted-RoBoost-PLS (LW-RoBoost-PLS), che va a combinare i vantaggi delle tecniche di modellazione locale, che gestiscono in modo efficace le non linearità, con quelli delle metodologie di regressione robusta, che riducono l'influenza degli outliers. Il metodo è stato inoltre esteso per l'applicazione a scenari multiblock.

I metodi proposti sono stati testati in due contesti industriali. Il primo caso studio riguarda la produzione di Acrilonitrile Butadiene Stirene (ABS) presso Versalis S.p.A., un impianto dotato di sensori in-linea che monitorano diverse fasi del processo produttivo. In questo contesto, i metodi LW-MB-

PLS, LW-ROSA e LW-RoBoost-PLS hanno mostrato significativi miglioramenti nelle prestazioni predittive rispetto agli approcci tradizionali. Per valutare la possibilità di predire accuratamente la qualità dell'ABS prima del completamento del processo, è stato implementato un approccio incrementale, basato sull'uso progressivo delle informazioni disponibili. Questo metodo ha consentito di analizzare l'evoluzione della capacità predittiva dei metodi proposti, fornendo indicazioni preziose sulla loro efficacia nelle varie fasi del processo.

Il secondo caso studio si focalizza sulla produzione di pesto alla genovese presso Barilla G. e R. F.lli S.p.A., un esempio di Industria 4.0 in cui l'infrastruttura non è completamente equipaggiata per una raccolta dati completa. In questo contesto, sono state affrontate sfide come la sincronizzazione dei dati, lo sviluppo di modelli predittivi in tempo reale ed un'analisi delle immagini per valutare la qualità delle materie prime impiegate nella preparazione del pesto.

Gli algoritmi e le metodologie presentati in questa tesi rappresentano un contributo significativo all'analisi di dati di processo complessi, migliorando le prestazioni predittive e superando molte delle limitazioni dei metodi tradizionali, come PLS. I risultati ottenuti possono contribuire all'evoluzione dell'Industria 4.0, migliorando il controllo dei processi, l'efficienza produttiva e la sostenibilità complessiva delle operazioni industriali.

RESUMÉ

Dans l'ère de l'Industrie 4.0, le contrôle en temps réel des processus et de la qualité est crucial pour optimiser la production, minimiser les déchets et améliorer la durabilité. La chimiométrie joue un rôle clé en exploitant de grandes quantités de données pour une prise de décision basée sur les données. La régression par moindres carrés partiels (PLS) est largement utilisée pour la prédiction en temps réel de la qualité, une étape essentielle du suivi des processus industriels. Cependant, lorsque l'on travaille avec des données collectées à différentes étapes d'un processus ou contenant des informations qui ne reflètent pas précisément les variations pertinentes du processus lui-même, la PLS peut rencontrer des défis importants en raison de fortes non-linéarités et de la présence d'outliers sévères. Ces problèmes peuvent nuire à la performance prédictive. De plus, la nécessité de traiter simultanément plusieurs blocs de données de nature diverse peut rendre ces scénarios encore plus complexes.

Cette thèse aborde ces problématiques en introduisant de nouveaux algorithmes basés sur la PLS.

Deux des méthodes proposées, la PLS multibloc pondérée localement (LW-MB-PLS) et l'alternance séquentielle orientée-réponse et pondérée localement (LW-ROSA), ont été conçues pour gérer les non-linéarités dans des situations impliquant plusieurs blocs de données. Pour la LW-MB-PLS qui ajuste un modèle de calibration linéaire distinct pour chaque nouvel échantillon entrant, une attention particulière a également été accordée à l'interprétabilité des sous-modèles locaux.

Pour traiter simultanément des non-linéarités et des outliers, cette thèse introduit une autre méthode novatrice appelée LW-RoBoost-PLS (Locally-Weighted-RoBoost-PLS). Cette méthode combine les avantages des stratégies de modélisation locale pour gérer efficacement les non-linéarités avec ceux des techniques de régression robuste pour atténuer l'impact des outliers. Elle a également été étendue au cadre multibloc.

Ces méthodes ont été spécifiquement développées et évaluées dans deux

scénarios industriels de processus continus.

Le premier concerne un cadre réel de l'Industrie 4.0, à savoir une usine chimique appartenant à Versalis S.p.A., où différents capteurs en ligne surveillent les neuf étapes de production de l'Acrylonitrile Butadiène Styrène (ABS). Dans ce contexte, les méthodes LW-MB-PLS, LW-ROSA et LW-RoBoost-PLS ont permis des améliorations significatives en termes de performance prédictive par rapport aux approches traditionnelles. En outre, afin de déterminer si des estimations satisfaisantes de la qualité de l'ABS pouvaient être obtenues avant la fin du processus de production, l'ensemble des blocs de données résultant de toutes les phases du processus a été modélisé selon un schéma incrémental. De cette manière, il a été possible d'évaluer l'évolution des capacités prédictives des méthodes de régression susmentionnées à mesure que la quantité d'informations disponibles augmentait.

Le second scénario concerne le processus de production de la sauce pesto alla Genovese chez Barilla G. e R. F.lli S.p.A. Ici, des défis typiques de l'Industrie 4.0 tels que la synchronisation des données ainsi que la construction de cartes de contrôle en temps réel et de modèles prédictifs ont été relevés. L'analyse d'images capturées par des caméras installées le long de la ligne de production a également été réalisée afin d'évaluer la qualité des matières premières utilisées pour la préparation de la sauce pesto.

Les algorithmes proposés apportent une contribution majeure à l'analyse des processus industriels complexes, améliorant les performances prédictives tout en surmontant les limitations des approches traditionnelles comme la PLS standard. Les résultats obtenus pourraient accélérer les avancées en Industrie 4.0, en renforçant le contrôle des processus et la durabilité.

List of Abbreviations

— **ABS** - Acrylonitrile-Butadiene-Styrene
— **AI** - Artificial Intelligence
— **CA** - Approximation Sub-Image
— **CD** - Diagonal Detail
— **CH** - Horizontal Detail
— **CNN** - Convolutional Neural Network
— **CV** - Cross Validation
— **CV-ANOVA** - Cross-Validation Analysis of Variance
— **CUSUM** - Cumulative Sum Control Chart
— **EWMA** - Exponentially Weighted Moving Average
— **FDA** - The Food and Drug Administration
— **FIR** - Far-Infrared
— **GPU** - Graphics Processing Unit
— **IoT** - Internet of Things
— **K-PLS** - Kernel Partial Least Squares
— **KNN** - K-Nearest Neighbors
— **KNN-LW-PLS** - K-Nearest Neighbors Locally Weighted Partial Least Squares
— **LCL** - Lower Control Limit
— **LDA** - Linear Discriminant Analysis
— **LLM** - Large Language Model
— **LV** - Latent Variable
— **MB** - Multiblock

— **MB-PLS** - Multiblock Partial Least Squares
— **MIR** - Mid-Infrared
— **MSPC** - Multivariate Statistical Process Control
— **NDSS** - Non-Destructive Spectroscopic Sensors
— **NIPALS** - Nonlinear Iterative Partial Least Squares
— **NIR** - Near-Infrared
— **NOC** - Normal Operate Condition
— **PCA** - Principal Component Analysis
— **PC** - Principal Component
— **PAT** - Process Analytical Technology
— **PLS-DA** - Partial Least Squares Discriminant Analysis
— **PLSR** - Partial Least Squares Regression
— **QP** - Quality Property
— **ReLU** - Rectified Linear Unit
— **RMedSECV** - Root Median Square Error in Cross-Validation
— **RMSECV** - Root Mean Square Error in Cross-Validation
— **RMSEP** - Root Mean Square Error in Prediction
— **ROSA** - Response-Oriented Sequential Alternation
— **SNV** - Standard Normal Variate
— **SO-PLS** - Sequential-Orthogonalised Partial Least Squares
— **SPC** - Statistical Process Control
— **SWT** - Stationary Wavelet Transform
— **UCL** - Upper Control Limit
— **UV** - Ultraviolet
— **VIP** - Variable Influence in Projection
— **VIS** - Visible Light
— **WT** - Wavelet Transform
— **WT-MIA** - Wavelet Transform Multivariate Image Analysis

List of Figures

List of Tables

Notation

In this work, the matrices are represented by bold capital letters ($\mathbf{A}$) and column vectors are represented by bold and lowercase characters($\mathbf{a}$) (the transposition operation is indicated by T in uppercase($\mathbf{a}^{\mathrm{T}}$)). The capital letters in italics define the constants (a), while for scalars and indexes the lowercase italic (a) character is used.

I

Overview of the Thesis

Chapter 1

Introduction

1.1 General Context

The industrial landscape has undergone significant transformations over the past three centuries, each marked by revolutionary advancements in technology and production methods. Known as Industrial Revolutions, these changes have reshaped manufacturing processes and have continually driven new approaches to quality control and process monitoring.

One of the critical aspects of this evolution has been the advancement of sensor technology through the so-called five eras [1]. In the off-line era, process monitoring relied on manual sampling and laboratory analysis, providing a limited understanding of real-time process variability. This evolved into the at-line era, where samples were analysed close to production line, reducing feedback times but still requiring manual intervention. The on-line era marked a significant step forward with automated systems capable of continuous sampling and data collection, improving response times to process changes. With the advent of the in-line era, sensors were placed directly within the process flow, allowing continuous, real-time data collection. Today, in the non-invasive era, advanced sensors, such as spectroscopic and optical tools, enable remote monitoring without physical contact, enhancing both the safety and efficacy of quality control [2]. Among the on/in-line sen-

sors, it has been widely demonstrated that NIR spectroscopy has a strong potential in monitoring production processes [3–6] due to its ability to detect both chemical and physical changes in samples.

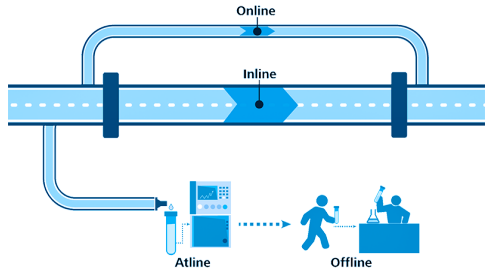


Figure 1.1 – Illustration of different ways of analysing and positioning sensors in industrial process monitoring. *Inline* analysis involves sensors integrated directly into the process flow, providing real-time measurements. *Online* analysis allows continuous monitoring with sensors positioned adjacent to the process flow, e.g. with a deviation. *At-line* analysis involves sampling nearby equipment for analysis with minimal delay. *Off-line* analysis requires samples to be sent to a laboratory for detailed examination, which typically involves longer timescales and more complex equipment. Source: https://www.inprocess-lsp.com/nanoflowsizer/system-configurations/.

A highly relevant approach in industrial development, introduced with the advent of Industry 3.0, is Process Analytical Technology (PAT) [7], a methodological framework that originated in the pharmaceutical industry and is now widely used in numerous other sectors [8] such as chemical, petrochemical and food. The Food and Drug Administration (FDA) has defined PAT as 'a system of production design, analysis, and control based on timely measurements (during processing) of critical quality and performance attributes of raw materials and intermediate products in order to ensure the quality of the finished product [7]. The adoption of this approach implied a shift from the traditional post-process quality assessment to quality as a results of a controlled process through the integration of in-line quality analysis of raw material, intermediate and finished products and control systems to ensure that products meet specifications during all the production phases, based on the principle that "quality cannot be verified after the fact, but must be designed in from the beginning" [7, 9]. Recent applications of PAT include continuous monitoring by advanced sensors, including spectroscopic sensors (e.g. near-infrared or Raman) and process sensors dedicated to the

detection of temperature, pressure and other fundamental parameters linked to machinery settings.

These elements represent the fundamental steps to fully embrace the Industry 4.0 era [10–12], which is driven by digital innovation and the integration of advanced technologies to create smarter, more adaptable and efficient systems. It introduced cyber-physical systems, the Internet of Things (IoT), artificial intelligence (AI) and big data analytics as key pillars for interconnected and autonomous production processes [13]. It challenges the traditional structure of the industrial automation pyramid [14], once defined by top-down or bottom-up information flows, by emphasising a more integrated and networked approach. Data now comes from all levels and are shared rapidly across the company ecosystem, enabling real-time decision making, rapid response to market changes. As a result, digital innovation and cutting-edge technologies are redefining the industrial landscape, pushing companies to move beyond rigid structures and adopt a more dynamic and adaptive model of managing industrial operations [15]. These developments are laying the foundations for next-generation manufacturing systems that will operate autonomously and self-regulate with minimal human intervention [12]. In this way, new operating structures that take advantage of the exponential increase in data generation will use big data techniques to analyse the data collected from IoT devices to optimise operations [16], predict when maintenance is needed and thus, improving product quality, providing a way for personnel to make faster, more accurate and objective decisions.

However, implementing this new paradigm in industrial practice is not straightforward and there are several aspects to consider. For example, in the context of a production line, understanding the Normal Operating Conditions (NOC) of a process can be challenging. As many interrelated parameters change simultaneously, plant operators often struggle to identify problems promptly when anomalies or deviations occur [17]. Latent variable based multivariate modelling (e.g. PCA, PLS) [18, 19] can greatly assist with this challenge, helping to extract the most meaningful information from process data and enabling on-the-fly detection of anomalies or deviations

[20–22].

Furthermore, when data from process sensors controlling machine settings (such as temperature, mixing rate, pressure, etc.) is fused with data from other plant stages or from sensors of a different type, such as NIR, it may be possible to achieve a better understanding of the process, which could help to design more efficient and environmentally friendly processes [23, 24]. While this "data fusion" approach is increasingly recognised as beneficial, it also poses new challenges in terms of handling, interpreting and modelling high-dimensional data in real time [25, 26]. The ability to effectively combine multiple data sources is a key step towards more efficient and environmentally friendly processes. In this sense, the development of multivariate control charts (based on latent variables) and real-time predictive models is beginning to be recognised as a significant advantage in the industry [27].

1.2 Objectives of the Thesis

This PhD project was funded by the Emilia-Romagna Region under the FSE+ 2021-2027 program, which aims to develop advanced skills to tackle the challenges of Industry 4.0. The research was conducted through a co-tutelle agreement between the University of Modena and Reggio Emilia and the University of Lille.

The thesis is set in the context of Industry 4.0, and among the several objectives of this field, the thesis focuses on the use of the available data to provide interpretation and decision support tools to optimise process control and product quality. In particular, latent variable-based multivariate statistical process control (MSPC-LV) was considered for data analysis, as these methods have contributed significantly to both process failure detection and diagnosis, and predictive modelling. However, the focus of this work is on predictive analytical approaches which, in line with PAT principles, are particularly suited to real-time prediction of product quality in industrial processes. However, when spectroscopic on-line or in-line data

and process sensor data need to be integrated and analysed simultaneously, different chemometrics tools may be needed and could be tailored to specific applications. In particular, when dealing with complex production involving several production phases data fusion and multiblock methodologies have proven to be efficient.

Therefore, the aim of this thesis was to develop multivariate chemometric models that handle high dimensionality, non-linearity and the presence of outliers in order to efficiently analyse and interpret data from heterogeneous sources (process sensors, spectroscopic data, operational parameters).

During the three-year research activity, carried out in collaboration with important industrial realities such as Barilla G. e R. Fratelli S.p.A and Versalis S.p.A, several objectives were pursued. On the one hand, work was carried out on the application and evaluation of chemometric methods in a context where the company (food sector) is still in the transition process toward adopting an Industry 4.0 approach. This included the critical analysis and selection of the most useful sensors for process monitoring, as well as the development of predictive modelling strategies for early estimation of product quality to ensure optimal production. As well as, enhanced information extraction from the existing sensor data, in this case in-line collected RGB images, in order to improve raw material monitoring. On the other hand, special attention has been paid to the development of new multiblock approaches capable of integrating data from different sources, overcoming the limitations of traditional models through techniques robust to the presence of non-linear phenomena and anomalous data.

The algorithms and methods presented in this thesis offer a contribution to solve issues arising in complex production scenarios where diverse sensors are active and continuously produce data potentially information rich but as well blurred by several variability sources. By providing improved predictive performance and addressing several key limitations associated with conventional multivariate regression methods (such as standard PLS), these approaches have the potential to significantly advance Industry 4.0 initiatives. In turn, these advances can promote more effective process control and support the drive towards greater sustainability in industrial operations.

1.3 Outline of the Thesis

The structure of this thesis is outlined below to guide the reader through the various aspects of the research. Section one built on two chapters provides an overview of the state of the art and illustrates the context of the study by describing the data structure and the main chemometric methods used for on-line prediction and process monitoring. In particular, **chapter 2** traces the evolution of data use in industrial contexts, detailing the shift from univariate approaches, focused on analysing one variable at a time, to multivariate techniques that consider the entire data structure simultaneously.

The second section, consisting of two chapters, focuses on the study and evaluation of methodologies aimed at supporting the transition to an Industry 4.0 approach in a food production company.

- **Chapter 3** explores how latent variables approaches can enhance the study of an industrial food production process, specifically *pesto alla genovese* at the Barilla G. e R. Fratelli S.p.A. production plant. This chapter highlights both the challenges and the opportunities offered by these methods in the transition to Industry 4.0.
- **Chapter 4** focus to a single sensor, an RGB camera from the same production plant. This chapter aims to improve the information extracted from sensors by comparing two different image analysis approaches: one rooted in chemometrics and the other based on neural networks.

The third section, consisting of 3 chapters, deals with multiblock models, addressing possible non-linearities and methods to reduce the impact of outliers.

- **Chapter 5** explores the benefits of multiblock approaches by integrating data from multiple sensors in an ABS production plant at Versalis S.p.A. It introduces an extension of the LW-PLS model, called LW-MB-PLS, designed to handle multiple blocks of data simultaneously in a non-linear context.

- **Chapter 6** presents an extension of the ROSA method, called LW-ROSA, specifically designed to deal with non-linearities also in a multi-block context, which, unlike the methods in Chapter 5, is based on a selection of the most influential blocks to achieve the prediction.
- **Chapter 7** introduces a novel approach, LW-RoBoost-PLS, which simultaneously handles non-linearities and the presence of outliers in the calibration data.

Finally, the thesis concludes with a summary of the contributions, an analysis of the limitations, and a discussion of potential future developments.

Chapter 2

Industry 4.0 and Data: From Univariate Analysis to Chemometrics

This chapter presents the basic concepts necessary to understand the contents of this thesis. We will begin with a brief overview of what is meant by an industrial plant, describing its main characteristics and the different types that exist. Since the thesis focuses on data analysis, we will then move on to analysing the data that an industrial process produces, illustrating the different typologies.

Next, we will get to the core of the matter by showing how this data is analysed to support operational and strategic decisions. We will first examine the simplest and most established univariate approaches and then introduce the main multivariate analysis methods. In this way, the chapter will provide the necessary background for understanding the concepts introduced and discussed in the following sections of the thesis.

2.1 Industrial process

Industrial processes comprise one or more stages through which a product (raw material) undergoes a gradual transformation until it reaches the desired final form. These processes are classified as either continuous or batch processes.

Continuous process

In a continuous process, raw materials are introduced at the start and flow continuously through conveyors, pipelines and other infrastructure, through each stage of production until the final product is obtained. Transformations therefore occur in space rather than time [28]. This type of production is characterised by a continuous change in the state of the process. As a result, the sensors used to control the process must continuously monitor these changes over time.

Common causes of deviations from NOC include changes in catalyst activity, variations in raw material quality, shifts in product demand and similar factors. As both the feed and the processing of the raw materials remain constant, there is a continuous flow of intermediate and final products. This continuous operation means that during changes in formulation or production, there may be partial overlap of different products, as well as longer start-up times and a higher likelihood of non-conformity in the initial and final stages.

A critical challenge in continuous processes is to synchronise the sensors data to maintain a clear view of product quality. As the product moves through its various transformation stages, it is measured by different sensors at different times. This introduces variable time delays that have a significant impact on process modelling. In continuous industrial processes, where data from spatially separated sensors must be integrated, these time delays must be taken into account to ensure the accuracy of the resulting models [29].

Batch Processes

In a batch process, production has defined start and end points [30]. Once a batch has been produced, the process stops and a new batch cannot start until the previous one has been completed. In this type of operation, the production line waits for an entire batch of raw material to arrive before processing it through the various manufacturing stages. As a result, production takes place at predetermined intervals, with raw materials arriving at the beginning and the finished product becoming available after a period of time. This is in contrast to continuous processes, where parts are processed continuously as they pass through the system.

Batch production also faces synchronisation issues, but these can vary depending on the setup. For example, in batch processes, all sensors may be located in the same place, or sensors may be positioned at different heights in a reactor, creating a three-way data structure: batches $\times$ sensors $\times$ time. Consequently, both synchronisation and alignment may be required, not only between batches, but also between sensors. Furthermore, since each batch may require a different amount of time to complete, batches may not overlap perfectly, making it difficult to align measurements and make accurate comparisons [31].

As this work focuses on continuous processes, batch processes will not be discussed further. However, the points outlined here highlight some of the unique challenges associated with batch production systems.

The "choice" between continuous and batch processes depends on the company's objectives and available resources. Furthermore, processes are not always strictly batch or continuous; hybrid situations can occur. For example, the introduction of raw materials into a mixer may resemble a batch phase even within an continuous operation.

2.2 Industrial Data

In industrial processes, it is essential to know if the process is operating at normal conditions, i.e. if the settings of the machineries that control the process are within the specifications and fluctuations with respect to the setted values are random and small so that the final product can meet the expected quality. To this aim several sensors are installed to both monitor and ensure the proper setting so that action can be taken if the process variability depart from the normal conditions. The sensors can be broadly divided into two main types: classical process sensors (which monitor/control the machineries settings) and spectroscopic sensors (which monitor the intermediate products). Both play a critical role in understanding and improving process performance, but are of different nature and furnish complementary information. This section explores the nature of the data collected by the two sensors categories, highlighting their characteristics, typical applications, and the challenges associated with their use.

2.2.1 Classical Process Data

Although there is continuous evolution in the field of process control, with substantial improvements in equipment, the techniques for measuring quantities such as pressure, temperature, flow, pH and levels have remained of fundamental importance to characterise the process. According to the UNICHIM standards [32], the sensors installed in the system are divided into different categories. For example, a sensor with the abbreviation TIC is a device that measures temperature (T), provides an indication (I) and is equipped with a control function (C). The first letter indicates the physical quantity measured (e.g. T for temperature, P for pressure, etc.), while the second and third letters define the type of sensor and its function, which can be alarm (A), control (C), indication (I) or recording (R). Control (C) means that the sensor, in addition to measuring the quantity, provides data to a control system capable of intervening in the process (e.g. modifying the flow rate of a fluid or the action of a heater) to keep the variable within the

desired value (set-point). The difference between an indicator and a recorder is that the recorder is equipped with a system for continuous data recording. However, this data often lacks the detailed compositional information required for quality control, and this is where spectroscopic data come in.

2.2.2 Spectroscopic Data

Non-destructive spectroscopic sensors (NDSS) [33], such as NIR spectroscopy, fluorescence, Raman or hyperspectral imaging, are becoming increasingly common, enabling rapid, non-destructive assessment of multiple parameters in a wide range of products [34–39]. Spectroscopy exploits how matter interacts with electromagnetic radiation. By analysing absorption, emission or scattering spectra, detailed information about the molecular structure, chemical composition and physical properties of samples can be obtained. The electromagnetic spectrum covers a wide range of wavelengths, see Figure 2.1. Consequently, different branches of spectroscopy focus on specific spectral regions.

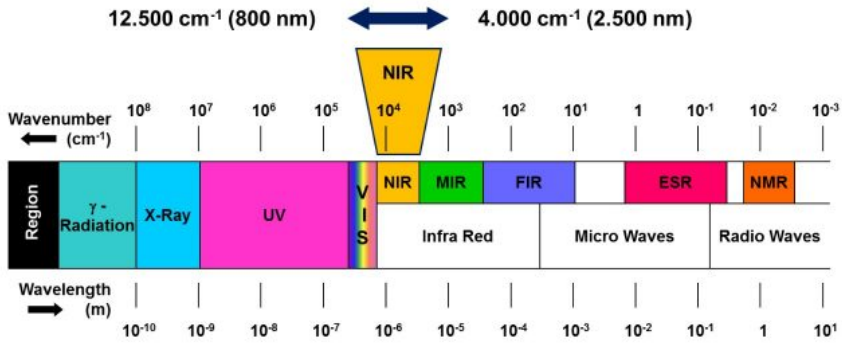


Figure 2.1 – The figure illustrates the various regions of the electromagnetic spectrum, including gamma radiation, X-rays, ultraviolet (UV), visible light (VIS), infrared (NIR, MIR, FIR), microwaves, and radio waves, highlighting the Near-Infrared (NIR) region ranging from 12,500 cm^{-1} (800 nm) to 4,000 cm^{-1} (2,500 nm). Source: Bruker [40]

NIR spectroscopy [41, 42] focuses on the spectral region between 780 and 2500 nm, which lies between the visible and mid-infrared, in which the absorption bands lies manly to overtones and combinations of fundamental

vibrations. It is part of the broader field of infrared spectroscopy, which is divided into three regions: the near infrared (NIR), the mid infrared (MIR) and the far infrared (FIR). These regions are particularly valuable for studying molecular vibrations, which provide key insights into chemical bonding and the internal structure of molecules.

Figure 2.1 highlights the NIR spectral region, one of the most impacting in the field of in-line sensors [41, 43], mainly for the following advantages of NIR spectroscopy:

- **Non-destructive analysis:** Samples, whether solid, liquid, or otherwise, can often be analysed directly without invasive chemical preparation.
- **Rapid measurements:** Spectral acquisitions can be completed in seconds, enabling real-time, in-line or at-line analysis.
- **Integration into production processes:** The instruments can be adapted to operate outside the tightly controlled conditions of the laboratory. The probes can be inserted directly into the process line or connected to a flow cell to analyse the sample directly from the production line, using either transmission or transflectance measurements.

Mention to vibrational spectroscopy

In first approximations, it is possible to describe the vibrations of chemicals bonds with the theory of a harmonic oscillator model. However, the energy that can be absorbed or emitted by a molecule is discrete. This means that energy is transferred in discrete packets called quanta, and that the energy levels derived from the molecule's potential are discrete levels defined by the quantum number v, which can take on integer values from zero and up.

$$E_{vib} = \left(v + \frac{1}{2}\right) \frac{h}{2\pi} \sqrt{\frac{k}{\mu}} \tag{2.1}$$

Equation 2.1 describe the energy vibration levels for a bi-atomic molecule, where v is the vibrational quantum number, h the Plank constant, k, the force constant and μ the reduced mass of the bonding atoms. According to potential harmonic model, only transitions between consecutive energy levels ($\Delta v = \pm 1$) that cause a change in dipole moment are possible, however, this model is not able to describe actually a molecule because it does not consider Coulombic repulsion between atoms or dissociation of bonds. Consequently, the vibrational behaviours of molecules is better described from an anharmonic oscillator model. The anharmonicity can result in transitions between vibrational energy states where $\Delta v = \pm 2$, $\Delta v = \pm 3 \ldots$ These transitions between non-successive vibrational states yield absorption bands known as *overtones* (first and second overtone, respectively) at, approximately, multiples of the fundamental vibrational frequency (from $v = 0$ to $v = 1$). The frequencies of many overtone bands appears in the NIR region of 780 nm and 2000 nm. Another type of absorption band occurs in the NIR region are the *Combination bands*. In polyatomic molecules, two or more vibrational modes can interact in such a way as to cause simultaneous energy changes and give rise to absorption bands the frequencies of which are the sums of multiples of each interacting frequency. NIR combination bands appear between 1900 nm and 2500 nm.

The intensity of the NIR band depends on the change in dipole moment and the anharmonicity of the bond. The hydrogen atom is the lightest, and therefore exhibits the largest vibrations and the greatest deviations from harmonic behaviour. Hence, the main bands typically observed in the NIR region correspond to bonds containing hydrogen and other and other light atoms (C–H, N–H, O–H and S–H). On the contrary, the bands for bonds such as C=O, C–C and C–Cl are much weaker or even absent.

The NIR spectroscopy technique provides information-rich, high-dimensional datasets that capture detailed information about the chemical and physical properties of a sample. These data can be collected in real time (since the NIR technique is nondestructive and implementable online), and can be directly related to the chemical properties of the product, allowing real-time quality control and process optimisation [44–46]. This approach has proven

to be advantageous across various industrial sectors. For example, in the food industry, NIR can be used to monitor parameters such as moisture, protein and fat content in cereals [47, 48], meat [49] and dairy products [50, 51]. In the pharmaceutical industry, it supports the assessment of active pharmaceutical ingredient concentrations and uniformity in tablets [52], reducing the need for time-consuming offline analyses. The petrochemical sector uses NIR to determine octane numbers and monitor propellant blending operations [6, 52, 53], while the polymer and plastic industries use it to evaluate the composition and detect impurities during processing [54, 55]. In general, these applications enable for improved process optimisation, waste reduction, and improved product consistency.

2.3 From Univariate to Multivariate Data Analysis

Traditionally, industrial process analysis has relied on univariate statistical process control (SPC) tools to monitor processes, detect anomalous behaviours, and support decision making. SPC aims to monitor quality or process variables over time, ensuring they remain close to their desired values within normal sources of variation. This approach relies on monitoring one variable at a time, making it impractical to control all variables simultaneously. Consequently, product quality is typically monitored using SPC charts such as Shewhart [56], CUSUM [57, 58], and EWMA[59] charts applied to only a few selected product variables $\mathbf{Y}$ measured via off-line laboratory analysis.

SPC charts focus solely on the magnitude of deviation in each variable. The Shewhart chart, one of the most widely used SPC techniques, evaluates whether the measured value of a variable falls within limits based on the natural variability of the process, defined by the Upper Control Limit (UCL) and Lower Control Limit (LCL). The UCL and LCL are typically set at ± 3 standard deviations from the target value. Observations that exceed these limits are considered "out of control".

Although effective in certain contexts, SPC methods have several draw-

backs. As mentioned above, by considering only a few variables and not taking into account their correlation, this approach doesn't fully describe product quality, which often depends on the simultaneous behaviour of several quality characteristics [20].

Figure 2.2 illustrates this limitation with an example involving two product quality variables, $\mathbf{y}_1$ and $\mathbf{y}_2$. Looking at the single Shewhart chart, the process appears to be within statistical control limits. The state where the variability is only caused by natural variation and production under NOC. Indeed, for $\mathbf{y}_1$ and $\mathbf{y}_2$ all the measurements are within the limits. However, moving from univariate to multivariate data analysis, it is clear that the final conclusion is different. In fact, the scatter plot $\mathbf{y}_1$ vs $\mathbf{y}_2$ shows that the measured point marked with a red circle (●) deviates from the expected correlation. Thus, it does not belong to the NOC population. Therefore, the use of multivariate approach, allows the detection of subtle trends and anomalies that may otherwise go unnoticed, leading to improved process understanding and more robust quality assurance.

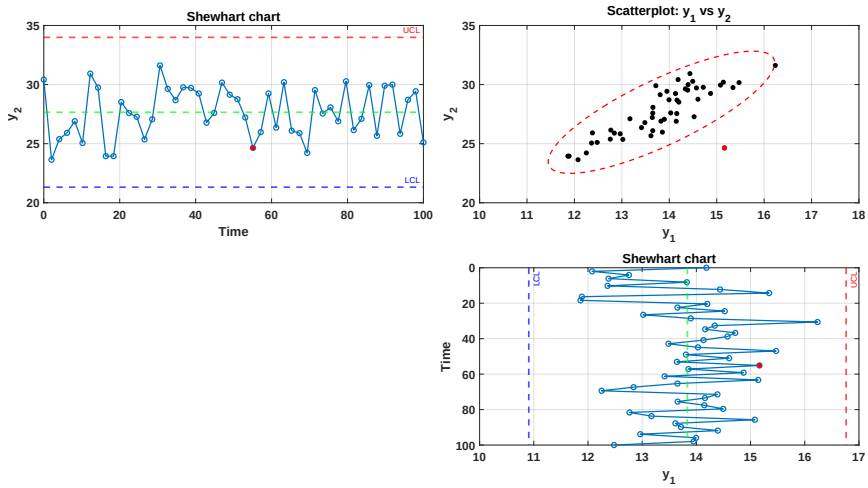


Figure 2.2 – Analysis of process data using Shewhart control diagrams and a bivariate scatter plot. The top left and bottom right panels show a Shewhart control diagram for $\mathbf{y}_2$ and $\mathbf{y}_1$ over time, respectively, with dashed lines representing the upper and lower control limits (UCL and LCL). The top right panel presents a bivariate scatter plot of $\mathbf{y}_1$ and $\mathbf{y}_2$, showing the correlation between the two variables and highlighting a point that appears to be within control limits in univariate charts but is outside the control region in the bivariate context (red ellipse). Adapted from [60]

Furthermore, SPC methods typically analyse product quality properties ($\mathbf{Y}$) separately from process variables ($\mathbf{X}$), which are continuously collected by the sensor system (temperature, flow, pressure), without taking into account that neither the process variables nor the product quality variables are independent [61]. Also, monitoring the process variables $\mathbf{X}$ would help to detect problems during production that could lead to an undesirable product. Moreover, even if $\mathbf{Y}$ measurements are frequently available, monitoring the process variables ($\mathbf{X}$) can help to diagnose the assignable causes of an undesirable event. When monitoring product quality, even if we can determine which quality variable is out of limits, it can be difficult to determine what went wrong in the process [61].

These limitations have paved the way for multivariate statistical process control (MSPC). Continuing the focus on the Shewhart chart, the literature has proposed direct extensions such as the *Hotelling's* T^2 and χ^2 charts [62–64].

Focusing on the quality control, at this time all the quality parameters ($\mathbf{Y}$) are considered simultaneously. Given a vector $\mathbf{y}_k$ of all measurements at time k normally distributed, with a in-control covariance matrix $\boldsymbol{\Sigma}$, is possible to detect how much the variables deviates from the population mean $\boldsymbol{\mu}$, computing the statistic:

$$\chi_k^2 = \left(\mathbf{y}_k - \boldsymbol{\mu}\right) \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_k - \boldsymbol{\mu}\right) \tag{2.2}$$

Given that the mean is $\boldsymbol{\mu}$, this statistic follows a central χ^2 distribution with n degrees of freedom. By plotting χ_k^2 versus time, one obtains the multivariate chart. To detect if there are anomalies, an upper control limit (UCL) is defined given a $\chi_{\alpha,n}^2$ where α is an appropriate significance level (for example $\alpha = 0.05$) to perform the test. However, the χ^2 statistic described in Equation 2.2 represents the Mahalanobis distance [65] of any points from the target $\boldsymbol{\mu}$.

Thus, for a two-variable case, as illustrated in Figure 2.2, a scatter plot can be used to visualize how products distribute with respect to the two

quality variables. In this representation, the control limits are depicted as an ellipse cantered on $\boldsymbol{\mu}$.

On the other hand, to monitor the temporal evolution of the process, particularly when dealing with more than two variables, due to limitations in the graphical representations, the classical multivariate control chart is used. In this case, the χ^2 statistic is plotted over time (time on the x-axis), allowing the detection of shifts or trends in the process performance relative to the control limits.

In case the covariance matrix $\boldsymbol{\Sigma}$ is unknown, it is necessary to estimate it from the N historical data. The use of *Hotelling's* T^2 statistics, in this case [66], will be used to construct the control chart:

$$T_k^2 = \left(\mathbf{y}_k - \bar{\mathbf{y}}\right) \mathbf{S}^{-1} \left(\mathbf{y}_k - \bar{\mathbf{y}}\right) \tag{2.3}$$

where $\bar{\mathbf{y}}$ is the mean estimated from the data and $\mathbf{S}$ is the estimation of $\boldsymbol{\Sigma}$

$$\mathbf{S} = (N-1)^{-1} \sum_{k=1}^{N} (\boldsymbol{y}_k - \bar{\boldsymbol{y}})(\boldsymbol{y}_k - \bar{\boldsymbol{y}})^T \tag{2.4}$$

to obtain the control chart the T^2 values can be plotted against the time as well as for χ_k^2 charts. For these charts, the upper control limit T_{UCL}^2, for the new samples, is estimated by Equation 2.5, based on F-distribution at α confidence level and with n and $n-q$ degrees of freedom [67].

$$T_{UCL}^2 = \frac{(n^2-1)q}{n(n-q)} F_\alpha(n, n-q) \tag{2.5}$$

Once an out-of-control point is detected, the main challenge lies in pinpointing which variables are responsible. This aspect is discussed in Section 2.4.3.

As in the univariate case, there are other types of multivariate charts,

such as Multivariate CUSUM and Multivariate EWMA [68, 69].

A limitation of this type of multivariate control chart is the need to invert $\boldsymbol{\Sigma}$ or $\boldsymbol{S}$. Indeed, as previously noted, industrial data are often highly correlated, which complicates the calculation of inverse matrix. To overcome this issue, latent variable-based methods have been introduced. By projecting the data onto an orthogonal subspace, these methods effectively bypass the inversion of large covariance matrices and thus avoid the instability inherent in such procedures.

2.4 Latent Variable Methods for Real-Time Process Monitoring and Quality Control

Latent variable methods, based on Principal Component Analysis (PCA) [70–72] and/or Partial Least Squares Regression (PLS) [73–75], have emerged as powerful alternatives for process monitoring and quality control, particularly when dealing with high-dimensional data holding highly correlated variables as in industrial datasets. These methods use dimensionality reduction techniques to compress the original variable space into a lower-dimensional subspace while retaining the critical information needed for monitoring or quality control. Latent variable-based methods fall into the category of process history-based approaches [76], as they use historical databases (data collected from previous years of production) to derive empirical models and establish acceptable operating limits.

Overall, this methodology consists in applying several steps [77]. Firstly, Exploratory Data Analysis [78] is applied, for instance by using PCA, to gather a first data inspection. The aims of this step is to get an overview of the data, identify if there is presence of drift or cluster structure, if there are any issues such as the presence of outliers. The outlier detection is indeed a very critical phase, because they have a deleterious influence in the model and consequently in the interpretation of the process data. Once the data structure has been explored, it is essential to identify the data representing the normal operating conditions, typically the stationary states

of the process. This data is fundamental to model building, as it is used to train the models. However, to ensure the reliability of the predictions and insights provided by the model, validation is required. This means that not all NOC data can be in the training set, but it is necessary to split the data into training and test sets [79, 80].

Following the splitting stage, a data preprocessing step is usually performed to remove any irrelevant information [81], such as noise, offsets or unwanted effects/defects, and to correct for instrumental drifts. Once the data are properly prepared, the most appropriate model can be applied for the specific analysis purpose.

When there is the need to optimise some model parameters, a third validation set would be required, but in most cases the available data might not be sufficient. In such cases, it is possible to overcome the need for a third set by using internal validation techniques, such as cross-validation [82].

Latent variable control charts can be constructed to monitor response variables ($\mathbf{Y}$) or predictor variables ($\mathbf{X}$), such as process sensor variables [76]. For this purpose, the most commonly used methods are PCA, which considers $\mathbf{X}$ and $\mathbf{Y}$ independently, or PLS, which analyses both groups ($\mathbf{X}$ and $\mathbf{Y}$) together. The latter approach allows for real-time predictions, enabling the estimation of product quality in real-time based on $\mathbf{X}$ data.

2.4.1 Principal Component Analysis

Principal component analysis (PCA) [70, 83] is probably the most known multivariate statistical tool for exploratory data analysis to compress and interpret large data sets. Taking into account a matrix $\mathbf{X}$ with N row (measurements) and V columns, in our case they represent the sensors data and as aforementioned are correlated, it is possible to condense the entire information in a few dimensions identifying a low dimensional subspace that best describe the original data. Mathematically is equivalent to operate a projection of the data onto a latent variables subspace along which the variance is maximized. This new set of A latent variables are called Principal Com-

ponents (PCs). The PCs are derived by a linear combination of the original variables where the first component has to explain the largest possible variance. Each following component under the constrain to be orthogonal to the precedence components, has to capture the maximum possible residual variance. The PCA model can be represented by the decomposition Equation 2.6:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.6}$$

The $\mathbf{T}$ matrix ($N \times A$), called the scores matrix, represents the sample coordinates in the new low-dimensional space, and allow us to understand the structure of the data, it is composed by N rows and A number of columns equal to the number of PCs. The loadings matrix $\mathbf{P}$ ($V \times A$) is composed by a number of columns equal to the number of variables (V) and rows equal to the number of PCs (A). The loadings values corresponded to the weights by which each original variable entered the linear combination, thus defining the PCs, representing the contribution of each variable to each PC. The analysis of loadings matrix allow us to understand the correlation structure of the variables. The residual matrix $\mathbf{E}$ ($N \times V$), which represented the unmodeled information, has the same dimension of $\mathbf{X}$, and it is obtained by the subtraction of recalculated data from the PCA model $\widehat{\mathbf{X}} = \mathbf{T}\mathbf{P}^T$ from the original data $\mathbf{X}$.

2.4.2 Partial Least Squares

Partial Least Squares Regression (PLS) [73, 84] is a linear regression method widely used in multivariate calibration to model linear relationships between independent variables (predictors or regressors, $\mathbf{X}$ block) and dependent variables (responses, $\mathbf{Y}$ block). PLS excels at handling high-dimensional data sets with a high degree of collinearity. This is achieved by operating in a low-dimensional space defined by latent variables (LVs) obtained by simultaneously decomposing $\mathbf{X}$ and $\mathbf{Y}$. These LVs are oriented in directions that maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$. While orthogo-

nality in the $\mathbf{X}$ scores space is always ensured, orthogonality in the $\mathbf{Y}$ scores space depends on the specific PLS implementation. Mathematically, PLS model can be summarized by the outer relations, a PCA-like decomposition of the predictors and responses matrix:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.7}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2.8}$$

where the outer relations are linked by the inner relation.

$$\mathbf{U} = b\mathbf{T} \tag{2.9}$$

Re-expressing this as a regression model:

$$\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{B} \tag{2.10}$$

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T = \mathbf{R}\mathbf{Q}^{-1} \tag{2.11}$$

Let us define the matrix $\mathbf{X}$ with dimensions $N \times V$ and the matrix $\mathbf{Y}$ with dimensions $N \times M$. From the PLS model of A latent variables, we obtain the following matrices: $\mathbf{T}$ $(N \times A)$, $\mathbf{P}$ $(V \times A)$, and $\mathbf{E}$ $(N \times V)$, which represent the $\mathbf{X}$-scores, $\mathbf{X}$-loadings, and $\mathbf{X}$-residuals, respectively. Similarly, for the matrix $\mathbf{Y}$ with dimensions $N \times M$, the PLS model provides $\mathbf{U}$ $(N \times A)$, $\mathbf{Q}$ $(M \times A)$, and $\mathbf{F}$ $(N \times M)$, corresponding to the $\mathbf{Y}$-scores, $\mathbf{Y}$-loadings, and $\mathbf{Y}$-residuals, respectively. Additionally, $\mathbf{B}$ is the regression coefficients matrix, which directly enables the prediction of $\mathbf{Y}$ from $\mathbf{X}$. Finally, $\mathbf{W}$ $(V \times A)$ represents the weights matrix used in the decomposition of $\mathbf{X}$.

A key advantage of PLS lies in its ability to provide not only a predictive model linking independent and dependent variables, allowing real-time prediction of the quality parameter $\mathbf{Y}$, but also a comprehensive representation of both datasets in terms of latent factors. This dual capability facilitates the interpretation of results, enabling the extraction of meaningful insights from complex data structures.

In this thesis, PLS model is used to build multivariate calibration models for the prediction of quality parameters to estimate in real time the final product quality in different processes.

2.4.3 Latent Variables Control Charts

The use of projection methods has revolutionised the approach of statistical process monitoring. The operator to detect if the process work in NOC instead of checking several control charts in this case can only check two multivariate controls charts [20].

Once a dimensional reduction, e.g. PCA, has been performed, two distances are defined: one in the scores space (scores distance) and one in the residuals space (sum of squared prediction errors, SPE). Using the respective statistics Hotelling's T_A^2 and the χ^2 distribution, control charts can be obtained. The scores distance for each training sample is defined by the following equation:

$$\mathbf{t}_i^2(A) = \mathbf{t}_i^\top \Theta^{-1} \mathbf{t}_i = \sum_{a=1}^{A} \frac{t_i^2}{\lambda_a}. \tag{2.12}$$

Here, the equation defines the distance, while the corresponding statistic T_A^2 is used for the control chart. In this equation, Θ is the covariance matrix of the scores matrix $\mathbf{T}$ (which only has diagonal terms holding the scores variance for each component), obtained during the training phase by building a PCA model on the NOC data. The term $\mathbf{t}_i$ represents the scores vector for the i-th observation. For each new observation (test) $\mathbf{x}$, it is possible to calculate the $\mathbf{t}_{\text{new}}^2(A)$ values by projecting the new data onto the NOC PCA space using $\mathbf{t}_{\text{new}} = \mathbf{xP}$, where $\mathbf{P}$ is the matrix of loadings. Also in this case, to evaluate if there are deviations from NOC is necessary to define an UCL, and the statistical confidence limits for T^2 can be calculated by assuming the F-distribution according the following equation:

$$T^2_{A,UCL} = \frac{(N^2-1)A}{N(N-A)} F_{\alpha(A,N-A)} \tag{2.13}$$

where A is the number of latent variables, N is the number of observations used to perform a PCA model, and $F_{\alpha(A,N-A)}$ is the upper $100\alpha\%$ critical point of the F-distribution with $(A, N-A)$ degrees of freedom. This chart detects if a new observation projected in the hyperplane defined by PCs is within the limits determined by the reference data.

Moreover, multivariate charts based on latent variables introduce an additional control tool: the Q (or SPE) chart. In this chart, each point represents the sum of squares of each row of $\mathbf{E}$ (equation 2.14). This measures how well each sample conforms to the latent variable model. Specifically, it accounts for the amount of information not explained by the model. Geometrically, it represents the Euclidean distance of the observation from the hyperplane formed by the A principal components (PCs).

$$Q_i = \mathbf{e}_i \mathbf{e}_i^T \tag{2.14}$$

The **Q**-limits define the distance from the hyperplane in PCA that is considered unusual for a system operating under NOC [85]. Several methods exist to establish the upper control limits (UCL) for the weighted χ^2 statistic, which is assumed for Q. For example, one of the most common approaches was suggested by Jackson and Mudholkar (1979) [86], who developed a methodology based on the moments of the residual eigenvalues. Jackson's framework leverages the first three moments of these eigenvalues (θ_1, θ_2, θ_3) to compute the UCL as:

$$UCL = \theta_1 \left[z_\alpha \frac{\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{2.15}$$

Here, θ_1, θ_2, and θ_3 are the first three moments of the residual eigenvalues, calculated using the eigenvalues λ_j of the PCA residual covariance

matrix $\theta_k = \sum_{j=A+1}^{\text{rank}(\mathbf{X})} (\lambda_j)^k$. The parameter h_0 is defined as $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$. The value z_α represents the critical value of the standard normal distribution corresponding to the desired significance level α.

In addition to Jackson's methodology, other techniques have been proposed to define the Q-statistic's control limits. For example, Box's method [87, 88] uses the weighted χ^2 distribution to establish thresholds, while the DModX metric [89] provides a complementary approach by quantifying the sample-wise residual distance normalized to the model's critical limits.

The two multivariate control charts act as complementary indices, providing a comprehensive view of the overall health and performance of the process.

This Control Charts (T_A^2 and Q) can be applied also to PLS model using the scores and residual provided from the equation 2.8. In addition, for real-time quality control, the estimated quality properties can be plotted over time to monitor their quality behaviour. This approach allows the detection of potential deviations from control limits based on model predictions, avoiding the delays associated with waiting for laboratory analysis results.

Contributions Plots

Monitoring diagrams may detect deviations from normal operating condition data, indicating that something is wrong with the process, but they do not provide information on the cause of the problem by indicating which variable v or subset of variables is responsible. To solve this problem, contribution diagrams can be used to identify the variables that are driving the process out of control. [66, 90–92].

The contribution of the v-th variable to the latent variables model,denoted as c_v^T, can be expressed as:

$$c_v^T = \mathbf{t}_{\text{new}} \cdot \Lambda^{-1} \cdot \mathbf{p}_v^T \cdot \mathbf{x}_v^{\text{new}} \tag{2.16}$$

where $\mathbf{p}_v$ is the vector corresponding to the v-th row of the loading matrix associated with the first A selected principal components. Λ is a diagonal matrix containing the eigenvalues of the first A components.

For residual analysis, the contribution of the v-th variable to the Q-statistic (c_v^Q) is calculated as:

$$c_v^Q = (x_v^{\text{new}} - \mathbf{p}_v \cdot \mathbf{t}_{\text{new}})^2 \tag{2.17}$$

Here, $\mathbf{x}_v^{\text{new}}$ represents the observed value of the v-th variable, and $\mathbf{t}_{\text{new}}$ is the score vector of the new observation. This calculation helps identify how much each variable contributes to the residual variance not explained by the PCA model.

Contribution plots, typically displayed as bar charts with the x-axis representing variables and the y-axis showing their contributions (c_v^T or c_v^Q), highlight variables with the largest contributions as likely sources of deviations or process anomalies.

II

The food industry towards Industry 4.0

Chapter 3

Moving towards on-line quality assessment in a food plant

This work was carried out in collaboration with Dr. Alessandro D'Alessandro, who completed his PhD under the supervision of Professor Marina Cocchi. In particular, I focused on solving the data synchronisation issue, find an automated way for data cleaning, tailoring the chemometrics pipeline, especially suitable preprocessing.

This chapter is based on the article listed below, with selected content and results.

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

3.1 Introduction

The food industry has been slower to move towards Industry 4.0 than other sectors, due to its specific characteristics. While chemical and Pharma companies have made significant progress in adopting digital technologies, food companies remain very diverse. Differences in size, organisational structure and available resources highlight the challenges of implementing digital solutions across the sector. In addition, the inherent variability of food products, including differences in quality, shape, origin and perishability, adds another layer of complexity to their handling and processing. Historically, these challenges have been addressed through the knowledge and skills of experienced employees and master manufacturers. However, as the scale and complexity of production increases, traditional approaches are becoming less adequate and more targeted research and innovation is urgently needed to support the digital transformation of this industry and ensure food quality that meets consumer expectations, perceptions and acceptance [93].

The present work concerns a feasibility study to set up a model for on-line monitoring of the pesto production process in the Barilla G. e R. Fratelli S.p.A. company, where currently a vision system (RGB camera) monitors the main raw material, i.e. basil, and an in-line NIR probe monitors the first semi-finished product. The main objective of this preliminary feasibility study is to evaluate the possible advantages that MSPC-LVs based on in-line acquired data can provide, both in terms of the possibility of estimating the quality of the final product in real time, and in terms of capturing the evolution of the process and the possible deviation from the normal operating conditions (NOC). In this context, PCA models have been used to study the structure of the data and the information it provides. Multivariate control charts for process monitoring based on NOC data were also constructed. Finally, a first attempt was made to obtain predictive models for real-time prediction of the main pesto quality parameters. Emphasis was also placed on discussing the steps that were more critical for the development of the models. Although the results are very preliminary, some interesting indications and directions for improvement could be formulated.

3.2 Materials and Methods

3.2.1 Process Description

The data analysed in this study were collected from the pesto sauce production line during the 2020 harvest season at a facility owned by *Barilla G. e R. Fratelli S.p.A.*, situated near Parma, Italy. During this period, two distinct varieties of basil (*Ocimum basilicum*), the primary ingredient in the sauce, were supplied by five local providers and delivered continuously to the production line. Each basil variety underwent four harvests at different stages of maturity: the initial harvest took place at 40 days of growth, followed by subsequent harvests at intervals of 20 days.

At the beginning of the process line, a vision system (RGB camera) is installed to capture images of basil plants as they pass along the conveyor belt. The system is configured to provide real-time parameters such as the average and standard deviation values (every 15 seconds) of the R, G, and B channels, as well as a rough estimation of the basil leaves' area in the captured images. However, this estimation is not always available at consistent time intervals, and the raw images are not consistently stored. Therefore, only the R-G-B parameters were considered in this study.

Following this step, the basil is mixed with salt and oil to form an intermediate product, which is monitored online using a NIR probe. Subsequently, the remaining ingredients of the sauce are added to the intermediate product, completing production and yielding the final product. The quality of this final product is evaluated through off-line laboratory analyses. A schematic representation of the process is shown in Figure 3.1.

A critical challenge in modelling on-line data for a continuous process is establishing a timeline to align sensor data collected at different stages with the same material. In other words, to construct a row in the data matrix, the variables must correspond to the same sample. This alignment proved particularly complex in this case because the intermediate product is mixed with other ingredients in three distinct mixers. These mixers are

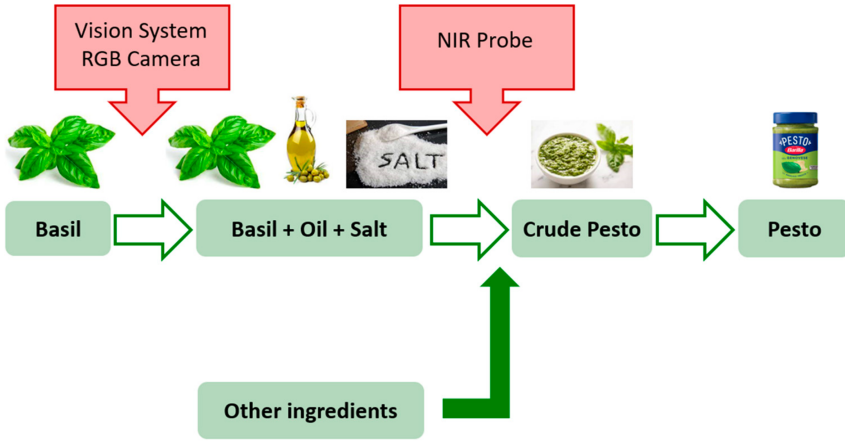


Figure 3.1 – Schematic representation of Pesto sauce production process.

emptied sequentially, transferring the crude pesto to the subsequent processing steps while maintaining a continuous material flow. To address this, the residence time was determined in consultation with plant experts, ensuring that the NIR spectra, which correspond to the intermediate material, could be matched with the final pesto product at the end of the line, where quality parameters are measured.

Data collected between May and August 2020 were analysed for this study. However, not all data recorded during this period were used for model development due to production interruptions, instrument maintenance, and unreliable measurements. Ultimately, 459 data points were selected for analysis.

Interruptions also posed another critical challenge when constructing the data matrix, as stoppages were relatively frequent. During conveyor belt stoppages at the raw material stage, the on-line RGB system continued capturing images of the same basil. To address this, RGB parameter time trends were inspected, with constant values identified as indicators of stopping periods. Additionally, the activation of the pump transferring the intermediate product to the NIR probe was logged and used as a marker for stoppages.

Finally, the data were further cleaned by removing anomalies identified in the RGB parameter trends and NIR spectra.

3.2.2 Reference Analysis

The final quality of the pesto sauce is assessed through off-line analysis of samples collected right after production is completed. For simplicity and brevity, only two parameters, consistency and lipid content, are considered in detail in this study, although several other parameters, such as water activity, pH, and dry residue, are also used to evaluate product quality.

The consistency of pesto is evaluated by measuring the flow of a standard sample volume (100 cm^3) under its own weight. The flow is related to the sample's viscosity. The measurement is performed using a Bostwick consistometer (ASTM F1080-93), a stainless-steel device consisting of a reservoir (5 x 5 x 4 cm), a movable gate, two adjusting screws for leveling, and a track marked with a ruler. The sample, conditioned to a temperature of 20°C, is placed into the reservoir. After the gate is opened and the timer started, the sample flows along the track. Consistency is determined by measuring the distance, in centimeters, traveled by the sample in 30 seconds. Prior to the measurement, the consistometer is leveled using the dedicated adjusting screws.

The total lipid content is determined through solvent extraction from a weighed aliquot of the sample (5 to 10 g). The extraction is carried out using ethyl ether in a Soxhlet apparatus for 4 hours. After extraction, the sample is placed in a rotary evaporator, followed by drying in an oven at 105°C for 2 hours to remove residual solvent. The extracted fat is then weighed at room temperature, and its content is expressed as a percentage of the initial sample weight.

3.2.3 On-line Instrumentation

A Sensure prototype camera (Sensure, Bergamo, Italy) is installed above the conveyor belt, positioned immediately after the basil plants are supplied. The camera captures RGB images every 15 seconds. The R, G, and B values are extracted from the images and treated as separate variables.

A ProFoss spectrometer (Foss, Hillerød, Denmark) was used to collect spectra of the basil, salt, and oil mixture, referred to as the intermediate product. The spectrometer is equipped with an optical fiber probe installed at the acquisition site on the process pipe. Spectra were acquired over the 1100–1650 nm spectral range in transmission mode, with a nominal resolution of 0.5 nm and 64 scans per sample.

3.2.4 Data Analysis

The semi-finished pesto, a mixture of oil, salt and chopped basil, produced in the first part of the production process, before the other ingredients were added, and the data from the vision system were analysed with two objectives: on the one hand, we wanted to evaluate the possibility of building an on-line monitoring model (Multivariate Control Charts) capable of describing the natural variability inherent to the process and capturing any anomalous fluctuations; on the other hand, we wanted to build predictive models (Predictive Models) to evaluate the feasibility of predicting the quality characteristics of the pesto sauce in real time. However, as described in section 2.4, prior to model building, selection of the appropriate preprocessing and the multivariate data exploration (Principal Component Analysis) was a mandatory step to check the data structure, the presence of outlying samples and to identify the time points corresponding to normal operating conditions of the plant.

Preprocessing

To ensure the effectiveness of the models, the data were pre-processed to remove effects that introduce variability. not related to the information to be retrieved. The pre-processing methods were adapted to the specific type of data and the objectives of each step, as illustrated below:

- **Vision System Data**
  The RGB data were preprocessed using autoscaling to standardize the variance among the different color channels.

- **NIR Spectra Prior to PCA and MSPC**
  NIR spectra were preprocessed to remove effects, such as scattering, that introduce variability unrelated to the information of interest, and to enhance the extractable information. Specifically, Savitzky-Golay second derivative and mean centering were applied before performing exploratory Principal Component Analysis (PCA) and constructing multivariate control charts.
- **NIR Spectra Prior to PLS Regression**
  The same preprocessing approach, Savitzky-Golay second derivative and mean centering, was applied to compute the Partial Least Squares (PLS) regression model for lipid content.
- **Preprocessing for Consistency PLS Model**
  A different preprocessing strategy was required to build the PLS model for consistency. Unlike lipids, which are chemically linked to specific absorption bands that guide the modeling, consistency is not directly associated with a chemical component. This makes it more challenging to model, particularly given the influence of process fluctuations on spectra collected online.

  To address this, the Dynamic Orthogonal Projection (DOP) algorithm [94] was applied to remove spectral variability that hindered the development of a satisfactory calibration model. DOP algorithm is part of the family of orthogonalization techniques and was originally introduced in the context of process monitoring, particularly to address situations where new process conditions influence the signals provided by sensors.

  The algorithm relies on four key elements: $\mathbf{X}_{\text{source}}$ and $\mathbf{y}_{\text{source}}$, representing the source scenario, and $\mathbf{X}_{\text{target}}$ and $\mathbf{y}_{\text{target}}$, corresponding to a separate set of samples in new conditions.

  In an ideal scenario $\mathbf{X}_{\text{source}}$ should include spectral data which are not affected by spourious/unwanted sources of variability, however this is not the case since spectra were collected at very high rate in-line during the process. Thus, the average spectra corresponding to the same consistency values in the calibration set were used as source data

($\mathbf{X}_{\text{source}}$). This approach assumes that such averaged spectra represent the true underlying spectral profile, unaffected by uncontrolled conditions. Conversely, the raw calibration spectra were used as target data ($\mathbf{X}_{\text{tar}}$). The DOP algorithm operates under the principle that samples with the same (or very similar) y-values should exhibit identical spectral profiles.

This allows for the estimation of "virtual" target spectra ($\mathbf{X}^*_{\text{tar}}$) that are unaffected by uncontrolled conditions. The estimation process uses a distance or association matrix ($\mathbf{M}$), calculated based on the y-values of the source ($\mathbf{y}_s$) and target ($\mathbf{y}_t$) domains. The singular value decomposition (SVD) [95] of the difference matrix between the measured and virtual target spectra is then employed to determine the components ($\mathbf{A}$) required for orthogonalization.

$$\mathbf{X}^*_{tar} = \mathbf{M}\,\mathbf{X}_{source} \tag{3.1}$$

$$\mathbf{D} = \mathbf{X}_{tar} - \mathbf{X}^*_{tar} \tag{3.2}$$

$$[\mathbf{U}_A, \mathbf{S}_A, \mathbf{V}_A] = SVD(\mathbf{D}, A) \tag{3.3}$$

$$\mathbf{X}_{source,corrected} = \mathbf{X}_{source}\,(\mathbf{I} - \mathbf{V}_A\mathbf{V}_A^T) \tag{3.4}$$

In our specific case, A=4 was used, after testing using from 1 to 5. Once the average spectra are corrected by orthogonal projection can be directly used to predict the validation set since the correction is embedded in the model, in this case only mean centering (of both $\mathbf{X}$ and $\mathbf{y}$) is applied prior to PLS.

Analysis

Principal Component Analysis (PCA) was used both for exploratory data analysis and to build multivariate control charts for MSPC. For the latter, the data set was manually divided into calibration and test sets, concentrating on observations under NOC. Each uninterrupted production period

was split as follows: approximately 65% of the temporally contiguous data points were assigned to the calibration set, while the remaining 35% were assigned to the test set. This split was aimed at emulating real-world continuous monitoring scenarios where the samples to be predicted follow the calibration data sequentially over time. Observations identified as outside the NOC based on exploratory PCA were fully included in the test set. To estimate the correct number of PCs cross-validation was performed with *venetian blind* scheme with ten splits. To determine whether a sample is extreme or anomalous, indicating a deviation from normal operating conditions, acceptance limits must be established for both T^2 and Q control charts. The T^2 limit is derived from Hotelling's T^2 distribution, while the Q limit is based on the χ^2 distribution. The Q limit can be computed using either the Jackson and Mudholkar approximation or the Box method.

For the PLS models Data has been partitioned into calibration (70%) and validation (30%) sets by means of Duplex algorithm [96]. The PLS model dimensionality, i.e. the number of PLS components, has been assessed by the Root Mean Square Error in Cross-Validation (RMSECV), while the Root Mean Square Error in Prediction (RMSEP) has been used to evaluate the models predictive capability. Residuals plots were also inspected.

3.3 Results and Discussion

3.3.1 Exploratory Data Analysis

Each type of data, RGB parameters and NIR spectra, was analysed separately to visualise and explore the data structure. PCA applied to the NIR spectra (acquired over 459 time points) revealed a distinct cluster of samples with negative values of PC1 and positive values of PC2, as shown in Figure 3.2a, very far from, and therefore different from, all other samples. The plot of PC1 versus time (Figure 3.2b) clearly shows that these samples consistently correspond to production restarts after periods of inactivity.

Figure 3.2d shows the loadings plot for PC1 and PC2, blue and red

lines respectively, where it is possible to see the absorption bands mainly responsible for this difference. However, in order to interpret the scores and loadings plots together, a PC1 vs. PC2 loadings scatter plot has also been generated (Figure 3.2c). In both figures, the wavelengths contributing to the separation between NOC and anomalous samples are highlighted in purple. In particular, the band at 1400 nm, although the most intense in PC1, does not contribute to the description of anomalous samples, but is associated with extreme NOC samples with high PC1 scores, as observed in Figure 3.2a.

Conversely, the bands at 1170, 1213, 1236 and 1410 nm describe the behaviour of the anomalous samples. These bands fall along the separation direction, indicating that these samples have significantly different absorptions at these wavelengths. In detail, the bands at 1178 nm and 1410 nm can be attributed to lignin: the second overtone of the C-H bond stretching in CH_3 and the first overtone of the O-H bond stretching in the ROH group, respectively. In addition, the bands at 1213 and 1236 nm correspond to the first and second overtones of the C-H bond stretching in oleic and linoleic acids (CH_2) found in olive oil [97, 98].

Since these samples show outliers behaviour, as they clearly do not represent the NOC, were removed and a new PCA model was built in order to obtain a better visualization of possible differences among NOC samples.

For the new model, the first PC (79.36% of variance explained) did not show any interesting trend, so PC2 and PC3 were inspected. Figure 3.3 a,b shows the score plot of PC2 vs. PC3, where the samples are coloured according to the different additional information available, i.e. suppliers and different cuts. The names of the suppliers have not been disclosed due to confidentiality restrictions agreed with the company. PC2 discriminated the samples by supplier, as almost all samples from supplier number two had positive PC2 values, and the samples from suppliers three and four had negative PC values, suggesting that they were more similar to each other with respect to number two. Only the samples coming from supplier five did not clearly differentiate from the others, whereas the number of samples from supplier one were too low to judge. In addition, PC2 and PC3 were

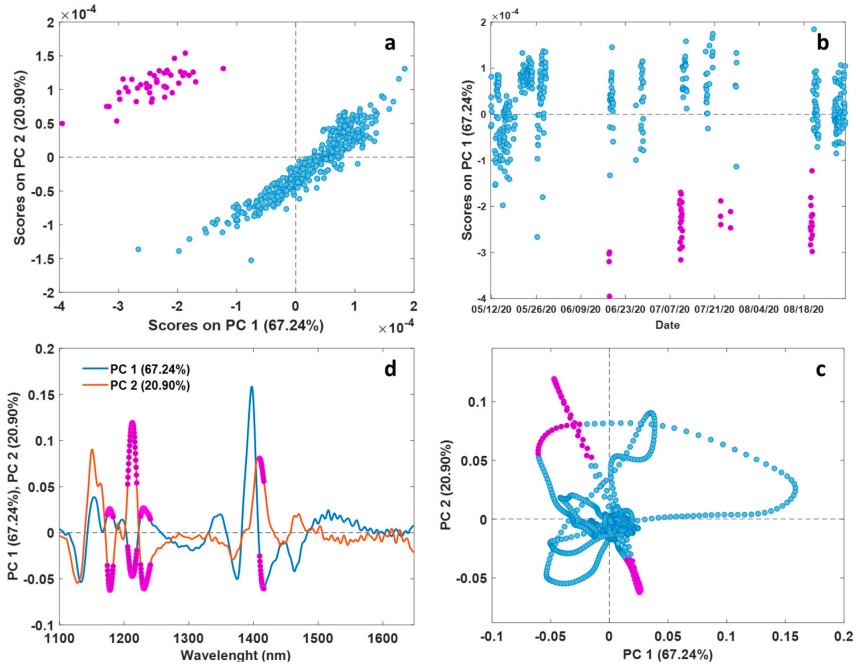


Figure 3.2 – Results of the Exploratory Data Analysis performed on NIR data: PC1 vs. PC2 scores plot (**a**), scores on PC1 as a function of time (**b**), loadings on PC1 and PC2 as a function of wavelength (**c**), and loadings on PC1 vs. PC2 (**d**). In (**a**) and (**b**), purple points represent anomalous samples, while in (**c**) and (**d**), purple points highlight wavelengths that primarily characterize the differences between anomalous samples and the other ones.

able to discriminate between samples related to cuts one and two (negative values of PC2 and positive values of PC3) with respect to samples related to cuts three and four.

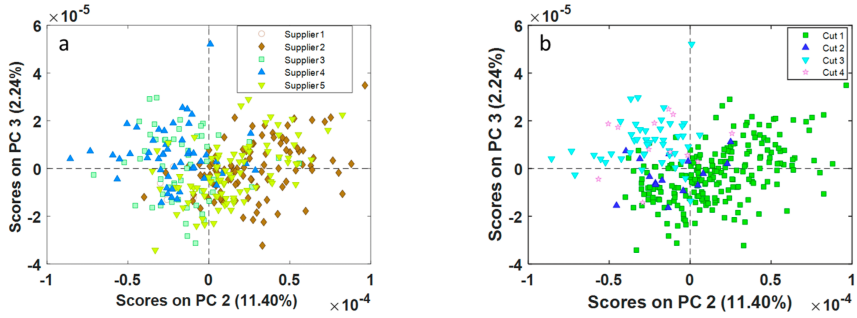


Figure 3.3 – Results of the Exploratory Data Analysis performed on NIR data. PC2 vs PC3 Scores plots colored by different suppliers (**a**) and cuts (**b**).

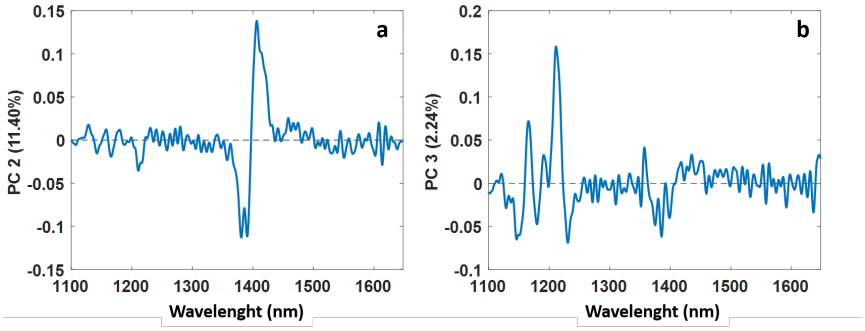


Figure 3.4 – Loadings plot of PC2 and PC3, respectively.

The possibility of distinguishing between different cuts is relevant for the company, as younger basil plants generally give a higher quality product. However, looking at the two plots simultaneously, it is clear that only certain suppliers, namely number three and four, have supplied samples characterised by low cuts. Figure 3.4 a and b show the loading plots of PC2 and PC3 respectively, showing the NIR bands responsible for these differences. Even if it is not possible to assess whether suppliers or cuts influence them, the PCA results showed to be a valuable tool to assess whether the incoming information about the raw material could be linked to the intermediate product characteristics, obviously a more systematic planning of the next harvesting campaigns could clarify whether the cut or the supplier is the influencing factor.

The PCA analysis performed on the data collected by the RGB camera was not able to detect the anomalous behaviour of the samples highlighted in Figure 3.2. A possible explanation is that after a stop, when production is restarted, the process needs time to return to NOC and it may happen that the NIR spectra refer to material that is probably a residue of the old process (before the restart) and therefore the acquired spectra do not initially reflect the newly produced intermediate product. In addition, the observation of sample separation due to different cuts or suppliers is less efficient than the corresponding analysis performed on the NIR spectra. Therefore, these differences are not related to colour variations, but mostly to the "chemical" profile of the basil.

3.3.2 MSPC Charts

The most notable results from the MSPC charts based on PCA were obtained using only the NIR data, as the inclusion of RGB parameters did not provide additional insights. The PCA model, which explains 93% of the data variance using four principal components, was constructed by including in the calibration set (294 samples) only those samples identified as belonging to NOC according to plant experts. The test set (165 samples) comprised both NOC and anomalous samples.

The T^2 chart, shown in Figure 3.5a, represents the distance of each sample from the origin within the model space. Black circles indicate the calibration samples used to build the PCA model, while red diamonds represent the test samples projected onto the model. This chart identified five groups of samples with high T^2 values, corresponding to NIR spectra collected during production restarts. No other test samples exceeded the T^2 limit.

In the Q chart (Figure 3.5b), which quantifies the distance of each sample from the model space, the same samples associated with production restarts were identified as anomalous, similar to the T^2 chart. This indicates that the model does not adequately describe these samples. A few additional, non-consecutive test samples, representing less than the nominal 5% of the total, exceeded the chart limits.

To further investigate, samples were colour-coded based on cut, supplier, consistency, and lipid values to determine if these features influenced their behaviour. However, no significant trends were observed.

These results demonstrate the effectiveness of the MSPC charts in detecting deviations from NOC, which correspond to differences in intermediate products. This capability facilitates faster identification of potential plant issues or, as in this case, monitoring the process adaptation as it returns to NOC conditions after a production stop. NIR spectroscopy proved to be highly sensitive to variability in intermediate product samples, whether caused by process resetting (as in this case), process drift, or also to varia-

tion of NIR instrumentation setting/performance. Interpretation of loadings and analysis of previous production campaigns data may help discerning the different situations.

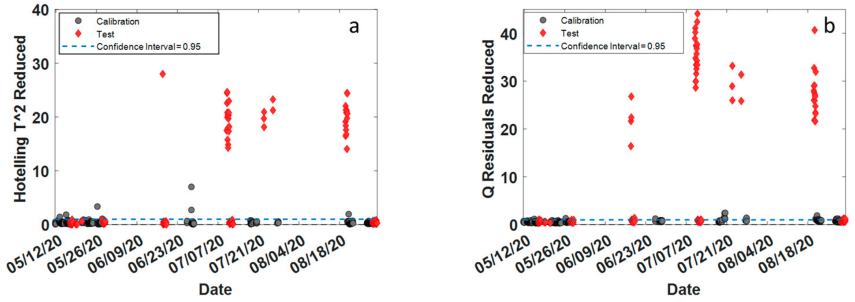


Figure 3.5 – **a)** T^2 and **b)** Q based MSPC charts.

3.3.3 Predictive Models

An attempt was made to develop predictive models that could potentially be used to estimate the quality properties of the final product in real-time. Since the RGB data did not yield reliable prediction models for the parameters, only results obtained from NIR data are presented. In Table 3.1 summarised the results for all properties even if only consistency and lipid content are discussed in detail.

Table 3.1 – PLS regression results for multivariate calibration of pesto quality parameters by using on-line NIR (70/30% calibration/validation split by duplex)

Quality parameters	LVs	RMSECV	RMSEP	% average error
Consistency (cm)	9	0.64	0.68	9.88
Lipids (w/w%)	5	1.6	2.0	2.5
pH	8	0.056	0.065	1.1
Water activity	4	0.001	0.0044	0.37
Dry residue (w/w%)	4	0.4254	0.5745	0.63

A critical challenge in this context is ensuring the correct matching between the intermediate product sample (at a specific production time), for which the NIR spectrum is acquired, and the corresponding finished pesto product at the end of the production line. This matching requires careful

consideration of the residence time, as the quality parameters of the finished product are assessed off-line using reference methods.

Before model computation, data were split by using the duplex algorithm with a 70/30% proportion in calibration and test sets, giving 142(cal)/61(test) and 33(cal)/12(test) for consistency and lipids, respectively. Afterwards, four samples belonging to the anomalous group of observations, detected by using T^2 and Q distance, were removed from the test set for consistency. The prediction model for consistency was built using 9 latent variables (LVs), selected based on the minimum RMSECV value determined through a *venetian blind* cross-validation scheme with 10 splits. The RMSEP value was found to be close to the RMSECV (Table 3.1), corresponding to an average relative percentage error of 10% in prediction. This level of error was considered acceptable by the company for an early, on-line quality estimation of the intermediate product.

Samples in the test set exhibited considerably higher variability compared to those in the calibration set (Figure 3.6a and 3.6b). Nevertheless, the residuals versus measured values plot for consistency (Figure 3.6b) showed that errors for both calibration and test samples were randomly distributed, with no visible trends or systematic bias.

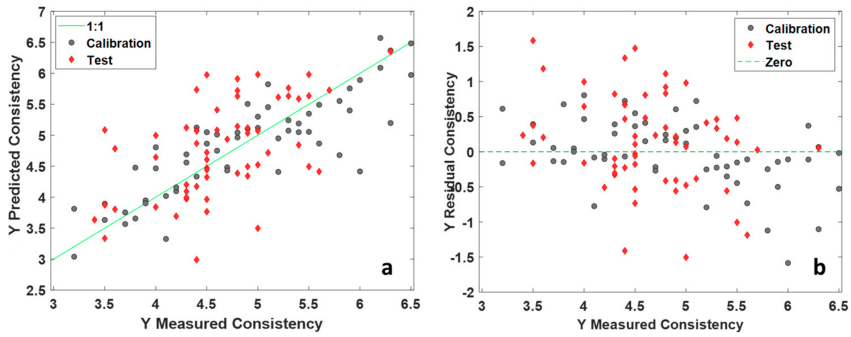


Figure 3.6 – PLS results on NIR data for consistency. Predicted vs measured values plot (**a**), residuals vs measured values plot (**b**).

The prediction model for total lipid content was constructed using a smaller number of samples compared to the consistency model, as this parameter is assessed less frequently. In this case, 5 LVs were selected based on

the minimum RMSECV value obtained using the same *venetian blind* cross-validation approach (10 splits). As depicted in figure 3.7a, the majority of the samples had a lipid content ranging between 46% and 49%, with only a few samples exhibiting higher values. This distribution reflects the common real-time production scenario, where maintaining consistent product quality is a primary objective.

Although a couple of samples in the test set were predicted with higher errors, the majority of predictions fell within a 2% error range, which the company considers acceptable for determining if the product meets the specification for this parameter. Notably, one of the two test samples with high lipid content was predicted accurately, while the other was underestimated (Figure 3.7b).

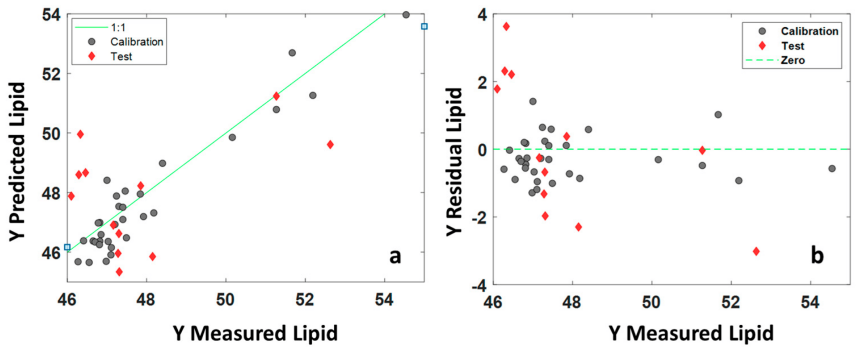


Figure 3.7 – PLS results on NIR data for lipids content. Predicted vs. measured values plot (**a**), residuals vs. measured values plot (**b**).

Variable Influence in Projection (VIP) scores were used to identify which variables, in this case specific bands, contributed most significantly to the predictive model. The analysis revealed that the band at 1166 nm, attributable to the olive oil's second overtone of the CH stretching of CH_3 [97, 98], was the most influential for predicting total lipids content. Additionally, other bands associated with lipids in olive oil [97, 98] were observed at 1422 nm and 1461 nm, corresponding to the CH stretching and deformation of CH_2, both exceeding the significance threshold [99].

Briefly commenting on the results of the other properties. The pH model shows similar values of RMSECV and RMSEP, with average prediction er-

rors of 1.1% acceptable for on-line quality monitoring. The most influential spectral bands include those associated with CH_2 and CH_3, O-H of ROH and aromatic groups, and N-H.

In the water activity model, the test samples showed higher errors than the calibration set (RMSECV 0.001; RMSEP 0.004), but the errors were within an acceptable range of ±0.01. The key bands for this parameter are related to O-H around 1450 nm.

The model for the dry residue shows comparable RMSECV and RMSEP values, with an average error of 0.63% considered acceptable by the company. However, the residuals show a linear trend, suggesting a possible underfitting of the model. The most influential bands include those associated with C-H and N-H.

3.4 Conclusion

This study explores the feasibility of real-time monitoring in an industrial food processing context, specifically focusing on a pesto production line. Due to the unavailability of historical data, the findings are limited to a single basil harvesting campaign. The analysis addressed two main objectives: the use of latent variable-based multivariate control charts to monitor process stability and identify deviations exceeding natural variability, and the development of real-time predictive models for quality parameter estimation. Despite the constraints of limited data, the results offered valuable insights, summarized as follows.

Multivariate Statistical Process Control

i) The RGB parameters collected via the vision system, although potentially beneficial, did not contribute additional information beyond what was captured by the NIR spectroscopy. This limitation may stem from the restricted number of features extracted from the images. Further research is needed to enhance this aspect, for instance,

by leveraging image analysis tools to detect and quantify damaged leaves, branches, and stem percentages.

ii) NIR-based multivariate control charts successfully identified changes in the intermediate product caused by production restarts after temporary stops. This finding highlights the sensitivity of NIR spectroscopy to even minor process changes. Additionally, the control charts demonstrated their ability to determine when fluctuations reverted to the natural variability of the process, ensuring product consistency.

On-line Predictive Models

iii) Predictive models developed to estimate the final product quality of pesto, based on the NIR spectra of intermediate products, yielded errors in external predictions deemed acceptable for real-time quality assessment by the company.

iv) A notable challenge in building predictive models for final product quality parameters lies in the constrained variability of the responses, which are naturally bound to the predefined specification ranges. Without the possibility of extending the calibration range through pilot studies, the models are best used as early indicators rather than precise predictors of quality values. Even with limited calibration data, these models can provide preliminary checks to confirm whether product specifications are being met for determining whether the product remains within acceptable specification ranges. The percentage error observed in the predictions was within the company's tolerances, supporting the practical utility of these models in real-time monitoring scenarios.

In conclusion, it is important to highlight the key challenges encountered during the study, including the absence of systematic recording for online-acquired data, the complexities involved in establishing an effective synchronization scheme, and the essential role of spectral preprocessing in addressing the numerous sources of variability inherent in the process environment. Despite these challenges, the analysis of the available data and the

initial results of this work have demonstrated the feasibility of implementing process monitoring strategies. These findings pave the way for the design and development of a systematic data storage framework that would significantly advance efforts in this direction, enabling more robust and reliable monitoring and optimisation of industrial processes.

Chapter 4

A Comparative Study of Chemometrics and Deep Learning on Semantic Segmentation Classification

4.1 Introduction

Images are one of the most common and versatile data formats, conveying visual information in an immediate and intuitive way. Technically, an image can be thought of as a collection of numerical values arranged on a surface, values captured by any device capable of capturing data in the X and Y directions [100]. Beyond their mere visual representation, images contain a wealth of complex information that can be extracted using advanced processing techniques for recognition, segmentation and feature identification. These methods play a key role in a wide range of industrial and research applications, from diagnostic analysis to process control.

Visual perception is crucial for both humans and computers. Humans exhibit remarkable performance in visual tasks, as we are able to recognise

fine details, isolate objects and separate them from the background with ease [101]. The goal of computer vision is to replicate this ability to extract meaningful information from images through a variety of techniques that mimic human visual perception. One particularly powerful application area is the industry [102], where artificial vision systems are revolutionising processes such as quality control, product selection and packaging [103]. For example, these systems can inspect products to detect defects, measure dimensions and verify compliance with specified quality standards [104–106]

In computer vision, image segmentation is one of the most important steps in image analysis and understanding. By segmenting an image into meaningful, non-overlapping regions, each containing objects with similar properties, it becomes possible to isolate the target features or components of interest [107]. Whether using classical techniques, such as threshold-based segmentation and edge detection [108, 109], or modern deep learning approaches, segmentation remains a core process in numerous applications. In food quality control, for example, image segmentation can be used to detect directly online defective apples [110]

Building on the concepts discussed in the previous chapter, this work aims to increase the amount of information extracted from vision systems used to inspect basil leaves destined for the production of traditional Genoese pesto. The current output, in the form of average RGB values, has proved insufficient for real-time quality estimation and Multivariate Statistical Process Control (MSPC). However, previous knowledge suggests that the ratio of stems to leaves has a significant influence on the final characteristics of the product. Therefore, this study aims to derive novel features, specifically the percentage of stems, leaves and background in RGB images, using two complementary data analysis paradigms. Under the "classical" approach, we initially used various image segmentation methods, such as the Otsu method [108], and object detection tools available in the Matlab Image Analysis Toolbox. However, the highly variable illumination conditions made it difficult to identify robust segmentation thresholds that worked consistently across all images. Therefore, we adopted another method combining wavelet transform (WT) filtering for feature enhancement [111, 112],

coupled with PLS-DA for pixel classification. Moreover, following this philosophy of approaches, feature extraction (characterizing the spatial distribution of pixel intensities in images to obtain statistical parameters that can differentiate images containing the same elements but with different spatial arrangements, known as texture features [113, 114]) followed by PLS-DA has been considered. In parallel, a second approach that takes advantage of the power of deep learning networks. [115].

This work does not aim to identify which approach is the best, but by comparing these two methodologies, we can provide a broader view of how different analytical strategies can be exploited, in this case, to improve real-time monitoring and quality control of basil leaves for a pesto production process.

4.2 Material and Methods

4.2.1 Basil RGB Images

An RGB vision system (Sensure, Orio Al Serio, BG, Italy) [116] was installed in the factory for on-line analysis of the basil plants. Although the system does not normally store the acquired images due to limited storage space, several hundred images were manually archived during the 2021 and 2022 production periods with the aim of improving the relevant information that could be extracted from the collected data, such as the percentage of stems and leaves.

The RGB images captured have dimensions of 1280×1020 pixels, and an example of such an image is shown in Figure 4.1. Three images are shown, illustrating different degrees of coverage depending on the amount of basil on the conveyor belt, as well as different lighting conditions influenced by the characteristics of the vision system. In Figure 4.1a it can be seen that the conveyor belt is more covered by basil plants compared to the other images. In addition, a comparison between Figure 4.1a and Figure 4.1b highlights differences in the lighting system. In particular, Figure 4.1c shows

a reflection line on the right side that is not present in the other images.

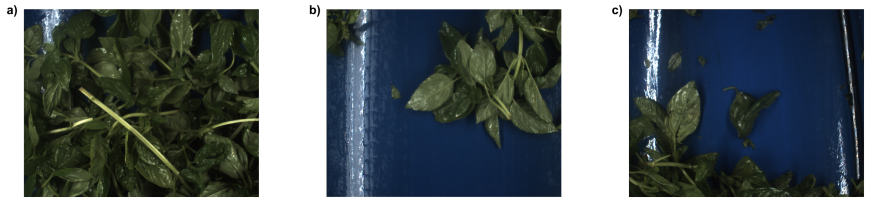


Figure 4.1 – RGB images of basil leaves acquired in-line on the blue conveyor belt. **a), b), c)** show the different degree of coverage and variation in illumination

To train the classification models described in the following sections, each pixel had to be labelled. This was done using a simple graphics editor such as Paint, where different colours were assigned to each class: red for the branches and yellow for the stems, but the background was not explicitly coloured. Instead, it was assigned by exclusion: once the labelled images were imported, any pixel that was neither red nor yellow was automatically designated as the background. The following figure shows the images labels shown in Figure 4.1.

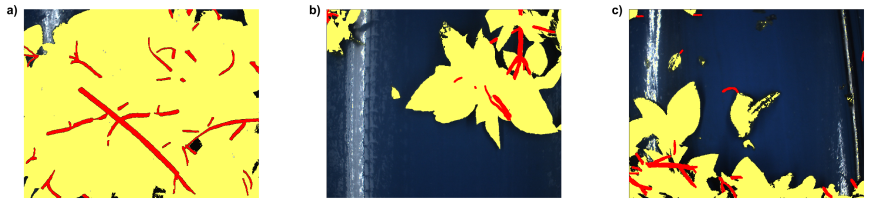


Figure 4.2 – Labeled images used for training the classification model where red pixels represent branches, yellow pixels correspond to stems.

4.2.2 Partial Least Squares – Discriminant Analysis (PLS-DA)

Partial Least Squares – Discriminant Analysis (PLS-DA) [117, 118] is a classification methods based on Partial Least Squares (PLS) regression. In PLS-DA, the relationship between the predictor matrix $\mathbf{X}$ and the response matrix $\mathbf{Y}$ is modelled by a PLS model as described in section 2.4.2. However, unlike standard PLS regression, where $\mathbf{Y}$ contains continuous real values, in PLS-DA $\mathbf{Y}$ consists of categorical data encoded in a binary format. Each

column of $\mathbf{Y}$ corresponds to one class, and for any given row representing the n-th sample, a value of 1 indicates membership in the corresponding class, while 0 indicates non-membership. For example, in a three-class problem, if the row for the n-th sample is $[0, 1, 0]$, this implies that the sample belongs to the second class.

Once the optimal PLS-DA model is built with A latent variables, the regression coefficients $\mathbf{B}$ can be used to predict $\hat{\mathbf{y}}$ values for new samples ($\mathbf{x}_{\text{new}}$). However, these predicted values are continuous and do not directly indicate class membership. To assign samples to specific classes, a classification rule must be applied to convert the continuous $\hat{\mathbf{y}}$ values into categorical outputs.

Several classification strategies can be employed:

- **Maximum Value Rule:** Assign the sample to the class with the highest predicted $\hat{\mathbf{y}}$ value. While simple, this approach may not be ideal for problems involving more than two classes [119].
- **Linear Discriminant Analysis (LDA)** [120] or **Quadratic Discriminant Analysis (QDA)** [121]: These methods can be applied to the $\mathbf{X}$-scores or $\mathbf{Y}$-scores to determine class membership [122].
- **Class Thresholds** [123]: A specific threshold can be defined for each class. Thresholds are typically optimized based on classification performance estimated through cross-validation.

Each of these strategies has its strengths and limitations, and the choice of method often depends on the specific characteristics of the dataset and the classification problem at hand.

4.2.3 Approach 1: 2D WT-MIA coupled with PLS-DA

2D WT-MIA (2D-wavelet decomposition and multivariate image analysis)[111] combines two powerful techniques: two-dimensional wavelet decomposition (2D WT) and multivariate image analysis (MIA). This approach is aimed at extracting and analysing essential features from images by taking advantage of the ability of wavelets to decompose an image into compo-

nents at different levels of detail, followed by the application of multivariate methods to analyse, classify or predict based on the extracted features.

In the first step, two-dimensional wavelet decomposition (2D WT) works on each colour channel of an image by breaking it down into a series of sub-images representing different levels of spatial detail. First, a combination of low-pass and high-pass filters are applied along the rows of the image. The resulting outputs are then filtered again along the columns. This process produces four distinct sub-images. The approximation sub-image (CA) is computed using low-pass filters in both directions and captures smooth and large scale changes, such as tones. The other three sub-images represent horizontal detail (CH), computed using a combination of low-pass and high-pass filters; vertical detail (CV), computed using high-pass filtering in the rows and low-pass filtering in the columns; and diagonal detail (CD), resulting from high-pass filtering in both directions. These three detail subimages capture sharp and oriented changes such as stripes or edges.

The decomposition process can be iterated multiple times by reapplying the filters to the approximation subimage, with the number of iterations referred to as the decomposition level. The level is determined by the dimensions of the image, and in our analysis we performed the decomposition up to the maximum level (9) allowed by the image size. We used the Haar wavelet filter [124] (also known as db1) and implemented the stationary wavelet transform (SWT) [125].

As shown in Figure 4.3, after the wavelet decomposition, the extracted sub-images were subjected to multivariate analysis techniques, in this case with the aims of classifying each pixel classification model PLS-DA has been considered. For a single channel image, each decomposition sub-image of dimensions $n_1 \times n_2$, is unfolded pixel by pixel, obtaining a matrix of size $n_1 \times n_2$ rows $\times$ 4 columns $\times$ level decomposition. Then the matrices corresponding to the different channels are concatenated column by column. Later, the data were autoscaled to compensate for the different scales of the WT coefficients at different decomposition levels, and finally analysed using a hierarchical PLS-DA strategy. In particular, a first PLS-DA was developed to identify the basil plant's background. For example, by assigning class 0 in

figure 4.3 to the background and class 1 to the basil plant (leaves + stems), later on the pixel identified as basil plant, another PLS-DA was performed to classify the leaves from the stems.

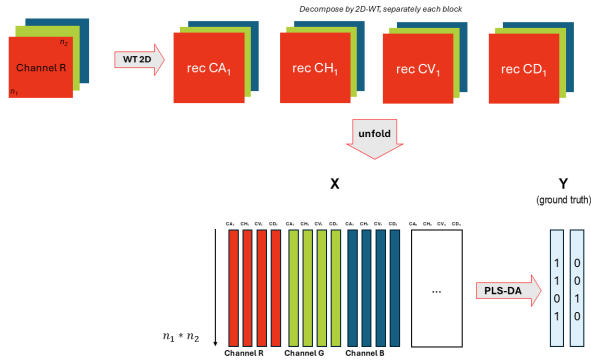


Figure 4.3 – Workflow of the approach. The wavelet decomposition of a RGB channels image at the first decomposition level followed by PLS-DA model 1 to predict background vs (stems + leaves).

The calibration models were developed using four different images as a calibration set. These images were chosen to represent different degrees of conveyor belt coverage and different lighting conditions. The number of components for both PLS-DA models was determined by selecting the number that minimised the classification error during cross-validation, which was performed using a five-split venetian blind approach. The classification decision rule was to assign each pixel to the class with the highest predicted **Y** probability.

After optimising the two PLS-DA models, they were tested on an independent set of images. By refolding the predicted class membership vector, the spatial arrangement of the correctly classified pixels could be visualised.

4.2.4 Approach 2: Features Extraction + PLS-DA

This strategy is inspired by the feature extraction process typical of convolutional neural networks (CNNs). Specifically, it uses a 5×5 sliding window approach (called a kernel). To preserve the original image size, a padding (frame around the image) of 2 pixels is applied to all sides of the

image before filtering. This kernel moves through the image pixel by pixel, and within each 5×5 window, four features are calculated from the twenty-five pixel values: mean, median, standard deviation and entropy [114]. In particular, entropy is a measure of the degree of disorder or randomness within the window, providing insight into the local texture; mathematically, it is calculated according to the equation 4.1, where p denote the single bin frequency of the normalized distribution, within the kernel.

$$H = -\sum p \log_2(p) \tag{4.1}$$

This feature extraction procedure is applied independently to each channel of the image, resulting in four different feature maps per channel. Notably, these kernels do not require optimization, making the approach computationally efficient under this point of view. Once the features have been computed for all channels, these feature maps are unfolded into a single matrix, where each row represents a pixel and each column corresponds to one of the extracted features. This matrix is the input for the subsequent classification phase. In this study, a Partial Least Squares Discriminant Analysis (PLS-DA) is performed with the aim of classifying three different classes (background, leaves, branches).

The calibration model for this first approach was developed using four different images as calibration sets. These images are the same as those used in the first approach and were chosen to represent different degrees of conveyor coverage and different lighting conditions. Even in this case, it is necessary to define the number of Latent Variables, which is determined through a cross-validation step.

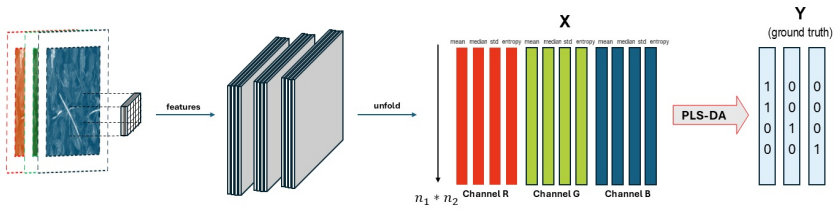


Figure 4.4 – Schematic representation of the approach. On the left, the three channels of the original image with the padding frame. A 5×5 kernel is applied to extract four features (mean, median, standard deviation, and entropy) for each channel. The extracted features are organized into feature images. Next, the data is transformed into a matrix $\mathbf{X}$ through an unfolding. Finally, a PLS-DA model is applied to classify the pixels.

4.2.5 Approach 3: Convolutional Neural Network

In recent years, deep learning [126–128] has emerged as one of the most fashionable and revolutionary approaches in the field of artificial intelligence. It is based on the use of a large number of artificial neurons, organised in different layers, to learn complex data structures. Although the theoretical basis of deep learning dates back several decades, it has only recently become viable for a wide range of applications due to advances in hardware performance, particularly graphics processing units (GPUs).

Deep neural networks consist of three main components: an input layer, one or more intermediate layers (hidden layers) and an output layer. The input layer receives raw data, such as images or signals, while the intermediate layers transform and combine the received signals through non-linear activation functions, enabling the network to learn complex relationships in the data. Finally, the output layer generates the final result, which can be a continuous value, a probability or a class.

Among the most widely used deep learning architectures, convolutional neural networks (CNNs) [127, 129] have gained enormous popularity, especially in image and video processing. CNNs are designed to exploit the spatial structure of visual data, making them extremely effective for tasks such as object recognition, image classification and semantic segmentation.

They recognise objects regardless of their position in the image, making them ideal for dealing with complex scenarios such as variations in lighting or visual disturbances.

Its architecture is based on the use of filters (or kernels) that automatically extract relevant features from the input data through the convolution operation. These filters allow the network to identify local patterns, such as edges, textures and other fundamental structures, without the need to manually specify these features.

Two approaches are commonly used: One is to use existing deep learning architectures and, if necessary, optimise their performance by fine-tuning them according to actual needs in order to save time while ensuring good performance. The other is to define new deep learning architectures and train them on image datasets. This approach is generally more time-consuming, but allows the model to be better adapted to the case at hand.

CNN - Architectures

For this case study, we opted to use a CNN built from scratch, as represented in Figure 4.5. The architecture is inspired by the U-Net structure [130], featuring a two-phase design: a convolutional stage followed by a deconvolutional phase. The convolutional stage is designed to act as a feature extractor, identifying and encoding the most relevant information from the input data. The transposed convolutions are used on the extracted features to reconstruct the pixelwise labels, this step can be thought of as an attempt to interpolate/upscale from feature space to pixelwise label space (original space).

The model starts with a convolutional stage aimed at extracting features, and progressively moves to a reconstruction stage to produce pixel-wise labels. The input to the network consists of a patch of the original RGB image with dimensions $256 \times 256 \times 3$.

The first convolutional layer uses 64 3×3 filters to extract features such as edges, textures, and gradients from the input image. Each filter slides

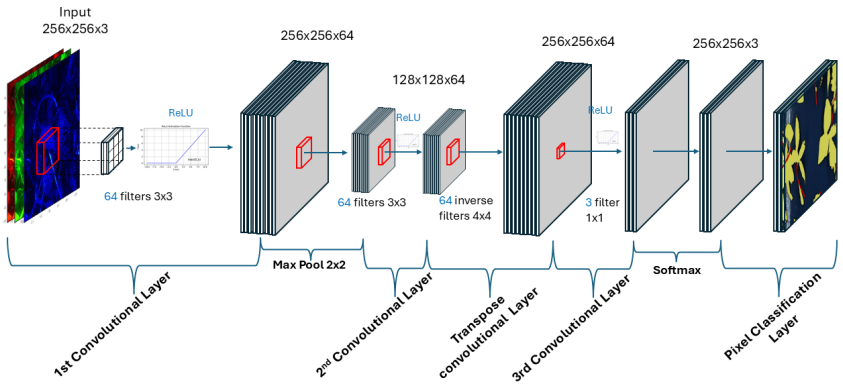


Figure 4.5 – Architecture of the Convolutional Neural Network (CNN) designed and implemented for this study.

across the image, performing a dot product between the filter values (or kernel weights) and the corresponding local region of the image. This process effectively highlights specific patterns or features within the image, based on the learned filter weights. The output of this operation is a set of feature maps that represent the detected patterns at various spatial locations. This is followed by a ReLU activation function. The resulting feature maps, after applying ReLU (equation 4.2), retain the original 256×256 spatial resolution while encapsulating the essential non-linear features needed for the subsequent layers.

$$f(x) = \begin{cases} x, & \text{for } x > 0, \\ 0, & \text{for } x \leq 0. \end{cases} \tag{4.2}$$

ReLU activation function introduces non-linearity into the model, enabling it to learn more complex patterns beyond linear combinations. ReLU has several key advantages, first, it helps avoid the vanishing gradient problem, where gradients become very small during back propagation, slowing down or halting weight updates [131]. Additionally, by keeping gradients significant for positive inputs, ReLU ensures more efficient training, improves convergence by allowing faster and more stable optimization compared to

other activation functions like sigmoid or tanh [132].

Moreover, ReLU promotes sparsity in the network by setting negative values to zero, effectively deactivating certain neurons. This sparsity acts as a form of implicit feature selection, making the network less dependent on all connections and reducing the risk of over fitting [131].

In the second layer, the model continues the feature extraction process by applying another set of 64 convolutional filters of the same size. In order to reduce the spatial dimensions, a max-pooling operation with a 2×2 kernel is applied, downscaling the feature maps to 128×128. This pooling operation helps to preserve the most salient features while reducing the computational complexity of subsequent layers [133]. However, this operation cannot be inverted as some of the information is lost during the pooling process. In the case of max-pooling, for example, only the maximum values of each window are retained, while all other values are discarded, making it impossible to reconstruct the original input from the output.

Next, the architecture includes a transposed convolutional layer of 64 4×4 filters. This layer is responsible for upscaling the feature maps [134] back to their original 256×256 spatial resolution. The final convolutional layer uses three 1×1 filters to map the high-dimensional feature space to the three output classes corresponding to the pixel labels. This step preserves the spatial resolution while reducing the depth of the feature maps. This last two steps were necessary to have the spatial resolution in the original image to do a proper pixel classification and to have the third dimension as the number of classes to predict, in our case 3 classes background leaves and branches, necessary for the softmax.

The output is then passed through a softmax activation function, which calculates the probability of each pixel belonging to each class. The pixel classification layer assigns the final class label to each pixel based on the highest predicted probability. In detail, to cover the class imbalance, a classification layer has been used that takes the values from the softmax function and assigns each input to one of the classes [135], taking into account the class weight. In our case, the weight has been calculated as the inverse of

the frequency.

Training and optimization

The training phase of the Deep Learning has the aims to optimize the weights of all layers in the Convolutional Neural Network (CNN). This optimization was achieved using the backpropagation algorithm [136], a method that calculates the gradient of the loss function with respect to each weight in the network. The algorithm propagates the error backward through the network, starting from the output layer, allowing the model to iteratively adjust its weights to minimize the error.

For the training of Deep Learning approach 104 images have been used and a data augmentation strategy was applied. The data augmentation process involved creating random transformations of the input images, such as random horizontal and vertical reflections and rotations. In addition, the training datasets was prepared extracting 128 random patch from the original images. This process has been repeated also for the validation of 15 images used for the optimization of the neural network (net) parameters.

The Adam optimizer [137], a widely used optimization algorithm in deep learning, was employed during training due to its ability to combine the benefits of adaptive learning rates and momentum. It determines how the weights are updated according to the gradients provided by backpropagation. The model was trained for a maximum of 200 epochs, where each epoch represents a complete pass through the training dataset. To avoid unnecessary computations and mitigate the risk of overfitting, early stopping was implemented with a validation patience of 50 epochs.

The hyper-parameters for the convolutional neural network (see Table 4.1) were optimized using a random grid search strategy [138]. The parameters considered [126] in this process included the initial learning rate, batch size, L2 regularization factor, gradient decay factor, and the number of filters in the convolutional layers selecting the combination that provide the best values of accuracy. The initial learning rate determines the step size

taken by the optimiser during gradient descent. A carefully chosen value provides a balance between convergence speed and stability. The batch size defines the number of samples used to compute the gradient and update the weights in each iteration. Smaller batch sizes provide more frequent updates but may introduce noise, while larger batches stabilise gradients at the cost of higher memory and computational requirements.

L2 regularisation, also known as weight decay, adds a penalty to large model weights, reducing the risk of overfitting and encouraging simpler, more generalisable solutions.

Finally, the number of filters determines the model's ability to extract features in each convolutional layer. A higher number of filters allows the model to learn more complex patterns, but also increases computational cost and the potential for overfitting.

The values for each hyperparameter tested and the one selected after the optimisation procedure as described above are shown in the following table.

Table 4.1 – Optimized Hyperparameters

Parameter	**Options**	**Selected Value**
Initial Learning Rate	$0.0001, 0.0003, 0.0010, 0.0032, 0.0100$	0.0003
Batch Size	$32, 64, 128$	128
L2 Regularization	$0.0001, 0.0003, 0.0010, 0.0032, 0.0100$	0.0003
Gradient Decay Factor	$0.8, 0.9, 0.99$	0.9
Number of Filters	$16, 32, 64, 128$	64

4.3 Results and Discussion

For all the approaches considered, the performance evaluation has been done on an external validation set of 15 images (Figure 4.6) acquired on the 2022 harvest production, despite to the calibration/validation that contain the image acquired during the 2020-2021.

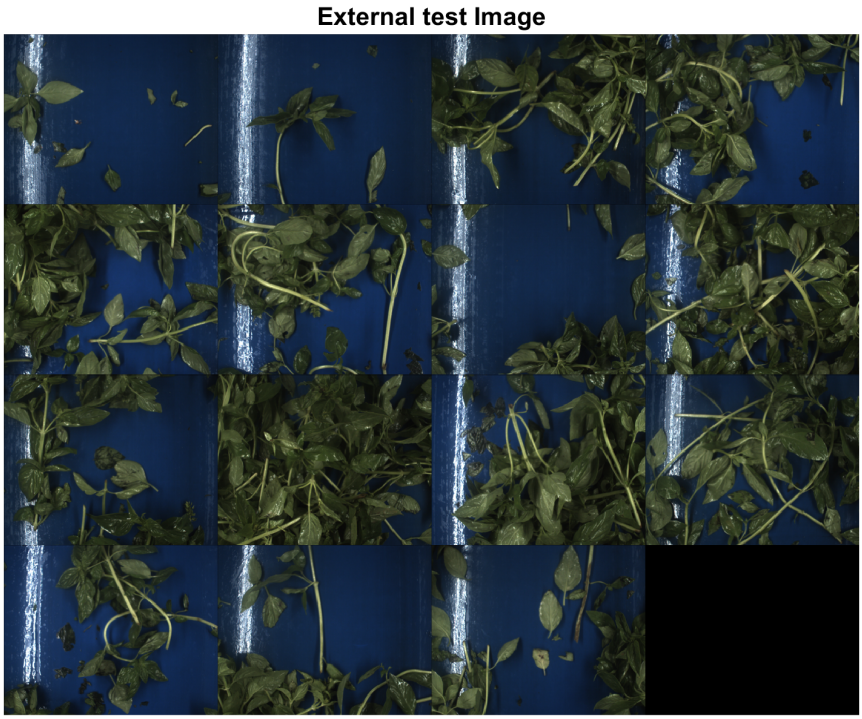
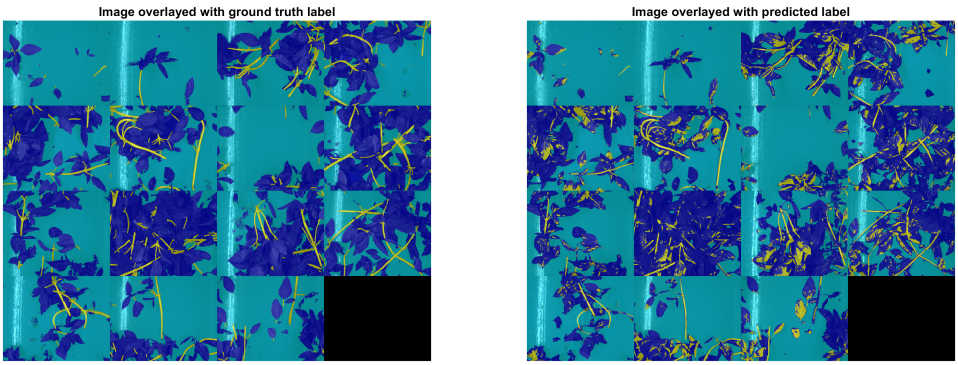


Figure 4.6 – Illustration of the 15 images used as external test set

4.3.1 Approach 1

The comparison between ground truth (Figure 4.7a) and the results obtained using the proposed approach, WTMIA + PLS-DA (Figure 4.7b), with most likely classification criterion, highlights the method's ability to identify the classes of interest in the analysed images. However, while background recognition (light blue) is very good in all test images, some discrepancies are observed in the prediction of stems (yellow) and leaves (blue). Some difficulties emerge in distinguishing thin stems from leaves, especially in regions where these structures are less visually homogeneous or with a denser leaf mass. In details, the classification stage is based on a two-step PLS-DA strategy. The first PLS-DA model separates the background from basil plants using 6 latent variables (LVs), selected through cross-validation to ensure optimal model performance. Subsequently, a second PLS-DA model is applied to separate stems from leaves, using 4 LVs also chosen through cross-validation. It should be underlined that, at this preliminary stage,

the problem that discriminat classification methods [123] such as PLS-DA may encounter when faced with an imbalance between classes [139] has not been addressed. Therefore, the following result does not include any strategy (such as subsampling) to overcome this problem, as leaf pixels are more frequent than branch pixels.



(a) Ground Truth **(b)** Approach 1 Prediction

Figure 4.7 – Comparison of ground truth and WT-MIA + PLSDA prediction.

The results in Table 4.2 show the percentages of correctly classified pixels for stems, leaves and background in the test images using Approach 1. The background classification is extremely accurate, with correctness rates above 94% in most images and peaks reaching values above 99%. This result confirms the effectiveness of the algorithm in separating the background from the other elements, something that was already evident in the visual overlay of the predictions (Figure 4.7).

Leaf classification is also good, with average values above 85%. However, there are some images in which the performance falls below 80% (e.g. images 13, 14 and 15), probably due to variations in illumination that complicate the distinction of this element. Looking at figure 4.6, it can be seen that here the leaves have a shade of green very similar to that of the stems.

Stems, on the other hand, are the most problematic element to classify, with significantly lower percentages than leaves and background. The correctness for stems varies between 37.42% (image 1) and 65.70% (image 2), demonstrating a difficulty of the model in distinguishing thin, less ho-

mogeneous structures that are often covered by leaves. This limitation is clearly visible in Figure 4.7b, where the stems appear frequently interrupted or partially classified compared to ground truth (Figure 4.7a).

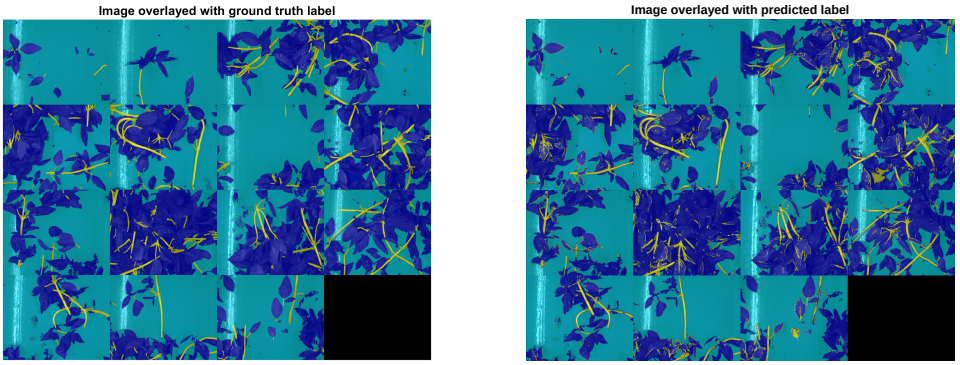
Table 4.2 – Percentages of correctly classified pixels for branches, leaves, and background in test images using the Approach 1.

Image number	Branches (%)	Leaves (%)	Background (%)
1	37.42	85.11	99.56
2	65.70	86.85	99.32
3	52.26	82.25	99.47
4	52.10	86.42	94.71
5	49.17	86.08	94.37
6	60.07	85.49	98.49
7	52.55	81.85	94.71
8	51.59	90.50	83.31
9	54.84	90.06	95.55
10	46.78	93.25	89.81
11	48.12	86.59	99.02
12	59.25	86.59	96.33
13	62.86	77.89	97.61
14	52.51	79.28	96.74
15	52.51	79.28	96.74

These results suggest that Approach 1 is particularly robust for classifying dominant features such as background and leaves, but requires further improvement to deal with the complexity of fine details such as stems. Overall, the method shows good application potential, but further optimisation could improve its performance.

4.3.2 Approach 2

Figure 4.8 show a comparison between the ground truth and the prediction obtained by the second method considered. In the visualization, the predicted regions overlap well with the actual structures of the image, indicating that the model successfully identifies the expected patterns. However, some inconsistencies may still be present, particularly some difficulties are encountered in the correct assignment of branches or leaves.



(a) Ground Truth **(b)** Approach 2 Prediction

Figure 4.8 – Comparison of ground truth and approach 2 prediction.

A quantitative metrics is integrated to the visual inspection. Table 4.3 provide the percentage of pixels correctly classified for each class.

Leaves show the highest classification accuracy, with most images showing over 90% correctly classified pixels, peaking at 95.29%. This suggests that the model is highly effective at recognising leaf structures. Similarly, the background classification performs well, with accuracy values consistently above 86%. This indicates that the model is successful in distinguishing background regions from basil plants.

However, the classification of branches shows greater variability, with values ranging from 59.45% to 97.54%. In some test images (e.g. image 7 with 59.45% and image 10 with 61.06%) the model struggles to distinguish branches from other classes.

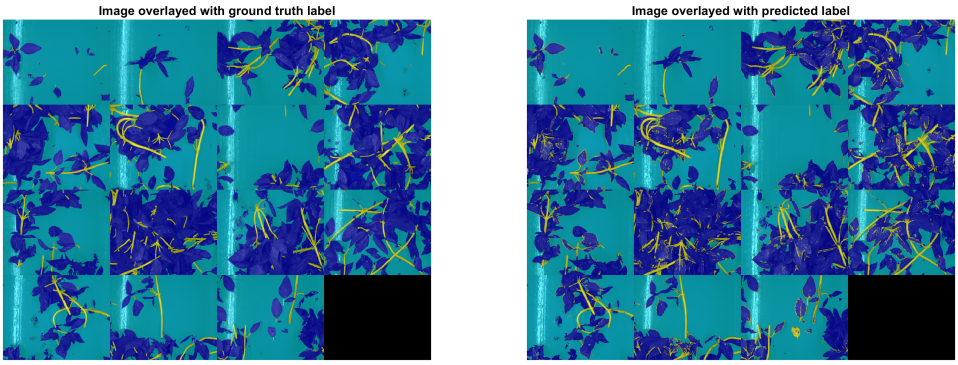
Table 4.3 – Percentages of correctly classified pixels for branches, leaves, and background in test images using the Approach 2.

Image number	Branches (%)	Leaves (%)	Background (%)
1	69.07	90.30	98.97
2	97.54	94.52	98.66
3	69.77	93.51	91.11
4	77.46	92.71	95.17
5	73.58	92.16	94.85
6	82.31	92.55	94.19
7	59.45	94.71	96.99
8	79.52	91.94	86.15
9	79.95	93.19	94.86
10	61.06	94.19	82.62
11	74.31	93.20	89.77
12	84.11	93.14	90.90
13	75.82	95.29	95.95
14	82.81	89.98	96.04
15	74.30	90.08	95.46

4.3.3 Approach 3

Figure 4.9 shows a comparison between the ground truth (a) and the model predictions (b). The results show that the deep learning model provides good overall performance for pixel-wise segmentation of the analysed images. However, there are some limitations.

The visual comparison shows that the model reproduces the ground truth in most areas. However, the boundaries between stems and leaves present a challenge, as the thin and less visually homogeneous stems are often difficult to distinguish from the overlying leaves. This problem can be seen, for example, in images 7 and 11, where the accuracy for stems is significantly lower than for other images. On the other hand, a high performance can be observed for the background.



(a) Ground Truth **(b)** Approach 1 Prediction

Figure 4.9 – Comparison of ground truth and CNN's prediction.

Table 4.4 shows the results obtained by the deep learning model for the classification of the test images. The values shown are the percentages of correctly classified pixels for three categories: stems, leaves and background.

For leaves, the accuracy is particularly high, peaking at 94.17% in image 5 and remaining above 90% in most of the images analysed. The average accuracy for this category is around 92%, which shows that the model is well trained to distinguish this class even under conditions of higher complexity.

Background is the category with the most consistent and best performance: the accuracy is consistently above 90% in all images, with an overall average for this category of around 95%, confirming that the model handles this class effectively. This is an expected result considering that the background is mostly homogeneous.

For branches, the accuracy varies between 71.40% (image 7) and 98.26% (image 2), with an overall average of around 85%. These results indicate that the model is generally effective in recognising branches, although difficulties are observed in some of the more complex images, such as images 7 and 11, due to overlapping or a similar colour with the leaves due to illumination problems.

Table 4.4 – Percentages of correctly classified pixels for branches, leaves and background in test images using the Deep Learning model.

Image number	Branches (%)	Leaves (%)	Background (%)
1	77,60	91,61	98,87
2	98,26	91,78	99,12
3	78,79	92,93	91,30
4	85,30	91,72	94,93
5	79,95	94,17	94,26
6	89,05	86,91	93,66
7	71,40	92,86	97,25
8	88,43	87,56	84,93
9	85,03	93,31	95,31
10	76,63	92,07	81,45
11	83,13	93,37	89,69
12	90,42	91,28	91,64
13	85,33	92,42	96,11
14	90,27	90,56	96,49
15	85,04	92,62	95,14

4.4 Conclusion

The aim of this work is not to define which approach is definitively the best for pixel-wise classification tasks. Although the deep learning model achieved the best performance in terms of percentage of correctly classified pixels for each class and produced results closer to the ground truth, there were significant challenges associated with this approach. Training the model required considerable effort, not only in terms of computational time, but also in terms of manual labour. A large number of images had to be manually annotated to create the ground truth. In addition, training the model required more advanced and expensive hardware (GPUs) compared to traditional machine learning models, as well as a significant amount of time; parameter optimisation alone took two days. Despite these costs, the deep learning model can effectively handle complex data, although at the

expense of interpretability.

Given that this study represents an early stage of investigation, the approaches have considerable potential for improvement. On the deep learning side, it would be possible to use pre-existing models designed for similar tasks [140, 141] and apply fine-tuning [142]. However, this was not possible in this case as such models are not available in MATLAB. Another potential improvement could be to move to next-generation AI methods, such as transformer-based architectures [143], which were originally developed for large language models (LLMs) [144] but are now being applied to image analysis tasks [145]. However, these models are inherently more complex, requiring a much larger calibration set and significantly more computational resources to train.

On the other hand, classification models based on discriminant analysis, such as PLS-DA, face challenges such as class imbalance. Effective strategies need to be developed to address this issue. Current solutions, such as subsampling or class weighting, have proven to be inadequate. One possible solution could be to modify the classification criteria, for example by applying LDA or QDA to the $\mathbf{X}$ or $\mathbf{Y}$ scores. Alternatively, selecting a more representative calibration image set could help improve model performance.

In summary, deep learning offers superior performance, especially in the detection of branches, but at a high cost in terms of interpretability and computational resources. In contrast, chemometrics approach offers lower performance but is more cost-effective and retains interpretability, making it a viable alternative depending on the specific requirements of the task. For instance, a possible strategy could involve combining these features in the context of process monitoring or real-time quality control, as presented in Chapter 3. In terms of prediction speed, while both the Deep Learning approach and WT-MIA + PLS-DA process each image in under a second, the approach designed to emulate CNNs requires several seconds per image.

Based on the results, it would then be possible to evaluate which approach offers the best compromise for the company, considering how well each fits the expectations and the cost of implementation.

III

Real-time quality control: Multi block non linearities and outliers

Chapter 5

Real-Time Quality Control: Benefits of Multimodal Sensor Fusion and Nonlinear Modeling

This chapter is based on the article listed below, with selected content and results.

1. Strani, L., Vitale, R., Tanzilli, D., Bonacini, F., Perolo, A., Mantovani, E., Cocchi, M. (2022). A multiblock approach to fuse process and near-infrared sensors for on-line prediction of polymer properties. *Sensors*, 22(4), 1436.
2. Tanzilli, D., Strani, L., Bonacini, F., Ferrando, A., Cocchi, M., Durante, C. (2024). Implementing Multiblock Techniques in a full-scale plant scenario: On-line Prediction of Quality Parameters in a Continuous Process for Different Acrylonitrile Butadiene Styrene (ABS) Products. *Analytica Chimica Acta*, 342851.

5.1 Introduction

Companies want to be able to monitor the quality of the end product in real time so that they can intervene quickly in the event of a fault, thus reducing chemical waste and manpower costs associated with the numerous laboratory analyses that are carried out on a daily basis. In various sectors (precision agriculture, pharmaceuticals, food, chemicals), it is common practice to use analytical sensors for the complete characterisation and continuous monitoring of processes [146]. As mentioned in section 2.4, the analysis of the data provided by these sensors requires advanced statistical tools to extract physic-chemical information useful for process control, such as Multivariate Statistical Process Control (MSPC), in particular Latent Variable (LV) approaches [17, 18, 147–150]. In general, LV MSPC tools rely on so-called 'engineering process variables' [77] collected from on-line sensors (temperature, flow, pressure) to build multivariate reference models characterising normal operating conditions (NOC). Recent technological advances have also favoured the extensive use of spectroscopic probes, in particular near-infrared (NIR) spectroscopy [149–154], to monitor the ongoing process or predict intermediate and final quality parameters [151, 153–155].

It is easy to deduce that in modern production scenarios, data often come from multiple sources [155–159]. For example, a single sample may be analysed by different instruments to obtain a more complete view of its properties (as is the case in raw material characterisation) or, in industrial lines, several sensors measuring temperature, pressure, flow rates, etc. are installed at different stages of a production process to capture its temporal evolution [155]. Consequently, the data sets are not only multivariate but also multi-source [159]. An example is when two different spectroscopic techniques are used, such as NIR and ultraviolet-visible (UV-Vis): each provides a multivariate profile (acquired at different wavelengths), and these therefore represent two separate data sources [160]. Multi-source data can also result from different operating conditions, for example when data is collected from different process batches run with different parameters [159, 161, 162]. Or, as mentioned earlier, product quality can be monitored on-line by

spectroscopic measurements. Illustrative examples can be found in the food [163, 164], pharmaceutical [165, 166] and chemical [6, 167] industries. By fusing data from multiple sensors, ongoing processes can be monitored and product quality attributes, normally assessed by off-line laboratory analysis, can be predicted in real time [149, 155, 168, 169].

Handling complex, multi-source, multivariate datasets without appropriate chemometric methods can lead to suboptimal data interpretation [170]. In this regard, multiblock data analysis techniques can be highly beneficial as they draw on complementary information from multiple data sources [170, 171]. These techniques allow for deeper insights and better data visualisation, improve predictive accuracy and facilitate the identification of critical variables that most influence models [170–175].

We therefore investigated multiblock chemometric techniques [6, 167, 168, 170–174] designed to perform low-level data fusion [175, 176], which can offer distinct advantages over medium- or high-level integration approaches [175], particularly in terms of model training, maintenance and interpretability. By using the original variables directly, without any compression steps, it is possible to assess the importance of each block/type of sensor in the model, i.e. to check their degree of uniqueness or redundancy. In general, data fusion approaches, particularly those based on low-level strategies, could improve prediction accuracy and facilitate the identification of critical variables.

In the predictive context, Multiblock Partial Least Squares (MB-PLS) [176, 177] was first proposed and is one of the most widely used. This prevalence is largely due to its simplicity and integration with numerous instrumentation and statistical software platforms. However, several other methods have been developed that focus more on interpreting the role of the different blocks [171], such as highlighting the common [178, 179] and/or specific information carried by each data block [180, 181].

Sequential methods, such as Sequential-Orthogonalised Partial Least Squares (SO-PLS) [172] or Response-Oriented Sequential Alternation (ROSA) [182], extract non-redundant information that is most relevant for predic-

tion from each different data block analysed. The distinguishing features of ROSA, which is also based on PLS regression, are: (i) being invariant to block scaling and not affected by the spurious bias resulting from the combination of data blocks of different sizes (similar to sequential orthogonal PLS - SO-PLS); and (ii) being computationally efficient and capable of dealing with any number of blocks, even a very large number (unlike SO-PLS).

The first phase of the work focused on real-time ABS quality estimation in a continuous production plant with sequential product formulations to produce different product types. This application is an excellent proof of concept for real-time prediction on an industrial scale in a multiblock context, as the plant has numerous sensors, including four NIR sensors, and different process stages.

Nevertheless, continuous large-scale industrial processes can be particularly complex, not only because of the many different sensors, but also because multiple products can be produced on the same production line at different times by changing operational settings and formulations without stopping production. Adapting to new conditions can temporarily produce inferior products, while the different characteristics of each product add another layer of variability that can affect prediction accuracy. In addition, the range of parameters to be predicted can change significantly. Training a unique model for all product could be sub-optimal; for example, attempting to predict the quality of a sporadically produced product may prove inaccurate due to the lack of information on the evolution of the process over time. In such cases, local regression methods can help to improve model robustness by focusing on local data characteristics rather than assuming a single global relationship, thereby incorporating real-time process evolution information. This often provides a more flexible representation of complex patterns [183–185].

To date, no method has been proposed that integrates a multivariate local regression approach with a multiblock framework. To bridge this gap, a novel multiblock extension of the local regression method Locally-Weighted Partial Least Squares (LW-PLS) [183], here called Locally-Weighted Multiblock Partial Least Squares (LW-MB-PLS), has been developed.

In conclusion, the multiblock strategies resulted useful for the estimation of quality characteristics. The ROSA method demonstrated strong predictive performance for their quality parameters, identifying critical data blocks and highlighting the role of both early and late process sensors. Although LW-MB-PLS provided similar predictive results, it offered tangible benefits in mitigating systematic prediction errors on specific products, suggesting a distinct advantage in highly variable and non-linear industrial processes.

Furthermore, for each predicted sample, the block specific contribution (explained block variance and significant VIPs) was explored to exploit the local nature of LW-MB-PLS.

5.2 Data

5.2.1 Process Description

The data collection was carried out in a continuous industrial production plant for Acrylonitrile Butadiene Styrene (ABS), owned by Versalis (ENI Group). The plant produces nine different types of ABS, which differ slightly in formulation and/or operating conditions. These products are referred to as "Product 1-9". The process can be described in terms of six different stages, as shown in Figure 5.1. In the first, called "PRE-REACTION", the monomers, styrene, butadiene and acrylonitrile, are mixed together. In the next three, called "REACTION 1", "REACTION 2" and "REACTION 3", the monomers react and begin to form the ABS polymer. In the last two sections, called "DEVO" and "END", the residual monomers are removed and the final product is cut. Each section has between 5 and 30 process sensors (PS) that measure temperatures, flow rates, pressures and motor speeds, for a total of 118 sensors. Four NIR probes are also installed along the process line. The first two, called "NIR PRE-1" and "NIR PRE-2", are placed at the beginning and end of the PRE REACTION section, before the reaction takes place, to monitor the reagents before and after their mixing. The third NIR probe is placed between the REACTION 1 and REACTION 2 sections to monitor the state of the reaction. Finally, a fourth NIR probe

is placed in the END section, at the very end of the process, just before the product is cut. A schematic representation of the production line is shown in figure 5.1. In the current work, process and NIR sensor data collected on-line from this plant during the period January 2020 to April 2022, as well as quality data collected off-line and analysed by the company laboratory during the same period, have been considered and analysed.

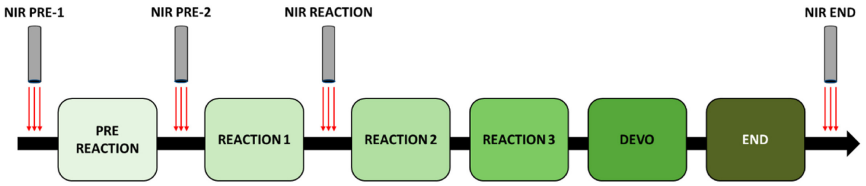


Figure 5.1 – Schematic diagram of the ABS production line. The green blocks represent the six different sections into which the PS has been divided, while the grey bars and red arrows represent the positions where the four on-line NIR probes have been placed.

5.2.2 Reference Analysis

Due to confidentiality agreements with the company, the specific names of the two ABS quality parameters assessed cannot be disclosed and are referred to as QP1 (Property 1) and QP2 (Property 2) respectively. Both parameters are assessed offline by taking samples of the final product. Specifically, QP1/Property 1 is measured three times a day, while QP2/Property 2 is measured twice a day.

Parameter QP1 (Property 1) provides information on the physical characteristics of the product, with reference values expressed in grams, and is related to the rheological behaviour of the polymer. Parameter QP2 (Property 2), on the other hand, determines the impact resistance of the product and is expressed in Joule.

For each of these two parameters, the company defines threshold values (minimum and maximum) above which the quality of the product is considered inferior and is sold at a reduced price. In the course of the study, 2184 tests were carried out on QP1 and 1349 tests were carried out on QP2. The values for QP1 (log-transformed in some tests) ranged from 1.6 to 11.1 g,

while those for QP2 ranged from 4.1 to 38.9 J.

5.2.3 NIR measurements

Spectra were collected on-line from the four distinct acquisition points using a Matrix FT-NIR spectrometer (Bruker Optics, Milan, Italy). The instrument was equipped with optical fibers (length of 100 m and a diameter of 600 μm). These fibers were linked directly to the acquisition sites on the process pipe through HT immersion probes (Drawing-no. 661.2350_1, Hellma GmbH and Co. KG, Müllheim, Germany). The acquisition was conducted in transmission mode, spanning the spectral range of 12500 to 4000 cm^{-1}, with a nominal resolution of 4 cm^{-1} (64 scans per sample).

5.3 Data Analysis

The collected data was organised into ten different data blocks, categorised by data type and area of collection (see Table 5.1). Specifically, PS measurements were organised into five data blocks, each corresponding to a specific area of the plant. On the other hand, the NIR spectra were divided into four blocks, each associated with a single optical probe. Figure 5.1 shows the names and abbreviations (which will be used hereafter) of all the blocks, together with their respective positions within the plant. This also serves as an indication of their temporal sequence, given the continuous nature of the process.

Table 5.1 – Data block description

Block Full Name	Block Abbreviated Name	Data Type	No. of Variables[1]	Order
NIR Pre Reaction 1	NP1	NIR Spectra	390	1
Pre Reaction	PR	PS	7	2
NIR Pre Reaction 2	NP2	NIR Spectra	390	3
Reaction 1	R1	PS	15	4
NIR Reaction	NR	NIR Spectra	390	5
Reaction 2	R2	PS	10	6
Reaction 3	R3	PS	8	7
Devolatilizer/End zone	DE	PS	30	8
NIR End zone	NE	NIR Spectra	390	9

[1] For NIR data blocks, the number of variables is equal to the number of spectral wave numbers; for PS data blocks, it is the number of process variables collected in the respective plant area.

5.3.1 Data synchronization

For each multiblock technique used, the data blocks used for analysis were constructed following the chronological progression of the ABS production process, taking into account the placement of the various sensors throughout the production line. In simple terms, each data point within the datasets corresponds to information collected at different times, but is accurately associated with the same processed material, ensuring data synchronisation. The time delay between the different sections of the plant, which indicates the time taken for the same material to pass from one section to another, has been determined using the flow rate values derived from the pumps installed throughout the plant. These specific PS provide information on the flow of material ($Kg\,h^{-1}$) passing through a reactor or tank. By knowing their volumes and ensuring that they are consistently full, it is possible to approximate the time taken for the material to pass from one section to another.

5.3.2 Preprocessing

Each block of data was pre-processed differently. Specifically, each PS data block was scaled so that all variables had unit variance, taking into account their different nature and scales. While in each NIR data block

the spectra were trimmed to consider only the spectral range from 6500 to 5000 cm^{-1}, which shows spectral bands attributable to either reactants or products, and then treated with Standard Normal Variate (SNV) [81] for the analysis of QP1 and with Savitzky-Golay First Derivative [81] (1D) using a 15 point window for the analysis of QP2.

After the individual block pre-processing, since MB-PLS and the proposed methods LW-MB-PLS are sensitive to block variance, due to the row concatenation of the blocks, each dataset was scaled to unit block variance (including column mean centering). While ROSA, as explained in section 5.3.4, doesn't require the row concatenation and the blocks were just mean centred.

5.3.3 Multiblock Partial Least Squares

This section explains the principles of the MB-PLS implementation originally proposed by Westerhuis and Coenegracht [176], which can be considered as a standard PLS with appropriate block scaling steps as described in [177]. Thus, MB-PLS is an extension of classical PLS regression [84] for applications involving different blocks of data that share the same number of rows (observations), with respect to the data matrix $\mathbf{X}$, resulting from the row-wise concatenation of different data blocks.

$$\mathbf{X} = [\,\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B\,] \tag{5.1}$$

This method provides both global parameters (also called *super*), such as scores, weights, loadings, and regression coefficients, and local parameters (referred to as *block*), such as scores and weights for each data block, as described by Equations 5.3 – 5.5:

$$\mathbf{w}_b = \mathbf{X}_b^{\mathrm{T}} \cdot \mathbf{u}(\mathbf{u}^{\mathrm{T}}\mathbf{u})^{-1} \tag{5.2}$$

$$\mathbf{t}_b = (\mathbf{X}_b \cdot \mathbf{w}_b)(\sqrt{V_b})^{-1} \tag{5.3}$$

$$\mathbf{w} = \mathbf{T}^{\mathrm{T}} \cdot \mathbf{u}(\mathbf{u}^{\mathrm{T}}\mathbf{u})^{-1} \tag{5.4}$$

$$\mathbf{t} = \mathbf{T} \cdot \mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{w})^{-1} \tag{5.5}$$

where V_b represents the number of variables in a specific data block, and $\mathbf{t}_b$ and $\mathbf{w}_b$ refer to a single block scores and weights, respectively. Conversely, $\mathbf{t}$ and $\mathbf{w}$ denote the global (*super*) scores and weights. The matrix $\mathbf{T}$ is obtained by concatenating all $\mathbf{t}_b$ vectors.

This approach enables the assessment of each data block's contribution (by analyzing $\mathbf{w}_b$) to the prediction of the response variable(s), $\mathbf{y}$ or $\mathbf{Y}$, improving process understanding.

5.3.4 Response-Oriented Sequential Alternation

Response-Oriented Sequential Alternation (ROSA) is a multiblock regression approach introduced by Liland et al [182] based on Partial Least Squares (PLS) regression. ROSA operates as a sequential algorithm, computing one PLS component at a time from a single block, thus the method is invariant to block scaling (blocks are only mean-centred), which distinguishes it from MB-PLS, and also to block ordering, which distinguishes it from other sequential multiblock methods [172]. These features allow ROSA to handle many blocks of different dimensions. In addition, ROSA has high computational efficiency. In fact, it bypasses the need for iterative convergence in the optimisation criteria, and it deflates only the response variable rather than all blocks.

Specifically, each PLS component is selected from a single block by selecting the block that provides a single PLS component with the smallest prediction residuals relative to the other candidate blocks. Subsequent components are constrained to be orthogonal to the subspace spanned by the

previously selected components, ensuring orthogonality in scores and loadings. The main steps of the ROSA algorithm are described by the following equations:

$$\mathbf{w}_b = \mathbf{X}_b^{\mathrm{T}} \mathbf{y} \tag{5.6}$$

$$\mathbf{t}_b = \mathbf{X}_b \mathbf{w}_b \left(\mathbf{X}_b \mathbf{w}_b\right)^{-1} \tag{5.7}$$

$$\mathbf{r}_b = \mathbf{y} - \mathbf{t}_b (\mathbf{t}_b^{\mathrm{T}} \mathbf{y}) \tag{5.8}$$

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

$$\mathbf{q}_a = \mathbf{y}^{\mathrm{T}} \mathbf{t}_a \tag{5.9}$$

$\mathbf{t}_a$ are the previously selected scores for the a-th LV. For the calculation of subsequent LVs, steps 5.6 to 5.9 are repeated, updating $\mathbf{y}$ with $\mathbf{y}$-residual relative to the winning block ($\mathbf{r}_{b,\mathrm{winning}}$).

The $\mathbf{X}$-loadings ($\mathbf{P}$) and PLS regression coefficients ($\mathbf{b}$) (potentially including a constant term b_0) can be computed using Equations 5.10–5.12, once the optimal number of LVs has been determined, and the corresponding scores, weights, and $\mathbf{y}$-loadings are gathered in matrices $\mathbf{T}$, $\mathbf{W}$, and $\mathbf{q}$.

$$\mathbf{P} = \mathbf{X}^{\mathrm{T}}\mathbf{T} \tag{5.10}$$

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{q} \tag{5.11}$$

$$b_0 = y_m - \mathbf{x}_m^{\mathrm{T}}\mathbf{b} \tag{5.12}$$

Here, $\mathbf{x}_m$ is a vector containing the mean of every variable of $\mathbf{X}$, whereas y_m is the mean of $\mathbf{y}$. In ROSA, every chosen LV carries information exclusively from the winning $\mathbf{t}_{b,\text{winning}}$-block (the one with the smallest residuals as per Equation 5.8), and all LVs are orthogonal. It is crucial to emphasize that all blocks are always considered as candidates at every step of the algorithm. Consequently, successive LVs may contain information from the same previously chosen block or from a different one.

5.3.5 Locally Weighted Multiblock Partial Least Squares regression

The Locally Weighted Partial Least Squares (LW-PLS) method [183, 186] extends the traditional PLS approach to achieve accurate predictions even when dealing with complex data structures such as clusters and nonlinear relationships [186–188] between independent ($\mathbf{X}$) and dependent ($\mathbf{Y}$) variables. In this study, we have considered a K-Nearest Neighbours Locally Weighted (KNN-LW) strategy [183] to develop a multiblock. For a single dataset (single block), this approach involves identifying the K nearest neighbours from the calibration set for each new observation to be predicted. These neighbours are weighted using a function [183] that takes into account a dissimilarity measure ($\mathbf{d}_i$), such as the Euclidean or Mahalanobis distance, between the selected K neighbours and the observation to be predicted. The weighting function is defined as:

$$w_{local,i} = \exp\left(-\frac{d_i^*}{h \cdot \sigma(\mathbf{d}^*)}\right) \tag{5.13}$$

where d_i represents the normalized dissimilarity of the i-th neighbour (among the K nearest ones), $\sigma(\mathbf{d}^*)$ is the standard deviation of the vector $\mathbf{d}^*$ (containing the dissimilarity values for all K nearest neighbour), and h is a parameter that controls the shape of the weighting function. A higher h value reduces the impact of dissimilarities on the weights.

After the weights are determined, a local PLS model is built. In this model, each calibration sample among the neighbouring observations is assigned a specific weight as defined by equation 5.13. Both $\mathbf{X}$ and $\mathbf{Y}$ are weighted-centered, and, similar to standard PLS, the covariance between the $\mathbf{X}$-scores and $\mathbf{Y}$-scores is maximized while ensuring that the scores of different components remain orthogonal. The locally weighted PLS model follows these equations:

$$\mathrm{Cov}(\mathbf{t}_a, \mathbf{u}_a) = \mathbf{t}_a^\top \mathbf{D} \mathbf{u}_a \tag{5.14}$$

$$\mathbf{t}_a^\top \mathbf{D} \mathbf{t}_{a+1} = \mathbf{u}_a^\top \mathbf{D} \mathbf{u}_{a+1} = 0 \tag{5.15}$$

Here, $\mathbf{t}_a$ and $\mathbf{u}_a$ are the $\mathbf{X}$-scores and $\mathbf{Y}$-scores vectors for the a-th latent variable (LV), respectively, and the diagonal matrix $\mathbf{D}$ contains the local weights for each sample.

For regression modelling, the predictions for new samples ($\hat{\mathbf{Y}}$) are calculated using the equation:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B} \tag{5.16}$$

where $\mathbf{B}$ contains the regression coefficients.

In this work, we propose a straightforward implementation to extend this locally weighted approach to the multiblock case, i.e. the development of Locally Weighted Multiblock Partial Least Squares Regression (LW-MB-PLS), while retaining the core of both methods and computational efficiency. The proposed algorithm first performs a low-level data fusion of all blocks

by concatenation and applying block scaling. This ensures that a single data block does not dominate the others simply because of a larger number of variables. The local weighting scheme is then applied to the fused data set. This ensures a unique set of neighbours for each new sample to be predicted, and a single set of weights to be optimised by tuning the h parameter. A possible downside of this unique selection could be that, hypothetically, if the neighbours for a given sample were computed independently for each block of data, they might not necessarily be the same, resulting in a suboptimal local model. However, optimising both the weights and the number of neighbours could compensate for this, providing good predictive performance while maintaining a simpler model.

Tuning of model parameters

The various parameters, such as the number of latent variables (A), the number of nearest neighbours (K), and the shape factor (h), were optimized through Cross-Validation. The set of values explored for each parameter are reported in Table 5.2, and all possible combinations were tested (Grid-Search). The optimal values were then established by inspection of the corresponding Mage plot [181]. This plot is employed to identify the optimal combination of factors (i.e. those yielding the lowest prediction error) for the input blocks [189].

Table 5.2 – Parameters considered for optimization in Cross-Validation with their respective tested values.

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

Evaluating the block salience

In order to assess the contribution of each block to the LW-MB-PLS model, we calculated the explained variance for each block, analogous to

that proposed by Westerhuis et al [177]. In addition, VIP values for each block (VIP_b) were obtained by summing the VIP values of the variables belonging to the block. In this case, a significance threshold equal to the number of variables in a block was used for VIP_b, taking into account the threshold of one usually set for each individual variable. However, since a specific model is used for each sample to be predicted, both parameters have a different value per block and per sample, which makes it possible to study if and how the local models vary when different quality products are considered.

5.3.6 Model building

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

The reliability of the prediction models was assessed using the root mean square error in prediction (RMSEP) and compared with the root mean square error in cross-validation (RMSECV). The CV-ANOVA approach [190] was used to assess which models gave significantly different RMSECV and RMSEP. This was done in two ways: i) comparing models obtained using the

same technique but computed with different blocks used for model building, and ii) comparing models obtained using different techniques but computed with the same blocks used for model building. This approach made it possible to investigate the importance of both the prediction method used and the different blocks of initial data used.

5.4 Results and discussion

The following section present the results of the prediction models generated by both the ROSA and LW-MB-PLS methods using different numbers of blocks. The ROSA results are examined first, followed by the LW-MB-PLS results. The MB-PLS results are commented on only briefly to highlight the differences between the ROSA and LW-MB-PLS methods. For consistency with the following chapters, where only QP1 is considered, the discussion of QP2 results is omitted in this section, and reference is made to Paper 3 in the Appendix for further details.

5.4.1 ROSA results

The initial ROSA prediction models were constructed using all available blocks of data. Based on the RMSECV values (up to 20 LVs were examined), 13 LVs were selected for the QP1 model. As explained in section 5.3.4, the ROSA algorithm identifies a 'winning block' for each LV, revealing the most influential sections of the plant for predicting the parameters of interest.

Figure 5.2 illustrates the frequency and order in which blocks were selected by the algorithm across the LVs during QP1 model development. Notably, no NIR blocks were selected. Instead, the DEVO-END block was selected 10 times out of 13, with the remaining selections being REACTION 2 (once, at the fifth LV) and REACTION 3 (twice, at the first and tenth LVs), indicating that QP1 estimation relies heavily on PS data from the final stages of the process.

The resulting model achieved a RMSEP of 0.74 g. Figure 5.3a shows

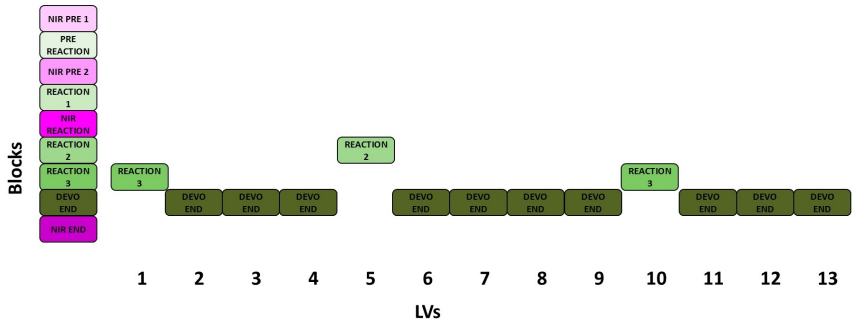


Figure 5.2 – ROSA model (using all the available blocks) for QP1 prediction. The winning block selected for each LVs is shown in correspondence of the component number. The left bar reports the time order of the blocks along the process.

a uniform distribution of predictions for the validation set, with all objects falling within the expected QP1 range. Figure 5.3b shows the same distribution, colour coded to represent eight different ABS products. Each product occupies a different range of QP1 values. For example, product 7 has QP1 values around 10g, higher than the other products. Conversely, Product 1, the most commonly produced, typically shows values between 4 and 6 g, although a few samples are spread across the entire QP1 range.

Figure 5.4 shows the PLS regression coefficients for the three blocks selected by ROSA. Variables with VIP scores greater than one are marked with red diamonds, indicating significant contributions. Almost all PS variables in each block exceed this threshold, except for a few in REACTION 2 (variables 1, 2, 6 and 7). Specific sensor names are withheld due to confidentiality agreements with the company. Certain sensors, such as 5, 9, 11 and 12 in REACTION 2, have significantly higher regression coefficients than others in the same block. Similar trends are observed in "REACTION 3" and "DEVO-END". In general, variables with high VIP scores but low absolute regression coefficients are influential for only a few LVs, suggesting limited overall significance. This information can help plant operators to identify critical sensors to monitor to ensure that the final product meets QP1 thresholds. Uncontrolled variation in these sensors could have a significant impact on product quality.

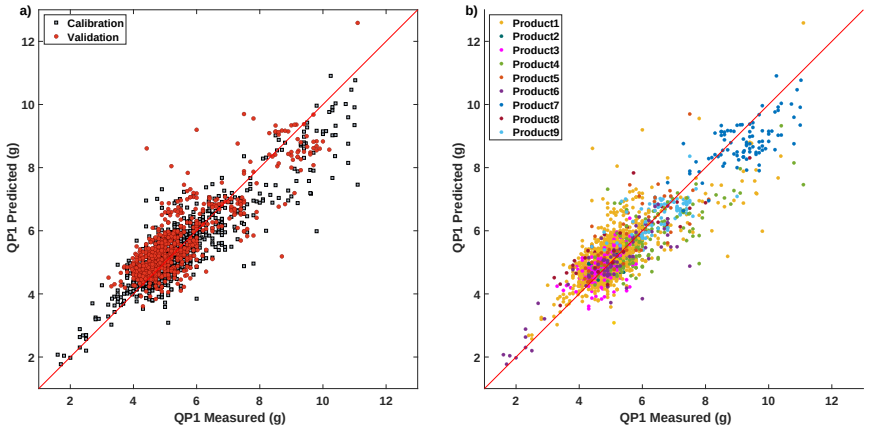


Figure 5.3 – Plots of predicted vs measured values of QP1 obtained by the ROSA model using all the available blocks. In (**a**) Samples are colored according to calibration (gray) and validation (red) and in (**b**) according to ABS product type.

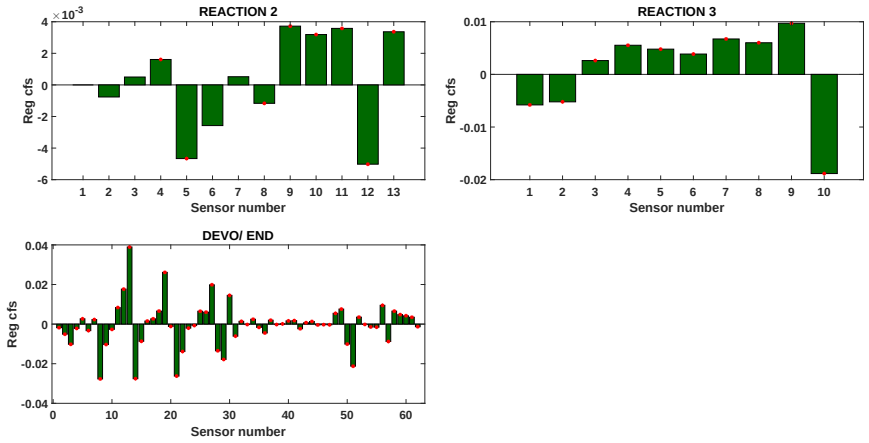


Figure 5.4 – Regression coefficients for the selected block. The red diamonds indicate variables with VIP scores exceeding one.

While these results demonstrate strong predictive performance, two key aspects warrant further investigation: (i) the possibility of achieving accurate predictions before the product is finalised, and (ii) whether reliable predictions can be achieved using only spectral or process sensor data. To address these questions, additional ROSA models were developed using subsets of the data: blocks prior to the END zone, PS data only, and NIR data only (with and without NIR END spectra). The results, summarised in Table 5.3, show that models using all available blocks gave the lowest RMSEP

values. Reducing the number of blocks significantly increased the prediction error ($p < 0.05$), as expected due to the loss of critical information, particularly from late stage process data. Notably, models using only NIR blocks performed the worst, confirming that QP1 is not strictly correlated with ABS chemical composition, but rather reflects performance characteristics assessed through mechanical and physical testing. These parameters are more sensitive to processing variations, which can cause significant changes independent of chemical composition.

Table 5.3 – Results obtained by applying ROSA.

Blocks used for model building[a]	**LVs**	**RMSECV (g)**	**RMSEP (g)**
NP, PR, NP2, R1, NR,**R2**, **R3**, **DE**, NE[b]	13	0.55^{a}	0.74^{a}
NP1, **PR**, NP2, **R1**, NR, **R2**, **R3**	6	0.72^{b}	0.81^{b}
NP1, **PR**, **NP2**, **R1**, NR, **R2**	10	0.75^{bc}	0.83^{bc}
NP1, **PR**, NP2, **R1**, NR	8	0.83^{c}	0.86^{cd}
PR, R1, **R2**, **R3**, **DE**	13	0.55^{a}	0.74^{a}
PR, **R1**, **R2**, **R3**	6	0.72^{b}	0.81^{b}
PR, **R1**, **R2**	10	0.74^{bc}	0.85^{cd}
PR, **R1**	8	0.83^{c}	0.86^{cd}
NP1, **NP2**, NR, **NE**	11	0.84^{c}	0.89^{d}
NP1, NP2, **NR**	7	1.00^{d}	1.19^{e}
NP1, **NP2**	6	1.12^{e}	1.27^{h}

In a column, values with the same letter are not statistically different between each other ($p > 0.05$).
[a]D = DEVO, DE = DEVO-END, E = END, NE = NIR END, NP1 = NIR PRE 1, NP2 = NIR PRE 2, NR = NIR REACTION, PR = PRE REACTION, R1 = REACTION 1, R2 = REACTION 2, R3 = REACTION 3.

Models excluding final stage blocks showed higher RMSEP values, but remained acceptable to process operators, demonstrating the feasibility of obtaining reasonable estimates prior to the completion of the ABS production process. This opens up two promising avenues for real-time prediction and control: (i) using both spectral and process sensor data to comprehensively monitor critical process areas and sensors across the plant, or (ii) focusing solely on PS data to streamline data management and reduce noise interference. Both approaches have significant industrial relevance. Accurate quality predictions can reduce reliance on off-line analysis, saving labour and minimising waste. Simpler data management improves interpretability, enabling all plant operators to use the results effectively.

5.4.2 LW-MB-PLS results

Data analysis using the LW-MB-PLS method followed the same procedure as the ROSA method. The first model examined was the one built with all available blocks and. The results obtained are shown in Figure 5.5, it is possible to note that in this case, even if the prediction is more spread around the fit line (red line), in particular in the range 8-10 g, the trend in the products are more aligned with it. A possible explanation could be that there are few calibration points in this range, so the sub-calibration set may be suboptimal due to the choice of point in a different range.

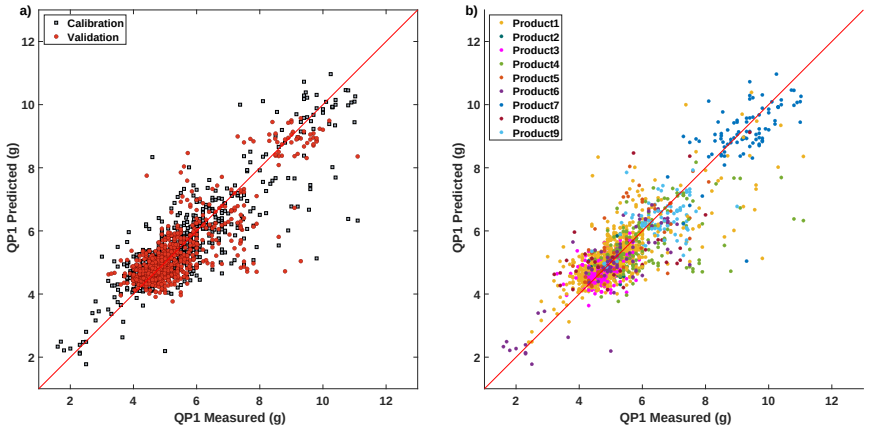


Figure 5.5 – Plots of predicted vs measured values of QP1 obtained by the LW-MB-PLS model using all the available blocks. In (**a**) Samples are colored according to calibration (gray) and validation (red) and in (**b**) according to ABS product type.

Figure 5.6 shows the Måge plot used to assess which combination of h and k parameters gave the lowest RMSECV for a given LV. Combinations that gave very high RMSECV values were not included to improve the clarity of the figure. It can be seen that on the Pareto front only combinations with h covering the higher values tested (1-4) are present, while almost all k values are present (except the smallest value of 100) and there is no interaction between h and k (similar low RMSECV values are obtained with either small or high k, regardless of the value of h), and therefore the most influential parameter is the number of LVs, in fact there is a clear increase in error after 3 LVs. Among the combinations giving a similar RMSECV

value, the most parsimonious was chosen, both in terms of the number of LVs and the number of neighbours. The minimum RMSECV corresponds to the following settings: one LV, 300 neighbours (k) and an h value of 1.

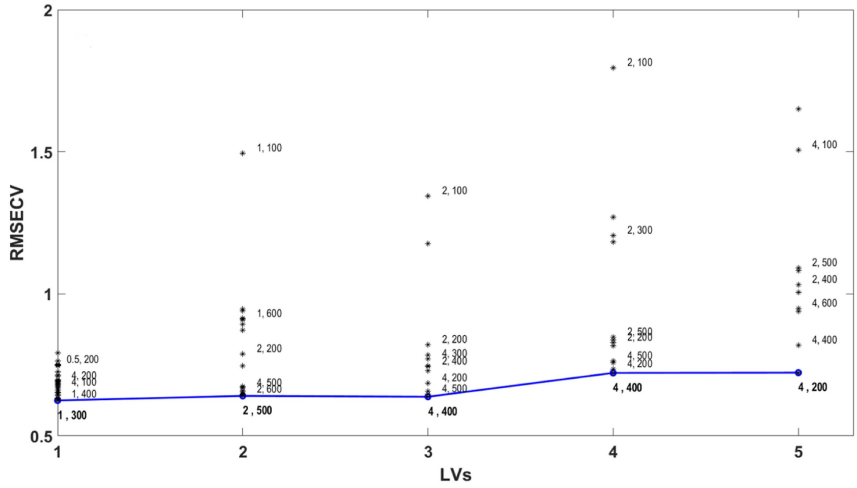


Figure 5.6 – Måge plot for the LW-MB-PLS QP1 model. The point label report first the value of h, then that of k. The points on the Pareto front have labels in bold.

In addition, the neighbours selected by the algorithm using the Euclidean distance in PCA space for a given sample (exploring several different samples) were examined to assess whether the neighbourhood included samples with the same type of product or products with similar QP1 values. As an example, Figure 5.7 shows a plot of all QP1 values obtained in the laboratory (reference method) versus production time, where the 300 neighbours of validation sample number 544 (black filled triangle from 7 March 2022) are represented by the filled squares. It can be seen that a significant proportion of the samples selected by the algorithm as neighbours belong to the same ABS product category as the sample to be predicted. Similarly, the majority of the QP1 values are close to the QP1 value of the sample to be predicted. This means that the selected sample is predicted almost exclusively by samples that are similar to it, without taking into account very different samples that could negatively influence the prediction performance.

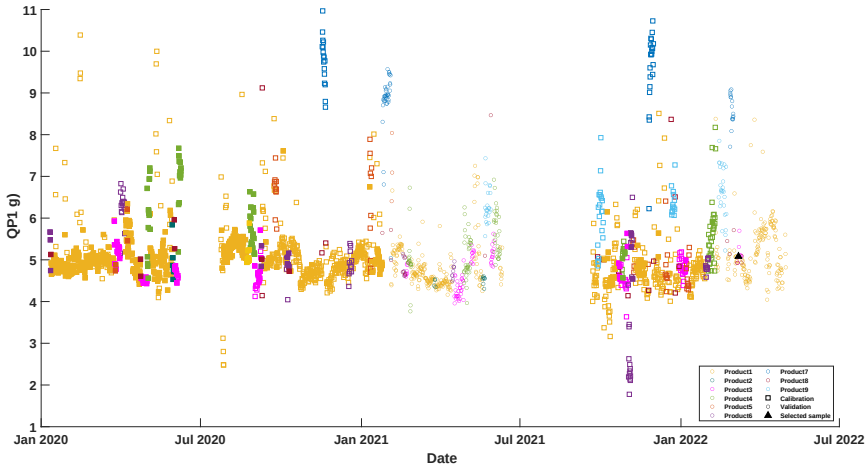
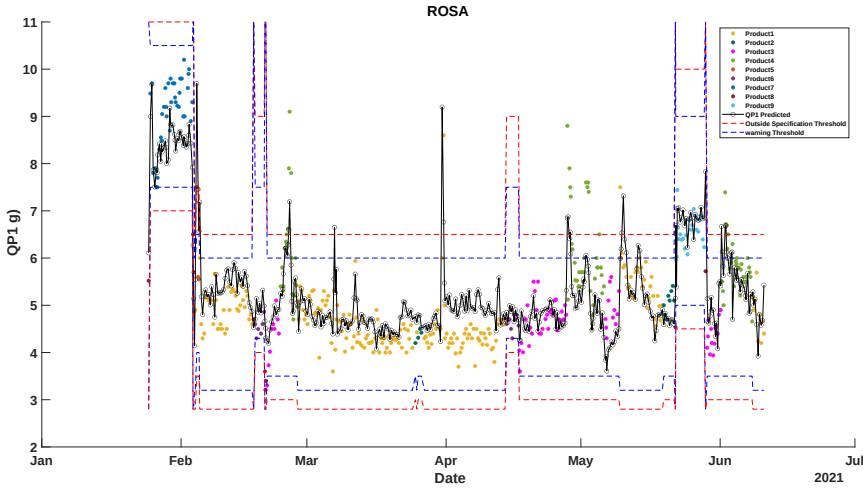


Figure 5.7 – QP1 values vs time, coloured by ABS product. Square refer to calibration samples, whereas circles refer to validation samples. The samples represented by the filled square denote the selected neighbours to build the predictive model for the sample depicted by the black triangle (which belong to Product 1 type). Non-filled symbols represent samples that have not been selected by the model as neighbours.

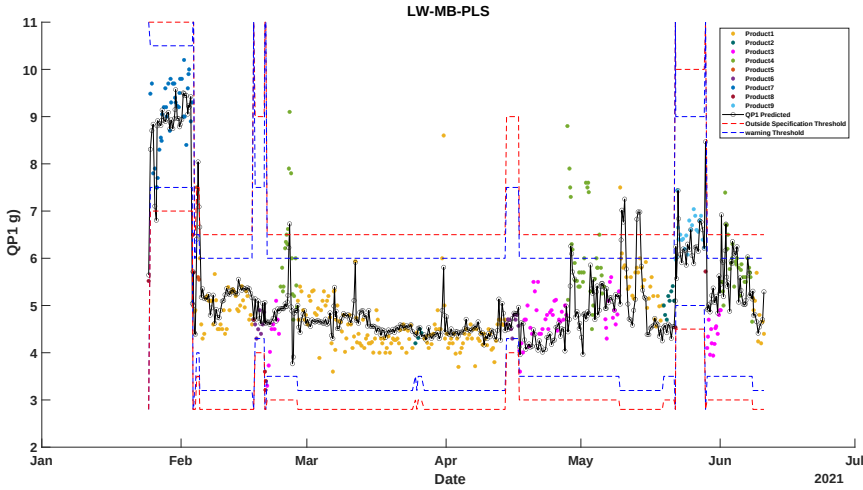
The model obtained in this way showed an RMSEP of 0.75 g. The prediction trend, in the form of a control chart, can be observed in Figure 5.8b. Here, the predicted values of QP1 obtained by the LW-MB-PLS model using all available blocks are represented by the black unfilled circles, while the filled circles, coloured according to the different product types, represent the QP1 reference values obtained from the off-line laboratory analysis. The figure specifically shows data from January to June 2021, which corresponds to the data included in the validation set. The model's predicted values cover a range very close to that of the validation set, following the production changes. In fact, even in the case of a formulation change or a shift in plant operating parameters, the model follows the trend of the predictions well. The blue and red dashed lines represent the two thresholds set by the company to assess whether a product is out of specification or not. In particular, the blue line represents a warning value, where the product is still considered to be of high quality, but close to the out of specification threshold represented by the red line. Obviously, these thresholds vary according to the different ABS products. The predictions follow the trend of the reference analysis when they are above the thresholds, even if it sometimes seems that

the model underestimates some of these values.

For comparison purposes, the prediction versus time for QP1 obtained by the ROSA model, using all the blocks available for modelling, is shown in Figure 5.8. The general trend is similar, but for some products and periods there is evidence of systematic errors, even if the unit is within the warning thresholds and therefore acceptable. The only exception is April, where the sample for product 1 is well predicted by LW-MB-PLS (Figure 5.8b) but not by ROSA (Figure 5.8a).



(a) Measured off-line values and predicted values using the ROSA model, which employed all the available data blocks.



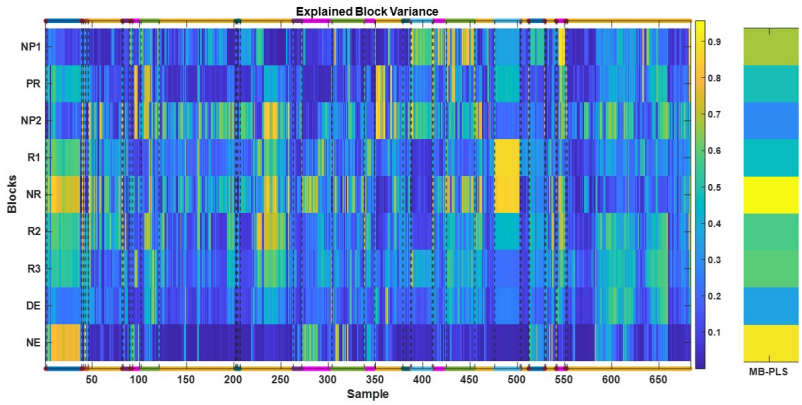
(b) Measured off-line values and predicted values obtained using the LW-MB-PLS model, which employed all the available data blocks.

Figure 5.8 – Time evolution of the measured (coloured filled circles) and predicted values (black non-filled circles) of QP1 for the January–June 2021 validation period. The predictions were obtained using two different models: ROSA and LW-MB-PLS. Blue and red dashed lines represent the warning thresholds and the actual low-quality threshold, respectively.

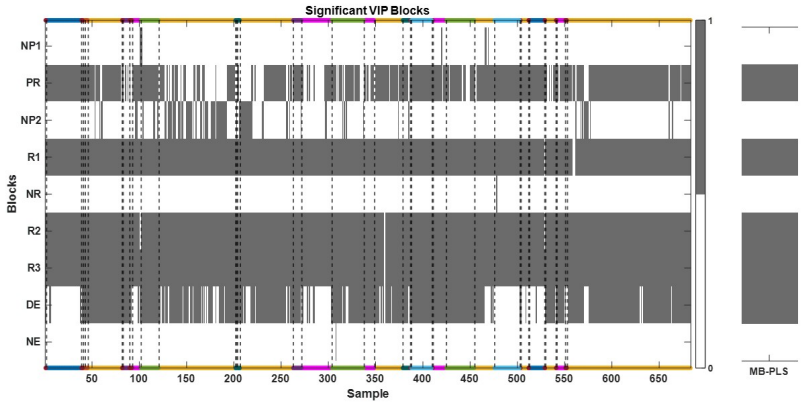
Local Model Interpretation

One of the complications of local models is the loss of interpretation, as one will no longer have one model for all tests, but several models, one for each sample to be predicted. In this section, a procedure is proposed to assess the diversity between the various models, in other words, to understand how they use the information from the calibration data differently, locally. The variance explained for each block and the VIPs will be examined.

Figure 5.9 show the explained variance for each block and the block VIPs of the model examined (i.e. QP1, all available blocks). At the top and bottom of the figures there is a coloured line corresponding to each product, while the dashed black lines indicate a product change. The examples shown are orders per production time and include both validation sets (before and after the production stop). The right part of the figures also shows, for comparison, the explained variance per block from the MB-PLS model computed on the same blocks. In general, when comparing the figures, it can be seen that in the LW-MB-PLS model there is a certain consistency between the VIP values or the explained block variance for the same type of product. Therefore, depending on the product, the relevance of the blocks changes (e.g. for product 7, blue, block NE contributes to the model, explained variance above 70%, while for product 1, yellow, it does not), and sometimes also for the same product with time (i.e. block R1 and NR contribute a lot for product 9 in the period June 2022 (about 500 as sample number in Fig. 5.9a), while not at the end of May 2021 (about 400 as sample number in Fig. 5.9a). It is noteworthy that the production stop took place between the two time periods. In addition, considering the VIP values, although the NIR blocks contribute to the model for some products/periods, for the same products/periods the VIPs are below the significance threshold, which means that these blocks contribute to components that explain a small percentage of QP1. This is consistent with the fact that ROSA does not select them.



(a) Explained variance for each block



(b) block VIPs

Figure 5.9 – The results shown refer to validation sets, i.e. covering the whole production time, for QP1. Explained variance for each block (**a**) and block VIPs (**b**) related to the LW-MB-PLS model built with all the available data blocks; values are shown in coded colour according to the colour bar. Coloured lines at the top and the bottom of the figure indicate the product grade, whereas the dashed black lines indicate a product change. On the right of the figures, for comparison, are shown the results of the MB-PLS model computed with the same blocks. In (b), dark gray areas indicate a significant VIP value for the specific block.

5.4.3 Comparison between the multiblock methods

Table 5.4 – Results obtained through LW-MB-PLS and MB-PLS.

Blocks for model building	LVs	RMSECV (g)	RMSEP (g)	MB-PLS RMSEP (g)
NP1, PR, NP2, R1, NR, R2, R3, DE, NE[3]	1	0.62^{a}	0.75^{a}	0.82^{a}
NP1, PR, NP2, R1, NR, R2, R3	1	0.64^{ab}	0.91^{bc}	0.99^{b}
NP1, PR, NP2, R1, NR, R2	1	0.67^{ab}	0.97^{c}	0.97^{b}
NP1, PR, NP2, R1, NR	1	0.73^{bc}	1.28^{d}	0.97^{b}
PR, R1, R2, R3, DE	2	0.57^{a}	0.78^{a}	0.80^{a}
PR, R1, R2, R3	1	0.59^{a}	0.77^{a}	0.87^{a}
PR, R1, R2	2	0.63^{ab}	0.85^{b}	0.84^{a}
PR, R1	1	0.67^{ab}	0.85^{b}	1.05^{b}
NP1, NP2, NR, NE	2	0.74^{c}	1.34^{d}	2.15^{d}
NP1, NP2, NR	3	0.81^{d}	1.67^{e}	2.64^{e}
NP1, NP2	3	0.98^{e}	1.31^{d}	1.26^{c}

In a column, values with the same letter are not statistically different between each other ($p > 0.05$).
[a]D = DEVO, DE = DEVO-END, E = END, NE = NIR END, NP1 = NIR PRE 1, NP2 = NIR PRE 2, NR = NIR REACTION, PR = PRE REACTION, R1 = REACTION 1, R2 = REACTION 2, R3 = REACTION 3.

According to the results in Table 5.4, as in the case of the ROSA method, also for LW-MB-PLS, the models that show the most accurate prediction performance in terms of RMSEP are those computed with all the blocks available, and reducing the number of blocks tends to significantly increase the prediction error ($p < 0.05$), except for QP1 when the excluded blocks are the NIR ones (i.e. the models that include all the process sensor blocks and all except the last DE have the same performance). This confirms that they are not relevant for the prediction of QP1. In general, the same consideration made in section 5.4.1 can be confirmed here. Table 5.4 also shows the results obtained with standard MB-PLS, which in most cases show higher RMSEP values than those obtained with LW-MB-PLS.

The differences between the predictive performance of the three methods, ROSA, LW-MB-PLS and MB-PLS, were evaluated according to the ANOVA performed taking into account the prediction error, as described in section 5.3.6. The results are shown in Table S1 in the supplementary material to PAPER III. In general, MB-PLS shows significantly worse prediction performance than the other two methods in almost all cases (i.e. blocks used for modelling), with a few exceptions where it performs equally to LW-MB-PLS. In the case of QP1, it can be observed that ROSA and LW-MB-PLS show similar performance mostly when NIR blocks are not present in the blocks considered for modelling. In general, ROSA performs better

when noisy blocks are present because it can select only a few of the blocks and only non-redundant information. However, for some ABS products, LW-MB-PLS helped to reduce a systematic prediction error that was quite evident in ROSA. This can be seen in Figure 5.10, where the last part of the validation period corresponding to March-April 2022 for QP1 is shown. Here, the application of LW-MB-PLS reduced the model bias, making the prediction trend more accurate. The mean prediction error is the same or slightly lower for ROSA, meaning that LW-MB-PLS outperforms ROSA for the prediction of certain products, such as product 1 in the figure.

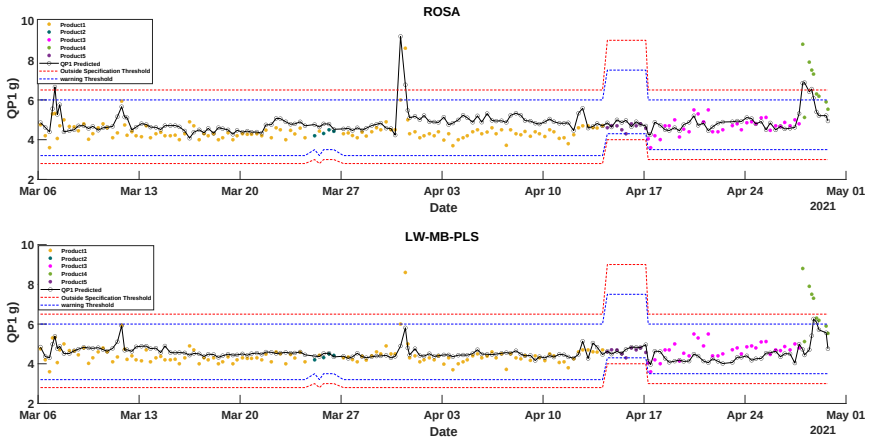


Figure 5.10 – Time evolution of the measured (filled circles) and predicted values (black circle) of QP1 for the final portion of March–April 2022 validation period by ROSA and LW-MB-PLS. The predictions were obtained by means of the models that employed all the available data blocks.

A first general remark is that ROSA and LW-MB-PLS are based on different methodologies. LW-MB-PLS, being based on MB-PLS, does not provide a clear extraction of the common and distinctive information retrievable from each block, since block importance is evaluated only in terms of block weights in the final model. On the other hand, ROSA aims at retrieving unique complementary information by applying block orthogonalisation to the previously extracted component before moving on to the next component, whereas MB-PLS does not remove already used information in a block. From an application point of view, which is what this paper is concerned with, we can observe that both methods provide models with good predictive capability. ROSA has the advantage of providing a single model

and is therefore very easy to implement in a real-time scenario (only the **b** coefficients need to be stored and used for prediction). LW-MB-PLS requires the calculation of the distances between the sample to be predicted and all the calibration samples (slow step) and the fitting of a PLS model with the selected neighbours before the prediction step (fast step). Another attractive feature of ROSA is its ability to filter out redundant information between blocks, which can guarantee more robust models. However, the method depends on the choice of the winning block and often several blocks have similar errors, so this aspect needs further investigation. However, in cases such as the process studied with multiple product grades, or in the presence of non-linearities, a local approach is required to reduce systematic errors.

5.5 Conclusion

This work analysed the application of two multiblock regression methods, Response Oriented Sequential Alternation (ROSA) and Locally-Weighted-Multiblock Partial Least Squares (LW-MB-PLS), a new extension of Locally-Weighted Partial Least Squares introduced in this paper, for the online prediction of quality parameters (QP1 and QP2) in a full-scale styrene polymer production plant. Comparing these two approaches with a traditional MB-PLS

In general, it was found that the integration of several sensors contributes to a good real-time estimation of the characteristics that define the quality of a product. When also discussing QP2 (see Paper III), which is not covered in this chapter, it is also important to highlight how NIR sensors contribute significantly to the estimation of QP2 and not QP1.

The ROSA method showed promising predictive performance for both QP1 and QP2, with the selection of influential blocks providing information on critical plant sections. The importance of sensors in the initial and final stages of the process was highlighted and the impact of specific sensors on the quality of the final product was clarified. The results indicated the

possibility of obtaining reasonable estimates of QP1 and QP2 values prior to the completion of the production process, offering potential approaches for real-time prediction and control. On the other hand, the LW-MB-PLS method, while generally showing comparable predictive accuracy, demonstrated effectiveness in reducing systematic errors for some products. The computational efficiency of ROSA was recognised, although LW-MB-PLS presented advantages in mitigating prediction biases for specific ABS products.

From an application point of view, both methods are feasible for real-time prediction. LW-MB-PLS can be recommended when non-linearity is observed or, as in the present case, when different grades of product have to be handled. ROSA is particularly fast and can be used to evaluate the relevance of each block sequentially, in addition it can lead to a more robust model by filtering out redundant information between blocks. In perspective, ROSA can be used in the process understanding phase to exploit the possible scenarios and then, if a prediction bias is observed, it can be resorted to local modelling using only the most salient block. However, a drawback of ROSA that requires further investigation is how to deal with blocks with similar errors in the selection phase. Overall, this study contributes to the understanding of multiblock regression techniques in the context of continuous production processes, providing valuable insights for plant operators and paving the way for further advances in online quality prediction and control.

Chapter 6

Improving ROSA: From Global to Local Models

6.1 Introduction

As highlighted in Chapter 5, it has become increasingly common in various research and practical domains to use multiple analytical platforms for the comprehensive characterisation of specific systems. This trend has stimulated the development of numerous approaches for the simultaneous analysis of the different data sets (blocks) provided by these platforms. Such approaches aim both to improve the prediction of certain properties (responses) and to elucidate how these properties are related to the collected measurements (predictors or regressors).

Response-Oriented Sequential Alternation (ROSA) [182] has recently been introduced as an innovative Partial Least Squares (PLS) based regression tool for multiblock data analysis. Unlike traditional methods (e.g. Multiblock Partial Least Squares-MB-PLS [177] or Sequential and Orthogonalised Partial Least Squares-SO-PLS [172]), ROSA is characterised by high computational speed and enhanced robustness to differences in block variance and permutations of block order. This robustness results from a competition-based "winner-takes-all" strategy in which each model compo-

nent is computed from the block of predictors that yields the lowest residual error at that time.

Despite these advantages, ROSA has certain limitations.

A notable weakness is that ROSA relies on the global minimum of residuals to select the "winning block" for each component, even though there may be other blocks with residuals that are not statistically different. This criterion could find a local minimum. As investigated in PAPER III, an attempt was made on the ROSA model with all blocks by forcing the selection of a different block for the first component, each time selecting one of the blocks with equivalent residuals. As shown in figure 6.1 for the model applied to QP2 of the ABS plant (Chapter 5), none of these alternative models performed significantly better in terms of RMSEP, although some performed worse. However, the initial choice of block affects subsequent block choices at higher components and also affects the number of components giving the lowest RMSECV. If model interpretation is based on the block selected by ROSA, this ability to have an alternative block selection approach will affect interpretability.

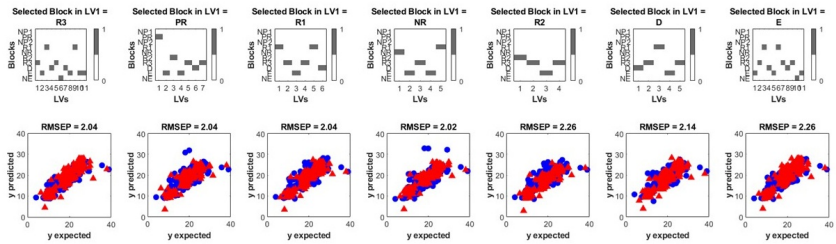


Figure 6.1 – The first row shows the blocks selected by different ROSA models, varying the selection of the first block by using others that are not statistically significant. Note how the selection of subsequent blocks changes depending on the initial block selection. The second row shows the measured versus predicted plots for the QP2 property. Calibration points are shown in blue, while test points are shown in red.

Various strategies could be adopted to mitigate this problem. For example, in an industrial setting, blocks measured earlier in the process (when several blocks are statistically equivalent) could be prioritised to allow earlier predictions before the final product is complete. More generally, it might be

beneficial to develop a strategy that simultaneously includes all blocks that are not statistically different.

In addition, as a linear modelling technique, ROSA may struggle with non-linear relationships in the data. This was confirmed in chapter 5, where ROSA's predictions were biased for certain products. In this chapter, we continue to address the challenge of nonlinearity that characterised this section of the thesis, and propose improvements to the ROSA algorithm that aim to overcome this second limitation. Specifically, we introduce a locally weighted version of ROSA that is capable of handling non-linear relationships between predictors and responses, while preserving its distinctive property of independence among blocks. The advantages of this solution are demonstrated and evaluated using both simulated and real case studies.

6.2 Data

6.2.1 Simulated Data

A four-block simulated dataset, representing different analytical techniques used in sample analysis, was generated using an R software package developed by Metz et al. [191] (available at `https://github.com/maxmetz/data_simulation`).

For each of the first three blocks ($\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$), spectral-type data were created (see Equation 6.1) by multiplying a 1000×10 matrix $\mathbf{R}$ whose columns were drawn from a normal distribution with mean 0 and standard deviation 0.5 by a 10×1001 matrix $\mathbf{B}^T$, containing ten distinct pseudo-spectral profiles obtained through combinations of Gaussian curves. Each block was generated independently to ensure unique spectral profiles.

$$\mathbf{X} = \mathbf{R}\,\mathbf{B}^T \tag{6.1}$$

A fourth dataset, $\mathbf{X}_4$, intended to represent non-continuous variables

(e.g. pH, temperature), was generated only from the 1000×10 matrix $\mathbf{R}$ (standard deviation 0.5, using the Matlab *normrnd* function). This approach can emulate measurements of process variables or other discrete data.

A single dependent variable, $\mathbf{y}$, was then simulated according to the following equation to introduce a non-linear dependence on $\mathbf{X}$:

$$\mathbf{y} = 0.8(\mathbf{t}_1^2 + 15) + 0.3(\mathbf{t}_3^2 + 11) + 0.1(\sin(\mathbf{t}_4) + 10) \tag{6.2}$$

Here, $\mathbf{t}_b$ $(N \times 1)$ contains the first principal component scores from the PCA decomposition of the b-th block, $\mathbf{X}$. To demonstrate the ROSA property of discarding information not relevant to the prediction, the second block $\mathbf{X}_2$ is not correlated with $\mathbf{y}$.

Afterward, both $\mathbf{X}$ (all four blocks) and $\mathbf{y}$ were corrupted by Gaussian white noise. Finally, the data were split into calibration and test sets using a 70/30 ratio.

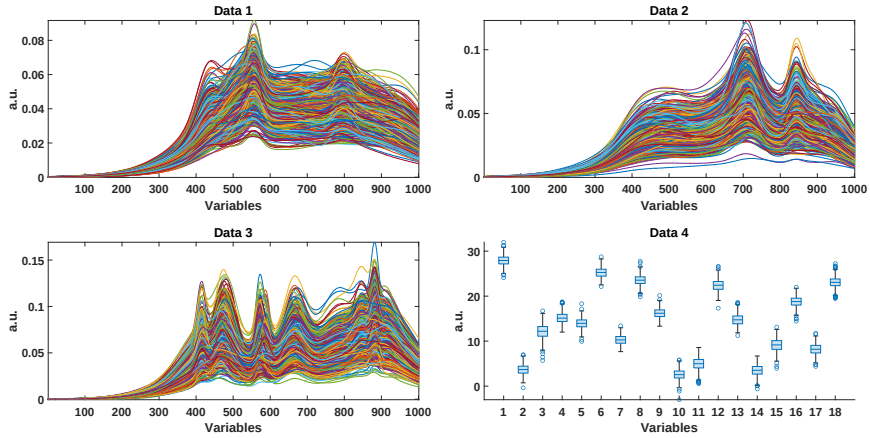


Figure 6.2 – The figure represents simulated data for four blocks of data. The first three plots correspond to simulated spectra, while the fourth plot represents simulated discrete variables.

6.2.2 ABS production data

The Real case of study take into account for this study is the production of Acrylonitrile butadiene styrene (ABS) in Versalis S.p.A explained in section 5.2.1.

6.3 Model

This section will present the extension of ROSA to handle non-linear calibration points, which closely resembles case calibration points. Starting with the motivation on the calibration set weighting strategy and then moving on to the illustration of the pipeline from the algorithm. The results obtained with LW-ROSA on both simulated and real-case datasets will be compared with those achieved using ROSA, whose explanation is provided in Chapter 5.

6.3.1 Locally Weighted ROSA

Locally Weighted ROSA adopts a local modelling philosophy, meaning that for each new sample $\mathbf{x}_{\text{new}}$, the prediction is based on a model trained on a calibration set that emphasizes the contribution of samples most similar to $\mathbf{x}_{\text{new}}$. The rationale behind this approach is that only calibration points closely resembling the new sample contribute meaningfully to the relationship between $\mathbf{X}$ and $\mathbf{y}$. As noted by Lesnoff et al. [183], various strategies can be used to emphasize the influence of the most relevant points, including selecting a sub-calibration set, weighting calibration points according to their dissimilarity, or combining both methods.

In LW-MB-PLS, introduced in Chapter 5, a combination of K-Nearest-Neighbour selection and weighting is employed. Neighbours are identified and weighted according to their Euclidean distance in the PCA (score) space of the matrix obtained through low-level data fusion and block variance scaling of all variables. Although the number of nearest neighbours and

the weighting function are optimized, this approach may still be suboptimal because each block spans a different variable space. Consequently, distances between $\mathbf{x}_{\text{new}}$ and calibration samples may differ across blocks, potentially yielding different sets of nearest neighbours in each space.

This issue is illustrated in the figure below, where the first calibration sample (red triangle) is projected onto the two principal components defined for each block. The calibration points (circles) are colour-coded according to the weight they receive, computed from the Euclidean distance and a Gaussian function (see equation) 5.13 with $h = 1$.

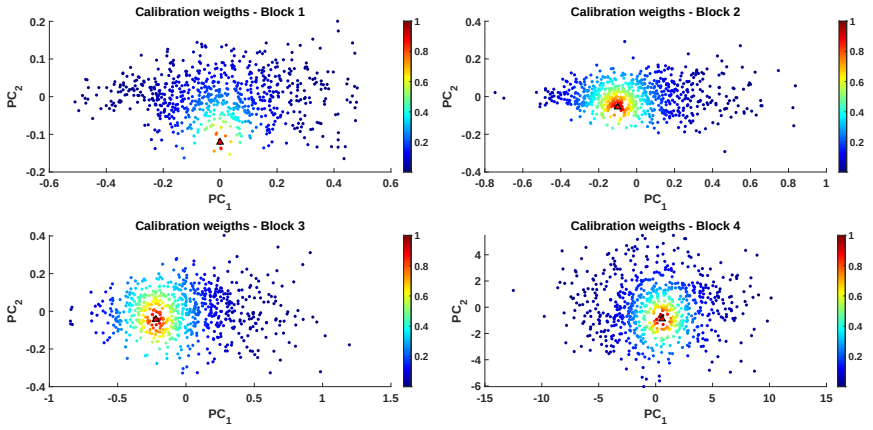


Figure 6.3 – This figure illustrates the distribution of calibration weights across four different data blocks. Each plot represents the calibration points projected in the PCA space of the respective block, with the first two principal components as axes. The calibration points are shown as circles, and their colours indicate the assigned weights, The red triangle in each plot represents a new sample projected into the PCA space. This visualization emphasizes how the calibration weights vary across the blocks, reflecting the different importance assigned to calibration points relative to the new sample in each case.

This aspect is further emphasised in table 6.1, which shows the top ten nearest neighbours for each block. The table highlights the different order of neighbours across blocks, allowing a clear comparison of their diversity.

Table 6.1 – Top 10 nearest neighbors for each block.

	Block 1	**Block 2**	**Block 3**	**Block 4**
1	700	144	498	34
2	240	221	177	306
3	59	320	491	659
4	276	32	64	548
5	382	405	214	582
6	503	121	441	366
7	292	87	501	47
8	101	52	191	53
9	557	348	163	181
10	637	161	244	151

Weighting strategy

Based on the difference shown between the blocks, it is not possible to select a calibration subset with the nearest neighbours, because all blocks could have different samples and the ROSA model would not be consistent. For instance, the scores values of n-th rows will refers to different samples.

To overcome this, the use of all the calibration set weighted according the dissimilarity has been adopted. With this strategy the non linearity has been addressed considering that the most influent calibration point will have a higher weights while the other will have a low weights and consequently their impact will be reduced. In literature are present different weighting function [192], the ones explored in this work will be listed in the following tables 6.2:

When the dataset is large, to drastically reduce the influence of the most dissimilar samples, an approach inspired by activation functions in the Deep Learning domain has been adopted. To perform this truncation of the weights, the following steps are necessary:

- Identify the most similar samples within a specific percentile n of the distance distribution. This percentile parameter needs to be optimized based on the data.

Table 6.2 – Distance-based weighting schemes.

Weighting function	W
Uniform	1
Triangular	$1 - d_i$
Quadratic	$(1 - d_i^2)^2$
Cubic	$(1 - d_i^3)^3$
Gaussian	$e^{-\frac{d_i^2}{h \cdot \sigma(\mathbf{d}^*)^2}}$
Exponential	$e^{-\lvert d_i \rvert}$

Here, d_i represents the distance of a training set sample from the new sample under examination.
$* \ \sigma(\mathbf{d}^*)$ is the standard deviation of the distances of the entire calibration set from the new point to pred
that modifies the shape of the function, modulating how much the distance impacts the weights.

- Apply a LeakyReLU-like function [193], where the weights within the percentile remain unchanged, while the weights beyond the percentile are scaled by a multiplicative factor α (generally $\alpha = 0.1$ or $\alpha = 0.01$).

The LeakyReLU function is defined as:

$$W' = \begin{cases} W & \text{if } W \leq W_{\text{threshold}} \\ \alpha \cdot W & \text{if } W > W_{\text{threshold}} \end{cases}$$

where $W_{\text{threshold}}$ corresponds to the weight cut-off determined by the n-th percentile, and α is the scaling factor that adjusts the weights of less similar samples. Using a strategy that considers samples within a percentile [194] has the advantage of dynamically adjusting the number of samples based on the density of different regions in the space. This approach ensures that a suitable number of samples is selected depending on the local data distribution, which is not feasible when using a fixed K number of samples.

Algorithm Pipeline

As anticipated, the proposed method is based on the LW-PLS algorithm within a local weighting scheme, particularly following the framework described by Lesnoff *et al.* [183].

The algorithm involves the following key steps:

1. Compute the distances between the new sample to be predicted and the calibration set for each block. Distances can be calculated in different spaces, such as the original variable space, PCA space, or PLS space.
2. Calculate the weights based on the dissimilarity measures.
3. Perform a weighted ROSA procedure.
4. Repeat the process for each new sample.

Given the general framework identified by these 4 steps, the proposed locally weighted ROSA algorithm can be implemented in different ways, depending on how the similarity (type of distance used Euclidean, Mahalanobis) between the samples and the distance-based weighting scheme are defined. This aspect provides enormous flexibility in this context, allowing the model to adapt to different cases. However, the construction of a LW-ROSA model requires the identification of the optimal complexity, i.e. the number of latent variables to be used, the choice of metric and distance-based weighting scheme, and whether or not to apply truncation of the data can also be optimised. In the present study, this choice was made on the basis of the results of cross-validation with a grid search.

6.4 Results and Discussion

6.4.1 Simulated case

Starting from the analysis of the simulated case using the original ROSA method, the model was optimised by cross-validation with a 10-fold *venetian blind* split. The optimal model was selected using 3 latent variables (LVs), representing the best compromise between RMSECV and model complexity, with a resulting RMSECV of 0.057.

Figure 6.4 shows the prediction results on the test set. The final model achieved a RMSEP of 0.075. The order of block selection was as follows [1 - 3 - 4]. As expected, the second block, which does not contribute to $\mathbf{y}$, was

not selected. In a test to check the trend of block selection, a dive was made to increase the number of LVs to 10 LVs and the second block was also never selected by the model.

However, by examining the predicted vs. expected plots and the residual plots, it is possible to observe the presence of a trend. This trend indicates that ROSA may struggle with data sets characterised by the presence of non-linearities.

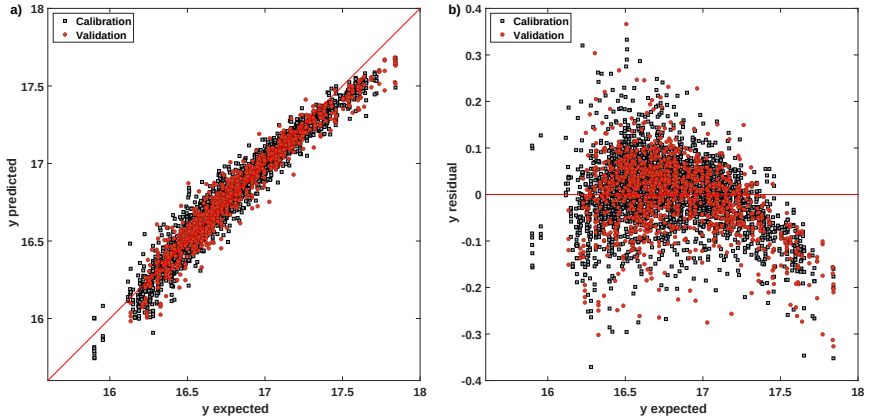


Figure 6.4 – Prediction results of the ROSA method on the simulated dataset. **a)** shows the **y** predicted vs. **y** expected values, with calibration points represented as black squares and validation points as red circles. **b)** presents the **y** residuals, highlighting the trends in the residuals.

To further investigate and attempt to resolve the observed trend in the residuals, the same simulated data set was analysed using the proposed LW-ROSA method.

The primary objective was to assess whether the incorporation of locally weighted approaches could mitigate the identified trend and improve the model's ability to handle the non-linearities present in the data. To this end, a LW-ROSA model with 3 LVs, using a Euclidean distance on the PCA score space of each block and a cubic weighted function with a decrease of the weights after the 20° percentile using the activation function explained in section 6.3.1, has been built. This model performed with an RMSEP of 0.062. This combination was chosen through a grid search cross-validation, testing all the possible combinations of the different parameters listed in

table 6.3 and selecting the one that presented the minimum on the Måge plot [181], as reported in section 5.4.2.

Table 6.3 – Parameters Considered in the Cross-Validation

Parameter	Values
Shape function (h)	0.1, 0.5, 0.8, 1, 2
Percentile	5, 10, 15, 20, off
Weighting function	Gaussian, Quadratic, Cubic, Exponential
Distance metric	Euclidean
Space	Score space (2 PCs)

The shape function (h) is a parameter specific to the Gaussian weighting function.

Comparing the results obtained with LW-ROSA (see Figure 6.5) with those of the original ROSA method, it is possible to note that the trend is almost absent, except for a few points at the extremes of the higher values of **y**. A possible explanation could be the presence of few points in the calibration that describe this range.

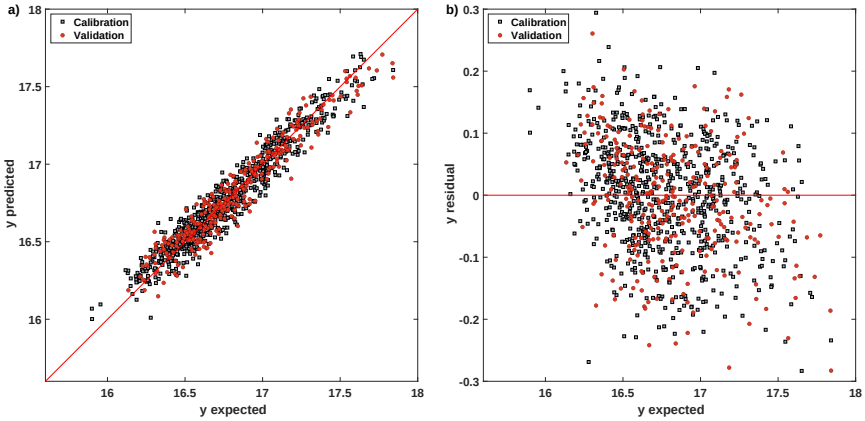


Figure 6.5 – Prediction results of the LW-ROSA method on the simulated dataset. **a)** shows the **y** predicted vs. **y** expected values, with calibration points represented as black squares and validation points as red circles. **b)** presents the **y** residuals, highlighting the trends in the residuals.

As in the global case, data from block 2 are not considered by the model for the purpose of predictions. However, unlike the global model, here the selection of blocks varies depending on the sample analysed. From Figure 6.6, this selection does not seem to show a clear correlation with the value of the **y**. Further investigation will be needed to understand the underlying

local logic driving the selection of blocks.

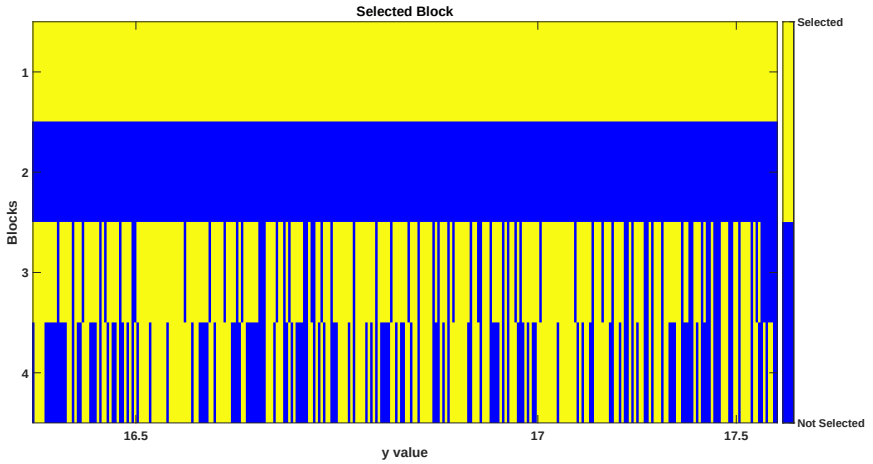


Figure 6.6 – Visualization of the blocks selected by the different local models. The selected blocks are highlighted in yellow, while the unselected ones are in blue. The y-axis represents the blocks, and the x-axis corresponds to the test samples, sorted in ascending order of their y-values.

6.4.2 ABS production data

For the real-world case study of ABS production at Versalis S.p.A., the objective is to address the bias observed for some products, as highlighted in Section 5.4.1. Although this issue has already been addressed with the LW-MB-PLS model (Section 5.4.2), an additional objective is to achieve predictions that rely only on the most influential blocks, similar to the approach used in ROSA something that is not achieved with LW-MB-PLS.

For this purpose, the LW-ROSA model was constructed using all available blocks as input. The model was optimised using a cross-validation strategy with a grid search for the best combination of parameters, as shown in Table 6.4. The chosen configuration represents the best compromise between RMSECV and model complexity. Specifically, the model used 5 LVs, with dissimilarity calculated from the Euclidean distance in PCA space (2 PCs), weights determined by a Gaussian function ($h = 0.1$) and a weight decrease after the 5th percentile using the activation function described in 6.3.1. In this case, where the calibration set is more rich, it seems necessary to have

a more drastic reduction of the weights for the more dissimilar samples and, contrary to the simulated case, a more severe weighting option was chosen.

Table 6.4 – Parameters Considered in the Cross-Validation

Parameter	Values
Shape function (h)	0.1, 0.5, 0.8, 1, 2
Percentile	5, 10, 15, 20, off
Weighting function	Gaussian, Quadratic, Cubic, Exponential
Distance metric	Euclidean
Space	Score space (5 PCs) for all blocks

The shape function (h) is a parameter specific to the Gaussian weighting function.

Figure 6.7 compares the predictions on the test set obtained with the ROSA and LW-ROSA models. Although both models achieved an equal RMSEP of 0.74 (g), the LW-ROSA model showed a slight improvement in the predictions for Product 1 (yellow). These predictions are better aligned with the 1:1 red line, despite a greater dispersion observed for Product 3 (purple).

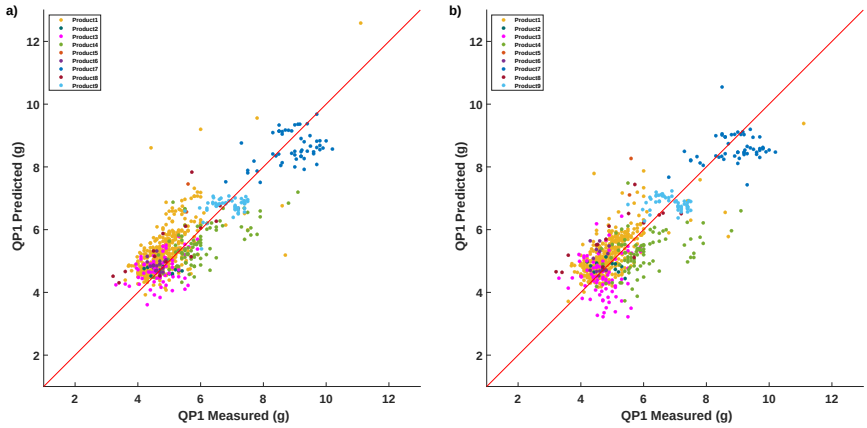


Figure 6.7 – Prediction results for the test set using **a)** the ROSA model and **b)** the LW-ROSA model. The data points are represented as colored circles, where the color corresponds to the product type. The x-axis represents the measured values of GPI, while the y-axis represents the predicted values.

Out of the 9 blocks available, the LW-ROSA model consistently selected only last three process sensor blocks (Reaction 2 - Reaction 3 - Devo/END) across all products. While the selective use of influential blocks is an advantage, in this case study it did not lead to a noticeable improvement in

performance. This result may be due to the weighting strategy, which may not yet be fully optimised to adequately reduce the impact of less influential samples.

6.5 Conclusion

In this chapter, an extension of the ROSA algorithm, called Locally Weighted ROSA (LW-ROSA), is introduced to address the limitations of the original ROSA method in handling non-linearities. The LW-ROSA approach is based on the strategy of local methods that emphasise the contribution of the most relevant samples, in this case using a weight function. One of the novelties and contributions to the framework of multiblock nonlinear models is that with this version of ROSA, independence between blocks is maintained.

The application of LW-ROSA to both simulated and real case studies demonstrated its potential benefits. In the simulated case, the method successfully mitigated the trends observed in the residuals of the original ROSA model, reducing the bias in the prediction.

In the real-world case of ABS production at Versalis S.p.A., LW-ROSA demonstrated its ability to select only the most influential blocks, consistently selecting three process sensor blocks across all products. Even though it achieved comparable RMSEP values to the original ROSA model, LW-ROSA showed a smoother fit for certain product types, such as product 1. However, the overall improvement in prediction performance was limited.

Although still at an early stage of development, the LW-ROSA approach shows promising potential. The weighting strategy, combined with deep learning inspired activation functions, has proven to be an effective tool to dynamically adjust the influence of samples based on local density. However, further refinement may be required to better suppress the influence of less influential samples. Ideally, the method should emulate a system that operates with a single calibration subset. In others word, where only the nearest neighbours serve as inputs and contribute to the modelling of a single latent

variable, with all other samples effectively assigned zero weight.

A potential improvement would be to use a standard ReLU function to assign zero weights, rather than a smooth activation function such as LeakyReLU. However, there is a challenge with this approach: the presence of a large number of zero-weighted samples could interfere with the calculation of scores by affecting the $\mathbf{XY}$ covariance. Consequently, this could lead to suboptimal score calculations.

Chapter 7

Locally-Weighted-RoBoost-PLS: a multivariate calibration approach to simultaneously cope with non-linearities and outliers

7.1 Introduction

Partial Least Squares regression (PLS) [73, 84] is a widely used multivariate calibration tool for modelling linear relationships between independent variables and dependent variables (responses). In contrast to classical multilinear regression, it excels in scenarios in which high-dimensional datasets exhibiting a high degree of collinearity are involved. The main advantage PLS brings regards the fact that it provides not only a model linking the regressors and responses but also a full description of both in terms of latent factors, thus enabling an in-depth interpretation of the final results it yields. In spite of all these benefits, though, PLS can suffer from severe limitations when the relationships between regressors (usually encoded in a matrix de-

noted as $\mathbf{X}$ of dimensions $N \times V$) and responses (usually encoded in a matrix denoted as $\mathbf{Y}$ of dimensions $N \times M$) are non-linear. This situation can occur quite frequently, e.g. when dealing with agronomic samples collected during different harvesting campaigns [195, 196], and in industry when the same plant manufactures different products through smooth formulation changes [197] or undergoes temporal drifts due to raw material or catalyst degradation [198]. If non-linearities are moderate, non-linear (e.g. logarithmic) transformations of the response(s) [199] can be used to model them, otherwise, in more complex situations, non-linear implementations of PLS should be considered. These extensions of PLS include, for example, Kernel PLS (K-PLS [200–202]) or local PLS [203–205]. K-PLS handles non-linearities by applying a specific kernel function (polynomial, gaussian, sigmoidal, etc.) to $\mathbf{X}$ for mapping the original regressors into a higher-dimensional feature space where a linear PLS model can be constructed. On the other hand, the main idea behind local PLS is to build for each new specimen whose responses are to be predicted an individual PLS model on a reduced subset of calibration samples that are most similar to it [194, 203–207]. Local PLS implementations are definitely the most commonly utilized in chemometrics. This is partly due to the fact that i) K-PLS may struggle to adequately capture the non-linearities of a given dataset if the kernel function is not properly tuned [208], ii) K-PLS model interpretation is complicated by the fact that information on the relevance or importance of the original variables is lost when performing the aforementioned mapping operation and iii) K-PLS might require long processing times when the sample size becomes particularly large [201]. Among local PLS implementations, K-Nearest-Neighbours-Locally-Weighted-PLS-Regression (KNN-LW-PLS) [183] is undoubtedly the one that has lately attracted more attention from users and practitioners. More in detail, KNN-LW-PLS trains local models exactly as outlined before, but, in addition, it weighs the samples belonging to the calibration subset according to their distance to the one to be assessed. This allows effectively capturing and describing strong non-linearities and complex patterns in data, such as the presence of distinct observation clusters, that may hamper the application of classical linear PLS to the entire dataset at hand. Another issue that can dramatically jeopardize the predictive performance of standard PLS is the presence of outliers in the calibration data. In this

regard, in the last decades, many robust versions of PLS have been developed in an attempt to downweigh outlying samples and, thus, reduce their influence in the PLS model calibration stage [209–212]. In this chapter, we particularly focus on a recent robust PLS implementation, RoBoost-PLS [213, 214], which stands out for its ability to reduce the influence of outliers during calibration by weighting the investigated samples differently for each extracted latent factor and according to three distinct criteria: **X**-residuals, **Y**-residuals and leverage. This method has proven to be effective when it comes to dealing with outliers in both **Y** and **X**. Although all these PLS extensions (non-linear and robust) perform well when trying to tackle the specific target problem for which they have been originally proposed, they may encounter difficulties in situations where both non-linearities and outliers coexist. Notwithstanding, to the best of our knowledge, no PLS algorithm capable of handling both these issues simultaneously has been devised yet. For this reason, we propose here a novel approach based on a rational combination of KNN-LW-PLS and RoBoost-PLS and named Locally-Weighted-RoBoost-PLS (LW-RoBoost-PLS), designed to deal with similar scenarios. Although the method still uses KNN for local weighting, the name has been simplified by dropping the 'KNN-' for ease of reading. The performance of LW-RoBoost-PLS was here evaluated on simulated and real datasets and compared with that of the two native approaches from which it originates.

7.2 Data

LW-RoBoost-PLS was tested on simulated data generated by using an `R` software package developed by Metz et al. [191] (available at `https://github.com/maxmetz/data_simulation`) and on a challenging real-world dataset related to the production process of Acrylonitrile-Styrene-Butadiene (ABS). Both simulated and real-world data exhibit outliers and non-linear variable relationships, making them suitable for assessing the effectiveness of the proposed approach.

7.2.1 Simulated data

A spectral-like dataset characterized by the presence of both **X**- and **Y**-outliers and by a non-linear dependence between **X** and **Y** was simulated as detailed below.

A dataset **X** of dimensions $N \times V$ was generated by multiplying a 900×10 matrix (say **R**) whose columns carried values drawn from a normal distribution with mean equal to 0 and standard deviation equal to 0.5 by a 10×1001 array (say $\mathbf{B}^T$) whose rows contained ten different pseudo-spectral profiles (see Equation 7.1), the pseudo-spectral profiles were obtained by the combination of 10 distinct Gaussian curves:

$$\mathbf{X} = \mathbf{R}\mathbf{B}^T \tag{7.1}$$

An individual dependent variable, **y**, was afterwards simulated based on the following equation to ensure a non-linear dependence with **X**:

$$\mathbf{y} = \mathbf{t}_1^2 + 15 \tag{7.2}$$

where $\mathbf{t}_1$ $(N \times 1)$ gathered the first principal component scores resulting from the Principal Component Analysis (PCA) decomposition of **X**. Both **X** and **y** were afterwards augmented with outlying observations.

Thirty moderate **X**-outliers were produced as in Equation 7.1 but drawing the values along each column of **R** from a normal distribution with mean equal to 1 and standard deviation equal to 0.8. Forty extreme **X**-outliers were instead produced altering both **R** and $\mathbf{B}^T$, i.e. drawing the values along the columns of **R** from a normal distribution with mean equal to 1.5 and standard deviation equal to 0.5, and using different pseudo-spectral profiles for $\mathbf{B}^T$. Thirty **Y**-outliers were, finally, obtained modifying Equation 7.2 as:

$$\mathbf{y}_{\text{out}} = \mathbf{t}_1 + 17 \tag{7.3}$$

Ultimately, Gaussian white noise was added to both $\mathbf{X}$ and $\mathbf{y}$.

For the sake of a fair comparison among KNN-LW-PLS, RoBoost-PLS and LW-RoBoost-PLS, the entire dataset was split into a calibration set of 700 samples and a test set of 300 samples. It is important to notice here that outliers were exclusively kept in the calibration set and that they constituted approximately 15% of it. Figure 7.1 displays the simulated calibration data and their corresponding $\mathbf{y}$-value distribution.

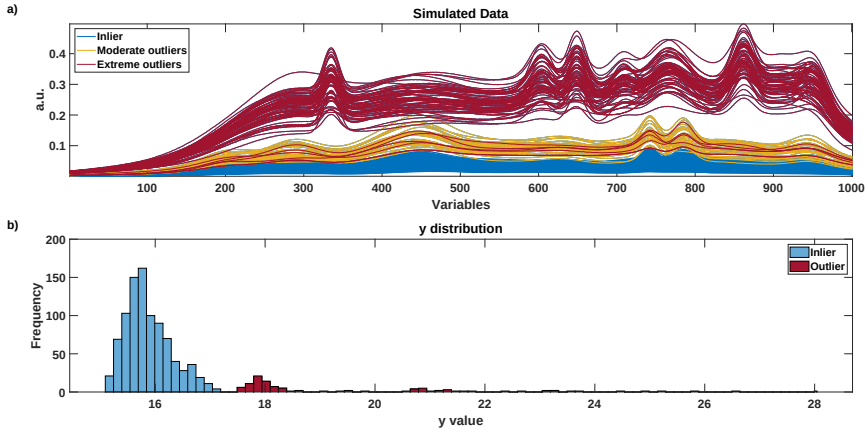


Figure 7.1 – **a)** Simulated calibration spectra and **b)** their corresponding $\mathbf{y}$-value distribution

7.3 ABS production data

The real dataset analysed relates to a production campaign of ABS carried out between January 2020 and April 2022 in an industrial plant owned by ENI Versalis, located in Mantova (Italy). More details can be found in section 5.2.1.

In contrast to the previous sections, the entire available data set was divided into calibration and test sets, corresponding to the first two years of production (2020-2021, 1851 measurement points) and the remaining manufacturing period (2022, 344 measurement points), respectively. In order to evaluate the performance and develop the KNN-LW-PLS, RoBoost-PLS and LW-RoBoost-PLS regression models for predicting a single quality pa-

rameter, referred to as Property 1, the calibration set was not cleaned of potential outliers. On the other hand, the test set underwent a data cleaning step to minimise the presence of outliers to ensure a fair comparison of results between the different approaches.

7.4 Model

This section provides an overview of PLS, KNN-LW-PLS and RoBoost-PLS as well as a comprehensive description of the algorithmic scheme underlying LW-RoBoost-PLS and of the double cross-validation procedure designed for the optimization of its tuneable parameters.

7.4.1 Partial Least Squares regression (PLS)

Among the available PLS algorithms, we here refer to Nonlinear Iterative Partial Least Squares (NIPALS, on which also RoBoost-PLS is based), originally proposed by Herman Wold in the 1970s [84] and later adapted and modified by Svante Wold and Harald Martens [73]. NIPALS is an iterative approach that calculates PLS components one at a time by sequentially deflating from the data at hand the variability accounted for by the one estimated in the previous computational step.

Algorithm 1: NIPALS

Input: A predictor matrix $\mathbf{X}\,(N \times V)$ and a response matrix $\mathbf{Y}\,(N \times M)$, both centered.

1. Initialize $\mathbf{Y}$-scores, $\mathbf{u}_a$, as a column of $\mathbf{Y}$.
2. Calculate $\mathbf{X}$-weights for the a-th latent variable:
$$\mathbf{w}_a = \mathbf{X}^T\mathbf{u}_a \left(\mathbf{u}_a^T\mathbf{u}_a\right)^{-1}$$
3. Calculate $\mathbf{X}$-scores:
$$\mathbf{t}_a = \mathbf{X}\mathbf{w}_a$$
4. Define $\mathbf{Y}$-loadings:
$$\mathbf{q}_a = \mathbf{Y}^T\mathbf{t}_a \left(\mathbf{t}_a^T\mathbf{t}_a\right)^{-1}$$
5. Update $\mathbf{u}_a$:
$$\mathbf{u}_a = \mathbf{Y}\mathbf{q}_a \left(\mathbf{q}_a^T\mathbf{q}_a\right)^{-1}$$
6. Repeat steps 2 to 5 until convergence (i.e., until the difference between consecutive estimations of $\mathbf{u}_a$ is below a user-defined threshold).
7. Calculate $\mathbf{X}$-loadings:
$$\mathbf{p}_a = \mathbf{X}^T\mathbf{t}_a \left(\mathbf{t}_a^T\mathbf{t}_a\right)^{-1}$$
8. Deflate $\mathbf{X}$ and $\mathbf{Y}$:
$$\mathbf{X} = \mathbf{X} - \mathbf{t}_a\mathbf{p}_a^T$$
$$\mathbf{Y} = \mathbf{Y} - \mathbf{t}_a\mathbf{q}_a^T$$
9. Repeat steps 2 to 8 until all required latent variables (A) are extracted.
10. Calculate regression coefficients:
$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{Q}^T$$

7.4.2 K-Nearest-Neighbours-Locally-Weighted-PLS (KNN-LW-PLS)

The essential idea behind local PLS approaches is i) to select, among all the calibration samples available, a calibration subset made up of those that are most similar to each new specimen whose responses are to be estimated and ii) to calibrate a linear PLS regression model on such a reduced subset of observations for prediction purposes. Even if computationally demanding (notice that this way an individual PLS model is constructed for every new incoming test sample), this strategy permits to readily handle complex non-linearities linking predictors and responses.

Among the various local PLS implementations described in literature [194, 205, 215, 216], in the present study we focus on the one proposed by Lesnoff et al., KNN-LW-PLS [183]. However, for the sake of simplicity and consistency, the scheme illustrated in Algorithm 2 is based on NIPALS rather than on the Dayal-MacGregor algorithm [217] as originally reported in [183].

KNN-LW-PLS consists of three fundamental steps. First, the distance between a test sample and each training sample is calculated and a user-defined number of nearest neighbours, K, of such a test sample is selected. Then, these nearest neighbours are weighted according to their similarity (distance) to the test sample to be assessed using the following weighting function:

$$w_{\text{local},n} = \exp\left(-\frac{d_n^*}{h\,\sigma(\mathbf{d}^*)}\right) \tag{7.4}$$

where $\mathbf{d}^*$ is a vector containing all the max-normalized distance values calculated, d_n^* is the max-normalized distance computed for the n-th neighbour, σ denotes the standard deviation operator, and h represents a parameter that influences the shape of the weighting function. The higher h, the less d_n^* affects the weights. For infinite values of h, every selected calibration sample has the same weight.

In a nutshell, once the calibration subset has been identified, a weighted-PLS model is constructed as outlined below:

Algorithm 2: KNN-LW-PLS

Input: A predictor matrix $\mathbf{X}\,(N \times V)$, a response matrix $\mathbf{Y}\,(N \times M)$, and the observation related to an incoming test sample $\mathbf{x}_{\text{new}}\,(1 \times V)$.

1. Calculate the distances between $\mathbf{x}_{\text{new}}$ and all samples in the calibration set $\mathbf{X}$.
2. Select the K nearest neighbours of $\mathbf{x}_{\text{new}}$ and construct the calibration subsets $\mathbf{X}_{\text{sub}}$ and $\mathbf{Y}_{\text{sub}}$.
3. Assign the weight $w_{\text{local},n}$ to each n-th row of the matrix $\mathbf{X}_{\text{sub}}$:

$$\mathbf{w}_{\text{local},n} = \exp\left(-\frac{d_n^*}{h\,\sigma(d^*)}\right)$$

4. Calculate the diagonal matrix $\mathbf{D}$ by scaling and placing the elements of the vector $\mathbf{w}_{\text{local}}$ along its diagonal:

$$\mathbf{D} = \text{diag}(\mathbf{w}_{\text{local}}) \cdot \frac{1}{K}$$

5. Perform weighted mean-centering on $\mathbf{X}_{\text{sub}}$ and $\mathbf{Y}_{\text{sub}}$, with $\mathbf{1}$ being a vector of ones of appropriate size:

$$\mathbf{X}_{\text{sub}} = \mathbf{X}_{\text{sub}} - \mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{X}_{\text{sub}}$$

$$\mathbf{Y}_{\text{sub}} = \mathbf{Y}_{\text{sub}} - \mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{Y}_{\text{sub}}$$

6. Initialize $\mathbf{u}_a$ as a column of $\mathbf{Y}_{\text{sub}}$.
7. Calculate weighted $\mathbf{X}$-weights:

$$\mathbf{w}_a = \mathbf{X}_{\text{sub}}^T\mathbf{D}\mathbf{u}_a\,(\|\mathbf{X}_{\text{sub}}^T\mathbf{D}\mathbf{u}_a\|)^{-1}$$

8. Calculate $\mathbf{X}$-scores:

$$\mathbf{t}_a = \mathbf{X}_{\text{sub}}\mathbf{w}_a$$

9. Calculate $\mathbf{Y}$-loadings:

$$\mathbf{q}_a = \mathbf{Y}_{\text{sub}}^T \mathbf{D}\mathbf{t}_a \left(\mathbf{t}_a^T \mathbf{D}\mathbf{t}_a\right)^{-1}$$

10. Update $\mathbf{u}_a$:

$$\mathbf{u}_a = \mathbf{Y}_{\text{sub}}\mathbf{q}_a$$

11. Repeat steps 7 to 10 until $\mathbf{u}_a$ converges.
12. Calculate $\mathbf{X}$-loadings:

$$\mathbf{p}_a = \mathbf{X}_{\text{sub}}^T \mathbf{D}\mathbf{t}_a \left(\mathbf{t}_a^T \mathbf{D}\mathbf{t}_a\right)^{-1}$$

13. Deflate $\mathbf{X}_{\text{sub}}$ and $\mathbf{Y}_{\text{sub}}$:

$$\mathbf{X}_{\text{sub}} = \mathbf{X}_{\text{sub}} - \mathbf{t}_a \mathbf{p}_a^T$$

$$\mathbf{Y}_{\text{sub}} = \mathbf{Y}_{\text{sub}} - \mathbf{t}_a \mathbf{q}_a^T$$

14. Repeat steps 7 to 13 until all the required latent variables (A) are extracted.
15. Calculate regression coefficients:

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{Q}^T$$

16. Repeat steps 1 to 15 for each new test sample.

7.4.3 RoBoost-PLS

RoBoost-PLS [213, 214] is a variant of classical PLS proposed by Metz et al. with the aim of reducing the impact of outliers during the model calibration phase. RoBoost-PLS is initialised with weights identical for all the N investigated samples and equal to $1/N$. Once the first RoBoost-PLS latent variable or component has been extracted, each one of these weights is adjusted as:

$$w_{\text{RoBoost},n} = \frac{1}{N}\, g(\|\mathbf{e}_n\|, \alpha) \cdot \prod_{j=1}^{m} g(f_{n,j}, \beta) \cdot g(l_n, \gamma) \tag{7.5}$$

with $\|\cdot\|$ denoting the Euclidean norm, $\mathbf{e}_n$ being the $\mathbf{X}$-residual vector associated to the n-th calibration sample, $f_{n,j}$ the j-th element of the $\mathbf{Y}$-residual vector associated to the n-th calibration sample, l_n the leverage computed for the n-th calibration sample and g a bisquare function defined for the generic variable k as:

$$g(z_n) = \begin{cases} (1 - z_n^2)^2, & \text{for } |z_n| < 1, \\ 0, & \text{for } |z_n| \geq 1. \end{cases} \tag{7.6}$$

where:

$$z_n = \frac{k_n}{c\tilde{k}} \tag{7.7}$$

For the sake of clarity, when calculating $g(\|e_n\|, \alpha)$, then $k_n = \|\mathbf{e}_n\|$, $c = \alpha$ and $\tilde{k}$ represents the median of all the $\|\mathbf{e}_n\|$ values retrieved for the N samples under study. When calculating $g(f_{n,j}, \beta)$, instead, $k_n = f_{n,j}$, $c = \beta$ and $\tilde{k}$ connotes the median of all the $f_{n,j}$ values retrieved for the N samples under study. A similar reasoning holds also for $g(l_n, \gamma)$.

The bisquare function is one of the weighting functions commonly used in robust calibration methods [209]. This weighting scheme is based on the assumption that outliers tend to have extreme values in terms of residuals and/or leverage. By normalising these quantities with respect to their median distribution, observations that deviate significantly from the median will show normalized values greater than 1. As a result, they are effectively down weighted by the bisquare function (equation 7.6), which assigns zero weight.

It is important to notice here that:

1. such an operation of weight readjustment is conducted every time a

new RoBoost-PLS latent variable is derived,

2. the readjusted weights obtained for a given RoBoost-PLS component are used as weight initial estimates for the successive one, and
3. α, β, and γ are parameters of the function g that need to be tuned.

Algorithm 3: RoBoost-PLS

Input: A predictor matrix $\mathbf{X}\,(N \times V)$ and a response matrix $\mathbf{Y}\,(N \times M)$.

1. Initialize $\mathbf{u}_a$ as a column of $\mathbf{Y}$.
2. Assign an equal initial weight $w_{\mathrm{RoBoost},n}$ to each n-th row of the matrix $\mathbf{X}$ as:
$$w_{\mathrm{RoBoost},n} = \frac{1}{N}$$
3. Calculate the diagonal matrix $\mathbf{D}$ by placing the elements of the vector $\mathbf{w}_{\mathrm{RoBoost}}$ along its diagonal.
4. If $a = 1$, perform weighted mean-centering on $\mathbf{X}$ and $\mathbf{Y}$, with $\mathbf{1}$ being a vector of ones of appropriate size:
$$\mathbf{X} = \mathbf{X} - \mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{X}$$
$$\mathbf{Y} = \mathbf{Y} - \mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{Y}$$
5. Calculate weighted $\mathbf{X}$-weights:
$$\mathbf{w}_a = \mathbf{X}^T\mathbf{D}\mathbf{u}_a\,(\|\mathbf{X}^T\mathbf{D}\mathbf{u}_a\|)^{-1}$$
6. Calculate $\mathbf{X}$-scores:
$$\mathbf{t}_a = \mathbf{X}\mathbf{w}_a$$
7. Calculate $\mathbf{Y}$-loadings:
$$\mathbf{q}_a = \mathbf{Y}^T\mathbf{D}\mathbf{t}_a\,(\mathbf{t}_a^T\mathbf{D}\mathbf{t}_a)^{-1}$$
8. Update $\mathbf{u}_a$:
$$\mathbf{u}_a = \mathbf{Y}\mathbf{q}_a$$
9. Calculate a convergence parameter proposed by Metz et al. [214]:
$$\phi_a = \mathbf{u}_a\mathbf{D}\mathbf{t}_a\left(\mathbf{t}_a^T\mathbf{D}\mathbf{t}_a\right)^{-1}$$

10. Calculate **X**-loadings:

$$\mathbf{p}_a = \mathbf{X}^T\mathbf{D}\mathbf{t}_a \left(\mathbf{t}_a^T\mathbf{D}\mathbf{t}_a\right)^{-1}$$

11. Calculate **X**-residuals (**E**), **Y**-residuals (**F**) and leverage values (l):

$$\mathbf{E} = \mathbf{X} - \mathbf{t}_a\mathbf{p}_a^T$$

$$\mathbf{F} = \mathbf{Y} - \mathbf{t}_a\mathbf{q}_a^T$$

$$l = \mathbf{t}_a$$

12. Update the weight of each n-th calibration sample:

$$w_{\text{RoBoost},n} = \frac{1}{N} g(\|e_n\|, \alpha) \cdot \prod_{j=1}^{m} g(f_{n,j}, \beta) \cdot g(l_n, \gamma)$$

13. Update **D** accordingly.

$$\mathbf{D} = \text{diag}(\mathbf{w}_{\text{RoBoost}})$$

14. Repeat steps 4 to 13 until convergence of ϕ_a.
15. Deflate **X** and **Y**:

$$\mathbf{X} = \mathbf{X} - \mathbf{t}_a\mathbf{p}_a^T$$

$$\mathbf{Y} = \mathbf{Y} - \mathbf{t}_a\mathbf{q}_a^T$$

16. Repeat steps 5 to 15 until all the required latent variables (A) are extracted.
17. Calculate regression coefficients:

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{Q}^T$$

7.4.4 Locally-Weighted-RoBoost-PLS (LW-RoBoost-PLS)

Algorithmic scheme

LW-RoBoost-PLS relies on a synergistic combination of KNN-LW-PLS and RoBoost-PLS. This novel approach is mainly based on the algorithmic pipeline in Algorithm 2, but once the local calibration subset $\mathbf{X}_{\text{sub}}$ is defined, a RoBoost-PLS model (see Algorithm 3) is constructed instead of a standard PLS one. More specifically, LW-RoBoost-PLS encompasses the following 4 computational steps:

1. For each new sample whose responses are to be predicted, say $\mathbf{x}_{\text{new}}$, its distance from any calibration sample is calculated. This distance can be calculated in the original variable space, in the subspace of a certain number of principal components of $\mathbf{X}$, or directly within the subspace of a global PLS model trained on $\mathbf{X}$ and $\mathbf{Y}$.
2. The K closest neighbours of $\mathbf{x}_{\text{new}}$ are then identified and gathered in a new data matrix $\mathbf{X}_{\text{sub}}$. Their respective responses are also collected in a new data array $\mathbf{Y}_{\text{sub}}$.
3. The samples in $\mathbf{X}_{\text{sub}}$ are weighted according to their distance to $\mathbf{x}_{\text{new}}$.
4. RoBoost-PLS is applied to $\mathbf{X}_{\text{sub}}$ and $\mathbf{Y}_{\text{sub}}$. In this way, the local weights imposed in the previous step can be adjusted but not increased (to preserve the distance contribution) according to the degree of "outlyingness" of the corresponding observations. This step is crucial to mitigate the impact of possible abnormal samples present among the selected neighbours.

Algorithm 4: LW-RoBoost-PLS

Input: A predictor matrix $\mathbf{X}\,(N \times V)$, a response matrix $\mathbf{Y}\,(N \times M)$, and the observation related to an incoming test sample $\mathbf{x}_{\text{new}}\,(1 \times V)$.

1. Calculate the distances between $\mathbf{x}_{\text{new}}$ and all samples in the calibration set $\mathbf{X}$.
2. Select the K nearest neighbours of $\mathbf{x}_{\text{new}}$ and construct the calibration subsets $\mathbf{X}_{\text{sub}}$ and $\mathbf{Y}_{\text{sub}}$.
3. Assign the weight $w_{\text{local},n}$ to each n-th row of the matrix $\mathbf{X}_{\text{sub}}$:

$$w_{\text{local},n} = \exp\left(-\frac{d_n^*}{h\,\sigma(d^*)}\right)$$

4. Calculate the diagonal matrix $\mathbf{D}$ by scaling and placing the elements of the vector $\mathbf{w}_{\text{local}}$ along its diagonal:

$$\mathbf{D} = \text{diag}(\mathbf{w}_{\text{local}}) \cdot \frac{1}{K}$$

5. If $a = 1$, perform weighted mean-centering on $\mathbf{X}_{\text{sub}}$ and $\mathbf{Y}_{\text{sub}}$:

$$\mathbf{X}_{\text{sub}} = \mathbf{X}_{\text{sub}} - \mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{X}_{\text{sub}}$$

$$\mathbf{Y}_{\text{sub}} = \mathbf{Y}_{\text{sub}} - \mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{Y}_{\text{sub}}$$

6. Initialize $\mathbf{u}_a$ as a column of $\mathbf{Y}$.
7. Calculate weighted $\mathbf{X}$-weights:

$$\mathbf{w}_a = \mathbf{X}_{\text{sub}}^T\mathbf{D}\mathbf{u}_a\,(\|\mathbf{X}_{\text{sub}}^T\mathbf{D}\mathbf{u}_a\|)^{-1}$$

8. Calculate $\mathbf{X}$-scores:

$$\mathbf{t}_a = \mathbf{X}_{\text{sub}}\mathbf{w}_a$$

9. Calculate $\mathbf{Y}$-loadings:

$$\mathbf{q}_a = \mathbf{Y}_{\text{sub}}^T \mathbf{D} \mathbf{t}_a \left(\mathbf{t}_a^T \mathbf{D} \mathbf{t}_a\right)^{-1}$$

10. Update $\mathbf{u}_a$:

$$\mathbf{u}_a = \mathbf{Y}_{\text{sub}} \mathbf{q}_a$$

11. Calculate a convergence parameter ϕ_a as proposed by Metz et al. [214]:

$$\phi_a = \mathbf{u}_a \mathbf{D} \mathbf{t}_a \left(\mathbf{t}_a^T \mathbf{D} \mathbf{t}_a\right)^{-1}$$

12. Calculate $\mathbf{X}$-loadings:

$$\mathbf{p}_a = \mathbf{X}_{\text{sub}}^T \mathbf{D} \mathbf{t}_a \left(\mathbf{t}_a^T \mathbf{D} \mathbf{t}_a\right)^{-1}$$

13. Calculate $\mathbf{X}$-residuals ($\mathbf{E}$), $\mathbf{Y}$-residuals ($\mathbf{F}$), and leverage values ($\mathbf{l}$):

$$\mathbf{E} = \mathbf{X}_{\text{sub}} - \mathbf{t}_a \mathbf{p}_a^T$$

$$\mathbf{F} = \mathbf{Y}_{\text{sub}} - \mathbf{t}_a \mathbf{q}_a^T$$

$$\mathbf{l} = \mathbf{t}_a$$

14. Update the weights according to the degree of "outlyingness" of each n-th calibration sample:

$$w_{\text{RoBoost},n} = \frac{1}{N} g(\|\mathbf{e}_n\|, \alpha) \cdot \prod_{j=1}^{m} g(f_{n,j}, \beta) \cdot g(l_n, \gamma)$$

15. Apply the following condition:

$$w_{\text{RoBoost},n} = w_{\text{local},n} \quad \text{if } w_{\text{RoBoost},n} > w_{\text{local},n}$$

16. Update $\mathbf{D}$ accordingly:

$$\mathbf{D} = \text{diag}(\mathbf{w}_{\text{RoBoost}})$$

17. Repeat steps 5 to 16 until convergence of ϕ_a.

18. Deflate $\mathbf{X}_{\text{sub}}$ and $\mathbf{Y}_{\text{sub}}$:

$$\mathbf{X}_{\text{sub}} = \mathbf{X}_{\text{sub}} - \mathbf{t}_a \mathbf{p}_a^T$$

$$\mathbf{Y}_{\text{sub}} = \mathbf{Y}_{\text{sub}} - \mathbf{t}_a \mathbf{q}_a^T$$

19. Repeat steps 6 to 18 until all the required latent variables (A) are extracted.
20. Calculate regression coefficients:

$$\mathbf{B} = \mathbf{W} \left(\mathbf{P}^T \mathbf{W}\right)^{-1} \mathbf{Q}^T$$

21. Repeat steps 1 to 20 for each new test sample.

Parameter Tuning

Several parameters need to be fine-tuned to obtain an optimal LW-RoBoost-PLS model: the number of latent variables to extract, the number of nearest neighbours, K, as well as the values of h, α, β and γ. For this purpose, in this study, a double cross-validation procedure [218, 219], consisting of two nested loops, was implemented: for each LW-RoBoost-PLS component, the inner loop was exploited to find the values of α, β and γ minimising the root median square error related to the responses of interest and defined as:

$$\text{RMedSECV} = \sqrt{\text{median}\left[(y_{\text{CV}} - y_{\text{true}})^2\right]} \tag{7.8}$$

where y_{CV} is a vector containing the response values predicted in cross-validation and y_{true} carries the reference response values.

The use of RMedSECV mitigates, in fact, the impact of potential outliers present in the calibration set. On the other hand, the outer loop was utilised to set h and K given the optimised α, β and γ. A Måge plot [181] was ultimately resorted to for determining the final complexity of the

LW-RoBoost-PLS model under study, i.e. the number of LW-RoBoost-PLS latent variables associated to the best combination of h, K, α, β and γ.

It has to be noticed here that, in order to guarantee the statistical rigour of the developed cross-validation approach, the data splitting scheme adopted here was designed so as to ensure the independence of the subsets on which the inner and outer loops were run, respectively. . A schematic representation of the cross-validation strategy is provided in appendix C.

7.4.5 Model Performance

The performance of the models resulting from the different approaches compared was evaluated in terms of root mean square error in external validation (RMSEP). It is important to note that the evaluation assumes that there are no outliers in the test set.

$$\text{RMSEP} = \sqrt{\frac{\sum_{n=1}^{N} \left(y_{n,\text{predicted}} - y_{n,\text{true}}\right)^2}{N}} \tag{7.9}$$

and bias:

$$\text{bias} = \frac{\sum_{n=1}^{N} \left(y_{n,\text{predicted}} - y_{n,\text{true}}\right)}{N} \tag{7.10}$$

7.5 Results and Discussion

In this section, we compare the performance of LW-RoBoost-PLS with that of RoBoost-PLS and KNN-LW-PLS, focusing also on the way in which the different approaches assign weights to the calibration samples.

7.5.1 Simulated Data

The optimal KNN-LW-PLS and RoBoost-PLS parameter combinations were determined through a 10-split Venetian blind cross-validation approach, while for LW-RoBoost-PLS the double cross-validation procedure described in Section 7.4.4 was run with 5 data splits for both the inner and the outer loop. α, β and γ were all varied in the interval $[3-7]$ with increments of 2 units, which resulted in 27 possible combinations to test.

On the other hand, the range $[30-100]$ with increments of 10 units and the values 1 and 2 were investigated for the local parameters h and K (for a total number of 16 possible combinations). Table 7.1 summarises the outcomes obtained for the three models optimised as detailed before, while figure 7.2 displays their corresponding prediction plots related to the external validation.

Table 7.1 – Parameter settings of the compared models. LVs denotes the number of latent variables.

Model	**Local Parameters** (k, h)	**RoBoost Parameters** (α, β, γ)	**LVs**	**RMSEP**	**Bias**
RoBoost-PLS	-	5, 3, 3	1	0.081	-0.035
KNN-LW-PLS	100, 1	-	2	0.062	0.097
LW-RoBoost-PLS	70, 2	3, 7, 5	2	0.049	0.000

Figure 7.3, instead, represents the weights assigned to all the training samples by the three different methods compared in the present chapter (mind that for KNN-LW-PLS and LW-RoBoost-PLS only the weights of a single local model are given). As one can clearly see, setting α, β and γ at 5, 3 and 3 allows RoBoost-PLS to correctly downweigh all the anomalous observations which were included in the calibration set (see Figure 7.3a). It is worth noticing here that lower values of these parameters would have led to a more severe outlier detection, increasing the number of data items recognised as anomalous. However, being RoBoost-PLS based on a linear modelling strategy, as expected, the prediction plot it returns exhibits an evident nonlinear trend (see Figure 7.2a).

Conversely, by selecting a calibration subset of 100 neighbours per test sample (in this case, the distance values were calculated in the subspace of the first two principal components of $\mathbf{X}$, explaining approximately 98% of the

variance of the entire calibration data), KNN-LW-PLS easily accommodates non-linearities, which is reflected by the fact that it leads to a lower RMSEP than RoBoost-PLS. Moreover, the nonlinear trend originally observed in Figure 7.2a is significantly reduced when KNN-LW-PLS is exploited (see Figure 7.2b). Nonetheless, the weights estimated by KNN-LW-PLS do not readily consider the existence of outliers among the k-nearest neighbours identified for each test observation (see, as an example, Figure 7.3b which refers to the sample denoted with a red triangle in Figure 7.2b). This might be the root cause inducing the higher prediction bias affecting KNN-LW-PLS.

Finally, LW-RoBoost-PLS permits achieving a zero bias by accurately down-weighing the outliers in the local calibration subsets selected. In their optimal combination, α, β, and γ were here found to be equal to 3, 7, and 5, respectively. Overall, these higher values compared to those resulting from the application of RoBoost-PLS may originate from the smaller number of outlying samples with which LW-RoBoost-PLS has to deal iteratively when each of the aforementioned calibration subsets is handled. In addition to its inherent robustness against local outliers, LW-RoBoost-PLS can directly cope with the nonlinearities simulated in this circumstance (see also Figure 7.2c), yielding the lowest RMSEP value among the three methodologies under study.

Furthermore, having a look at the first LW-RoBoost-PLS latent variable weights for a specific test sample (red triangle in Figure 7.2c), it can be observed how this technique reduces the influence of samples exhibiting outlying behaviours even if they were chosen as members of the local calibration subsets. This is made possible by step 14 of Algorithm 4, which decreases and even zeroes (see the yellow stars in Figure 7.3c) the initial local weights imposed based only on the distance between the training observations and this test sample (blue stars in Figure 7.3c), provided that such training observations show abnormal $\mathbf{X}$-residuals, $\mathbf{y}$-residuals, and/or leverage (see, for instance, the red circles in Figure 7.3c for the sake of illustration, the spectral profiles and the $\mathbf{y}$-values simulated for these 4 samples are displayed in Figure 7.4 together with those of the rest of the calibration subset considered

in this specific contingency).

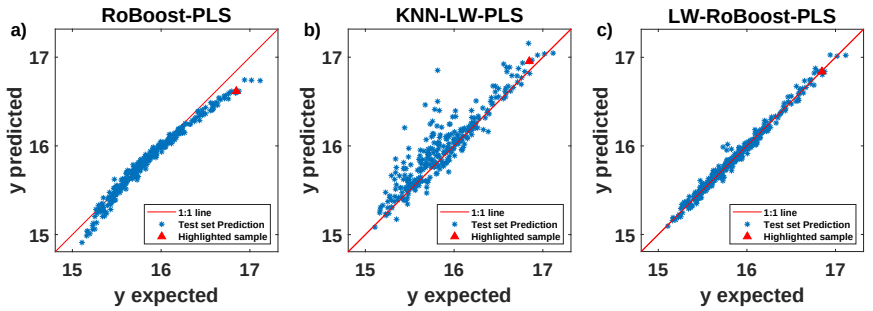


Figure 7.2 – Simulated data: predicted **y**-values versus measured **y**-values plots resulting from the application of the optimal **a)** RoBoost-PLS, **b)** KNN-LW-PLS and **c)** LW-RoBoost-PLS models. The displayed predictions relate to the samples of the external validation (test) set. For comparison, the predictions obtained using a PLS model are reported in appendix C.

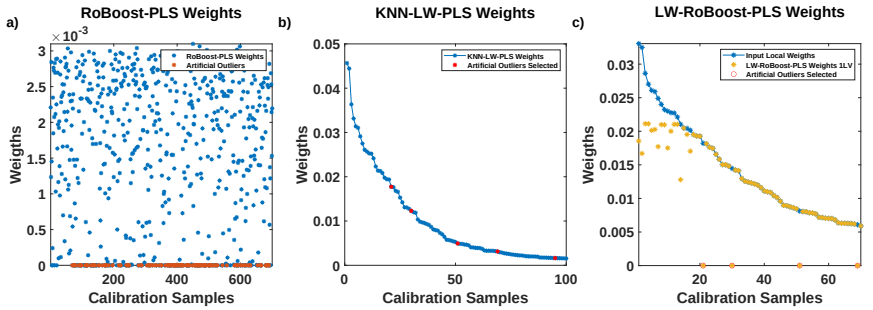


Figure 7.3 – Simulated data: calibration sample weights estimated by **a)** RoBoost-PLS, **b)** KNN-LW-PLS and **c)** LW-RoBoost-PLS. Notice that for KNN-LW-PLS and LW-RoBoost-PLS only the weights of the local models constructed for the sample denoted with a red triangle in Figures 2b and 2c are given. In Figure 3c, the blue stars represent the initial weights assigned based on the distance between the training observations and this test sample, while the yellow ones correspond to the final weights calculated at the end of the LW-RoBoost-PLS computational procedure.

7.5.2 ABS data

The ABS data analysed in chapter 5 were analysed using multiblock and local predictive approaches in an attempt to determine the influence of each step in the production process on the quality of the copolymer produced. These included data cleaning, outlier removal and response linearisation

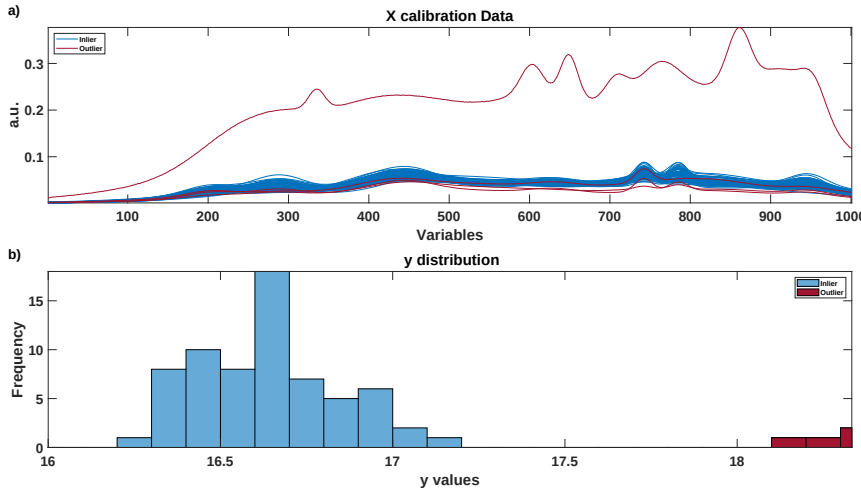


Figure 7.4 – Simulated data: **a)** pseudo-spectral profiles and **b)** **y**-values of the samples belonging the local calibration subset identified for the test observation denoted with a red triangle in Figure 2b and 2c. Aberrant behaviours can be observed for the samples to which LW-RoBoost-PLS finally assigns a zero weight (see red solid lines and bars).

steps, which were skipped here in order to test the three methods under investigation in a rather complex industrial scenario. The raw data collected in the plant and merged as outlined in section 5.2.1 were then fed directly into RoBoost-PLS, KNN-LW-PLS and LW-RoBoost-PLS for comparison purposes.

It has to be noticed here that KNN-LW-PLS was replaced by K-Nearest-Neighbours-Locally-Weighted-Multiblock-PLS (KNN-LW-MB-PLS), applied keeping the nearest neighbours identified for each test sample unchanged across all the investigated data blocks of course, the same modification was carried out also for LW-RoBoost-PLS, at least when steps 1 to 3 of Algorithm 4 were concerned and that such data blocks underwent distinct preprocessing operations, i.e. NIR spectra were pretreated by Standard Normal Variate (SNV [81]) and mean centering, while the measurements resulting from each process sensor employed along all the duration of the production process were autoscaled. Prior to multiblock data analysis, the single blocks were finally scaled to unit block variance. In order to keep the calibration and cross-validation sets truly independent, preprocessing was

performed after the cross-validation data splitting.

As for the simulated data, we report in Table 7.2 the outcomes related to the prediction of the response values associated with the external validation set samples returned by a RoBoost-PLS, a KNN-LW-MB-PLS, and a LW-RoBoost-PLS model optimised as outlined in Section 7.5.1.

Table 7.2 – Parameter settings of the compared multiblock (MB) models (for the sake of simplicity, MB has been omitted from the algorithms' acronyms in the first column). LVs denotes the number of latent variables.

Model	**Local Parameters** (k, h)	**RoBoost Parameters** (α, β, γ)	**LVs**	**RMSEP (g)**	**Bias (g)**
RoBoost-PLS	-	5, 5, 5	4	0.85	-0.39
KNN-LW-PLS	200, 2	-	4	0.82	0.16
LW-RoBoost-PLS	200, 1	7, 3, 7	5	0.64	-0.05

RoBoost-PLS extracts in total 4 LVs, but the retrieved combination of α, β, and γ values (all set equal to 5) in this case seems not to effectively decrease the influence of all the outliers present in the calibration data as several predicted $\mathbf{y}$-values it yields are far away from the ideal 1:1 fitting line (see Figure 7.5a).

On the other hand, KNN-LW-MB-PLS requires 200 neighbours and 4 LVs to build local models that allow for a significant improvement in the prediction quality (see, for instance, the blue dots in Figure 7.5b) and guarantee a more reasonable $\mathbf{y}$-residuals distribution (results not shown).

Last but not least, LW-RoBoost-PLS outperforms both RoBoost-PLS and KNN-LW-MB-PLS in terms of predictive ability (it, indeed, returns the lowest RMSEP), as also the prediction plot in Figure 7.5c highlights. At a closer look, it is also possible to notice how LW-RoBoost-PLS is capable of reducing the prediction bias affecting specifically the samples of product (grade) 5 and 8 as well as the dispersion observed when KNN-LW-MB-PLS was exploited, for example, for product 3 at high $\mathbf{y}$-values.

In addition, Figure 7.6 displays the weights assigned to all the training samples by the three considered approaches (once again, for KNN-LW-MB-PLS and LW-RoBoost-PLS only the weights of a single local model, the one constructed for the test sample denoted with a red triangle in Figures 7.5b and 7.5c, are given).

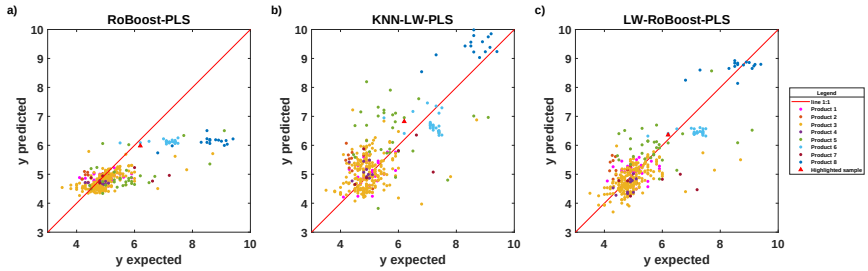


Figure 7.5 – ABS data: predicted **y**-values versus measured **y**-values plots resulting from the application of the optimal **a)** RoBoost-PLS, **b)** KNN-LW-PLS and **c)** LW-RoBoost-PLS models. The displayed predictions relate to the samples of the external validation (test) set. The colour coding reflects the manufactured ABS grade.For comparison, the predictions obtained using a PLS model are reported in appendix C

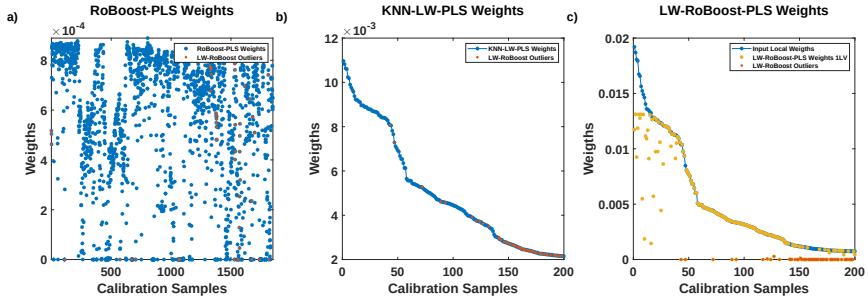


Figure 7.6 – ABS data: calibration sample weights estimated by **a)** RoBoost-PLS, **b)** KNN-LW-PLS and **c)** LW-RoBoost-PLS. Notice that for KNN-LW-PLS and LW-RoBoost-PLS only the weights of the local models constructed for the sample denoted with a red triangle in Figures 5b and 5c are given. In Figure 5c, the blue stars represent the initial weights assigned based on the distance between the training observations and this test sample, while the yellow ones correspond to the final weights calculated at the end of the LW-RoBoost-PLS computational procedure. In Figures 5a and 5b, the weights of the samples identified as outliers by LW-RoBoost-PLS (i.e. for which the LW-RoBoost-PLS weight was found to be approximately zero) are represented as red dots.

In this circumstance, LW-RoBoost-PLS manages to zero some of the weights mistakenly kept higher by both RoBoost-PLS and KNN-LW-MB-PLS (see the red dots in Figures 7.6a and 7.6b) and evidently associated with observations characterised by aberrant NIR spectral profiles and/or **y**-values (see Figure 7.7).

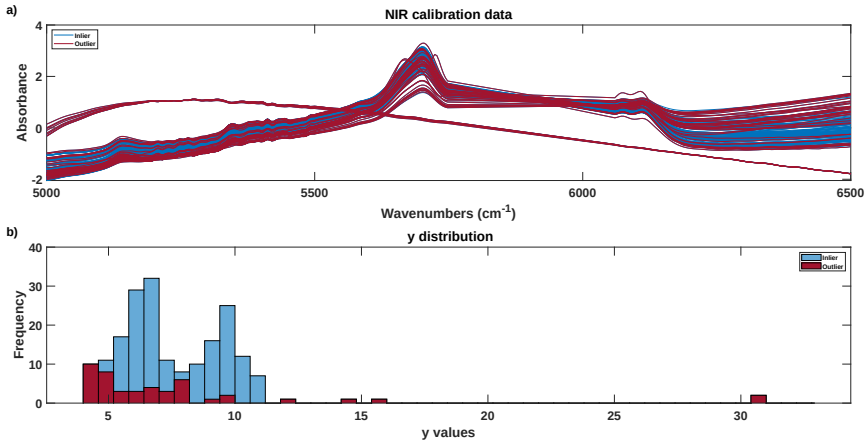


Figure 7.7 – ABS data: **a)** spectral profiles and **b)** **y**-values of the samples belonging to the local calibration subset identified for the test observation denoted with a red triangle in Figure 5b and 5c. Aberrant behaviours can be observed for most samples to which LW-RoBoost-PLS finally assigns a zero weight (see red solid lines and bars).

7.6 Conclusion

We proposed a novel PLS-based multivariate calibration approach capable of handling non-linearities between predictors and responses while mitigating the detrimental influence of outliers during the calibration of a regression model. This approach, named LW-RoBoost-PLS, relies on a double weighting scheme resulting from a rational combination of the operational principles of KNN-LW-PLS and RoBoost-PLS. This combination is the key feature that makes it possible to simultaneously overcome the challenges posed by non-linearities and outliers. LW-RoBoost-PLS was here tested on both simulated and complex real-world data and was found to outperform both the native methods from which it originates in terms of predictive power when the data under study exhibited strong non-linear variable intercorrelations and contained aberrant observations in $\mathbf{X}$ and/or $\mathbf{Y}$. More specifically, LW-RoBoost-PLS proved to be particularly suitable for real-time prediction in industrial scenarios where the collected measurements are usually characterised by complex non-linearities related to seasonal variations or production scale-ups and by the presence of severe outliers due to

sensor failures and/or drifts.

Regarding the choice of hyperparameter ranges, in this study the tested values were chosen based on prior knowledge and empirical experience with similar models. In general, for the robust parameters (α, β, γ), higher values are typically used when a low number of outliers is expected, while lower values are preferred when a higher contamination is expected. If no prior information is available, it is advisable to include both low and high values in the grid search to ensure robustness across different scenarios. For local parameters, the number of neighbours k should typically range from a few tens to a few hundreds, depending on the size and structure of the dataset.

Nonetheless, for a full on-line implementation, several aspects will have to be investigated in the future. One of them regards the detection and identification of abnormal test observations for which a computational solution is currently being explored and will ideally be described in a forthcoming publication.

IV

General discussion and conclusions

Chapter 8

Conclusion

The focus of this doctoral thesis has been on the use and development of the chemometric approach in a competition of Industry 4.0, which promotes the use of data to improve the industrial process under several aspects, efficiency, achieving more green production for example reducing the waste and keeping data-driven decision. In relation to the established partial objectives, this discussion chapter is divided into two different parts, a first one based on the case the production process of *pesto alla genovese* at Barilla G. e R. Fratelli S.p.A. and the second one focused on obtaining real-time information on product quality trends for the production of acrylonitrile butadiene styrene (ABS) at Versalis S.p.A..

Food Industry

This section, which focuses on the production of *pesto alla genovese*, highlights the challenges that a food industry may face in its transition to Industry 4.0. In particular, the **chapter 3**, the production process was examined from the perspective of process monitoring and real-time product quality assessment.

This task proved to be very challenging, due to the lack of historical data

and the absence of systematic storage of process sensor data. In addition, the process is predominantly continuous, characterised by multiple tanks and frequent pauses, stops and restarts. Minor variations in flow or stop times added complexity to the synchronisation phase. In order to address and mitigate the effects introduced by unwanted sources of variability, the choice of pre-processing tools was also challenging, leading to the use of non-traditional spectral pre-processing.

Despite these challenges:

- the multivariate control charts showed sensitivity to process variations, such as restarts after temporary interruptions, and were effective in indicating when the process had returned to normal operating conditions (NOC).
- The predictive models developed to estimate the quality of the pesto provided results that were considered acceptable by the company, highlighting their potential as tools for real-time quality monitoring of the finished product.

Ultimately, this preliminary study provides a basis for defining strategies to support and improve the transition to Industry 4.0. For example, it identified the value of installing NIR sensors to monitor the process, as well as the need to improve existing sensors, such as the RGB camera used to monitor the quality of raw basil. In addition, the improvement of process data storage and retrieval systems was highlighted as essential, along with the automatic registration of supplementary information.

Chapter 4 further explores how to improve the potential of the RGB camera by comparing different approaches to image analysis. The comparison between deep learning models and traditional chemometric approaches revealed both the advantages and limitations of each method. Deep learning demonstrated superior accuracy, but at a significant cost in terms of computational resources and development time. On the other hand, chemometrics based approaches, although less accurate, provided a more cost-effective and interpretable alternative. This highlights the potential for tailoring the choice of methods to specific needs, balancing performance, cost and inter-

pretability.

Multiblock, non linearities and outliers

This section focuses on three key areas: multi-block regression, handling non-linearities in multi-block frameworks, and simultaneous handling of non-linearities and outliers using ABS production as a case study. Each of these areas represents a significant step forward in the field of process monitoring and quality control.

- The use of **multiblock regression** methods, such as MB-PLS and ROSA, has demonstrated the potential of integrating data from multiple sensors to improve real-time quality prediction. By capturing complementary information that is often overlooked in traditional single-block approaches. In the case study of styrene polymer production, this approach has shown how the co-operation of multiple sensors from different parts of the process helps to predict quality parameters. In addition, the use of ROSA, which selects information only from specific blocks, also makes it possible to understand which blocks are most influential in predicting the properties under investigation. For example, it was found that NIR sensors were influential in predicting property 2 and not 1.
- To address the challenge of nonlinearities, two innovative extensions of multiblock regression have been proposed: Locally Weighted Multiblock PLS (LW-MB-PLS) and Locally Weighted ROSA (LW-ROSA). These methods represent a new direction in multiblock modelling, combining the effectiveness of local approaches with the advantages of multiblock frameworks. LW-MB-PLS extends MB-PLS by incorporating a local weighting scheme that allows the model to adapt to non-linear patterns that occur under different product types or process conditions. This method has proven effective in reducing prediction bias for some ABS products, highlighting its suitability for scenarios where non-linearities play a key role. On the other hand, LW-ROSA is based on the ROSA framework, which is also based on a weighting

scheme to deal with non-linearities. It is built on a similar pipeline to LW-MB-PLS, but retains the same advantages of ROSA, i.e. independence between blocks. However, while this method was successful in reducing bias in the simulated data, it did not provide a significant advantage in the real case, probably due to a sub-optimal weighting scheme. Taken together, LW-MB-PLS and LW-ROSA demonstrate the importance of local strategies in a multiblocks framework especially when the global (linear) version of is not able to provide a $\mathbf{X}$ $\mathbf{Y}$ relation that is valid for all the products.

A further step forward in the context of local models has been made in terms of interpretability. In general, one of the disadvantages of local models is that they have to be built for each new point to be predicted, so in addition to the computational cost, there is a loss of interpretability. To overcome the latter limitation, a strategy for interpreting local models based on the visualisation of model parameters is presented. This allows a deeper understanding of how different local models use information in different ways, an area that has not been addressed in the literature. For example, this work has been used to understand how information from different blocks contributes differently to different products.

- A third method, Locally-Weighted RoBoost-PLS (LW-RoBoost-PLS), was introduced to simultaneously deal with non-linearities and mitigate the negative impact of outliers. This method combines the principles of KNN-LW-PLS and RoBoost-PLS, resulting in a double weighting scheme, one to handle the non-linearities and later one to prevent outliers from compromising the calibration model. Tested on both simulated and real data, LW-RoBoost-PLS demonstrated superior predictive performance in scenarios characterised by non-linear dependencies and strong outliers, making it suitable for real-time applications in complex industrial environments.

The present study attempted to make a significant leap forward in the area of multiblock analysis. Crucial aspects that had not yet been addressed were analysed and this led to a number of important results.

8.1 Future perspective

As indicated in the previous paragraph, with regard to the process of *pesto alla genovese*, a step forward could be made by recalibrating the models with an enlarged dataset of historical data and, above all, being able to integrate also the information extracted from the monitoring of basil through the RGB camera.

As for the local models, in all the Euclidean distance has been used in a PCA space, in the literature there are several studies where in terms of real-time predictions, it is suggested to use different distances to calculate the dissimilarity among the samples [197, 220, 221]. Furthermore, compared to the traditional LW-PLS pipeline taken as reference, there are various modifications in the literature [222] and it would be worth testing if they are integrable in a multiblock context to improve computational efficiency, generally the slow phase is the definition of the calibration subset. Finally, the most challenging, due to the initial stage of development, is the improvement of LW-ROSA.

Local models have the ability to dynamically adapt to changing conditions to ensure optimal performance [223]. This means that, given the time-varying nature of industrial processes, local models can adapt to new conditions based on the historical data used for training. However, an aspect not addressed in this work is how these models could be extended to accommodate entirely new setups not previously described in the training data. This opens up opportunities for future research, particularly in exploring the integration of recursive or incremental modelling approaches [198] to allow continuous updating and adaptability to evolving process scenarios. In addition, a significant challenge lies in determining whether a new condition falls within NOC or represents an anomaly.

To conclude, as the topic of the thesis is Industry 4.0, the proposed tools have only been tested in this context, but it would be worth testing them in other contexts.

Chapter 9

Acknowledgements

Now that I have reached the end of this journey, which began just over three years ago, and I'm finally turning a corner, it is time to thank all the people who have made this journey possible.

Fortunately, I have had the chance to meet many colleagues and friends who have shaped all this time, so being very difficult to decide the order, I will leave to my friend MATLAB (sorry Python users) to decide it.

```
load('people.mat', 'people_list',
    'acknowledgements_list');

% 1. Sort the list in alphabetical order
[sorted_list, sort_index] = sort(people_list);

% Sort acknowledgements accordingly
sorted_acknowledgements =
    acknowledgements_list(sort_index);

% 2. Shuffle the list randomly
random_index = randperm(length(sorted_list));
random_people = sorted_list(random_index);
random_acknowledgements =
    sorted_acknowledgements(random_index);

```

```
% 3. Loop through each person and display their name
    and acknowledgement
for i = 1:length(random_people)
    fprintf('%s: %s\n', random_people{i},
        random_acknowledgements{i});
    pause
end
```

» I would like to express my sincere gratitude to the Modena group and especially to Caterina, Samuele, Alessandro and Lorenzo. Caterina, thank you for giving me the opportunity to grow through different experiences, both in the laboratory and in teaching. Your trust and support in difficult moments were invaluable. Samueli, thank you for being the best lab brother I could have wished for. Alessandro, thank you for keeping me well-fed with countless jars of pesto and for sharing the exploration of southern France with me. I am grateful for your kindness and courtesy. Lorenzo, thank you for being not only a mentor but also a true friend. Thank you for the chess breaks and for always taking the time to support me when things got tough.

» Thank you, Francesca (🐤), for you this whole thesis would not be enough, nor would a whole chapter of thanks. Without you I would not have been able to finish anything: you have accompanied me and shared this journey with me, as well as bearing the worst of it. Thank you for celebrating with me all the successes, but above all, thank you for allowing me to face the difficulties, for understanding me and making my weaknesses stronger by ensuring that they were not limitations. I am sure I could not have done it without you. Any applause at the end of a journey needs two hands: thank you from the bottom of my heart for being the other essential hand that made it possible. ♡

» I would like to sincerely thank my supervisors, Cyril, Raffaele and Marina, for giving me the opportunity to embark on and complete this journey. Each of you has contributed to my professional development in unique and essential ways, helping me form the foundation of the researcher I hope to be one day. I am particularly grateful for the way in which you have gone beyond academic guidance to establish a

connection outside of work, which has certainly been of great personal value to me.

» To Dalelux for a hope

» A big thank you to my family: my parents, Federica, Luca, Nonna Margherita, and zio Mariano. There is no need to explain why, each of you already knows. Your affection and your unwavering belief in me have been fundamental to fulfilling all the goals I have set for myself. A special mention goes to Grandma for her endless supply of *polpette* and *limoncello*.

» I would also like to thank the *Ghega office*, or rather its members, for making Modena feel like home.

A special mention goes to Alessio, for accompanying me through the various stages of my life between bulking and cutting; to Biagio, Valeria (ti voglio bene) and Simone, for the dinners on the theme of southern Italy, always marked by "*se voglio*" and "*mossa del puma*"; to Lorenzo, for being the true source of inspiration behind Loppo GPT project ("Would you rather...") , to Mirco for being an extremely kind director but above all for becoming the new Matlab expert in the office; and to Manuel, for always bringing good vibes.

A heartfelt thank you to Alessandra, a worthy travel companion: the strength and perseverance that characterised this journey were certainly nourished by our common roots, *Abruzzesi*.

I would like to take this opportunity to remind you of the time I fainted (Alessio, I told you that I shouldn't have done the bench press during the cutting phase), and although you did not worry about me in the slightest, even with a message, I can tell you that the dark moment I went through did not affect our friendship, but it might do so now, because I have never actually fainted

» A special thanks goes to all the scientists I have collaborated with during these years at the University of Modena and Reggio Emilia, the University of Lille, the University of the Basque Country and ChemHouse in Montpellier. I am also very grateful to Federico Marini for

introducing me to chemometrics and getting me interested.

» A heartfelt thank you also goes to all my friends: Puppi, Ciao Luca, Linda, Samvele, Silvia and Francesco, Peppe, Alessandro and Roberta, Momo, Adri, Glenn, Lorenza and Francesco, for being by my side throughout this journey.

» Finally, special thanks to Eugy various and sundry

```
??? Error using ==> life_experience Line 42
Index exceeds matrix dimensions.
```

V

Appendix

Appendix A

Multiblock-PLS results on Versalis Plant

This appendix presents the results of the multiblock-PLS (MB-PLS) model using all the available blocks. The MB-PLS model required optimization of the number of latent variables (LVs), which was performed using 10-fold cross-validation with a *venetian blind* split. The optimal model was determined with 9 latent variables (LVs), corresponding to a minimum RMSECV of 0.63 g.

The resulting model achieved an RMSEP of 0.82 g. Figure A.1a shows the distribution of predictions for the validation set, with all objects falling within the expected QP1 range. Figure A.1b shows the same distribution, colour coded to represent eight different ABS products, each occupying a different QP1 range. In particular, the predictions show an irregular distribution around the 1:1 line (red line). Product 1 (yellow) shows less dispersion compared to product 7 (blue) at higher QP1 levels. For the latter, the predictions for the test set are unsatisfactory and show two distinct clusters, one overestimated and the other underestimated.

In Figure A.2, the model regression coefficients are presented. All process sensor blocks were identified as important contributors to the prediction of QP1. Regarding the NIR blocks, each of them also showed spectral regions

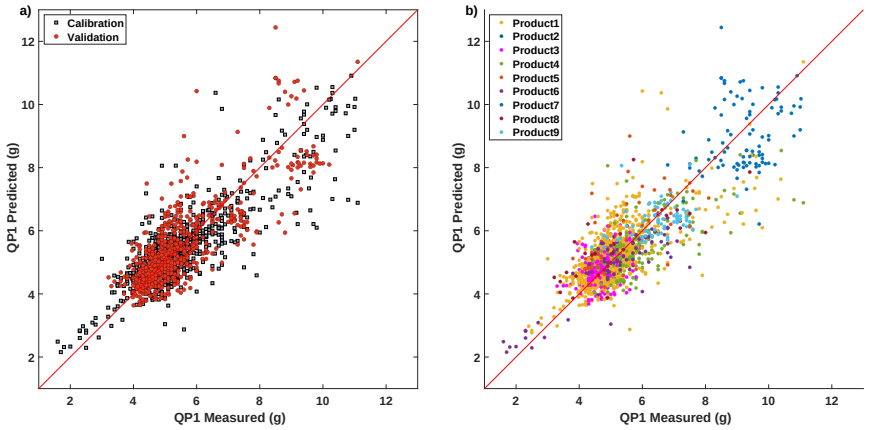


Figure A.1 – Plots of predicted vs measured values of QP1 obtained by the MB-PLS model using all the available blocks. In **a)** Samples are coloured according to calibration (gray) and validation (red) and in **b)** according to ABS product type.

that are relevant for the prediction. In particular, the *NIR Reaction* block highlights a key region between 5400 cm^{-1} and 5250 cm^{-1}, which can be attributed to the stretching vibration of a functional group from one of the three precursor compounds. For the other NIR blocks, relevant regions of interest were observed in the absorption bands between 6000 cm^{-1} and 6500 cm^{-1}. These results emphasize the complementary contributions of process sensors and NIR spectroscopy in capturing the critical information required for accurate predictions.

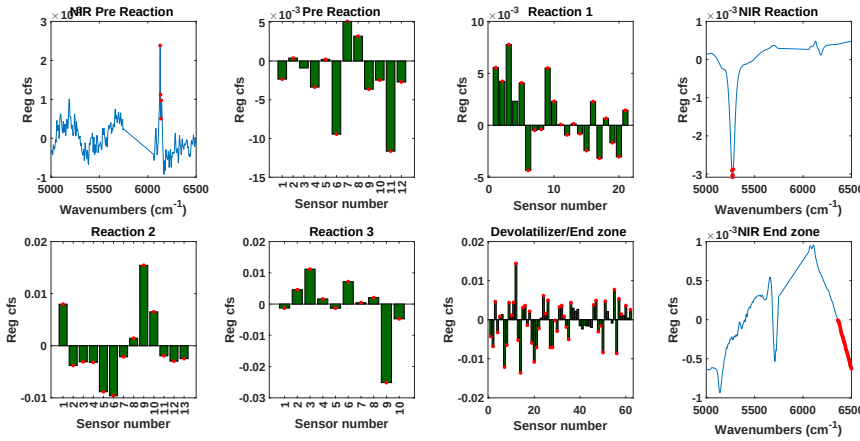


Figure A.2 – Regression coefficients resulting from the MB-PLS model for each data block, different block name is reported on top. Red stars indicate variables exhibiting VIP scores higher than one.

Appendix B

LW-ROSA Algorithm

The following scheme shows the main steps of the LW-ROSA algorithm

Algorithm 5: LW-ROSA

Input: A cell array `predictors` containing B predictor matrices $\mathbf{X}_b$ $(N \times V_b)$;

A response matrix $\mathbf{y}$ $(N \times 1)$;

A cell array `test_sample` containing B test blocks $\mathbf{x}_{\text{new},b}$ $(1 \times V_b)$;

(Optional) The type of weight function `weight_func` (e.g., `gaussian`);

(Optional) A parameter h for the Gaussian weight function;

(Optional) Whether to truncate weights using `LeakyReLU` and the percentile threshold.

1. Compute distances and weights

   for $b \leftarrow 1$ **to** B **do**

   a) Calculate the distances between $\mathbf{x}_{new}$ and all samples in the calibration set $\mathbf{X}_b$

   b) Compute the weights w_b for the b-th block according to `weight_func` and parameter h;

   c) Optional weight truncation with LeakyReLU:

$$\mathbf{w}_b' = \begin{cases} \mathbf{w}_b & \text{if } \mathbf{w}_b \leq w_{b,\text{threshold}} \\ \alpha \cdot \mathbf{w}_b & \text{if } \mathbf{w}_b > \ w_{b,\text{threshold}} \end{cases}$$

d) Calculate the diagonal matrix $\mathbf{D}_b$ by scaling and placing the elements of the vector $\mathbf{w}_{b,\text{local}}$ along its diagonal.

2. Perform weighted mean-centering on $\mathbf{X}_\text{b}$ and $\mathbf{y}$, with $\mathbf{1}$ being a vector of ones of appropriate size:
   for $b \leftarrow 1$ **to** B **do**

$$\mathbf{X}_\text{b} = \mathbf{X}_\text{b} - \mathbf{1}\mathbf{1}^T\mathbf{D}_b\mathbf{X}_\text{b}$$

$$\mathbf{y} = \mathbf{y} - \mathbf{1}\mathbf{1}^T\mathbf{D}_b\mathbf{y}$$

3. Compute a weight ROSA

   a) Calculate the block weights and the scores for each blocks:
   for $b \leftarrow 1$ **to** B **do**

$$\mathbf{v}_b = \mathbf{X}_b^\top\mathbf{D}_b\mathbf{y}$$

$$\mathbf{t}_b = \mathbf{X}_b\mathbf{v}_b\left(\mathbf{X}_b\mathbf{v}_b\right)^{-1}$$

   b) Define the $\mathbf{y}$ residual for each block:

$$\mathbf{r}_b = \mathbf{y} - \mathbf{t}_b\left(\mathbf{t}_b^\top\mathbf{y}\right)$$

   c) Define the winning block according the block that provide the lowest $\mathbf{r}_b$

   d) Assign as scores of the a-th LV the scores corresponding to $\mathbf{t}_b$ of the winning block

   e) Orthogonalization of $\mathbf{t}_a$ with respect to the preceding components:
   if $a > 1$ **then**

$$t_a = t_a - T_{a-1}\left(T_{a-1}^\top\mathbf{D}_aT_{a-1}\right)^{-1}T_{a-1}^\top\mathbf{D}_at_a$$

   f) Calculate the $\boldsymbol{y}$-loadings:

$$\mathbf{q}_a = \mathbf{y}^T\mathbf{D}_a\mathbf{t}_a$$

g) Update $\mathbf{y}$ with the smallest residual $\mathbf{r}_b$

h) Orthogonalize and normalize the winning weights $\mathbf{v}_a$

i) Repeat step from $3a$ to $3h$ until all the required latent variables (A) are extracted.

j) Extract the $\mathbf{X}$-loadings:
for $b \leftarrow 1$ **to** B **do**

$$P_b = \mathbf{X}_b^T \mathbf{D}_b \mathbf{T}$$

k) row-wise concatenate $\mathbf{V}$ and row-wise concatenate $\mathbf{P}$

l) Calculate regression coefficients:

$$\mathbf{B} = \mathbf{V} \left(\mathbf{P}^T \mathbf{V} \right)^{-1} \mathbf{q}^T$$

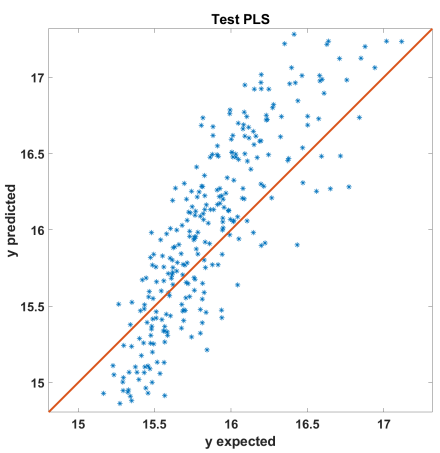
4. Repeat steps 1 to 3 for each new test sample.

Appendix C

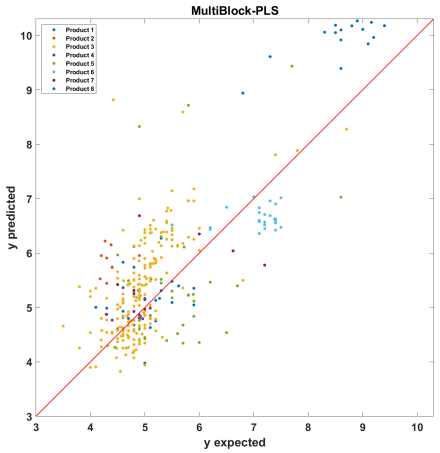
Locally-Weighted-RoBoost-PLS

This section provides additional material related to Chapter 7, including the results of the predictions obtained with the PLS models, the double cross-validation scheme adopted, and the Måge plot used to select the optimal number of latent components in the LW-RoBoost PLS models.

Figure C.1a shows the predictions on the simulated dataset obtained using a PLS model with 3 latent variables selected by cross-validation. The model gives an RMSEP of 0.40 and a prediction bias of 0.12. This result highlights the presence of non-linearities and outliers in the data set. In fact, deviations from the 1:1 straight line are observed, in particular a non-linear trend and a significant dispersion in the predictions, suggesting the need for more flexible or robust modelling strategies. On the other hand, figure C.1b shows the real industrial case with predictions on the test set obtained using a PLS model with 6 latent variables selected by cross-validation. The model gives an RMSEP of 0.32 and a prediction bias of 0.89. As can be seen in the figure, the prediction bias is highly dependent on the product type.



(a) Simulated data: predicted **y**-values versus measured **y**-values plots resulting from the application of the optimal PLS model.



(b) ABS data: predicted **y**-values versus measured **y**-values plots resulting from the application of the optimal PLS model. The colour coding reflects the manufactured ABS grade.

Figure C.1 – predicted **y**-values versus measured **y**-values plots

Figure C.2 shows the double cross-validation scheme used to optimise the parameters of the LW-RoBoost-PLS model, as described in section 7.4.4.

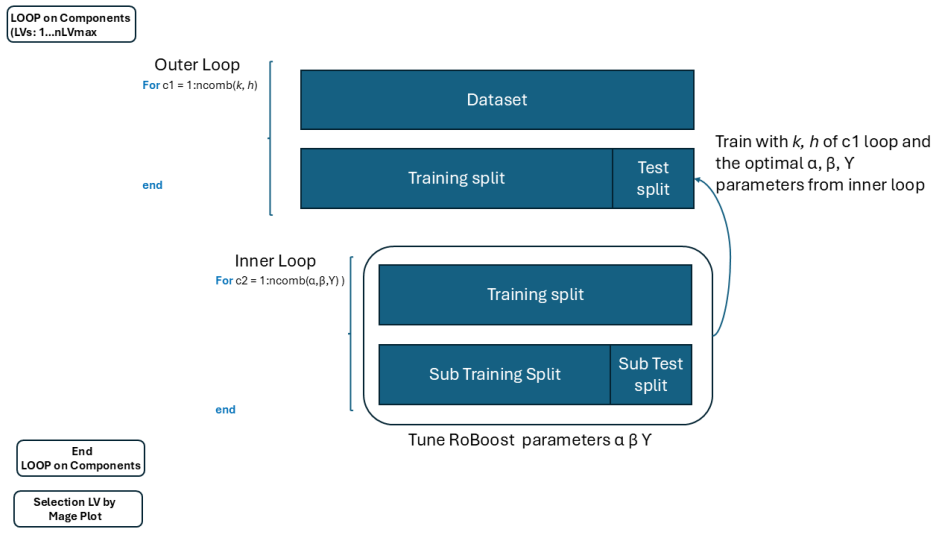
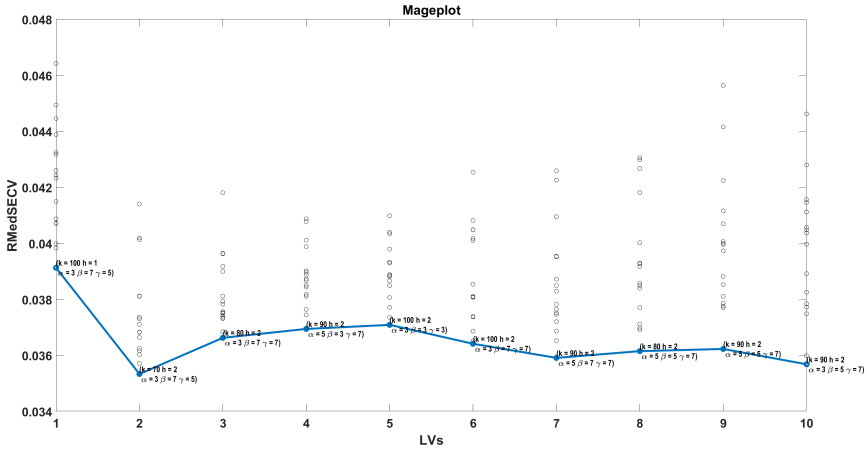
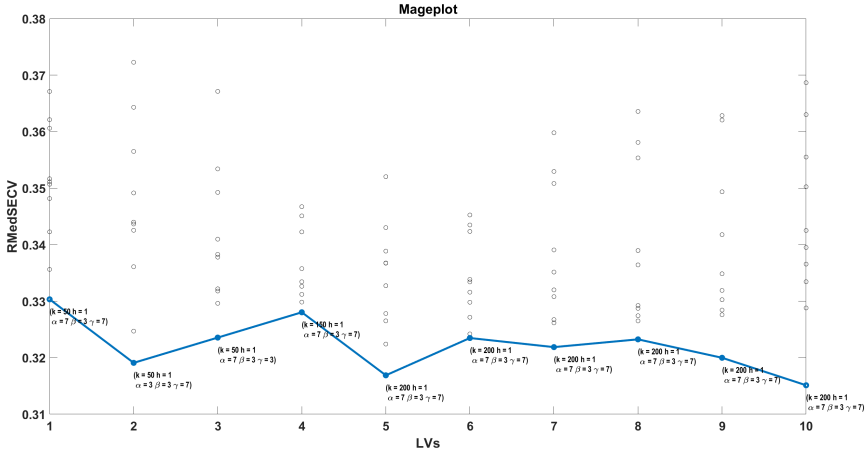


Figure C.2 – Double cross-validation scheme

The result of the double cross-validation used to select the optimal number of latent variables (LVs) is represented by the Måge plot (figure C.3b). This graph shows, for each number of LVs considered, the combination of parameters that gave the best predictive performance.



(a) Måge plot obtained from the simulated dataset.



(b) Måge plot obtained from the ABS dataset.

Figure C.3 – Måge plot, each black point represents the RMedSECV value associated with a particular combination of the optimizable hyperparameters (k, h, α, β, γ) for a given number of latent variables (LVs). The blue line connects the lowest RMedSECV values across the different LV values, indicating the optimal configuration for each model complexity. For each minimum point, the corresponding hyperparameter values are displayed

Scientific Contributions

Published Papers

Publications derived from and included in the thesis work:

- **Paper I:** Daniele Tanzilli, Alessandro D'Alessandro, Samuele Tamelli, Caterina Durante, Marina Cocchi, Lorenzo Strani, *"A feasibility study towards the on-line quality assessment of pesto sauce production by NIR and chemometrics"*, Foods, 12(8), 1679, 2023. DOI: 10.3390/foods12081679.
- **Paper II:** Lorenzo Strani, Raffaele Vitale, Daniele Tanzilli, Francesco Bonacini, Andrea Perolo, Erik Mantovani, Angelo Ferrando, Marina Cocchi, *"A multiblock approach to fuse process and near-infrared sensors for on-line prediction of polymer properties"*, Sensors, 22(4), 1436, 2022. DOI: 10.3390/s22041436.
- **Paper III:** Daniele Tanzilli, Lorenzo Strani, Francesco Bonacini, Angelo Ferrando, Marina Cocchi, Caterina Durante, *"Implementing Multiblock Techniques in a full-scale plant scenario: On-line Prediction of Quality Parameters in a Continuous Process for Different Acrylonitrile Butadiene Styrene (ABS) Products"*, Analytica Chimica Acta, 342851, 2024. DOI: 10.1016/j.aca.2024.342851.
- **Paper IV:** Daniele Tanzilli, Lorenzo Strani, Maxime Metz, Maxime Ryckewaert, Jean-Michel Roge, Matthieu Lesnoff, Raffaele Vitale, Cyril Ruckebusch, Marina Cocchi, *"Locally-Weighted-RoBoost-PLS: a new multivariate calibration approach to simultaneously cope with*

non-linearities and outliers", Analytica Chimica Acta, Submitted.

Publications not included in the thesis:

- **Paper V:** Lorenzo Strani, Francesco Bonacini, Angelo Ferrando, Andrea Perolo, Daniele Tanzilli, Raffaele Vitale, Marina Cocchi, *"Real Time Quality Assessment of General Purpose Polystyrene (GPPS) by means of Multiblock-PLS Applied on On-line Sensors Data"*, Chemical Engineering Transactions, 100, 175–180, 2023. DOI: 10.3303/CET23100030.
- **Paper VI:** Daniele Tanzilli, Marina Cocchi, José Manuel Amigo, Alessandro D'Alessandro, Lorenzo Strani, *"Does hyperspectral always matter? A critical assessment of near infrared versus hyperspectral near infrared in the study of heterogeneous samples"*, Current Research in Food Science, 9, 100813, 2024. DOI: 10.1016/j.crfs.2024.100813.
- **Paper VII:** Lorenzo Strani, Marina Cocchi, Daniele Tanzilli, Alessandra Biancolillo, Federico Marini, Raffaele Vitale, *"One class classification (class modelling): state of the art and perspectives"*, TrAC Trends in Analytical Chemistry, 118117, 2024. DOI: 10.1016/j.trac.2024.118117.
- **Paper VIII:** Lorenzo Strani, Barbara Benedetti, Marina Cocchi, Caterina Durante, Guido Perra, Mattia Pietropaolo, Samuele Pellacani, Daniele Tanzilli, *"Optimization of an analytical method based on the use of zwitterionic-phosphorylcholine-HILIC column for the determination of multiple polar emerging contaminants in reclaimed water"*, Journal of Chromatography A, 465605, 2024.
  DOI: 10.1016/j.chroma.2024.465605.

Conference Contributions

Oral Presentations

1. Daniele Tanzilli, Lorenzo Strani, Cyril Ruckebusch, Marina Cocchi, Raffaele Vitale, *"On some possible improvements to the ROSA algorithm"*, XIX Chemometrics in Analytical Chemistry (CAC), Santa Fe, Argentina, 9–12 September 2024.
2. Daniele Tanzilli, Lorenzo Strani, Maxime Metz, Maxime Ryckewaert, Jean-Michel Roge, Matthieu Lesnoff, Raffaele Vitale, Cyril Ruckebusch, Marina Cocchi, *"Locally-Weighted-RoBoost-PLS: a novel approach to simultaneously handle non-linearities and outliers in multivariate regression scenarios"*, The 8th International Chemometrics Research Meeting (ICRM), Soesterberg, the Netherlands, 24–27 June 2024.
3. Daniele Tanzilli, Lorenzo Strani, Maxime Metz, Maxime Ryckewaert, Jean-Michel Roge, Matthieu Lesnoff, Raffaele Vitale, Cyril Ruckebusch, Marina Cocchi, *"Some recent advances in non-linear modelling of industrial process data"*, Workshop di Chemiometria, Ravenna, Italy, 27–29 May 2024.
4. Daniele Tanzilli, Alessandro D'Alessandro, Lorenzo Strani, Caterina Durante, Marina Cocchi, *"Improve the industrial process understanding through chemometrics"*, Workshop dei giovani ricercatori chimici abruzzesi, Online, 12–13 July 2022.
5. Daniele Tanzilli, Lorenzo Strani, Maxime Metz, Maxime Ryckewaert, Jean-Michel Roge, Matthieu Lesnoff, Raffaele Vitale, Cyril Ruckebusch, Marina Cocchi, *"Locally-Weighted-RoBoost-PLS: a new multivariate calibration approach to simultaneously cope with non-linearities and outliers"*, 18th Scandinavian Symposium on Chemometrics, Gothenburg, Sweden, 1–4 October 2023.
6. Daniele Tanzilli, Alessandro D'Alessandro, Lorenzo Strani, Caterina Durante, Marina Cocchi, *"Different methods to monitor the quality of pesto sauce in an industrial process"*, Workshop di Chemiometria,

L'Aquila, Italy, 30 May–1 June 2022.

Poster Presentations

1. Daniele Tanzilli, Lorenzo Strani, Maxime Metz, Jean-Michel Roge, Matthieu Lesnoff, Raffaele Vitale, Cyril Ruckebusch, Marina Cocchi, *"Locally-Weighted-Roboost-PLS: a regression model for simultaneous handling of non-linearities and outliers"*, XXIII Giornata della Chimica dell'Emilia Romagna, Modena, Italy, 19 December 2024.
2. Daniele Tanzilli, Marina Cocchi, Cyril Ruckebusch, Raffaele Vitale, *"Enhancing Product Quality Control in Industrial Processes through the Identification of Critical Process Steps"*, SCI2024 XXVIII Congresso Nazionale della Società Chimica Italiana, Milano, Italy, 26–30 August 2024.
3. Daniele Tanzilli, Alessandro D'Alessandro, José M. Amigo, Marina Cocchi, *"When do heterogeneous samples really need hyperspectral imaging techniques?"*, XI Colloquium Chemiometricum Mediterraneum, Padova, Italy, 27–30 June 2023.
4. Daniele Tanzilli, Lorenzo Strani, Alessandro D'Alessandro, Caterina Durante, Marina Cocchi, *"Monitoring of pesto sauce production through multiblock approaches"*, XXI Giornata della Chimica dell'Emilia Romagna, Bologna, Italy, 19 December 2022.
5. Daniele Tanzilli, Alessandro D'Alessandro, Lorenzo Strani, José M. Amigo, Jesper Løve Hinrich, Marina Cocchi, *"IMAGINE NIR to monitor Pesto sauce industrial production"*, XVIII Chemometrics in Analytical Chemistry (CAC), Rome, Italy, 29 August–2 September 2022.
6. Daniele Tanzilli, Lorenzo Strani, Maxime Metz, Maxime Ryckewaert, Jean-Michel Roge, Matthieu Lesnoff, Raffaele Vitale, Cyril Ruckebusch, Marina Cocchi, *"Developing a Local RoBoost PLS framework"*, XI Colloquium Chemiometricum Mediterraneum, Padova, Italy, 27–30 June 2023.

Bibliography

(1) Callis, J. B.; Illman, D. L.; Kowalski, B. R. Process analytical chemistry. *Analytical Chemistry* **1987**, *59*, 624A–637A.

(2) Miller, C. E. Chemometrics for on-line spectroscopy applications—theory and practice. *Journal of Chemometrics: A Journal of the Chemometrics Society* **2000**, *14*, 513–528.

(3) Grassi, S.; Strani, L.; Alamprese, C.; Pricca, N.; Casiraghi, E.; Cabassi, G. A FT-NIR Process Analytical Technology Approach for Milk Renneting Control. *Foods* **2022**, *11*.

(4) França, L.; Grassi, S.; Pimentel, M. F.; Amigo, J. M. A single model to monitor multistep craft beer manufacturing using near infrared spectroscopy and chemometrics. *Food and Bioproducts Processing* **2021**, *126*, 95–103.

(5) Zhou, Q.; Dai, Z.; Song, F.; Li, Z.; Song, C.; Ling, C. Monitoring black tea fermentation quality by intelligent sensors: Comparison of image, e-nose and data fusion. *Food Bioscience* **2023**, *52*, 102454.

(6) He, K.; Zhong, M.; Li, Z.; Liu, J. Near-infrared spectroscopy for the concurrent quality prediction and status monitoring of gasoline blending. *Control Engineering Practice* **2020**, *101*, 104478.

(7) Food; Administration, D., et al. Guidance for industry, PAT-A framework for innovative pharmaceutical development, manufacturing and quality assurance. *http://www. fda. gov/cder/guidance/published. html* **2004**.

(8) Cullen, P.; O'Donnell, C. P.; Fagan, C. C. Benefits and challenges of adopting PAT for the food industry. *Process analytical technology for the food industry* **2014**, 1–5.

(9) Murphy, T.; O'Mahony, N.; Panduru, K.; Riordan, D.; Walsh, J. In *2016 27th Irish Signals and Systems Conference (ISSC)*, 2016, pp 1–7.

(10) Hermann, M.; Pentek, T.; Otto, B. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 2016, pp 3928–3937.

(11) Cañas, H.; Mula, J.; Díaz-Madroñero, M.; Campuzano-Bolarín, F. Implementing industry 4.0 principles. *Computers & industrial engineering* **2021**, *158*, 107379.

(12) Arden, N. S.; Fisher, A. C.; Tyner, K.; Lawrence, X. Y.; Lee, S. L.; Kopcha, M. Industry 4.0 for pharmaceutical manufacturing: Preparing for the smart factories of the future. *International Journal of Pharmaceutics* **2021**, *602*, 120554.

(13) Hasnan, N. Z. N.; Yusoff, Y. M. Short review: Application areas of industry 4.0 technologies in food processing sector. **2018**, 1–6.

(14) Lucizano, C.; de Andrade, A. A.; Facó, J. F. B.; de Freitas, A. G. In *2023 15th IEEE International Conference on Industry Applications (INDUSCON)*, 2023, pp 1195–1198.

(15) Rijal, A. *Industry 4.0 and World Economic Divergence-A novel perspective on the impact of fourth industrial revolution on the world economy*; tech. rep.; Jacobs University Bremen, 2022.

(16) Raptis, T. P.; Passarella, A.; Conti, M. Data Management in Industry 4.0: State of the Art and Open Challenges. *IEEE Access* **2019**, *7*, 97052–97093.

(17) Ferrer-Riquelme, A. Statistical control of measures and processes. **2009**.

(18) MacGregor, J.; Bruwer, M.; Miletic, I.; Cardin, M.; Liu, Z. Latent variable models and big data in the process industries. *IFAC-PapersOnLine* **2015**, *48*, 520–524.

(19) Chen, A.; Zhou, H.; An, Y.; Sun, W. In *2016 IEEE 25th International Symposium on Industrial Electronics (ISIE)*, 2016, pp 1022–1027.

(20) Kourti, T. Process analytical technology beyond real-time analyzers: the role of multivariate analysis. *Critical reviews in analytical chemistry* **2006**, *36*, 257–278.

(21) MacGregor, J. F. In *Computer Aided Chemical Engineering*; Elsevier: 2004; Vol. 18, pp 87–98.

(22) MacGregor, J. F.; Yu, H.; Muñoz, S. G.; Flores-Cerrillo, J. Data-based latent variable methods for process analysis, monitoring and control. *Computers & chemical engineering* **2005**, *29*, 1217–1223.

(23) Pérez-Beltrán, C. H.; Jiménez-Carvelo, A. M.; Torrente-López, A.; Navas, N. A.; Cuadros-Rodríguez, L. QbD/PAT—State of the art of multivariate methodologies in food and food-related biotech industries. *Food Engineering Reviews* **2023**, *15*, 24–40.

(24) Grassi, S.; Alamprese, C. Advances in NIR spectroscopy applied to process analytical technology in food industries. *Current Opinion in Food Science* **2018**, *22*, Foodomics Technologies 2018 * Innovations in Food Science, 17–21.

(25) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal* **1994**, *40*, 826–838.

(26) Awhangbo, L.; Bendoula, R.; Roger, J.-M.; Béline, F. Multi-block SO-PLS approach based on infrared spectroscopy for anaerobic digestion process monitoring. *Chemometrics and Intelligent Laboratory Systems* **2020**, *196*, 103905.

(27) Kourti, T. Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of adaptive control and signal processing* **2005**, *19*, 213–246.

(28) Cole, K. P.; Johnson, M. D. Continuous flow technology vs. the batch-by-batch approach to produce pharmaceutical compounds. *Expert Review of Clinical Pharmacology* **2018**, *11*, 5–13.

(29) Cattaldo, M.; Ferrer, A.; Måge, I. Variable time delay estimation in continuous industrial processes. *Chemometrics and Intelligent Laboratory Systems* **2024**, *246*, 105082.

(30) Betz, G.; Junker-Bürgin, P.; Leuenberger, H. Batch and continuous processing in the production of pharmaceutical granules. *Pharmaceutical development and technology* **2003**, *8*, 289–297.

(31) González-Martínez, J.; Camacho, J.; Ferrer, A. MVBatch: A matlab toolbox for batch process modeling and monitoring. *Chemometrics and Intelligent Laboratory Systems* **2018**, *183*, 122–133.

(32) UNICHIM, *Impianti chimici. Simboli e sigle per schemi e disegni. Manuale N. 6*; UNICHIM: Milano, 1994.

(33) Scotter, C. N. Non-destructive spectroscopic techniques for the measurement of food quality. *Trends in food science & technology* **1997**, *8*, 285–292.

(34) Gredilla, A.; Fdez-Ortiz de Vallejuelo, S.; Elejoste, N.; de Diego, A.; Madariaga, J. M. Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: A review. *TrAC Trends in Analytical Chemistry* **2016**, *76*, 30–39.

(35) Rodriguez-Saona, L.; Aykas, D. P.; Borba, K. R.; Urtubia, A. Miniaturization of optical sensors and their potential for high-throughput screening of foods. *Current Opinion in Food Science* **2020**, *31*, 136–150.

(36) Sanchez, P. D. C.; Arogancia, H. B. T.; Boyles, K. M.; Pontillo, A. J. B.; Ali, M. M. Emerging nondestructive techniques for the quality and safety evaluation of pork and beef: Recent advances, challenges, and future perspectives. *Applied Food Research* **2022**, *2*, 100147.

(37) Silva, S.; Guedes, C.; Rodrigues, S.; Teixeira, A. Non-destructive imaging and spectroscopic techniques for assessment of carcass and meat quality in sheep and goats: A review. *Foods* **2020**, *9*, 1074.

(38) Strani, L.; Durante, C.; Cocchi, M.; Marini, F.; Måge, I.; Biancolillo, A. Data fusion strategies for the integration of diverse non-destructive spectral sensors (NDSS) in food analysis. *TrAC Trends in Analytical Chemistry* **2024**, 117957.

(39) Pu, H.; Sun, D.-W.; Ma, J.; Cheng, J.-H. Classification of fresh and frozen-thawed pork muscles using visible and near infrared hyperspectral imaging and textural analysis. *Meat Science* **2015**, *99*, 81–88.

(40) Bruker What is FT-NIR Spectroscopy? `https://www.bruker.com/en/products-and-solutions/infrared-and-raman/ft-nir-spectrometers/what-is-ft-nir-spectroscopy.html`.

(41) Blanco, M.; Villarroya, I. NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry* **2002**, *21*, 240–250.

(42) Skoog, D. A.; Holler, F. J.; Crouch, S. R., *Instrumental analysis*; Brooks/Cole, Cengage Learning Belmont: 2007; Vol. 47.

(43) Osborne, B. G. Near-infrared spectroscopy in food analysis. *Encyclopedia of analytical chemistry: applications, theory and instrumentation* **2006**.

(44) Sarraguça, M. C.; Lopes, J. A. Quality control of pharmaceuticals with NIR: From lab to process line. *Vibrational Spectroscopy* **2009**, *49*, 204–210.

(45) Sánchez, M.; Bertran, E.; Sarabia, L.; Ortiz, M.; Blanco, M.; Coello, J. Quality control decisions with near infrared data. *Chemometrics and Intelligent Laboratory Systems* **2000**, *53*, 69–80.

(46) Rosas, J. G.; Blanco, M.; Gonzalez, J. M.; Alcalà, M. Real-time determination of critical quality attributes using near-infrared spectroscopy: A contribution for Process Analytical Technology (PAT). *Talanta* **2012**, *97*, 163–170.

(47) Aykas, D. P.; Ball, C.; Menevseoglu, A.; Rodriguez-Saona, L. E. In situ monitoring of sugar content in breakfast cereals using a novel FT-NIR spectrometer. *Applied Sciences* **2020**, *10*, 8774.

(48) Chadalavada, K.; Anbazhagan, K.; Ndour, A.; Choudhary, S.; Palmer, W.; Flynn, J. R.; Mallayee, S.; Pothu, S.; Prasad, K. V. S. V.; Varijakshapanikar, P., et al. NIR instruments and prediction methods for rapid access to grain protein content in multiple cereals. *Sensors* **2022**, *22*, 3710.

(49) Dixit, Y.; Casado-Gavalda, M. P.; Cama-Moncunill, R.; Cama-Moncunill, X.; Markiewicz-Keszycka, M.; Cullen, P.; Sullivan, C. Developments and challenges in online NIR spectroscopy for meat processing. *Comprehensive Reviews in Food Science and Food Safety* **2017**, *16*, 1172–1187.

(50) O'Callaghan, D.; O'Donnell, C.; Payne, F. A comparison of on-line techniques for determination of curd setting time using cheesemilks under different rates of coagulation. *Journal of Food engineering* **1999**, *41*, 43–54.

(51) Lyndgaard, C. B.; Engelsen, S. B.; van den Berg, F. W. Real-time modeling of milk coagulation using in-line near infrared spectroscopy. *Journal of Food Engineering* **2012**, *108*, 345–352.

(52) Palmer, J.; O'Malley, C.; Wade, M.; Martin, E. B.; Page, T.; Montague, G. Opportunities for process control and quality assurance using online NIR analysis to a continuous wet granulation tableting line. *Journal of Pharmaceutical Innovation* **2020**, *15*, 26–40.

(53) Fan, C.; Liu, T.; Hu, G.; Yang, M.; Long, J. Online Determination on the Properties of Naphtha as the Ethylene Feedstock Using Near-Infrared Spectroscopy. *Petroleum Chemistry* **2023**, *63*, 1069–1079.

(54) Bassey, U.; Rojek, L.; Hartmann, M.; Creutzburg, R.; Volland, A. The potential of NIR spectroscopy in the separation of plastics for pyrolysis. *Electronic Imaging* **2021**, *33*, 1–14.

(55) Santos, A. F.; Silva, F. M.; Lenzi, M. K.; Pinto*, J. C. Monitoring and control of polymerization reactors using NIR spectroscopy. *Polymer-Plastics Technology and Engineering* **2005**, *44*, 1–61.

(56) Shewhart, W. A. Economic control of quality of manufactured product van Nostrand. *New York* **1931**.

(57) Page, E. Cumulative sum charts. *Technometrics* **1961**, *3*, 1–9.

(58) Woodward, R. H.; Goldsmith, P. L., *Cumulative Sum Techniques*, Edizione Inglese; Oliver and Boyd: London, 1964.

(59) Hunter, J. S. The exponentially weighted moving average. *Journal of quality technology* **1986**, *18*, 203–210.

(60) Kourti, T.; MacGregor, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and intelligent laboratory systems* **1995**, *28*, 3–21.

(61) MacGregor, J. F.; Kourti, T. Statistical process control of multivariate processes. *Control engineering practice* **1995**, *3*, 403–414.

(62) Hotelling, H. Multivariate quality control-illustrated by the air testing of sample bombsights. *Techniques of statistical analysis* **1947**.

(63) Alt, F. B. In *Encyclopedia of Statistical Sciences*, Kotz, S., Johnson, N. L., Read, C. B., Eds.; John Wiley & Sons: New York, 1985; Vol. 6, pp 110–122.

(64) Montgomery, D. C., *Introduction to Statistical Quality Control*, 3rd; John Wiley & Sons: New York, 1996.

(65) Zerzucha, P.; Walczak, B. Concept of (dis) similarity in data analysis. *TrAC Trends in Analytical Chemistry* **2012**, *38*, 116–128.

(66) Kourti, T.; MacGregor, J. F. Multivariate SPC methods for process and product monitoring. *Journal of quality technology* **1996**, *28*, 409–428.

(67) Mason, R. L.; Tracy, N. D.; Young, J. C. Decomposition of T^2 for Multivariate Control Chart Interpretation. *Journal of Quality Technology* **1995**, *27*, 99–108.

(68) Lowry, C. A.; Woodall, W. H.; Champ, C. W.; Rigdon, S. E. A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics* **1992**, *34*, 46–53.

(69) Wierda, S. J. Multivariate Statistical Process Control—Recent Results and Directions for Future Research. *Statistica Neerlandica* **1994**, *48*, 147–168.

(70) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, 37–52.

(71) Dumarey, M.; Hermanto, M.; Airiau, C.; Shapland, P.; Robinson, H.; Hamilton, P.; Berry, M. Advances in continuous active pharmaceutical ingredient (API) manufacturing: real-t ime monitoring using multivariate tools. *Journal of Pharmaceutical Innovation* **2019**, *14*, 359–372.

(72) Tahir, F.; Islam, M. T.; Mack, J.; Robertson, J.; Lovett, D. Process monitoring and fault detection on a hot-melt extrusion process using in-line Raman spectroscopy and a hybrid soft sensor. *Computers & Chemical Engineering* **2019**, *125*, 400–414.

(73) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems* **2001**, *58*, 109–130.

(74) André, S.; Saint Cristau, L.; Gaillard, S.; Devos, O.; Calvosa, É.; Duponchel, L. In-line and real-time prediction of recombinant antibody titer by in situ Raman spectroscopy. *Analytica Chimica Acta* **2015**, *892*, 148–152.

(75) Frank, I.; Feikema, J.; Constantine, N.; Kowalski, B. Prediction of product quality from spectral data using the partial least-squares method. *Journal of Chemical Information and Computer Sciences* **1984**, *24*, 20–24.

(76) Kourti, T. Multivariate statistical process control and process control, using latent variables. **2009**.

(77) Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven soft sensors in the process industry. *Computers & chemical engineering* **2009**, *33*, 795–814.

(78) Tukey, J. W., *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, 1977.

(79) Lopez, E.; Etxebarria-Elezgarai, J.; Amigo, J. M.; Seifert, A. The importance of choosing a proper validation strategy in predictive models. A tutorial with real examples. *Analytica Chimica Acta* **2023**, *1275*, 341532.

(80) Westad, F.; Marini, F. Validation of chemometric models – A tutorial. *Analytica Chimica Acta* **2015**, *893*, 14–24.

(81) Rinnan, Å.; Van Den Berg, F.; Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* **2009**, *28*, 1201–1222.

(82) Camacho, J.; Ferrer, A. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical aspects. *Chemometrics and Intelligent Laboratory Systems* **2014**, *131*, 37–50.

(83) Bro, R.; Smilde, A. K. Principal component analysis. *Analytical methods* **2014**, *6*, 2812–2831.

(84) Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Analytica chimica acta* **1986**, *185*, 1–17.

(85) Wise, B. M.; Gallagher, N. B. The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control* **1996**, *6*, 329–348.

(86) Jackson, J. E.; Mudholkar, G. S. Control procedures for residuals associated with principal component analysis. *Technometrics* **1979**, *21*, 341–349.

(87) Box, G. E. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The annals of mathematical statistics* **1954**, 290–302.

(88) Nomikos, P.; MacGregor, J. F. Multivariate SPC charts for monitoring batch processes. *Technometrics* **1995**, *37*, 41–59.

(89) Eriksson, L.; Byrne, T.; Johansson, E.; Trygg, J.; Vikström, C., *Multi-and megavariate data analysis basic principles and applications*; Umetrics Academy: 2013; Vol. 1.

(90) Westerhuis, J. A.; Gurden, S. P.; Smilde, A. K. Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and intelligent laboratory systems* **2000**, *51*, 95–114.

(91) Fuentes-García, M.; Maciá-Fernández, G.; Camacho, J. Evaluation of diagnosis methods in PCA-based Multivariate Statistical Process Control. *Chemometrics and Intelligent Laboratory Systems* **2018**, *172*, 194–210.

(92) Mnassri, B.; Adel, E. M. E.; Ananou, B.; Ouladsine, M. Fault Detection and Diagnosis Based on PCA and a New Contribution Plot. *IFAC Proceedings Volumes* **2009**, *42*, 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, 834–839.

(93) European Commission Food Quality. Knowledge for Policy, 2022.

(94) Zeaiter, M.; Roger, J.; Bellon-Maurel, V. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. *Chemometrics and Intelligent Laboratory Systems* **2006**, *80*, CHIMIOMÉTRIE 2004, 227–235.

(95) Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1936**, *1*, 211–218.

(96) Snee, R. D. Validation of regression models: methods and examples. *Technometrics* **1977**, *19*, 415–428.

(97) Casale, M.; Simonetti, R. Review: Near Infrared Spectroscopy for Analysing Olive Oils. *Journal of Near Infrared Spectroscopy* **2014**, *22*, 59–80.

(98) Galtier, O.; Dupuy, N.; Le Dréau, Y.; Ollivier, D.; Pinatel, C.; Kister, J.; Artaud, J. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. *Analytica Chimica Acta* **2007**, *595*, Papers presented at the 10th International Conference on Chemometrics in Analytical Chemistry, 136–144.

(99) Wold, S.; Johansson, E.; Cocchi, M., et al. In *3D QSAR in Drug Design: Theory, Methods and Applications.* Kluwer ESCOM Science Publisher: 1993, pp 523–550.

(100) Amigo, J. M. In *Data handling in science and technology*; Elsevier: 2019; Vol. 32, pp 3–16.

(101) Behnke, S., *Hierarchical neural networks for image interpretation*; Springer Science & Business Media: 2003; Vol. 2766.

(102) Duchesne, C.; Liu, J.; MacGregor, J. Multivariate image analysis in the process industries: A review. *Chemometrics and Intelligent Laboratory Systems* **2012**, *117*, Special Issue Section: Selected Papers from the 1st African-European Conference on Chemometrics, Rabat, Morocco, September 2010 Special Issue Section: Preprocessing methods Special Issue Section: Spectroscopic imaging, 116–128.

(103) Zhao, Z.; Wang, R.; Liu, M.; Bai, L.; Sun, Y. Application of machine vision in food computing: A review. *Food Chemistry* **2024**, 141238.

(104) Hu, G.; Zhang, E.; Zhou, J.; Zhao, J.; Gao, Z.; Sugirbay, A.; Jin, H.; Zhang, S.; Chen, J. Infield apple detection and grading based on multi-feature fusion. *Horticulturae* **2021**, *7*, 276.

(105) Lintvedt, T. A.; Andersen, P. V.; Afseth, N. K.; Heia, K.; Lindberg, S.-K.; Wold, J. P. Raman spectroscopy and NIR hyperspectral imaging for in-line estimation of fatty acid features in salmon fillets. *Talanta* **2023**, *254*, 124113.

(106) Pham, V. H.; Lee, B. R. An image segmentation approach for fruit defect detection using k-means clustering and graph-based algorithm. *Vietnam Journal of Computer Science* **2015**, *2*, 25–33.

(107) Zheng, C.; Sun, D.-W. Image segmentation. *Computer Vision Technology for Food Quality Evaluation* **2011**, 37.

(108) Otsu, N. et al. A threshold selection method from gray-level histograms. *Automatica* **1975**, *11*, 23–27.

(109) Jain, A. K.; Ratha, N. K.; Lakshmanan, S. Object detection using Gabor filters. *Pattern recognition* **1997**, *30*, 295–309.

(110) Liang, X.; Jia, X.; Huang, W.; He, X.; Li, L.; Fan, S.; Li, J.; Zhao, C.; Zhang, C. Real-time grading of defect apples using semantic segmentation combination with a pruned YOLO V4 network. *Foods* **2022**, *11*, 3150.

(111) Li Vigni, M.; Prats-Montalban, J. M.; Ferrer, A.; Cocchi, M. Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA). *Journal of Chemometrics* **2018**, *32*, e2970.

(112) Prats-Montalbán, J. M.; Cocchi, M.; Ferrer, A. N-way modeling for wavelet filter determination in multivariate image analysis. *Journal of Chemometrics* **2015**, *29*, 379–388.

(113) De Juan, A. In *Data Handling in Science and Technology*; Elsevier: 2019; Vol. 32, pp 115–150.

(114) Haralick, R. M.; Shanmugam, K.; Dinstein, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics* **1973**, 610–621.

(115) Aloysius, N.; Geetha, M. In *2017 international conference on communication and signal processing (ICCSP)*, 2017, pp 0588–0592.

(116) Tanzilli, D.; D'Alessandro, A.; Tamelli, S.; Durante, C.; Cocchi, M.; Strani, L. A Feasibility Study towards the On-Line Quality Assessment of Pesto Sauce Production by NIR and Chemometrics. *Foods* **2023**, *12*.

(117) Barker, M.; Rayens, W. Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society* **2003**, *17*, 166–173.

(118) Marini, F. Classification methods in chemometrics. *Current Analytical Chemistry* **2010**, *6*, 72–79.

(119) James, G.; Witten, D.; Hastie, T.; Tibshirani, R., et al., *An introduction to statistical learning*; Springer: 2013; Vol. 112.

(120) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics* **1936**, *7*, 179–188.

(121) McLachlan, G. J., *Discriminant Analysis and Statistical Pattern Recognition*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: 1992.

(122) Indahl, U. G.; Martens, H.; Næs, T. From dummy regression to prior probabilities in PLS-DA. *Journal of Chemometrics: A Journal of the Chemometrics Society* **2007**, *21*, 529–536.

(123) Cocchi, M.; Biancolillo, A.; Marini, F. In *Comprehensive analytical chemistry*; Elsevier: 2018; Vol. 82, pp 265–299.

(124) Lai, Y.-K.; Kuo, C.-C. J. A Haar wavelet approach to compressed image quality measurement. *Journal of Visual Communication and Image Representation* **2000**, *11*, 17–40.

(125) Nason, G. P.; Silverman, B. W. In *Wavelets and statistics*; Springer: 1995, pp 281–299.

(126) Géron, A., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*; " O'Reilly Media, Inc.": 2022.

(127) Alzubaidi, L.; Zhang, J.; Humaidi, A. J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M. A.; Al-Amidie, M.; Farhan, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data* **2021**, *8*, 1–74.

(128) Li, Y.; Zhang, H.; Xue, X.; Jiang, Y.; Shen, Q. Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2018**, *8*, e1264.

(129) Gupta, J.; Pathak, S.; Kumar, G. In *Journal of Physics: Conference Series*, 2022; Vol. 2273, p 012029.

(130) Ronneberger, O.; Fischer, P.; Brox, T. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015, pp 234–241.

(131) Bai, Y. In *SHS Web of Conferences*, 2022; Vol. 144, p 02006.

(132) He, K.; Zhang, X.; Ren, S.; Sun, J. In *Proceedings of the IEEE international conference on computer vision*, 2015, pp 1026–1034.

(133) Scherer, D.; Müller, A.; Behnke, S. In *International conference on artificial neural networks*, 2010, pp 92–101.

(134) Dumoulin, V.; Visin, F. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* **2016**.

(135) Bishop, C. M.; Nasrabadi, N. M., *Pattern recognition and machine learning*; 4; Springer: 2006; Vol. 4.

(136) Rojas, R.; Rojas, R. The backpropagation algorithm. *Neural networks: a systematic introduction* **1996**, 149–182.

(137) Diederik, P. K. Adam: A method for stochastic optimization. *(No Title)* **2014**.

(138) Passos, D.; Mishra, P. A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemometrics and Intelligent Laboratory Systems* **2022**, *223*, 104520.

(139) Ali, A.; Shamsuddin, S. M.; Ralescu, A. L. Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl* **2013**, *5*, 176–204.

(140) Khanam, R.; Hussain, M.; Hill, R.; Allen, P. A comprehensive review of convolutional neural networks for defect detection in industrial applications. *IEEE Access* **2024**.

(141) Liu, B.-Y.; Fan, K.-J.; Su, W.-H.; Peng, Y. Two-stage convolutional neural networks for diagnosing the severity of alternaria leaf blotch disease of the apple tree. *Remote Sensing* **2022**, *14*, 2519.

(142) Sladojevic, S.; Arsenovic, M.; Anderla, A.; Culibrk, D.; Stefanovic, D. Deep neural networks based recognition of plant diseases by leaf image classification. *Computational intelligence and neuroscience* **2016**, *2016*, 3289801.

(143) Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.

(144) Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* **2023**.

(145) He, K.; Gan, C.; Li, Z.; Rekik, I.; Yin, Z.; Ji, W.; Gao, Y.; Wang, Q.; Zhang, J.; Shen, D. Transformers in medical image analysis. *Intelligent Medicine* **2023**, *3*, 59–78.

(146) Bowler, A. L.; Bakalis, S.; Watson, N. J. A review of in-line and on-line measurement techniques to monitor industrial mixing processes. *Chemical Engineering Research and Design* **2020**, *153*, 463–495.

(147) Kourti, T. Multivariate statistical process control and process control, using latent variables. **2020**.

(148) Wold, S.; Kettaneh-Wold, N.; MacGregor, J. F.; Dunn, K. G. Batch process modeling and MSPC. **2009**.

(149) Morris, J.; Martin, E.; Stewart, D. Batch Process Monitoring through the integration of Spectral and Process Data. *IFAC Proceedings Volumes* **2005**, *38*, 13–18.

(150) Aguado, D.; Ferrer, A.; Seco, A.; Ferrer, J. Comparison of different predictive models for nutrient estimation in a sequencing batch reactor for wastewater treatment. *Chemometrics and intelligent laboratory systems* **2006**, *84*, 75–81.

(151) Gabrielsson, J.; Jonsson, H.; Trygg, J.; Airiau, C.; Schmidt, B.; Escott, R. Combining process and spectroscopic data to improve batch modeling. *AIChE journal* **2006**, *52*, 3164–3172.

(152) Lourenço, N. D.; Lopes, J.; Almeida, C.; Sarraguça, M.; Pinheiro, H. M. Bioreactor monitoring with spectroscopy and chemometrics: a review. *Analytical and bioanalytical chemistry* **2012**, *404*, 1211–1237.

(153) Avila, C.; Mantzaridis, C.; Ferré, J.; de Oliveira, R. R.; Kantojärvi, U.; Rissanen, A.; Krassa, P.; de Juan, A.; Muller, F. L.; Hunter, T. N., et al. 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.

(154) Sousa, B. V.; Silva, F.; Reis, M. A.; Lourenço, N. D. Monitoring pilot-scale polyhydroxyalkanoate production from fruit pulp waste using near-infrared spectroscopy. *Biochemical Engineering Journal* **2021**, *176*, 108210.

(155) Strani, L.; Mantovani, E.; Bonacini, F.; Marini, F.; Cocchi, M. Fusing NIR and Process Sensors Data for Polymer Production Monitoring. *Frontiers in Chemistry* **2021**, *9*, 748723.

(156) Ge, Z. Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems* **2017**, *171*, 16–25.

(157) Silva, B. S.; Colbert, M.-J.; Santangelo, M.; Bartlett, J. A.; Lapointe-Garant, P.-P.; Simard, J.-S.; Gosselin, R. Monitoring microsphere coating processes using PAT tools in a bench scale fluid bed. *European Journal of Pharmaceutical Sciences* **2019**, *135*, 12–21.

(158) Wubshet, S. G.; Wold, J. P.; Afseth, N. K.; Böcker, U.; Lindberg, D.; Ihunegbo, F. N.; Måge, I. Feed-forward prediction of product qualities in enzymatic protein hydrolysis of poultry by-products: A spectroscopic approach. *Food and bioprocess technology* **2018**, *11*, 2032–2043.

(159) Strelet, E.; Peng, Y.; Castillo, I.; Rendall, R.; Wang, Z.; Joswiak, M.; Braun, B.; Chiang, L.; Reis, M. S. Multi-source and multimodal data fusion for improved management of a wastewater treatment plant. *Journal of Environmental Chemical Engineering* **2023**, *11*, 111530.

(160) Biancolillo, A.; Bucci, R.; Magrı, A. L.; Magrı, A. D.; Marini, F. Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication. *Analytica chimica acta* **2014**, *820*, 23–31.

(161) Vitale, R.; de Noord, O. E.; Westerhuis, J. A.; Smilde, A. K.; Ferrer, A. Divide et impera: How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding. *Journal of Chemometrics* **2021**, *35*, e3266.

(162) Campos, M. P.; Reis, M. S. Data preprocessing for multiblock modelling–a systematization with new methods. *Chemometrics and Intelligent Laboratory Systems* **2020**, *199*, 103959.

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

(164) Gorla, G.; Ferrer, A.; Giussani, B. Process understanding and monitoring: A glimpse into data strategies for miniaturized NIR spectrometers. *Analytica Chimica Acta* **2023**, *1281*, 341902.

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

(166) Velez, N. L.; Drennen, J. K.; Anderson, C. A. Challenges, opportunities and recent advances in near infrared spectroscopy applications for monitoring blend uniformity in the continuous manufacturing of solid oral dosage forms. *International Journal of Pharmaceutics* **2022**, *615*, 121462.

(167) De Oliveira, R. R.; Pedroza, R. H.; Sousa, A. O.; Lima, K. M.; 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.

(168) Diez-Olivan, A.; Del Ser, J.; Galar, D.; Sierra, B. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Information Fusion* **2019**, *50*, 92–111.

(169) De Oliveira, R. R.; Avila, C.; Bourne, R.; Muller, F.; 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.

(170) Smilde, A. K.; Måge, I.; Naes, T.; Hankemeier, T.; Lips, M. A.; Kiers, H. A.; Acar, E.; Bro, R. Common and distinct components in data fusion. *Journal of Chemometrics* **2017**, *31*, e2900.

(171) Mishra, P.; Roger, J.-M.; Jouan-Rimbaud-Bouveresse, D.; Biancolillo, A.; Marini, F.; Nordon, A.; Rutledge, D. N. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends in Analytical Chemistry* **2021**, *137*, 116206.

(172) Biancolillo, A.; Næs, T. In *Data handling in science and technology*; Elsevier: 2019; Vol. 31, pp 157–177.

(173) Alinaghi, M.; Bertram, H. C.; Brunse, A.; Smilde, A. K.; Westerhuis, J. A. Common and distinct variation in data fusion of designed experimental data. *Metabolomics* **2020**, *16*, 1–11.

(174) Måge, I.; Smilde, A. K.; van der Kloet, F. M. Performance of methods that separate common and distinct variation in multiple data blocks. *Journal of Chemometrics* **2019**, *33*, e3085.

(175) Song, Y.; Westerhuis, J. A.; Smilde, A. K. Separating common (global and local) and distinct variation in multiple mixed types data sets. *Journal of Chemometrics* **2020**, *34*, e3197.

(176) Westerhuis, J. A.; Coenegracht, P. M. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics: A Journal of the Chemometrics Society* **1997**, *11*, 379–392.

(177) Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics: A Journal of the Chemometrics Society* **1998**, *12*, 301–321.

(178) El Ghaziri, A.; Cariou, V.; Rutledge, D. N.; Qannari, E. M. Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K+ 1) datasets. *Journal of Chemometrics* **2016**, *30*, 420–429.

(179) Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics* **1996**, *10*, 463–482.

(180) Næs, T.; Tomic, O.; Mevik, B.-H.; Martens, H. Path modelling by sequential PLS regression. *Journal of Chemometrics* **2011**, *25*, 28–40.

(181) Måge, I.; Menichelli, E.; Næs, T. Preference mapping by PO-PLS: separating common and unique information in several data blocks. *Food quality and preference* **2012**, *24*, 8–16.

(182) Liland, K. H.; Næs, T.; Indahl, U. G. ROSA—a fast extension of partial least squares regression for multiblock data analysis. *Journal of Chemometrics* **2016**, *30*, 651–662.

(183) Lesnoff, M.; Metz, M.; Roger, J.-M. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. *Journal of Chemometrics* **2020**, *34*, e3209 CEM-19-0063.R1, e3209.

(184) Ge, Z.; Song, Z. A comparative study of just-in-time-learning based methods for online soft sensor modeling. *Chemometrics and Intelligent Laboratory Systems* **2010**, *104*, 306–317.

(185) Song, Y.; Ren, M. A Novel Just-in-Time Learning Strategy for Soft Sensing with Improved Similarity Measure Based on Mutual Information and PLS. *Sensors* **2020**, *20*.

(186) Kim, S.; Kano, M.; Nakagawa, H.; Hasebe, S. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International Journal of Pharmaceutics* **2011**, *421*, 269–274.

(187) Pérez-Marín, D.; Garrido-Varo, A.; Guerrero, J. Non-linear regression methods in NIRS quantitative analysis. *Talanta* **2007**, *72*, 28–42.

(188) Hazama, K.; Kano, M. Covariance-based locally weighted partial least squares for high-performance adaptive modeling. *Chemometrics and Intelligent Laboratory Systems* **2015**, *146*, 55–62.

(189) Menichelli, E.; Almøy, T.; Tomic, O.; Olsen, N. V.; Næs, T. SO-PLS as an exploratory tool for path modelling. *Food Quality and Preference* **2014**, *36*, 122–134.

(190) Indahl, U. G.; Naes, T. Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling. *Journal of Chemometrics* **1998**, *12*, 261–278.

(191) Metz, M.; Biancolillo, A.; Lesnoff, M.; Roger, J.-M. A note on spectral data simulation. *Chemometrics and Intelligent Laboratory Systems* **2020**, *200*, 103979.

(192) Bevilacqua, M.; Marini, F. Local classification: locally weighted–partial least squares-discriminant analysis (LW–PLS-DA). *Analytica Chimica Acta* **2014**, *838*, 20–30.

(193) Parhi, R.; Nowak, R. D. The role of neural network activation functions. *IEEE Signal Processing Letters* **2020**, *27*, 1779–1783.

(194) Allegrini, F.; Pierna, J. F.; Fragoso, W.; Olivieri, A. C.; Baeten, V.; Dardenne, P. Regression models based on new local strategies for near infrared spectroscopic data. *Analytica Chimica Acta* **2016**, *933*, 50–58.

(195) Archontoulis, S. V.; Miguez, F. E. Nonlinear regression models and applications in agricultural research. *Agronomy Journal* **2015**, *107*, 786–798.

(196) Aastveit, A. H.; Marum, P. Near-infrared reflectance spectroscopy: different strategies for local calibrations in analyses of forage quality. *Applied spectroscopy* **1993**, *47*, 463–469.

(197) Fujiwara, K.; Kano, M.; Hasebe, S.; Takinami, A. Soft-sensor development using correlation-based just-in-time modeling. *AIChE Journal* **2009**, *55*, 1754–1765.

(198) Kadlec, P.; Grbić, R.; Gabrys, B. Review of adaptation mechanisms for data-driven soft sensors. *Computers & chemical engineering* **2011**, *35*, 1–24.

(199) Martens, H.; Næs, T., *Multivariate calibration*; John Wiley & Sons: 1992.

(200) Rosipal, R.; Clancy, D. Kernel Partial Least Squares for Nonlinear Regression and Discrimination, NTRS Author Affiliations: NASA Ames Research Center, NTRS Document ID: 20030014609 NTRS Research Center: Ames Research Center (ARC), 2002.

(201) Bennett, K.; Embrechts, M. An optimization perspective on kernel partial least squares regression. *Nato Science Series sub series III computer and systems sciences* **2003**, *190*, 227–250.

(202) Zhang, X.; Yan, W.; Shao, H. Nonlinear multivariate quality estimation and prediction based on kernel partial least squares. *Industrial & engineering chemistry research* **2008**, *47*, 1120–1131.

(203) Centner, V.; Massart, D. L. Optimization in locally weighted regression. *Analytical Chemistry* **1998**, *70*, 4206–4211.

(204) Pérez-Marín, D.; Garrido-Varo, A.; Guerrero, J. Non-linear regression methods in NIRS quantitative analysis. *Talanta* **2007**, *72*, 28–42.

(205) Naes, T.; Isaksson, T.; Kowalski, B. Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry* **1990**, *62*, 664–673.

(206) Shenk, J. S.; Westerhaus, M. O.; Berzaghi, P. Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy* **1997**, *5*, 223–232.

(207) Kim, S.; Kano, M.; Nakagawa, H.; Hasebe, S. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International journal of pharmaceutics* **2011**, *421*, 269–274.

(208) Duma, Z.-S.; Susiluoto, J.; Lamminpää, O.; Sihvonen, T.; Reinikainen, S.-P.; Haario, H. Kf-pls: Optimizing kernel partial least-squares (k-pls) with kernel flows. *Chemometrics and Intelligent Laboratory Systems* **2024**, *254*, 105238.

(209) Pell, R. J. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems* **2000**, *52*, 87–104.

(210) Filzmoser, P.; Todorov, V. Review of robust multivariate statistical methods in high dimension. *Analytica chimica acta* **2011**, *705*, 2–14.

(211) Filzmoser, P.; Nordhausen, K. Robust linear regression for high-dimensional data: An overview. *Wiley Interdisciplinary Reviews: Computational Statistics* **2021**, *13*, e1524.

(212) Cummins, D. J.; Andrews, C. W. Iteratively reweighted partial least squares: a performance analysis by Monte Carlo simulation. *Journal of Chemometrics* **1995**, *9*, 489–507.

(213) Metz, M.; Abdelghafour, F.; Roger, J.-M.; Lesnoff, M. A novel robust PLS regression method inspired from boosting principles: RoBoost-PLSR. *Analytica Chimica Acta* **2021**, *1179*, 338823.

(214) Metz, M.; Ryckewaert, M.; Mas-Garcia, S.; Bendoula, R.; Dardenne, P.; Lesnoff, M.; Roger, J.-M. RoBoost-PLS2-R: an extension of RoBoost-PLSR method for multi-response. *Chemometrics and Intelligent Laboratory Systems* **2022**, *222*, 104498.

(215) Davies, A. M.; Britcher, H. V.; Franklin, J. G.; Ring, S. M.; Grant, A.; McClure, W. F. The application of fourier-transformed near-infrared spectra to quantitative analysis by comparison of similarity indices (CARNAC). *Microchimica Acta* **1988**, *94*, 61–64.

(216) Schaal, S.; Atkeson, C. G.; Vijayakumar, S. Scalable techniques from nonparametric statistics for real time robot learning. *Applied Intelligence* **2002**, *17*, 49–60.

(217) Pan, H.-L.; Huang, C.-M.; Huang, C.-y. Mapping aboveground carbon density of subtropical subalpine dwarf bamboo (Yushania niitakayamensis) vegetation using UAV-lidar. *International Journal of Applied Earth Observation and Geoinformation* **2023**, *123*, 103487.

(218) Baumann, D.; Baumann, K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *Journal of cheminformatics* **2014**, *6*, 1–19.

(219) Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated double cross validation. *Journal of Chemometrics: A Journal of the Chemometrics Society* **2009**, *23*, 160–171.

(220) Song, Y.; Ren, M. A novel just-in-time learning strategy for soft sensing with improved similarity measure based on mutual information and pls. *Sensors* **2020**, *20*, 3804.

(221) Zhang, X.; Kano, M.; Song, Z. Optimal weighting distance-based similarity for locally weighted PLS modeling. *Industrial & Engineering Chemistry Research* **2020**, *59*, 11552–11558.

(222) Zhang, X.; Wei, C.; Song, Z. Fast locally weighted PLS modeling for large-scale industrial processes. *Industrial & Engineering Chemistry Research* **2020**, *59*, 20779–20786.

(223) Groenevelt, H. The just-in-time system. *Handbooks in operations research and management science* **1993**, *4*, 629–670.

Published Papers Included in the Thesis

PAPER I

A Feasibility Study towards the On-Line Quality Assessment of Pesto Sauce Production by NIR and Chemometrics

Daniele Tanzilli, Alessandro D'Alessandro, Samuele Tamelli, Caterina Durante, Marina Cocchi and Lorenzo Strani

Foods 2023, 12, 1679.
https://doi.org/10.3390/foods12081679

foods

MDPI

Article

A Feasibility Study towards the On-Line Quality Assessment of Pesto Sauce Production by NIR and Chemometrics

Daniele Tanzilli [1,2], Alessandro D'Alessandro [1], Samuele Tamelli [1], Caterina Durante [1], Marina Cocchi [1,*] and Lorenzo Strani [1]

[1] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy; daniele.tanzilli@unimore.it (D.T.); alessandro.dalessandro@barilla.com (A.D.); samueletamelli1997@gmail.com (S.T.); caterina.durante@unimore.it (C.D.); lostrani@unimore.it (L.S.)
[2] Université de Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, 59000 Lille, France
* Correspondence: marina.cocchi@unimore.it

Abstract: The food industry needs tools to improve the efficiency of their production processes by minimizing waste, detecting timely potential process issues, as well as reducing the efforts and workforce devoted to laboratory analysis while, at the same time, maintaining high-quality standards of products. This can be achieved by developing on-line monitoring systems and models. The present work presents a feasibility study toward establishing the on-line monitoring of a pesto sauce production process by means of NIR spectroscopy and chemometric tools. The spectra of an intermediate product were acquired on-line and continuously by a NIR probe installed directly on the process line. Principal Component Analysis (PCA) was used both to perform an exploratory data analysis and to build Multivariate Statistical Process Control (MSPC) charts. Moreover, Partial Least Squares (PLS) regression was employed to compute real time prediction models for two different pesto quality parameters, namely, consistency and total lipids content. PCA highlighted some differences related to the origin of basil plants, the main pesto ingredient, such as plant age and supplier. MSPC charts were able to detect production stops/restarts. Finally, it was possible to obtain a rough estimation of the quality of some properties in the early production stage through PLS.

Keywords: MSPC charts; on-line; process monitoring; NIR; Basil; pesto production; PCA; PLS

Citation: Tanzilli, D.; D'Alessandro, A.; Tamelli, S.; Durante, C.; Cocchi, M.; Strani, L. A Feasibility Study towards the On-Line Quality Assessment of Pesto Sauce Production by NIR and Chemometrics. *Foods* **2023**, *12*, 1679. https://doi.org/10.3390/foods12081679

Academic Editors: Jordi Riu and Barbara Giussani

Received: 15 March 2023
Revised: 10 April 2023
Accepted: 14 April 2023
Published: 18 April 2023

1. Introduction

One of the most important aspects in food industrial production, in addition to basic safety and compliance requirements, is the capability to guarantee a constant quality of the final product, including all aspects from composition to appearance and taste. To achieve this aim, a lot of effort is spent monitoring the process, usually through univariate control charts and focusing most of the effort on monitoring the quality of the final product. However, operating in this way is not optimal when the food processing is complex, and production is massive. In fact, it may be difficult in this way to understand which are the Normal Operative Conditions (NOC) of the process. Since many parameters can change simultaneously and can be correlated, it is not easy for plant operators to detect the problem in a fast way in case anomalies or deviations occur [1]. In addition, although reference analyses are reliable and efficient in assessing the final product's quality, they provide slow responses as the sample must be collected, brought into the laboratory, and analyzed, being, at the same time, expensive in terms of money, operators' effort, and waste. For this reason, different types of sensors that can provide timely information are becoming more and more used for on-line quality probing from raw materials to the semi-finished and final products. It has been widely demonstrated that NIR spectroscopy has a powerful potential in monitoring food production processes [2–11], due to its ability to detect both chemical and physical changes in the samples. To cite a few applications: NIR has been used for

process monitoring in the dairy industry, from the prediction of raw milk composition to milk coagulation in cheese production and yogurt fermentation [11]; the fermentation processes in the wine and brewery industries; and the powdered ingredients mixing stage in different food matrices [10]. Thus, the on-line implementation of a NIR monitoring system is desired for several reasons: the timely handling of any possible faults, reducing products out of specification, thus reducing waste and economical loss. Moreover, if in addition to the data coming from process sensors controlling the machinery settings (such as the temperature, mixing rate, pressure, etc.) fused with NIR, it could become feasible to achieving a better understanding of processes, which could aid in designing more efficient and environmentally friendly processes [12,13]. However, it is still not so common in food production to have implemented systems for the data storage of retrieval process sensors. Nonetheless, companies are becoming increasingly interested in developing models that can achieve real-time monitoring and improve industrial processes.

However, it is difficult to handle, fuse, and interpret sensor data, as it is not possible to rapidly extract useful information from spectra and images without proper statistical tools. Thus, developing multivariate control charts based on latent variables and real-time prediction models, benefitting from the chemometric development in this area, is starting to be a recognized advantage in the industry.

It has been extensively demonstrated how Multivariate Statistical Process Monitoring/Control based on Latent Variables (MSPC-LVs) can lead to an efficient process monitoring [14–22].

The present work concerns a feasibility study to set up a model for the on-line monitoring of the pesto production process in the company Barilla, where, at the moment, a vision system (RGB camera) is monitoring the main raw material, i.e., basil, and a NIR probe installed in-line is monitoring the initial semi-finished product. The main aim of this preliminary feasibility study is the evaluation of the possible advantages that MSPC-LVs based on in-line acquired data can furnish both in terms of the possibility of estimating the quality of the finite product in real time and capturing the process evolution and the eventual departure from NOC. In this context, PCA models have been used to explore the data structure and the information they furnish. Furthermore, multivariate control charts for process monitoring based on NOC data were built. Lastly, a first attempt to obtain predictive models for the real-time prediction of main pesto quality parameters has been also carried out.

The focus has been on discussing the steps that were more critical for the models' development. Although the results are very preliminary, some interesting indications and directions for improvement could be formulated.

2. Materials and Methods

2.1. Process Description

The analyzed data were collected from the pesto sauce line during the 2020 harvesting season in a production plant owned by the company Barilla G. e R. Fratelli S.p.A., located near Parma, Italy. In this campaign, two different varieties of basil (*Ocimum basilicum*), the main ingredient of the sauce, have been provided by five local suppliers and continuously delivered to the process line. Each basil variety was harvested four times at different plant ages: the first cut was performed at 40 days, whereas the successive cuts were each carried out every 20 days.

At the beginning of the process line, a vision system (RGB camera) was installed that acquired images of basil plants while passing on the conveyor belt. The system was set to deliver some parameters in real time, such as the average and standard deviation values (every 15 s) of the R, G, and B channels and a rough estimation of the basil leaves' area in the acquired image (not always available at the same time intervals); however, the raw images were not always stored. Thus, in this work, only the R, G, and B parameters could be considered.

After this step, basil was mixed with salt and oil, forming an intermediate product, which was monitored on-line by a NIR probe. Then, all the other ingredients of the sauce were added to the intermediate product to complete the production and obtain the final product, whose quality was assessed by off-line laboratory analyses. A schematic representation of the process is reported in Figure 1. A critical issue when modeling on-line data for a continuous process is to establish the process timeline to match the sensors data acquired at different time steps with the same material; in other words, the considered variables should refer to the same sample to assemble a row of the data matrix. In this case, this step revealed particularly challenging, since the mixing of the intermediate product with the other ingredients (taking place after the NIR probe) was achieved in three distinct mixers that were emptied, transferring the crude pesto to the following processing steps sequentially, ensuring a continuous material flux. Thus, the residence time was established with the experts at the plant in order to correctly match the NIR spectra, corresponding to the intermediate material with the finished pesto at the end of the line, on which the quality parameters were acquired.

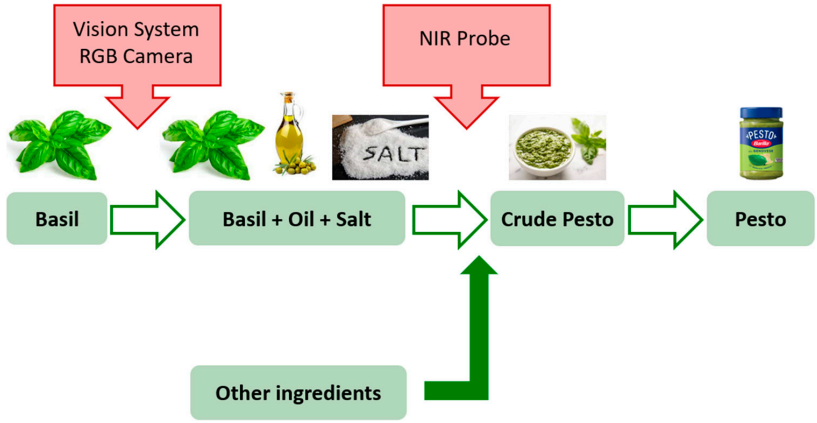


Figure 1. Schematic representation of pesto sauce production process.

In this study, data collected from May to August 2020 were analyzed, but not all the data recorded during this period were considered for model building, due to production pauses, instrument maintenance, and unreliable acquisitions. Finally, 459 data points were considered.

This is a second critical issue when assembling the data matrix since interruptions could be quite frequent. Since the on-line RGB continue to acquire the image of the same basil when a stop occurs at the raw material conveyor belt, an inspection of the RGB parameters' time trends with the identification of constant values as the indication of the stopping period was used. Moreover, the activation of the pump transferring the intermediate to the NIR probe was registered and was also used as an indication of stopping periods.

Finally, a data cleaning based on anomalous RGB values of the spectra was also accomplished.

2.2. Reference Analysis

Consistency parameters and lipids content have been considered for the assessment of pesto sauce quality. These parameters were assessed off-line by collecting pesto samples right after their production was complete.

The Consistency of pesto is evaluated measuring the flow of a standard volume of sample (100 cm^3) under its own weight. The flow could be related to the sample viscosity. To perform the measure, a Bostwick consistometer was used (ASTM F1080-93). This is a

stainless-steel slide with a reservoir of 5 × 5 × 4 cm, a mobile gate, two adjusting screws for planarity, and a track with ruler markings. The sample, conditioned to the temperature of 20 °C, was loaded into the reservoir. Then, the gate was opened, the timer was started, and sample flowed on the track. The consistency of the pesto was evaluated, measuring the distance in centimeters flowed in 30 s. Before the measure, the dedicated adjusting screws leveled the consistometer.

The total lipids content was determined by solvent extraction on a weighed sample aliquot (5 to 10 g). The extraction was conducted with an ethyl ether in a Soxhlet apparatus for 4 h. The sample was placed in a rotary evaporator and placed in an oven at 105 °C for 2 h to remove the solvent. The fat extracted was weighed at room temperature, and its content was expressed in a percentage, divided by the initial weight of the sample.

2.3. On-Line Instrumentation

A Sensure prototype camera (Sensure, Bergamo, Italy) was installed above the conveyor belt right after the basil plants were supplied, acquiring RGB images every 15 s. R, G, and B values were extracted by images and treated as separate variables.

A ProFoss spectrometer (Foss, Hillerød, Denmark) was used to collect the spectra of the basil, salt and oil mixture, namely, the intermediate product. The instrument was equipped with an optical fiber, whose probe was installed at the acquisition site on the process pipe. The spectra were acquired over the 1100–1650 nm spectral range in the transmission mode, with a nominal resolution of 0–5 nm and 64 scans per sample.

2.4. Data Analysis

The data analysis objectives were twofold: on the one hand, we evaluated the potentiality of establishing an on-line monitoring model (Section 2.4.3: Multivariate Control Charts; the results discussed in Section 3.2) capable of describing the natural variability inherent to the process and of capturing any eventual anomalous fluctuation, and, on the other hand, we aimed at establishing predictive models (Section 2.4.4: PLS Regression; the results discussed in Section 3.3) to evaluate the feasibility of the prediction of quality properties of the pesto sauce in real time.

However, prior to the model building, multivariate data exploration (Section 2.4.2: Principal Component Analysis; the results discussed in Section 3.1) has been a mandatory step to inspect the data structure and presence of deviating samples and to establish the time points corresponding to the normal operating conditions for the plant.

To ease readability, the applied preprocessing has been enclosed and detailed in Section 2.4.1: Preprocessing.

2.4.1. Preprocessing

The applied preprocessing is listed per the type of data and modelling phase:

- Vision System Data

The RGB data were preprocessed with autoscaling to uniformly model the variance among the different color channels.

- NIR spectra prior to PCA and MSPC

NIR spectra were pre-processed to remove effects, such as scattering, introducing variability not linked with information to be retrieved, and/or to enhance extractable information. In particular, Savitzky–Golay 2nd derivative and mean centering were applied prior to exploratory Principal Component Analysis and multivariate control charts building.

- NIR spectra prior to PLS regression

Savitzky–Golay 2nd derivative and mean centering were also used as preprocessing to compute the Partial Least Squares (PLS) regression model for the lipids content.

A different preprocessing strategy was needed to obtain the PLS model for consistency. This property was not directly linked to a chemical component, as the lipids show

specific absorption bands that can guide the modeling; thus, it was more difficult to model, especially considering how many registered on-line spectra were influenced by any process fluctuations. Thus, in order to remove spectral variability hindering the possibility of obtaining a satisfactory calibration model, a Dynamic Orthogonal Projection (DOP) [23] algorithm was applied, using the average spectra corresponding to the same consistency values (in the calibration set) as the source data (X_{source}) and the raw calibration spectra (X_{tar}) as the target. The main concept in DOP is that samples showing the same (or very close) y values should show the same spectral profile; thus, the "virtual" target spectra (X^*_{tar}), unaffected by the influence of uncontrolled conditions, could be estimated based on a distance or association matrix (M), calculated based on the y values of the source (y_s) and the target (y_t) domain. The singular value decomposition (SVD) of the difference matrix among measured and virtual target spectra was then used to determine the components (A) for orthogonalization:

$$X^*_{tar} = M^* \times X_{source} \quad (1)$$

$$D = X_{tar} - X^*_{tar} \quad (2)$$

$$[U_A\ S_A\ V_A] = \text{svd}(D, A) \quad (3)$$

$$X_{source\_corrected} = X_{source}\ (I - V_A V_A^T) \quad (4)$$

In our specific case, A = 4 was used after testing using from 1 to 5.

Once the average spectra were corrected, orthogonal projection could be directly used to predict the validation set, since the correction was embedded in the model. In this case, only mean centering (of both **X** and **y**) was applied prior to PLS.

2.4.2. Principal Component Analysis

Principal Component Analysis (PCA) is a method that by decomposition of the original data **X** into two matrices **T** and **P**, [24] according to Equation (5), allow reducing the dimensionality of the data set with a large set of variables, simplifying the exploration phase and the data visualization. PCA performs a projection of data from the original variables into new variables orthogonal to each other, the Principal Components (PCs), which are a linear combination of the original ones.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (5)$$

If the **X** matrix was composed of *n* rows (samples) and m columns (variables), the **T** matrix, called the scores matrix, which allowed us to understand the structure of the data, was composed by *n* rows and a number of columns equal to the number of PCs, and the loadings matrix **P** was composed by a number of rows equal to m and columns equal to the number of PCs. The loadings values corresponded to the weights by which each original variable entered the linear combination, thus defining the PCs, representing the contribution of each variable to each PC. The analysis of loadings matrix allowed us to understand the correlation structure of the variables [25]. The residual matrix **E**, which represented the unmodeled information, had the same dimension of **X,** and it was obtained by the subtraction of recalculated data from the PCA model ($\mathbf{TP}^T$) from **X**.

2.4.3. Multivariate Control Charts

PCA was also used to build multivariate control charts for MSPC. The dataset had been split in each calibration and test set manually, considering NOC observations, subdividing each period without production stops, as follows: the first part (about 65%) consisted of temporally contiguous points in the calibration set; and the second part (about 35%) was in the test set. In this way, we mimicked the real situation of continuous monitoring where samples to be predicted came after in time for each period. Observations not in NOC, as highlighted by exploratory PCA, were all included in the test set.

To estimate the correct number of PCs, cross-validation was performed with a *venetian blind* scheme with ten splits. The MSPC charts were based on two parameters: Hotelling T^2, which described the distance of a sample in the model space, and Q, which defined the distance of a sample from the model space. In other words, if a sample had high T^2 values, the model was able to describe it, but the distance between the sample and the center of the model was high, i.e., it showed an extreme behavior. On the other hand, if a sample was characterized by high Q values, the model was not able to describe the sample properly, hence the correlation structure of variables was different from the other samples. To assess if a sample was extreme or anomalous, signifying a departure from normal operative conditions for both control charts, the acceptance limits had to be estimated. The T^2 limit was obtained based on Hotelling's T^2 distribution, whereas the Q limit was based on χ^2 distribution and was calculated either with Jackson and Mudholkar approximation or the Box method [26,27].

2.4.4. PLS Regression

PLS is a linear regression method that allows predicting one or more response variables (Y block) from a predictor matrix (X block), establishing a multivariate linear relationship. It operates in a low-dimensional space defined by the Latent Variables (LVs), obtained from the simultaneous decomposition of **X** and **Y**, which are oriented on directions of maximum covariance between **X** and **Y** [28]. A PCA-like decomposition of **X** and **Y** is achieved (outer relation):

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{6}$$

$$\mathbf{Y} = \mathbf{U}\,\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{7}$$

where an inner relation links the outer relation:

$$\mathbf{U} = b^{*}\mathbf{T} \tag{8}$$

Hence, re-expressing this as a regression model:

$$\grave{\mathbf{Y}} = \mathbf{X}\,\mathbf{B} \tag{9}$$

where **T** and **U** are **X** and **Y** scores, **P** and **Q** are **X** and **Y** loadings, and **E** and **F** are the residual matrices, respectively. B holds the regression coefficients that allow the prediction of **Y** from **X** directly.

Data were partitioned into calibration (70%) and validation (30%) sets by the means of a Duplex algorithm [29]. The PLS model dimensionality, i.e., the number of PLS components, was assessed by the Root Mean Square Error in Cross-Validation (RMSECV), while the Root Mean Square Error in Prediction (RMSEP) was used to evaluate the models' predictive capability. Residual plots were also inspected.

3. Results and Discussion

3.1. Exploratory Data Analysis

Each type of data, RGB parameters, and NIR spectra were analyzed separately to visualize and explore the data structure. PCA analysis carried out on NIR spectra (acquired for 459 time points) had highlighted the presence of a cluster of samples at the negative value of PC1 and positive value of PC2, as shown in Figure 2a, as very far and different from all the other samples. Observing the PC1 versus time plot (Figure 2b), it was evident that these samples always corresponded to restarts, where production started after a period of inactivity. In Figure 2c, the loadings line plots for PC1 and PC2 are shown as the blue and red lines, respectively, where it is possible to see the absorption bands as mainly responsible for this difference. However, to jointly interpret scores and loadings plots, a PC1 vs. PC2 loadings scatter plot was also generated (Figure 2d). In the two figures, highlighted in purple, the wavelengths that describe the separation between the NOC and anomalous

samples are shown. It can be observed that the band in PC1 at 1400 nm, despite being the most intense, is not involved in the description of anomalous samples but just in extreme NOC samples with high values of PC1 scores in Figure 2a. On the other hand, the bands at 1170, 1213, 1236, and 1410 nm describe the behavior of the anomalous samples, as they fell in the separation direction, meaning that these samples had very different absorptions at these wavelengths. In detail, the bands at 1178 and 1410 nm can be ascribable to lignin, namely, the second overtone of C-H bond stretching of CH_3, and to the first overtone of the O-H bond stretching of the ROH group, respectively. Whereas, the band at 1213 and 1236 nm are related to the first and second overtone of C-H bond stretching of oleic and linoleic acid in olive oil CH_2 [30,31].

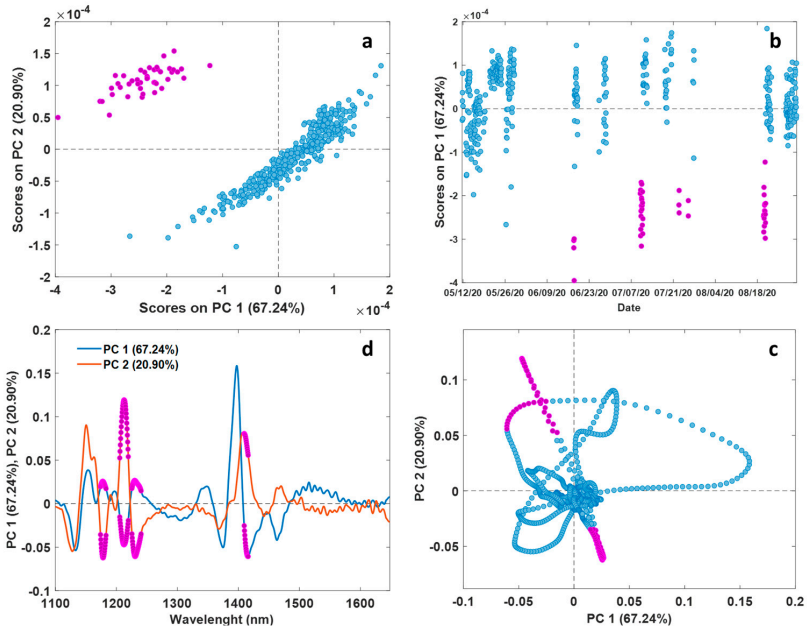


Figure 2. Results of the Exploratory Data Analysis performed on NIR data. PC1 vs. PC2 Scores plot (**a**), Scores on PC1 as a function of time (**b**), Loadings on PC1 and PC2 as a function of time (**c**), and Loadings on PC1 vs. PC2 (**d**). In (**a**,**b**), purple points represent anomalous samples; in (**c**,**d**), purple points represent wavelengths that mainly depict the difference of anomalous samples from the other ones.

Since these samples show the outliers' behavior, as they clearly do not represent the Normal Operative Conditions (NOCs), they were removed, and a new PCA model was built in order to obtain a better visualization of the possible differences among NOC samples.

The first PC (79.36% of variance explained) did not show any interesting trend, thus PC2 and PC3 were inspected. In Figure 3a,b, the scores plot of PC2 vs. PC3 is reported, where samples are colored according to the different additional information available, i.e., suppliers and different cuts, respectively. The suppliers' names have not been disclosed because of confidential agreement restrictions with the company. PC2 discriminated samples according to suppliers, as almost all samples of supplier number two had positive PC2 values, and the samples of suppliers three and four had negative PC values, suggesting that they were more similar to each other, with respect to number two. Only the samples coming from supplier five did not clearly differentiate from the others, whereas the number of samples from supplier one were too low to judge. Furthermore, PC2 and PC3 could distinguish between samples related to cut one and two (negative values of PC2 and positive values of PC3), with respect to samples related to cut three and four. The possibility to

discriminate against different cuts is relevant for the company, as younger basil plants generally give a higher quality product. However, observing the two plots simultaneously, it is evident that only certain suppliers, namely, number three and four, had delivered samples characterized by low cuts. In Figure S1a,b, the loadings plots of PC2 and PC3 are reported, respectively, which show the NIR bands responsible for these differences. Even if it is not possible to assess if suppliers or cuts influence them, the PCA resulted in a valuable tool to assess if incoming information about raw materials could be linked to the intermediate product characteristics; evidently, a more systematic planning of the next harvesting campaigns could clarify if a cut or supplier were the influential factors.

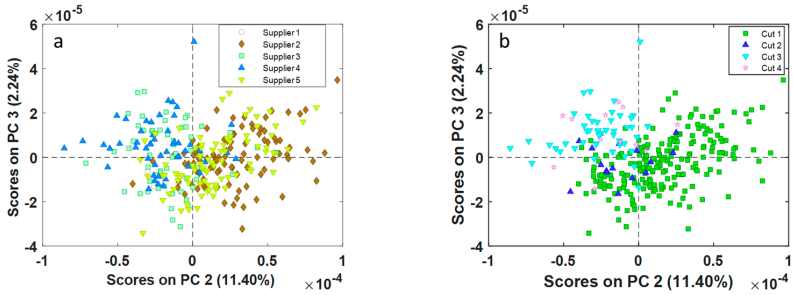


Figure 3. Results of the Exploratory Data Analysis performed on NIR data. PC2 vs. PC3 scores plots colored by different suppliers (**a**) and cuts (**b**).

PCA analysis carried out on data collected by an RGB camera was not able to detect the anomalous behavior of the samples highlighted in Figure 1. A possible explanation is that the process needed time to return to NOCs after a stop when the production restarted, and it could happen that the NIR spectra referred to material that was probably a residue of the old process (before the restart), and thus the acquired spectra did not depict the intermediate product newly produced at the beginning. Moreover, the observation of the samples' separation due to different cuts or suppliers was less efficient than the respective analysis performed on the NIR spectra. Thus, these differences were not linked to color variation but mostly to the basil's "chemical" profile.

3.2. MSPC Charts

The most interesting results related to the MSPC charts based on PCA were obtained by using the NIR data only (inclusion of RGB parameters did not provide additional insights). The PCA model, which explains 93% of the data variance with 4 Principal components, was calculated by inserting only the samples that were considered in NOCs according to plant experts in the calibration set (294 samples), whereas the test set (165 samples) comprised both NOCs and anomalous samples. The T^2 chart, reported in Figure 4a, describes the distance of each sample from the origin within the model space. Black circles represent the calibration samples used to build the PCA model, whereas red diamonds represent the test samples projected on the model. This chart detected five groups of samples with high T^2 values, which, again, corresponded to the NIR spectra acquired at the different restarts of the production. No other test sample exceeded the T^2 limit. Regarding the Q chart (Figure 4b), which describes the distance of each sample from the model space, the same samples corresponding to the restart are seen anomalous as for the T^2 chart, meaning that the model did not properly describe these samples. The charts' limits include few non-consecutive samples and inside of the nominal 5% of the total.

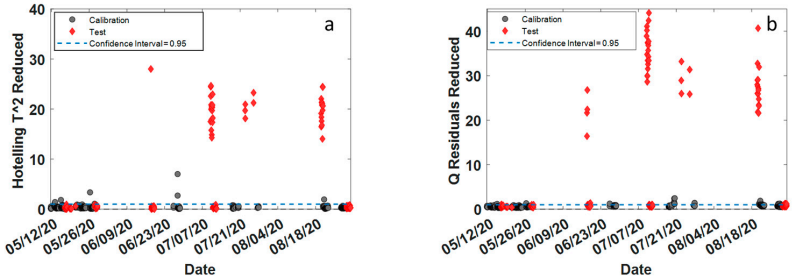


Figure 4. T^2-(**a**) and Q-(**b**) based MSPC charts.

Samples were also colored according to cut, supplier, consistency, and lipids values to observe if their behavior was related to these different features, but no particular trends were detected.

Nonetheless, the results obtained show how these charts are efficient in detecting possible departure from NOC, which translate to differences in intermediate products, accelerating the identification of possible plant issues or, as in this case, the adaptation of the process while returning to NOCs after a stop period. NIR is a very sensible technique to signal any variability occurring in intermediate production samples that can be due to process resetting (actual case), process drift, or variation in the NIR instrumentation setting/performance. The interpretation of the loadings and analysis of previous production campaigns data may help in discerning the different situations.

3.3. Predictive Models

An attempt to obtain predictive models, which can then be possibly used to estimate the consistency and lipids content of the final product in real time, was undertaken. Since RGB data were not able to provide reliable prediction models for both parameters, only results obtained by NIR data are presented, as summarized in Table 1.

Table 1. Results obtained by PLS Regression.

Method	LVs	RMSECV	RMSEP
Consistency (cm)	9	0.64	0.68
Lipids (%)	5	1.59	2

Before model computation, data were split by using a duplex algorithm with a 70/30% proportion in the calibration and test sets, giving 142(cal)/61(test) and 33(cal)/12(test) for consistency and lipids, respectively. Afterwards, four samples belonging to the anomalous group of observations, detected by using the T^2 and Q distances, were removed from the test set for consistency.

The prediction model for consistency was built using 9 LVs, corresponding to the minimum RMSECV (*venetian blind*, 10 splits) value. The RMSEP value was close to RMSECV (Table 1) and corresponded to an average relative percentage error of 10% in prediction, which was considered acceptable by the company for an early (intermediate product) on-line quality estimation. The samples in the test set showed a rather high variability compared to the ones in the calibration (Figure 5a,b). Nonetheless, the residuals vs. measured values of the consistency plot (Figure 5b) highlighted that the errors on both the calibration and test samples were randomly distributed, not showing any visible trend, excluding any bias.

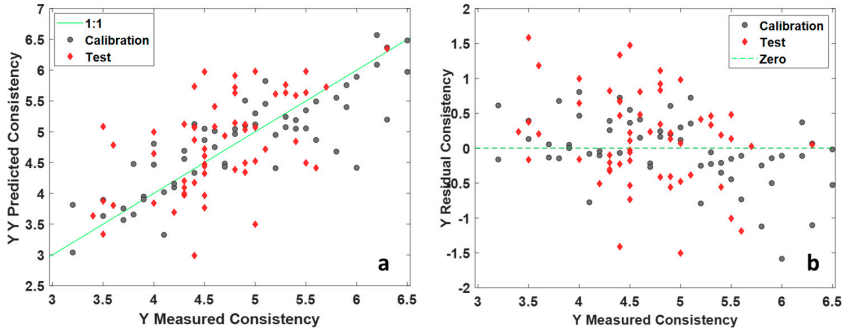


Figure 5. PLS results on NIR data for consistency. Predicted vs. measured values plot (**a**), residuals vs. measured values plot (**b**).

The prediction model of total lipids content was built using a lower number of samples than the previous model, as this parameter was assessed less frequently than consistency. In this case, 5 LVs were selected according to the minimum RMSECV (*venetian blind*, 10 splits) for the model's construction. As shown in Figure 6a, the majority of the samples had a lipid content included in the range 46–49%, and only a few samples presented higher values. This is a quite common situation in real time production, where a consistent quality of the product is pursued. In this case, a couple of samples in the test set were predicted with a higher error but, in general, the error values comprised the 2% range, which the company considered acceptable for controlling if the product was within specification for this parameter. One of the two samples with a high lipid content in the test set was predicted accurately, whereas the other one was underestimated (Figure 6b).

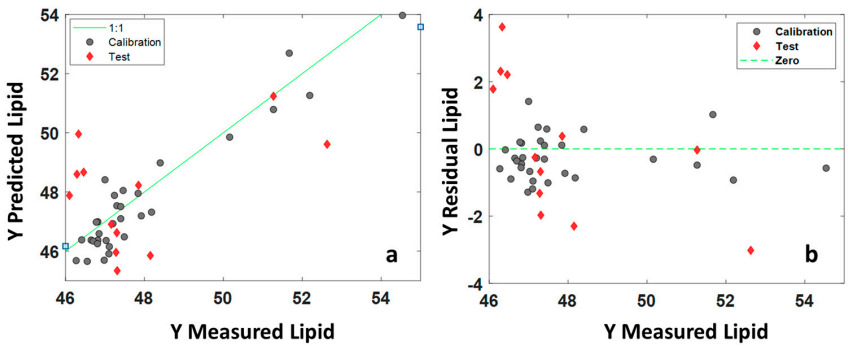


Figure 6. PLS results on NIR data for lipids content. Predicted vs. measured values plot (**a**), residuals vs. measured values plot (**b**).

In Figure 7, the Variable Influence in Projection (VIP) scores are shown [32], which highlight that the band at 1166 nm, ascribable to the olive oil's second overtone of the CH stretching of CH_3 [30,31], is the most influent for the prediction of total lipids content. Moreover, other bands linked to lipids in olive oil [30,31] can be found at 1422 and 1461 nm, typical of the CH stretching and deformation of CH_2, both above the significance threshold [32].

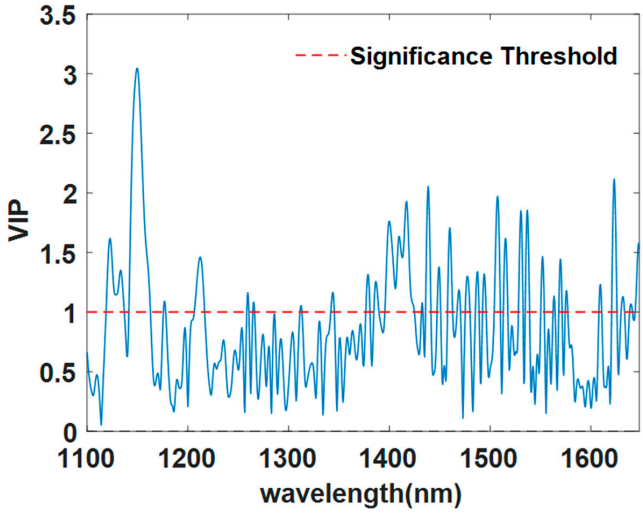


Figure 7. VIP scores of PLS on NIR data for lipids content.

4. Conclusions

This study presents a feasibility study towards the real-time monitoring of an industrial food process line (pesto production). Since historical data were not available, the obtained results referred to a single basil harvesting campaign. The modeling effort concerned both latent variables based multivariate control charts, aimed at monitoring the stability of process conditions and the eventual detecting of fluctuations exceeding the natural variability of the process as well as the quality properties' prediction in real time. Despite the fact that the collected data were limited, the results gave interesting insights, which are summarized in the following.

4.1. MSPC Results

(i) the RGB parameters obtained by the vision system, albeit potentially very useful, were not increasing information retrieved from NIR. We think this is due to the limited number of features extracted by the image, which could otherwise provide a good characterization of raw material; further work is in progress in this direction (e.g., detecting the percentage areas of damaged leaves, branches, and stems by an image analysis tool);

(ii) NIR-based multivariate control charts could detect restarts after temporary production stoppages, underlining that some changes occur in the intermediate product. On one hand, this is an indication of how sensible NIR spectroscopy is to monitor any changes, and, on the other hand, a monitoring system can clearly indicate when process fluctuations return to natural process variabilities and to the constancy of the product.

4.2. On-Line Predictive Models

(iii) The predictive model to estimate the pesto's consistency and total lipids content, based on the NIR spectra of the intermediate product, gave errors in the external predictions, which are considered acceptable by the company for on-line quality estimation.

(iv) It is worth noting that while building predictive models of final product quality parameters based on on-line sensors data is highly desirable, they suffer from the limited response variability (which, evidently, should be confined in the in-specific ranges). When, as in this case, it is not possible to expand the calibration range by pilot studies, the models can be, nonetheless, used as a timely rough indication of the property's value. In this respect, more than an estimation of the quality values, they may give a preliminary check about respecting specifications Within this framework, the obtained models seem promising.

Finally, it is worth mentioning the main issues encountered, such as the lack of systematic recording of acquired on-line data, the difficulties in recovering a sound synchronization scheme, and the critical role of spectral preprocessing to cope with the many sources of variabilities intrinsic in a process framework.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/foods12081679/s1, Figure S1: Results of the Exploratory Data Analysis performed on NIR data. Loadings Plot (a) PC1 vs. wavelengths; (b) PC2 vs. wavelengths.

Author Contributions: Conceptualization, D.T., A.D., M.C. and L.S.; methodology, D.T., C.D., M.C. and L.S.; software, D.T. and L.S.; validation, D.T., A.D., C.D. and L.S.; formal analysis, M.C. and L.S.; investigation, D.T., S.T. and A.D.; resources, A.D. and M.C.; data curation, D.T., S.T. and A.D.; writing—original draft preparation, D.T., C.D. and L.S.; writing—review and editing, D.T., A.D. M.C. and L.S.; visualization, D.T. and S.T.; supervision, C.D., M.C. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are unavailable due to privacy restrictions.

Acknowledgments: L. Strani aknowledges R. Vitale (University of Lille) for suggestions and support in the course of a Virtual Mobility Grant in the frame of the COST Action CA19145 "European Network for Assuring Food Integrity using Non-Destructive Spectral Sensors (SENSORFINT).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferrer-Riquelme, A. Statistical Control of Measures and Processes. In *Comprehensive Chemometrics*; Elsevier: Amsterdam, The Netherlands, 2009; pp. 97–126.
2. Grassi, S.; Strani, L.; Alamprese, C.; Pricca, N.; Casiraghi, E.; Cabassi, G. A FT-NIR process analytical technology approach for milk renneting control. *Foods* **2022**, *11*, 33. [CrossRef] [PubMed]
3. Franca, L.; Grassi, S.; Pimentel, M.F.; Amigo, J.M. A single model to monitor multistep craft beer manufacturing using near infrared spectroscopy and chemometrics. *Food Bioprod. Process.* **2021**, *126*, 95–103. [CrossRef]
4. Zhou, Q.; Dai, Z.; Song, F.; Li, Z.; Song, C.; Ling, C. Monitoring black tea fermentation quality by intelligent sensors: Comparison of image, e-nose and data fusion. *Food Biosci.* **2023**, *52*, 102454. [CrossRef]
5. Catelani, T.A.; Santos, J.R.; Páscoa, R.N.; Pezza, L.; Pezza, H.R.; Lopes, J.A. Real-time monitoring of a coffee roasting process with near infrared spectroscopy using multivariate statistical analysis: A feasibility study. *Talanta* **2018**, *179*, 292–299. [CrossRef]
6. Hao, Y.; Lu, Y.; Li, X. Study on robust model construction method of multi-batch fruit online sorting by near-infrared spectroscopy. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2022**, *280*, 121478. [CrossRef]
7. Strani, L.; Grassi, S.; Alamprese, C.; Casiraghi, E.; Ghiglietti, R.; Locci, F.; Pricca, N.; De Juan, A. Effect of physicochemical factors and use of milk powder on milk rennet-coagulation: Process understanding by near infrared spectroscopy and chemometrics. *Food Control* **2021**, *119*, 1074. [CrossRef]
8. Maléchaux, A.; Le Dréau, Y.; Artaud, J.; Dupuy, N. Control chart and data fusion for varietal origin discrimination: Application to olive oil. *Talanta* **2020**, *217*, 121115. [CrossRef]
9. Pérez-Beltrán, C.H.; Jiménez-Carvelo, A.M.; Torrente-López, A.; Navas, N.A.; Cuadros-Rodríguez, L. QbD/PAT—State of the Art of Multivariate Methodologies in Food and Food-Related Biotech Industries. *Food Eng. Rev.* **2023**, *15*, 24–40. [CrossRef]
10. Grassi, S.; Alamprese, C. Advances in NIR spectroscopy applied to process analytical technology in food industries. *Curr. Opin. Food Sci.* **2018**, *22*, 17–21. [CrossRef]
11. Pu, Y.Y.; O'Donnell, C.; Tobin, J.T.; O'Shea, N. Review of near-infrared spectroscopy as a process analytical technology for real-time product monitoring in dairy processing. *Int. Dairy J.* **2020**, *103*, 104623. [CrossRef]
12. Baines, T.; Brown, S.; Benedettini, O.; Ball, P.D. Examining green production and its role within the competitive strategy of manufacturers. *J. Ind. Eng. Manag.* **2012**, *5*, 53–87. [CrossRef]
13. Rico-Rodriguez, F.; Strani, L.; Grassi, S.; Lancheros, R.; Serrato, J.C.; Casiraghi, E. Study of Galactooligosaccharides production from dairy waste by FTIR and chemometrics as Process Analytical Technology. *Food Bioprod. Process.* **2021**, *126*, 113–120. [CrossRef]
14. Strani, L.; Mantovani, E.; Bonacini, F.; Marini, F.; Cocchi, M. Fusing NIR and Process Sensors Data for Polymer Production Monitoring. *Front. Chem.* **2021**, *9*, 785. [CrossRef] [PubMed]
15. Westerhuis, J.A.; Gurden, S.P.; Smilde, A.K. Generalized contribution plots in multivariate statistical process monitoring. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 95–114. [CrossRef]

16. Avila, C.; Mantzaridis, C.; Ferré, J.; de Oliveira, R.R.; Kantojärvi, U.; Rissanen, A.; Krassa, P.; De Juan, A.; Muller, F.L.; Hunter, T.; et al. 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. [CrossRef] [PubMed]
17. Macho, S.; Rius, A.; Callao, M.P.; Larrechi, M.S. Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy: Standardisation of the calibration model. *Anal. Chim. Acta* **2001**, *445*, 213–220. [CrossRef]
18. Joshi, K.; Patil, B. Multivariate statistical process monitoring and control of machining process using principal component-based Hotelling T2 charts: A machine vision approach. *Int. J. Product. Qual. Manag.* **2022**, *35*, 40–56. [CrossRef]
19. Biancolillo, A.; Scappaticci, C.; Foschi, M.; Rossini, C.; Marini, F. Coupling of NIR Spectroscopy and Chemometrics for the Quantification of Dexamethasone in Pharmaceutical Formulations. *Pharmaceuticals* **2023**, *16*, 309. [CrossRef]
20. de Oliveira, R.R.; Pedroza, R.H.; Sousa, A.O.; Lima, K.M.; 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. [CrossRef]
21. Kourti, T. Application of latent variable methods to process control and multivariate statistical process control in industry. *Int. J. Adapt. Control Signal Process.* **2005**, *19*, 213–246. [CrossRef]
22. Strani, L.; Vitale, R.; Tanzilli, D.; Bonacini, F.; Perolo, A.; Mantovani, E.; Ferrando, A.; Cocchi, M. A Multiblock Approach to Fuse Process and Near-Infrared Sensors for On-Line Prediction of Polymer Properties. *Sensors* **2022**, *22*, 1436. [CrossRef]
23. Zeaiter, M.; Roger, J.M.; Bellon-Maurel, V. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 227–235. [CrossRef]
24. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
25. Bro, R.; Smilde, A.K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [CrossRef]
26. Jackson, J.E.; Hearne, F.T. Hotelling's T_M^2 for Principal Components—What about Absolute Values? *Technometrics* **1979**, *21*, 253–255.
27. Nomikos, P.; MacGregor, J.F. Multivariate SPC charts for monitoring batch processes. *Technometrics* **1995**, *37*, 41–59. [CrossRef]
28. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
29. Snee, R.D. Validation of regression models: Methods and examples. *Technometrics* **1977**, *19*, 415–428. [CrossRef]
30. Galtier, O.; Dupuy, N.; Le Dréau, Y.; Ollivier, D.; Pinatel, C.; Kister, J.; Artaud, J. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. *Anal. Chim. Acta* **2007**, *595*, 136–144. [CrossRef]
31. Casale, M.; Simonetti, R. Near infrared spectroscopy for analysing olive oils. *J. Near Infrared Spectrosc.* **2014**, *22*, 59–80. [CrossRef]
32. Wold, S.; Johansson, E.; Cocchi, M. PLS: Partial least squares projections to latent structures. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kluwer ESCOM Science Publisher: Dordrecht/Leiden, The Netherlands, 1993; pp. 523–550.

PAPER II

A Multiblock Approach to Fuse Process and Near-Infrared Sensors for On-Line Prediction of Polymer Properties

Lorenzo Strani, Raffaele Vitale, Daniele Tanzilli, Francesco Bonacini, Andrea Perolo, Erik Mantovani, Angelo Ferrando and Marina Cocchi

Sensors 2022, 22, 1436.
https://doi.org/10.3390/s22041436

sensors

MDPI

Article

A Multiblock Approach to Fuse Process and Near-Infrared Sensors for On-Line Prediction of Polymer Properties

Lorenzo Strani [1], Raffaele Vitale [2], Daniele Tanzilli [1], Francesco Bonacini [3], Andrea Perolo [3], Erik Mantovani [3], Angelo Ferrando [3] and Marina Cocchi [1,*]

[1] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via 4 Campi 103, 41125 Modena, Italy; lostrani@unimore.it (L.S.); daniele.tanzilli@unimore.it (D.T.)
[2] Centre National de la Recherche Scientifique (CNRS), Laboratoire de Spectroscopie pour les Interactions, la Réactivitè et l'Environnement (LASIRE), Cité Scientifique, University Lille, F-59000 Lille, France; raffaele.vitale@univ-lille.fr
[3] Research Center, Versalis (ENI) S.p.A., Via Taliercio 14, 46100 Mantova, Italy; francesco.bonacini@versalis.eni.com (F.B.); andrea.perolo@versalis.eni.com (A.P.); erik.mantovani@versalis.eni.com (E.M.); angelo.ferrando@versalis.eni.com (A.F.)
* Correspondence: marina.cocchi@unimore.it; Tel.: +39-059-205-8554

Abstract: Petrochemical companies aim at assessing final product quality in real time, in order to rapidly deal with possible plant faults and to reduce chemical wastes and staff effort resulting from the many laboratory analyses performed every day. In order to answer these needs, the main purpose of the current work is to explore the feasibility of multiblock regression methods to build real-time monitoring models for the prediction of two quality properties of Acrylonitrile-Butadiene-Styrene (ABS) by fusing near-infrared (NIR) and process sensors data. Data come from a production plant, which operates continuously, and where four NIR probes are installed on-line, in addition to standard process sensors. Multiblock-PLS (MB-PLS) and Response-Oriented Sequential Alternation (ROSA) methods were here utilized to assess which of such sensors and plant areas were the most relevant for the quality parameters prediction. Several prediction models were constructed exploiting measurements provided by sensors active at different ABS production process stages. Both methods provided good prediction performances and permitted identification of the most relevant data blocks for the quality parameters' prediction. Moreover, models built without considering recordings from the final stage of the process yielded prediction errors comparable to those involving all available data blocks. Thus, in principle, allowing final ABS quality to be estimated in real-time before the end of the process itself.

Keywords: Acrylonitrile-Butadiene-Styrene; low-level data fusion; multiblock-partial least squares (MB-PLS); multivariate statistical process control; polymer production; quality prediction; real-time monitoring; response-oriented sequential alternation (ROSA)

Citation: Strani, L.; Vitale, R.; Tanzilli, D.; Bonacini, F.; Perolo, A.; Mantovani, E.; Ferrando, A.; Cocchi, M. A Multiblock Approach to Fuse Process and Near-Infrared Sensors for On-Line Prediction of Polymer Properties. *Sensors* **2022**, *22*, 1436. https://doi.org/10.3390/s22041436

Academic Editor: Natividad Duro Carralero

Received: 11 January 2022
Accepted: 7 February 2022
Published: 13 February 2022

1. Introduction

Nowadays, in several different domains like precision agriculture as well as pharmaceutical, food and chemical manufacturing, it is very common to utilize many analytical sensors to comprehensively characterize complex systems under study and to monitor processes while they evolve over time [1]. Analyzing the data yielded by such sensors by means of appropriate statistical tools is challenging but crucial in order to obtain meaningful physico-chemical information and design efficient production monitoring and control schemes. In particular, in industrial applications, a relevant issue is how to integrate or fuse the data resulting from sensors of different nature, potentially installed at different locations in the plant and in real time.

Multivariate Statistical Process Control (MSPC) is a well-established tool to accomplish real time monitoring and control of industrial production, in particular Latent Variables-Based MSPC (LV-MSPC) [2–7]. Most LV-MSPC relies on so-called engineering process

variables [8], i.e., measured by on-line sensors controlling machinery settings (such as flow-meters, temperature and pressure probes, etc.) to build reference multivariate models for normal operating conditions (NOC), which are afterwards used to derive multivariate control charts and/or predicting quality attributes of finite product. More recently, thanks to technological developments, spectroscopic probes, especially near-infrared (NIR) ones, are extensively exploited [6,7,9–13] to monitor process evolution, or, in other words, to determine intermediate and final product quality parameters. Many studies in literature report on these aspects. Their results mainly refer to pilot scale plants [9,11,12,14] as well as to batch types of processes and seldom are engineering process variables and NIR measurements combined for constructing LV-MSPC models [6,10,14,15].

Fusing spectra with engineering variables is not a trivial task. However, process monitoring and control can greatly benefit from fusing these diverse data types, since, in this way, chemical composition-related information and physical and mechanical behavior/properties can be integrated.

This work focuses on a continuous styrenic polymer production process [16], monitored by means of NIR probes installed on-line in a production plant, as well as by standard process sensors. The main aim is to build real-time monitoring models to predict two of the main quality attributes of the final polymeric product by fusing NIR and process sensors' data. A preliminary feasibility study was recently conducted by the authors at the pilot-plant level [14].

Two aspects are particularly relevant for industry: (i) the possibility of estimating in real time the quality of a finite product, thus reducing the operational time and the amount of chemicals commonly required for laboratory off-line assessments by reference methods; and (ii) to reach the anticipated assessment of departure from desired quality before the end of production itself, in order to plan possible early modifications of the operating settings.

To this end, we investigated the application of multiblock chemometric methods [17–25] which are suitable to accomplish data fusion at low-level [26,27] and might bring interesting advantages with respect to alternative mid-level and high-level data integration strategies [26] especially in terms of model training, maintenance and interpretability. In fact, original variables are directly used without any compression steps, and it is possible to assess the salience of each block/type of sensors in the model, i.e., inspecting their degree of uniqueness or redundancy.

In particular, we compared a well-established multiblock MSPC approach, such as MultiBlock Partial Least Squares (MB-PLS) regression [21], with Response-Oriented Sequential Alternation (ROSA) [22]. The distinctive features of ROSA, which is also based on PLS regression [28,29], are: (i) to be invariant to block scaling and not to be affected by the spurious bias resulting from the combination of data blocks of different size (similarly to sequential orthogonal PLS (SO-PLS) [20]); and (ii) to be computationally efficient and capable of dealing with any number of blocks, also a very high number (differently from SO-PLS).

We tested models constructed on measurements yielded by sensors that were active at all different process stages (up to the process production end), as well as models where measurements from the last stage were excluded. This was in order to evaluate if polymer quality could be forecasted prior to the end of production. The results achieved, by both MB-PLS and ROSA, show satisfactory predictive performance for the determination of the two quality parameters investigated. At the same time, the most relevant data blocks were assessed.

2. Materials and Methods

2.1. Process Description

Data presented in the current work were collected on-line in an Acrylonitrile-Styrene-Butadiene (ABS) industrial production plant (full scale) operating in continuous process, owned by Versalis (ENI group). For the sake of simplicity, the plant can be regarded as divided into five different areas: (i) pre-poly/mixer, where the three precursor monomers (acrylonitrile, styrene and butadiene) are mixed together; (ii) reaction point A; (iii) reaction

point B; (iv) reaction point C; and (v) devolatilizer/cut zone, where the finite product is cut. Throughout all these areas seventy process sensors (PS), which measure temperatures, pressures, flow rates and motor speed, and four NIR probes are installed. The NIR probes are placed in four specific and crucial areas of the production plant: one where dissolution of butadiene in styrene occurs, before the addition of acrylonitrile; one in the pipe for the recovery of condensed reagents; one between the first and the second reaction points; and one at the very end of the process, just before the cut zone. Overall, both PS and NIR probes record data/spectra with a frequency of about one minute. In this study, data registered from January 2020 to May 2021 were analyzed, even if not all the data recorded during this period were considered in model building, due to production pauses and deviations from the operative conditions relevant for the current study.

2.2. Reference Analysis

Two different parameters have been considered for the evaluation of ABS quality. Nonetheless, because of confidential agreement restrictions with the company, their actual names will not be disclosed, but they will be referred to as Property 1 and Property 2. Properties 1 and 2 are assessed off-line by collecting ABS samples, i.e., final product, two (Property 1) and three (Property 2) times per day. Property 1 is related to ABS composition, i.e., the percentage of a certain chemical compound in the final product. On the other hand, Property 2 gives information about physical features of the product and the values of the related reference analysis are expressed in grams. In the period covered by this study 597 and 904 laboratory tests (homogeneously distributed all over the time period) were carried out to determine Property 1 and Property 2, respectively. Property 1 values ranged from 20 to 21.8%; Property 2 values ranged from 3.9 to 6.1 g.

2.3. NIR Spectroscopy

A Matrix FT-NIR spectrometer (Bruker Optics, Milan, Italy) was used to acquire spectra in the four different acquisition sites. The instrument was equipped with optical fibers (length: 100 m, diameter: 600 μm), whose probes (HT immersion probe, Drawing-no. 661.2350_1, Hellma GmbH and Co. KG, Müllheim, Germany) were directly connected to the four different acquisition sites located on the process pipe. Spectra were collected in transmission mode over the 12,500–4000 cm^{-1} spectral range, with a nominal resolution of 4 cm^{-1} (64 scans per sample).

2.4. Data Analysis

2.4.1. Data Block and Multiblock Arrangement

The ensemble of collected data was arranged into nine distinct data blocks, according to the data type and the acquisition area along the process: on the one hand, PS measurements were gathered in five blocks, one per every area of the plant (see also Section 2.1); on the other hand, NIR spectra were arranged into four blocks, each corresponding to an individual optical probe. In Table 1, the names and abbreviations (which will be hereafter used) of all the blocks are shown, together with their size and the location along the plant. This is also an indication of how they are ranked in time, being a continuous process.

For both multiblock approaches, the data blocks were assembled considering the chronological progression of the ABS production process and, therefore, based on the location of the different sensors along the production line. In other words, each data point present in the datasets refers to information collected at different times, but it is correctly matched to the same processed material (i.e., data are synchronized).

Figure 1 displays a schematic representation of the low-level data fusion strategy adopted.

Table 1. Data block description.

Block Full Name	Block Abbreviated Name	Data Type	No. of Variables [1]	Order
NIR dissolution	NIR-diss	NIR Spectra	390	1
Prepoli/Mixer	Prep/mix	PS	7	2
NIR condensation	NIR-cond	NIR Spectra	390	3
Reaction Point A	RP-A	PS	15	4
NIR Reaction Point A	NIR-RP-A	NIR Spectra	390	5
Reaction Point B	RP-B	PS	10	6
Reaction Point C	RP-C	PS	8	7
Devolatilizer/cut zone	Devo/cut	PS	30	8
NIR cut zone	NIR-cut	NIR Spectra	390	9

[1] For NIR data blocks, the number of variables is equal to the spectra wave numbers, whereas for PS data blocks it is equal to the number of PS present in the respective plant area. The column "Order" highlights how the process evolves chronologically.

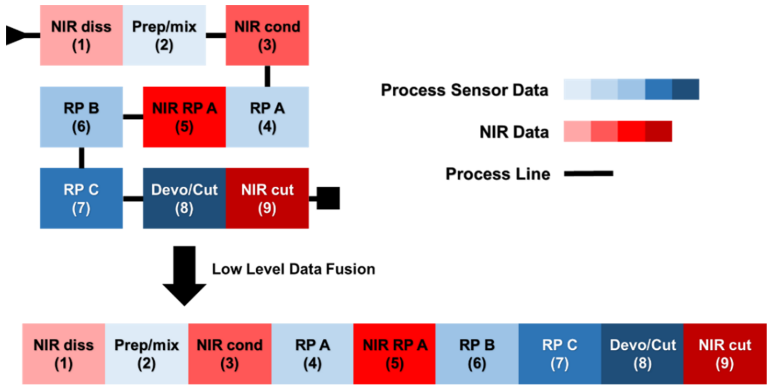


Figure 1. Schematic representation of the low-level data fusion approach resorted to in this study. Values in brackets indicate the chronological order of the data blocks.

2.4.2. Preprocessing

Individual block preprocessing

Prior to the multiblock modeling phase, each data set was preprocessed individually. In particular, variables in each PS data block were scaled to unit variance (different in nature and scales) whereas spectra, in each NIR data block, were baseline-corrected by using automatic weighted least squares [30]. Moreover, only the spectral range from 6500 to 5000 cm^{-1} (the sole one exhibiting spectral bands ascribable to either reactants or products) was taken into account for subsequent model training. Figure 2 shows the effect of the baseline correction executed on the NIR spectra of the NIR-RP-A data block.

Multiblock preprocessing

After the individual preprocessing of the single blocks, each data set was scaled to unit block variance (including column mean-centering) prior to MB-PLS [21]. In fact, MB-PLS operates directly on row-wise concatenated data blocks and a fair block contribution has to be assured.

Concerning ROSA, the individual pre-processed blocks were just mean-centered since such a method treats one block at a time, as it will be detailed in the following sections.

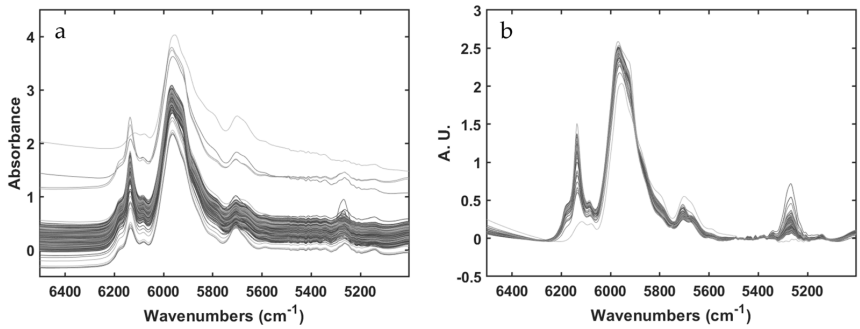


Figure 2. Spectra collected at NIR-RP-A, data block before (**a**) and after (**b**) baseline correction using automatic weighted least square method.

2.4.3. MB-PLS

We exploited here the MB-PLS implementation originally proposed by Westerhuis and Coenegracht [31] which can be looked at as standard PLS with appropriate block scaling steps as described in [21]. Thus, MB-PLS is an extension of the classical PLS regression [28] for applications involving different data blocks that share the same number of rows (observations), relating to the data matrix $\mathbf{X}$, resulting from the row-wise concatenation of N different data blocks (Equation (1)):

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N] \tag{1}$$

to the response(s) of interest.

This method provides global (also called *super*-) scores, weights, loadings and regression coefficients, as well as local (also called *block*-) scores and weights for each data block, as it is shown in Equations (2)–(5):

$$\mathbf{w}_b = \mathbf{X}^{\mathbf{T}}{}_b * \mathbf{u}/\mathbf{u}^{\mathbf{T}}\mathbf{u} \tag{2}$$

$$\mathbf{t}_b = (\mathbf{X}_b * \mathbf{w}_b)/\sqrt{\mathbf{n}_b} \tag{3}$$

$$\mathbf{w} = \mathbf{T}^{\mathbf{T}} * \mathbf{u}/\mathbf{u}^{\mathbf{T}}\mathbf{u} \tag{4}$$

$$\mathbf{t} = \mathbf{T} * \mathbf{w}/\mathbf{w}^{\mathbf{T}}\mathbf{w} \tag{5}$$

where $\mathbf{n}_b$ is the number of variables in a given block, $\mathbf{t}_b$ and $\mathbf{w}_b$ are the local scores and weights, respectively, whereas $\mathbf{t}$ and $\mathbf{w}$ are the global (*super*) scores and weights. $\mathbf{T}$ is yielded by the concatenation of all $\mathbf{t}_b$.

This way, it is possible to assess the contribution of each data block (analyzing $\mathbf{w}_b$ for the prediction of the response variable/s $\mathbf{y}/\mathbf{Y}$, improving the process understanding).

2.4.4. ROSA

Response-Oriented Sequential Alternation (ROSA) is a multiblock regression method proposed by Liland et al. [22] that is also based on PLS regression. Different from MB-PLS, in that ROSA is a sequential algorithm, similar to, e.g., SO-PLS [20], which renders the method invariant with respect to block-scaling (blocks are just mean centered), as well as to block ordering, differently from SO-PLS. These features allow dealing with a large number of blocks of different dimensions.

Moreover, ROSA exhibits a high computational efficiency, as it does not require the iterative convergence of an optimization criterion, and because only the response is deflated, not all the blocks. In fact, each PLS component is selected from a single block, picking among the various covariance-maximizing candidate components, estimated from each data block, the one returning the smallest prediction residuals. Successive components

are constrained to be orthogonal to the subspace spanned by the previously winning components. Thus, scores' and loadings' orthogonality is ensured.

The ROSA algorithm for a single response variable, **y**, is summarized in the following equations:

$$\mathbf{w}_b = \mathbf{X}_b{}^{\mathrm{T}} * \mathbf{y} \quad (6)$$

$$\mathbf{t}_b = \mathbf{X}_b * \mathbf{w}_b \quad (7)$$

$$\mathbf{r}_b = \mathbf{y} - \mathbf{t}_b \, \mathbf{t}_b{}^{\mathrm{T}} \mathbf{y} \quad (8)$$

where $\mathbf{X}_b$ is a single data block, while $\mathbf{w}_b$, $\mathbf{t}_b$ and $\mathbf{r}_b$ are block weights, scores and residuals, respectively. The first component is selected as the one computed from the b_{th}-block yielding the smallest residuals ($\mathbf{r}_b$), and $\mathbf{t}_1$ are taken to be equal to $\mathbf{t}_b$ of the winning block. The corresponding weights and scores are normalized (and also orthogonalized with respect to the preceding components from the second component on). The **y**-loadings are finally estimated as:

$$\mathbf{q}_a = \mathbf{y}^{\mathrm{T}} \, \mathbf{t}_a \quad (9)$$

where $\mathbf{t}_a$ are the scores previously selected for the a_{th} LV.

X-loadings (**P**) and PLS regression coefficients (**b**) (and possibly a constant term b_0) can be estimated according to the Equations (10)–(12), after selecting the number of optimal LVs and collecting the corresponding scores, weights, **y**-loadings in matrix array **T**, **W** and **Q**.

$$\mathbf{P} = \mathbf{X}^{\mathrm{T}} \, \mathbf{T} \quad (10)$$

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{Q} \quad (11)$$

$$\mathbf{b}_0 = \mathbf{y}_{\mathrm{m}} - \mathbf{x}_{\mathrm{m}} * \mathbf{b} \quad (12)$$

where $\mathbf{y}_{\mathrm{m}}$ is the mean of **y** and $\mathbf{x}_{\mathrm{m}}$ is a vector with the mean for each variable of **X**.

Thus, each selected LV in ROSA encodes information proceeding only from the winning b_{th}-block (the one achieving smallest residuals according to Equation (8)), and all LVs are orthogonal. It is important to notice that all blocks are always candidates at each algorithmic step. Therefore, consecutive LVs can depict information from the same block previously selected, or from a different one.

2.4.5. Multiblock Models Building

With the aim of developing predictive models for the two parameters taken into account in this study and assessing which are the most important data blocks for their estimation, both MB-PLS and ROSA were investigated.

All the available data were split into calibration and validation sets for both Property 1 and Property 2. In order to assess models' performance in a scenario mimicking a real-time application, the calibration sets comprised data collected during the year 2020 (~70% of total data), whereas the validation sets comprised data collected in 2021. Clearly, only samples, i.e., time points, for which the offline reference measurement were available were taken into account.

The two optimized best-performing models were finally utilized for assessing the values of Property 1 and 2 at time points where no reference data were acquired, in order to check whether the resulting estimations spanned a similar properties values range with respect to close time points.

In order to establish the complexity, i.e., number of PLS components, of each model, venetian blinds cross-validation with ten cancellation groups for Property 1 and four cancellation groups for Property 2 was resorted to. Model reliability was determined in terms of both root mean square error in cross-validation (RMSECV) and root mean square error in prediction (RMSEP).

Data blocks were preprocessed as described in Section 2.4.2.

For both MB-PLS and ROSA, the contribution of each block and block variables in the final predictive model was assessed by investigating the PLS regression coefficients and

Variable Importance in Prediction (VIP) [32,33]. PLS block-weights were also inspected but, for the sake of brevity, the related figures are not reported, as the provided information was similar to that obtained by regression coefficients.

2.5. Software

All the chemometric analyses were performed using routines and toolboxes implemented in the MATLAB environment (the Mathworks Inc., Natick, MA, USA).

MB-PLS has been calculated through the PLS-Toolbox version 8.9 (Eigenvector Research Inc., Wenatchee, WA, United States).

ROSA (with options for venetian blind cross-validation, VIP calculation and validation sample response prediction) was implemented by the authors based on the MATLAB code provided in ref. [22].

3. Results

3.1. Property 1 Prediction

When all the available data blocks (PS and NIR measurements for all plant areas) were simultaneously modelled ROSA resulted to be the most performant method for the prediction of Property 1, yielding a RMSEP of 0.14%. On the other hand, MB-PLS returned a RMSEP value of 0.2%. This difference, however, is not substantial. The results are shown in Table 2 and Figure 3. ROSA selected only three of the nine blocks under study, two of which, Devo/cut and NIR-cut, relate to the last stage of the process, where the polymerization is over and the product is ready to be cut. Furthermore, among the 13 latent variables selected through the cross-validation procedure (aimed at minimizing RMSECV), eight were calculated from the NIR-cut block, which highlights a crucial relevance of the final NIR sensor, in this case, for the quality prediction. Figure 3a shows how the predictions for the objects of the validation set are homogeneously distributed within the expected range of the quality parameter concerned. In Figure 3b–d the PLS regression coefficients associated to the three blocks selected by ROSA are represented (the red stars denote variables/spectral regions whose VIP scores were higher than one). In the RP-A data block (selected only one time out of 13) only three temperature sensors were found relevant for Property 1 prediction, whereas in Devo/cut and NIR-cut data blocks all the sensors and nearly all the spectral regions sampled were somewhat important. In Figure 3d it is evident that the largest (in absolute value) regression coefficients are those corresponding to bands centered at 5900 cm^{-1} and 5250 cm^{-1} that can be ascribed to the investigated ABS compound.

Table 2. Results yielded by MB-PLS and ROSA for the prediction of Property 1.

Model ID	Blocks Entering the Model	LVs	RMSEC (%)	RMSECV (%)	RMSEP (%)
MB PLS all	All	11	0.12	0.16	0.20
MB PLS no cut zone	1 to 7	11	0.13	0.17	0.23
MB PLS only PS	2–4–6–7–8	11	0.24	0.26	0.38
MB PLS only NIR	1–3–5–9	10	0.13	0.15	0.22
MB PLS only NIR no cut zone	1–3–5	8	0.14	0.15	0.22
ROSA all [1]	4(1)–8(4)–9(8)	13	0.11	0.14	0.13
ROSA no cut zone	3(6)–4(1)–5(3)–6(2)	12	0.15	0.18	0.2
ROSA only PS	2(1)–4(6)–7(3)	10	0.23	0.25	0.31
ROSA only NIR	9(8)	8	0.12	0.13	0.14
ROSA only NIR no cut zone	3(12)–5(2)	14	0.16	0.18	0.19

[1] the values in brackets indicate the number of times a certain block was selected by the ROSA algorithm.

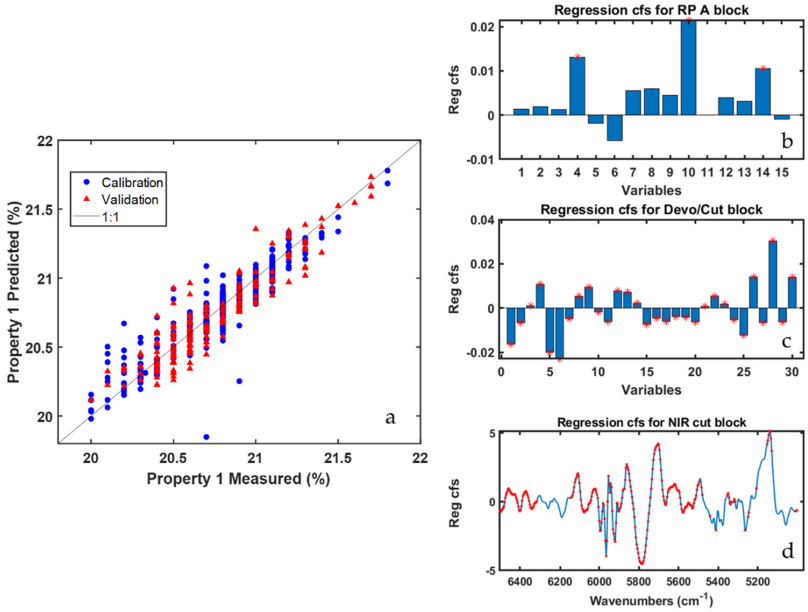


Figure 3. ROSA results for Property 1 prediction (all data blocks were modelled simultaneously). Predicted vs. measured value plot (**a**); regression coefficients for the RP-A (**b**); Devo/cut (**c**); and NIR cut (**d**) data blocks. Red stars indicate variables having VIP scores higher than one.

Although such results might already be considered relatively satisfactory from a predictive point of view, two additional aspects would be worth investigating: (i) whether reasonably good quality prediction of Property 1 values could be obtained before the product is cut (i.e., without relying on sensors installed within the cut area); and (ii) whether the exclusive use of spectral sensors or process sensors could be sufficient for a reliable estimation of this quality index. To this end, in addition to the dataset containing all the blocks, MB-PLS and ROSA models were calculated using fused datasets comprising only the blocks before the cut zone, only PS data and only NIR data (both including and excluding the spectra contained in the NIR-cut block), respectively.

Table 2 reports the results of all the computed multiblock prediction models related to Property 1. It is possible to observe that prediction errors resulting from ROSA are systematically lower than the one obtained by means of MB-PLS. It is also clear how NIR data are far more important for the prediction of Property 1 than PS data. In fact, when ROSA is run on both block types, components from NIR data sets are more often selected than those computed from PS data sets. Moreover, in MB-PLS models, variables related to NIR blocks are always relevant for Property 1 prediction. In addition, the RMSEP of models that are calculated using only NIR data is comparable to that of models using both PS and NIR data, while using only PS data blocks results in a significant increase of the prediction error in calibration, cross-validation and external validation. This is somehow expected, as Property 1 is linked to ABS chemical composition and, therefore, an analytical technique like NIR spectroscopy is definitely more suitable for its determination than more standard engineering PS probes, which only indirectly reflect how fluctuations in the process operating conditions may affect the polymer characteristics.

Since ROSA models always selected components estimated from the blocks located on the plant cut area, i.e., blocks eight and nine, we also decided to calibrate ROSA models (using both PS and NIR data and only NIR data) excluding completely such blocks from the computational procedure (see 'ROSA no cut zone' and 'ROSA only NIR no cut zone' in Table 2, respectively). In both cases, RMSEP values for models not including the cut area,

were found higher, yet acceptable by process operators. This clearly makes it possible to retrieve reasonable Property 1 value estimate before the completion of the ABS production process. Moreover, similar prediction errors were obtained by using only NIR blocks or when combining NIR and PS blocks. Hence, two possible pathways can be envisioned for the real-time prediction and control of Property 1: (i) resorting to both data types and getting a clearer idea of the important process areas/sensors all along the production plant; or (ii) just exploiting NIR spectra for more efficient data management and to deal with less noisy data.

In order to evaluate the role of all types of sensors, Figure 4 displays the results yielded by the 'ROSA no cut zone' model. It is worth mentioning that half of the blocks selected by the ROSA algorithm relate to the reaction points A and B, whereas the other half to the NIR-cond data block, whose respective probe is right before these reaction points. Looking at the order (not reported for the sake of brevity) in which blocks were selected by ROSA, it can be observed how the winning blocks for the first five latent variables were RP-B (picked only one time) and NIR-RP-A (picked four times). For the remaining model dimensions, NIR-cond was selected six times in a row, while RP-A and RP-B one each. Details about the selection order are useful to assess which blocks, i.e., areas of the plant, encode the most important information for the prediction of the investigated quality parameter.

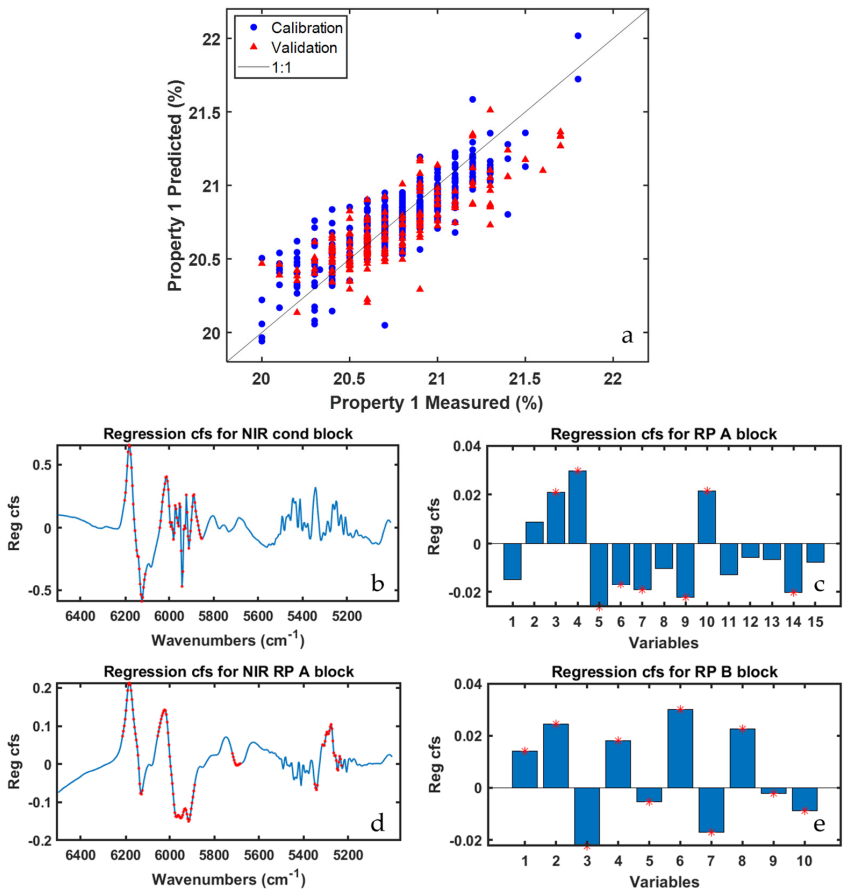


Figure 4. ROSA results for Property 1 prediction ('ROSA no cut zone' model). Predicted vs. measured value plot (**a**); regression coefficient for NIR cond (**b**); RP A (**c**); NIR RP A (**d**); and RP B (**e**) data blocks. Red stars indicate variables having VIP scores higher than one.

Figure 4b,d show the regression coefficients for the two aforementioned NIR blocks, with NIR-RP-A exhibiting a larger number of spectral variables characterized by VIP scores higher than one, especially in the region between 5400 cm^{-1} and 5250 cm^{-1}, that are ascribable to the stretching of a functional group of one of the three precursor compounds on which Property 1 directly depends. Conversely, in Figure 4c,e the regression coefficients for the PS data blocks are graphed: the most significant variables, according to their respective VIP values, are almost all related to temperature and motor speed sensors installed in different subzones of the reaction points A and B.

3.2. Property 2 Prediction

The same model building strategy described before was finally followed for the prediction of Property 2. Table 3 reports the results obtained by means of both MB-PLS and ROSA. ROSA, when all the available data blocks were simultaneously modelled, did not select any cut area block, therefore the 'ROSA no cut zone' model was not trained in this case.

Table 3. Results yielded by MB-PLS and ROSA for the prediction of Property 2.

Model ID	Blocks Entering the Model	LVs	RMSEC (g)	RMSECV (g)	RMSEP (g)
MB PLS all	All	10	0.25	0.27	0.34
MB PLS no cut zone	1 to 7	8	0.27	0.29	0.37
MB PLS only PS	2–4–6–7–8	9	0.27	0.29	0.35
MB PLS only NIR	1–3–5–9	7	0.34	0.34	0.48
MB PLS only NIR no cut zone	1–3–5	6	0.36	0.37	0.5
ROSA all [1]	2(1)–4(1)–5(1)–6(1)	4	0.32	0.33	0.46
ROSA only PS	2(1)–4(1)–6(1)	3	0.32	0.33	0.45
ROSA only NIR	5(6)–9(3)	9	0.33	0.34	0.52
ROSA only NIR no cut zone	5(8)	8	0.33	0.34	0.52

[1] The values in brackets indicate the number of times a certain block was selected by the ROSA algorithm.

MB-PLS models calibrated by using (i) all the data blocks or (ii) only PS data returned the most satisfactory results, contrary to the results obtained for Property 1. In fact, the influence NIR spectra have on the estimation of Property 2 prediction is not predominant, except for the NIR-RP-A block, which was selected many times by the ROSA algorithm and whose variables always showed VIP scores higher than one in MB-PLS. These results can be interpreted in the light of the fact that Property 2 is not linked to the chemical composition of ABS but evaluates the performance of the finite product as determined by mechanical/physical tests. Subsequently, it is undoubtedly more affected by variability occurring in the processing steps, and can change significantly even if the aforementioned chemical composition does not change. RMSEP increased up to 0.52 g when no PS block was considered. However, for models built without PS data, MB-PLS achieved a slightly better performance than ROSA (0.48–0.5 g vs. 0.52 g). These results suggested how the exclusive use of NIR sensors is not sufficient for a reliable estimation of Property 2.

Overall, MB-PLS showed a better prediction performance for Property 2. The best results were obtained by the 'MB PLS all' model (RMSEP = 0.34 g), even though 'MB PLS no cut zone' and 'MB PLS only PS' provided similar results.

In Figure 5 is where the predicted vs. measured value plot resulting from the 'MB-PLS all' model is shown. By inspecting the corresponding residuals plot (not shown for the sake of brevity) it can be observed that, on average, the 2021 production campaign (validation set), yielded lower values of Property 2 than that conducted in 2020 (calibration set). This deviation explains the relatively high difference between RMSEP and RMSEC and RMSECV. However, the presence of a reasonable amount of validation samples in the whole calibration range was guaranteed and the company deemed the prediction error acceptable for routine monitoring.

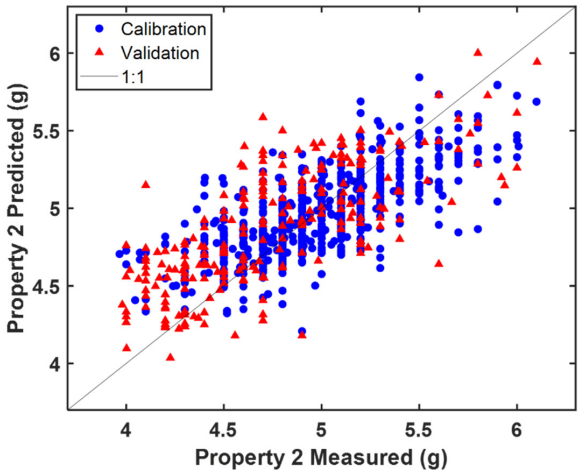


Figure 5. Predicted vs. measured value plot resulting from the 'MB-PLS all' model.

In Figure 6 the 'MB-PLS all' model regression coefficients are reported. All PS were found to be important for the prediction of Property 2 based on their VIP scores values. For what concerns the NIR blocks regression coefficients, the NIR-RP-A is confirmed to be the block with the largest number of highly predictive spectral regions, which are mainly related to the three precursors monomers of ABS. For the other NIR blocks, relevant regions of interest were found in correspondence of the absorption bands centered at 5900 cm^{-1} and 6100 cm^{-1}, respectively.

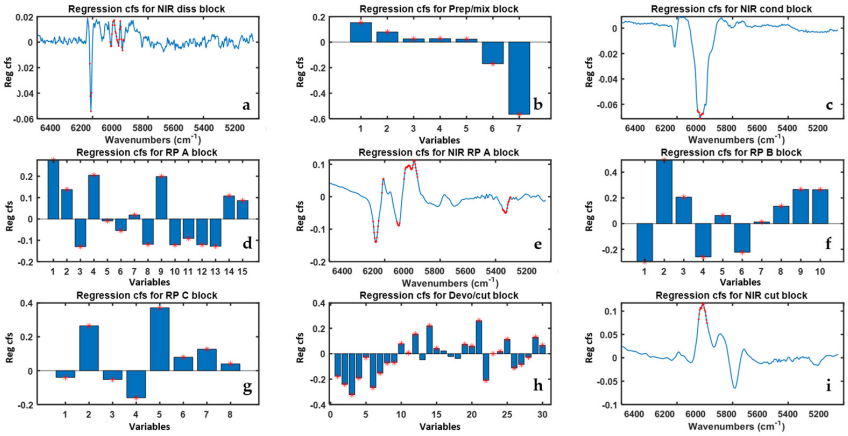


Figure 6. Regression coefficients resulting from the 'MB-PLS all' model for each data block the letters (**a**–**i**) refer to the different block whose name is reported on top. Red stars indicate variables exhibiting VIP scores higher than one.

3.3. Real-Time Predictions

Finally, Figure 7 illustrates the predicted values of Property 1 obtained through the ROSA model constructed on all data blocks (Table 2, row 1) for the time points for which reference response measurements were not acquired.

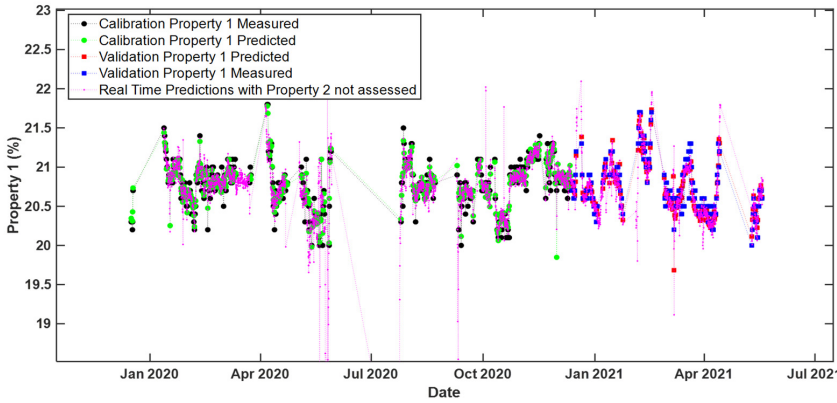


Figure 7. Real time predictions of Property 1 (i.e., time evolution of the measured and predicted values). The predictions were obtained by means of the 'ROSA all' model. Legend: black circles—calibration set measured values; green circles—calibration set predicted values; blue squares—validation set measured values; red squares—validation set predicted values; magenta dots—predicted values related to time points for which no reference response measurements were available. For ease of visualization only every 2 h predictions during the considered time period are shown.

These predicted values span a range very similar to that covered within both the calibration and the validation set. A few slight deviations were observed, interestingly right after specific shut-down time periods: such deviations may, in fact, arise from the fact that many industrial processes (including polymerization processes) take a certain time to readapt to NOC conditions after particular external interventions (e.g., cleaning, maintenance, etc.).

Similar results were obtained for real-time predictions with the model 'MB-PLS no cut zone' for Property 2, as shown in Figure 8.

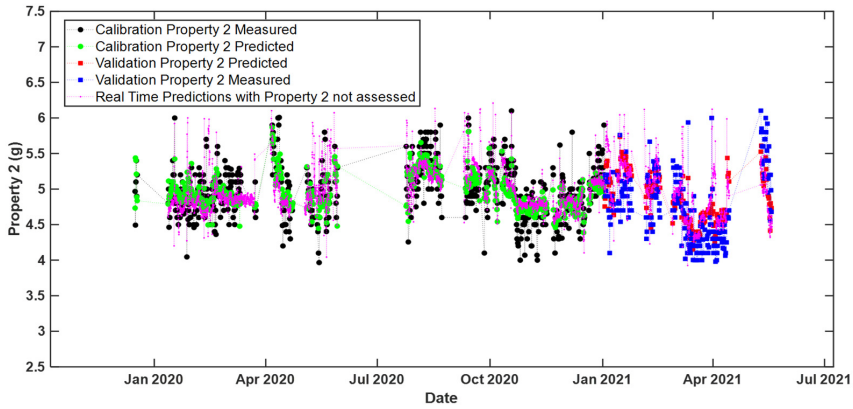


Figure 8. Real time predictions of Property 2 (i.e., time evolution of the measured and predicted values). The predictions were obtained by means of the 'MB PLS no cut zone' model. Legend: black circles—calibration set measured values; green circles—calibration set predicted values; blue squares—validation set measured values; red squares—validation set predicted values; magenta dots—predicted values related to time points for which no reference response measurements were available. For ease of visualization only every 2 h predictions during the considered time period are shown.

4. Conclusions

This work demonstrated how multiblock approaches could be used for the construction of reliable and robust real-time monitoring models for the on-line prediction of industrial quality parameters of ABS. In fact, the data partition in different blocks and the low-level data fusion strategy adopted here permitted to improve ABS production process understanding, enabling the assessment of the most crucial plant areas and the relevant sensors for the prediction of such specific parameters. Moreover, the application of these approaches is essential when two or more different analytical platforms of different nature, like the NIR spectrometer and more standard engineering process sensors, are simultaneously used to control any generic production process.

More specifically, in this article, both MB-PLS and ROSA allowed performant predictive models to be constructed for the two properties under study (i.e., Property 1 and 2). In particular, for the prediction of Property 1, ROSA resulted in a lower RMSEP compared to MB-PLS, highlighting the importance of NIR data over process sensor data when a chemical composition-related quality index is to be estimated. On the other hand, Property 2 was more efficiently predicted by a MB-PLS method, which pointed out a higher relevance of process sensors compared to NIR data when, instead, physical features need to be assessed.

Furthermore, models computed without taking into account measurements related to the final area of the plant (cut zone) provided comparable prediction errors with respect to the best models built on all the ensemble of available data. This is of great industrial interest, since, in principle, ABS quality could be determined before its production is completed, which might allow possible modifications of the plant settings and/or changes in the operating conditions to be planned in advance and with reduced costs.

In conclusion, these approaches could help in: (i) accelerating decision making and troubleshooting; (ii) reducing the amount of chemical waste generated in full-scale plants; (iii) decreasing the number of off-line laboratory tests required for quality control; and (iv) facilitating any type of operation along the production line as well as possible fault detection and diagnosis.

Author Contributions: Conceptualization, R.V., F.B., E.M., A.F. and M.C.; methodology, L.S., R.V., D.T., F.B., A.P., E.M., A.F. and M.C.; software, L.S., D.T., A.P. and M.C.; validation, R.V., F.B., A.P., E.M., A.F. and M.C.; investigation, L.S., R.V., D.T., F.B., A.P., E.M., A.F. and M.C.; resources, F.B., A.P., E.M. and A.F.; data curation, L.S., D.T., F.B., E.M. and A.P.; writing—original draft preparation, L.S. and D.T.; writing—review and editing, L.S., R.V., F.B., A.F. and M.C.; supervision, R.V., F.B. and M.C.; funding acquisition, A.F. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: LS post doctoral position was granted by Emilia Romagna region under POR FSE project "Data analytics per la REAlizzazione di sistemi predittivi e Monitoraggio real TIME di processi produttivi in industria 4.0 (DREAMTIME)" PA n° 2019-13551/RER.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because of confidential agreement restrictions with the company. Requests to access the datasets should be directed to erik.mantovani@versalis.eni.com.

Acknowledgments: The authors acknowledge Angelo Ferrando of Versalis (ENI) Company for supplying data used for the current study and fruitful discussion of the results, and Federico Marini for useful suggestions and discussion concerning ROSA.

Conflicts of Interest: Authors A.F., F.B., A.P. and E.M. are employed by Versalis (ENI) SpA. The remaining 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.

References

1. Bowler, A.L.; Bakalis, S.; Watson, N.J. A review of in-line and on-line measurement techniques to monitor industrial mixing processes. *Chem. Eng. Res. Des.* **2020**, *153*, 463–495. [CrossRef]
2. MacGregor, J.F.; Bruwer, M.J.; Miletic, I.; Cardin, M.; Liu, Z. Latent Variable Models and Big Data in the Process Industries. *IFAC-PapersOnLine* **2015**, *48*, 520–524. [CrossRef]
3. Kourti, T. Multivariate Statistical Process Control and Process Control, Using Latent Variables. In *Comprehensive Chemometrics*, 2nd ed.; Brown, D.S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 4, pp. 275–303.
4. Ferrer-Riquelme, A.J. Statistical Control of Measures and Processes. In *Comprehensive Chemometrics*, 2nd ed.; Brown, D.S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 4, pp. 215–236.
5. Wold, S.; Kettaneh-Wold, N.; MacGregor, J.F.; Dunn, K.G. Batch Process Modeling and MSPC. In *Comprehensive Chemometrics*, 2nd ed.; Brown, D.S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 4, pp. 305–332.
6. Morris, J.; Martin, E.; Stewart, D. Batch Process Monitoring through the integration of Spectral and Process Data. *IFAC-PapersOnLine* **2005**, *38*, 3–18. [CrossRef]
7. Aguado, D.; Ferrer, A.; Seco, A.; Ferrer, J. Comparison of different predictive models for nutrient estimation in a sequencing batch reactor for wastewater treatment. *Chemom. Intel. Lab. Syst.* **2006**, *84*, 75–81. [CrossRef]
8. Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven Soft Sensors in the process industry. *Comput. Chem. Eng.* **2009**, *33*, 795–814. [CrossRef]
9. 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*, 3164–3172. [CrossRef]
10. Lourenço, N.D.; Lopes, J.A.; Almeida, C.F.; Sarraguça, M.C.; Pinheiro, H.M. Bioreactor monitoring with spectroscopy and chemometrics: A review. *Anal. Bioanal. Chem.* **2012**, *404*, 1211–12137. [CrossRef] [PubMed]
11. Avila, C.; Mantzaridis, C.; Ferré, J.; de Oliveira, R.R.; Kantojärvi, U.; Rissanen, A.; Krassa, P.; de Juan, A.; Muller, F.L.; Hunter, T.N.; et al. Acid number, viscosity and end-point detection in a multiphase high temperature polymerization process using an online miniaturised MEMS Fabry-Pérot interferometer. *Talanta* **2021**, *224*, 121735. [CrossRef] [PubMed]
12. Sousa, B.V.; Silva, F.; Reis, M.A.M.; Lourenço, N.D. Monitoring pilot-scale polyhydroxyalkanoate production from fruit pulp waste using near-infrared spectroscopy. *Biochem. Eng. J.* **2021**, *176*, 108210. [CrossRef]
13. Grassi, S.; Strani, L.; Casiraghi, E.; Alamprese, C. Control and monitoring of milk renneting using FT-NIR spectroscopy as a process analytical technology tool. *Foods* **2019**, *8*, 405. [CrossRef]
14. Strani, L.; Mantovani, E.; Bonacini, F.; Marini, F.; Cocchi, M. Fusing NIR and Process Sensors Data for Polymer Production Monitoring. *Front. Chem.* **2021**, *9*, 748723. [CrossRef] [PubMed]
15. De Oliveira, R.R.; Avila, C.; Bourne, R.; Muller, F.; de Juan, A. Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control. *Anal. Bioanal. Chem.* **2020**, *412*, 2151–2163. [CrossRef] [PubMed]
16. Scheirs, J.; Priddy, D. *Modern Styrenic Polymers: Polystyrenes and Styrenic Copolymers*; Scheirs, J., Priddy, D., Eds.; J. Wiley and Sons, Ltd.: Chichester, UK, 2003.
17. Mishra, P.; Roger, J.M.; Jouan-Rimbaud-Bouveresse, D.; Biancolillo, A.; Marini, F.; Nordon, A.; Rutledge, D.N. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC* **2021**, *137*, 116206. [CrossRef]
18. Vitale, R.; de Noord, O.E.; Westerhuis, J.A.; Smilde, A.K.; Ferrer, A. Divide et impera: How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding. *J. Chemometr.* **2020**, *35*, e3266. [CrossRef]
19. Campos, M.P.; Reis, M.S. Data preprocessing for multiblock modelling—A systematization with new methods. *Chemom. Intel. Lab. Syst.* **2020**, *199*, 103959. [CrossRef]
20. Biancolillo, A.; Næs, T. The sequential and orthogonalized PLS regression for multiblock regression: Theory, examples, and extensions. In *Data Fusion Methodology and Applications*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 157–177.
21. Westerhuis, J.A.; Kourti, T.; MacGregor, J.F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **1998**, *12*, 301–321. [CrossRef]
22. Liland, K.H.; Næs, T.; Indahl, U.G. ROSA—A fast extension of partial least squares regression for multiblock data analysis. *J. Chemom.* **2016**, *30*, 651–662. [CrossRef]
23. El Ghaziri, A.; Cariou, V.; Rutledge, D.N.; Qannari, E.M. Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K + 1) datasets. *J. Chemom.* **2016**, *30*, 420–429. [CrossRef]
24. Löfsted, T.; Trygg, J. OnPLS—A novel multiblock method for the modelling of predictive and orthogonal variation. *J. Chemometr.* **2011**, *25*, 441–455. [CrossRef]
25. Tauler, R.; Maeder, M.; de Juan, A. Multiset data analysis: Extended multivariate curve resolution. In *Comprehensive Chemometrics*, 2nd ed.; Brown, D.S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2020; Volume 2, pp. 305–336.
26. Cocchi, M. Introduction: Ways and Means to Deal with Data from Multiple Sources. In *Data Fusion Methodology and Applications*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 1–25.
27. Smilde, A.K.; van Mechelen, I. A Framework for Low-Level Data Fusion. In *Data Fusion Methodology and Applications*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 27–50.
28. Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [CrossRef]

29. Ferrer, A.; Aguado, D.; Vidal-Puig, S.; Prats, J.M.; Zarzo, M. PLS: A versatile tool for industrial process improvement and optimization. *Appl. Stoch. Models Bus. Ind.* **2008**, *24*, 551–567. [CrossRef]
30. Wise, B.M.; Gallagher, N.B.; Bro, R.; Shaver, J.; Winding, W.; Scott-Coch, R. *Chemometrics Tutorial for PLS Toolbox and Solo*; Eigenvector Research, Inc.: Wenatchee, WA, USA, 2008; p. 173.
31. Westerhuis, J.A.; Coenegracht, P.M. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *J. Chemom.* **1997**, *11*, 379–392. [CrossRef]
32. Wold, S.; Johansson, E.; Cocchi, M. PLS: Partial least squares projections to latent structures. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp. 523–550.
33. Favilla, S.; Durante, C.; Li Vigni, M.; Cocchi, M. Assessing feature relevance in NPLS models by VIP. *Chemom. Intell. Lab. Syst.* **2013**, *129*, 76–86. [CrossRef]

PAPER III

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

Daniele Tanzilli, Lorenzo Strani, Francesco Bonacini, Angelo Ferrando, Marina Cocchi, Caterina Durante

Analytica Chimica Acta 1316 (2024) 342851
https://doi.org/10.1016/j.aca.2024.342851

Analytica Chimica Acta 1316 (2024) 342851

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

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

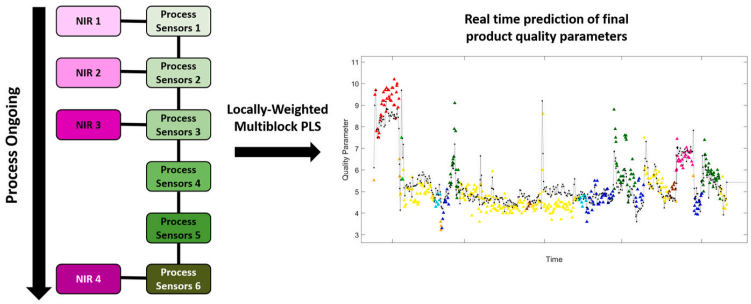
Daniele Tanzilli [a,b], Lorenzo Strani [a,*], Francesco Bonacini [c], Angelo Ferrando [c], Marina Cocchi [a], Caterina Durante [a]

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

HIGHLIGHTS

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

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Prof. L. Buydens

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

ABSTRACT

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

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

https://doi.org/10.1016/j.aca.2024.342851
Received 22 December 2023; Received in revised form 5 May 2024; Accepted 7 June 2024
Available online 8 June 2024

stage sensors play crucial roles in predicting product quality. A reasonable estimation of quality parameters before production completion has been achieved. On the other hand, the proposed LW-MB-PLS, while comparable in predictive performances, allowed reducing systematic prediction errors for specific products.
Significance: This study contributes valuable insights for continuous production processes, aiding plant operators and paving the way for advancements in online quality prediction and control. Furthermore, it is implemented as a locally weighted extension of MB-PLS.

1. Introduction

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

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

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

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

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

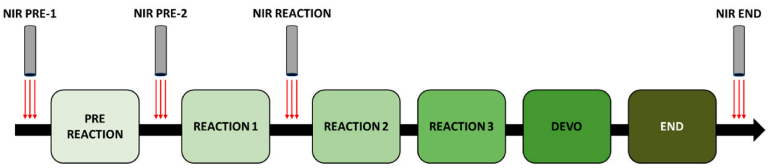


Fig. 1. Schematic representation of the ABS production line. The green blocks represent the six different sections into which the PS have been divided, whereas the gray bars and the red arrows represent the positions where the four on-line NIR probes were placed. (For interpretation of the references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

2. Materials and methods

2.1. Process and sampling

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

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

2.2. ABS quality parameters

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

2.3. NIR measurements

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

2.4. Data analysis

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

2.4.1. Data synchronization

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

2.4.2. Single block data preprocessing

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

2.4.3. Multiblock methods

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

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

Specifically, each PLS component is selected from a single block,

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

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

$$\mathbf{w}_b = \mathbf{X}_b^{T*} \mathbf{y} \quad (1)$$

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

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

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

$$q_a = \mathbf{y}^T \mathbf{t}_a \quad (4)$$

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

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

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

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

$$\mathbf{b}_0 = \mathbf{y}_m - \mathbf{x}_m^* \mathbf{b} \quad (7)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

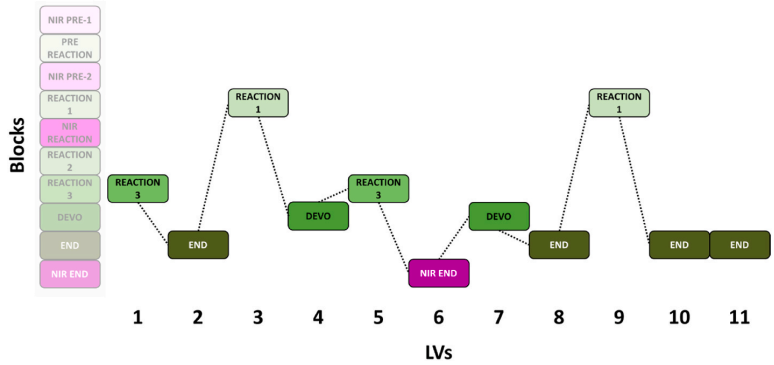


Fig. 2. ROSA model (using all the available blocks) for QP2 prediction. The winning block selected for each LVs is shown in correspondence of the component number. The left bar reports the time order of the blocks along the process.

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

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

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

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

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

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

2.5. Software

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

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

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

3. Results and discussion

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

3.1. ROSA results

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

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

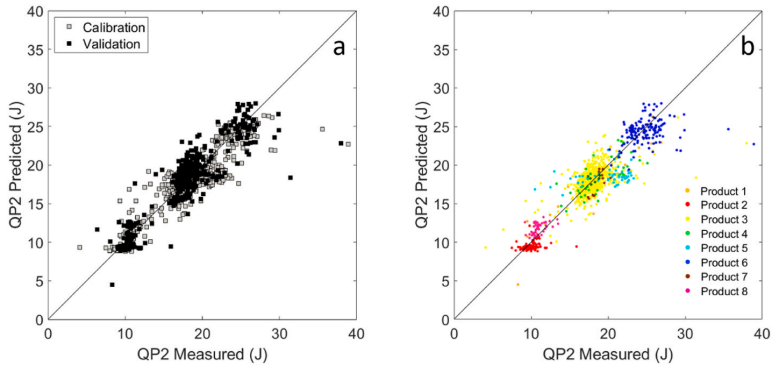


Fig. 3. Plots of predicted vs measured values of QP2 obtained by the ROSA model using all the available blocks. In (a) Samples are colored according to calibration (gray) and validation (black) and in (b) according to ABS product type.

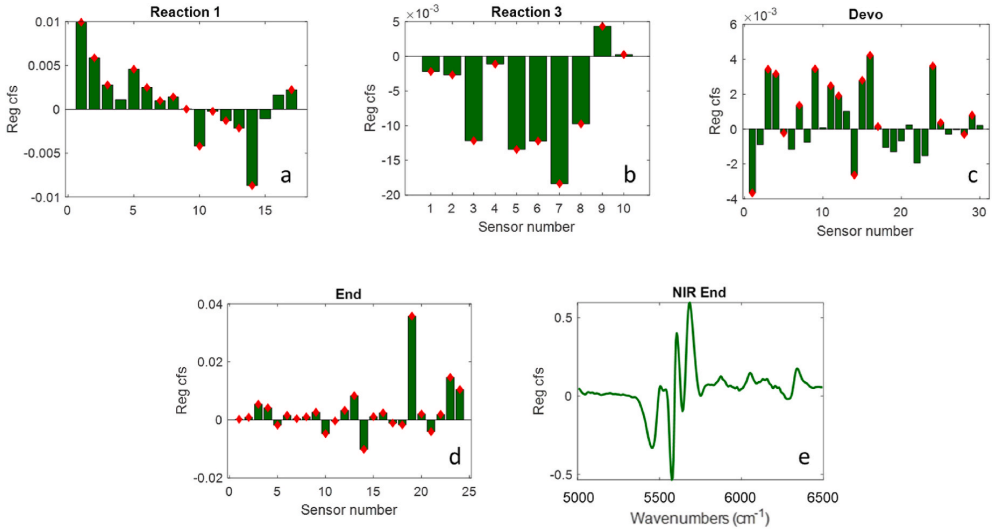


Fig. 4. Regression coefficients for REACTION 1 (a), REACTION 3 (b), DEVO (c), END (d) and NIR END (e). The red diamonds indicate variables with VIP scores exceeding one. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

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

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

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

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

Table 2
Results obtained by applying ROSA.

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

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

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

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

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

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

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

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

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

3.2. LW-MB-PLS results

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

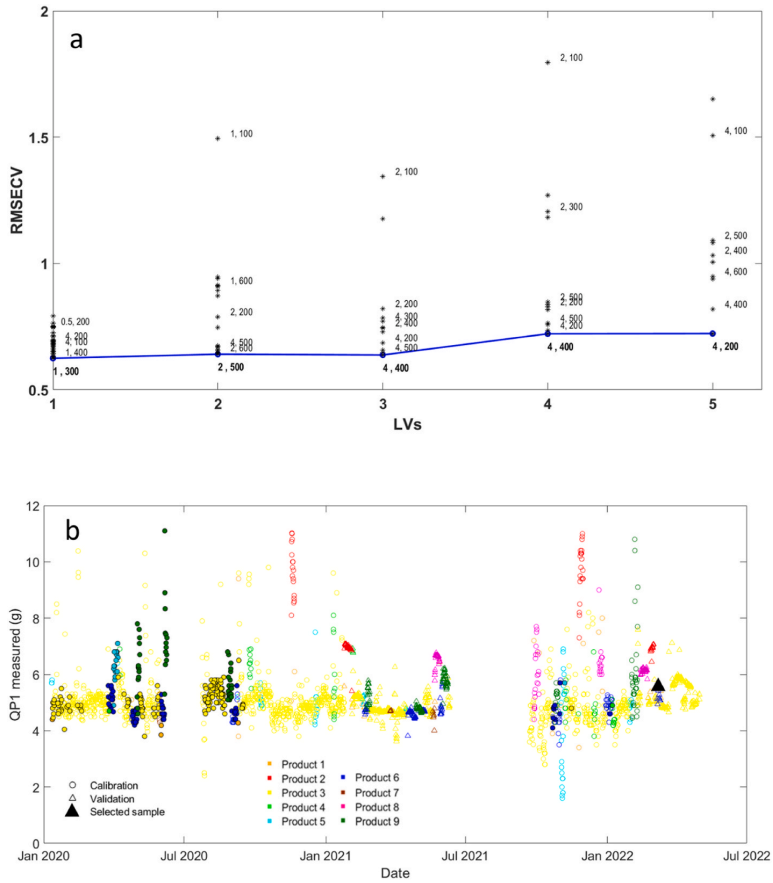


Fig. 5. (A) Måge plot for the LW-MB-PLS QP1 model. The point label report first the value of *h*, then that of *k*. The points on the Pareto front have labels in bold; (b) QP1 values vs time, colored by ABS product. Circles refer to calibration samples, whereas triangles refer to validation samples.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

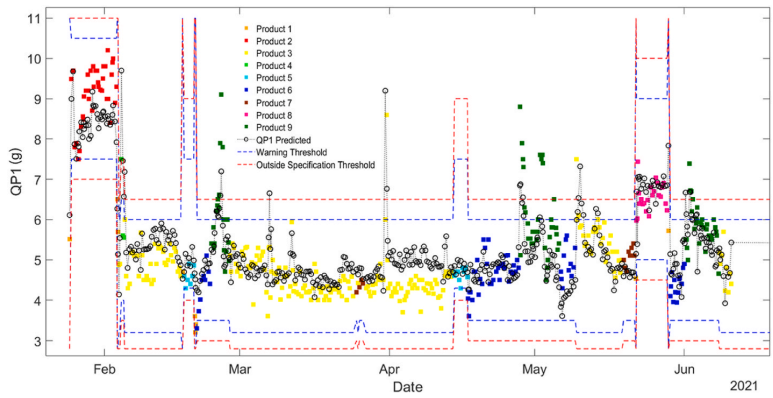


Fig. 6. Time evolution of the measured (colored filled squares) and predicted values (black non-filled circles) of QP1 for the January–June 2021 validation period. The predictions were obtained by means of the LW-MB-PLS model that employed all the available data blocks. Blue and red dashed lines represent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

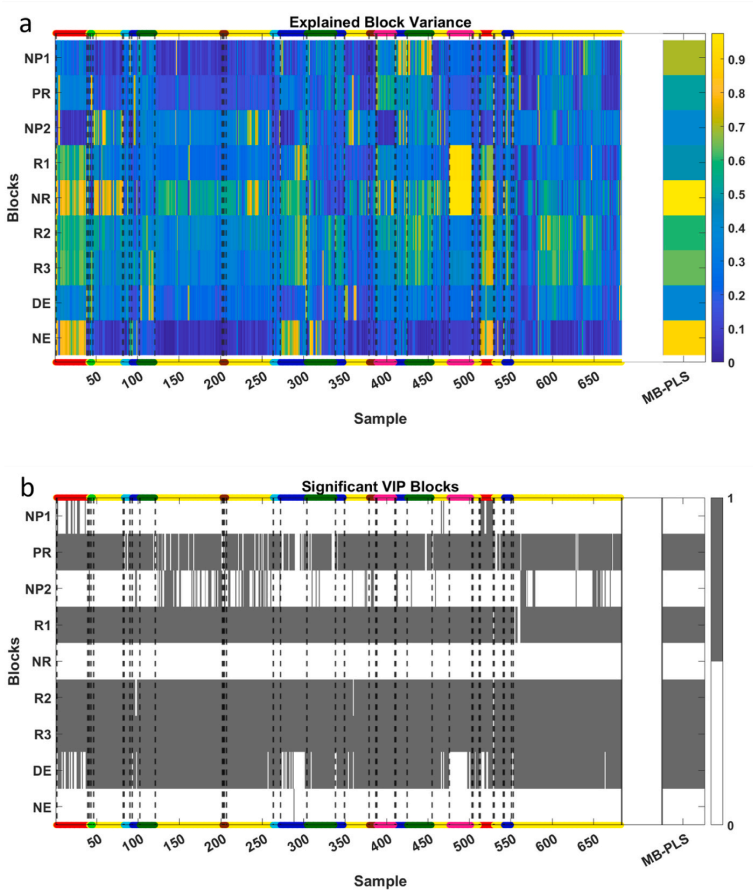


Fig. 7. The results shown refer to validation and prediction sets, i.e. covering the whole production time, for QP1. Explained variance for each block (a) and block VIPs (b) related to the LW-MB-PLS model built with all the available data blocks; values are shown in coded color according to the color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

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

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

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

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

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

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

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

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

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

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

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

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

3.3. Comparison between the multiblock methods

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

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

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

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

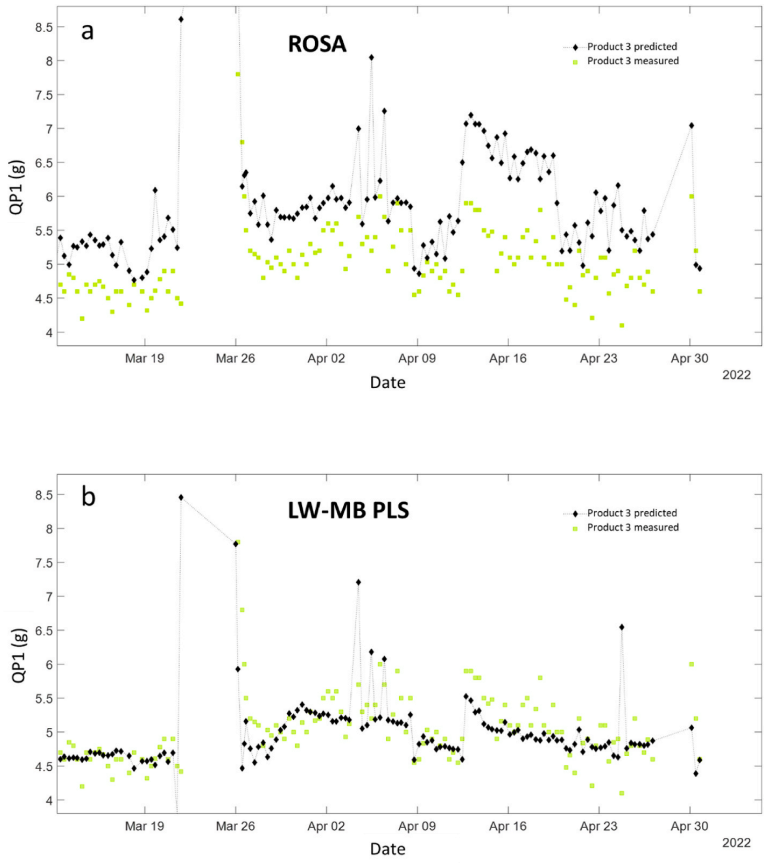


Fig. 8. Time evolution of the measured (green squares) and predicted values (black diamonds) of QP1 for the final portion of February–April 2022 validation period by ROSA (a) and LW-MB-PLS (b). The predictions were obtained by means of the models that employed all the available data blocks. (For interpretation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

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

4. Conclusions

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

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

From an applicative point of view, both methods are implementable for real time predictions. LW-MBPLS can be recommended when nonlinearity is observed, or as in the present case when different grade of products must be handled. ROSA is especially fast and can be used to

sequentially assess the relevance of each block, in addition it may bring to more robust model by filtering redundant information among blocks. In perspective, ROSA can be used in the process-understanding phase to exploit the possible scenarios and then if a prediction bias is observed it can be resorted to local modelling using only the most salient block. However, a drawback of ROSA, which require further investigation, is how to deal with blocks showing similar error in the selection phase.

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

Declaration of generative AI in scientific writing

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

CRediT authorship contribution statement

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

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgements

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

Appendix A. Supplementary data

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

References

[1] L. Strani, E. Mantovani, F. Bonacini, F. Marini, M. Cocchi, Fusing NIR and process sensors data for polymer production monitoring, Front. Chem. 9 (2021) 748723.
[2] Z. Ge, Review on data-driven modeling and monitoring for plant-wide industrial processes, Chemometr. Intell. Lab. Syst. 171 (2017) 16–25.
[3] B.S. Silva, M.J. Colbert, M. Santangelo, J.A. Bartlett, P.P. Lapointe-Garant, J. S. Simard, R. Gosselin, Monitoring microsphere coating processes using PAT tools in a bench scale fluid bed, Eur. J. Pharmaceut. Sci. 135 (2019) 12–21.
[4] S.G. Wubshet, J.P. Wold, N.K. Afseth, U. Böcker, D. Lindberg, F.N. Ihunegbo, I. Måge, Feed-forward prediction of product Qualities in enzymatic protein hydrolysis of poultry by-products: a spectroscopic approach, Food Bioprocess Technol. 11 (2018) 2032–2043.
[5] E. Strelet, Y. Peng, I. Castillo, R. Rendall, Z. Wang, M. Joswiak, B. Braun, L. Chiang, M.S. Reis, Multi-source and multimodal data fusion for improved management of a wastewater treatment plant, J. Environ. Chem. Eng. 11 (2023) 111530.
[6] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31.
[7] R. Vitale, O.E. de Noord, J.A. Westerhuis, A.K. Smilde, A. Ferrer, How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding, J. Chemom. 35 (2) (2020) e3266.
[8] M.P. Campos, M.S. Reis, Data preprocessing for multiblock modelling – a systematization with new methods, Chemometr. Intell. Lab. Syst. 199 (2020) 103959.
[9] D. Tanzilli, A. D'Alessandro, S. Tamelli, C. Durante, M. Cocchi, L. Strani, A feasibility study towards the on-line quality assessment of pesto sauce production by NIR and chemometrics, Foods 12 (8) (2023) 1679.
[10] S. Grassi, A. Giraudo, C. Novara, N. Cavallini, F. Geobaldo, E. Casiraghi, F. Savorani, Monitoring chemical changes of coffee beans during roasting using real-time NIR spectroscopy and chemometrics, Food Anal. Methods 16 (5) (2023) 947–960.
[11] G. Gorla, A. Ferrer, B. Giussani, Process understanding and monitoring: a glimpse into data strategies for miniaturized NIR spectrometers, Anal. Chim. Acta 1281 (2023) 341902.
[12] S. Grassi, L. Strani, C. Alamprese, N. Pricca, E. Casiraghi, G. Cabassi, A FT-NIR process analytical technology approach for milk renneting control, Foods 11 (1) (2021) 33.
[13] C.V. Möltgen, T. Puchert, J.C. Menezes, J.C.D. Lochmann, G. Reich, A novel in-line NIR spectroscopy application for the monitoring of tablet film coating in an industrial scale process, Talanta 92 (2012) 26–37.
[14] M. Verstraeten, D. Van Hauwermeiren, M. Hellings, E. Hermans, J. Geens, C. Vervaet, I. Nopens, T. De Beer, Model-based NIR spectroscopy implementation for in-line assay monitoring during a pharmaceutical suspension manufacturing process, Int. J. Pharm. 546 (1–2) (2018) 247–254.
[15] A.Q. Vo, H. He, J. Zhang, S. Martin, R. Chen, M.A. Repka, Application of FT-NIR analysis for in-line and real-time monitoring of pharmaceutical hot melt extrusion: a technical note, AAPS PharmSciTech 19 (2018) 3425–3429.
[16] N.L. Velez, J.K. Drennen, C.A. Anderson, Challenges, opportunities and recent advances in near infrared spectroscopy applications for monitoring blend uniformity in the continuous manufacturing of solid oral dosage forms, Int. J. Pharm. 615 (2022) 121462.
[17] R.R. de Oliveira, R.H. Pedroza, A.O. Sousa, K.M. Lima, A. de Juan, Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy, Anal. Chim. Acta 985 (2017) 41–53.
[18] K. He, M. Zhong, Z. Li, J. Liu, Near-infrared spectroscopy for the concurrent quality prediction and status monitoring of gasoline blending, Control Eng. Pract. 101 (2020) 104478.
[19] L. Strani, F. Bonacini, A. Ferrando, A. Perolo, D. Tanzilli, R. Vitale, M. Cocchi, Real time quality assessment of general purpose polystyrene (GPPS) by means of multiblock-PLS applied on on-line sensors data, Chem. Eng. Trans. 100 (2023) 175–180.
[20] A. Diez-Olivan, J. Del Ser, D. Galar, B. Sierra, Data fusion and machine learning for industrial prognosis: trends and perspectives towards Industry 4.0, Inf. Fusion 50 (2019) 92–111.
[21] A.K. Smilde, I. Måge, T. Naes, T. Hankemeier, M.A. Lips, H.A. Kiers, E. Acars, R. Bro, Common and distinct components in data fusion, J. Chemom. 31 (7) (2017) e2900.
[22] P. Mishra, J.M. Roger, D. Jouan-Rimbaud-Bouveresse, A. Biancolillo, F. Marini, A. Nordon, D.N. Rutledge, Recent trends in multi-block data analysis in chemometrics for multi-source data integration, TrAC, Trends Anal. Chem. 137 (2021) 116206.
[23] A. Biancolillo, T. Naes, The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions, in: M. Cocchi, Data Handling in Science and Technology, Elsevier, Amsterdam, pp. 157-177.
[24] M. Alinaghi, H.C. Bertram, A. Brunse, A.K. Smilde, J.A. Westerhuis, Common and distinct variation in data fusion of designed experimental data, Metabolomics 16 (2019) 2.
[25] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemom. 33 (2019) e3085.
[26] Y. Song, J.A. Westerhuis, A.K. Smilde, Separating common (global and local) and distinct variation in multiple mixed types data sets, J. Chemom. 34 (2020) e3197.
[27] J.A. Westerhuis, P.M. Coenegracht, Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares, J. Chemom. 11 (1997) 379–392.
[28] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemom. 12 (1998) 301–321.
[29] A. El Ghaziri, V. Cariou, D.N. Rutledge, E.M. Qannari, Analysis of multiblock datasets using ComDim: overview and extension to the analysis of (K+ 1) datasets, J. Chemom. 30 (8) (2016) 420–429.
[30] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemom. 10 (5-6) (1996) 463–482.
[31] T. Næs, O. Tomic, B.H. Mevik, H. Martens, Path modelling by sequential PLS regression, J. Chemom. 25 (2011) 28–40.
[32] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, Food Qual 24 (1) (2012) 8–16.
[33] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemom. 25 (8) (2011) 441–455.

[34] K.H. Liland, T. Næs, U.G. Indahl, ROSA—a fast extension of partial least squares regression for multiblock data analysis, J. Chemom. 30 (2016) 651–662.
[35] L. Strani, R. Vitale, D. Tanzilli, F. Bonacini, A. Perolo, E. Mantovani, A. Ferrando, M. Cocchi, A multiblock approach to fuse process and near-infrared sensors for on-line prediction of polymer properties, Sensors 22 (4) (2022) 1436.
[36] M. Lesnoff, M. Metz, J.M. Roger, Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data, J. Chemom. 34 (5) (2020) e3209.
[37] S. Kim, M. Kano, H. Nakagawa, S. Hasebe, Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection, Int. J. Pharm. 421 (2) (2011) 269–274.
[38] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Non-linear regression methods in NIRS quantitative analysis, Talanta 72 (1) (2007) 28–42.
[39] K. Hazama, M. Kano, M. Covariance-based locally weighted partial least squares for high-performance adaptive modeling, Chemometr. Intell. Lab. Syst. 146 (2015) 55–62.
[40] E. Menichelli, T. Almøy, O. Tomic, N.V. Olsen, T. Naes, SO-PLS as an exploratory tool for path modelling, Food Qual. Prefer. 36 (2014) 122–134.
[41] U.G. Indahl, T. Naes, Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling, J. Chemom. 12 (4) (1998) 261–278.
[42] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design. Theory, Methods and Applications, ESCOM Science Publishers, Leiden, 1993, pp. 523–550.
[43] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemometr. Intell. Lab. Syst. 129 (2013) 76–86.

Supplementary Materials

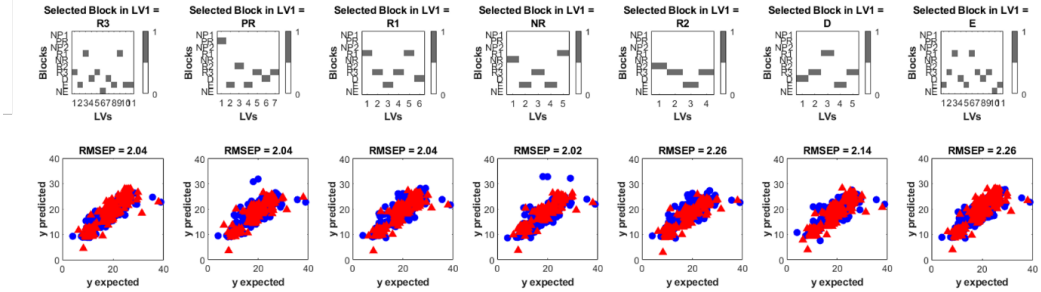


Figure S1. Prediction results for property QP2 obtained by ROSA by changing the selected block for the first component. Top: each subplot shows the selected blocks (in black) for each component. The first from left, reports the ROSA model obtained by taking as winner in the first component the block with the lowest error, i.e. R3. In the following, the first selected block was forced to be the one indicated in the title, e.g. in the second (from left) block PR instead of R3. The blocks selected in the following components were the winning ones according to the standard ROSA algorithm. Bottom: actual vs. predicted values for the calibration (blue circles) and test (red triangles) sets along with their corresponding RMSEP.

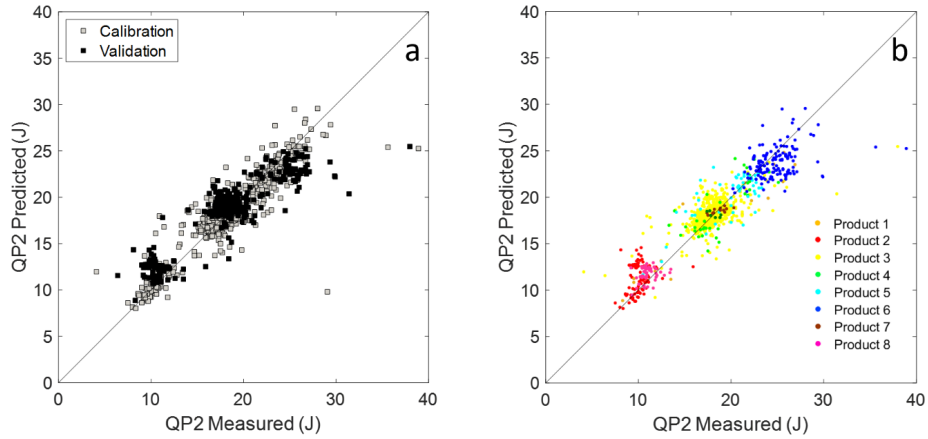


Figure S2. Plots of predicted vs measured values of QP2 obtained by the LW-MB-PLS model using all blocks. In (a) Samples are colored according to calibration (gray) and validation (black) and in (b) according to ABS product type.

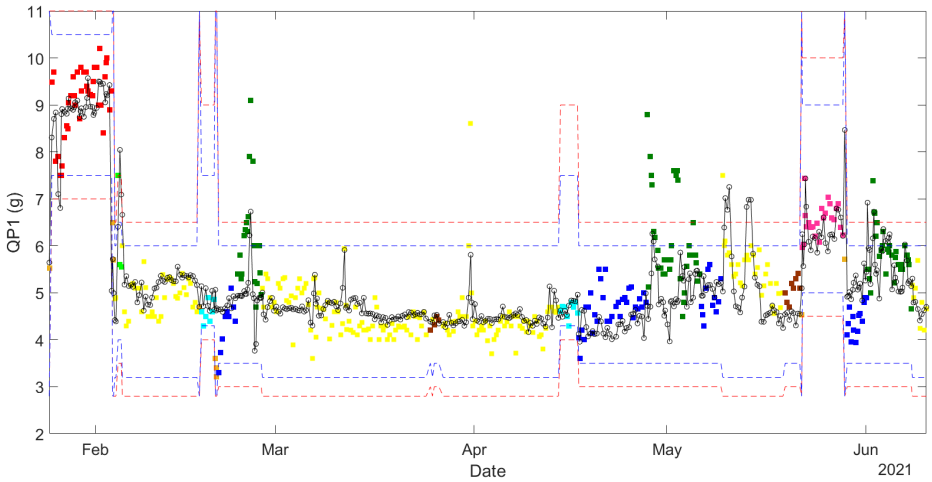


Figure S3. Time evolution of the measured (colored filled squares) and predicted values (black non-filled circles) of QP1 for the January-June 2021 validation period. The predictions were obtained by means of the ROSA model built starting from all the available data blocks. Blue and red dashed lines represent the warning thresholds and the actual low-quality threshold, respectively.

Table S1. RMSEP values obtained by ROSA, LW-MB-PLS and MB-PLS

QP1	**ROSA (g)**	**LW-MB-PLS (g)**	**MB-PLS (g)**
NP1,PR,NP2,R1,NR,**R2**,**R3**,**DE**,NE*	0.74	0.75	0.82
NP1,**PR**,NP2,**R1**,NR,**R2**,**R3**	0.81	0.91	0.99
NP1,**PR**,**NP2**,**R1**,NR,**R2**	0.83	0.97	0.97
NP1,**PR**,NP2,**R1**,NR	0.86	1.28	0.97
PR,R1,**R2**,**R3**,**DE**	0.74	0.78	0.8
PR,**R1**,**R2**,**R3**	0.81	0.77	0.87
PR,**R1**,**R2**	0.85	0.85	0.84
PR,**R1**	0.86	0.85	1.05
NP1,**NP2**,NR,**NE**	0.89	1.34	2.15
NP1,NP2,**NR**	1.19	1.67	2.64
NP1,**NP2**	1.27	1.31	1.26
QP2	**ROSA (J)**	**LW-MB-PLS (J)**	**MB-PLS (J)**
NP1,PR,NP2,**R1**,NR,R2,**R3**,**D**,**E**,**NE**	2.04	2.13	2.37
NP1,**PR**,NP2,**R1**,**NR**,R2,**R3**,**D**	2.46	3.12	3.66
NP1,**PR**,NP2,**R1**,**NR**,**R2**,**R3**	2.62	3.11	3.17
NP1,**PR**,NP2,**R1**,NR,**R2**	3.52	3.45	4.07

NP1,**PR**,NP2,**R1**,**NR**	3.93	4.35	4.64
PR,**R1**,**R2**,**R3**,**D**,**E**	2.06	2.1	2.74
PR,**R1**,**R2**,**R3**,**D**	2.12	2.3	7.14
PR,**R1**,**R2**,**R3**	2.67	2	3.92
PR,**R1**,**R2**	2.69	2	2.06
PR,**R1**	3.25	4.06	4.25
NP1,NP2,**NR**,**NE**	2.57	3.8	4.14
NP1,**NP2**,**NR**	3.28	4.36	4.51
NP1,**NP2**	3.4	4.98	4.78

In a row, cells filled by the same color indicate that RMSEP values are not statistically different between each other ($p > 0.05$). Green = lowest RMSEP, red =highest RMSEP, orange = middle value between the other two. Block names in bold indicate which blocks have been selected by ROSA.

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