UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

School of Graduate Studies
Multiscale Modelling, Computational Simulations and Characterization in Material and Life Sciences

Multivariate process monitoring in polymers production

PhD candidate:

Dr. Erik Mantovani

Tutor:

Prof. Marina Cocchi

Co-Tutor:

Dr. Francesco Bonacini

School director: Prof. Ledi Menabue

XXVIII cycle

Once we accept our limits,

We go beyond them

Albert Einstein

Preface

This PhD thesis describes my activity along the last three years and shows some of the advantages in the use of multivariate approach in petrochemical production. A lot of work has been done and many people played a role and have made successful this project. Every one of them deserves a "Thank you" for his/her specific help and supports during such a special opportunity I had.

I did not have a regular career and student path; so, I would describe briefly the way in which data analysis became part of my job and a little bit of my life. When I began to work in the petrochemical site of Mantua, November 2004, the concept of PhD was incredible far away from me, only my dearest friends know how far it was, and I am not lying but probably I did not know what PhD was. Just after the college, Versalis hired me as "research plant operator"; in a couple of days I started to move valves, to manage medium scale reactors, to learn the smell of each dangerous substance, etc., in short to understand what chemistry is in industrial and research environments: a quite different thing from the idea that books and teachers gave me. The next year, when I was 20, my student career started again but I did not leave my job, maybe because it was not too bad and some money made me independently happy. University took six and a half years of my life, environmental engineering bachelor and master degrees, but I was pleased for the achieved result. In the same periods, I got new position in spectroscopy laboratory, firstly as technician and then as researcher; it was there where chemometrics and multivariate analysis became less strange words to my ears. Month after month, my colleagues Christian Sappino and Francesco Bonacini, and my chief Angelo Ferrando, taught me how to make a good NIR calibration and how to extrapolate, probably not in the best way, information with principal component analysis. Their help and enthusiasm, that they have not yet lost and that goes further the chemometrics, was fundamental for my results and it still is. I am grateful for all and I know that the first "Thank you" shall be for them.

In 2013 I began the PhD in which were involved production and process of Versalis, company for which I am still working continuously from 2004. For this possibility, I must say "Thank you", probably a big one, to Nicola Fiorotto, the R&D chief of Mantua, which strongly believes in me and in data analysis and made enormous efforts to give me this possibility (the first time in Versalis).

Many departments were involved in this project and I can honestly say that every one of them cooperated for the success of application. For a plenty of reasons, I could find diffidence and distrust but every one of them has been of great help. Thus, another “thank you” is for them. Despite the end of doctorate, the good results convinced managements on multivariate data analysis; currently, my colleagues and I are still working in data analysis on new tasks. Therefore, the effort that the Versalis employees did in multivariate applications returned great results. Thank you again. Of course, I cannot even imagine such results without proper theoretical support. I met extraordinary people in the last three years that shared with me many good ways to address issues with multivariate techniques and even more important a plenty of priceless thing that have seriously modified my way of being. First, I’m deeply thankful to the Chemometrics Research Group of the University of Modena and Reggio Emilia: Dr. Lucia Bertacchini, Dr. Caterina Durante, Dr. Mario Li Vigni and Dr. Michele Silvestri and in particular to my tutor, Dr. Marina Cocchi, that followed the tasks evolution, encouraged me in front of difficulties and believed from the beginning in this project. I am grateful to every one of you for the small or big help you gave to me, “Thank you”.

The PhD course was improved by a traineeship at the University of Valencia in which I met the professor Alberto Ferrer and his PhD students Raffaele Vitale, Abel Folch Fortuny and Daniel Gonzalo Palací López. I immediately felt home with them and I really appreciate the way in which they collaborated with me in order to achieve the best results and to make my stay in Spain a marvellous experience. “Thank you” for your time and your passion in our job that is, again, priceless.

My father, my friends, my whole family and, I did not forget you, my girlfriend, sometimes met a tired, nervous and angry version of me but fortunately, they are all still here. I was not so present as before, I know, and for every one of you is necessary a special and lovely “Thank you”. I am far away to mention all people I met and that shared with me their ideas that gave me some suggestions or, why not, only spent good spare time; I will do personally, as soon as I can, during a conference or maybe in a dirty pub. Strange coincidences have allowed this opportunity and I am sure that it will become one of the best things I did in my life.

Thank you

Abstract

The impressive technological innovation taking place in the last decades among others consequences has brought to the generation of a huge amount of data. Nowadays, hundreds of sensors constantly monitor production, belonging to many different kinds, and this condition is common to various fields: from the agricultural to the refining, from food to the waste treatment industry. The chemical and petrochemical industries, that concern my PhD, have most to gain from this innovations and this explain why companies invest large sums of capital in process control and data management. At the same time, the classic statistics tools employed in process control were not sufficient/efficient to meet the emerging challenges of managing a growing number of diverse sensors, integrating their data and convert it to information. Thus, new tools are being developed, most of which belong to the multivariate data analysis (MVDA) techniques. These include decomposition techniques that allow data compression and shift the analysis focus from the time trends of individual variables to the correlation pattern of the whole data, in other words focusing on and revealing data structure; this is the main advantage of multivariate approach. Another MVDA feature, one of fundamental importance, is to generate confidence limits related to the whole system variance and not linked on a single variable. Multivariate techniques take into account the fundamental correlation that exists between the variables. Notwithstanding the enormous improvement of multivariate algorithms with respect to calculation efficiency, ease of use and portability, the multivariate data analysis is not widespread applied in the industry in despite of hardware and software evolution that allow calculation in a reasonable time and most often in real time. This could be mainly due to a cultural gap and this thesis tried to afford it, at least in the specific working context where it has been developed. The abundance of sensors and the chemical process features return data matrix with strongly correlated variables, a significant noise amount and a considerable percentage of missing data. Moreover, petrochemical lines suffer a lot of different effects: flow rate variation, items aging and fouling factor; this characteristics make multivariate data analysis even more adapt to process monitoring.

The Versalis company which is a partner in this thesis project, has historical database that made possible to me tackling some of the main problematic related to polymer production. PhD project take into account the three following issues:

- Batch production monitoring
- Quality parameter monitoring in a continuous production
- Trouble shooting analysis

These tasks differ for data structure and in relation to the applied multivariate methods. Each issue has its own peculiarities.

Batch production consists in the repetition in time of operations that lead from the raw material to the final product: raw materials charge, manufacturing, product discharge. It is a discontinuous process, which generates a three dimensional database that needs suitable techniques able to preserve and extract the information in such a multidimensional dataset.

Continuous process uses laboratory analysis to control the quality of production, despite the frequency is not suitable for real time monitoring. Plant controllers set conditions basing on their knowledge and experience in order to keep plant variability inside the acceptance limits (fixed as optimal for product quality), changing plant parameters when according to laboratory analysis deviance from normal operating condition is observed. In such a context would be extremely advantageous the application of sensors/models able to return real time quality parameters. In fact, product quality depends on plant set up and the process sensors (flow meters, thermocouples, pressure gauges) correlation structure impacts on the material features. These correlations might allow the extrapolation of production quality via process variables, eventually with dedicated sensors that could monitor intermediate production (i.e. spectroscopic data), in this way a virtuous cycle of monitoring, evaluation and re-setting of operative conditions can be set.

In petrochemical field, the economic damage caused by a plant shutdown or long term bad quality might cover the profits margin. The identification of problem causes assumes high importance to ensure normal operating processes and furthermore a complete understanding allows finding a permanent solution, i.e. optimal settings and monitoring tools. This multivariate application shows, even to the most sceptical, that the evaluation of whole process behaviour highlights phenomena that could not be observable by looking at single variables. Either the global plant examination or block analysis allows finding global variations, which depend to the single effect that have correlation among themselves. My PhD study, along the three past years, proofs the importance of multivariate data analysis and multivariate process monitoring in petrochemical environment and in the other context in which data suffers the same problems: correlation, noise and missing data.

Sommario

L'incredibile evoluzione tecnologia che abbiamo vissuto negli ultimi 20-30 anni, ha avuto come una delle sue conseguenze la generazione di un enorme flusso di dati. Centinaia di sensori dei più svariati tipi monitorano costantemente la produzione e questa condizione è comune ai campi più disparati: dal settore agricolo alla raffinazione, dall'industria alimentare alla gestione dei rifiuti. Il settore chimico e quello petrolchimico, su quest'ultimo verte il mio dottorato, possono trarre molti vantaggi da questa innovazione e questo spiega perché le aziende hanno investito grandi capitali nelle strumentazioni di controllo processo e nella gestione dei dati. Al contempo gli strumenti di statistica classica, normalmente impiegati nel controllo di processo, non erano sufficienti e adatti alla gestione di un crescente numero di sensori, di integrare tra loro i differenti dati e di convertire tutto ciò in informazioni utili. Perciò sono stati sviluppati nuovi strumenti, molti dei quali sono compresi nelle tecniche di analisi multivariata dei dati (MVDA). Queste includono le tecniche di decomposizione che permettono la compressione dei dati e spostano l'attenzione dall'andamento nel tempo della singola variabile alla struttura di correlazione dei dati, in altre parole mostrano la struttura dei dati nel complesso; questo è il principale vantaggio dell'approccio multivariato. Un'altra caratteristica relativa alla MVDA, anch'essa di fondamentale importanza, è quella di generare limiti di confidenza basati sulla variazione del sistema e non sul valore di ogni sensore d'impianto. Le tecniche multivariate tengono conto della fondamentale correlazione che esiste tra le variabili. Nonostante l'enorme miglioramento degli algoritmi per quanto concerne l'efficienza di calcolo, la facilità d'uso e l'applicabilità, l'analisi multivariata non è largamente applicata in industria, malgrado l'evoluzione di hardware e software permetta calcoli in tempi ragionevoli, spesso in tempo reale. Questo può essere dovuto ad un lacuna culturale che questa tesi cerca di affrontare, quantomeno in relazione al contesto dove è stata applicata. La natura dei processi nel settore petrolchimico e l'abbondante numero di sensori restituiscono dati fortemente correlati, caratterizzati da rumore non trascurabile e con considerevoli percentuali di missing data. In questo modo l'informazione disponibile è spesso celata e non viene individuata con tecniche classiche basate su un approccio univariato. Inoltre l'industria petrolchimica risente di problematiche legate alle variazioni dei prodotti, dei carichi di impianto, dell'invecchiamento e dello sporcamento delle linee produttive, aspetti che rendono ancora di più le tecniche multivariate adatte al monitoraggio del processo stesso. I database storici sono ormai presenti in molte realtà industriali e mi hanno permesso di esplorare alcune delle principali problematiche che si possono trovare.

Il progetto di dottorato, svolto in collaborazione con Versalis-eni attingendo ai loro database, prende in considerazione i seguenti tre punti:

- Monitoraggio di una produzione batch
- Monitoraggio della qualità del prodotto in una produzione in continuo
- Individuazione delle cause di un problema

Queste problematiche hanno caratteristiche fortemente diverse sia in termini di strutture dati che in relazione alle metodologie di analisi applicate. La produzione batch ha come peculiarità la totale ripetizione delle operazioni che portano dalle materie prime al prodotto finito per ciascun lotto di materiale in ingresso: carico, lavorazione e scarico. Si tratta di una produzione discontinua che genera una base dati molto differente da quella di un processo in continuo, in particolare ha tre dimensioni: lotto, tempo, parametri di processo. Quindi deve essere trattata con tecniche idonee, atte alla conservazione ed estrazione dell'informazione.

I processi continui si avvalgono delle analisi di laboratorio per controllare la qualità della produzione, anche se la loro frequenza non è sufficiente per un monitoraggio real time; i responsabili della gestione d'impianto sulla base della loro conoscenza ed esperienza fissano i valori dei parametri di processo cercando di mantenere controllata la variabilità entro limiti accettabili (fissati come ottimali per la qualità del prodotto finito) correggendoli anche in funzione dei risultati delle analisi di laboratorio. In questo contesto, sarebbe fondamentale avere sensori/modelli in grado di dare responsi in real-time sulla qualità della produzione. In effetti, la qualità del prodotto dipende dalle condizioni operative dell'impianto e la struttura di correlazione tra i sensori di processo, P, T, flussi, etc., si riflette sulle caratteristiche dei materiali prodotti. Queste relazioni possono consentire di estrapolare la qualità del prodotto dalle variabili (sensori) di processo (eventualmente integrandole con sensori on-line che monitorino la qualità degli intermedi di produzione, es. dati spettroscopici) stabilendo un ciclo virtuoso di monitoraggio, valutazione e re-setting delle condizioni operative di processo.

Nell'industria petrolchimica i danni economici causati da un fermo impianto o da un lungo periodo di produzione fuori norma sono tali da coprire i margini di guadagno. L'individuazione delle cause di questi eventi anomali è quindi di fondamentale importanza per cautelare il processo e per permettere alla gestione la risoluzione definitiva del problema. Questo tipo di indagine multivariata mostra anche ai più scettici che valutare il comportamento globale del processo restituisca informazioni che la singola variabile non è in grado di individuare. Analizzare il processo produttivo, o blocchi di quest'ultimo, per intero permette di cogliere le variazioni globali che a loro volta sono funzione degli effetti singoli, tra loro dipendenti.

Il dottorato di ricerca da me sostenuto in questi tre anni vuole mostrare come l'analisi multivariata dei dati possa essere uno strumento di rilevante importanza nel contesto petrolchimico in particolare, e in tutti i contesti nei quali le strutture dati hanno le medesime caratteristiche: alta correlazione, elevato rumore e presenza di dati mancati

Table of contents

I

Introduction

Content

1.1 Aims and outline of the thesis

Nowadays, polymers are very common materials used in our daily life but actually, the term polymer is quit young. It has been used the first time by Berthelot [1] in 1866 who noted that "*styrolene (styrene), heated at 200° during a few hours, transforms itself into a resinous polymer*": he was in front of the first synthetic polymer. Natural polymers were already known, e.g. in the form of textile fibres, and at the end of 19th century researchers started to study how to obtain them by synthesis. The interest in polymers, at the beginning, came from a growing demand of materials with their characteristic, rapidly natural compounds started to be insufficient and synthetic activity grown. Synthetic polymers were used as substitutes of many traditional materials even if materials of poor quality were obtained. Example of widely used synthetic polymers replacing the natural one are nylon, polyacrylate and a lot of newer tissue developed in the textile industry. The polymers industry expanded fast and stimulated, upstream, academic research to improve quality. Actually, the global interest on polymers is well established and the state of art research allowed such an improvement of polymers quality that has changed the thinking that polymers are poor quality alternative to natural materials. Now it could be stated that there maybe only unappropriated applications of polymers while the quality issue is an overcome problem. Petrochemical field [2] is the core of polymer production. It is in the middle of polymers production chain: before it, except for the renewable resources, there is the refinery where oil is extracted and the main compounds separated, after there are plenty of transformation processes that realize the final items.

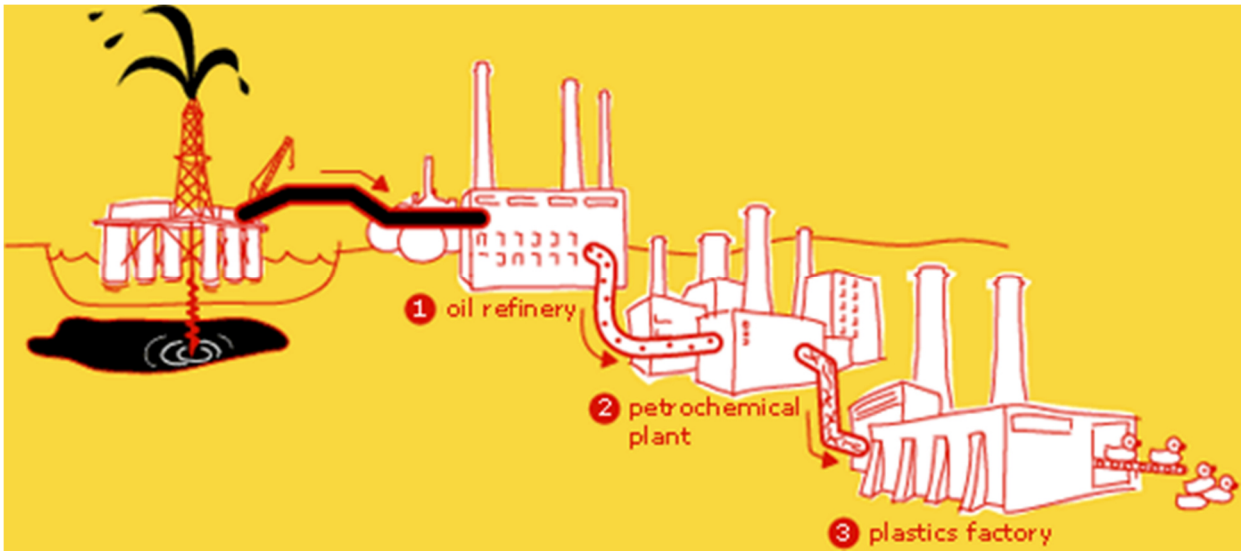

Figure 1.1 From oil to final items, "http://www.nobelprize.org/educational/chemistry/plastics/readmore.html"

Thesis focus on petrochemical process and involves both monomer and polymer production. Despite they belong to the same field, monomers and polymers have a completely different production process and features [3]. Plant managers and operators are very skilled in conducting production notwithstanding the complexity of the production steps; however, many problematic still affect petrochemical production. Major issues are: variability of raw materials, undesired variations in plant settings and dirtiness, that among many, create lacks in product quality and make the companies to decrease the gain. The applications presented in the Thesis show how a multivariate approach can tackle these issues. Indeed, several variations sources are inherent to production from the beginning to the end of manufacturing, from the inlet materials to the final products and the focus of the adopted multivariate approach has been to assess main variability sources and explain their origin (possible causes). The studied plants are located at the petrochemical site of Mantova, one of the production sites of Versalis S.p.A., which is briefly presented into the next chapter together with a concise company description. Thesis aims at building process monitoring/control tools starting from the study of Versalis historical databases, process and laboratory, and considering as well a new near infrared (NIR) application as process analytical technology to monitor intermediate product quality. In all presented applications, process variables were considered to gain information about process state and evolution. Process variables originate from a very high number of sensors measurement taken at high time frequency. Beside these spectral signals (NIR-on line) and additive concentrations were considered to describe the behaviour of the system. Various multivariate algorithms, in chapter 2 they are presented in some details, allowed data decomposition, extraction of information from such databases and building multivariate monitoring charts. Chemometric methods have been selected according to challenge, considering the availability of data and the peculiarity of the studied processes. The Thesis, after Introduction and a Methods section, is organized in three main chapters, each one concerning a different process representative of main issues in multivariate process monitoring:

Batch production monitoring

- Quality parameter monitoring in a continuous production
- Trouble shooting analysis

These tasks differ for data structure and each issue has its own peculiarities.

Batch production, chapter 3, consists in the repetition in time of operations that lead from the raw material to the final product: raw materials charge, manufacturing, product discharge. It is a discontinuous process, which generates a three dimensional database that needs suitable techniques [4 5] able to preserve and extract the information from such a multidimensional dataset.

The considered batch process was studied by analysis of historical data exploiting the several issues that batch process complexity poses in the perspective of building up a monitoring system based on process variables.

Continuous process (chapter 4) are run continuously feeding raw materials to the reactors. In the studied one, off-line laboratory analysis of intermediates are used to control the quality of production, despite the frequency of these tests is not suitable for real time monitoring. Personnel super visioning the plant set the operative conditions on the basis of their knowledge and experience in order to keep plant variability inside the acceptance limits (fixed as optimal for product quality), when according to laboratory analysis deviance from normal operating condition is observed the parameters are changed, again on previous experience basis trying to come back to previous conditions. In such a context, would be extremely advantageous the application of sensors/models able to return quality evaluation in real time. In fact, product quality depends on plant set up and the process sensors (flow meters, thermocouples, pressure gauges) correlation structure impacts on the material features. These correlations might allow the extrapolation of production quality via process variables [6], eventually with dedicated sensors that could monitor intermediate production (i.e. spectroscopic data), in this way a virtuous cycle of monitoring, evaluation and re-setting of operative conditions can be set.

In petrochemical field, the economic damage caused by a plant shutdown or long term bad quality might cover the profits margin (chapter 5). The identification of problem causes assumes high importance to ensure normal operating processes and furthermore a complete understanding allows finding a permanent solution, i.e. optimal settings and monitoring tools. The application presented in this chapter is a stringent example of how multivariate modelling can aid trouble shooting and showed, even to the most sceptical personnel at the plant, that the evaluation of whole process behaviour highlights phenomena that could not be observable by looking at single variables [7]. Either the global plant examination or block analysis allows finding global variations, which depend to the single effect that have correlation among themselves.
Thus, it has been possible to select production lines representative of both continuous and batch processes with different problematic to highlight how much a multivariate statistical approach can aid finding solutions.

In the following Section 1.2 a brief overview of Versalis company and petrochemical context from the perspective of process data handling are presented.

1.2 Versalis Company overview

Versalis [8], Eni's chemicals company, is Italy's largest Italian chemical company by sales, production volumes and number of employees. It is based in Italy and in a number of other countries with avant-garde production plants and a wide-ranging sales network that enable efficient customer assistance, within an integrated organization able to meet all market needs. With production levels of 6 million tonnes and a turnover of € 6 billion in 2013, in particular, the company holds market stewardship in manufacturing:

- Intermediates
- Polyethylene
- Styrenics
- Elastomers

To expand presence on a global scale, particularly towards new markets, the company have developed an internationalization process focusing on making the most of all possible opportunities in order to create synergies while maintaining the commitment towards quality and sustainable development for the environment and the communities in the surroundings of Versalis plants. Company have a wide range of proprietary technologies, an R&D keeping pace with the industry evolution, a comprehensive product portfolio, a wide-reaching distribution network, customized solutions and after-sales assistance. Therefore, Versalis has been recently interfacing with markets through a market-oriented focus on a more diversified and added value product portfolio; by optimizing some less competitive production sites in order to ensure greater integration and flexibility; by emphasizing research and patents achievements and further expanding its technological and commercial footprint globally. Versalis has also entered the bio-based chemicals and polymers industry partnering with leading global biotech companies. Through Matrìca S.p.A., a fifty per cent joint venture with Novamont S.p.A., the company has turned an existing petrochemical site into a leading production and R&D centre for green chemistry at Porto Torres, Sardinia. It has also collaborated with Genomatica to produce butadiene from renewable feedstocks; with Yulex Corporation, for production of guayule-based natural rubber and with Pirelli & C. S.p.A. for a joint research project aimed at utilizing guayule-based natural rubber in tire production. Amongst other green undertakings, it is worth noticing the R&D project on "yeast oils" from biomass to develop new bio-based products being conducted at the Versalis green chemistry centre at Novara. Other important projects based on technological innovation using renewable sources are included within the scope of significant transformation projects regarding a few production sites.

These include the agreement with Elevance Renewable Sciences for production from vegetable oils at Porto Marghera, and the one entered into with Neville Venture for the production of hydrocarbon resins at the Priolo site. On the international scene, the partnership with the Malaysian concern Petronas is particularly significant, as is the joint venture Lotte Versalis Elastomers, the latter established with the Korean Lotte Chemical.

1.2.1 Mantua petrochemical site

The plants involved in MVDA are in the petrochemical site of Mantua [9]. Quite near to the city centre, the petrochemical site occupies a strategic position with a lot of transportation infrastructure: railways, highways and a river port. It measures 130ha and every year moves about 2 millions of raw material and final product. For someone not familiar with petrochemical industry it consists in a self-sustaining area. So, inside Mantua site are the water and waste treatment plant, water purification plant, R&D department, control laboratories, fire station, medical station, HSE department, administrative offices and a storage area with capacity of 170.000 m^3.

Three different productions coexist in Versalis site:

- Styrene production

- Polymers production

- Phenol and by-products production

The styrene production uses ethylene and benzene as raw materials, obtains ethyl benzene, which is an intermediate material and via dehydrogenation produces styrene monomers. Two lines produce SM (Styrene Monomer): the so-called ST20 and the ST40 with a daily production of 1500 tons/day.

The polymers production realizes polymerization of styrene and the copolymerization with acrylonitrile or/and rubber. This materials family feeds various sectors such as automotive, packaging, toys, household products and so on. Eight plants in Mantua make polystyrene omo-polymers and co-polymers (PS) with a nominal capacity of 850 tons/day.

The last production, phenol and derivate, uses as raw materials hydrogen, from styrene monomers cycle, and cumene. The main product is phenol but a plenty of by-products covers a wide range market. Phenol, acetone, acetophenone, cyclohexanol and other substances belong to the production of nylon, pharma, resins and other applications. Two lines have a daily global production of 7500 tons. Previous description is non-exhaustive and aims to present which productions interest the site of Mantua. A detailed and dedicated description is provided in all chapter applications.

1.3 The multivariate statistical process control concept

A high number of sensors are usually installed on production plants, for instance thermocouples, pressure gauges and flow meters and they generate a huge amount of data. These instrumentations were installed with the aim of controlling production and support monitoring of normal operating condition: the aim is to reduce lacks of production given by accidental malfunctioning or sub-optimal conditions that may be encountered due to the complexity of process. However, database offers many data and users would like to use this historical information to monitor process, to control the production and even more to predict quality in real time extrapolating in some way useful information from the past operations. Furthermore, in all industrial processes it is requested to deeply taking into account energy saving, optimization of production planning, efficient utilization of raw materials and measurements taken at the plant by the sensors can be used for these tasks. The production control in Petrochemical industry, as in many others, is mainly based on the expert personnel knowledge, vitally important, supported by single parameter control chart developed for few selected sensors/monitoring points and laboratory material controls. In fact, the statistical process control (SPC) assumed this meaning in production context: monitoring of production quality via some key process variables in a univariate way. SPC concept and methods are fundamental in process industry and even more in the case of petrochemical production [10]. SPC monitors performance of repetitive operations, such as polymers production for instance, in order to check if they are in a state of control or not. Control state is defined in terms of similarity to the desired values for the specific set of monitored parameters, hence if the distance from measured and reference values is inside the confidence limits or not. However, process is always subjected to variability [11]; depending on the specificity of each process, variability is characteristic of the given plant and related, for instance, to the items used for production or to the external conditions. Since a certain degree of variability cannot be avoided it is important to define the normal operation condition (NOC) under which the process can be considered stable around its own "natural" variability, in other word internal to the confidence limits of the monitored process parameters (process variables). Operators perfectly know such variation and monitor it, as far as NOC hold no action is required on the plant. Traditionally SPC focus on the monitoring of few key process variables or product quality variables in a univariate way, as plant controller actually does. However, more than a variable represent the whole process and it results in a large number of control charts that plant operator shall attend to. Occasionally, different sources of variability occur and process goes out of control condition and when an anomaly occurs, several parameters change simultaneously due to the correlation between variables. An abnormal event probably affects more than a sensor and if such a situation occurs, operator cannot easily detect the source of the problem.

Variation along production might be related to impurities, plant aging, sensors shut down, leakages and to many other possible causes, and no one single measurable variable directly can reveal a phenomenon that affects several sensors because such effects are not directly measurable.

In spite of focusing on single variables changes, multivariate statistical process control (MSPC) focus attention on the whole set of process variables and their correlation structure, in this way it allows anomalies detection, changes due to raw materials, to plant set up, etc. in other words, all possible events that modifies the conditions of the process. Operators should be interested in this variation and in how much a single measured variable contributes to this variation.

The (MSPC) [12] allows a completely different way to monitor production: a unit or even a plant can be monitored looking at few multivariate charts. MSPC is based on the concept of variable correlation: in a complex production process whit a lot of measured variables, likely many of them are interrelated and does not behave independently. In univariate control chart a single variable profile appears and thus how it may correlate to others is just not considered, in any case is not visualized. Multivariate methods, described in further details on chapter2, allow compressing the huge number of sensors monitoring a production plant in a set of few parameters (latent variables) [13] describing the sources of variation in the process while keeping track of the correlations among sensors; in this way few trajectories instead of hundreds or thousands shows the movement of production and confidence limits describe how much such condition is away from the desired or normal operating situation.

In order to increase the economic competitiveness, especially considering the low margins for cost reduction, the only way is to increase production quality. Most of resources and efforts are spent in monitoring quality, particularly of finished products, to a less extent of intermediates, however discovering departure from quality at this stage has two main drawback: it is often too late to prevent loss of production and it is difficult to identify the causes of low quality, especially at which stage of production they took place and which parameters are involved.

Industrial process should remember that product quality comes from raw material quality, a perfect knowledge of process and a proper control of production conditions. In each one of these points, a multivariate approach improves results; historical database contains information that latent methods are able to extract improving the direct understanding of NOC and abnormal situations. Multivariate control charts allow controlling the whole process instead of the single variable to guarantee the stability of production and to detect changes. Nonetheless, multivariate check of raw materials might improve the process stability.

The multivariate approach in petrochemical industry, from the raw material to the product quality passing through the process design [14], could be compared to the process analytical technology (PAT) in pharmaceutical environment; PAT [15] is method to design and control manufacturing through raw material and process measurement in order to ensure the product quality. MSPC should help either process to achieve the desired production using the available measurement or to guarantee the same product quality through a systematic procedure, controlled via multivariate charts.

Historical database availability and multivariate techniques might open interesting perspective on these tasks. Every process has historical database where the values read by the monitoring instrumentation for a huge number of sensors are stored. Also the data acquired on raw materials and quality controls is stored. These huge databases usually fulfil legal requirements, e.g. traceability, and/or have to comply with internal organization. Most commonly, only a very small part of data is looked at for monitoring.

Some companies already use multivariate statistical process control (MSPC) methods and report successful stories [16 17 18 19 20]; refinery, pharmaceutical company, agriculture and food industries have opened to the multivariate technique in the lasts decades and have taken advantages from their experiences. Through these cases of studies, it can be seen that the main difference between multivariate and univariate analysis consists on the focus of analysis: multivariate data analysis pays attention to the global variation, then highlights each variable contribution and extract information on variable to variable, observation to observation and variable to observation correlation. Univariate method detects specific effect in single variables that are usually determined by a phenomenon that is hardly self-explained by one single parameter deviance. Thus, the concept of MSCP could be condensed in global variation control whereas the concept of SCP lies in single variables check.

1.4 References

1. Gnanou, Y., & Fontanille, M. (2008). *Organic and physical chemistry of polymers*. John Wiley & Sons.
2. R. Smith (2005), *Chemical Process: Design and Integration - Chapter1: The Nature of Chemical Process Design and Integration*, John Wiley & Sons
3. Robert A. Meyers (2005). *Handbook of Petrochemicals Production Processes*, from Chapter 17.3 to 17.5, 2005 McGraw-Hill Education
4. Camacho, J., Pico, J., & Ferrer, A. (2008). Bilinear modelling of batch processes. Part I: theoretical discussion. *Journal of Chemometrics*, *22*(5), 299-308.
5. García-Muñoz, S., Kourti, T., MacGregor, J. F., Mateos, A. G., & Murphy, G. (2003). Troubleshooting of an industrial batch process using multivariate methods. *Industrial & engineering chemistry research*, *42*(15), 3592-3601.
6. Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and intelligent laboratory systems*, *28*(1), 3-21.
7. Kresta, J. V., MacGregor, J. F., & Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance. *The Canadian Journal of Chemical Engineering*, *69*(1), 35-47.
8. Versalis: the new face of chemical, chemistry to evolve (2010) [https://www.eni.com/en_IT/attachments/azienda/attivita-strategie/petrolchimica/polimeri-europa/pubblicazioni/Versalis-Brochure-istituzionale-per-il-Plast12.pdf]
9. Stabilimento di Mantova dichiarazione ambientale 2012 dati 2013 [https://www.eni.com/it_IT/attachments/azienda/attivita-strategie/petrolchimica/polimeri-europa/pubblicazioni/DA-2012-dati-2013-versalis-MN.pdf]
10. Chaudhry, S. S., & Higbie, J. R. (1989). Practical implementation of Statistical Process Control in a chemicals industry. *International Journal of Quality & Reliability Management*, *6*(5).
11. A.J. Ferrer-Riquelme (2009). *1.04 - Statistical Control of Measures and Processes*, Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier
12. Kourti, T. (2006). Process analytical technology beyond real-time analyzers: the role of multivariate analysis. *Critical reviews in analytical chemistry*, *36*(3-4), 257-278.

13. T. Kourti (2009), *4.02 - Multivariate Statistical Process Control and Process Control, Using Latent Variables,* Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier
14. Pomerantsev, A. L., & Rodionova, O. Y. (2012). Process analytical technology: a critical view of the chemometricians. *Journal of Chemometrics*, *26*(6), 299-310.
15. R. Phan-Tan-Luu, R. Cela (2009). *1.09 - Experimental Design: Introduction* Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier
16. Macho, S., & Larrechi, M. S. (2002). Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. *TrAC Trends in Analytical Chemistry*, *21*(12), 799-806.
17. Bonacini, F., Ferrando, A., Montovani, E., Sappino, C., Arcidiacono, G., Ardizzone, D., & Rossi, E. (2013). Fourier transform near infrared application for advanced process control of an ethylene cracking plant. *NIR news*, *24*(6), 9-11.
18. Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, *19*(4), 213-246.
19. Ferrer, A. (2007). Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process. *Quality Engineering*, *19*(4), 311-325.
20. Skagerberg, B., MacGregor, J. F., & Kiparissides, C. (1992). Multivariate data analysis applied to low-density polyethylene reactors. *Chemometrics and intelligent laboratory systems*, *14*(1), 341-356.

II

Method

Content

2.1 Introduction

Multivariate analysis has been applied in various fields and with different aims on the last decades [1 2 3 4 5]. The wide application range generates methods that differ for the purposes and a lot of minor modifications that mainly concern to the field (i.e. medicine and food industries) and/or the data types (i.e. imagines and signals). In petrochemical industry, MVDA has been applied to the raw material control, in order to define if inlet materials are inside or outside the limits [6]; many of these applications involve near InfraRed Spectroscopy (NIR) [7]. Moreover, in product development multivariate analysis has found application with the Design of Experiment (DoE) [8 9 10] that allows a wide exploration of variables range taking into account as well the factor correlations. In molecular structure study, multivariate analysis plays a fundamental role: for instance, the development of new drug molecules with improved efficacy benefits from Quantitative Structure Activity/Propriety Relationship (QSAR/QSPR) [11] methods that apply multivariate technique. Nonetheless, in many industrial productions MVDA support plant monitoring and production via control chart [12 13 14 15 16 17]; multivariate regression allows parameters prediction and improves historical data analysis understanding. These are few among the plenty of possible situations in which data analysis and, in particular, multivariate methods improve results.

The three applications, object of this thesis, are developed by using methodology and algorithms that belong to the following:

- Exploratory data analysis
- Modeling, both regression and discrimination
- Multivariate control charts

Other fundamental issues are: data pre-treatment, which is often the clue to obtain good models, it is off course data set dependent, and models diagnostic and validation. In fact, a multivariate model gives results that should be checked; the consistency of the given information is fundamental to avoid all the possible consequences: bad prediction, wrong solution for a problem, etc. Analyst must verify model consistency and predictive capability. To this aim, internal validation and external validation have to be applied. In this section, the methodologies used in the Thesis will be briefly outlined, while the specific methods/approaches connected with batch analysis/monitoring will be introduced in Chapter 3.4.

2.2 Pre-processing methods

A good data scientist knows how much important data pre-processing methods are. Pre-processing is aimed at removing irrelevant and systematic variation that might affect multivariate analysis while it should leave untouched the information contained in dataset [18]. It may result in many advantages, such as parsimony, i.e. less principal components needed, better interpretable latent variables in a Principal Component Analysis (PCA) and better calibration with more stable predictions in a Partial Least Square (PLS) model. Data pre-treatment includes from simple methods [19] to quite complicate transformation algorithms [20]. Some of these are dedicated to signals treatment, for instance spectroscopic (NIR, NMR), and others aid managing process data. Literatures give an impressive number of different pre-processing methods [21] and the reason lies in the wide range of application that needs proper data pre-treatment to improve results.

Two basic data pre-treatment belongs to many applications and have been applied to dataset used in this thesis: mean centring and auto-scaling. Mean centring is self-explaining; it consists in shifting all the observation to the centroid of the data. It removes variables (columns) mean from every observations and does not affect samples distance in latent variables space. The mean centring removes the principal source of variation that is otherwise identified into the first latent variables, concerning PCA model. Autoscaling is a combination of mean centring and unit variance scaling. Unit variance scaling applies to the columns of data matrix a weight equals to the inverse of its standard deviation, so that the variance of each variable will be equal to one after transformation. Scaling is useful in a system where the variables have different units and/or scales in order to avoid the magnitude effect, hence variable with larger scales to dominate data analysis. Autoscaling makes all variables comparable but a side effect is the noise enhancement. For example, a level gauge in a vessel or in a storage tank keeps the set value but oscillating around it in a range which is within instrument error and or due to normal operative conditions “natural” variability, thus we do not want this level values to become as important as other sensors variables, while in the case of spectroscopic data we do not want baseline variation to be as important as signal regions corresponding to peaks/band. Different solution may be adopted, depending on the nature of data and data sets, such as using block scaling or Pareto scaling and or removing variables whose variance is of the same order of magnitude of experimental error/uncertainty.

In the applications considered in this thesis, the variables are generally sensors, such as thermocouples, rotation per minute counter, impedance measures, etc. and only one application is based on near infrared spectra. Plant historical databases collect data where variables are of very different units and scales and most often autoscaling is the pretreatment method of choice.

However a special case is represented by analysis of batch data, in this case several considerations have to be taken into account because the batch data are originally three dimensional arrays (time point x variables x batch) and both unfolding and centring/scaling issues are critical. These will be discussed in Chapter 3.4 In spectral dataset, where the different variables are the intensity values at each wavelength, auto-scaling is generally dangerous. Spectroscopy data have some peculiarities that need proper pre-treatment algorithms. Typical issues concerns background, baseline, scatter that generally introduce vertical shift (constant, proportional or non-linearly varying with wavelength), these effects can be due to several experimental conditions or sample features, e.g. path length, presence of bubbles, different particle size, colour, etc.

Scattering effect is one of the main concerns when working in near infrared spectral region: due to scattering effect the radiation reflected by the sample is partly lost. To remove scattering effects several pre-treatments have been proposed, most used are Standard Normal Variate (SNV) and Multiplicative Scattering Correction (MSC) [22]. SNV actually corresponds to row autoscaling, eq. 2.3, and allows removing constant vertical shift:

$$x_{i,j\ SNV} = \frac{x_{i,j} - \bar{x}_i}{\sqrt{\frac{\sum\left(x_{i,j} - \bar{x}_i\right)^2}{n-1}}} \tag{2.1}$$

Where $x_{i,j}$ is the *j-th* wavelength frequency of *i-th* spectrum, $\bar{x}_i$ is the average absorbance of *i-th* spectrum, n is the number of wavelength in the spectrum and the $x_{i,j\ SNV}$ is the intensity of the *j-th* wavelet of *i-th* spectrum after SNV.

The MSC [23] can remove vertical shift proportionally increasing with wavelength. Correction is accomplished by regressing each measured spectrum against a reference one, e.g. the average spectrum, estimating slope and intercept and correct for them as shown:

$$\boldsymbol{x} = \boldsymbol{r}b + a \tag{2.2}$$

$$b = (\boldsymbol{r}^T\boldsymbol{r})^{-1}\boldsymbol{r}^T\boldsymbol{x} \tag{2.3}$$

$$\boldsymbol{x}_{MSC} = \frac{(\boldsymbol{x} - \boldsymbol{a})}{b} \tag{2.4}$$

Where $\boldsymbol{r}$ is the reference spectrum and $\boldsymbol{x}$ corresponds to the original and $\boldsymbol{x}_{MSC}$ is the corrected one.

2.3 Exploratory data analysis methods

Each plant manager and every process operator know that sensors give a mix of necessary and unnecessary information and that time by time a sensor may fail on the other hand plant workers rely on that instrumentation to ensure a good and safe production. At maximum point (samples) can be displayed in three-dimensional graphs, so at maximum three variables can be inspected simultaneously in a single plot. Furthermore, a qualified plant controller can monitor maybe up to 20-30 sensors a time, a notable ability that comes from experiences and knowledge throughout the years. Anyway, in each plant hundreds or thousands of sensors are installed, much more than the plant operators may look at. Thus, multivariate exploratory data analysis [24] has the purpose to extract and plotting information taking into account all thousands of variables by using compression/decomposition techniques, so that few bi-three dimensional plots can be sufficient to look at.. Coming back to the data, there are other common situations in which multivariate exploration might help process understanding. Thousands of variables for every day monitoring correspond to a huge amount of data that are a mixture of information and noise.

$$Dataset = Information + Noise$$

Most of variance in process data is not due by systematic changes in operative conditions or process state that are induced by specific phenomena but consists of noise. In multivariate analysis model residuals, i.e. the un-modelled part of data, are usually inspected to understand if correspond to unstructured noise, or still contains information that model lefts in the noise, also called structured noise. One of the most used tools in multivariate exploratory data analysis is Principal Component Analysis [25]. This unsupervised tool allows revealing pattern in data without any a priori assumption. Moreover, PCA offers some easy way to visualize the results and plot is suitable for a complete interpretation of both, and simultaneously, variables and samples space.

2.3.1 Principal component analysis

Principal component analysis [26] could be considered as the basic tool in multivariate data analysis. It provides an approximation of an **X** matrix by a product of two small matrices **T**, scores, and **P**, loadings that together capture the essential patterns of **X**. The scores return the pattern between objects and, in the same way, loadings shows the pattern between variables [27]. PCA has more than one objective [17] and is suitable for many tasks. Analysts may be interested in similarity between object in order to classify experiments or whatever and, in the same way, they could verify class hypothesis and find out differences from their idea or nonetheless PCA is a common way to identify outliers.

In addition, data reduction assumes high importance when large amounts of data are taken into account, PCA decomposition preserves data variance, hence information, while achieving data reduction to a few dimensional space thus improving much data representation and interpretation. The simplest way to visualize how PCA works is by geometric interpretation. Data **X** has *m* objects and *n* variables that defines features of an observation along the variables, in row direction, and describes variable changes through the observations, in columns direction. Original matrix can be represented as a series of *m* points in an *n* dimensional space; the space will have many axes as the number of variables, frequently more than three despite we are not able to visualize such a multidimensional structure.

The first Principal Component (PC) corresponds to an axis of maximum variance, i.e. an axis on which the point corresponding to the projection of the *m* points (scores, $\mathbf{t_1}$) have maximum variance. The $\mathbf{t_1}$ scores are the coordinates of the *m* points when projected on the PC1 line, the $\mathbf{l_1}$ loadings correspond to the angles that PC1 (first latent variable) forms with the original *n* variables axes. The next principal component (PC2) will be as well a direction of maximal "residual" variance (this time with respect to the variance left after PC1 has been derived) with the additional constraint of being orthogonal to the previous component. This process might continue until the number of latent variables is equal to the mathematical rank of the data matrix (i.e., *n* or *m* depending if observations are less than variables or vice versa). Orthogonality condition implies that the following component explains new behaviour (phenomenon) of data and account for smaller portion of variance.

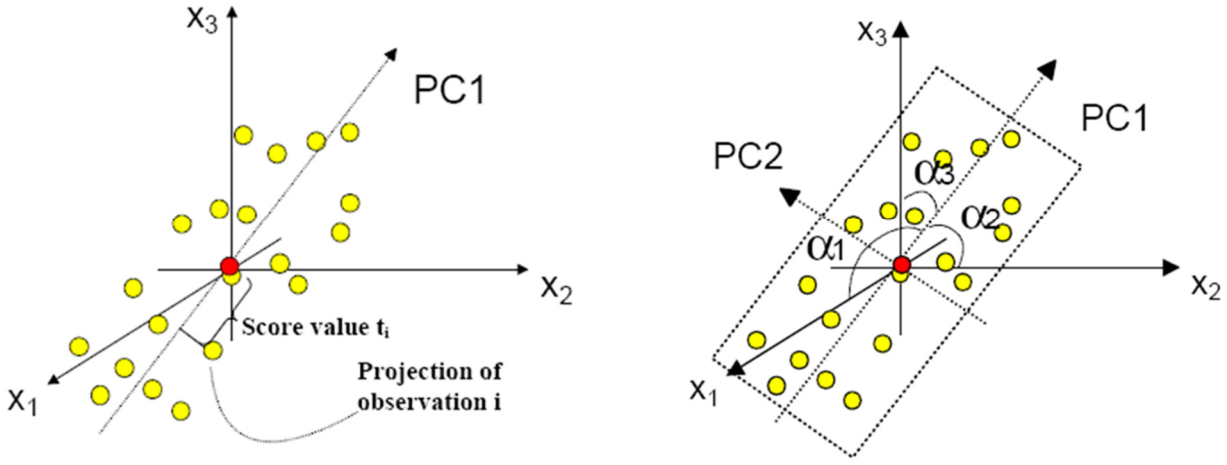

Figure 2.1 Graphical representation of principal component analysis in a 3D space. On the left the first PC and a score value, on the right the PC1-PC2 plane and loading values

Mathematically, PCA is an optimization problem, i.e. maximization of var. ($\boldsymbol{t_1}$) ($\boldsymbol{t_1}$= $\mathbf{X}\boldsymbol{l_1}$), with constraints ($\|\mathbf{l}_1\|$ =1, normalization, and $\mathbf{l}_1^{\mathrm{T}}$ $\mathbf{l}_2$= 0, orthogonality) and can be solved by eigenvalues/eigenvector decomposition of the covariance matrix of **X** (given that $\boldsymbol{t_1}\mathbf{T}\boldsymbol{t_1}$ = $\boldsymbol{l_1}'\mathbf{XTX}\boldsymbol{l_1}$):

$$cov(\boldsymbol{X}) = \frac{\boldsymbol{X}'\boldsymbol{X}}{m-1} \tag{2.5}$$

$$cov(\boldsymbol{X})\boldsymbol{l_i} = \boldsymbol{\lambda_i}\boldsymbol{l_i} \tag{2.6}$$

Eigenvectors corresponds to loadings and scores can be derived by definition above. Alternatively, PCA can be obtained by singular value decomposition (SVD) of **X**:

$$[\boldsymbol{U}\ \boldsymbol{S}\ \boldsymbol{L}] = SVD(\boldsymbol{X}) \tag{2.7}$$

where **S** is diagonal and holds the squares of eigenvalues with respect to eigenvalue decomposition of covariance matrix, **L** correspond to loadings and scores can be obtained by: **T** = **US**. Thus, PCA decomposes **X** matrix in the products of the matrix **T** score and **L** loading except for the residual matrix **E**:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{L}' + \boldsymbol{E} \tag{2.8}$$

The matrix **E** contains the variance of **X** that scores and loadings did not describe.

Residuals information is useful and used in diagnostic tool for observations and variables, e.g. identification of outliers, nonlinear trends, etc. The scores are of dimension *m* x *k* (*k* is the number of PCs) and the loadings of dimension *n (*columns of the original matrix) x *k*;

$$\boldsymbol{X} = \boldsymbol{t_1}\boldsymbol{l_1}' + \boldsymbol{t_2}\boldsymbol{l_2}' + \cdots + \boldsymbol{t_k}\boldsymbol{l_{1k}}' + \boldsymbol{E} \tag{2.9}$$

t and **l** pairs are related to one latent variable and in a descending order of importance from the first to the last component. The amount of explained variance is proportional to the eigenvalue **λ** from diagonalization of the covariance of **X**. The most widely diffused and used algorithm for PCA computing are the Non-linear Iterative Least Square (NIPALS) algorithm by H. Wold [28] and SVD. The principal difference between the two algorithms is that in SVD all-possible components are derived simultaneously while NIPALS is a sequential algorithm; the calculated components are obtained one at time a time iteratively. Various plots are obtained in principal component analysis and for the exploratory application, some of them shall be introduced.

The scores plot shows observations (samples) in a 2D space in which axes are two principal components, for instance the first against the second (the same could be done in a 3D space with three PCs), and obviously, the point coordinates are the projection of observation on the corresponding latent variables. This plot allows the visualisation of clusters, eventual outliers and to inspect the relation among samples, similarity or dissimilarity. The loadings plot shows variables instead of observations and their coordinates reflects the correlation with new latent variables: a high loading value corresponds to a high influence of that variable in the principal component. In loadings plot variable correlations and significances are highlighted, how much a variable affects a specific latent variable and in which way a sensor behaves in comparison to another one, directly, opposite or not correlated. Scores and loadings plots, if concern the same latent variables, might be represented together in order to discuss samples similarity/trends in terms of variables relevant to observe them. A so-called biplot combines scores and loadings plot in one graph but in my personal opinion, the two separate plots offer a more readable overview.

A PCA model is a good approximation of the raw data matrix. Two parameters are particularly significant to describe the compliance of data to the PCA model: these are two distances, defined for each sample, [13] the distance of the projected sample from the centre of the model (scores distance) and the distance of the sample from the model hyper plane (orthogonal distance). The first, known as Hoteling's T^2, measures variation within the model and corresponds to the sum of normalized squared scores for each *i* sample:

$$\boldsymbol{T_i^2 = \frac{t_i * t_i'}{\lambda_i}} \tag{2.10}$$

The eigenvalue normalize the T^2 values; it means that each score value is weighted by the explained variance in order to evaluate properly the distance, weighted on PC relevance. Such values return how much observation values differ from the model in terms of magnitude.

Assuming a normal distribution of the scores, the confidence limits are obtained using the F-distribution:

$$T_{limit}^2 = \pm \frac{\boldsymbol{A(N^2 - 1)}}{\boldsymbol{N(N - A)F_{critical}}} \tag{2.11}$$

Where N is the number of observations in the model training set and A is the number of components in the model. F critical uses A and N-A degrees of freedom.

The level of significance is normally set to 95%. It can be interpreted as measuring the systematic variations of the process, and out of limits data indicate that the systematic variations are out of control. The variable contribution for T^2 comes from the following equation:

$$t_{cont,i} = \boldsymbol{t_i}\lambda^{-1/2}x_{ik}\boldsymbol{p_k}^t \tag{2.12}$$

Where λ are the eigenvalues from SVD decomposition, x_{ik} is the value of deviating sample and $\boldsymbol{p}$ the loadings for variable k._The second distance parameter corresponds to the squared residuals, called Q, or SPE. It describes the distance from the hyper plane and indicates the conformity of each sample to the PCA model. Q returns the amount of unexplained variance that the decomposition did not capture with the k principal components. The Q value for each sample is:

$$Q_i = \boldsymbol{e_i}\boldsymbol{e_i'} \tag{2.13}$$

Where $\boldsymbol{e_i}$ corresponds to a sample residuals vector. The Q limits come from the χ^2-distribution. The contribution of the original variable to the Q value is:

$$Q_{cont,ik} = e_{ik}2 \tag{2.14}$$

Q and T^2 contribution plots for ease of interpretation by plant operator are frequently normalized by considering the 95 percentile. Sample with only high Q value, look 'well behaving' once projected on model space, because they share some features with the modelled observations, but it is not completely well modelled because part of their variation is not accounted by the PCA model.

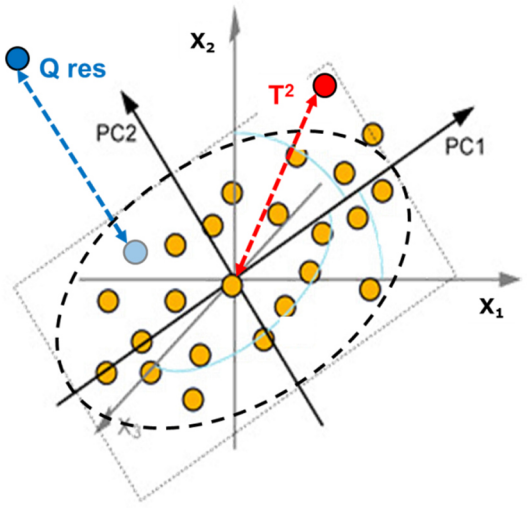

Figure 2.2 Graphical representation of the PCs space for a two components model. In red a sample with high T^2 value and in blue an observation with high Q residual value. Light blue dot represents the projection of blue observation on the components plane

From the contribution plots, it is possible to highlight the effect of single original variables to the deviating observation/batch and improve the understanding of the "abnormalities" causes Concerning the process monitoring, contribution plot describes which variables affect the observation projection and helps the problem solution via single sensors identification. T^2 and Q are fundamental in process monitoring and control: these two values are used to build multivariate control charts in which operators can recognize faults, drifts, plant aging and other phenomena that may affect production.

A drift in T^2 profile consists in one or more variables that move away from the mean plant conditions, the scores plot centre. For instance, consider a model that includes observations of a thermal exchanger that works for a while around the same temperature value and does not vary more than few decimal grades. A PCA model considers all the available observations. In a certain moment the set point is moved some grades above in which the items perfectly works; the projection of this new condition returns an high T^2 value because the model centre corresponds to another set point but projection gives a quite normal Q residual, in fact the variable correlation structure, i.e. correlation among sensors, did not change. After few observations, the set point returns at the previous value but the control valve of heating fluid shows a slightly different value; it depends on the previous movement and to the bad maintenance. Such observation returns a quite normal T^2, because in past observations valve have a similar degree of opening, due to the little temperature variation, but Q residuals increases because one variables behaves different from all the other and the correlation between them changed.

2.4 Regression methods

A regression method consists in the definition of a relation b between a measured variable x and a property y:

$$y = bx + e \tag{2.15}$$

Literature offers an incredible variety of regression methods [29 30] but everyone has the same purpose, the proper prediction of a property from one or more measured variables. When regression method are used for predicting concentration of chemicals either from spectroscopic measurement or other methods, the task is called calibration. In petrochemical industries many calibration are developed at laboratory scale. Laboratory take advantage from the use of calibrations: complicate analysis or the ones characterized by expensive procedures could be replaced from secondary methods such as infrared (IR) or near infrared (NIR) spectroscopy [31 32 33 34]. Regressions might involve a single measured variable, such as in linear regression or many [29] as in the multi linear regression (MLR) [35]. Univariate calibrations are not part of this thesis in which regression involved NIR spectroscopy and process monitoring applications that rarely find univariate regression the proper method. More than one variable take part in NIR calibrations, in this case more than a frequency, as in process, almost always, a sensor does not directly explain a product feature. Therefore, MLR can be used to build multivariate regression concerning the ability to manage high number of variable of this method. However, if the number of variables is particularly high, higher than the number of observation, a multilinear regression methods cannot be fit (not enough degree of freedoms). A number of variables higher than observation return an undetermined system because in the inversion operation that is performed in order to predict a new sample, there are not enough samples to solve the equation system. Another limit of multilinear regression is that the dependent variables have not to be correlated. Thus, NIR spectra, unless few wavelengths are selected, cannot be handled by MLR.

The partial least squares (PLS) regression [36] method approaches the problem in the latent variables space instead of the original ones; the idea is to extract the components that describe the system variation and at the same time are best correlated with the property response. The model shifts to the latent structure and for that, PLS is frequently called prediction in latent structure.

2.4.1 Model Validation

It is an intrinsic characteristic of every supervised model (regression, linear and non-linear, discrimination, classification, etc.) that model fit will increase with model complexity, i.e. augmenting the number of adjustable parameters in the model, such as number of descriptors variable in MLR and number of component in PCR, PLS, etc.. Hence, achieving the minimum model error, RMSEC (root mean squares error in calculation/fit), is not a criterion that can guarantee optimal model selection and in order to avoid model overfitting the predictive capability of the model has to be assessed, as well as the model stability, e.g. constancy of model parameters estimates when resampling, and consistency with the assumptions, e.g. residuals analysis. These issues are well described in a recent tutorial paper [37]. Here, the issue of model predictive ability is considered. In general, the set of samples used in model building (calculation) are called calibration or training set, these should be carefully selected (theory of sampling) as being representative of the whole variability domain in which we want our model to hold. In order to test predictive capability we need to use the model on samples different and mostly important truly independent, from the calibration ones, also called validation samples or test set. Also in this case they should span the whole experimental domain (both X and Y) on which the model is built.

Two phases have to be clearly distinguished: i) estimation of model meta-parameters, e.g. assessing the number of PLS components, and ii) model validation (predictive power). In the first phase, internal validation, Cross Validation (CV) is widely used. Since seldom the number of samples is sufficient to split them in three sets, calibration, internal validation and validation, the trick is to exclude in turn part of the calibration samples for internal validation and re-calculate the model until each sample has been excluded once and hence predicted. Root Mean Square Error in Cross Validation (RMSECV) is then used to estimate the CV error:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{cv,i} - y_{orig,i}\right)^2}{n}} \tag{2.16}$$

Equation 2.1 shows the cross validation error where $y_{cv,i}$ and $y_{orig,i}$ are respectively the predicted and the original property value for the *i-th o*bservation and *n* the number of samples. The original observation is compared with the one predicted and the sum of square residuals give the errors in cross validation. In the context of latent variables based model RMSECV estimation is repeated by changing the number of components from one to a predefined maximum so that the best model dimensionality (best compromise between minimum RMSECV and model parsimony) can be estimated.

To accomplish CV the dataset is split in groups and the cross validation takes into account one of them each iteration as external observations while the remaining samples are used for model calibration. Different splitting schemes can be adopted; three of the most commons are mentioned. The leave one out selection (CV-LOO) consists in removing a single observation at time; it is unique and useful for dataset in which the number of observation is quite low, otherwise tend to give a too optimistic estimation of prediction error. The Venetian blind scheme is very used due to the fact that is unique and good estimations are usually achieved. In this case, an observation every *k* (where k is the number of splits) observations is cancelled and predicted. In this case, Cross validation iterationss are round of *k/n*. This scheme fails in sorted dataset, ordered by time, samples preparation, or everything else; in such a case, *contiguous block* becomes the solution.

Here, a block of contiguous observations is cancelled in each split. Cross validation runs *b* iterations, where *b* is equal to the number of blocks/splits. It has to be mentioned that random subsets can be also adopted in this case however the scheme is not unique and splitting has to be repeated a certain number of times. In the second phase, an external validation set of observations, possibly from "in situation" use of the model has to be used, e.g. new batches, in batch production, further days, weeks in continuous production, further years in environmental/agro-food monitoring and so on. In this case the root mean square error in prediction parameter is used to assess predictive capability:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{pred,i} - y_{meas,i}\right)^2}{n}} \tag{2.17}$$

Further considerations have to be done, when CV is used for PCA models in supervised methods, such as SIMCA classification or , which is relevant to my Thesis, multivariate process monitoring. In these cases, the PCA models represent the reference description of the category of interest, e.g.in control samples, and thus the number of PCs is crucial. Not always CV is the best method to asses this [38 39] also it has to be taken into account that to obtain truly samples estimation in prediction it is not enough to adopt a splitting scheme for samples but this has to be coupled with splitting scheme for X-variables in appropriate way [40].

2.4.2 Partial least square regression

PLS regression [41] is a method for relating two data matrices, **X** and **Y**, by a linear multivariate model. It can predict more than one property thus can simultaneously handle Y matrix prediction (in this case the method is also called PLS2). PLS derives its usefulness from its ability to analyse data with many, noisy, collinear, and even incomplete variables in both **X** and **Y**. PLS operates simultaneously a PCA-like decomposition of **X (**descriptors) and **Y** (responses or properties matrix) in order to explain as much as possible of **X** and in order to find the best correlation with **Y**, i.e. the PLS criterion is to maximize XY covariance. More than one algorithm performs PLS but the following description relates to the NIPALS developed by H. Wold [29]. In the same way as the algorithm do in principal component analysis, decomposition of X and **Y** is accomplished:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E} \tag{2.18}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}' + \boldsymbol{F} \tag{2.19}$$

The **T** and **P** matrices are the **X** scores and loading and **U** and **Q** the score and loading of **Y**. In PCA, latent variables explain the main source of variation in the process data, but in PLS the latent variables give variation in **X** that is most predictive of the response **Y**. So the **X** scores are predictor of **Y** meanwhile are **X** latent variables. Geometrically, the principal component rotates components to improve the regression; it determines a new vector for each component that defines how much a variable deserves on this rotation. The algorithm calculates weights iteratively until the convergence:

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}' \tag{2.20}$$

$$\boldsymbol{W} = \boldsymbol{U}'\boldsymbol{X} \tag{2.21}$$

The weight matrix **W** comes from the **U** scores of **Y** and it shows the iterative process that generate **T** scores oriented to prediction. As in PCA, $\boldsymbol{t}$ vector is orthogonal along iterations. For each component a so-called inner relationship holds, i.e. X-scores **t** are linearly related to Y-scores u:

$$\mathbf{U}=\mathbf{T}b \tag{2.22}$$

The PLS model can finally be reformulated as a mulitlinear regression model where the PLS regression coefficients **B** directly relate the measured variables with responses:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{XB} \tag{2.23}$$

Where **B**:

$$\boldsymbol{B} = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}\boldsymbol{Q}' \tag{2.24}$$

In regression, the aim is to perform the best as possible prediction; an easy way to check performance is the comparison of predicted and measured value, in which the optimal fit follow the 1:1 regression line with R^2 equal to one. Often, the ideal solution is far away from the modelled.

The error in cross validation and external validation are more suitable than R^2 to verify the prediction capability because are in the measurement units and do not depend on the range of variation of the response spanned by validation samples, a comparison with primary error returns a preliminary information on calibration. The calibration bias shows the presence of systematic error in the model that should be carefully verified. As in PCA scores plots offer information on the sample quality; observation might be distributed following a secondary effect and scores values could highlight this behaviour with clustering in scores hyper plane. Sample quality might drift for many reasons, i.e. sampling and measurement condition, and the scores plot supports the identification of such problems. In PLS both X and Y scores are available to interpret.

In PLS model the **X**-weights values assume greater importance then **X**-loadings: weights represent the importance of single variables for prediction and the relationships between variables. The knowledge on system on which regression is applied could be verified with weights as, on the other side, they might suggest unknown behaviours of predicted system. In addition, the residuals variance offers some diagnostic on regression, e.g. if the model is adequate or not, if there are trends in residuals while they are assumed normal, etc. In the case of multiple responses, in which **Y** has more than a column, the **Y**-loadings describe correlation between responses variables; a PLS2 model returns information also on how dependent variables are inter-correlated.

2.4.3 Locally weighted regression

Polymer productions, maybe more than other chemical and petrochemical manufacturing, suffer of aging effect usually due to the smooth deterioration taking place in the system, e.g. exhaustion of catalyst that behave differently from start to end of its life. This source of variation affects all sensors and their correlation and reflects in shifting (introduce a drift) the whole set up to different normally operation condition (NOC). It means that the quality of production is still acceptable but the sensors register different settings. When this occurs, if predictive models, based on process variables, are active for on-line monitoring of specific properties it is likely that they will start to furnish biased predictions. In fact, the shift in NOC means that the new samples (time points) will move systematically in a different direction in the latent variables space with respect to where calibration samples were located. Another condition that could introduce bias in the predictions is changing the production type, this is frequently done in polymers production because it is often needed to switch to a different polymer grade while the process and plant used is the same. A different polymer grade consists in the same monomers and comonomers, eventually some additives change, with new final proprieties that come from another plant setting; final polymer differs not from the main component but for their physical characteristics that strongly relates with the production conditions. In this case, as well operative settings change systematically from a certain time point on. One solution might be to build distinct predictive models for each type of polymer of different grade; but in continuous production there is a transitions between different polymer grade and thus if different models are adopted there will be a time period that none of the models will be able to predict properly. Moreover, it has to be consider that not always the drift or production changes are constantly shifting in the same direction but there can be cycles, in other words it might happen that operative settings will be more close to a period in the past than to nearest time points.

From a data analysis point of view Locally Weighted Regression (LWR) [42 43] may deal with these problems. Locally weighted regression is based on local instead of global regression models, by re-considering model building each time a prediction has to be done, selecting as calibration set a given number of closest neighbours, in original or latent variables space, to the sample (time point) to be predicted. This feature seems appealing to better describe the current process conditions, since model the calibration model adapts himself to the samples, and it ideally removes all the calibration points that does not resemble the new operative condition. Thus, we expect that the application of LWR to on-line predictive models should cope with variance due to the product modifications and long-term drift improving the prediction quality.

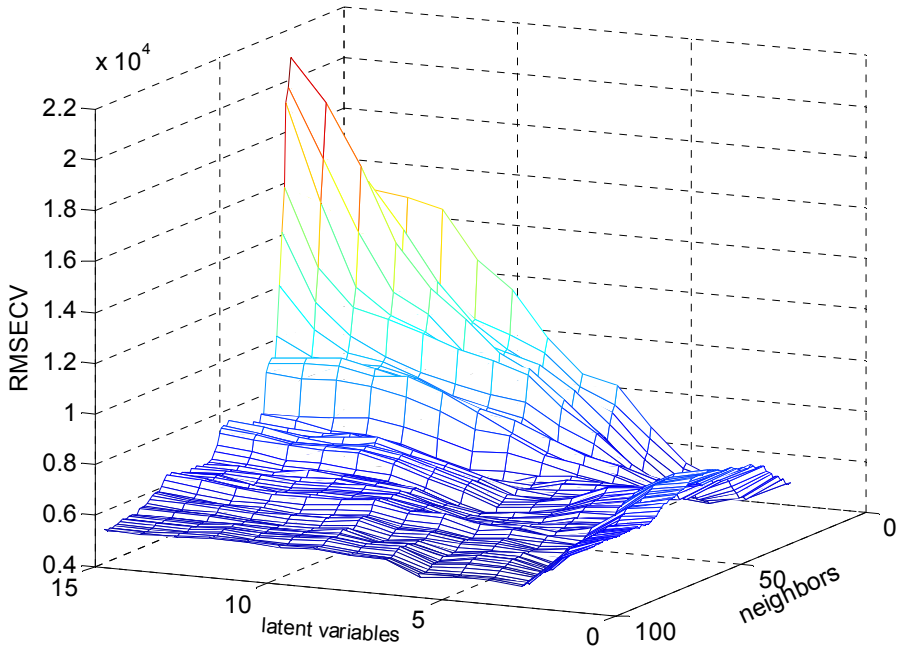

Figure 2.3 Example of RMSECV surface for LWR models. RMSECV is reported as function of the number of latent variables (in the examples varying from 1 to 15 in step of 1) and the number of neighbours (varying from 1 to 100 in step of 1).

LWR method, when PLS is used as regression method, needs to set the parameters: (1) the number of neighbours, *n*, to be considered (that will be used as calibration set); (2) the weight function, i.e if neighbours are to be weighted according to their distance, and (3) the number of PLS components, *A*. It consists in a setting phase, where these parameters are defined, and in a successive one where predictions are performed based on these settings. In other words, every time a new sample has to be predicted: (4) the *n*-neighbours are identified form preceding observations, by calculating the distances between the samples and each previous observations; (5) a PLS model with *A* latent variables is calculated where the *n* calibration samples are not used as such, but properly scaled according to distances in (5), and (6) the sample is predicted according to this model. The parameters selection, first phase, can be operated by considering as fitness function the root mean square error in cross validation (RMSECV), internal cross-validation, or root mean squares error of an external (monitoring) test set (RMSEPm). Practically, *n* and *A* are varied on a grid of values, whose range is user defined, and left out (CV) or monitoring test samples are predicted to obtain corresponding RMSECV/RMSEPm, then the error surface is inspected, Fig. 2.3, to find a minimum. The weight function is defined a priori. Step 4 might vary between applications, for instance could be either in the original space or into the latent variables space.

Concerning the petrochemical environment, locally weighted regressions aim at improving prediction for on-line PLS models that are affected by variability that cannot be avoided, such as the plant aging or that belongs to the specific way process is operated, e.g. the transition from one polymer grade to another during production. Likely, LWR models can fix also the flow rate disturbance in which a change in variables relationships and magnitude occur.

2.5 Classification methods

Classification problem are encountered every times there id the need to discriminate samples, to identify sample origin and in all possible situations in which observation have to be categorized [44]. For instance, in chemical productions it is required to discriminate good and bad productions or material degree. At laboratory scale, there are plenty of examples in which classification is used; often it is necessary to recognize samples origin, to discriminate raw materials between suppliers, to identify the compounds type and so on. Classification methods use a priori class information and are supervised methods, just to name some representative for kind: Soft Independent Modelling of Class Analogies (SIMCA) [45] is a class-modelling tool handling one-class model situation and allowing assignation or not of sample to a given category, as well as multiple assignations; Partial Least Square Discriminant Analysis (PLS – DA) [46], as linear classifier; and Support Vector Machines (SVM) [47] as non-linear classifier (provided the kernel function used is non-linear). Sometimes analysts do not know a priori categories but are interested in finding if and in how many classes samples may group; in this case, unsupervised methods are used, and mostly are distance-based such as Cluster Analysis [48] and k-Nearest Neighbours (kNN) [49].

PLS-DA has been used in batch process analysis, Chapter 3.6 and is briefly described in the next section.

2.5.1 PLS discriminant analysis

As his name suggests, PLS-DA derives from partial least square and allows discriminant analysis [50]. The algorithms does not differs from the PLS and the discriminant part consist into the **y** vector. Instead to a continuous parameter, **y** consists in a dummy variable with 0/1 or -1/1 codification in which zero (minus one) defines not-belonging to a given category or class and one defines belonging to it. More than two classes return a dummy **Y** matrix with a number of columns equal to the number of classes, each one codified as described above. The model hyperspace is divided in as many regions as the number of classes imposed in calibration dataset. In PLS-DA to choose the number of components CV is, as well, used but the fitness function to minimize in this case is the number of misclassifications in CV. In fact, prediction returns a real value for each class, not only 0 and 1, and using RMSECV would be misleading. The classification rule in PLS-DA can be implemented differently considering the available software. One straightforward way is to associate the unknown sample to the class with the highest y-predicted value. A second approach defines a proper threshold on y-predicted values for each class to establish if sample belongs or not to the specific class.

A model returns prediction that could be false negative and false positive and the limits definition must consider sensitivity and specificity. False negative means that sample of a class is predicted as not belonging to its class and shows a low sensitivity. False positive describes a situation in which sample is associated to a class but it belongs to another and consists in low specificity.

Such as the classic PLS, the PLS-DA returns loadings and weights in which analyst might find variables related to a good discrimination and with high significance in observation clustering. The variables selection methods might improve sample discrimination as it does in regression models.

2.6 Variable selection methods

In analytical methods such as the NIR spectroscopy and in polymers production context, variables are highly correlated, noisy and sometimes useless for the aims. Latent variables based methods are generally robust to noise unless extreme imbalanced situations among informative and uninformative variables are the case. However, often for interpretative purposes or to simplify models can be useful to rank or select features according to their relevance in modelling, especially in regression/classification tasks. Often, there is a pressure from spectroscopist or plant engineers to select features a priori on the basis of their expertise and their knowledge has always supported application development, nonetheless their view is most often based on univariate approach and especially in the analysis of complex systems, such as a production plant and when impurity, interfering species, etc. affect spectra, it might be really sub-optimal to discard feature a priori. The risk is to miss relevant information and to be misleading in evaluation of causal relationships. In addition to the a priori knowledge that, in my opinion, should be still considered maybe after learning from data, literature offers plenty of variable selection methods [51 52 53] that fit various kinds of data and problems. This section presents the Variables Importance in Prediction (VIP), Forward/Backward selection and Genetic Algorithm (GA), the methods involved in the thesis applications. Anyway, in order to use the proper selection method, it is necessary to realize that each one works under certain assumptions.

2.6.1 VIP

Variable Importance in Prediction (VIP) [51 55 56] is a variable ranking method, i.e. it point out the relevance of variables in a PLS model, but does not compare models built on different number/combinations of features. It asses the influence of variables by combining measurement of how much a variable contributes to data description, both in the dependent (X) and the independent (Y) latent variables spaces In other words, VIP indicates which of the variables, independent ones, are the most involved in parameters prediction, dependent variables.

The VIP value for the *j-th* variable is given as:

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 * SSY_f * J}{SSY_{total} * F}} \tag{2.25}$$

w_{jf} is the weight of variable *j* in the *f* principal component, SSY_f is the explained variance of the **Y** by the *f-th* component and *J* is the number of X-variables.

Numerator describes the relevance of a variable in predicting the Y (i.e. through the PLS weight) weighted by how much Y variance each principal component account for. In other words, if a variable attain a high weight values in a component that account for low amount of Y variance it is considered less important than a variable who attain the same weight but in a component describing higher Y variance. SSY_{total} is the total explained variance of the dependent variable and f is the number of components. Denominator serves for normalization, in fact sum of squared VIP over all j variables close to J. Given this closure, generally a threshold of 1 is used to select the most influential descriptors for the model and lower VIP values as indicator of poor significance. However, recent literature suggest several ways to refine the choice of threshold [57].

2.6.2 Interval PLS

Interactive or interval PLS [58], iPLS, has been proposed initially to select relevant spectral regions in NIR spectroscopy data set. The method is based on calculating several PLS models including only variables corresponding to small regions of the spectrum, called blocks or intervals. The number of blocks/intervals is user defined and intervals overlap is allowed. The evaluation of insertion/deletion of a block is done sequentially and both forward and backward selections are applied. Block dimension can be a single variable; in this case dimension should be imposed to one. In forward direction, for the first cycle, algorithm builds as many model as the number of defined intervals. Cross validation is used to compare models and the one with lowest error is selected. This is the first cycle of variable selection. When more than one interval is desired, iPLS performs a second cycle in which the first interval is coupled with the other remaining intervals, one at time, to create a new set of PLS models. Comparison bases again on RMSECV and the best pair of intervals is selected. Such cycle is performed as many times as requested. Reverse mode operates removing intervals from the entire dataset; in this case the worst block, the one with highest error, is removed. User might set the iPLS parameter (block size, number of iteration and block overlapping) considering many sides of their dataset; as an example, a complex system suggest more intervals, in order to keep more information of a complicate set of data.

Another important aspect is the variable reduction that aims to the selection of few important variables and this case orients user on low iteration number. Block size and overlapping strictly depends on samples and vary case by case. The three parameters offer different solution on the same task and user have to set iPLS to obtain a suitable selection for their problem.

2.6.3 Genetic algorithm

Genetic algorithm (GA) [59] is an optimization or search methodology inspired by genetic selection theory. This algorithm simulates the evolution process, in which the best individual has more chance to transmit its genome and to survive and as in nature, these steps are coded: initial population, mutations, cross-over and next generations.

Among the different context of use, GA can be used for variable selection tasks [60]. Concerning the variable selection, gene codification is binary thus a, one indicates that the variable corresponding to this gene has to be selected, while zero indicates non-selection. A chromosome comprises as many genes (entries) as variables, and thus points to a subset of the variables. In our process context, a chromosome consists in 0s and 1s where zero means sensors exclusion and one inclusion. The GA algorithm used in the Thesis application is the one of R. Leardi [61], in which, based on long experience, most of the usual parameters of choice, such as number of individual in the starting population, percentage of mutations and number of generations are fixed. Each chromosome in initial population is randomly filled. As fitness function, to evaluate the best individuals in the population, RMSECV of the PLS model built with the selected variables is used, adopting venetian blind as CV scheme. As in nature, next generation comes from the previous one and each offspring has usually half of the parent chromosomes. From a pair of chromosomes comes a new chromosome; a random assignment gives to each gene in the child chromosome the value of one parent, in relation to the same gene. It means that offspring bring to a different subset of variables. The probability of an individual of being selected for next generation is associated to his fitness function value, so the best ones have a greater probability of being picked up than the worst. Moreover, elitism is applied, i.e. the two absolute best individuals are kept in next generation, in other words never die. Further, as in nature mutations take place, i.e. random changing a small percentage, about 1%, of some genes value. The algorithm ends when the fixed number of generations has been achieved. Since GA is based on random initialization, restarting a run will produce different results. Thus, GA is restarted from scratch one hundred times, i.e. 100 runs. The frequency of selection in the final model (corresponding to the end of each GA run) for each variable is stored. Then a stepwise PLS procedure is used to finally select how much of the most frequently selected variables are to be retained. The entire procedure is repeated five time (for a total of 500 hundred GA runs from scratch) and the variables belonging to the five selected sets (hence the most frequently selected in the 500 hundred runs) are considered altogether to enter the final stepwise procedure.

This algorithm has the advantage of balancing exploration and exploitation and restarting many times avoid being stuck in local minima. Moreover, using frequency of selection lower the possibility of chance selection. In order to avoid over-fitting, in the algorithm are implemented two heuristic: i) the total number of variables to be selected do not to exceed two hundred (in case there are more binning is applied) and ii) a preliminary check for chance correlation is operated by adding a significant number of random variables to the data set and verifying that the frequency of selection of this variables is well below the one of "real" variables. The R. Leardi GA has been modified in our research group in order to work in two steps when the number of variables, as in spectra, is huge. In the first step, a gene codifies an interval instead of a single variable. So the heuristic is that no more than 200 intervals are defined. In the second steps the data sets is formed by the most selected intervals of the previous 500 hundred GA runs (without using stepwise selection) stopping to add intervals when the sum of the number of variables in all selected intervals exceed 200. Then the algorithm is used again with normal codification for single variable.

2.7 References

1. Tauler, R., Walczak, B., & Brown, S. D. (2009). *Comprehensive chemometrics: chemical and biochemical data analysis*. Elsevier.
2. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke (1998). *Data Handling in Science and Technology Volume 20, Part A, Pages 1-867 (1998),* Handbook of Chemometrics and Qualimetrics: Part A, ISBN: 978-0-444-89724-4.
3. Wold, S., Berglund, A., & Kettaneh, N. (2002). New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P (research, development and production)—with examples from pharmaceutical research and process modeling. *Journal of chemometrics*, *16*(8-10), 377-386.
4. Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., & Andersson, C. A. (1998). Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometrics and Intelligent Laboratory Systems*, *44*(1), 31-60.
5. Hair, Joseph F., et al. *Multivariate Data Analysis: A Global Perspective*. 7th ed. Upper Saddle River: Prentice Hall, 2009.
6. Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, *6*(6), 329-348.
7. Workman Jr, J. (1993). A review of process near infrared spectroscopy: 1980–1994. *J. Near Infrared Spectrosc*, *1*(4), 221-245.
8. Park, G. J. (2007). Design of experiments. *Analytic Methods for Design Practice*, 309-391.
9. Eriksson, L. (Ed.). (2008). *Design of experiments: principles and applications*. MKS Umetrics AB.
10. Kettaneh-Wold, N. (1992). Analysis of mixture data with partial least squares. *Chemometrics and Intelligent Laboratory Systems*, *14*(1), 57-69.
11. Wold, S., Johansson, E., Cocchi, M., *3D QSAR in Drug Design: Theory, Methods and Applications*. ESCOM Science, Publishers, Leiden, Netherlands, 1993, p.523
12. Kourti, T., *Multivariate Statistical Process Control and Process Control, Using Latent Variables.* Brown, S. D.; Tauler, R.; Walczak, B. (Eds.); *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*. Ed. Elsevier Science Ltd., Amsterdam, Netherlands, 2009.

13. Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and intelligent laboratory systems*, *28*(1), 3-21.
14. Kourti, T. (2006). Process analytical technology beyond real-time analyzers: the role of multivariate analysis. *Critical reviews in analytical chemistry*, *36*(3-4), 257-278.
15. Macho, S., & Larrechi, M. S. (2002). Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. *TrAC Trends in Analytical Chemistry*, *21*(12), 799-806.
16. Liang, J., & Qian, J. (2003). Multivariate statistical process monitoring and control: Recent developments and applications to chemical industry. *Chinese Journal of Chemical Engineering*, *11*(2), 191-203.
17. Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, *19*(4), 213-246.
18. Seasholtz, M. B., & Kowalski, B. (1993). The parsimony principle applied to multivariate calibration. *Analytica Chimica Acta*, *277*(2), 165-177.
19. Van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, *7*(1), 142.
20. González-Martínez, J. M., Noord, O. E., & Ferrer, A. (2014). Multisynchro: a novel approach for batch synchronization in scenarios of multiple asynchronisms. *Journal of Chemometrics*, *28*(5), 462-475.
21. Kowalski, B. R., & Seasholtz, M. B. (1991). Recent developments in multivariate calibration. *Journal of Chemometrics*, *5*(3), 129-145.
22. Fearn, T., Riccioli, C., Garrido-Varo, A., & Guerrero-Ginel, J. E. (2009). On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems*, *96*(1), 22-26.
23. Isaksson, T., & Næs, T. (1988). The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. *Applied Spectroscopy*, *42*(7), 1273-1284.
24. Vigni, M. L., Durante, C., & Cocchi, M. (2013). —Exploratory data analysis. In *Data handling in science and technology chemometrics in food chemistry* (Vol. 28, pp. 55-126). Elsevier.
25. Jackson, J. E. (2005). *A user's guide to principal components* (Vol. 587). John Wiley & Sons.
26. Jolliffe, I. (2002). *Principal component analysis*. John Wiley & Sons, Ltd.

27. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1), 37-52.
28. Wold, H. *Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach*. In: Gani, J.(ed.): *Perspectives in probability and statistics*. Applied Probability Trust, Sheffield, England, 1975.
29. Kutner, Michael H. *Applied linear statistical models*. Vol. 4. Chicago: Irwin, 1996.
30. Noorossana, R., Eyvazian, M., Amiri, A., & Mahmoud, M. A. (2010). Statistical monitoring of multivariate multiple linear regression profiles in phase I with calibration application. *Quality and Reliability Engineering International*, *26*(3), 291-303.
31. Zagonel, G. F., Peralta-Zamora, P., & Ramos, L. P. (2004). Multivariate monitoring of soybean oil ethanolysis by FTIR. *Talanta*, *63*(4), 1021-1025.
32. Chung, H. (2007). Applications of Near-Infrared Spectroscopy in Refineries and Important Issues to Address. *Applied Spectroscopy Reviews*, *42*(3), 251-285.
33. Reboucas, M. V., dos Santos, J. B., Domingos, D., & Massa, A. R. C. (2010). Near-infrared spectroscopic prediction of chemical composition of a series of petrochemical process streams for aromatics production. *Vibrational Spectroscopy*, *52*(1), 97-102.
34. Bonacini, F., Ferrando, A., Mantovani, E., Sappino, C., Arcidiacono, G., Ardizzone, D., & Rossi, E. (2013). Fourier transform near infrared application for advanced process control of an ethylene cracking plant. *NIR news*, *24*(6), 9-11.
35. Martens, H., & Næs, T. (1984). Multivariate calibration. I. Concepts and distinctions. *TrAC Trends in Analytical Chemistry*, *3*(8), 204-210.
36. Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, *185*, 1-17.
37. Westad, F., & Marini, F. (2015). Validation of chemometric models–A tutorial. *Analytica chimica acta*, *893*, 14-24.
38. Camacho, J., & Ferrer, A. (2012). Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects. *Journal of Chemometrics*, *26*(7), 361-373.
39. Camacho, J., & Ferrer, A. (2014). Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: practical aspects. *Chemometrics and Intelligent Laboratory Systems*, *131*, 37-50.
40. Bro, R., Kjeldahl, K., Smilde, A. K., & Kiers, H. A. L. (2008). Cross-validation of component models: a critical look at current methods. *Analytical and bioanalytical chemistry*, *390*(5), 1241-1251.

41. Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.
42. Bevilacqua, M., Bucci, R., Materazzi, S., & Marini, F. (2013). Application of near infrared (NIR) spectroscopy coupled to chemometrics for dried egg-pasta characterization and egg content quantification. *Food chemistry*, *140*(4), 726-734.
43. Cortizo, M. S., Larsen, D. O., Bianchetto, H., & Alessandrini, J. L. (2004). Effect of the thermal degradation of SBS copolymers during the ageing of modified asphalts. *Polymer Degradation and Stability*, *86*(2), 275-282.
44. Kurt Varmuza, Modelling of Clusters, Chapter Pattern Recognition in Chemistry, Volume 21 of the series Lecture Notes in Chemistry pp 88-91
45. Candolfi, A., De Maesschalck, R., Massart, D. L., Hailey, P. A., & Harrington, A. C. E. (1999). Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA. *Journal of pharmaceutical and biomedical analysis*, *19*(6), 923-935.
46. Chiang, L. H., Russell, E. L., & Braatz, R. D. (2000). Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics and intelligent laboratory systems*, *50*(2), 243-252.
47. Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.
48. Hastie, T.; Tibshirani, R.; Friedman, J. *Hierarchical clustering. The Elements of Statistical Learning (2nd ed.).* Springer, New York, USA, 2009.
49. Sharaf, M. A., Illman, D. L., & Kowalski, B. R. (1986). *Chemometrics* (Vol. 82). John Wiley & Sons.
50. Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of chemometrics*, *17*(3), 166-173.
51. Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, *24*(11-12), 728-737.
52. Chong, I. G., & Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, *78*(1), 103-112.
53. Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62-69.
54. Weisberg, S. (2005) Variable Selection, in Applied Linear Regression, Third Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0471704091.ch10

55. Wold, S., Johansson, E., & Cocchi, M. (1993). PLS—partial least squares projections to latent structures. *3D QSAR in drug design*, *1*, 523-550.
56. Favilla, S., Durante, C., Vigni, M. L., & Cocchi, M. (2013). Assessing feature relevance in NPLS models by VIP. *Chemometrics and Intelligent Laboratory Systems*, *129*, 76-86.
57. Gosselin, R., Rodrigue, D., & Duchesne, C. (2010). A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemometrics and Intelligent Laboratory Systems*, *100*(1), 12-21.
58. Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, *54*(3), 413-419.
59. Leardi, R. (2007). Genetic algorithms in chemistry. *Journal of Chromatography A*, *1158*(1), 226-233.
60. Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of chemometrics*, *6*(5), 267-281.
61. Leardi, R., Seasholtz, M. B., & Pell, R. J. (2002). Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Analytica Chimica Acta*, *461*(2), 189-200.

III

Abnormal situation detection

Batch process analysis

Content

3.1 Introduction

In chemical and petrochemical production, continuous and non-continuous processes belong to a completely separated word [1]. To give an idea of a Batch process we can take example from our everyday life. Consider the cake cooking: a recipe describes which ingredients it needs, how to stir the mixture, how long get it into the oven and the proper temperature. Industrial batch processes are similar, a formulation defines the raw materials and their quantity, a proper set-up and scheduled operations are used to conduct the process until the final product is obtained. In continuous production, the different position on the production-chain represents a different time, i.e. the time evolution of the "intermediate products" takes place at different position in space. As an example, let consider an industrial cake production: cakes pass through a long oven, with an increasing temperature from the beginning to the end, at low velocity in order to get properly cooked; then candies, creams and other products are added after, each one in a precise point of the conveyor belt. Thus cooking time/stage of the cake correspond to position on the oven.

In petrochemical continuous process is the same: the main steam come across different condition along pipe-line and the additives enter in specific position Petrochemical industries favour continuous processes because with respect to batch processes present several advantages: time saving, less operations, lower variability and others related to the easier management of safety and environmental issues. Discontinuous production has long been the accepted procedure for the manufacturing of many types of polymers, but nowadays the tendency is to replace it, mainly batch processes concern the pharmaceutical industry in which smaller quantities are worked. Despite this general trend, some products in the chemical processing industry are still realized via batch or both kinds of processes may coexist in some polymers production due to the specific requirement of final products that are made starting from the same raw materials. Moreover, batch process allows flexibility in processing of multiple products by accommodating the diverse operating conditions or additives receipts associated with each product. Again, in spite of the growing change toward continuous production, the batch type remains the only alternative for plenty of sectors in processing industry. As a result, research development in multivariate data analysis of process data focused most efforts in algorithms/approaches for batch process understanding and control [2 3], at the same time many applications have been published [4 5]. Such as in a continuous production, data are stored for each sensor with a certain frequency. Apparently, the data set structure appears similar between the two process types, continuous and discontinuous (batch).

Thus, what is peculiar in batch production? Plant managers wish to compare products and production and, in a continuous process, it is quite straightforward to relate the every second or minute, data acquired by process sensors to the plant set-up at that time points and the quality of the final material obtained, since the flow is continuous. In batch, the situation is different, since each batch exists in a time laps that goes from raw material loading to the discharging operation. Thus, the final products have to be carefully related to the specific starting batch and process conditions of that time period. In practice, the collected sensor data have a three dimensional structure: batches x sensors x time points (number depending on batch duration). At Versalis a specific batch production, namely the EPS plant, ST14, that will be described in detail in next sub section, has been of particular interest because serious troubles were caused in the past due to abnormal growth of polymer in big aggregates while it should be homogenously dispersed in smaller particles.

This phenomenon is defined as instability of the polymer dispersion and when it occurs, the production has to be stopped as soon as possible to preserve the reactor. The precipitation of monomer/polymer styrene causes the spheres aggregation that cannot be separated again. Plant management affirms that the instability is the biggest problem in EPS production and actually, it is true considering that a dump batch costs more than 100k euros. An intrinsic difficult in control is that the instability phenomenon is not signalled by any of the process parameters examined as far. Usually, to prevent instability to take over, the particles size in suspension is monitored. This is done by an expert operator that looks at the suspension through a window made in a derivation loop and observe how much the spheres increase their size. Observation is done at the expected critical time interval during batch production. Moreover, as a preventive measure, in order to avoid the particles growing up too much, Tricalcium Chloro Phosphate (TCP) in powder is added. The added quantity depends on how big the particles size in the observed suspension is; it means that to an instable batch more suspending agent (TCP) will be added with respect to a stable one.

For few batches, if compared to a daily production of two or three, was not possible to remedy instability and the batch dump, i.e. styrene went up into the reactor and consequently all the raw materials cannot be reused and shall be treated as waste, a further undesired cost. That is the motivation that brings the focus of multivariate analysis on the period where variability of the suspension was observed. Initially, I focused on the dump batch as the undesired critical event but at second time, it seems more effective to analyse the production that required high TCP addition, i.e. the period in which batches reacted completely, but with high instability degree during reaction.

3.2 ST14 Process description

Expandable polystyrene (EPS) [6] consists of little spheres, with a diameter from 0.2 mm to 3.0 mm. It contains a mixture of iso-pentane and n-pentane, in concentration between 4% and 7%, which acts as expanding agent. In dedicated items, EPS expands and the size increase up to 50-100 times; expansion happens gradually and at controlled temperature of about 80°C. ST14 line produces the expandable pearls, while the final items, such as safety packages or building insulation panels are manufactured by customers buying EPS pearls form Versalis. Essentially Versalis portfolio [7] include three groups of EPS:

- Normal EPS
- Self-Extinguish EPS
- Ecological EPS

The so-called "Normal" uses pentane as expanding agent and coating to improve processability. The second product differs from the classical EPS due to addition of an anti-flame agent; it cannot propagates the flame, so it cannot burn up without an external and continuous flame. The last EPS, ecological grade, use less expanding agent and its production is more sustainable than for normal EPS type. A lot of products can be manufactured by this three basic polymers, as function of particles size, type of additives and other material characteristics.

The polymerization takes place in water suspension. In a first vessel styrene and additives are mixed together to realize the so-called monomer mixture. A second vessel contains water and receives the mixture: the high-speed rotation and the suspending agent, first suspending agent addition, allows drops formation and reaction. As in a continuous flow stirred-tank reactor (CSTR), polymerization cycle contemplates that the whole mass should attain the same set temperature, pressure and stirring. Reactions undergoes various steps; phase one) the monomer to polymer conversion proceeds of about 70% at this point a second aliquot of suspending agent is added to stops the spheres size growing and to stabilize suspension. When EPS particles reach the expected conversion grade, bead identity point, a third addition of suspending agent make the solution/suspension ready to pass to the next step: the addition of the expanding agent. At this stage, the plant operator increases the temperature and the batch completes reaction (conversion of about 70%). At this point, the washing and packaging steps begins. In order to facilitate the comprehension of the results, the two next sub-sections describe in deeper details the main production steps.

3.2.1 Monomer mixture preparation

The first phase of ST14 happens into the vessel, named D401 (Fig. 3.1), this is a CSTR system in which the monomer mixture is homogenised. The due styrene quantity fills D401 and operators execute these actions:

- Add liquid and solid additives
- Purge D401 with nitrogen
- Increase temperature up to 60°C

Monomer mixture preparation takes 4 hours in isothermal conditions without requiring any operator intervention. After this time, the mixture is transferred into the vessel called R401, and the transfer line is washed with demineralized water.

3.2.2 Polymerization

Styrene polymerization happens in a dedicated reactor. In order to obtain a semi-continuous production polymerization is accomplished in parallel in three identical vessels, called R401A/B/C. They have the same equipments and instrumentations and work in parallel. Operator prepares each reactor, before the mixture is transferred, by filling with water (final water-styrene ratio shall be about 0.5), adding the first catalyst, NaOH (to regulate pH) and the first dose of suspending agent solution. Then monomer mixture can be transferred. Styrene polymerization requires another catalyst that works at higher temperature than the first one, and also sodium metabisulphite (MBS), a sort of soap agent, is added which operate with the suspending agent, tricalcium chloro phosphate (TCP), to ensure suspension stability. Then, vents are automatically closed and the reactor temperature increase until 90 °C; during heating plant operator checks pressure (on univariate chart) that should exceed one bar and, if necessary, they will introduce nitrogen. At time *zero*, i.e. at the time point in which the thermocouple registered exactly 90 °C, the operator controls the suspension quality by sight (looking at the glass window positioned in the derivation loop); the reaction is checked with a defined frequency in order to monitor particles size until the end of the 90°C phase. The so-called "*separation zero"* time point corresponds to when a certain conversion degree is reached where the polymer has a weight quite similar to the water and this is the correct time to feed the second amount of suspending agent. This addition definitely stops pearls growing and determines the final dimension of EPS product. The rotation per minute (RMP) value augments and the temperature passes from 90°C to more or less 120°C, in function of the desired EPS grade.

Stirring is fundamental in reaction that took place in suspension and probably is the most important parameter for the plant operators. Last step consists of expanding agent addition: just before the heating ramp begins, pentane mixture is pumped slowly into R401; it take half an hour. Pressure increases and reach about 10 bars due to the combined effect of temperature and expanding agent; it is a positive effect that helps pentane permeation; in case of low pressure, a nitrogen aliquot could be entered. EPS stays a given period at high temperature to be sure that complete conversion of styrene in polymer take place; complete means that the monomers left in water should be less than 0.05%. Then, the jacket cool down reactor to 60°C and pearls are send to the finishing section in which they are washed and packed in special packaging so called “octabins” (with an octagonal base).

To summarize the polymerization reaction includes five temperature steps:

- Heating from 60°C to 90°C
- A period at stable 90°C temperature
- Heating from 90°C to 120°C (in relation to EPS grade)
- A period at stable 120°C temperature
- Cooling from 120°C to 60°C

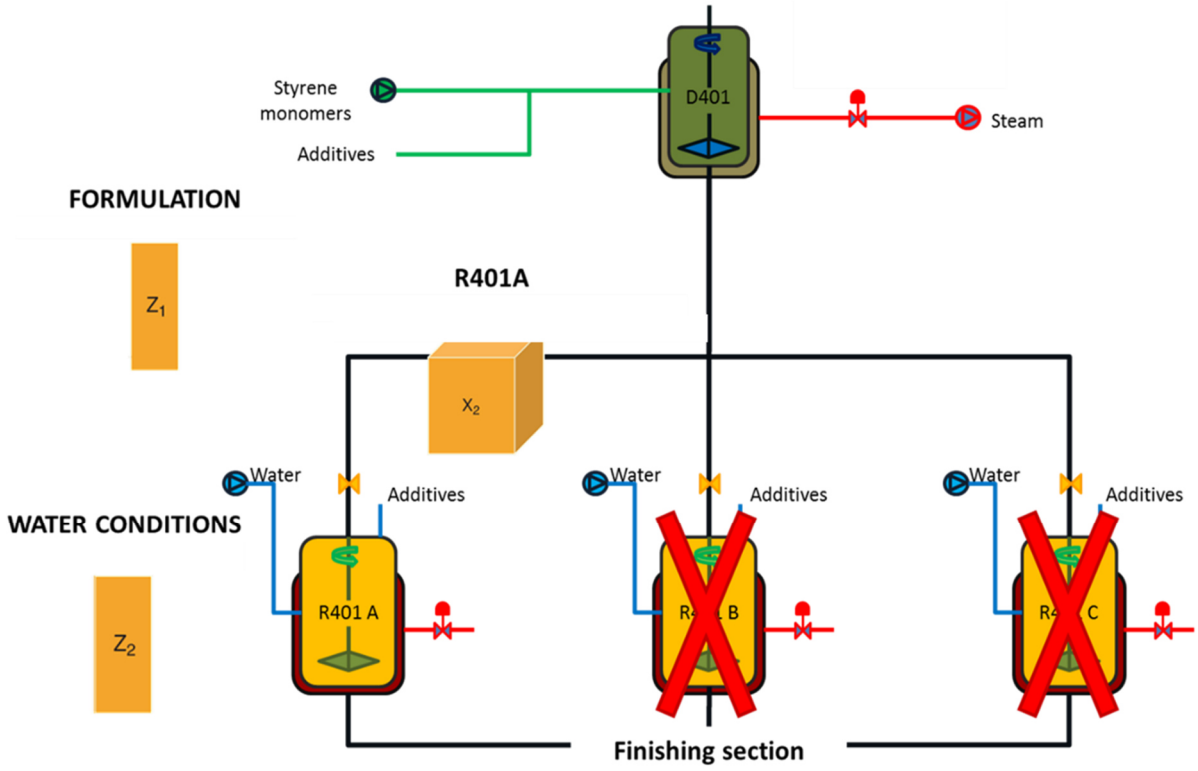

Figure 3.1 Basic ST14 plant scheme

3.2.3 Suspension instability

EPS production concerns a simple and well-known reaction. In terms of conversion and chemical knowledge, there are no critical points. Despite this, suspension is incredible complicate due to a sensible equilibrium between two phases: polymers and water. When instability occurs, plant come across many issues. Firstly, styrene and polystyrene pearls condense until they form a single undispersed phase. Suspension is lost and there is no procedure to recover it. Then, polymerization continues due to the presence of catalyst, temperature and monomers. Even more, such unwanted polymerization generates heat that accelerate conversion rate as in a runaway reaction [8]. In order to avoid all these phenomena, plant operators cool down the reactor jacket that absorbs heat of reaction, then they add a high amount of TCP that favours the pearls formation and finally they add a solvent to preserve stirrer functionality and decrease temperature. This short explanation highlights the dump phenomenon and its considerable consequences. Some batches can present process parameters that highlight possible instability: pearls grow up too fast, jacket temperature decreases, in order to catch more heat from the process, etc.. Evidences of an incoming instability suggest the addition of TCP; a small amount of stabilizer might decrease the spheres dimension to the desired size and recover a perfect suspensions. For these reasons, TCP supplementary addition will be related to an instable production, during data analysis. After careful evaluation with plant managers and operators, the observations have been divided in four groups according to the amount of suspending agent second addition:

- No TCP
- Minor than 80 litres of TCP solution
- Between 250 and 80 litres of TCP solution
- Above 250 litres of TCP

These clusters improved the readability of results in both exploratory and discrimination analysis and have been used to perform PLS-DA models.

3.3 Data structure

Data Analysis focused on three data sets taken from the ST14 database and on two sub sets of data taken from one of previous mentioned ones. All the data refer to the period from May 2011 to February 2013. In the following description, the number of time points is not indicated because depends on the data analysis aim, i.e. to if the whole batch or only the first reaction phase was considered.

In the following is reported a brief description of the five analysed data sets, with the name they will be referred to in the rest of the Chapter:

- **R401A Batch (177 Batches x 10 Variables x Time points)**
  The data coming from reactor A are in a 3D array where the three modes correspond to: the observation time point, the different sensors and the batches, Fig.3.2, where each horizontal slab is a batch.
  The objective of multivariate data analysis for this data set is batch comparison.
  The process sensors in R401A are monitoring the following parameters:
    - Two temperature values inside the reactor
    - Jacket temperature
    - Stirrer rotation per minute
    - Motor stirrer absorbance
    - Pressure inside reactor
    - Pressure control inside reactor
    - Pressure stirrer oil

  These eight variables are the ones involved in plant control.

- **Warping information (177 Batches x Reference time points)**
  A batch production suffers of alignment problem: despite the repetition of procedure, each batch takes different time to be discharged both globally and along single phases. In order to compare batches thus an alignment is mandatory to make them congruent. This is also the only way to confront batch trajectories, observation shall have the same length in order to compare them and mostly important the corresponding time points should represent the same "time" of reaction. The alignment process, if done by warping methods, returns also a vector that describes how each batch has been aligned on the reference one, i.e. the movement of time observation to match the same reaction points/phases This is an important information and should be included in data analysis, e.g. in block PCA and block PLS-DA [9].

- **Batch trajectories features (177 Batches x 27 Variables)**
  One way to manage 3D batch structure can be to resume the batch trajectories in their salient characteristics, e.g.: maximum and minimum values, slopes of specific phases, standard deviation, etc. This manual process takes long time but frequently returns interesting results. Briefly are reported the trajectory features used in this application:

  - *Batch time group*: First phase duration, second phase duration, starting point of the first phase, starting point of the second phase
  - *Reactor temperature*: Mean temperature of the first phase, mean temperature of the second phase
  - *Jacket temperature*: Jacket max temperature during first heating, time point of the max temperature during first heating, first heating duration, jacket max temperature during second heating, time point of the max temperature during second heating, second heating duration
  - *RPM stirrer*: Rpm mean value of the first phase, rpm standard deviation of the first phase
  - *Absorbance stirrer*: Mean absorbance of the first phase, standard deviation absorbance of the first phase, final point absorbance of the first phase, mean absorbance of the second phase, standard deviation absorbance of the second phase, final point absorbance second of the phase, interpolation of the first phase stirrer absorbance
  - *Reactor pressure*: Max pressure of the first phase, max pressure during batch
  - *Pressure stirrer oil*: Mean oil pressure of the first phase, standard deviation oil pressure of the first phase, mean oil pressure of the second phase, standard deviation oil pressure of the second phase, interpolation of the first phase oil pressure

- **Formulation (2305 Obs. x 20 Variables)**
  Each EPS grades have many additives. The concentration gives information on additives influence and on the dependence of products from them. The formulation data matrix represents the additives concentration but includes also the powder involved in process stabilization such as soap and suspending agent among the others.

- **Water plant treatment (177 batches x 40 Variables)**
  In many productions, water consists in a utility fluid and does not affect the results of reaction; in EPS suspension, it should be treated as a raw material. Water quality influences, for instance, soap formation and suspension stability. The matrix of water plant treatment collect the water quality parameters and takes also into account some other fundamental parameters that operate the water treatment plant.

Multivariate data analysis on ST14 production considered all the matrices mentioned above. Reactor sensors and warping information matrices include the time point direction, and the length is function of the time frequency. The frequency considered for assembling the data is two minutes. Higher frequencies return an enormous quantity of data that poses memory problem for data elaboration on the actual available PC and on the other hand may be too much reflecting instant spikes and other spurious effects. Moreover, if a higher than 2 minutes frequency is considered, many values are missing and would have been interpolated from Versalis database; in fact, with the aim of space saving, company database stores only a part of the acquired data following rules imposed by database administrator, while the remaining are deleted. In any case, database allows querying of any frequencies and in case of missing data it returns an interpolation of real data acquired. Then I decided to extract data every 120 seconds.

3.4 Batch methods

3.4.1 Issues about preprocessing in batch analysis

A three-way array can be pre-processed in different ways and the choice of preprocessing influences the results. The main reason of scaling is to make comparable variables that have very different scales, e.g. on-line measurement and laboratory analysis, with different ranges and/or intensity. In batch analysis, things are complicated by the fact that preprocessing and unfolding procedures cannot be considered independently. In a 3D array, two different main approaches might be selected for the scaling and centring [10].

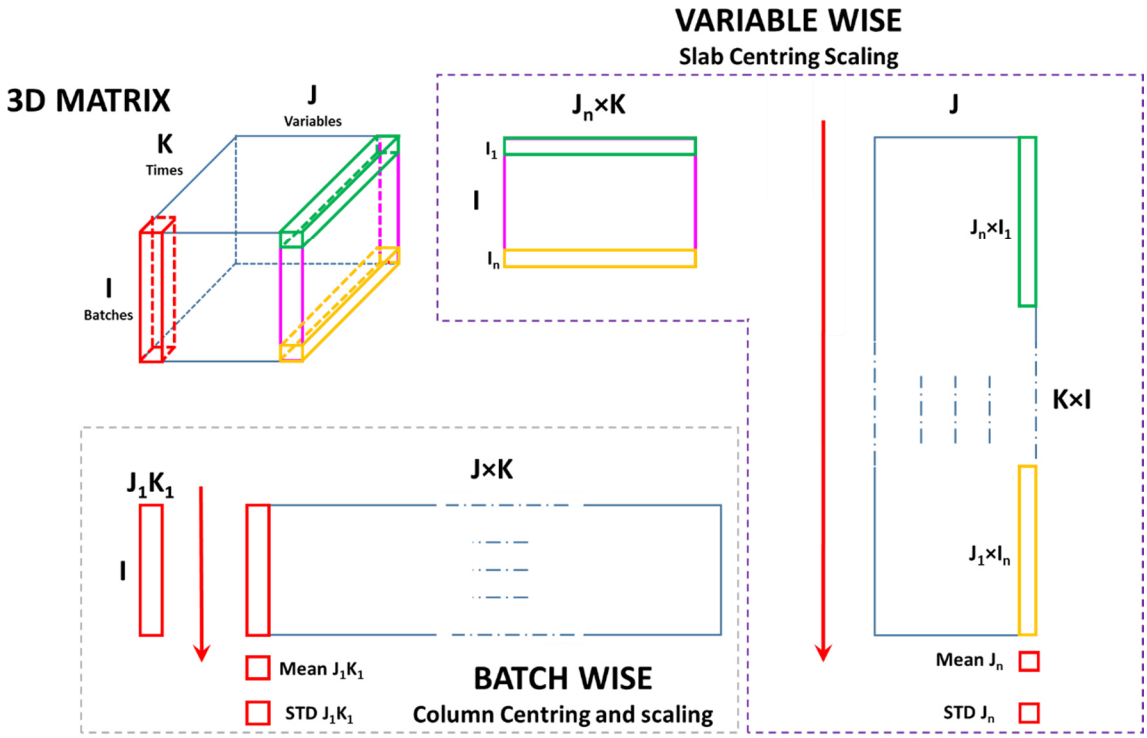

Figure 3.2 Centering and Scaling batch data. Column centering and scaling refers to batch-wise (BW) unfolding and correspond to "autoscaling" of a single time point for each variables across all batches (red arrow in left bottom sketch). Slab centering and scaling refers to variables-wise (VW) unfolding and corresponds to "autoscaling" a variable across all time points and batches (red arrow in right sketch)

The first one, the slab centring and scaling, considers the whole trajectory (in time) of a variable. It results in a translation of the trajectories around the global mean and in a successive normalization by standard deviation. This preprocessing is equal to Autoscaling (Chapter 2.1) on the variable wise unfolded matrix. Trajectories keep their original slope but values are comparable between variables. Equation 3.1 shows the calculation of mean centred and scaled data:

$$x_{i,j,k\ slab} = \frac{\left(x_{i,j,k} - \bar{x}_j\right)}{\sqrt{\frac{\sum\left(x_{i,j,k} - \bar{x}_j\right)^2}{n-1}}} \tag{3.1}$$

Where, $x_{i,j,k}$ is a value of the batch *i* for the *j* variable at the *k* time point and $\bar{x}_j$the mean value of variable *j* for all available batches and time points. Value *n* is equal to the time points amount, for all batches.

The other approach considers a single column of a variable, the same time point for all available batches, to perform the centring and scaling. In this way, trajectories are transformed and their profiles completely change but preprocessing has taken into account the relation between trajectory time points. This preprocessing corresponds to Autoscaling of the batch wise unfolded matrix. The second method, not the first one, removes non-linearity between time points.

$$x_{i,j,k\ Column} = \frac{\left(x_{i,j,k} - \bar{x}_{jk}\right)}{\sqrt{\frac{\sum\left(x_{i,j,k} - \bar{x}_{jk}\right)^2}{n-1}}} \tag{3.2}$$

Where $\bar{x}_{jk}$is the mean value of variable *j* for the k time point. Value *n* is equal to the batches number. In variables wise unfolding centring and scaling might be applied to the single batch block; it could be useful in order to observe the variation between batches concerning the mean points of variable trajectories. Batch alignment is part of batch preprocessing and, for its importance and complexity, is discussed into a separate section, the next one. Anyway, alignment is not always necessary; in the case of variable wise unfolding, in which time point are evaluated as single observation, alignment is not necessary to perform decomposition of the data, e.g. by PCA, however if the PCA scores obtained are then refolded in 3D, alignment matter, because if it was not applied some odd interpolation, modifying the original data structure, can take place at this stage. Also in the case of feature trajectory, in which instead of the entire variables profiles only some key features are considered, alignment is unnecessary, as far as the correspondence in time/event of the points used to define the features is carefully checked.

3.4.2 Batch alignment

In the ideal batch production a batch always takes the same time, follow the same trajectory and realize a one single product of fixed quality. Actually, the production reality is far from the ideal situation, sometimes very far away from it. Batch process steps are frequently not best described using time as reference measure but, for instance, the conversion ratio can be more suitable to establish correspondence among different batches, or to match production phases. Moreover, slightly differences in formulation might returns considerable duration changes. These situations and many others give datasets with different sizes. Moreover, batches might have the same duration but it does not mean that alignment is not necessary: in order to compare productions, the phases have to overlap, at least in the starting and ending points, more than the time values. Same length might depend from scheduled time but it does not necessarily consists of synchronized production. For Data analysis to be effective and not misleading a proper aligned dataset is needed, in which the batches evolution can be correctly compared.

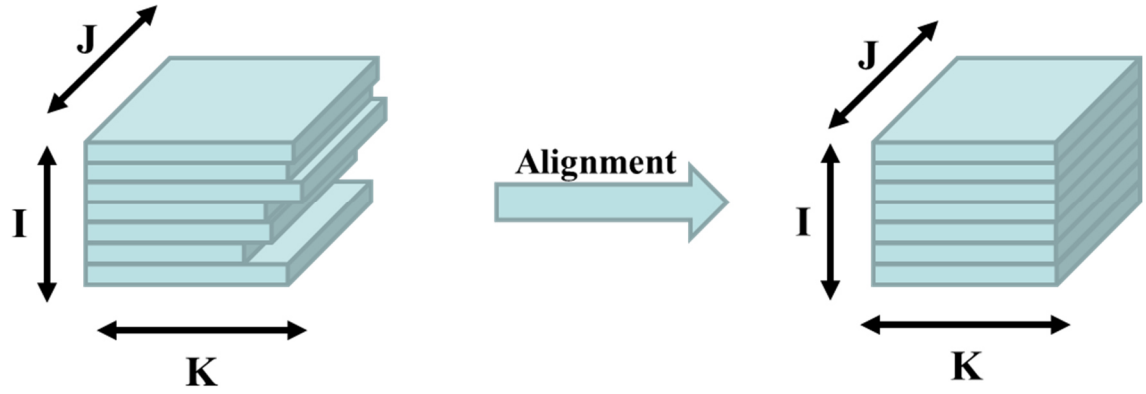

Figure 3.3 Batches before and after alignment procedure. *I* defines the number of the batches, *J* corresponds to the variables and *K* is the times point direction

Different alignment procedures allow a suitable data transformation. The PLS-Toolbox for MATLAB developed by Eigenvector [11] has implemented the linear alignment and the correlation optimization warping (COW) [12]. The linear alignment might work for simple cases in which the asynchronism consists of different duration and of a single step production. In linear transformation, data are stretched or compressed such a single string, in a linear way. User selects a single variable that direct the alignment; algorithms operate in order to make as similar as possible the trajectories. This alignment method does not modify the values but shifts, replies or deletes some of them. This point is common to all the methods here mentioned. COW was introduced by Nielsen et al. [13] to fit the misalignment in discrete data signals.

The main idea is to match each batch to a reference signal, one of the dataset vector or the mean vector, modifying the entire batch trajectory segment by segment. The so-called slack parameters limits the segment modification, how many points is possible to add or remove in each segment. When the number of time-points differs from the reference signal, linear interpolation is used to reach the same length. COW procedure implies the division of sample vector in *I* block of *i* length and then one or more points are added or removed; the quantity depends to the slack parameters, anyway algorithms aims to maximize the correlation between the two segments.

The result is a batch matrix in which observations have the same length of the reference vector; quality of alignment might change with the two parameters segment length and slack. To manage process misalignment, dynamic time warping (DTW) is a widely applied method. DTW warps in nonlinear way two trajectories, in such a way that events are aligned and the distance between them minimized. The algorithm was presented by Sakoe and Chiba [14]; it found application also in batch process monitoring. The dynamic time warping algorithm assigns to each data point of the reference batch a point of the batch to be aligned (test batch) meanwhile not all data points of test batch must be coupled with a point in reference batch. It implies that every test batch will have the same length after alignment, equal to the reference batch, but an independent warping transformation has been applied. Global problem might be defined as:

$$argminD(\boldsymbol{f}) = \frac{\sum_{k=1}^{k} d_{rs}[m(k), n(k)]w(k)}{\sum_{k=1}^{k} w(k)} \tag{3.3}$$

Where d_{rs} is the distance between reference *m* and sample *n*. Such distance must be minimized. The weight is useful to remove bias but not mandatory for the warping. As in linear and in COW algorithms, DTW returns a matrix in which batches have the same time points and the main phases aligned in a suitable way. Furthermore, dynamic time warping returns the warping info. It consists in a couple of vectors; one indicates the reference time points and the other the correspondent sample time points. Every batch has a descriptor with different length; in order to include this information in data analysis another transformation is necessary. As well laid out by J.M. González-Martínez [15], a simple transformation solves this misalignment and makes feasible the warping information. For each reference time point, only the maximum sample time point is selected. So, if a reference time point refers to two or more sample time points only the biggest will take part into the warping information vector.

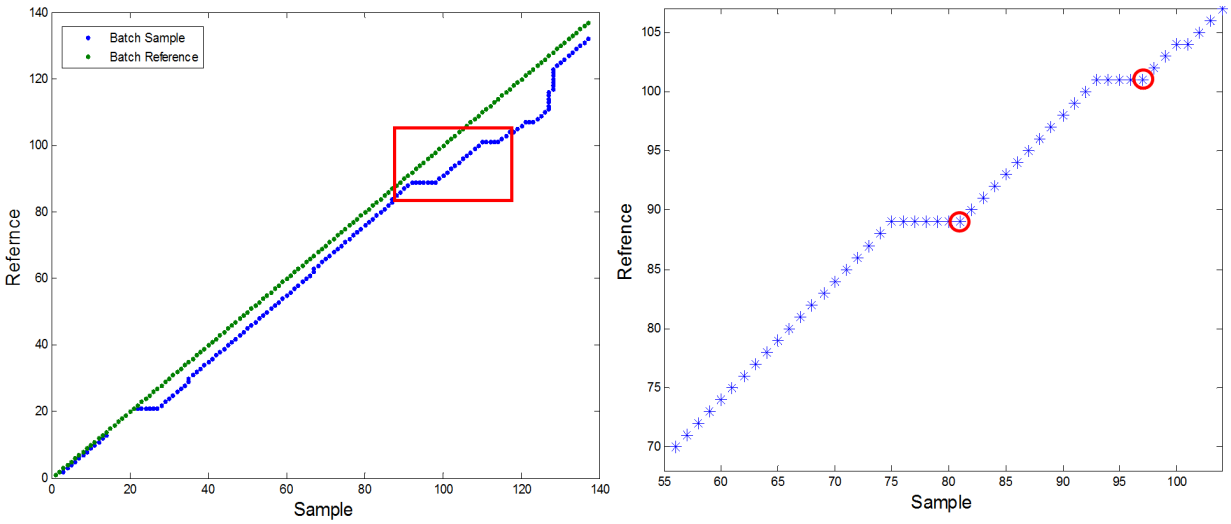

Figure 3.4 Example of a DTW alignment, on the left the sample (blue) and the reference (green) trajectories, on the right an example on how to use time point for the warping information

The ST14 application shows data that contain various kind of asynchronisms. The most frequent is the delay in starting of the batch trajectory but also the misalignment between phases often appears; permanence at 90 and 120°C takes longer as a function of conversion degree.

Because of these varying situations, the recently proposed method developed by José M. González-Martínez in collaboration with Onno E. de Noord and Alberto Ferrer, with whom I had the pleasure to work, has been appied. The method is called *Multisynchro* [16] and has the advantage to assess the kind of asynchronisms present from batch to batch and to selects the suitable procedure for each misalignment. Asynchronisms are separated into four classes [17]: batches with equal duration but with key events misaligned (Class I), batches interfered by external factors that generate different duration (Class II), incomplete batches (Class III) and batches that differ for the starting point but with the same evolution (Class IV). An algorithm detects the asynchronism type via DTW profiles in which the profile changes in relation to the asynchronisms. In the low-level routine, Multisyncro applies the specific synchronization procedure that is based on DTW but fits the specific problem. All the algorithms were developed by the authors in MATLAB™ and a GUI simplify the method application. Further details are available in the reference.

3.4.3 Batch matrix unfolding

Alignment process returns a three-way data matrix in which every batch has the same number of row, time points, and columns, variables. A three dimensions dataset might be evaluated with algorithms that can manage more than two dimensions such as the Parallel Factor Analysis (PARAFAC) [18] and the Tucker model [19]. These methods preserve the data structure and extract information in the three directions at the same times and may have some benefits in terms of sample correlation in both directions. While PARAFAC has been widely used in second order calibration the process monitoring applications in which is used are much less. A discussion about the most suitable approach is beyond the scope of this Thesis, just to mention, often the problem is that process data do not fulfil trilinearity assumptions. The most common procedures for batch analysis uses unfolding [20 21 22].

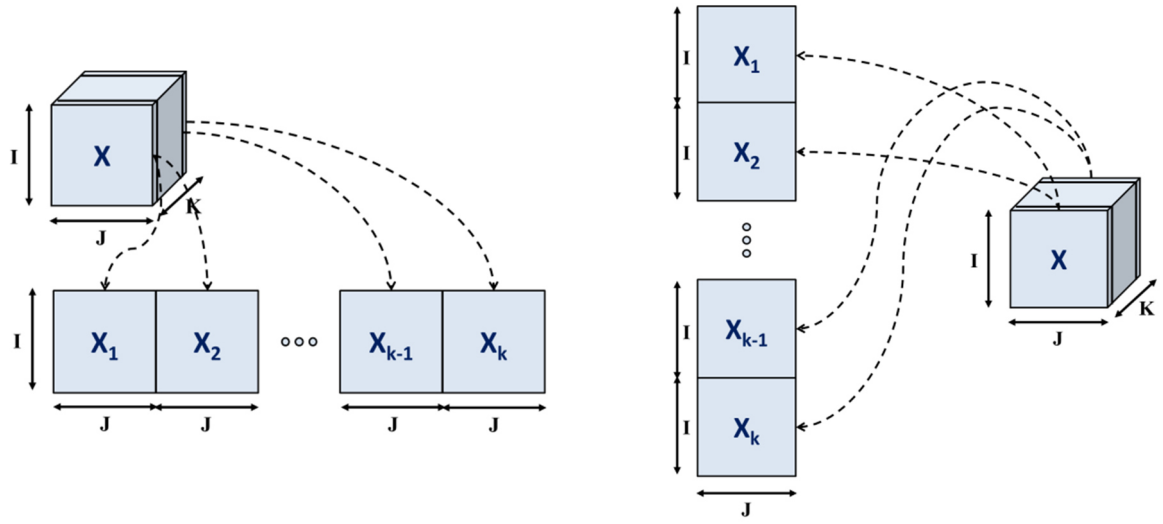

Figure 3.5 Different arrangements of three-way batch data in two-way forms. On the left, batch-wise unfolding and, on the right, variable-wise unfolding

Unfolding consists into the separation and reorganization of every batch sheets. Single production dataset has a defined number of variable and times point that, after the alignment, become equal through all batches. In Fig. 3.5 are shown the two main approaches used for unfolding: Variable-Wise unfolding (VW-Unfolding) and Batch Wise Unfolding (BW-Unfolding). In VW-Unfolding, the single batch slab (time points x variables) are put one below the other, hence the number of variables is preserved and the number of rows become equal to the product of the number of batches per time points. This unfolding procedure returns a "fat-matrix" in rows direction, i.e. the number of observations largely exceeds the number of variables. In BW-Unfolding, the single batch slab are put one besides the other, hence the number of rows is equal to the number of batches, and the number of columns become equal to the product of the number of variables per time points.

This unfolding returns a “fat-matrix” in width directions. A single column, in BW-Unfolding, corresponds to the value of a process sensor at a determined time point.

Ferrer et al. [20] have discussed as VW and BW Unfolding are extreme cases and that it may be worth to consider as well an unfolding approach that uses both. In particular,VW-Unfolding does not take into account the dynamic of the process, i.e. the possibility that during batch evolution the correlation structure of process sensors may change. In other words that specific sensors may show correlation (or not) at different time points. In fact, the way data are arranged implies that there is a correspondence (in a row) of each time point for all the sensors, thus only the instantaneous cross-covariance among sensor is considered. Moreover, the latent variables models are based on variance calculated throughout the whole time points for each sensor. Thus, the loadings plot will show correlation patterns among sensors that are based on average effect along batch time evolution. VW-unfolding has the advantage of low uncertainty in estimating the LV models given the huge number of observations, and it is the suitable for stable batch processes, where dynamic is not relevant. In other terms, the sensors correlation structure has to be constant to achieve proper results with VW. Another assumed advantage of this unfolding is that batches need not to be aligned prior to PCA, PLS analysis. However, not aligning batches will have drawback in data pre and post processing.

BW-Unfolding needs batch alignment (each time point, being a column, needs to correspond in each batch) and returns a single row for each batch. The main disadvantage is that models are estimated with very low number of observation compared to huge number of variables. However, the main advantage in BW consists of possibility to catch the dynamics of the process. In fact, correlation among sensors might change during batch evolution and these changes are modelled since every time point for a specific sensor can be linked to any time point of other sensors (each column is a time point). BW allows the exploitation of batches trajectories for each sensor and showing how they correlate during batch evolution with themselves and with the other sensors. This information might be fundamental to discover difference in process, to highlight the time points in which sensors correlation structure changes, and to highlight production trouble of limited duration. In the following application, I applied both variable wise and batch wise approaches and compare the retrieved information. However, EPS production features suggest that BW unfolding could be the most suitable one.

3.4.4 Multi-Block analysis

Continuous and discontinuous production often needs the addition of some ingredients, powder or liquids, to improve the quality of final products. Additives can drastically change the behaviour of materials and could determine the success of a material. In continuous production, flow meters indicate how much of an additive is added into the main stream. In batch process is quite different; frequent manual operation and discontinuous steps make difficult to recover the concentration level from the trend of process sensors data. Additive weights are thus stored in another way, maybe in a excel sheet. This example is one of the many situation in which information comes from different sources but concerns the same process and must be integrated in order to not lose information. Another situation is a multistep process in which the different sensors and phases have to be linked to the final product. In general, all different kind of data, e.g. formulation data and sensors data, or sensors data for different phases, and quality data of intermediate or of the final product, constitute a data block from the data analysis point of view. Different block can be combined in data analysis if they share one dimension, e.g. the same samples or the same variables, or same time points. In batch data, usually the shared dimension holds batches.

Blocks have their meaning and importance, data treatments and/or proper models preserve such information and help the results understanding. Literature offers different methods [9] to manage more than one matrix at the same time in order to extract from each one as much information as possible.

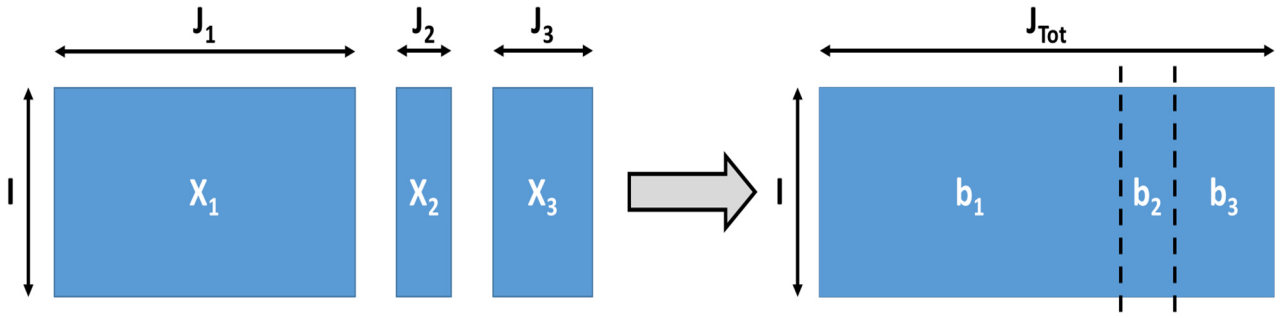

Figure 3.6 Multiblock scheme for three matrices. Number of samples *I* is the same in each block. The resulting new matrix will have a number of variables, J_{Tot}, equal to the sum of J_1, J_2 and J_3

In the batch application, I used a scaling factory that makes variables comparable and block variance equal and independent from the number of variables in each block (block autoscaling).

$$X = [X_{B1}, X_{B2}, \dots, X_{Bk-1}, X_{Bk}] \tag{3.4}$$

$$X_{Bk,yi} = \frac{\left(X_{Bk,yi} - \overline{X_{Bk,y}}\right)}{SD\left(X_{Bk,y}\right)} * \frac{1}{\sqrt[2]{N_{VarBk}}} \tag{3.5}$$

Where **X** data concatenates some *Bk* matrices, where each one represents a block, for instance *Bk* can be the formulation or the water treatment conditions. Matrices must have the same number of rows, and the resulting **X** has as number of columns equal to the sum of the number of variables of each block. Similarly to autoscaling, block-scaling subtracts from every column, $X_{Bk,yi}$, the mean value,$\overline{X_{Bk,y}}$, and then divides it for the column standard deviation $SD\left(X_{Bk,y}\right)$. Blockscaling differs because of the second part of eq.3.5: i.e. data are then scaled in relation to the size of block (the number of variables in a block N_{VarBk}).

Even if the number of columns is quite different, this scaling makes block to have equal variance. In a multi block model, the significance of each single block is important to assess and the Block Importance in the Prediction (BIP) index permits a fast and proper evaluation [23]. BIP calculation is analogous to VIP (Chapter 2.5.1):

$$BIP_k = \sqrt{\frac{\sum_{f=1}^{F} w_{kf}^2 * SSY_f * j}{SSY_{total} * f}} \tag{3.6}$$

Where w_{kf} is the weight of block *k* in the *f* principal component, SSY_f is the explained variance of the **Y** by the *f-th* component and *j* is the number of blocks. Denominator contains SSY_{total} that is the total explained variance of the dependent variable and *f* that corresponds to the number of components.

3.5 Exploratory data analysis

The five data sets potentially contain a lot of information on ST14 EPS production. Explorative data analysis (Principal Component Analysis) is useful to extract all the possible knowledge on this batch production.

3.5.1 Formulation data

Formulation data, could give an overview of the different polystyrene grades and on their dependence from additives. Data were autoscaled prior to PCA. A three components model explains 65% of the total data variance. Some clusters are observable in the scores plot, after the exclusion of few observations for which were not available information on reactor and product type, and of a batch whit high leverage (it is the only one without a specific additive). The PC1 vs PC2 score plots, Fig. 3.7, shows batches coloured according to the reactors (A , B or C) in which EPS has been produced. Reactors do not cluster, therefore, in relation to the formulation, expandable polystyrene polymerization does not depend on reactor. The picture, on the right side, highlights the observed clusters: each one defines quite well a product. The LN grade, black dots, and the AE, orange dots, differs for self-extinguish attitude of the second one. EH batches appear similar to the LN grade and separation depends on wax and other specific EH additives. Separation between AE and VERDI is not clear and concern mainly the expanding agent quantity.

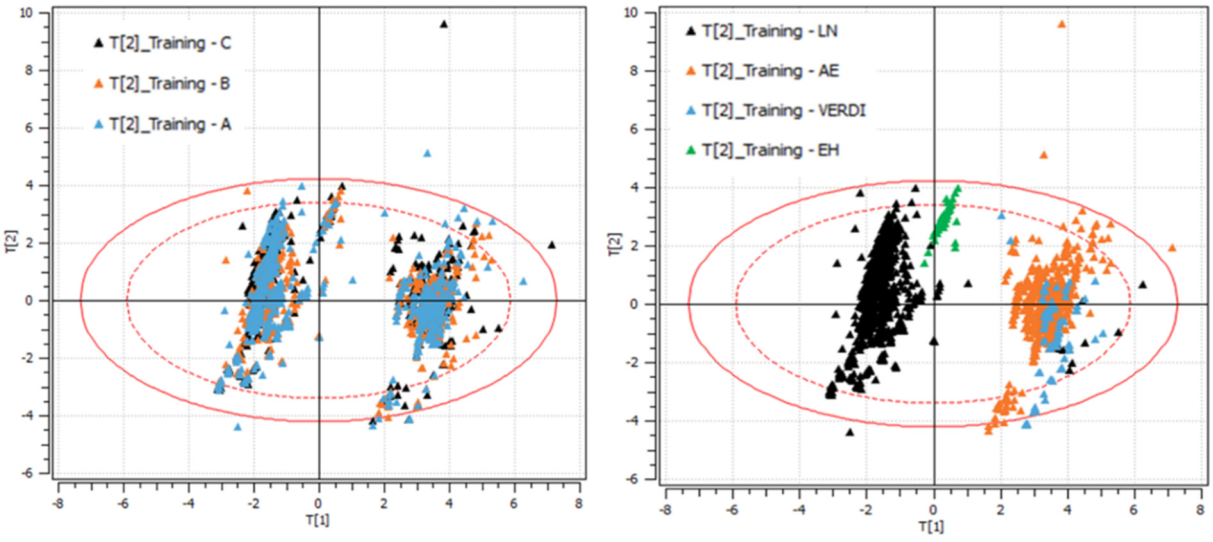

Figure 3.7 Score plot PC1 vs PC2 PCA on formulation dataset. Dots colour relates to reactor (left) and to production (right). Red dashed and continuous lines are respectively the confidence limits at 95% and 99%

Having established the similarity of reactors and the cluster with respect to produced materials, data analysis considered a single reactor dataset, the reactor R401A and the self-extinguish material.

Considering only one EPS grade is necessary to exclude the variability due to reaction set up that depends on the different formulations used. Thus, analysis focused on variation inside the AE EPS.

3.5.2 R401A Batch

Data from Versalis database (PI System) were organized variable wise. It means that matrix is variable oriented and the number of variables is equal to the number of considered sensors. The number of observations corresponds to the sum of time point for every batch; it is equal to the product between points and batches number. For each sensors, batches follow a trajectory that is interesting to recover and compare with multivariate analysis. If some differences between batches exist, they should appear also into single sensors trajectories.

Concerning of the worst production situation, the dump batch trajectories, for several sensors, are totally different from the rest of production. At certain time point, plant operators added a high quantity of TCP to avoid styrene polymerization and added solvent to preserve stirrer functionality and to decrease the viscosity. The three sensors profiles of bath, jacket temperature and pressure, represented by blue lines, show the end of reaction. In chronological order controller stopped heating, decreased pressure and the mass of reaction cool down. This batch will be considered in discrimination analysis in order to evaluate if dump productions are predictable.

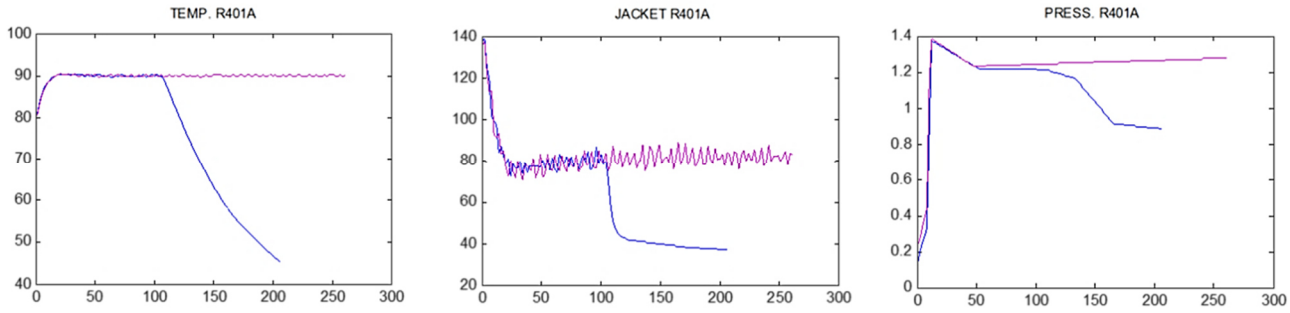

Figure 3.8 Comparison of good (purple) and dump batch trajectory (blue) for reactor temperature, jacket temperature and pressure inside reactor

Considering the other batches, around two hundreds, they are too much to detect variation in trajectories comparison. To facilitate the readability of the plots the batches are coloured, Fig. 3.9, by TCP amount added: for the first groups (blue) the second TCP addition was not performed, it means that the production perfectly reach the second heating phase, from the second to the fourth cluster the second TCP addition took place with increasing concentration. These distinctions are not clearly observable in the figure so the groups mean trajectories were considered, in that way only four profiles have to be compared. Temperature plots are quite similar and both return the same information: batches with less suspending agent addition have longer delay and the first heating phase begins later than the ones with higher TCP. Jacket temperature shows slight variation between groups but blue lines are lower than the others are so it means that "no addition" cluster seems linked to a lower peak in jacket profiles. RPM trajectories are similar and the values along time do not return peculiarities, maybe the black line, highest TCP concentration, is more scattered. This last difference concerns reactor pressure in which higher suspending agent comports higher pressure.

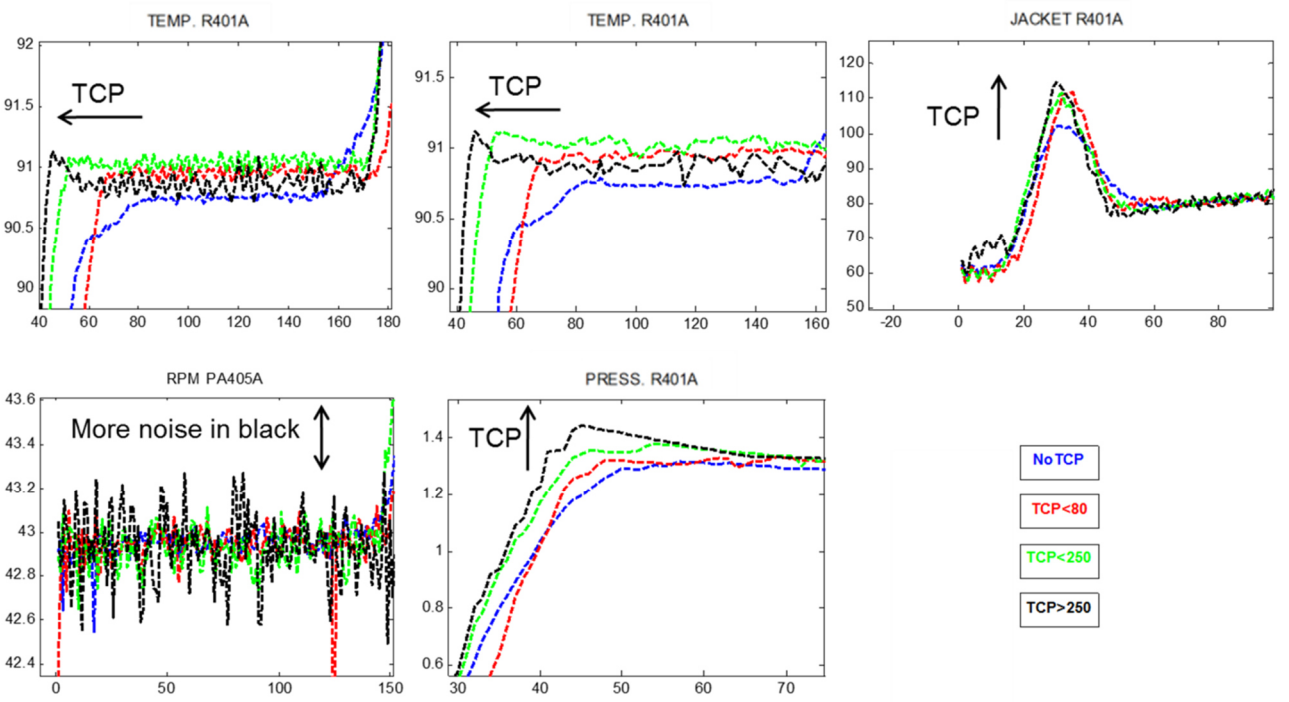

Figure 3.9 Mean trajectories of significant raw data. Lines colour relates to the TCP amount added to recover instability. Arrows suggest TCP effects

These observations come from the raw data, without pretreatment. The alignment procedures adjusted the data and removed, for examples, the differences in temperatures profiles; it confirms that warping information should be retained in batch comparison.

Next, PCA on the same dataset, VW-unfolded, with the exclusion of dump batches data has been done. A two components model describes more than 90% of variation. The scores plot shows batch trajectories, the data were centred and not scaled in order to appreciate better batches trajectories.

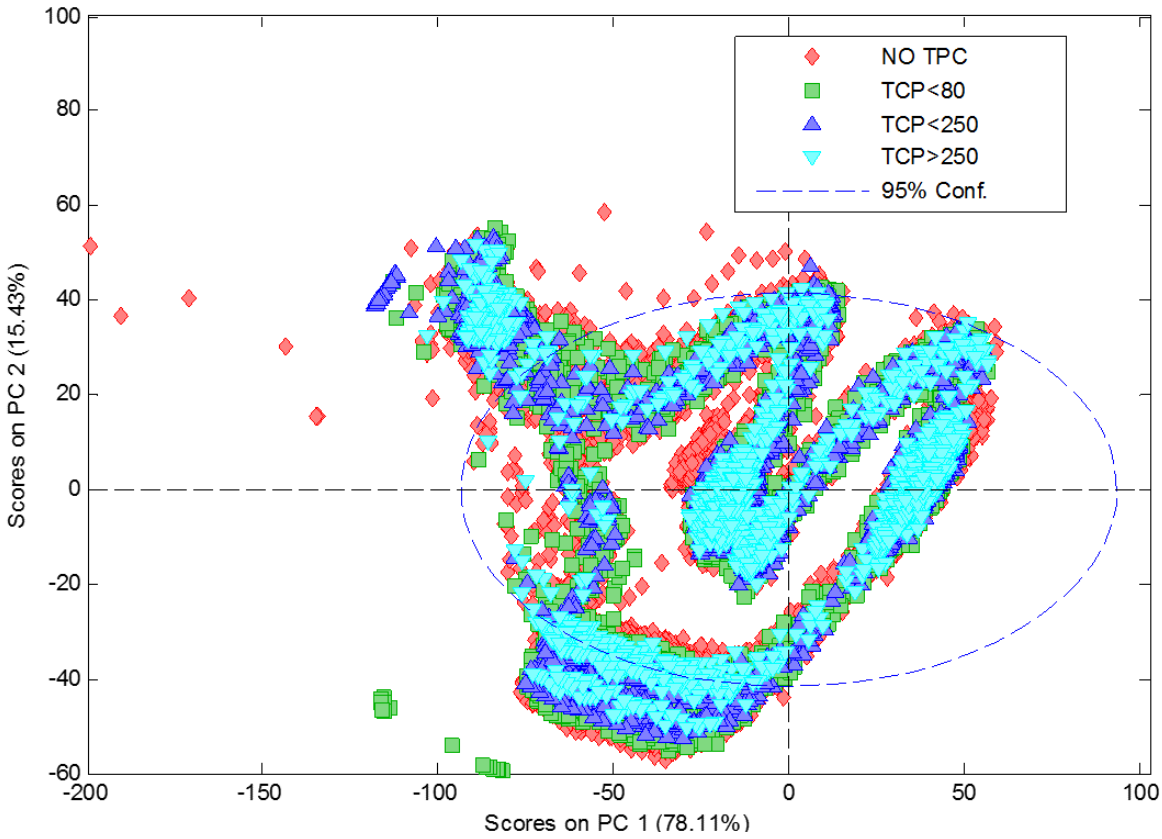

Figure 3.10 Score plot PC1 vs. PC2, PCA on R401A Batch data variable wise unfolding. Dots colour relates to the TCP amount added to recover instability

A misalignment of the batches, especially in the first part of the reaction, could be observed in the fourth quadrant: the width of the first phase shows how much batches differs from each other. Profiles show the five temperature steps that start from the fourth quadrant and move toward the third with five segments. This model might be interesting to verify the process timing or to detect a strong outlayer but is not satisfactory for a process monitoring. Variation among good batches covers the hypothetic variation related to the bad batches.

The loading plot into the next picture, figure 3.11, supports the user in variables relationship understanding. It shows two behaviours: the first consist into reaction temperatures that vary together along the first component as the pressure does; the second mainly consists into jacket variation that change a lot and opposite to the reactor pressure. The tree stirrer variables act equally and are not relevant; they are too close to the origin. So, the separation between variables mainly belongs to the first principal component. Into the second latent variables only reactor pressure and jacket have a high values.

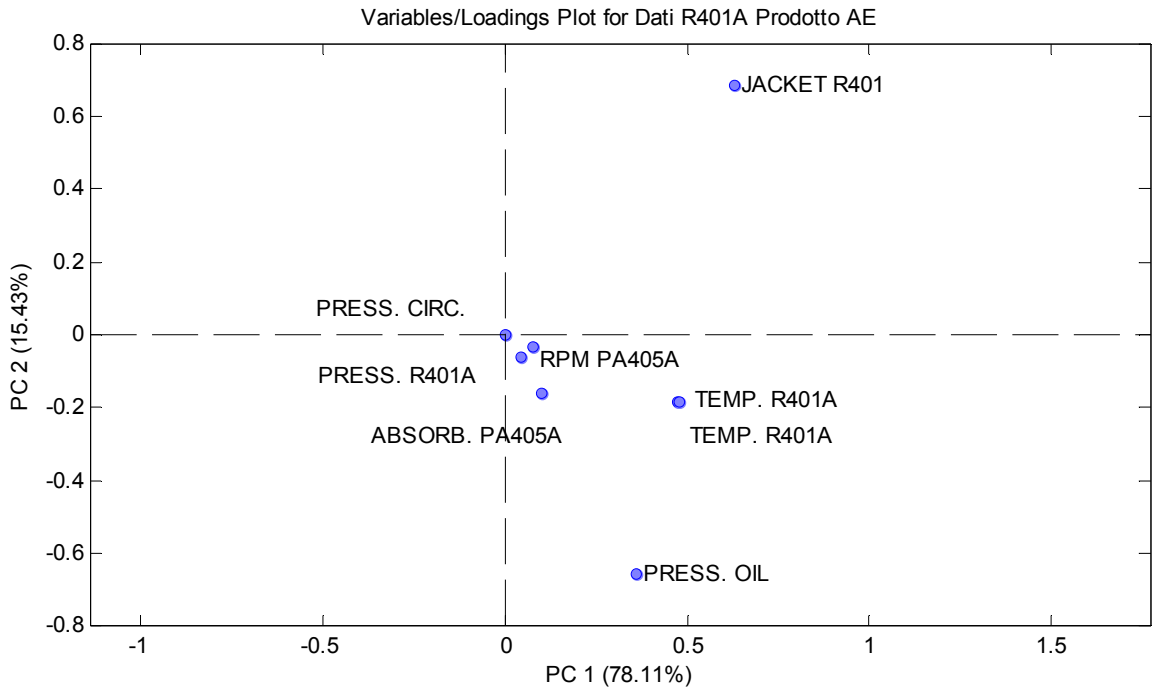

Figure 3.11 Loadings plot PC1 vs. PC2, PCA on R401A Batch data variable wise unfolding

A variable wise unfolding allows a comparison between whole trajectories and ignores the batch dynamic. It could be useful for problem solving in which bad effects are along all production and do not affect variables at different time.

In order to evaluate dynamic process data must be unfolded in batch wise direction. A PCA model with five components explains around 50% of data variance; it seems a low quantity of information but such a "fat" in width, dataset contains a lot of noise and would needs a lot of batches to decrease it.

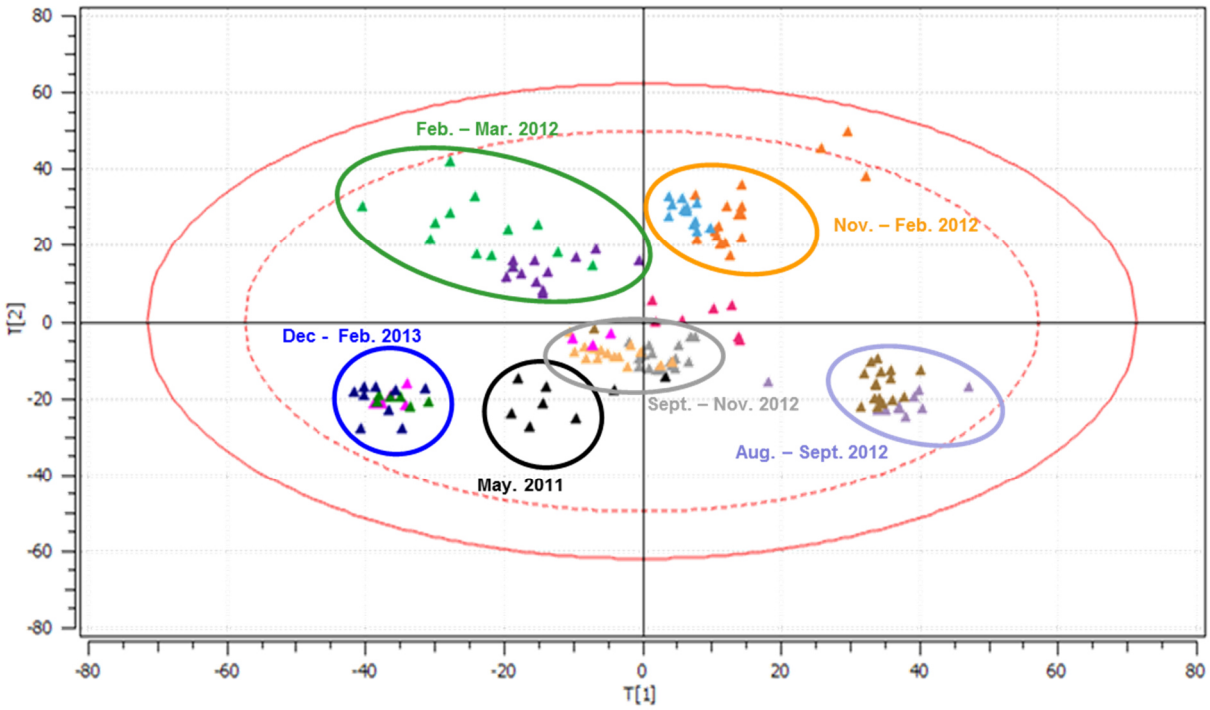

Figure 3.12 Score plot PC1 vs PC2, PCA on R401A Batch wise unfolding, triangles colour relates to production months. Ellipses indicate production range. Red dashed and continuous lines are respectively the confidence limits at 95% and 99%

In the PC1 vs PC2 score plot, 19% of explained variance, are visible some clusters related to specific production months. The first principal component describes difference related to the reaction temperature, as can be seen by the loadings plots. Temperatures R401A have the higher loadings values on PC1. PC2 describes again a phenomena related to the batches temperature but only the first stable phase (first part of time points of Temp. R401A) defines this effect. Loop pressure and RPM value are important variables mainly in the second reaction step and it could be appreciated in the second component. Also the controlled pressure of stirrer oil has an effect during whole suspension process.

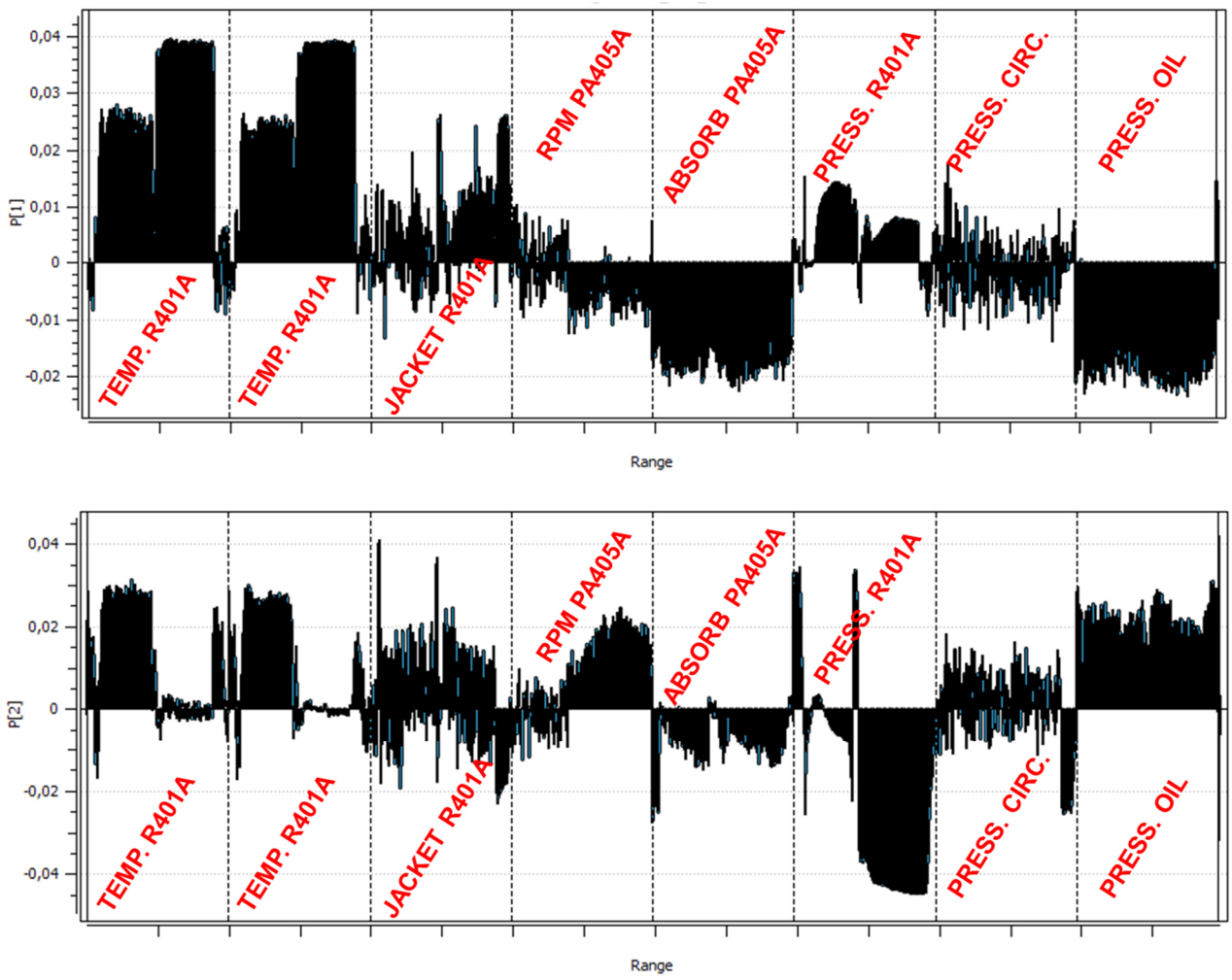

Figure 3.13 Loadings plot PC1 and PC2, PCA on R401A Batch wise unfolding

Taking into account both score and loading plots, batches appear separated due to the production period but not strongly related to the external temperature. From left to right production happens at higher temperature and from top to down with higher pressure during the 120°C permanence. In order to discover as much as possible on instability, I coloured observation in relation to the suspending agent added during reaction but it did not returns evidences.

The scores plot, figure 3.14, in which batches are coloured in relation to the initial TCP addition, shows a trend in scores plot. Initial suspending agent concentration grows up in both components and it suggests that when temperatures are higher more TCP is needed, and lower pressure during the second stable phase is observed from raw profiles; probably plant managers increase temperature and suspending agent in order to prevent dumps and it reflect in a second phase with less pressure. As previously stated, instability mainly regards the reaction before the second heating phase, during the permanence at 90°C. For that reason, batches were split into this two phases and only the first part, the so-called first reaction phase, have been aligned. These data were subsequently BW-unfolded and then analysed by PCA; this model explains 40% of variance with 5PC's. The covariance map supports this choice: figure 3.14 shows two completely independent zones, the two phases does not appear correlated or maybe the correlation inside the phase is greater than the one between phases. Covariance matrix might help process comprehension; a production in which phases are strictly correlated shows positive or negative covariance, following the direct or inverse correlation, due to an influence of the past action in the behaviour of the newer ones. Map in figure 3.14 clearly describe two phases extremely correlated inside them, red squares, but with poor influenced from the other, grey squares.

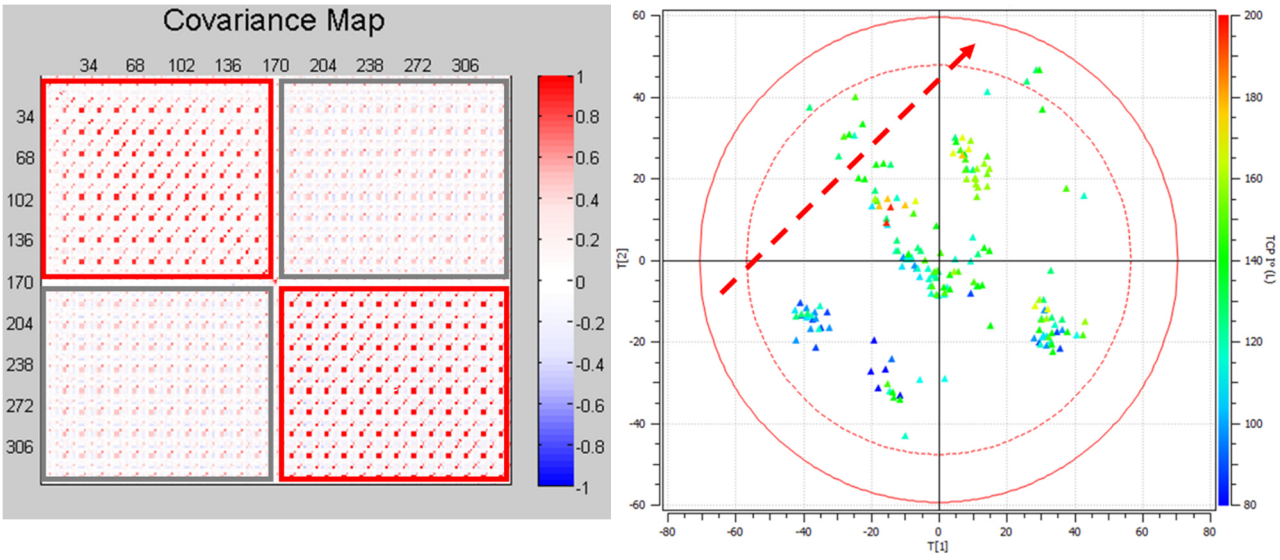

Figure 3.14 Alignment covariance Map (left). Dots colour refers to the direct (red) and indirect (blue) correlation. white means no correlation. Grey and red squares highlight uncorrelated and correlated zone. Score plot PC1 vs PC2, PCA on R401A Batch BW unfolding, first phase alignment. Data colours relates to the TCP initial concentration. Red dashed and continuous lines are respectively the confidence limits at 95% and 99%. Red arrow indicates the direction in which TCP augment

Separation does not change among the batches and the clusters seem the same of the previous PCA so the behaviour of the first phase seems the same of the global one. The T^2 and Q plots have also been checked in order to detect eventual high values for observations, which could correspond to high instability (assuming that high TCP addition is always and strictly linked to this issue), but no relation to the TCP quantity appears. The batches T^2 and Q values do not seem to depend to the instability but from production variability. Anyway, exploratory data aims at discovering as much information as possible form the available datasets and not at the discrimination of the batches. Looking further components, the second principal component, figure 3.15, describes a seasonal variation. Such phenomenon derives from the stirrer features, absorbance and oil pressure, and from the pressure. It means that lower temperature makes agitation difficult and, less intuitive, the pressure reach higher values. Even so, variation follows external effect and suffers of the temperature changes.

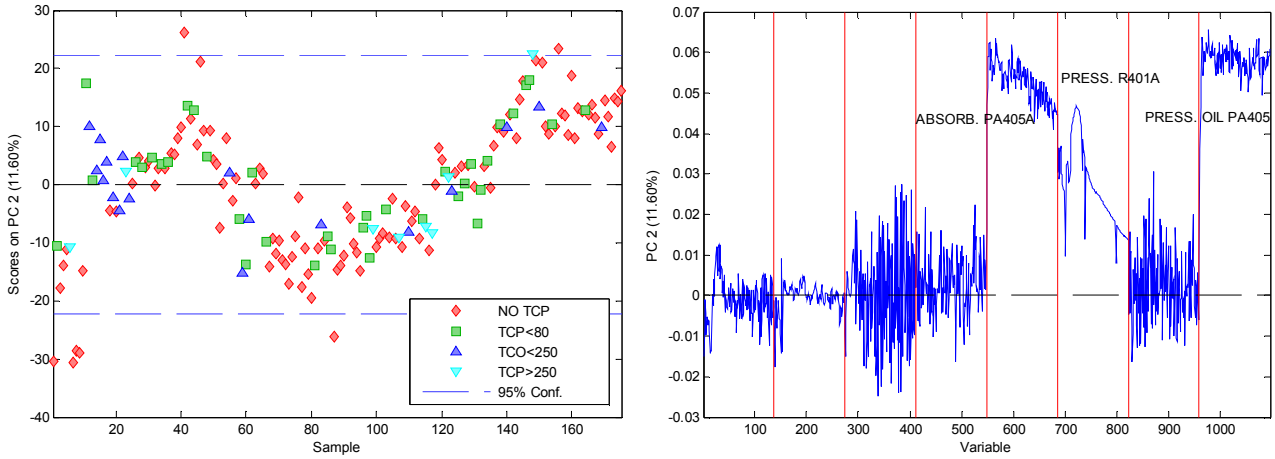

Figure 3.15 On the left, samples ordered vs. PC2 score, on the right, loading PC2. PCA model on R401A batch wise unfolding, first phase alignment. Dots colour relates to the TCP amount added to recover instability

3.5.3 Batch trajectories features

In order to check if additional insights could be gathered by a different approach, the trajectories feature matrix has been analysed, i.e. the one in which the time evolution is condensed as shapes information by defining a set of features from the time profile of each sensor. In BW-Unfolding trajectories of the same variable but from different batches are one above the other and the trajectories from different variables but of the same batch are one beside the other. During features extraction the singe sensor trajectory, for each batch (the columns in BW that correspond to the same process variable) are replaced by some of their characteristics (min, max, std, etc..) and as a result a simpler and smaller matrix in which the number of variables is quite a lot less than in BW is obtained. From the point of view of process control, this trick has multiple advantages:

- No alignment required
- Data is available during all the process
- Easy to understand

Compared to the whole trajectories it does not required alignment because it is a propriety of the trajectory and does not depends on the single time point. A disadvantage is that, complete shape appears only at the end of production for each batch, thus in on-line monitoring (not relevant for explorative data analysis of historical data) should be predicted as the reaction proceeds, but even if features might change along the process they can be predicted by imputation from the LV model.

The features matrix is composed from 177 observations (the same batches as before) and 27 variables (features) described in section 3.3 *"Batch trajectories features"*.

A PCA model with two latent variables accounts for 35% of variance. In the first principal component (Fig. 3.16), the main information point to external temperature influence. A similar scores profile is obtained by analysing only the first phase of the batch process: this phenomenon concerns the stirrer, and not the temperature as supposed from the high loading value; it could be observed in the second PC loadings (Fig. 3.15 right) of the PCA model for first phase. Also the rpm and oil pressure features describe the same behaviour for the first phase but some new information come from other parameters collected. There is influence on the second phase by temperature and the maximum value of the jacket; maybe the winter external temperature does not permit the same setting of values or it influences jacket more than in the other seasons.

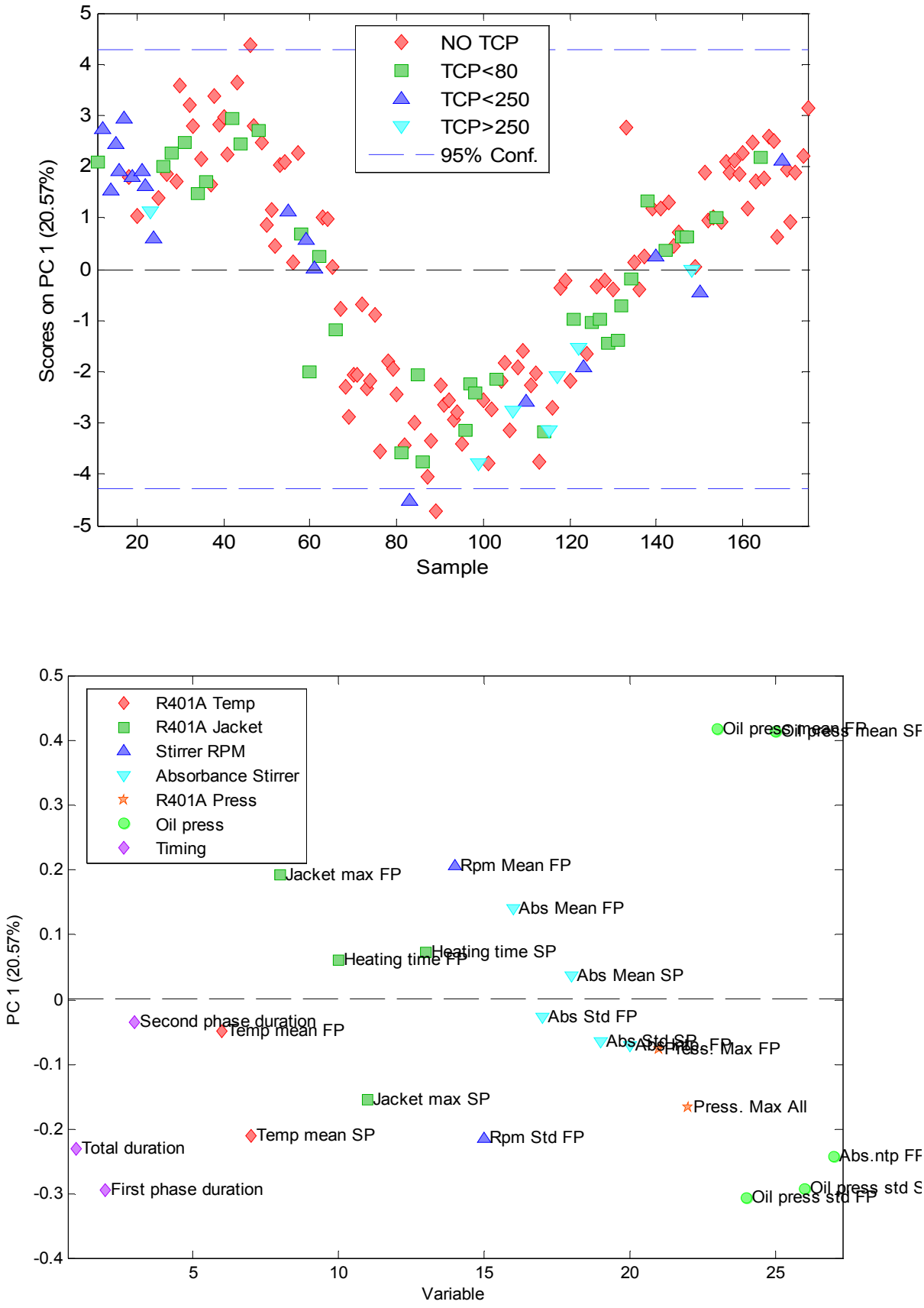

Figure 3.16 Above, samples ordered vs. PC1 score, PCA model on trajectories features. Dots colour relates to the TCP amount added to recover instability. Below, PC1 loading plot, Dots colour relates to trajectory from which feature has been obtained

The second principal component of PCA on trajectories features, figure 3.17, shows a progressive change of the plant: PC2 is highly scattered but the behaviour appears clearly on the component; batches progressively decrease scores values, from a positive to a negative. In order to explain the meaning of this second latent variable the loadings plot are inspected: the batches temperatures get down along the years and this maybe could be good for production, but the second stage takes longer than before and obviously it generates higher production costs. These phenomena seems inversely correlated from the loadings plot and considering the styrene monomer reaction, the information is correct: lower temperatures mean longer reaction time. The second latent variable loadings plot shows also other effects: the absorbance in the first phase increases as does the pressure inside reactor. Considering the well-known relationship between temperature and pressure, their inverse correlation in loadings plot seems quite strange; probably this reflects a different effect that has some indirect correlation with these sensors. Nonetheless, the absorbance mean value increases: either stirrer suffers of aging or plant controller modified reaction parameters along the years. One of these situations could explain the constant drift. For sure, the last plot gives interesting information unknown so far. Next components do not show behaviours related to the instability phenomenon.

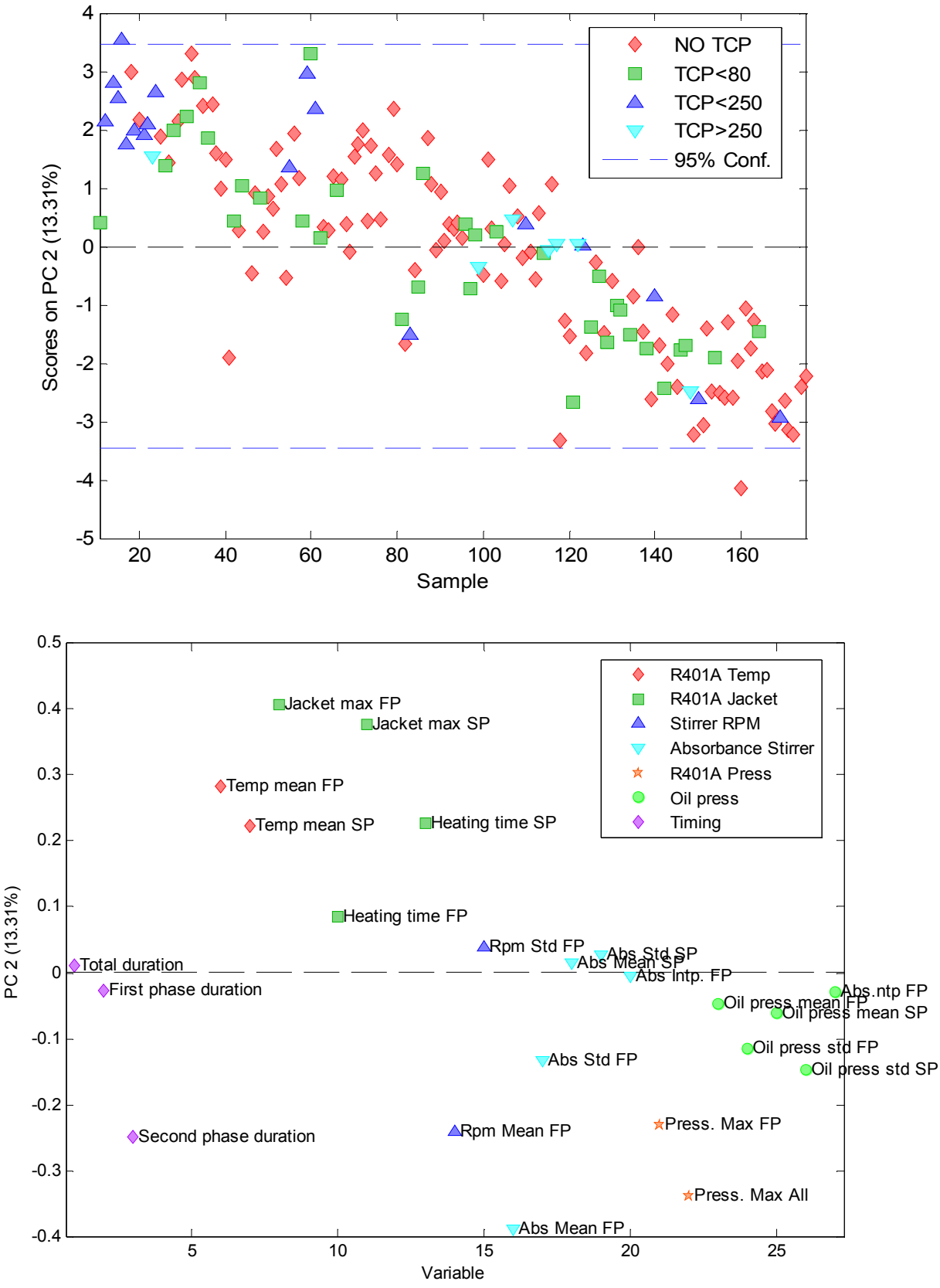

Figure 3.17 Above, samples ordered vs. PC2 score, PCA model on trajectories features. Dots colour relates to the TCP amount added to recover instability. On the bottom, PC2 loading plot, Dots colour relates to trajectory from which feature has been obtained

3.5.4 Water plant treatment

The water pretreatment matrix was also analysed to get any other information that could help the ST14 process understanding. As previously stated, this dataset concern the quality of water used in styrene polymerization. The PCA model explains the 61% of variance with five latent variables; dots colour relates to the production periods as into the previous models (Fig 3.10). Each components shows a change along the two years and quite constant water parameters inside the same month: so the water quality might change in long-term and could be a determinant parameter. PC1 vs PC3 score plot results the clearest ones in terms of clustering; the older productions (May 20111) are the most strange, on the downright of the plot. Moreover, the black triangles group includes the dump batch and it is the one with the higher suspending agent quantity needed to preserve the suspension. Anyway, the scores plot in which colours correspond to the added quantity do not disclose any clear clusters according to TCP but anyhow highlighted the dump batch in 2011 (orange colour). This information thus can be useful and has been added to formulation and reactor matrices in a block model as will be discusses further.

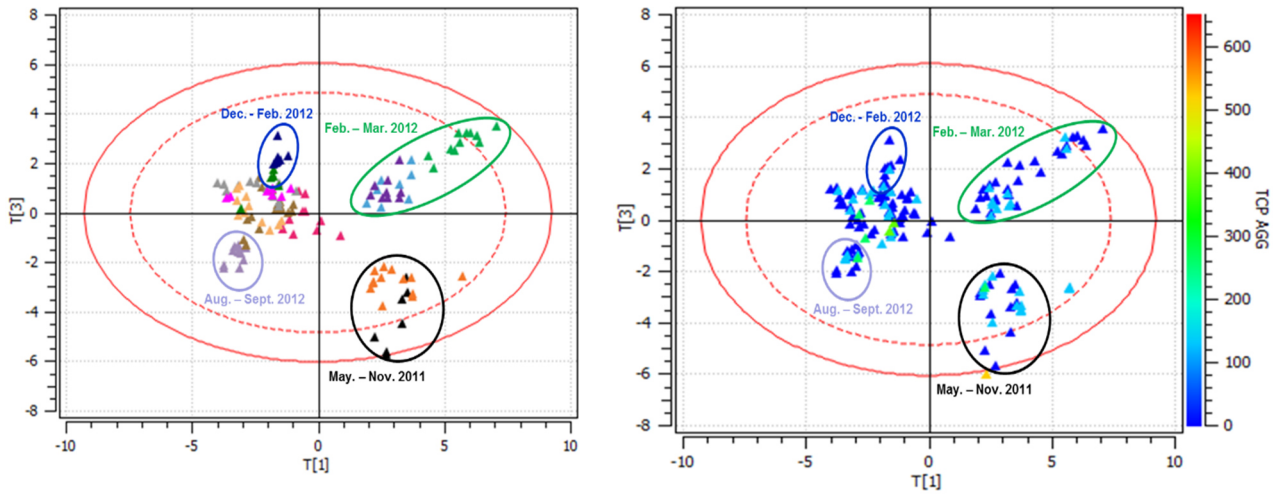

Figure 3.18 Score plot PC1 vs PC3, PCA on water treatment matrix, on the left observations colour represent the different production periods, on the right colours the TCP amount added to recover instability

The exploration analysis gives various suggestions on ST14 plant and was fundamental for the next step, the discriminant analysis.

3.6 PLS discriminant analysis

Expandable polystyrene process suffers from instability problem. The discriminant analysis aims at detecting differences between normal operative condition (NOC) batches and fault production. The first step concerns the definition of "good" and "bad" productions that for ST14 can be gathered by the concentration of suspending agent (TCP); the minor the quantity added, the more stable the solution hence best production, the higher the quantity added the higher the instability of the solution, to be noticed the dump batch belongs to this condition. However, it has to be observed, that just because of the addition of increasing quantity of suspending agent a "bad" situation is generally recovered and does not bring to "fault/stop" of production. This renders more difficult to observe differences in LV models because of the correction made. Moreover, into the database is available a dump batch, the one observed in exploratory data analysis. Its trajectories are so different from the other batches due to the operation that preserves plant items, such as the stirrer for example; jacket cools reactors, pressure decreases and a solvent stops reaction. Immediately a model recognizes the dump production but it is not interesting because the fault already occurred and the raw material already lost, since production stuck. More interesting is the comparison of the stable batches with the dump one, in the time points before the instability become out of control; in order to make this, all trajectories (i.e. time points) have been considered from start to the time point that corresponds to the last stable situation for the dump batch. This data set was first analysed by PCA .Also, this model did not return a clear separation of the dump batches and it means that just before the suspension instability went out of control everything into the reactor was fine.

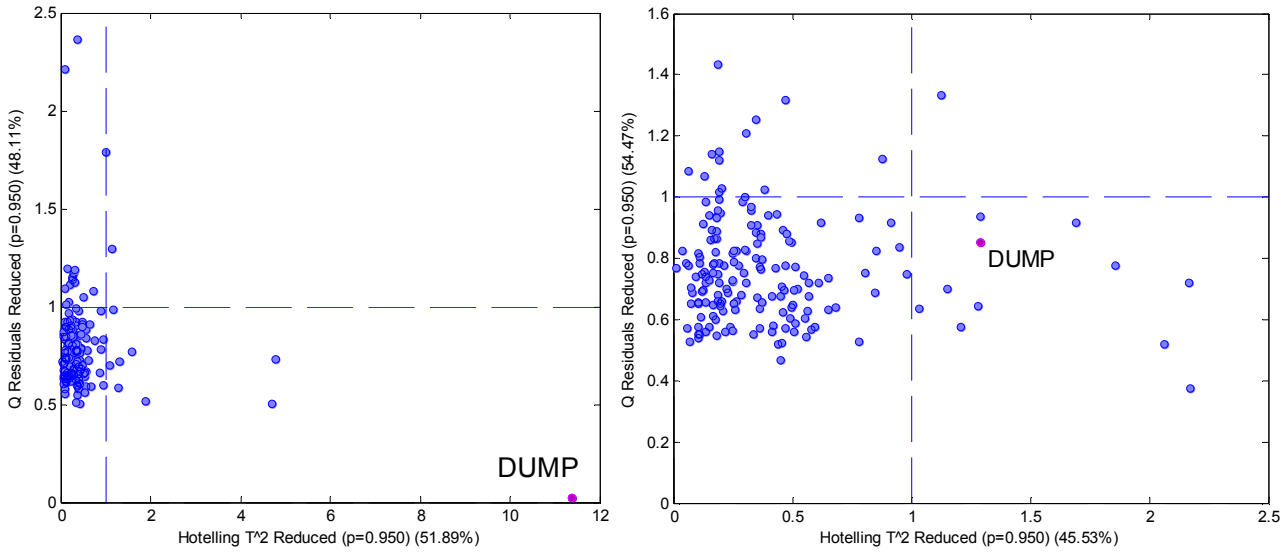

Figure 3.19 vs Q residuals, PCA model on entire process trajectories (left), and on process timing before DUMP (right)

Scores plot shows that dump trajectories are similar to the ones of the batches produced during the same month. The batch where production stopped was removed from the dataset and PLS –DA model built with the remaining observations. The best and worst conditions, i.e. the batches without any supplementary addition of TCP and the batches in which TCP addition overcome 250 litres, have been used to define the two categories; this choice has been taken in order to see if a PLS discriminant model considering only the two most different batches categories could work. The reaction matrix is undoubtedly the most relevant in process monitoring and must be involved in fault detection; a suitable way to include such a fundamental information in data analysis is the batch wise unfolding. I took data of the entire production and made a PLS-DA model with No-TCP and high TCP feeding as categories. The model explain 74% of the global variance in X with 3 PC's but returns an error around 0.5 in prediction; it means that the model is not able to distinguish between the batches. Looking through the principal components, the second latent variable shows a separation that seems related to the period of production and that all the high TCP batches have higher values. The other effect visible into PC2 is a drift from left to right, the variables mainly involved into this separation are the ones related to the stirrer. Only two batches of the No TCP cluster have been removed due to their high value in T^2 and Q.

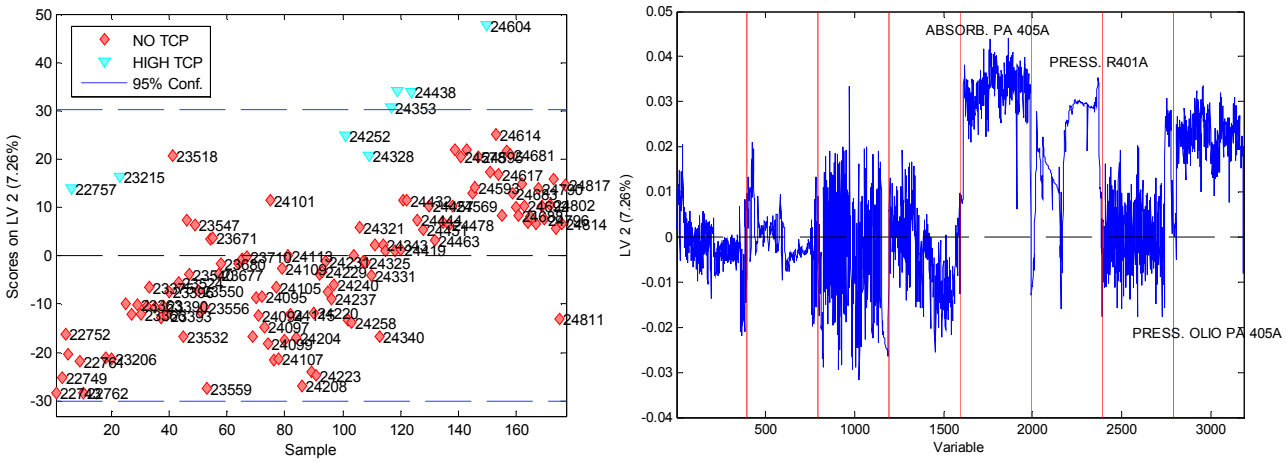

Figure 3.20 On the left, samples ordered vs. PC2 score, on the right, PC2 loading plot. PLS-DA on the reactor dataset BW unfolded. Dots colour relates to the TCP amount added to recover instability

In order to remove possible noisy information, PLS-DA was done only on the first reaction part, with reactor data until the end of the 90°C phase (Chapter 3.2). Again, few "Good" batches are excluded, because out of the confidence limits and the high TCP batches have been used as the "bad" production class. Discrimination model still suffer of poor prediction capability and it is not reliable for a fault individuation. Anyway, the information gathered is consistent with the previous one; the pressure into the reactor seems related to instability as the variables related to the stirrer.

The discrimination models cannot be used in batch discrimination because of the poor quality of prediction but a separation between productions appears on PLS-DA model and it might return some information related to instability. So, the profile difference on raw and aligned data and the contribution plot have been compared, the "Bad" batches trajectories and the mean of the "Good" ones. Beginning from the contribution plot, high TCP observation has higher coefficient on temperature and lower for pressure, figure 3.21 left plots. Unfortunately, raw trajectories on these sensors do not show the same behaviour and good production seems more dissimilar that bad is. Batches discrimination needs some differences in trajectories to separate groups and to predict an external observation; the previous example does not return any evident variation between raw trajectories related to the contribution plot and it makes discrimination evaluation useless.

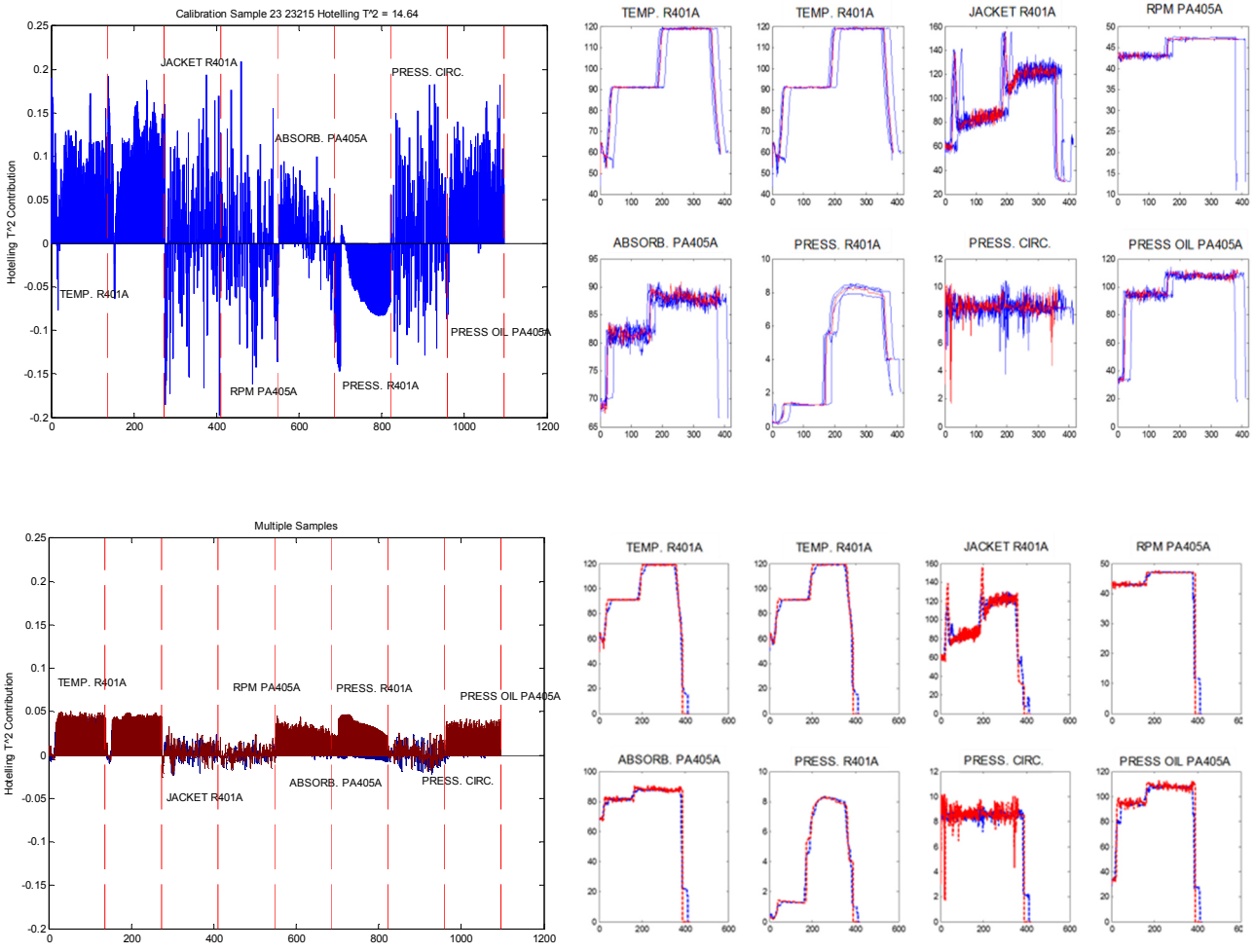

Figure 3.21 T^2 contribution plot of 23215 batch (top left) and of two referring batches (down left). The raw profiles (top right) and the mean trajectories (down right) of no TCP, in red, and TCP, in blue, batches

In fact, the model error suggests that model has poor reliability. Other information might come from the residuals plot and residuals profiles of "good" and "bad" batches were compared. Again, the correlation between variables seems similar between batches of the same production periods and differences among them does not relate with the batch quality but more to the month of production.

In the figure 3.9, some effects were found into the raw trajectories and maybe aligned procedure removed this variation. To improve discrimination into the next model also the warping information has been used. The warping alignment returned two columns relative to the pair: reference (trajectory used to align) and sample (the batch trajectory to be aligned in turn) from this warping descriptors can be derived that might help to recover the sources of information lost by alignment step. Also the PLS-DA model including warping descriptors is not able to discriminate the productions with high suspending agent concentration form the low ones: warping vector was properly scaled but its effect was too little to change the behaviour of the model.

Therefore, discrimination analysis took into account also the formulation dataset. Data are block scaled in order to preserve the contribution of each block in spite of the number of variables is so different. Despite the effort, the result did not change: PLS-DA model shows an interesting separation along components and on T^2 Hotelling plot, but cross validation confirm that model cannot predict new data. The information that comes from the formulation data matrix does not match with to the instability.

Water and trajectories dataset has been used in the same way but the prediction quality does not improve as expected. A lot of variance came from the exploratory analysis and I supposed that part of such variation was correlated to the instability phenomenon. Moreover, with the Block Importance in Prediction (BIP) I discovered that the most important matrix in data separation is still the reactor sensor matrix independently from the scaling.

The score plots of all PLS-DA models performed returns a similar trend: as in figure 3.20, light blue triangles, "Bad" batches, have higher values than the "Good" production into the second component, or into another among the first three. Such condition suggests that observation differs inside the same production period; the separation belongs to a specific group and not to the whole dataset. Therefore, data where centred per the production period: mean values were calculated for each production months and instead to use the global mean in which all the batches are taken into account each production has been centred on its own mean; then PLS-DA re-run. Scaling might decrease variation between groups and makes them more similar; the expected effect is to highlight more clearly the variation among batches.

This approach resulted unsuccessful and in spite of a scores plot in which high TCP batches appears different (Fig. 3.22), the model has a poor discrimination quality. Instable batches have high positive values into the first component, and this is the only feature that distinguishes high TCP from No TCP productions. Loadings plot shows that the pressure along the first production phase is the most important parameter, the only with a considerable effect along reaction; pressure inside tank vary considerably from batch to batch in a such way that the other sensors have no importance into the first latent variable.

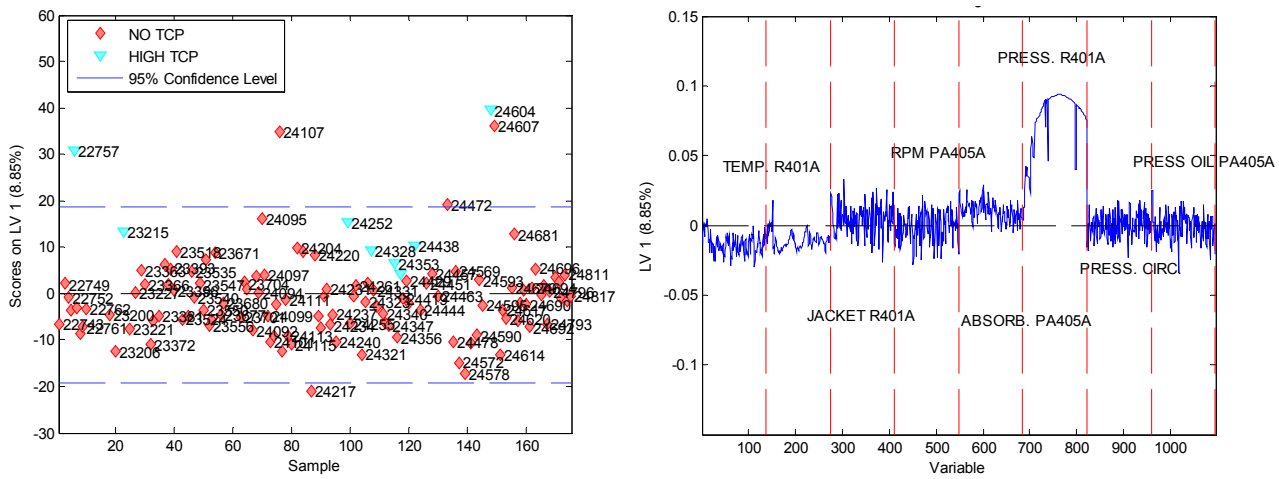

Figure 3.22 On the left, samples ordered vs. PC1 score, on the right, PC1 loading plot. PLS-DA on the reactor dataset BW unfolded and block centred. Dots colour relates to the TCP amount added to recover instability

Various approaches tried to model the instability in batch production with very little results in terms of prediction. The information came from the formulation matrix, the water parameters and from the warping information that took part in multi-block discrimination model.

3.7 Conclusion

In this study was not found any clear phenomenon to discriminate batches that needs suspending agent addition and batches stable during whole production process. This is the first honest conclusion on the ST14 application, for the self-extinguish EPS produced from 2011 to the 2013. The poor prediction ability, when discriminant models were applied, might be related to the relevance of the sensors with respect to the phenomenon taken into account or to the time points frequency that could be too low to capture the moment when dispersion became instable to an extent that could not be controlled by stabilization agent addition.

Actually, all the available sensors were the ones involved in multivariate analysis and no more are installed into the ST14 plant. The most important propriety not considered in the analysis is the particle size dimension; operator measures the dispersed drops dimension with a reference sample and these data are wrote down manually and it helps the operators to control the reaction. Particle sizes were not available for many of the considered batches. In relation to the frequency, that might be not the suitable one, the database actually acquires data with higher frequency but store these for a short period, then only a part of the acquired data is saved for long-term evaluation. Recovering data my interpolation is reasonable for NOC but when accidental high frequency events happen it is impossible to recover them in this way. In order to increase the stored time points it is necessary to modify database setting.

Despite the poor discrimination quality, the multivariate analysis returns information throughout the reactor dataset, the formulation matrix, the water treatment parameters; also the trajectories dataset and the warping information, in which data comes from the elaboration of sensors matrix, helped the study. Exploration analysis shows that the evolving of the batch is independent form the reactor in which production happens but it obviously depends on formulation, i.e. desired grade; EPS field produces four different products that mainly differ for the additives. It does not mean that reactors are equal but describe a semi continuous production in which reactors produces grades with the same formulation and that about additives there are not peculiarities related to reactors.

Another meaningful analysis is the one on water treatment. Water is a fundamental raw material in EPS production and the analysis returns that its quality varies along the years. Such variation did not improve the predictive ability of discrimination models, anyway, clusters in PCA water treatment model looks like the clusters of PCA model on R401 batch wise unfolding and this gives evidence of the dependence of production on the season.

The external temperature contributed a trend to the batch trajectories features that is individuated by the first latent variable, i.e. a seasonal effect. Productions suffer the external temperature that increases the stirrer absorbance in the winter months and similarly the other parameter related to stirring, the oil pressure, changes its value. The second latent variable describes an aging effect: maximum jacket temperature decreases as the first temperature does and the stirrer makes more effort years after years coherently to the reactor pressure.

The most process-correlated data set is the R401 batch. In order to compare batches alignment was mandatory. In that way, length and main events match and a multivariate analysis compares properly batches. Exploratory analysis returns a variation related to the production periods: reaction temperature varied from a production to another due to the manual setting of this temperature.

Again, in sensors matrix, season affects production and in PCA score plot appears this behaviour; external temperature effect might disturb or even interferes with the other phenomenon between batches. Global trajectories and the first phase reaction have been used separately in order to build an efficient discrimination model. PLS-DA gave some results that are not acceptable for data prediction. Discrimination model involved also the other matrices but the error in prediction did not change enough.

The MVDA is not satisfactory in relation to the discrimination process. Frequencies of selection, sensor involved in analysis and some other hypothesis might explain why models were not able to predict instability. Operators control production via mean particle size and TCP additions are function of those values. The particle size information is probably necessary to keep under control the instability and cannot be inferred by the multivariate models based on the other sensors. This is the reason that prompted us, to, realize a quantitative and repeatable measurement system to follow the mean particle size (MPS) based on NIR spectroscopy; this has been patented and in the next section there are more details about it. Such method is not actually installed on ST14 due to the bureaucracy involved in new equipment setting on industrial plant. Nonetheless, in very near future multivariate data analysis will benefit from the use of reactor sensors coupled with NIR information, to this aim data fusion modelling will be needed [24].

3.8 Application for on-line process control

In suspension reactions, the Particle Size Dimension (PSD) has an important role; it determines the size of final product pearls and gives to operator a qualitative information on process stability. Various methods have been developed for the particle size measurement, among the others the Focused Beam Reflectance Measurement (FBRM) [25], the Particle Image Analysis [26] and the application of Raman spectroscopy [27]. The last one is the most similar to the NIR application. An advantage of spectroscopy, as method to estimate the PSD, is that spectroscopic data relate primarily to the chemistry of the process, such as the conversion degree and the additives concentration besides physical effect as the pearl size. Thus, they allow control both composition and size of a suspension reaction. The application is partly innovative [28 29 30 31 32 33] and company decided to patent it internationally. I report in the last chapter the patent text in order to describe the application.

3.9 References

1. R. Smith (2005), *Chemical Process: Design and Integration - Chapter1: The Nature of Chemical Process Design and Integration*, John Wiley & Sons
2. García-Muñoz, S., Kourti, T., & MacGregor, J. F. (2004). Model predictive monitoring for batch processes. *Industrial & engineering chemistry research*, 43(18), 5929-5941
3. Kourti, T., Nomikos, P., & MacGregor, J. F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of process control*, *5*(4), 277-284
4. Tates, A. A., Louwerse, D. J., Smilde, A. K., Koot, G. L., & Berndt, H. (1999). Monitoring a PVC batch process with multivariate statistical process control charts. *Industrial & engineering chemistry research*, *38*(12), 4769-4776.
5. Zhang, H., & Lennox, B. (2004). Integrated condition monitoring and control of fed-batch fermentation processes. *Journal of Process Control*, *14*(1), 41-50.
6. Castellani L., Longo A., Pasquali F., (2005), Encyclopaedia of hydrocarbons, Volume II Refining and petrochemicals, Thermal conversion processes, Istituto della enciclopedia italiana fondata da giovanni treccani S.p.A
7. Eni Versalis portfolio, Styrenics,[https://www.versalis.eni.com/irj/portal/anonymous?NavigationTarget=navurl://f9ab9bac7fb70463e2c2e9d94807359c&guest_user=anon_en]
8. Nemeth, S., & Thyrion, F. C. (1995). Study of the runaway characteristics of suspension polymerisation of styrene. *Chemical engineering & technology*, *18*(5), 315-323.
9. Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of chemometrics*, *12*(5), 301-321.
10. Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, *17*(1), 16-33.
11. Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S. *PLS_Toolbox for use with MATLAB™*. Eigenvector Research, Inc., Wenatchee, USA, 2006.
12. Tomasi, G., van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, *18*(5), 231-241.

13. Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, *805*(1), 17-35.
14. Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *26*(1), 43-49.
15. González-Martínez, J. M., Westerhuis, J. A., & Ferrer, A. (2013). Using warping information for batch process monitoring and fault classification. *Chemometrics and Intelligent Laboratory Systems*, *127*, 210-217.
16. González-Martínez, J. M., Noord, O. E., & Ferrer, A. (2014). Multisynchro: a novel approach for batch synchronization in scenarios of multiple asynchronisms. *Journal of Chemometrics*, *28*(5), 462-475.
17. González-Martínez, J. M., Vitale, R., de Noord, O. E., & Ferrer, A. (2014). Effect of synchronization on bilinear batch process modeling. *Industrial & Engineering Chemistry Research*, *53*(11), 4339-4351.
18. Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, *38*(2), 149-171.
19. Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*(3), 279-311.
20. Camacho, J., Pico, J., & Ferrer, A. (2008). Bilinear modelling of batch processes. Part I: theoretical discussion. *Journal of Chemometrics*, *22*(5), 299-308.
21. Camacho, J., Picó, J., & Ferrer, A. (2008). Bilinear modelling of batch processes. Part II: a comparison of PLS soft-sensors. *Journal of Chemometrics*, *22*(10), 533-547.
22. Kourti, T. (2003). Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *Journal of Chemometrics*, *17*(1), 93-109.
23. Tomba, E., De Martin, M., Facco, P., Robertson, J., Zomer, S., Bezzo, F., & Barolo, M. (2013). General procedure to aid the development of continuous pharmaceutical processes using multivariate statistical modeling–An industrial case study. *International journal of pharmaceutics*, *444*(1), 25-39.
24. Silvestri, M., Bertacchini, L., Durante, C., Marchetti, A., Salvatore, E., & Cocchi, M. (2013). Application of data fusion techniques to direct geographical traceability indicators. *Analytica chimica acta*, *769*, 1-9.

25. Hukkanen, E. J., & Braatz, R. D. (2003). Measurement of particle size distribution in suspension polymerization using in situ laser backscattering. *Sensors and Actuators B: Chemical*, *96*(1), 451-459.
26. Larsen, P. A., Rawlings, J. B., & Ferrier, N. J. (2006). An algorithm for analyzing noisy, in situ images of high-aspect-ratio crystals to monitor particle size distribution. *Chemical Engineering Science*, *61*(16), 5236-5248.
27. Reis, M. M., Araújo, P. H., Sayer, C., & Giudici, R. (2003). Evidences of correlation between polymer particle size and Raman scattering. *Polymer*, *44*(20), 6123-6128.
28. Santos, A. F., Lima, E. L., & Pinto, J. C. (1998). In-line evaluation of average particle size in styrene suspension polymerizations using near-infrared spectroscopy. *Journal of Applied Polymer Science*, *70*(9), 1737-1745.
29. Santos, A. F., Lima, E. L., & Pinto, J. C. (2000). Control and design of average particle size in styrene suspension polymerizations using NIRS. *Journal of Applied Polymer Science*, *77*(2), 453-462.
30. Lousberg, H. H. A., Boelens, H. F. M., Le Comte, E. P., Hoefsloot, H. C. J., & Smilde, A. K. (2002). On-line determination of the conversion in a styrene bulk polymerization batch reactor using near-infrared spectroscopy. *Journal of applied polymer science*, *84*(1), 90-98.
31. Chicoma, D. L., Sayer, C., & Giudici, R. (2011). In-Line Monitoring of Particle Size during Emulsion Polymerization under Different Operational Conditions using NIR Spectroscopy. *Macromolecular Reaction Engineering*, *5*(3-4), 150-162.
32. Reis, M. M., Araújo, P. H., Sayer, C., & Giudici, R. (2003). Correlation between Polymer Particle Size and in-situ NIR Spectra. *Macromolecular rapid communications*, *24*(10), 620-624.
33. Gossen, P. D., MacGregor, J. F., & Pelton, R. H. (1993). Composition and particle diameter for styrene/methyl methacrylate copolymer latex using UV and NIR spectroscopy. *Applied spectroscopy*, *47*(11), 1852-1870.

IV

Trouble shooting and process monitoring
Continuous process analysis

Content

4.1 Introduction

The monomer production might show some advantages with respect to polymer production plant: the process concerns a single product or a single stream that may split in successive sections in order to obtain specific and pure compounds. Monomer production is ideally simple, similarly to a distillation process, liquid enters in column, and either from the top or from the bottom a more pure product come out, the process is continuous and the same monomer should be obtained with best purity and with the cheapest process. However, the integration of side products from different processes, i.e. hydrogenation section that gives hydrogen to a cracking process, forces plant management to otherwise unnecessary setting changes. Anyhow, monomer production does not suffer the problems due to grade change as polymerization plants do. In principle, a monomer plant should not change setting from the beginning to the end of production, unless because of catalyst exhaustion or of different raw materials. Obviously, the ideal production is far away from the reality. Even monomer productions suffers the flow rates changes, external temperature and other undesired variation that might affect production.

Styrene monomer (SM) production was taken into account as one the Thesis benchmark. Last production campaign, that covered about 4 years, had a serious undesired polymerization problem that caused an earlier and most expensive maintenance. Styrene must not polymerize inside the monomer plant and proper settings have been adopted to avoid conversion in polystyrene; production is undertaken at low temperature, additives are used to avoid polymerization and many other resources are employed to keep production under control. Despite that, in the 2011 – 2014 campaign, a high quantity of polymer was formed; it was enough to block the monomer fluid and seriously affected the maintenance.

In this application, multivariate data analysis aims to the understanding of why this particular phenomenon occurred, for which plant management did not found a reason with standard univariate approach and on the basis of previous knowledge. The troubleshooting analysis focused on more than one campaign and on different periods to be sure that every source of variability that could match with phenomena under consideration have been considered. The following paper (submitted for publication) contains details about the applied methods and the obtained results. Furthermore, the last section describes the MATLAB interface I developed to support the plant monitoring with multivariate control charts.

The paper does not include a proper process description but enlarge upon the multivariate analysis and conclusions; for that reasons an exhaustive description is included in this introduction.

4.1.1 Monomer production plant (ST20)

ST20 is the acronym of the styrene plant that produces the styrene monomer from benzene and ethylene. The process reaction goes through an intermediate product, the ethyl-benzene (EB), followed by a dehydrogenation step to obtain the carbons double bond.

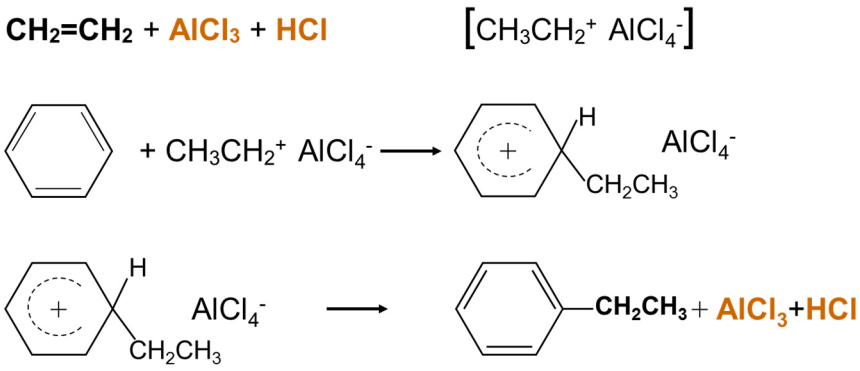

Figure 4.1 Monomer synthesis scheme

Styrene Monomer (SM) line includes two separate phases: alkylation and dehydrogenation. During the alkylation step, benzene and ethylene react according to a Friedel-Crafts reaction. Aluminium trichloride catalyses the ethyl benzene production inside reactor R1105. An ethyl chloride stream comes into the same reactor and promotes reaction whit in the previous one. The alkylation phase consist of an exothermal reaction and for the control of Temperature inside the reactor a thermal exchanger is used, i.e. controls temperature: an EB aliquot pass through the thermal exchanger and the amount of the flux regulates the temperature. Although, EB steam partly condenses in R1105 it is however sufficient to decrease the reactor temperature. Alkylates mixture is composed from Ethyl benzene, Benzene, Toluene, Poly Ethyl benzene, Paraffin high boiling compounds and catalyst. Downstream from reactor, catalyst decants and flows back to the reactor in order to save raw catalyst. Alkylates and a certain amount of water go to the C1008 that divides organic compound, from top of the column, and $AlCl_3$ solution, from the bottom. Then the main stream receives a second treatment: operator adds a certain amount of NaOH water solution in order to neutralize pH plant. Again, the undesired basic solution flow whit the HCl water solution from aluminium trichloride plant to water treatment section. At this point alkylate shall be separated and here begin the distillation sub phase.

Distillation section is composed from five columns:

- C102 recover non-converted benzene.
- C103 dry the benzene flow to the reaction.
- C104 divide benzene and toluene coming from dehydrogenation
- C105 extract ethyl benzene from reaction fluid and send the residuum on the next column
- C106 separate di-ethyl benzene from the high weight compounds

Alkylate mixture feeds C102 column: upper vapour goes to C103 column such as C104 vapours, alkylate liquid goes to C105 from the bottom. The column C104 separates benzene from toluene; the second one is stored as final product. Benzene from C102 and C104 flows into C103, which removes water until a residual humidity of 20ppm, then the dried benzene is transferred to the reactor R1105N.

In C105 alkylate are separated from di-ethyl benzene (DEB) and poly-ethyl benzene (PEB) compounds, alkylate goes out from the top. The residuum from C105 feeds column C106 that recovers di-ethyl benzene, recycled into alkylation section, and send PEB to the gas treatment (PEB is commonly used as gas absorber). The ethyl benzene, intermediate product, goes to the dehydrogenation section from the C105 passing to C202: steam and benzene are mixed to form the so-called steam oil, water vapour facilitates dehydrogenation reaction that take place into the R3201A reactor. There are three reasons to use steam:

- It provides the heat to the dehydrogenation reaction (endothermic);
- It decrease the partial pressure of ethyl-benzene that shift reaction balance towards the formation of styrene;
- It increase catalyst life removing the carbon (cracking of hydrocarbons).

The second reactor maximizes conversion and returns a mixture composed by ethyl-benzene, steam, styrene, toluene, benzene, hydrogen and high boiling compounds (TAR) as consequence of thermal cracking into the reactors. Therefore, that mixture partially condenses in a series of thermal exchanger and, finally, in a water quench. A decantation process separates organic phase from water phase and the last is recovered for the quench. Residual gas (off-gas), composed by hydrogen and organic vapour, are treated in a shared section dedicated to the washing and purification process. Dehydrogenated mixture contains mainly styrene with residual ethyl-benzene and benzene; a series of columns separates the final product, styrene, from unconverted, ethyl-benzene and benzene.

The four columns are briefly described:

- C201 splits benzene and toluene that go to the C104, purification section.
- C202 recovers ethyl-benzene and send it to dehydrogenation section.
- C203A purifies styrene
- C203B separates residual styrene from high boiling substances

Styrene polymerizes with high temperature so all columns in purification section work in vacuum in order to decrease the bottom temperature and to reduce fouling factor due to polymerization. The C201 column separates on top benzene, water and toluene and send it to C104 (refer to distillation section). Lower C201 product feeds C202: ethyl-benzene, lower boiling liquid, come back to dehydrogenation section then the rest goes to the C203A. At that stage styrene flows from the top and goes to the storage tank; to inhibit polymerization, operators add few ppm of TBC. The C203B recover residual styrene and, by a controlled combustion, dumps TAR to recover energy.

The unwanted polymerization involved mainly the condensation section and the trouble shooting analysis take into account only this zone of the plant. The purpose of condensation phase is to remove undesired gas and to reduce stream temperature. At this production step, monomer mixture consists in steam, styrene, ethyl benzene, benzene, toluene and hydrogen. Mixture arrives to the condensation section from reactor R3201B; this section includes two parallel and identical lines so called “new” and “old” line. Whole section works in vacuum in order to decrease the condensation temperature. The mixture are higher than 500°C when it arrives to the first thermal exchanger in both lines: here the temperature decreases at maximum 150°C. A series of water quench cools mixture at more or less 60°C then the two lines join together in a collective exchanger that partly condenses the mixture; the remaining gases go to the compressor and are then treated by another plant. Organics compounds arrive to the D202 tank which separates the monomers form water before sending it to distillation section.

4.2 Paper

Troubleshooting of a continuous styrene process and multivariate monitoring

Erik Mantovani[a], Leonardo Trentini[a] and Marina Cocchi[b]

a: Research and Development department, eni Versalis, Via Taliercio, 14 Mantua Italy

b: Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi, 103 Modena Italy

E-mail: erik.mantovani@versalis.eni.com

Phone: +39 0376 305501

Fax: +39 0376 305639

Keywords: Trouble shooting, Multivariate control charts, Styrene production plant, Multivariate statistical process monitoring

Abstract

In this article, we show the application of multivariate statistical process monitoring to a continuous process concerning styrene monomer production, in particular a case of troubleshooting. The approach consists in multivariate analysis of historical data from plant database retrieval in order to assess normal operating condition and to highlight time trends possibly linked or anticipating the plant fault. In monomer production troubles are likely related to unwanted side reaction. Principal component analysis (PCA) as exploratory tool allowed to depict the main behaviour of the plant, for instance the flow changes, the catalyst ageing as well as abnormal changes. The obtained results support the importance of using a multivariate data analysis approach in fault detection and show that correlation structure and process variability are the driving effects that influence the whole production. Moreover, the results obtained stimulate the interest of plant management in continuous process monitoring and thus a software application based on the derived multivariate models has been implemented for plant monitoring. This has been tailored on company database and ICT system and already supports styrene monomer production.

Introduction

In the last decades multivariate data analysis (MVA) has been applied in chemical and petrochemical industry to improve [1] and monitor production [2] and many examples come from literature [3 4 5]. When aiming at trouble shooting analysis the analyst focus attention and effort on a particular time period where a negative event, either a departure from normal operative conditions provoking a plant stop or a departure from attended quality of intermediate or end products, took place; the objective is to discover as much as possible on the operative conditions close to process fault in order to gain useful information [6] possibly allowing detection of the origin of the problem so to avoid it in successive productions. Sometimes cause – effect relation is not so clear considering only the experience due to previous faults. Moreover, usually the knowledge acquired by the management personnel strictly depends on few keys variables that they check daily or in normal process condition. However, it is common that negative events (i.e. the plant shut down) are the result of many side effects, maybe of small entity, involving several process parameters and their correlations. This is why monitoring single variables, most often, does not highlight the main problem. Decomposition methods, such as Principal Component Analysis (PCA), are most suited to address these problems, in fact they take into account all sensors and their correlation while deriving only a small number of latent variables. Another advantage is that they can work well despite the signal to noise ratio is generally very low process data and often there is a considerable presence of missing data. Applications in literature report successful multivariate study for troubleshooting, e.g. Muñoz et al. [7] were able to recover fundamental information on a batch drying process and to identify the causes of poor quality of production. An interesting example of PCA application for process control is the one from Wise et al. [8] that analysed variation in a slurry-red ceramic melter process in which normal operative conditions were used to discriminate disturbed production. Also the quality might be controlled via latent method and the application in polymer process from Skagerberg et al. [9] proves that plant setting defines final product characteristics and that multivariate analysis is able to extract this information.

This application concerns styrene monomer production, in which both high quality and quantity matter, as in many other manufacturing. In this particular process for the monomer synthesis, high quality is almost always achieved and it does not represent a so critical point, consequently plant setting does not change frequently and operative conditions remain rather stable. The main variation in production is often related to the changes in flow rate. Such changes are due to variability of raw materials cost, monomer and polymer demand and other market based reasons.

The flow rate change may affect stability, anyhow nobody notice clear side effects since a lot of years. Furthermore, a long term drift in the plant is observed and it is due to the catalyst exhaustion: reaction needs more heat to reach the proper conversions and temperature will be increased to correct for that. This well-known behaviour affects the entire campaign. Instability may also take place as a consequence of special maintenance or substitution of some items. That imply to stop the entire production, anyway item breakdown is not so frequent. In general, monomers production may suffer from many even small events/effects and could benefit from a multivariate process monitoring system. The aim of this work is two-folds at exploratory level the historical plant data, from previous campaigns, may allow rationalizing plant behaviour in terms of catalyst aging and how it affects the plant set-up to maintain the desired product quality, and in the modelling phase multivariate control charts are built for real time monitoring. Historical data analysis showed some production evidences that the plant management was not aware of and brings to increasing knowledge of this specific process/plant. These results allowed then to develop a user interface based on moving windows PCA (MVPCA) [10] for real time process monitoring. The paper is organized as follows: first the main workflow of the process and main type of fault is presented, then data and methods are described and results are discussed in sub-sections specific to each analysed data sets.

1 Process description

ST20 is the acronym of the styrene plant that produces the styrene monomer from benzene and ethylene [11] that was considered in this study. The process reaction goes through an intermediate product, the ethyl-benzene (EB), followed by a dehydrogenation step to obtain the carbons double bond. Styrene Monomer (SM) line includes two separate phases: alkylation and dehydrogenation. During the alkylation step, benzene and ethylene react according to a Friedel-Crafts reaction. Aluminium trichloride catalyses the ethyl benzene production inside a plug and flow reactor. The alkylation phase is an exothermal reaction and for the control of temperature inside the reactor a thermal exchanger is used. The out coming alkylates mixture is composed by Ethyl benzene (EB), Benzene, Toluene, Poly Ethyl benzene, Paraffin, high boiling compounds and catalyst. A series of columns separate the intermediate product (EB) from the others that should be treated in order to recover raw material and increase the conversion. EB goes to the dehydrogenation section through a column in which steam is added to form the so-called steam oil; water vapour facilitates dehydrogenation reaction that take place into a successive reactor.

There are three reasons to use steam:

- It provides the heat to the dehydrogenation reaction (endothermic);
- It decrease the partial pressure of ethyl-benzene that shift reaction balance towards the formation of styrene;
- It increase catalyst life removing the carbon (cracking of hydrocarbons).

The second reactor maximizes conversion and returns a mixture composed by ethyl-benzene, steam, styrene, toluene, benzene, hydrogen and high boiling compounds (TAR) as consequence of thermal cracking into the reactors. Therefore, another separation step segregate the final product, styrene monomer. Styrene must not polymerize in ST20 and for that reasons the entire monomer production happens at temperature in which styrene might not polymerize. Monomers will be used for various successive production and an eventual polymers particles might create serious problem in that phase, nonetheless into ST20 pipes and instruments. Unfortunately, such unwanted polymerization happened during last campaign and involved mainly the condensation section in which condition are critical; the trouble shooting analysis take into account only this zone of the plant. For that reasons some details on this section are given. The purpose of condensation phase is to remove undesired gas and to reduce stream temperature. Mixture arrives to the condensation section from reactor this section includes two parallel and identical lines so called “new” and “old” line. Whole section works in vacuum in order to decrease the condensation temperature. The mixture are higher than 500°C when it arrives to the first thermal exchanger in both lines: here the temperature decreases at maximum 150°C. A series of water quench cools mixture at more or less 60°C then the two lines join in a collective exchanger that partly condenses the mixture; the remaining gases go to the compressor and are then treated by another plant. Organics compounds arrive to a tank called D202, which separates the monomers form water before sending it to distillation section.

2 Main fault description

Plant management and operator use sensors data in order to improve the quantity and quality of production, to reduce plant variability and to ensure the safety of whole process. Often, this huge amount of data are not taken into proper account and are not analysed in the optimal way as a result potentially useful information flow away making then very difficult to unravel the causes of faults and preventing the possibility of anticipate them. Trouble shooting analysis-taking advantage of MVDA can assist to find the causes of faults, of worst production quality and to identify the plenty of situations that generate lack in efficiency. Styrene production undergoes 3 - 4 years of continuous production before a deep maintenance is programmed and this is rather expensive. On last campaign from 2011 to 2014, few months before the scheduled plant stop for maintenance, a high amount of polymer forced the interruption of monomers production. The quantity of particles, so called "polystyrene flake", was enough to shut the main pipe. Such an unexpected phenomena extended the maintenance period, and needed more steps to restore the monomer production line, so that there was a consistent loss of money. The personnel in charge of plant management did not notice any significant change in the monitored parameters, nor worst production quality or anything that could warn the personnel to take actions in order to prevent this anomalous accumulation of polymer, at least not before it became of such dimension as to be observed into the pipes and tanks. This polystyrene flake, due to its characteristic, floated in the process fluid and reached many items before any changes in temperature or pressure were identified. Management experience returns different hypothesis related to the causes and the starting point of polymerization trouble based on their experience and previous knowledge but not supported by data. The usual univariate approach have been applied but it gave poor information as it did during the production process. These are the starting basis for applying a multivariate process monitoring approach starting from analysis of historical data. We compared different period of production to evaluate the behaviour of the plant along the four years. Without assuming any *a priori* hypothesis, we could highlight drift, spike and other variation not detected into the previous univariate analysis.

3 Data and methods

3.1 Data retrieval

As for all sensor installed in Versalis company, ST20 instrumentation sends continuously values to the database. It means that database stores all the temperature, level, pressure, etc. for whole production line. An Excel add-in called PI DataLink allows to extract this data from the database, according to different settings, such as the time frequency, the digits to consider and other parameters that allow customization of data extraction. The first point that should be address was the dataset assembly, i.e. which variables to include. More than one thousand sensors monitor the entire styrene production, some of these control the core of the monomer line and other the utilities, some the storage tank and many other the auxiliary devices. So far, it is only possible to extract data pertaining to each plant section distinctly, this implementation suited the plant management univariate approach that select sensors to monitor on the basis of their prior experience. The different available sections that is possible to extract form database includes set up of “old” and “new” parallel lines, reactor and water recycling parameters. The on-line chromatography data, as amount of reactant/products, have also been considered, but the whole chromatogram was not available.

Thus, different data sets have been extracted to take into account all possible source of process variability, the data structure is common: on the data set rows are observations which describe the plant condition in a defined time range (daily mean, hourly mean, etc.) whereas columns correspond to a specified sensor, a level in a tank for instance, with the values aligned to the corresponding observation time. So a single entry in the data table contains the value corresponding to a sensor in a particular and defined time.

Three data sets have been considered, using the same sensors, time period and frequency of acquired data.. The first one called “last campaign” includes only the production from 2011 to the plant stop, i.e. August 2014. The second matrix, called “campaigns comparison”, includes also the previous production, from 2008 to the 2011; this second period allows a comparison with a "good" production and help the results rationalization. Every observation describes the mean condition of a daily production for both data sets. The third data set refers to the last two months of monomer production; management assumes that the undesired polymerization problems happened during this period. In this data set, every 30 minutes a new observation has been stored. Data analysis description and discussion of results is organized in sub sections corresponding to these three data sets.

3.2 Data analysis

Principal component analysis (PCA) can be considered the basic tool of multivariate data analysis [12]. Here is used for exploratory data analysis and to build a reference model for multivariate control charts development.

A typical dataset, from a continuous production process, consists of an ***X*** matrix of ***n x m*** dimensions where ***n*** is the number of time observations and ***m*** the number of variables (sensors output). PCA is a decomposition method that compress the original data (*observations x variables)* to few latent variables and thus returns a new data matrix (*scores x loadings)*. So, concerning observations, sample are no longer defined by the sensor value but by scores values; it means that a new space, the principal components space, describes all the observations and shall be used to compare the plant set up along the years. Trends in time (observations) could be highlighted by scores plot, either scatter or line plots; the loadings plot allows an evaluation of sensors correlation and relevance. By comparison of scores and loadings plots it is possible to identify the group of sensors that most influence the time evolution of the process and also the observations that correspond to specific trends. Before PCA data were autoscaled.

PCA models were obtained by the PLS toolbox [13].

Multivariate process monitoring consists of four steps:

1. Data collection: collect good quality observations, which correspond to normal operative conditions (NOC) as assessed by analysis of historical data;
2. Building a reference PCA model on NOC data: this model defines the good quality space and allows building T^2 and residuals (Q or SPE) multivariate control charts
3. Model Validation: external test with good and bad external observations
4. On line monitoring: it consist on projecting on-line acquired data (new observations/ time points) on the reference PCA model and estimate their T^2 and Q values and check if they are inside model limits; if not generate the corresponding contribution plots and identify possible deviance causes.

In general, NOC are identified as the production setting in which final product has the desired characteristics and no fault or anomalous production stops happened. In a continuous process it might also corresponds to the various set points and the PCA model should be created in order to accept or not this different condition. NOC refers, e.g., to a specific production ratio and the other available set up will be identified as abnormal condition despite the quality of final product is still good.

Thus, a reference model should include the desired production conditions, even more than one, in order to define a space able to discriminate undesired process set-up. These NOC often correspond to a specific period of production in which also abnormal condition might be included; such samples shall be removed from calibration set because they are not NOC observations and affect the model quality.

Once a reference PCA model on NOC data has been obtained production monitoring consists in checking two multivariate control charts. The T^2 chart that is based on Hotelling-T2 statistics [14] and monitors the variations within the model; these values return how much observation values differ from the model in terms of magnitude. It means that T^2 is able to identify observations in which covariance structure among variables is the same but some of their sensors present unusual values, too high or too low compared to model samples. In process monitoring, observations with high T^2 values frequently correspond, e.g., to a higher hourly production or to another imposed variation. In order to identify abnormal data the limits are calculated using the F-distribution with the desired significance level, normally set to 95%.

The Q-chart, also known as SPE [15], describes the distance of observation from the model hyper plane. It means that abnormal observations do not belong to the model covariance despite the single variables values might be similar to the one in the calibration dataset. High Q value should be considered carefully in plant monitoring because it might be related to the physical fault or change of a measured parameter, e.g. a dirty thermal exchanger with lower efficiency or anomaly in a regulation valve; but even a misreading returns high Q value. So, the operator experience is fundamental to identify the alarm significance. The Q limits come from the χ^2-distribution and, as in T^2 limits, significance level should be selected.T^2 and Q contribution plots are generated only for the out of limits observations. This plot supports controller in problem identification throughout the weight of each sensors in unusual plant condition.

4 Troubleshooting

Principal component analysis considered historical data from the last monomer production (from 2011 to 2014), the preceding campaign (form 2008 to 2011) and the data, which focus on July and August 2014. Some clear phenomena have been identified and the results are presented considering each data set. Table 1 reports the performance of the PCA models.

Model name	**PC number**	**X variance (%)**	**Obs. x Var.**	**Period (Freq.)**
Campaign 2011- 2014	4	72	1100 x 49	4 years (1 day)
Campaigns comparison 2008 -2014	3	63	2100 x49	8 years (1 day)
Chromatographic data	2	61	2100 x 6	
Focus on July – August 2014	3	66	1460 x 49	2 months (30 min.)
Low flow rates	3	48	1050 x 49	1 month (30 min.)

Table 1 PCA models of the three different data sets

4.1 “2011-2014 campaign”

In figure 1 is reported the first principal component (PC) from which it is possible to observe a different behaviour for the observations corresponding to all the 2014, circles and diamonds markers. In addition, the observations of year 2013 highlight the start of a drift (second half of down oriented triangles) when compared to the preceding time points. In general, changes in the production flow rate strongly influence plant setting and this influence is usually reflected in the first principal component. In this case, as can be gathered by the loadings plot, figure 2, the flow rate is not one of the relevant variables in the first PC1. Thus, the 2013 and 2014 variation does not seem to depend on flow variability. It means that some source of variation, larger than the flow rate change, influenced monomer production. It has to be verified if observed trends are due to a fault or to a varied parameters setting programmed for some reasons in that period. Between the second semester of 2013 and the first of 2014 the vacuum compressor shut down twice and the plant stops for some hours. In the scores plot it is possible to highlight the effect of both: the first one is observable because there is a change in the orientation of the 2013 observations (star symbol) and the other one because there is a drastic increase of the scores value on PC1 at the beginning of the 2014 (circles). It means that the plant stop and restart changed the plant behaviour and this effect can be observed by the strong variation of scores values in PC1.

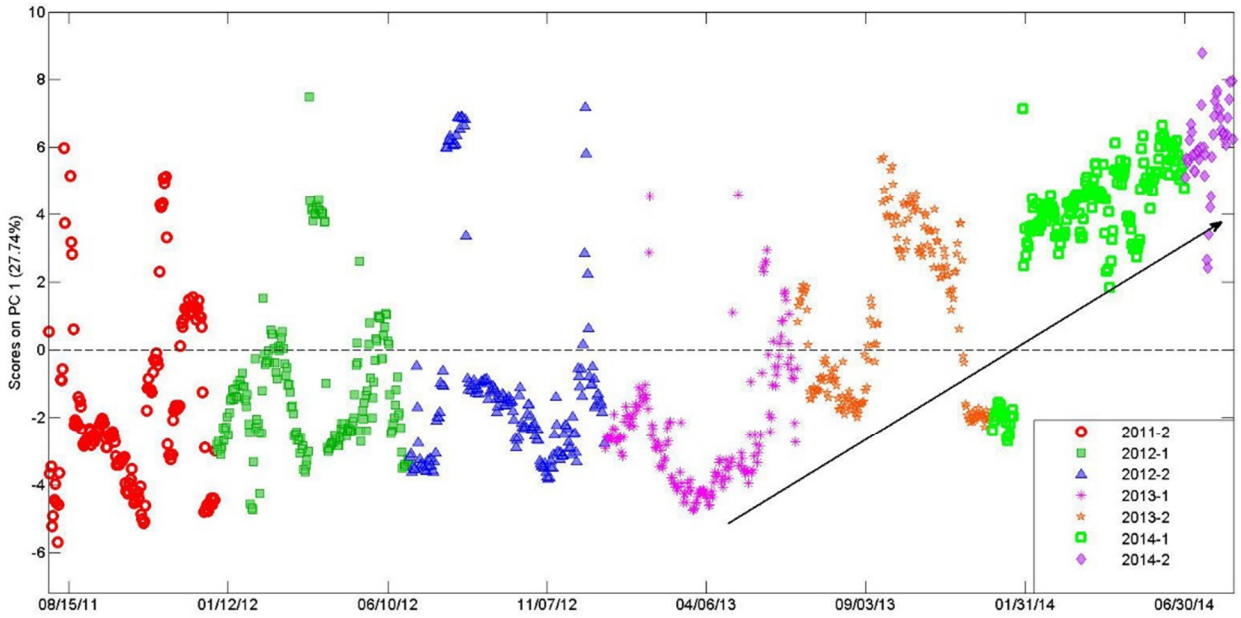

Figure 1 Samples in time order vs. PC1 scores value. PCA on 2011-2014 campaign. Dots change colour every six months. The arrow has been drawn to highlight the drift

The first component loadings plot clearly point out the relevance of each sensor. The sensors with high and positive loadings value are responsible for the shift of the 2014 time points at high positive scores and thus should have a variable profile similar to what observed in the scores plot; on the contrary high negative loadings values indicate variables whose profile goes in the opposite way.In figure 2 two red circles have been added to highlight sensors with high loadings values: temperature increases for D3202, D2202 and D202 during 2014 and this is coherent with the higher pressure measured in reactors. This resulted to be due to a small variation in the setting: the condensation section shows a global change, a different set up related to the higher pressure.

The blue circles drawn in figure 2 highlight sensors with high negative loadings values: the first thermal exchangers of both parallel lines, E3202 and E203N, decrease their temperature apparently without reason. The shut-down on first semester of 2014 caused the temporary stop of production and after that temperature setting definitely changed.

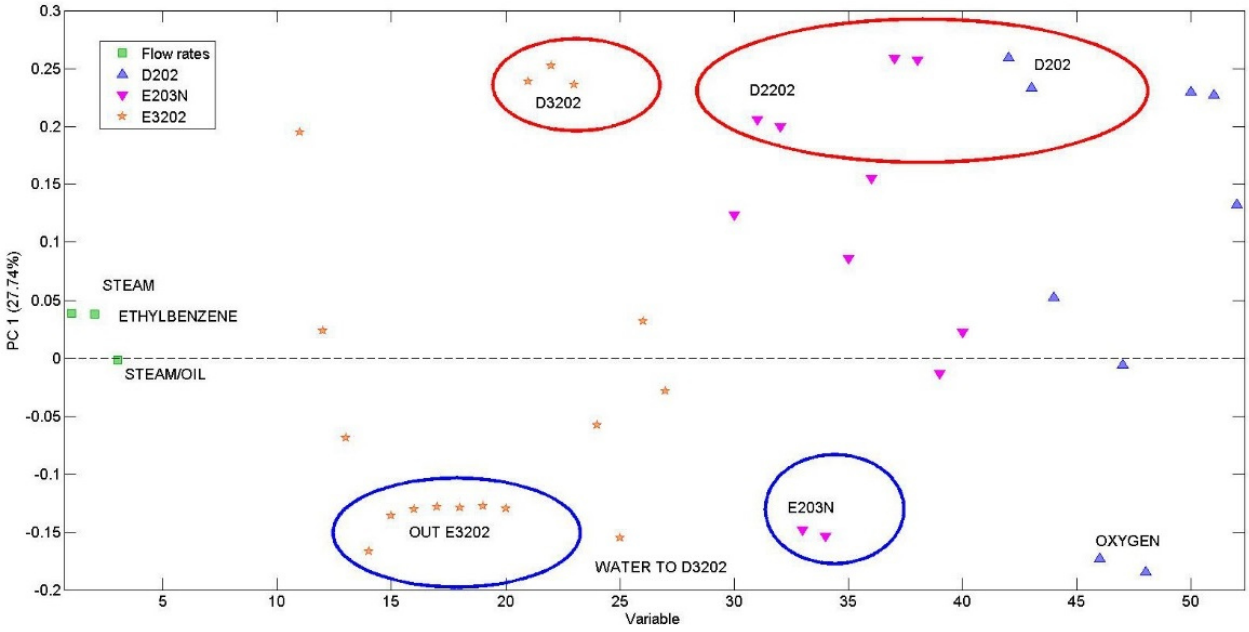

Figure 2 Variables number vs. PC1 Loadings. PCA on 2011 – 14 campaign

Thus, important information comes from the first principal component: the stop generates some variation and the global set up changed. Temperature relation (dots) with flow rates (line) changed subsequently to this event.

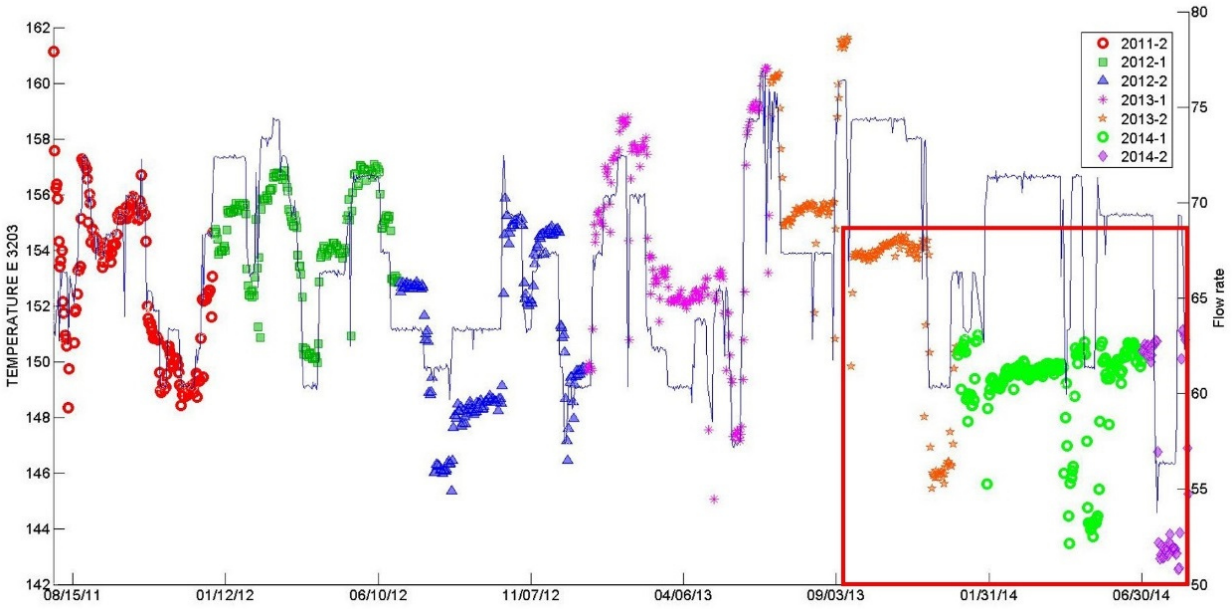

Figure 3 Samples in time order vs. Thermal exchanger temperature values vs. flow rates values (blue line). Dots change colour every six months

Among the variables with a high absolute loading value, there is one of the water flowmeters, water to D3202. This flow rates decreases while the temperature in figure 3 increases. As explained in process description, condensation section consists in two separate lines and only one of them underwent while the other remained more or less constant during all the production. Temperature, by contrast, changed in both lines so the phenomena appear not to be correlated, meaning that a lower water flow did not cause a temperature increment. Anyway, this information can be useful to assess the functionality of water supply items and the necessary water quantity, maybe less than usually used. The hourly production changes are highlighted by the second principal component.

Obviously, a change in styrene rate influence all plant set up. In normal condition, this variation would be prominent but in this case as shown by PCA it is only seen in the second principal component, figure 4, hence lower as contribution to total variance. From the point of view of plant manager this observation is not interesting because the styrene rate was intentionally modified to produce the desired quantity of styrene.

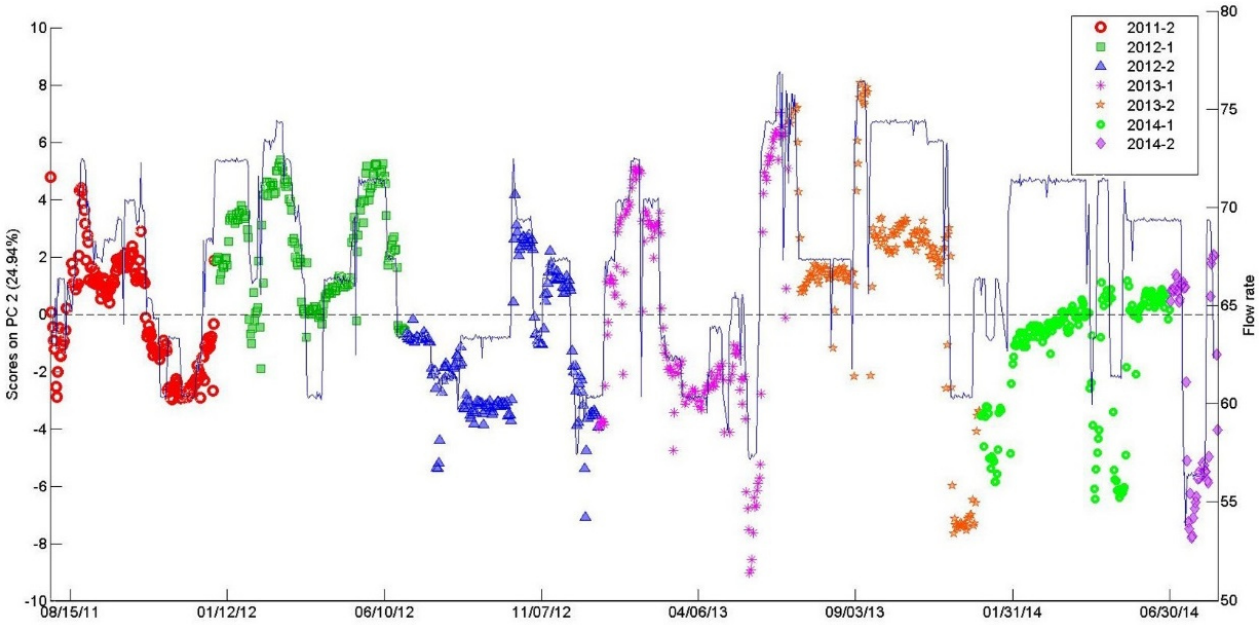

Figure 4 Samples in time order vs. PC2 scores (left). PCA on 2011 – 14 campaign. Dots change colour every six months. Samples in time order vs. Ethyl benzene flow rate (right), continuous line

In data analysis and in particular with continuous production, flow rate is fundamental to distinguish variation due to programmed changes in production set up and that due to accidental events. The catalytic reaction, core of styrene production, changes as catalyst ages along production.

PCA decomposition captures ageing information on the third PC. Again, that phenomenon is well known by the researchers and manager but it is worth noticing the ability of principal component model to describe all phenomena occurring during the production campaign; looking only at the third latent variable the plant operator could follow the ageing of the plant and planning for instance when the production should be stopped on an objective and quantitative basis. The PC3 profile clearly displays in figure 5 a plant drift form the 2011 to the middle of 2013. The last year appears stable or decreasing on the PC3 and describes a new behaviour after the first stop.

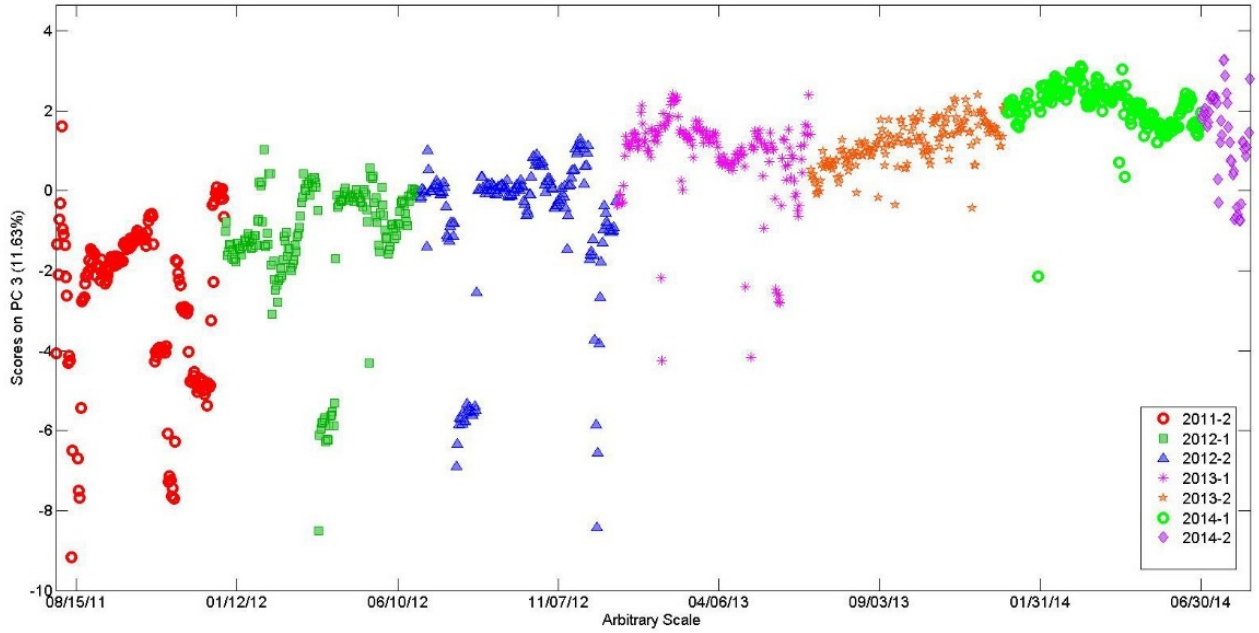

Figure 5 Samples in time order vs. PC3 scores value. PCA on 2011-2014 campaign. Dots change colour every six months

Information comes from the last modelled component, PC4, in which delta pressure sensor has a high influence as it could be observed in figure 6. The filter that separate solid residuum from the water are controlled by the delta pressure gauge that indirectly returns the level of dirtiness; from 2011 to 2013 every few days the delta reached a value above 200 mbar and the filter was cleaned. In 2014 this procedure appears less regular than before and takes more time to overcome 200 mbar, as shown in figure 7 where the values of 2014 observations are different from the rest .

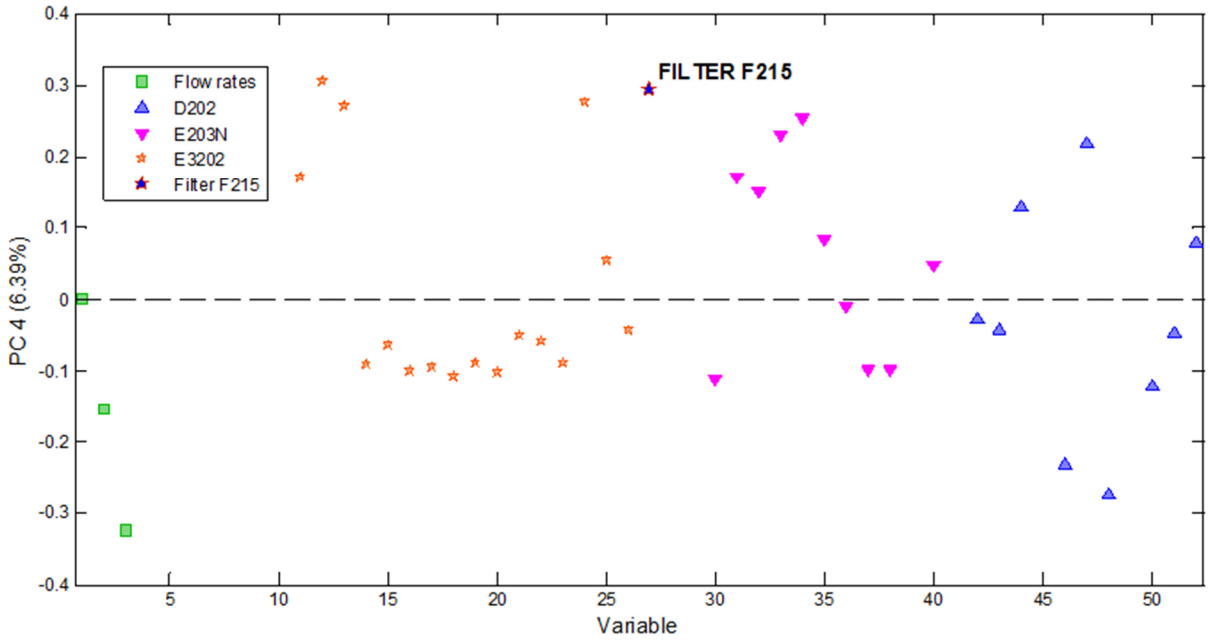

Figure 6 Variables number vs. PC4 Loadings plot PC1, PCA on matrix 2011 -– 14 campaign

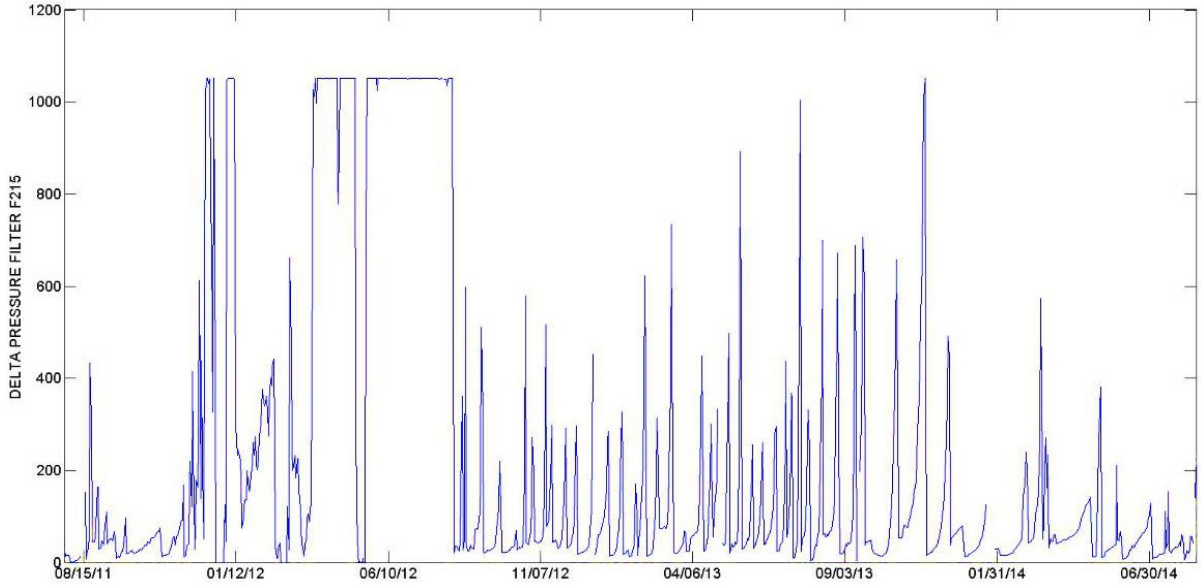

Figure 2 Samples in time order vs. delta pressure filter F215 A/S pressure

4.2 Campaigns comparison

This data analysis considered both productions, a production without polymers accumulation (2008 - 2011) and one with the polymerization problem that caused the plant stop. The aim is the detection of the fault, how it relates to the process parameters and possibly to discover the causes of polymers accumulation. To evaluate better the single campaign behaviour, the observations corresponding to the 2008 - 11 time points are represented with black circles and 2011 - 14 by diamonds. The scores plot PC1 vs PC2 in figure 8 (left), shows two well separated clusters. Considering that these campaigns cover a similar period, four years, and the production involved is obviously styrene monomers whit the same catalyst and reaction it is not expected to observe two distinct clusters but eventually two similar overlapping time trends for each campaign. Thus, PCA shows not ideal conditions and the first principal component scores plot, figure 8 (right), better shows the two clusters behaviour.

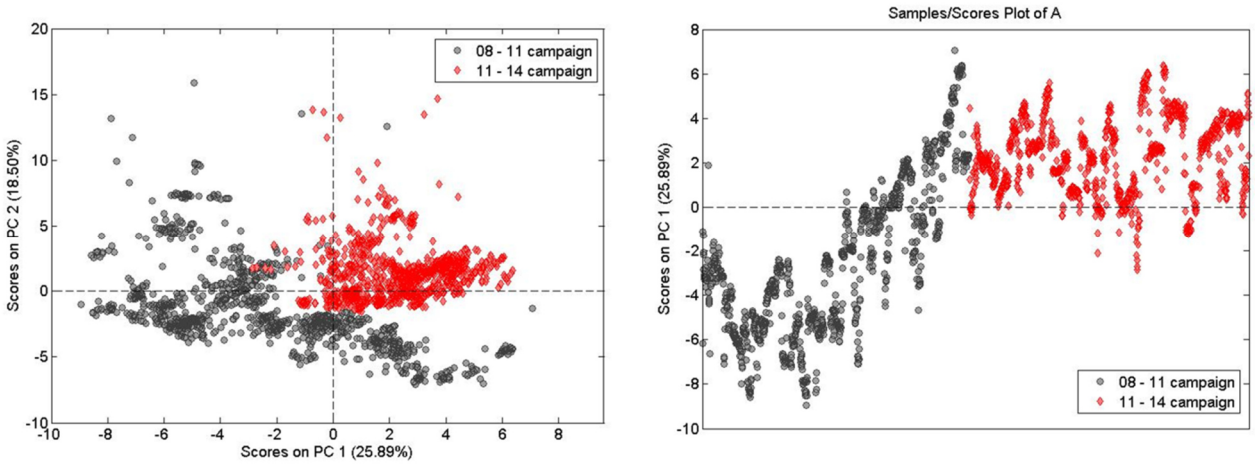

Figure 3 Score plot PC1 vs. PC2 (left) and samples in time order vs. PC1 scores value (right) PCA on 2008-2011 and 2011 - 2014 campaigns. Dots colour relates to campaign.

The first campaign shows a drift during the four years, at variance the second ones have almost coincident start and end production points considering their scores values. Moreover, all the 2011-14 points have a scores value similar to that of the last part of the 2008-11 observations; it means that, concerning the first principal component, during 2011 - 14 styrene pipeline had the same set up of 2008 - 11 ending conditions. As the drift follows catalyst exhaustion and controls production along ageing the bad production keeps condition equal to the final part of the first campaign, which is usually the most critical production period, due to lousy catalyst performance since it is close to exhaustion.

Therefore, the variables show increasing or decreasing time trends, according to their loadings in figure 9, for 2008-2011 campaign, while fluctuate around the same values during 2011-2014 campaign. Many sensors show a time trend as the first PC scores and it means that overall plant conditions changed between the two campaigns. For instance, temperature of the new line in 2011-14 was around the maximum value reached in 2008-11. The change in plant conduction seems linked to the decision of changing one setting: the original set flows 80 tons/hour of hot water, around 100 °C, and heats with steam in a second heating step. The condensed water comes out to the exchanger and generates a cooling effect, temperature decrease up to 130 °C. In order to decrease exchanger-fooling effect, that the plant manager supposed related to water flow, the setting has been modified so that the discharge of only 5 tons/hours of condensed water was programmed and this made the temperature to increase up to 160 °C.

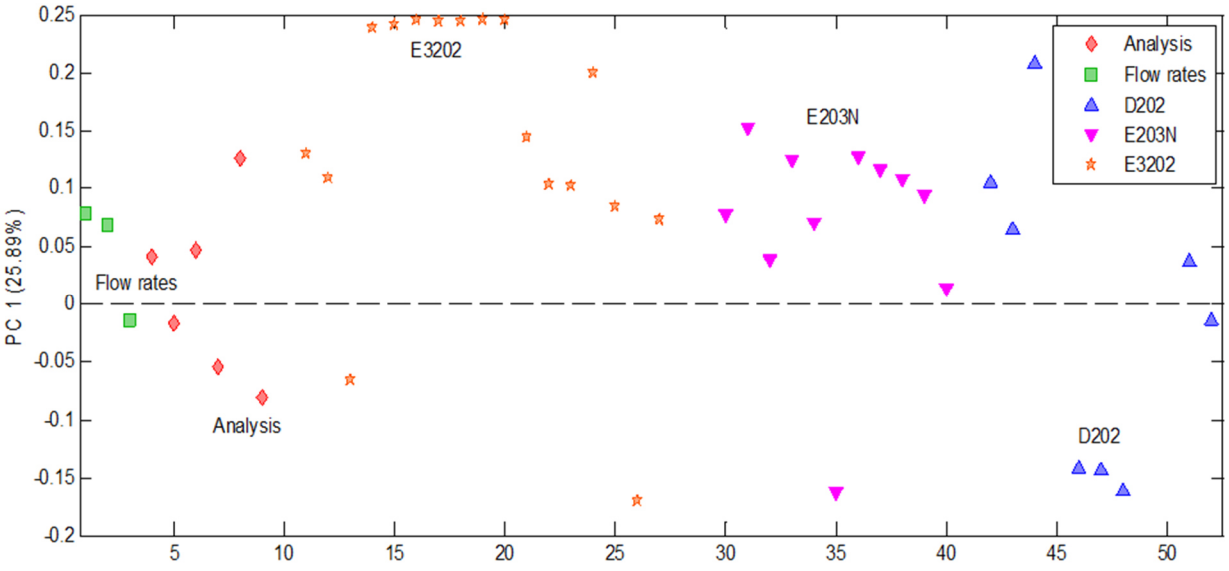

Figure 4 Variables number vs. PC4 Loadings plot PC1, PCA on 2008-2011 and 2011 - 2014 campaigns.

Many of the other sensors showed the same behaviour because of this modification: the higher temperature thus generated new production set, a different setting that returns an equal production quality. The lower discharge, from 80 to 5 tons/hour, imposed a partially new equilibrium, in order to achieve the same production quality. In addition, the main water thank level, D202, shows a profile coherent with the scores profiles of first PC. During 2014 campaign the water level was 10% more with respect to the level imposed on precedent campaign. Tank level kept the last value imposed in 2011, just before plant maintenance. PCA was done also on the on line chromatographic data collected during the two campaigns. The data consist of di-ethyl benzene, paraffin, benzene, toluene, ethyl benzene and styrene amount values.

The loadings plot is shown in figure 10; it can be observed that DEB and EB vary in the same direction and opposite to toluene and styrene. This behaviour matches with course of reaction in the process. In fact, in monomer process a higher amount of DEB comes from a higher ethyl benzene EB concentration and at a certain point in production DEB amount was set to double. The styrene and toluene amount decrease as expected since DEB and EB could not convert completely into styrene or toluene in the new conditions set. This second PCA analysis describes an imposed variation, the change in DEB limits ordered by plant management.

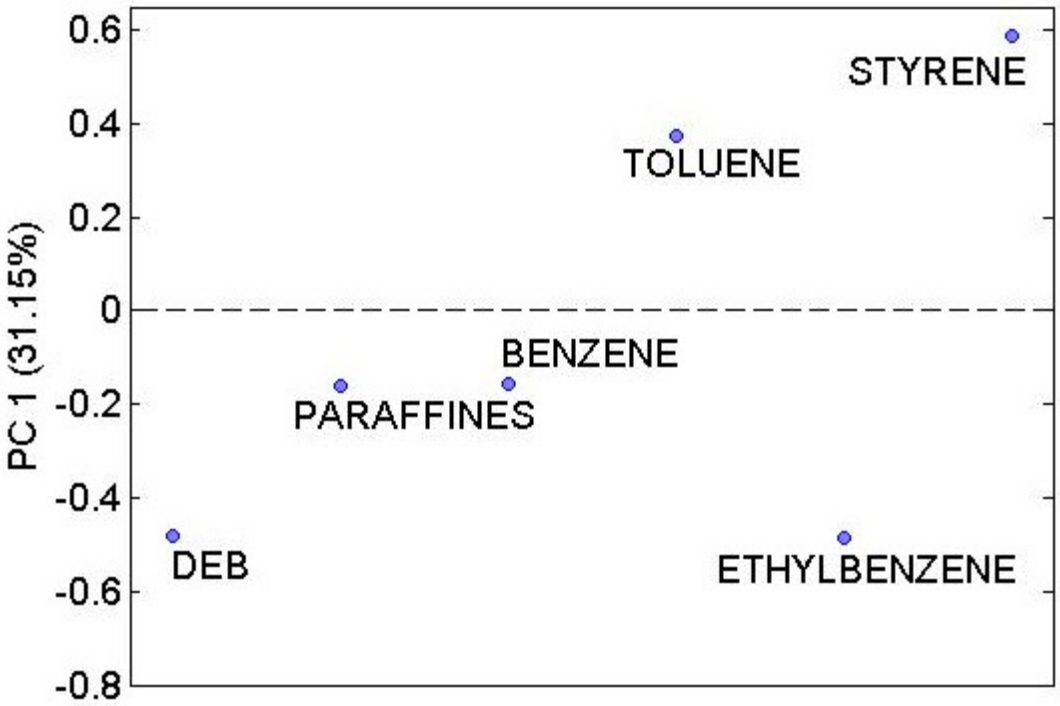

Figure 10 Loading plot PC1. PCA on 2008-2011 and 2011 - 2014 campaigns of chromatographic data

When chromatographic data are added to the campaign comparison data the resulting PCA did not show substantial differences with respect to what can be inferred by the separate analysis of the two data sets.

4.3 Focus on July- August 2014

The plant management was interested in focusing data analysis specifically on the production period close to the end of the 2014 campaign; they supposed that accumulation of polymer problematic became serious or even started between July and August 2014. Their prior idea is that this problem is connected to the imposed low flow rates, setting that occurred in concomitance with approaching the end of catalyser life, these conditions, according to experience, might provoke styrene polymerization. Thus, a dataset has been assembled concerning only last production period, from July to August. The scores plot of the first versus second latent variable shows that three observations are quite different from the rest, figure 11.

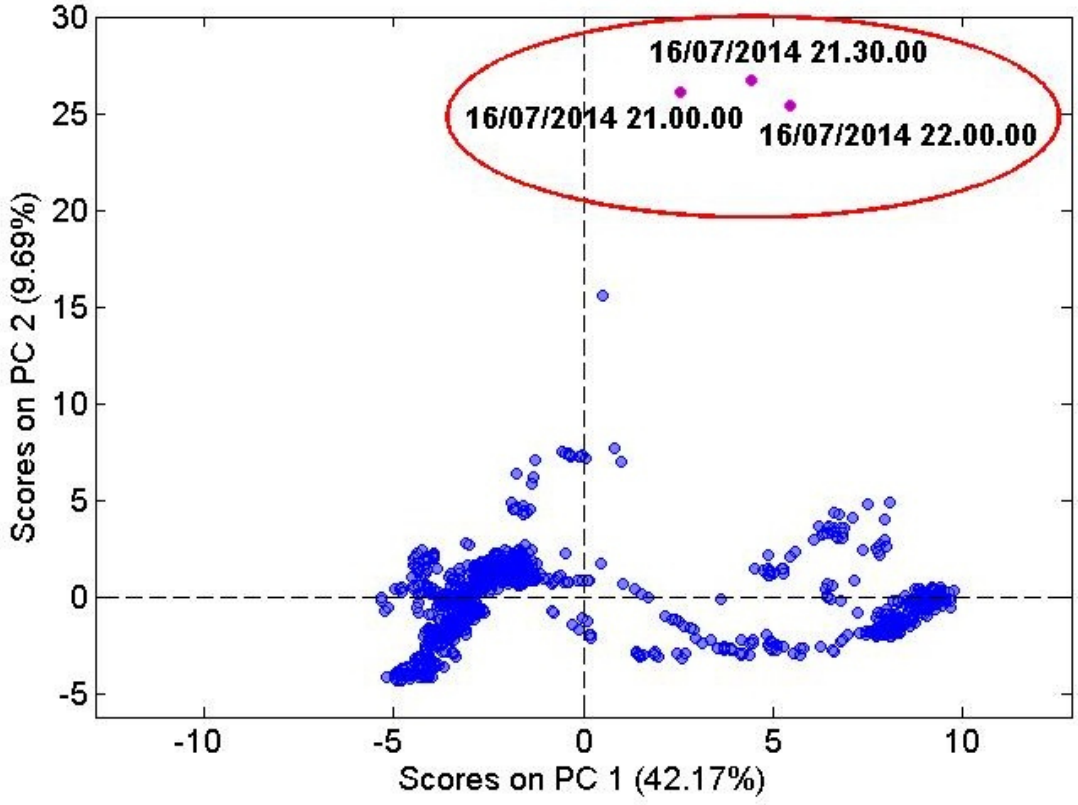

Figure 11 Score plot PC1 vs PC2. PCA on July - August 2014 focus. Circle highlights observation of 16/07/2014 relative to a sudden stop

For one hour, the old pipe temperature reached 105°C instead of a normal set of 60°C. Plant controllers well know this phenomenon: an unwanted compressor stop increases pressure and as a consequence temperature increased. A pressure gauge, sensor not included in the ones considered in data analysis, shows a higher pressure in correspondence to the 16th of July and in particular on the three observations indicated. The pressure gauge does not belong to the condensation section and for that reasons was excluded from the analysis and on operator suggestion, afterwards checked.

As supposed, a priori selection of variable might create a lack of fit in dataset that some time is covered by process knowledge. Anyway, the proposed explanation seems confirmed. On the other hand, the new pipe line did not behave in the same way, in fact during the compressor stop the thermocouple sensor on this line registered the same temperature as before the stopping. The two lines are equal in flow rates and pressure and should act identically. Also, the registered temperature of 105°C in old line is higher than normally expected if the temperature rise should be ascribed only to the higher pressure observed. A different explanation could be that something was blocking the monomer flow in one line, i.e. the new ones which did not show any change in temperature, so that the other, the old one, was heated more and a different behaviour took place. Thus we decided to split the data set, in two groups corresponding to high and low production rate and to consider the latter in a new PCA. This could be useful in order to compare the productions in equal conditions, since flow rates deeply change the overall set up. A PCA model built on time points corresponding to low flow rate, which are the ones supposed critical for accumulation of polymer, show a cluster on the PC2 vs. PC3 scores plot (green points on figure 12, left) corresponding to the 18th and 19th July and to lower monomer temperature, 20 degrees less, as shown in figure 12 (right). During that period the plant operators feed the ethyl-benzene in liquid form instead of gas. The styrene production keeps the same production quantity but from different raw material form, liquid instead gas; the liquid alimentation decreased temperature because gas provides heat. Quality still remains inside the requested limits.

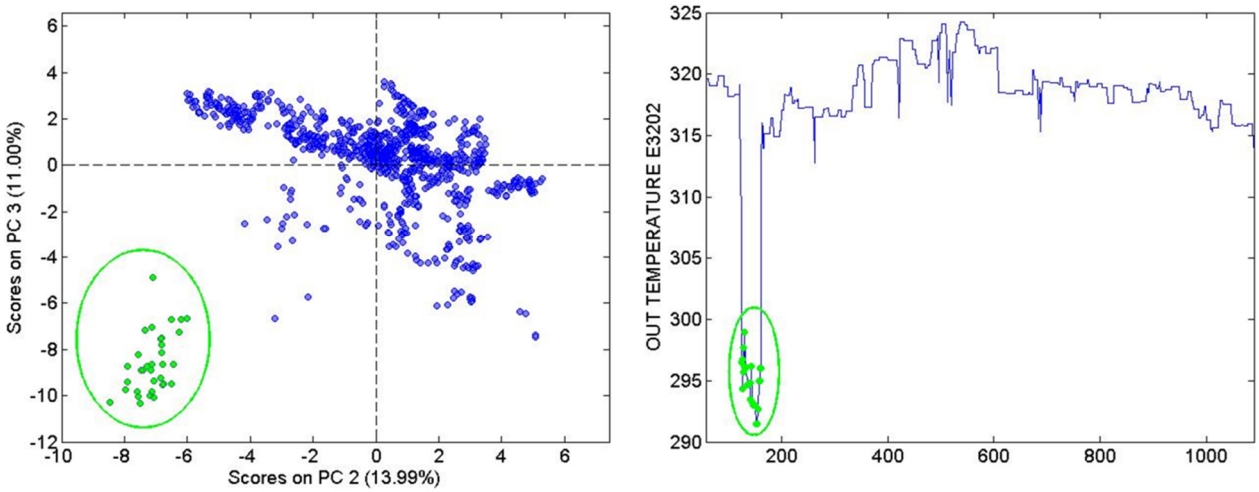

Figure 12 Score plot PC2 vs PC3 (Left). PCA on low flow rates, July - August 2014 focus. Samples in time order vs. temperature exchanger 3202 (Right). Circles highlight liquid ethyl benzene

4.4 Final remarks

Summarizing the data analysis results taking into account the knowledge of the process and of the different items some important conclusion can be drawn, as reported below.

The 2011 to 2014 campaign highlights different plant set up in the middle of 2013 and on the beginning of 2014. Both changes are due to sudden plant shut downs that degenerate, each time, in different operative conditions for styrene production. No correlation with flow rates was observed and it could be stated that plant stops had an impact in monomers production; next to the events production had a new set up that realized a good quality monomers but from a new plant equilibrium.

After the stop the sensors revealed big differences in the two lines parameters. Following the 2013 shut down, many temperatures changed the set values. The pressure in reactor strongly influences the temperature but that correlation does not explain all observed variations. The inversion in first principal component profiles (Fig.1), corresponds to the first stop, and the high scores values reveal the high impact provoked by shut down, such as the irregular variation into the third PC (Fig.5), assumed as proportional to the catalyser exhaustion, following the previous consideration.

Other two phenomena might be correlated to the observed variations. From the beginning of 2013 water flow rates decrease, in particular the amount of water that fills tank D3202. The other phenomenon is the delta of pressure in water filter: from the beginning of 2013 the pressure profile starts to change and this can be explained assuming that it took more time for dirty to accumulate and shut the filter surface, so 250 mbar of pressure were reached in longer time period than in previous years. Both events cannot be ascribed to plant management, i.e. they do not depend on plant modifications programmed or actuated as actions.

The availability and comparison of the two campaigns gives much information on historical plant behaviour. Principal component analysis immediately returns an interesting phenomenon: many sensors drift, as expected, during the 2008-2011 campaign while kept their values oscillating around the one reached at the end of the period in the 2011-2014 campaign. It means that monomer production during 2011-2014 went on with operative conditions close to final set up of 2008 – 2011, for most of the sensors that were considered in data analysis. A modified item, thermal exchange in the so called new line, induced a change in the production setting and the modified set up appears to be identic to the most critical plant conditions experimented at the end of the preceding campaign. The first principal component perfectly follows the catalyst consumption from the 08’ time period until the substitution with the new ones, for the next campaign, 11’ period. Notwithstanding the

catalyst renewal, the operative settings in the new production campaign 2011-2014 stayed for nearly four years close to final conditions of the preceding campaign.

The D202 level remained at 55% in 2011 - 14 instead of 45%, which has been set in 2008 - 11. D202 stores water that is used in condensation phase and the same tank is used to separate polymers particles. Plant manager considers not critical the D202 level and assumes that such a variation should not be enough to affect production. This is why controller imposed the new value at the beginning to 55% and did not modify it until maintenance.

Also the on line chromatography measurement shows differences between the two campaigns: diethyl benzene and ethyl benzene had higher concentrations in the 2011-2014 campaign whereas styrene and toluene concentrations were lower with respect to the preceding campaign. In that case, the values change in order to reach the new standard in DEB concentration (imposed by production management) and consequently EB, styrene and toluene got a new concentration to. In ethyl benzene stream, DEB amount doubled and, as consequence, styrene and toluene content decreased.

The focus on July and August 2014 offers some interesting information also because of the higher frequency of time points, an observation every 30 minutes. Principal component analysis immediately highlights an unexpected temperature augment in the old line and a constant value in the new one. This effect was due to a compressor shut down. The two lines had a different behaviour as a consequence of this shut-down, most probably the new line was already dirty, so experienced no variation in temperature, while the older one was almost clean and strongly changed its temperature. Therefore, according to the hypothesis, the polymerization took place before 16 July and involved mainly the new line.

The observations corresponding to low flow rate return information on plant set up: the change of ethyl benzene alimentation from gas to liquid lead the second exchanger temperature to decrease. The physical state of raw EB frequently changes but it does not appear strictly related to the encountered polymerization problem.

The plant manager and the researcher department have evaluated positively the outcome of the multivariate analysis conducted on ST20. In fact, as shown, on this basis it was possible to derive a lot of suggestions for plant controller. It is worth noticing, that the information could be derived by adopting different data analysis frameworks with respect to period and frequency of time points considered as function of the sought objective. For instance, focusing on July and August with higher time frequency it has been possible to observe the different lines behaviour; this phenomenon was not evident when analysing the whole 2011-14 data due to the lower time frequency and longer period taken into account.

In the analysis of 2011-14 data, nothing strange was detected concerning the D202 level but comparing this data with the 2008-2011 ones, by using a global PCA model, the D202 level immediately appears as very influent in the first component loading plot, thus responsible of the main difference between the two campaigns. Even more important, multivariate data analysis allows detecting the effects due to variability always occurring during production, either due to natural fluctuation or programmed modification of set-up to better handle requirements, e.g. changes in material grade and variation in flow rates, and that frequently are not detected nor understood in terms of how propagate/involve different sensors. With multivariate data analysis each phenomena could be managed taking advantages of all different tools from data organization, pre-processing and appropriate modelling. Data analysis might help more than actually does in problem solving and in plan monitoring as demonstrated in this application.

5 Adaptive process control whit MVPCA

On the basis of the previous results it seems feasible to monitor ST20 production implementing a multivariate monitoring tool, whose implementation is described in this section. To support on-line control the known variability should be managed in continuous in order to obtain the desired information. To this aim we developed a homemade interface in MATLAB environment that was then compiled to create a Windows executable file that could be used in the plant without requiring MATLAB installation in the PC present there and suitable for end-user. This general scheme for building process monitoring models (multivariate control charts) outlined in the method section, need some modification to take into account that production set-up may undergoes changes not necessarily due to undesired drift or accidental faults but also to “natural” system evolution or programed fluctuations around reference values for some specific production need. This varying conditions, generally small, reflect in a continuous change of the NOC, i.e. the "good" quality set-up need to be updated to continue to well represent the actual space of the desired production operative conditions. This can be accomplished by reference model updating, i.e. in re-adjustment of multivariate control charts confidence limits. In particular, we decided to use moving windows principal component analysis, i.e. the input matrix (reference set) on which PCA model and hence multivariate charts are based continuously change including new time points and discarding a corresponding amount of the oldest ones. In this specific application a new observation is added every time point, so that it become the first row in the input data matrix and the oldest observation in time, from the previous reference set, is deleted.

The monitoring tool developed allows flexibility. The user can set the frequency of predictions, i.e. how often the acquired data should be projected on the model, and the time range on which to build the reference model, i.e. the size of time window in PCA or in other words how many past observations are to be included in PCA reference model building.

The analysis of trouble shooting above, shows that if one want to monitor a new campaign, the 2008-2011 data can be used to build a reference model since no fault or anomalous conditions occurred. However if one want to have a warning of catalyst being close to life end , the reference model should consider the observation before the last steep in the observed drift, so that out of limits observation could warn on catalyst exhaustion being close. As well in correspondence of programmed changes in production setting, e.g. amount of reactant, reference set has to be updated until observations will start to re-enter limits.

Thus, once frequency and time range for reference model parameters are set, then the model continuously update as explained above. As an example, if plant surveillance personnel are interested in long-term drift, they should select a week or a month as time range and to perform prediction every 6 hours; on the contrary, if focus is on plant conduction a daily time range and prediction every few minutes are to be recommended. In this way, it will be possible with moving windows PCA detecting slow drift, as fouling factor, or suddenly events, as decreasing/increasing of flow rate. Other scripts, besides this main routine, allow input/output management, in a user-friendly interface.

Plant variations are normally detected by both parameters for sure at the beginning of a change, but Q returns inside limits faster than T^2. Imposed variation implies one or more operator actions that increase or decrease variable values, maybe more than described by calibration set; in the same way, variations might modify sensors covariance. After a while, plant finds a new equilibrium in which probably the variable correlations is quite similar to the previous process condition but the values are still different from the modelled ones. For this reasons T^2 takes longer than Q to returns inside confidence limit. Operator should pay attention in all cases in which one or both distances go out of limits without an imposed variation. Anyway, Q returns information on plant behaviour much difficult to detect with classical univariate monitoring; valve aperture, pump absorbance and other parameters that show how items are working strictly relates to their function and the covariance between these represents their efficiency and correct operation. Alarms in Q parameter might suggest in advance an equipment troubles or worse working condition through change in their correlation structure.

5.1 Data flow and code description

The MWPC (Moving Window Process Control) concern three main step:

- Database and system indexing
- Multivariate model building and limits calculation
- New observations prediction and plot updating

The first step runs only once at the program starts. End-users have to prepare, as they usually do when they need to obtain univariate charts, or other statistical analysis to be run with commercial software available at the company, an excel file reporting the sensors name (i.e. variables to be considered in PCA model building) and optionally some historical data. The MWPC applicative on the basis of this information stores the sensors addresses in Versalis database, and assign a unique link for each variables to be used in multivariate analysis. A graphical interface open where the user can enter the time range and the prediction time frequency.

During the first iteration, if the excel file provided contains historical data (or if they are provided in Matlab file format from previous multivariate monitoring) these are loaded as input data. The software thus performs PCA and automatically defines a proper number of components. Many automatic methods have been proposed for setting the optimal number of latent variables [14 15], each one with some advantages and peculiarities. Our choice, is based on heuristic, and considers that a component has to be retained if it account for 5% more variance with respect to the preceding one.

To improve the adaptation to new conditions, while program calculates PCA model with all available samples eventual outliers in T^2 and Q are detected. These observations are excluded from the calibration set and a new PCA is performed and used for prediction of the subsequent observations.

The implementation of the MWPC model allows to detect anomalies/faults and to adapt to changes in plant behaviour until newer set-up becomes the NOC. Transition phases are better detected and new conditions accepted in less observations points. The transition from two, different but equally "normal" conditions of the plant is summarized in three steps:

- New good production conditions are at first detected as abnormal by the reference model in use because different from the previous NOC
- Moving the window, i.e. including new observations in the reference set, allows the Model to enlarge its domain since both old NOC and observations corresponding to the new good production conditions become sampled and distance limits are updated accordingly.
- Progressing in time, old NOC conditions come out of the distance limits because the "new good production" observations are currently more represented in the reference set until the model is based on the "new" NOC

The PCA results stored by the MWPC applicative are: scores, loadings, T^2 and Q (SPE) confidence limits values, and T^2 and Q (SPE) contributions for all variables (sensors). At prediction stage, the software automatically download new data from the server, with the frequency set by the user, and project it on the reference PCA model. The following three graphs are automatically shown to the end-user through the GUI:

- On up-left the T^2 values of historical data and the predicted ones. Values appears in chronological order and scaled on the limits, it means that an observation higher than one is out of the set confidence limit (95%).
- On down-left the SPE values of historical data and the predicted ones; such as T^2 scaled data appears time ordered.
- On the right the PC1 vs. PC2 scores plot shows observations score values.

Data colour depends on date: oldest points are blue and colour vary to yellow than to red with gradient. This "jet" colour, as defined in MATLAB code, facilitates reading and are the same into the three graphs.

In figure 13, there is an example of MWPC with parameters set to 3 months' time range for reference set and 1 hour as frequency of acquisition, i.e. a new observation each 60 minutes. On the left side, figure 13 shows that some data were out of T^2 and Q limits. Such limits have been calculated on the NOC observations from one hour before the predicted sample to 3 months before it. In these cases program returns the contribution plot in which operator finds the effect of each sensor and might understand the reasons of the out of limits observations. Every red dot in the figure indicates an observation with at least one distance out of limits.

Referring to the T^2 chart, operator observed that observations from 168 to 215 were out of the limit and their distances constantly decreased until it returns below the confidence limits. Observations until the point 216 have been excluded but the subsequent observations forced model to accept such condition as the normal operative. These kind of situation describes a controlled change in which plant passes from a production setting to a new one; e.g. from a higher to a lower production ratio. Initially observations will be detected as anomalous due to their absence or smaller number in the reference set but after a sufficient period, they will become the new NOC as T^2 control charts shows. Moreover, T^2 values keep decreasing and it means that PCA is modelling this new condition and detecting as abnormal the older one. The spikes in Q chart describes changes in covariance structure and shall be controlled by operator via contribution plot in order to understand if such variation is real, and not a bad reading, and their consequences in plant operation. The PC1 vs PC2 plot in the same figure helps the global understanding of production using their respective loadings plot that are shown as interactive plot, figure 13. Scores plot shows changes trough production and some spikes, e.g. points 1112, 2229 and 2327, that operator easily detects via Q control chart.

Spikes are frequently due to bad reading or sudden events that modify drastically sensors value but for a short period, often few hours.

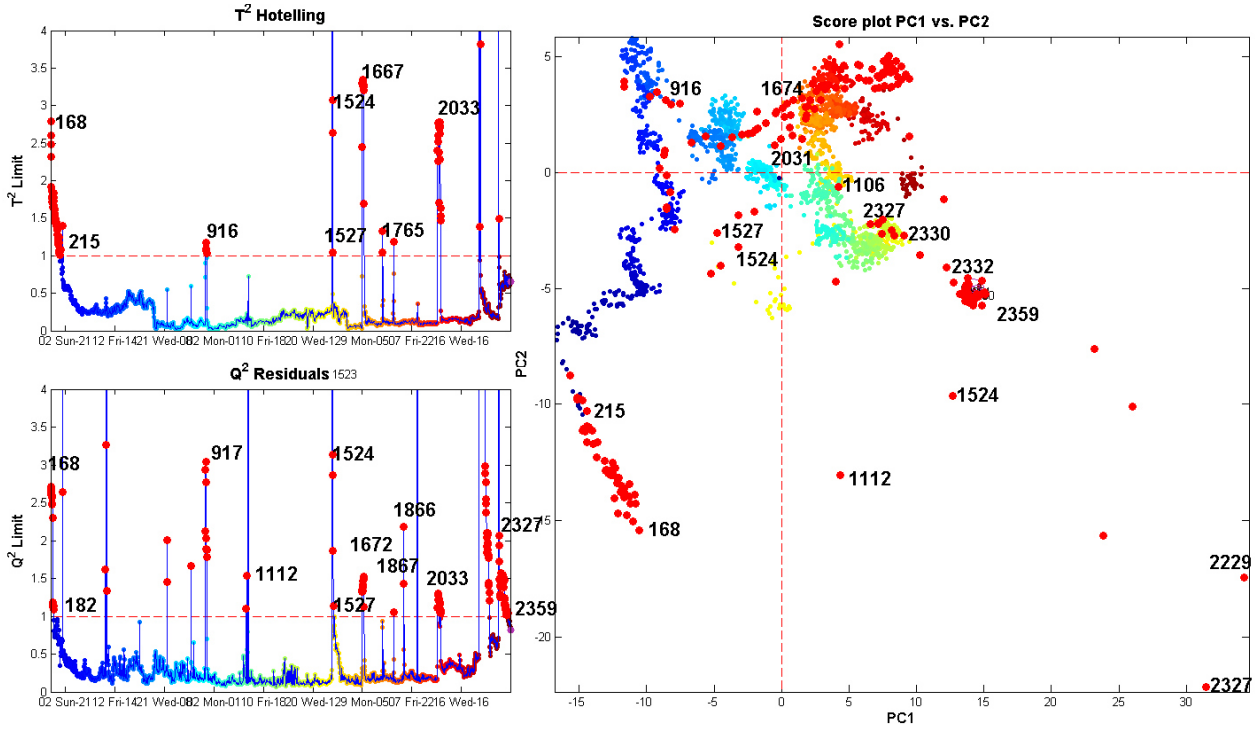

Figure 135 MWPC main figure. On the left sample in time order vs. T^2 and Q2 charts. On the right PC1 vs PC2 score plot. Example of three months of production monitored with frequency of one hour. Red dots correspond to the observations out for at least one of the two limits.

Software creates a second figure; an example comes from figure 14, in order to help the user in observation analysis. Two plots, on top of each other, allows visualization of six different graphs:

- First principal component score profile
- Second principal component score profile
- First principal component variables loadings
- Second principal component variables loadings
- T^2 Contribution plot
- SPE Contribution plot

The first and second latent variables describe variation in production and this observations vs. time plot highlights change in plant setting or in production behaviour. Principal component plot coupled with sensors loading values returns an exhaustive explanation of plant condition: this two plots give information on process drifts and on implied sensors. Contribution bar plots are normalized with the reference set (95% percentile) for each variable so every value respect specific below unit and vice versa. In default setting, contribution plot shows the most recent observation outside the confidence limits; the *TQ_Value* button generates a dialog window and user could select any abnormal (outside of model limits) past data by observation numbers.

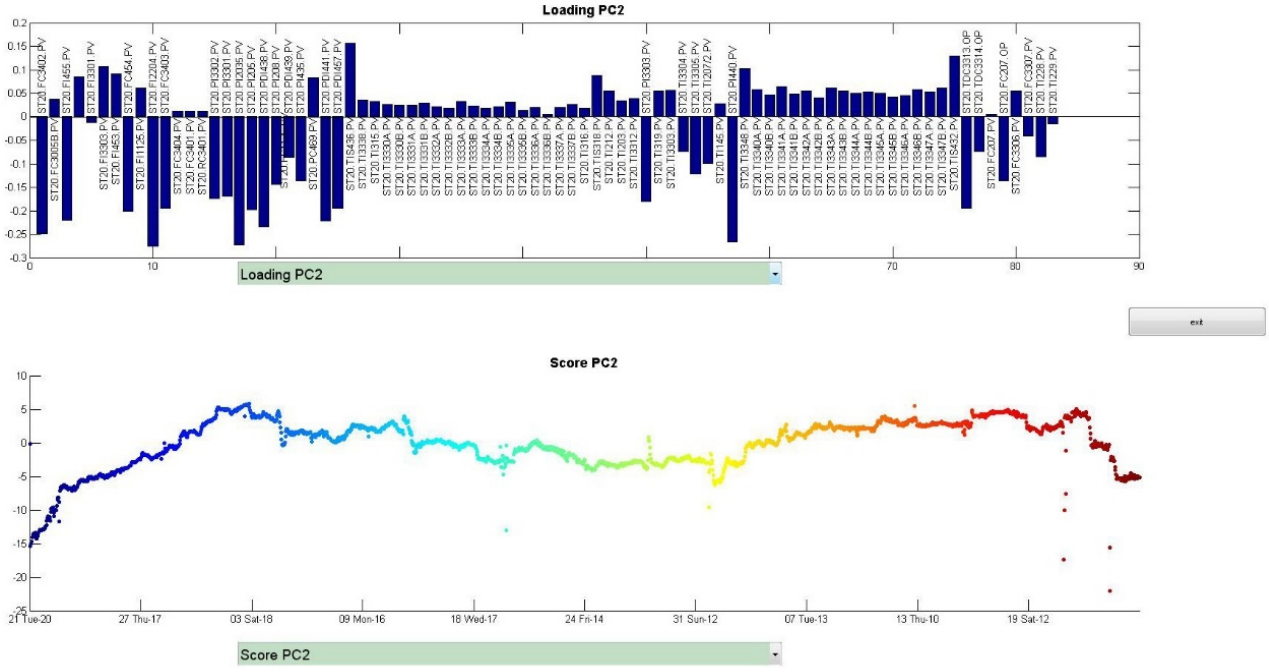

Figure 14 MWPC interactive figure, loading PC2 (above) and score PC2 (below) has been selected from menus

The PCA model is automatically updated and the data for the next iteration are moved one time point as explained above. A so-called exit button allows the shutdown and ends program iterations if the end-user need to do this.

References

1. Kresta, J. V., MacGregor, J. F., & Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance. *The Canadian Journal of Chemical Engineering*, *69*(1), 35-47.
2. Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., & Yin, K. (2003). A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers & chemical engineering*, *27*(3), 327-346.
3. Tomba, E., De Martin, M., Facco, P., Robertson, J., Zomer, S., Bezzo, F., & Barolo, M. (2013). General procedure to aid the development of continuous pharmaceutical processes using multivariate statistical modeling–An industrial case study. *International journal of pharmaceutics*, *444*(1), 25-39.
4. Russell, E. L., Chiang, L. H., & Braatz, R. D. (2000). Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *51*(1), 81-93.

   Singh, K. P., Malik, A., Mohan, D., Sinha, S., & Singh, V. K. (2005). Chemometric data analysis of pollutants in wastewater—a case study. *Analytica Chimica Acta*, *532*(1), 15-25.

   Kourti, T. (2002). Process analysis and abnormal situation detection: from theory to practice. *Control Systems, IEEE*, *22*(5), 10-25.

   García-Muñoz, S., Kourti, T., MacGregor, J. F., Mateos, A. G., & Murphy, G. (2003). Troubleshooting of an industrial batch process using multivariate methods. *Industrial & engineering chemistry research*, *42*(15), 3592-3601.
8. Skagerberg, B., MacGregor, J. F., & Kiparissides, C. (1992). Multivariate data analysis applied to low-density polyethylene reactors. *Chemometrics and intelligent laboratory systems*, *14*(1), 341-356.
9. Wang, X., Kruger, U., & Irwin, G. W. (2005). Process monitoring approach using fast moving window PCA. *Industrial & Engineering Chemistry Research*, *44*(15), 5691-5702.
10. L.Trentini, A. Longo, F. Pasquali, Encyclopedy Treccani, Vol. II refining and petrochemicals, *"Thermoplastic styrenic polymers"*.
11. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1), 37-52.
12. Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S. *PLS_Toolbox for use with MATLAB™*. Eigenvector Research, Inc., Wenatchee, USA, 2006.

13. Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and intelligent laboratory systems*, *28*(1), 3-21.
14. Westerhuis, J. A., Gurden, S. P., & Smilde, A. K. (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, *51*(1), 95-114.
15. Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, *20*(4), 397-405.

V

Quality monitoring

Continuous parameters estimation

Content

5.1 Application introduction

Product quality is one of the most relevant parameter that defines price of a material and its popularity on the market. For sure, advertising, good market position and other conditions might increase or decrease the selling quantity but, without any doubts, quality seriously affects the success of a material in long-term period. Producers measure and certify their product before purchasing in order to guarantee at least the minimum quality level of material. Anyway, the quality is defined during production and should be monitored in that phase, as many companies currently do; with a certain frequency, operator takes a small amount of production, sufficient to characterize the material. In this way, plant controllers perfectly know which the quality of production is. Management decides analysis schedule taking on consideration measurement costs, laboratory technician availability and normal plant variation. This returns often a lower frequency than desired but is a widespread situation in petrochemical field. A lack of information in production might causes bad quality product, clients complains and consequently the reduction of gain margins. Such an unwanted situation finds solution in on-line measurement: instead to perform control analysis in laboratory, a dedicated on-line instrumentation measures one or more parameters, some examples come from the literature [1 2 3 4 5 6]. The out coming information has a higher frequency: for on-line chromatography, analysis takes less than an hour, spectroscopic methods measures sample in few minutes and in the simplest case of titration and pH meter data returns in the order of seconds. In Versalis only few productions have on-line instrumentation and nearly all of them supports environmental controls or security procedure. Continuous measurement certainly helps plant controllers, might reduce lower quality production and grade transitions. On the other hands, it has considerable cost, needs expertise, maintenance and some critical productions require only custom or particular (and more expensive) instrumentation.

The expandable polystyrene (EPS) production fits the last case for its peculiarities: some particular instrument might help production but the product/process features do not permit classical on-line monitoring apparatus (NIR, UV, Raman). Thus, the idea was to estimate the main quality parameter, the molecular weight (Mw), by using the available information from the plant: the sensor database and the available past laboratory analysis.

Thus, this application concerns prediction of polymers Mw by comparing the process variation with the materials change. Unless an obvious lower precision with respect to off-line laboratory measurement, there are a lot of advantages related to the estimation of quality from plant sensors:

- It does not need any investment, new sensors installation or specific items are not necessary
- Calibration is built up with the existing measurements and the new scheduled, it means without any extra laboratory costs
- The prediction frequency depends on sensors and/or controller requirement because the calculation time is irrelevant

This Chapter focus on Mw prediction by PLS regression based on process variables data and highlights the main issues experienced during analysis and calibration set up. Although it may seem a straightforward approach there are not many applications in which product quality is actually estimated by process variables.

5.2 Process description

Expandable polystyrene (EPS) is among the biggest commodity polymers produced in the world. It is solid foam with a unique combination of characteristics, like lightness, insulation properties, durability and an excellent processability. Particle foams based on EPS have proven to be a suitable material for the building thermal insulation over the past more than 50 years and are the most widely used insulation material in the market after glass wools [7]. This large demand is satisfied through two main production technologies: the aqueous suspension and the continuous mass production. The first one is a batch suspension polymerization and more details have been provided in chapter 3. The newest technology has the advantages and disadvantages of a continuous production. The hourly production is quite higher, start and stop operations are required only at the beginning and end of production instead of every batch that need a proper operation sequence. Less manual operations are required: in a continuous process and potentially dangerous environment, it sounds advantageous. On the other hand, batch process is more flexible and allows a rapid change in production; in continuous production, if polymer grade changes frequently, as it is the case, often intermediate and end-product quality decrease during the transition among grades.

Anyhow the balance was in favour of advantages of a continuous process and Versalis few years ago invested money and resources in the continuous mass EPS production, also known as ST11 plant described following.

5.2.1 EPS continuous mass production

The process might be considered as composed from two main parts: the polystyrene (PS) production phase and the polymer-mixing phase in which expanding agent and other additives enter the reactor. Polystyrene production [8 9] is the same independently if the result is the general-purpose polystyrene (GPPS) or the expandable polystyrene. PS process uses styrene monomers and ethyl benzene that has two principal proprieties: it is a good solvent for PS and it works as chain transfer. Polystyrene has high viscosity that makes pumping difficult; ethyl benzene decreases viscosity and allows high conversion rate. Moreover, it acts as chain transfer and its quantity regulate the molecular weight, maybe the most important feature of PS.

Polymerization happens in one or more subsequent reactors, plugging-flow (PF) or Continuous-flow Stirred-Tank Reactor (CSTR), it depends on selected technologies; the heat of reaction is partly absorbed by itself for polymerization and the remaining absorbed by a cooling oil jacket. It is quite strange that an oil jacket absorbs heat instead to provide it but there are many advantages if compared, i.e., to water; furthermore, it supports plant start-up, a quite critical procedure.

Conversion is controlled via initiator, peroxide that starts the radical reaction and increases its velocity. Along one or more reactors, fused polymer reaches the conversion ratio in which viscosity, near 80%, permits next operation. This stream is known as pre-polymer and has about 20% of non-converted monomers, solvent and impurities. Last step in polystyrene production consists in the separation of polymers; it happens in a so-called devolatilization tank. Prepolimer enters in a vessel in vacuum condition and, for a combined temperature and pressure effect, monomer and solvent are separated. This fluid, after some specific treatments, comes to the beginning of the plant in order to save raw materials. First phase of ST11 production finishes here.

Classical PS process ends with a cutting item that generates polymers beads, whose dimension is about 5 millimetres; maybe some other additives are added to facilitate successive transformation but it is not the case of EPS mass continuous production in which polymer undergoes to another section.

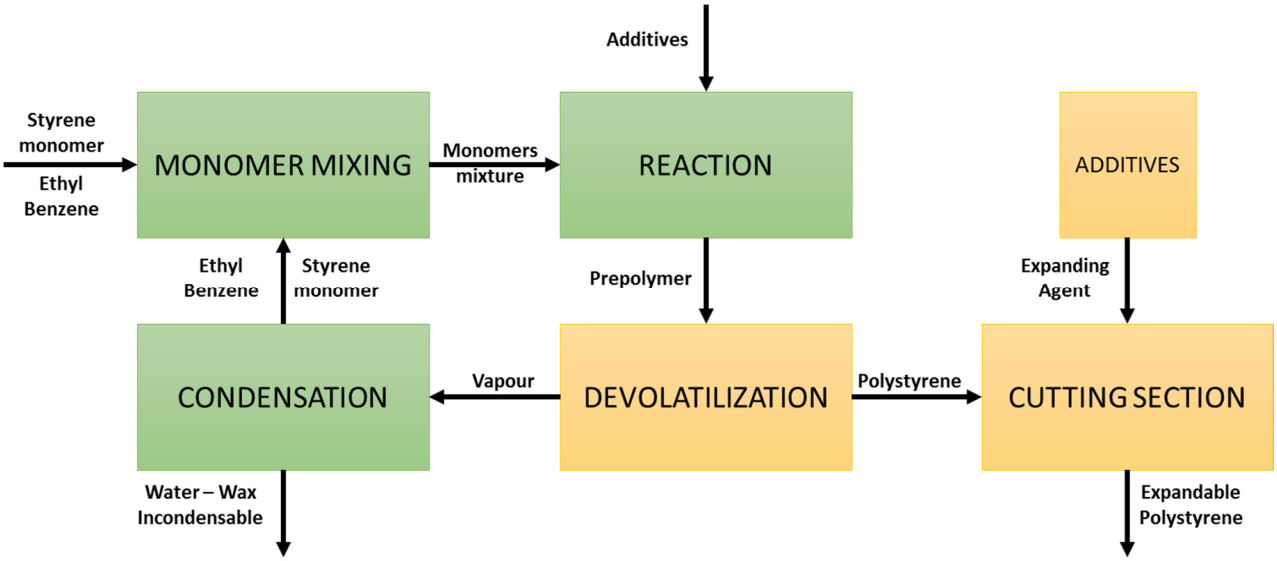

Figure 5.1 ST11 block scheme, only yellow block were involved in multivariate analysis

Melted polymer goes through a dedicated item in which expanding agent and various additives are mixed. High temperature and pressure make complex these phase and Versalis is proprietary of a dedicated equipment and technology; for that reasons the description is not completely exhaustive. Final products is rather similar to the suspension ones, but with a more regular dimension, in any cases about 1-2 millimetres. Final users probably do not distinguish mass EPS from the suspension one due to their similarity. For sure, mass production allows a custom additivation and products grades with suitable applications.

5.2.2 Molecular weight – Propriety and measurement

A specificity of this process is the possibility to manage the molecular weight (Mw) in such way that EPS might be modified in order to satisfy dedicated market. Mw defines the number of monomer units that compose a single polymers macromolecule. Mw is the main polymer structure indicator and, therefore, correlated with the expandability of the EPS. The other polymer features, such as the melt flow index (MFI) and, the Izod impact among others, severely depend on the molecular weight. Furthermore, a correct Mw allows a good workability and the weight values vary according to the final customer items. It could be managed either during polystyrene process or with additives concentration. Regulation is possible because of the close correlation that links the molecular weight and the production temperature: there is a strong influence of temperature, which, basically, decreases the molecular weights value [10**]**. Moreover, the chain transfer modifies or even defines the molecular weight [11]. Finally, compounds mixed or added in EPS, of course, have an influence on the final Mw that relates to their concentrations. In order to monitor the EPS production the Mw measure is a fundamental parameter, it is the main structure indicator and, therefore, correlated with the expandability of the EPS.

Gel Permeation Chromatography (GPC) is the reference measure for Mw and returns the mean value and its distribution [12]. GPC/LS (Light Scattering) is among the most powerful methods in polymer characterization [13]. A GPC column allows separation of macromolecules according to their size, in the usual, big first, order of elution. The pore sizes design lets the large solute particles to pass through uninhibited. However, the small particles permeate the gel and are slowed down so the smaller particles are the slower to get through the column.

This separation gives the molecular weight distribution. For the EPS molecular weight analysis, the Waters Alliance E2695 separation module has been used. Four linear columns, with 10^6 to 10^3 Angstrom porosity, perform separation and the Waters 2414, refractive index detector, measures the scattering and correlates it with particle size. Mw analysis via GPC needs a proper calibration; monodisperse polystyrene standards make instrument calibration quite easy and fast. The measurement error is about 3000 units and was obtained by repeated sample measurements.

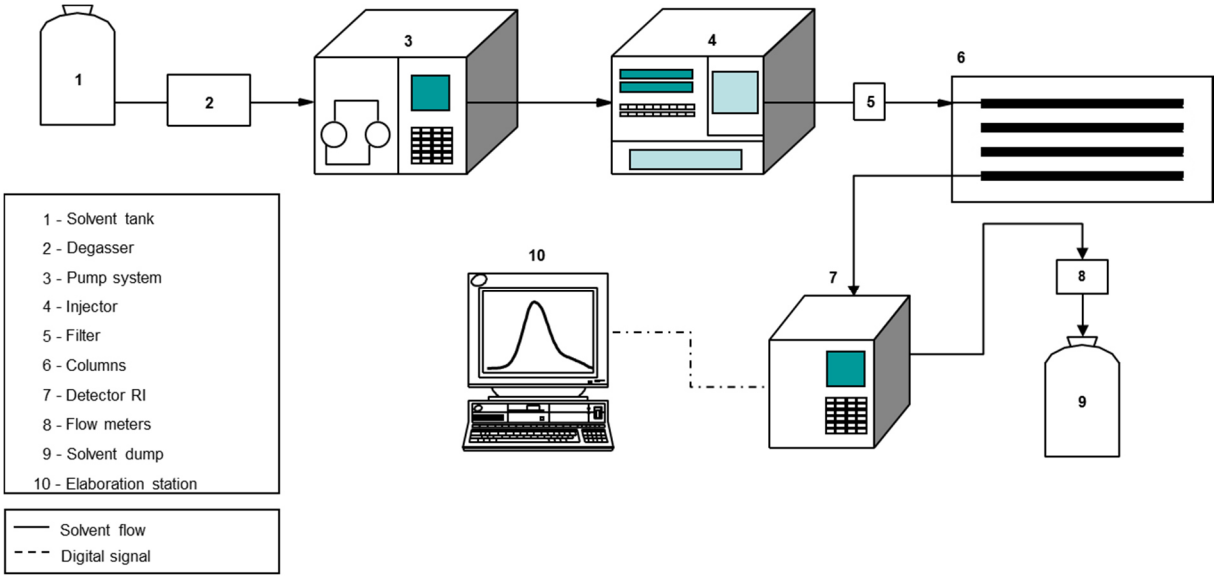

Figure 5.2 GPC/LS scheme for Mw measurement

The scheduled analysis for Mw measurement is daily and only in particular cases, plant re-starting or grade change, frequency increases up to 2-3 analyses per day. It allows a rather punctual quality control but extremely low frequency material monitoring. EPS production is not a unique and particular situation in petrochemical industry in which laboratory checks must be as few as possible to reduce costs, especially with such kind of analysis.

5.3 Data collection and preliminary analysis

The data collected from process sensors have more or less the same features and acquisition/storage frequency issue as for the ST11 process. To obtain a calibration model, and validate it, the values measured by the reference method, i.e. molecular weight estimated via GPC, are needed. Thus, laboratory dataset is the response vector **y** with *mx1* dimensions, where *m* is equal to the available laboratory tested samples. Calibration process dataset, **X**, has been synchronized with laboratory samples tested, and consists of *m* rows and of as many columns, *n*, as sensors on ST11. A row of **X**, *1xn* dimensions, can be considered as a portrait that shows the process in that moment. Sensors are reported in Appendix 1. Variables description refers to the items installed in ST11; in order to protect Versalis knowledge, description limits to the items name and sensors type (thermocouple, level gauge, etc.).Available data were split into calibration (115 observations) and validation samples (36 observations). Known values are form the end of 2011 to the 2014 and include more or less three years production.

For assessing the model predictive capability in routine situation (as simulation of continuous monitoring), a further dataset has been considered covering the period October - December 2014 , data were extracted from database with hourly frequency in order to simulate a continuous prediction but obviously only the daily laboratory Mw measurements could be used to verify the goodness of predictions.

5.3.1 Dataset overview

In data driven modelling the concept of garbage in, garbage out is crucial [14]; a noise, incomplete and full of non-sense variables dataset returns bad prediction ability. Raw database contains a lot of variability and it must be treated in a suitable way to catch all the interesting information. An accurate analysis on single variables gives a good preliminary overview, especially if coupled with plant operator knowledge. Experience is of utmost importance in this phase because only understanding of the system ensures proper considerations. After a careful check, ninety-eight sensors out of hundreds were selected: they are thermocouples, flow meters, level gauges and so on. This a priori pruning of sensors to include in the model was done independently to the supposed correlation with the molecular weight; the selection was done according to sensors functionality in process control, by plant operator experience, by the availability of historical data and by checking the amount of missing/spurious data.

Differently from batch process, where sensors in the same position of the plant register the same conditions for each batch and batch evolution is given by time, every sensors of a continuous plant "sees" a different material while time pass by as it is better detailed in the following.

Sensors measure features along EPS field, lying quite distant from each others and measuring, obviously, the melt next to the apparatus. Each sensors ideally considers the same delta volume (dv) but at a different time instant: the feeding point corresponds to time zero and, taking into consideration a dv, each sensor have a related delay. It could be imagined as a little square that touch every sensors during its travel from the beginning to the end of the plant. Continuous production evolves along plant and unless side-phenomenon occur or product change, at the same point in space (position of the sensor on the production line) there pass the same material, time independently.

Thus, there is the need to take into consideration the spatial distance between sensors and to establish a delay time proportional to the flow rate to correctly match the values registered by sensor with the measured product Mw; it means that each sample was related to the plant conditions in which it was produced. A further available variable was PS weight estimation by simulation. Two phases, as explained earlier, compose EPS continuous mass line: the polystyrene (PS) production and polymer-additive mixing phase. Few years ago in the R&D centre of Versalis, the polystyrene research group built a simulation model to calculate the mean molecular value of polystyrene that comes out from the devolatilization section (first phase), this model is based on polymer theory considerations and elaboration takes as input first phase plant set-up. The PS Mw estimated data might be interesting in order to evaluate the correlation between PS and EPS molecular weight and which is the effect on Mw due to the second production phase. PS estimated molecular weight was both taken and not taken into account as variable in the X-block for EPS regression model building. When taken into account the **X** has 99 columns, 98 sensors plus the PS Mw.

5.3.2 Exploratory Analysis

5.3.2.1 Univariate

Essential information comes from the molecular weight comparison between PS and EPS. A simple difference plot shows that the second phase of production, from PS to EPS, involves a reduction of the Mw; it can be easily detected on Fig. 5.3: the measured value from GPC is always less than the value estimated by first principles model. For both values, there are two production ranges: first group refers to observation from 1 to 11, December 2011and the second includes the remaining samples. The lower molecular weight corresponds to a different polymer grade, i.e. the presence of a second monomer. This kind of EPS is now out of production and those groups of observation relates to the last produced campaign. This production having lower Mw range could help in building the calibration (the spanned domain is larger) but because is no longer on Versalis portfolio and, even more, because the presence of the second monomers modified production set-up, it has been decided to exclude observations from 1 to 11.

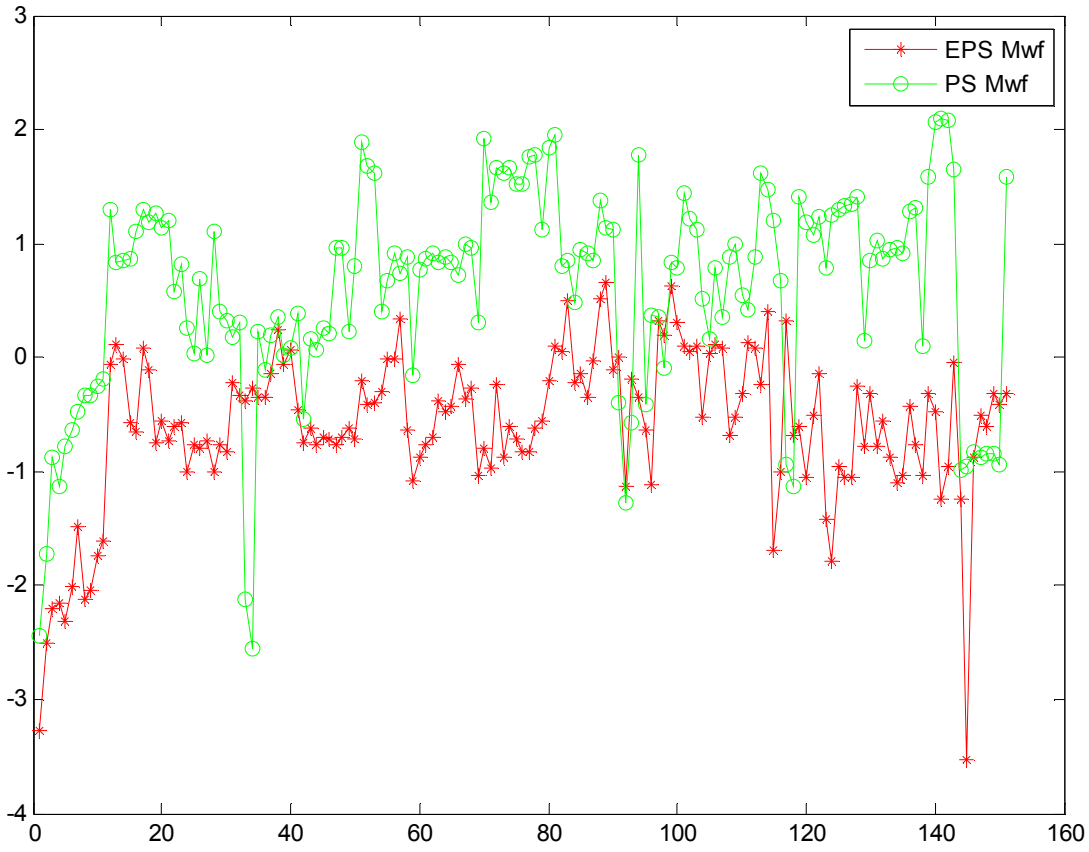

Fig. 5.3 Polystyrene and expandable polystyrene molecular weight comparison

Variables might contain missing data and also their distribution should be checked. Miss read values become significant when a variable has a considerable percentage of missing data.

NIPALS algorithm [15] might "safely" impute missing data if this are randomly spread in the data matrix and not too high percentage with respect to the number of available data [16]. Thus, variables have been checked for missing data and those with an inappropriate number of missing data. Some sensors were excluded because present a constant profile. Some sensors had very far from normal distribution, considering historical data, for instance the flow ratio of some additives injection in the system. A bimodal distribution was observed and these variables behave as sort of categorical data: zero-one, no-flow and normal-flow, and variance is limited to these two levels. They should have been excluded however, in order to keep good relationships with plant manager, that based on his prior knowledge considers very important those flow rates, they were kept in the starting matrix.

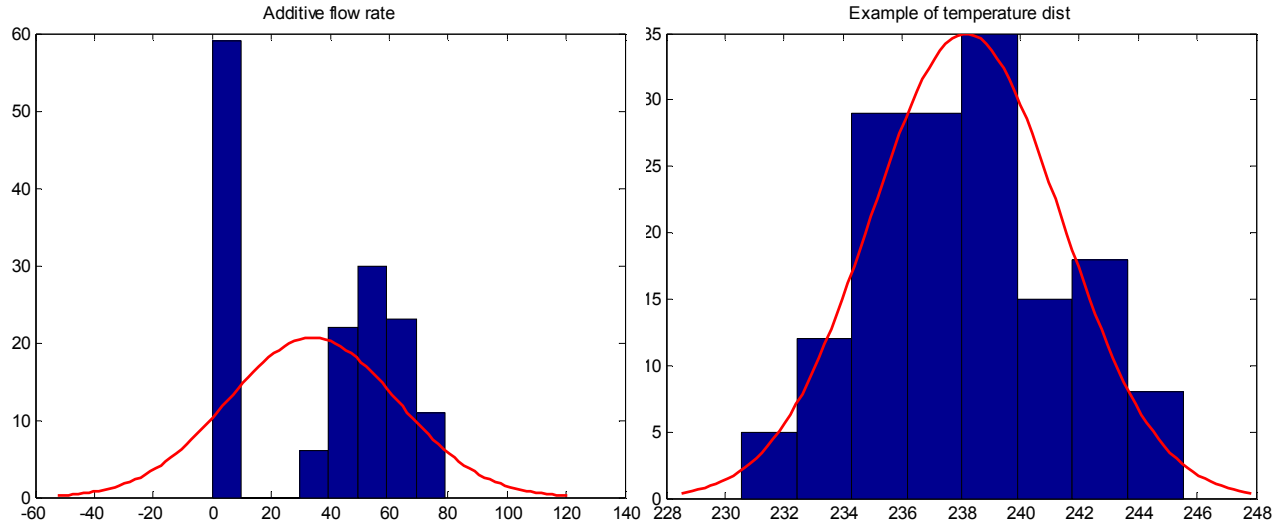

Figure 5.4 Bar plots shows frequencies of data and the unimodal distributions in red. On the lefts red line does not make sense and two maximum (0 and 60) are easily found. On the right, example of unimodal distributed temperature.

5.3.2.2 Multivariate

Univariate analysis returned some information on single variable behaviour; multivariate methods (PCA) might give information on plant variation during the three considered years. In PCA also PS Mw was included. A six components model describes the 70% of variance. As usual in the polymerization plant, production is not stable and this appears in scores plots. In PC1 vs. PC2 scores plot it can be observed: 2011 and 2012 conditions are different from the more recent observations 2013 and 2014; they differ mainly on the first PC as the line scores plot shows (Fig. 5.5). ST11 stopped production for some months in 2012 and production start again on 2013.

Probably in the meantime something was changed in plant setting. From loadings plot (Fig. 5.6), among the other effects, it might be observed that tank temperature D5501 (n° 25 in the plot) increased in years 2013 and 2014. In the same direction the thermal exchanger E5505B (n° 91 and 94 in the plot) augmented its level and temperature. Many variables have high loadings values on PC1 describing a global variation in plant setting. The second latent variable drifts along years: in polymer plant, such phenomenon frequently is due to fouling of the system.

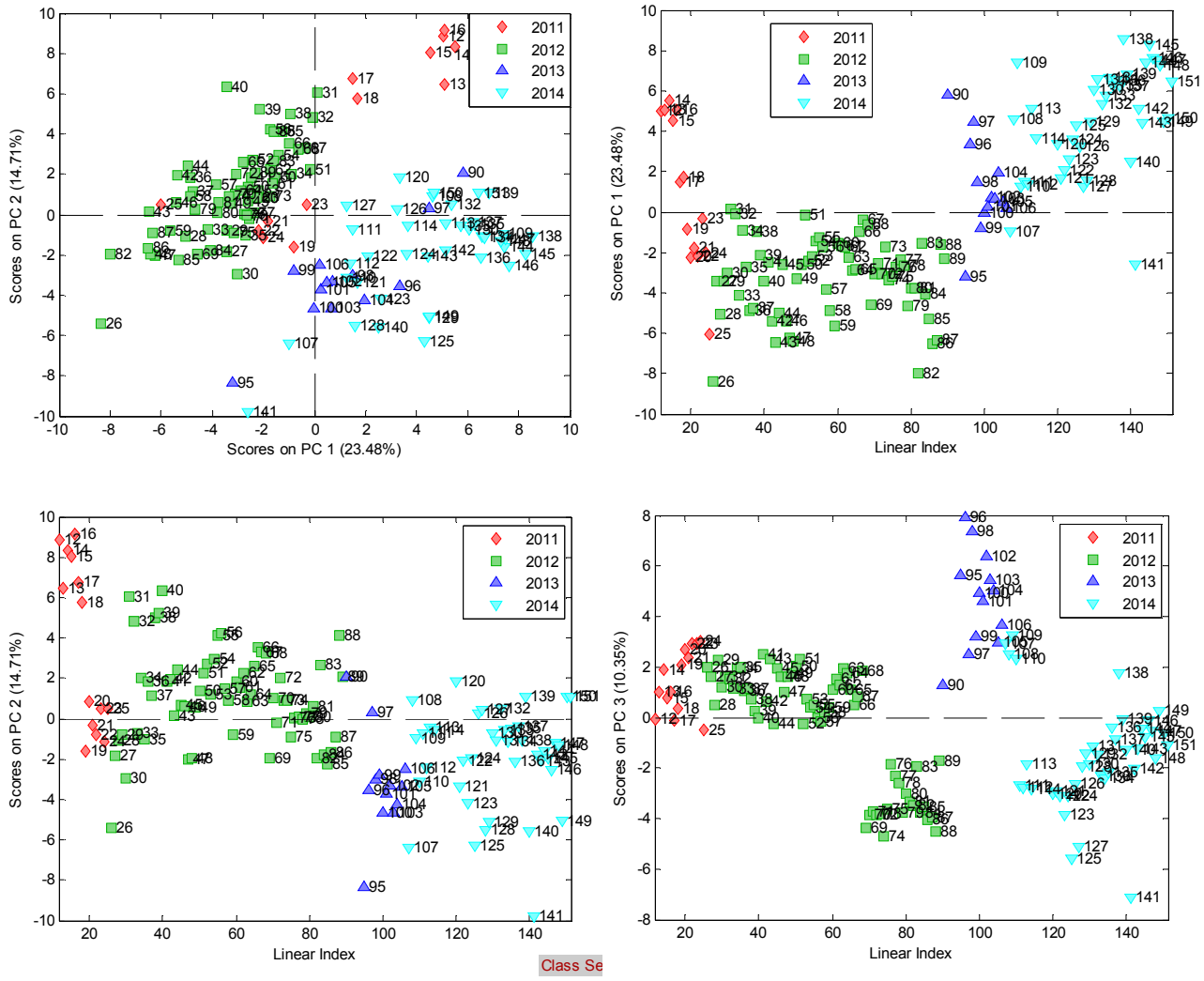

Figure 5.5 Scores plot plane PC1 vs PC2 (top left), Scores plot line PC1 (top right), PC2 (down left) and PC3 (down right). PCA model on Observation matrix. Colour vary according to the year of production

Variables in loading plots shows that the progressive decreasing of second latent variable is related to a minor pressure (n° 28, 41 and 42 in the plot) along years; probably it relates also to a change in flow rate. The third component seems related to a seasonal effect but the shape is more complicated with respect to the external temperature trend; maybe a combined effect of external temperature and plant set-up might explain this trend.

D5514 (to n° 76 from 81 in the plot) is the item more related to that behaviour; all the thermocouples installed describe such change in the years but its temperature, above 200 °C, seems more related to plant set up that external temperature variation.The production line suffers from uncontrolled variations in the medium and long terms. Those unwanted variability sources coexist in many petrochemical productions. Changes over time could make much more difficult to obtain a good predictive model due to the variance that does not relate to the molecular weight. Multivariate explorative analysis is not mandatory but highly useful in a regressions task too. In literature most of regression model used in process monitoring are based on on/in-line spectroscopic sensors or other instrumentation that monitor materials composition [17 18 19]; such a kind of information relates both to the chemistry and physic features. A PLS model based on process sensors only is more challenging since may capture chemistry but indirectly. Nonetheless, the advantages are in terms of money, maintenance and application feasibility.

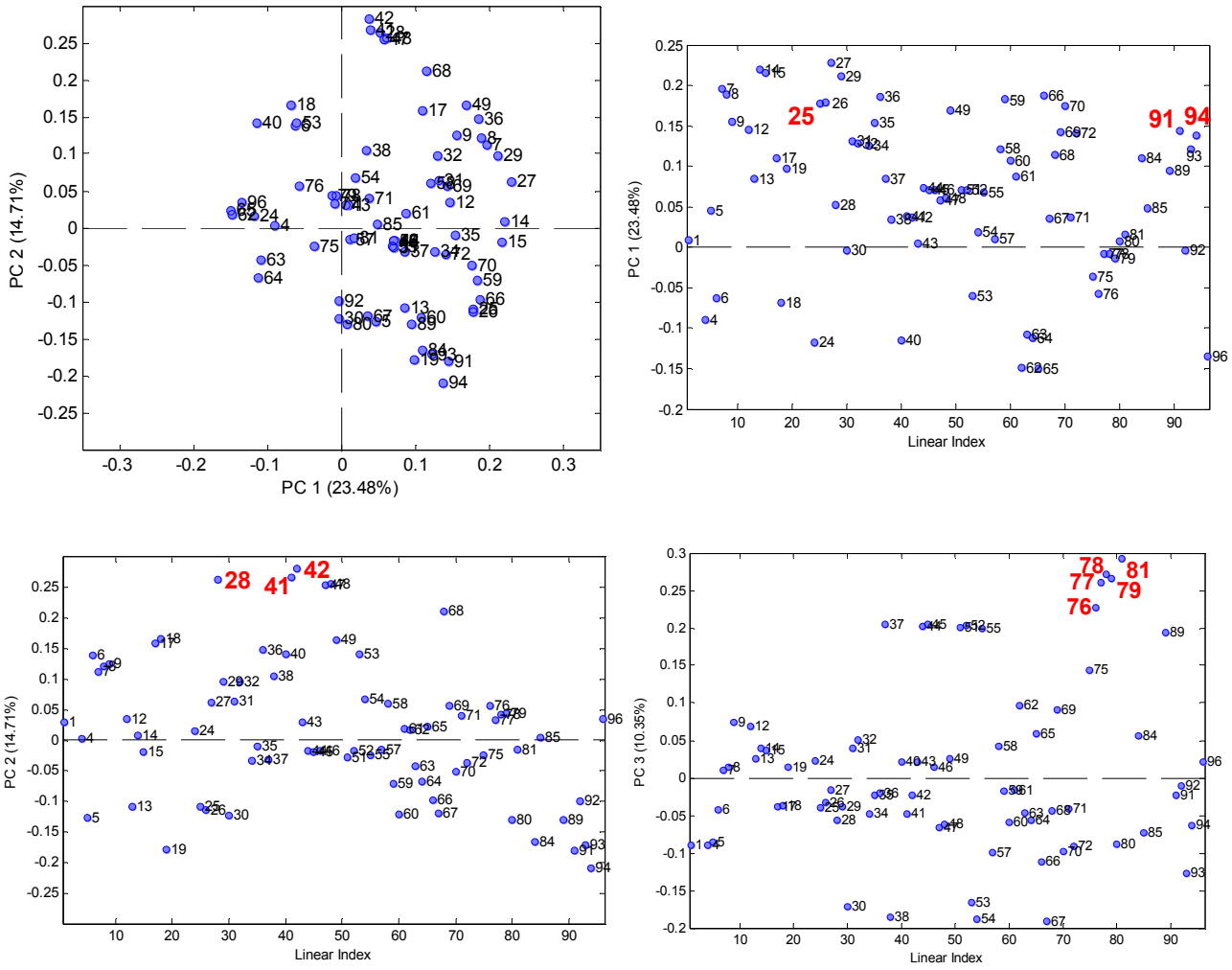

Figure 5.6 Loadings plot PC1 vs PC2 (top left), Loadings line plots: PC1 (top right), PC2 (down left) and PC3 (down right). PCA model on Observation matrix

5.4 PLS Model

The **X** and **y** datasets were pretreated using autoscaling. About 100 samples and three years makes the PLS model robust enough. It is expected that most of the variables will not be correlated with molecular weight, given their "indirect" accounting of "chemical" behaviour. A model with all variables returns a cross validation error of 3550 units, but the prediction error exceeds the 8000 units, too high for a usable estimation of Mw for production control purposes. A proper variable selection might improve regression and prediction results. The VIP, forward selection and Genetic algorithms were applied to evaluate the best and more stable results in prediction capability. Genetic algorithm returns the lowest error in cross validation with a low selected variables number. By its nature, genetic algorithm models too much the calibration dataset generating over fitting problems [20]: it means that the error obtained in cross validation does not reflects the real estimation capability. To evaluate this unwanted effect the data from December 2011 to March 2014 were used for the calibration and the remaining data, from April to August, as external validation dataset. A summary of main information of regression model (Tab 5.1) shows an adequate prediction ability: the molecular weight error in cross validation and primary analysis, GPC extraction, are comparable. Moreover, the explained **X** and **Y** variance reach a considerable values.

PC's	Calibration Validation samples	Variables number	X var.	Y var.	RMSEC	RMSECV	RMSEP
6	**97/36**	**27**	**77.3%**	**71.5%**	**2420 u**	**3120 u**	**5290 u**

Table 5.1 PLS model statistics, Genetic algorithm variable selection

Six principal components exhibit a complex plant behaviour concerning the 27 calibration variables. VIP and forward selection shows worse results than GA. The first model with all sensors had a poor prediction ability and VIP, for its peculiarities, cannot improve regression results from a bad model [21]. Forward selection model might obtain satisfactory prediction and the model quality does not differ too much from the GA in cross validation but it suffers a lack of quality during external prediction; the higher number of sensors the more GA is prone to overfitting. The PLS statistic and GA variables selection make the external validation mandatory to evaluate the multivariate regression efficiency.

By using the PLS model built on subset of sensors selected by GA, the external prediction returns an appropriate error for a plant control, to follow product drift and in order to segregate "strange" production. The aim of this study was to have an "easy" and robust way to estimate Mw of EPS hence controlling production. This is rather different aim with respect to product certification that shall be verified by primary analysis. However, the achieved results were considered satisfactory from the plant management and polystyrene research group of Versalis.

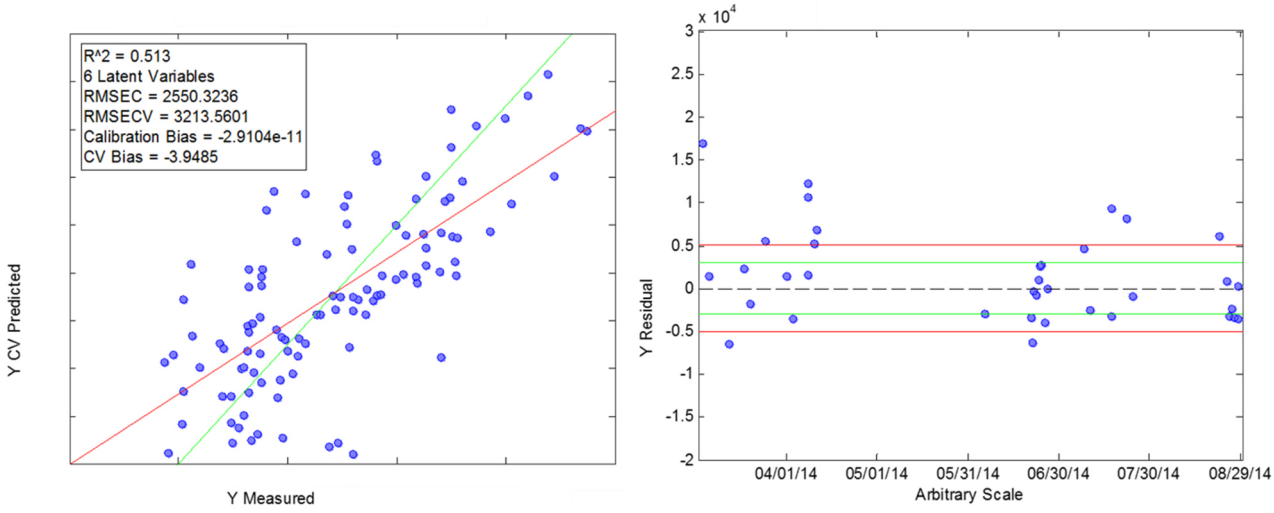

Figure 5.7 On the left measured vs CV predicted Mw values, green line represents the ideal regression (1:1) and red line is the real ones. On the right residuals values in prediction, green line shows 3000u error and red is 5000u error.

Despite the good results, some issues still remain about model prediction quality and its stability. ST11 production line suffers drift and production as highlighted by exploratory analysis. A continuous polymer plant has a life period going from the plant start to the plant global maintenance. This period might vary from few months to 4-5 years, in which pipes get dirty and instruments decreases their efficiency. The long-time variation is really similar to a batch process that affects all production in relation to the aging of the plant. Actually, EPS database is quite young, less than six years. Initially many stops were necessary in order to set up correctly the plant and now is actually its first long living production period. Thus, old "batches" in strictly sense, i.e. resembling actual production, are not available but referring to the exploratory PCA, first and third components show some repetitive behaviour. This behaviour might suggests the segregation in short periods with similar samples conditions in relation to the variation unrelated to the Mw. In this way, such unwanted variation seems to not affect Mw prediction by the model. However, In order to verify alternative way to improve prediction quality the locally weighted regression was applied to the same dataset.

5.4.1 Locally weighted PLS

Local methods, as the name suggests, select a subset of data instead all at available one, defines a sub domain instead the global space in order to improve dataset understanding and regression results; for LWR explanation the reader is referred to section 2.3.2. ST11 application bases on PLS regression and in order to obtain comparable results also the locally weighted model uses PLS. In the classical partial least square regression, the only parameter that could vary is the number of latent variables, when variables, samples and pretreatment are defined (in any case, a lot of work to do). In LW-PLS more parameters should be defined: the maximum number of neighbours, the weight function and the maximum number of latent variables to perform regression. Maximum number of neighbours was imposed equal to the all data minus the necessary amount of samples for validation; the aim was to explore all possible range for neighbour selection. The maximum number of latent variable was set to 20, more than necessary.

Observation weights might be function of time distance between points, for instance, but in this exploratory phase the selected neighbours were not weighted. Parameters selection uses an external dataset that might be either part of the original (cross-validation) or a separated one. The algorithm calculates iteratively error, cross or external validation, for every combinations of latent variables and neighbours number. It returns an error surface in which xy-space corresponds to the neighbours and latent variables numbers. The minimum error is ideally the best results but an high number of latent variables should be selected carefully. Other parameter to set up is the distance, rather the most important because it guides the sample selections. The distance, such as the similarity measure, might be both calculated in to the sensors space and the latent variables space; PCA returns drift and behaviour in latent coordinates and for that, even the distances in principal component space were chosen.

Model name	PC's	Neighbours	Distance type	Validation Type	RMSECV	RMSEP
Classical PLS	**6**	**97/97**	**-**	**-**	**3120 u**	**5800 u**
LWR	**6**	**69/97**	**Normal**	**Internal**	**3300 u**	**6500 u**
LWR	**3**	**75/97**	**PCA**	**Internal**	**3100 u**	**6000 u**
LWR	**3**	**93/97**	**Normal**	**External**	**-**	**5900 u**
LWR	**3**	**89/97**	**PCA**	**External**	**-**	**6000 u**

Table 5.2 Comparison between classical and local PLS. For LWR are interna and external cross-validation and even the distances calculated in normal and latent space

Concerning EPS production, dataset was examined both with internal and external validation. In general, the LWR models did not improve predictive capability, Also the differences among the validation approach, internal CV or external set, gave rather similar results. The LWR models required a number of neighbours close to the total number of samples, thus it seems that all considered time period is needed to built the PLS model to have good Mw prediction. The most parsimonious LWR model is the one on third row of table 5.2: 3LV, 75 neighbours selected on latent space and with 6000 units of prediction error. It could be stated that in EPS application, local method does not improve prediction, at least for the available dataset. However, every surface error describes the same condition: bad prediction with too many latent variables and small progresses for model with more than 30 samples, surfaces were more or less flat.

Figure 5.8 shows how prediction varies with growing number of neighbours; the same latent variables number, three, were used. It could be observed that for each external observation, few samples are enough to reach a prediction as much reliable as the global model: in fact the prediction residuals show that more neighbours improves regression for high residual samples, such as samples number 9, 10, 13 and 21, and has no effect for the other ones in which residuals are inside the 5000u limits, also with 20 neighbours. Thus, the neighbours selection was oriented on a lower number despite the error is a little bit higher, in this way the effect of local regressions are more evident.

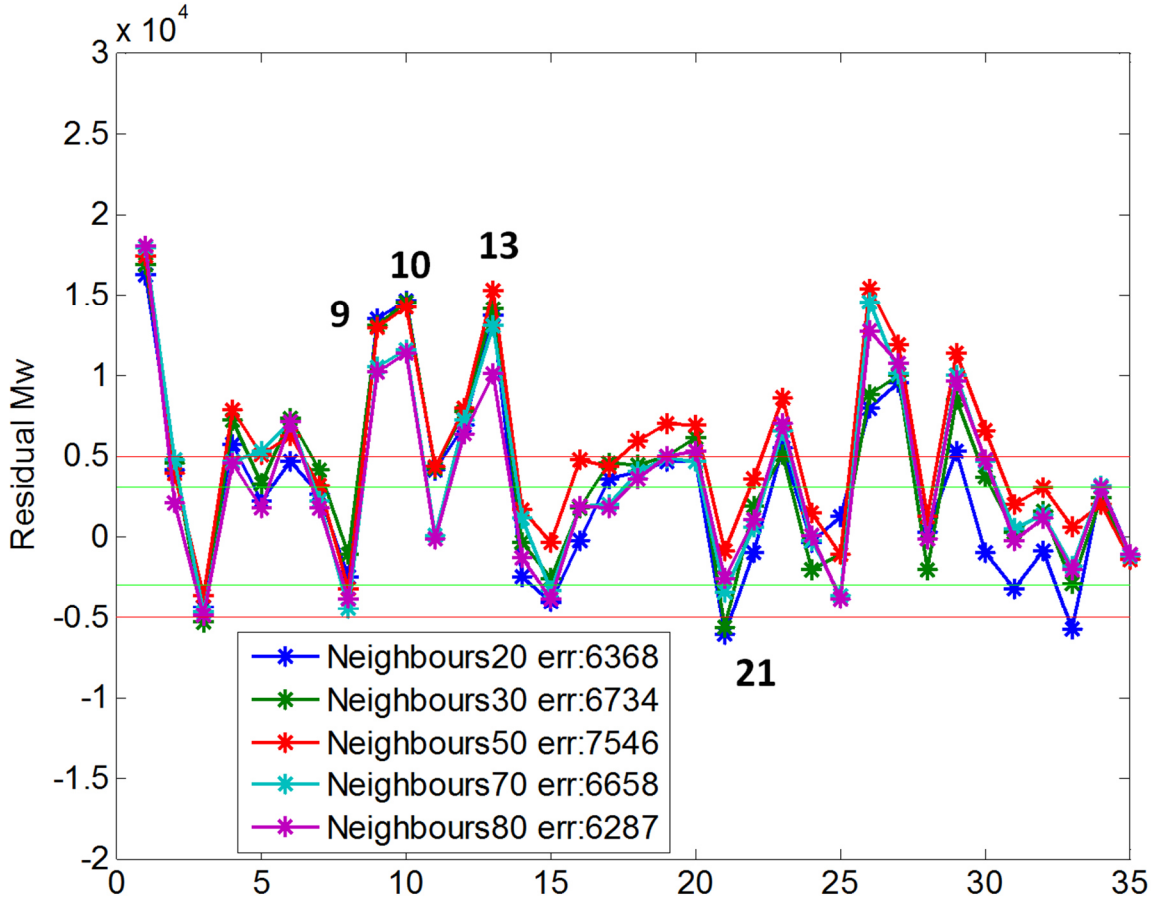

Figure 5.8 External dataset prediction with LWR-PLS for growing number of neighbours. Sample's number vs. its residual value for predicted Mw. Green horizontal line shows 3000u error and red horizontal is 5000u error.

Furthermore, the nearest 20 neighbours include samples of the same production period but even sample temporally distant. Both for good and bad predicted samples, neighbours belong to various periods and it means that plant change might affect prediction or the same conditions return as a sort of cycle. Example in figure 5.9 shows neighbours of test sample 16, April 2014; part of them are immediately before validation sample, February and March 2014, but the remaining belong to the beginning of 2012. To evaluate the capability of LWR model has been used 20 neighbours and 3 latent variables.

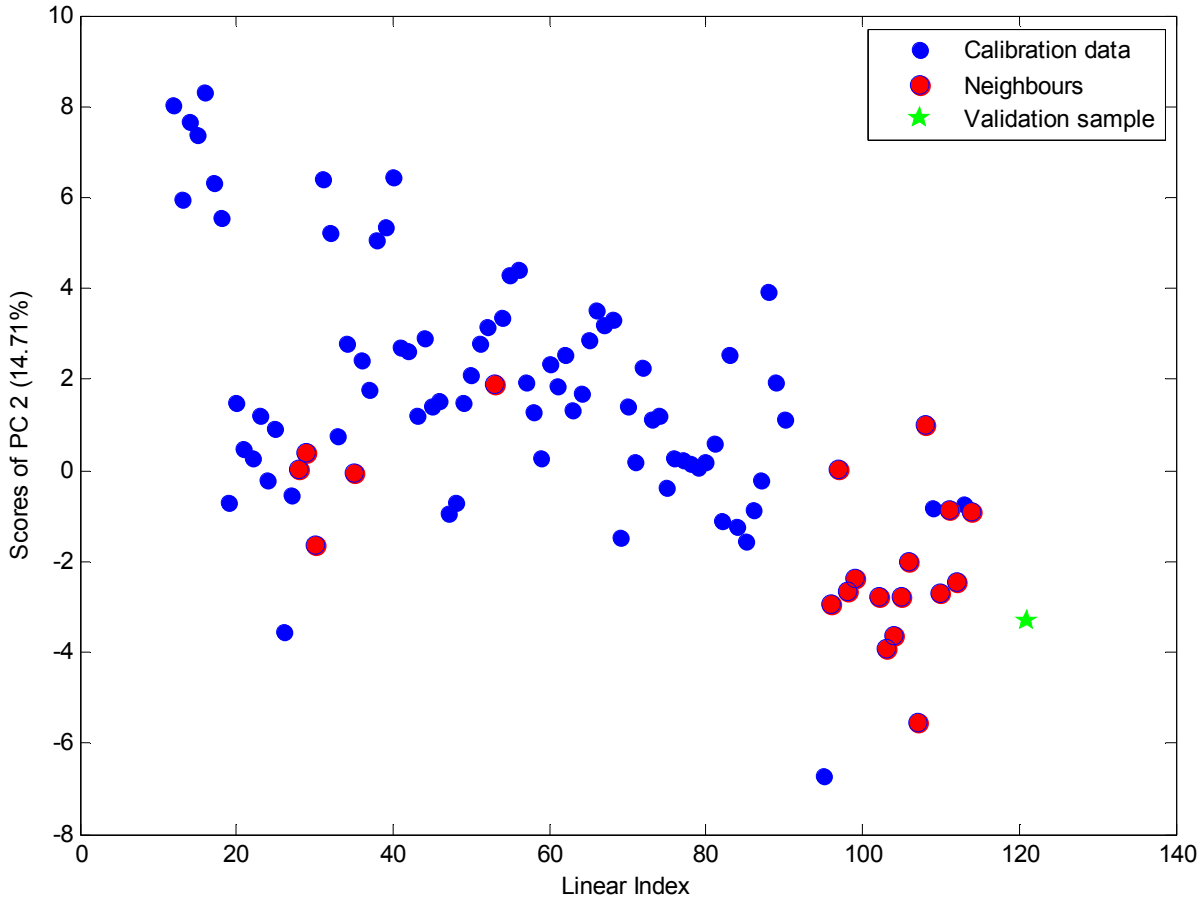

Figure 5.9 Neighbours (red) of validation sample number 16 (green star) shown on PC2 scores of PCA on Observation matrix

5.4.2 Continuous monitoring simulation

The test data, simulating the monitoring on-line situation, include all the data between 20 October and the 01 November 2014. The references Mw laboratory measurement are available only for four samples. Both for classical and local regression, molecular weight was estimated hourly and results appear in figure 5.10; it shows that LWR has a better prediction than PLS and that have also a quite stable signal, a positive note for real time monitoring. In December 2014, five GPC samples allowed a second continuous test that returns less successful prediction in which the two profiles overlap and LWR seems a little bit noisier.

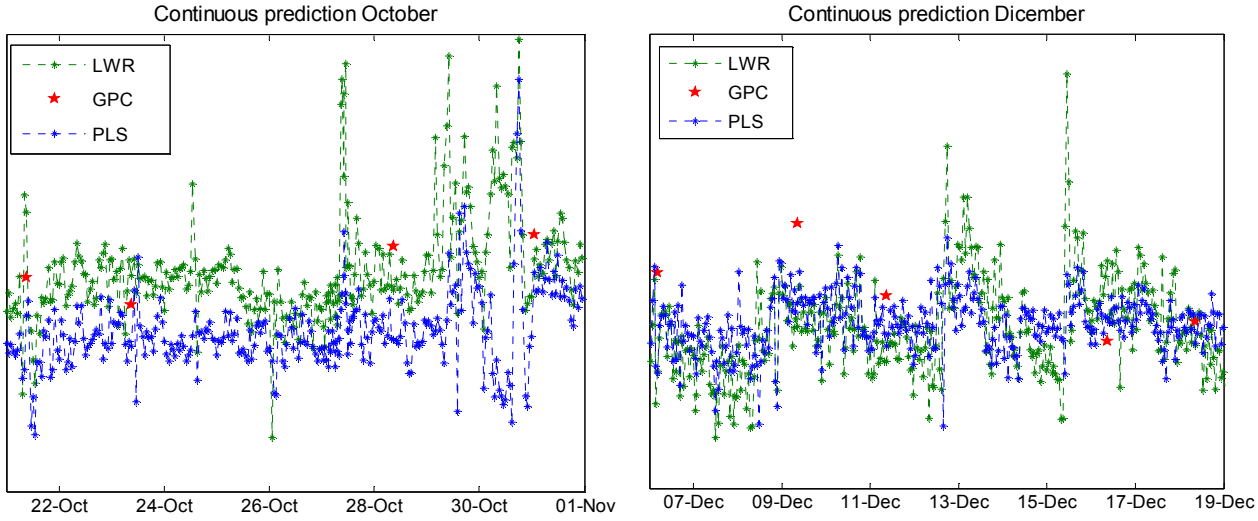

Figure 5.10 On-line prediction simulation. Red stars represent real measurements, green and blue lines show local and classical PLS regressions.

Shapes describes similar prediction ability, LWR appears more confident and in any case as good as classical PLS. The EPS application is a good example of MSPC and it has a successful conclusion.

5.5 Conclusion

The molecular weight prediction appeared complicate from the beginning. The importance of Mw estimation and the high cost of analysis guided me and plant managers to that choice but at the same time everybody recognized how challenging it was. The ST11 features do not allow classical spectroscopic application and in any case a cheaper solution was preferred.

The various multivariate techniques applied returned a reliable regression in which the relatively low error gives to the plan controller precious information. Too many times applications end here, and a good model does not find proper use. Fortunately the constant collaboration with Versalis technology department promotes application. A standalone computer calculates both polystyrene molecular weight with first principles model and the EPS Mw with classical PLS every 30 minutes. The software stores also the data in ST11 database and plant controllers looks estimation directly in the control room. In terms of software building and investments it is a trivial application, but the high importance lies in the continuous use of a multivariate parameter for plant control.

LWR might improve prediction but for simplicity the PLS model has been applied; local weighted regression needs, in addition to estimation, the neighbour identification and the PLS model calculation. First step does not involve heavy calculation and simple excel sheets might evaluate distances; latent variables calculation concerns iterative process that could be easily solved with a MATLAB function but initially technology department preferred the simple application.

Anyway, prediction is performed with regression coefficient and a control on sensors data to avoid missing or meaningless data. Every three months a calibration check will suggest if it is necessary to update model with newer samples. Alongside this, LWR is monitored in order to recognize its advantages and the convenience in the application of local model.

5.6 References

1. Nagata, T., Ohshima, M., & Tanigaki, M. (2000). In-line monitoring of polyethylene density using near infrared (NIR) spectroscopy. *Polymer Engineering & Science*, *40*(5), 1107-1113.
2. DeThomas, F. A., Hall, J. W., & Monfre, S. L. (1994). Real-time monitoring of polyurethane production using near-infrared spectroscopy. *Talanta*, *41*(3), 425-431.
3. Santos, J. C., Reis, M. M., Machado, R. A., Bolzan, A., Sayer, C., Giudici, R., & Araújo, P. H. (2004). Online monitoring of suspension polymerization reactions using Raman spectroscopy. *Industrial & engineering chemistry research*, *43*(23), 7282-7289.
4. Özpozan, T., Schrader, B., & Keller, S. (1997). Monitoring of the polymerization of vinylacetate by near IR FT Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *53*(1), 1-7.
5. Haddleton, D. M., Perrier, S., & Bon, S. A. (2000). Copper (I)-mediated living radical polymerization in the presence of oxyethylene groups: Online 1H NMR spectroscopy to investigate solvent effects. *Macromolecules*, *33*(22), 8246-8251.
6. Mignard, E., Leblanc, T., Bertin, D., Guerret, O., & Reed, W. F. (2004). Online monitoring of controlled radical polymerization: nitroxide-mediated gradient copolymerization. *Macromolecules*, *37*(3), 966-975.
7. Klodt, R. D., & Gougeon, B. R. A. D. (2003). Particle Foam Based on Expandable Polystyrene (EPS). *Modern Styrenic Polymers: Polystyrenes and Styrenic Copolymers*, *6*, 165.
8. L.Trentini, A. Longo, F. Pasquali, Encyclopedy Treccani, Vol. II refining and petrochemicals, *"Thermoplastic styrenic polymers"*.
9. Brydson, J. A. (1999). *Plastics materials, Cap.16 Plastic based on styrene*. Butterworth-Heinemann.
10. Murzagaliev, N. F., Rakhimkulov, R. A., Kiryukhin, A. M., & Alekseev, S. V. (2009). Change in molecular mass distribution of polystyrene during its processing. *Oil and Gas Business*, 9 (1).
11. Mayo, F. R. (1943). Chain transfer in the polymerization of styrene: the reaction of solvents with free radicals1. *Journal of the American Chemical Society*, *65*(12), 2324-2329.
12. Kissin, Y. V. (1995). Molecular weight distributions of linear polymers: detailed analysis from GPC data. *Journal of Polymer Science Part A: Polymer Chemistry*, *33*(2), 227-237.

13. Radke, W., Simon, P. F., & Müller, A. H. (1996). Estimation of number-average molecular weights of copolymers by gel permeation chromatography-light scattering. *Macromolecules*, *29*(14), 4926-4930.
14. Møller, S. F., von Frese, J., & Bro, R. (2005). Robust methods for multivariate data analysis. *Journal of Chemometrics*, *19*(10), 549-563.
15. Nelson, P. R., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and intelligent laboratory systems*, *35*(1), 45-65.
16. Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S. *PLS_Toolbox for use with MATLAB™*. Eigenvector Research, Inc., Wenatchee, USA, 2006.
17. Fischer, D., Bayer, T., Eichhorn, K. J., & Otto, M. (1997). In-line process monitoring on polymer melts by NIR-spectroscopy. *Fresenius' journal of analytical chemistry*, *359*(1), 74-77.
18. De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J. P., & Vervaet, C. (2011). Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *International Journal of Pharmaceutics*, *417*(1), 32-47.
19. Vanarase, A. U., Alcalà, M., Rozo, J. I. J., Muzzio, F. J., & Romañach, R. J. (2010). Real-time monitoring of drug concentration in a continuous powder mixing process using NIR spectroscopy. *Chemical Engineering Science*, *65*(21), 5728-5733.
20. Leardi, R., Seasholtz, M. B., & Pell, R. J. (2002). Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Analytica Chimica Acta*, *461*(2), 189-200.
21. Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, *24*(11-12), 728-737.

5.7 Appendix 1

The table of sensors indicates in red the ones included in regression via GA selection.

Num.	Description	Num.	Description
1	**Molecula_weight FPM**	2	Old Comonomers
3	**Press. R5001A**	**4**	**Press. R5001B**
5	**Press. R5001C**	**6**	**Press. R5001D**
7	Temp. R5101	8	Temp. R5101
9	Press. R5101 to E5106	10	RPM G5110A
11	RPM G5110B	**12**	**Temp. E5101**
13	Temp. from E5101	**14**	**Temp. Oil E5101**
15	**Temp. Oil E5101**	16	Temp. G5111
17	Level D5100	**18**	**Press. D5100**
19	**Temp. D5100**	20	RPM G5103A
21	RPM G5103B	22	Temp. D5101
23	Level D5101	24	Level D5101
25	Temp. D5501	**26**	**Temp. D5101**
27	RPM G5521	**28**	**Press. G5521**
29	**Flow G5504**	30	Flow Additive 1
31	Flow Additive 2	32	Flow Additive 3
33	Flow Additive 4	34	Flow Additive 5
35	Flow Additive 6	**36**	**Flow Additive 7**
37	**Abs. G5504**	38	Press. G5504
39	**Press. H5518**	40	Press. H5518
41	**Press. to H5501**	**42**	**Press. G5504**
43	Press. G5504	44	Press. to H5508
45	Press. to H5508	46	Press. to Y5504
47	**Press. to EY5515/1**	**48**	**Press. to EY5515/1**
49	Press. from G5505/S	50	Press. Y5504
51	Press. to EY5515/2	52	Press. to EY5515/2
53	**Press. to H5501**	**54**	**Press. to H5501**
55	Press. to H5508	56	RPM G5504
57	Temp. G5504	58	Temp. EY5515/1
59	Temp. EY5515/2	60	Temp. EY5515/2
61	Temp. EY5515/2	62	Temp. Oil EY5515/1
63	Temp. Oil to H5501	64	Temp. Oil to H5508
65	Temp. Oil to EY5515/2	66	Temp. Oil G5504
67	Temp. add H5518	**68**	**Temp. G5504**
69	**Temp. G5504**	**70**	**Temp. H5508**
71	Press. to E5510	72	Press. to E5510
73	Temp. Oil to E5510	74	Level D5514
75	Press. D5514	76	Temp. H5514
77	Temp. D5514	78	Temp. D5514

Num.	Description	Num.	Description
79	Temp. D5514	80	Temp. from D5514
81	Temp. Oil D5514	82	Abs. G5103A
83	Abs. G5103B	84	Temp. G5103A/B
85	Press. G5103A/B	86	RPM G5103A
87	Temp. Oil G5517	88	Press. FY5506_1
89	Temp. FY5506/1	90	Press. PY5506_1
91	Level E5505B	92	Press. E5505B
93	Temp. D5504	94	Temp. E5505B
95	RPM G5103B	96	Level D5101
97	Press. Y5504	98	RPM G5504
99	RPM G5103A	100	Total R5001
101	**Additive 3+4**	102	RPM main stream

VI

Conclusion and future developments

Thesis collects three different applications that involves petrochemical productions related to the polymer manufacturing. These tasks have been fundamental to demonstrate the usefulness of multivariate data analysis together with polymer production issues such as the process monitoring, troubleshooting and monitoring of production quality. The consistency and effectiveness of the derived multivariate models, most of which have been implemented on-line and are being used for monitoring at the plant, made Versalis management confident in the multivariate approach. In fact, they supported the prosecution of projects started during the doctorate and actually are sponsoring new projects following the approaches developed in the Thesis. This point makes certainly PhD project successful, the high competition in petrochemical market highlights the necessity of process improvement and the applied multivariate methods allowed better production understanding, often without additional costs. Each single application returns interesting conclusions and shall have further developments. A brief summary for every tasks follows.

Concerning batch EPS production, the analysis of historical data depicted critical issues, moreover it has been proposed an innovative solution to monitor the main parameter influecing process quality based on infrared spectroscopic for which Versalis obtained an international patent. Plant management approved the online NIR application and in the near future shall be installed on line on the R401A, the reactor dealt with in Chapter 3. Further developments, would concern merging information from on-line spectroscopy monitoring and process data collected by plant sensors. It is expected that a global (data-fused) model would be capable of preventing dumps and to optimize the duration of production.

The ST20 process routinely use the monitoring software developed as result of troubleshooting study, Chapter 4 An immediate feedback of increasing production variance makes easier the plant monitoring also for researchers and management, which often do not have either direct or constant access to the control rooms. The *ad hoc* graphical interface developed during Thesis, where the monitoring model is implemented and automatically fed by Versalis database, allows any operator to monitor continuously the ST20 production.

About the last application, Chapter 5, continuous EPS production is actually supported by the value of inlet and outlet molecular weights, respectively by a theoretical and multivariate model. After an experimental phase in which the data have been downloaded on a stand-alone computers, from mid-2015 plant controllers are reading these two data on at-line computers. The future aim is to verify how much the timely Mw prediction advantages the production, in order to prove the effect of better process understanding in terms of saved money.

Finally, the main outcome has been the recognition of the support given to the various polymers production plants and to emphasize the amount of information that controllers lose monitoring one variable at time. Companies store mountain of data that are too often wasted instead of transforming them in precious information to improve process and product quality. Each applications proved that in every processes in which data are abundant, structured and noisy, and for sure this is the case of chemical and petrochemical production, the multivariate approach results far more successful than classical univariate methods.

Patent WO 2015/075629 A1

APPLICANT: VERSALIS S.P.A.

INVENTORS: BONACINI, Francesco; MANTOVANI, Erik; FERRANDO, Angelo;

Method for monitoring a control parameter of a polymerization reaction and relative apparatus for implementing said method

PUBBLICATION DATE: 28 May 2015

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 May 2015 (28.05.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/075629 A1

(51) International Patent Classification:
G01N 21/85 (2006.01)

(21) International Application Number:
PCT/IB2014/066129

(22) International Filing Date:
18 November 2014 (18.11.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
MI2013A001916 19 November 2013 (19.11.2013) IT

(71) Applicant: **VERSALIS S.P.A.** [IT/IT]; Piazza Boldrini, 1, I-20097 San Donato Milanese (MI) (IT).

(72) Inventors: **BONACINI, Francesco**; Via Gioacchino Basevi, 18, I-46100 Mantova (IT). **MANTOVANI, Erik**; Via G. Romano, 19, I-46036 Revere (MN) (IT). **FERRANDO, Angelo**; Via Arnaldo Terzi, 8/2, I-16039 Sestri Levante (GE) (IT).

(74) Agent: **BOTTERO, Carlo**; c/o BARZANO' & ZANARDO MILANO S.P.A., Via Borgonuovo, 10, I-20121 Milano (IT).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) Title: METHOD FOR MONITORING A CONTROL PARAMETER OF A POLYMERIZATION REACTION AND RELATIVE APPARATUS FOR IMPLEMENTING SAID METHOD

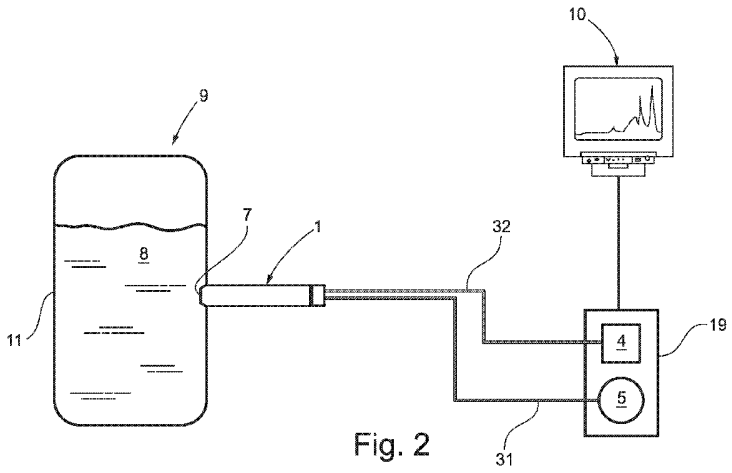

Fig. 2

(57) Abstract: The present invention relates to a method for monitoring a control parameter of a polymerization reaction mixture in heterogeneous phase comprising the following steps: (a) acquiring at least one NIR reflectance spectrum of said mixture; (b) calculating a value of said control parameter by means of a calibration curve which correlates the NIR reflectance spectrum with the values of said control parameter measured with a reference measurement method. The present invention also relates to an apparatus for implementing said method.

WO 2015/075629 A1

METHOD FOR MONITORING A CONTROL PARAMETER OF A POLYMERIZATION REACTION AND RELATIVE APPARATUS FOR IMPLEMENTING SAID METHOD

The present invention relates to a method for monitoring a control parameter of a polymerization reaction and the relative apparatus for implementing said method.

As is known, in processes for the industrial production of polymers, the possibility of continuously controlling the conditions at which the polymerization reaction takes place in order to obtain a high-quality product and high production yields, in addition to guaranteeing the management of the production plant under safety conditions, is of fundamental importance.

The control of an industrial polymerization process is generally based on the monitoring of some physical and chemical parameters of the reaction mixture, such as for example, the temperature and pressure inside the reactor, and the state of progress of the reaction.

In the case, for example, of the suspension synthesis of expandable polystyrene (EPS), the most important control parameters to be monitored include the size of the particles or beads of polymer which are formed and grow as the reaction proceeds and the conversion degree of the styrene monomer into polymer.

Various techniques and devices are available in the state of the art, which allow the control parameters of polymerization reactions to be accurately monitored.

Among these, in particular, monitoring techniques based on Near Infrared Spectroscopy (NIR) are receiving increasing attention, as they allow the above parameters to be rapidly and precisely measured in-line, i.e. directly on the reaction mixture inside the reactor.

In-line measurement devices based on NIR spectroscopy generally comprise a measurement probe connected by means of optical fibers to a light source and to a spectrophotometer. The probe is housed in the reactor, with the measurement head immersed in the polymerization mixture. The measurement head comprises a cavity inside which the polymerization mixture subjected to analysis flows. During the measurement, the probe irradiates the polymerization mixture through a window transparent to NIR radiation (sampling window) present in the measurement cavity, with a light radiation having a wavelength in the near infrared (incident radiation) produced by the light source and, at the same time, collects the light radiation scattered by the reaction mixture – created by the effect of irradiation with the incident radiation – sending it to the spectrophotometer.

In the technical field of the present invention, NIR spectroscopy is substantially applied in so-called "transmittance" or "in transflectance" modes.

In transmittance spectroscopy, the radiation analyzed by the spectrophotometer is the fraction of

incident radiation that passes through the sample, i.e. the fraction which is neither absorbed nor reflected thereby.

In "in transflectance" spectroscopy, the radiation analyzed by the spectrophotometer is the fraction of incident radiation which, after passing through the sample, is reflected by a specific reflecting screen situated in the measurement cavity along the pathway of the radiation, beyond the sample; the radiation reflected by the screen passes through the sample a second time, before reaching the spectrophotometer.

By means of chemometric calibration methods, the absorption characteristics of a NIR spectrum of a polymerization mixture can be correlated with the desired control parameter (e.g. average size of the polymer particles or conversion degree of the starting monomer), obtaining a predictive mathematical model of the control parameter. After an appropriate validation of the calibration method, the predictive model can be used for estimating the value of the control parameter of a polymerization mixture from a NIR spectrum measured on it in-line.

The monitoring methods of control parameters of polymerization reactions in heterogeneous phase by means of NIR spectroscopy known in the art have various critical aspects.

In these reactions, the polymerization mixture tends to stick to the sampling window of the immersion

probe due to its highly viscous nature. A residue of material consequently accumulates on the window as the reaction proceeds, which interferes with the measurement, preventing the correct determination of the control parameter. The fouling of the sampling windows therefore requires frequent stoppages of the plant to allow cleaning or substitution of the probe.

NIR probes, moreover, generally consist of a metallic body having an elongated form which is inserted in a wall of the reactor so that the measurement head remains constantly immersed in the polymerization mixture. The body of the probe extends from the wall of the reactor towards its interior, in some cases protruding as far as a distance of approximately 20 centimetres. Consequently the probe, with its dimensions, affects the fluid-dynamics of the polymerization reaction.

An example of the use of transmittance NIR spectroscopy for the in-line monitoring of the control parameters of the synthesis of polystyrene in aqueous suspension and the critical aspects indicated above, is described in A.F. Santos *et al.*, Journal of Applied Polymer Science, Vol. 70, 1737-1745 (1998).

The drawbacks of in-line measurements can be partially overcome by using NIR measurement systems operating on-line. In on-line measurement systems, the sample to be analyzed is removed, either continuously or discontinuously, from the reactor during the

polymerization reaction and transferred to the measurement instrument through a transfer line. In on-line measurements, the absence of a probe immersed in the polymerization mixture, however, does not eliminate the problem of fouling of the sampling windows, as the polymerization mixture must in any case pass into a measurement cell (positioned outside the reactor and connected to this through a transfer line) with problems completely analogous to those of the measurement cavities of immersion probes. Furthermore, in many cases, during the transfer of the sample of the reaction mixture from the reactor to the measurement instrument, there may be a structural alteration in the sample itself, which can also significantly influence the result of the determination of the control parameter (for example, coalescence phenomena of the particles may arise with a consequent overestimation of the real dimensions of the particles of the polymerization mixture).

The critical aspects illustrated above have so far prevented an adequate exploitation of the potentiality of NIR spectroscopy in the monitoring of control parameters of polymerization reactions in heterogeneous phase. In the case of the synthesis of EPS, for example, in the light of the drawbacks discussed above, the use of evaluation methods of the average size of polymer beads during the reaction, based on visual inspection of the polymerization mixture on the part of

an operator or on an analysis of samples of the reaction mixture effected in a laboratory (so-called off-line measurements), is still frequent.

The objective of the present invention is to overcome or at least to reduce the drawbacks of the state of the art indicated above.

Within this general objective, an objective of the present invention is to provide a method for monitoring a control parameter of a polymerization reaction in heterogeneous phase which is reliable, capable of guaranteeing a monitoring of the above parameter also continuously, and which does not require frequent maintenance interventions of the equipment.

In the light of these objectives and others that will appear more evident hereunder, according to a first aspect, the present invention relates to a method for monitoring a control parameter of a polymerization reaction mixture in heterogeneous phase comprising the following steps:

(a) acquiring at least one NIR reflectance spectrum of said mixture;

(b) calculating a value of said control parameter by means of a calibration curve which correlates the NIR reflectance spectrum with the values of said control parameter measured with a reference measurement method.

According to a second aspect, the present invention relates to a control method of a polymerization

reaction in heterogeneous phase which comprises the phase of monitoring at least one control parameter of said reaction according to the above-mentioned monitoring method.

According to a further aspect, the present invention relates to an apparatus for implementing the above monitoring method which comprises at least one polymerization reactor equipped with at least one acquisition system of NIR reflectance spectra comprising:

- at least one probe for irradiating a light radiation onto a polymerization reaction mixture contained in said reactor and substantially only collecting the radiation reflected from said mixture in response to said irradiation,

- at least one detection system optically coupled with said probe for detecting said radiation reflected.

The Applicant has surprisingly found that the drawbacks of the state of the art can be overcome using a method for determining the control parameters of a polymerization reaction in heterogeneous phase, wherein the NIR spectra of the polymerization mixture are acquired in reflectance, rather than in transmittance or transflectance. NIR reflectance spectroscopy, frequently used in the state of the art for the characterization of solid surfaces, in fact, allows optical fiber probes to be used, which do not require being immersed in the polymerization mixture to allow

the NIR spectra to be acquired in-line, but only require that they be in contact with it.

Thanks to the fact that the probes for NIR reflectance spectroscopy are based on a different spectroscopic analysis procedure with respect to that of probes for transmittance and transflectance spectroscopy, they do not have measurement cavities in which the reaction mixture must flow, and they are consequently much less subject to fouling phenomena. Probes for NIR reflectance spectroscopy, in fact, acquire the spectra through a sampling window which only requires being in contact with the polymerization mixture. These probes can therefore be substantially aligned with the internal wall of the polymerization reactor.

The advantages that can be obtained with the use of the reflectance spectroscopy according to the present invention are even more surprising considering that the intensity of the radiation reflected by the polymerization mixture that can be collected by a probe (and subsequently processed by the detection system) is much lower with respect to what can be collected with transmittance or transflectance measurement probes.

The method according to the present invention and the relative apparatus for implementing it are described hereunder for a better understanding of the characteristics of the present invention, with reference to the following figures:

- Figure 1, which shows a schematic representation of a probe for NIR reflectance spectroscopy according to the state of the art, which can be used for the purposes of the present invention;

- Figure 2, which shows a schematic representation of an apparatus for implementing the present invention.

With reference to Figure 1, the probe 1 for NIR reflectance spectroscopy comprises a body 2, generally in cylindrical form, in which at least one bundle of optical fibers 3 is housed. The above body 2 is coupled at one end with a measurement head 6, which, in turn, is coupled with a sampling window 7.

A first fraction 32 of said bundle of optical fibers 3, for example those positioned in the internal part of the bundle, is optically connected, at a first end, to a light radiation source 4 (for example a tungsten halogen lamp) and, at a second end, to the sampling window 7. A second fraction 31 of said fibers, for example, that at the outermost part of the bundle 3, is, on the other hand, optically connected, at a first end, to a detection system 5, to which the radiation reflected from the sample due to the irradiation effected with the first fraction of fibers 32, is conveyed, and, at a second end, to the sampling window 7.

In an alternative configuration, the fraction of optical fibers connected to the light radiation source 4 can be at the outermost part of the bundle 3, whereas

the fraction of optical fibers that conveys the radiation reflected to the detection system 5, is in the innermost part.

The sampling window 7 is made of a material transparent to light radiation (for example, quartz or sapphire) coming from the light radiation source 4.

When functioning, the probe 1 can irradiate, through the sampling window 7, the light radiation coming from the source 4, through the optical fibers 32, onto a polymerization mixture 8 positioned in a polymerization reactor 9, and, at the same time, substantially collect only the radiation reflected from said mixture 8 in response to said irradiation. The radiation reflected is then sent by means of the optical fibers 31 to the detection system 5, which processes the light signal and sends it to an electronic processor 10 where it can be processed and visualized in the form of a NIR spectrum (for example, an absorption unit spectrum with respect to a wavenumber).

The light radiation source 4 and the detection system 5 can be both advantageously housed in a spectrophotometric measurement device 19 (figure 2).

The probe 1 is inserted inside the polymerization reactor 9 so that the sampling window 7 is in contact with the polymerization mixture 8. The probe 1 is preferably inserted inside a wall 11 of the reactor 9, in a substantially transversal direction with respect

to this. In this configuration, the sampling window 7 is substantially parallel to said wall 11 of the reactor 9. The surface of said window 7 facing the polymerization mixture is also substantially aligned with the internal surface of the wall 11 of the reactor 9. Said surface of said window 7, in fact, protrudes with respect to the surface of the internal wall 11 of the reactor by less than 1 cm, preferably less than 0.5 cm, even more preferably less than 0.2 mm. This arrangement of the probe in addition to guaranteeing a significantly low fouling of the window 7, does not influence the fluid-dynamics of the polymerization reaction inside the reactor 9, as the probe 1 occupies a negligible fraction of the internal volume of the reactor 9.

It cannot be excluded, however, that the probe 1 can be inserted in the reactor 9 also in other configurations (for example, with the measurement head 6 tilted with respect to the wall 11 of the reactor 9), as the absence of a measurement cavity in any case guarantees a limited fouling of the sampling window 7 during the reaction.

With the monitoring method of the present invention, up to a hundred polymerization cycles (batch) can be consecutively effected in the same reactor without having to intervene for cleaning the probe between one cycle and the next.

The method according to the present invention can

be used for continuously or discontinuously monitoring the trend of one or more control parameters of a polymerization reaction in heterogeneous phase. The values of the control parameter acquired in the monitoring can be advantageously used for controlling the progress of the reaction and possibly for intervening on the same, for example by modifying the operating parameters of the reactor and/or equipment connected therewith.

The method according to the present invention can be advantageously applied to the monitoring of polymerization reactions of α-olefins having general formula $R_1R_2C=CH_2$, wherein:

- R_1 is hydrogen or methyl;
- R_2 is a group selected from: C_1-C_{10} alkyl, C_1-C_6 aryl possibly substituted with one or more groups selected from halogen, C_1-C_4 alkyl and C_1-C_4 alkoxyl.

The application of the method according to the present invention to the monitoring of a synthesis reaction of polystyrene, preferably in aqueous suspension, is particularly preferred.

The control parameters that can be monitored with the present method substantially include parameters correlated to the physico-chemical properties of the polymerization mixture that influence the characteristics of its NIR spectrum.

Preferred control parameters are: the average particle size (or diameter) of the polymer formed

during the polymerization reaction and the conversion degree of at least one monomer of the polymerization mixture (reference monomer).

In the case of the monitoring of the average particle size, the method is particularly effective for determining the average dimensions of polymer particles within the range of 100 μm – 3 mm, preferably 300 μm – 1.5 mm. Particles having dimensions within the above range, in fact, cause interference phenomena with the incident light radiation which generate NIR spectra with a reduced background noise.

The invention is further described hereunder with reference to the case of the monitoring of a control parameter of the synthesis reaction of expandable polystyrene (EPS) in aqueous suspension, in particular the monitoring of the dimensions of polymer particles formed during the reaction. This embodiment should in any case be considered as being a preferred and non-limiting embodiment of the application scope of the invention.

The synthesis reaction of EPS is a radical polymerization reaction of styrene. The reaction is generally carried out at a temperature within the range of 80-200°C in the presence of one or more radical initiators.

The suspension polymerization reaction is effected by dispersing styrene in water inside a reactor (e.g. autoclave). The reactor is typically equipped with a

stirring system, a heating system of the polymerization mixture and a cooling system for removing the heat formed during the reaction.

The reaction mixture generally comprises styrene and water in a styrene:water weight ratio within the range of 0.5:1-1:0.5.

In order to favour the dispersion of styrene in the water and/or avoid coalescence phenomena between the polymer particles, the polymerization mixture can comprise suspending agents, such as inorganic salts of phosphoric acid (e.g. calcium triphosphate) and/or anti-caking agents, such as surfactants (e.g. dodecylbenzenesulfonate), polyvinyl alcohol, and polyvinylpyrrolidone.

The suspending agents are generally present in the polymerization mixture in an overall quantity ranging from 0.001% to 1% by weight with respect to the weight of the polymerization mixture, said quantity also being selected in relation to the dimensions of the reactor and relative fluid-dynamics.

The anti-caking agents are generally present in the polymerization mixture in an overall quantity ranging from 0.1 to 1,000 ppm by weight with respect to the weight of the polymerization mixture.

Peroxide compounds, such as benzoylperoxide, tert-butylperoxy-2-ethylhexanoate and tert-butylperbenzoate, can be used, for example, as radical initiators.

Each radical initiator is generally present in the

polymerization mixture in a quantity ranging from 0.05 to 1.0% by weight with respect to the weight of the polymerization mixture.

In order to obtain expandable polymer beads, the polymerization mixture also comprises at least one expanding agent. Examples of expanding agents are: aliphatic or cycloaliphatic hydrocarbons containing from 3 to 6 carbon atoms (e.g. n-pentane, iso-pentane, cyclopentane), halogenated derivatives of aliphatic hydrocarbons containing from 1 to 3 carbon atoms (e.g. dichlorodifluoromethane, 1,2,2-trifluoroethane, 1,1,2-trifluoroethane), carbon dioxide, water and ethyl alcohol.

The expanding agent is generally present in the polymerization mixture in a quantity ranging from 2 to 10% by weight with respect to the weight of the polymerization mixture.

The polymerization mixture can also comprises chain transfer agents, expansion adjuvants, nucleating agents, plasticizers, etc.

In a preferred embodiment, the polymerization reaction is carried out in the presence of at least two peroxide initiators, the first active at a first temperature (e.g. 85-95°C), the second active at a second temperature, higher than the above first temperature (e.g. 110-140°C).

In this embodiment, the polymerization mixture is initially heated to the temperature at which the first

initiator is active (e.g. about 90°C) and kept at this temperature until the so-called "zero separation" point is reached.

The "zero separation" point corresponds to the point in which the density of the polymer which has been formed in the polymerization mixture is substantially equal to that of the water. After reaching the "zero separation" point, the temperature of the polymerization mixture is raised to the temperature at which the second initiator is active (e.g. about 115°C) and then kept at this temperature until the polymerization reaction has been completed.

The expanding agent can be added to the polymerization mixture from the beginning together with the styrene monomer or during the reaction, for example when the "zero separation" point has been reached.

During the reaction, within the polymerization mixture, particles of EPS (beads) are formed, which increase in size as the reaction progresses. The growth of the particles can be regulated, for example, by adding more or less high quantities of dispersing agents and/or anti-caking agents, which prevent the particles from coalescing with each other.

At the end of the reaction, the EPS beads are separated from the reaction mixture and subjected to washing, in order to eliminate the residues of reaction mixture, and drying.

The final product consists of EPS beads having a

substantially spherical form and an average diameter generally ranging from 0.2 to 2 mm, preferably from 0.5 to 1.0 mm.

In order to apply the method according to the present invention to the monitoring of a control parameter of a synthesis reaction of EPS, such as, for example, the above size of polymer beads, it is necessary to have a calibration curve capable of correlating a NIR reflectance spectrum of a polymerization reaction mixture of EPS with the values of this control parameter.

The calibration curve can be prepared with methods known to a skilled person in the field. The calibration curve is preferably obtained with univariate regression methods or, more preferably, by means of chemometric multivariate regression methods.

In order to obtain the calibration curve, for example, a plurality of sample polymerization mixtures (also indicated hereafter as "batch samples") can be prepared, each of which is subjected to polymerization according to the same pre-established temperature profile.

For the purposes of the present invention, temperature profile refers to the time sequence of the raising and lowering phases of the temperature of the polymerization mixture and relative maintenance periods of said mixture at each temperature. The temperature profile used for obtaining the calibration curve is

preferably that used for the same polymerization reaction in the production process to which the monitoring method according to the present invention is to be applied.

In general, the greater the number of batch samples used for preparing the calibration curve, the more accurate the determination of the control parameter will be during the polymerization reaction subjected to monitoring.

The number of batch samples is generally selected in relation to the number of variables that can influence the control parameter to be monitored.

The number of batch samples used for defining the calibration curve is preferably equal to at least 5, more preferably at least 10. In a particularly preferred embodiment, the number of batch samples ranges from 10 to 50.

A plurality of NIR reflectance spectra is acquired on each batch sample in order to obtain the calibration curve. Each of said spectra is acquired at a different advancement degree of the polymerization reaction.

The value of the control parameter(s) is determined in correspondence with the acquisition of each NIR spectrum, using a reference measurement method.

For the purposes of the present invention, a reference measurement method is a measurement method, different from that of the present invention, which allows the value of the control parameter of interest

to be determined on the polymerization mixture. For the purposes of the present invention, the value determined with the reference method is also indicated as "reference value" of the parameter.

Basically, methods known in the art commonly used for the monitoring of control parameters of polymerization reactions in heterogeneous phase can be substantially used as reference methods. The reference method used is preferably a method which allows the value of the control parameter to be determined with a low margin of uncertainty.

In the case of the average particle size of the polymer, the reference value can be determined by subjecting, for example, the polymerization mixture to the following analyses: acoustic spectroscopy, spectrophotometry (e.g. laser diffraction, dynamic light scattering) and image analysis.

The reference value of the average size of the polymer particles can also be obtained by visually comparing the polymer particles present in the polymerization mixture with standard samples of particles of the same polymer having known dimensions.

The reference method for determining the average particle size is preferably acoustic spectroscopy, more preferably ultrasonic spectroscopy.

When the control parameter is the conversion percentage of a reference monomer present in the polymerization mixture (in the case of EPS, the styrene

monomer) said parameter can be determined, for example, by means of refraction index analysis or thermogravimetric analysis of the polymerization mixture according to techniques known in the art.

The determination of the control parameter with the reference method can, for example, be effected by taking an aliquot of the batch sample of which the NIR spectrum has been acquired and subjecting it to an off-line measurement of the control parameter. The determination of the control parameter with the reference method can also be effected by means of on-line and in-line measurement methods.

As already specified, the NIR spectra used for the calibration are acquired on each batch sample at different advancement degrees of the polymerization reaction. The advancement degree, hereinafter also indicated as "reaction time", is the period of time that has elapsed between the beginning of the polymerization (time "zero") and a given moment during the reaction (for example, the moment of acquisition of a NIR spectrum).

In the case of the synthesis of EPS described above, the beginning of the polymerization is generally associated with reaching the lowest temperature at which one of the radical initiators present in the polymerization mixture is active.

The reaction times at which the NIR spectra of a first batch sample are acquired can be the same as a

second or further batch sample or different.

In the case of the synthesis of EPS, in a preferred embodiment, the NIR spectra and the determinations of the control parameter (e.g. average particle size) with the reference method are effected, for all the batch samples, at the temperature at which the first radical initiator is active (e.g. 90°C) and at different conversion percentages of the styrene monomer (for example: 10%, 30%, 50% and 65% of conversion).

In accordance with the present invention, in preparing the calibration curve, it is generally preferable to acquire the NIR spectra and determine the corresponding reference values of the control parameter of interest on a series of batch samples wherein the polymerization is carried out under such conditions that the value of said parameter is substantially independent of the conversion degree of the reference monomer.

In particular, in the case of the monitoring of the average particle size of EPS, the polymerization reaction is preferably carried out in each of the batch samples under such conditions that, with the same conversion degree of the monomer, each of the batch samples contains particles of EPS having different dimensions.

For the suspension synthesis of EPS, it has been observed that the particle size mainly depends on the stirring degree of the polymerization mixture and

quantity of suspending agent, whereas the conversion percentage of the styrene monomer mainly depends on the temperature profile adopted for the reaction and the quantity of radical initiator.

In the case of EPS, the above condition of independence of the particle size from the conversion degree can therefore be obtained by stirring the batch samples with a different intensity and/or using different concentrations of suspending agents and/or anti-caking agents.

The NIR spectra, both for obtaining the calibration curve (calibration spectra) and for determining the control parameter in the application of the method according to the present invention (measurement spectra), are preferably acquired within the range of wavenumbers 4,000-15,000 cm^{-1}, preferably within the range of 6,000-10,000 cm^{-1}.

The NIR calibration and measurement spectra are acquired in-line, using a NIR reflection probe of the type known in the art, for example as previously described with reference to figures 1 and 2 enclosed. For this purpose, the probe can be connected, for example, to a Fourier transform or to a diode array spectrophotometer.

The light radiation reflected from the sample analyzed and collected by the detection system can be advantageously processed, according to the techniques known to experts in the field, in the form of

absorption spectrum (A) in relation to the wavenumber (cm^{-1}) of the incident radiation. The absorption (A) is calculated starting from the reflectance value (R) measured on the basis of the relation A = log (1/R).

Once acquired, the calibration and measurement spectra can be pre-processed with methods known in the art in order to correct any possible spectral distortions due for example to shifts of the base line.

The calibration spectra and reference values of the particle sizes of EPS (determined with the reference measurement method) are analyzed with known univariate and/or multivariate linear regression mathematical-statistical methods in order to determine a mathematical correlation (calibration curve) between the spectroscopic characteristics of the NIR spectra and the values of the average dimensions of the polymer particles.

The multivariate linear regression method is preferably selected from: multiple least-square method, partial least-square method, method of the main components and combinations thereof.

The calibration curve resulting from the application of the above multivariate regression methods can be represented, for example, by the equation:

$$P = K_0 + K_1L(\lambda_1) + K_2L(\lambda_2) + \ldots + K_nL(\lambda_n),$$

wherein P is the value of the control parameter,

$K_{0, 1, 2, \ldots, n}$ are the linear regression coefficients

and

$L(\lambda_1)$, $L(\lambda_2)$,…, $L(\lambda_n)$ are the absorbance values (L) at the wavelength λ_n, or other value derivable from the absorbance.

The calibration curve obtained from the multivariate regression analysis is subsequently subjected to validation using a series of control batch samples prepared analogously to the batch samples used for the calibration curve and subjected to polymerization according to the same temperature profile adopted for the same.

Once validated, the calibration curve can be used for calculating the value of the control parameter (e.g. average size of the polymer particles), by applying it to a NIR spectrum acquired in-line on a polymerization mixture of which the evolution of said parameter during the reaction is to be monitored.

In another preferred embodiment, the method of the present invention can be used for monitoring the conversion degree of a reference monomer during a polymerization reaction in heterogeneous phase.

For this purpose, the calibration curve is constructed by correlating the values of the conversion percentage of the reference monomer, measured with a reference method, with the corresponding NIR spectra acquired in-line on the same batch samples.

The preparation of the batch samples, the acquisition of the spectra and corresponding reference

measurements of the control parameter and the mathematical-statistical treatment of the data collected for obtaining the calibration curve are the same as those previously described for monitoring the particle size.

In the case of the synthesis of EPS, the monitoring of the conversion degree of a reference monomer (i.e. styrene monomer) is particularly important for determining both the attainment of the "zero separation" point and also the completion of the reaction and consequently the moment for proceeding with the recovery of the reaction products.

In this respect, it should be noted that in some cases it may be difficult or even impossible to measure reference values of a control parameter under the same acquisition conditions of the NIR spectra. In the case of the synthesis of EPS, for example, NIR reflectance measurements can also be effected when the reaction mixture is at a temperature of about 200°C, whereas reference values of the conversion degree cannot be obtained at the same temperature, as samples of the polymerization mixture cannot be collected from the reactor at temperatures generally higher than approximately 90°C.

In these cases, a calibration curve which adequately predicts the values of the control parameter can only be obtained at a low temperature, i.e. within the temperature range at which the NIR calibration

spectra were acquired and at which the samples to be subjected to analysis with the reference measurement method could be collected, whereas at temperatures higher than the maximum limit of said range, the value of the control parameter predicted can deviate - in absolute terms - even significantly from the actual value.

The method, object of the present invention, however, in any case advantageously allows the maximum conversion point of the reference monomer to be identified with a good approximation. When the polymerization reaction approaches the maximum conversion point, in fact, it can be observed that the value predicted for the conversion percentage through the application of the calibration curve, even if it has significantly deviated from the actual value (generally higher than a 100% conversion value), it reaches in any case a maximum value which remains constant with time. Reaching this maximum value is therefore an index of having reached the maximum conversion degree of the reaction.

The following embodiment example is provided for further illustrating the invention.

EXAMPLE 1

In the case of a synthesis reaction of EPS in aqueous suspension, 13 batch samples were prepared for setting up a calibration curve.

Each batch sample subjected to polymerization

contained:

- water and styrene monomer in a weight ratio equal to 1:1 (3.5 kg each);

- benzoylperoxide (peroxide initiator active at 85-95°C), 0.4% with respect to the total weight of the polymerization mixture;

- tert-butylperbenzoate (peroxide initiator active at 110-120°C), 0.2% with respect to the total weight of the polymerization mixture;

- calcium triphosphate (suspending agent), 0.02% with respect to the total weight of the polymerization mixture;

- Na dodecylbenzenesulfonate (anti-caking agent) 20 ppm with respect to the total weight of the polymerization mixture;

- pentane (expanding agent), 8.5% by weight with respect to the total weight of the polymerization mixture.

For all the batch samples, the polymerization reaction was carried out at 90°C for 3.5 hours and was then continued at 115°C (temperature rise rate 0.5 °C/min) for 4.5 hours. Each polymerization reaction was effected in an autoclave having a capacity of 8 litres, equipped with a stirring system and a heating and cooling system of the polymerization reaction.

At the end of the reaction, the reaction mixture was cooled and the beads of EPS were separated, washed with water and dried.

During the reaction, the batch samples were kept under stirring by means of a blade stirrer, whose rate was set at a value ranging from 180 to 220 revs/minute (rpm).

The concentration of calcium triphosphate of each batch sample was selected within the range of 0.01-1% with respect to the weight of the polymerization mixture.

Each of the 13 batch samples was characterized by a different combination of stirring rate and calcium triphosphate values, so as to have, with the same conversion degree, a different average size value of the beads. In this way, the dimension of the beads was substantially prevented from depending on the advancement degree of the reaction.

5 or 6 NIR spectra were acquired for each batch sample at a temperature of 90°C for a total of 60 spectra. The spectra were acquired in correspondence with the following conversion percentages of styrene monomer: 10%, 30%, 50% and 65%.

The spectra were acquired by means of a NIR reflection spectra probe (Model GLADIUS of the company Hellma), by putting the sampling window in contact with the polymerization mixture. The probe was introduced transversally with respect to a wall of the autoclave, so as to have an internal surface of the sampling window substantially aligned with the internal surface of the wall of the autoclave.

A NIR Fourier transform spectrophotometer (BRUKER MATRIX F) was used for the acquisition of the spectra. The processing of the spectra for the calibration took into consideration the frequency range of 9,400 - 6,000 cm^{-1}.

The corresponding particle-size value was determined on the batch samples subjected to the NIR spectra measurements, using, as reference method, a visual comparison of a sample collected from each of said batch samples with a series of 10 reference samples of EPS beads having known dimensions ranging from 0.1 mm to 1.7 mm.

The calibration curve was obtained by applying a multivariate linear regression method based on the partial least square method, identifying 8 main components. The calibration curve was validated by means of cross-validation, obtaining an average square error equal to 0.05 mm.

Figure 3 shows the correlation graph between reference value (abscissa) and calculated value by means of the calibration curve (ordinate).

Figure 4 shows a prediction graph of the value of the average particle size (in mm) with respect to the reaction time obtained by applying the calibration curve, in which: the full circles (●) indicate the estimated value of the parameter; the empty circles (o) indicate the value determined with the reference method.

The results show that the method according to the present invention is sensitive to a variation in the average particle sizes and therefore suitable for effectively and reliably monitoring their growth during the polymerization reaction.

Furthermore, the polymerization cycles of the batch samples were effected in the same autoclave without the necessity of any cleaning or maintenance intervention of the probe.

After validation of the calibration, the method was used for continuously monitoring the growth of the EPS beads during the polymerization. Figure 5 shows some of the NIR spectra measured in-line during the polymerization reaction, in which the raising of the spectral base line linked to an increase in the average dimensions of the EPS beads, passing from about 300 micrometers (spectrum A below) to about 1.2 mm (spectrum B above), can be clearly seen.

CLAIMS

1. A method for monitoring a control parameter of a polymerization reaction mixture in heterogeneous phase comprising the steps of:

(a) acquiring at least one NIR reflectance spectrum of said mixture;

(b) calculating a value of said control parameter by means of a calibration curve which correlates the NIR reflectance spectrum with the values of said control parameter measured with a reference measurement method.

2. The method according to the previous claim, wherein said polymerization reaction is a polymerization reaction of one or more monomers selected from α-olefins having general formula $R_1R_2C=CH_2$, wherein:

- R_1 is hydrogen or methyl;

- R_2 is a group selected from: C_1-C_{10} alkyl, C_1-C_6 aryl possibly substituted with one or more groups selected from halogen, C_1-C_4 alkyl and C_1-C_4 alkoxyl.

3. The method according to the previous claim, wherein said polymerization reaction is a synthesis reaction of polystyrene in aqueous suspension.

4. The method according to one or more of the previous claims, wherein said control parameter is the average size of the polymer particles which are formed during said polymerization reaction.

5. The method according to the previous claim, wherein said average particle size is within the range of 100

μm – 3 mm, preferably 300 μm – 1.5 mm.

6. The method according to claim 4 or 5, wherein said reference measurement method is selected from: acoustic spectroscopy, spectrophotometry, image analysis and visual comparison of said polymer particles with standard samples of particles of the same polymer having known dimensions.

7. The method according to one or more of the previous claims, wherein said control parameter is the conversion degree of at least one monomer of the polymerization mixture.

8. The method according to the previous claim, wherein said reference measurement method is selected from refraction index analysis or thermogravimetric analysis.

9. The method according to one or more of the previous claims, wherein said calibration curve is obtained by correlating, by means of a multivariate linear regression method, a plurality of NIR reflectance spectra with a corresponding plurality of values of said control parameter; said spectra being acquired on a series of sample polymerization mixtures at different advance degrees of the polymerization reaction; said values of said control parameter being measured on said sample polymerization mixtures with a reference measurement method.

10. The method according to one or more of the previous claims, wherein said multivariate linear regression

method is selected from: multiple least-squares method, partial least-squares method, method of the main components and combinations thereof.

11. The method according to one or more of the previous claims, wherein said NIR reflectance spectra are acquired within the range of wavenumbers 4,000-15,000 cm^{-1}, preferably within the range of 6,000-10,000 cm^{-1}.

12. A method for the control of a polymerization reaction in heterogeneous phase which comprises the phase of monitoring at least one control parameter of said reaction according to one or more of the previous claims.

13. An apparatus for implementing the method according to claim 1 comprising at least one polymerization reactor 9 equipped with an acquisition system of NIR reflectance spectra comprising:

- at least one probe 1 for irradiating a light radiation onto a polymerization reaction mixture 8 contained in said reactor 9 and substantially only collecting the radiation reflected from said mixture 8 in response to said irradiation,

- at least one detection system 5 optically coupled with said probe 1 for detecting said radiation reflected.

14. The apparatus according to the previous claim, wherein said probe 1 comprises at least one measurement head 6 and at least one sampling window 7; said probe 1 being included in a wall 11 of said reactor 9, the

internal surface of said sampling window 7 being substantially aligned with the internal surface of said wall 11.

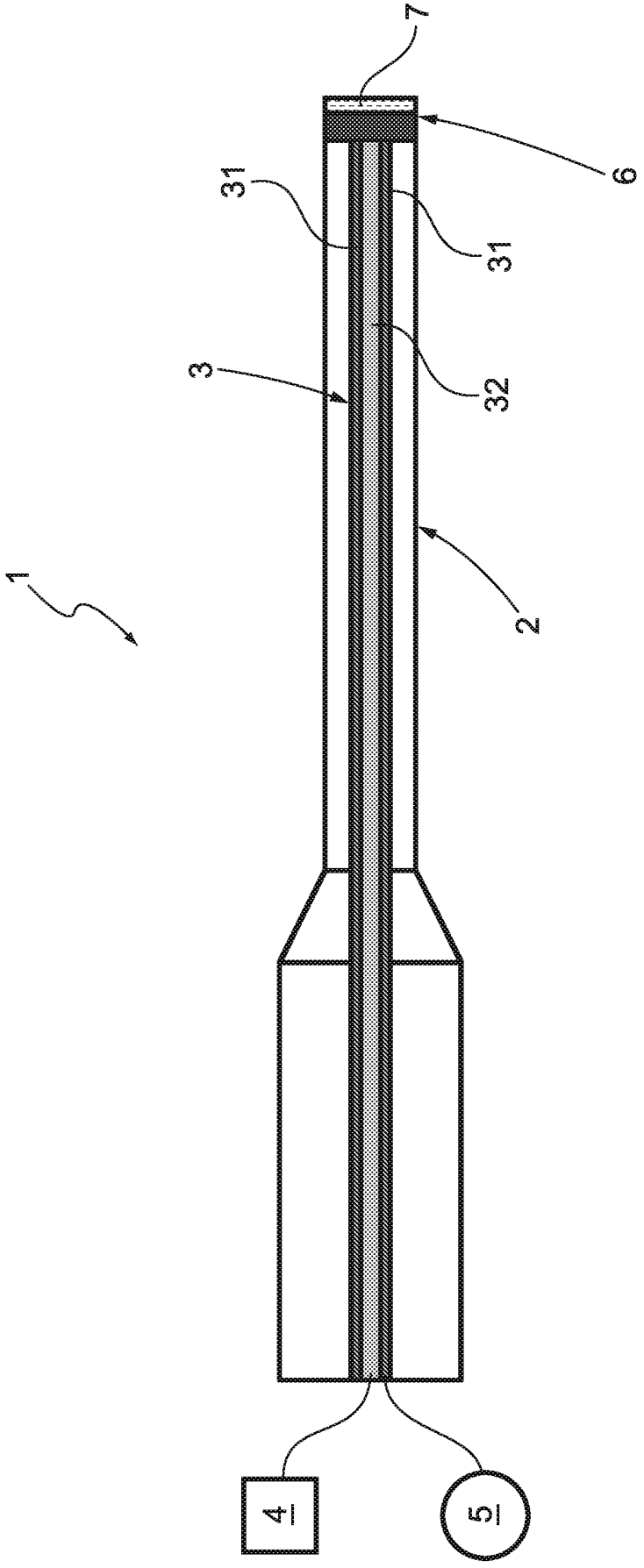

Fig. 1

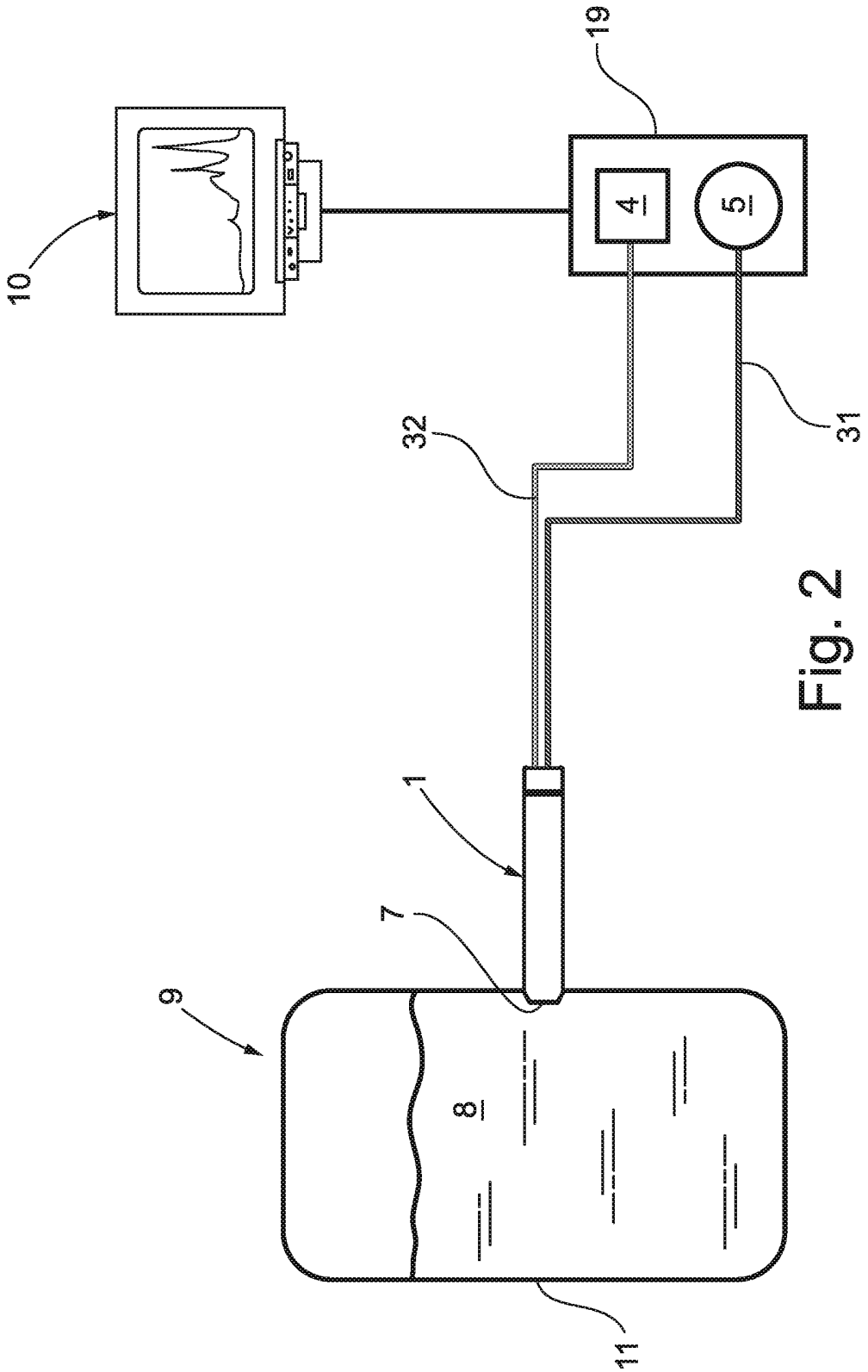

Fig. 2

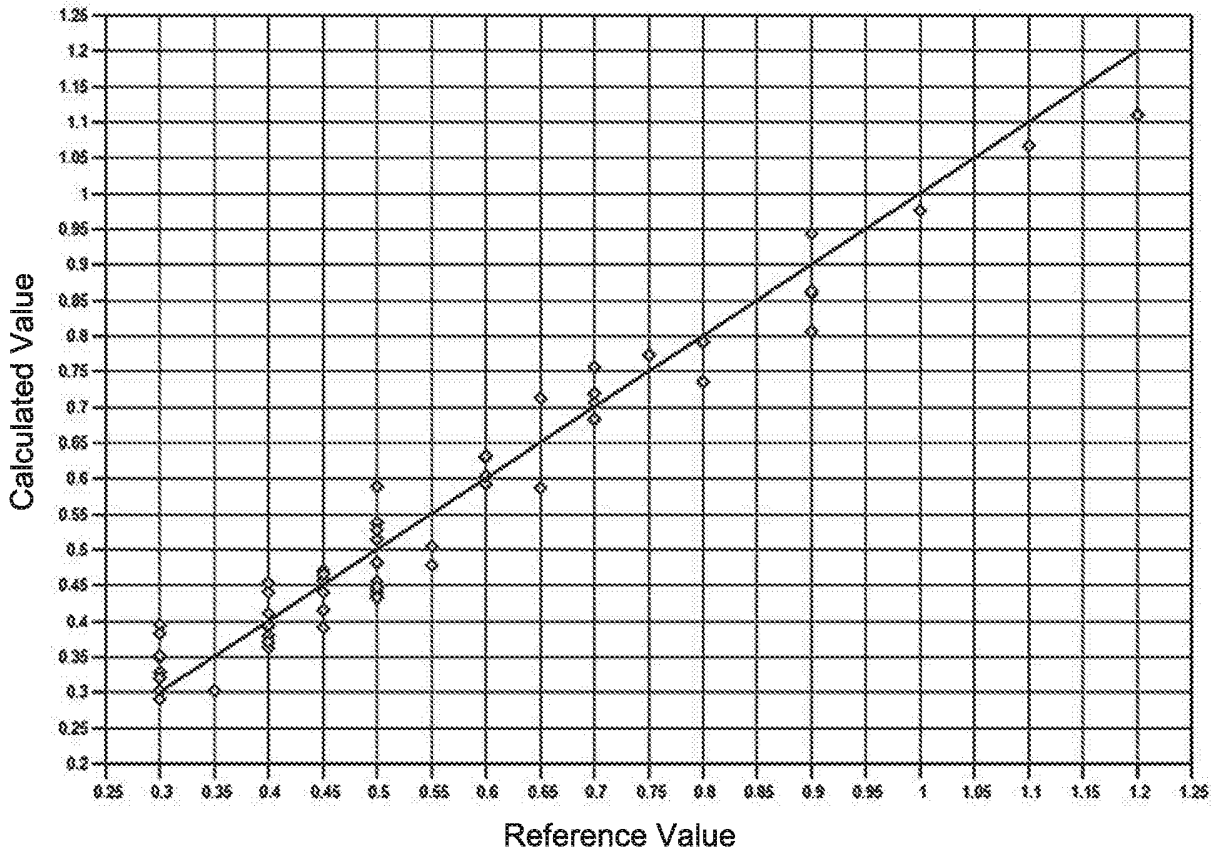
Calculated Value
Reference Value

Fig. 3

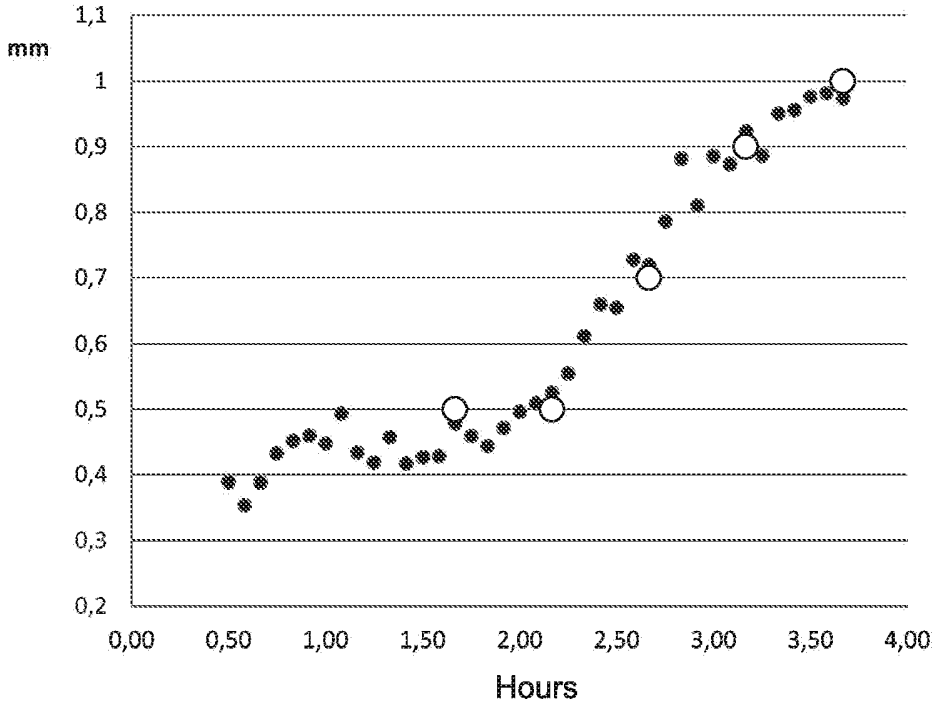
mm
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,00
0,50
1,00
1,50
2,00
2,50
3,00
3,50
4,00
Hours

Fig. 4

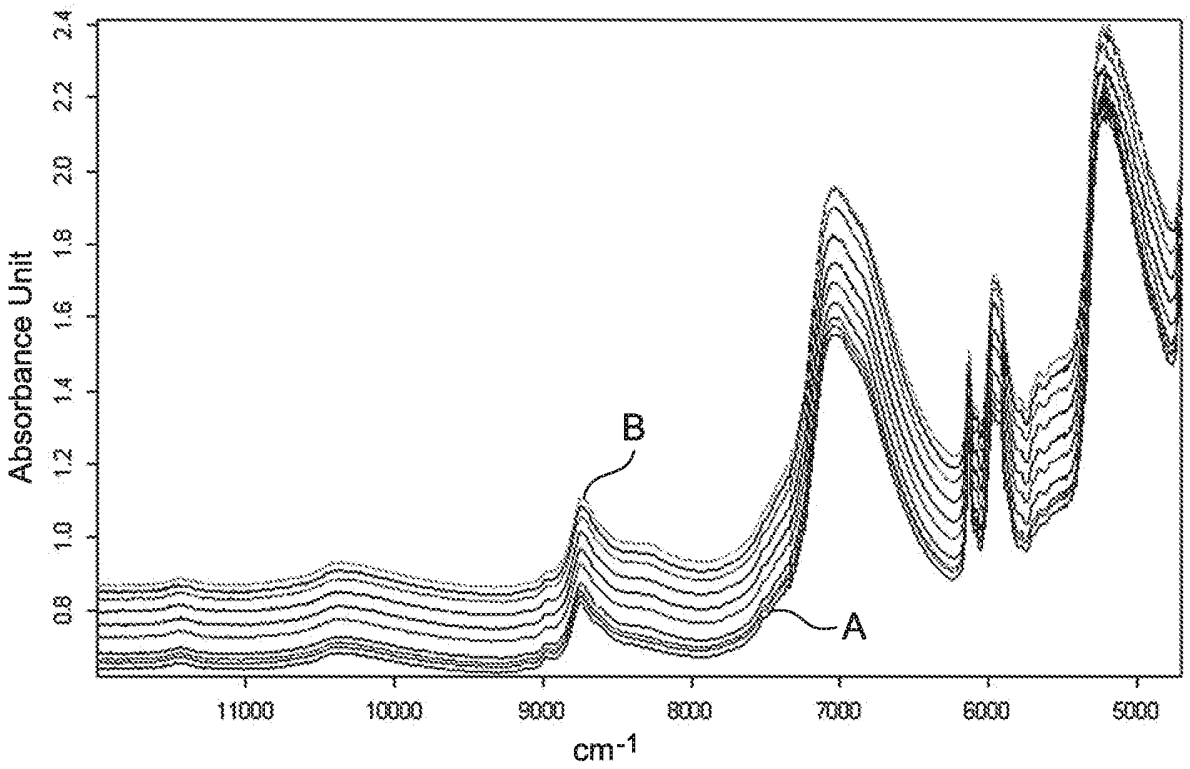
Absorbance Unit
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
B
A
11000
10000
9000
8000
7000
6000
5000
cm-1

Fig. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2014/066129

A. CLASSIFICATION OF SUBJECT MATTER
INV. G01N21/85
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2002/082383 A1 (KUROSE HIDEYUKI [JP] ET AL) 27 June 2002 (2002-06-27) abstract; figures 1,2 paragraph [0055] - paragraph [0058] paragraph [0074] -----	1-14
A	US 2012/203472 A1 (LACOMBE YVES [CA]) 9 August 2012 (2012-08-09) abstract; figure 1 -----	1-14

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 January 2015	20/01/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Vorropoulos, G

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/IB2014/066129

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2002082383 A1	27-06-2002	DE 60110360 D1	02-06-2005
		DE 60110360 T2	06-10-2005
		EP 1223186 A1	17-07-2002
		JP 2002194079 A	10-07-2002
		US 2002082383 A1	27-06-2002
US 2012203472 A1	09-08-2012	NONE	