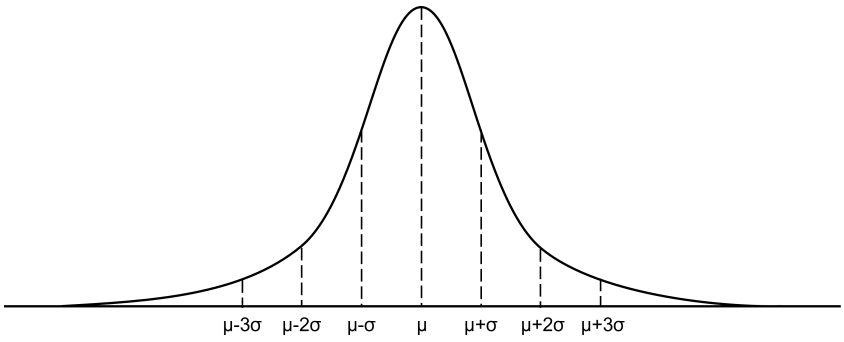
TÉCNICO LISBOA



API production process improvement project: a Six Sigma approach

Bernardo Maria Fernandes Ferreira

Thesis to obtain the Master of Science Degree in

Chemical Engineering

Supervisor(s): Prof. Henrique Anibal Santos de Matos
Dr. Pedro Miguel Guerreiro Felizardo

Examination Committee

Chairperson: Prof. Maria Cristina De Carvalho Silva Fernandes
Supervisor: Dr. Pedro Miguel Guerreiro Felizardo
Member of the Committee: Prof. Pedro M Castro

November 2021

“It is not necessary to change. Survival is not mandatory”

William Edwards Deming

Acknowledgments

This project would have not been so exciting and motivating without the help and contribution of the mentioned people and institutions.

Firstly, I would like to thank Hovione for the internship opportunity. It is safe to say it was a dream come true to end my academic journey at such a challenging company. To Pedro Felizardo for all the guidance and insights given on multivariate data analysis. It is truly a fascinating topic that I hope to master on the upcoming years. For all the continued support, the good reprimands, the tips on how to get what I needed and on how to ”move” within the company, a sentiment of sincere gratitude goes to Marilena Ornelas. She was crucial to the success of this project and without her help things would have moved at a rather slow pace.

I also would like to thank Professor Henrique Matos, whom I really appreciated working with, for the monitoring, feedback and encouragement provided since the beginning and throughout all the thesis duration.

To the production team, Pedro Santos, Bruno Marta, Susana Medinas and João Sequeira, thanks for the open-minded spirit to embrace the improvements and suggestions. You were relentless and it was really a pleasure working side by side with you. I also need to thank everyone at Hovione that in a way or another helped me during the project: Bernardo Matias, Rita Reis, Paulo Glória, Manuel Raposo and Ricardo Sousa among many others.

To all my friends made during my journey at Técnico. Things would have been boring and way more difficult without you guys. To the ”wonder-trio”, Rodrigo Amorim and José Fonseca, thank you for the unforgettable memories created during these five years. Each one of us shall follow his path but the moments we spent will stay forever on our memory. My friends outside of Técnico also played a big role on my journey through university. Thank you for the support when things looked unfeasible and for all the good moments and laughs.

I feel that words are not enough to thank my family. To my father, mother, brother and sister, thank you for everything, you made me what I am today. Thank you for the education, the values transmitted, the good and lesser good moments, for the tireless support and above all, for always believing in me. Also, to my extended family, my cousins, aunts and uncles and to my grandmother and grandfather. I am truly blessed.

Last but not least, I want to thank God for the gift of life and for everything and everyone that He has been putting in my path. I shall face the new challenges with Faith and Hope.

Resumo

A presente tese teve como objetivo a melhoria de um processo de produção de um API genérico no site da Hovione em Sete Casas. Pretendeu-se optimizar o rendimento do passo final através da metodologia do ciclo DMAIC. Durante o período analisado (de Julho de 2018 a Janeiro de 2021), o rendimento apresentou um valor médio de 81,83% com uma amplitude total relativa de 11%. Foi contabilizado que, em média e por ano, se perde um lote de produto final devido à variabilidade do rendimento. Foram utilizadas técnicas de análise de dados multivariada no sentido de se encontrar correlação estatística entre as características das matérias primas e as variáveis de processo com as variáveis dependentes consideradas. Verificou-se que a impureza H e G presentes na matéria prima para a etapa final do processo tinham um impacto negativo no rendimento. A análise estatística das variáveis de processo revelou que a cristalização é a operação mais crítica no passo final. Os modelos foram reconstruídos considerando a pureza do produto final em vez do rendimento como variável de resposta. A contribuição das variáveis de processo para a qualidade do produto estava na mesma linha do que para o rendimento. O processo conducente ao intermediário de entrada do passo final foi também analisado tomando as impurezas H e G como variáveis dependentes. Embora os dados só estivessem disponíveis para 6 lotes de produção, algumas acções de melhoria puderam ser retiradas a partir dos modelos. As acções de melhoria foram filtradas com base no seu impacto e esforço e foi elaborada uma folha de controlo interactiva, bem como um fluxograma contendo o conhecimento adquirido, a fim de manter as melhorias.

Palavras-chave: Indústria farmacêutica, Processo de produção de API, Ciclo DMAIC, Análise de dados multivariada.

Abstract

This thesis aimed at the improvement of a generic API production process at Hovione´s Sete Casas site. Yield optimization of the final product production step was the problem to solve through the DMAIC cycle methodology. Over the analysed period (from July 2018 to January 2021), yield had an average value of 81.83% with a relative range of 11%. It was accounted that, on average and per year, a full batch throughput is lost due to yield variability. Multivariate data analysis techniques were used in order to find statistical correlation between input quality attributes and process variables with the response variables. Impurity H and G present in the input material to the final product process step were found to negatively impact the yield. The statistical analysis of the process variables revealed that the crystallization is the most critical to yield operation. The models were re-built considering the final product´s assay instead of yield as response variable. The contribution of the process variables to the quality of the product was on the same line as for yield. The process leading to the input intermediary of the final product process was also analysed taking impurity H and G as response variables. Although data was only available for 6 production batches, some actions could be taken from the models. The improvement actions were screened based on their impact and effort and an interactive control sheet as well as a summary flowchart of the generated process understanding were elaborated in order to maintain the improvements.

Keywords: Pharmaceutical industry, API production process, DMAIC cycle, Multivariate data analysis.

Contents

List of Tables

List of Figures

Nomenclature

Greek symbols

μ	Mean.
σ	Standard deviation.

Roman symbols

$\overline{x}_i$	Sample mean.
$\tilde{x}_{ij}$	Data point after preprocessing.
CV	Coefficient of variation.
Max	Maximum value for a random variable.
Min	Minimum value for a random variable.
$Q2$	Predictive coefficient of determination.
Q_1	First quartile.
Q_2	Second quartile.
Q_3	Third quartile.
R	Range.
$R2$	Coefficient of determination.
s_i	Sample standard deviation.
x_{ij}	Data point before preprocessing.

Glossary

API is the Active Pharmaceutical Ingredient, which is used to make the drug product. It is the ingredient of any drug product that produces the intended effects. 3

CDMO *Contract Development Manufacturing Organization* is a company that serves other companies in the pharmaceutical industry on a contract basis to provide comprehensive services from drug development through drug manufacturing thus allowing the major company to focus on drug discovery and drug marketing. 3

drug product is the final product that combines the drug substance and excipients. 1

drug substance the same as API. 1

FDA *United States Food and Drug Administration* is a federal entity of the Department of Health and Human Services responsible for protecting and promoting public health through the control and supervision of food safety, tobacco products, dietary supplements, pharmaceutical drugs, vaccines, biopharmaceuticals, blood transfusions, medical devices, electromagnetic radiation emitting devices, cosmetics, animal foods and feed and veterinary products. 1

ICH *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* is an organization that connects regulatory agencies and pharmaceutical industry to discuss scientific and technical aspects of pharmaceutical product development and registration. 1

lead time the total time (calendar time) between the batch start and its completion. 5

LSL and USL lower specification limit (LSL) and upper specification limit (USL) are the limits imposed by the customer for a variable of interest. Outside of these limits, the customer or the health authority acceptance of the product or service will probably be impaired. 8

ppm parts per million. 8

SRM limiting Starting Raw Material used for a batch production. At Hovione, it can also be referred as the respective batch size for production.. 5

Chapter 1

Introduction

1.1 Motivation

The pharmaceutical industry has the noble mission of improving the quality of human life, striving to find new medicines and new ways to alleviate the burden of disease.

The *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (ICH) quality guideline Q8 (ICH Q8) defines quality in the pharmaceutical industry as ”the suitability of either a drug substance or drug product for its intended use. This term includes such attributes as the identity, strength, and purity” [1]. Further consideration can be added to this definition in terms of reliable clinical performance: a quality drug product ”delivers clinical performance per label claims and does not introduce additional risks due to unexpected contaminants” [2]. Integration of these two approaches is done considering that the clinical parameters that are crucial to good clinical performance are derived from the quality attributes of the drug product or substance [3]. Generally speaking, the well-defined quality attributes are a more reliable measure of drug quality since the control of these parameters is tighter and more straightforward to perform than an evaluation of clinical performance [3].

Since the *U.S. Food and Drug Administration* (FDA) report publication *Pharmaceutical Current Good Manufacturing Practices for the 21st Century* [4] on 2004, the industry´s approach to quality began to change from Quality by Testing (QbT) to Quality by Design (QbD) [5, 6]. In the traditional paradigm, quality is assured by a series of testing on raw materials and the final process output. Only when all specifications are met can the product be released to the market or proceed to the next step on the value chain (*e.g.* from chemical synthesis of the drug substance to formulation of the drug product) [3, 5]. When all specifications are not met, the batch has to be reprocessed or can even be rejected, leading to failure in meeting customer demand. It is estimated that, in the early two-thousands, 5 to 10% of the total batches produced in the industry needed reprocessing or were discarded [7]. The root causes for such failures are usually not fully understood due to poor process understanding. Under this standard, the manufacturers risk continuous losses until the root causes are identified and resolved or the criteria for batch approval are widen consequently promoting poor drug safety [3, 5]. The inflexibility of the manufacturing process [3] and extensive testing is what ensures drug product quality on the

described paradigm halting innovation on an industry that constantly needs to revamp itself.

Contrary to this traditional notion, comes the approach developed and coined by the quality pioneer, Dr. Joseph Juran, Quality by Design (QbD). This systematic, scientific and holistic perspective assures product quality on the design of the process thus eliminating the need for extensive testing on the final product [3, 8, 9]. ICH Q8 defines QbD as ”a systematic approach to development that begins with predefined objectives and emphasizes product and process understanding and process control, based on sound science and quality risk management” [1]. With an emphasis on process understanding and control, causes for variability can then be better identified and resolved thus increasing, in a consistent manner, product quality. Based on this definition, one could infer that the QbD approach may seem exclusive of process development [9] but that is not the case as process optimization comes in as critical to ensure process robustness over time, *i.e.*, that the process operates on a reliable way over a range of inputs (materials and process parameters) [1, 10].

Another initiative pushed by regulatory agencies, deeply related to ensuring process robustness over time [11, 12] is process analytical technology, PAT. As defined by ICH Q8 , PAT is ”a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality” [1]. It deepens the understanding and control of the process which are pillars of the QbD approach to quality, in order to develop and operate robust processes [10, 12].

The recent pursuit for quality in the pharmaceutical industry was combined with a quest for productivity increase that translates into effective use of the company‘s resources. Lean and Six Sigma have proven to succeed on the matter and although slower than other industries [13–15], began to gain ground on the turn of the century in the top pharmaceutical companies [14, 15].

The successful integration of quality and productivity, based on a scientific understanding of manufacturing processes, is, nowadays, on the top of the agenda of pharmaceutical companies [16, 17] paving the way for operational excellence programs in an attempt to manage cost, quality and time while at the same time focusing on customers needs [18, 19]. For some companies, the term OPEX may simply mean isolated initiatives linked with cost reduction or with Lean and Six Sigma, for others it may be a top-to-bottom cultural mindset with the engagement of every employee.

The need for a continuous improvement culture in the pharmaceutical industry, that relies on process understanding [10], is evident based on the number of methodologies and programs that have been launched and pushed by regulatory agencies and applied (or are still to be) since the beginning of the 21st century [20]. All these methodologies and programs have one ultimate and common objective: to continuously enhance the quality of drug products that will eventually lead to an improvement in the quality of human life. The tagline of Hovione, ”In it for life”, fits perfectly on this end goal of every pharmaceutical company.

1.2 Hovione

Hovione is a contract development and manufacturing organization, CDMO, that provides contract manufacturing services and licensing opportunities for proprietary products. Hovione is also present in the field of generic API products.

The company was founded in Portugal in 1959 by Ivan Villax and his wife, Dianne Villax, and two more Hungarian refugees: Nicholas de Horthy and Andrew Onody. Initially operating on the basement of Villax´s home in Lisbon, the company expanded and Hovione´s first industrial plant was built in 1969 at Sete Casas, Loures. At the time, being Japan the main market of Hovione´s products, Ivan Villax decided to open an office in Hong Kong in 1979 and later, a site in Macau in 1986. In 2001, Hovione opens a third manufacturing site in New Jersey, USA, reinforcing their global position. It was also during the turn of the century that Hovione extended its services by offering advanced and innovative solutions in the particle engineering field, becoming the leader in spray drying technology. In 2008 Hovione acquires Pfizer´s API site in Cork, Ireland. Reaffirming the company´s global footprint, offices were opened in Japan and India.

Figure 1.1: Hovione´s presence worldwide [21].

Today, Hovione employs 1,600 people worldwide and offers more than 590 m^3 of manufacturing capacity. The company owns more than 400 patents recognized worldwide and is the biggest employer of Ph.D. workers in Portugal.

Its core values are excellence, rigor, innovation, quality, and a strong customer commitment. The Sete Casas production department is divided into several areas according to the product´s destination (if it is produced exclusively to a specific client or if it is a generic) and grouped by the chemical similarity between molecules.

1.3 Topic Overview

Ivan Villax´s research that led to the creation of Hovione was focused on two classes of active pharmaceutical ingredients (API): tetracyclines and corticosteroids. The former class of molecules is produced naturally on the adrenal gland (located above the kidneys) and has a direct impact on stress and immune response, protein and carbohydrate metabolism, blood electrolyte levels, the regulation of inflammation, and behavior [22]. Since their discovery, corticosteroids have been used in almost every area of medicine and administered by nearly every route [23]. They are one of the most prescribed classes of drugs worldwide with an estimated 10 billion dollars per year in sales [24].

Fluticasone propionate (figure 1.2, $C_{25}H_{31}F_3O_5S$) is a corticosteroid that can be administrated via oral, nasal or topical route [25]. The route of administration depends on the condition to treat [25–27]:

- oral route (inhaled): to treat asthma and chronic obstructive pulmonary disease (COPD);
- nasal route: allergic and non-allergic rhinitis, nasal polyps, and allergies;
- topical route: eczema and dermatoses.

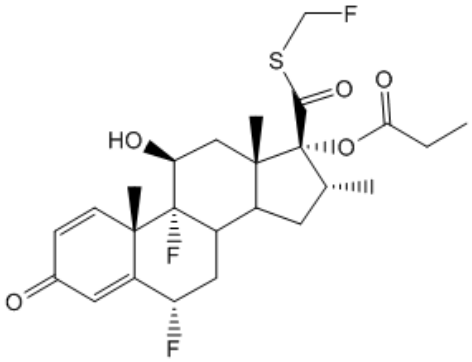


Figure 1.2: Fluticasone propionate chemical structure, $C_{25}H_{31}F_3O_5S$.

The compound was firstly patented by Glaxo in 1980 [28]. Nowadays Hovione holds the patent, producing it on Sete Casas installations by means of a process developed in its R&D center. The process can be divided into 5 steps, giving each an isolated intermediate as output.

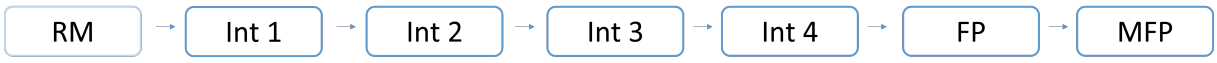


Figure 1.3: Process overview displaying the several steps that lead to the final API product.

RM is the raw material that starts the production train. The steps leading to intermediary 4 are chemical steps (chemical transformation occurs) while the step leading to the final product (FP) is a purification step and the step leading to MFP (micronized final product) is a size reduction step. Each chemical step involves dissolving the respective intermediate product with a designated solvent, followed by reacting with one or more reactants. At the end of each chemical reaction, the new intermediate is isolated either by pH adjustment, cooling, antisolvent addition, solvent evaporation, or a combination of these procedures, followed by a filtration, drying and packaging steps. The only difference to the purification step is

the absence of reaction. The output, in order to accomplish a subsequent successful formulation of the API, *i.e.*, improved drug solubility thus increasing bioavailability [9, 29–31], goes through a particle size reduction step on a jet-mill, leading to MFP.

The API (further referred to as FP) is produced by campaign on a multipurpose installation (also intended for APIs of the corticosteroids family) spread across two adjacent buildings on the plant. General metrics concerning the production of FP are given in table 1.1.

Table 1.1: Overview metrics for all the steps of the production process of FP on a time frame from 2018 to 2020. For average and expected yield, the standard deviation is also presented.

Process step	**Scheduled lead time (days)**	**SRM (kg)**	**Avg. yield (%w/w)**	**Expected yield (%w/w)**
Int 1	3.9	110	94 ± 0.6	92 ± 5
Int 2	10.8	102	102 ± 2.6	99 ± 8
Int 3	7.5	104	106 ± 2.2	107 ± 5
Int 4	5.1	112	103 ± 2.4	101 ± 10
FP	9.4	35	78 ± 15.6	70 ± 28
MFP	11.3	26	94 ± 2.9	95 ± 5

The yield of the production steps can be viewed as the throughput ratio. To better understand this variable (%w/w), the following formula should be considered:

$$Yield(\%) = \frac{Net\ weight\ obtained}{Net\ weight\ loaded} \times 100 \tag{1.1}$$

The formula used for yield calculation does not consider the purity of the final substance obtained nor the changes in molecular weight of the isolated intermediates or final product. As such, the ratio obtained can be higher than 100%, which should not be considered abnormal.

The uncertainty in the expected yield is increased moving up on the production train and stopping on the final API before size reduction (table 1.1). This increase is roughly accompanied by an increase in the standard deviation of the average obtained yield since 2018. High variability in the yield obtained ultimately leads to ineffective use of the production resources (equipment, personnel, utilities, etc) and so, in accordance with the production team, this was the primary problem to target. An in-depth analysis of this problem statement will be done in section3.1.

The work methodology followed a backwards approach, as illustrated in figure 1.4.

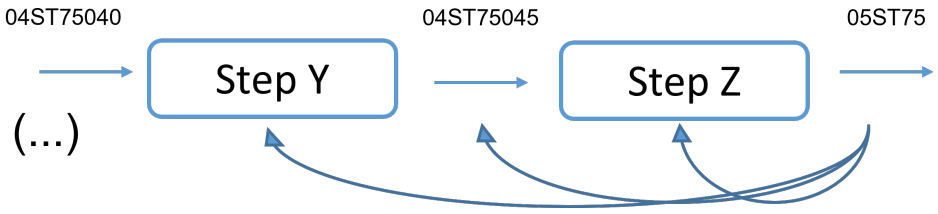


Figure 1.4: Work methodology followed on this thesis. The identification of the steps as step Z or step Y is only indicative that these are the final steps of the overall API production process and will not be used in this thesis.

The problem is identified on the process output and the cause(s) for the problem are searched firstly on the process that leads to the output. It is important to mention that for the present work, the process output was considered to be FP and not MFP. The step leading to the latter is just a size reduction step that does not involve any chemical operation. Losses of yield do happen due to particle deposition in the equipment chamber. According to table 1.1, these losses can be up to 10% and although not negligible they were not considered.

If the variability is explained by the conditions in which the last step runs, the cause for the problem is identified and improvement actions can be drawn from the analysis. If the cause for variability is not totally identified in the last step, then the input to this step must be analysed and so forth. In order to have concrete improvement measures, in the case that a certain attribute (or a set of attributes) on a raw material for a step explain the greatest part of the variability, the step leading to the production of that intermediate must also be analysed to uncover what causes that critical attribute.

In this way, the causes for variability on the output parameter will be spotted, improvement actions will be launched and tested. Process understanding will be further developed that leads, due to the implemented improvement actions, to an increase in process robustness.

1.4 Objectives

Inserted in the site´s continuous improvement plan, which ultimately aims at a throughput increase, this project will lead to the development of process understanding on a generic API product. Through a clear definition of the problem to be solved (yield optimization), process mapping and data collection, statistical analysis of the historical data, and the suggestion of improvements, the project will follow the DMAIC cycle as part of the Six Sigma approach to process improvement and problem-solving, that will be explained in detail further on.

The expected outcome is the establishment of improvement actions with a view to yield optimization drawn from an increased process knowledge (that should be transmitted to the production team) and the creation of sustainment tools in order to maintain and preserve what was uncovered during the project.

Chapter 2

Background

2.1 Six Sigma Origins

Ideas are never born out of the void. New ways of doing things arise from a set of events and conditions that enable change. Six Sigma origin was no different. By the early 1980s, Motorola faced serious issues related to poor product quality that was leading to a decrease in market share and the inability of staying on top of Japanese competitors [32, 33]. Bill Smith, a Motorola engineer at the time, was studying the correlation between a product performance indicator and the variability and rework percentage on the respective process concluding that processes with high variability and rework rate often led to higher failure in the field [33]. Dr. Mikel Harry was responsible to help Bill Smith formulate a tangible and applicable approach that could be used within the company to achieve near quality perfection through variability elimination [33, 34] and so, Six Sigma was born.

With the extreme support of the then chairman of the company, Bob Galvin, Six Sigma helped the company achieve enormous improvements that resulted on (between 1987 and 1997): five-fold increase in sales with profits going up 20% every year; $14 billion on savings in Six Sigma projects and a stock price increase with a 21.3% rate every year [35]. After Motorola´s extreme success, Six Sigma began to be applied by other top companies such as General Electric and Allied Signal (known as Honeywell) followed by an immense wave of adherence (in the early 2000s): Ford, Dupont, Dow Chemical, Johnson & Johnson, Kodak, Sony, Toshiba and many more [35].

2.2 Six Sigma System

The term ”Six Sigma” originated from the very core of the philosophy itself: reduce variability to improve quality. Assuming that a random process variable (X) follows a normal distribution with a designated mean, μ, and standard deviation, σ, then $X \sim \mathcal{N}(\mu, \sigma^2)$. The probability density function is given by the following equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})} \tag{2.1}$$

Graphically this mathematical relation can be represented as follows in figure 2.1.

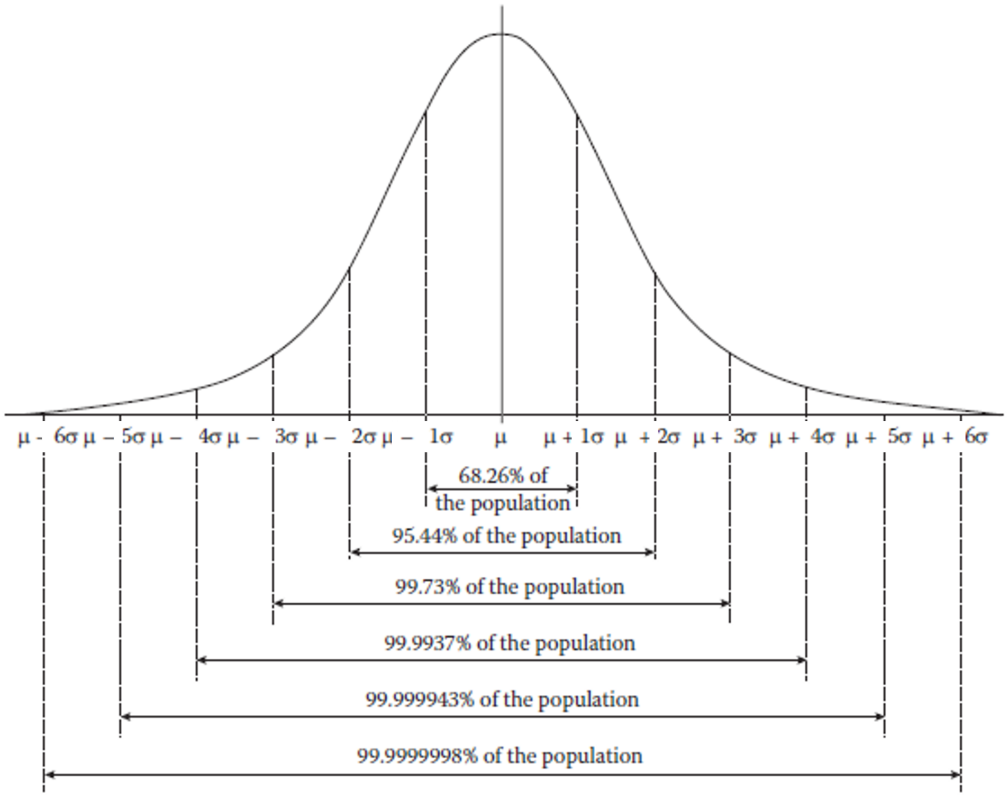


Figure 2.1: Normal distribution curve with a mean μ and standard deviation σ, exhibiting the percentage of area that lies within the $\pm\sigma$ intervals from the mean [32].

The percentage of data that lies within $\mu \pm 6\sigma$ is 99.9999998%. If the lower and upper specification limits for a certain process or product parameter (LSL and USL) are located at $\pm 6\sigma$ from the mean, then the proportion of defectives, *i.e.* the proportion falling outside the specification limits, would be 0.002 ppm. Allowing for a 1.5σ shift on the mean, one would get 3.4 ppm defectives (99,9996% within specification) [32]. This is the metric that the pioneers of Six Sigma at Motorola set out to achieve on all of their processes reflecting the goal of near perfection in terms of quality [32].

Under these circumstances, a process that (always allowing for a 1.5σ shift on the mean) produces 3.4 ppm defectives is said to have a 6σ capability. For example, if a process has its specification limits located at $\pm 3\sigma$ from the mean, then it will produce 66,800 ppm defectives: decreasing the sigma capability of the process increases the proportion of defectives as shown in table 2.1.

Table 2.1: Process sigma capability and proportion of defectives in ppm (proportion that falls outside the specification limits of the process) [32].

Sigma Capability	Proportion Defectives (ppm)
1	697,700
2	308,700
3	66,800
4	6,200
5	230
6	3.4

The traditional view of quality stated that if certain output attribute falls within the previously established specification limits there is no loss in terms of quality. Contrary to this approach, the Taguchi loss function, developed by the Japanese engineer Genichi Taguchi in the 1950s [36], states that every deviation from the target value for the attribute has a loss in the value of the product and so, a function was developed to quantify the loss as proportionate to the deviation [37].

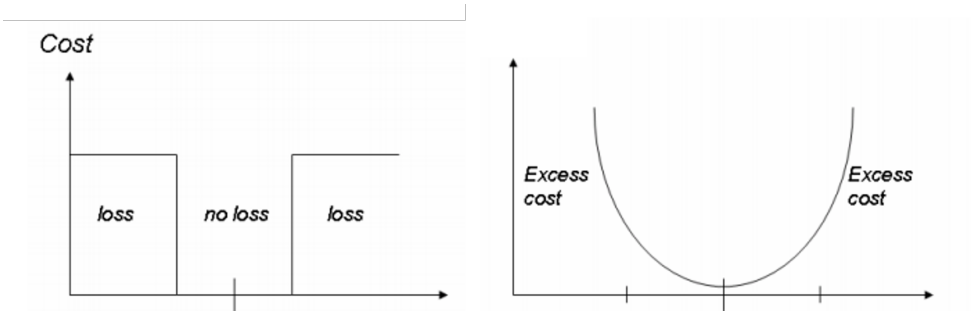


Figure 2.2: Traditional view and the Taguchi view on quality [38].

Taguchi loss function is credited for the increase in continuous improvement projects throughout the world and therefore, one of the foundations of the Six Sigma system[36]: reduce variation to achieve quality perfection.

2.3 Six Sigma Philosophy

Six Sigma is more than a statistical concept, used to measure the performance of products or processes against customer requirements in order to ensure quality perfection. Six Sigma can be seen as ”a comprehensive and flexible system for achieving, sustaining and maximizing business success” [35]. To accomplish business excellence, Six Sigma seeks to achieve this by carefully understanding customer expectations, use of facts through data collection and statistical analysis, and improving processes based on facts and data [32]. Business success is a broad concept and can be materialized in many fields such as cost reduction; productivity improvement; market share increase; cycle time reduction; defect reduction; corporate cultural change and many more [32, 35]. Six Sigma is more of a management philosophy, that translates into a cultural belief within a company than just a set of statisti-

cal tools. However, these tools are the foundation of the system and are what provide the reasoning for managerial decisions.

2.4 DMAIC Cycle

Every process can be defined, measured, analysed, improved, and controlled (DMAIC). Six Sigma views all work as processes and so, all work can be defined, measured, analysed, improved, and controlled [39]. This sequence of actions is at the very basis of Six Sigma, the DMAIC improvement cycle or problem-solving strategy. Underlined in this problem-solving strategy is the simple equation [35, 40]:

$$Y(CTx) = f(X - influencers) \tag{2.2}$$

Measurable parameters have to be defined on the process output that can quantitatively describe the problem - these are the critical-to-x (CTx) variables of the project. Here the term ”x” means any area that has an impact on the customer. Some examples of areas that are often subject to problem-solving projects are: quality (CTQ); cost (CTC); delivery (CTD); safety (CTS) [39]. These response parameters are determined by a set of variables, the X-influencers. If these influencing variables (X-influencers) are controlled, then the process outputs parameters are controlled as well.

2.4.1 Define Phase

A problem is measured on the output and can be defined as ”an undesirable situation which may be solvable by some agent although probably with some difficulty” [41]. The first step of the DMAIC problem-solving cycle fundamentally aims to answer the two following questions:

1. What is the problem?
2. How big is the problem?

Together with the strategic goals, these two questions form the project statement that should help the project team focus on the core issues and establish a common starting point [39].

The mathematical translation of the problem statement is to be made during this initial stage of the process: a CTx variable (or more than one) has to be identified on the process output that significantly describes the problem. This will be the response or ”Y” variable throughout the whole project. A basic and very high level of the process flow can also be drawn on this initial phase of the process improvement project.

2.4.2 Measure Phase

The second phase of the DMAIC cycle deals mainly with in-depth process mapping and data collection.

Process mapping can start at a high level, with an IPO (input-process-output) diagram. This diagram lists the principal unit operations of the process as well as their respective inputs and outputs. It is

a suitable tool to brainstorm all the variables (the X-influencers) that can negatively or positively have an impact on the response variables (Ctx). Applied to chemical processes, the Process Flow Diagram (PFD) can be drawn on this phase of the project.

The data to be analysed in the subsequent phase is collected in the current stage. Batch processes originate data that can be arranged as a 3-way data table as illustrated in figure 2.3.

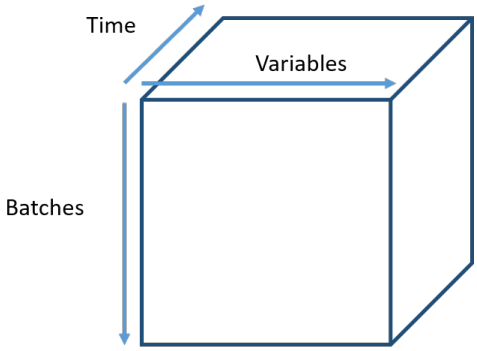


Figure 2.3: Three-way batch process data table.

For chemical synthesis processes in the pharmaceutical industry, *i.e.*, batch processes, two categories of data can be defined: process photograph and process film. The process photograph data is represented by the front face of the cube (figure 2.3) where to each completed batch a single value of a certain variable is attributed. These variables are for example: duration of operations; flow rates and agitator and jacket set-points. Process photograph data does not give a complete picture of the batch since it only displays a shot of the process and therefore cannot be considered for robust improvement actions derived from the analysis.

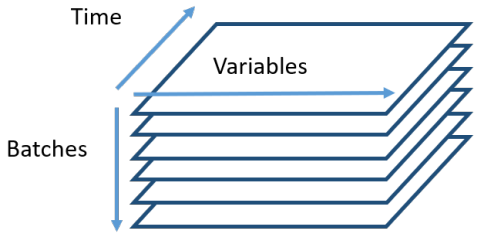


Figure 2.4: Three-way batch process data table showcasing the third dimension that is given with time. For process film data, each completed batch is represented by a horizontal slice on the cube (figure 2.3).

The process film data gives a more complete overview of the process. It consists, for every batch, of the data time points of variables such as temperature, pressure, speed of the agitator, and pH among many more. It is important to mention that the statistical tools used to analyse process photograph data are the same as the ones used to analyse this type of data set, the only difference being the arrangement of the data, as will be explained further.

Under this phase of the project, a Measurement System Analysis (MSA) can also be performed in order to ensure that the measuring instruments are adequate and measure the process parameters truthfully, *i.e.*, with accuracy and precision [32, 40]. Accuracy (as opposed to bias) is measured by the

difference between the average of the readings and the true value of the measurement and precision (as opposed to variability) is measured by the standard deviation of the readings. Two other measuring properties can be drawn from the ones previously stated: linearity which relates to the change of bias over a range of values and stability that is linked to how the precision remains constant with time [32]. These are the parameters to be considered when performing an MSA to a measuring instrument, usually before the measures are taken.

2.4.3 Analyse Phase

During the Analyse phase, the data collected during Measure is statistically analysed. It is during this phase that process understanding is consolidated: the X-influencers that mostly impact the identified Y parameters are signaled and through correlation analysis it is understood how does the variation in X influences the behavior of Y.

Variability on the output of a chemical process (throughput and quality) can derive from the variation on the conditions in which the process is running, variation on the input quality data, or operational variability inherent to processes not fully automatized. To fully understand a process is to identify all the critical sources of variability that can have an undesired impact on the attributes of the output [10]. The relationships between the input material data and the process variables with the final product quality data or the yield of the process are to be understood during this phase of the DMAIC cycle.

2.4.3.1 Multivariate Data Analysis

In the beginning of the past century, process engineers were lucky enough to get a few measures of their processes as well as input and output data. Nowadays, the paradigm has changed. It is estimated that the large pharmaceutical company has around 6 PB (6 thousand TB or 6 million GB) of data stored [6]. Accompanied by this explosion of data being collected during manufacturing, comes the notion that pharmaceutical processes are complex and problems that can arise during either development or manufacturing are often explained by a wide set of variables. In order to fully understand these systems, a new set of statistical analysis tools had to be brought up [10].

Multivariate data analysis (MVDA) tools derive from Chemometrics. Chemometrics is a field of Chemistry ”that uses statistical and mathematical models to design or select optimal measurement procedures and experiments, and provide maximum chemical information of the studied process with the analysis of collected data” [42]. Furthermore, MVDA as an integrated tool of Chemometrics is recognized in ICH Q11 [43] guideline as a crucial mechanism for process development and optimization [9]. The discipline has its focus on the following areas [10]:

- exploratory analysis;
- pattern recognition;
- classification and/or discrimination analysis;
- multivariate calibration;

- process modelling;
- monitoring and control.

Of the following applications of Chemometrics and MVDA, only exploratory analysis, pattern recognition, and process modeling were employed in the present process improvement project.

Software Used

Before entering into a detailed explanation of the algorithms, data arrangement, and statistical models used during the Analyse phase, a brief section is dedicated to the software used.

In a first instance, Minitab was utilized. Minitab is a statistical program that focuses, in a broad sense, on data analysis and statistical process improvement being a very common program to be used in Six Sigma and Lean initiatives. However, the need for a more incisive and applied, to the pharmaceutical industry and to MVDA, software was manifested early on during the project and so, SIMCA started to be used for the more advanced MVDA algorithms and data arrangements. SIMCA is a bio-pharmaceutical MVDA software with the main goal of creating value out of big data and focused on process monitoring and process optimization always with the premise of increased process understanding.

For univariate characterization and some early data displaying Minitab was used, but for the most part of multivariate modeling SIMCA was the chosen software.

Preprocessing Methods

Raw data usually comes with noise, missing values, and unwanted variation. In this way, data preprocessing, as the first step in data analysis, is necessary in order to bring forth the true and chemically relevant underlying information [44, 45]. However, if not performed accordingly, data preprocessing can induce unwanted variation and so, proper preprocessing is crucial to a good analysis outcome [45].

Several types of data preprocessing are available in the software used tor MVDA.

Table 2.2: Types of preprocessing methods considered for the project [46]. $\tilde{x}_{ij}$ represents the data point after preprocessing; x_{ij} represents the data point before preprocessing; $\overline{x}_i$ represents the average value of the variable being considered and s_i represents the standard deviation of the variable being considered.

Method	Formula
Centering	$\tilde{x}_{ij} = x_{ij} - \overline{x}_i$
Unit-Variance (UV) Scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_i}{s_i}$
Pareto Scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_i}{\sqrt{s_i}}$

Centering is usually applied together with other preprocessing methods [46]. However, it can be useful when all variables to be analysed are of the same kind (and so have a similar scale) [47].

Unit-variance scaling (UV scaling) is a very common method [45] which, after centering, divides the data value by the variable´s standard deviation. This method is particularly useful when the variables

are not of the same kind and therefore are not comparable [47]. This method results in the expansion of small values, such as noise values, that may lead to an increase in the influence of measurement errors that are usually relatively large for small values [45].

Another scaling method, Pareto scaling, consists in dividing the centered value by the square root of the variable standard deviation. It is placed in between the two extremes of UV scaling and no scaling at all [47].

For the analysis performed during the project, UV scaling was always chosen as the preprocessing method. The models will consider different types of variables, with different scales (pressure, temperature, operations duration, agitator speed), and so, UV scaling was deemed as the best choice.

Principal Components Analysis

Principal Components Analysis (PCA) is a mathematical procedure that transforms a large set of variables, that may be correlated or not, into a lower-dimensional set of new variables designated as principal components [10]. These principal components are extracted from the data set according to the directions of greatest variability: the first principal component (PC1) will explain the greatest amount of variability, the second principal component (PC2, which is orthogonal and therefore uncorrelated to PC1) will follow the first and so forth.

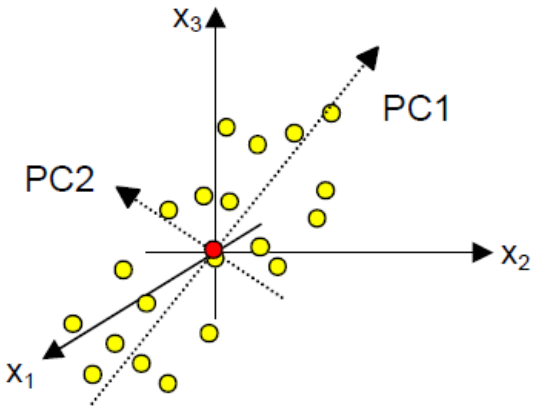


Figure 2.5: Extraction of two principal components (PC1 and PC2) from an arbitrary data set with three variables (x_1, x_2, x_3) showcasing the directions of most variability in the data set [47].

There are several criteria to choose the ideal number of principal components for a PCA analysis. Two undesired situations can occur: under-fitting and over-fitting.

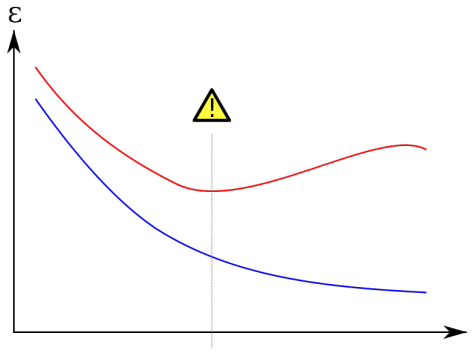


Figure 2.6: Model error (in blue) and validation error (in red) with the increase in model complexity represented in the x-axis [48].

Looking closely at figure 2.6, it is notorious that the validation error is always bigger than the model error. There is a certain point in model complexity where the validation error starts to increase with the increase in model complexity and from this point on there is a situation of overfitting: the model performs well on the training set (the data set in which the model was built) but performs poorly with the test set (validation data set) [48]. This leads to inaccurate predictions by including too many components that describe random error and noise [10, 48]. Before the inflection point in the validation error curve, the situation is called under-fitting.

For PCA models, one of the most common rules is to include all the PCs that have an eigenvalue higher than 1 [49]. The characteristic eigenvalue of a PC is a quantitative measure of how good does the component summarizes the data and is used in most cases as the rule to select the ideal number of PCs.

This type of analysis fits in the first group of Chemometrics applications: exploratory analysis. It is mostly used as an initial approach to large and complex data sets providing an initial overview of the data (*e.g.* identify clusters, outliers, and deviating patterns) [10].

The PCA analysis yields two important outputs, the scores, and loadings of the data set. The scores give information about the similarity between observations and are usually represented in a bi-plot of two PCs (for example PC2 Vs. PC1) [50]. The loadings are usually also represented on a bi-plot of two PCs but, contrary to the scores, give information about the relationship between the original variables [50]. The scores and loadings bi-plots can be used together to figure out, for example, which original variables are contributing more for a specific pattern on the observations.

In figure 2.7, the scores plot and the loadings plot for a training exercise, adapted from [47] are shown. The dataset contains the consumption indexes of certain food products (the variables) among the European countries (the observation or samples).

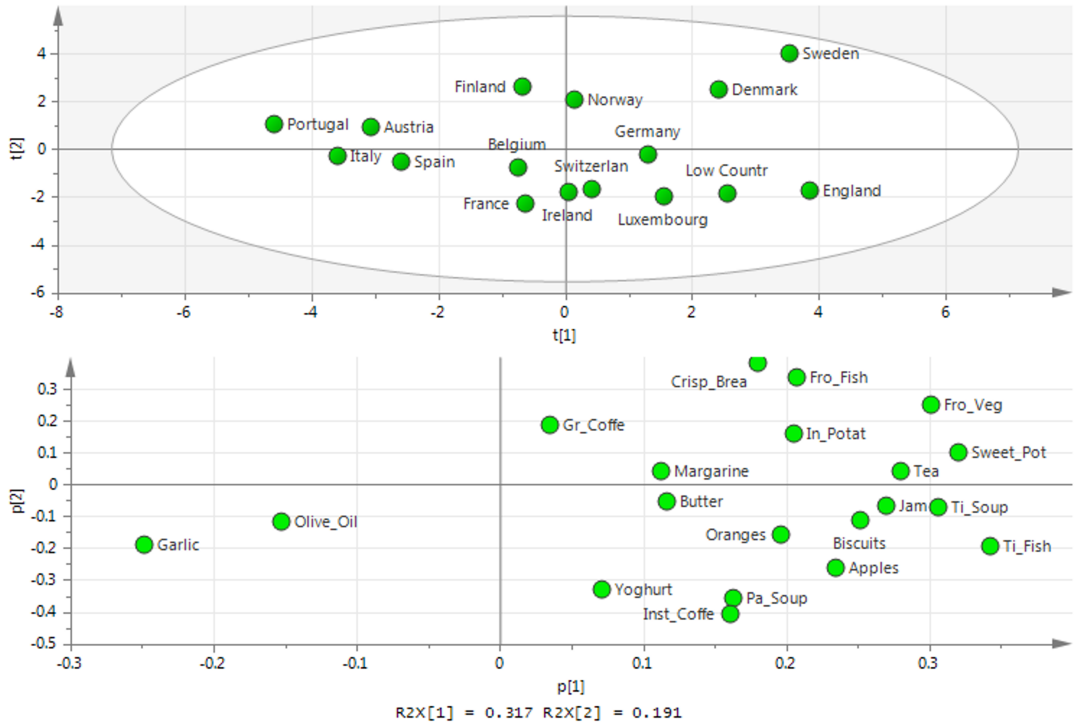


Figure 2.7: Scores plot (above) and loadings plot (below) for PC1 and PC2 for the consumption of some provisions on European countries. Adapted from [47].

A PCA model was adjusted with 3 PC´s: the $R2$ for each component, *i.e.*, the variance explained by each component, was for the first 31.7%, for the second 19.2%, and for the third, 13.8%, yielding a cumulative variance explained of 64.8%.

Looking at the scores plot, it is visible that all the observations fall inside Hotelling´s T2 ellipse, which is equivalent to the 95% confidence interval for univariate analysis. This statistical test provides the distance from the observation to the center of the model and can be used to detect deviating observations: observations that fall outside this ellipse can be considered outliers and therefore can have a large influence on the model obtained [10]. Three groups of countries can be seen: the southern European countries in the left-hand region of the plot, the Scandinavian countries on the top-right region, and the central European countries in the lower-center region of the plot indicating that each country in the group has similar consumption patterns than the others in that group (southern, central and Scandinavian countries) and that the groups have different consumption patterns among themselves. Turning over to the loadings plot, the relationship between the variables can be drawn. Garlic and olive oil are shown very far apart from all the other variables indicating a discriminating group of variables (variables that discriminate the observations). Analysing both graphs at the same time, some conclusions can be drawn: firstly, that olive oil and garlic are mostly consumed in the Mediterranean countries and that these are the variables that differentiate these countries from the rest. Crispbread and frozen fish differentiate the Scandinavian countries while the central European countries can be differentiated with high consumption of instant coffee and powder soup (Pa_Soup).

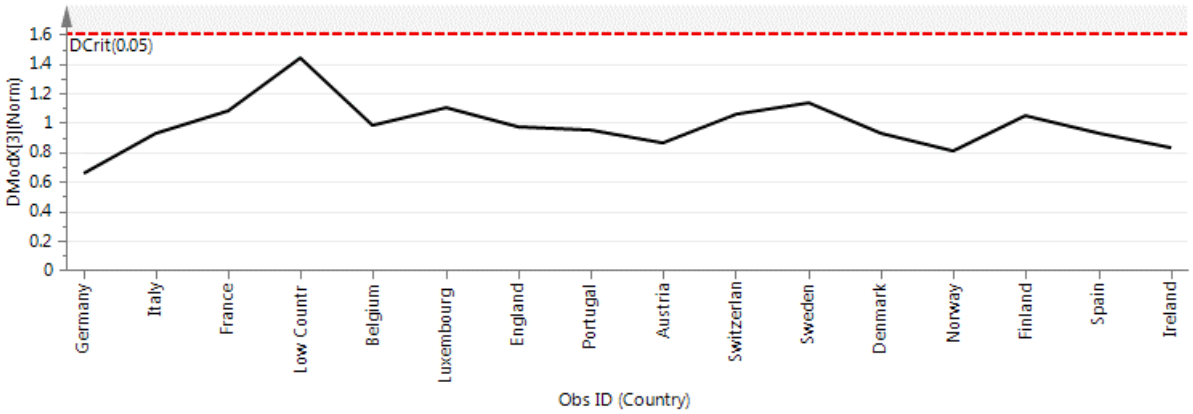


Figure 2.8: DModX parameter or residuals to the observations considered. Adapted from [47].

Another important parameter that is depicted in figure 2.8 is the residuals (or DModX parameter). The residuals represent the unexplained variation in the model [10, 47]. A critical value for the residuals is also plotted and observations that fall outside this value can be coined as outliers [47]. For the example being considered, no observation had residuals above the critical value.

Partial Least Squares

After obtaining an overview of the data set with PCA (exploratory analysis), the need for a more incisive and applied algorithm arises: to model the relationship between a given set of independent variables X and one or more dependent variables or responses Y.

Simply put, Partial Least Squares (PLS) is an extension of PCA but, instead of looking at the directions of greatest variability on the whole data set, it extracts the components in the X data set that explain the most variability in the Y data set [9, 42, 51].

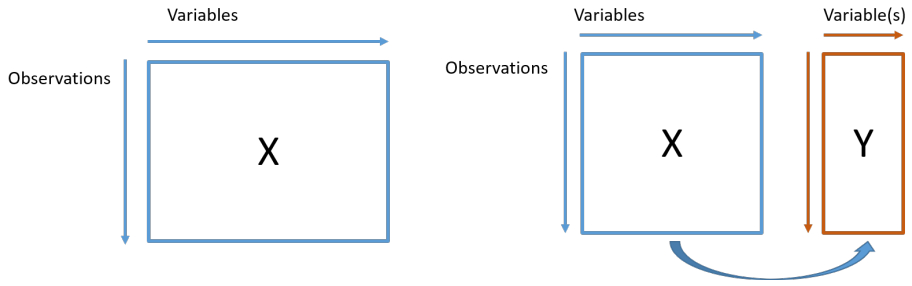


Figure 2.9: Scheme of the data set configuration in PCA (to the left) and PLS (to the right).

The new latent variables extracted from the data set have the same mathematical proprieties as the PCs in PCA modeling, hence, the output of a PLS analysis, although different in its content due to the differences in the algorithm, are also the scores and loadings. One additional useful output of a PLS analysis is the scaled regression coefficients, which give the relative importance of the original variables to the response variable in question [10]. Very negative coefficients indicate that a variable has a more negative effect on the response variable than less negative coefficients and the same for positive coefficients.

In order to evaluate the model's predictive ability, two important indicators can be used. The first one is $Q2$ which indicates how well the model predicts the responses for a new data set [10]. This value is always lower than $R2$. Another way of evaluating a model´s predictive capacity is to plot the actual values for the response variable in the Y-axis (observed) versus the values predicted by the model on the X-axis. A line with a slope of 1 and a y-intercept of 0 is the target for a good model [52] where the predicted values are the same as the observed one.

The ideal number of latent variables is chosen based on the best predictive ability of the model, the highest $Q2$ value which is computed through cross-validation [47, 53]: some rows of the data set are kept out of the model and predicted by the model then to be compered with the actual values. This is done until all rows have been excluded from the model [47].

Batch Modelling

As stated before, batch processes give rise to a 3D data table where the variation of the variables over time is given for each completed batch (figure 2.4). The same methodology can be applied: firstly, exploratory analysis is conducted in order to detect deviating tendencies or outliers and to get an initial overview of the data (this type of modeling is called Batch Evolution Model, BEM); secondly, the evolution of the variables, the X data table, is modeled against the response variables, the Y data table (this type of modeling is called Batch Level Model, BLM).

Under exploratory analysis, a regular batch control chart can be done to one variable at a time, displaying the evolution of the variable in the analysed batches. The problem arises when a lot of variables are to be seen simultaneously in order to establish a good batch trajectory. The arrangement of the 3D process data table is done as showed in figure 2.10 and is called observation-wise unfolding [6]: batch (B1, B2, Bn) data is vertically stacked.

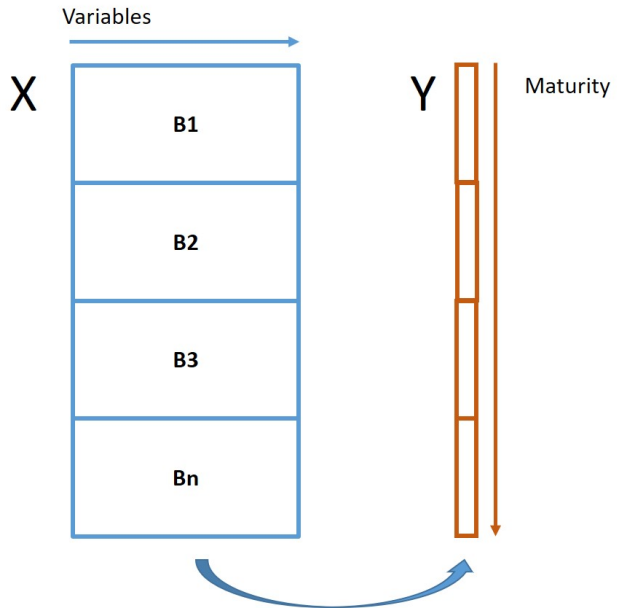


Figure 2.10: Scheme of data layout for a Batch Evolution Model (BEM). Maturity is used to give the model a direction and is a variable that is very descriptive of batch evolution (it is usually the time at which the samples are drawn). This deconstruction method of the 3D process data table is called observation-wise unfolding.

This condensation of all the information is done by plotting a batch trajectory through the scores of a PLS model - the Batch Evolution Model (BEM). As stated earlier in this section the aim of this model is to detect anomalies in the process trajectory and investigate which are the variables responsible for those deviations on the overall process trajectory. An initial overview of the way the batches are being run is also given. Below is an example of a well-controlled process, displaying a similar scores evolution for all analysed batches.

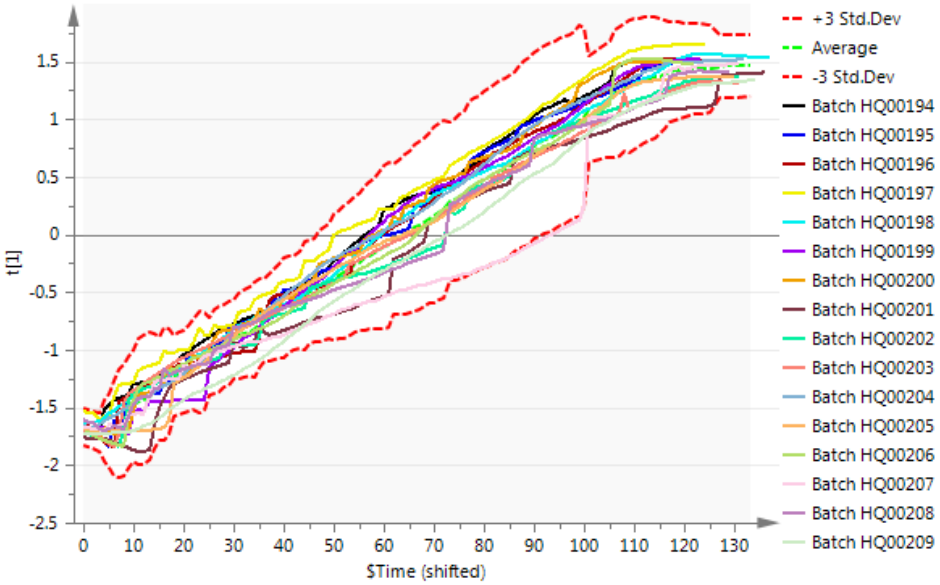


Figure 2.11: Scores control chart for a crystallization step during the FP process showcasing that all analysed batches have a similar batch trajectory to each other. Control charts will be discussed in more detail in section 2.4.5.

The challenge arises when a relationship has to be drawn between the 3D process data table and the 2D data table that is representative of the final parameter measured upon batch completion. This challenge is demonstrated on figure 2.12.

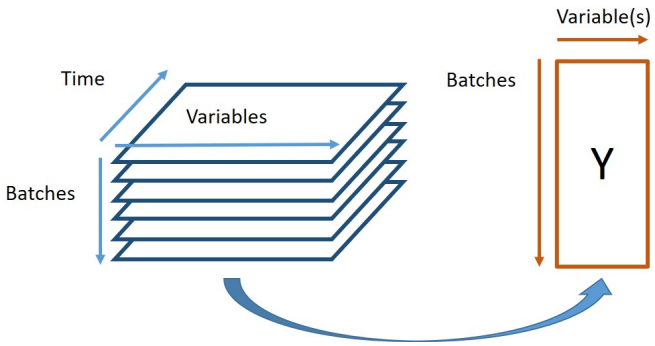


Figure 2.12: 3D process data table and 2D final conditions data table.

As was the case for PCA and PLS, after obtaining an overall look at how the batches are running by comparing their batch trajectories, a more precise and incisive approach is needed. For successful modeling, the deconstruction of the 3D process data table to a 2D matrix has to be made differently than it was done for BEM. As shown in figure 2.13, the variable profiles for each batch are horizontally stacked on the same line which corresponds to a single batch. This rearrangement is called batch-wise unfolding [6].

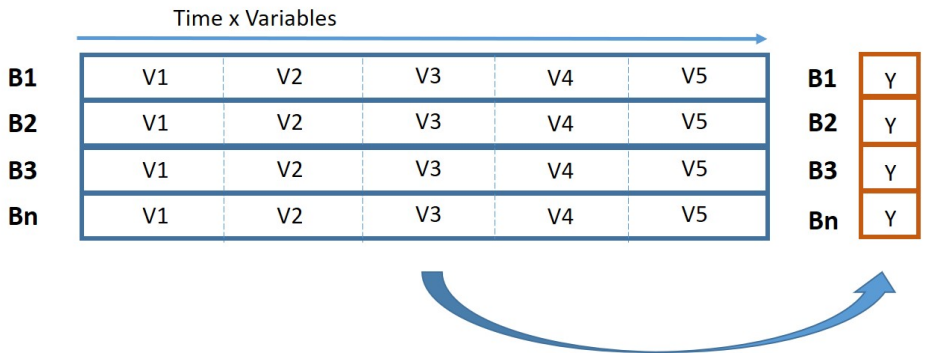


Figure 2.13: Process data table deconstruction for BLM termed batch-wise unfolding.

A PLS model is created between the unfolded process data table and the final conditions data table providing an effective and robust method to tune any sections of the process in order to optimize the Y parameter(s) by carefully understanding which are the variables and time sections that have the greatest impact on the final parameter. This study is carried out by an analysis of the loadings of such models [6].

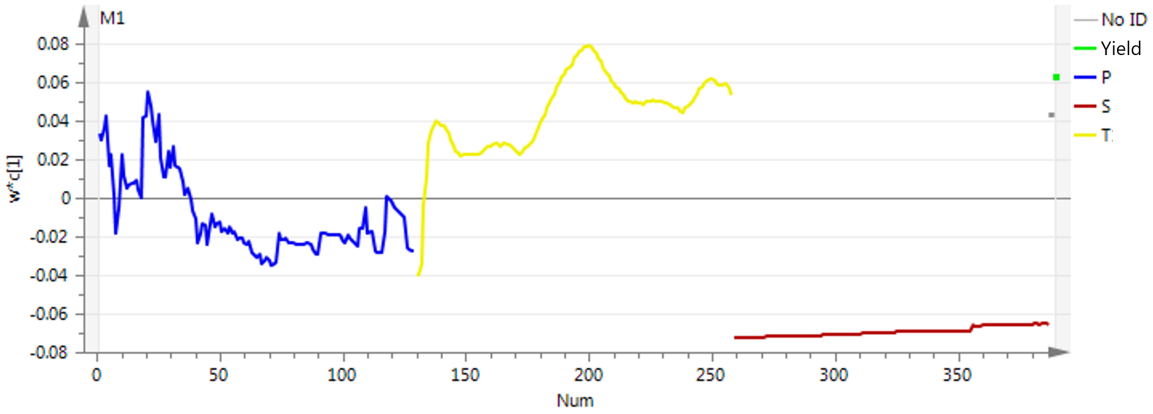


Figure 2.14: Loadings of the first component given against batch maturity, colored by variable (pressure as blue, temperature as yellow, and agitation as red), for a crystallization step during the FP process.

In figure 2.14 the loadings of the first component for a crystallization step are presented against batch maturity for each variable. For example, the blue line represents, throughout the entire operation being analysed, the contribution of that variable to the response variable (displayed in green on the left part of the plot), in a similar way to the PLS model coefficients. The loadings are colored by variable so it is easier to differentiate the contributions of the analysed parameters. For temperature (yellow variable), although on some sections more than others, higher temperatures give higher yields, because the loadings for this variable are always positive and of the same order of magnitude as the loading of the Y variable. These graphs present themselves as a strong and robust basis for concrete improvement actions with a focus on the way the process is being run.

As previously stated, variability in a process output can be explained by the input quality attributes or in the way the process is running. In the case that the quality attributes of the raw materials explain most part of the variability on the process output variable, it is still necessary to analyse the step that leads to the formation of the raw materials in order to either minimize or maximize those critical quality attributes that have an impact on the final Y. In this way, the study of the loadings through BLM is of extreme importance and is what can give a more solid foundation for a successful resolution of the problem to be solved.

2.4.4 Improve Phase

The goal of the fourth phase of the DMAIC cycle can be divided into three consecutive parts.

The first one is to successfully materialize the results of the statistical analysis conducted on the previous phase into tangible, concrete, and feasible actions for problem resolution and process improvement.

Secondly, is idea prioritization. Usually, the effort to implement the actions that come out of the Analyse phase surpasses the time or resources available to implement them and so, prioritization is imperative [54]. This action prioritization can be done by placing them on an Impact Vs. Effort matrix as shown in figure 2.15.

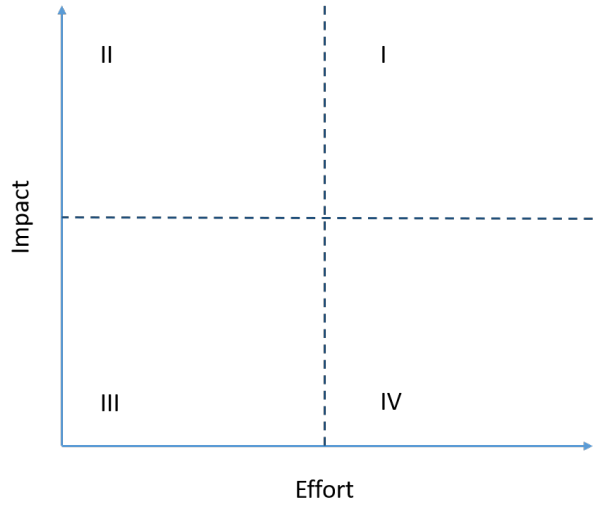


Figure 2.15: Impact Vs. Effort matrix template displaying the 4 categories of actions.

The actions are placed on the quadrants according to their impact and effort classification. The impact categorization can be performed using a set of statistical indicators that reflect the strength of the model: $R2$ and $Q2$ as already mentioned. The effort classification is done empirically with the aid of the process experts.

Actions placed on the first quadrant of the matrix have a high impact but also require high effort. Only a small number of actions should be placed in this category. The second quadrant covers actions that have high impact and require a low effort for completion and are coined ”quick wins”. These are the actions to focus on. Actions with low impact and low effort are placed on the third quadrant. These actions are only to be considered when there is a surplus of time and resources. Finally, on the fourth quadrant, the actions with high effort and low impact are placed and are usually actions that are not worth completing.

Still under the second objective of the Improve phase, an action plan with feasible timelines and accountable people for each task is drawn and is to be used in the third objective of the Improve phase: action plan implementation. Only on this stage is the process actually improved and the *status quo* is changed into a revamped version of the process.

2.4.5 Control Phase

Dr. Walter Shewhart, working in the early 20´s, was the pioneer of a set of tools that aimed at statistical quality control on mass production [32]. According to him, variability on a process variable or in an output attribute is due to the following two causes [32, 39]:

- common or chance causes: these are inherent to the processes and usually are not controlled though can be classified as unavoidable;
- special or assignable causes: these arise from a specific occurrence that the operator when adequately alerted can remove or minimize the impact.

The control charts (figure 2.16) arose in order to differentiate between these two causes of variation. Two undesirable situations can occur when this differentiation is not performed: when trying to eliminate common cause variation the result ends up being more variation instead of less and if special cause variation is not spotted and the effort is not put to eliminate it the outcome is usually an exponential growth of variation [39].

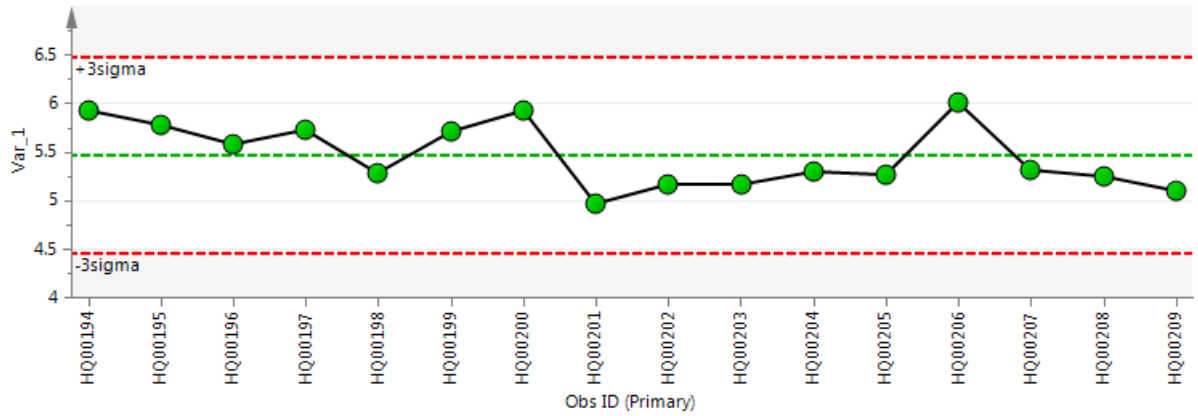


Figure 2.16: Control chart for a random variable over some production batches. The 3σ limits are outlined as red and the mean for the process variable as green.

Dr. Shewhart established the differentiating criterion as plus and minus 3σ from the variable mean [32, 39, 55]. If an observation falls outside these control limits (upper control limit, UCL and lower control limit, LCL), the process can be coined as out of control and so, probably, one or more special causes have occurred and actions must be taken in order to discover and eliminate them [32, 55].

The control chart presented in figure 2.16 addresses process photograph type of data, where for each complete batch the variable takes just one value. Control charts for process film type of data can also be established like the one presented in figure 2.11 (although not regarding the actual values for a process variable but rather the scores for a PLS model).

Maintaining the improvements achieved is also a part of this final phase of the DMAIC cycle. A rigorous process of documentation of the lessons learned during the entire project should be performed and a clear identification of how the improvements can be replicated and applied to other processes [39]. An often-used practical way for the improvements sustain is the development of training materials in order to ensure continued support for the people involved with the process on a daily basis.

2.5 The Pharmaceutical Industry and Six Sigma

As previously stated, Six Sigma has proven to succeed in many industries with countless tangible and measurable benefits [35]. However, the pharmaceutical industry has been reluctant or at least slower to apply Six Sigma tools to their business processes [13, 14, 20].

Drug recalls have shown alarming trends over the last few years [20], as seen on figure 2.17, reminding, the pharmaceutical industry of the long way to achieve quality excellence, which is critical to attain since the customers of the industry are patients that rely on the industry to alleviate their condition.

The 6σ capability can be considered the future of drug quality [20] that is translated into 3.4 defectives per million opportunities [20, 40]. An enormous improvement will have to take place since currently the pharmaceutical industry´s sigma capability is around 3σ [20].

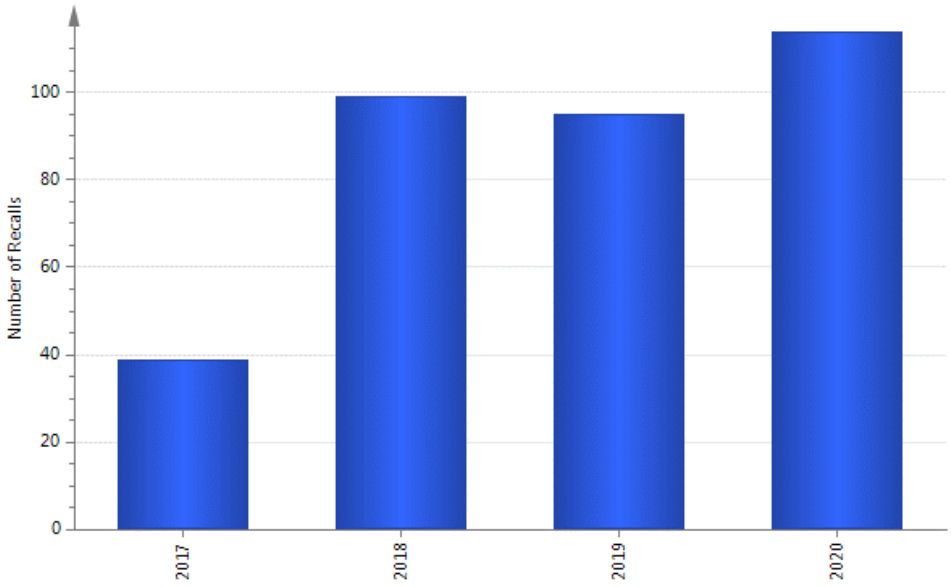


Figure 2.17: Number of FDA drug product recalls on the last 3 completed years [56]. For 2017, data was unavailable for the first 8 months of the year. A direct proportionality ratio was conducted as an approximation to the overall drug recalls on that year.

A fundamental mindset to achieve quality excellence (the Six Sigma target) is continuous improvement that can be carried off through many corporate programs. Presented below is an example of application and the measured improvements on a top pharmaceutical company, showcasing that, although there is still a long path ahead, the effort is being put in place and the incremented improvements are being achieved.

Taking part in this strive to achieve quality excellence, the discipline of Chemometrics has been established as a tool to be used in every step of product development and manufacturing by enabling process understanding thus increasing process robustness that eventually leads to an increase in quality [10]. Some examples are also presented of how top pharmaceutical companies progressively adopting MVDA as an integrated tool on their process improvement projects.

2.5.1 Pfizer´s *Right First Time*

Pfizer is one of the top research-based pharmaceutical companies worldwide. Pfizer Global Supply (from now referred to as PGS), a Pfizer subsidiary, ensures manufacturing operations and supply network, making sure that products are produced to the highest standards of quality, safety, and efficacy, and are available when and where needed [19].

As an industry leader, PGS has been at the vanguard of a decade-long, industry-wide movement to drive performance by fostering a continuous improvement culture throughout the organization. It all started in 2003 with the *Right First Time* program that aimed to ensure high-quality products while

at the same time reducing costs [19]. Their primary target, in order to achieve the initial goal of the program, was to aim at systematically uncovering root causes for unwanted and uncontrolled variations in manufacturing processes [19]. As for the side-effects of this main target, the following can be listed: improved effectiveness and efficacy with the elimination of non-value-added activities, gained process understanding, and increased process robustness. Processes that were consistently recording a 2 to 3 σ capability began to perform at a 4 to 5 σ capability [19] with the result being a more reliable process in terms of yield, output quality, and process speed. The increased predictability of their manufacturing processes also enabled an inventory level decrease. By 2008, Lean initiatives were incorporated into PGS *Right the First Time* strategy providing enhanced opportunities for lead time optimization, increased efficiency, and reduced inventory.

Pfizer soon understood that sustaining the initial momentum gained with successful initiatives was critical [19]. The company leadership noted that the early phases of *Right First Time* and Lean initiatives had to be integrated into a full system and top to bottom approach in order to keep the good results in a rapidly evolving pharma environment. This full system approach would eventually lead to [19]:

- Company-wide focus on value-adding activities instead of single point process improvements;
- Coordinated approach to cost and capabilities instead of increased process robustness;
- Organization transformation and full cost reduction instead of improved product quality and increased manufacturing productivity.

In order to support the transformation process, a common set of principles and metrics were needed to align change at every level of the organization. The Network Performance Principles (NPP´s) emerged as a common basis to orient the efforts to make Pfizer best in class [19]. This set of principles describes the vision of a best in class company (in an ideal state) by answering the following set of questions [19]:

- How operations and supply chains should operate in an ideal state;
- How elements within and across PGS will operate together;
- How balanced metrics drive high performance;
- How highly capable colleagues deliver operational performance.

To measure, qualitatively, the performance across Pfizer sites relative to this transformation process, the Network Performance Assessment (NPA) was created. The NPA supported the assessment on where the site is in the transformation process, which elements are progressing and which are not, constituting these the areas that should be prioritized [19].

Transformation is a conscious and reliable transition to an upper state of business performance. The business processes employed by Pfizer over the past two decades and its results clearly highlight that, nowadays, a continuous improvement culture is what can foster the competitive advantage among top pharmaceutical companies.

2.5.2 MVDA Applications

Multivariate data analysis has turned into a reliable set of tools to ensure monitoring and process optimization during manufacturing. Below are a few cases of the successful use of the toolkit and its gains among some pharmaceutical companies.

In tablet manufacturing, variability in output quality is mostly explained by the variability of the inputted raw materials [57]. In order to optimize the tableting process, a group of scientists in Pfizer built PLS models that combined raw materials attributes and process variables with output quality features like hardness, disintegration time, dissolution [10, 57]. Before a new batch, with specific raw material with certain attributes, the models are used to find optimal process conditions in order to optimize the desired final product features. This reduces the need for experimentation when a raw material with different attributes is used in tablet manufacturing [10, 57].

In Novartis, multivariate statistical process control (MSPC) was used for process monitoring [58]. A Batch Evolution Model (BEM) was built and new batches were compared in order to evaluate whether the process trajectory falls into the expected ranges [10, 58]. Over the course of one year, the batches produced did not fell out of the model ranges indicating the analysed process was in statistical control [58].

A more holistic approach was taken by GSK to support the development of a continuous process for the manufacture of Paracetamol tablets [10, 59]. The analysis occurred in two sequential steps. Firstly, exploratory data analysis (PCA) was used in order to identify the critical process parameters and raw material attributes. Then, regression analysis (PLS) was applied to relate final product quality with those critical process parameters and raw material attributes [59].

Chapter 3

Results

3.1 Define

The first phase of the problem-solving DMAIC cycle sets out the tone and the boundaries for the entire project. Two crucial and structural questions are meant to be answered: what is the problem and how big is it.

3.1.1 What is the problem?

According to table 1.1, the expected uncertainty on the yield of each step leading to any intermediary in the API production train is increasing moving from the starting raw material to the final product. The increase in the uncertainty of the predicted yield is accompanied by an actual increase in the standard deviation of the obtained yields. High variability on any process output parameter can be translated into a poorly controlled process and lack of robustness, where process robustness can be defined as the lack of sensitivity of the process outputs to fluctuations in the process inputs and process variables [9].

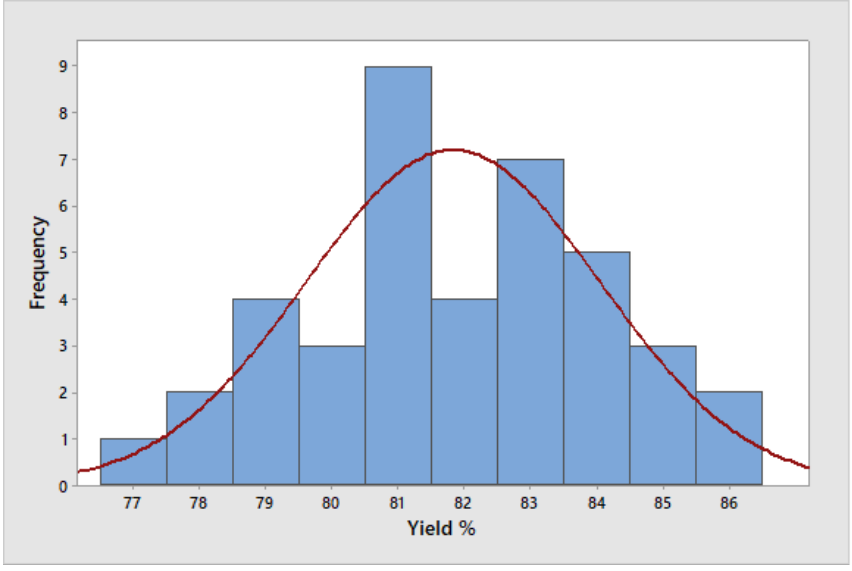


Figure 3.1: Histogram of the yield of FP process from July 2018 to January 2021 with normal distribution fitting.

As seen in figure 3.1, ranging from July 2018 to January 2021 (40 batches were included in the analysis), high variability on the final step of the API production process is verified leading to the conclusion that the process could be in tighter control. Univariate statistics for this variable are presented in table 3.1.

Table 3.1: Univariate statistics for the yield of FP production step.

μ	81.83
σ	2.22
CV (%)	2.71
Min	77.30
Max	86.30
R	9.00
$RelativeRange$ (%)	11.00
Q_1	80.20
Q_2	81.91
Q_3	83.48

The timeline trend of the variable being analysed can also be represented. It is important to mention that only batches with the previous intermediary (intermediary 4) as the starting raw material were included in the analysis, as there were some batches present in the time range (July of 2018 to January 2021) that were reprocessing batches. Reprocessing batches have more pure input materials and so, since the FP production step is a purifying step, higher yields were observed that were not considered.

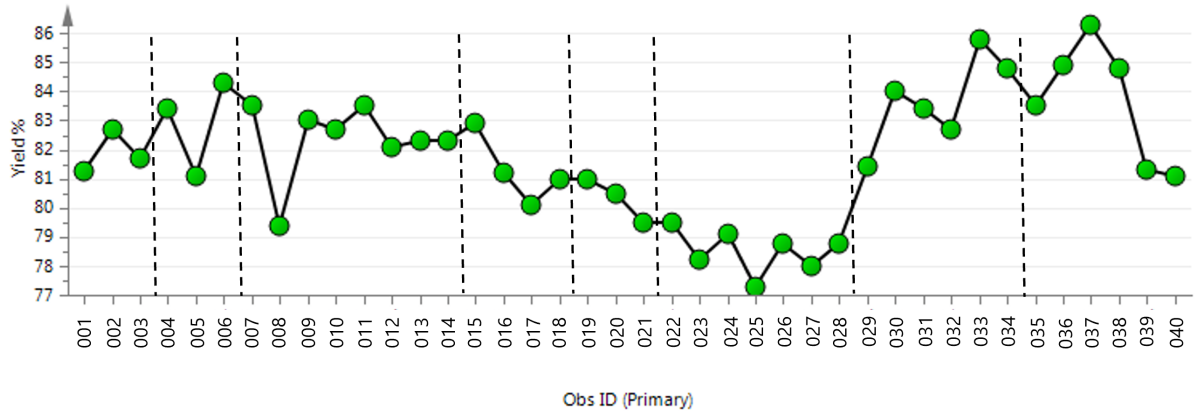


Figure 3.2: Run chart of the yield of FP production step from July 2018 to January 2021. The data points are divided into production campaigns.

Up until batch 14, although with some deviations, a constant trend was observed. Starting from batch 15 a negative tendency is evident up until batch 27 and from that batch to batch 37 the trend reverses. One could argue that the more recent trend is positive and therefore, no improvement project should be put in place. However, the variability is clearly present: the reasons for such low yields on batches 22 to 28 and for such high yields on batches 33 and 37 should be identified and understood in order to

prevent and replicate them respectively in a view to an optimized yield for FP process.

As previously stated, high variability on any output parameter can be translated into a poorly controlled process. Applying this key idea to high variability on yield, one can infer that it can be translated into poorly controlled throughput. A process in which the throughput is not predictable leads to an unknown number of batches needed to satisfy a client´s order that is inevitably accompanied by biased and uncertain production planning. High variability in yield also leads to ineffective use of the company´s resources since the cost of equipment, personal, raw materials, and utilities do not change according to the throughput. By standardizing and preferably optimizing the yield of the final step, the process advances to a state of tighter control, the throughput is increased and the company resources are used at a higher utilization rate. As side (but desirable) effects of the success of the improvement project, process understanding is gained and a culture of continuous improvement is fostered among the company.

3.1.2 How big is the problem?

With the first guiding question of the Define phase answered, the impact of the problem is to be computed through the answer of the second guiding question: ”How big is the problem?”. In order to correctly calculate the impact of the project, a holistic metric (preferably financial) has to be identified.

As stated in the previous section, high variability on yield causes biased and uncertain planning that ultimately leads to missed opportunities in terms of throughput that can then be converted to missed opportunities in terms of revenue considering an average price of the final product and that on average, 17 batches of FP are produced per year.

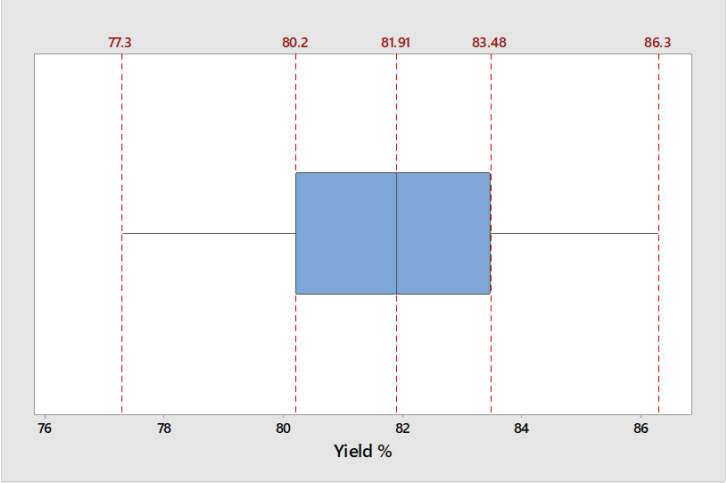


Figure 3.3: Box plot of the yield of FP process. The first, second, and third quartiles are showed as well as the minimum and maximum values.

The calculation for the possible impact of the succeeded project was performed considering several optimization scenarios: if all batches had a yield equal to Q_2; if all batches had a yield equal to Q_3 and if all batches had a yield equal to the maximum ever achieved (during the timeline considered).

$$MO = \frac{\sum_{n}^{\infty}(SP - Yield_n)}{n} \tag{3.1}$$

The calculation formula is expressed in equation 3.1, where SP designates the optimization set-point considered (Q_2, Q_3 or maximum ever achieved) and MO the calculated missed opportunities.

Table 3.2: Averaged missed opportunities (MO) in terms of throughput and revenue on a batch and yearly basis. The values are removed due to confidentiality reasons.

Optimization Set-Point	Avg. kg/batch	Avg. k$/batch	Avg. kg/year	Avg. M$/year
Q_2	–	–	–	–
Q_3	–	–	–	–
Max	–	–	–	–

With the maximum ever achieved optimization set-point, the missed opportunities in terms of throughput on a year basis represent, roughly, one FP production batch which is a tangible and considerable amount of product being lost to yield variability.

Instead of just calculating the averaged missed opportunities for batches running at Q_2, Q_3, and Max, the parameter was computed for a range of yield optimization set-points from Q_2 to Max. Below is the representation of the averaged missed opportunities in terms of revenues per year against the yield optimization set-point.

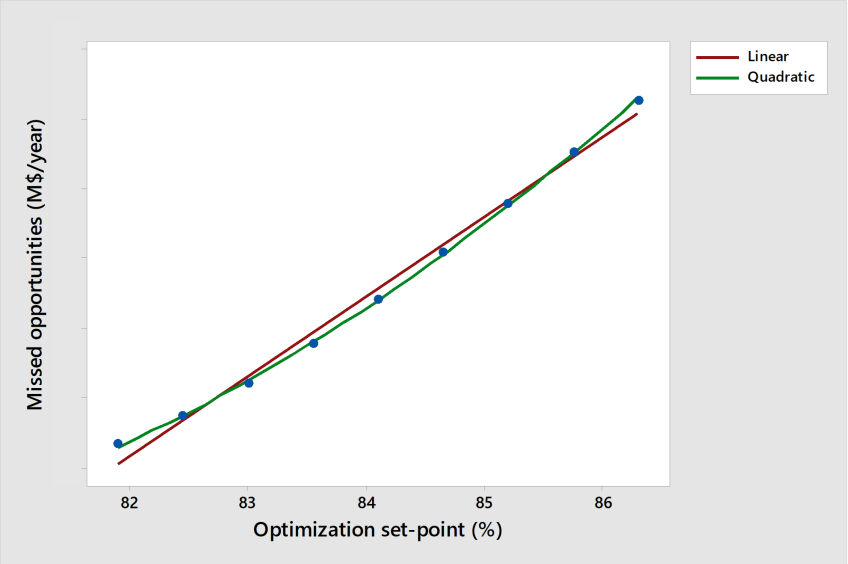


Figure 3.4: Graphical representation between the averaged missed opportunities in terms of revenue per year and the optimization set-point. A linear model ($R2$=99.15%) and a quadratic one ($R2$=99.97%) were fitted.

Although with a very small difference in the values of $R2$, the quadratic model displays a better fit to the data, showcasing that the missed opportunities in terms of revenue are exponentially dependent on the yield optimization. Towards the higher end of the optimization set-point, the possible gains will be higher than those of the lower end.

3.2 Measure

During the second phase of the process improvement cycle, a better understanding of the problem and of the process is done through process mapping. The IPO diagrams of each step of the production train (figure 1.3) were drawn. These diagrams can be found in annex B.

Regarding data collection, both process photograph and process film types of data were collected. The first category of data is usually present in databases filled by the process engineers and was kindly provided. For the collection of the second data category, the automation system records of the site had to be accessed. In the concerned production area the installations are not fully equipped with up-to-date sensors that actually record in an automation system, making it possible only to collect data from a small number of variables.

As stated before, a problem can be considered ”an undesirable situation which may be solvable by some agent although probably with some difficulty” [41]. Problems are caused by the inputs or the process variables and can be measured in a process output. As such, in order to control the desired variable(s) on the process output, one must manipulate the input and/or the controlled process variables. Total elimination of output variability will never be possible due to the contribution of uncontrolled factors, such as equipment malfunctioning or failure, uncontrolled human-related issues, and inherent common-cause variability [32, 39].

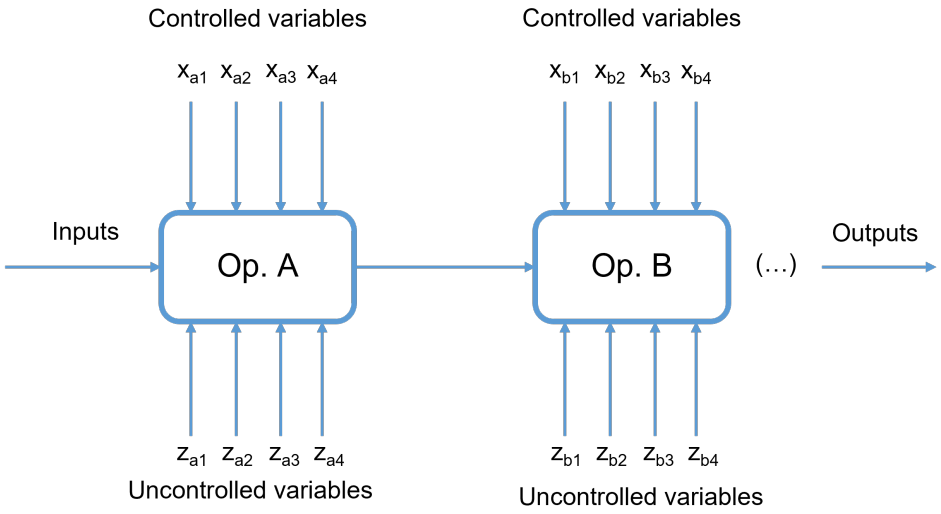


Figure 3.5: Scheme of a process (with just two operations, Op. A and Op. B) showcasing that the output is a function of the inputs and process variables (controlled and uncontrolled).

During the Measure phase of the project, process understanding is gained. Probable critical process variables and raw material attributes were listed and prioritized according to the discrete and empirical scale presented in table 3.3. This variable prioritization was made with the production team in order to increase process knowledge.

Table 3.3: Variable and input attribute discrete and empirical classification method employed.

Classification	Description
0	It is proved that there is no correlation with the response variable
1	Probably there is no correlation with the response variable but it is not proved so
3	Probably there is some correlation with the response variable
9	It is proved that there is a strong correlation with the response variable

The response variable (the Y variable) depends upon the production step being analysed. The yield is the final and ultimate parameter to be standardized and optimized but, according to figure 1.4, the raw material attributes can explain the problem. If that is the case, the production step leading to that material should be also analysed, in order to figure out what is the cause for the identified critical attribute(s). In this type of analysis, the response variable will no longer be yield but rather that critical attribute.

3.2.1 Process Description

The present project leaned over the analysis of FP and intermediary 4 production steps and so the response variables that will be present in results regarding the Analyse phase will be the yield of FP step and material attributes of intermediary 4. Only these two processes will now be subject to a detailed description since they are the ones being analysed.

3.2.1.1 Intermediary 4

Section content changed due to confidentiality reasons.

Intermediary 3 is dissolved and two inorganic salts are added. A final and gaseous reactant is added and reaction takes place. Multiple degasing steps take place in between the load of the several reactants. Precipitation happens due to antisolvent addition and cooling. The suspension is then filtered and dried.

3.2.1.2 FP

Section content changed due to confidentiality reasons.

Intermediary 4 is dissolved. Precipitation happens through solvent evaporation, antisolvent addition and cooling. The suspension is then filtered and dried.

3.3 Analyse

This phase of the DMAIC cycle covers all the statistical analysis of the collected data. As stated before, multivariate data analysis techniques were used as a more holistic, robust, and feasible approach to univariate statistics. From the areas of Chemometrics, only exploratory analysis (with pattern recognition included) firstly and process modeling secondly were employed in the project.

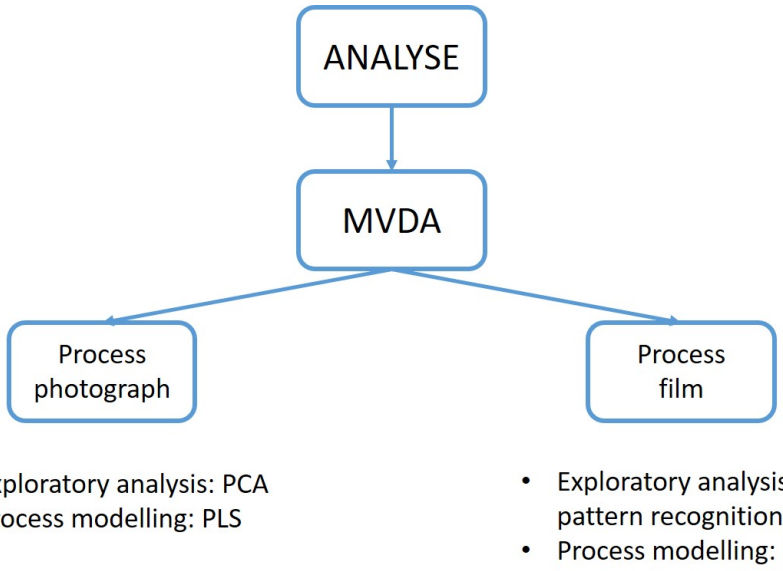


Figure 3.6: Scheme of the strategy followed during Analyse phase of the DMAIC cycle. Process photograph and process film types of data refer to the content explained in section 2.4.2.

The results will be presented following the backwards approach (figure 1.4): starting on the last production step and moving in reverse on the production train. For each production step, firstly process photograph type of analysis for process variables and quality data of the raw materials and then process film type of analysis to the operations that were coined as relevant in the process photograph analysis will be performed and also following the strategy presented in figure 3.6, *i.e.*, primarily exploratory analysis followed by process modeling.

All the models that will be presented during the following section are termed black-box models or statistical models. Within this framework, systems are only viewed in terms of their inputs (stimulus) and outputs (responses), without any knowledge of their internal workings, and are built based only on historical or experimental data [60].

Figure 3.7: Black-box model scheme.

Contrary to white-box models, or mechanistic models, which are entirely based on mathematically expressed universal natural laws, black-box models only need the specification of the system´s inputs and outputs and are particularly useful when the system is poorly understood [60] which is the case for the chemical-pharmaceutical industry, where systems are often too complex and the simplifications made to achieve a mechanistic model often compromise and surpass the advantages for this type of modeling.

3.3.1 FP process step analysis

Following the backwards approach methodology, illustrated in figure 1.4, the final step of the production train will be analysed, both in terms of process photograph and process film type of data. For the directed models (such as PLS and BLM) the response variable that will give the model its direction will be the yield of this final production step since it is the primary problem to solve.

3.3.1.1 Quality

The yield on the final step (the response variable considered) was modeled against the quality data of the starting raw material of the final step (intermediary 4). Two impurities of this intermediate were removed from the analysis due to the fact that their values, over the considered production batches were always much lower than the limit of quantification (LoQ). This value is characteristic of the measuring instrument and is the lowest value to which measurements of the designated substance can be reliably detected with some predefined minimums for bias and imprecision [61] and so, the values reported in the release report of intermediary 4 were not meaningful and were left out. It is also important to mention that, as seen in table 1.1, the size of intermediary 4 batches (SRM around 112 kg) is much higher than the size of FP batches (SRM of 35 kg), and so, one batch of intermediary 4 is used in more than one batch of FP. For batches of the latter that had more than one batch of intermediary 4 as SRM, in order to figure out the final value for the designated impurities for the overall SRM, a simple mass balance was performed. This technique was used for the analysis of quality data on all the production steps.

Firstly, PCA analysis was performed to the data set (exploratory analysis). Three principal components were chosen to explain the most relevant part of the variability of the data set: PC1 explains 44%, PC2 explains 23% and PC3 explains 19% of the variability which yields a cumulative percentage of 86%, the rest is considered noise. Through the scores of the model, no deviating batches, or outliers were detected neither were clusters. The loadings of the model are presented in the following figure.

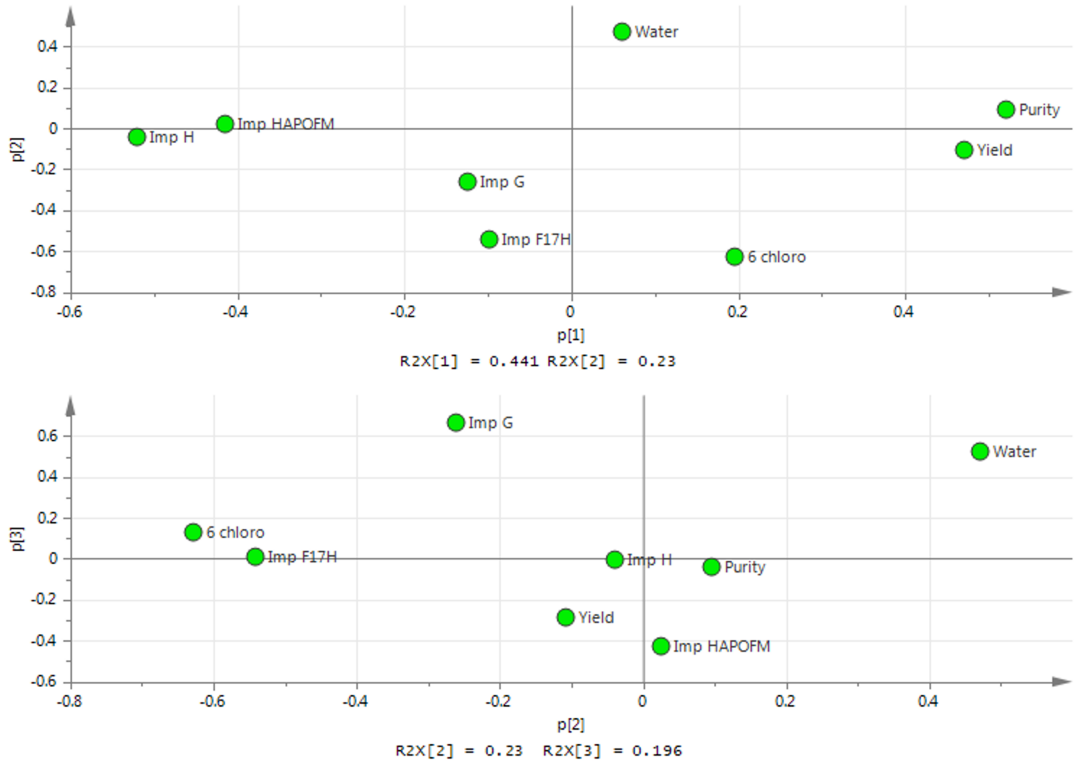


Figure 3.8: Scatter plot of the loadings for the PCA analysis of yield of FP production step and the quality data of intermediary 4 (impurities and purity). On the top, the second component (PC2) is plotted against the first component (PC1) and on the bottom, the third component (PC3) is plotted against the second component (PC2). The variability explained by each component is shown at the bottom of each graph.

Firstly, it is important to mention that, as seen in the plot of PC1 Vs. PC2, purity of intermediary 4 is, almost perfectly, negatively correlated with impurity H. Variables with such a high correlation can be termed inter-changeable variables, *i.e.*, if purity is swapped with impurity H or vice-versa there is no change in the problem or its solution. The variable to be excluded in the subsequent analysis was purity as this variable is a mere difference between the totality and all the impurities and does not represent a real substance being detected in the analytical methods.

Looking at both plots, the yield of FP step is strongly explained by PC1 and not well explained by PC2 and PC3 since the projection of this variable on these components is close to zero (although higher on PC3). Impurity H and impurity HAPOFM seem to be the two impurities that have a more negative impact on yield according to PC1. Looking at PC3, impurity G can also be signaled as having a negative impact on yield, and impurity HAPOFM with a positive impact. These contrary effects regarding impurity HAPOFM will be explained in regression analysis.

Fitting a PLS model to the data and excluding purity from the variable set, the following model was obtained.

Table 3.4: PLS model statistics for the yield of FP as the response variable and the quality data of intermediary 4 as independent variables.

	R2	Q2
Comp. 1	0.800	0.744
Comp. 2	0.063	0.192
Cumulative	0.863	0.793

Considerable high values for both $R2$ and $Q2$ are shown in table 3.4 showcasing that the variability on the yield of FP is almost completely explained by the variability on the quality data of the input material for that production step. However, this fact does not exempt an analysis of the process variables.

The coefficients of the model are presented in figure 3.24.

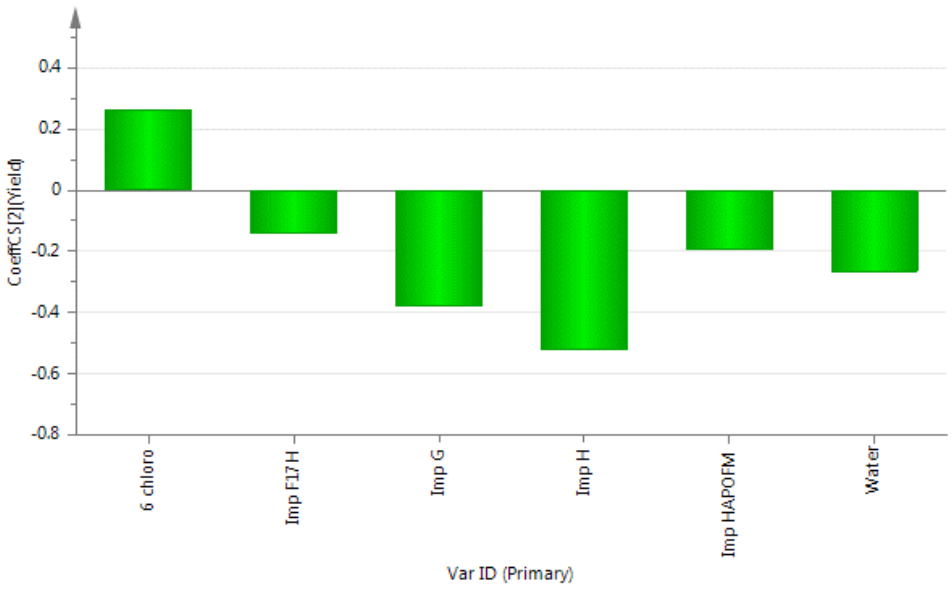


Figure 3.9: PLS model coefficients for the several impurities present in the intermediary 4 quality data against the yield of FP.

It is important to mention that in this type of analysis, the coefficients give the relative contribution of the variables and not the actual contribution or regression coefficients. As such, these figures presented are completely independent of the actual scale of values of each variable.

Some results are in the same line as the ones obtained with PCA analysis and some are not. This is expectable as the way the two algorithms work is different, as already explained. Impurity H appears to be the substance that has the most negative impact on the yield followed by impurity G and the water content. Impurity HAPOFM, which had a high relevance on the loadings of PCA analysis has a small contribution in PLS analysis together with impurity F17H. Impurity 6 chloro has a positive impact on yield. This odd relation is explained in the fact that this substance is present in the production process since the beginning, in the purchased material, and does not purge in any of the production steps.

3.3.1.2 Process photograph

The quality data of the input material explains most part of the variability in the yield of the final step. However, this fact does not exempt a deep analysis of the process itself as the percentages of variability explained are not additive and so, an also very high value for the way the process is being run can be obtained. As such, process photograph type of data was analysed which consists of variables that have a single value for each completed batch (duration of operations, flow rates, and quantities of reactants used). This type of data, as explicit in the name, only gives a shot of the process, not the full unfolding of the batch, and is used mainly to give some initial insights on the process and a starting point for process film type of data.

Exploratory data analysis (PCA model) was performed to the data set and non-relevant variables were excluded, in other words, variables that did not correlate at all with the response variable in question. With the remaining factors, regression analysis (PLS model) was performed to the reduced data set. Through a process of adding and removing variables from the analysis, the best model (with the given data set) was constructed.

Table 3.5: PLS model statistics for the yield of FP as the response variable and the process photograph type of variables for FP process step.

	R2	**Q2**
Comp. 1	0.319	0.206
Cumulative	0.319	0.206

Although with a very low percentage of variability explained by the model, some conclusions can be drawn that were roughly expected. The coefficients for the model are presented below, considering only the relevant variables.

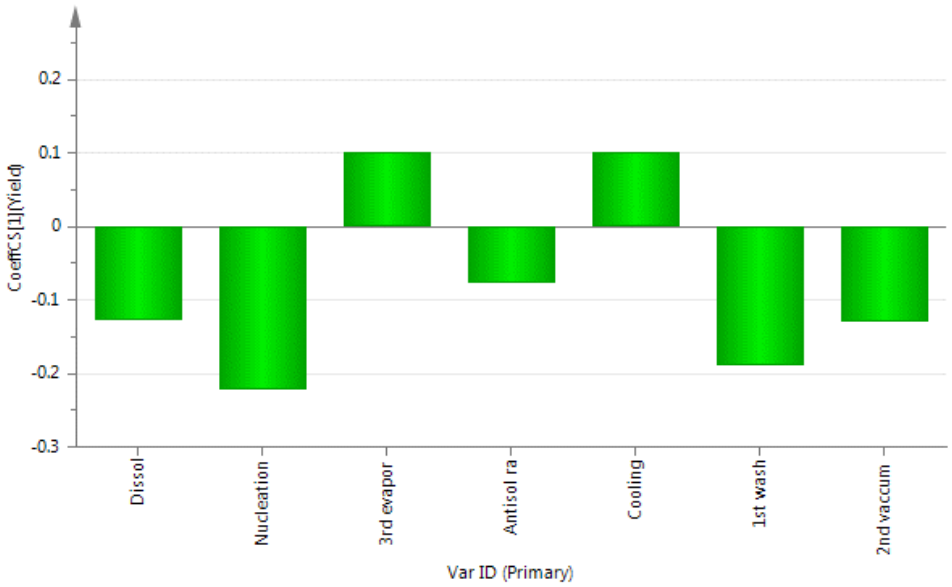


Figure 3.10: PLS model coefficients for the process variables (process photograph type of data) against the yield of FP. ”Dissol” is the time took for the complete dissolution of the input material; ”Nucleation” is the time took for nucleation with solvent evaporation; ”3rd evapor” is the time took to evaporate the remaining solvent after the API nucleation; ”Antisol ra” is the flow rate of antisolvent addition; ”Cooling” is the time spent in the cooling of the suspension after solvent evaporation and antisolvent addition and ”1st wash” and ”2nd vacuum” are related with the filtration step.

The time spent on the nucleation is the variable that has the most negative impact on the yield. This operation stops upon a visual inspection of the mixture by the operators. Assuming that there is no intrinsic variability on this ending point from batch to batch and that the temperature profiles (meaning the evaporation rate) are the same, more time needed to nucleate the suspension means more solvent evaporated which in a broad way can be translated into higher solubility of the solid material in the mixture. This higher solubility can cause incomplete nucleation (whether on the solvent evaporation step or in the antisolvent and cooling steps) thus causing yield losses. A study on the temperature profiles during solvent evaporation is to be performed.

Regarding the cooling step of the crystallization, one can see that more time spent on this operation means higher yields. The cooling rate is specified and deep analysis on possible differences between batches on this variable should be conducted since it is a critical value in cooling crystallization [62].

For the antisolvent rate of addition, higher values of this variable lead to lower yields. This is expectable as lower rates of addition favor nucleation contrary to higher rates that favor crystal growth [62]. Since the amount of antisolvent loaded is specified in the batch production record and does not vary from batch to batch, a more in-depth analysis was conducted on the antisolvent addition time. This duration is specified as CONFIDENTIAL minutes in the operations manual but there were some batches where it was not followed.

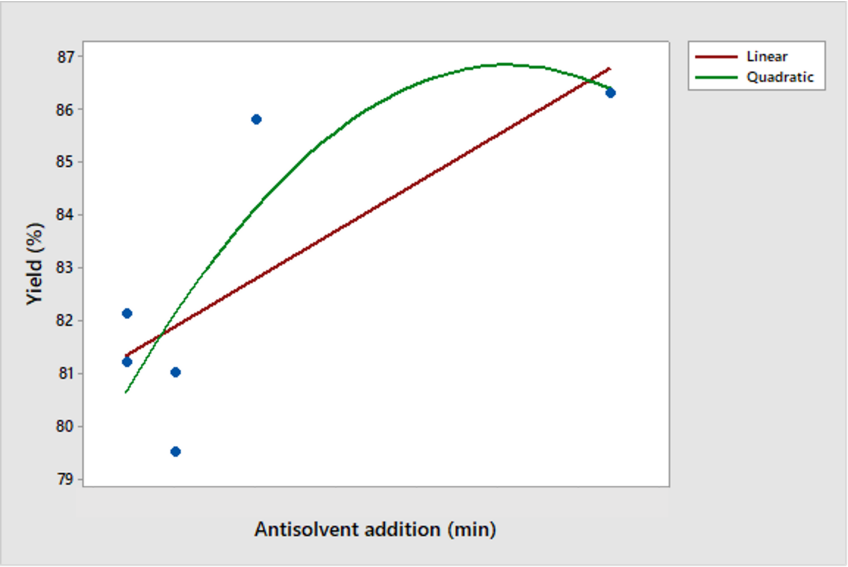


Figure 3.11: Linear and quadratic regression of yield of the final step against the duration of antisolvent addition on the batches that the indication for CONFIDENTIAL minutes of addition time was not followed. For linear regression, $R2$ equals 0.573 and for quadratic regression $R2$ equals 0.652.

As seen in the graph of figure 3.11, the higher the time for addition (translated into a smaller addition rate), the higher the yield. As stated before, this type of analysis is only enough to spark discussion and deeper analysis, in this case, the level profile of the reactor during antisolvent addition.

For the remaining variables, no comments will be made as these operations (crystallization and filtration) will be analysed at a deeper level with process film type of data.

3.3.1.3 Process film

For the more rigorous and incisive analysis of process film type of data, both the crystallization (and all its successive steps) and the filtration will be the subject of study. As illustrated in figure 3.6, firstly, exploratory data analysis and pattern recognition will be performed through BEM and then process modeling through BLM, to figure if the differences spotted in the variable profiles from batch to batch do in fact influence the response variable.

Batch Evolution Model

Since the considered production area is one of the oldest areas on the site, not a lot of variables are actually recorded by the automation system. As such, there is actually no use in displaying the scores on a batch control chart as this method is useful when a lot of variables are to be seen simultaneously and so, for this particular case, the variable profiles can be analysed individually in each crystallization step and during the filtration.

Under exploratory analysis the profiles for each variable were studied. For the solvent evaporation part of the crystallization, the pressure profile has some fluctuations towards the end of the process in roughly all analysed batches. This is due to minor fluctuations in the vacuum pressure applied. The temperature is practically always constant in order to maintain a specific evaporation rate. For the antisolvent

step, the pressure profile has some minor fluctuations also due to differences in the room pressure. Regarding the temperature profile the addition of antisolvent is accompanied by a temperature increase. This is because the addition of the antisolvent increases the reflux temperature of the suspension and so, in order to keep the solvent evaporation, the jacket temperature is increased. For the agitator speed , there is the indication to stay at CONFIDENTIAL during the entire crystallization step. This indication is roughly followed as there were some batches that the agitator speed was considerably different from the stipulated. After the addition of antisolvent, the suspensions is cooled at a specified rate of about CONFIDENTIAL. The pressure profile is on the same line as in the evaporation and antisolvent steps. The specified cooling rate is being followed. The agitator speed during this operation has some fluctuations.

Still, under exploratory data analysis and pattern recognition, the filtration variable profiles are also worth commenting on. From an operational point of view, this is a difficult operation to control, which is visible in the completely random pressure profile during this operation. For the agitator speed in the filter dryer, random peaks appear. The presence of an agitator in the filtration process is crucial due to the aid in mixing and the cake smoothing in case of crack formation. This last phenomenon can cause channeling of the washing solvents which originates a non-effective filtration operation [9]. In the temperature profile a decreasing trend is verified in all batches that are connected with the transfer of the multiple washings at a low temperature.

Batch Level Model

The next step, and the one that actually gives a solid ground to establish any improvement actions, is Batch Level Modelling, where the differences in between batches on the variable profiles are modeled against the Y variable and the impact of such differences on the particular problem to be solved is uncovered. A BLM model was established for each crystallization step and for the filtration as a whole.

Crystallization

Starting with the solvent evaporation step on the crystallization, the following model was obtained.

Table 3.6: BLM model statistics for the solvent evaporation step in FP crystallization and yield as the response variable.

	R2	**Q2**
Comp. 1	0.655	0.426
Comp. 2	0.057	0.051
Cumulative	0.712	0.455

The loadings for the first component during the solvent evaporation step are given in figure 3.12. The second component does not explain the response variable considered. Also, due to this fact, the model coefficients do not vary from the loadings of the first component and so the analysis will be based solely on this metric.

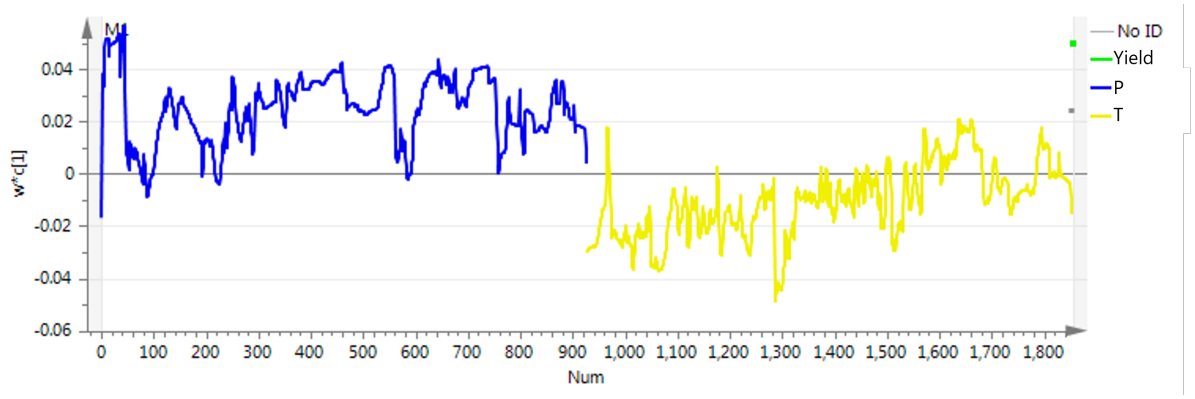


Figure 3.12: Loadings of the first component are given against batch maturity for each variable (pressure as blue and temperature as yellow) for the solvent evaporation step in the crystallization of FP process.

All the crystallization steps occur at atmospheric pressure and so this is not a controlled (nor can it be manipulated) variable during the process. Regarding temperature, the loadings show a lot of sudden fluctuations during the course of the operation which means that the impact of the variable on yield is changing very fast. There is no clear pattern on this contribution and so, no conclusion or action can be drawn from the BLM model of the solvent evaporation analysis.

The BLM model for the antisolvent addition step has the following characteristics.

Table 3.7: BLM model statistics for the antisolvent addition step in FP crystallization and yield as the response variable.

	R2	**Q2**
Comp. 1	0.647	0.459
Cumulative	0.647	0.459

Since the model only has one component, the loadings of this first component are the same as the model coefficients.

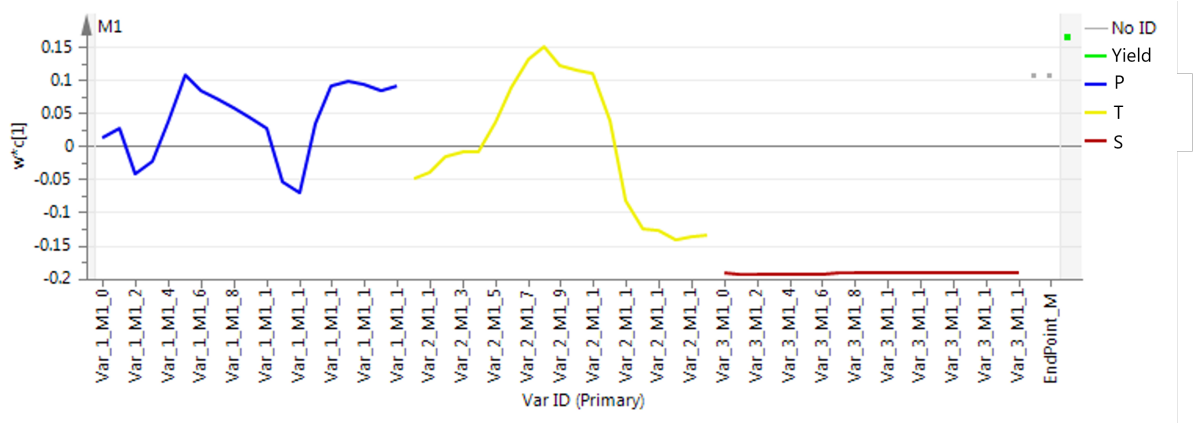


Figure 3.13: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the antisolvent addition step in the crystallization of FP process.

No comment will be made on the pressure contribution (blue line in figure 3.13) because the op-

eration is carried out at atmospheric pressure. Regarding the temperature impact on the yield, it can be observed that at the beginning of the operation it rises to be strongly positive (meaning that high values of temperature lead to high values of yield) and then decays to be strongly negative. Combining the yellow line in figure 3.13 with the variable profile two improvements could be proposed based on this model: firstly, to promote faster heating at the beginning of the operation since up until the middle of the duration, high temperatures favor high yields and secondly, to lower the final temperature target since the impact of the variable is negative, high temperatures favor low yields. The agitator speed has, during the entire antisolvent addition, a strong negative impact on yield, that is to say, high values favor low yields. The only conclusion that can be withdrawn is that the indication to stay on CONFIDENTIAL should be followed.

On both profiles, temperature and agitator speed, during antisolvent addition, batch 025 seems an outlier. This is verified by Hotelling´s T^2 plot for the model established.

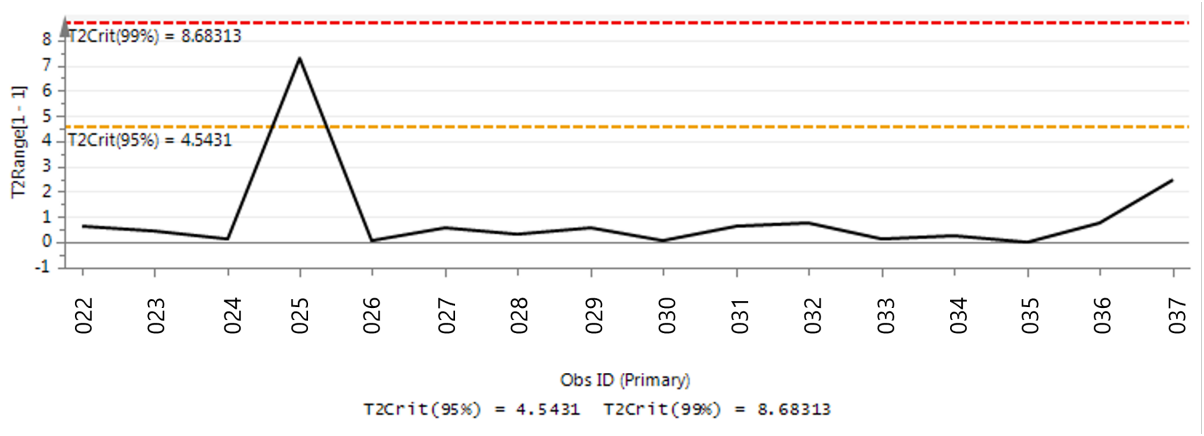


Figure 3.14: Hotelling´s T^2 for the BLM model built for the antisolvent addition.

An outlier has a huge effect on the model [10] and so, in order to evaluate if the same conclusions could still be drawn, batch 025 was removed and a new BLM model was constructed.

Table 3.8: BLM model statistics for the antisolvent addition step in FP crystallization and yield as response variable without batch 025.

	R2	**Q2**
Comp. 1	0.669	0.553
Cumulative	0.669	0.553

An increase of 3.4% in $R2$ and of 20.5% in $Q2$ is observed when batch 025 is removed meaning that a more accurate and robust model is actually obtained. The loadings for the first component are presented below.

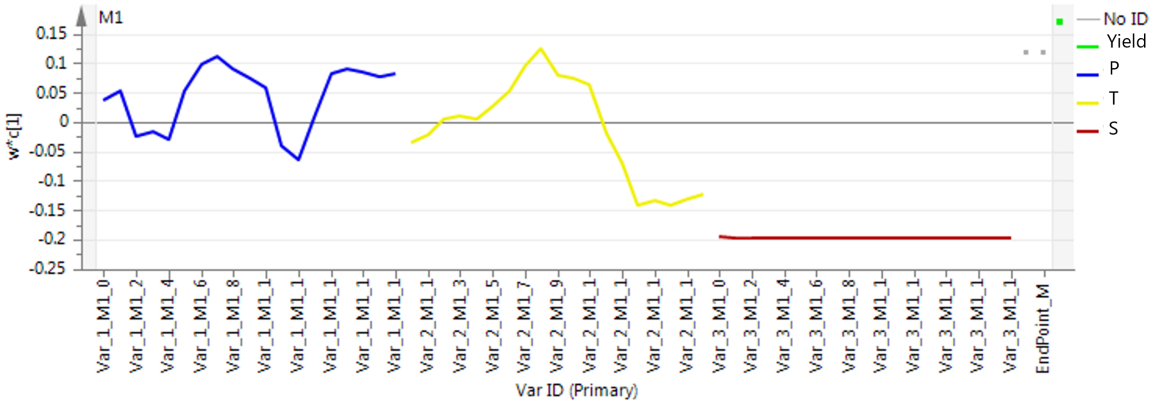


Figure 3.15: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the antisolvent addition step in the crystallization of FP process with batch 025 removed from the model.

With the removal of the outlier from the model, both the agitator speed and temperature impact remain unaltered meaning that the outlier observation was not twisting the conclusions previously stated.

The final crystallization step BLM model has the following fitting, presented in table 3.9.

Table 3.9: BLM model statistics for the cooling step in FP crystallization and yield as the response variable.

	R2	**Q2**
Comp. 1	0.648	0.535
Cumulative	0.648	0.535

A similar amount of variability explained by the model ($R2$) is obtained to the antisolvent addition model. Values in this order of magnitude are considered good and therefore the models being presented can be considered as robust models. Once again, since the model only has one component, the loadings for the first component are presented.

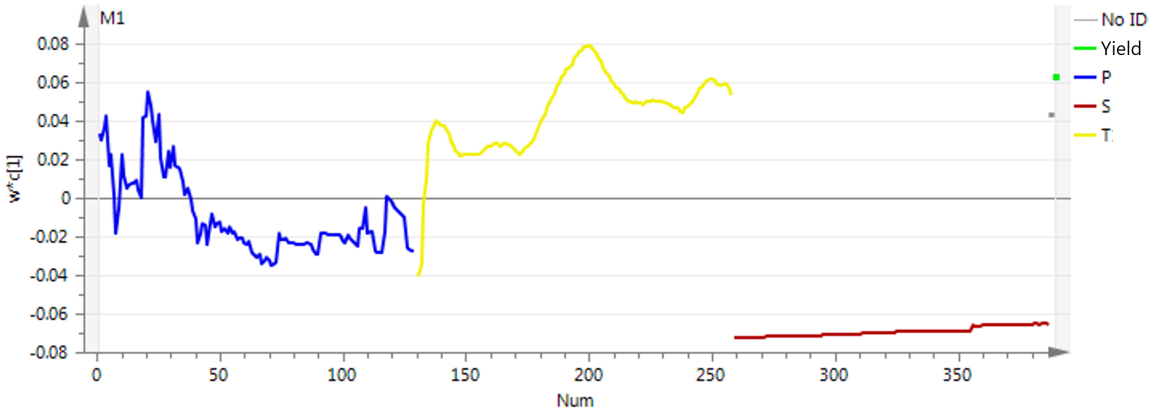


Figure 3.16: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the cooling step in the crystallization of FP process.

Once more, the contribution of the agitator speed is firmly negative and so, high values of agitator speed lead to low values of yield during the entire operation. Still, the only conclusion that can be withdrawn is to follow the indication to stay on CONFIDENTIAL during the operation. For pressure, there are also no comments to add. Regarding the temperature´s impact on yield, it is clear that it stays strongly positive during the entire cooling operation, apart from the very beginning in which there is an almost vertical ascending contribution. The cooling rate for this operation is specified as CONFIDENTIAL and overall, there are not many differences in the cooling ramp profile from batch to batch. However, what the loadings of the model are exhibiting is that, although in some parts of the operation more evident than others for example the loading peak at x-axis value around 200, lower cooling rates (meaning higher temperatures) lead to higher yields. The contribution of temperature is positive during the final end of the operation, meaning that, higher final temperatures lead to higher yields. The final temperature has to be within the process interval: CONFIDENTIAL but there is the indication to aim at CONFIDENTIAL.

Filtration

The filtration step of FP process was also analysed. The drying step was not since it was an operation classified as non-critical to the yield of the process. A BLM model with three components was obtained.

Table 3.10: BLM model statistics for the filtration step in FP process and yield as the response variable.

	R2	Q2
Comp. 1	0.497	0.364
Comp. 2	0.341	0.165
Comp. 3	0.122	0.564
Cumulative	0.961	0.768

Since the BLM model for the filtration step has three components, the loadings will be presented on a scatter plot like the loadings for a PCA model.

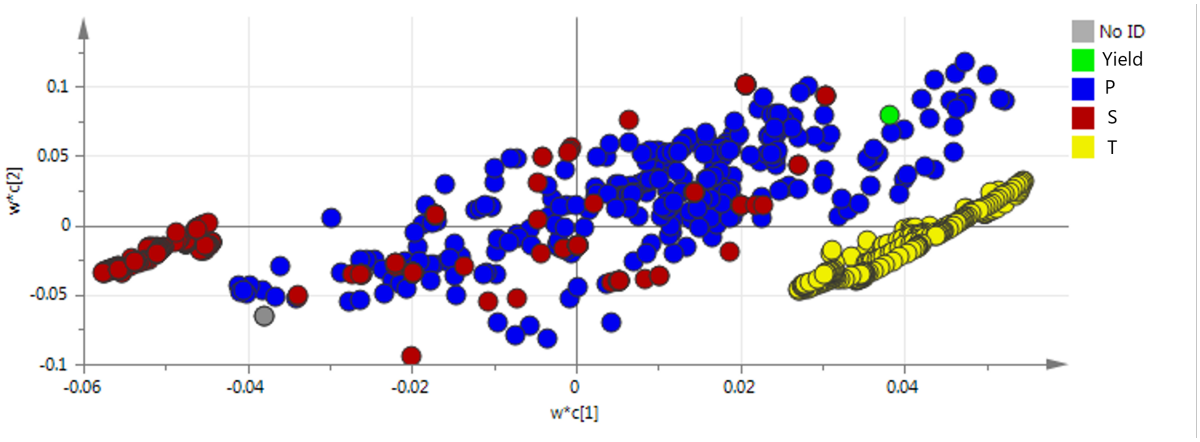


Figure 3.17: Loadings of the first component versus the second component for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the filtration step in the FP process.

The Y variable is explained by both the first and second components. However, this does not happen

to the third component and so the analysis will be focused only on the two first. Looking only at the projections on the first component (x-axis), only temperature has a positive impact on yield during the entire operation. However, looking at the projections for the second component (y-axis), the impact of the variable in yield is close to null, sometimes being weakly negative and others weakly positive. This way of displaying the loadings does not allow to identify which sections of the entire filtration had a positive, negative, or null impact on yield. This is possible when displaying the loadings one component at a time like it has been done until this point.

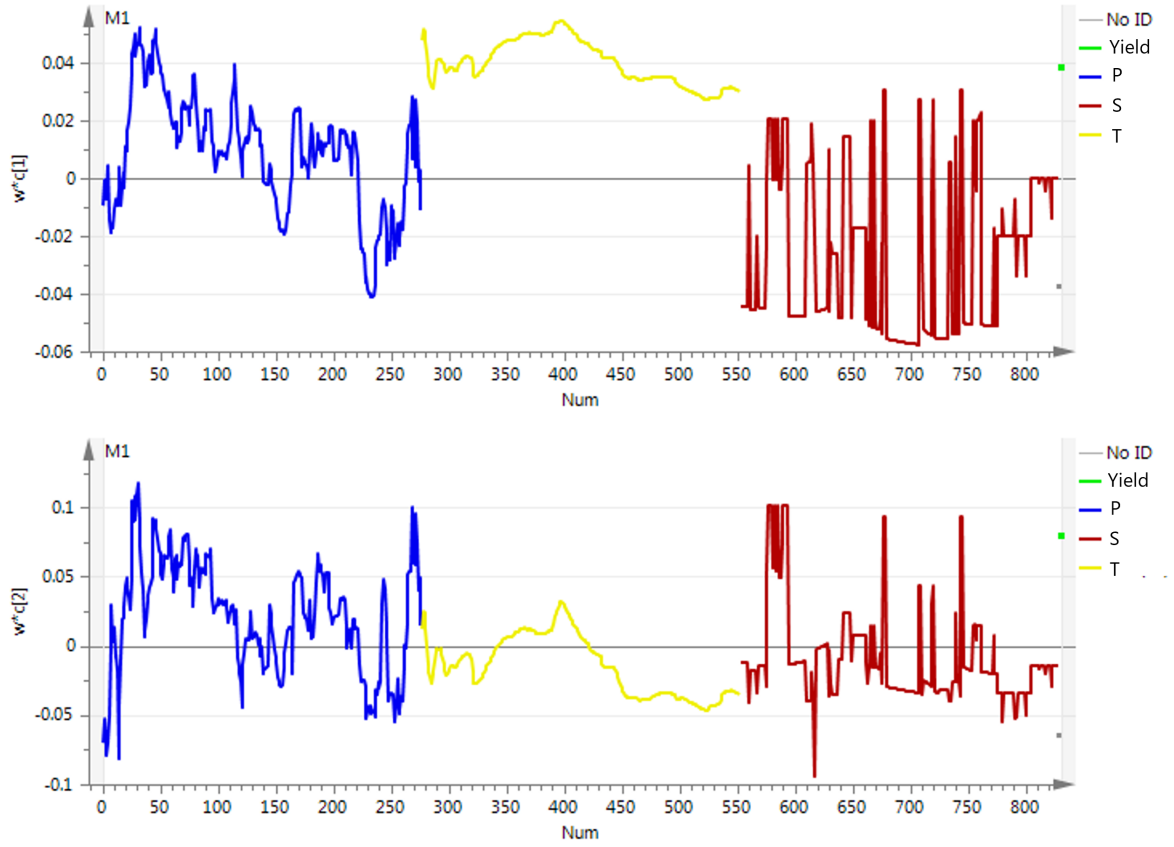


Figure 3.18: Loadings of the first component (top) and of the second component (bottom) given against batch maturity for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the filtration step in the FP process.

A closer look at both graphs on figure 3.18 tells that both pressure and agitator speed have a very unstable contribution to yield according to the two components. This unstable relationship with the Y variable is confirmed due to the also unstable and unstandardized variable profiles. Prior to optimization, there must be a standardization step and these two cases are the proof of that. Regarding temperature, confirming what was concluded with figure 3.17, according to the first component there is a positive and strong relationship during all the extent of filtration with yield, meaning that higher temperatures favor yield. That is not the case with the second component, where, although the yield is more poorly explained than by the first component, there is a clear weak relation to the end of the operation with yield. During the filtration, the temperature is not controlled, and the decreasing trend observed for all batches is due to the transfer of the washings, which is done between CONFIDENTIAL and CONFIDENTIAL

with no clearer indication. Attending only to the relation shown in the first component an improvement action to be proposed could be to transfer the washings during filtration closer to the upper end of the stipulated process interval.

3.3.1.4 Assay

FP process is the last chemical process in the production train of the API. Therefore, all operations have tighter control and batches are of a much smaller size (much less input material quantity) than the batches of the remaining intermediaries. All the analysis presented had, as response or Y variable, the yield of this production step. Since it is the last chemical step before the final product, it is important to check if the implementation of such measures to optimize yield will damage the process performance in other equally relevant variables.

Process performance, in the context of chemical-pharmaceutical processes, can be viewed in terms of throughput (the yield of the process) and quality. The variable found to cover quality performance was the product´s assay. The assay is a different measure from purity. The latter is a quantitative measure, a mere difference between the unit value and the sum of all the impurities. Contrary to this, the assay is a measure of the potency or activity of a certain analyte (in this case, the desired API) on a substance [63].

Once again, exploratory data analysis was employed in the first approach to the problem. A PCA model for all the quality release results of FP (assay and impurities) and yield was constructed with two PC´s: PC1 explains 0.505 of the variability in the data set and PC2 explains 0.202. No outliers were found through Hotelling´s T^2 and through DModX. The loadings scatter plot is presented below.

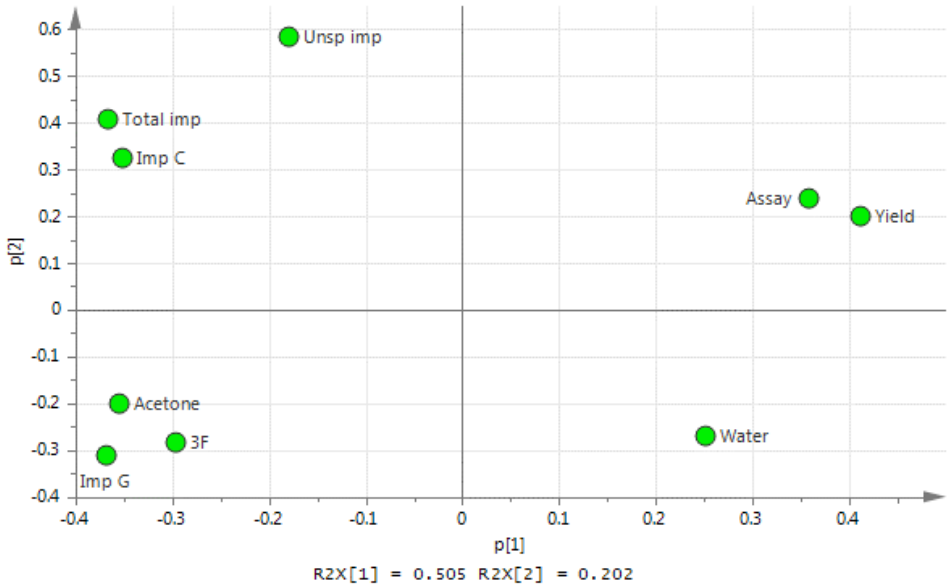


Figure 3.19: Scatter plot of the loadings for the PCA analysis of yield of FP production step and the quality data of FP (impurities and assay).

Assay and yield are strongly and positively correlated, which, in theory, means that process changes in order to maximize yield will also lead to a maximization of the assay. Nevertheless, the relevant BLM

models were re-built considering the new response variable.

Table 3.11: BLM model statistics for the first component of the antisolvent addition, cooling, and filtration step in FP production process considering yield and assay as response variables. The variation in percentage from the models considering yield as response variable to the models considering assay is also presented.

Response variable	**Yield**		**Assay**		**Variation (%)**	
Parameter	**R2**	**Q2**	**R2**	**Q2**	**R2**	**Q2**
Antisolvent addition	0.647	0.459	0.612	0.497	-5.410	8.279
Cooling	0.648	0.535	0.873	0.516	34.722	-3.551
Filtration	0.497	0.364	0.754	0.656	51.710	80.220

There is never a strong decrease in both $R2$ and $Q2$ when modeling assay instead of yield. Also, the increase in percentage is considerable when looking at the filtration step (only the first component statistics are presented in table 3.11).

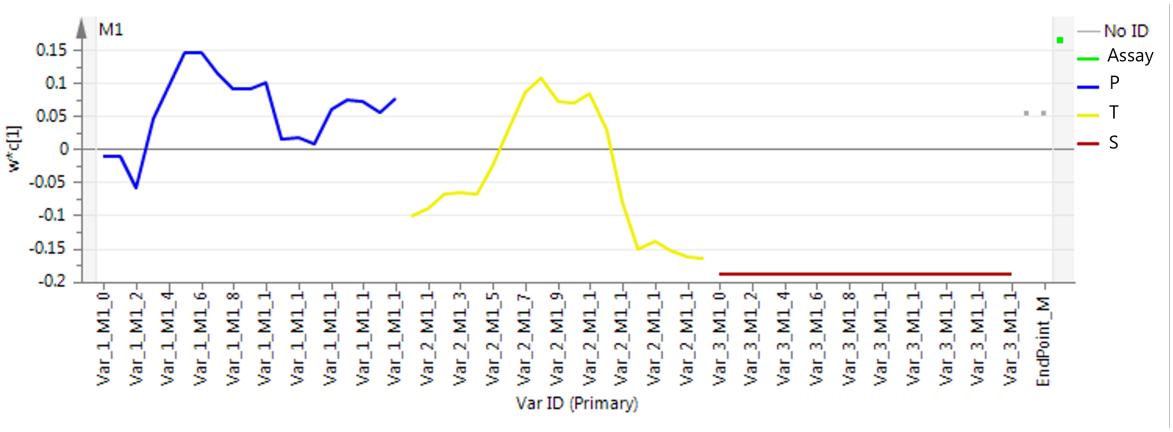


Figure 3.20: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow and agitator speed as red) for the antisolvent addition step in the crystallization of FP process with assay as the response variable.

The loadings profile for temperature displayed is very similar to the loadings profile displayed in figure 3.13. For the assay, it seems that in the middle of the operation, the positive impact of temperature is less notorious than to yield since, in the latter, the loadings for temperature are in the same order of magnitude as the response variable. However, the important fact to retain is that there is no contrary relationship in the loadings of both temperature and agitator speed

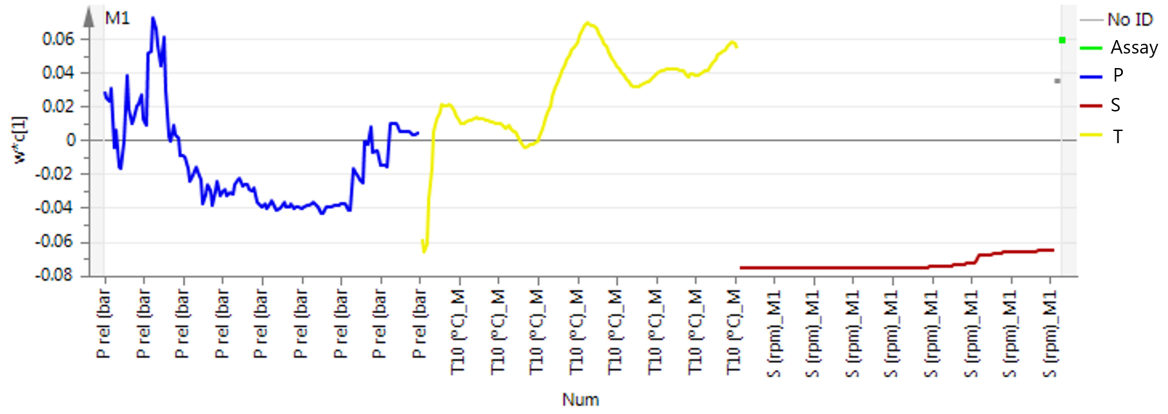


Figure 3.21: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the cooling step in the crystallization of FP process with assay as the response variable.

The same situation is verified for the cooling step of the crystallization. The temperature loadings pattern for assay is similar to the one for yield (figure 3.16). Once more, the important thing to be retained is that the contribution of both temperature and agitator speed to yield is not contrary to the contribution to assay.

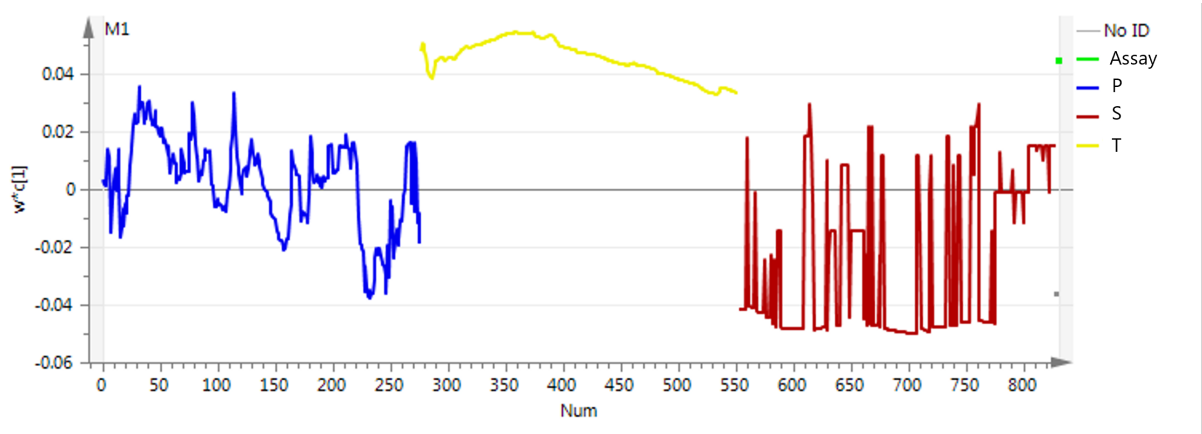


Figure 3.22: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow and agitator speed as red) for the filtration step in the FP process with assay as the response variable.

In the filtration step, the temperature loadings are positive and on the same order of magnitude as the loading for assay, as it happens for yield (figure 3.18, top position).

The problem to be solved in this process improvement project was not related to the assay of the final product, FP. However, it was deemed important to evaluate if the actions being proposed to optimize the process in terms of throughput would not jeopardize the process performance in terms of quality. The findings are that yield and quality go on the same way, *i.e.*, optimizing throughput performance will also lead to an optimization of quality performance.

3.3.2 Intermediary 4 process step analysis

According to the PLS analysis conducted to exploit the dependence of yield with intermediary 4 quality data (analysis coefficients in figure 3.24), impurity H has the strongest negative impact on yield followed by impurity G. The quality data of the input material on the last production step explains most of the variability on yield, 86.3%, and the model has an excellent predictive ability, 79.3%, as shown by the PLS model statistics in table 3.4. As such, and following the work methodology, the production step that leads to the formation of intermediary 4 will also be modeled regarding impurity H and G as response variables.

3.3.2.1 Quality

Only impurity H and impurity G, from the release data of intermediary 4 were considered for the subsequent analysis. PCA modeling was performed with two PC´s: PC1 explaining 37.6% and PC2 27.5% yielding a cumulative fraction of variability explained by the model of 65.1%. No outliers were found on the scores plot neither on the DModX plot. The loadings of the first two components are shown below.

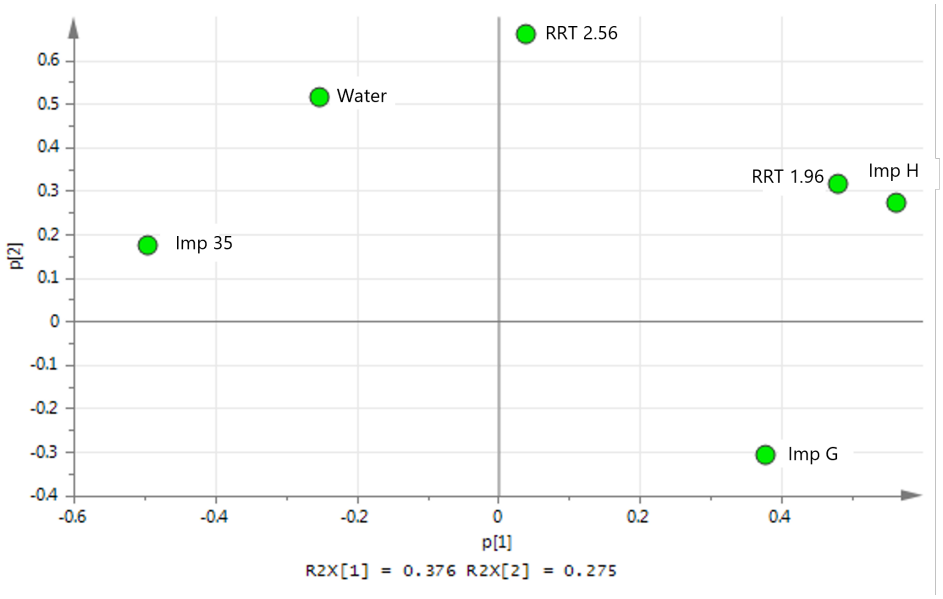


Figure 3.23: Scatter plot of the loadings for the PCA analysis of impurities H and G of intermediary 4 and the quality data of intermediary 3. ”RRT” means relative retention time which is an analytical variable that can be used to classify unknown molecules.

Both intermediary 4 impurities are equally described by PC1 and impurity with RRT of 1.96 seems to have a strong positive effect on them. Impurity 35, according to PC1 has a negative relation with impurity H which means that high values of impurity 35 lead to lower values of impurity H.

A more incisive method is necessary and a PLS model is fitted to the data set, considering firstly impurity H as the response variable and then impurity G.

For impurity H, a PLS model with just one component was obtained and the model statistics are presented in the table below.

Table 3.12: PLS model statistics for impurity H of intermediary 4 as the response variable and the quality data of intermediary 3 as independent variables.

	R2	Q2
Comp. 1	0.664	0.215
Cumulative	0.664	0.215

Such a big difference between the model performance (described by $R2$) and its predictive ability ($Q2$) may indicate over-fitting of the data. Model complexity cannot be reduced since the PLS model only has one component.

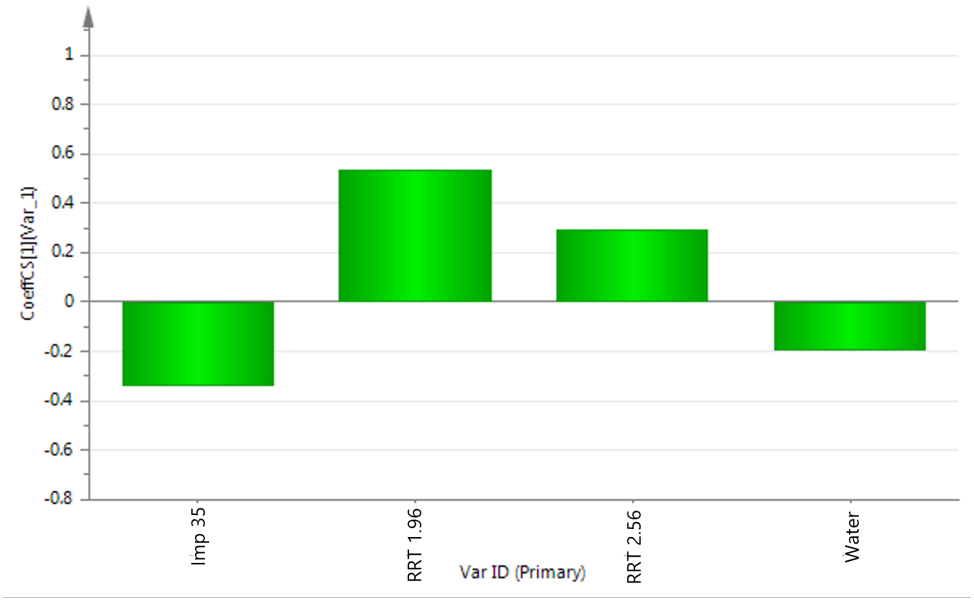


Figure 3.24: PLS model coefficients for the several impurities present in the intermediary 3 quality data against impurity H of intermediary 4.

Impurity 35 still has a negative impact on impurity H which can be seen as an odd relationship. Both impurities that appear with an RRT of 2.56 and 1.96 can be seen as precursors of impurity H.

Impurity G is now taken as the response variable, instead of impurity H and the intermediary 3 quality data is modeled against the new variable of interest.

Table 3.13: PLS model statistics for impurity G of intermediary 4 as the response variable and the quality data of intermediary 3 as independent variables.

	R2	Q2
Comp. 1	0.255	-0.068
Comp. 2	0.031	-0.525
Cumulative	0.286	-0.175

A very small fraction of the variability explained in the Y-variable by the model, together with negative predictive ability lead to the conclusion that impurity G does not depend on any of the substances present

in the release data of intermediary 3 and so the elimination of the intermediary 4 impurity G content will have to derive from the analysis of the production process in question.

3.3.2.2 Process photograph

As seen in table 3.4, the impurities present in the input material to the FP process step explain most part of the variability in the yield which is the primary problem to be solved on this project. However, in order to establish concrete improvement actions, a further step back has to be taken in order to evaluate which sections of the intermediary 4 process step are causing the increase in the identified impurities that lead to a decrease in yield.

Exploratory data analysis was conducted to process photograph type of data. For the subsequent analysis (also regarding process film type of data) only 8 production batches will be analysed. Although being considered a very small number of batches, data was only available for these batches and so, the analysis was carried on. With the non-relevant variables excluded, regression analysis was performed following the same strategy already described: adding and eliminating variables until the best model was found. For impurity H the following model was obtained.

Table 3.14: PLS model statistics for impurity H as the response variable and the process photograph type of variables for intermediary 4 process step.

	R2	**Q2**
Comp. 1	0.940	0.650
Comp. 2	0.041	0.504
Cumulative	0.981	0.826

An excellent fitting is observed of the selected independent variables against impurity H as Y-variable. The model coefficients are presented below in figure 3.25.

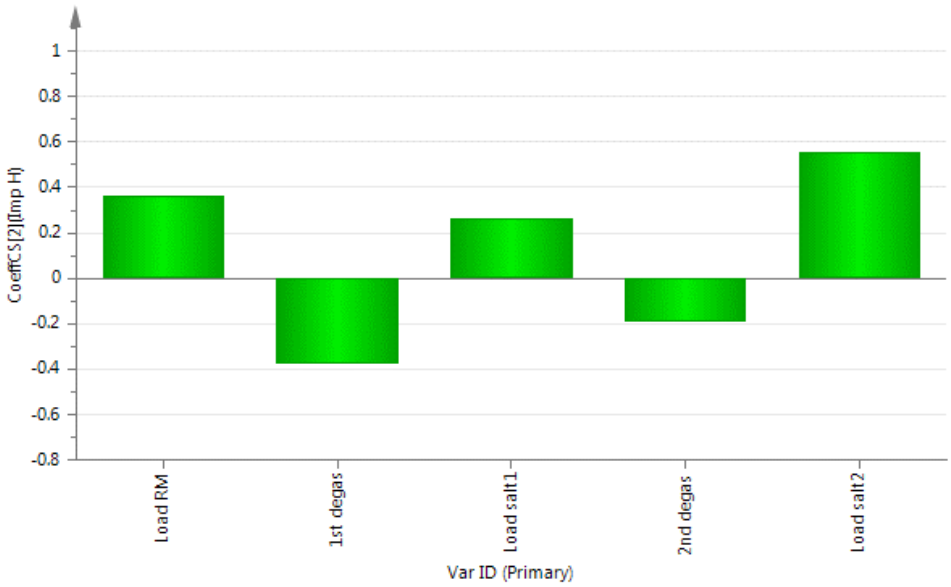


Figure 3.25: PLS model coefficients for the process variables (process photograph type of data) against impurity H. ”Load RM” is the time took for the charging of the raw material (intermediary 3); ”1st degas” is the time took on the first degassing step; ”Load salt1” is the time took to charge the first inorganic salt; ”2nd degas” is the time took on the second degassing step and ”Load salt2” is the time took to charge the second inorganic salt.

Impurity H is formed whenever an oxidizing agent is present and so, the process step being analysed has multiple degassing steps prior to the main reaction. The most relevant terms included in the model are precisely the duration of operations deeply related to the entrance or exiting of oxygen in the process vessel, the degassing steps, and the loading of the reactants: a positive relation of the time took to load the reactants is observed meaning that more time spent on loading the reactants yields more impurity H content in intermediary 4 and a negative relation of the degassing duration is observed meaning that less time spent on this operations, more impurity H content will be detected. All these findings are natural since the major motif for the impurity‘s formation is the presence of an oxidizing agent, in this case, oxygen.

The timely evolution of the variables being analysed is depicted in figure 3.26.

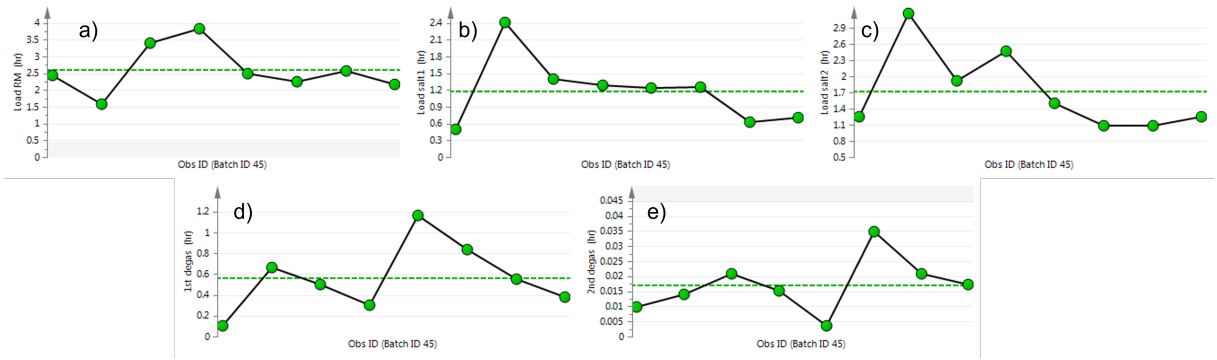


Figure 3.26: Evolution of the duration of the degassing and loading of reactants. Figure a), b) and c) are the duration of the loading raw materials, the first and second innorganic salt respectively, and figures d) and e) are the first and second degassing steps respectively.

For the degassing steps, figure 3.26 part d) and e), a negative trend on the more recent production batches is observed which is not the desired state since more time spent on these operations leads to less impurity H content in the process output. In the same way, a decreasing trend is observed on more recent batches for the time took to load the reactants, figure 3.26 part a), b) and c). Contrary to the degassing steps, where low operational times favor high impurity content, here, low duration favors low impurity content due to the lower atmosphere exposition, thus reducing the oxygen level.

For impurity G, no set of meaningful terms was found to be statistically relevant so, no analysis of the process photograph type of data was conducted considering impurity G as the response variable. However, the process film type of data will be analysed for both impurities since the effort needed for that is very low and the fact that, although important to get an overview of some critical operations and gain process knowledge, the process photograph data analysis results do not fully condition process film analysis.

3.3.2.3 Process film

Process film type of data was analysed, firstly at an exploratory level with BEM and then, considering both impurities H and G as response variables for BLM. Data was only available for 6 production batches.

Batch Evolution Model

The process equipment for the intermediary 4 production step is different from the FP step. Still, there is no use in displaying the scores for evaluating deviating batch trajectories since the variables being recorded are few (for intermediary 4 equipment the jacket temperature is also recorded). The variable profiles will be analysed separately for the pre-reaction, reaction, crystallization and filtration steps.

The pre-reaction operations comprise the degassing steps and the loading of the reactants (the inorganic salts). For the first degassing step, the pressure profile presents random negative peaks, which are deeply related with the vacuum applied in the degassing step. During these steps, the temperature should be between CONFIDENTIAL and CONFIDENTIAL, but preferably closer to the lower limit. The agitator speed is supposed to be at CONFIDENTIAL but there were two batches that the variable was

set at CONFIDENTIAL. After the first degassing, the load of the first salt takes place followed by another degassing step. The temperature shows an overall increasing tendency. This is due to the fact that the jacket temperature is reduced to zero and remains constant during the entire subsequent operations, until the end of the reaction. Regarding, the agitator speed profile, the scenario for the first degassing step is repeated with two deviating batches not following the CONFIDENTIAL indication. The charge of the second salt is followed by another degassing step. The temperature is kept under the stipulated limit and the scenario for the agitator speed is the same as for the first degassing step and for the charge of the first salt. Finally, the reaction takes place with the slow addition of the final reactant in the gaseous state but liquefied upon entering the reactor. This is the cause for the decreasing tendency in the pressure profile for all analysed batches, at the beginning of the operation. The random peaks of pressure observed are due to the several purges carried out to rinse out the gas bottle. The temperature during the reaction is supposed to be between CONFIDENTIAL and CONFIDENTIAL with no clearer indication, which generates some variability in the temperature profiles for the operation. The scenario regarding the agitator speed is the same as in the previous steps: two batches had the agitator set-point at CONFIDENTIAL instead of the stipulated CONFIDENTIAL.

The intermediary 4 crystallization is performed over two sequential steps: antisolvent addition and cooling. On the reactor where crystallization takes place, the sensors for pressure, agitator speed, and temperature had a different recording in the automation system on the vast majority of the analysed production batches and so, point values of the variables were rounded to units. During the antisolvent addition, pressure oscillates randomly between CONFIDENTIAL and CONFIDENTIAL. This variable profile is not to be regarded given the problem in the sensors. In the level sensor, there was no problem and so the variable profile is worth analysing. All batches fulfilled the indication of adding the antisolvent in at least CONFIDENTIAL, however, some batches took way longer. The level profiles follow the same pattern up until CONFIDENTIAL of addition when the differences start to occur: in some batches, the remaining quantity of antisolvent is added abruptly and on other batches, the smooth profile is maintained. Concerning the agitator speed profile, strongly different strategies were followed on the considered production batches. Since the antisolvent addition is an exothermic operation at the beginning, it is normal to observe a slight increase in temperature which is then followed by an overall decrease because the temperature is supposed to be within the process limits (CONFIDENTIAL$<$T(ºC)$<$CONFIDENTIAL). The strategy to maintain the temperature in the reactor is irregular since the jacket temperature profile does not show any clear and identifiable tendency. Regarding the cooling step of the crystallization, the pressure was set, during the entire operation as CONFIDENTIAL. The agitator speed assumed different and sometimes changing values for the considered production batches. These differences will be analysed in terms of BLM with the response variables. The cooling ramp on intermediary 4 is less tightly controlled than in the FP process, however, on all analysed batches, the final temperature was inside the process limits (CONFIDENTIAL$<$T(ºC)$<$CONFIDENTIAL). To complete this cooling step, the jacket temperature presents a completely random evolution with time. The cooling ramp is controlled in an arbitrary manner, indicating the way for standardization to take place, before any optimization takes place.

The variable profiles for the filtration step, especially pressure and agitator speed appear as random as the ones for FP process. Regarding temperature, a clear increasing trend is observed which is due to the fact that the washings transferred are at a higher temperature than the cake inside the filter-dryer.

Batch Level Model

Following the same analysis path as to FP production process, BLM was applied to the data considering the two impurities identified as critical, impurity H and G. The results will be presented firstly for impurity H and secondly for impurity G. Only operations with relevant models, that actually can be the basis for optimization actions, will be presented: the variable profiles during the crystallization and the filtration were found to not have any relevant influence on both response variables considered, impurity H and G.

Impurity H

The charge of the first salt was modeled against impurity H and the following model was obtained.

Table 3.15: BLM model statistics for the charge of the first salt in intermediary 4 reaction step and impurity H as the response variable.

	R2	Q2
Comp. 1	0.877	0.630
Comp. 2	0.107	0.695
Cumulative	0.984	0.887

Since the best model has two components, the loadings will be presented on a scatter plot, colored by variable.

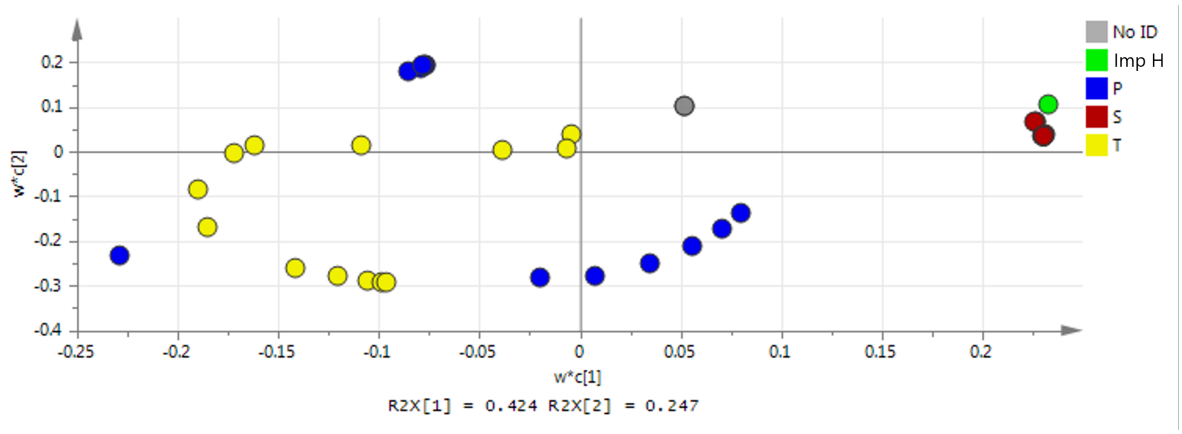


Figure 3.27: Loadings of the first component versus the second component for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the charge of the first salt during the reaction step of intermediary 4 process.

Given the very low relevance of the second component to the current response variable, only the loadings of the first component will be analysed versus batch maturity, as it has been done until this point.

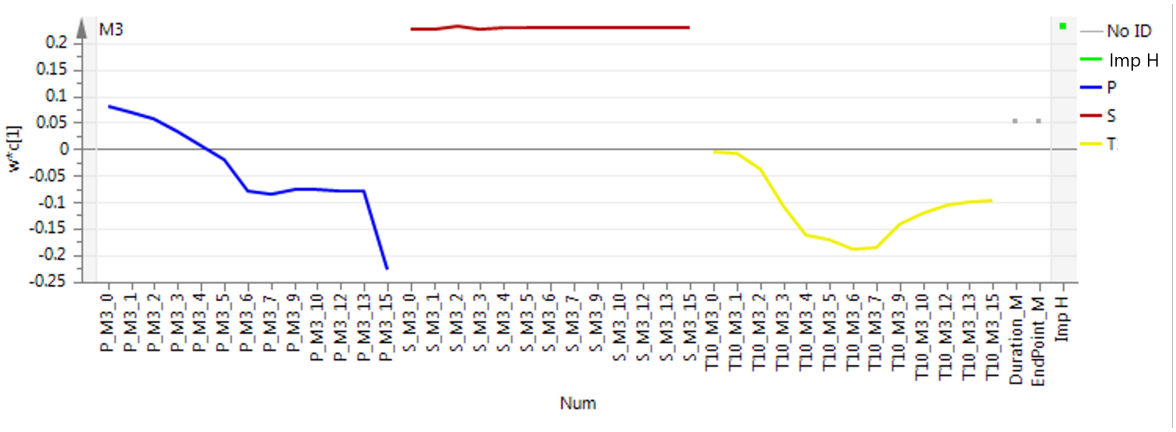


Figure 3.28: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the charge of the first salt during the reaction step of intermediary 4 process.

A strong positive relationship of the agitator speed with impurity H during the entire operation is verified: high values of agitator speed, during the entire operation, yield high values of impurity H in intermediary 4. Only two batches had agitation profiles higher than CONFIDENTIAL, which is the process indication for this variable. As such, in order to lower impurity H content in the process output, this indication should be followed. Throughout the steps that comprise the reaction of intermediary 4 (the degassing steps and loading of reactants), this positive relation with impurity H is verified, due to the outlier batches that had the agitator set-point at CONFIDENTIAL. As such, the indication to stay at CONFIDENTIAL should be followed during all these steps. In regards to temperature, a negative relation is verified, which is especially relevant (more negative) in the middle of the charge but never actually gets positive. The temperature is kept within the process limits (CONFIDENTIAL<T(ºC)<CONFIDENTIAL) but there is the indication of staying as close as possible to the lower limit of the interval. Given the relation presented above, lower values for temperature during the entire operation will increase the content in impurity H thus decreasing the yield in FP process.

A similar model was built to check if the temperature relation with impurity H was maintained during the load of the second salt. The following model with two components was obtained.

Table 3.16: BLM model statistics for the charge of the second salt in intermediary 4 reaction step and impurity H as the response variable.

	R2	**Q2**
Comp. 1	0.883	0.500
Comp. 2	0.096	0.651
Cumulative	0.979	0.825

It is important to point out that both models exhibit a great fitting to the data, giving more robustness to the improvement actions that are withdrawn from them.

The loadings scatter plot for the model concerning the second salt charge are presented below.

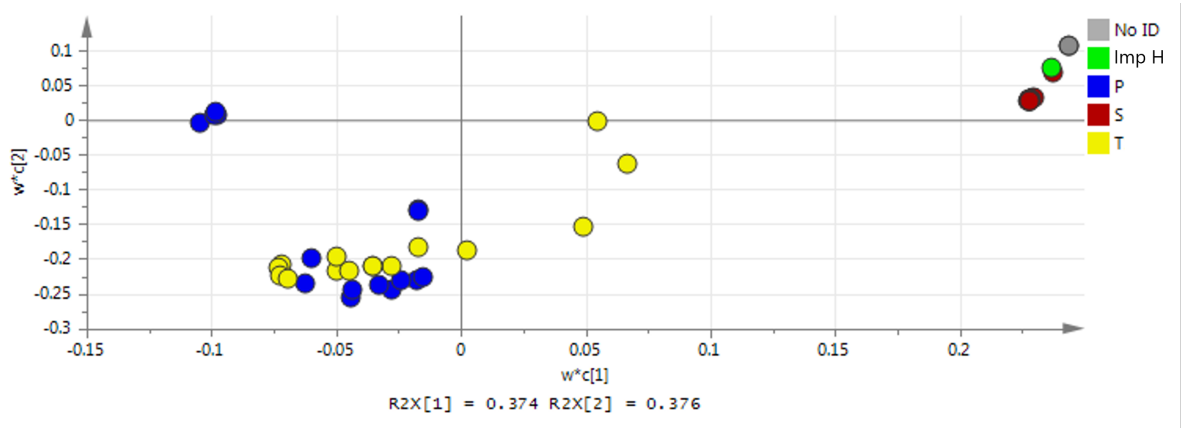


Figure 3.29: Loadings of the first component versus the second component for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the charge of second salt during the reaction step of intermediary 4 process.

The first component is the only one that is actually relevant to the response variable. The main thing to withdraw from this analysis is that temperature does not present a strong positive relation with impurity H. There are some parts of the operation where in fact there is positive relation with the response variable but it is not to be concerned since it is a weak relation. Consequently, the improvement action taken from the modeling of the load of the first salt can be extended to the load of the second salt. It is worth recalling that these models are only considering one response variable and that the relations presented can be changed when considering other variables, like other impurities or the process throughput.

Finally, the reaction itself (the addition of the main reactant) is modeled. A model with two components and a good fitting was obtained.

Table 3.17: BLM model statistics for the reaction step in intermediary 4 and impurity H as the response variable.

	R2	**Q2**
Comp. 1	0.805	0.561
Comp. 2	0.147	0.445
Cumulative	0.952	0.757

Due to the fact of having two components, the loadings are presented firstly on a scatter plot, has it has been done.

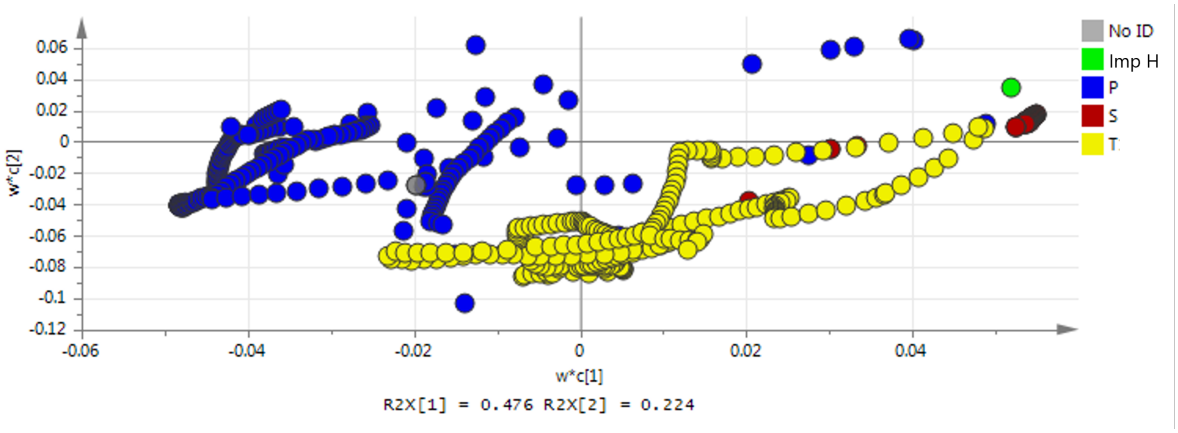


Figure 3.30: Loadings of the first component versus the second component for each variable (pressure as blue, temperature as yellow and agitator speed as red) for the reaction step of intermediary 4 process.

Once again, the second component is much less relevant to the response variable than the first component, allowing it to analyse, on a timely basis, only the loadings for the first component, as they are presented in figure 3.31.

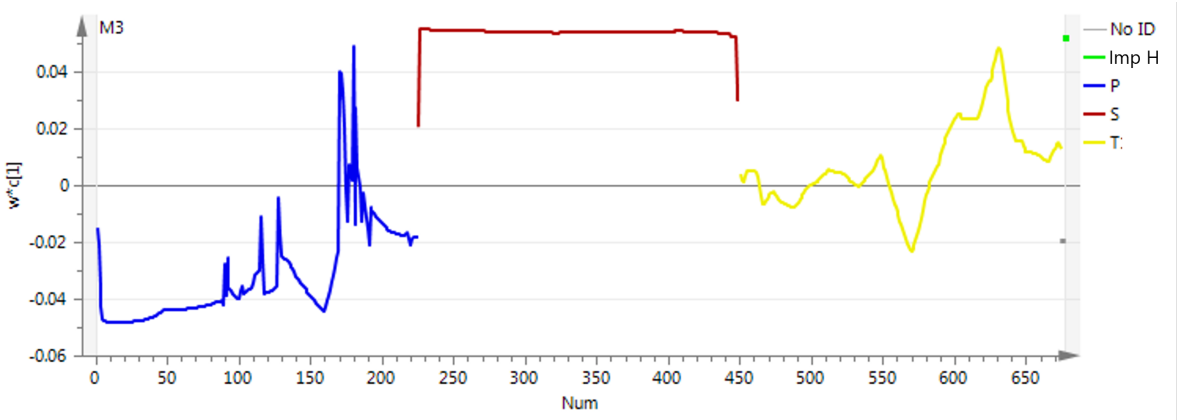


Figure 3.31: Loadings of the first component given against batch maturity for each variable (pressure as blue, temperature as yellow and agitator speed as red) for the reaction step of intermediary 4 process.

The relation of agitator speed with the response variable is maintained and the indication to stay at CONFIDENTIAL should be followed. For temperature, a non-relevant impact is observed through the vast majority of the reaction, with the loadings being close to zero, which is followed by a sudden negative and positive peak. These changing contributions cannot yield any action to optimize the response variable. A closer and combined look at the pressure profile and the loadings for pressure reveals that the peaks in the variable profile coincide with the peaks in the loadings. Setting aside the positive peaks on the loadings, a clear negative and strong relation, especially at the beginning of the addition of the main reactant is observed between the variable and impurity H. In order to decrease impurity H content in the process output, higher pressures should be kept during the initial phase of addition.

Impurity G

The charge of the first salt was modeled against the new response variable, impurity G. A two-component model was obtained.

Table 3.18: BLM model statistics for the charge of the first salt in intermediary 4 reaction step and impurity G as the response variable.

	R2	Q2
Comp. 1	0.584	0.255
Comp. 2	0.370	0.135
Cumulative	0.954	0.355

A big difference is actually observed between $R2$ and $Q2$ which might indicate over-fitting. However, the variation in $Q2$ from the first to the second component is positive and so, a two-component model will be analysed. The loadings scatter plot are presented below.

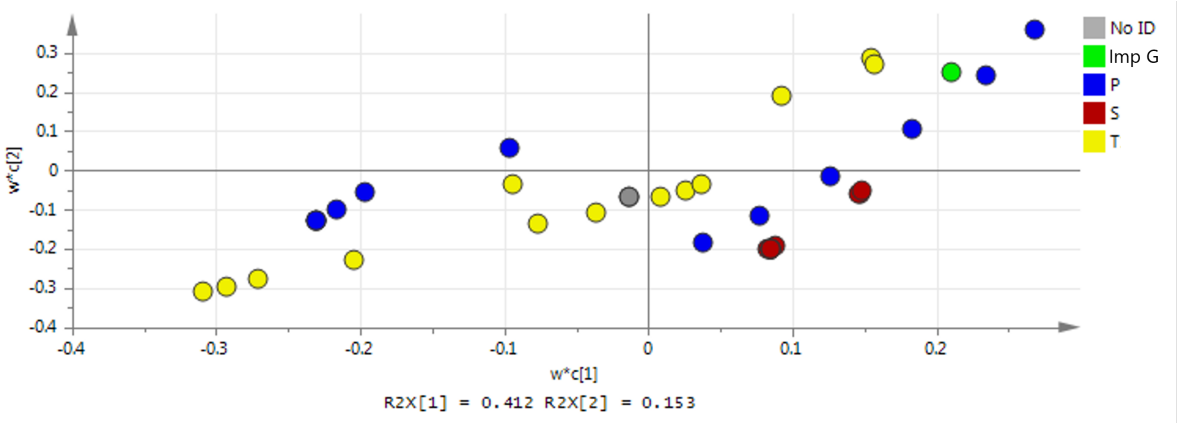


Figure 3.32: Loadings of the first component versus the second component for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the charge of the first salt during the reaction step of intermediary 4 process.

The response variable is explained by both components in a similar way, given its projections on the y and x-axis. Consequently, for the complete analysis of this operation, both components will be plotted against batch maturity.

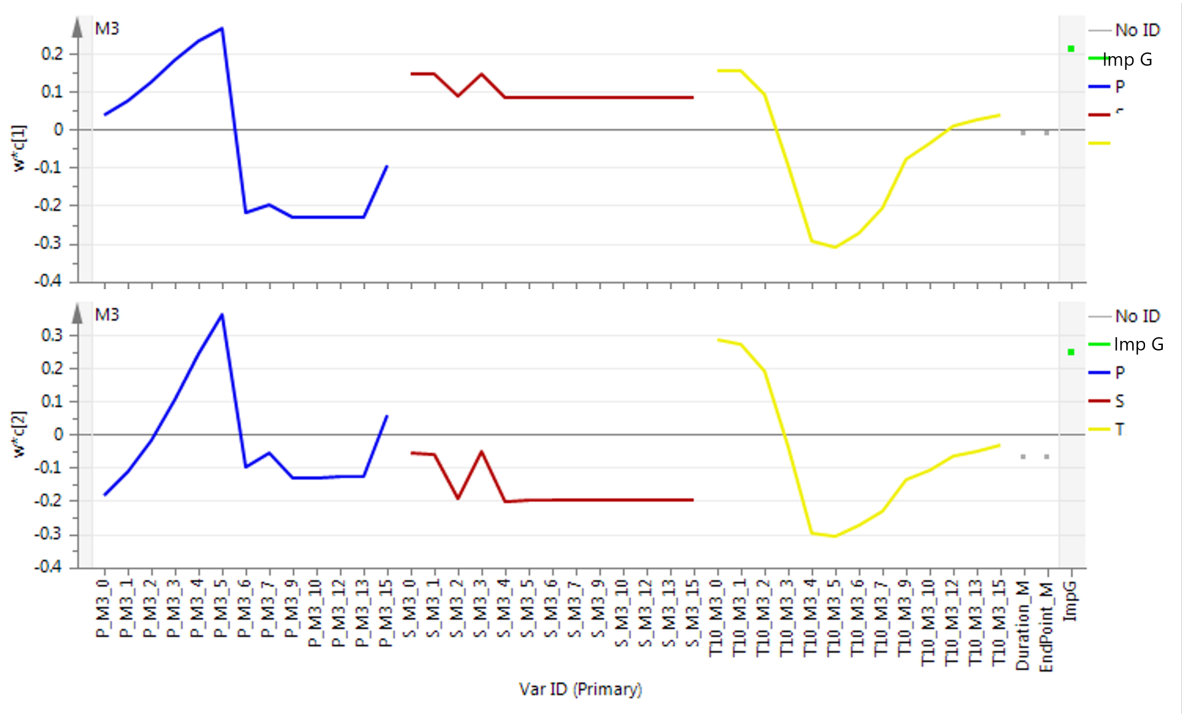


Figure 3.33: Loadings of the first component (top) and the second component (bottom) given against batch maturity for each variable (pressure as blue, temperature as yellow, and agitator speed as red) for the charge of the first salt during the reaction step of intermediary 4 process.

The agitator speed contribution to the response variable differs from the first component to the second component. Looking at figure 3.32, no concrete conclusion can be taken concerning this variable due to the low relevance to the response variable according to each individual component and also due to different contributions (positive for the first component and negative for the second component). A closer look at the contribution of temperature on each component reveals a similar pattern. Although positive in the beginning, the impact of temperature abruptly becomes negative according to both components and stays negative (on some parts of the operation more than others) during the rest of the operation. As such, lower values of temperature favor the increase in impurity G content, which decreases yield in FP. It is important to mention that this relation is the same for impurity H (figure 3.28), adding up robustness to the improvement action taken from the models: to stay at higher temperatures in order to decrease both impurities content in intermediary 4.

3.4 Improve

A systematization of all the statistical conclusions took during the previous DMAIC cycle phase and a translation of those conclusions into tangible and feasible measures are the main goals of this fourth stage of the process improvement project.

3.4.1 Statistical Analysis Summary

3.4.1.1 FP process step

The analysis conducted to FP process had as response variable the yield of the final production step.

From process photograph analysis, it was concluded that higher time spent on the addition of the anti-solvent during crystallization, favored higher yields, as seen in figure 3.11. The level profile is mandatory to analyse in this case since differences in the addition pattern (which are unknown) are also crucial to the response variable being considered. The improvement action that can be taken out from this analysis is that a level sensor should be integrated into the reactor where FP crystallization takes place.

The filtration step was found to have random profiles for both pressure and agitator speed. As previously stated, this operation is hard to standardize in an operational point of view. Before any optimization can take place, an effort for standardization should be put in place.

Under BLM analysis, during the antisolvent addition in the crystallization step, an higher heating rate in the first half of the operation leads to higher yields, as seen on the loadings given for the first component, in figure 3.15. It is also noticeable that higher final temperatures lead to lower yields. However, the targeted final temperature during this step corresponds to the process reflux temperature, to which the solvent continues to evaporate and therefore cannot be changed. For the subsequent cooling step in the crystallization, a lower cooling rate (but still under the process indication of *around* CONFIDENTIAL) favors higher yields (figure 3.16). Also represented in the loadings of this operation, higher final temperatures also favor higher yields. The process interval for the final temperature in this operation is between CONFIDENTIAL and CONFIDENTIAL with the indication to target CONFIDENTIAL. This indication could be increased, but still within the process limits, in order to increase yield. The filtration BLM model has two components with despairing contributions of temperature for the response variable (figure 3.18): the first component shows a positive relationship and the second shows a changing but mainly weak negative relation. Considering that the change in temperature inside the filter is due to the transfer of the washings at a low temperature and attending only to the relation shown in the first component an improvement action to be proposed could be to transfer the washings during filtration closer to the upper end of the stipulated process interval. All these actions were found, as already exposed, to affect product quality in the same way, that is to say, yield and quality go hand in hand.

3.4.1.2 Intermediary 4 process step

Impurity H and impurity G were the two response variables considered for intermediary 4 process step statistical analysis. It is important to mention that all the analysis conducted to intermediary 4 production step were carried with a small number of observations leading to an inevitable loss in terms of robustness.

No set of relevant process photograph variables were found to correlate to impurity G with meaningful statistical parameters ($R2$ and $Q2$). For impurity H, operations where the reactor‘s content can be in contact with the atmosphere and the degassing operations, were regarded as very explanatory as seen in the model statistics presented in table 3.14. More time spent on degassing operations and less time

spent on the charge of the reactants will strongly diminish impurity H's content in intermediary 4, as shown in the model coefficients in figure 3.25.

Once again, the filtration presents itself as a highly variable operation regarding pressure and agitator speed and standardization actions should be put in place.

Through BLM analysis it was found that, taking impurity H as the response variable, the agitator speed during all the reaction steps (loading of reactants and reaction itself) should be set as 90 rpm. During the inorganic salts load, higher temperatures favor low impurities content (both H and G) in the output, as seen in figures 3.28, 3.29 and 3.32. There is the indication to preferably stay at the lower limit of the process temperature interval (CONFIDENTIAL$<$T(ºC)$<$CONFIDENTIAL) which, considering just these two impurities as response variables, should be changed in order to target the higher end of the interval. The reaction itself was also modeled and it was found that higher pressures led to lower impurity H content (figure 3.31). The main reactant load is done through an auxiliary gas cylinder which is heated in order to keep its content in the gaseous state. A straightforward way to increase pressure in the reactant addition would be to increase the temperature of the gas cylinder. However, this measure could attach countless safety-related additional risks that were not studied. The implementation of this improvement action was left for the production team to carefully analyse.

3.4.2 Prioritization

After the statistical analysis results translation into concrete improvement actions, a prioritization method is imperative to apply in order to sort which actions will be tackled first and the number of resources needed. This classification is based on two parameters: impact and effort. The classification based on the former is empirical, based on process knowledge that the production team has been gathering through the batches, and is done on a scale of 1 to 10. The classification based on the latter is performed according to statistical parameters ($R2$ and $Q2$) that are assigned by performing a statistical model considering just the variable of interest. The values obtained for $R2$ and $Q2$ are summed in order for the impact classification to reflect both the fitting of the model but also its predictive ability.

Table 3.19: Improvement actions summary and classification in terms of their impact and necessary implementation effort.

Action ID	Process step	Response variable	Operation	Description	R2	Q2	Impact (R2 + Q2)	Effort
1	FP	Yield	Crystallization	Faster heating in the beginning of antisolvent addition	0.564	0.165	0.729	3
2	FP	Yield	Crystallization	Cool the suspension at a slightly lower rate than CONFIDENTIAL	0.458	0.405	0.863	4
3	FP	Yield	Crystallization	Increase the final temperature target in the cooling towards higher end of process interval	0.458	0.405	0.863	4
4	FP	Yield	Filtration	Transfer the washings closer to higher end of process interval	0.424	0.275	0.699	4
5	Intermediary 4	Imp H	Degassing	Take more time on the first degassing step	0.165	0	0.165	3
6	Intermediary 4	Imp H	Degassing	Take more time on the second degassing step	0.389	0	0.389	3
7	Intermediary 4	Imp H	Degassing	Agitator speed set as CONFIDENTIAL	0.779	0.585	1.364	1
8	Intermediary 4	Imp H	Reactants load	Take the least amount of time on the first salt load	0.259	0	0.259	5
9	Intermediary 4	Imp H	Reactants load	Take the least amount of time on the second salt load	0.744	0.194	0.938	5
10	Intermediary 4	Imp H	Reactants load	Agitator speed set as CONFIDENTIAL	0.759	0.553	1.312	1
11	Intermediary 4	Imp H	Reactants load	During inorganic salts load stay closer to the higher end of temperature process interval	0.467	-0.100	0.367	4
12	Intermediary 4	Imp G	Reactants load	During inorganic salts load stay closer to the higher end of temperature process interval	0.716	0.105	0.821	4
13	Intermediary 4	Imp H	Reaction	Agitator speed set as CONFIDENTIAL	0.704	0.46	1.164	1
14	Intermediary 4	Imp H	Reaction	Increase pressure at the early stages of reaction	0.531	0.211	0.742	8

Based on the impact and effort classification presented in table 3.19, the improvement actions are now placed on an Impact Vs. Effort matrix, divided into four quadrants that guide the prioritization process.

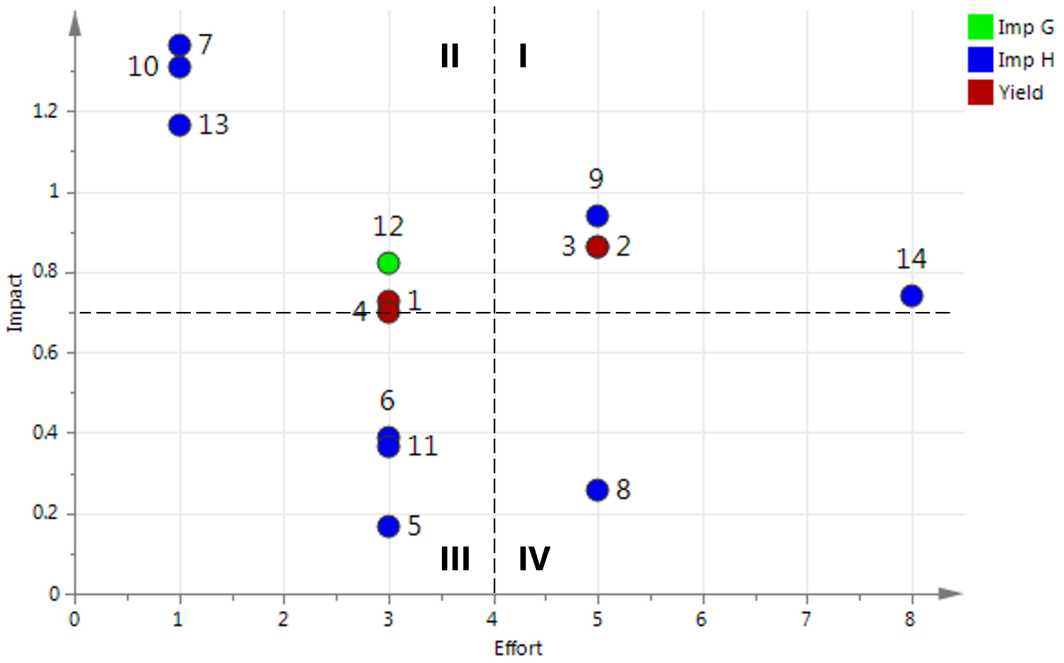


Figure 3.34: Impact Vs. Effort matrix with the improvement actions identified by their ID defined on table 3.19 and colored by response variable.

The top-left quadrant (second quadrant) comprises actions with high impact and low effort. Actions 7, 10, and 13 are placed in this quadrant with the highest impact and the lowest effort and aim at the reduction of impurity H content in intermediary 4 through the following of the indication to stay at CONFIDENTIAL during all pre-reaction and reaction steps. Also on this quadrant is the only improvement action that focuses on reducing impurity G content by staying at the higher end of the process temperature interval during the inorganic salts' loading. Considering the increase in yield of FP, two more actions are placed on the second quadrant: action 1 which relates to faster heating up until the middle of antisolvent addition, and action 4 which relates to the transfer of the washings (during filtration) closer to the higher end of the process temperature interval. All these actions that fall in the second quadrant of the matrix can be coined "Quick Wins" and are the ones that should be tackled first because they yield the best return based on the effort.

Actions placed on the first quadrant require a high effort but have also a higher impact. This group of actions can be coined "Major Projects". Included in this group are the two actions for the yield of FP optimization concerning the cooling step of the crystallization, and also two actions for the decrease in impurity H content on intermediary 4: action 9 that is related to faster loading of the second salt during the pre-reaction steps and action 14 that is related to the increase in pressure during the early stages of the reaction.

The third quadrant comprises actions that require little effort but also provide fewer returns and can be called "Incremental Actions" because these are actions to pursue whenever there are spare resources.

All the actions in this group are meant to reduce impurity H´s content through either more time on the degassing steps or by staying closer to the higher end of the process temperature interval during the inorganic salts´ load.

Finally, there are the ”Money Pits”, improvement actions that require a lot of effort and do not give a substantial amount of return.

3.5 Control

The final phase of the DMAIC process improvement cycle has one major goal: to sustain the improvements. There are several ways of achieving this goal, however, in this project, only control charts and a summary flowchart will be used.

3.5.1 Flowchart

After an extensive analysis of the primary problem (high variability on the FP process yield) with the root causes being uncovered, it is necessary to implement frameworks within the organization to make sure that the process understanding gained will not be lost. In this line, an internal KPI value for the FP process yield will be established by the production team. Whenever a FP batch has a lower than the KPI yield, an internal investigation is triggered. In order to systematize all the knowledge created during the project and to serve as an aiding tool for the internal investigations, a flowchart was elaborated and is presented in annex C.

3.5.2 Control Charts

One of the most common tools to apply in the final phase of DMAIC cycles are the control charts that are mainly used for process monitoring.

The flowchart for the internal investigations is a great tool for guiding the process of analysing concluded batches and spot what was the cause for a deviation that has already happened. Instead of acting on the problem, a preventive tool like the control charts is essential to actively monitor the process and sustain the improvements in real time. In this way, an interactive Excel file was created, to be filled by the operators with quick and easy to obtain process data (like temperatures displayed and starting and ending times for the operations). Control charts were created and are updated automatically, as the data is filled in. Bellow, an example of one of these control charts is presented. It is important to mention that Dr. Shewart suggested that in order to build a control chart, *i.e.*, to establish the control limits, 25 samples should be used [32]. However, in the present case, only 16 samples (or observations or batches) will be used since the data was only available over the considered period of time.

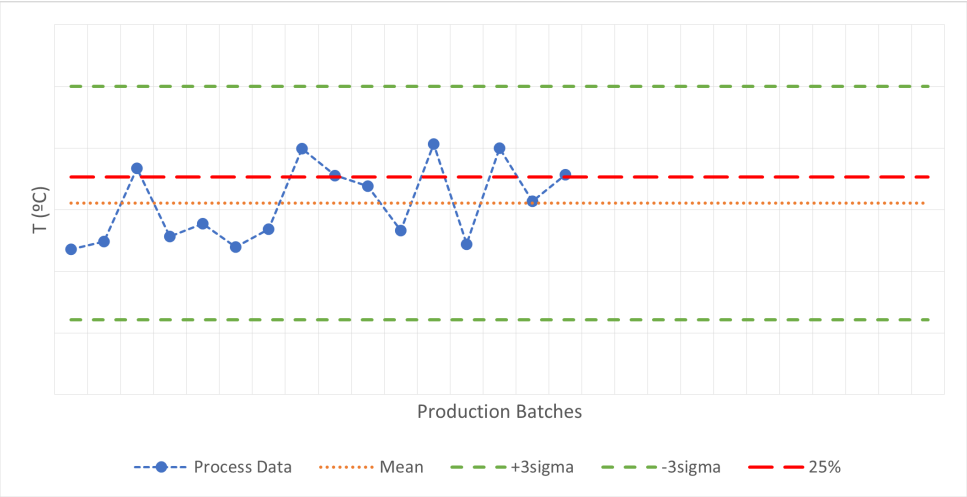


Figure 3.35: Control chart of the final temperature in the cooling step of the crystallization of FP. Above the red dashed line lies 25% of the data assuming normal distribution.

The example shown in figure 3.35 shows the final temperature in the cooling step of the crystallization during FP process. The 3σ limits are displayed as well as the process mean. In this specific operation, it was found through the statistical analysis performed, that higher final temperatures would lead to higher yields (recall figure 3.16; action number 3 in table 3.19). Consequently, a red line was drawn in the chart that marks the 25% highest point values. In order to achieve higher yields (and still within the process temperature interval, CONFIDENTIAL<T(ºC)<CONFIDENTIAL), final temperatures for this operation should be systematically above the 25% mark. It is important to mention that the choice of the percentage was empirical and can be changed when deemed convenient by the production team.

Chapter 4

Conclusions

Framed on Hovione´s Sete Casas production site continuous improvement plan, the present work aimed at the yield optimization of a production process of a generic corticosteroid API, fluticasone propionate.

Six Sigma´s DMAIC process improvement and the problem-solving cycle was the chosen methodology to approach the problem. The project was divided into five separated and clearly defined phases: define; measure; analyse; improve and control, which have proven to succeed on continuous improvement projects. Combined with the DMAIC cycle, a backwards methodology was applied: after identifying the problem in the process output (the high variability on FP process yield), the root causes will be uncovered going in reverse through the API production train. In this way, a channeling of the efforts to the process steps where the principal causes of the problem are is performed.

Ranging from July 2018 to January 2021, 40 production batches were included in the analysis. Over this period, yield took an average value of 81.83% with a standard deviation of 2.22%. The minimum and maximum values were 77.30% and 86.30% which is 11% of the variable mean. High yield variability leads to biased and uncertain production planning (especially in a multi-purpose installation as is the case) and poor use of company resources since the batch costs do not change with the throughput obtained. The parameter used to quantify the problem was the missed opportunities per year (considering that 17 batches of FP are produced per year and the product is sold for CONFIDENTIAL per gram) if during the analysed timeline all batches had a yield equal to the chosen optimization set-point. If all batches had a yield equal to 86.83%, the losses to yield variability on a yearly basis account roughly as one production batch which is translated into CONFIDENTIAL million dollars.

Process mapping was followed in order to further increase process familiarization. The two types of batch data were collected: process photograph and process film types of data.

Only the two last steps on the API production train were statistically analysed, the step leading to FP and to intermediary 4. Starting off with the FP step, it was found that the input quality data explained 86.3% of yield variability, being impurity H firstly and impurity G secondly, the ones with the most negative impact. Nevertheless, given the high percentage of variability explained by the input quality data, the process variables were also analysed. It was found that the antisolvent addition step could have a big

impact on yield and so, a level sensor should be installed on the crystallization reactor. It was also found that both pressure and agitator speed, the controlled variables on the filtration process, presented random profiles and so, before any optimization action, a standardization effort should be put in place. Under BLM analysis (where only 16 production batches were considered), a heating rate during the first half of antisolvent addition was found to favor yield as well low agitator speeds. These conclusions are supported by a model that explains 64.7% of yield variability. On the subsequent crystallization step, the cooling step, it was noticed that lower cooling rates (but still *around* CONFIDENTIAL) and higher final temperatures (but within the process interval, CONFIDENTIAL<T(ºC)<CONFIDENTIAL) gave higher yields. On the filtration step, temperature, which is not manipulated during the operation, correlated positively with yield: The washings transfer could be done closer to the higher end of the process interval in order to stay at higher temperatures inside the equipment. All the relevant models were re-built considering the final product´s assay as the response variable instead of yield. It was found that the variables' contributions were not different from those considering yield as the response variable and so, in the case of FP process step, quality performance and throughout performance go hand in hand.

The analysis on the intermediary 4 process step had as response variables impurity H and G. For impurity H, which is formed whenever an oxidizing agent is present, the time of the degassing operations was found to be negatively correlated (higher time on these operations led to lower impurity content) and the time for the reactants load was found to be positively correlated (higher time on these operations led to higher impurity content). The two components model built for this set of variables explains 98.1% of impurity H´s variability on the analysed batches. As was the case for FP filtration step, on intermediary 4 both pressure and agitator speed presented random profiles. On the BLM models constructed for this process step, data was only available for 6 production batches. For both impurity H and G, during the load of the inorganic salts, the temperature was found to be negatively correlated. This being said, in order to reduce impurity H content in intermediary 4, during these steps, the temperature should be closer to the higher end of the process interval (CONFIDENTIAL<T(ºC)<CONFIDENTIAL). These models explain a big part of the response variable variability (98.4% for the first salt load and 97.9% for the second). For impurity G, the temperature contribution during these operations was no different. During the beginning of the reaction with BMF, the pressure was discovered to be negatively correlated with impurity H given the two components model that explains 95.2% of the response variable variability. Pressure could be increased by the increase of the reactant´s gas cylinder temperature. Various health and safety risks could rise up and were not studied. Still, regarding impurity H as the response variable, the agitator speed was found to be, through all pre-reaction and reaction steps, positively correlated. The process indication to stay at CONFIDENTIAL should be followed.

After concluding the statistical analysis phase of the project, an Impact Vs. Effort matrix was built for the improvement actions taken. The actions with low effort and high impact, termed "Quick Wins" are related to the agitator speed during the pre-reaction and reaction step regarding impurity H as response variable; to stay at the higher end of the process temperature interval during the load of inorganic salts regarding impurity G as response variable; to transfer the washings, during FP filtration, closer to the

higher end of the process temperature interval and to promote faster heating in the first half of antisolvent addition, during the FP crystallization regarding both yield as response variables.

An interval KPI minimum yield value will be established by the production team and whenever a batch performs under the established KPI, an internal investigation will take place, based on the process understanding generated. As such, an aiding flowchart was constructed to help out the team in these investigations and an interactive Excel file with univariate control charts of relevant operations.

For future work, in order to increase the robustness of intermediary 4 models, more production batches should be incorporated since data was only available for 6. Taking a more holistic approach, this project should serve as the basis for a demystification of MVDA applied to chemical synthesis pharmaceutical processes. These processes are very complex and the black-box statistical approach is the one to take for process improvement.

Bibliography

[1] I. C. on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH Q8(R2) - Pharmaceutical Development*. 2009.

[2] J. Woodcock. The Concept of Pharmaceutical Quality. *American Pharmaceutical Review*, 2004.

[3] L. X. Yu. Pharmaceutical quality by design: Product and process development, understanding, and control. *Pharmaceutical Research*, 25(4):781–791, 2008.

[4] Department of Health, Human Services U.S Food, and Drug Administration. *Pharmaceutical cGMPs for the 21st Century - a risk-based approach*. 2004.

[5] L. Zhang and S. Mao. Application of quality by design in the current drug development. *Asian Journal of Pharmaceutical Sciences*, 2016.

[6] A. Ferreira, J. C. Menezes, and M. Tobyn. *Multivariate Analysis in the Pharmaceutical Industry*. Elsevier, 2018.

[7] L. Abboud and S. HensleyStaff. New Prescription For Drug Makers: Update the Plants. *The Wall Street Journal*. Consulted on 23/04/2021.

[8] L. X. Yu, G. Amidon, M. A. Khan, S. W. Hoag, J. Polli, G. K. Raju, and J. Woodcock. Understanding Pharmaceutical Quality by Design. *The AAPS Journal*, 16(4), 2014.

[9] D. J. Am Ende and M. T. Am Ende. *Chemical Engineering in the Pharmaceutical Industry: Active Pharmaceutical Ingredients*. Wiley, 2nd edition, 2019.

[10] A. P. Ferreira and M. Tobyn. Multivariate analysis in the pharmaceutical industry: Enabling process understanding and improvement in the PAT and QbD era. *Pharmaceutical Development and Technology*, 20(5):513–527, 2015.

[11] T. Kourti. Quality by design in the pharmaceutical industry: Process modelling, monitoring and control using latent variable methods. 7:36–41, 2009.

[12] B. Scott and A. Wilcock. Process analytical technology in the pharmaceutical industry: A toolkit for continuous improvement. *PDA Journal of Pharmaceutical Science and Technology*, 60(1):17–53, 2006.

[13] J. Woodcock. Reliable drug quality: An unresolved problem. *PDA Journal of Pharmaceutical Science and Technology*, 66(3):270–272, 2012.

[14] J. A. Garza-Reyes, I. E. Betsis, V. Kumar, and M. A. Radwan Al-Shboul. Lean readiness – the case of the European pharmaceutical manufacturing industry. *International Journal of Productivity and Performance Management*, 67(1):20–44, 2018.

[15] B. Chatterjee. *Applying Lean Six Sigma in the Pharmaceutical Industry*. Routledge, 2014.

[16] D. Bellm. Operational excellence in the pharmaceutical industry – an architecture for emerging markets. Master's thesis, University of St.Gallen, 2015.

[17] H. Gebauer, M. Kickuth, and T. Friedli. Lean management practices in the pharmaceutical industry. *International Journal of Services and Operations Management*, 5(4):463–481, 2009.

[18] T. Friedli and D. Bellm. Pharmaceutical OPEX - The Next Generation. *Life Science Leader Magazine*, 2013.

[19] T. Friedli, P. Basu, D. Bellm, and J. Werani. *Leading Pharmaceutical Operational Excellence*. Springer, 2013.

[20] L. X. Yu and M. Kopcha. The future of pharmaceutical quality and the path to get there. *International Journal of Pharmaceutics*, 528(1-2):354–359, 2017.

[21] Comunidade Científica — Hovione. `https://www.hovione.pt/inovacao-e-qualidade/inovacao-na-hovione/comunidade-cientifica`. Consulted on 23/08/2021.

[22] S. Nussey and S. Whitehead. *Endocrinology: An Integrated Approach*. BIOS Scientific, 1st edition, 2001.

[23] W. Ericson-Neilsen and A. D. Kaye. Steroids : Pharmacology , Complications , and Practice Delivery Issues. *Oschner Journal*, 14(2):203–207, 2014.

[24] H. Schacke, W.-D. Docke, and K. Asadullah. Mechanisms involved in the side effects of glucocorticoids. *Pharmocology and Therapeutics*, 96:23–43, 2002.

[25] K. Remien and A. Bowman. *StatPearls: Fluticasone*. 2021.

[26] Fluticasone Propionate Cream WebMD. `https://www.webmd.com/drugs/2/drug-8786/fluticasone-propionate-topical/details`, . Consulted on 01/04/2021.

[27] Fluticasone Propionate Nasal WebMD. `https://www.webmd.com/drugs/2/drug-77986-245/fluticasone-propionate-nasal/fluticasone-spray-nasal/details`, . Consulted on 01/04/2021.

[28] J. Fischer and C. R. Ganellin. *Analogue-based Drug Discovery*. Wiley, 1st edition, 2006.

[29] N. Midoux, P. Hošek, L. Pailleres, and J. R. Authelin. Micronization of pharmaceutical substances in a spiral jet mill. *Powder Technology*, 104(2):113–120, 1999.

[30] P. Khadka, J. Ro, H. Kim, I. Kim, J. T. Kim, H. Kim, J. M. Cho, G. Yun, and J. Lee. Pharmaceutical particle technologies: An approach to improve drug solubility, dissolution and bioavailability. *Asian Journal of Pharmaceutical Sciences*, 9(6):304–316, 2014.

[31] N. Esfandiari and S. M. Ghoreishi. Ampicillin Nanoparticles Production via Supercritical CO2 Gas Antisolvent Process. *AAPS PharmSciTech*, 16(6):1263–1269, 2015.

[32] K. Krishnamoorthi and V. R. Krishnamoorthi. *A First Course in Quality Engineering: Integrating Statistical and Management Methods of Quality*. CRC, 3rd edition, 2011.

[33] B. K. Nunnally and J. S. McConnell. *Six Sigma in the Pharmaceutical Industry: Understanding, Reducing, and Controlling Variation in Pharmaceuticals and Biologics*. CRC, 1st edition, 2007.

[34] J. Antony, R. Snee, and R. Hoerl. Lean Six Sigma: yesterday, today and tomorrow. *International Journal of Quality and Reliability Management*, 34(7):1073–1093, 2017.

[35] P. S. Pande, R. P. Neuman, and R. R. Cavanagh. *The Six Sigma Way: How GE, Motorola, and Other Top Companies are Honing Their Performance*. McGraw-Hill, 1st edition, 2000.

[36] D. Kiran. *Quality Loss Function*. Elsevier, 2017.

[37] M. S. Boorla, T. Eifler, C. McMahon, and T. J. Howard. Product robustness philosophy – A strategy towards zero variation manufacturing (ZVM). *Management and Production Engineering Review*, 9 (2):3–12, 2018.

[38] G. Nyrén. A Six Sigma project at Ericsson Network Technologies. Master's thesis, Luleå University of Technology, 2007.

[39] T. M. Kubiak and D. W. Benbow. *The Certified Six Sigma Blackbelt Handbook*. ASQ Quality Press, 2nd edition, 2009.

[40] H. C. Ott. Six Sigma and Lean: Quantitative Tools for Quality and Productivity - Yellow Belt Certification handouts. *Technical University of Munich - edX Platform*, 2020.

[41] G. P. Agre. The Concept of Problem. *Educational Studies: A Journal of the American Educational Studies Association*, pages 121–142, 1982.

[42] M. Otto. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*. Wiley, 3rd edition, 2016.

[43] I. C. on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH Q11 - Development and Manufacture of Drug Substances* . 2012.

[44] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC - Trends in Analytical Chemistry*, 132, 2020.

[45] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. Buydens. Breaking with trends in pre-processing? *TrAC - Trends in Analytical Chemistry*, 50:96–106, 2013.

[46] R. A. V. D. Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. V. D. Werf. Centering , scaling , and transformations : improving the biological information content of metabolomics data. *BMC Genomics*, 15:1–15, 2006.

[47] Umetrics. Multivariate Data Analysis and Modelling Basic Course. 2002.

[48] X. Ying. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168 (2), 2019.

[49] H. F. Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151, 1960.

[50] T. Næs, T. Isaksson, T. Fearn, and T. Davies. *A user-friendly guide to Multivariate Calibration and Classification*. NIR Publications, 2nd edition, 2017.

[51] S. Wold and M. Sjostrom. PLS-regression : a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130, 2001.

[52] H. G. Gauch, J. T. Hwang, and G. W. Fick. Model Evaluation by Comparison of Model-Based Predictions and Measured Values. *Agronomy Journal*, 95(6):1442–1446, 2003.

[53] P. P. Roy and K. Roy. On some aspects of variable selection for partial least squares regression models. *QSAR and Combinatorial Science*, 27(3):302–313, 2008.

[54] F. Voehl, H. J. Harrington, C. Mignosa, and R. Charron. *The Lean Six Sigma Black Belt Handbook: Tools and Methods for Process Acceleration*. 2013.

[55] W. A. Shewhart. *Economic Control of Quality of Manufactured Product*. Wiley, 1933.

[56] Drug Recalls-FDA. `https://www.fda.gov/drugs/drug-safety-and-availability/drug-recalls`. Consulted on 20/06/2021.

[57] K. Muteki, V. Swaminathan, S. S. Sekulic, and G. L. Reid. De-risking pharmaceutical tablet manufacture through process understanding, latent variable modeling, and optimization technologies. *AAPS PharmSciTech*, 12(4):1324–1334, 2011.

[58] M. Machin, L. Liesum, and A. Peinado. Implementation of modelling approaches in the QbD framework: examples from the Novartis experience. *European Pharmaceutical Review*, 16, 2011.

[59] Y. Cui, X. Song, K. Chuang, C. Venkatramani, S. Lee, G. Gallegos, T. Venkateshwaran, and M. Xie. Variable selection in multivariate modeling of drug product formula and manufacturing process. *Pharmaceutical Technolgy*, 101, 2012.

[60] S. Estrada-Flores, I. Merts, B. De Ketelaere, and J. Lammertyn. Development and validation of ”grey-box” models for refrigeration applications: A review of key concepts. *International Journal of Refrigeration*, 29(6):931–946, 2006.

[61] D. Armbruster and T. Pry. Limit of Blank, Limit of Detection and Limit of Quantitation. *The Clinical Biochemist Reviews*, 29:49–52, 2008.

[62] H. H. Tung, E. L. Paul, M. Midler, and J. A. McCauley. *Crystallization of Organic Compounds: An Industrial Perspective*. Wiley, 2008.

[63] W. Hewitt. *Microbiological assay for pharmaceutical analysis: a rational approach. Boca Raton*. CRC, 2003.

Appendix A

Variable Profiles

CONFIDENTIAL

Appendix B

Input-Process-Output Diagrams

CONFIDENTIAL

Appendix C

Internal Investigation Flowchart

CONFIDENTIAL