Optimization of a Membrane Production Process according to PAT

Rafael Alexandre Mota Batista Santos

Thesis to obtain the Master of Science Degree in

Biological Engineering

Supervisors: Prof. Dr. José Monteiro Cardoso de Menezes; Dr. Marek Hoehse

Examination Committee

Chairperson: Prof. Dr. Duarte Miguel de França Teixeira dos Prazeres

Supervisor: Prof. Dr. José Monteiro Cardoso de Menezes

Member of the Committee: Prof. Dr. Helena Maria Rodrigues Vasconcelos Pinheiro

December 2014
Acknowledgments

The journey in which resulted this master dissertation was a fantastic opportunity that will hardly ever be exceeded. More than just for the opportunity, I have to thank Prof. José Cardozo Menezes for its coordination and advice throughout this all of these months.

Then I have to express my gratitude to Dr. Christian Grimm for the warm welcome in his department. I have to thank Dr. Marek Hoehse for the guidance, the support and all time invested in me. And of course I have to thank the other Marek, the one that threw me a birthday party, the one that helped me for everything and make it all possible, the friend. I also have to thank José, Patricia, Matthias and all the department of Chemometrics, for making this portuguese very comfortable in Germany.

I also have to thank Bernardo, João Fernandes and all the people that help me, not just in the master thesis, but during all my passage through IST. And of course I cannot go without thank my dearest friend Rita not just for listening and caring, but for all the dreams, the laughs, the kisses, the cries, our rights, our wrongs and all her sweet disposition, that always keeps me in the good mood.

Não posso deixar de agradecer ao meu irmão Bruno não só pela ajuda na edição da presente tese, mas por ser o meu best buddy de sempre. Além de toda a minha família, para terminar, tenho de agradecer aos meus pais, Boaventura e Ana Cristina. Nada seria possível sem eles e assim como nada teria o mesmo gosto. Por todos os momentos, experiencias e amor incondicional, um obrigado.
Abstract

The quality control of a polyethersulfone membrane production process is a time consuming procedure that involves the repeated measurement of its critical quality attributes (CQA) until they fit within specification. To optimize such delays, it has been proven that the combination of Process Analytical Technology (PAT) tools with near infra-red spectrometry can monitor the stability of critical process parameters (CPP) in a production line and even predict CQAs.

In the present thesis, it is proposed to independently apply this principle to membrane 15458, 15457, 15404-H and 15407. This study aims to evaluate the relationships between selected CPPs and membrane’s CQAs, using design of experiments, and to build predictive models for each membrane. Through multivariate data analysis, these models were obtained by merging the NIR spectral data and the CQAs for each membrane.

This work showed a good correlation of the CQAs: bubble point (BP), thickness and air flux, with the most important CPPs. The fraction for the variation of the most important CQA, BP, was between 0.79 – 0.98 with a prediction ability between 0.64 – 0.9. For the model building, the standard error of prediction of all BP models appears to be within the same range (0.05 – 0.16 bar) and lower than half the standard deviation of the in-specification range. This study proves that this optimization can be expanded to different kinds of membranes, which contributes for the improvement of the monitoring of this type of manufacturing processes.

Keywords: Process analytical technology; monitoring; Design of Experiments; Near-Infrared Spectroscopy; membrane production;
Resumo

O controlo de qualidade de um processo de produção de membrana de polietersulfona é moroso e envolve a medição repetida dos atributos críticos de qualidade (ACQ) até que estes se encaixem dentro da especificação. Para optimizar estes atrasos, provou-se que a combinação de ferramentas de tecnologia de processo analítico com espectrometria do quase-infravermelho pode monitorar a estabilidade dos parâmetros críticos do processo (PCC) numa linha de produção e, mesmo prever ACQs.

Na presente tese, propõe-se aplicar este princípio de forma independente para as membranas 15458, 15457, 15404 e 15407. Este estudo tem como objectivo avaliar as relações entre os PCCs selecionados e os ACQs das membranas, utilizando planeamento de experiencias, e construir modelos préditivos para cada membrana. Através da análise multivariada de dados, estes modelos foram obtidos por fusão dos dados espectrais do quase-infravermelho e os ACQs para cada membrana.

Este trabalho mostrou uma boa correlação dos ACQs: ponto de bolha (PB), espessura e fluxo de ar, com os PCCs mais importantes. A percentagem para a variação do ACQ mais importantes, PB, foi entre 0,79 - 0,98, com uma capacidade de previsão entre 0,64 - 0,9. Para a construção do modelo, o erro padrão de previsão de todos os modelosPB parece estar no mesmo intervalo (0,05 - 0,16 bar) e menor do que metade do desvio padrão do intervalo de especificação. Este estudo prova que esta optimização pode ser expandida para diferentes tipos de membranas, o que contribui para a melhoria da monitorização deste tipo de processos de fabrico.
Index

1. Introduction ........................................................................................................................................... 1
   1.1. Motivation ......................................................................................................................................... 2
   1.2. Thesis Objective ................................................................................................................................. 4
   1.3. State of Research ............................................................................................................................... 5
   1.4. Thesis outline ..................................................................................................................................... 6
2. Theoretical Background .......................................................................................................................... 7
   2.1. Process Analytical Technologies (PAT) and Chemometrics ............................................................. 7
   2.2. Design of Experiments (DoE) ........................................................................................................... 10
      2.2.1. Introduction ................................................................................................................................. 10
      2.2.2. Implementation Procedure ......................................................................................................... 12
   2.3. Near-Infrared Spectroscopy .............................................................................................................. 14
      2.3.1. Fundamentals .............................................................................................................................. 14
      2.3.2. Instrumentation .......................................................................................................................... 17
      2.3.1. Applications .................................................................................................................................. 18
   2.4. Multivariate Analysis ......................................................................................................................... 19
      2.4.1. Data Pre-Treatment ...................................................................................................................... 19
      2.4.2. Models ......................................................................................................................................... 22
      2.4.3. Model Validation .......................................................................................................................... 24
3. Membrane technology and Production Process ......................................................................................... 27
   3.1. Membrane Types and Processes ....................................................................................................... 27
   3.2. Membrane Production Process ...................................................................................................... 28
   3.3. Critical Quality Attributes .............................................................................................................. 29
   3.4. Critical Process Parameters ............................................................................................................ 31
   3.5. Reference Analysis ............................................................................................................................ 34
   3.6. NIR Spectrometer ............................................................................................................................. 35
4. Experimental Procedure ......................................................................................................................... 37
4.1. Project Framework ................................................................................................................. 37
4.2. RA and CPP definition ........................................................................................................... 38
4.3. DoE Building ......................................................................................................................... 39
4.4. DoE Rearranging .................................................................................................................. 40
4.5. Membrane Referencing ........................................................................................................ 41
5. Experimental Results and Data Analysis ............................................................................... 43
  5.1. Sampling and Datasets ......................................................................................................... 43
  5.2. Evaluation of CPP-CQA relationships .................................................................................. 43
    5.2.1. Statistical parameters .................................................................................................... 43
    5.2.2. Membrane 15458 ........................................................................................................... 44
    5.2.3. Membrane 15457 ........................................................................................................... 46
    5.2.4. Membrane 15404-H ....................................................................................................... 49
    5.2.5. Membrane 15407 ........................................................................................................... 50
    5.2.6. Analysis considerations ................................................................................................. 52
  5.3. Model Building ................................................................................................................... 53
    5.3.1. General Aspects ............................................................................................................. 53
    5.3.2. Membrane 15458 ........................................................................................................... 55
    5.3.3. Membrane 15457 ........................................................................................................... 59
    5.3.4. Membrane 15404-H ....................................................................................................... 63
    5.3.1. Membrane 15407 ........................................................................................................... 67
6. Conclusions .............................................................................................................................. 72
7. Future work .............................................................................................................................. 76
8. Bibliographic References .......................................................................................................... 77
Appendix A. Instrumentation ......................................................................................................... 80
Appendix B. Statistical Parameters ............................................................................................... 82
  B.I. R² – Correlation Factor ......................................................................................................... 82
  B.II. Q2 – Predictive Factor ......................................................................................................... 82
  B.III. MV – Model Validity .......................................................................................................... 82
  B.IV. Rep – Reproducibility ........................................................................................................ 82
Table Index

Table 1 – Membrane CPPs used for DoE building for each membrane.................................................................39
Table 2 – Time changes /unit for all the CPPs used in the DoEs for the 4 membranes ...........................................40
Table 3 – Standard evaluation values for the parameters R², Q², MV and Rep according to [22].................................44
Table 4 – Statistic evaluation of each model. Nexp is the number of experiments. Green marked values meet the specifications of a good model exposed in Table 3 (R² ≥ 0.80). Red marked values do not meet such specifications. ........................................................................................................................................53
Table 5 – In-specification limits used at the membrane production site to evaluate the process performance and their SD and SD/2 ........................................................................................................................................75
Table 6 – Summary of resulting data of CPP-CQA relationship analysis and model building. All statistical parameters are dimensionless quantities with the exception of SEC, SECV and SEP that use the dimension of each CQA (BP- bar, Tk - µm, Df - ml/min; AF - L/m².s)........................................................................................................................................83
Figure 1 – Scheme of the membrane production process. The CPP influence is highly focused on casting, conditioning and immersion ................................................................. 3
Figure 2 – Scheme of the membrane production process with the NIRS implemented ................................................................. 3
Figure 3 - Symbioses between all the concepts that constitute PAT. The upper half of the scheme in Figure 2 relates to process understanding (state estimation and identification) while the lower half deals with fault detection and process control, essential to QbD. [18] .......................................................................................... 9
Figure 4 – Factors that influence the system to generate responses. The most influent factors are known as CPP and responses that are more affected by the system are known as CQA ................................................................. 10
Figure 5 – An example of application of Chance Method (A), COST Method (B), Grid Model (C) and DoE Method (D). The system has two parameters (Oven temperature vs. Baking time) [23] .......................................................................................... 11
Figure 6 – Standard example of the sequence for implementation of DoE [22]. .................................................................................. 12
Figure 7 – Different configurations of DSs: full factorial (row A), fractional factorial (row B) and composite design (row C). Each row shows DSs for 2 factors, 3 factors and more than 3 factors [22]. .................................................................................. 13
Figure 8 – NIR frequency range of overtones. [26] .................................................................................................................. 16
Figure 9 – Variation of the error of prediction with the increasing number of components. The estimation error (long dashed line) and model error (short dashed line) contribute to the prediction error (continuous line).g [4] ................. 23
Figure 10 – Membrane processes and their separation characteristics [34] .................................................................................. 27
Figure 11 – Detailed description of the conditioning and precipitation step. As the membrane is formed the solvent in the casting solution is evaporated and a non-solvent is condensed in the membrane. [1] .................................................................................. 28
Figure 12 – (1) Variation of the flow with gas pressure. As the pressure increases, the flow increases linearly (Diffusive flow) until the BP. Higher pressures indicate higher flows (Bulk flow) (2) - Scheme of BP-method at the membrane scale (grey shapes) with the wet filling (blue shapes). (a) At room pressure, the diffusion through the wet membrane is limited. (b) As the pressure increases, diffusion increase. (c) Once obtained the BP pressure, the bulk air flows through the membrane without the impediment of the wet liquid. [1] .................................................................................. 30
Figure 13 – Scheme of the membrane production process. The CPPs analyzed in this work are marked where they have most impact on the process (except CSp) .................................................................................. 32
Figure 14 – Effect of an increasing T_{band} (direction of the arrow) on the pore size and morphology .................................................................................. 32
Figure 15 – Effect of an increasing T_{Pb} (direction of the arrow) on the pore size and morphology .................................................................................. 33
Figure 16 – Measurement scheme of the BioPAT® Spectro. The halogen lamp produces a broad band light that hits the sample. The resulting light is separated in the grating and detected in the Diode array. [1] .................................................................................. 36
Figure 17 – Spectral data of BioPAT® Spectro before the optical optimization (alt) and after (neu) [1] .................................................................................. 36
Figure 18 – Stages of the project “Automation of Casting Machine”. Current stage is marked with a yellow circle (6. Automatic model building for each membrane type). [1] .................................................................................. 37
Figure 19 – Overview of the thesis objective .................................................................................................................. 38
Figure 20 – Sets of membrane sheets divided by CPP state .................................................................................................................. 42
Figure 21 – Scheme of the referencing of a membrane sheet. The blue bar indicates the area in which the NIR scans the membrane. The circle indicates the cutted piece of membrane required by the measurement with Sartocheck. .................. 42
Figure 22 – Summary fit plot for membrane 15458 using 3 CPPs and MLR. Blue bars denote R². Orange bars represent Q². Grey bars signify MV and yellow bars represent Rep. ................................................................................. 44
Figure 23 – Correlation plots for membrane 15458. A is BP (Bubble Point), B is Tk (Thickness), C is AF (Air Flux) and D is Df (Diffusion). Green bars represent each CPP contribution of the model. Black lines represent the associated error. All data used was scaled and centered and error has a confident interval of 95%. ..............................................45
Figure 24 – Summary fit plot for membrane 15457 using 5 CPPs and MLR. Blue bars denote R². Orange bars represent Q². Grey bars signify MV and yellow bars represent Rep. ..................................................................................47
Figure 25 – Correlation plots for membrane 15457. A is BP (Bubble Point), B is Tk (Thickness), C is AF (Air Flux) and D is Df (Diffusion). Green bars represent each CPP contribution of the model. Black lines represent the associated error. All data used was scaled and centered and error has a confident interval of 95%..........................48
Figure 26 – Summary fit plot for membrane 15404-H using 3 CPPs and MLR. Blue bars denote R². Orange bars represent Q². Grey bars signify MV and yellow bars represent Rep. .................................................................49
Figure 27 – Correlation plots for membrane 15404-H. A is BP (Bubble Point), B is Tk (Thickness) and C is AF (Air Flux). Green bars represent each CPP contribution of the model. Black lines represent the associated error. All data used was scaled and centered and error has a confident interval of 95%..........................50
Figure 28 – Summary fit plot for membrane 15407 using 4 CPPs and MLR. Blue bars denote R². Orange bars represent Q². Grey bars signify MV and yellow bars represent Rep. ..................................................................................51
Figure 29 – Correlation plots for membrane 15407. A is BP (Bubble Point), B is Tk (Thickness) and C is AF (Air Flux). Green bars represent each CPP contribution of the model. Black lines represent the associated error. All data used was scaled and centered and error has a confident interval of 95%..........................52
Figure 30 – Loadings plot for BP model of 15458 for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line). ........................................................................................................55
Figure 31 – Estimates plot for BP model of 15458. Black dots represent raw data and red dots represent outliers. SEC=0,19 bar and R²=0,806..................................................................................55
Figure 32 – Residues plot for BP model of 15458. Residues are divided according to higher than 2 SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1 SD (green dots)..........................55
Figure 33 – Estimates plot for BP model of 15458 (cleaned data). It shows normal data (dots, SEC=0,05 bar and R²=0,983), cross-validation data (rhombus, SECV=0,08 bar and R²=0,956) and validation data (squares). SEP=0,11 bar and R²=0,991 ..................................................................................56
Figure 34 – Residues plot for BP model of 15458 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1 SD (green dots)..........................56
Figure 35 – Press plot for BP model of 15458 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line), SECV (grey line) and Balanced (yellow line)..................................................................................56
Figure 36 – Loadings plot for Tk model of 15458 for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line). ........................................................................................................56
Figure 37 – Estimates plot for Tk model of 15458. Black dots represent raw data and red dots represent outliers. SEC=1,6 µm and R²=0,855..................................................................................56
Figure 38 – Residues plot for Tk model of 15458. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1 SD (green dots)..........................56
Figure 39 – Estimates plot for Tk model of 15458 (cleaned data). It shows normal data (dots, SEC=0,90 µm and R²=0,954), cross-validation data (rhombus, SECV=1,35 µm and R²=0,9) and validation data (squares). SEP=0,47 µm and R²=0,986 ..................................................................................57
Figure 40 – Residues plot for Tk model of 15458 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1 SD (green dots)..........................57
Figure 41 – Press plot for Tk model of 15458 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line), SECV (grey line) and Balanced (yellow line) ................................................................. 57
Figure 42 – Loadings plot for AF model of 15458 for 5 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line), 4 PC (yellow line) and 5 PC (green line) ................................................................. 58
Figure 43 – Estimates plot for AF model of 15458. Black dots represent raw data and red dots represent outliers. SEC=0,05 L/m².s and R²=0,887 ........................................................................................................ 58
Figure 44 – Residues plot for AF model of 15458. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) ................................................................. 58
Figure 45 – Estimates plot for AF model of 15458 (cleaned data). It shows normal data (dots, SEC=0,02 L/m².s and R²=0,981), cross-validation data (rhombus, SECV=0,05 L/m².s and R²=0,938) and validation data (squares). SEP=0,08 L/m².s and R²=0,804 ........................................................................................................ 58
Figure 46 – Residues plot for AF model of 15458 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) ................................................................. 58
Figure 47 – Press plot for AF model of 15458 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line), SECV (grey line) and Balanced (yellow line) ................................................................. 58
Figure 48 – Loadings plot for BP model of 15457 for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line), 4 PC (yellow line) and 5 PC (green line) ................................................................. 59
Figure 49 – Estimates plot for BP model of 15457. Black dots represent raw data and red dots represent outliers. SEC=0,16 bar and R²=0,876 ........................................................................................................ 59
Figure 50 – Residues plot for BP model of 15457. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) ................................................................. 59
Figure 51 – Estimates plot for BP model of 15457 (cleaned data). It shows normal data (dots, SEC=0,08 bar and R²=0,966), cross-validation data (rhombus, SECV=0,11 bar and R²=0,929) and validation data (squares). SEP=0,16 bar and R²=0,867 ........................................................................................................ 60
Figure 52 – Residues plot for BP model of 15457 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) ................................................................. 60
Figure 53 – Press plot for BP model of 15457 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line), SECV (grey line) and Balanced (yellow line) ................................................................. 60
Figure 54 – Loadings plot for Tk model of 15457 for 5 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line) ........................................................................................................ 60
Figure 55 – Estimates plot for Tk model of 15457. Black dots represent raw data and red dots represent outliers. SEC=4,45 µm and R²=0,946 ........................................................................................................ 60
Figure 56 – Residues plot for Tk model of 15457. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) ................................................................. 60
Figure 57 – Estimates plot for Tk model of 15457 (cleaned data). It shows normal data (dots, SEC=2,06 µm and R²=0,986), cross-validation data (rhombus, SECV=2,93 µm and R²=0,977) and validation data (squares). SEP=3,03 µm and R²=0,974 ........................................................................................................ 61
Figure 58 – Residues plot for Tk model of 15457 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) ................................................................. 61
Figure 59 – Press plot for Tk model of 15457 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line), SECV (grey line) and Balanced (yellow line) ........................................................................................................ 61
Figure 60 – Loadings plot for AF model of 15457 for 5 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line), 4 PC (yellow line) and 5 PC (green line). .................................62
Figure 61 – Estimates plot for AF model of 15457. Black dots represent raw data and red dots represent outliers. SEP=0,13 L/m².s and R²=0,881 .................................................................62
Figure 62 – Residues plot for AF model of 15457. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................................62
Figure 63 – Estimates plot for AF model of 1547 (cleaned data). It shows normal data (dots, SEC=0,13 L/m².s and R²=0,96), cross-validation data (rhombus, SECV=0,19 L/m².s and R²=0,924) and validation data (squares). SEP=0,12 L/m².s and R²=0,88 .................................................................62
Figure 64 – Residues plot for AF model of 15457 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................................62
Figure 65 – Press plot for AF model of 15457 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line), SECV (grey line) and Balanced (yellow line) .................................................................62
Figure 66 – Loadings plot for BP model of 15404-H for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line). .................................................................63
Figure 67 – Estimates plot for BP model of 15404-H. Black dots represent raw data and red dots represent outliers. SEC= 0,14 bar and R²=0,881 .................................................................63
Figure 68 – Residues plot for BP model of 15404-H. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................................63
Figure 69 – Estimates plot for BP model of 15404-H (cleaned data). It shows normal data (dots, SEC= 0,13 bar and R²=0,926) and validation data (squares). SEP=0,11 bar and R²=0,802 .................................................................64
Figure 70 – Residues plot for BP model of 15404-H (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................................64
Figure 71 – Press plot for BP model of 15404-H (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line) and Balanced (yellow line) .................................................................64
Figure 72 – Loadings plot for Tk model of 15404-H for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line), 4 PC (yellow line) .................................................................64
Figure 73 – Estimates plot for Tk model of 15404-H. Black dots represent raw data and red dots represent outliers. SEC=6,47 µm and R²=0,781 .................................................................64
Figure 74 – Residues plot for Tk model of 15404-H. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................................64
Figure 75 – Estimates plot for Tk model of 15404-H (cleaned data). It shows normal data (dots,SEC=2,53 µm and R²=0,977) and validation data (squares). SEP=3,52 µm and R²=0,904 .................................................................65
Figure 76 – Residues plot for Tk model of 15404-H (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................................65
Figure 77 – Press plot for Tk model of 15404-H (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line) and Balanced (yellow line) .................................................................65
Figure 78 – Loadings plot for AF model of 15404-H for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line). .................................................................66
Figure 79 – Estimates plot for AF model of 15404-H. Black dots represent raw data and red dots represent outliers. SEC=1,92 L/m².s and R²=0,718 .................................................................66
Figure 80 – Residues plot for AF model of 15404-H. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................. 66
Figure 81 – Estimates plot for AF model of 15404-H (cleaned data). It shows normal data (dots, SEC=0,19 L/m².s and R²=0,998), and validation data (squares). SEP=1,56 L/m².s and R²=0,499 ................................................................. 66
Figure 82 – Residues plot for AF model of 15404-H (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................. 66
Figure 83 – Press plot for AF model of 15404-H (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line) and Balanced (yellow line) ................................................................. 66
Figure 84 – Loadings plot for BP model of 15407 for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line). ........................................................................................................... 66
Figure 85 – Estimates plot for BP model of 15407. Black dots represent raw data. SEC=0,11 bar and R²=0,968 ............................................................................ 67
Figure 86 – Residues plot for BP model of 15407. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow dots) and lower than 1SD (green dots) .................................................. 67
Figure 87 – Estimates plot for BP model of 15407 (cleaned data). It shows normal data (dots, SEC=0,06 bar and R²=0,993), and validation data (squares). SEP=0,14 bar and R²=1 ..................................................................................................... 68
Figure 88 – Residues plot for BP model of 15407 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................. 68
Figure 89 – Press plot for BP model of 15407 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line) and Balanced (yellow line) ................................................................. 68
Figure 90 – Estimates plot for the old BP model of 15407 (cleaned data). It shows DoE data (blue dots, SEC=0,06 bar, R²=0,993; SEP=0,14 bar and R²=1), and system data (black dots, SEC=0,10 bar, R²=0,923; SEP=0,08 bar and R²=0,921) ..................................................................................................... 68
Figure 91 – Residues plot for the old BP model of 15407 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow dots) and lower than 1SD (green dots). Blue dots represent the residues of the new data .................................................................................. 68
Figure 92 – Loadings plot for Tk model of 15407 for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line). ................................................................. 69
Figure 93 – Estimates plot for Tk model of 15407. Black dots represent raw data. SEC=24,48 µm and R²=0,549 ............................................................................ 69
Figure 94 – Residues plot for Tk model of 15407. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................. 69
Figure 95 – Estimates plot for Tk model of 15407 (cleaned data). It shows normal data (dots, SEC=18,04 µm and R²=0,781 and validation data (squares). SEP=51,71 µm and R²=1 ..................................................................................................... 69
Figure 96 – Residues plot for Tk model of 15407 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................. 69
Figure 97 – Press plot for Tk model of 15407 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line) and Balanced (yellow line) ................................................................. 69
Figure 98 – Load plot for AF model of 15407 for 5 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line), 4 PC (yellow line) and 5 PC (green line). ........................................................................................................... 70
Figure 99 – Estimates plot for AF model of 15407. Black dots represent raw data. SEC=0,19 L/m².s and R²=0,823 ............................................................................ 70
Figure 100 – Residues plot for AF model of 15407. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................................. 70
Figure 101 – Estimates plot for AF model of 15407 (cleaned data). It shows normal data (dots, SEC=0,12 L/m².s and $R^2=0,945$) and validation data (squares). SEP=0,09 L/m².s and $R^2=1$ ................................. 71
Figure 102 – Residues plot for AF model of 15407 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots) .................................. 71
Figure 103 – Press plot for AF model of 15407 (cleaned data). The errors presents according to PC number is SEC (blue line), SEP (orange line) and Balanced (yellow line) ................................................................. 71
Figure 104 – Estimates plot for the old AF model of 15407 (cleaned data). It shows DoE data (blue dots, SEC=0,12 L/m².s, $R^2=0,945$; SEP=0,09 L/m².s and $R^2=1$), and system data (black dots, SEC=0,116 L/m².s, $R^2=0,898$; SEP=0,073 L/m².s and $R^2=0,954$ ........................................................................................................ 71
Figure 105 – Residues plot for the old AF model of 15407 (cleaned data). Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow dots) and lower than 1SD (green dots). Blue dots represent the residues of the new data ........................................................................................................ 71
Figure 106 – Decrease of SEP (red bars) and increase of $R^2$ (yellow bars) for the membranes 15458, 15457, 15404-H and 15407 ................................................................. 72
Figure 107 – Comparison between SEC (blue bars), SECV (orange bars) and SEP (yellow) for the models of the 4 membranes by CQA (Plot A – BP; Plot B – Tk; Plot C – AF). Due the low number of experiments, cross-validation was not used in membranes 15404-H and 15407 which explains the absent values. In the plot B, the TK of 15407 is beyond the max value of the axis, 51,71 µm ........................................................................................................ 74
Figure 108 – Comparison between SEC (blue bars), SECV (orange bars) and SEP (yellow) for the models of the 4 membranes according to the BP. Due the low number of experiments, cross-validation was not used in membranes 15404-H and 15407 which explains the absent values ........................................................................................................ 75
Figure 109 – Sartocheck and stainless steel in-line filter holder ........................................................................................................ 80
Figure 110 – Bubble Point manual measurer ........................................................................................................ 80
Figure 111 – Caliper (Prüfmittel) used to measure thickness ........................................................................................................ 80
Figure 112 – Digital Optic Microscope used to measure thickness ........................................................................................................ 80
Figure 113 – Air permeability device used to measure Air Flux ........................................................................................................ 81
Figure 114 – NIR device: BioPAT® Spectro ........................................................................................................ 81
<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{y}_i$</td>
<td>Model-estimated property value for the sample $i$</td>
</tr>
<tr>
<td>$I_o$</td>
<td>Intensity of incident radiation</td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>Vector that contains the mean response values</td>
</tr>
<tr>
<td>$y_i$</td>
<td>Values of $Y$-data</td>
</tr>
<tr>
<td>$1_N$</td>
<td>Vector of ones that is $N$-elements long</td>
</tr>
<tr>
<td>$a$</td>
<td>Multiplicative correction factor</td>
</tr>
<tr>
<td>$A$</td>
<td>Number of components of the model</td>
</tr>
<tr>
<td>$a_0$</td>
<td>Standard deviation of the sample spectrum</td>
</tr>
<tr>
<td>$a_1$</td>
<td>Average value of the sample spectrum to be corrected</td>
</tr>
<tr>
<td>Abs</td>
<td>Absorbance</td>
</tr>
<tr>
<td>ACS</td>
<td>Amount of Casting Solution (kg/m²)</td>
</tr>
<tr>
<td>AF</td>
<td>Air Flux (L/m².s)</td>
</tr>
<tr>
<td>API</td>
<td>Active Pharmaceutical Ingredient</td>
</tr>
<tr>
<td>$b$</td>
<td>Addictive correction factor</td>
</tr>
<tr>
<td>BP</td>
<td>Bubble Point (bar/mbar)</td>
</tr>
<tr>
<td>CA</td>
<td>Cluster Analysis</td>
</tr>
<tr>
<td>$C_A$</td>
<td>Concentration of an Absorber</td>
</tr>
<tr>
<td>COST</td>
<td>Changing One Single factor at the Time</td>
</tr>
<tr>
<td>Cov</td>
<td>Coverage</td>
</tr>
<tr>
<td>CP</td>
<td>Center Points</td>
</tr>
<tr>
<td>CPP</td>
<td>Critical Process Parameter</td>
</tr>
<tr>
<td>CQA</td>
<td>Critical Quality Attribute</td>
</tr>
<tr>
<td>CSp</td>
<td>Casting Speed (m/s)</td>
</tr>
<tr>
<td>$CV$</td>
<td>Cross Validation</td>
</tr>
<tr>
<td>$Df$</td>
<td>Diffusion (ml/min)</td>
</tr>
<tr>
<td>DoE</td>
<td>Design of Experiments</td>
</tr>
<tr>
<td>DOM</td>
<td>Digital Optical Microscope</td>
</tr>
<tr>
<td>DS</td>
<td>Design Space</td>
</tr>
<tr>
<td>$E$</td>
<td>Matrix of Residuals applied to PCA/PLS</td>
</tr>
<tr>
<td>EMA</td>
<td>European Medicines Agency</td>
</tr>
<tr>
<td>EMEA</td>
<td>European Agency for the Evaluation of Medicinal Products</td>
</tr>
<tr>
<td>EMSC</td>
<td>Extended Multiplicative Scatter Correction</td>
</tr>
<tr>
<td>$f$</td>
<td>Residuals of the $y$-data</td>
</tr>
<tr>
<td>FA</td>
<td>Factor Analysis</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>FIR</td>
<td>Far Infra-Red</td>
</tr>
<tr>
<td>FMEA</td>
<td>Failure Mode Effect Analysis</td>
</tr>
<tr>
<td>ICH</td>
<td>International Conference on Harmonization</td>
</tr>
<tr>
<td>IPC</td>
<td>In-Process Control</td>
</tr>
<tr>
<td>IR</td>
<td>Infra-Red</td>
</tr>
<tr>
<td>$L^i$</td>
<td>Matrix of Loadings applied to PCA/PLS</td>
</tr>
<tr>
<td>MC</td>
<td>Mean Centering</td>
</tr>
<tr>
<td>MIR</td>
<td>Mid Infra-Red</td>
</tr>
<tr>
<td>MLR</td>
<td>Multiple Linear Regression</td>
</tr>
<tr>
<td>MS corrected</td>
<td>Mean square of the pure error</td>
</tr>
<tr>
<td>MS total SS corrected</td>
<td>Mean square of the total SS corrected</td>
</tr>
<tr>
<td>MSC</td>
<td>Multiplicative Scatter Correction</td>
</tr>
<tr>
<td>MV</td>
<td>Model Validation</td>
</tr>
<tr>
<td>MVDA</td>
<td>Multivariate Data Analysis</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of variables in the spectrum/ Number of calibration samples</td>
</tr>
<tr>
<td>$N_{CV}$</td>
<td>Number used for cross-validation</td>
</tr>
<tr>
<td>$N_{Exp}$</td>
<td>Number of experiments (Initial)</td>
</tr>
<tr>
<td>$N_{fExp}$</td>
<td>Final number of experiments (after outlier detection)</td>
</tr>
<tr>
<td>NIPALS</td>
<td>Non-linear iterative partial least squares</td>
</tr>
<tr>
<td>NIR</td>
<td>Near Infra-Red</td>
</tr>
<tr>
<td>NIRS</td>
<td>Near Infra-Red Spectroscopy</td>
</tr>
<tr>
<td>$N_{ou}$</td>
<td>Number of outliers</td>
</tr>
<tr>
<td>$N_{p}$</td>
<td>Number of external validation samples</td>
</tr>
<tr>
<td>$n_{PC}$</td>
<td>Number of Principal Components</td>
</tr>
<tr>
<td>NPS</td>
<td>Normal Production State</td>
</tr>
<tr>
<td>$N_{val}$</td>
<td>Number used for validation</td>
</tr>
<tr>
<td>NW</td>
<td>Norris Williams method</td>
</tr>
<tr>
<td>OEM</td>
<td>Original Equipment Manufacturer</td>
</tr>
<tr>
<td>OOS</td>
<td>Out of Specification (Product)</td>
</tr>
<tr>
<td>PA</td>
<td>Process Analysis</td>
</tr>
<tr>
<td>PAC</td>
<td>Process Analytical Chemistry</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PASG</td>
<td>Pharmaceutical Analytical Sciences Group</td>
</tr>
<tr>
<td>PAT</td>
<td>Process Analytical Technology</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Components</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>Principle Component Regression</td>
</tr>
<tr>
<td>PCS</td>
<td>Process Control System</td>
</tr>
<tr>
<td>PES</td>
<td>Polylethersulfone</td>
</tr>
<tr>
<td>PI</td>
<td>Proportional-Integral controller</td>
</tr>
<tr>
<td>PID</td>
<td>Proportional-Integral-Derivative controller</td>
</tr>
<tr>
<td>plof</td>
<td>p for lack of fit</td>
</tr>
<tr>
<td>PLS</td>
<td>Partial Least Squares</td>
</tr>
<tr>
<td>PRESS</td>
<td>Prediction residual sum of squares</td>
</tr>
<tr>
<td>Q²</td>
<td>Percentage of the variation of the response predicted by the model. Predictive factor</td>
</tr>
<tr>
<td>QbD</td>
<td>Quality by Design</td>
</tr>
<tr>
<td>qᵢ</td>
<td>Transposed matrix of the depending loadings</td>
</tr>
<tr>
<td>R²</td>
<td>Percentage of the variation of the response explained by the model; correlation factor</td>
</tr>
<tr>
<td>RA</td>
<td>Risk Analysis</td>
</tr>
<tr>
<td>Rep</td>
<td>Reproducibility</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean square error</td>
</tr>
<tr>
<td>RMSEC</td>
<td>Root mean square error of calibration</td>
</tr>
<tr>
<td>RMSECV</td>
<td>Root mean square error of cross-validation</td>
</tr>
<tr>
<td>RMSEP</td>
<td>Root mean square error of prediction</td>
</tr>
<tr>
<td>r-Out</td>
<td>Ratio between the outliers and the initial number of experiments</td>
</tr>
<tr>
<td>r_p</td>
<td>Radius of capillary shaped pore</td>
</tr>
<tr>
<td>RPN</td>
<td>Risk Priority Number</td>
</tr>
<tr>
<td>SCADA</td>
<td>Supervisory Control and Data Acquisition</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SEC</td>
<td>Standard Error of Calibration</td>
</tr>
<tr>
<td>SECV</td>
<td>Standard Error of Cross Validation</td>
</tr>
<tr>
<td>SEM</td>
<td>Scanning Electron Microscope</td>
</tr>
<tr>
<td>SEP</td>
<td>Standard Error of Prediction</td>
</tr>
<tr>
<td>SG</td>
<td>Savitzky-Golay filters method</td>
</tr>
<tr>
<td>S_M</td>
<td>Matrix containing the standard deviations for each of the variables in the X/Y data</td>
</tr>
<tr>
<td>SNVT</td>
<td>Standard Normal Variate Transformation</td>
</tr>
<tr>
<td>SPC</td>
<td>Statistical Process Control</td>
</tr>
<tr>
<td>SS</td>
<td>Total sum of squares of Y corrected for the mean</td>
</tr>
<tr>
<td>SSREG</td>
<td>Partial sum of squares of Y corrected for the mean</td>
</tr>
<tr>
<td>T</td>
<td>Matrix of Scores (applied to (T_{PCA}/T_{PLS}) PCA/PLS)</td>
</tr>
<tr>
<td>T_Band</td>
<td>Temperature of the band/belt (°C)</td>
</tr>
<tr>
<td>T_col</td>
<td>Temperature of the (Distillation) Column (°C)</td>
</tr>
<tr>
<td>Tk</td>
<td>Thickness (µm)</td>
</tr>
<tr>
<td>T_PB</td>
<td>Temperature of the precipitation bath (°C)</td>
</tr>
<tr>
<td>t_Roll</td>
<td>Time between changing the roll</td>
</tr>
<tr>
<td>v</td>
<td>Vibrational level</td>
</tr>
<tr>
<td>X/Y</td>
<td>Data set containing information about a system</td>
</tr>
<tr>
<td>X-H</td>
<td>Chemical bounds of the specific specie X to hydrogen</td>
</tr>
<tr>
<td>X_j</td>
<td>Data of the variable j</td>
</tr>
<tr>
<td>X_MC</td>
<td>Data set obtained after using mean centering</td>
</tr>
<tr>
<td>X_Ref</td>
<td>Data set of the reference</td>
</tr>
<tr>
<td>X_SCAL</td>
<td>Data set obtained after using scaling</td>
</tr>
<tr>
<td>X_SNVT</td>
<td>Data set obtained after using Standard Normal Variate Transformation</td>
</tr>
<tr>
<td>γ</td>
<td>Surface tension at the liquid/air interface</td>
</tr>
<tr>
<td>ΔP</td>
<td>Applied pressure</td>
</tr>
<tr>
<td>ε</td>
<td>Absorptivity</td>
</tr>
<tr>
<td>θ</td>
<td>Contact angle</td>
</tr>
<tr>
<td>I</td>
<td>Intensity of transmitted radiation</td>
</tr>
<tr>
<td>l</td>
<td>Optical path length</td>
</tr>
</tbody>
</table>
1. Introduction

Since the beginning of industrial times, mankind usually urged to understand the full potential of a production process. In order to achieve the best value in the desired end product, it is necessary to control it. An optimized process control can save the most significant inputs, i.e. time, raw materials, utilities expenses and human resources. This lures the industry into investing in this sector. This is particularly important in biotechnology industries. On one hand, the ongoing growth of biotechnology market demands a constant innovation and pursuit of new competitive advantages. On the other hand, this type of industry requires high precision control technologies. Therefore, it is expected that this particular niche of companies tend to be more aware and focused in the implementation of optimized control systems in order to obtain the best results.

The Sartorius Stedim Biotech GmbH is a German business corporation focused on manufacturing laboratory and biotechnology solutions. The product range goes from weighting technologies to cell fermenters and control sensor devices. Their main products are filter membranes used in many direct applications such as pre and sterile filtration in aerated vessels. They also produce syringe filters as raw materials for more complexes application such as pregnancy tests. In the past few years, Sartorius become a Process Analytical Technology (PAT) solutions provider by supplying process analyzers and software tools capable of optimizing manufacturing processes.

The PAT approach combines process knowledge with analyzers, sensors, and software for automation and control. This approach usually starts with risk assessment (RA) which requires prior knowledge of the process. Using Design of Experiments (DoE) it is possible to know which parameters might have an impact on the product. However this implies that the quality of the process/product can be analyzed. This raises the need for analyzers and sensors. Such control instrumentation will generate a significant amount of data which raises the need for Multivariate Data Analysis (MVDA). Finally, the process needs to automatically react on sensor/MVDA data which raises the need for control software (e.g. Supervisory Control and Data Acquisition - SCADA). Only then every single substep of the process is controlled and fully understand. In these conditions, it is not necessary to control the quality of the final product anymore, since it is built-in the process. This is the cornerstone of Quality by Design (QbD).
1.1. Motivation

In 2010, Sartorius Stedim Biotech has planned to implement a PAT approach in a membrane casting machine, as one of their projects. The objective of this project is to improve product quality control, through the use of spectroscopy, DoE and MVDA.

Since most of these membranes are used to assemble filter cartridges, deviations in the quality may compromise the proper functioning and final throughput of this component. Thus it may lead to inappropriate aeration and sterilization of target environments. The quality of the final membranes relies on the production process, more specifically on critical process parameters (CPPs).

The continuous production of Polyethersulfone (PES) membranes involves several stages. The CPPs have a special influence in the beginning of the process. So far the process control is only monitoring the stability of CPPs. No monitoring in product and process quality has been implemented. Quality control is only performed on a small fraction of the membrane roll at the end of the process. So far, the sampling practice is only performed at the end of a 180m roll. This raises a problem since there is not a fixed set of CPP for each membrane. These parameters may vary over lot and time. Therefore, if quality control traces a membrane that is out of specification (OOS), adjustments have to be made in CPP. The quality control also relies on the assumption that there is no variability within the roll. Therefore quality deviations within a roll may slip through, if the end of the roll fits within specification. OOS membranes have to be resent to quality control. This process is repeated until the membrane fits within specification. There is a substantial time delay between defining process parameters and the proper adjustment of critical quality attributes (CQA) of each membrane. The lack of full control raises the need to monitoring the process stability. This will significantly reduce the amount of samples that are sent to quality control, which reduces the time delay. Instead of sampling every roll, it is only necessary to sample in the beginning of each batch while monitoring the process stability. Considering that a batch can produce several rolls, the sampling process is significantly reduced. In case of detection of OOS parameters, additional samples need to be sent to quality control, which makes the sampling a guided practice.
In the beginning of this project, the main challenge was how to implement the process monitoring. A prior knowledge of the process will tell the optimum control instrument and its location on the production line (Figure 1). Having this in mind, it is necessary to consider that initially the casting solution is applied to a band. Then it is immersed in a precipitation bath. After precipitation agent extraction, the membrane is washed and dried. A good control instrument has to be able to predict physical and/or chemical proprieties in the stages that are more influenced by CPPs. In this case, it will be between the casting and immersion stages. It has been shown that using Near Infrared Spectroscopy (NIRS) over the belt edge, makes it possible to trace the physical properties of the final membrane. Therefore a BioPAT® Spectrometer has been installed in the production line (Figure 2).
Feasibility studies for one membrane have been followed after this implementation. A good process monitoring has to be able to predict not only optimum CQA ranges but also deviations. Therefore it was necessary to put the system in different conditions. A good option was to use DoE. This statistical approach accounts for all possible process variants. Using the DoE generating tools, it is possible to create a recipe with different ranges of parameters. Once applied to the process, this procedure will generate near infrared (NIR) spectral data adequate to be imported into MVDA software tools. Through this software, the spectral data is correlated to the respective generated CQA, in order to create a model. This model can be introduced in the production line system making the monitoring of process an instant experience for the user.

Instead of being based on stability of CPPs, the monitoring of process lies on variations of the generated membrane once it is formed. Even without the prediction of membrane quality, it is possible to reduce the sampling by guiding it:

1. Sampling in the beginning of a batch and ensuring that production is within specification.

2. Monitoring the process stability (changes in NIR data).

3. If the process is stable, no sampling is required. Otherwise, additional samples are required to be sent to quality control.

If the prediction of the quality is successful, the parameters can be altered without any delay for quality control. This change can be done as many times as required until the quality is within specification. Once the process is stable, quality is controlled with NIRS. Therefore the process is 100% controlled instead of once every 180 m. The NIRS becomes a valuable instrument of process control that is capable to predict OSS regions. As this instrument acquires more spectra, the more data is reintroduced in the system recalibrating the model. This contributes for an increasing robustness of the model.

A feasibility study for one membrane showed that product and process stability can likely be monitored. It also unveiled that the prediction of that membrane quality might be feasible. It is needed to take in consideration that the used casting machine can produce different membranes. Therefore, more feasibility studies have to be performed. This will guarantee that the process monitoring is active as long as the machine is running.

1.2. Thesis Objective

It is proposed in the current thesis to make a feasibility study for membranes 15458, 15457, 15404-H and 15407. The choice of these types lies on the importance in terms of quantity of production. Each feasibility study starts by defining the most relevant CPP for each membrane. This will allow to generate four DoEs that consider the change of the chosen CPPs. Each DoE will be performed in the used casting machine. This will generate four sets of NIR spectral data. The NIR data is gathered using an automated data
acquisition software. This data will be used in model building. To build these models, it is necessary to measure all CQA. This will allow to:

- Evaluate relationships between CPPs and CQAs for each membrane, using multiple linear regression (MLR). Also it can expose less relevant CQAs relatively to the changes of the selected CPPs.
- Build models using MVDA for each CQA of each membrane, using a modified version of Partial Least Square regression (PLS). These models will be generated using pre-treatments and outlier detection.

Once models are built and automation completed, the results are displayed in process control software (SCADA). In order to reduce parameters variability (e.g. lot-to-lot variability) the data obtained must be acquired in the most uninterrupted way possible. For this reason, each membrane will be considered as a different system. This will also make the data gathering a more feasible process.

As part of the project initiated in 2010, it is expected that this thesis will help to contribute for a new approach in the monitoring and control of casting machines for membrane production.

1.3. State of Research

The development of powerful computers in last decades permitted a systematic management of large quantities of data. This allowed for the multivariate methods to be used in analysis, monitoring and diagnostics of processes [1], [2]. Typical MVDA included statistical methods such as Principal Component Analysis (PCA) and PLS [1], [3].

All of this development allowed instrumentation like NIRS to use chemometrics and MVDA in analytical chemistry and process monitoring. NIR had its breakthrough starting from the 1980, when computers got fast enough to treat NIR spectra in a way that the chemical information could be extracted. NIR spectra produced can carry information on the physical and chemical constituents of samples. However, the NIR data are complex spectra with overlapping bands of chemical bonds [4]. This makes interpretation impossible using univariable analysis.

More recently, NIRS is becoming increasingly widespread in production processes in the biotechnology and pharmaceutical industries. [5]–[7]. It is not a surprise that lately, there have been numerous publications on applications of NIRS.

For instance, in the pharmaceutical industry, NIRS is now used both to trace physical parameters [8] and determination of chemical content. Chemical components, such as active pharmaceutical ingredients (APIs)
can now be measured. [7]. NIRS can also be used to detect polymorph structures that can modify the dissolution properties of the final drug. Among others, NIR can trace moisture content in powders or granulates, tablets, capsules, as well as lyophilized vials or in solutions [7], [9].

Most recently it was reported that NIRS could also predict biogas yield in anaerobic digestion by the analysis of the starting materials [10], [11]. NIRS has so far been applied only in a very limited application to the analysis of thin layers such as paper, films or membranes [1].

Considering the present work, only 3 companies are able to produce this type of membrane worldwide. Therefore, it is very difficult to collect statistical data of production process that can be relevant for comparison. This is aggravated by the fact that each company protects its own data, which makes most of this information impossible to reach.

Prior to the present work, Mathias Bode has used statistical methods for process optimization in his Master project work “Multivariate data analysis of a casting machine” at Sartorius. After identified CPPs, Mathias was able to merge parameters and spectral data. Mathias then applied MVDA on the fused data and compared the combined models with the models on spectroscopy data only. This certainly allows the exploration of new perspectives in real-time control and expand understanding of the process.

1.4. Thesis outline

This thesis is organized as follows. In chapter 2 the main concepts adjacent to this thesis are introduced, such as PAT, DoE, NIRS and MVDA. Chapter 3 is used to describe the membrane technology and its production process, including factors/responses and tools for characterization of membranes. Further, in Chapter 4, it is described the methodology used to gather experimental data. In the first part of chapter 5, the obtained results are used to evaluate relationships between CPPs and CQAs. Later in this same chapter, the same results are used for model building using MVDA. Finally, this dissertation is concluded in chapter 6 with a summary of the final results and findings. The chapter 8 is dedicated to perspective the future work.
2. Theoretical Background

This chapter is dedicated to explore the main concepts connected to the objective of this thesis. The chapter starts with a contextualization of PAT and Chemometrics (Section 2.1) and discusses the current viewpoint of QbD framework. In Section 2.2, focus on of the DoE and what factors affect its implementation. The Section 2.3 is dedicated to Near Infrared Technology. In section 2.3 it is explained the techniques that allow the interpretation of spectral data with MVDA.

2.1. Process Analytical Technologies (PAT) and Chemometrics

For the last several decades, many industries have used Process Analytics (PA) to achieve a deeper understanding of dynamics of their processes. This is a consequence of the increasing demanding for efficient control systems. According to R. Guenard and G. Thurau, PA has been generally described as:

*Chemical or physical analysis of materials in the process stream through the use of an in-line or on-line analyzer.*[12]

The classical scope of a process analytical method consist in adding physical or chemical information of a process analyzer to the control scheme of manufacturing process. [12]. However this definition does not comprehend all the factors influencing the efficiency of these processes. Due to the developments in operational excellence software, the definition was extended in such items as:

- Development of the process, particularly the identification of CQA and their relationship to the quality of the product.
- Design of a robust process to control these attributes
- Simple sensors and more complex process analyzers.
- An approach that uses and correlates all significant process information.
- Data mining approaches to detect long-term trends and interactions.
- Powerful data management systems to process the large amounts of data generated.

Considering the use of tools from Analytical Chemistry applied to PA, it is possible to define Process Analytical Chemistry (PAC). This multi-disciplinary field combines analytical chemistry techniques, process engineering, process chemistry and MVDA. A preliminary application of PAC can be used to gain a descriptive knowledge of a process [13]. This will allow a deeper understanding of parameters that impact a process and the quality of the product. The determination of trends in these parameters allows to establish the process signature. Such can be used to make changes in order to keep a process running with some set limits.
Ultimately, PAC leads to more consistent products, reduced waste, enhanced manufacturing efficiencies, overall improvement in the use of resources, better safety, and reduced costs. [12]

Although it is a widely known application since decades, PAC has only been used recently in pharmaceutical industry [12]. Drug manufacturing is an industrial activity comprehensively regulated by worldwide regulatory institutions. This include pharmacopoeias, Food and Drug Administration (FDA), International Conference on Harmonisation (ICH) and European Medicines Agency (EMA) [14][15]. In this industry, applicable regulations require strict control of not only raw materials, but also production processes. Such control has been mainly performed by using chromatographic techniques despite some inner disadvantages. These include sluggishness, the need for careful sample preparation and the production of potentially hazardous waste. [14][15]. Having this in mind, in 2004, FDA introduced PAC in pharmaceutical industry as Process Analytical Technology (PAT). It defined as:

“(...) a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality” [16]

PAT starts to consider Risk Analysis (RA), which consists in the identification of hazards and estimation of the risk associated with them. This component of PAT can be qualitative or quantitative and it is responsible for linking the likelihood of occurrence and severity of harms[17]. RA uses techniques as Failure Mode Effect Analysis (FMEA). This technique is used to identify factors that have impact on process and product. Potential hazards are quantified in levels of severity, observability and detectability. The combination of these 3 factors translates in a specific number, that can be compared with other hazards (Risk Priority Number - RPN) [17]. Other tools like the Ishikawa diagram can be used to identify root causes that potentially contribute to a particular effect.

PAT also takes advantage of process analyzers and process control tools. Process analyzers can generate substantial amounts of data. This raises the need for PAT to consider knowledge assimilation through constant data gathering [18][19]. This data can be better understand with the use of chemometrics. This discipline uses mathematical and statistical methods for the selection of the optimal experimental procedure and data treatment of chemical analyses. Chemometrics regroups several topics such as DoE, information extraction methods (modelling, classification and test of assumptions) allowing the understanding of chemical mechanisms [7]. Chemometrics uses classification and regression methods. Classification methods include discriminant linear analysis, principal component analysis (PCA), factor analysis (FA) and cluster analysis (CA). Regression methods include MLR, principle component regression (PCR), and partial least squares (PLS). Chemometrics can also takes advantage on non-linear techniques such as neural networks for calibration and Kohonen networks for classification. Although in general these last are less robust than linear techniques because of lack of extrapolation capacity.
PAT is focused on the production process. This prevents the idea that product quality should be tested into products. Instead, it should be built-in or should be by process design. This perspective is known as Quality by Design (QbD). The connection between PAT and QbD lies on the fact that as the process makes the product, the better the process, the better the product. Before QbD, end-product testing was the established standard practice to record quality. This avoids OOS products to reach end-users, at the expense of rejecting or reworking these products. With QbD, all upstream critical-to-quality variability sources are controlled and kept in check. This leads to well-designed processes with all the necessary feed-forward and feed-back control.

In practical terms a good application of PAT requires a good monitoring of a system, in order to supervise all its range of action. This will allow to make a valid diagnosis. This outcome will determine what type of control it is necessary to obtain the desired process state/end product. The control addresses back the changes so that they can be monitored. As shown in Figure 3, PAT can be described as a synergy between these four components.

![Figure 3 - Symbioses between all the concepts that constitute PAT. The upper half of the scheme in Figure 2 relates to process understanding (state estimation and identification) while the lower half deals with fault detection and process control, essential to QbD.](image)

PAT can be implemented in different industrial scenarios:

- To replace an existing in-process control (IPC);
- To be used in process development; to scale-up and full-scale production of a new product and process;
- Integrated in an existing industrial process to improve process understanding since it can provide CPPs of various operations. [19]–[21].
2.2. Design of Experiments (DoE)

2.2.1. Introduction

As part of a good use of RA, it is necessary to identify the right variables in order to have a proper insight of the system. The input variables, also known as factors, are the key parameters that dynamically exert a relevant influence within the system. They can be used to manipulate towards a specific objective. Factors can be controllable or not. If the changes within factors occur in continuous scale, these factors are known as quantitative. If these changes occur according to a specific propriety, these factors are referred as qualitative. The system process adjusts itself due the influence of the factors. From all factors that can impact a system, there are different levels of influence (Figure 4). The most influent factors are known as Critical Process Parameters (CPP). Once altered, the system produces output information, also known as responses. The responses that are more affected by the system deviations are known as Critical Quality Attributes (CQA).[22]

![Figure 4 – Factors that influnce the system to generate responses. The most influent factors are known as CPP and responses that are more affected by the system are known as CQA.](image)

The main objective is to find the optimum CPP that can make the system respond with an optimum CQA. Most of the times, with just RA, it is impossible to know the main relations between CPPs and CQAs. Also the relation between different CPPs can be relevant for a good evaluation of the system. This constitutes a significant limitation in process understanding and the improvement of its predictability and efficiency. To solve this, it is necessary to define a Design Space (DS) of experiments. This can be achieved by varying the values of CPP(s) and registering the respective values of CQA(s).

Traditionally this has been done in many different ways. The most intuitive method is changing the values by chance (Figure 5A). Good results can be randomly achieved by arbitrary selecting CPPs. Although, this method never guarantees the finding of the optimal combination of CPP-CQA. This raises the need for a systematic approach. A good option relies on changing one factor at a time, until there is no further improvement. This is called the COST (Changing One Single factor at the Time) approach (Figure 5B). Although it represents the intuitive way of performing experiments, it is inefficient since it does not necessarily provide information about the optimum conditions. This is particularly true when the information is about interactions between CPPs. Changing one factor at a time, does not lead to any further
improvement in the results, since it produces different results with different starting points. This makes it almost impossible to predict system behaviors with untested values of CPPs. So maybe performing experiments for every possible combination of CPPs would solve the problem. In fact, this method called “Grid-method” does unveil the main effects and interactions between CPP (Figure 5D). However, for more complex systems, this method requires and generates a huge amount of data. Moreover, most of these data can be redundant. This consists in a problem if the number of experiments is conditioned. Another problem with both COST and Grid methods is the lack of significant statistical parameters as reproducibility. In order to estimate reproducibility, it takes in consideration repeated values, known as center points (CP). It is statistically known that the measurements will produce some variation in the results due to minor systematic (inner to the measuring device/measurer) or unsystematic (stochastically) errors. This will even happen in the same identical experimental conditions. This phenomenon occurs from laboratory experiments to full-scale production and it will create a small dispersion in the results that is often described as noise. In order to proper understanding of the system output, one must know the size of the experimental noise in order to establish control limits that indicate an acceptable variation. The data should vary in a limited interval which the size can be measured by the standard deviation. It should also tend to group around a central value which the location can be estimated by the average. Both of these parameters are sufficient to characterize the properties of the variability and they can be used to monitor the behavior of the system.[23]

Figure 5 – An example of application of Chance Method (A), COST Method (B), Grid Model (C) and DoE Method (D). The system has two parameters (Oven temperature vs. Baking time) [23]
One good alternative to both COST and Grid methods is to carefully construct a set of selected experiments in which all the factors vary simultaneously. This approach is known as Design of Experiments (DoE). This method highlights impact, interactions and significance over a broad experimental range of all CPP with the lowest number of experiments. Using DoE approach, it is possible to sharpen the estimate of the real effect. It is also possible to estimate the size of the noise considering the part of the variation which the mathematical model leaves unexplained, also known as residuals. This constitutes another advantage towards COST/Grid approach, since this last one is limited to capture some coincidental noise effect.[22]

2.2.2. Implementation Procedure

The following paragraphs guide the reader through the DoE procedure, explaining every step in detail. In order to obtain the desired results, the problem formulation is of central importance in DoE. A correct problem exposition makes it clear, the intentions underlying and experimental investigation. L. Erikson [22] exposed the problem formulation as a 6 step stage starting with a familiarization of the process.

- In this first approach, it is required for an initial contact with the experiment scenario considering the needed resources. In this step, it is used RA in order to evaluate the experiment feasibility.
- The second stage is known as screening. In this stage, the impact of CPPs and its own optimal ranges are identified in a DoE with a limited number of experiments. It covers numerous parameters over a broad range. Screening identifies the most relevant experiment contributors and leaves out factors that can also contribute with noise.
- The third step is described as optimization. After the screening step, now the impact of parameters is known. In this DoE, the parameter region is normally reduced compared to the original screening DoE. However more parameter levels are selected. The multi levels of the parameters allow for a precise determination of interactions and slope of the CPP-CQA relation. This allows to build surface plots where optimal parameter regions (also known as sweet spots) can be identified.
- The fourth step is known as robustness testing and it is used to challenge the sweet spot with common variations. The factors may depend on a statistical range of states instead of just one optimal. Therefore, it is important to test these changes and see how they will influence the responses.

Figure 6 – Standard example of the sequence for implementation of DoE [22].
As shown in Figure 6 the general sequence of the application of DoE starts by selecting the type of DoE (Screening, Optimization and Robustness). After this step, factors and responses are defined. Then, there is the selection of a specific DS.

The DS generated can be full factorial (Figure 7A) or fractional factorial (Figure 7B). These have two investigation levels of design and they are normally used in screening test. For optimization, the recommended design is the composite design (Figure 7C). All of these designs present a regular geometry inherited by the experimental region’s symmetrical shape. If the experimental region becomes irregular due to inaccessible experimental zones, other types of designs have to be employed. A full factorial design may describe the full extreme regions of the design space. However, it increases the number of factors and therefore, the number of experiments. In terms of designing a specific experiment protocol, the increased number of experiments can be time/resource consuming and optimization in terms of shortening the set of trials is sometimes beneficial. This means that there tends to be a redundancy in a $2^k$ full factorial design, that is, an excess number of estimated experiments that lack relevance. This leads to fractional factorial designs where the design space is divided (e.g. in designs with 3 parameters, the cube design is split into two half cubes). According to L. Erikson [22], useful fractional factorial designs tend to have $2^{n-1}$ trials and $2^{n-2}$ with five or more factors. This technique, although advantageous, can lead to confounded effects. This means that the solutions of the DoE evaluation are not unique. Therefore, multiple DoE configurations may express impacts/interactions between CPP and CQA of different natures.

![Diagram of different configurations of DSs](image)

Figure 7 – Different configurations of DSs: full factorial (row A), fractional factorial (row B) and composite design (row C). Each row shows DSs for 2 factors, 3 factors and more than 3 factors [22].

However in most cases, relations between CPP and CQA are independent of the experimental design. Therefore, fractional factorial designs may be a good approach to apply at the screening level. [22]
After defining the DS, experiments are performed according to the values CPPs established. Response values are measured and inputted. DoE is then evaluated and interpreted. This interpretation is based on the relations of CPPs-CQAs and how this relation fits. There are hundreds of models. However for practical reasons simplification has to be done.

Depending on the objective of the DoE (screening, optimization, robustness) the corresponding model type will limit the fit functionality. The simplest model is the linear and it only takes into account individual behavior of the several factors. For simpler system analysis (e.g. Screening DoEs) it is usually assumed that the models are linear. The system is considered well fitted if, after considering the interaction or quadratic terms, these terms have no significance. More complex models allow for quadratic terms. However the model still might be linear and the additional levels might also confirm that there is no quadratic but linear relation. If the relation is linear, quadratic terms will have no significance in the distribution plot. Cubic and consequent more complex models are considered rarely relevant in industrial practice.

After a proper evaluation, the DoE may not meet the targeted objectives. This can happen because:

- optimum responses may require different ranges of CPP;
- optimization of the DS may be required:
  - different DoE objectives (screening, optimization, robustness) may be required;

Additional experiments may be added in further developments. This will lead to the restart of the sequence (Figure 6) until the results meet the objectives. DoE consists a very useful tool for evaluation of systems. As its main application, DoE highlights optimal values of responses for a system. Although, DoE can generate significant statistical data that can be used as a support for the development of other applications.

2.3. Near-Infrared Spectroscopy

2.3.1. Fundamentals

As light passes through matter there are 5 possible interactions possible to take place: absorption, reflection, transmission, scattering and fluorescence. The measurements of this interactions can render information about the physical and chemical properties of a material. In case of absorption, instrumentation is not capable of direct absolute measure. Instead, it measures the radiation that is not absorbed, i.e. transmitted radiation. The direct way to measure transmission is by putting a sample between the detector and the light source. The reflected transmission (also known as diffuse reflection) can also be measured. Both these types of measurements can be deducted from the incident radiation. Taking only direct transmission into account, it is possible to obtain absorption in its most simple form only with intensity of incident radiation, \( I_o \), and the intensity of transmitted radiation, \( I \) (equation 1).
Once obtained, according to Beer-Lambert Law, absorbance \( (\text{Abs}) \) of a sample is directly proportional to the concentration of its absorber \((C_A)\) and to optical path length \((l)\). The linear factor \((\varepsilon)\) is known as absorptivity (equation 2).

\[
\text{Abs} = \log \left(\frac{I_0}{I}\right)
\]  
(1)

\[
\text{Abs} = \varepsilon C_A l
\]  
(2)

For accurate measurements, this correlation should be strictly applied to transmission values on samples that have no scatter effects. However, it can also be applied in light scattering materials taking into account their diffuse reflectance. In this case, the path length of radiation can vary and it is affected by light scattering. The scattering effect occurs when radiation transmitted through the surface and emerging after partial absorption is diffused by stochastic reflections, refractions and diffractions in others surfaces. This phenomenon is intrinsically related with moisture content and other physical properties of the sample, such as geometry, temperature and refractive indices. In this scenario, absorbance and concentration are no longer a linear relationship. [4]–[7].

The absorption of energy increases the vibrational energy in a molecule. The absorption occurs by three mechanisms: fundamental absorption, overtones and combinations of fundamental vibrations. As molecules are irradiated, they absorb photons with the specific amount of energy coincident with the characteristic vibrations. The energy curve of an oscillating molecule is affected by intermolecular interactions and vibrations around the equilibrium position are variable. Moreover, the distances between energy levels that the molecule contains, decreases with increasing energy. This situation is often compared with the quantum mechanical model of an anharmonic oscillator. Absorption of energy results in an excitation of the molecule to a higher vibrational level \((\nu)\). Quantum mechanics do not exclude transitions between vibrational levels higher than 1 for anharmonic systems. Thus transitions between vibrational states of \(\Delta \nu > 1\) are possible, although their probability decreases with an increase of the vibrational level number. Excitations to the consequent higher vibrational level are known as fundamental vibrations and they can be observed in Infrared (IR) spectroscopy at 2500-15000 nm or 4000-300 cm\(^{-1}\). If the excitation process results in photons on levels that are multiples of fundamental vibrations of IR, the resulting frequency is known as overtone. If the same photon excites two or more vibrations simultaneously, it gives rise to a combination of fundamental vibrations. These can be multiples of one or more of the vibrations and there is no theoretical limit to the number of absorptions that can be involved. This number of combinations increases with the number of vibrations involved. These situations occurs in the Near-Infrared (NIR) region of the electromagnetic spectrum, situated between 780 and 2500 nm or from 12800 to 400 cm\(^{-1}\) if measuring in wavenumbers. In NIR region, the band position of combinational and overtone vibrations deviate with increasing multiplicity from the exact multiples of their fundamentals. The reason can be found in the anharmonicity of the vibrations. Regions of absorption result in broad peaks caused by the overlapping of a multitude of different absorptions. The absorption bands occurring in the NIR region are mostly associated to overtones and combinations of fundamental vibrations of functional groups of \(-\text{CH}, -\text{NH}, -\text{OH}\) and \(-\text{SH}\). One reason for
this phenomenon is the fact that most of the X-H fundamentals absorb wavelengths > 2000 cm\(^{-1}\) so that their first overtones already appear in the NIR frequency range (Figure 8).[7], [14], [24], [25]. The intensity of the successive absorptions will decrease by a factor between 10 and 100 in comparison with fundamental vibrations that occur in the mid-infrared absorption bands. The overlapping of the many overtone and combination bands in NIR region will cause a very low structural selectivity compared to Mid Infrared (MIR) or Raman. For this reason, NIR region was not used until late 1950’s mainly because of its complexity, compared to MIR and Far Infrared (FIR) regions. The discovery of near-infrared energy was made by Herschel in the 19th century, while measuring the change in temperature of colors of light diffracted by a prism. He observed that the temperature continued to increase when he placed a thermometer beyond the red end of visible spectrum. Although reported, his work was essentially forgotten. Nowadays with the use of chemometrics and MVDA, NIRS generated data can be decoded. This will allow to relate the NIR spectral information to sample properties. This is important since NIR can be advantageous in comparison with other analytical methods.

![Near Infrared Band Assignment Table](image)

Figure 8 – NIR frequency range of overtones. [26]

In fact, NIR incident energy also is characterized by a low absorption coefficient which allows the beam to penetrate deep within the sample. This confers a high efficiency of this technique in a wider range of applications such as solids or turbid mixtures. [4], [24]. Also, the capability of handling both diffusing scattering effects and overlapping spectral bands can be an advantage in the aftermath. This is because NIRS with MVDA can trace a specific compound on a mixture and the effect of its presence in heterogeneous mixtures. This last one is an important advantage towards other analytic methods. More
conventional analytical methods usually require full purification of the samples. Although, most of the times effective, these methods do not take in consideration important synergetic interactions between compounds. This constitutes a major drawback when interactions between compounds are decisive for the purposes of the analysis. This fact gains major importance in an industrial environment where most of the materials have different levels of purity. Using the appropriate data analysis, NIR spectra can serve as an unique fingerprint that can be used both in laboratory and process scale.[24]

Despite all these benefits, there are also drawbacks in NIR technology. It can only cover all organic compounds with X-H chemical bonds. This includes water, which can have a major impact in spectra. Water generates interference that can make it difficult to trace compounds with lower concentrations. This problem is relevant in biological systems since culture media are mainly composed by water.

In order to take advantage of NIRS, it is necessary to choose the right instrumentation based on the situation and objective of the analysis.

2.3.2. Instrumentation

The first NIR spectrometers were based on modified UV-Vis spectrometers with a scanning grating and a NIR sensitive detector. Since then, multiple configurations have been developed and all of them vary based on the required analyte sensitivity, reliability, easy to manage, calibration transferability and implementation needs. This will condition the measurement principle which the selected NIR spectrometer can have. The measurement scheme may differ according to the way in which the information is collected. [24]

In a basic system, the light beam travels from a light source, hits the sample and is finally detected. Since the information on the beam depends on wavelength selection, its separation is necessary. This can either take place before or after the contact with sample. [24]

In the first case, the incident light is screened in monocromator (a scanning grating) to select specific wavelengths. The generated monochromatic wave reaches the sample and it is detected afterwards. The final spectrum is composed by a sequential segments of wavelengths. Since these segments will be generated as the grating is repositioned, the measurement cannot take place instantly. This constitutes a drawback for process implementation if the sample changes over time. Another disadvantage is the constant need of a standard wavelength to correct for the positioning of the grating. One way of reducing measurement times is replace the monochromator with a wheel of interference filters. These filters speed up the measurement, since they change hundreds of times per second. Also, this new configuration does not require a precise positioning which results in almost instant measurements for a complete spectrum. Nonetheless, the number of filters is limited and must be preselected, this raises the need to design a specific instrument according to the target application. Another example of a possible configuration is Fourier-Transform. This type of spectroscopy uses a two-beam interferometer. A beam splitter divides the
incident light into two beams that hit a fixed mirror and a movable mirror respectively. Upon returning to the splitter, the two beams are superimposed (interfered) and that effect generates an image of light that is projected on the sample. The non-absorbed light is then photodetected. The movable mirror will generated differences in the optical path lengths which alters the wavelengths detection. The light that results from the interference between the two beams is deciphered using the Fourier transform algorithm [24]. Fourier-Transform Spectroscopy may need several scans to improve signal to noise ratio and it can be very sensitive to disturbances.[24]

As mentioned above, the beam can also be separated after hitting the sample. In these type of configurations the incident beam that strikes the sample has a broader range of wavelengths. It is necessary to employ a diffractive element on the detector. One typical configuration with this multichannel spectrometers can be found in diode array spectrometers. The light that results from the interaction with sample is guided to a grating mirror and the diffracted light is imaged on a linear plate of photodiodes, known as diode array. This configuration is particularly useful for process implementation due to the lack of moving parts. This allows the whole spectra to be acquired in a single step, which makes small sample changes more detectable over time. Thus, Diode array spectrometers are a good choice for process implementation. [24]

Many other configurations are possible. Although the NIR instrumentation core is simple. Accessories can be added to the NIR depending on the instrument/sample geometry and final application.

2.3.1. Applications

NIRS can be used in laboratory or process analyses. Laboratory analyzers are used in off-line or at-line measurements (e.g. quality control, research and plant laboratories). These instruments require high analyte sensitivity and reliability. Therefore, important aspects of these spectrometers are: a high signal-to-noise ratio, instrument stability, sufficient resolution and optimum sample conditions. The speed of the measurement is less important which makes scanning grating spectrometers, a feasible choice. Due to its high sensitivity, Fourier Transform Spectroscopy is optimal for these applications.[6], [27]

Process analyzers are used for in-line or online measurements in order to provide real-time process data while operating in the harshest conditions. These analyzers can therefore be used as process controllers if prior correct calibration. For this propose is required static, fast and rugged instruments that can perform numerous readings per second without being sensitive to process vibrations. Diode array spectrometers can be used in this type of applications.

The versatility of this technology increases the potential of NIRS. Nowadays it can used to determine a variety of quality parameters in a large number of biological materials, such as grains, feed, meat, vegetables and fruit. NIR can be applied, not only to food or agriculture industry, but also to a variety of
different fields such as chemical or oil industry. Such wide range of applications, raises the need for NIRS to be regulated with guidelines from European Agency for the Evaluation of Medicinal Products (EMEA/CVMP/961/01) and the Pharmaceutical Analytical Sciences Group (PASG). [7] Of course none of NIR applications are possible without the fundamental support of chemometrics and multivariate analysis methods.[4], [7].

2.4. Multivariate Analysis

As seen in the previous chapter Lambert Beer’s Law is not valid for scattering samples. However the NIR spectrum can contain hidden patterns and relations between the sample’s attributes and the generated spectral data. On this case, using univariate analysis is not a viable choice. This is because this type of analysis deals with one variable at the time in order to understand each roll of a target parameter on a data set. Such analysis generates multiple results that can be redundant or even contradictive. As an alternative, Multiple Data Analysis (MVDA) searches for variables with similar patterns and groups them in fewer variables. This allows for MVDA to identify and separate meaningful data from data that lacks importance (also known as noise). Therefore, MVDA can:

- Deal with dimensionality problems
- Handle as many variables and observations as required/identified;
- Cope with exact linear relationships (multicollinearity) and missing data;
- Allow a better understanding of processes, and their relationships with variables;
- Recognize process conditions, trends and particularities;
- Realize the effects of the process in the outputs and how to manage inputs in order to improve product quality e optimize process control; [15], [28], [29]

MVDA can be used for classification or quantification purposes. Classification methods are based on cluster and discriminant analysis.[4] The framework of this thesis is centered in quantification purposes. Therefore, in the rest of this chapter, it is described as common quantification methodology. In this methodology, the data is compressed by the use of pre-treatments. Then, models like PCA and PLS are used to reduce the number of variables. These models are validated and possible outliers are identified and removed.

2.4.1. Data Pre-Treatment

Some raw data from analytical instruments need to be modified, in order to obtain the optimal results using MVDA. It is therefore essential to use pre-treatments to account for variations background and/or scatter.
effects. These operations will improve multivariate regressions, classification or exploratory analysis.[15], [30]

2.4.1.1 Scaling methods

Since variables often have different numerical ranges, the ones with a larger range will present a larger variance. Thus, variables with the largest range will present a largest contribution to the model. In some cases, it is therefore necessary to scale the data. This will equalize the impact of each variable on the model. Mathematically, scaling generates X-data (X) (or Y-Data (Y)) by the inverse square diagonal matrix containing the standard deviations for each of the variables in the data (S_M)

\[ X_{Scal} = X(S_M)^{-1} \]  

Another common used pre-treatment is mean-centering. This pre-treatment removes the absolute intensity information of the variables. Mean-centering consists in calculating the average value of each variable and then subtracting it from the original data. The variance of each variable remains untouched. Both scaling and mean centering have a considerable impact in MVDA methods especially in variables with different units. The mean centering (MC) can be described in equation 4 where 1_N is a vector of ones that is N-elements long, and \( \bar{x} \) is a vector that contains the mean response values for each variable.

\[ X_{MC} = X - 1_N \cdot \bar{x} \]  

2.4.1.2 Scatter Correction

As seen in Section 2.3, NIR data is affected by scattering effects. In one hand, when dealing with chemical parameters, these effects can add a significant amount of noise and in most cases, its removal improves the quality of the MVDA. When dealing with physical parameters, scattering corrections remove important information and its use is not recommended. This is because, these methods are designed to reduce the physical variability between samples due to scatter.

Multiplicative Scatter Correction (MSC) is a common pretreatment that can deal with scattering effects. It removes artifacts or imperfections (e.g. undesirable scatter effect) from the data matrix using a calibration spectrum of a calibration data set. This correction method assumes that any sample spectrum can be simply estimated as a multiple of a white reflector (used as reference spectrum), plus an offset. Therefore, MSC method assumes that offset and multiplicative spectral effects are much larger than effects from changes in chemical composition of the sample. Such limitation places MSC in a method that can lead to poor modelling results if the sample and reference spectral data present significant chemical differences. MSC starts by decomposing the spectral data in the reference spectrum and in a multiplicative (a) and additive (b) correction factor.
\[ X = aX_{ref} + b1_N \]  

(5)

Where \( X_{ref} \) is the data obtained from the reference spectrum and \( 1N \) is the vector of ones that as the length equal to the number of variables. Using linear regressions, the correction factors \( a \) and \( b \) are estimated and then used to create a new corrected spectra (\( X_{MSC} \)).

\[ X_{MSC} = \frac{(X - b1_N)}{a} \]  

(6)

The main challenge in MSC is to define an appropriate reference spectrum. Moreover, since reference and sample spectra not always present a linear relationships, it is necessary to decompose the X-data in more elaborate functions, known as extensions of MSC (EMSC). Such extensions are designed to provide improved pre-treatment results in cases where spectral effects of sample chemistry are relatively large.

Another common pre-treatment in spectral data is Standard Normal Variate Transformation (SNVT). Like MSC method, SNVT performs both an additive and multiplicative adjustment. However, the determination of factors is performed differently. For each sample’s spectrum, the offset adjustment is simply the standard deviation of the values over all the variables.

\[ X_{SNVT} = \frac{X}{a_1} = \frac{(X - x\bar{1}_N)}{\sqrt{\sum_{j=1}^{N}(x_j - x)^2 / N - 1}} \]  

(7)

Where \( N \) is the number of variables in the spectrum and \( x_j \) is the value of the variable \( j \). The parameters \( a_0 \) and \( a_1 \) are respectively the average value of the sample spectrum to be corrected and the standard deviation of the sample spectrum. If only considered the standard deviation (\( a_0 \) is equal to zero) the method is known as Normalization. Unlike MSC, both SNVT and Normalization only required the given spectral data so each observation is processed on its own, isolated from the remainder of the set. Normalization can be used at both the beginning and end of correction, although it is easier to evaluate its effect when it is performed prior to any other operation.

### 2.4.1.3 Spectral derivatives

Derivatives can be used as a pretreatments to remove offset and background slope variations between samples. This is only possible to apply to variables that are expressed as continuous physical property (e.g. spectroscopy data) where the property is a wavelength. The first derivate removes effectively only baseline offset variations. The second derivate additionally removes differences in baseline slopes between spectra. This last one is particularly important in NIR spectra due to the diffuse reflectance effect. This second derivative can attenuate this effect highlighting more relevant information in the spectra. Derivatives of higher orders have no relevant effect on the spectra which makes them irrelevant. This is because as the order increase, a larger amount of information is lost. Spectral derivatives can be applied using using Norris Williams (NW) and Savitzky-Golay filters (SG). [6]
In NW derivation, the spectral data is previously smoothed by averaging over a given number of consecutive points. The first order derivation takes the difference between two smoother values with a given gap size. The smoothing step in the first derivative is inherited on the second order derivation which is also smoothed. This method avoids noise inflation when deriving the data.

The SG method starts by using polynomial fitting and then derive this generated polynomial. Regarding the discrete nature of the spectral data, this polynomial is an approximation to the continuous behavior of the spectral wave. After the fitting, the parameters of this polynomial are calculated and the analytical derivation allows to obtain a derivative at any order. [4], [7], [15], [29]

2.4.2. Models

A proper data preparation is followed by a data compression, by the use of models. The simplest way, for quantification analysis, is using Multiple Linear Regression (MLR). This technique it is an extension of linear regression and considers all variables as independent. It was shown to be rather effective during the early years of analytical NIRs but it is based on the univariate inverse method. As mentioned above, univariate methods conditions the interpretation of spectral data since correlation between the data is not considered. This raises the need for other MVDA models. The following section will give a basic overview of some of these models.

2.4.2.1 PCA

The ultimate goal of the use of models is to reduce the number of variables. Such can be achieved, either by selecting some variables and discarding the remaining ones, or transform all variables into linear combinations. This can be achieved with Principal Component Analysis (PCA). PCA is a multivariate projection method design to extract and display the systematic variation in a data set. PCA transforms an original set of correlated variables to a new set of linear combinations of the original variables, known as Principal Components, PCs. These PCs are represented as orthogonal vectors. Statistically, this analysis will find directions in the given set of samples of biggest variation and inserts new axis, i.e. a new Principal Component according to the least squares sense. This means that each sequential PC is determined such that it explains the most remaining variance in the data set. Therefore it is expected that the initial PC contains the largest amount of variability inside the data set (maximum variance) [4], [7], [15].

PCA not only simplifies the interpretation but also the identification of patterns, such as groups of observations, trends and outliers. Mathematically, the PCA implementation in a data set (matrix X) can be described in equation 3.

\[
X = T_{PCA}L_{PCA}^T + E_{PCA} \tag{8}
\]
The data set matrix $X$ is decomposed in a scores $T$, the transposed matrix of loadings $L^t$ and residuals $E$. The scores matrix $T$ shows how data from the initial variables is projected in the space defined by the generated PCs. The contribution of each variable in the new space is exposed in the loadings matrix $L$. Compressing the data requires a number of PCs lower than that of the initial variables. This necessary involves ignoring a small fraction of variation of the original data $X$. This data faction is presented in the residuals matrix $E$. The reduction to an optimal number of PC is the main problem in data compression methods. As shown in Figure 9, using too few components leads to [4], [15], [29]:

- Underfitted model - the model is not complex enough to comprehend the important variability of data which leads to an increasing model error.
- Overfitted model – the model becomes too dependent on the number of samples. This reduces the variability which leads to an increasing estimation error. The model becomes extremely sensitive to uncalibrated conditions.

![Figure 9 - Variation of the error of prediction with the increasing number of components.](image)

Therefore it is necessary to reach a balance between “the need to explain as much of the original data” with the “need to avoid incorporating too much noise into the model”. [4], [15].

### 2.4.2.2 PLS

Once obtained a suitable PCA, it is possible to use a regression step, known as Principal Components Regression (PCR). As a regression step, the primary objective is to correlate a matrix data $X$ to a matrix $Y$. In PCR, the matrix $Y$ does not influence the $X$ data of PCA. This raises the problem of deciding which components to use in the regression equation.

The Partial Least Squares (PLS) was created exactly to avoid the dilemma. Instead of using the PCA compressing step and attribute a corresponding data $Y$, PLS uses both matrix $X$ and $y$ from scratch to create components. Such PCs are obtained by maximizing the covariance between $y$ and all possible functions of
This leads to components, which are more directly related to variability in $y$ than PCs. PLS can be defined mathematically, just like PCA, with an additional equation that considers the $y$-data

$$X = T_{PLS}L_{PLS}^T + E_{PLS}$$  \hspace{1cm} (9)

$$y = T_{PLS}q^T + f$$  \hspace{1cm} (10)

The data set matrix $X$ is decomposed in a scores $T$, the transposed matrix of the independent loadings, $L^T$ and residuals of $X$, $E$. This is taken into account the $y$-data that is decomposed in scores $T$, the transposed matrix of the depending loadings, $q^T$, and the residuals of $y$, $f$. There are several different PLS algorithms such as the non-linear iterative partial least squares (NIPALS) or SIMPLS. Depending on the algorithm choose, PLS can be condensed in regression coefficients. As the $y$ variance is taking into account with PLS, the data compression can result in an increasing complexity of this model. Thus, is expected that PLS tends has a great potential to overfit the model, especially if the if the $y$-data contains a considerable amount of noise. [4], [15], [31], [32]

### 2.4.3. Model Validation

#### 2.4.3.1 Error quantification

After building the model, it is necessary to test its ability to predict unknown $y$-values. This will validate the model and evaluate its viability for the stabilized purpose. Validation will verify the model's robustness and reproducibility. To validate models, it is possible to use common linear parameters like correlation factor ($R^2$). This parameter takes into account of the range of $y$-values used. This unitless factor only provides information about the correlation of the data with the model. For further analysis is necessary to calculate the error estimation, known as root mean square error (RMSE). This error is based on the average difference between actual and predicted $y$-values and the advantage of having the same units of the observations. The model fit will be increasingly good as RSME value decreases. RSME can be applied to either the calibration data set (RMSEC) or to the predicted data in a validation set (RMSEP). When applied to the calibration data, RMSEC can be described by:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N - A - 1}}$$

Where $y_i$ is the known $y$-values, $\hat{y}_i$ the model-estimated property value for the sample $i$, $N$ the number of calibration samples and $A$ is the number of components of the model. When building the model, the $y$-data used for validation can be included in the model (internal validation) or used exclusively for testing (external validation).
2.4.3.2 External validation

External validation relies on the principle that if samples used in the model are sufficiently representative then by adding external samples, the model should be able to predict them. By using these samples it is possible to calculate the root mean square error of prediction (RMSEP) can be described by the equation:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N_p} (y_i - \hat{y}_i)^2}{N_p - 1}} \quad (12)$$

Where $N_p$ is the number of external validation samples. RMSEP measures accuracy and estimates the prediction ability of the actual predictor to be used. In NIR spectroscopy, another prominent performance measure is the standard error of prediction (SEP). This statistical factor measures precision and it can be obtained by the equation.

$$\text{SEP} = \sqrt{\frac{\sum_{i=1}^{N_p} (y_i - \hat{y}_i - \text{bias})^2}{N_p - 1}} \quad (13)$$

And bias can be calculated with:

$$\text{bias} = \sqrt{\frac{\sum_{i=1}^{N_p} (y_i - \hat{y}_i)^2}{N_p}} \quad (14)$$

SEP and RMSEP can easily relate with the equation.

$$\text{RMSEP}^2 = \text{SEP}^2 + \text{bias}^2 \quad (15)$$

2.4.3.3 Internal Validation

Instead of adding external samples, it is possible to use calibration samples to validate the model. This methodology is known as internal validation. The most internal validation technique is cross-validation (CV). When using CV, one or more internal validation procedures are performed. Each procedure (also known as sub-validations) involves the removal of a part of the calibration data. The remaining data is used in a subset of calibration model. Since there are numerous combinations in the removal process, CV will present different types. The whole procedure will cease when all the predefined combinations are performed. The most used types of CV are:

- Selected subset: a single sub-set of calibration data is manually removed for prediction.
- Leave-one-out: in a series of a predefined number of sub-validations, one sample in each sub-set is removed for prediction.
- Random: in a series of a predefined number of sub-validations, a random number of samples in each sub-set is removed for prediction.
• Block-wise: similar to Random but the samples, to be removed for prediction, are contiguous sample blocks.
• Alternating sample: also similar to Random the samples, to be removed for prediction, are in a pre-specified ordinal number of each subset.

Like external validation, CV also uses RMSEP. The difference is that this error is determined from each sub-validation experiments. The averaged RMSEP's from all sub validations translates in a root mean square error of cross-validation (RMSECV). Although useful, CV can be influenced if the selected subset of validation samples is not representative of the samples in the rest of the data. Also if there are too many replicates, both calibration and predictive sub-sets of data can englobe a copy of these values. This makes the CV performance decreases significantly. CV ensures that the model developed does not capture variance specific to the calibration set unrelated to the variation of the parameter of interest.[4], [15]

2.4.3.4 Outlier Detection

When building a model, it is often that not all observations will fit in the model. The observations that present a different behavior from the rest of the data are known as outliers. Outliers can be variables, X-data (x-outliers) or y-data (y-outliers) and can be found either in calibration or prediction data sets. X-outliers are those x-vectors that in some way are abnormally positioned relative to the majority of x-data and y-outliers are defined as those observations that have a different relationship between y and x.

Outliers’ existence can be explained mostly by errors or incorrect actions like formatting errors, failure in instrument of acquisition of data or non-representative sampling. A sample can also be identified as an outlier if it is statistically irrelevant (e.g. if the sample lies more than a few standard deviations away from the mean in a Gaussian distribution). The main objective of outlier detection to decrease or eliminate the effect of non-fitting data points and allow the remainder to determine the best results.

Objective detection of outliers depends on the level of its influence on the model. This can be detecting using leverage and residual statistics. Leverage statistics reflects the extremeness of samples or variables within the model. In contrast, residual statistics can be used to detect samples and variables that have a high amount of variance outside the model. Residuals are a good indicator for both x and y-outliers since they indicate the amount of variance outside the model. Residuals also vary according to the number of PCs which makes them a good indicator for optimizing the model. As there is no actual rules for distinguish the outliers from samples to retain, most authors use multiples of the residual standard deviation to objectively select outliers. [4], [15]
3. Membrane technology and Production Process

3.1. Membrane Types and Processes

A membrane is a selective barrier between two phases. This barrier can be classified according to its thickness, homogeneity, type of transport within it and all the factors that influence this last phenomenon. Membranes can also be natural and synthetic. Synthetic membranes can be organic or inorganic and they are widely used in medicine and bioprocessing. The most important classification in this type of membrane is morphology since it defines its application. Therefore membranes can be either symmetric or asymmetric. Asymmetric filter membranes combine the high selectivity of a dense membrane with the high permeation rate of a very thin membrane. The flow rate increases with the asymmetry, since the layer of small pores is thinner. The flow may be orthogonal to filter (known as dead-end filtration) or tangential (cross-flow filtration). [33]–[35]

The resistance towards the mass transfer is determined mainly by the total membrane thickness and porosity. Membranes can be made in a wide variety of pore sizes according to the final application (Figure 10). [33]

![Figure 10 – Membrane processes and their separation characteristics](image-url)

For the framework of the present thesis, microfiltration is especially relevant. Microfiltration membranes have pores in the 1 – 10 µm range and can be characterized using integrity tests. There are various forms of integrity tests such as: using scanning electron microscopy (SEM), bubble-point (BP) method, mercury
intrusion porometry and permeation measurements. Industrially, the BP method is the most used integrity test. The BP method will be further discussed below.[33]

3.2. Membrane Production Process

In casting machine, the membranes are produced in a precipitation reaction. The starting material is a casting solution composed mainly by polymers of polyethersulfone (PES). After prepared, this solution is poured into a metal belt where the formation starts. The formation process varies with the type of produced membrane. Although the core process be can divided into two steps: conditioning and precipitation (Figure 11).

During the conditioning, the casting solution is spread on a band that contacts with a gas stream saturated with precipitating agent (non-solvent). The membrane formation rate increases with the concentration of precipitating agent in the stream. The membrane pores are formed in this step. As the precipitation agent condenses in the membrane, the solvent is evaporated. The band belt turns in a gyratory drum and submerges in a precipitation bath, initiating the precipitation step. This precipitation bath is constituted mainly by organic compounds. The asymmetric structures of pore sizes are mainly formed in this step. In this process step, the solvent is washed out and the symmetry of the pore size is significantly decided [1]. After the conditioning step, the membrane has to be stretched during the whole process, since if wrinkled, it can lead to its disruption. The process proceeds with the membrane cleaning, drying and finally, rolling up. After rolled up, the membranes are sent to quality control. Additionally the membranes are examined in a screening projector.

![Figure 11 – Detailed description of the conditioning and precipitation step. As the membrane is formed the solvent in the casting solution is evaporated and a non-solvent is condensed in the membrane.[1](image)](image)

After the production, membranes are stored and used for various applications. For the filter cartridge assemble, the membranes are folded and subsequently welded and inserted into filter housings. These filters are used to avoid contamination in biopharmaceutical production processes, particle reduction and sterile filtration. They should be stored in a dust-free place not above room temperature. This storage must
avoid exposing them to sunlight, solvent or other chemical vapors. The PES membranes can be sterilized by autoclaving (121-143°C, 1-2 bar, during 30 to 60 min in wet environment), sterilization with ethylene oxide, γ-irradiation or by chemical disinfection. [36]

In casting machine, the membrane characterization relies on the size of the retained particles, filtrate flow and operational lifetime. These characteristics can be determined by several quality attributes such as pore size, its distribution and porosity. Morphologically, the membrane has different sides, since one was formed in contact with the steel belt and the other one with the air. The air side is smoother in texture since it is relatively less porous. Also the membrane is not homogeneous in its width (considering the length as the length of the roll). Thus, in integrity tests, the membrane is divided into equal parts. The NIR is placed in center of the membrane (in length).

The total length of the process is variable since during the cleaning step the membrane passes through a specific number of vertical boards. The membrane rolls in a semi-continuous process at a certain speed, until rolled up in the end of the production line.

Every roll change is followed by a sampling of the current membrane. The wind-up of 1 roll takes up to 2 hours, depending on the casting speed. The casting speed defines all the speed of the production line.

The quality of the membrane is defined in the beginning of the process. However it is controlled at the end of the process with delay times between 45 and 120 minutes. Having in account that every OOS roll has to be entirely discarded, it is obvious to conclude that the monitoring of these CPPs constitutes a major improvement in this production process. Both CPPs and CQAs of this production process are identified in the following sections of this chapter.

3.3. Critical Quality Attributes

In order to ensure a desired product quality in a process, certain physical, chemical or biological properties must lie within certain limits, ranges or distributions. These responses known as critical quality attributes (CQAs) justify the product value for the consumer.

In casting line the CQAs are related to thickness and permeability. These parameters are measured in quality, and hence with a large time delay relative to the moment of the membrane formation. In this section these parameters are presented.

- **Bubble Point (BP)**

The membrane pores and its distribution constitute a fundamental attribute in which lies the capacity of retention. A good away to evaluate the retention can also be by detecting the lack of it, i.e. quantifying the flow that passes through diffusion. The BP method is based on this principle. This method starts by wetting
the membrane. The wetting liquid is held inside the porous structure by capillary force. This force increases as the pore size decreases. A specific gas pressure dependent on the pore size of the membrane is necessary to force out the liquid from the pores. This pressure is generally designated as the bubble point (BP). It indicates the largest pores of a filter since liquid is first expelled from them. According to Laplace it is possible to relate the radius of a capillary shaped pore \( r_p \) and the pressure applied \( (\Delta P) \).\textsuperscript{[33]}–\textsuperscript{[35]}

\[
 r_p = \frac{2\gamma}{\Delta P \cos \theta}
\]

Where \( \gamma \) is the surface tension at the liquid/air interface and \( \theta \) the contact angle. As the \( \gamma \) is relatively high for water/air interfaces (72.3x10\(^{-3}\) N/m) it is necessary to apply high pressures. On the other hand, \( \gamma \) is lower for alcohols, which makes these substances good alternatives if higher pressures are difficult to achieve. The awareness of the largest pores of a filter allows defining the membrane application (i.e. microfiltration) which makes it one of the most important CQA. If the BP is below a defined limit, it can no longer be ensured that a filtration process will do the job intended (e.g., remove all bacteria in sterile filtration).\textsuperscript{[33]}

![Figure 12](image)

**Figure 12** – (1) Variation of the flow with gas pressure. As the pressure increases, the flow increases linearly (Diffusive flow) until the BP. Higher pressures indicate higher flows (Bulk flow) (2) - Scheme of BP-method at the membrane scale (grey shapes) with the wet filling (blue shapes). (a) At room pressure, the diffusion through the wet membrane is limited. (b) As the pressure increases, diffusion increase. (c) Once obtained the BP pressure, the bulk air flows through the membrane without the impediment of the wet liquid.\textsuperscript{[1]}

- **Diffusion** (Df)

When pressure is applied to the upstream side of a wet membrane, a diffuse gas stream starts to flow through the filter membrane after a short time. The diffusion (Df) is the value of flow at the BP. It is measured in flow units (ml/min).

- **Thickness** (Tk)

The membrane thickness (Tk) can also be quantified and it is also an indicator of the final product quality. Generally, the lower the Tk, the weaker the structure of the membrane, which makes it more prone to break.
On the other hand, increasing the thickness will eventually form a second layer which is not beneficial to the quality of the membrane. For this reason, it is an important CQA and it can be measured in µm.

- **Air Flow (AF)**

When exposed to a certain amount of air, the membrane will generate a specific amount of resistance that also depends on its structure. This resistance can be traced by measuring the pressure of the air flow (AF) through the membrane. This air flow can be measured in volume pressure units (L/(m².s)) and it is an indicator of the membrane permeability.

### 3.4. Critical Process Parameters

All of membranes' attributes result from the influence of critical process parameters (CPP) in both the conditioning and precipitation steps. These CPP are manipulated by the operators from a local control system. As these variables are industrial process based, their regulation uses Proportional-Integral (PI)/Proportional-Integral-Derivative (PID) controller types. Therefore every change of parameters will make the measured value revolve around the set-point until the difference between two consecutive values is minimal. This stabilization process is time consuming and since the process is semi-continuous, some membrane meters have to be discarded. The CPP used in this part of the project are presented in this section. Their influence on the process can be observed in Figure 13.

- **Temperature of the Band (T\text{Band})**

The temperature of the band is measured in degrees Celsius (°C). Due to prior experimentation it is considered the most important CPP in membrane production. This is verified due to its strong connection with the most relevant CQA, BP. The different temperature of the belt defines the temperature of the coating solution during the precipitation. A high T\text{Band} leads to a better absorption of the precipitation agent (Figure 14). This results in the formation of larger pores which leads to lower Bubble Point and Air Flow values.
The gyratory drum is emerged from the precipitation bath. The temperature of the band is regulated with three different parts by a local cooling/heating water system. The casting solution is charged in the band in the second zone.

- **Temperature of the Precipitation Bath** ($T_{PB}$)

The $T_{PB}$ affects the pore size in the band side of the membrane. High temperatures leads to a slower precipitation process and thus to larger pores. Unlike the $T_{Band}$, increasing $T_{PB}$ shapes the pores in a funnel-shape like morphology. This can produce high flow and membrane capacity.
Figure 15 – Effect of an increasing $T_{PB}$ (direction of the arrow) on the pore size and morphology.

- **Casting Speed ($CSp$)**

The speed of the casting ($CSp$) process is fundamental since it defines the velocity of the whole process. It is connected with $T_k$ in combination with the amount of casting solution. $CSp$ also defines the precipitation time and all subsequent steps (e.g. temperature of drying). According to each membrane type, the casting speed and the number of boards in the process, the total time of production may vary. Nevertheless, each roll can carry 180m of membrane which make the changing times between rolls dependent of the casting speed only (equation 16). $CSp$ is measured in m/s.

$$t_{Roll\ (min)} = \frac{180}{CSp} \times 60 \quad (17)$$

- **Amount of Casting Solution (ACS)**

A controlled flow pours the casting solution onto the band. This flow is time dependent just like the casting speed. This means the control process has to, sometimes, compensate the deviations in casting speed in order to guarantee a fixed amount of casting solution (ACS) in the band. Such is achieved by adjusting this solution flow. The amount of casting solution on the band is strongly dependent on the thickness of the membrane. It is measured in kg/m².

- **Temperature of the Column ($T_{col}$)**

The organic solvent from the precipitation bath is regenerated in a distillation column. This column separates the organic solvent from the other precipitation bath components at a certain temperature. The sequent stream of this solvent is recycled into the process. Just like in the $T_{PB}$ case, high temperatures of organic solvents leads to a slower precipitation process and thus to larger pores.

- **Coverage ($Cov$)**

Finally, the contact time of the organic solvent gas stream and the membrane can also be an interesting variable to use in the DoE approach. This time can be shortened by adding a covering plate between the membrane band belt and the gas flow streaming feed. There are 3 different coverage plates which makes this variable primarily quantitative. Although what differs from plate to plate is their size, so it will be worked
as a qualitative parameter. This will significantly reduce the amount of experiments in DoE. This is expected to have the same effect as $T_{PB}$ and $T_{col}$ in the membrane porous structure.

3.5. Reference Analysis

The membrane referencing, or measuring, consists in the detailed identification of the CQAs of every single membrane. This identification will use different measuring tools depending on the CQA. Inside each CQA, there are also different options that translate in different levels of precision of the measurement. The measuring tools relevant for the framework of this project are presented in this section and the respective photos are in Appendix A.

- **Sartocheck® 3 Plus**

Sartocheck is a membrane filter integrity test system from Sartorius, which is used among other things for the identification of defective cartridges in the process (Appendix A – Figure 109). It is used at-line in the process to measure the BP and Df. The setup includes a hardware control device and a stainless steel in-line filter holder where the membrane is placed after previously being cut in a circle with 243mm of diameter. Sartocheck® 3 Plus can measure Df from 0,1 - 999,9 ml/min and BP from 100-8000 mbar. The accuracy of these measurements is ±5% of the measured value for Df and ±50 mbar for BP. [37]

- **Manual Measurements**

The BP can be measured manually with the structure showed in Appendix A – Figure 110. The graduated cylinder is filled with water in the beginning of the experiment and the filter membrane subjected to an air flow with increasing pressure. When the first air bubble appears in the membrane surface, the correspondent pressure is the bubble point. The accuracy of this measurement depends on the rate of pressure change. The membrane’s flow rate can also be measured with the same device (with e.g. water).

- **Caliper (Prüfmittel)**

The thickness can be measured with a portable caliper by placing the membrane between two small plates (Appendix A – Figure 111). A normalized pressure is applied on the surface to increase reproducibility and to reduce variations due to different handling. Although fast and easy to use, this equipment is not very sensitive. The uncertainty associated with this thickness is ±5 µm.

- **Scanning Electron Microscope (SEM)**

In Scanning Electron Microscopy, electrons are usually accelerated to high energies in a beam that hits a sample. Unlike Transmission Electron Microscopy (TEM), the interaction between the sample and the electron beam is not transmitted through the sample. Instead it generates emerging electrons that can be
detected.[1], [38]. The generated information can be used to map the topography of the membrane. The resolution of the data obtain is limited to the diameter of the beam. Although this technique produces very accurate images of the membrane, the vacuum generation and sputtering are time consuming and not feasible for a large amount of samples.

- **Digital Optical Microscope (DOM)**

The digital optical microscope (DOM) is a variant of the standard optical microscope that can be used to obtain digital images of the samples (Appendix A – Figure 112). This method is a good alternative to SEM, since it can measure small deviations in membrane thickness with an error ±0.5 µm (see below how). Also this method does not require the vacuum generation and sputtering which makes it far less time consuming.

- **Permeability device**

The air flux can be measured with an air permeability device that can also measure roughness (Appendix A – Figure 113). In this device, the operation starts by placing the sample below the measuring head. The measuring head lowers onto the sample and clamps it. Air flows through the sample at a selected pressure. As soon as the preset flow rate is stable, the measuring value is displayed. The test time is adjustable from 1-40 seconds. This non-destructive test allows the same membrane to retain its main functionality. The error of this measurement is ±5% of the measured value for Df. [39]

### 3.6. NIR Spectrometer

The NIR spectrometry technology was first introduced in the casting machine in the project “Automation of Casting Machine”. As mentioned before, this device was placed in this production line in order to monitor the process. The NIR used is a BioPAT® Spectro, former PMD500 - NIR spectrometer (Appendix A – Figure 114). Process analyzers such as this one, are Sartorius Original equipment manufacturer (OEM) systems for online monitoring of CPPs.

The BioPAT® Spectro is a multichannel spectrometer that uses a polychromator after the beam hits the sample. The polychromator outputs multiple beams over a range of wavelengths simultaneously. Also has a concave diffraction grating, which functions as both as diffraction grating and a converging mirror. After being diffracted, the light is detected in a diode array (Figure 16). The number of photodiodes in this linear plate will determine the resolution of the spectrum. [24]
A good spectrometer to implement should be small in size and easy to assemble in order to minimize the interaction with the process. It should also be placed in a well-accessed area for maintenance purposes. The placing of this sensor is also crucial for the definition of which CPP is taken in effect in the spectral data. As mentioned before, the middle of the membrane (3rd zone) is more representative of the membrane’s CQA in quality control. The NIR is placed in the center of the membrane, relatively to the width of the casting machine. In the beginning of the project, the BioPAT® Spectro was introduced adjacent to the pouring of the casting solution and it detected the chemical and physical properties of the precipitating membrane. Later it was placed at the end of the belt where the membrane formation is almost finished. This made the spectrometer shift further from the conditioning step, to the beginning of the precipitation step. In this zone, the precipitation is already more advanced, so it could be possible to correlate the spectral data to the quality attribute \( T_{PB} \). Also the optics was optimized to detect more light for a distance NIR-belt of approximately 8 cm. As shown in Figure 17 for membrane 15407, the spectral peak resolution increased and the mean absorbance dropped from 1.5 to 1. By using the device in this new position, the intensity of the signal was increased eightfold compared to the reference spectrum. Using an optimized optic for BioPAT® Spectro, the signal quality could be significantly improved. [1]
4. Experimental Procedure

4.1. Project Framework

The “Automation of Casting Machine” is a project started in 2010. The project was initiated for automatic model building and to display the NIR results in the process control system. As shown in Figure 18, the project is constituted by several stages. First the operator has to sample the membranes according to the different DoE states (step 1). These samples will be characterized by a time stamping (stage 2). Furthermore the samples are referenced and added to a database (stage 3). This data is selectively associated with NIR spectral data and exported for calibration (stage 4 and 5). For the next stage, the objective is to repeat the previous stages for each different membrane type (stage 6). Finally, the predictions / statistical parameter are exported in the casting line’s control system (PCS) to show process stability. At the current point, the last accomplished step was converting the data into an NIR command.

Figure 18 – Stages of the project “Automation of Casting Machine”. Current stage is marked with a yellow circle (6. Automatic model building for each membrane type). [1]

It is now necessary to build a different model for each relevant membrane type. Such objective can be achieved by using a methodology as seen of Figure 19:

- The prior knowledge is used in RA to define CPPs;
These CPPs are used to build a DoE Map that is implemented in the production line;
- This will generate membrane samples and NIR data;
- The samples are referenced and used for evaluation of CPP/CQA relations;
- CQAs are then merged with the NIR spectral data in experiments.
- Through MVDA, these experiments are used to create a predictive model that can be implemented in the PCS of the casting machine.

The following chapter will describe in detail the experimental procedure applied in the 4 membranes.

![Diagram](image)

**Figure 19 – Overview of the thesis objective.**

### 4.2. RA and CPP definition

Since Sartorius produces several kinds of membranes, it is relevant to build different models for each membrane type. These types are produced with distinct values of CPPs. Therefore each DoE map will have a unique structure with only the most important set of CPPs for each membrane. It is also important to notice that a relevant CPP for a type of membrane, can be irrelevant for another. Such fact can change the number of variables in the system, which will be determined for the choice of model design and subsequent number of experimental trials.

As seen in Section 2.1, it is necessary to fully understand the process in order to establish the correct ranges of CPP’s variability. These decisions are made according to prior process/membrane technology know-how allied with PAT’s RA. After deliberating, it was chosen 4 types of membranes. The respective CPPs used in DoE are summarized in Table 1.
### Table 1 – Membrane CPPs used for DoE building for each membrane.

<table>
<thead>
<tr>
<th></th>
<th>$T_{\text{Band}}$ (°C)</th>
<th>$T_{\text{PB}}$ (°C)</th>
<th>$C_{\text{Sp}}$ (m/h)</th>
<th>ACS (kg/m$^2$)</th>
<th>$T_{\text{col}}$ (°C)</th>
<th>Cov</th>
</tr>
</thead>
<tbody>
<tr>
<td>15458</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15457</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>15404-H</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15407</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

### 4.3. DoE Building

The DoE building is a very important task in this step of the project, since it defines all the data that it is going to be used in the model. The objective in these DoEs is to build quantitative models that are robust and independent from normal process variations. Since the calibration file will be produced with other software, the only contribution of this step will be the map itself.

As mentioned above, each membrane will have a different DoE. This will yield slight changes CPP-CQA relationships in each membrane type. As referred in section 2.2.2, further analysis would include optimization and robustness tests. This implies to build a different DoE and introduce the results, for each of these steps. These protocols would not be viable in this work because:

- Each DoE for each membrane must be entirely performed without interruptions. Such requires an exclusive use of the production line. Therefore the membrane manufacturing schedule is affected. This limits the time in which a DoE can take place and raises the need for time optimizations;
- The objective of each DoE is not to obtain optimum ranges of production. Instead, it will only produce statistically relevant deviations in CQAs/NIR spectral data that are useful for the model building.

Having these issues in mind, all models were built using screening analysis in order to produce the DoE. As mentioned above, screening identifies the most relevant experiment contributors and leaves out factors that can also contribute with noise. This will be sufficient to analyze CPP-CQA interactions and for the fusing between CQAs and NIR data.

Each DoE requires a custom made design, especially for time optimization. This makes the Full Factorial design the best initial choice. This design may be redundant but it ensures correct models and guarantees the main influence of CPPs and CQA-CPP interactions are revealed.

Once selected this design, each DoE were selectively changed by adding/removing more convenient points. In some cases, it was relevant to highlight the impact of $T_{\text{Band}}$, since this CPP has the most influence in the most important CQA, BP.
4.4. DoE Rearranging

The production line has to stop its normal production during the DoE’s trials which make the time an important factor. This can be a challenge since the DoE is based on the change of several set points of CPP in the production line. Not all the changes to new stationary states in these CPPs take the same time, so it was necessary to define an average time for each one of them based on prior process know-how (Table 2).

Table 2 – Time changes /unit for all the CPPs used in the DoEs for the 4 membranes

<table>
<thead>
<tr>
<th>CPP</th>
<th>Time changing/unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB (°C)</td>
<td>20</td>
</tr>
<tr>
<td>TPB (°C)</td>
<td>10</td>
</tr>
<tr>
<td>CSp (m/h)</td>
<td>0,25</td>
</tr>
<tr>
<td>ACS (kg/m²)</td>
<td>Instantaneously</td>
</tr>
<tr>
<td>TC (°C)</td>
<td>30</td>
</tr>
<tr>
<td>Cov</td>
<td>0,5</td>
</tr>
</tbody>
</table>

The DoE maps were rearranged depending on the minimum time transition between each experiment state. This will reduce the amount of discarded membrane. It will also maximize the number of experiment states reached in a feasible working period, which can lead ultimately to more data and subsequently a better model. This time optimization has to be done with some considerations.

- The process has to start and end in the normal production state (NPS). Since the CPP’s ranges revolve around this NPS, these states constitute perfect CPs (see section 2.2). Although, not all possible symmetric ranges (imposed by the design) are possible and the mean value can be slightly different from the original NPS.
- TB (°C) is the most important CPP and so, it is a priority during the DoE rearrangement. This means that changes in TB are preferential in case of a draw of this CPP with other one.
- For temperature related CPP, heating is generally faster than cooling.
- More time consuming transitions should use lower values of CSp, in order to produce less discarded membrane per time.

In a DoE context, the time minimization is relatively easy to execute with a visual interpretation of points in space. This interpretation is only possible with up to 3 variables (3D). For more variables it is possible to use the 3D interpretation with the 3 more relevant variables and keep the other ones constant.

After obtaining the DoE, it is now necessary to know the values of each CQA that corresponds to the established CPP. This will allows the analysis of CPP-CQA interactions and the model building using CQAs and NIR data. The CQA data can be obtained by referencing each set of membrane.
4.5. Membrane Referencing

Once the DoE is finished, it must be performed in the production line. Each DoE state is introduced in the casting machine by changing the set-points of the selected CPPs. After these set-points are reached, the CQAs of the produced membranes are already affected by these changes. When these membranes reach the cleaning step in the process, they are marked with a pen for recognition in the roll-up. The time of this marking procedure is recorded for the NIR integration, for each membrane.

In the roll-up, each set of membranes is divided in several sets of 10-20 sheets. Each set of sheets will represent a moment where the CPPs were aligned according to the designed DoE. Ultimately, all the sheets within the same set would have the same values of CQAs (Figure 20). Although, since the difference between the set-point and the measured value of CPP is not always zero, there may be some variation between CQA´s sheets within the same set of sheets. In order to reduce these measuring errors within the referencing, each CQA considered is an average of (at least) three measures. Each sheet is 53.8 cm in length and an average of 35 cm wide.

The referencing protocol requires skilled personal that can operate with the instruments. Some of these measurements required previous sample preparation. The referencing process is divided in pre-referencing and refining referencing. The pre-referencing step is facultative for this project and it’s executed at-line by the operators with single measures. This type of referencing is faster and uses instrument settings without taking account for wider ranges of CPP.

The refining referencing is performed later with triplets. The NIR data is obtained from the center of the membrane which makes it the zone 3, the only possible place to extract data attributes. The CQA measured are presented below.

- **Bubble Point (BP)**

The bubble point is measured with manual apparatus and Sartocheck (see Section 3.5). Both instruments are used in pre-referencing but only Sartocheck is used in refining referencing due its higher accuracy. For the manual measuring only a small circle of membrane is required. The Sartocheck requires cutting a larger circle of the membrane with 243mm of diameter (Figure 21). Each membrane will have a specific range of pressure where the bubble point is reached which makes the Sartocheck settings vary within the same experiment set.
Figure 20 – Sets of membrane sheets divided by CPP state.

Figure 21 – Scheme of the referencing of a membrane sheet. The blue bar indicates the area in which the NIR scans the membrane. The circle indicates the cutted piece of membrane required by the measurement with Sartocheck.

- **Diffusion (Df)**

Diffusion is also measured in Sartocheck alongside Bubble Point, both in pre-referencing and refining referencing.

- **Thickness (Tk)**

Membrane thickness was measured in a caliper (Prüfmittel) and in an optical microscope (DOM). The Prüfmittel was used in both referencing moments and the DOM only in the refining referencing. In DOM, the membrane has to be cut in order to measure the transversal side. A standard blade tends to squeeze the membrane and form smooth edges which make the visualization more difficult to quantify and less accurate. Therefore, each membrane is wet and then dipped in liquid nitrogen. The water in the membrane freezes the structure which allows to break it and form a sharp edge without affecting the Tk. The resulting shards are stuck between 2 plates with the thinner face of the membrane exposed upwards to the lens. Then, the Tk is measured.

- **Air Flow (AF)**

The AF measuring is performed in a simple and quick way using instrument PTA-Line Bendtsen. The sheet is placed below a measuring head that lowers and clamps the membrane. The air flows through the sheet at a selected pressure during 2 seconds.[39]
5. Experimental Results and Data Analysis

5.1. Sampling and Datasets

The DoEs of the four membranes were performed in the casting line and generated a different number of samples. For the membranes 15458 and 15457, it were gathered 23 and 24 samples, respectively. For the membrane 15404-H it were only gathered 14 samples. The lower number of samples for this last membrane is explained by the fact that the process failed to reach some CPP ranges, in the middle of performing the DoE. This forced to discard some of the initially planned DoE states. In the membrane 15407, it was only possible to obtained 10 experiments. This is because the stipulated casting solution for the current batch of this membrane was finishing, which reduced time to perform the DoE to only half a day. For model building this was not a problem, since this membrane was already modeled before. Therefore, this last DoE served as a validation procedure of the old model. After performing the DoE, the membranes’ CQAs were measured. As referred in the previous chapter, the final value of CQA is an average between the measurements made within each set of membranes. In some cases, the standard deviation (SD) was too high and some repeated measurements had to be done (data not shown). This guaranteed that there were enough samples for an outlier detection. As one of the main objectives of this thesis, it was now necessary to evaluate relationships between CPPs and CQAs for each membrane. Each set of CQAs was introduced in each DoE on the software Modde. This software generated statistical information about CPP/CQA relations for each membrane. This information is exposed in section 5.2. Furthermore, in section 5.3, the models built with MVDA are presented for each CQA set of each membrane.

5.2. Evaluation of CPP-CQA relationships

5.2.1. Statistical parameters

The software Modde uses a MLR approach to create statistical models with CPPs and CQAs. This will generate different statistical plots. Since the main objective is to analyze and evaluate the impact of CPPs on the process, the more relevant plots are the fit plot and correlation coefficient plots. The fit plots expose 4 different statistical indicators that provide information about the relationship between the data and the model generated by Modde. These indicators are $R^2$, $Q^2$, Model Validity (MV) and Reproducibility (Rep) and their equations can be found in Appendix B. Just like in the linear analysis, $R^2$ parameter is a fraction of the variation of the response explained by the model. $Q^2$ is an estimate of the predictive ability of the model according to cross validation. Model validity (MV) tells about how strong the model is, considering the experimental data. Rep corresponds to the variation of the response under the same conditions, often at the center points, compared to the total variation of the response. All these four indicators are presented in
a scale from 0 to 1 which can evaluate the model. The evaluation proposed by Modde can be seen in Table 3.[22], [40].

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Reference values</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>~1</td>
<td>Good fitness of the data</td>
</tr>
<tr>
<td>$Q^2$</td>
<td>&gt;0,5</td>
<td>Good predictability</td>
</tr>
<tr>
<td>MV</td>
<td>&gt;0,25</td>
<td>No lack of fit of the model</td>
</tr>
<tr>
<td>Rep</td>
<td>&gt;0,5</td>
<td>Good reproducibility</td>
</tr>
</tbody>
</table>

5.2.2. Membrane 15458

The membrane 15458 was modeled with 23 experiments. Two of these experiments are center points. As shown in Figure 22, all CQA except Df, present a $R^2$ approximately or higher than 0,79. In BP, Tk and AF, the $Q^2$ is slightly lower which means the model fitting is slightly higher than the model predicting. In these three, the MV is also high with the lower value at 0,31. When this parameter is higher than 0,25 there is no lack of fit of the model. This means that the model error is in the same range variation of the reproduced experiments. The Rep indicator in these CQAs is also above 0,74. That means that the repeated measures (in this case the CP) present a small deviation compared with the total variation of the response. The model generated with Df is poorly fitted with a low $R^2$, around 0,09. This model has also no prediction capacity with a $Q^2$ indicator below 0. Although MV and a Rep are both above 0,53, the poor model fitting shown by $R^2$ and $Q^2$, leads to conclude that these indicators have no significant impact in the model analysis.

Figure 22 – Summary fit plot for membrane 15458 using 3 CPPs and MLR. Blue bars denote $R^2$. Orange bars represent $Q^2$. Grey bars signify MV and yellow bars represent Rep.
In the coefficient correlation plots show the impact of each CQA in the model. All of the plots for all membranes are scaled and centered in order to unveil CPP-CQA relationships. The confidence interval is also a constant in the following analysis and equal to 95%. The contribution of each model is as high as the absolute value of each CPP. Positive values translates a direct relation between factors and responses. Negative values mean that factors and responses are inversely correlated. Contributions of the interactions between CPPs were not considered in any of the following cases (e.g., $T_{\text{Band}} \cdot T_{\text{PB}}$ contribution to $BP$). This is because in most cases, these contributions were negligible comparing with the contribution of standalone CPPs.

As seen in Figure 23, for this membrane the BP, Tk and AF the $T_{\text{Band}}$ is the factor with more impact. The $T_{\text{PB}}$ has a low absolute coefficient and an error interval that englobes the coefficient itself in both BP and Tk. In these cases it is plausible to conclude that the contribution of $T_{\text{PB}}$ is insignificant. On the other hand, $T_{\text{PB}}$ has a significant contribution in AF.

Figure 23 – Correlation plots for membrane 15458. A is BP (Bubble Point), B is Tk (Thickness), C is AF (Air Flux) and D is Df (Diffusion). Green bars represent each CPP contribution of the model. Black lines represent the associated error. All data used was scaled and centered and error has a confident interval of 95%,
In the considered models CSp has a low contribution with exception of Tk and AF. Since the model Df was poorly fitted to the data, the coefficients and the respective error fail to be statistically significant.

For membrane 15458, the models based on BP, AF and Tk present a consistent validity. The Df data poorly fits in a model. This is possibly because Df strongly depends on input parameters in Sartocheck that are not taken in consideration (e.g. diffusive gas stream). BP negative correlates with T_{Band}. As seen on section 3.4 higher values of temperature of the band produce a membrane with wider pores. This means a lower pressure required to reach the BP state. Simultaneously, wider pores translates a higher flow of air, which explains a positive correlation with AF.

When measuring with DOM, the membrane is stuck between two plates that may slightly compress it. This can affect the membrane’s Tk in the range shown in the coefficient plot. Membrane with wider pores can be more hollow, and thus more prone to be affected by this. This possible explains why thicker membranes show a negative correlation with T_{Band}.

The T_{PB} is only significantly correlated with AF. Having such in mind, it is possible to conclude that the pore formation is mainly performed in the conditioning step. The contribution observed in the AF can be explained by the high sensivity of measurement. This justification is the same for residual contribution of CSp. In this last CPP the high contributor is Tk. For a constant amount of casting solution speed, the faster the process runs, the less solution is cast in the band. So it is expected that this thickness decreases with increasing casting speed.

5.2.3. Membrane 15457

The membrane 15457 was modeled with 24 experiments. Two of these experiments are center points. As shown Figure 24, all CQA except Df, present a R^2 approximately or higher than 0.87. In BP, Tk and AF the Q^2 is slightly lower which means the model fitting is slightly higher than the model predicting. In BP the MV indicator is below 0. This means that the model error is superior of the measuring error. For Tk and AF, the MV is superior of 0.55. The Rep indicator in these CQAs is also above 0.99. That means that the repeated measures (in this case the CP) present a small deviation compared with the total variation of the response. In membrane 15457, the model generated with Df is poorly fitted. Although the R^2 indicator is higher than in membrane 15458, around 0.51. In 15457, this model has also a low prediction capacity with a Q^2 indicator of 0.06. MV is 0.93 and Rep around 0.14. Once again the poor model fitting shown by R^2 and Q^2, leads to conclude that this indicators have no significant impact in the model analysis.
Figure 24 – Summary fit plot for membrane 15457 using 5 CPPs and MLR. Blue bars denote $R^2$. Orange bars represent $Q^2$. Grey bars signify MV and yellow bars represent Rep.

In the models of BP and AF the $T_{\text{Band}}$ is the factor with more impact (Figure 25). $T_{\text{Band}}$ has also impact on Tk, although the ACS is more influent in this CQA. For the BP, $T_{\text{col}}$ and ACS have no significant contribution. For the Tk model, $T_{\text{Band}}$ has a relative contribution followed by CSp and $T_{\text{col}}$. $T_{\text{PB}}$ appears to have no significant contribution. In the AF model, $T_{\text{PB}}$, CSp and ACS have a small contribution. On the other hand, $T_{\text{col}}$ appears to have no substantial influence. For membrane 15457, the models based on BP, AF and Tk present a consistent validity. Like for membrane 15458, Df data poorly fits in a model. Since the model Df was poorly fitted to the data, the coefficients and the respective error fail to be statistically significant. Also, it is possible because Df strongly depends on uncontrolled input parameters in Sartocheck.

$T_{\text{Band}}$ correlates negatively with BP and correlates positively with AF. The possible pressing effect provoked between the two plates mention for membrane 15458, is also visible in membrane 15457 which explains the negative correlation between Tk and $T_{\text{Band}}$. $T_{\text{PB}}$ only correlates negatively in significant values with BP and AF. This strongly suggests that the pore formation process is divided between the end of condition and beginning of precipitation step. In the Tk model, it was expected that CSp would have a more significant influence. Such contradiction can be explained with the auto-regulation of the flow of solution that is poured in the band. Both CSp and ACS can be regulated with this flow. By default, the control system will auto-correct this flow according to different speeds to match the thickness of the membrane. This makes sense in case the process needs to be fasten or slowed conserving the physical without affecting thickness. Hence, to test CSp influence this auto correction needs to be disabled. According to the correlation obtained it plausible to believe that such didn’t happen in this membrane. Although the ACS is not autocorrected which explains the high contribution of this parameter in Tk. ACS positively correlates with Tk because the more casting solution is poured in the band, the thicker the membrane will be.
$T_{col}$ appears to have limited influence in $Tk$ and $Df$, and insignificant effect on the other CQA. In a first glance it is plausible to admit that this parameter has no direct effect on the membrane. Although it is necessary to consider that the range of this CPP is very low comparing to the range inputted in the same type of variable, temperature.
5.2.4. Membrane 15404-H

The membrane 15404-H was modeled with 14 experiments. Three of these experiments are center points. As mentioned and explained below, the analysis of DI was not carried out in the remaining membranes.

As shown in Figure 26 all CQA present a $R^2$ approximately or higher than 0.74. BP appears to have a better indicators with a $R^2$ of 0.93, $Q^2$ of 0.9, MV of 0.83 and Rep of 0.94. In both Tk and AF, the fit is higher than 0.74. the prediction ability above 0.6 and Rep above 0.91. MV is about 0.63 in Tk. However this indicator is 0 in AF. As a result, this model may require a more exhaustive validation process.

![Figure 26 – Summary fit plot for membrane 15404-H using 3 CPPs and MLR. Blue bars denote $R^2$. Orange bars represent $Q^2$. Grey bars signify MV and yellow bars represent Rep.](image)

In the models of BP and AF the $T_{Band}$ is the factor with more impact (Figure 27). The negative correlation between BP and $T_{Band}$ is also seen in this model. CSp has an impact on Tk, although the $T_{Band}$ is slightly more influent. In fact, unlike in 15458 and 15457, $T_{Band}$ corroborates positively with Tk and AF. CSp also appears to have an inversed correlation for AF, when compared with membranes 15457 and 15458. This inversion can be justified by:

- Possible errors in the membrane production and/or referencing measurements.
- Possible different chemical composition of casting solution/precipitation agent;
- Possible specific physical properties for this membrane.

Despite these inversions, BP model appears to show the same type of correlation for all the CPP, comparing with the previous membranes. The negative CSp correlation is also conserved for the Tk model. These facts
suggest that the deviations may have origin on the measurements instead of the process. $T_{PB}$ appears to have no significant contribution for all the CQA detected for this membrane. It is plausible to conclude that the pore formation on this membrane is more related to the conditioning step.

![Figure 27](image_url)

**Figure 27** – Correlation plots for membrane 15404-H. A is BP (Bubble Point), B is Tk (Thickness) and C is AF (Air Flux). Green bars represent each CPP contribution of the model. Black lines represent the associated error. All data used was scaled and centered and error has a confident interval of 95%.

### 5.2.5. Membrane 15407

The membrane 15407 was modeled with 10 experiments. This membrane was the only one with a model already implemented. Therefore, the following analysis serves more as a validation procedure. As shown in (Figure 28) all CQA present a $R^2$ approximately or higher than 0.98. Again in all CPPs $Q^2$ is slightly lower with means the model fitting is slightly higher than the model predicting, with the lowest value at 0.84. MV
is above 0.73 in all cases. This makes the model error in the same range of the measuring error for all the models. The Rep indicator in these CQAs is also above 0.98. That means that the center points present a small deviation compared with the total variation of the response.

![Figure 28 – Summary fit plot for membrane 15407 using 4 CPPs and MLR. Blue bars denote $R^2$. Orange bars represent $Q^2$. Grey bars signify MV and yellow bars represent Rep.](image)

In the models of BP and AF the $T_{Band}$ is the factor with more impact (Figure 29). Actually in the BP model, no other CPP has a relevant impact since the error bars comprise all the contribution. Since $T_{PB}$ has a non-significant contribution in the BP model it is possible that the pore size is not affected in the precipitation step. Like in the models of 15457 and 15458, ACS is highly and positively correlated with Tk. Again, the more casting solution is poured in the band, the thicker will be the membrane. Negative correlation between AF and $T_{Band}$ is also observed in this membrane. $T_{PB}$ shows a slight contribution in both Tk and AF models. Cov appears to have no significant contribution in any of the models.

In general, the models for membrane 15407 appears to have a good fit with the data and also good predictable abilities. It is also worth mentioning that for this models, it was used only 10 experiments. There are 2 center points, which leaves only 8 points for the model building. This explains the high error bars in the contribution plots. Despite such fact, the main contributions are visible only with such few experiments. This suggests a good robustness of the model. A more detailed relationship analysis would require more experiments.
5.2.6. Analysis considerations

After analyzing the membranes it is possible to make some considerations that can be useful in the next step of the model building. The analysis of all membranes is summarized in Table 4 Table 3.[22], [40].

Table 3

Considering the analysis and the evaluation of the results, the main conclusions are presented next:
• Df is not a relevant CQA for the membranes where it were analysed. Both low $R^2$ and $Q^2$ constitute primary indicators of the lack of success in the model building. This is why the analysis were not carried on further membranes.

• The low $R^2$ of BP in 15458 is not significant since it stays right near limit value.

• The low MV in BP for membrane 15457 indicates this model needs a more intense validation procedure.

• BP and AF are strongly correlated with the same CPPs. This suggest that both these CQAs are related. Better error intervals for the AF model suggest that this CQA can be a better predictor. Nevertheless the prediction of the BP, the primary CQA, appears to be more feasible for these membranes types.

• CSp, ACS and $T_{Band}$ appear to be the most influential CPPs in thickness.

Table 4 – Statistic evaluation of each model. $N_{Exp}$ is the number of experiments. Green marked values meet the specifications of a good model exposed in Table 3 ($R^2 \geq 0.80$). Red marked values do not meet such specifications.

<table>
<thead>
<tr>
<th>$N_{Exp}$</th>
<th>CQA</th>
<th>$R^2$</th>
<th>$Q^2$</th>
<th>MV</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>15458</td>
<td>BP (bar)</td>
<td>0.79</td>
<td>0.67</td>
<td>0.89</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>Tk (µm)</td>
<td>0.88</td>
<td>0.82</td>
<td>0.31</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>AF (L/m².s)</td>
<td>0.99</td>
<td>0.98</td>
<td>0.62</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Df (ml/min)</td>
<td>0.09</td>
<td>-0.2</td>
<td>0.73</td>
<td>0.53</td>
</tr>
<tr>
<td>15457</td>
<td>BP (bar)</td>
<td>0.87</td>
<td>0.76</td>
<td>-0.05</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Tk (µm)</td>
<td>0.97</td>
<td>0.94</td>
<td>0.79</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>AF (L/m².s)</td>
<td>0.96</td>
<td>0.92</td>
<td>0.55</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Df (ml/min)</td>
<td>0.51</td>
<td>0.06</td>
<td>0.93</td>
<td>0.14</td>
</tr>
<tr>
<td>15404-H</td>
<td>BP (bar)</td>
<td>0.93</td>
<td>0.9</td>
<td>0.83</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>Tk (µm)</td>
<td>0.74</td>
<td>0.59</td>
<td>0.63</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>AF (L/m².s)</td>
<td>0.84</td>
<td>0.68</td>
<td>-0.2</td>
<td>1</td>
</tr>
<tr>
<td>15407</td>
<td>BP (bar)</td>
<td>0.98</td>
<td>0.84</td>
<td>0.73</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Tk (µm)</td>
<td>1</td>
<td>0.99</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>AF (L/m².s)</td>
<td>0.99</td>
<td>0.86</td>
<td>0.85</td>
<td>0.98</td>
</tr>
</tbody>
</table>

5.3. Model Building

5.3.1. General Aspects

The model building was performed with software SX-Center and SX-Plus from NIR Online. To obtain a predictive model, each set of measured CQAs are assigned to the respective spectral data. This combination is hereafter called experiment. Such merging is accomplished by attributing the marking time
of each membrane to NIR spectral data with the same corresponding time. Although, the registered marking time corresponds to the time where the membrane is on the cleaning step. So it was necessary to add a delay (of some minutes) that can be calculated using CSp and the length of the production line from the marking time to the NIR detection.

After both sets of data are merged, the software uses a specified model to reduce the number of variables (as seen in Section 2.3). This can generate multiple outcomes depending on the algorithm of the chosen model. For the building of these predictive models, it was used a modified version of PLS. The algorithm of this model is protected by NIR-Online and specifically designed to provide a better output when using spectral data. This software also uses a spline function to match the specific NIR polichromator with the spectral data.

After defining these initial parameters, the software can now present a predictive model with the raw data. This is followed by deciding an initial number of principal components (nPC). Such requires to choose/analyze the most relevant plot in all statistical plots available in the software. The initial nPC can be established by using a Loadings Plot. This plot is often used in NIR spectroscopy to identify a spectral band of pure component in a sample. So it is expected that this plot is significantly relevant for the identification of chemical properties. Normally for physical parameters, like BP, there is not a specific band. Loadings for physical parameters will be based on scattering (background effects) without the need of specific bands. It is important to refer that the pores are a result of some organic compounds that are caught in the casting solution. Therefore chemical spectral bands can correlate with the BP. Nevertheless it is expected scattering as the main effect.

After establishing an initial nPC, the estimates and residues plot give an overview of the model. The estimates plot puts the original CQA set against the predicted set. This plot provides statistical insight of the SEC and the fit of the prediction ($R^2$). The residues plot can expose the residues of the predicted data. It also highlights the data above standard deviation (SD) and two times the SD. Both these plots are fundamental when choosing the necessary pre-treatments and in outlier detection. If necessary both these data modifications are performed.

The model without outliers can be now be validated using simple validation and/or CV. For this step, it is necessary to define a number (for validation ($N_{val}$) or for cross-validation ($N_{CV}$)). Afterwards the software takes out one experiment in every multiple of that number. Using estimates plot, it is possible to observe changes in SEC, SECV, SEP and their respective parameter of fit, $R^2$. Finally, using a Press plot it is possible to evaluate the initial nPC chosen. Once finalized, the model is exported in a form of a calibration file. The model building for each CQA of each membrane is presented in the next sections.
5.3.2. Membrane 15458

For the all the 3 CQA models (BP, Tk and AF) it was used the 23 experiments. For the BP model, in the loadings plot in Figure 30, it is clear that in the 4th component, the spectral data has some noise. Hence, a good first approach is to initially consider the number of principal components as 3. In the estimates plot in Figure 31, there is a small dispersion. In this plot there were identified 6 outliers. Looking for the residues plot in Figure 32, it is clear that the residues with a higher SD were considered as outliers.

The model was validated using 4 points. These points were used to calculate the SEP. The model was also cross-validated using 5 points. The model seems to keep its basic integrity after validation and cross-validation. No additional pre-treatment was applied to the data. The SEC decreased 75% and $R^2$ increased 18% (Figure 33). The residues plot in figure Figure 34 shows that the SD range also decreased with only one point higher than 2 times the SD. Press plot (Figure 35) shows that there are small deviations of the different errors with a nPC higher than 3. This is also verified in the balanced error. Therefore 3 PC guaranties a low model error and estimation error. This is because low nPC can increase model error and higher nPC affects the estimation error.
For the Tk model, in the Loadings plot, the spectral data already has some noise in the 4th component. (Figure 36). This justifies that the first approach for the nPC components as 3. There were identified 3 outliers in the estimates plot (Figure 37). In residues plot it is clear that all residues with a higher SD comprise the identified outliers.
The validation used 4 points and cross-validation used 5 points. The model seems to keep its basic integrity after validation and cross-validation. No additional pre-treatment was applied to the data. The SEC decreased 43% and $R^2$ increased 10% (Figure 39). The residues plot shows that the SD range also decreased with all the points lower than 2SD (Figure 40). Press plot shows that there are small deviations of the different errors with a nPC higher than 3 (Figure 41). In fact, SEP have its lower value with this nPC. Therefore, it is plausible to conclude that 3 PC guarantees a low model error and estimation error.

In the AF model the Loadings plot show that the spectral data already has some irrelevant noise in the 5th component (Figure 42). Hence the first approach for the nPC components is 4. There were identified 2 outliers in the estimates plot (Figure 43). In residues plot it is clear that all residues with a higher standard deviation comprise the identified outliers (Figure 44).
The validation used 4 points and cross-validation used 5 points. The model seems to keep its basic integrity after validation and cross-validation. No additional pre-treatment was applied to the data. The SEC decreased 67 % and $R^2$ increased 10% (Figure 45). The residues plot shows that the SD range also decreased with all the points lower than 2SD (Figure 46). Press plot shows that there are small deviations of the different errors with a nPC higher than 4 (Figure 47). Press plot shows that SEP increases with the nPC while SEC decreases. Both SECV and the balanced error increase significantly with nPC higher than 5. In this scenario, nPC of 5 could have been a better choice. Although choosing a nPC of 4, reduces the probability of error in the estimation process.
5.3.3. Membrane 15457

For the all the 3 CQA models (BP, Tk and AF) it was used the 24 experiments. For the BP model, the Loadings plot shows that relevant information about the data can be shown until the 5th component (Figure 48). With 5 PC, the loads appear to revolve around 0 bar. A good first approach is to initially consider the number of principal components as 4. Estimate plot shows the presence of 3 potential outliers (Figure 49). Looking for the residues plot it is clear that the residues with a higher SD were considered as outliers (Figure 50).

![Figure 48 – Loadings plot for BP model of 15457 for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line), 4 PC (yellow line) and 5 PC (green line)](image1)

![Figure 49 – Estimates plot for BP model of 15457. Black dots represent raw data and red dots represent outliers. SEC=0.16 bar and $R^2=0.876$](image2)

![Figure 50 – Residues plot for BP model of 15457. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow dots) and lower than 1SD (green dots)](image3)

The model was validated using 4 points. These points were used to calculate the SEP. The model was also cross-validated using 5 points. The model seems to keep its basic integrity after validation and cross-validation. No additional pre-treatment was applied to the data. The SEP decreased 50% and $R^2$ increased 12% (Figure 51). SEP is also in the same range of the SEC which indicates a good predictability of the model. The residues plot shows that the SD range also decreased (Figure 52). Press plot corroborates the initial choice of nPC of 3 was a good choice (Figure 53). Although the error is not minimum with this nPC. Lower nPC would increase model error. Higher nPC would decrease the balanced error but could also add noise to the system, which affect the estimation error.
Looking for the Loadings plot for the Tk model in Figure 54, it is clear that in the 5th component, the spectral data already has some noise. A good first approach is initially consider the number of principal components as 4. In the estimates plot, there is a small dispersion. In this plot there were identified 3 outliers (Figure 55). In residues plot it is clear that all residues with a higher SD are considered outliers with the exception of the sample number 13 (Figure 56). This sample was kept since it didn’t improve the model if taken out.
The model was validated using 4 points. These points were used to calculate the SEP. The model was also cross-validated using 5 points. The model seems to keep its basic integrity after validation and cross-validation. No additional pre-treatment was applied to the data. The SEP decreased 54% and $R^2$ increased 4% (Figure 57). The residues plot shows that the SD range also decreased (Figure 58). The press plot shows that all errors follow a trend (Figure 59). This means that those errors decrease with the increasing nPC. The initial choice of nPC of 4 was a good choice. This nPC does not compromise the estimation and model errors.

For the AF model, in Loadings plot in Figure 60, it is clear that in the 5th component, the spectral data already has some noise. A good first approach is initially consider the number of principal components as 4. In the estimates plot, there were identified 3 outliers (Figure 61). In residues plot it is clear that all residues with a higher standard deviation (SD) are considered outliers (Figure 62).
The model was validated using 4 points. These points were used to calculate the SEP. The model was also cross-validated using 5 points. The model seems to keep its basic integrity after validation and cross-validation. No additional pre-treatment was applied to the data. The SEP decreased 60% and $R^2$ increased 10% (Figure 63). The residues plot shows that the SD range also decreased (Figure 64). The press plot shows that between the 4th and the 6th component the balanced error does not show a relevant change (Figure 65). A nPC higher than 6 increases significantly the SECV and SEP. The initial choice of nPC of 4 was a good choice.
5.3.4. Membrane 15404-H

For the all the 3 CQA models (BP, Tk and AF) it was used the 14 experiments. For the BP model, the Loadings plot shows that increasing nPC also increases the noise of data (Figure 66). This can be explained with the low number of experiments. As first approach initially is considered a number of principal components of 2. Estimate plot shows the presence of 1 potential outliers, experiment number 12 (Figure 67). Looking for the residues plot it is clear that the experiment 12 has a higher SD (Figure 68).

![Figure 66 – Loadings plot for BP model of 15404-H for 4 components for 1 PC (blue line), 2 PC (orange line), 3 PC (grey line) and 4 PC (yellow line).](image)

![Figure 67 – Estimates plot for BP model of 15404-H. Black dots represent raw data and red dots represent outliers. SEC= 0,14 bar and $R^2=0,881$](image)

![Figure 68 – Residues plot for BP model of 15404-H. Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots)](image)

The model was validated using 4 points. These points were used to calculate the SEP. With low number of experiments, cross-validation can generate sub-models with different behaviors. For this reason, cross-validation was not used in this model. For these membrane models it was used EMSC. This pre-treatment proved to improve significantly the main results. This is only possible because MSC methods assume that offset and multiplicative spectral effects are much larger than chemical composition effects. In fact, in this case, it is expected that chemical composition stays constant within experiments.

Low number of experiments do not allow a very complex cleaning. For this reason it was decided take only 1 outlier. A modest data cleaning shows a not very significant improvement in SEC and $R^2$. SEC only decreased 4% and $R^2$ improved 5% (Figure 69). SEP is also in the same range of the SEC which indicates a good predictability of the model. The data cleaning proved to not reduce significantly the range of the residues. Although all residues present a value lower than 2 times the SD (Figure 70). Press plot shows that the balance error decreases in the initial 2 first nPC (Figure 71). SEC and SEP are equal between 2 and 3 nPC. A nPC of 3 could also be a good option. Although this could increase SEP in a range that SEC decreasing could not follow. This is not advisable in such short number of experiments.
For the Tk model, the Loadings plot in Figure 72 shows that spectral data is not relevant in the 4th PC. Initially it was consider the number of principal components as 3. Estimate plot shows the presence of 1 potential outliers, experiment number 12 (Figure 73). Looking for the residues plot it is clear that experiment 12 has a higher value than 2 times the SD (Figure 74).
To calculate the SEP, the model was validated using 4 points. Like in the BP model, cross-validation was not used in the model. The pre-treatment EMSC was also taking in consideration. The low number of experiments forced to take out only 1 outlier. Taking out 12 proved to produce a better model. The data cleaning shows a decreased of 61% in SEC and an increased $R^2$ improved 20% (Figure 75). SEP is also in the same range of the SEC which indicates a good predictability of the model. The data cleaning proved to not reduce significantly the range of the residues (Figure 76). Press plot shows that the balance error do not change significantly in the initial 3 nPCs (Figure 77). A nPC of 2 could also be a good option. Although the model fit is significantly lower for 2 PC (data not shown).

![Estimates plot](image)

**Figure 75 – Estimates plot for Tk model of 15404-H (cleaned data).**

It shows normal data (dots, SEC=2.53 $\mu$m and $R^2=0.977$) and validation data (squares). SEP=3.52 $\mu$m and $R^2=0.904$

![Residues plot](image)

**Figure 76 – Residues plot for Tk model of 15404-H (cleaned data).**

Residues are divided according to higher than 2SD (red dots), lower than 2SD and higher than 1SD (yellow point) and lower than 1SD (green dots)

![Press plot](image)

**Figure 77 – Press plot for Tk model of 15404-H (cleaned data).**

The errors presents according to PC number is SEC (blue line), SEP (orange line) and Balanced (yellow line)

For the AF model, the Loadings plot in Figure 78 shows that spectral data is not relevant in the 4th PC (Figure 78). Initially it was consider the number of principal components as 3. Estimate plot shows the presence of 2 potential outliers, experiments number 4 and 12 (Figure 79). Looking for the residues plot it is clear that experiment 12 has a higher value than 2 times the SD (Figure 80).
To calculate the SEP, the model was validated using 4 points. Once again, cross-validation was not used and EMSC was also taking in consideration. The low number of experiments forced to take out only 1 outlier. Although in this case, experiments 4 and 12 show a significant deviation in the model building. Therefore both experiments were taken out. The data cleaning shows a decreased of 90% in SEC and increased $R^2$ improved 28% (Figure 81). SEP is also much higher than SEC which indicates a poor predictability of the model. The data cleaning proved to reduce significantly the range of the residues (Figure 82). Press plot shows that all error decrease significantly in the first 2 PC (Figure 83). A nPC of 2 could also be a good option. Although the model fit is significantly low for 2 PC (data not shown). Therefore a nPC of 3 is the best option considering the framework.
5.3.1. Membrane 15407

The membrane 15407 was the only one with a model already implemented in the production line. These allows to use the data already run by the previous model. These data can be compared with the new models with the exception ofTk. The Tk model previously obtained was generated with membranes measured with Thickness tool (Prüfmiteel). As seen on Section 3.5, this tool has a higher uncertainty comparing with the DOM. Therefore a comparison between old data and new data would be unwise. For the BP model, it was used the 10 experiments. Loadings plot shows that spectral data is not relevant in the 4 th PC (Figure 84). Initially it was consider the number of principal components as 3. Estimates plot shows a trend in the data (Figure 85). The low number of experiments discarded the possibility for using outlier detection. Although, Residues plot shows that experiment number 4 could be a potential outlier (Figure 85). This outlier was keep in the model due to reduced number of experiments.

To calculate the SEP, the model was validated using 4 points. Like in 15404-H models, cross-validation was not used. The pre-treatment EMSC was also taken to consideration. The data cleaning shows a decreased of 46% in SEC and increased $R^2$ improved 3% (Figure 87). SEP is in the same range of SEC which indicates a good predictability of the model. The data cleaning proved to reduce significantly the range of the residues (Figure 88). Press plot shows that SEP increases until the 2 PC, decreasing abruptly after the 3 rd PC (Figure 89). Choosing a higher nPC could be an option. Although, considering the low number of experiments, such option would not be wise.
This new model was put against the data already gathered at the production line. These data is composed by 287 experiments taken since the model implementation. In the estimates plot there is good fit within the old data (Figure 90). Comparing the SEC and SEP of both sets of data, it is clear that the new data prediction error falls within the same range. Looking to the residues plot, the new data set also is equal or lower than 1 SD of all residues (Figure 91).

For the Tk model the Loadings plot shows that spectral data is not relevant in the 4th PC (Figure 92). For this reason it was consider the number of principal components as 3. Estimates plot shows lack of fit within
predicted and observed data (Figure 93). This can be confirmed with a low $R^2$ and high SEC. Residues plot shows no sign of potential outlier (Figure 94).

To calculate the SEP, the model was validated using 4 points. Once again, cross-validation was not used and EMSC was also taken into consideration. The data cleaning shows a decreased of 26% in SEC and increased $R^2$ improved 30% (Figure 95). Yet, even after the data cleaning, the model showed to have a substantial lack of fit. Also the model’s SEP is very superior to the uncertainty of the measuring tools ($ \pm 0.8 $ µm). The data cleaning proved to slightly reduce the range of the residue (Figure 96). Press plot shows that the SEP and Balanced error is minimum at the 4th PC (Figure 97).
The generated predictive model did not produce the best prediction results. In a general way, Tk is a parameter with high variability within the same sample. Also, the DoE was performed with a casting solution at the end of the batch. Possibly precipitation and stirring problems in casting solution could be in the origin of the problem. Also possibility of errors in the measuring are not discarded.

For the AF model the Loadings plot shows that spectral data is not relevant in the 5th PC (Figure 98). For this reason it was consider the number of principal components as 4. Estimates plot shows a trend within predicted and observed data (Figure 99). Residues plot shows no sign of potential outlier (Figure 100).

For SEP estimation, the model was validated using 4 points. Once again, cross-validation was not used and EMSC was also taken into consideration. The data cleaning shows a decreased of 38% in SEC and increased \( R^2 \) improved 13% (Figure 101). Estimates plot shows that validation points are well predicted by the model SEC and SEP are both within the same range. Residues range also decreased with none higher than 2 times the SD value (Figure 102). Press plot shows that SEC decreases with the number of components (Figure 103). SEP is more inconstant although presents a minimum value at the 4th component.
This cleaned model was put against the data already gathered at the production line. These data is composed by 278 experiments taken since the implementation of the first model. Just like the BP model, the estimates plot there is good fit within the old data (Figure 104). Comparing the SEC and SEP of both sets of data, it is clear that both parameters are very similar. Looking to the residues plot, the new data set also is equal or lower than 1 SD of all residues (Figure 105).
6. Conclusions

Both the resulting data from CPP-CQA relationship analysis and model building are summarized in Table 6 - Appendix C. All the conclusions taken hereafter are applied to the 4 studied membranes, 15458, 15457, 15404-H and 15407.

Regarding the CPP-CQA relationship analysis, BP, Tk, AF data seen to have a good fit within the selected CPPs. BP is strongly correlated with T\text{Band}, which corroborates with the prior-knowledge of the process. The T\text{PB} correlation with pore size indicators, such as BP, seems to be a good indicator of in which step does the pore development is more relevant (condition or precipitation step). As predicted from prior-knowledge, it was also verified that Tk is highly affected by CSp and ACS. It was seen that Tk influences T\text{Band}. TCol and Cov don’t appear to have a significant impact in the membranes where they were tested. Diffusion is irrelevant as a membrane CQA for the selected membranes and group of CPPs.

Regarding the model building, it is clear that the SEP and R\text{2} improved with the data pre-treatment (Figure 106). Therefore, it is possible to conclude that all models presented a good response to the data cleaning. The percentage improvement of these statistical parameters is in the same range of membranes 15457 and 15458. Such does not happen with membranes 15404-H and 15407 with ranges from 4-90% for SEP decrease and 3-30% for R\text{2} increase.

![Figure 106 – Decrease of SEP (red bars) and increase of R\text{2} (yellow bars) for the membranes 15458, 15457, 15404-H and 15407.](image)

When it comes to pre-treatments, EMSC appears to have a good effect on the models of membranes 15404-H and 15407. This pre-treatment adds an additive and multiplicative adjustment to the spectral data of all data set. Although effective, the EMSC pre-treatment uses data already stored, which can create a “bias effect” that can induce errors in the model building. In fact, the high improvement in R\text{2} in some models of
these membranes can reflect this fact. Therefore, there must be a special caution when applying these pre-treatments to avoid misinterpretations.

The outlier detection removed up to 26% of the experiments. The low number of experiments in membranes 15404-H and 15407 conditioned other types of data cleaning such as cross-validation and outlier detection (15407 only). In the light of these facts, the use of EMSC was important to ensure a reasonable model. The model produced by the data with 10 experiments and pre-treated with EMSC proved to be fitted in a much larger data set.

The quality of the model seems to correlate differently when analyzing by CQAs and membrane type. When analyzing by CQAs (Figure 107), the SEP of all BP models appear to be within the same range (0,05 – 0,16 bar), which is positive since this is the main CQA. Tk models for the membranes 15458, 15457 and 15404-H appear to be within the same range (0,47 – 3,52 µm). The errors for membrane 15407 are too high (SEC=18,04 µm and SEP=51,71 µm). Since it appears to exist no miscorrelation between Tk and the CPPs for this membrane (R²=1, Q²=0,99; MV=0,97; Rep=0,99), the problem must be possibly related to precipitation and stirring problems in casting solution tank and/or errors in the referencing. More data should be acquired in order to take further conclusions. When it comes to AF, the membranes 15458, 15457 and 15407 appear to be within the same error range (0,02 – 0,42 L/m².s). The membrane 15404-H has a low SEC (0,19 L/m².s) but a high SEP (1,56 L/m².s). In fact, when evaluated the relation between Tk-CPPs of this membrane, the statistical parameters do not all appear to be relevant, with a very low value of MV (R²=0,84, Q²=0,68;MV=-0,2;Rep=1). In the light of these facts it is plausible to conclude that maybe AF is not a good CQA for this type of membrane. Although, as mentioned above, models of 15404-H and 15407 were built with a low number of experiments. Analyzing the quality of the model according to membrane type, models generated by 15404-H and 15407 seem to be very dependent on pre-treatments (like EMSC) and lack consistent results. On the other hand, membranes 15458 and 15457 more reliable results. This can be directly associated with the size of each data sets. Evidently, bigger data sets carry more information about the system and prevent poor model building. Normally, small data sets translate in lower range of variability which restricts the ability of predict better results.
Figure 107 – Comparison between SEC (blue bars), SECV (orange bars) and SEP (yellow) for the models of the 4 membranes by CQA (Plot A – BP; Plot B – Tk; Plot C – AF). Due the low number of experiments, cross-validation was not used in membranes 15404-H and 15407 which explains the absent values. In the plot B, the TK of 15407 is beyond the max value of the axis, 51.71 µm.

The model evaluation can also be complemented having in mind the final application of the process. Since this is an industrial process, this evaluation should consider standard in-specification limits used in the production line. Having in mind these ranges, it is possible to set limits with SD and SEP values for each membrane. These in-specification limits are documented for the BP, since it is the most important CQA. (Table 5).
Table 5 – In-specification limits used at the membrane production site to evaluate the process performance and their SD and SD/2

<table>
<thead>
<tr>
<th></th>
<th>BP Min (bar)</th>
<th>BP Máx (bar)</th>
<th>SD (bar)</th>
<th>SD/2 (Bar)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15458</td>
<td>4,031</td>
<td>4,999</td>
<td>0,484</td>
<td>0,242</td>
</tr>
<tr>
<td>15457</td>
<td>2,580</td>
<td>3,386</td>
<td>0,403</td>
<td>0,202</td>
</tr>
<tr>
<td>15404-H</td>
<td>1,451</td>
<td>2,096</td>
<td>0,322</td>
<td>0,161</td>
</tr>
<tr>
<td>15407</td>
<td>5,724</td>
<td>6,611</td>
<td>0,443</td>
<td>0,222</td>
</tr>
</tbody>
</table>

As showed in Figure 108, all the errors of all the BP models are lower than the SD and SD/2 of the each in specification range. This guarantees that it is possible to use the NIR spectrometer to predict if these 4 membranes will stay inside specification. This will allow to use BP to monitor the process and to reduce the number of sampling. Of course when it comes to other CQAs, this monitoring may fail for some of the models like 15404-H and 15407.

In sum, this study shows that is possible to identify the prior knowledge of these membranes from the main relations between CPP-CQAs. It also shows that it is possible to expand model building to different kinds of membranes. Since Sartorius has other membranes produced by a precipitation process, this study contributes for the improvement on the monitoring of this type of manufacturing processes.
7. Future work

Considering the results obtained, it is clear that to expand the monitoring for other CQAs than BP, the models of 15404-H and 15407 may require additional improvement. This can be done by adding samples to the existing model or performing an entirely new DoE. The last option is preferential since other CPPs and CPPs’ range can be tested. Having in mind the models obtained from membranes 15458 and 15457, it is possible to think about a minimum number of samples to guarantee a good model. Although, this cannot be generalized since the number of samples for a DoE increases with the number of CPPs. Also each CPP has a different impact on the process which does not necessarily guarantee the building of a good model. For this reason it is difficult to establish a minimum number of experiments. Although it is worth to notice that 15458 proved to successfully generate good models yielding 23 experiments with only 3 CPPs. Having in mind that both 15404-H and 15407 use 3 CPPs, it is plausible that the outcome would improve with 23 samples. Additional samples from membrane 15458 and 15457 can be gathered and added to the model for validation. This will increase the reliability on these models.

The next step of the Project “Automation of Casting Machine” concludes by introducing this models in PCS, as soon as they are fully operational. These models can be used for monitoring for a long period. Of course, different process tweaks, such as different suppliers or lots of casting solution, may require for this models to be updated over time.
8. Bibliographic References


Appendix A. Instrumentation

Figure 109 – Sartocheck and stainless steel in-line filter holder.

Figure 110 – Bubble Point manual measurer

Figure 111 – Caliper (Prüfmittel) used to measure thickness

Figure 112 – Digital Optic Microscope used to measure thickness
Figure 113 – Air permeability device used to measure Air Flux

Figure 114 – NIR device: BioPAT® Spectro
Appendix B. Statistical Parameters

B.I. $R^2$ – Correlation Factor

$$R^2 = \frac{SS_{REG}}{SS}$$ \hspace{1cm} (18)

Where $SS_{REG}$ is the sum of squares of Y corrected for the mean, explained by the model and SS is the total sum of squares of Y corrected for the mean.

B.II. $Q^2$ – Predictive Factor

$$Q^2 = 1 - \frac{PRESS}{SS}$$ \hspace{1cm} (19)

Where PRESS is the prediction residual sum of squares and SS is the total sum of squares of Y corrected for the mean.

B.III. MV – Model Validity

$$MV = 1 + 0.57647 \log(plof)$$ \hspace{1cm} (20)

Where $plof = p$ for lack of fit.

B.IV. Rep – Reproducibility

$$Rep = 1 - \frac{MS_{pure\ error}}{MS_{total\ SS\ corrected}}$$ \hspace{1cm} (21)

Where MS is mean squares or variance.
Appendix C. Statistic Data

Table 6 – Summary of resulting data of CPP-CQA relationship analysis and model building. All statistical parameters are dimensionless quantities with the exception of SEC, SECV and SEP that use the dimension of each CQA (BP - bar, Tk - µm, Df - ml/min; AF - L/m².s)

<table>
<thead>
<tr>
<th>N Exp</th>
<th>CQA</th>
<th>DoE Evaluation</th>
<th>Raw data Calibration</th>
<th>Cleaned data Calibration</th>
<th>Cleaned data Cross-Validation</th>
<th>Cleaned data Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N Exp</td>
<td>CQA</td>
<td>R²</td>
<td>Q²</td>
<td>QC</td>
</tr>
<tr>
<td>15458</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,79</td>
<td>0,67</td>
<td>0,89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,88</td>
<td>0,82</td>
<td>0,31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,99</td>
<td>0,98</td>
<td>0,62</td>
</tr>
<tr>
<td>15452</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,87</td>
<td>0,76</td>
<td>-0,05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,97</td>
<td>0,94</td>
<td>0,79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,96</td>
<td>0,92</td>
<td>0,55</td>
</tr>
<tr>
<td>15404-H</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,93</td>
<td>0,9</td>
<td>0,83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,74</td>
<td>0,59</td>
<td>0,63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,84</td>
<td>0,68</td>
<td>-0,2</td>
</tr>
<tr>
<td>15407</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,98</td>
<td>0,84</td>
<td>0,73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,98</td>
<td>0,84</td>
<td>0,73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,98</td>
<td>0,84</td>
<td>0,73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,98</td>
<td>0,84</td>
<td>0,73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,98</td>
<td>0,84</td>
<td>0,73</td>
</tr>
</tbody>
</table>