UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Tesi di Laurea Magistrale in
Ingegneria Chimica e dei Processi Industriali

Development of transfer learning techniques for the scale-up of biopharmaceutical processes

Relatore: Prof. Pierantonio Facco
Correlatore: Dr. Gianmarco Barberi

Laureando: LORENZO COTALINI

ANNO ACCADEMICO 2023 – 2024

Abstract

The aim of this Thesis is to evaluate whether a multivariate multi-block latent variable regression model, JYPLS (García-Muñoz et al., 2005), is as an effective tool to predict and optimize product quality through transfer learning from a pilot-scale plant to an industrial-scale one, ascertaining, at the same time, what is the appropriate number of pilot-scale batches required to transfer information from the pilot scale to the industrial scale. The case study considered in this Thesis is a simulated fed-batch process for penicillin fermentation. In particular, Pensim (Birol et al., 2002) is used to simulate the pilot scale, while Indpensim (Goldrick et al., 2014) is used for the industrial scale. The specific objectives are: *i*) to obtain the most accurate and precise estimations of the final penicillin concentration from process variables collected online, and *ii*) to optimize the product quality achieving the highest penicillin concentration possible at the end of the batch in the industrial scale through model inversion. Two methodologies for designing the experimental campaign for data collection in the pilot-scale plant are also evaluated: a full factorial design and happenstance design.

It has been demonstrated that JYPLS is an effective transfer learning model, particularly in the case of well-controlled processes where random disturbances on the process variables are limited, even if low amount of data are available from a low number of batches (starting from 2 pilot-scale batches the model can be improved). Model accuracy and precision are improved by 5%, and productivity increased by 0.03%, which is equivalent to more than 1 kg of penicillin per batch.

Table of contents

Introduction

The term *scale-up* describes the procedure of increasing the size of a plant from the laboratory scale of the pilot scale to the industrial scale. The engineering challenge is to maintain process stability and product quality during this transition (Barberi et al., 2022 Processes 2022, 10, 1796.). The selection of the proper operating conditions for the new industrial plant is of critical importance to the purpose of reducing costs and accelerating the process development (Tomba et al., 2012). For this reason, mathematical models are used during process development to describe the process (Barberi et al., 2021). These models are constructed using theoretical relationships derived from existing literature or historical data from the considered plant or from similar facilities (Chu et al., 2021). However, in the case of biopharmaceutical industries, where high value-added products are manufactured, it is challenging to find the necessary information (Botton et al., 2022; Facco et al., 2020). In fact, only a limited number of experiments are typically performed in the industrial scale, given the high costs involved. Instead, extensive lab-scale or pilot-scale experimental campaigns can be conducted prior to scale-up. However, this information is not incorporated into mathematical models, but rather serves as a general guideline for the industrial plant. In 2020, Facco et al. proposed a framework for the use of data analytics biopharmaceutical process scale-up. Following this paper, promising results are obtained from studies about monoclonal antibodies models (Barberi et al., 2021; Barberi et al., 2022; Botton et al., 2022). The aim of this Thesis is investigating the feasibility of transferring data from pilot-scale experiments to industrial-scale models and to quantify the number of batches required to transfer this information. Additionally, it seeks to determine whether different experimental designs in the pilot-scale on the pilot scale affect the resulting data. The accuracy and the precision of a model that utilizes both pilot-scale data and industrial-scale data Joint-Y Partial Least Squares (JYPLS; García-Muñoz et al., 2005) as a transfer learning approach to predict the final product quality from process data taken online is studied and compared with a state-of-the-art multivariate regression model that employs only industrial-scale data, Partial Least Squares (PLS; Wold, 1975). This study is carried out in the case of two simulated penicillin production process: Pensim (a 100-L pilot-scale plant) and Indpensim (a 100,000-L industrial-scale plant), two benchmarks which is highly suitable for the objectives of this Thesis. In the initial Chapter of this Thesis, the mathematical models are presented in a theoretical manner with the most significant equations. The second Chapter provides an overview of the key features of the processes. The subsequent Chapter focuses on the experimental strategies to collect data from the processes. The fourth and fifth Chapters present a comparative analysis of the models performance, with the former examining predictive capabilities and the latter assessing optimized operating conditions. The concluding chapter presents the findings of the study and offers insights into their implications.

Chapter 1

Mathematical methodologies for transfer learning

This Chapter presents the theoretical basis of the multivariate methodologies used in this Thesis, namely Partial Least Squares (PLS; Wold, 1975) and Joint-Y Partial Least Squares (JY-PLS; García-Muñoz et al., 2005). These are regression models to predict the penicillin concentration at the end of the production batch and to suggest the optimal operating conditions to maximize the productivity of the industrial-scale process.

1.1 Partial least squares

Partial least squares (PLS; Geladi & Kowalski, 1986) is a bilinear regression model that deals with 2 two-dimensional matrices, the predictor matrix (**X**) and the response matrix (**Y**) (both with batches along the rows and process variables along the columns for the predictors, and the product quality indices along the columns for the predicted variables) and reduces their dimensionality by identifying latent variables (LV) that best predict the responses from the predictors. Latent variables are hidden, unobservable factors that are inferred from the data, and represent underlying phenomena describing correlations within and between datasets. They are constructed as linear combinations of the original variables. PLS identifies the latent space that maximizes the covariance between **X** and **Y**. This assumes that both datasets are influenced by the same underlying factors, which are captured by the latent variables. By focusing on these shared latent structures, PLS helps to better understand the relationships between the variables in the two datasets.

In this Thesis, multiway-PLS (MPLS; Bro, 1996) is used because the regressor data, $\underline{\mathbf{X}}\ [N \times V \times T]$, are collected in a three-dimensional matrix (where the dimensions represent the batch numbers N, the number of variables V and the number of time instants T). MPLS consists in the unfolding of the three-dimensional matrix, followed by a standard PLS. In this Thesis, batch-wise unfolding is used. In the batch-wise unfolding, variables measured in all the batches in a predetermined time instant t, $\mathbf{X}_t\ [N \times V]$, are horizontally concatenated to obtain the matrix $\mathbf{X}\ [N \times V \cdot T]$. Accordingly, this means considering every variable in every time instant as a distinct column of the matrix and studying the correlation between different time instants of a single variables and its correlation also with all the time instants of all the variables.

Nonlinear iterative partial least squares (NIPALS; Wold, 1975) algorithm is used for calculating the latent variables, with the objective of optimising the covariance of **X** and **Y**.

1.1.1 Mathematical formulation

PLS comprises two outer relations (one for the predictors and one for the responses) and an inner relation that relates the predictors **X** $[N \times V \cdot T]$ to the response **Y** $[N \times P]$, were P is the number of the dependent variables. The outer relations are:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} + \mathbf{E} = \sum_{\mathrm{a}=1}^{\mathrm{A}} \mathbf{t}_{\mathrm{a}} \cdot \mathbf{p}_{\mathrm{a}}^{\mathrm{T}} + \mathbf{E} \quad , \tag{1.1}$$

$$\mathbf{Y} = \mathbf{U} \cdot \mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \sum_{\mathrm{a}=1}^{\mathrm{A}} \mathbf{u}_{\mathrm{a}} \cdot \mathbf{q}_{\mathrm{a}}^{\mathrm{T}} + \mathbf{F} \quad ; \tag{1.2}$$

where A is the number of latent variables, **T** $[N \times A]$ and **U** $[N \times A]$ are the score matrix for the X-block and Y-block, respectively, **P** $[V \cdot T \times A]$ and **Q** $[P \times A]$ are the loading matrix, **E** $[N \times V \cdot T]$ and **F** $[N \times P]$ are the residual matrix. The scores are the projections of the original data in the latent variables space, allowing to study the relationships among observations (i.e., batches). The loadings represent the contribution of each original variable to each latent variable, thus, providing information on the correlation among variables. Residuals are obtained by the difference between the original data and the reconstructed data, namely the information that is not captured by the PLS model.

The inner relationship is:

$$\mathbf{U} = \mathbf{T} \cdot \mathbf{B} + \mathbf{F} \quad ; \tag{1.3}$$

where **B** is the regression coefficients matrix. This relation connects the scores matrices of the **X** and **Y**.

1.1.1.1 NIPALS algorithm

The NIPALS algorithm (Wold et al., 2001) deals with data which are mean-centered and scaled to unit variance. The NIPALS algorithm is composed by eight steps:

1) initialize the score **u** as a column of the matrix **Y**:

$$\mathbf{u} = \mathbf{y}_p \quad ; \tag{1.4}$$

2) the weights of the X-block are calculated and normalize as follows:

$$\mathbf{w}^{\mathrm{T}} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{X}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}} \quad , \tag{1.5}$$

$$\mathbf{w}^{\mathrm{T}} = \frac{\mathbf{w}^{\mathrm{T}}}{||\mathbf{w}^{\mathrm{T}}||} \quad ; \tag{1.6}$$

3) the scores of the X-block are calculated as

$$\mathbf{t} = \mathbf{wX} \;\; ; \tag{1.7}$$

4) for the Y-block the loadings are calculated and normalize as:

$$\mathbf{q}^{\mathrm{T}} = \frac{\mathbf{t}^{\mathrm{T}}\mathbf{Y}}{\mathbf{t}^{\mathrm{T}}\mathbf{t}} \;\; , \tag{1.8}$$

$$\mathbf{q}^{\mathrm{T}} = \frac{\mathbf{q}^{\mathrm{T}}}{||\mathbf{q}^{\mathrm{T}}||} \;\; ; \tag{1.9}$$

5) finally, the score for the Y-block is updated as

$$\mathbf{u} = \frac{\mathbf{qY}}{\mathbf{q}^{\mathrm{T}}\mathbf{q}} \;\; ; \tag{1.10}$$

6) the convergence of the algorithm is tested with the X-block scores $\mathbf{t}$: the algorithm proceeds to step 7 if the normalized difference between the scores of two consecutive iterations is sufficiently small (typical thresholds are 10^{-8} or 10^{-10}), otherwise return to step 2.

7) Calculate the loadings of the X-block and deflate the matrices to remove the calculated component (subscript h) from the original regressor and response matrixes, to calculate a new component only on the information that has not already been modelled by previous latent variables:

$$\mathbf{p}^{\mathrm{T}} = \frac{\mathbf{t}^{\mathrm{T}} \cdot \mathbf{X}}{\mathbf{t}^{\mathrm{T}}\mathbf{t}} \;\; , \tag{1.11}$$

$$\mathbf{p}^{\mathrm{T}} = \frac{\mathbf{p}^{\mathrm{T}}}{||\mathbf{p}^{\mathrm{T}}||} \;\; , \tag{1.12}$$

$$\mathbf{E}_{a+1} = \mathbf{X} - \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}} \;\; , \tag{1.13}$$

$$\mathbf{F}_{a+1} = \mathbf{Y} - \mathbf{u}_a\mathbf{q}_a^{\mathrm{T}} \;\; ; \tag{1.14}$$

8) return to step 1 to build the next component by considering $\mathbf{E}_{a+1}$ and $\mathbf{F}_{a+1}$ instead of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

This procedure is iterated until the desired number of latent variables is constructed.

1.1.2 Prediction algorithm

To predict the response loadings and weights are needed. The scores are related to $\mathbf{X}$ through the modified weights, $\mathbf{W}^*$, as:

$$\mathbf{T} = \mathbf{X} \cdot \mathbf{W}^* \;\; . \tag{1.15}$$

The scores of X-block are a good predictor for the Y-block:

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{Q}^{\mathrm{T}} + \mathbf{F} \quad , \tag{1.16}$$

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{W}^{*} \cdot \mathbf{Q}^{\mathrm{T}} + \mathbf{F} \quad ; \tag{1.17}$$

the regression coefficients $\mathbf{B}$ are defined as:

$$\mathbf{B} = \mathbf{W}^{*} \cdot \mathbf{Q}^{\mathrm{T}} \quad ; \tag{1.18}$$

and Equation 1.16 can be rewrite as:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{B} + \mathbf{F} \quad ; \tag{1.19}$$

and the prediction equation is:

$$\hat{\mathbf{Y}} = \mathbf{X} \cdot \mathbf{B} \quad ; \tag{1.20}$$

where $\hat{\mathbf{Y}}$ is the predicted response.

The modified weights, needed to calculate the regression coefficients, can be obtained from the weights (i.e., calculated during NIPALS) in this way:

$$\mathbf{W}^{*} = \mathbf{W} \cdot (\mathbf{P}^{\mathrm{T}} \cdot \mathbf{W})^{-1} \quad . \tag{1.21}$$

1.1.2.1 Number of latent variables selection

The selection of the number of latent variables is a crucial point in the construction of a reliable model, because it is important to model the phenomena involved in the system under study to be optimally predictive for the response without capturing noise. Many criteria could be used to select the proper number of latent variables to be considered, but the most important are the minimization of the prediction residual sum of squares (PRESS) and the minimization of the root mean square error in cross validation (RMSECV).

In this Thesis, since two different models are compared (i.e., PLS and JY-PLS), the number of LVs that explain 90% of $\mathbf{Y}$ is selected, constraining the maximum allowable number of latent variables to eight.

1.1.2.2 Coefficient of determination

In this Thesis the metric chosen to evaluate the performance of the models is the coefficient of determination.

The coefficient of determination is defined in this way:

$$\mathrm{R}^2 = 1 - \frac{\sum_{p=1}^{P}\sum_{n=1}^{N}\left(y_{n,p} - \hat{y}_{n,p}\right)^2}{\sum_{p=1}^{P}\sum_{n=1}^{N}\left(y_{n,p} - \bar{y}_{n,p}\right)^2} \quad ; \tag{1.22}$$

where $\hat{y}_{n,p}$ is the predicted response, $y_{n,p}$ is the true response and $\bar{y}_{n,p}$ is the average value of the true response. The coefficient of determination has values between 0 and 1, where 1 indicates a perfect prediction, while 0 indicates that the model is predicting the average value of the observation. Negative coefficient of determination values indicates extremely bad prediction performance.

1.1.2.3 Hotelling's T^2

The Hotelling's T^2 statistic explains how close a sample is to the average conditions described in the calibration dataset and represents the position of each sample with respect to the score space origin. It is calculated for each observation *n* as:

$$T_n^2 = \mathbf{t}_n \boldsymbol{\Lambda}^{-1} \mathbf{t}_n^{\mathrm{T}} \quad , \tag{1.23}$$

where $\mathbf{t}_{\mathrm{n}}$ is the n-th row of the score matrix $\mathbf{T}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues corresponding to A latent variables and is defined as:

$$\boldsymbol{\Lambda} = \frac{\mathbf{T}\mathbf{T}^{\mathrm{T}}}{(N-1)} \quad . \tag{1.24}$$

Confidence limit for the Hotelling T^2 can be established. The confidence limits define a confidence region in which the true population mean is expected to lie within a certain probability (i.e., confidence). It is essentially the maximum value of the Hotelling's statistic expected to see if the true population mean is statistically equal to the hypothesized mean with the predetermined confidence. Accordingly, if an observation lies outside the confidence region, it results to be far from the average conditions. The confidence limit of the Hotelling's T^2 is defined as:

$$T_{lim}^2 = \frac{V \cdot T \cdot (N-1)}{(N - V \cdot T)} \cdot \mathrm{F}_{\alpha}(V \cdot T, N - V \cdot T) \quad ; \tag{1.25}$$

where $V \cdot T$ is the number of columns of the X-block, and $\mathrm{F}_{\alpha}(V \cdot T, N - V \cdot T)$ is the critical value for the F-distribution with $V \cdot T$ and $(N - V \cdot T)$ degrees of freedom and α significance level (e.g., 0.05 for a confidence limit with 95% probability).

1.2 Joint-Y Partial Least Squares

JYPLS is a multi-block regression model which is used for the transfer learning problem, specifically to transfer information from a source plant to a target plant. The primary objective is to construct a unified response latent variable model based on the data from both plants. This model is particularly advantageous during the start-up phase, when data for the target are limited. The only limitation pertains to the Y-block. In fact, the response matrices of the two plants must come from the same population of the variables. However, no restrictions are present for the predictors block (**X**), which can have different dimensions, namely have different variables. The key assumption behind JY-PLS is the existence of shared chemical and physical phenomena between plants, which allows the construction of a common latent space for the response.

1.2.1 Mathematical formulation

The JY-PLS uses data from two sources. In this Thesis, they are indicated as “pil” for the pilot process, which is the source plant, and “ind” for the industrial process, which is the target plant. The model uses a joint **Y** matrix to determine common loading matrix $\mathbf{Q}_J$, which identifies a common latent space for the Y-blocks of both plants. To better understand this structure a schematic representation of JY-PLS is shown in Figure 1.1.

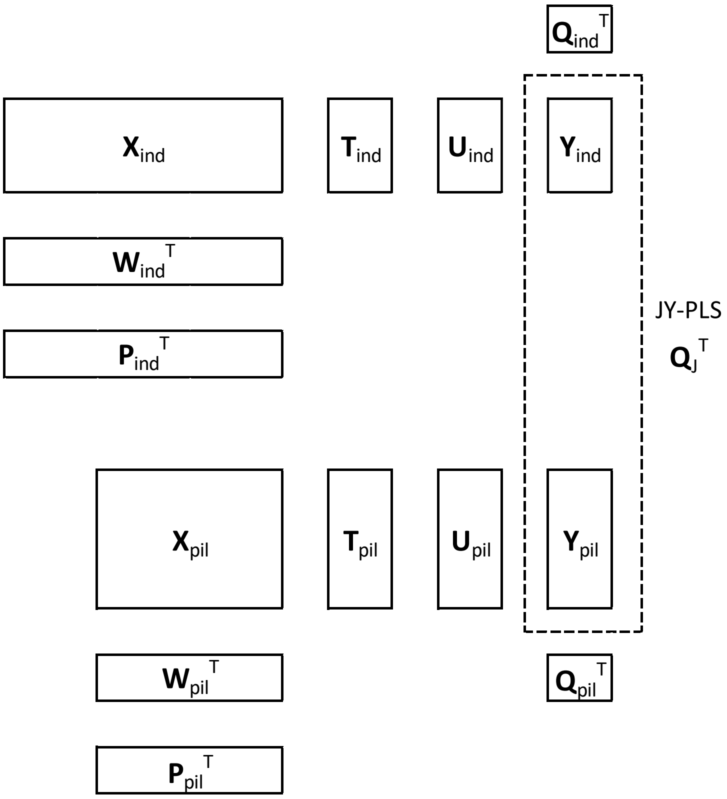


Figure 1.1. *Structure of the JY-PLS. Adapted from García-Muñoz et al. (2005)*

The common **Y** latent space is defined as:

$$\mathbf{Y}_\mathrm{J} = \begin{bmatrix}\mathbf{Y}_\mathrm{pil}\\ \mathbf{Y}_\mathrm{ind}\end{bmatrix} = \begin{bmatrix}\mathbf{T}_\mathrm{pil}\\ \mathbf{T}_\mathrm{ind}\end{bmatrix} \cdot \mathbf{Q}_\mathrm{J}^\mathrm{T} + \mathbf{E}_\mathrm{J} \quad ; \tag{1.26}$$

where $\mathbf{Y}_\mathrm{J}$ is the common matrix of the responses, where the response of the pilot scale $\mathbf{Y}_\mathrm{pil}$ is vertically concatenated with the one of the industrial scale $\mathbf{Y}_\mathrm{ind}$, and $\mathbf{E}_\mathrm{J}$ is the residual for this common matrix.

The single **X** data can be decomposed in its latent space as:

$$\mathbf{X}_\mathrm{pil} = \mathbf{T}_\mathrm{pil} \cdot \mathbf{P}_\mathrm{pil}^\mathrm{T} + \mathbf{E}_\mathrm{pil} \quad , \tag{1.27}$$

$$\mathbf{X}_\mathrm{ind} = \mathbf{T}_\mathrm{ind} \cdot \mathbf{P}_\mathrm{ind}^\mathrm{T} + \mathbf{E}_\mathrm{ind} \quad ; \tag{1.28}$$

weights are used to link the scores to the regressors in this way:

$$\mathbf{T}_\mathrm{pil} = \mathbf{X}_\mathrm{pil} \cdot \mathbf{W}_\mathrm{pil}^* \quad , \tag{1.29}$$

$$\mathbf{T}_\mathrm{ind} = \mathbf{X}_\mathrm{ind} \cdot \mathbf{W}_\mathrm{ind}^* \quad ; \tag{1.30}$$

the weight matrixes of Equations 1.28 and 1.29 are obtained from the modified ones as in Equation 1.19.

1.2.1.2. Modified NIPALS algorithm

A modified NIPALS algorithm (García-Muñoz et al., 2005) is usually employed to calculate the JY-PLS score, loading and weights. The main advantages of the NIPALS algorithm are an easy implementation and the possibility to manage missing data on both X and Y-blocks. The modified NIPALS comprises the following 8 steps.

1) Initialize the scores of Y-block as the first columns of the responses:

$$\mathbf{u}_\mathrm{pil} = \mathbf{Y}_\mathrm{pil,1} \quad , \tag{1.31}$$

$$\mathbf{u}_\mathrm{ind} = \mathbf{Y}_\mathrm{ind,1} \quad ; \tag{1.32}$$

where the subscript 1 is referred to the first column of the matrix.

2) Calculate the weights by regressing the X-block using the **u** scores:

$$\mathbf{w}_\mathrm{pil} = \mathbf{X}_\mathrm{pil}^\mathrm{T} \cdot \mathbf{u}_\mathrm{pil} \cdot \left(\mathbf{u}_\mathrm{pil}^\mathrm{T}\mathbf{u}_\mathrm{pil}\right)^{-1} \quad , \tag{1.33}$$

$$\mathbf{w}_\mathrm{ind} = \mathbf{X}_\mathrm{ind}^\mathrm{T} \cdot \mathbf{u}_\mathrm{ind} \cdot \left(\mathbf{u}_\mathrm{ind}^\mathrm{T}\mathbf{u}_\mathrm{ind}\right)^{-1} \quad ; \tag{1.34}$$

3) normalize the weights to unit length.
4) Calculate the scores of the X-block as:

$$\mathbf{t}_{\mathrm{pil}} = \mathbf{X}_{\mathrm{pil}} \cdot \mathbf{w}_{\mathrm{pil}} \cdot \left(\mathbf{w}_{\mathrm{pil}}^{\mathrm{T}} \mathbf{w}_{\mathrm{pil}}\right)^{-1} \quad , \tag{1.35}$$

$$\mathbf{t}_{\mathrm{ind}} = \mathbf{X}_{\mathrm{ind}} \cdot \mathbf{w}_{\mathrm{ind}} \cdot \left(\mathbf{w}_{\mathrm{ind}}^{\mathrm{T}} \mathbf{w}_{\mathrm{ind}}\right)^{-1} \quad ; \tag{1.36}$$

5) regress the joint $\mathbf{Y}_J$ onto the scores to obtain the joint loadings:

$$\mathbf{q}_{\mathrm{J}} = \mathbf{Y}_{\mathrm{J}}^{\mathrm{T}} \cdot \mathbf{t}_{\mathrm{J}} \cdot \left(\mathbf{t}_{\mathrm{J}}^{\mathrm{T}} \mathbf{t}_{\mathrm{J}}\right)^{-1} \quad ; \tag{1.37}$$

6) calculate new the scores from the joint loadings and check the convergence respect to the initial values of the scores (Equations 1.31 and 1.32):

$$\mathbf{u}_{\mathrm{pil}} = \mathbf{Y}_{\mathrm{pil}} \cdot \mathbf{q}_{\mathrm{J}} \cdot \left(\mathbf{q}_{\mathrm{J}}^{\mathrm{T}} \mathbf{q}_{\mathrm{J}}\right)^{-1} \quad , \tag{1.38}$$

$$\mathbf{u}_{\mathrm{ind}} = \mathbf{Y}_{\mathrm{ind}} \cdot \mathbf{q}_{\mathrm{J}} \cdot \left(\mathbf{q}_{\mathrm{J}}^{\mathrm{T}} \mathbf{q}_{\mathrm{J}}\right)^{-1} \quad ; \tag{1.39}$$

if convergence is not reached return to step 2 and iterate the procedure until convergence, otherwise continue to step 7.
7) Calculate the loadings for the X-block:

$$\mathbf{p}_{\mathrm{pil}} = \mathbf{X}_{\mathrm{pil}}^{\mathrm{T}} \cdot \mathbf{t}_{\mathrm{pil}} \cdot \left(\mathbf{t}_{\mathrm{pil}}^{\mathrm{T}} \mathbf{t}_{\mathrm{pil}}\right)^{-1} \quad , \tag{1.40}$$

$$\mathbf{p}_{\mathrm{ind}} = \mathbf{X}_{\mathrm{ind}}^{\mathrm{T}} \cdot \mathbf{t}_{\mathrm{ind}} \cdot \left(\mathbf{t}_{\mathrm{ind}}^{\mathrm{T}} \mathbf{t}_{\mathrm{ind}}\right)^{-1} \quad ; \tag{1.41}$$

8) deflate the predictors and the responses for each source and calculate the next component as described in point 7 of Section 1.1.1.1 (Equations 1.12 and 1.13).

1.2.2 Latent-variable model inversion

A calibrated JY-PLS model can be inverted to obtain the estimation of the conditions in term of regressors that ensure the desired responses. In this Thesis, JY-PLS inversion provides the operating conditions that maximize the productivity in the industrial process (García-Muñoz, 2004). First of all, the desired response ($\mathbf{y}_{\mathrm{ind,des}}^{\mathrm{T}}$) for the industrial process must be defined, then the inversion can be performed directly when there are no constraints on the new predicted scores vector ($\boldsymbol{\tau}_{\mathrm{ind,new}}$) and on the desired response ($\mathbf{y}_{\mathrm{ind,des}}$) as:

$$\boldsymbol{\tau}_{\mathrm{ind,new}}^{\mathrm{T}} = \left(\mathbf{Q}_{\mathrm{J}}^{\mathrm{T}} \cdot \mathbf{Q}_{\mathrm{J}}\right)^{-1} \cdot \mathbf{Q}_{\mathrm{J}}^{\mathrm{T}} \cdot \mathbf{y}_{\mathrm{ind,des}}^{\mathrm{T}} \quad ; \tag{1.42}$$

once the scores corresponding to the desired responses are obtained, it is possible to determine the original data from the scores vector as:

$$\hat{\mathbf{x}}_{\text{ind,new}} = \boldsymbol{\tau}_{\text{ind,new}} \cdot \mathbf{P}_{\text{ind}}^{\text{T}} \quad ; \tag{1.43}$$

instead, when a non-fully determined response and constraints are present on regressors and /or responses, the inversion is performed through an optimization problem (Tomba et al., 2012) and the problem is defined as:

$$\min_{\boldsymbol{\tau}_{\text{ind,new}}} \left[\left(\hat{\mathbf{y}}_{\text{ind,new}} - \mathbf{y}_{\text{ind,des}}\right) \cdot \boldsymbol{\Gamma} \cdot \left(\hat{\mathbf{y}}_{\text{ind,new}} - \mathbf{y}_{\text{ind,des}}\right)^{\text{T}} + g_1 \cdot \left(\sum_{a=1}^{A} \frac{\tau_a^2}{s_a^2} \right) \right] \quad , \tag{1.44}$$

st.

$$\hat{\mathbf{y}}_{\text{ind,new}} = \boldsymbol{\tau}_{\text{ind,new}} \cdot \mathbf{Q}_{\text{ind}}^{\text{T}} \quad , \tag{1.45}$$

$$\hat{\mathbf{x}}_{\text{ind,new}} = \boldsymbol{\tau}_{\text{ind,new}} \cdot \mathbf{P}_{\text{ind}}^{\text{T}} \quad , \tag{1.46}$$

$$\hat{y}_{p,\text{ind,new}} \leq b_p \quad , \tag{1.47}$$

$$lb_p^{\text{y}} \leq \hat{y}_{p,\text{ind,new}} \leq ub_p^{\text{y}} \quad , \tag{1.48}$$

$$lb_{v\cdot t}^{\text{x}} \leq \hat{\text{x}}_{v\cdot t,\text{ind,new}} \leq ub_{v\cdot t}^{\text{x}} \quad ; \tag{1.49}$$

in this set of equations, $\hat{\mathbf{y}}_{\text{ind,new}}$ is the predicted responses vector for the industrial process from the predicted scores vector $\boldsymbol{\tau}_{\text{ind,new}}$, τ_a is the a-th element of the solution $\boldsymbol{\tau}_{\text{ind,new}}$ and s_a is the standard deviation of the columns of the matrix $\mathbf{T}_{\text{ind}}$. $\boldsymbol{\Gamma}$ is a matrix in which on the diagonal elements there are the weights given to the specified equality constrains ($\mathbf{y}_{\text{ind,des}}$) in the solution, g_1 is a constant to balance the two terms in the summation (a good choice can be the reciprocal of the 95% confidence limit for the Hotelling's T^2, namely T_{lim}^2). b_p is the element of a vector **b** which force each column of the new predicted response ($\hat{y}_{p,\text{ind,new}}$) inside some specified ranges. lb_p^{y}, ub_p^{y}, are the lower and upper boundaries for each column of the new predicted response ($\hat{y}_{p,\text{ind,new}}$). $lb_{v\cdot t}^{\text{x}}$ and $ub_{v\cdot t}^{\text{x}}$ are the lower and upper boundaries for each column of the new predicted regressors ($\hat{\text{x}}_{v\cdot t,\text{ind,new}}$). Equations 1.48 and 1.49 are used to maintaining the proximity of the new inputs to the historical.

When the rank of the predictors is bigger than the one of the responses, some latent variables of **X** do not change the $\hat{\mathbf{y}}_{\text{ind,new}}$, the ensemble of these LV are called null space. Inside the null space $\boldsymbol{\tau}_{\text{ind,new}}$ can be changed without affecting the response. This space can be the target of an optimization problem in which between all possible solutions, which lead to the desired $\hat{\mathbf{y}}_{\text{ind,new}}$, an optimized solution is found imposing some specified criteria, such as economical or safety criteria. The condition that must be respected by the null space to guarantee the independence of the response is:

$$\hat{\mathbf{y}}_{\mathrm{ind,new}} = \mathbf{Q}_{\mathrm{ind}}^{\mathrm{T}} \cdot \left(\boldsymbol{\tau}_{\mathrm{ind,new}} + \Delta\boldsymbol{\tau}_{\mathrm{ind,new,null}}\right) \quad , \tag{1.50}$$

$$\mathbf{Q}_{\mathrm{ind}}^{\mathrm{T}} \cdot \Delta\boldsymbol{\tau}_{\mathrm{ind,new,null}} = 0 \quad . \tag{1.51}$$

Chapter 2

Penicillin process simulators at different scales

Chapter 2 presents an overview of two penicillin simulated processes employed in this Thesis for the problem of product/process scale transfer. The processes deal with the penicillin production at different production scales, namely, pilot scale and industrial scale. In particular, the key concepts, the structure and the most important model equations are presented.

2.1 Pilot-scale penicillin production process simulator Pensim

The Pensim process simulator (Birol et al., 2002) is a dynamic model of a pilot (100 L) fed-batch fermentation reactor for penicillin production.

The mechanistic model in Pensim, inspired by the Bajpai & Reuss (1980), is unstructured, meaning that no structural information about cellular activity is included and all the cellular physiology information is included into a single biomass term.

2.1.1 Pensim model structure

To better understand the structure of this model a simplified scheme in presented in Figure 2.1.

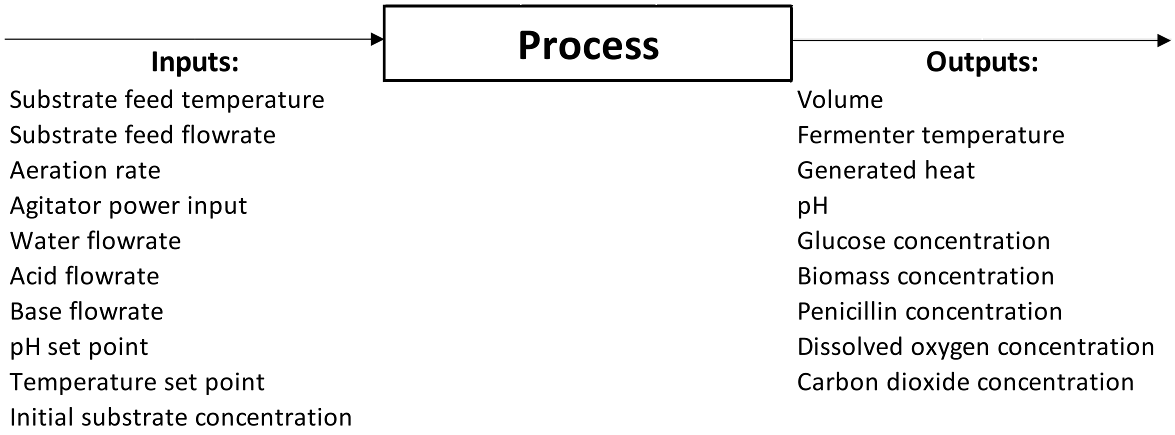


Figure 2.1. *Simplified Pensim model structure.*

The variables of the model are collected in the following table.

Table 2.1. *Variables list of the Pensim simulator.*

Variable #	Variable name	Type
1	Acid flowrate	Input
2	Aeration rate	Input
3	Agitation power	Input
4	Base flowrate	Input
5	Cooling water flowrate	Input
6	Substrate feed temperature	Input
7	Temperature controller set-point	Input
8	Initial substrate concentration	Input/design
9	pH controller set-point	Input/design
10	Substrate feed flowrate	Input/design
11	Generated heat	Output
12	Temperature	Output
13	Volume	Output
14	Biomass concentration	Output/state
15	Carbon dioxide concentration	Output/state
16	Oxygen concentration	Output/state
17	pH	Output/state
18	Substrate concentration	Output/state
19	Penicillin concentration	Output/target/state

Sampling time for process measurement is set to 1h, while for the concentrations, being more difficult to measure, is set to 12h. The feed addition starts when the glucose concentration reaches the threshold value of 0.3 (g/L). The length of the batches is 240h.

2.1.2 Mathematical model

The model structure is composed by six equations (2.1-2.6) which model the dynamic behaviour of the state variables and state the effects accounted for each of these variables:

$$X = f(X, S, C_{\mathrm{L}}, H, T) \ , \tag{2.1}$$

$$S = f(X, S, C_{\mathrm{L}}, H, T) \ , \tag{2.2}$$

$$C_{\mathrm{L}} = f(X, S, C_{\mathrm{L}}, H, T) \ , \tag{2.3}$$

$$C_{\mathrm{P}} = f(X, S, C_{\mathrm{L}}, H, T, C_{\mathrm{P}}) \ , \tag{2.4}$$

$$CO_2 = f(X, H, T) \ , \tag{2.5}$$

$$H = f(X, H, T) \ ; \tag{2.6}$$

where X (g/L) is the biomass concentration, S (g/L) is the substrate concentration, C_L (g/L) is the dissolved oxygen concentration, C_P (g/L) is the penicillin concentration, CO_2 (mmol/L) is the carbon dioxide concentration, H (mol/L) is the hydrogen ion concentration, namely the pH, and T (K) is the temperature. There are no hard constraints on input variables or kinetic parameters in the simulator, but the use of values outside the suggested ranges by Bajpai & Reuss (1980) could lead to outputs without physical meaning.

In this Thesis, the perfect control of the pilot scale is assumed, to simplify the problem. This means that no random disturbances on the input variables, nor measurement noise is considered. In the next Sub-subsections, the detailed formulations of the important pilot scale variables are shown for substrate concentration, pH, and penicillin concentration. All the other information can be found in the paper by Birol et al. (2002).

2.1.2.1 Substrate and penicillin

In this model the substrate utilization is strictly related to the feed flow rate, required for the maintenance of the organisms, the biomass growth and the penicillin production. The two major equations involve the mass balances of glucose and oxygen (dissolved) which are the nutrients for microorganisms:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\mu}{Y_{x/S}} \cdot X - \frac{\mu_{pp}}{Y_{p/S}} \cdot X - m_x \cdot X + \frac{F \cdot s_f}{V} - \frac{S}{V} \cdot \frac{\mathrm{d}V}{\mathrm{d}t} \quad , \tag{2.7}$$

$$\frac{\mathrm{d}C_L}{\mathrm{d}t} = -\frac{\mu}{Y_{x/O}} \cdot X - \frac{\mu_{pp}}{Y_{p/O}} \cdot X - m_O \cdot X + K_{la} \cdot (C_L^* - C_L) - \frac{C_L}{V} \cdot \frac{\mathrm{d}V}{\mathrm{d}t} \quad ; \tag{2.8}$$

where t is the time (h), μ is the specific growth rate (h^{-1}), $Y_{x/S}$ is the constant yield of biomass over glucose (g biomass/g glucose), μ_{pp} is the specific penicillin production rate (per h), $Y_{p/S}$ is the constant yield of penicillin over glucose (g penicillin/g glucose), m_x is the maintenance coefficient on substrate (per h), F is the feed flow rate of substrate (L/h), s_f is the feed substrate concentration (g/L), V is the culture volume (L), $Y_{x/O}$ is the constant yield of biomass over oxygen (g biomass/g oxygen), $Y_{p/O}$ is the constant yield of penicillin over oxygen (g penicillin/g oxygen), m_O is the maintenance coefficient on oxygen (h^{-1}) and C_L^* is the dissolved oxygen concentration at saturation (g/L).

The overall mass transfer coefficient K_{la} (L/h) is calculated with:

$$K_{la} = \alpha \cdot \sqrt{f_g} \cdot \left(\frac{P_w}{V}\right)^{\beta} \quad ; \tag{2.9}$$

where parameters α and β are estimated from experimental data (Bailey & Ollis, 1986), f_g is the flow rate of oxygen (L/h) and P_w is the agitation power (W).

2.1.2.2 pH

The mass balance of the hydrogen ions is:

$$\frac{\mathrm{d}[H^+]}{\mathrm{d}t}=\gamma\left(\mu\cdot X-\frac{F\cdot X}{V}\right)+\left[\frac{-B+\sqrt{(B^2+4\cdot 10^{-14})}}{2}-[H^+]\right]\cdot\frac{1}{\Delta t}\quad ; \tag{2.10}$$

where:

$$B=\frac{\left[10^{-14}/[H^+]-[H^+]\right]\cdot V-C_{a/b}\cdot(F_a+F_b)\cdot\Delta t}{V+(F_a+F_b)\cdot\Delta t}\quad ; \tag{2.11}$$

$[H^+]$ is the hydrogen ion concentration, γ is a constant (mol$[H^+]$/g biomass) determined from experimental data (Mou & Cooney, 1983), F_a and F_b are the acid and base flow rates (L/h), and $C_{a/b}$ is the concentration of both solutions assumed the same dilution (M).

In the specific growth rate of biomass (Equation 2.10) an inhibition effect of the concentration of hydrogen ions is included as:

$$\mu\propto f\left\{\frac{\mu_x}{1+[K_1/[H^+]]+[[H^+]/K_2]}\right\}\quad ; \tag{2.12}$$

μ_x is the maximum specific growth rate (h^{-1}) and the terms K_1 and K_2 are constants (Nielsen & Villadsen, 1994; Shuler & Kargi, 2002).

The pH of culture medium tends to decrease as the reaction proceeds. For this reason, a controller of pH is implemented to keep the acidity close to the set-point adjusting the flowrate of NH_4OH provided to the culture. Settings of the pH controller, as the one for the temperature controller can be found in (Birol et al., 2002).

2.1.2.3 Penicillin production

The penicillin mass balance is:

$$\frac{\mathrm{d}C_\mathrm{P}}{\mathrm{d}t}=\mu_{pp}\cdot X-K\cdot C_\mathrm{P}-\frac{C_\mathrm{P}}{V}\cdot\frac{\mathrm{d}V}{\mathrm{d}t}\quad ; \tag{2.13}$$

in the specific penicillin production rate, an inhibition effect of the substrate (Bajpai & Reuss 1980) is included as:

$$\mu_{pp}=\mu_p\cdot\frac{S}{\left(K_p+S+S^2/K_I\right)}\cdot\frac{C_L^p}{\left(K_{op}\cdot X+C_L^p\right)}\quad ; \tag{2.14}$$

where μ_p is the specific rate of penicillin production (h^{-1}), K_p is the inhibition constant (g/L), K_I is the inhibition constant for product formation (g/L), K_{op} is the oxygen limitation constant (-) and C_L^p is the dissolved penicillin concentration (g/L). Equation 2.14 illustrates that, up to a certain threshold, an increase in substrate favors penicillin production. Beyond this threshold, however, the increase in the substrate level led to a reduction in penicillin production.

2.2 Industrial-scale penicillin production process simulator Indpensim

Indpensim (Goldrick et al., 2014) simulates the behaviour of an industrial-scale (100000 L) fed-batch fermentation process for penicillin production. This simulator was developed for process control and optimization studies. The model, based on the work of Paul & Thomas (1996), is structured, meaning that the cellular information about the structure and the activity of the biomass are considered and it simulates the behaviour of an industrial strain of *Penicillium Chrysogenum*.

2.2.1 Model's structure

The structure of the IndPensim simulated process is presented in Figure 2.2.

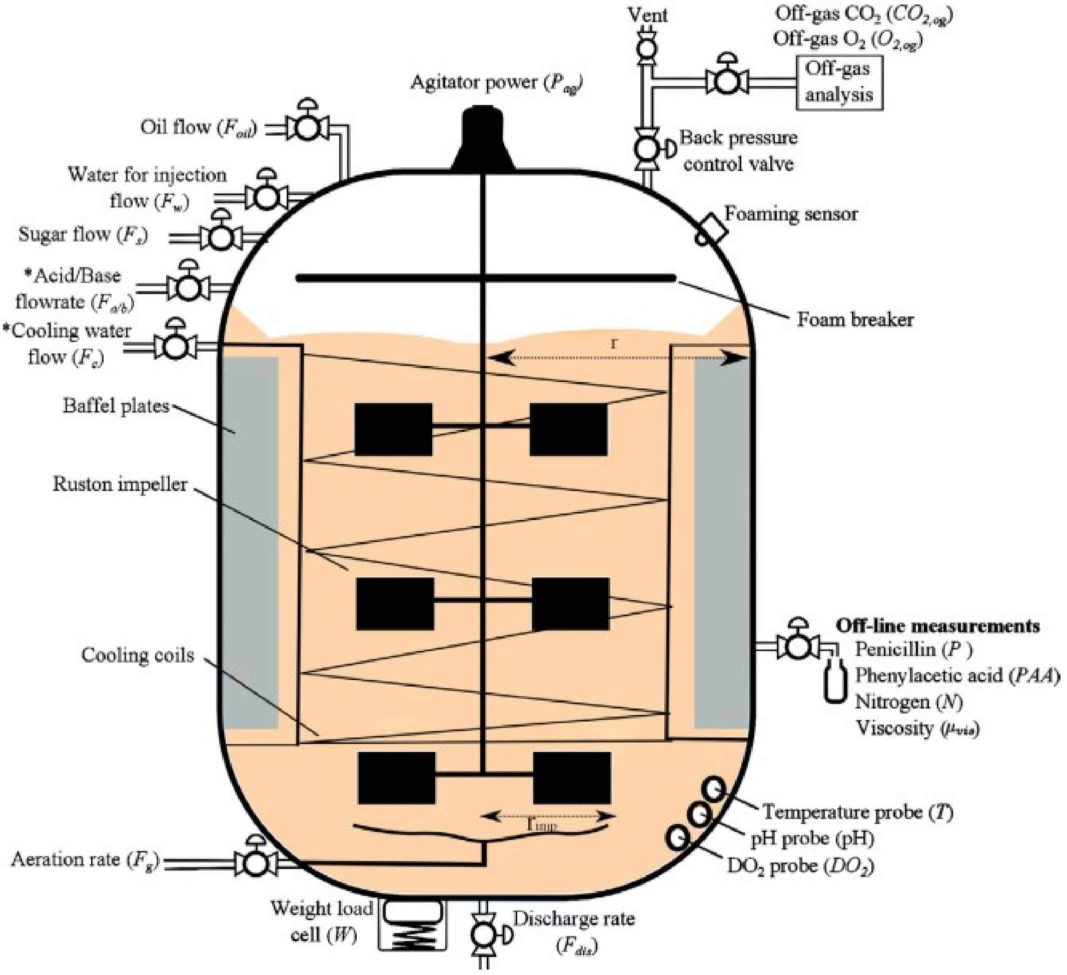


Figure 2.2. *Scheme of the Indpensim process with all inputs and outputs. Variables denoted with asterisk are not recorded by the batch records. (Goldrick, Ştefan, Lovett, Montague & Lennox 2014).*

As in Sub-section 2.1.1 the variables of the model are collected in a table.

Table 2.2. *Variables list of the Indpensim simulator.*

Variable #	Variable name	Type
1	Acid flowrate	Input
2	Aeration rate	Input
3	Agitation power	Input
4	Base flowrate	Input
5	Cooling water flowrate	Input
6	Heating water flowrate	Input
7	Injection water flowrate	Input
8	Oil flowrate	Input
9	Temperature controller set-point	Input
10	Initial substrate concentration	Input/design
11	pH controller set-point	Input/design
12	Substrate feed flowrate	Input/design
13	Ammonia concentration	Output
13	Biomass concentration	Output
14	Carbon dioxide % in off-gas	Output
15	Carbon evolution rate	Output/state
16	Dumped broth	Output
17	Generated heat	Output
18	Oxygen concentration	Output
19	Oxygen % in off-gas	Output
20	Oxygen uptake rate	Output
21	pH	Output
22	Phenylacetic acid concentration (offline)	Output
23	Phenylacetic acid concentration (online)	Output
24	Pressure	Output
25	Substrate concentration	Output
26	Temperature	Output
27	Viscosity	Output
28	Volume	Output
29	Weight	Output
30	Penicillin concentration	Output/target

Sampling time for online measurements is set to 1h, while off-line variables are recorded every 12h. The batch length is 240h congruent to one chose for the pilot scale simulator (Sub-section 2.1.2). The manipulated variables in this case are the initial substrate concentration, the recipe (which define the mixture flowrate of soybean oil and substrate) and the pH controller set-point.

2.2.2 Indpensim mathematical model

The mathematical model is composed by fourteen equations, the first four ones are the biomass balances in the different regions of the *Penicillium Chrysogenum Fungus*, which are: growing regions, non-growing regions, degenerated regions, and autolysed regions. The number of parameters is seventy-two. For simplicity only the most important balances are presented in the next sections, for the complete model and parameters values refers to the work by Goldrick and coworkers (Goldrick et al., 2014). The model adds random disturbances to the variables and parameters, the magnitude of these disturbances is shown in the following table.

Table 2.3. *Maximum random variability added to variables and parameters.*

Variable/Parameter list	Maximum percentage of variability (±%)	Type
Initial biomass concentration	30.00	Measured
Maximum specific growth rate of biomass	6.10	Measured
Maximum specific growth rate of penicillin	6.10	Measured
Initial substrate concentration	2.00-10.00	Manipulated
Initial dissolved oxygen concentration	3.33	Measured
Initial volume	0.86	Measured
Initial weight	0.81	Measured
Initial carbon dioxide concentration	2.63	Measured
Initial oxygen concentration	25.00	Measured
Initial pH	1.54	Measured
Initial temperature	0.17	Measured
Initial phenylacetic acid concentration	3.57	Measured
Initial nitrogen concentration	2.94	Measured
Overall mass transfer coefficient	11.76	Measured
Phenylacetic acid concentration in the feed	3.77	Measured
Nitrogen concentration in the feed	1.33	Measured

Furthermore, a low passing filter with a cut-off about 99.5% is implemented on the penicillin specific growth rate, biomass specific growth rate, oil mass inlet concentration, acid molar inlet concentration, base molar inlet concentration, phenylacetic mass inlet concentration, coolant inlet temperature, and oxygen percentage inlet concentration. This means that the 0.5% of the lowest values are discarded.

2.2.2.1 Substrate balance

The substrate is a mixture of sugars and soybean oil, the latter is the second carbon source and provides anti-foaming action. The substrate mass balance is:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -Y_{s/X} \cdot r_e - Y_{s/X} \cdot r_b - m_s \cdot r_m - Y_{s/P} \cdot r_P + \frac{F_s \cdot c_s}{V} + \frac{F_{oil} \cdot c_{oil}}{V} - \frac{F_{in} \cdot S}{V} \quad ; (2.15)$$

where S is the substrate concentration (g/L), $Y_{s/X}$ is the biomass substrate yield coefficient (g substrate/g biomass), r_e is the rate of extension (g biomass/(L·h)), r_b is the rate of branching (g biomass/(L·h)), m_s is the substrate maintenance term (g/(g·h)), r_m is the rate of maintenance (g/L), $Y_{s/P}$ is the penicillin substrate yield coefficient (g substrate/g penicillin), r_P is the rate of production (g penicillin/(L·h)), F_s is the sugar feed rate (L/h), c_s is the sugar concentration (g substrate/L), V is the vessel volume (L), F_{oil} is the soybean oil feed rate (L/h), c_{oil} is the soybean oil concentration (g soybean oil/L) and F_{in} represent all the process feed rate inputs except the discharge rate [L/h].

Two manipulated variables are influenced by the substrate: initial substrate concentration and recipe. The initial substrate concentration is the value of the concentration of this mixture at the beginning of the process. The recipe defines the amount of substrate to be fed at each moment of time along the batch. This substrate control strategy (Montague et al.,1986) shows a sharp increase in the substrate flow rate around 20h from the start of the batch to ensure excess of substrate that led to a rapid biomass growth at the beginning of the process. The flow rate is then reduced to low values to maximize penicillin production.

2.2.2.2 pH

The pH is modelled through a hydrogen ion (H^+) balance, which considers the generation of hydrogen ions during the growth phase, metabolic production, and maintenance activities, in addition to the introduction of acid/base and other process inputs:

$$\frac{\mathrm{d}[H^+]}{\mathrm{d}t} = \gamma_1 \cdot (\mu_x \cdot X + \mu_p \cdot C_\mathrm{P}) - m_{pH} \cdot X - \gamma_2 \cdot F + [H_1^+] \quad ; \qquad (2.16)$$

where γ_1 is the hydrogen ion production term during biomass and penicillin growth (-), μ_x is the growth rate of biomass (h^{-1}), μ_P is the growth rate of penicillin (h^{-1}), m_{pH} is the constant ion production related to the maintenance activities of the biomass, γ_2 is the hydrogen ion process inputs term (-) and account for the disturbances in inputs feed F (L/h), and the effect of acid/base additions is modelled by $[H_1^+]$ (mol/L). To model the effect of pH deviations on the growth rate of biomass, an inhibition term is included (Nielsen et al., 2003) as:

$$\mu_x \approx \mu_{x_{max}} \cdot \left[\frac{1}{1 + K_1/[H^+] + [H^+]/K_2}\right] \quad ; \qquad (2.17)$$

where K_1 and K_2 represent the higher and lower hydrogen ion concentration ($[H^+]$), respectively, at which the biomass growth rate is observed to be half its maximum value. Furthermore, $\mu_{x_{max}}$ is the maximum specific growth rate of biomass (h^{-1}). Note that these values are specific to the strain and process under consideration.

The influence of pH changes on fermentation processes is considerable and deviations of as little as 0.2 or 0.3 may have an adverse effect on a batch (Vogel and Todaro, 1997). For this reason, a PID control is implemented to keep the pH around an optimal set-point.

2.2.2.3 Penicillin production

The penicillin production mechanism is described through its mass balance:

$$\frac{\mathrm{d}C_\mathrm{P}}{\mathrm{d}t} = r_P - r_h - \frac{F_{in} \cdot C_\mathrm{P}}{V} \quad ; \qquad (2.18)$$

where r_P and r_h are the rate of product formation and product hydrolysis respectively (g/(L·h)) and F_{in} is the total flow in (L/h). Temperature and pH must be controlled to facilitate the cell growth and the product formation, because they influence the rate of product hydrolysis r_h. The degradation of penicillin is modelled using a second-order polynomial (Kheirolomoom et al. 1999):

$$\log(r_h) = B_1 + B_2 \cdot pH + B_3 \cdot T + B_4 \cdot pH^2 + B_5 \cdot T^2 \quad ; \qquad (2.19)$$

where constants B_1 - B_5 are equal respectively to: -67.8, -1.82, 0.36, 0.12, -4.9·10^{-4}; and are calculated to obtain a hydrolysis rate of 0.003 (h^{-1}) at temperature of 298 (K) and pH of 6.5 (-) (Paul and Thomas, 1996).

Chapter 3

Experimental data

This Chapter presents the available data, which are generated in silico by means of the simulated processes. The preprocessing to obtain data which can be treated through the methodologies presented in Chapter 2 is shown in terms of: data resampling, elimination of useless variables, and scaling.

3.1 Pilot-scale data

Recalling that the objective of this Thesis is to quantify the number of pilot scale batches that allows an improved description of the process at the industrial scale through transfer learning methods, to assess the effect of different data generation strategies, two independent datasets are created: *i*) using a full factorial design approach (FFD) and *ii*) using a gaussian design approach (GAU). Both the datasets are composed of 200 batches, which are used for calibrating the models, and 50 batches used to validate the models.

3.1.1 Design variables and ranges

For both the datasets, the design variables are: the initial substrate concentration, the substrate feed rate (that has the same effect of the recipe in the Indpensim simulator) and the pH set-point. These design variables are selected because they are: *i*) easy to manipulate, *ii*) important for the response variable; and *iii*) strictly linked to the manipulated variables of the industrial scale process simulator. The ranges for manipulating the design variables are reported in Table 3.1.

Table 3.1. *Ranges of manipulated variables for the Pensim simulator.*

Manipulated variable	Lower boundary	Upper boundary
Initial substrate concentration	1.0 (g/L)	61.0 (g/L)
Substrate feed rate	0.1 (L/h)	0.5 (L/h)
pH controller set-point	5.0 (-)	5.4 (-)

The ranges, taken from literature (Birol et al., 2001; Chu et al. 2018), are adjusted to avoid any numerical issue in the penicillin concentration values.

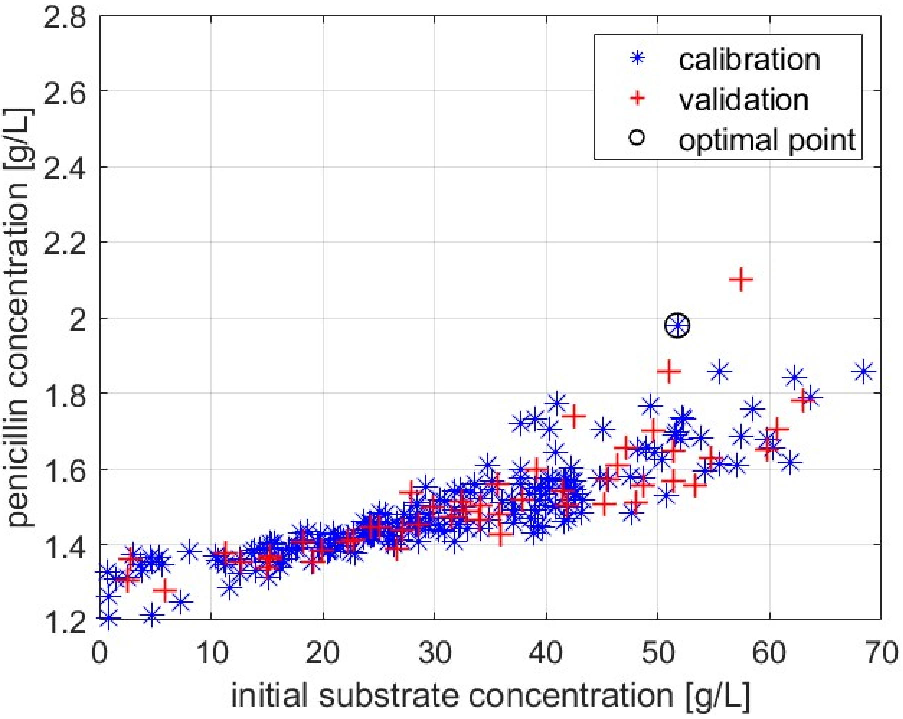
3.1.2 Gaussian design approach

In the gaussian design approach, the batches are generated sampling the design variables from independent gaussian distributions. The mean and the standard deviation of the gaussian distributions used in this Thesis are reported in Table 3.2. The means are selected as the mean value of the means in Table 3.1, and the standard deviations are set to one fourth of the range in Table 3.1, in this way 95% of the batches fall inside the ranges.

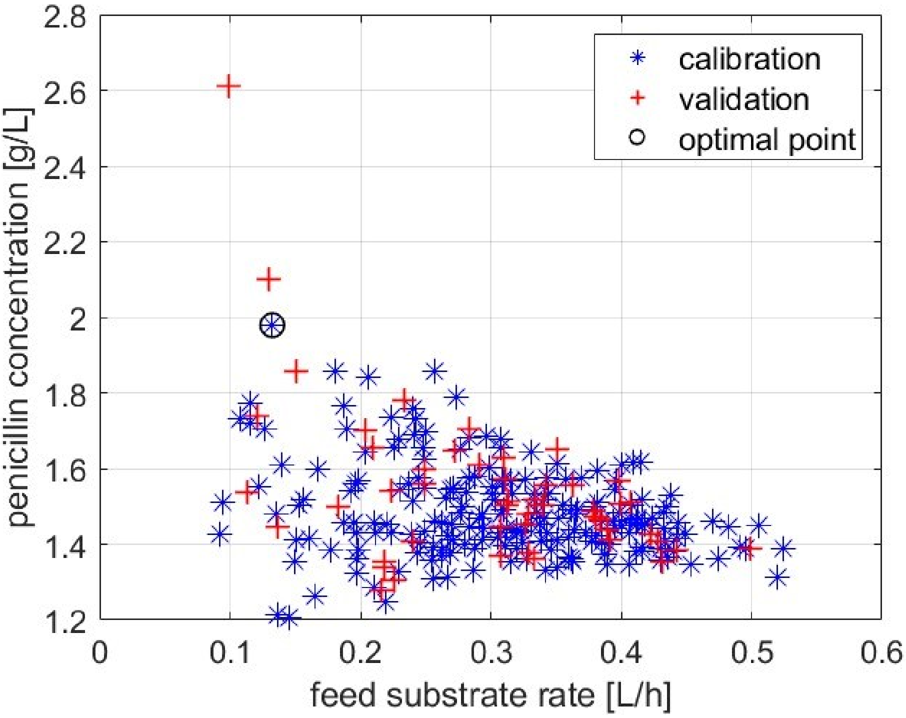
Table 3.2. *Means and standard deviations of the design variables used to generate data in the gaussian design approach on the Pensim simulator.*

Manipulated variable	Mean	Standard deviation
Initial substrate concentration	31.0 (g/L)	15.0 (g/L)
Substrate feed rate	0.3 (L/h)	0.1 (L/h)
pH controller set-point	5.2 (-)	0.1 (-)

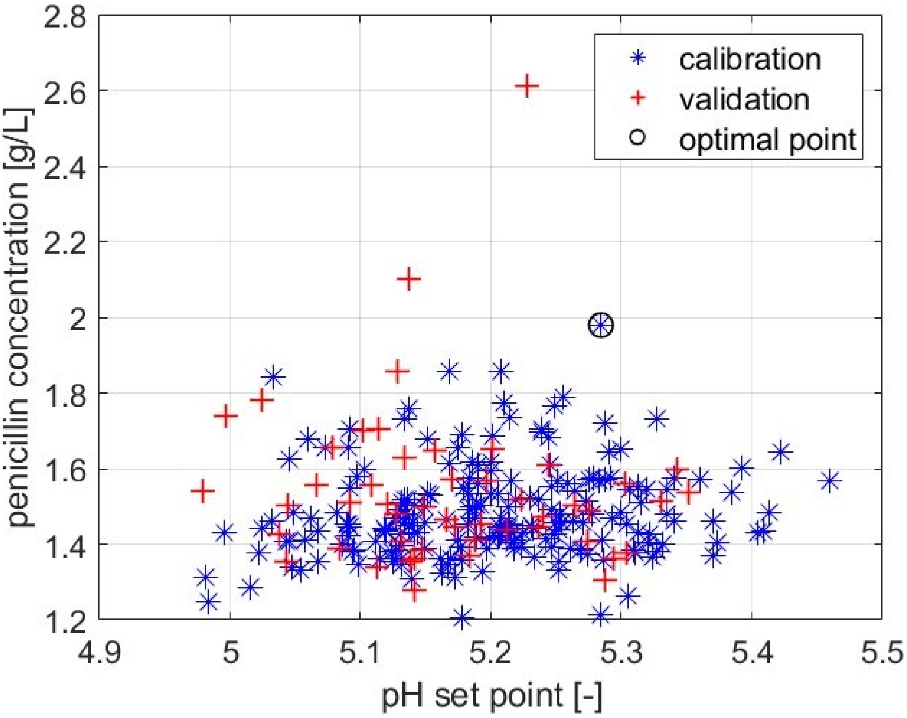
Using this approach, the operating conditions for 250 batches (200 for model calibration and 50 for validation) are generated (Figure 3.1). The validation batches were generated for an initial PLS model construction for the pilot scale.



(a)



(b)



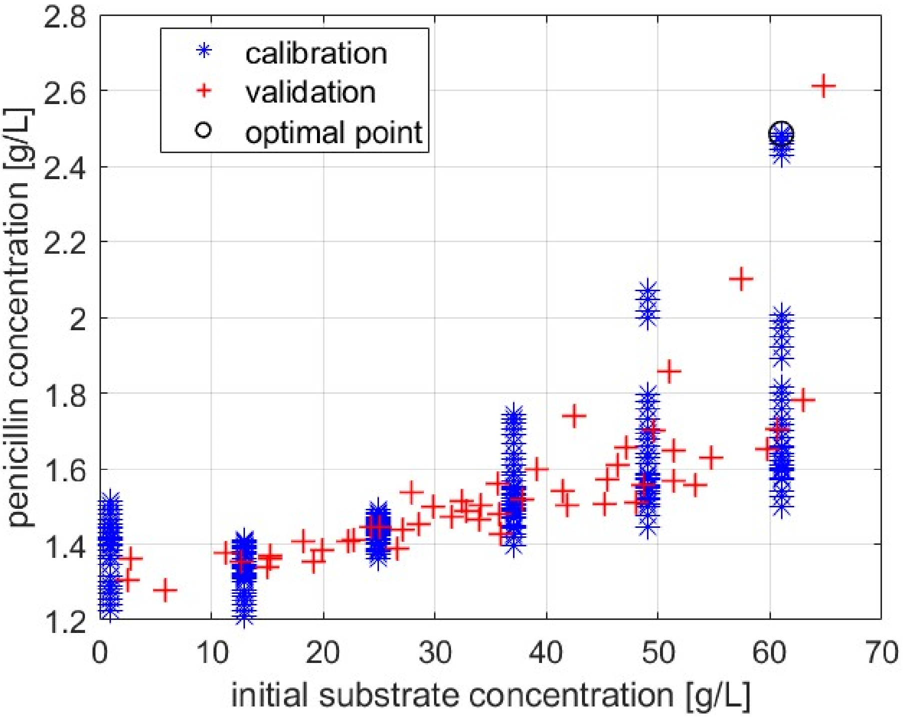
(c)

Figure 3.1. *Design variables values of calibration and validation batches generated with the Gaussian approach by the Pensim simulator: (a) Penicillin concentration values as function of the initial substrate concentration; (b) Penicillin concentration values as function of the feed substrate rate. (c) Penicillin concentration values as function of the pH set-point. Calibration batches are the blue stars, validation are the red crosses, and the black circle is the batch with the highest penicillin concentration in the calibration dataset.*

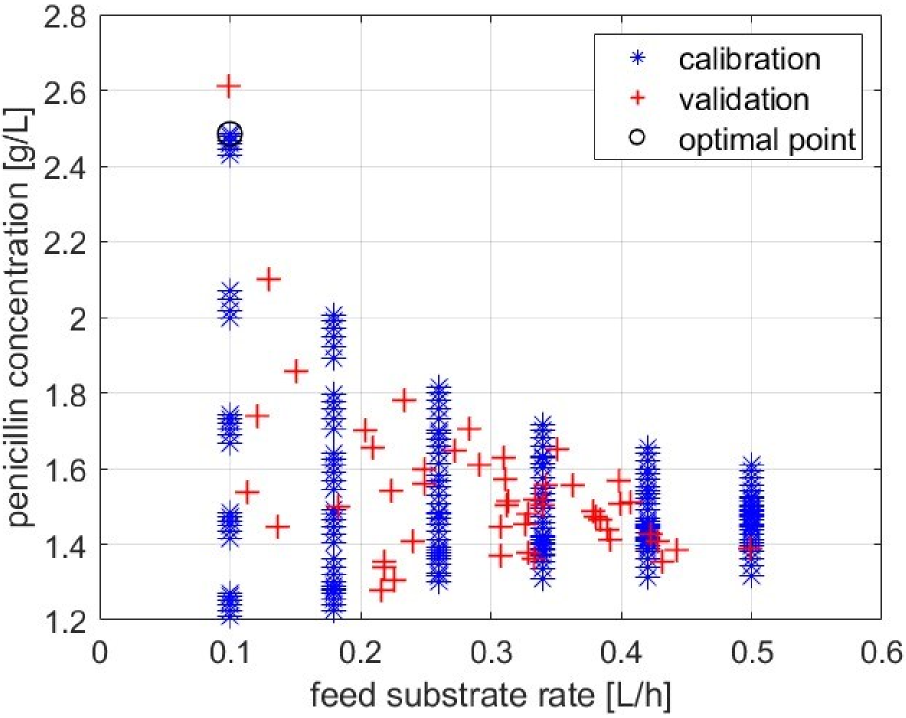
As previously mentioned in Sub-subsections 2.1.2.1 and 2.1.2.2, the importance of the initial substrate concentration and the feed substrate concentration is demonstrated in Figure 3.1 as they induce the largest changes in penicillin concentration. On the contrary, the pH set-point has a lower effect on the penicillin concentration than the other two design variables.

3.1.3 Full factorial design approach

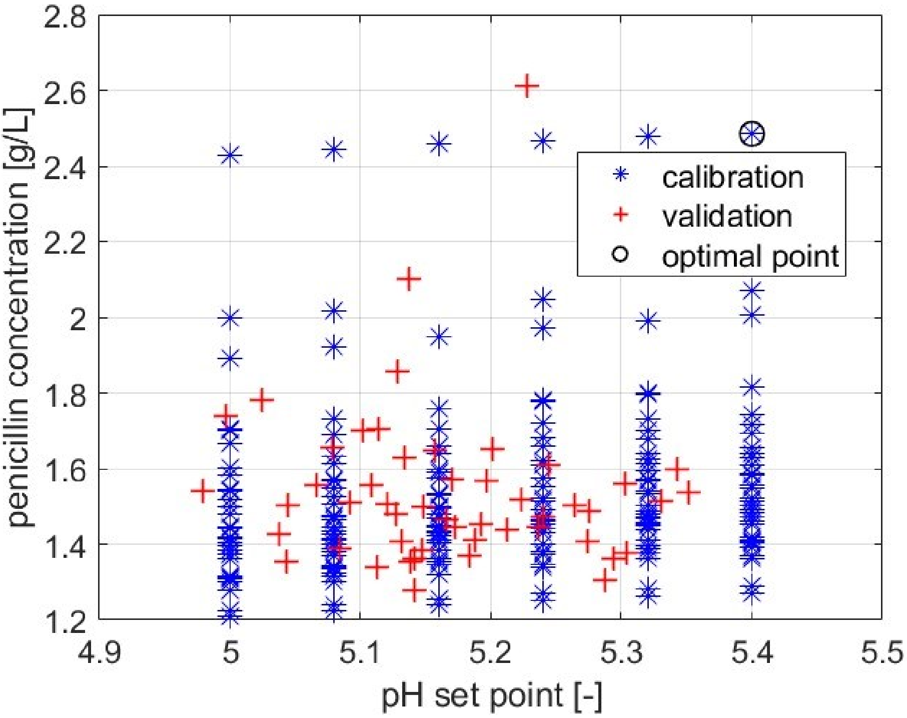
To generate the operating conditions for the full-factorial approach, a full-factorial design with 3 factors and 6 levels is used. Accordingly, the design span of each variable is divided into six segments, and the operating conditions result from a combination of the three factors in the 6 levels. The selected levels for the initial substrate concentration are: 1, 13, 25, 37, 49, and 61 (g/L). The selected levels for the substrate feed rates are: 0.10, 0.18, 0.26, 0.34, 0.42, and 0.50 (L/h). The selected levels for the pH controller are: 5, 5.08, 5.16, 5.24, 5.32, and 5.40 (-). Among the resulting 216 (6^3) batches, 200 were randomly selected to be used as calibration dataset, while the remaining 16 are discarded, this is done to match the number of calibration batches generated in Sub-section 3.1.2. No new validation batches are created with this approach. The same 50 validation batches crated in Section 3.1.2 for validation are used here as validation set.



(a)



(b)



(c)

Figure 3.2. *Design variables values of calibration and validation batches generated with the Full Factorial Design approach by the Pensim simulator: (a) Penicillin concentration values as function of the initial substrate concentration. (b) Penicillin concentration values as function of the feed substrate rate. (c) Penicillin concentration values as function of the pH set-point. Calibration batches are the blue stars, validation are the red crosses, and the black circle is the batch with the highest penicillin concentration in the calibration dataset.*

Even in this case the major effects on penicillin are related to the initial substrate concentration and the feed substrate rate, as already seen in Sub-subsections 2.1.2.1 and 2.1.2.2. Furthermore, from a deeper analysis on the effect of the variables, a small effect of the pH controller set-point is highlighted, especially when the initial substrate concentration is high, and the substrate feed rate is low.

3.1.4 Data treatment

In both approaches, 200 batches are generated, which are composed of 15 variables (aeration rate, agitator power, substrate feed rate, substrate feed temperature, volume, pH, temperature, generated heat, acid flow rate, base flow rate, and water flow rate) measured in 241 time instants (one sampling per hour from 0 to 240 hours) and 4 concentration variables (substrate, oxygen, biomass, and carbon dioxide) measured in 21 time instants (every 12h). The multiway methodologies used in this Thesis require the batch-wise unfolding of the matrices to properly manage the time evolution of batches (Section 1.1). In batch-wise unfolding the variables at measured at different time instants are horizontally concatenated to obtain two-dimensional matrices $\mathbf{X}_{pil,\mathrm{FFD}}$ and $\mathbf{X}_{pil,\mathrm{GAU}}$ of dimensions $[N \times V \cdot T] = [200 \times 2735]$, where 2735 results from 15·241+21·4. The response variable is organized in a vectors $\mathbf{Y}_{pil,\mathrm{FFD}}$ and $\mathbf{Y}_{pil,\mathrm{GAU}}$ of dimensions $[N \times P] = [200 \times 1]$, P is equal to 1 because only the last time instant is considered, for instance the outlet penicillin concentration.

3.1.4.1 Elimination of constant variables

Some variables, such as aeration rate, agitator power and the substrate feed temperature, are the process manipulated variables and they are kept constant all over the batch duration. Accordingly, since these variables do not add information to the model, they are excluded from the datasets. The resulting dataset used for modelling have dimensions $[200 \times 2012]$.

3.1.4.2 Scaling

Data are autoscaled prior the analysis by subtracting to each column its mean value and dividing for its standard deviation as:

$$\mathbf{x}_{v,n} = \frac{(\mathbf{x}_v - \bar{\mathbf{x}}_v)}{\sigma(\mathbf{x}_v)} \quad ; \tag{3.1}$$

where $\mathbf{x}_v$is the column v of the real values matrix $\mathbf{X}_{pil}$, $\mathbf{x}_{v,n}$ are the autoscaled values, $\bar{\mathbf{x}}_v$ is the mean of the column, and $\sigma(\mathbf{x}_v)$ is the standard deviation of the column.

3.2 Industrial-scale data

In the industrial case, the same number of batches is generated, namely 200 for model calibration and 50 for validation, using only the Gaussian design approach. The validation batches generated at the industrial scale are used to assess the reliability and performance of both PLS and JYPLS.

3.2.1 Design variables and ranges

The selected design variables are the initial substrate concentration, the recipe and the pH controller set-point, to manipulate factors and obtain effects like the pilot case. In the Table 3.3, the ranges in which the selected design variables are varied are presented.

Table 3.3. *Ranges of manipulated variables for the Indpensim simulator.*

Manipulated variable	Lower boundary	Upper boundary
Initial substrate concentration	1.0 (g/L)	5.0 (g/L)
Recipe	-1.0 (-)	1.0 (-)
pH controller set-point	6.0 (-)	6.5 (-)

The ranges for the initial substrate concentration and the pH controller set-point, taken from literature (Goldrick et al., 2014), are adjusted to avoid numerical issues.

3.2.1.1 Recipe

The recipe is a predetermined time profile of feed flowrate provided to the batch at each time point. The two default recipes of the simulator are considered and called -1 and 1 (-1 the recipe with the lowest values of the flowrate and 1 the one with the largest values). A third recipe, the average of the two default ones, is added and named 0.

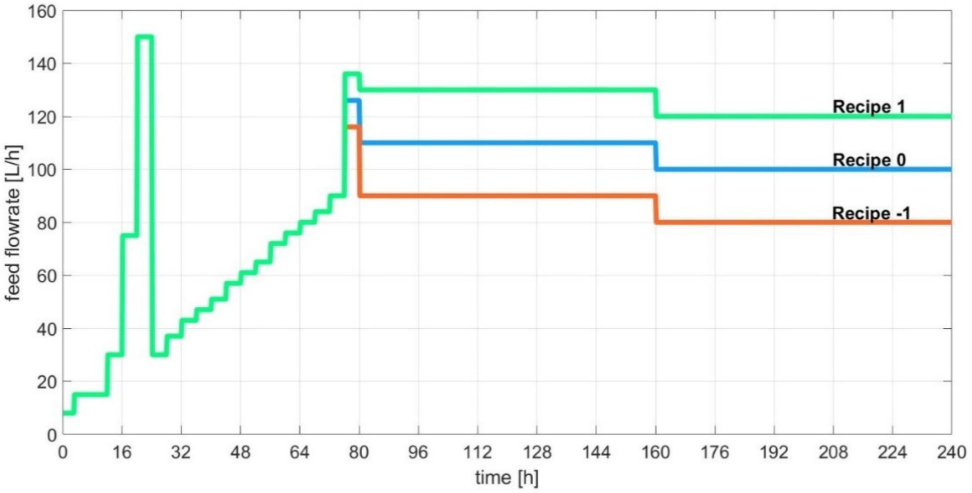


Figure 3.3. *Different recipes used in the generation of artificial data with the Indpensim simulator.*

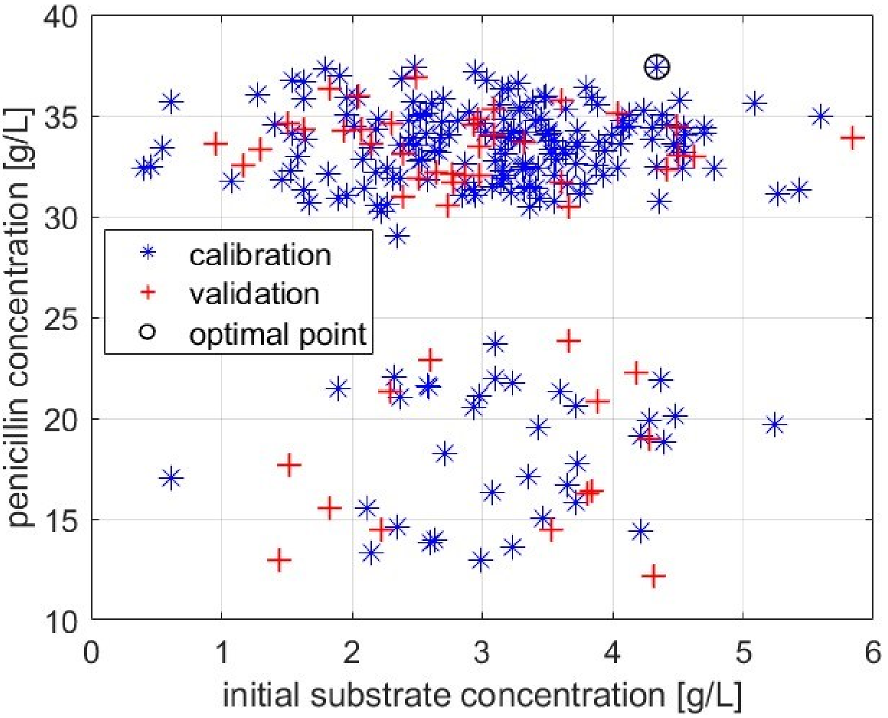
3.2.2 Generation of batch operating conditions

In the gaussian design approach, the batches are generated by sampling the values of the design variables from independent gaussian distributions. The mean and the standard deviation of the gaussian distributions used in this Thesis are reported in Table 3.4. The means are selected as the mean value of the ranges (Tables 3.3), and the standard deviations are set to one fourth of the range span for the same motivation of Section 3.1.2.

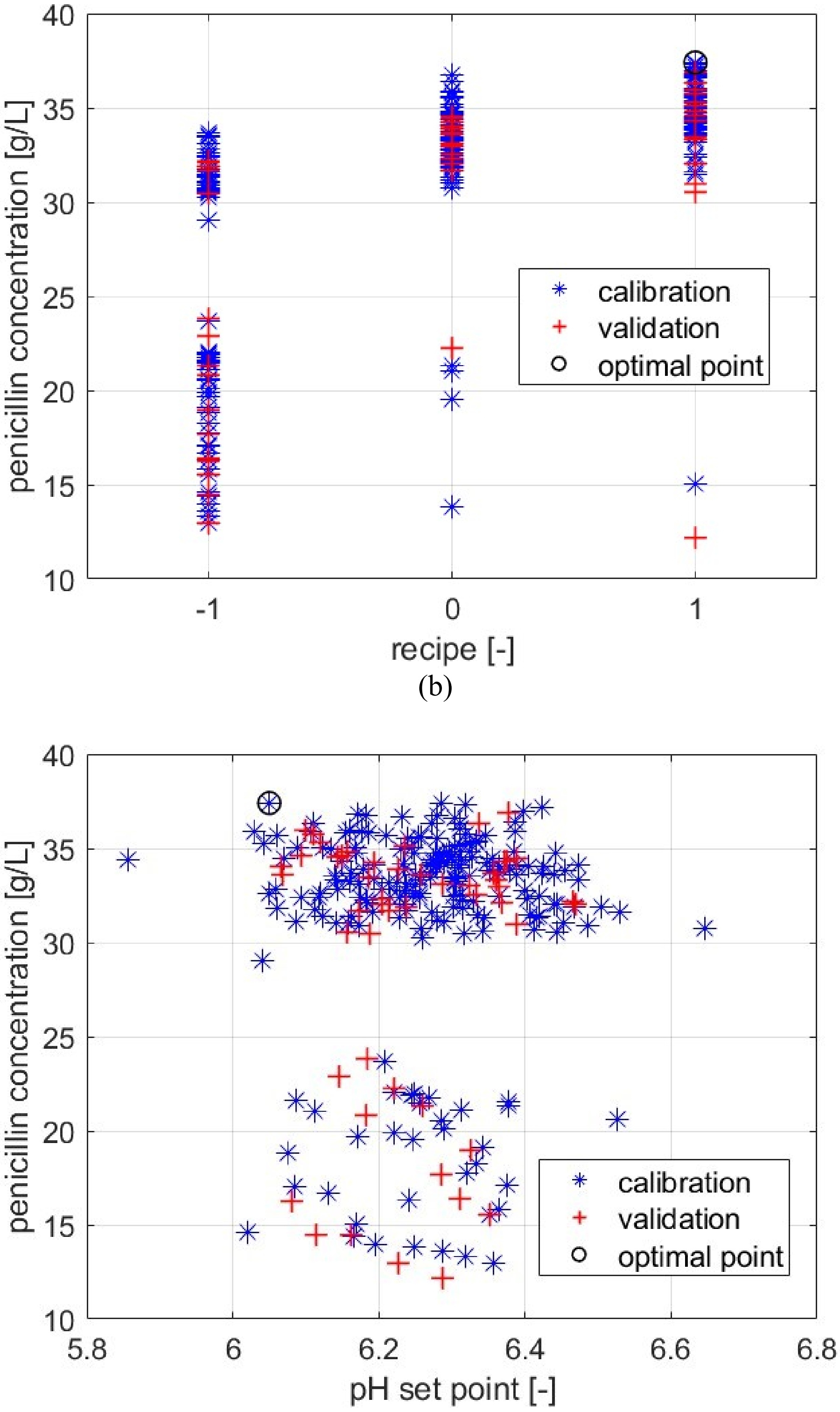
Table 3.4. *Means and standard deviations of the design variables used to generate data in the gaussian design approach on the Indpensim simulator.*

Manipulated variable	Mean	Standard deviation
Initial substrate concentration	3.00 (g/L)	1.000 (g/L)
pH controller set-point	6.25 (-)	0.125 (-)

The recipe is assigned randomly among the selected three recipes (i.e., $[1{,}0,-1]$), in such a way to obtain an equal number of batches with each recipe. Accordingly, 67 calibration batches follow the 1 recipe, 67 follow the 0 recipe, and 66 follow the -1 recipe. For the validation batches, 17 are follow the 1 recipe, 16 follow the 0 one, and 17 follow the -1 recipe.



(a)



(b)

(c)

Figure 3.4. *Design variables values of calibration and validation batches generated with the Gaussian approach by the Indpensim simulator: (a) Penicillin concentration values as function of the initial substrate concentration. (b) Penicillin concentration values as function of the feed substrate rate. (c) Penicillin concentration values as function of the pH set-point. Calibration batches are the blue stars, validation are the red crosses, and the black circle is the batch with the highest penicillin concentration in the calibration dataset.*

As previously mentioned in Sections 2.2.2.1 and 2.2.2.2, the importance of the recipe is demonstrated in Figure 3.4 as it induces the largest changes in penicillin concentration. On the contrary, the initial substrate concentration and the pH set-point have a lower effect on the penicillin concentration than the other two design variables. The two clusters of batches arise for the inhibition effects shown in Section 2.2.2.

3.2.3 Data treatment

At the industrial scale, 250 batches are generated, which are composed of 23 variables (aeration rate, agitator power, substrate feed rate, acid flow rate, base flow rate, cooling water flow rate, heating water flow rate, water inlet, pressure, dumped broth, substrate concentration, dissolved oxygen concentration, vessel volume, vessel weight, pH, temperature, generated heat, carbon dioxide percent in off-gas, oxygen uptake rate, on-line phenylacetic acid concentration, oil flow, oxygen in percent in off-gas, and carbon evolution rate) measured in 241 time instants (one sampling per hour from 0 to 240 hours) and 4 variables (off-line phenylacetic acid concentration, ammonia concentration, biomass concentration and viscosity) measured in 21 time instants (every 12h). The resulting data is unfolded in a batch-wise fashion to obtain the matrix $\mathbf{X}_{ind}$ of dimensions $[N \times V \cdot T] = [250 \times 5627]$. The response variable is organized in a vector $\mathbf{Y}_{ind}$ of dimensions $[N \times P] = [250 \times 1]$, again P is equal to 1 because only the final penicillin concentration is considered.

3.2.3.1 Elimination of constant variables

Some variables, such as aeration rate, agitator power, water inlet, pressure, dumped broth, on-line phenylacetic acid, and oil flow, are the process manipulated variables and are kept constant all over the batch duration. Accordingly, since these variables do not add information to the model, they are eliminated from the datasets. The resulting dataset used for modelling are $[250 \times 3940]$. At this point the 2D matrix is horizontally divided in two different matrices, one for the calibration batches containing the first 200 batches ($[200 \times 3940]$), and one for the validation batches containing the last 50 batches ($[50 \times 3940]$).

3.2.3.2 Scaling

Process variables in the industrial dataset $\mathbf{X}_{ind}$ are scaled in different ways according to the measurement type. The variables substrate feed rate, dissolved oxygen concentration, vessel volume, vessel weight, pH, temperature, generated heat, carbon dioxide percent in off-gas, oxygen uptake rate, oxygen in percent in off-gas, carbon evolution rate, off-line phenylacetic acid concentration, ammonia concentration, biomass concentration and viscosity, are autoscaled as explained in Section 3.1.4.2. The other variables, such as acid flow rate, base flow rate, cooling water flow rate, heating water flow rate, and substrate concentration, are scaled using a min-max technique. A different scaling method is used because these variables are the manipulated variables of the process and might induce numerical issues as they are typically set either at the minimum or at the maximum value. For the substrate concentration the same problem arises at time instant where almost all the batches have value close to zero. The equation used for the min-max scaling is:

$$\mathbf{x}_{v,n} = \frac{\mathbf{x}_v}{\max(\mathbf{x}_{v*}) - \min(\mathbf{x}_{v*})} \quad ; \tag{3.2}$$

where$\mathbf{x}_{v,n}$ is the column of the min-max scaled values of a real values column $\mathbf{x}_v$ of a variable v, $\max(\mathbf{x}_{v*})$ is the maximum value choice for a variable v, and $\min(\mathbf{x}_{v*})$ is the minimum value choice for the variable v. To use the min-max technique the minimum and maximum limits for these variables are identified (Table 3.5).

Table 3.5. *Minimum and maximum values used for the scaling in the Indpensim simulator.*

Variable	Minimum value	Maximum value
Acid flow rate	0 (L/h)	20 (L/h)
Base flow rate	0 (L/h)	225 (L/h)
Cooling water flow rate	0 (L/h)	700 (L/h)
Heating water flow rate	0 (L/h)	520 (L/h)
Substrate concentration	0 (g/L)	40 (g/L)

The values in Table 3.5 are obtained by rounding the maximum values of these variables, and using 0 for the minimum values, as it is the minimum physical value for these measurements.

Chapter 4

Penicillin concentration prediction performance: comparison between different strategies and sensitivity on the number of available data

This Chapter presents a comparison between multivariate models that predict the penicillin concentration at the end of the production batch. In particular, a PLS model built on industrial batches only is compared to a JY-PLS built on both pilot-plant data and industrial-plant data. Specifically, the sensitivity of the model performance to a varying number of experimental batches used for calibration is evaluated.

4.1 Procedure for performance comparison

The procedure presented in this Section is used to make a fair comparison between the PLS built on industrial data only, and JY-PLS, which considers both data from the pilot scale and the industrial scale to predict the penicillin concentration of the end product.

Specifically, the procedure used to calibrate and assess the performance of the JY-PLS is as follows:

1. Batches from the pilot scale simulator are chosen with the Kennard-Stone algorithm (Kennard & Stone, 1969). The Kennard-Stone algorithm is a space-filling sample selection method to identify a representative subset of data, which are typically used for model calibration or validation. The algorithm selects samples that are the most distant from each other in the feature space, guaranteeing at the same time a uniform distribution of the selected data. In this way, the most comprehensive representation of the entire data distribution is ensured. The number of pilot scale batches used in the JY-PLS are varied from 2 to 25.
2. A predefined number of industrial batches (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 40, 50, and 60) is randomly selected from the pool of 200 generated calibration batches. To improve the robustness of the results, this

3. procedure is repeated 100 times, each time selecting a different random subset of the data to calibrate the model. No more than 60 industrial batches are explored because itis a reasonable value in which the JYPLS does not add additional information to the models.
4. At each iteration, the model is built on the selected pilot and industrial data. The performance is evaluated, and the number of selected latent variables is recorded (Subsection 1.1.2.1). Performance is evaluated on the validation dataset composed by 50 industrial batches (Section 3.2).
5. The mean value and the standard deviation of the coefficient of determination are averaged over the 100 iterations to obtain a summary of the performance for each combination of the number of pilot scale and industrial scale batches.

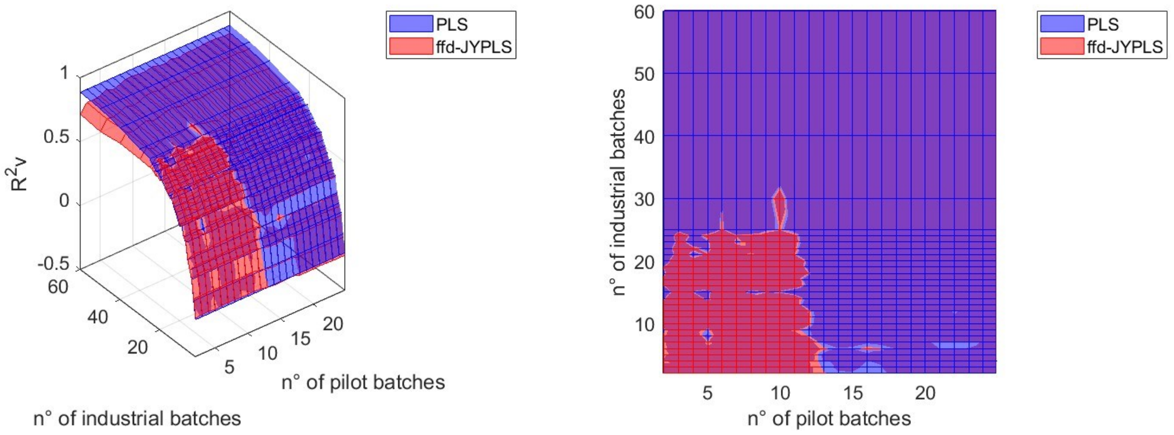
To calibrate and evaluate the PLS model the same procedure is followed, but considering only industrial batches (no information on the pilot scale is considered; step 1) and used to build the model (step 3). Furthermore, at step 2, the same industrial batches of JY-PLS are selected and used for the PLS, both for calibration and validation.

4.1.1 Analysis of the prediction performance with a varying number of calibration batches

In this Section, the comparison of model prediction performance is presented when a varying number of batches is used to calibrate the models. For the comparison between models, the coefficient of determination is used. This index results from an average over 100 iterations performed with a constant number of industrial calibration batches (step 4 of Section 4.1). For the JYPLS models the same pilot calibration batches (chosen as shown in step 1 of Section 4.1) are used for all the 100 repetitions.

4.1.1.1 ffd-JYPLS

The coefficient of determination is important because (as seen in Subsection 1.1.2.2) it gives a clear indication of the model accuracy. It is calculated on the industrial validation dataset.



(a) (b)

Figure 4.1. *Coefficient of determination R^2 on the industrial validation dataset at different numbers of pilot and industrial calibration batches: (a) isometric perspective. (b) view from above. In blue the PLS model, in red the ffd-JYPLS model.*

Figure 4.1 represents the values of the coefficient of determination evaluated on the validation industrial dataset. The blue surface is related with the PLS model, while the red one is related with the ffd-JYPLS (i.e., a model built on pilot scale batches generated with a full factorial design of experiments). In Figure 4.1b a view from above is presented to better highlight the region in which the ffd-JYPLS model has better performances (red zones) with respect to the PLS model. Clearly, the PLS model is independent with respect to the number of pilot batches, so the performance values are constant as the number of pilot batches varies. For both the models the coefficient of determination goes down decreasing the number of industrial batches, because the number of data that is not sufficient to have a good prediction of the process. The ffd-JYPLS until 13 pilot batches and 25 industrial batches works better than the PLS model, increasing the number of these batches two behaviors happen: *i*) if the number of industrial batches increases there is no need of using the pilot-scale batches because the industrial data provide sufficient information for the prediction; *ii*) if the number of pilot batches increases the model is tailored on the pilot scale losing performance on predicting the industrial-scale process.

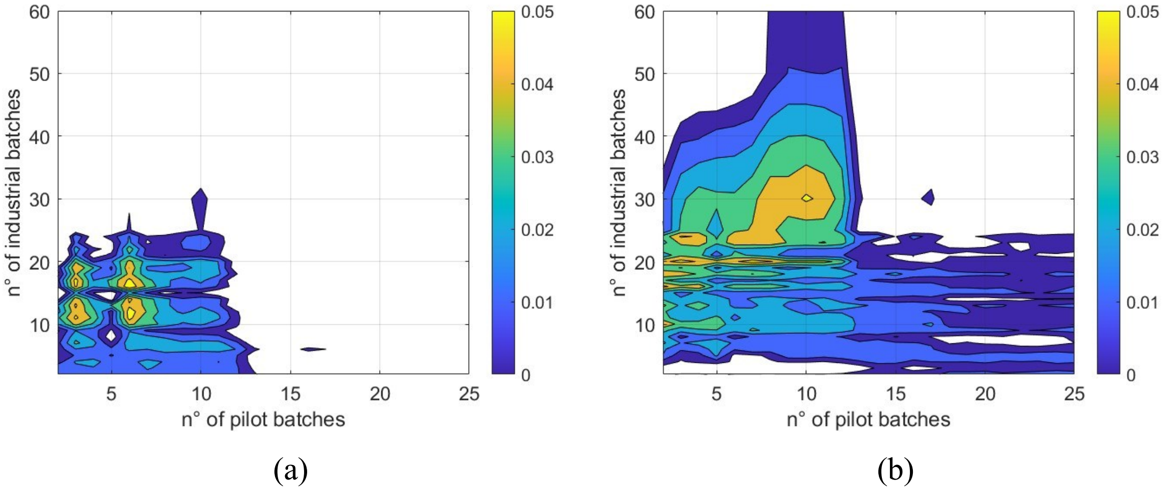


Figure 4.2. *Differences between the prediction performance (coefficient of determination R2) of PLS and the ffd-JYPLS: (a) R2 of ffd-JYPLS minus R2 of PLS view from above. (b) R2 standard deviation of PLS minus R2 standard deviation of ffd-JYPLS view from above.*

Figure 4.2 is useful to better understand the magnitude of the improvement in the predictive performance provided by the ffd-JYPLS. The white region of Figure 4.2a is the region in which the PLS has better predictive performances with respect to the ffd-JYPLS, namely the coefficient of determination of PLS is higher than the JY-PLS. Instead, the white region of Figure 4.2b is where the standard deviation (over the 100 iterations) is lower for the PLS, namely where the PLS performance is more precise (i.e., stable) than the ffd-JYPLS. In Figure 4.2a it emerges that until 5 pilot batches a zone with high variability is present, probably due to the low number of batches that in some cases are well distributed and sufficient to add information for a good prediction. Then a maximum is reached around 6 pilot batches, where the maximum improvement that can be reached with the ffd-JYPLS is 5% (yellow zone). Adding a larger number of pilot batches the performance decreases until the PLS model outperforms the JY-PLS (13 batches). It is interesting to highlight that using pilot scale batches makes the model more stable and less sensitive to the industrial batches selected (Figure 4.2b), because the standard deviation of R^2 in PLS is lower in the region until 25 industrial batches. This effect is not only confined to the region in which the ffd-JYPLS works better than the PLS, but for a larger range of the number of batches considered.

4.1.1.2 gau-JYPLS

For the gau-JYPLS (i.e., a model built on pilot scale batches generated with a gaussian distribution of the experiments) the same analysis as Subsection 4.1.1.1 is performed.

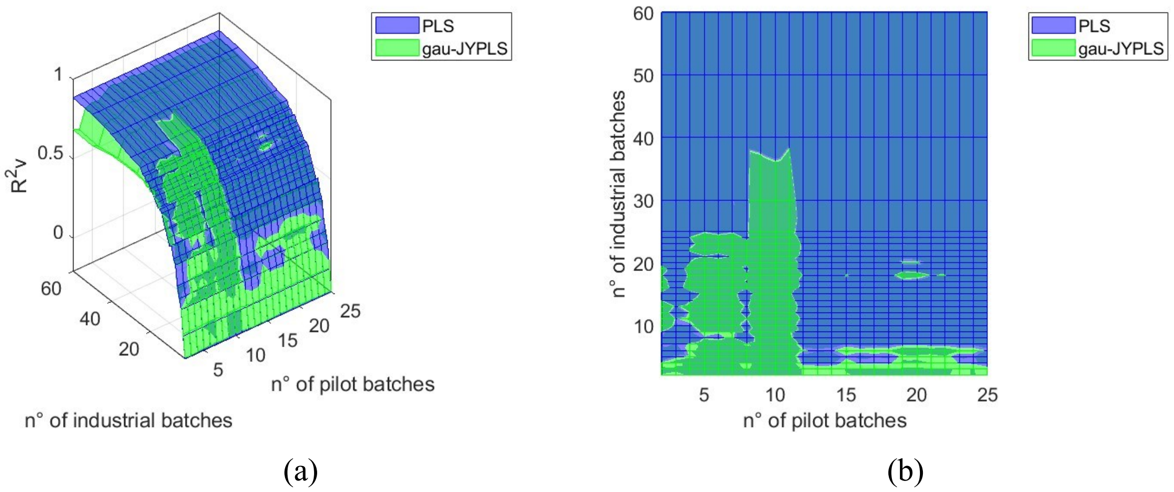
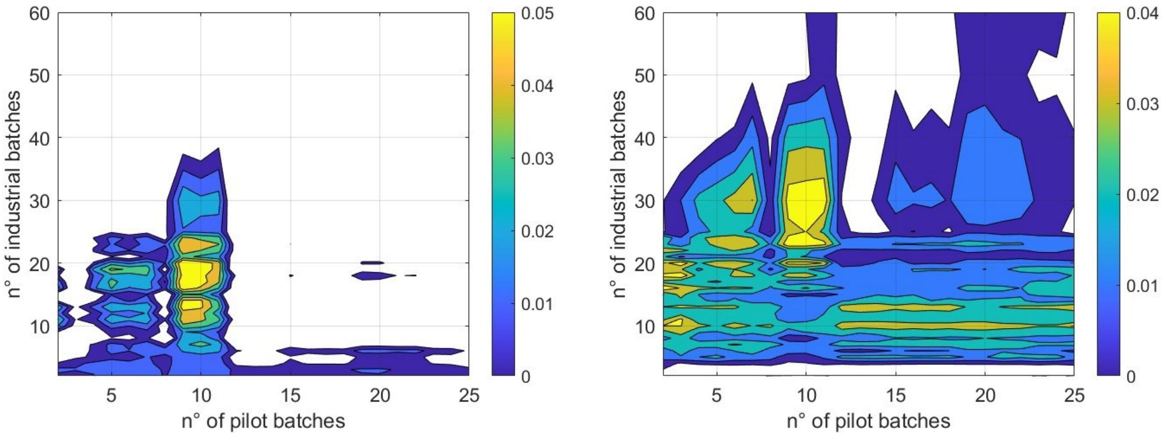


Figure 4.3. *Coefficient of determination R^2 on the industrial validation dataset at different numbers of pilot and industrial calibration batches: (a) Isometric perspective. (b) View from above. In blue the PLS model, in green the gau-JYPLS model.*

Figure 4.3 should be seen as Figure 4.1, the only difference is that in this case the green surface is the gau-JYPLS. As before, the coefficient of determination of both the models goes down decreasing the number of industrial batches for similar reasons as previously explained. This time a different area in which the JYPLS model works better than the PLS model is highlighted. A bigger area (up to 8 pilot batches) where the best performing model varies is present, this can be due to the generation of data through the gaussian distribution. In fact, with this approach 5% of the batches fall out the specified ranges of the design variables (Table 3.1), and these batches could be the first selected by the Kennard-Stone algorithm for the reasons shown in Section 4.1 step 1. Furthermore, gaussian generated data, explore a domain of variables conditions which is worse than the full factorial design ones, so the Kennard-Stone algorithm will also select less representative pilot calibration batches. Then up to 13 pilot batches and 30 industrial batches the gau-JYPLS model has better performances than the PLS model (green zones). Different from the ffd-JYPLS, beyond this point the performance of the PLS is better than JY-PLS, but in some spots JY-PLS model still outperforms PLS. This again can be due to the different method chosen to generate pilot scale data. In the blue areas the PLS model works better because either *i*) the number of industrial batches is too large and the pilot scales data does not add useful information to the model, or *ii*) a high number of pilot scale batches improves prediction of pilot scale data, but degrades the prediction of industrial scale one.



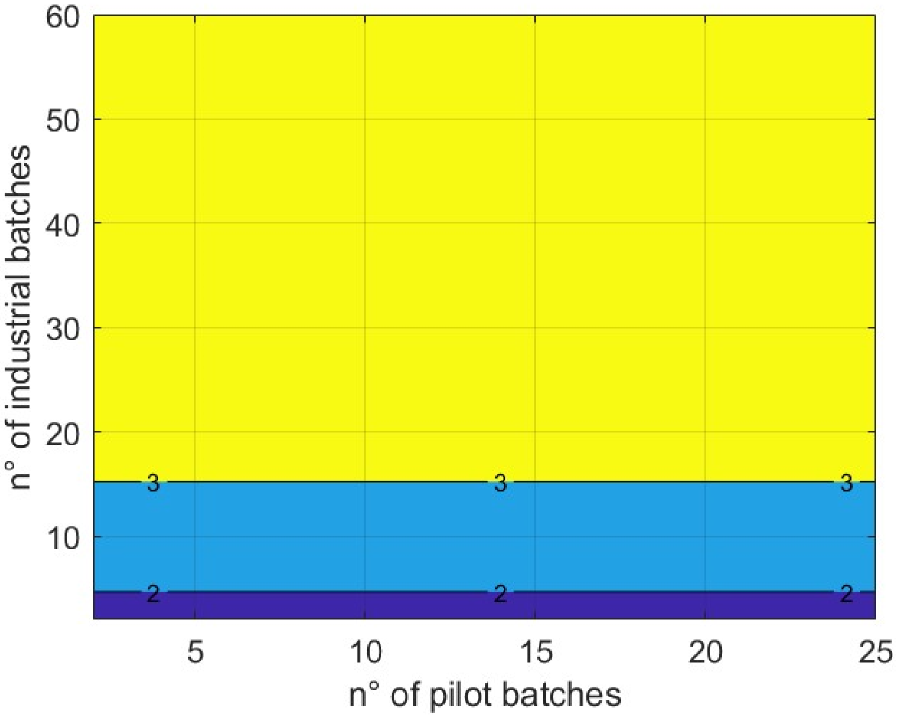
(a) (b)

Figure 4.4. *Differences between the prediction performance (coefficient of determination R^2) of PLS and the gau-JYPLS: (a) R^2 of ffd-JYPLS minus R^2 of PLS view from above. (b) R^2 standard deviation of PLS minus R^2 standard deviation of ffd-JYPLS view from above.*

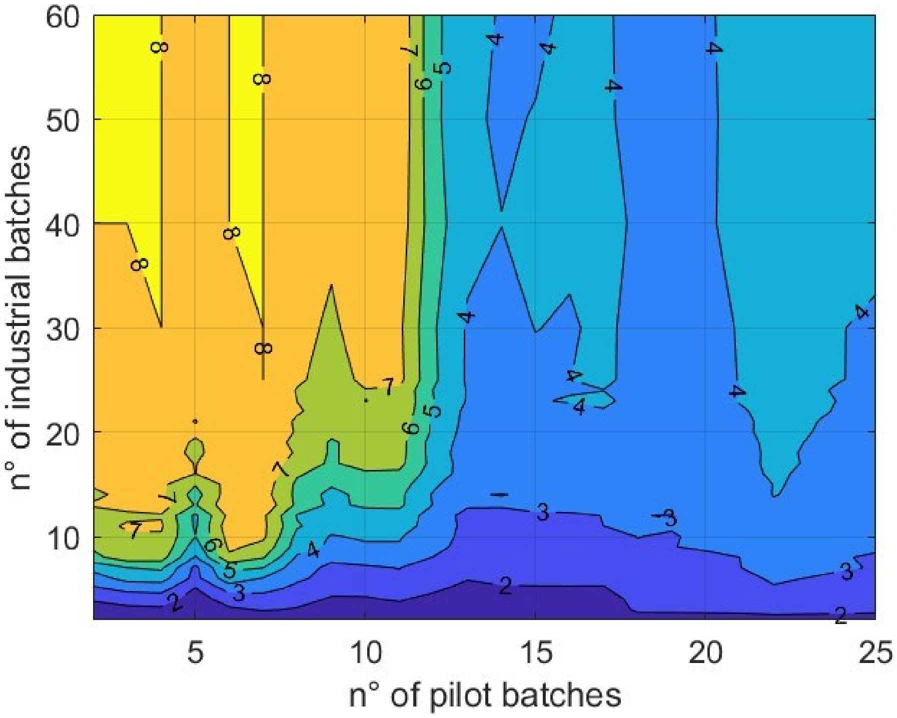
Figure 4.4 represents the magnitude of the improvement in the predictive performance provided by the gau-JYPLS and it is analogous to Figure 4.2. The maximum improvement in prediction performance of JY-PLS is reached at ~10 pilot batches, where maximum improvement, like previous case, of 5% is reached. In all the other areas where JY-PLS outperforms PLS the degree of improvement is lower compared to the maximum one (< 5%). Figure 4.4b shows that, in almost the entire experimental domain, the JY-PLS has a lower standard deviation and thus lower performance variability apart from a few areas reduced to high industrial batch numbers. The maximum is found at about ~10 pilot batches and 20-30 industrial batches. Furthermore, this effect is greater than the one that happens in the ffd-JYPLS case.

4.1.2 Analysis on the number of latent variables

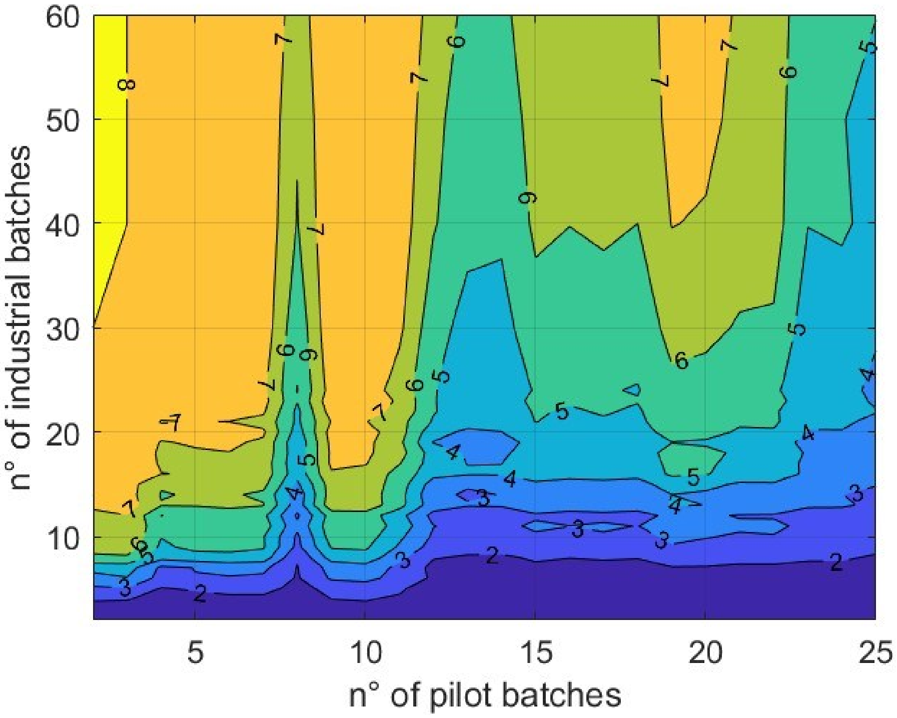
Is important to analyse the number of latent variables chosen for the model construction, because this can explain some of the behaviours observed in the previous Section.



(a)



(b)



(c)

Figure 4.5. *Mean number of latent variables chosen at different number of pilot and industrial batches: (a) PLS model. (b) ffd-JYPLS. (c) gau-JYPLS.*

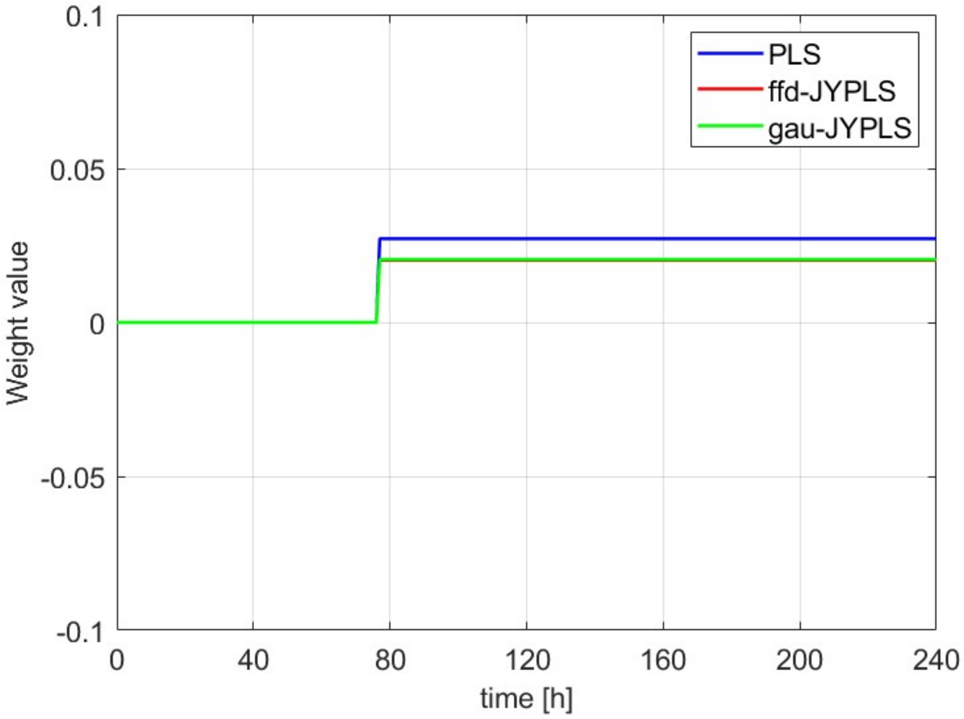
Figure 4.5 shows the number of selected latent variables, for all models, when the number of pilot and industrial batches varies. The JYPLS models use more latent variables to describe the process (Figure 4.5), but this can be expected because they take information from both the sources. In both the JY-PLS models (Figure 4.2a and 4.4a), a precise number of pilot batches (5 for the ffd-JYPLS and 8 for the gau-JYPLS) is present in which the number of latent variables chosen is not congruent with the rest of the plot. It is interesting because it is precisely the number of pilot scale batches that can be defined as the end of the initial variability region (the region in which the JYPLS performs better than the PLS before the more stable performance area). Furthermore, a great difference in the right side of the plots is observed. For the gau-JYPLS more latent variables are constantly chosen which can provide an explanation for the lower standard deviation of the R^2 indicating a more stable model.

4.2 Models weights analysis

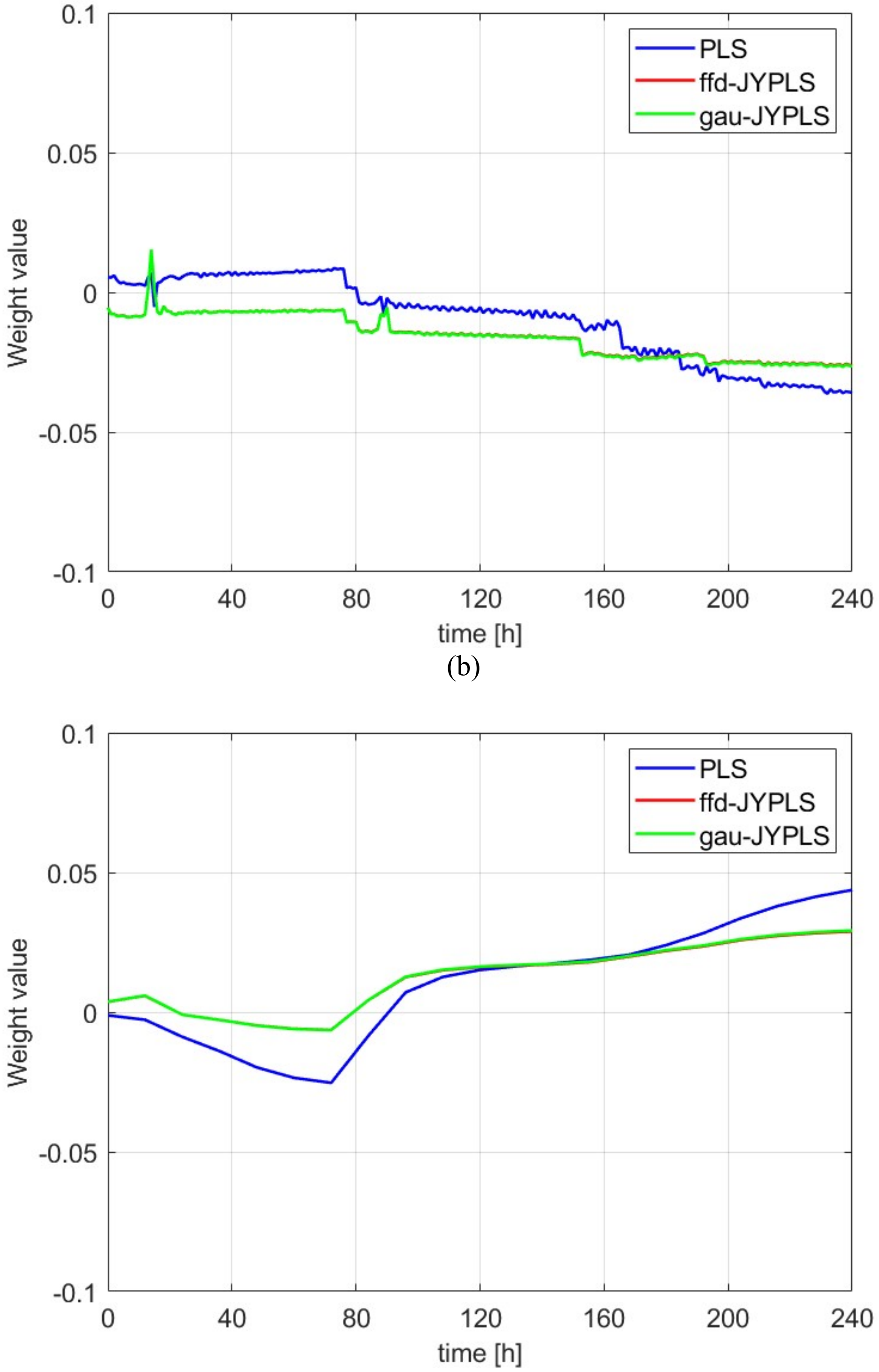
An analysis of the weights of the models is proposed in this Section. For the sake of simplicity, only one case is considered for each model, because the objective is to understand if there are differences in the PLS and in the JYPLS accounting for the variables. For the PLS 25 industrial calibration batches are chosen, instead for both the JYPLS model 8 pilot calibration batches and 15 industrial calibration batches are chosen. Three models are created with these specifications which are good enough and with similar performance (coefficient of determination around 0.85). Only the most important differences are exposed in the next Sub-sections.

4.2.1 Analysis of the first latent variable weights

For the first latent variable the value of the weights of the two JYPLS models are close. The PLS and the JYPLS have the similar weights, both in values and trends; the major differences are on the variables that the pilot simulator does not measure, as: phenylacetic acid concentration, ammonia concentration, and off-gas. For the variables which are measured in the simulators of both scales the most significant difference are captured in the following figures.



(a)



(b)

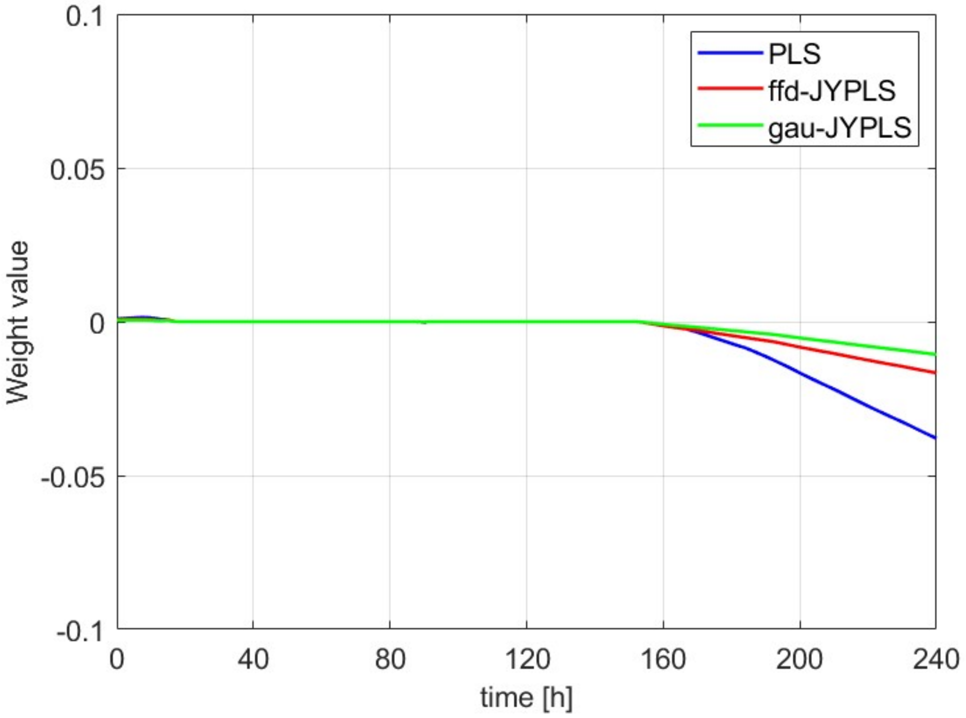
(c)

Figure 4.6. Weights values along batch duration for the first latent variable: (a) Substrate feed rate. (b) Dissolved oxygen concentration. (c) Biomass concentration. In blue the PLS model, in red the ffd-JYPLS model and in green the gau-JYPLS model.

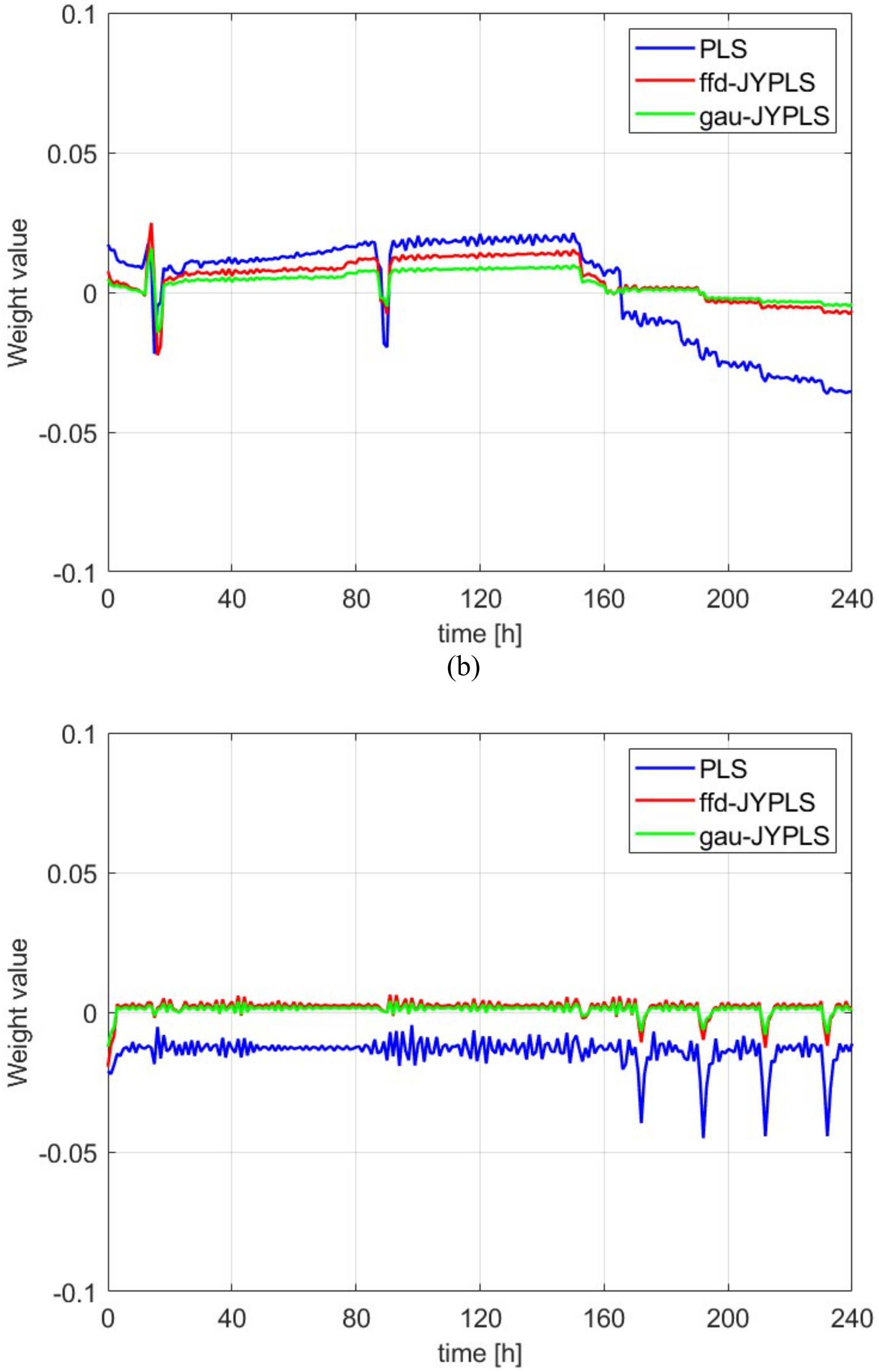
In Figure 4.6 the variables for which the weights of the first latent variable differ between the models are represented. Figure 4.6a represents the substrate feed rate (recipe), and the JYPLS models underestimate the importance of this variable with respect to PLS. For what concerns Figures 4.6b and 4.6c is clear that even the dissolved oxygen concentration (Figure 4.6b) and the biomass concentration (Figure 4.6c) are treated in different ways by the models.

4.2.2 Analysis of the second latent variable weights

In the second latent variable stronger differences emerge even between the two JYPLS models. These can be responsible to the different performances between the gaussian and the full factorial model. The general trend is that the ffd-JYPLS model is like PLS, while the gau-JYPLS appears to more dissimilar. As for the previous Sub-section the major differences are on the variables that the pilot simulator does not measure, as: phenylacetic acid concentration, ammonia concentration, and off-gas. For the variables which are measured in the simulators of both scales the most significant difference are captured in the following figures.



(a)



(b)

(c)

Figure 4.7. Weights values along batch duration for the second latent variable: (a) Substrate concentration. (b) Dissolved oxygen concentration. (c) pH. In blue the PLS model, in red the ffd-JYPLS model and in green the gau-KYPLS model.

In Figure 4.7, as in Figure 4.6, the variables for which the weights of the first latent variable differ between the models are represented. Figure 4.7 confirms the trend that the ffd-JYPLS is like PLS. Furthermore, even if the JYPLS models follow the trends of these weights, they underestimate the importance of all these three variables, this is the cause of the underperforming of the models. In fact, the substrate concentration (Figure 4.7a) and the pH (Figure 4.7c) are strictly related to the design variables. The same difference present in Figure 4.6b is committed in Figure 4.7b, but the magnitude of the difference is bigger.

Chapter 5

Optimizing operating conditions with variable number of batches

In this Chapter the objective is to evaluate which model proposes the best operating condition of the industrial process to maximize penicillin production. To do this an inversion problem (Sub-section 1.2.2) must be solved. In particular, two cases are analyzed: the one in which only industrial data are available and the other with the transfer learning from pilot plant to industrial plant. The sensitivity of the performance is evaluated for a varying number of available data.

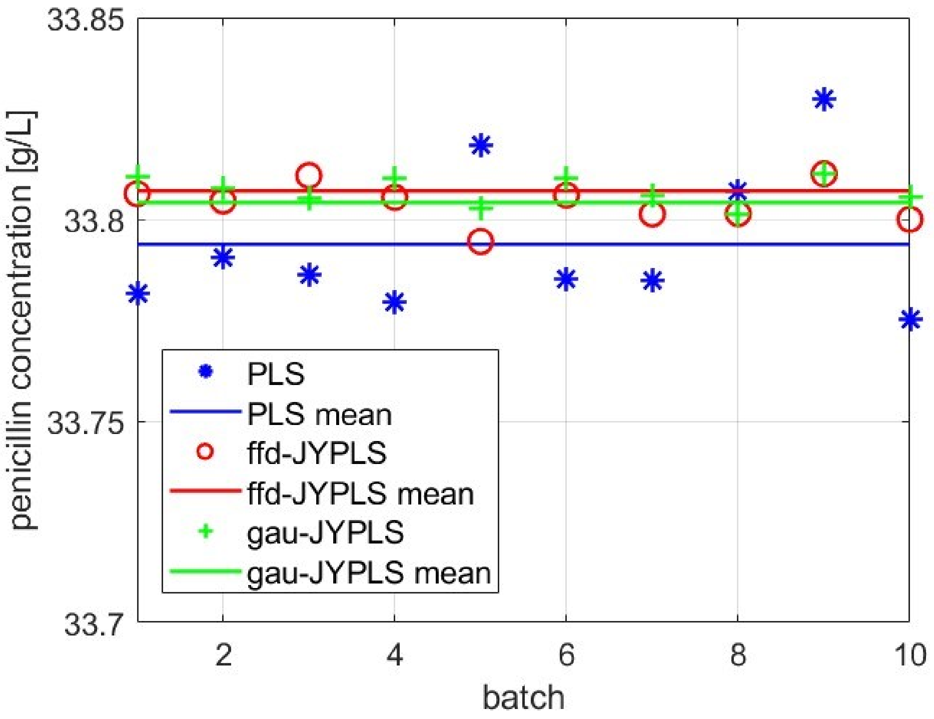
5.1 Product optimization through model inversion with and without transfer learning

To identify the optimal operating conditions, varying the number of batches, 4 different cases are studied: *i)* 8 pilot batches and 8 industrial batches, *ii)* 8 pilot batches and 25 industrial batches, *iii)* 25 pilot batches and 8 industrial batches, *iv)* 25 pilot batches and 25 industrial batches. These 4 cases are selected to show the different combinations of a limited and large number of batches for both the scales. The number of pilot batches are chosen in a way that can consider, in the full factorial design, the 8 vertices of the cube that represent the space of manipulated variables. The number of industrial batches are chosen to be in a condition of similar performance of the models. For all the tests the same industrial batches are used to calibrate the models.

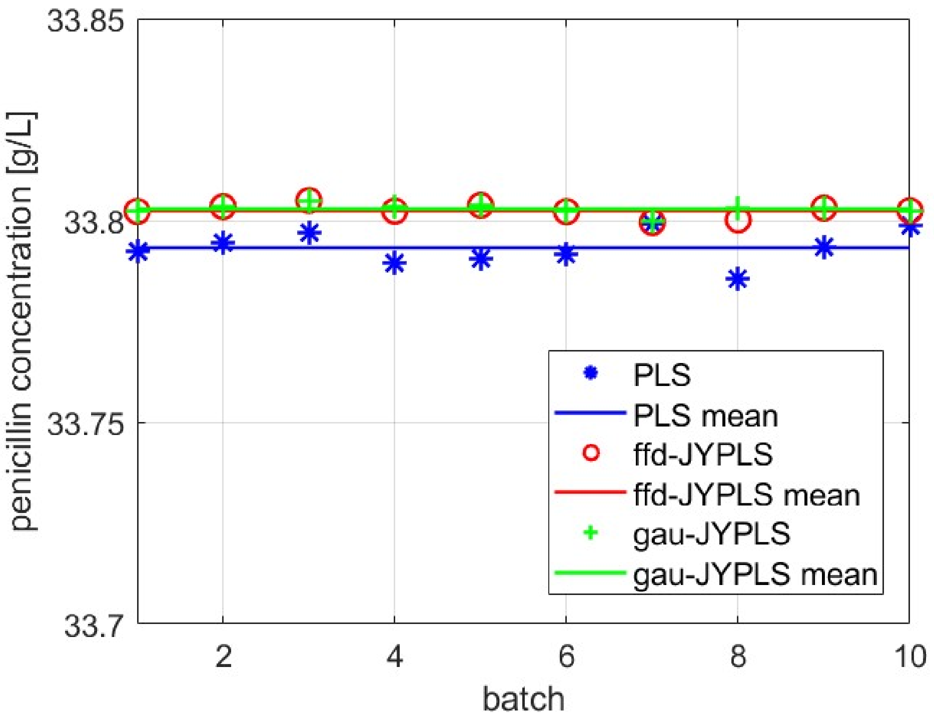
The optimal operating conditions, that correspond to the maximum penicillin produced, are obtained through the inversion of the model. The different models are built as in Section 4.1. Pilot calibration batches are chosen with the Kennard-Stone algorithm, instead the industrial scale calibration batches are chosen randomly. This time the choice of the industrial scale batches is not done 100 times but 10 times. Then, the batch corresponding to the identified optimal conditions are carried out in the industrial process to have an actual indication of the penicillin production. The final penicillin concentration are reported in the figures both as average values and as values of each specific case (i.e. each random selection of industrial calibration batches).

5.1.1 Industrial process without disturbances on process variables

The four cases of Section 5.1 are performed on the industrial process without disturbances (i.e. considering the values of Table 2.3 equal to 0). Excluding the disturbances means assuming a perfect control of the process, which can be an interesting case from a design perspective.



(a)



(b)

Figure 5.1. *End penicillin concentrations using the optimal point proposed by the inversion of the three models at different numbers of industrial calibration batches and pilot calibration batches, without random disturbances: (a) 8 pilot batches and 8 industrial batches. (b) 25 pilot batches and 25 industrial batches. Blue stars represent the penicillin concentration of the PLS model, red circles represent the penicillin concentrations of the ffd-JYPLS, and the green crosses represent the penicillin concentrations of the gau-JYPLS. The bold lines are the means (performed over the 10 iterations) of the three models.*

Figure 5.1 shows the penicillin concentration obtained using the optimal operating conditions determined through the inversion of the models. The results of the case 25 pilot batches and 8 industrial batches are not shown because equal to Figure 5.2a, same for the case 8 pilot batches and 25 industrial batches that gives the same results of Figure 5.2b. This is because the operating conditions proposed by the same model at different pilot batch numbers are very similar to each other, producing very similar results that would not be seen in figures. In all cases the JYPLS outperforms the PLS, providing operating conditions that allow a higher concentration of produced penicillin. Furthermore, the ffd-JYPLS works better than the gau-JYPLS. The t-Student test results say that the two JYPLS models are not statistically different. For 25 industrial batches both the JYPLS models are statistically different from the PLS model, instead, with 8 industrial calibration batches, only the gau-JYPLS is statistically different from the PLS model. In this last case the p-value of the t-Student test is investigated, the p-value of the models PLS and ffd-JYPLS is equal to 0.10 so they are statistically different if a 90% confidence limit is used. In Table 5.1 the mean penicillin concentration achieved in each case using the optimal process conditions is presented.

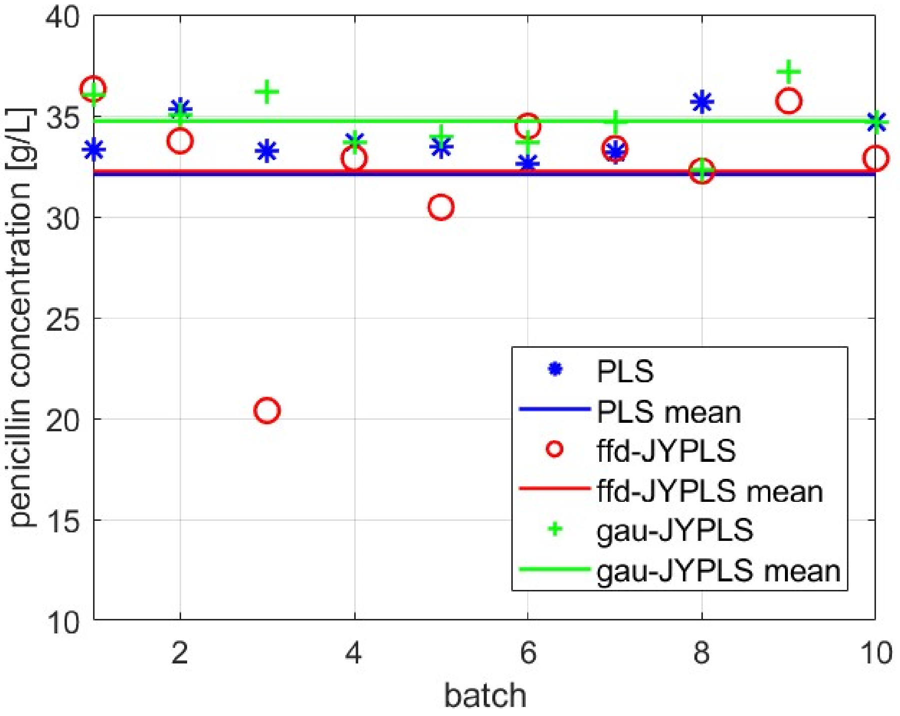
Table 5.2. *Mean values of the inversion for the three models in different combinations of industrial and pilot batches.*

n° of industrial batches	n° of pilot batches	PLS mean (g/L)	ffd-JYPLS mean (g/L)	gau-JYPLS mean (g/L)
8	8	33.7938	33.8071	33.8041
25	8	33.7932	33.8024	33.8028
8	25	33.7938	33.8071	33.8041
25	25	33.7932	33.8024	33.8028

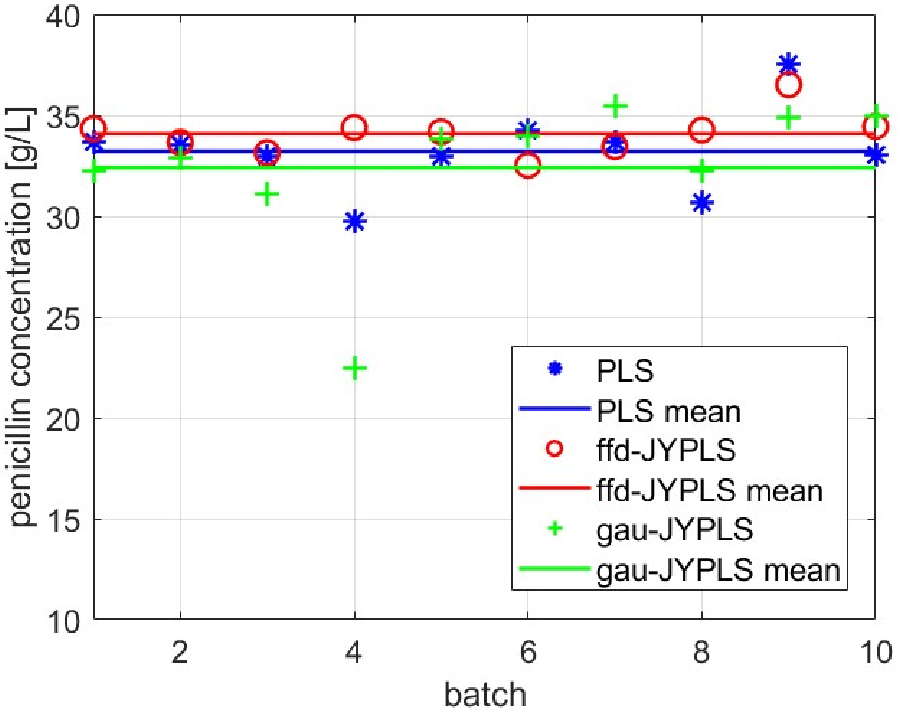
As expected, the PLS mean changes only with the number of industrial batches. Even the mean penicillin concentration achieved through the two JYPLS models change only with the number of industrial batches, meaning that an increase from 8 to 25 pilot batches does not affect the information added for the identification of the optimal operating conditions. The operating conditions proposed by the ffd-JYPLS improve the penicillin production of 0.04% (1.33Kg per batch) and 0.03% (0.92Kg per batch) with 8 and 25 calibration industrial batches, respectively. Clearly the improvement will decrease by increasing the number of industrial batches used for calibration. Accordingly, the JYPLS model ensures improved productivity with respect to the PLS one.

5.1.2 Industrial process with random disturbances in process variables

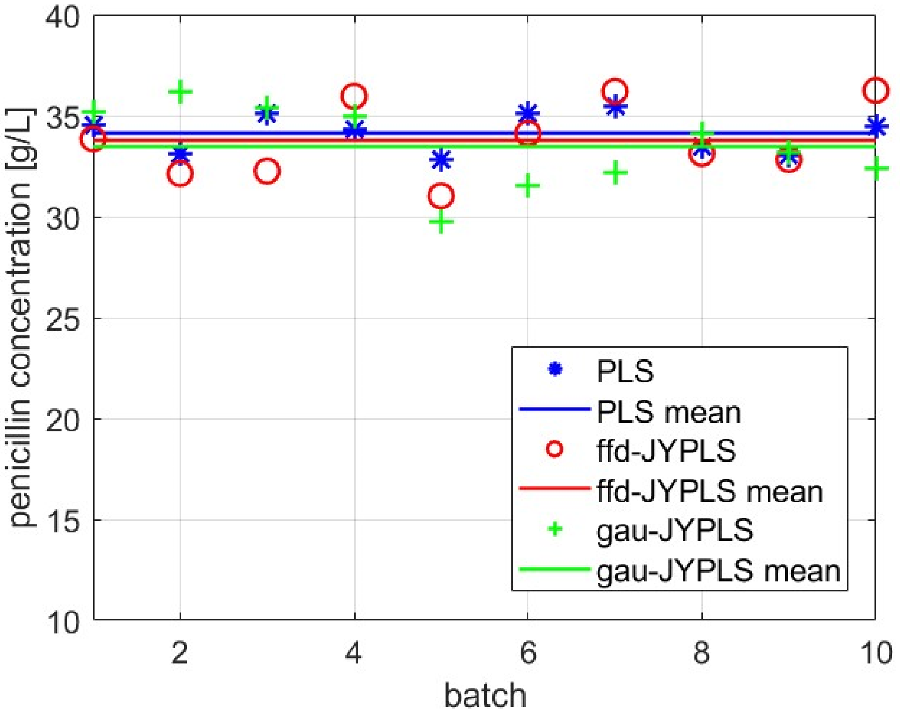
The results of the 4 cases, obtained with the disturbances added process disturbances (i.e. considering the values of Table 2.3) are shown in the following figures.



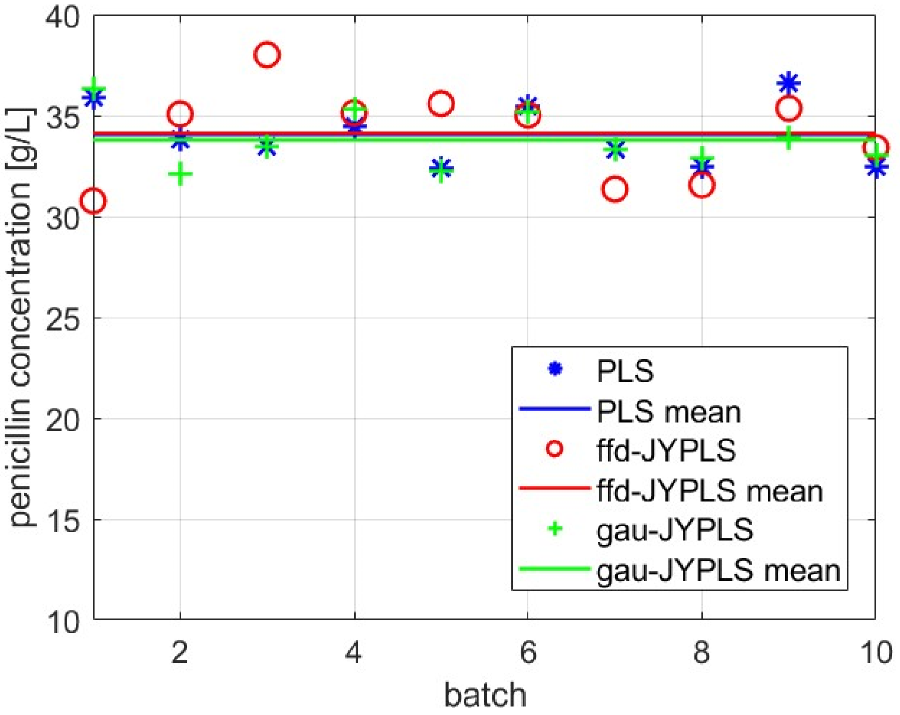
(a)


40
35
30
25
20
15
10
penicillin concentration [g/L]
2
4
6
8
10
batch
PLS
PLS mean
ffd-JYPLS
ffd-JYPLS mean
gau-JYPLS
gau-JYPLS mean

(b)


40
35
30
25
20
15
10
penicillin concentration [g/L]
2
4
6
8
10
batch
PLS
PLS mean
ffd-JYPLS
ffd-JYPLS mean
gau-JYPLS
gau-JYPLS mean

(c)



(d)

Figure 5.2. *End penicillin concentrations using the optimal point proposed by the inversion of the three models at different numbers of industrial calibration batches and pilot calibration batches, with random disturbances added: (a) 8 pilot batches and 8 industrial batches. (b) 8 pilot batches and 25 industrial batches. (c) 25 pilot batches and 8 industrial batches. (d) 25 pilot batches and 25 industrial batches. Blue stars represent the penicillin concentration of the PLS model, red circles represent the penicillin concentrations of the ffd-JYPLS, and the green crosses represent the penicillin concentrations of the gau-JYPLS. The bold lines are the means (performed over the 10 iterations) of the three models.*

Figure 5.2 shows, similarly to Figure 5.1, the penicillin concentration obtained using the optimal operating conditions determined through the inversion of the models. There is not a model that works better than the other, sometimes the JYPLS is the best choice, sometimes it is the PLS. This is due to the high random disturbances present in the process variables in this case. Performing a t-Student test, the penicillin concentration average values obtained in this case (i.e. adding the random disturbances in the models) are not statistically different, so it is not certain that the JYPLS model outperforms the PLS one.

Conclusions

This Thesis demonstrated that transfer learning in predictive models for scale up is not only possible, but also beneficial thanks to the implementation of JYPLS. This technique is 5% more accurate and 5% more precise with respect to a model built by the PLS, when the number of available data from pilot-scale batches and industrial-scale batches is low. This is a significant outcome, because transfer learning not only yields superior predictive accuracy, but also results in a model that is less susceptible to the industrial batches employed in calibration. The optimization of industrial operating conditions has been demonstrated to result in a 1 kg increase in penicillin production per batch, which represents a significant industrial outcome given that typical production plants utilize multiple chemostats, with each reactor producing over 30 batches per year. A further noteworthy aspect of these findings is that they can be readily implemented at no additional cost, assuming that an experimental pilot-scale campaign has already been conducted. The transfer of information has been demonstrated to enhance the performance of the PLS, particularly when the number of pilot-scale and industrial-scale batches is limited to 13 and 25, respectively, for the JYPLS. Another point supporting this thesis is that the results are obtained from simulators which describe the fermentation process in different ways; using similar simulators data could facilitate a greater transfer of information and more effective results. Since these results are obtained when the process is well controlled, in a simulated environment, and no random disturbances affect the process variables, when these disturbances cannot be neglected the advantage of using transfer learning techniques is not proven. About the experimental campaign in the pilot-scale plant, the results indicate that there is no difference between the case in which happenstance data are available or a full-factorial designed experimentation has been performed. The limitations of this work are related to the fact that the data were generated from simulated processes. For this motivation, future work will be oriented to utilize real plant data.

References

Bailey, J. E. & Ollis, D. F. (1986). Biochemical Engineering Fundamentals. New York: McGraw Hill.

Bajpai, R. & Reuss, M. (1980). A mechanistic model for penicillin production. Journal of Chemical Technology and Biotechnology 30, 330-344.

Barberi G., Benedetti, A., Diaz-Fernandez, P., Finka, G., Bezzo, F., Barolo, M. & Facco, P. (2021). Anticipated cell lines selection in bioprocess scale-up through machine learning on metabolomics dynamics. IFAC-PapersOnLine, Volume 54, Issue 3, 2021, pages 85-90.

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, C. D., Vappiani, J., P., Finka, G., Bezzo, F., Barolo, M. & Facco, P. (2022). Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development

Birol, G., Ündey, C. & Çinar, A. (2002). A modular simulation package for fed-batch fermentation: penicillin production. Computers and Chemical Engineering 26 (2002) 1553-1565.

Birol, G., Ündey, C., Parulekar, S. J., & Çinar, A. (2001). A morphologically structured model for penicillin production. Biotechnology and bioengineering, Vol. 77, No. 5, Marche 5, 2002.

Botton, A., Barberi, G. & Facco, P., (2022). Data Augmentation to Support Biopharmaceutical Process Development through Digital Models—A Proof of Concept. Processes 2022, 10, 1796.

Bro, R. (1996). Multiway calibration. Multilinear PLS. Journal of Chemometrics, January/February (1996), Volume 10, Issue 1, 47-61.

Chu, F., Cheng, X., Jia, R., Wang, F. & Lei, M. (2018). Final quality prediction method for new batch processes based on improved JYKPLS process transfer model. Chemometrics and Intelligent Laboratory Systems, 183 (2018), 1-10.

Facco, P., Zomer, S., Rowland-Jones, R. C., Marsh, D., Diaz-Fernandez, P., Finka, G., Bezzo, F., & Barolo, M. (2020). Using data analytics to accelerate biopharmaceutical process scale-up. Biochemical engineering journal, 164 (2020), 107791.

García-Muñoz (2004). Doctor of Philosophy Thesis. McMaster University, September (2004).

García-Muñoz, S., MacGregor, J. F., Kourti, T. (2005). Product transfer between sites using Joint-Y PLS. Chemom. Intell. Lab. Syst. 2005, 79, 101.

Geladi, P. & Kowalski, B. (1986). Partial least squares: a tutorial. Anal. Chim. Acta 185, 1.

Goldrick, S., Ştefan, A., Lovett, D., Montague, G. & Lennox, B. (2014). The development of an industrial-scale fed-batch fermentation simulation. Journal of Biotechnology 193 (2015) 70-82.

Kennard, R.W. & Stone, L.A. (1969). Computer aided design of experiments. Technometrics 11, 137-148.

Kheirolomoom, A., Kazemi-Vaysari, A., Ardjmand, M. & Baradar-Khoshfetrat, A. (1999). The combined effects of pH and temperature on penicillin G decomposition and its stability modeling. Process Biochem. 35, 205–211.

Montague, G.A., Morris, A.J., Wright, A.R., Aynsley, M. & Ward, A. (1986). Modelling and adaptive control of fed-batch penicillin fermentation. Can. J. Chem. Eng. 64, 567–580.

Mou, D. & Cooney, C. (1983). Modeling and adaptive control of fedbatch penicillin production. Biotechnology and Bioengineering, 25, 225/255.

Nielsen, J.H.i., Villadsen, J., Lidén, G. (2003). Bioreaction Engineering Principles. Kluwer Academic/Plenum Publishers.

Nielsen, J. & Villadsen, J. (1994). Bioreaction Engineering Principles. New York: Plenum Press.

Paul, G. C. & Thomas, C. R. (1996). A structured model for hyphal differentiation and penicillin production using Penicillium chrysogenum. Biotechnol. Bioeng.. 51, 558–572.

Shuler, M. & Kargi, F. (2002). Bioprocess Engineering Basic Concepts (2nd ed., Saddle River, NJ: Prentice Hall).

Tomba, E., Barolo, M. & García-Muñoz, S. (2012). Ind. Eng. Chem. Res., 51, 12886−12900.

Vogel, H. & Todaro, C. (1997). Fermentation and Biochemical Engineering Handbook. Noyes.

Wold, H. (1975). Path Models with Latent Variables: the NIPALS Approach. Quantitative sociology, international perspective on Mathematical and Statistical modelling, 307-357.

Wold, S., Sjöström, M. & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, Volume 58, Issue 2, 28 October 2001, Pages 109-130.