# Value chain planning optimization: a data driven digital twin approach

**F. Ferranti** * **F. Manenti** * **G. Vingerhoets** ** **M. Vallerio** **

* *Dipartimento di Ingegneria Chimica e Materiali, Politecnico di Milano, Milan, Italy*
** *BASF 4.0 - Data Analytics, BASF, Antwerpen, Belgium*

**Abstract:** The long term production planning for a large chemical production site, where 10+ different chemical plants share raw materials, infrastructure (e.g., tank farm, filling stations) and utilities (e.g. steam, electricity, technical gasses) might prove to be a challenging task. This paper introduces a data driven approach to build a digital twin of a chemical production site to aid the relevant decision makers in defining and evaluating the economic impact of a long term (i.e. several months ahead) production planning. Each chemical plant and energy production unit on site is represented by simple regression models relating the consumption of raw materials and utilities to its products. The resulting system of algebraic equations has been inserted in an optimization environment with the objective of maximizing the profit. In the optimization, also the electricity and steam generation were introduced to obtain a global energy balance of the production site. This combination resulted in a multi period Mixed-Integer Linear Programming (MILP) problem. The effect of electricity price and external temperature on the optimization results are also investigated.

*Keywords:* Value chain planning; Statistical models; MILP optimization; Digital twin

## 1. INTRODUCTION

In the last decades, the chemical industry has faced many challenges: the increase in competitiveness and the research for sustainability pushed many companies to further optimize production and reduce emissions. One way to reach this goal, for instance, is sharing facilities and creating interconnected processes. Nevertheless, this aggregation also increases the complexity of the production sites and more effort and attention are required to schedule and plan production.

This necessity arises mainly in large chemical production sites, like the BASF site in Antwerp, Belgium. This production site hosts 50 different chemical plants, owned by 4 different companies. All of the plants share common electricity, steam, and all other utilities nets as well as a wastewater treatment plant and all other logistic infrastructure. This paper focuses on the possibility of optimizing through a data driven digital twin two of the production value chains present on site.

A chemical value chain can be defined as a group of chemical plants interconnected between them, where the finished product of one plant is the raw material used by the next one. In our specific case, a value chain is a series of chemical processes that transforms a raw material into a valuable finished good. The focus is posed on stationary operations and long-term predictions, excluding dynamic and non-operating periods. In section 2 a brief state of the art for the used statistical model and the MILP problem is presented. Section 3 describes the investigated case study. Section 4 is dedicated to the description of the MILP problem and section 5 presents the obtained results. Finally, section 6 discusses this work conclusions.

## 2. STATE OF THE ART

### *2.1 Statistical models*

In this work only single and multiple regression models have been considered. The main reason is to evaluate the accuracy achieved with the most simple models. Moreover, a lower model complexity will result in an easier solution to maintain. The regression models chosen to find the correlations between finished good and consumption of raw materials and utilities can be represented in the form:

$$y = f(X, \beta) + \epsilon \tag{1}$$

where:

- y is the vector of the response variable.
- f is any function of the predictors matrix X and the unknown parameters vector $\beta$.
- $\epsilon$ is an n-by-1 vector of independent, identically distributed random disturbances.

To obtain these equations the software JMP SAS has been used. The model parameters are estimated through a least square minimization of the form:

$$\sum_{i=1}^{n} (y_i - f_\theta(x_i))^2$$

(Elster et al. (2015)). Where y represents the response variable, while x is the predictor vector. The quality of the obtained correlations can be statistically assessed based on different parameters, in order of importance,

residuals' distribution $\gamma$, RMSE (root mean square error) and $R^2$. The residuals, defined as the difference between the real values and the predicted ones, should be normally distributed. The RMSE is a measure of the average model error and it can be expressed as:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} \tag{2}$$

while $R^2$ indicates the fraction of the variance of the dependent variable that can be explained by the independent variables in the model and it has been obtained through the formulas:

$$SS_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$R^2 = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

### *2.2 Optimization*

There are many articles in literature that address the topic of value chain optimization for the chemical industry. Most of them use continuous variables for process parameters, and integer or binary values to indicate whether a plant is operating or not. Therefore, the mixed-integer programming (MIP) is the most used optimization method. The MIP problem can be divided into two categories: mixed-integer linear program (MILP) and mixed-integer non-linear program (MINLP). A MILP is a problem with:

- linear objective function $F(x)$, where x is the vector of unknown variables;
- bounds and linear constraints;
- restrictions on some elements of x to have integer values.

In mathematical terms it can be defined as:

$$min\ F(x) \text{ subject to } \begin{cases} x(intcon) \text{ are integers} \\ A * x \leq b \\ Aeq * x = beq \\ lb \leq x \leq ub \end{cases} \tag{3}$$

where intcon represents the indexes of the integer variables, while lb and ub the lower and upper bounds. A MINLP presents the same mathematical structure, but allows also non-linear equality and inequality constraints. Examples of MILP optimizations of utilities and a biofuel supply chain can be found respectively in Velasco-Garcia et al. (2011) and Awudu and Zhang (2013). In this article, a MILP optimization structure has also been performed, since all resulting models are linear or have been approximated to linear ones.

## 3. CASE STUDY

The case study represents two of the value chains present at the BASF Antwerp site. A schematic representation can be seen in Fig. 1, where:

- all plants used in the transformation of Product 1 are represented in yellow;
- all plants used in the transformation of Product 2 are coloured in green;
- Plant A and C process the raw materials to obtain Product 1 and 2.

The plants that process Product 1 belong to the first value chain, while the plants that transform Product 2 belong to the second value chain. Due to reasons of confidentiality, all names of chemicals and plants have been changed and all numerical values have been normalized. Most of the plants considered, from A to L, operate in continuous mode. The raw materials entering into the value chains are number from 1 to 6, while the final products and the intermediates are called "Product X", where X is a number between 1 and 23. For the continuous plants, the finished goods are represented in the figure, while for the batch plants they are omitted. Tanks are present between every plants and these have been numbered from 1 to 24 in Fig. 1. They have been introduced in the digital twin to visualize situations of scarcity or overproduction of a particular product. The last part of the digital twin consists of the steam net and the steam generation units. For the sake of simplicity these are not reported in Fig. 1.

### *3.1 Data*

The data considered to build the statistical models for this work goes from 1/11/2017 to 19/12/2019. The collected observations consist of hourly average of the measurements made by the flow meters on the site. Chemicals have been divided into three groups:

- raw materials, these are all the reagents of the plant under examination;
- utilities, considered as all the materials and energy flows that are used for different purposes, like cooling, heating or inertization of the reactors;
- final products, the chemicals obtained as a result of the chemical transformation in a plant.

### *3.2 Data cleaning*

The focus of modeling is to have a prediction of the consumption of raw materials and utilities while the plant is in stationary operating conditions. This implies that cleaning the data is fundamental to obtain good correlations. The adopted cleaning strategy consisted of the following steps:

- first of all, all the data intervals related to turnarounds and shut-downs have been excluded. The method used consisted in selecting a minimum hourly amount of raw materials flowing in the plant as a discriminant between producing and non-producing time intervals.
- then, the dynamic parts of the production periods have been excluded. This task has been accomplished by looking at the production stability and its variation from hour to hour. The objective consisted in selecting only the time periods when there are no big changes in production values. To do that, the following formula was used:

$$Mat(i) = Material\ produced\ at\ time\ i \tag{4}$$

$$Stability\ coefficient(i) = \\ = Standard\ deviation(Mat(i), ..., Mat(i-12)) \tag{5}$$

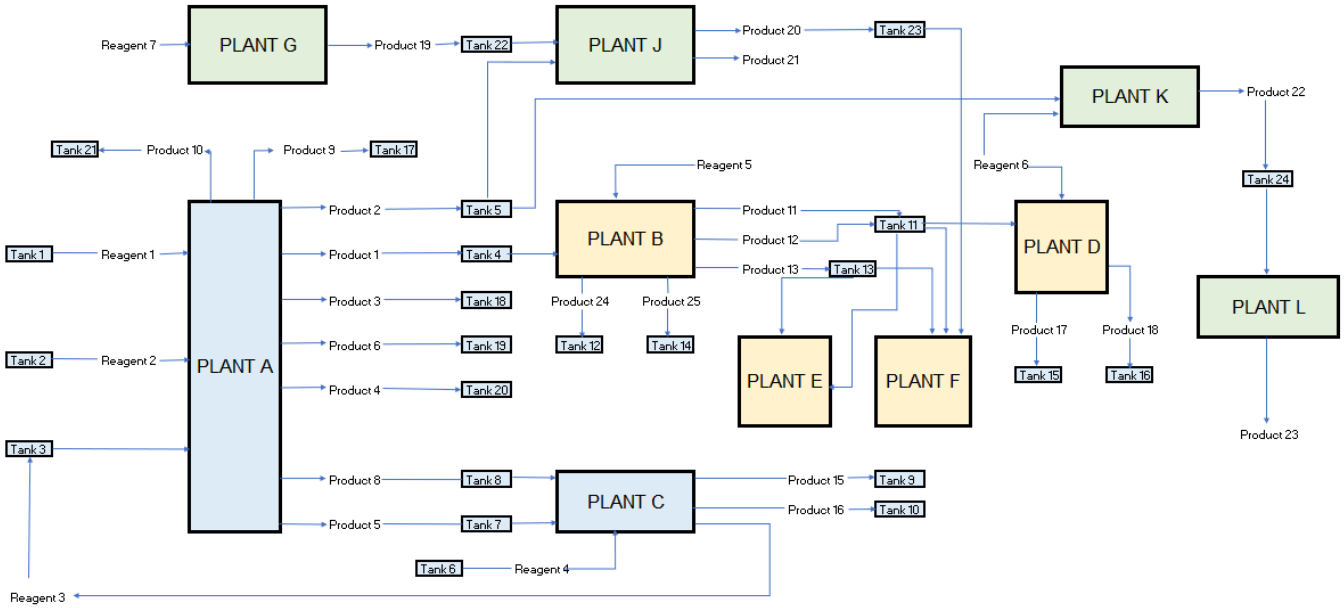


Fig. 1. High-level representation of the value chains

The standard deviation of the considered observation value and the 12 preceding ones have been chosen to select the stable operations. Therefore, it is possible to exclude the non-stationary modes in a systematic way by selecting an opportune limit value for this stability coefficient. The stability coefficient limit is selected on the basis of the sensor accuracy. In the intervals were the stability coefficient value is lower than the limit, the production is defined as stationary. In most of the plants, more than 80% of the raw data satisfied the criteria mentioned above and it was retained for analysis.

### *3.3 Statistical models*

The software JMP has been used to build the models. The objective is to predict the amount of raw materials and utilities needed to obtain the planned production volume. In the representation of plant B in Fig. 1 the raw materials we want to predict are the inlet flows consisting of Product 1 and Reagent 5. The utilities that has to be predicted in the specific case of plant B are the steam at 16 bar and the demineralized water entering the plant. The input for the regressions models are the finished goods produced by every single plant. The only additional parameter that has been used is the external temperature, since this affects utilities consumption for some plants. In Fig. 1 the finished goods used as input are Product 11, 12, 13, 24, and 25. At the end, it should be possible to predict raw materials and utility consumption per ton of produced finished good using only single and multiple regressions. One example of multiple linear regression is the correlation between reagent 1 in reactor A and the main products of that plant. In Fig. 2 it is possible to notice the points used to train the model, in black, and the excluded point corresponding to failures and a major turnaround, in red. Therefore, after having cleaned the data as described in the previous chapter the obtained prediction expression is:

$$R1 = -0.0965 + 1.235 * P3 - 0.09688 * P4 + 0.1103 * P5 \quad (6)$$

Where R1 is reagent 1 and P3, P4 and P5 are respectively product 3, 4 and 5. The regression algorithm was able to find a relation between the reagent and the products as it can be seen both from the statistical parameters and Fig. 3. From a statistical point of view, it results:

$$R^2 = 0.813 \quad (7)$$

$$RMSE\% = 6.24\% \quad (8)$$

The same approach has been used for all the other raw

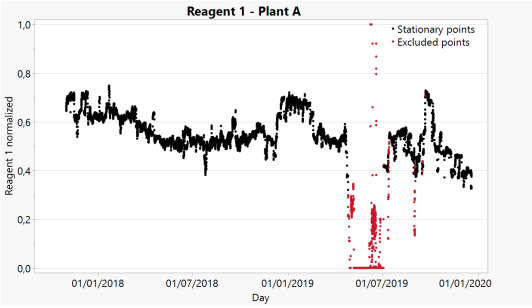


Fig. 2. Consumption of reagent 1 in the training period

materials, always trying to keep the models as simple as possible. This avoids over fitting and increases model robustness and maintainability.

The second step involves the modeling of the utilities used by the different plants. The utilities considered for the scope of this work are:

- steam, at multiple pressure levels, depending on the needed temperature and the site infrastructure constraints;
- demineralized water, used to produce steam, to clean the reactors and as a coolant;
- hydrogen, that is mainly used as a reagent;
- nitrogen, used for inertization reasons;
- electricity, needed for the mechanical equipment such as pumps and compressors.

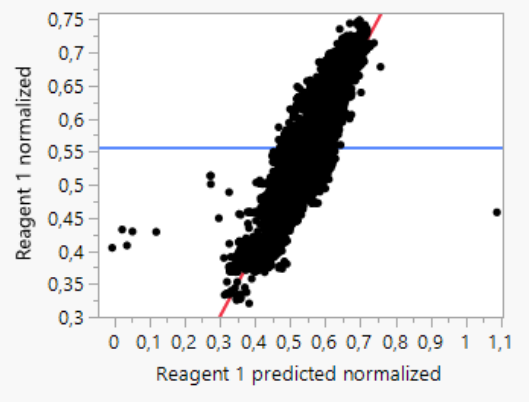


Fig. 3. Correlation between reagent 1 and the finished products present in the equation 6

The approach for modeling the utilities was the same applied to the reagents. Particular care was placed in data cleaning to exclude the peaks due to production stops and equipment failures. In most of the cases the parameter $R^2$ has shown lower values compared to the reagents. This is mainly due to the fact that a less strict correlation exists between energy and production.

### *3.4 Tank farm*

Tanks have been introduced to calculate the amount of materials stored on site. A simple material balance has been used:

$$Tank\ level = \sum inflow - \sum outflow + previous\ level \tag{9}$$

## 4. OPTIMIZATION STRUCTURE

The aim is to optimize the daily production to maximize the profit over the selected period. The optimizer takes into account the energy production cost as well as the inventory cost. Different plants on site produce energy through their exothermic reactions, while others are endothermic processes and need energy to convert reagents into products. The steam net does not allow for buffering and it needs to be kept in balance at all times while its production needs to be optimized by selecting the most cost effective way to produce it. A penalty has been introduce to keep inventories as low as possible. Most of the complexity comes from the selection of the energy production mix. Two different sources are available on site:

- 5 steam boilers, opeerating at different pressure levels;
- a Combined Cycle Gas Turbine (CCGT).

Lastly, there is the possibility to buy and sell the electricity directly on the market.

### *4.1 Constraints and objective function*

The optimization has been performed in MATLAB through a multi period Mixed-Integer Linear Programming (MILP). The selected time interval spans 49 days and the considered variables are:

- Raw materials used in the reactions;
- Steam consumed by the plants;
- Steam produced by reactor A, boilers and CCGT;
- Electricity generated by the CCGT, bought and sold in the market;
- Tanks, to store the raw materials before their usage.

The purpose of the optimizer is to maximize profit. In this work several terms are summed into a unique objective function. However, a multi-objective approach might be considered to better understand the possible trade-offs existing between minimizing energy consumption or emission while still maximizing production (Vallerio et al. (2015); Nimmegeers et al. (2019)). Every variable has boundaries based on the physical limitations of the respective unit. Moreover, the production is constrained to meet the planned amount at the end of the optimization period. The main source of steam is the combustion of natural gas in the boilers. They can produce steam at 16 bar pressure with a conversion:

$$P_{steam} = 0.01435/0.011 * NG \tag{10}$$

Where $P_{steam}$ is the amount of steam produced (Ton/day) and NG is the amount of natural gas ($Nm^3$/day). The CCGT is able to produce electricity as well as steam and it can operate at 5 different operating modes between 2820 and 4560 MWh. The conversion equation is:

$$Ep = NG * 0.60 \tag{11}$$

Where Ep is the electricity produced (MW) by the CCGT. Additionally, it is also possible to buy and sell electricity from the energy market. The prices for buying and selling electricity have been taken from the historical data of day-ahead Belgium market (see Fig. 4).

Finally for each ton of steam produced by the CCGT, the electricity production is reduced by 0.25 MW. All this information is used to build the global objective function that is shown here below:

$$\begin{aligned}
&\text{For i = 1:n days} \\
&CEb(i) = Eb(i) * PEb(i) \\
&CEs(i) = Es(i) * PES(i) \\
&C_CCGT(i) = 2500 * Switch(i) \\
&+ Ep(i)/0.60 * PNG \\
&+ 250 * Ep(i)/380 \\
&CS = \sum_{n=1}^{n\ days} (NG(n) * PNG) \\
&CI = \sum_{m=1}^{n\ tank} ( \sum_{n=1}^{n\ days} (tank\ level(m,n)) \\
&CE = \sum_{n=1}^{n\ days} (Price\ CCGT(n)) \\
&+ \sum_{n=1}^{n\ days} (CEb(n) - CEs(n)) \\
&CRM = \sum_{m=1}^{n\ RM} ( \sum_{n=1}^{n\ days} (RM(m,n) * PRM(m)) \\
&Revenue = \sum_{m=1}^{n\ prod} ( \sum_{n=1}^{n\ days} (P(m,n) * PP(m)) \\
&Profit = Revenue - CRM - CE - CI - CS
\end{aligned} \tag{12}$$

Where CEb, CEs represent the cost, Eb and Es, the amount, PEb and PEs the prices for electricity bought and sold. C_CCGT is the cost related to the operation of the CCGT. CE is the total cost for electricity. CS is the cost associated with the steam production. PNG is the price for natural gas. CI is the cost related to inventory. CRM is the cost related to the consumption of raw material, RM and PRM are the amount and the price of the respective raw material. P and PP are the amount and price of the specific product.The optimization was solved in MATLAB through the algorithm "intlinprog", more details on the Branch and Bound procedure can be found in Savelsbergh (1994) and Wolsey (1998).
Four different optimization periods have been considered. These were selected to estimate the effect of electricity prices and external temperature on the obtained solution (see Fig. 4). In particular, the four periods are reported in Table 1. Every performed optimization has been repeated

| | time | electricity price | external temperature |
|---|---|---|---|
| Period 1 | Jan/18 | middle | low |
| Period 2 | Jun/18 | middle | high |
| Period 3 | Oct/18 | high | middle |
| Period 4 | Feb/20 | low | middle |

Table 1. Different investigated periods

twice by using two different strategies:

- Global scheme: the optimization spans over the whole period of 49 days, with a final total production constraint;
- Weekly scheme: the optimization is performed per week with weekly production constraint.

The weekly scheme requires a reduced amount of computational time. However, it also presents a reduced number of degrees of freedom making it more difficult for the optimizer to adapt to unexpected production losses.

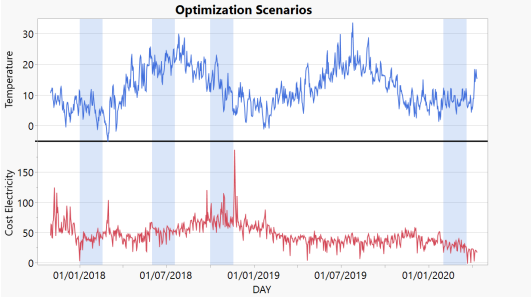


Fig. 4. Electricity price and external temperature trend from 1/11/2017 to 8/04/2020. in light blue the investigated periods 1 to 4.

## 5. RESULTS

### *5.1 Global and weekly optimization*

As we can see in Fig. 5 for Period 1 the internal production of electricity is preferred almost every day, due to the high cost of external energy in that period. This does not happen in Period 4 (Fig. 6), when it was more convenient to buy electricity from the market than producing it internally. In this case, the amount of steam produced by the CCGT is significantly smaller than the amount obtained by the boilers. The boilers production for all the intervals can be seen in Fig. 7. The minimum produced amount is 480 tons/day. This is the minimum value to keep the boilers in function and ready to provide the necessary heat in extraordinary situations. The differences between the different periods mostly depend on the electricity price. As it can be seen in Figure 7, the boilers are mostly used in period 3 and 4. However, the reason behind it is quite different. In period 4 this is due to low electricity price that make inefficient to activate the CCGT, While in period 3 the electricty price is so high that the CCGT is used to produce electricity and sell it on the market. For periods 1 and 2, the boilers are at their minimum capacity and all the remaining needed energy, in the forms of electricity and steam, is provided by the CCGT. Table 2 reports the results of the optimization for all the scenarios. The optimization is able to meet the predefined production planning in all investigated periods. The difference in profit between the different periods is mainly due to the energy cost. In particular, a trend can be spotted. In fact, the higher the electricity price, the higher the global profit. The difference between the global and weekly optimization scheme comes from the higher degrees of freedom available in the global scheme.

| Period | global optimization | weekly optimization |
|---|---|---|
| January 2018 | 191,00 | 190,96 |
| June 2018 | 191,61 | 191,59 |
| October 2018 | 193,61 | 193,58 |
| February 2020 | 191,10 | 191,12 |

Table 2. Profit (€ Mio) result for every interval of time and every optimization strategy

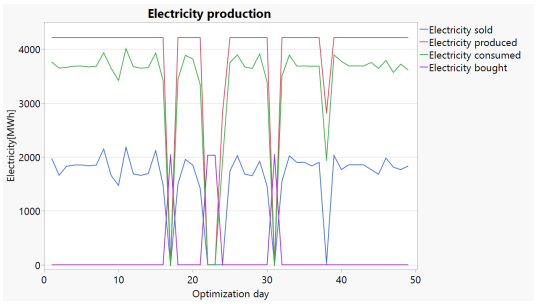


Fig. 5. Electricity produced and sold in Period 1. The red line represents the total amount of energy produced from the CCGT, the green one the part not transformed into steam. Blue and violet line are, respectively, the electricity sold and bought on the market

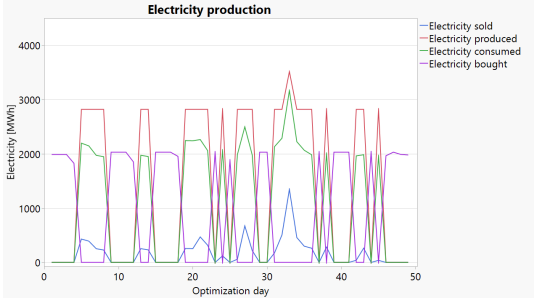


Fig. 6. Electricity produced and sold in Period 4.

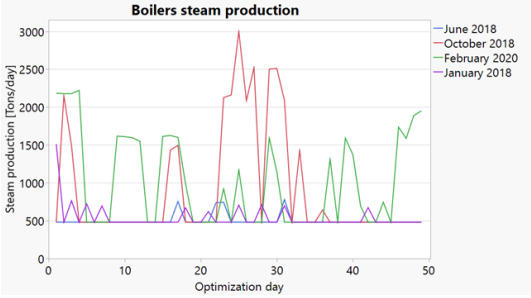


Fig. 7. Steam produced by the boilers in the different periods

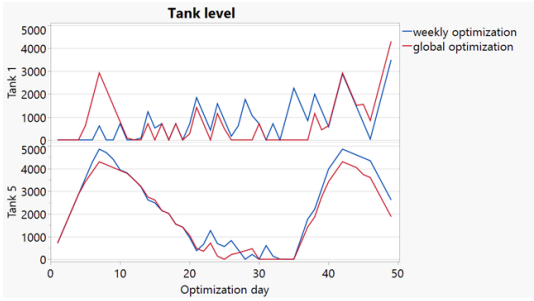


Fig. 8. Tank 1 and 5 level for June 2018 optimization.

### *5.2 Optimization with planned and unplanned shutdowns*

The impact of a planned and an unplanned shutdown is discussed in this section. Every period has been reevaluated twice. The difference between the results obtained with the global and weekly scheme are more significant in this situation. The shutdowns have been simulated by switching off plant K in Fig. 1. Two shutdowns have been imposed within the same period:

- in the first week, to see the effects of an unexpected failure. The lost production should be recovered in the remaining time;
- in the fifth week, to show the effect of a planned shutdown.

Every period has been optimized twice: once with the global scheme and once with the weekly one. The available tank volume plays an important role in both. In particular, it can be noticed that bigger tanks, 14% volume increase, are needed if the optimization is performed with the weekly scheme. This is again due to the reduced numbers of degrees of freedom available to the optimizer. This limits the possibilities to compensate for the lost production volumes. Fig. 8 shows the level of the tanks for Period 2. The tanks chosen are the ones corresponding to tank 1 and 5 in Fig. 1. This figure confirms the necessity of smaller tanks in the global optimization. In table 3 it is possible to see the outcome of every performed optimization.

| Period | global optimization | weekly optimization |
|---|---|---|
| January 2018 | 163,85 | 162,36 |
| June 2018 | 164,62 | 163,12 |
| October 2018 | 166,61 | 165,09 |
| February 2020 | 163,88 | 162,47 |

Table 3. Profit (€ Mio) result for every interval of time and every optimization strategy in the case of a plant shutdown

## 6. CONCLUSION

The objective of this work was to investigate the possibility to build a data-driven digital twin of a chemical production site based on simple regression model and use this to optimize the profit margin for long term production planning. The use of a Mixed Linear Integer Programming algorithm in many different situations has given the expected results. Both in normal operations and in presence of shutdowns it was possible to match the energy production with its consumption in the various processes. Finally, the difference between different lengths of the optimization interval has been tested. Shortening the time period is advantageous to reduce the computational time. However, this shortening can push the optimizer to find a lower profit value due to a reduction in available degrees of freedom. In the future, a continuous attention will be necessary to keep the models adherent to the modelled production processes. To ensure this and avoid potential model drifts and/or bias an automated control system, such as a Shewart's control chart should be implemented. A presentation of this method with a practical example can be found in Hossain and Masud (2016).

## REFERENCES

Awudu, I. and Zhang, J. (2013). Stochastic production planning for a biofuel supply chain under demand and price uncertainties. *Applied Energy*, 103, 189 – 196.

Elster, C., Klauenberg, K., Walzel, M., Wübbeler, G., Harris, P., Cox, M., Matthews, C., Smith, I., Wright, L., Allard, A., et al. (2015). A guide to bayesian inference for regression problems. *deliverable of EMRP Project NEW04 "Novel Mathematical and Statistical Approaches to Uncertainty Evaluation,"*.

Hossain, M.B. and Masud, M. (2016). Performance of t-square chart over x-bar chart for monitoring the process mean: A simulation study. *Journal of Mathematics and Statistical Science*, 2.

Nimmegeers, P., Vallerio, M., Telen, D., Van Impe, J., and Logist, F. (2019). Interactive multi-objective dynamic optimization of bioreactors under parametric uncertainty. *Chemie Ingenieur Technik*.

Savelsbergh, M.W. (1994). Preprocessing and probing techniques for mixed integer programming problems. *ORSA Journal on Computing*, 6(4), 445–454.

Vallerio, M., Vercammen, D., Van Impe, J., and Logist, F. (2015). Interactive nbi and (e)nnc methods for the progressive exploration of the criteria space in multi-objective optimization and optimal control. *Computers & Chemical Engineering*, 82, 186–201.

Velasco-Garcia, P., Varbanov, P.S., Arellano-Garcia, H., and Wozny, G. (2011). Utility systems operation: Optimisation-based decision making. *Applied Thermal Engineering*, 31(16), 3196 – 3205.

Wolsey, L.A. (1998). *Integer Programming*. Wiley-Interscience.