CHEMICAL ENGINEERING TRANSACTIONS
VOL. 86, 2021
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Sauro Pierucci, Jiří Jaromír Klemeš

ISBN 978-88-95608-84-6; **ISSN** 2283-9216
DOI: 10.3303/CET2186153

Data-Driven Digital Twin of a Chemical Production site For Production and Utilities Planning

Filippo Ferranti[a], Geoff Vingerhoets[b], Flavio Manenti[a], Mattia Vallerio[b,*]

[a]Dipartimento di Ingegneria Chimica e Materiali, Politecnico di Milano
[b]Data Analytics, BASF Antwerpen
mattia.vallerio@basf.com

The BASF Antwerp site is one of the most advanced and integrated chemical production sites of the BASF group and worldwide. The chemical plants on site are heavily interconnected, i.e., the product of one plant is the raw material used in the next one and some of the plants use the steam produced from neighboring ones. This high degree of interconnection makes it quite difficult to assess the feasibility let alone the economic optimality of production and maintenance plans for the entire site. This research explores the use of a data-driven digital twin to simulate and assess these plans. Two value chains, i.e., a collection of plants converting raw material to valuable end products have been selected as a proof-of-concept for the entire site. Each plant has been modeled by utilizing simple or multiple regression. Each regression model correlates the final product of a plant with the needed raw materials or utilities (e.g., steam or electricity). All regression models have been found using the software JMP. Additionally, all relevant tanks used to stock raw materials, intermediates and final products have been modelled. This allowed for the visualization and troubleshooting of a particular component excess or shortage. The resulting system of 76 variables has been solved in MATLAB in a multi-period fashion, where a period represents a day. The simulation has been first performed on the training data set, and then on a validation period to verify the models' performance.

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 optimize production and reduce emissions. One way to reach this goal is sharing facilities and creating interconnected processes. This aggregation also increases the complexity of the site and a greater effort is required to schedule and plan the production. This paper focuses on building a digital twin for two value chains located at the BASF site in Antwerp. A chemical value chain can be defined as a group of chemical plants, where the finished product of one serve as raw material for the next. From an industrial point of view, the main objective of this work is the creation of a tool to assess the effects of planning decisions on the whole site and therefore increasing operations' transparency. This will enable the plant and site management to make better decisions in different scenarios. From an academic perspective the purpose is to verify if it is possible to perform this task using only simple regression models. The aim is to simulate only stationary operations on a long-term period, excluding dynamic and non-operating periods. Several methods exist to create a model or a digital twin of an existing system. The main differentiation worth making is between physics-based and data-driven approaches (Wright and Davidson, 2020). A physics-based approach is achieved through the derivation of models based on all the physic-chemical phenomena occurring in the analyzed system, while a data-driven one aims at deriving the governing phenomena by studying the correlation observed on the available historic measurements. The first approaches are normally characterized by high complexity and computational effort while data-driven methods usually lump many parameters making the models simpler and limiting the amount of necessary information. The main drawbacks of these latter approaches are mainly that: (i) results might be difficult to interpret and (ii) their validity is limited to the input data range. The use of a data-driven digital twin for planning and production simulation in the chemical sector is not new.

Paper Received: 26 August 2020; Revised: 6 February 2021; Accepted: 3 May 2021
Please cite this article as: Ferranti F., Vingerhoets G., Manenti F., Vallerio M., 2021, Data-driven Digital Twin of a Chemical Productionsite for Production and Utilities Planning, Chemical Engineering Transactions, 86, 913-918 DOI:10.3303/CET2186153

A similar concept applied to batch plants can be found in Fumero et al. (2016), where the consumption of raw material is determined through fixed conversion. In our case, the resulting models render a variable consumption according to the, e.g., different production loads or outside temperature. A data-driven approach has also been investigated by Muteky and MacGregor (2008), where models were made for the purchasing of coals for coke making in the steel industry. Finally, in Li et al. (2016) a data-driven model approach is used for part of the units of a highly integrated refinery-petrochemical complex to reduce the computational effort. Similarly, this work aims at finding models correlating the raw material and energy flows entering a plant with the product and energy flows leaving it. The collection of all resulting models will give a complete digital twin of the two chemical value chains considered as use case. This article is structured as follows: section 2 discusses the relevant state of the art, section 3 introduces the use case, section 4 presents the results and finally section 5 discusses the conclusions.

2. State of the art

For this work only single and multiple regression models have been considered. The main reason is to evaluate the accuracy that could be achieved with the simplest model. The resulting model structure can be described as follows:

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

where *y* is a p-by-1 vector of measurements of the dependent variable, *f* is any function of *X* and *β; X* is a p-by-n matrix of predictors, with one row for each measurement, and one column for each predictor*; β* is an n-by-1 vector of unknown parameters that have to be estimated; s is a p-by-1 vector of identically distributed independent random disturbances. To obtain these equations the least-squares method has been used in JMP. In its most simple form, it consists in the estimation of the parameters through the minimization of:

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

with respect to the parameters (Elster et al., 2015). The quality of the obtained models has been assessed based on two parameters: the root mean square error (RMSE) and R^2. Where, the RMSE represents the average error between the model prediction and the actual values, while R^2 is the coefficient of determination which indicates the proportion of the variance in the dependent variable *y* that is predictable from the independent variable(s) x_i. Additionally, a detailed analysis of the residuals was also performed. Residuals are defined as the difference between the real values and the predicted ones. These should be normally distributed with zero mean. For more details on how these two terms are calculated the reader is referred to Steel R. and Torrie J. (1960) or any other book on statistics or regression.

3. Use case introduction

Due to confidentiality reasons, the names of plant and chemicals have been anonymized and all data and results have been normalized. The normalization has been realized by dividing all the hourly data by the highest value of the parameter. The results shown here are the summation of the normalized results into a daily value. The case study selected for this work is based on two production value chains of the BASF site in Antwerp. Figure 1 shows a graphical representation of the studied system. In particular:

- in yellow all plants related to the production value chain of Product 1;
- In green all plants related to the production value chain of Product 2;
- Plant A and C are a cluster processing the raw materials arriving on the site to obtain Product 1 and 2.

The two value chains are made of 10 plants indicated with a letter from A to L. The raw materials entering the value chain are called “Reagent X”, with a number between1 and 6, while the final products and the intermediates are called "Product X", where X is a number between 1 and 23. Tank farms are present between plants to collect and store the chemicals on site. The tanks have also been numbered from 1 to 24. The last part of the digital twin consists of the utilities net that serves every plant. As it is possible to notice, all plants are interlinked between them, with many products from a plant serving as reagents for the next one.

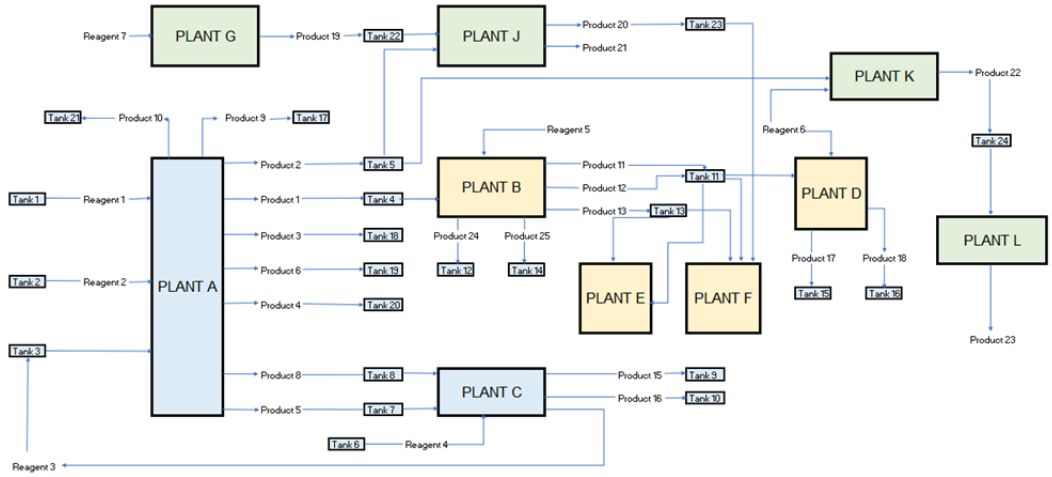

Figure 1: High-level representation of the value chains

3.1 Data

The data considered for this work has been divided into two groups: the training data goes from 1/11/2017 to 19/12/2019, while the validation data goes from 20/12/2017 to 12/04/2020. The collected observations consist of hourly averages. The focus of this work is to predict the consumption of raw materials and utilities while plants are in stationary operating conditions. Therefore, only data that match this description has been retained for analysis and validation, the remaining part was excluded in subsequent steps by removing: (i) the data related to plant turnarounds and shut-downs, (ii) data related to dynamic and transition periods, i.e., load changes, production ramp up and down and (iii) uni-variate and multivariate outliers.
The data related to the dynamic periods was identified by using a stability factor μ, defined as follows:

$$\mu = std(M(i,j), M(i,j-1), \dots, M(i,j-12)) \quad (3)$$

where std(M(i,j)) represents the standard deviation of material i at time j. It is possible to exclude the non-stationary modes in a systematic way by selecting a limit, based on the sensor accuracy, for the stability coefficient. The intervals showing a value lower than the defined limit are kept and the rest is excluded.

3.2 Data-driven modelling

All models have been built with the statistical software package JMP. For the sake of brevity, only one model will be discussed in detail (see Eq (4)). The equation describes the consumption of Reagent 1 in Plant A as a function of Product 3, 4 and 5 (see Figure 1). Figure 2a shows the data used for the model, in black, and the excluded one corresponding to failures and a major turnaround in red.

$$R1 = -0.0965 + 1.235 * \mathrm{P3} + -0.096875 * \mathrm{P4} + 0.11027 * \mathrm{P5} \quad (4)$$

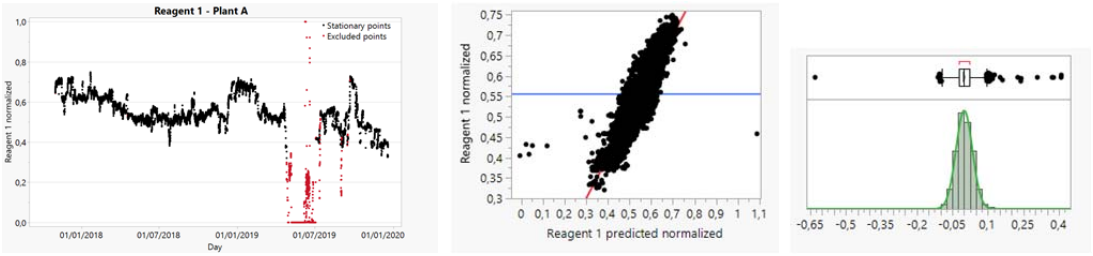

Figure 2: (a) Consumption of reagent 1 in the training period, (b) Correlation between Reagent 1 and the finished goods presented in the equation 4, (c) Distribution of the errors of the model

Figure 2b visualizes the actual values (black points) vs the predicted ones (red line) obtained through Eq (4). The resulting statistical parameters for this regression are:

$$R^2 = 0.813 \qquad RMSE = 6.24\,\%$$

Moreover, the residuals are normally distributed, as shown in Figure 2c, where the green line represents a normal distribution with the same mean and standard deviation of the sample of errors. The same approach has been used for all models.

3.3 Tank farm

The site is coping with enormous quantities of chemicals and tanks are needed to store materials before selling or using them in other plants. For that reason, every raw material and finished good relies on a certain number of tanks. The level in each tank is calculated as follows:

$$TL\ (i,t) = \sum inflow - \sum outflow + TL\ (i,t-1) \tag{5}$$

where $TL(i,t)$ and $TL(i,t-1)$ represent the tank level of Tank i at the current and previous period, $\sum inflow$ and $\sum outflow$ are the sum of all materials entering and exiting the tank, respectively.

3.4 Simulation platform

A multi-period simulation has been set up in MATLAB by dividing the entire time interval on a daily basis. The core of the simulation is the solution of the nonlinear algebraic system consisting of all the models and the material balances of the supply chain. In Table 1 the number of variables per simulation is shown. The simulation has been repeated for every day of the training and validation data-set.

Table 1: Number of variables and number of simulations per data-set

	Training data-set	Validation data-set
Variables number	76	76
Simulation days	760	115

The system of algebraic nonlinear equations is solved by minimizing the sum of squares of the components. The system is considered solved when the sum of squares approximates the value of zero. The method used by the solver is the Trust-Region-Dogleg Algorithm (see Conn et al., 2000, Nocedal and Wright, 2006 and Powell, 1968). Being the algebraic system nonlinear, it is necessary to provide an initial guess to the solver. This is randomly selected within the variables' defined boundaries.

4. Results

Two different simulations have been done to analyze the quality of the models and the whole simulation environment: (i) first the training data set and the results have been compared to the historical training data; (ii) then on the validation data-set. The results evaluation has been performed based on the following five criteria: (i) residuals normally distributed, (ii) residuals variance as small as possible, (iii) residuals mean as close as possible to zero, (iv) RMSE as small as possible and (v) no prolonged deviations from historical data. Training and validation results should respect all the criteria expressed before. If all the criteria are meet, the parameters of the results should be compared between them to see if there are significant differences. If that is the case, it will be necessary to understand the reasons behind it. The possible causes are: (i) an error in the sensor observations, or a (ii) a change in the studied process.

4.1 Raw materials

Looking at the two graphs in Figure 3a and 3b, it is immediately evident that the model is quite accurate. The distance between predicted and real values remains almost constant for the validation data set as well, confirming the validity of the model. This is also confirmed by the statistical reported in Table 2. These parameters are expressed as a percent of the average of the historical data for that same period. Their similarity shows that it is possible to use this model also in future predictions with similar results.

Table 2: Reagent 1 errors for training and validation data-sets expressed in % of average value

	Training data-set	Validation data-set
RMSE	5.72	7.09
Residuals mean	-0.05	-1.71
Residuals std	5.73	6.92

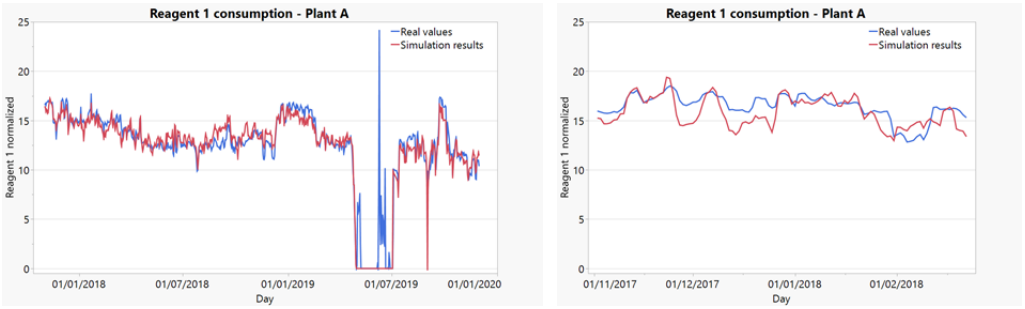

Figure 3: Comparison between simulation and training data (a) and validation data (b)for reagent 1

4.2 Utilities

For the utilities, let us consider the 16 bar steam consumption in plant B. The main reaction in this process is endothermic, while the side reactions are exothermic. In addition, the selectivity towards the main reaction it is inversely proportional to the catalyst age. Therefore, the steam is mainly consumed in the first part of the catalyst life while it is produced in the second part. Figure 4a shows a decreasing trend until May 2019, when the catalyst was changed. After that there is a steep increase, followed by another decreasing trend. Despite the complexity, the model is able to predict the whole period (see Figure 4a,b). The model equation is:

$$SB = 1.445 - 0.0000152 * CLT - 0.8415 * P24 - 0.655 * P11 \ - 0.2634 * P12 + 0.003506 * OT \quad (6)$$

where P24, P11 and P12 are the flows of Product 24, 11 and 12 entering plant B and OT is the outside temperature. SB is the 16 bar steam consumed, while CLT represents the catalyst lifetime, defined as the cumulative sum of Product 1 up till that day. The steam consumption peaks in Figure 4a are due unexpected stops of the plant, during which an excess of steam is introduced to guarantee safety. Figure 4b shows a constant difference between the real steam consumption and the simulation results for the validation period. The model based on data from the training period is not able to predict the data in the validation period. This bias mainly derives from a catalyst change and from a lower production load in the first months of 2020. An overview of the statistical parameters per period is reported here below in Table 3:

Table 3: Statistical parameters for the 16 bar steam model of plant B for the training and validation data sets

	Training data-set	Validation data-set
RMSE	32.41	66.51
Residuals mean	-2.03	62.30
Residuals std	32.38	23.38

The parameters are expressed as percentage of the mean value for the respective period. The deviation between the model and the actual values in the validation period appears significant. The best approach to correct this would be to retrain the model by including the validation period.

Figure 4: Comparison between simulation and training (a) and validation data (b) for steam used in plant B.

4.3 Tanks

Figure 5 shows the volume of resources stored in any given period throughout the simulation for three tanks. Tank 5 shows the accumulation of Product 2. The stored amount is continuously increasing except for the period when plant A undergoes a shutdown. In the shutdown interval the level of the tank decreases. The volume in Tank 8, which is storing Product 8, is instead mainly constant for the entire period. Note that only simulation results are showed since not all tanks are present on site.

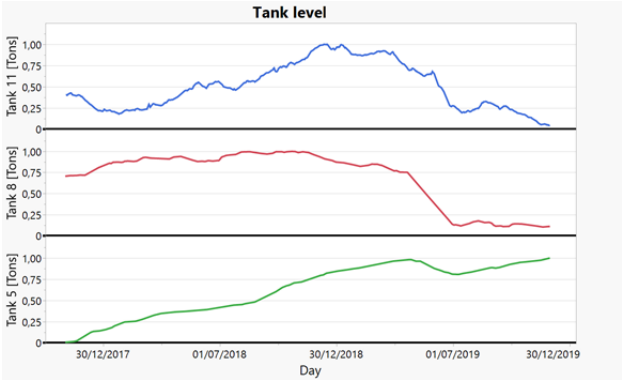

Figure 5: Level of the tanks number 5, 8 and 11 in normalized unit

5. Conclusions

This work introduces a data-driven digital twin of two production value chains. The goal of this work was dual: (i) from an industrial point of view, the main objective was to create a simulation tool for the production planning of a complex interconnected chemical site and (ii) from an academic perspective it was to verify the possibility to use a combination of simple regression models to achieve the industrial goal. The results showed in section 4 support the latter assumption. The created digital twin can be used to simulate and do a scenario-based analysis of the production planning for a complex chemical production site. The main challenge related to its implementation arises from the monitoring of the digital-twin accuracy and its maintenance. While the monitoring step can be automated by implementing statistical process control, the retraining steps will require human activity to assess the overall quality of the model. It is worth mentioning that this approach focuses on long term production planning, i.e., a month ahead or more, where the actual status of the production plant does not play a significant role. A short-term planning tool, on the other hand should incorporate this information, e.g., fouling level, mechanical wear, raw material quality. The next steps will introduce a multi-objective optimization framework in order to optimally plan production and energy consumption under different product and energy market conditions (Vallerio, M. et al., 2015 & Nimmegeers, P. et al., 2019).

References

Conn, A., Gould, N., and Toint, P.L., 2000, Trust-region methods. mps-siam series on optimization siam and mps. Society for Industrial and Applied Mathematics: Philadelphia, PA, USA.

Elster, C., Klauenberg, K., Walzel, M., W¨ubbeler, 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,”.

Fumero, Y., Moreno, M.S., Corsano, G., and Montagna, J.M., 2016, A multiproduct batch plant design model incorporating production planning and scheduling decisions under a multiperiod scenario. Applied Mathematical Modelling, 40(5), 3498 – 3515.

Li, J., Xiao, X., Boukouvala, F., Floudas, C.A., Zhao, B.,Du, G., Su, X., and Liu, H., 2016, Data-driven mathematical modeling and global optimization framework for entire petrochemical planning operations. AIChEJournal, 62(9), 3020–3040.

Muteki, K. and MacGregor, J., 2008, Optimal purchasing of raw materials: A data-driven approach. AIChEJournal, 54, 1554 – 1559.

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, 91: 349-362. https://doi.org/10.1002/cite.201800082

Nocedal, J. and Wright, S., 2006, Numerical optimization. Springer Science & Business Media.

Powell, M.J., 1968, A fortran subroutine for solving systems of nonlinear algebraic equations. Technical report, Atomic Energy Research Establishment, Harwell, England (United Kingdom).

Steel, R. G. D. and Torrie, J. H., 1960, Principles and Procedures of Statistics with Special Reference to the Biological Sciences.

Vallerio, M., Vercammen, D., Van Impe, J., 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,

Wright, L. and Davidson, S., 2020, How to tell the difference between a model and a digital twin. Adv. Model. and Simul. in Eng. Sci., 7(13).