\*mattia.vallerio@solvay.com

Abstract

Batch processes are the main production methodology for a large number of manufacturing industries: e.g., chemicals, food and beverage, pharma, Levenspiel (1998). However, these processes are subject to high variability: raw material composition, initial condition, unit degradation, and their intrinsic nonlinear and dynamic nature. All these reasons make batch processes challenging to analyse, control and optimize, Kourti (2005), Rawlings (2017). Nowadays, data can be considered an additional asset of all manufacturing processes and with the introduction of IoT devices, the economic burden to capture even more data has dropped. This data overload does not help on its own. It generates an added value when it is turned into actionable information (e.g., reduce batch time, increased first-time-right production). This contribution aims at introducing the need to establish the field of industrial data science (Mowbray et al. (2022)), and by doing so it highlights the use of two well-established data analytics methods in the context of batch manufacturing analysis. A well-known industrial example will be used to demonstrate the newly defined workflows to apply these two machine learning methods to convert this high variability and apparent excess of data into valuable information. The first method involves AutoML analysis, Wilson (2017). We will present how to automatically summarize its properties into features and identify the most relevant ones via non-linear correlation analysis (e.g. random forest) for a batch process. Then, trajectory analysis and functional data exploration will be covered. However, before doing so, we will discuss the need to align the data time-wise to effectively use these techniques (Dynamic Time Warping). In this last step, we will use the Functional Data Explorer in JMP Pro to monitor and identify deviations, Silverman (2002). These two novel workflows show how to apply ML method to industrial data successfully.

**Keywords**: Data Analytics, Machine Learning, Batch Manufacturing, AutoML, FPCA

* 1. Introduction

Batch processes are widely used in manufacturing, since they allow for minimum capex and maximum flexibility. An inherent characteristic of batch processes is the need to define and follow a recipe. This set of step by step rules guides the production and should ensure that a final product with a specific target quality should be obtained. Given the sometimes quite numerous steps involved in typical industrial recipes, it is quite evident how these processes will exhibit large variability between different batches with the same recipe. In recent years this drawback has been mitigated by the introduction of advanced automation and control methods specifically developed for batch processes (e.g., trajectory control, MPC). Kourti (2005); Rawlings et al. (2017). More recently the process industry has seen a sharp increase in interest and investments in Machine Learning (ML) and Artificial Intelligence (AI) Beck et al. (2016); Chiang et al. (2022); Sansana et al. (2021). It is now clear that methods and best practices used in other fields do not translate well to industrial data analysis in the process industry Mowbray et al. (2022). This is mainly due to three reasons: (i) the necessity to ensure safe operation, (ii) the lack of large and relevant industrial datasets and (iii) presence of legacy, outdated or underperforming automation and operational technology infrastructure. Clarke (2016); Shang and You (2019); Schweidtmann et al. (2021). In this work, we review the challenges in analyzing industrial batch data and present two novel workflows to obtain valuable insights using ML: Feature screening with AutoML (supervised learning) and anomaly detection with Functional Principal Component Analysis (FPCA) (unsupervised learning). The novelty of this work is in the methodology to apply ML on industrial data successfully. Additionally, we also introduce the use of noise as a feature to assess factor importance.

* 1. Batch data analytics: drying process use case

Analysing batch data is challenging due to varied model inputs (e.g., raw material properties, evolving conditions). This study explores two data-driven workflows for (i) screening for root causes and (ii) batch monitoring for anomalous events. The two workflow will be shown on an industrial batch drying process. The batch dataset is openly available and includes three phases: (i) Deagglomeration, (ii) Heating, and (iii) Cooling (see Fig.1). Structural and chemical reactions occur during drying. The batch begins with a variable cake amount and unknown solvent (Z). The operator adjusts dryer temperature and agitation speed and measures 10 variables (Xs). At the end, a sample determines the remaining solvent (Y) for quality control. (Garcıa-Munoz et al., 2003, 2004).



Fig.1 - Example of different type of variables shown in a trend over a sequence of five different batches.

* 1. Feature screening with AutoML

Batch data of varying time lengths is often analysed by summarizing each batch through statistics and process knowledge, such as peak temperature or its average rate of change during the reaction phase. These are referred to in the literature as landmark points or fingerprints. This approach assumes that subject matter experts (SMEs) know in advance the essential features to generate. Extending this method, one can compute statistics (e.g., max, min, average, range, std, quantiles) for every sensor, during each phase, for every batch, and grade. In ML, this process is known as feature engineering, often coupled with algorithms for optimal predictor selection.

* + 1. Auto ML for batch processes

AutoML aims to streamline machine learning problems by automating traditionally manual, complex analytical tasks that are often performed by data analysts. It operates under the assumption that there is an excess of data for fitting multiple models (training dataset), selecting model parameters to avoid overfitting (validation dataset), and finally assessing with unseen data (test dataset). In industrial applications, this assumption may not always hold, making AutoML mainly suited to screen for relevant sensors effortlessly.

* + 1. Feature engineering & selection

Automatic calculation of summary statistics for a batch process is possible. Once feature sets are generated per product, batch, and phase, these can serve as model inputs (X's). These features can be as granular as needed, summarizing statistics per automation step). They can also leverage all available process knowledge (e.g., pressure compensated temperature). ALAMO (Wilson and Sahinidis, 2017) is a well-known software package for this task. After generating the desired features, the next step is feature selection, achievable with any ML method capable of efficiently handling challenges like non-linear relationships, high co-linearity, and noise. Fig. 2 illustrates the results of this approach. Multiple statistics (mean, max, min, standard deviation, coefficient of variance) for every sensor, phase, and batch are calculated. A random forest model (termed Predictor Screening in JMP [SAS Inst.]) is used to list the contribution of each predictor. We introduce the use of an artificial noise signal as a cut-off to distinguish important factors.



Figure 2 - AutoML approach for batch process data. a) Statistics of all sensors per phase and batch are calculated, b) then a random forest model identifies the fingerprints with the strongest correlation to the target and c) exemplified by the torque sensor showing a correlation to the target.

* 1. Anomaly detection with FPCA

The variability in the drying process exemplifies the diverse durations encountered in batch processes (see Fig. 2a). In the dryer dataset, the primary source of variability arises from batch-to-batch differences in the loaded product amount and its solvent content. Typically, phases can extend due to various perturbations, catalyst deactivation, variations in raw material quality or quantity, reduced heating/cooling capacity, or simply from maintenance issues or scheduling decisions. This complexity introduces variability in the expected progression of a batch manufacturing process, whether executed through automated or manual procedures. Various techniques exist to align batch data with differing durations. The simplest and most efficient approach involves using automation triggers or, alternatively, a monotonously increasing or decreasing variable along the batch duration (e.g., conversion of the reaction or amount of water evaporated instead of time). When information for batch alignment is neither measured nor known, or real-time alignment is necessary, Dynamic Time Warping (DTW) prove useful. These methods statistically align batch trajectories and can be employed for classifying anomalous batches or identifying correlating parameters (Spooner et al., 2017; Zuecco et al., 2020).

* + 1. Functional or shape analysis (FPCA)

When batch-to-batch variability is minimal or exploring specific subsets of sensors is crucial, summary statistics might not be detailed enough and detailed trajectory analysis methods should be used. Functional Principal Components Analysis (FPCA) is an extension of PCA that analysis curves or trajectories. In batch processes, FPCA captures the primary sources of variation among multiple trends (see Fig.3) as a "function of batch time". FPCA applications extend to various domains: HPLC data (function of analysis time), spectroscopy data (function of wavelength), vibration (function of frequency), and battery degradation. Using FPCA, these trends or trajectories are summarized with a mean curve and a series of “shape functions” (eigenfunctions). These shapes have associated weights (loadings or eigenvalues). Each shape encapsulates the variability observed in latent trajectories (Silverman and Ramsay, 2002; Srivastava and Klassen, 2016). If the batch data is pivoted, FPCA insights might seem analogous to standard PCA. FPCA excels in identifying a set of component shapes that explain the maximum variation in the observed data. These can be interpreted as distinctive features seen in the process for certain batches, such as a different level rate in the deagglomeration phase or a temperature “shoulder” in the heating phase. Each batch trend can be reconstructed with the mean trajectory and a linear combination of these weighted shapes. The weights are called FPCs score and vary from batch to batch. As not all shape functions have equal contribution or importance, their weight must be considered when examining score plots.

* + 1. Functional statistical process control

Process engineers face constraints in creating, monitoring, and modifying Key Performance Indicators (KPIs) for their processes. Functional Statistical Process Control (FSPC) offers a crucial advantage by alleviating them of this task. The core concept analysis is to compare batch trajectories instead of relying on pre-selected KPIs. This process can be automated using FPCA to identify intrinsic trajectories for each sensor. Illustrated in Fig.4 a, a multivariate control chart can quantify the anomaly level in drying batches, irrespective of their time duration. Initially, batches undergo pre-alignment to eliminate any anomaly directly related to batch duration. Then, functional components serve as fingerprints for both tank level and drying temperature (Fig.4b & c). The Hotelling-T2 score indicates batch trajectories that shows anomalous behaviour. Examining the aligned batches in Fig. 4d reveals anomalous patterns in both temperature and tank level. Using FPCA, the only input required from process engineers is specifying which tags (sensors) to monitor closely. However, if a specific KPI, such as quality or production, is a priority, an AutoML screening analysis using correlation can be conducted first (refer to the previous example examining solvent content). This step can reduce the number of monitored tags improving focus on quality and production.





Figure 3 (top) & 4 (bottom) – Top: FPCA analysis of the tank level. This illustrates the description of the tank level a) using the average trajectory and a combination shape functions b) & c) with varying importance d), enabling the detection of anomalous batches e). Bottom: FSPC: A Hotelling-T2 control chart (a) is monitoring FPCA scores. Individual contributions for tank level and dryer temperature are depicted for an in-control (b) and an out-of-control batch (c). Colored batch trajectories for these sensors are shown d, not aligned left & aligned right.

* 1. Conclusions

In conclusion, this article emphasizes the growing importance of Industrial Data Science for analyzing, troubleshooting, and monitoring batch manufacturing processes. The challenges posed by the inherent variability and complexity of batch processes are addressed through the integration of data science and machine learning techniques. Two novel data analysis workflows are introduced along with the use of an artificial noise signal as a way to automatically discriminate import factors. The AutoML approach is introduced for feature screening, summarizing batch properties, and identifying relevant features through non-linear correlation analysis. The second method involves Functional Principal Component Analysis (FPCA) for anomaly detection, leveraging the entire trajectory of batch data. The alignment of batch data in terms of time and the use of FPCA shape functions are highlighted as essential steps in this data-driven methodology. The application of these methods is illustrated through a case study on an industrial batch drying/reaction process, showcasing the potential for actionable insights and continuous improvement in process control and optimization. The article underscores the need for a nuanced approach in applying machine learning methods to the process industry, considering safety, data limitations, and existing infrastructure challenges. Overall, the integration of data-driven methods with domain knowledge holds significant promise for advancing the efficiency and effectiveness of batch manufacturing processes.

References

D. A. C. Beck, et al. , 2016. Data science: Acc-elerating innovation and discovery in chemical engineering. AIChE Journal 62,1402–1416.

L. H. Chiang, et al., 2022. Towards artificial intelligence at scale in the chemical industry. AIChE Journal 68 (6).

S. Garcia-Munoz, et al., 2003. Troubleshooting of an industrial batch process using multivariate methods. Industrial & Engineering Chemistry Research 42 (15), 3592–3601.

S. Garcia-Munoz, T. Kourti, J. F. MacGregor, 2004. Model predictive monitoring for batch processes. Industrial & Engineering Chemistry Research 43 (18), 5929–5941.

T. Kourti, 2005. Application of latent variable methods to process control and multivariate statistical process control in industry. International Journal of Adaptive Control and Signal Processing 19 (4), 213–246.

O. Levenspiel, 1998. Chemical Reaction Engineering, 3rd Edition. John Wiley & Sons, Nashville, Tennessee.

M. Mowbray et al., 2022. Industrial data science – a review of machine learning applications for chemical and process industries. React. Chem. Eng. 7 (7),1471–1509.

J. B. Rawlings, D. Q. Mayne, M. Diehl, 2017. Model predictive control: theory, computation, and design. Vol. 2. Nob Hill Publishing Madison, Wisconsin.

J. Sansana, et al., 2021. Recent trends on hybrid modeling for industry 4.0. Computers & Chemical Engineering 151, 107365.

A. M. Schweidtmann, et al., 2021. Machine learning in chemical engineering: A perspective. Chemie Ingenieur Technik 93 (12), 2029–2039.

C. Shang, F. You, 2019. Data analytics and machine learning for smart process manufacturing: Recent advances and perspectives in the big data era. Engineering 5 (6), 1010 – 1016.

B. Silverman, J. Ramsay, 2002. Applied functional data analysis: methods and case studies.

M. Spooner et al., 2017. Selecting local constraint for alignment of batch process data with dynamictime warping. Chemometrics and Intelligent Laboratory Systems 167, 161–170.

M. Spooner et al., 2018. Harvest time prediction for batch processes. Computers & Chemical Engineering 117, 32–41.

A. Srivastava, E. P. Klassen, 2016. Functional and shape data analysis. Vol. 1. Springer.

Z. T. Wilson, N. V. Sahinidis, 2017. The alamo approach to machine learning. Computers & Chemical Engineering 106, 785–795.

F. Zuecco et al., 2020. Troubleshooting an industrial batch process for the manufacturing of specialty chemicals using data analytics. 30th ESCAPE. Vol. 48 of CACE. Elsevier, pp. 1129–1134.