Consolidating Industrial Batch Process Data for Machine Learning

Simon Mählkvist[1,2] Jesper Ejenstam[2] Konstantinos Kyprianidis[1]
[1]Future Energy Center, Mälardalen University, Sweden, simon.mahlkvist@mdh.se
[2]Kanthal, Sweden

Abstract

The paradigm change of Industry 4.0 brings attention to data-driven modeling and the incentive to apply machine learning methods in the process industry. Further, capitalizing on a great deal of data available is an adverse task. For batch processes, the dataset is in a three-way format (Batch × Sensor × Time). Depending on the process and the goal of the analysis, it might be necessary to aggregate batches together. For this reason, a campaign unfolding structure is applied. By grouping the batches under new labels relevant to the analytical goal, campaigns are created. These labels can be created from periodical occurrences, such as refurbishing the refractory lining in the case of the case study. In order to utilize the three-way batch format, it is necessary to align the batches. In order to address this, the feature-oriented approach Statistical Pattern Analysis (SPA) is applied. SPA derives statistics, e.g., mean, skewness and kurtosis from the time series, consequently aligning the batches. The SPA and the campaign approach create a dataset consisting of select statistics instead of an irregular three-way array. Functional data analysis (FDA) is used to smooth and extract first- and second-order derivative information from the sensors in which functional behavior can be observed before creating features. Principal Component Analysis (PCA) is used to examine the final dataset. Further, industrial processes are notoriously nonlinear, and even more so batch processes. Therefore, kernel-based principal component analysis (KPCA) is used to review the final dataset. The KPCA can accommodate different underlying characteristics by modifying the kernel function used.

Batch Process Analysis (BDA), Batch Preprocessing, Functional Data Analysis (FDA), Statistical Pattern Analysis (SPA), Kernel Principal Component Analysis (KPCA)

1 Introduction

Within the scope of industry 4.0, industries are determined to incorporate into their analytical framework machine learning methods. Despite the vast selection of turn-key solutions, the procedure often falls short on the neglected part of the analytical procedure: data acquisition and preprocessing. Legacy process industries suffer from the fallback of outdated infrastructure, making data-acquisition procedures cumbersome and preprocessing complex due to, e.g., lack of contextual information such as accurate timestamps.

Batch data analytics (BDA) is a field of study that focuses on analyzing industrial batch processes. A batch process produces products in a turn-based manner which repeats over the following phases: charging, operating, and discharging. Working with batch process datasets offers unique challenges. The dataset a batch process provides is a three-dimensional matrix (Batch × Sensor × Time, see Figure 1), which offers additional challenges. E.g., each batch is going to be of different lengths, leading to an uneven dataset. Also, the time interval between samples may not be uniform. To accommodate irregular sampling or uneven batches, batch synchronization or feature extraction can be applied. In this work, The Statistical Pattern Analysis (SPA) method is used to compile relevant statistics from the sensor data and create an aligned three-dimensional array. Further, there are many strategies to convert a three-dimensional array into a two-dimensional array. This procedure is commonly called unfolding. It is necessary to unfold batch data to make it suitable for a more comprehensive array of models. Also, a campaign structure is applied to understand and explain variables after several batches. In short, the campaign approach entails concatenating the batches into new batches before the feature extraction procedure.

Furthermore, industrial batch processes can be expected to be nonlinear, making them unsuitable for many conventional methods. In order to investigate the nonlinear phenomenon, the batch structured datasets are analyzed using kernel-PCA (KPCA), which can accommodate nonlinear behavior due to its variation of kernel functions. A more conventional Principal Component Analysis (PCA) supplements this analysis to investigate the linear behavior as well.

A common phenomenon in industrial process analysis is noisy data. Also, for some processes, it may be beneficial to investigate derivative information. Therefore, to smooth the sensor data and extract derivative information, functional data analysis (FDA) is utilized. FDA creates a battery of approximation functions that describe the underlying processes, allow extraction of derivative information, and, consequently, smoothing. There are many complex methods available regarding

DOI: 10.3384/ecp2118576

structuring the data, smoothen and extract derivative information, and select or reduce features. This work combines several methods to see the potential for this blend. This work intends to impart a perspective towards consolidating complex industrial batch process data with machine learning methods.

The rest of the paper is organized as follows. Section 2. contains the methodology of the paper, which in turn consists of several subsections. Section 2.1 describes the process from which the dataset is constructed. Section 2.2. explains the data acquisition procedure. Section 2.3. elaborates on batch data processing and the campaign structure. Section 2.4. is on the functional data analysis. Section 2.5. on the feature-oriented approach. Section 2.6 on PCA and KPCA. Section 2.7 describes the analytical framework.

2 Methodology

2.1 Steel Converter Dataset

The data in this paper is derived from a steel converter. The purpose of a steel converter is to enable the use of low-grade resources, e.g., scrap-based or low purity, to produce high-quality steel. A converter refines steel batch-wise and is used as a secondary procedure. The raw material is melted and then fed into the converter. The converters' refractory lining interacts with the melt directly, and when the lining is exhausted, the converter is put out of commission and requires relining before it can be reinstated. The chemical composition of the melt is registered before and after the process. Also, gas inlet temperature, flow, and pressure are stored. The number of batches produced in the variable of interest is referred to as the campaign metric. The methodology aims to explore the relations between the degradation, the sensor, and chemistry.

2.2 Data Acquisition

Acquiring data in a legacy process industry is a diverse procedure. The accessible data is likely limited to a static tabular view. Due to the large size of unconstrained sensor data, such tables are usually limited in scope, i.e., duration, time-resolution, or amount of sensors. The memory of the computer or server can be a limiting factor if analyzing high-resolution data with many sensors. By dynamically detecting relevant sensors or duration in time of interest, it is helpful to systematically collect similar or more data regarding this or even connect a model directly to the system. Being able to analyze and collect data from the same development platform is beneficial. There are several parameters to consider when working on data acquisition for process data. Extracting the duration of interest may prove difficult. For each batch, identifying the duration in which the process is working with respect towards the batch is called local batch time. The accuracy of the local batch times' start and stop timestamps should be reviewed for accuracy. The stored information about the batch may not initially be intended to be used for analytics purposes and, therefore, inaccurate. The workstation and database server has limited memory. Hence when working with multiple batches over a significant duration, it is necessary to extract the data in manageable segments.

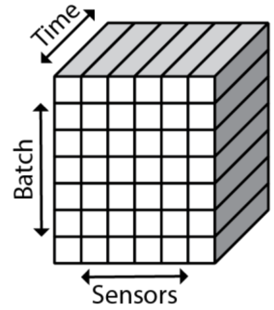

Figure 1. Aligned batch data

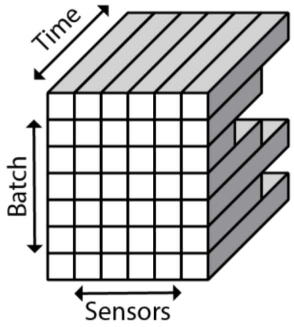

Figure 2. Uneven batch data

2.3 Structuring Batch Data

Batch data is designed in a three-way array structure with I, J, and K corresponds to a number of batches, variables, and time (local batch time) respectively, see Figure 1. (Nomikos & MacGregor, 1994) This results from the distinct structure of batch processes and, as a result, inhibits the utilization of conventional method without first transforming from three- to two-dimensions array. In practice, the batch duration deviates between batches, and batch data from industrial processes have different durations, hence producing an uneven dataset, as depicted in Figure 2. The procedure of aligning batch data is called batch synchronization or batch alignment.

By employing a campaign structure along with SPA, the need for time-sensitive alignment is circumvented. A description of how to combine and restructure the three-way array with data from upstream and downstream processes follows. The batch-wise unfolding approach transforms the dataset from an uneven three-way array to an uneven two-way array by concatenating the sensor trajectories (see Figure 3).

Wu et al. (2018) introduce a campaign-based batch unfolding structure which is further advanced by Wu et al. (2019). In batch processes where the metric or variable of interest varies or resets over several batches, the

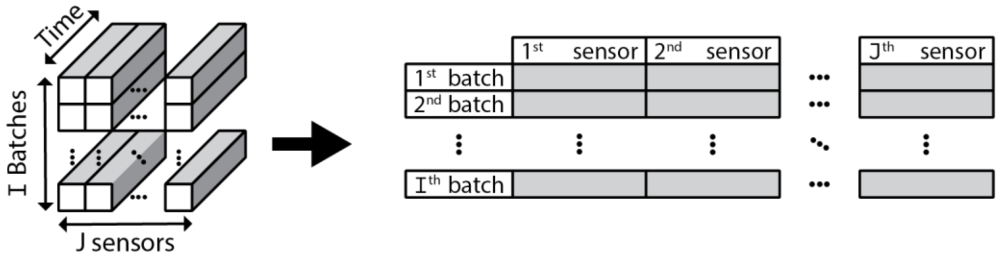

Figure 3. Batch-wise unfolding

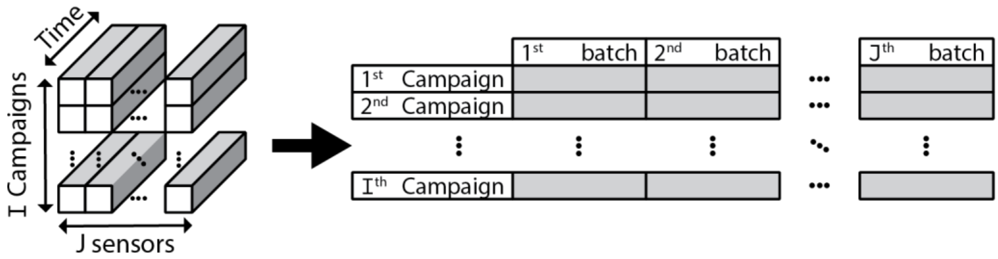

Figure 4. Campaign-wise unfolding

campaign approach coalesces metric-relevant data into a new type of batch (see Figure 4).

2.4 Functional Data Analysis

Functional data analysis (FDA), it applied in order to extract derivative information and to smooth the trajectories. FDA creates a battery of approximation functions that represents the underlying characteristic. Observing derivative information of physical variables, e.g., temperature and flow, can be beneficial since the derivative adds further information to the system. Using FDA, it is possible to interpret derivative information from time-series sensor data that show functional nature. (Ramsay & Silvermann, 1998) In Figure 5, the FDA is showcased where the original data is overlaid on top of the approximation function in subfigure (a). Further, it shows the 1st and 2nd derivatives of the approximation in subfigures (b) and (c), respectively.

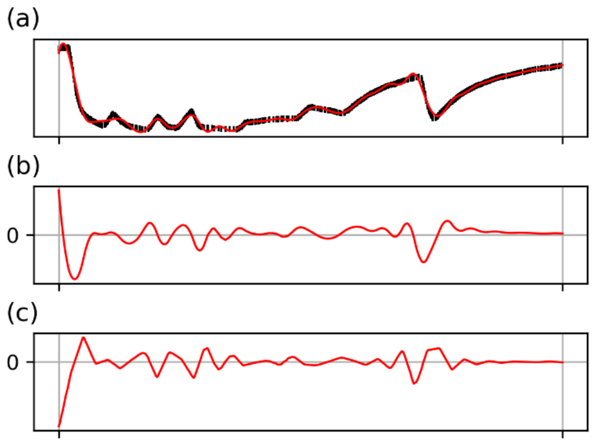

Figure 5. FDA example show smoothing and 1st and 2nd order derivative: (a) raw samples (black dots) and smoothing (red line), (b) 1st derivative, (c) 2nd derivative

Functional data representation is used in a multivariate functional kernel principal component analysis in (H. Wang & Yao, 2015). The functional local kernel principal component analysis is also in (F. He & Zhang, 2020). For more information about the FDA's fundamentals, see Ramsay et al. works on Functional Data Analysis (Ramsay & Silvermann, 1998).

2.5 Statistical Patten Analysis

The Feature-oriented method used in this approach is SPA. SPA was introduced in He and Wang (2011) as a fault detection framework and utilizes 1st, 2nd higher-order statistics derived from batch trajectories instead of the trajectories themselves to monitor the process. Like other feature-oriented methods, SPA alleviates the preprocessing step batch trajectory alignment by creating statistical metrics from the trajectories.

He and Wang (2011) monitor a semiconductor batch process and use the SPA statistics: mean, skewness, kurtosis, and covariance. Wang and He (2010) apply SPA with continuous processes and uses the following statistics: mean, variance, autocorrelation, cross-correlation, skewness, kurtosis. In this work, the mean, kurtosis, and skewness are used as statistics for pattern identification. Skewness is a measurement of distribution asymmetry. Kurtosis is a measurement of the spread of the data. In this work, SPA is used to transform a time series of data into a set of three statistics; mean, variance, kurtosis, and skewness.

For more information on feature-oriented methods for BDA, see Rendall et al. (2017) and Rendall et al. (2019). For a perspective on how the SPA framework relates to challenges purposed by smart manufacturing and other methods, see He et al. (2019)

2.6 PCA and Kernel-PCA

PCA is a commonly used method in process analytics for dimensionality reduction, feature selection, and unsupervised data exploration. The PCA is limited to investigating the linear relations in the dataset. Schölkopf et al. (1997) introduced KPCA, which provides further utility compared to the conventional, linear PCA. The KPCA aggregates the dataset, transforming it into a high-dimension feature space using a nonlinear mapping. Then, performing dimensionality reduction on the feature space and if a suitable kernel function and parameter is designed makes previously nonlinear data linearly interpretable. There are several kernel functions. The most commonly applied and used in this work are polynomial and radial basis functions (RBF). Both kernel functions have parameters that need to be configured. The KPCA can accommodate underlying nonlinear characteristics and show itself to outperform the PCA when performing feature extraction and classification on datasets with nonlinear behavior. (Lee et al., 2004) Works on fault detection using KPCA can be further viewed in the works by Lee et al. (2004), H. Want and Yao (2015), and He and Zhang (2020). For a fundamental look into Kernel methods, the reader is referred to the work of Schölkopf et al. (1997).

2.7 Analytical Design

Each batch has 15 sensors. FDA is applied 10 of these sensors, which creates 20 new trajectories per batch, 10 for both the 1st and 2nd order derivative. In total, this makes 35 trajectories per batch. Batches are concatenated to relevant campaigns. The chemical composition is measured for every batch before and after the process. 10 and 17 elements are registered before and after, respectively. For every sensor and chemical element, three statistical features are derived. Resulting in 186 features per campaign.

In order to investigate the impact of different methods, the campaign dataset is segmented into several different subsets. Every subset contains 93 campaign samples. The following list details the name, description, and number of features for the eight:

- Full: All data (186 features)
- 0th order: Sensor data without derivatives (45 features)
- 1st & 2nd order: Sensor data of both 1st and 2nd derivative (60 features)
- 1st order: 1st order derivative sensor data (30 features)
- 2nd order: 2nd order derivative sensor data (30 features)
- Chem: Both prior and post elemental composition (81 features)
- Pre chem: Elemental composition before the process (30 features)
- Post chem: Elemental composition after the process (51 features)

Using PCA, the explained variance for each of these dataset will be calculated and compared. The five most significant principal components (PC) will be further investigated, and each PC's five most significant features will also be compared. The fit of the significant PC to the campaign metric will be reviewed using ordinary least squares and the by investigating the coefficient of determination.

The KPCA is modeled for the RBF and polynomial kernel. The parameters for these are dynamic and is presented in results when relevant.

3 Results & Discussions

This section will show how the combined influence of the campaign structure, the FDA smoothing, the 1st and 2nd derivatives, and the SPA, by using PCA and KPCA to identify different phenomenon in the campaign dataset with respect to the campaign features.

The campaign dataset is segmented into eight different datasets, and PCA is performed on all constellations of the Campaign dataset. In Figure 6, the explained variance per principal component (PC) for 15 first components for each dataset is in a scree plot. Beyond 15 components, the explained variance approaches zero and is not included.

In Figure 7, the five samples of the largest explained variance of Figure 6 are extracted, and for each sample, again, the five features with the most significant explained variance are derived. There is overlap between the selected features, e.g., the features for the 1st PC in the Full dataset have similar significant features to the 1st component in the 1st & 2nd order dataset (see the top figure and second figure from the top in Figure 7).

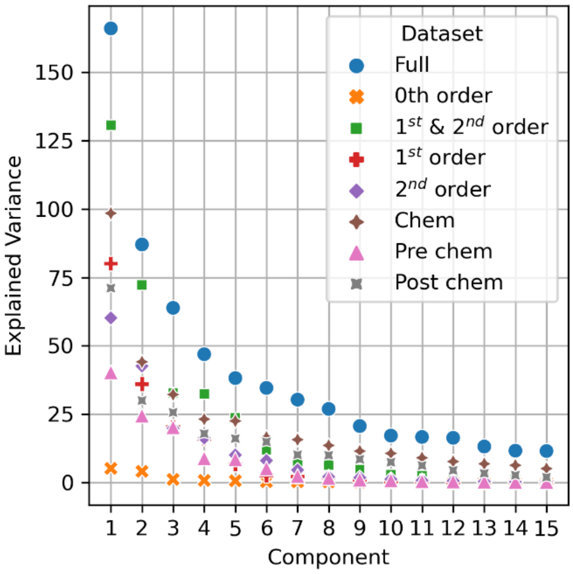

Figure 6. Explained Variance

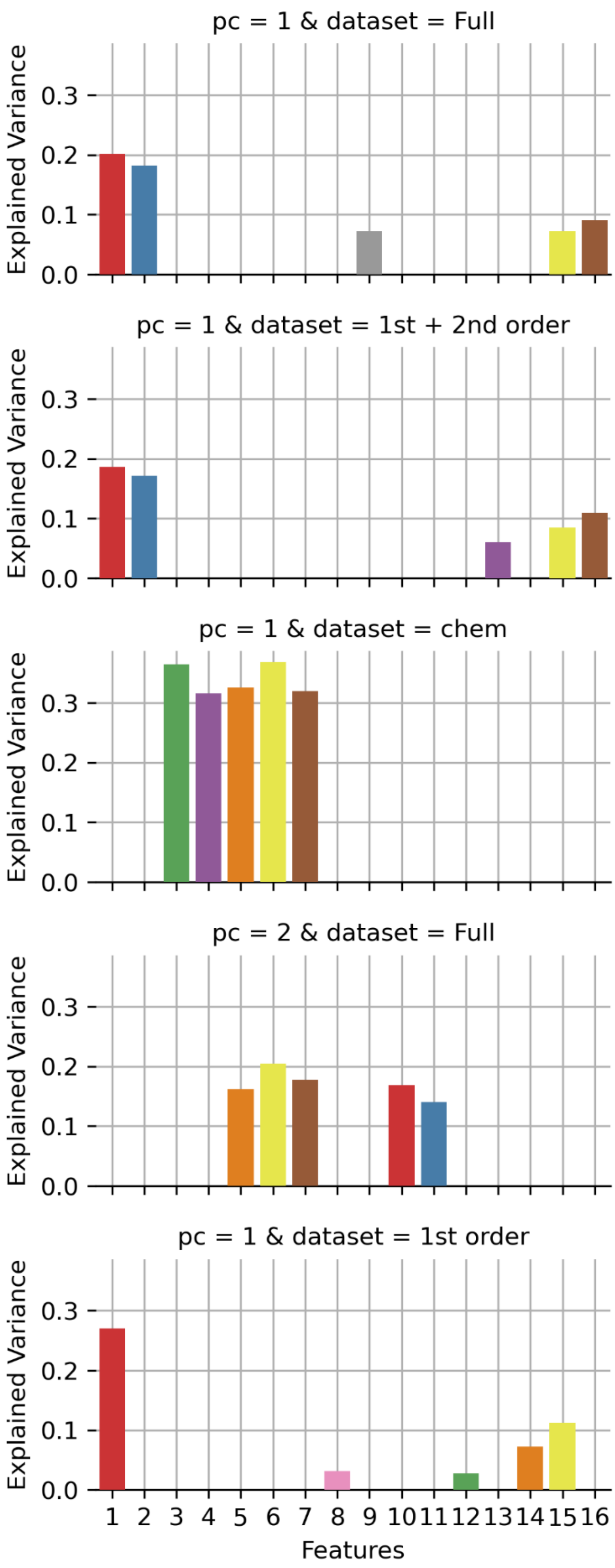

Figure 7. Feature variance of top 5 PCs

Two patterns are noteworthy when observing the data variance in Figure 6 and the variance of the features in Figure 7. First, the 1st PC of the Full dataset coincides with the 1st PC consisting of features from the 1st and 2nd order derivatives and indicates that the sensor derivatives influence the variance of the Full dataset. Further, the 1st PC from the dataset, consisting of only 1st order derivates, also shares features with both aforementioned datasets. The 1st feature seems to be the root of significant variance. Regarding the second discernable pattern, the 2nd PC of the Full dataset and the 1st component of the Chem dataset share three top features. Hence, the Full dataset explains the variation of two datasets with its 1st and 2nd PC.

The five PC is used to model the campaign metric. The coefficient of determination R^2 is calculated for each to determine how well the PCs are able to generalize the campaign metric. None show a significant R^2. Hence, while these PC describes the major variance in the datasets, they cannot be used to generalize the campaign metric.

The KPCA is used on the constellations of datasets. Several different KPCA is constructed using the polynomial and RBF kernels along with their corresponding parameters. Systematically investigating the pcs of the different KPCA constellations shows a weak linear correlation towards the campaign metric. No significant R^2. By observing the relations between the KPCA PC, two interesting patterns are visible.

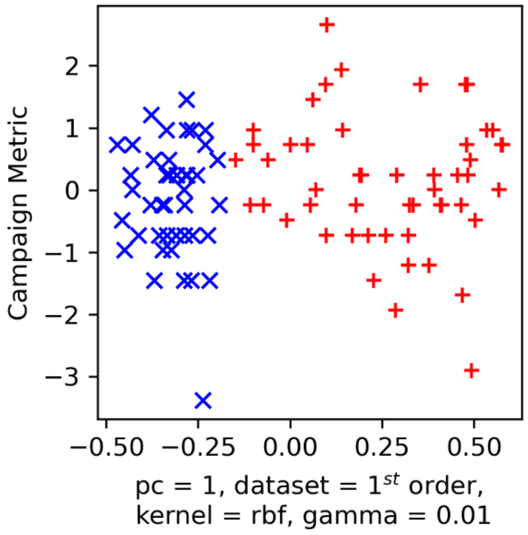

Figure 8. Clustering of campaign metric over PC from KPCA of the 1st order dataset

The first pattern is seen in Figure 8, where the PC is derived from the KPCA with RBF kernel and gamma 0.01 from the 1st order dataset. The 1st PC plotted over the campaign metric shows no significant correlation, but two clusters emerge, as illustrated by the shape and color difference. The features of the 1st order dataset are explored using a kernel density estimation (KDE) plot, and two features are distinct, as seen in Figure 9 and Figure 10, which, respectively, show the skewness and kurtosis of the same sensor. The sensor is the 1st derivative of the temperature sampled for a gas inlet.

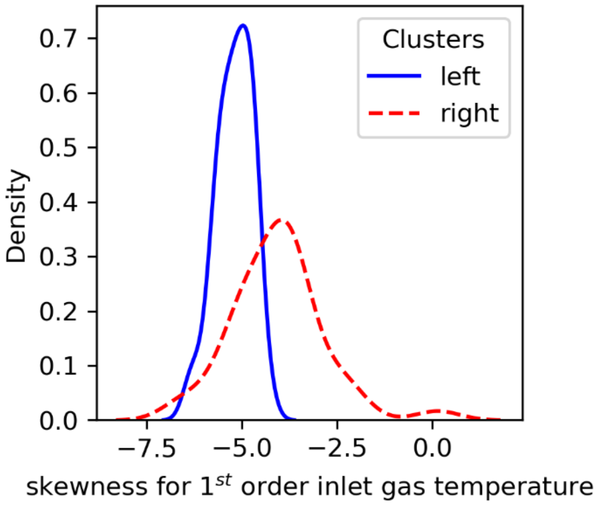

Figure 9. Density plot of skewness for 1st order derivative of inlet gas temp

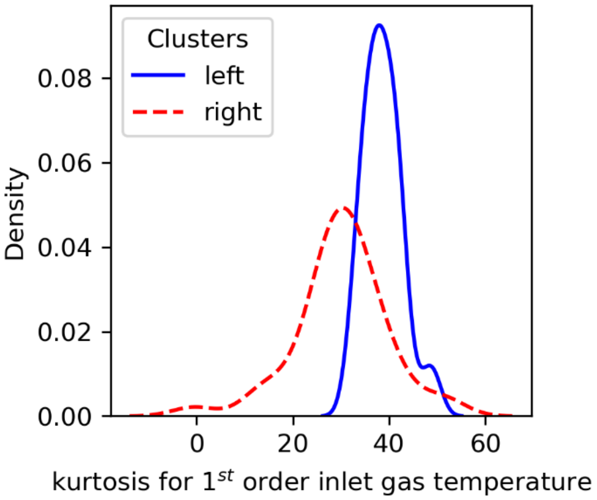

Figure 10. Density plot of kurtosis for 1st order derivative of inlet gas temp

The separation in Figure 9 and Figure 10 shows that the left cluster in Figure 8 has lower skewness and greater kurtosis than the right cluster.

The second pattern, seen in Figure 11, shows a clustering when comparing the campaign metric with the 1st PC derived from a KPCA with the RBF kernel and gamma of 0.001 of the Chem dataset. Figure 12 and Figure 13 further shows how this clustering transfers to the two features of the Chem datasets. Further, both features show the kurtosis for the chemical composition, of the same element, before and after the process. The clustering is overlapping since the distribution in Figure 12 for the right cluster shows two peaks, of which the line up with the other cluster. The right cluster for both figures aligns around where kurtosis is zero, which indicates a tighter spread. Hence, the left cluster is more random regarding the element's content, i.e., it has a more significant deviation.

Considering both the patterns identified via the KPCA (Figure 8 and Figure 11). Isolating the clusters and performing additional analysis could provide further information. Additional campaign samples would be beneficial as they would provide more data for analyzing data subsets. With a total campaign sample size of 93, further, partitioning can prove detrimental.

In general, the approach applied does not explain the campaign metric. Several variables may contribute to expanding the approach, and the rest of this section will reflect on this.

The PCA and KPCA transform the different datasets, and while the PCA is limited to a linear approach, the KPCA is not. However, even with several approaches by KPCA, a relevant interpretation concerning the campaign metric is not discovered. This may be because the analytical approach relies on a parametric framework that assumes the data to conform to underlying statistical distributions. Therefore it would be suitable to include non-parametric methods, such as variation of random forest, to analyze the relationship between the campaign data and metrics. The SPA is not limited to the statistics used in this work, and many other feature-oriented approaches have the potential to derive features that can explain other metrics.

The campaign-based approach unfolds the data batch-wise with respect to a campaign metric. Further, it is common in BDA to divide the batch into phases if the process has different operation modes. Also, it is possible that a campaign can have similar phases, e.g., the first batches are of specific interest and should be separated from the rest. Further, the analysis results of the campaign datasets may be understood if the batch mix is considered, e.g., the clustering is a result of different products or groups of products with similar pro-

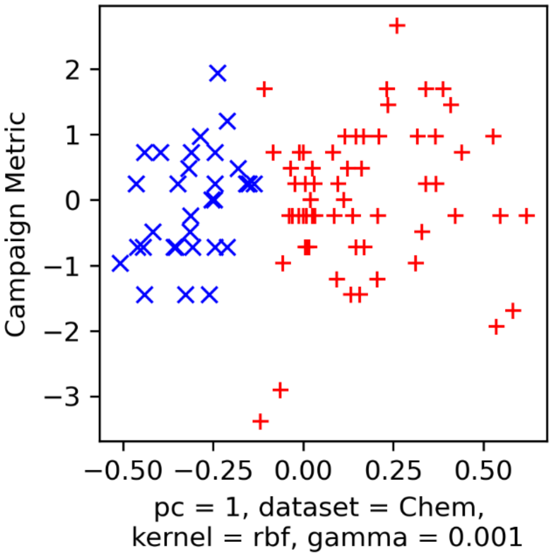

Figure 11. Clustering of Campaign metric over PC from KPCA of the Chem dataset

cess parameters, being overly represented in said campaign. Hence, designing the analysis so that the batch and campaign are divided into relevant phases is suggested. On the other hand, some of the datasets used are already high-dimensional. Considering batch and campaign phases, and adding additional statistical features, would further increase the number of features.

Therefore, an efficient and sustainable feature selection method should be applied so that a more complex and encompassing dataset can be considered.

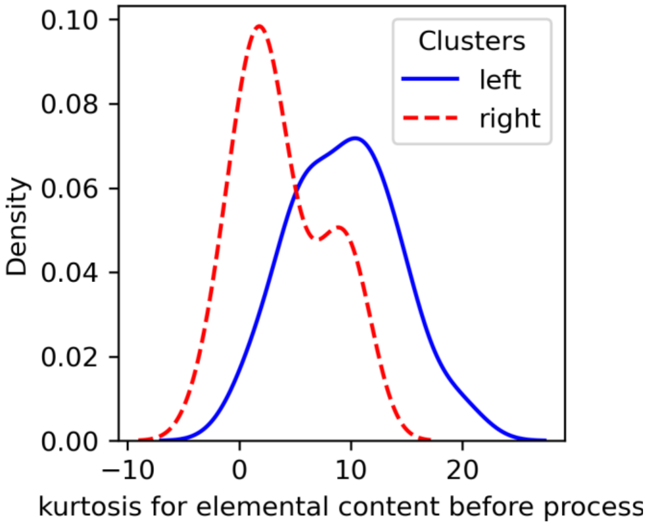

Figure 12. Density plot of kurtosis for elemental content before the process

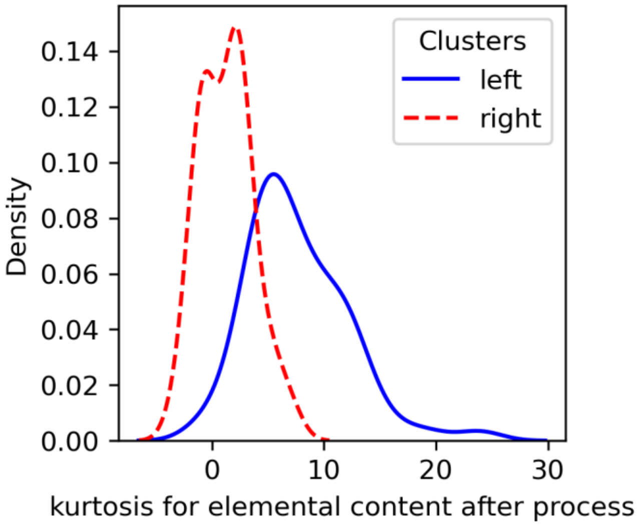

Figure 13. Density plot of kurtosis for elemental content after the process

4 Conclusion

This methodology provides a low complexity and a practical approach to batch process data preprocessing and synthesizes the unaligned three-way array into a two-way array. It is challenging to systematically evaluate the methodology due to a high number of design variables. The approach proves to be poor at generalizing the campaign metric, e.g., unable to explain the degradation mechanics. Several design improvements are discussed to enhance further the potential for the dataset to contain relevant information and increase the number of features, further aggravating the issues that high-dimensional datasets provide. Therefore, it would be valuable to implement a feature selection approach suitable for the campaign structure.

The KPCA approach uncovers interesting patterns in the data. These patterns manage to isolate different modes in the statistical features. The origin of these clusters is not determined, but their existence shows potential for the campaign structure to provide insights. It would be beneficial to increase the number of samples, I.e., increase the number of campaigns, to get a more accurate view of the underlying distributions by investigating data subsets. The challenge to this is that the rate at which data is generated is low. Hence the analytics has to rely on available historical data.

While the feature-oriented approach applied in this work is considered low complexity, the combination of campaign structure and FDA and KPCA makes it an elaborate construct. It shows potential to understand campaign-related phenomena, but further research into proper analysis methods is required.

Acknowledgements

The authors gratefully acknowledge Kanthal AB, Automation Region Research Academy (ARRAY), and the Swedish Knowledge Foundation (KKS) for their support.

References

F. He, and Z. Zhang. Nonlinear Fault Detection of Batch Processes Using Functional Local Kernel Principal Component Analysis. *IEEE Access*, *8*:1–1. 2020. doi:10.1109/access.2020.3004564

Q. P. He, J. Wang, and D. Shah. Feature space monitoring for smart manufacturing via statistics pattern analysis. *Computers and Chemical Engineering*, *126*:321–331. 2019. doi:10.1016/j.compchemeng.2019.04.010

Q. P. He, and J. Wang. Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes. *AIChE Journal*, *57*(1):107–121. 2011. doi:10.1002/aic.12247

J. M. Lee, C. K. Yoo, and I. B. Lee. Fault detection of batch processes using multiway kernel principal component analysis. *Computers and Chemical Engineering*, *28* (9):1837–1847. 2004. doi:10.1016/j.compchemeng.2004.02.036

P. Nomikos, and J. F. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, *40* (8):1361–1375. 1994. doi:10.1002/aic.690400809

J.O. Ramsay, and B.W. Silvermann. Functional Data Analysis. Springer Series in Statistics. *Biometrical Journal*, *40*

(1):56–56. 1998. doi:10.1002/(sici)1521-4036(199804)40:1<56::aid-bimj56>3.0.co;2-#

R. Rendall, L. H. Chiang, and M. S. Reis. Data-driven methods for batch data analysis – A critical overview and mapping on the complexity scale. *Computers and Chemical Engineering*, *124*:1–13. 2019. doi:10.1016/j.compchemeng.2019.01.014

R. Rendall, B. Lu, I. Castillo, S. T. Chin, L. H. Chiang, and M. S. Reis. A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes. *Industrial and Engineering Chemistry Research*, *56* (30):8590–8605. 2017. doi:10.1021/acs.iecr.6b04553

B. Schölkopf, A. Smola, and K. Müller. Kernel principal component analysis. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1327:583–588. 1997. doi:10.1007/bfb0020217

J. Wang, and Q. P. He. Multivariate Statistical Process Monitoring Based on Statistics Pattern Analysis. *Industrial & Engineering Chemistry Research*, *49* (17):7858–7869. 2010. doi:10.1021/ie901911p

H. Wang, and M. Yao. Fault detection of batch processes based on multivariate functional kernel principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *149*:78–89. 2015. doi:10.1016/j.chemolab.2015.09.018

O. Wu, A. E. F. Bouaswaig, S. M. Schneider, F. M. Leira, L. Imsland, and M. Roth. Data-driven degradation model for batch processes: a case study on heat exchanger fouling. *Computer Aided Chemical Engineering*, *43*:139–144. 2018. doi:10.1016/B978-0-444-64235-6.50026-7

O. Wu, A. E. F. Bouaswaig, L. Imsland, S. M. Schneider, M. Roth, and F. M. Leira. Campaign-based modeling for degradation evolution in batch processes using a multiway partial least squares approach. *Computers and Chemical Engineering*, *128*:117–127. 2019. doi:10.1016/j.compchemeng.2019.05.038