Data Pre-processing and Sensor-Fusion for Multivariate Statistical Process Control of an Extrusion Process

Frank Westad
frank.westad@idletechs.com
Idletechs AS
Trondheim, Trøndelag, Norway
Nowegian University of Science and Technology. Department of Engineering Cybernetics
Trondheim, Trøndelag, Norway

Lars Lodgaard
lars.lodgaard@benteler.com
Benteler Automotive
Raufoss, Innlandet, Norway

Torbjørn Pedersen
torbjorn.pedersen@idletechs.com
Idletechs AS
Trondheim, Trøndelag, Norway

ABSTRACT

In most manufacturing processes, data related to a product are collected across several process steps. Ensuring good data quality is essential for subsequent process modeling, monitoring, and control. Although data for a given process might already be available in digitized form in the process control systems or industrial databases, it is in most cases not so that the data can directly be used in its original form for process modeling. Pre-processing is often needed before modeling, which may include operations such as time alignment by handling different sampling frequencies and lag time, handling of missing values, and detection of sample outliers. Specific considerations must be made for processes with both continuous and batch process steps due to different data structures. This paper describes an industrial use case for extrusion monitoring starting from structured raw data and ending up with real-time multivariate statistical process control (MSPC) applying a sensor-fusion approach and feature extraction. The MSPC also enables in-depth analysis for identifying process variables in the case of samples lying outside of the normal operating conditions (NOC).

CCS CONCEPTS

• **Applied computing** → *Engineering*.

KEYWORDS

Data pre-processing, time alignment, sensor fusion, PCA, MSPC, outlier detection

ACM Reference Format:
Frank Westad, Lars Lodgaard, and Torbjørn Pedersen. 2023. Data Pre-processing and Sensor-Fusion for Multivariate Statistical Process Control of an Extrusion Process. In *Proceedings of the 3rd International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things (SEA4DQ '23), December 4, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3617573.3618029

SEA4DQ '23, December 4, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0378-2/23/12.
https://doi.org/10.1145/3617573.3618029

1 INTRODUCTION

In most industrial processes, data related to one product are collected across several process steps, including the characteristics of the raw materials. Although most data are present in a supervisory control and data Acquisition (SCADA) system, it is not given that the data are directly suitable for modeling. Optimal data preprocessing is often application-specific and cannot be done by pushing a button. Some reasons for this are missing values, uneven sampling frequencies, (varying) time lags between process steps, different sensor noise, and different dimensions of the data. Missing values can be handled by use of various methods for imputation [5], however, this operation should not be done unsupervised, especially in the case of weak correlation in time as well as between process parameters as it may introduce artifacts in the pre-processed data. Missing values can also be handled directly in multivariate methods such as PCA (see 2.7), by using e.g. the NIPALS algorithm [4]. The various process steps can have a combination of continuous, semi-batch, and batch behavior.

The overall objective in sensor fusion for industrial applications depends on the type of process and sensors. This article presents a sensor fusion approach with the aim of developing a holistic model for process monitoring that starts with raw materials and ends with critical quality attributes (CQA). This can be quantitative prediction or classification into e.g. pass/fail for the final product and, in some cases, the intermediate product. In this context, Process Analytical Technology (PAT) has been adopted as a generic methodology, especially in the pharmaceutical industry, but also in biopharma and polymer industries [1]. The idea behind PAT and Quality by Design (QbD) is to ensure that all critical process parameters (CPP) lie inside the specifications throughout the whole process. Ideally, there should be no need to perform quality control on the produced units if the process is under control at every step, or reduce this to a minimum as obtaining the CQAs is often a tedious task involving laboratory work. In this context, PAT and QbD are closely related to Zero Defect Manufacturing (ZDM) [2]. The concepts deal with controlling the process with so-called quality gates after critical process steps. According to Fragapane *et al*, in ZDM a majority of the published articles deal with the detection of defects (60%) or prediction (24%), whereas only a few articles deal with prevention, repair, and mitigation. As such, it seems there is still some way to go for holistic control of industrial processes given changes in raw materials and other uncontrolled sources of variation.

There exist many approaches for how to model the various steps in multi-step processes. Depending on the scientific disciplines one may call it sensor fusion and the actual models for multiblock models. Assuming that variables from two or more process steps are present in the data, several methods have been evaluated over the years [10], [6], [7]. This gives rise to several challenges, how to simultaneously model both continuous data and batch data; how to cope with time lags between process steps, which might be more or less constant; or how to combine univariate sensors with multichannel sensors and data such as spectroscopy measurements, time series represented in the frequency domain, and batch process trajectories of uneven length [9].

The aim of this article is to present important considerations in sensor-fusion and real-time Multivariate Statistical Process Control (MSPC) through an industrial use case.

2 MATERIALS AND METHODS

2.1 Data

As an example of data pre-processing, modeling, and real-time monitoring, a simplified use case including two process steps from an extrusion process at Benteler, Norway, is selected. In these steps a billet is pre-heated in a Permanent Magnet Heater (ZPE) before it is pressed into a profile in the extrusion press, this can be represented in 1. Both the heating process and the press are batch processes, with a duration of approximately two and four minutes, however, in the selected dataset the settings for the heaters were always fixed.

The effect of the heating is represented by distributed temperature measurements of the billets prior to loading onto the extruder. Every second billet is heated using alternating ZPEs installed in the production hall. The total energy during heating in kWh is included among the variables. All variables related to the extrusion and the exit temperature have a timeline trajectory and should ideally be analyzed as a 3D data structure (batch, time, and variable). However, by graphical inspection of these variables, they are almost constant within the time interval for the extrusion of individual billets. Therefore, only the stem pressure is considered as a time series in the pre-processing and analysis steps. See Appendix A.4 for the full list of variables and their abbreviations.

2.2 Matching the Various Sets of Variables

A time-matching algorithm was employed as there is a time lag between the two process steps. The available raw data for the temperatures after heating the billet were not given with billet IDs but by time stamps for the time of measurement. Continuous time series for the heating process itself was not available for analysis. As the extrusion data had both billet IDs and time stamps per second, matching the time stamps including an offset between billet loading and the start of the extrusion enabled data alignment. The sensor for the extruder exit temperature was positioned approximately two meters after the end of the extrusion compartment. Given the speed of the ram, this allowed time alignment from the continuous time stamps for subsequent analysis.

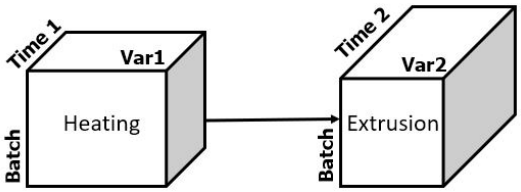

Figure 1: Schematic view of the heating and extrusion process steps

As these sensor data were acquired every second, time stamp matching was employed by finding the nearest point of time. A maximum offset in terms of seconds was employed to avoid spurious time alignment in the time series due to e.g. a change in the billet alloy or product on the extrusion line.

2.3 Aligning Extruder Press Curves

A function for aligning the stem pressure curves was implemented. The function calculates the difference between adjacent points in time to detect the start of a new cycle from the raw data. An example of the raw data is shown in Figure 2. As seen, the time gaps between the cycles, as well as the duration, are not constant. The function detects a new cycle when the difference between two consecutive points in time exceeds a given threshold. Another approach that was evaluated was to search for the maximum correlation between segments of the raw data based on an initial splitting. However, due to the rapid increase in pressure in the first seconds, this did not produce well-aligned data.

Stem pressure for selected samples after applying the procedure for finding the start and end of the individual cycles and reordering them into a data table for the two sample sets is shown in Figures 2 and 3. The duration of the press cycles varied from 240 to 300 seconds, with a one-second sampling rate. The most interesting part of the curve is at the start of the cycles, therefore the length of the curves was set to 220 points in time to avoid missing data. One could always time-warp the extrusion curves to a common length, however, this would introduce unwanted artifacts.

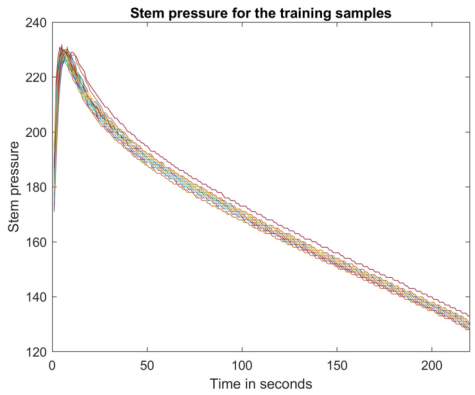

Figure 2: Stem pressure after alignment of the samples with normal trajectory

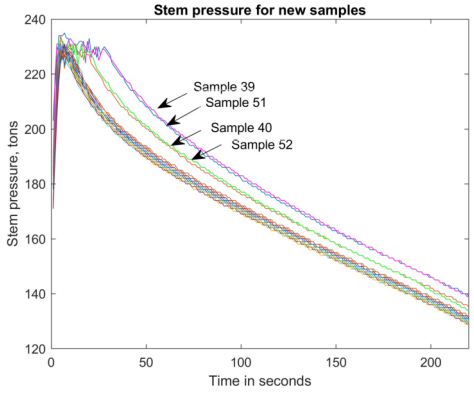

Figure 3: Stem pressure after alignment of the new samples

2.4 Feature Extraction

The aligned curves can be analyzed with multivariate methods directly or after extracting features from the curves. However, to improve the robustness of the following analysis, a layer of feature extraction is added. After investigation of the press curves as shown in Figure 2, the following features were derived:

- Standard deviation
- Maximum value
- Index of max value
- Sum of differences
- Max value of differences
- Number of positive differences
- Entropy

The derived features from the stem pressure curves are listed as variables 20-26 in Appendix A.4.

2.5 Final Sample Set and Selected Variables

Of 65 initial billet IDs, matching data were found for 61 billets after time alignment. Thus, the data available for the analysis consisted of 59 billets ("samples") from the extrusion process. For each of the individual time stamps the billet heating, time series per second for the extrusion process, and exit temperature are extracted during the pre-processing as described above, the variables are listed as numbers 15-19 in the table in Appendix B. The raw and derived temperature measurements of the billet after heating and the total energy during heating are listed as numbers 1-15. The delta from the set points is chosen rather than the actual temperatures for the sensors placed along the billets. A typical billet length is 1.2 meters.

2.6 Modeling Strategy

To demonstrate modeling and monitoring of industrial processes, 38 samples were assigned for establishing the model (training set) whereas 21 samples were assigned as a test set for monitoring. An initial analysis was performed on all samples so that the 38 samples selected for modeling were inside the normal operational range, thereby establishing the "basic" model for future projection and identification of out-of-specification situations. When starting from unformatted raw data, the overall pipeline for this use case is as follows:

(1) Retrieve data from the SCADA system for all relevant process variables, ref. Figure 1
(2) Match the various sets of variables using available categorical information (billet IDs), timestamps and known time lags between the measurements
(3) Align the press curves to have a common start point and length
(4) Extract features from the press curves
(5) Perform time-matching between the sets of variables
(6) Calculate the mean value for all other process variables that are observed per second in the extrusion cycles
(7) Combine all process variables and features in one table
(8) Develop a model on "good" batches/runs to represent the normal operating conditions (NOC)
(9) Project new samples onto the model and detect deviations for individual samples as well as variables for on-line MSPC

2.7 Methods

2.7.1 Principal Component Analysis (PCA). PCA is a multivariate method for decomposing a data matrix into underlying latent variables or principal components (PCs). [3] The criterion is to maximize the variance for the direction of each PC. Although PCA in itself is a simple mathematical operation and in some scientific communities merely used for noise filtering and yielding an orthogonal basis for further purposes, the underlying latent variables are often interpretable when applying domain knowledge. The general form of the PCA model is:

$$\mathbf{X} = \mathbf{T}_A \mathbf{P}_A^T + \mathbf{E} \tag{1}$$

The columns in $\mathbf{T}$ are called scores and can be visualized as a "map" of the samples. The score vectors are simply linear combinations of the original variables for the various principal components. The rows in $\mathbf{P}$ are called loadings and represent how the variables contribute to each principal component. The loadings can be visualized for interpretation of the partial correlations between all variables in the so-called correlation loading plot. Once a PCA model has been established on a training data set, a new sample x_{new} may be projected onto this model, giving the projected score for each component a:

$$\hat{t}_{new,a} = \mathbf{x}_{new}\mathbf{p}_a \tag{2}$$

The projection of a new sample is the basis for detecting out-of-control situations in MSPC. Critical limits for outlier detection can be estimated once a model has been established. For multivariate methods such as PCA, one distinguishes between two types of outliers: i) Inside the model space, ii) In the residual space. The so-called Hotelling's T^2 statistic is applied for identifying outliers of the first type whereas residual analysis is performed with the use of Q- or F-residual statistics. Given these limits, samples that deviate from the normal situation can be detected.

2.7.2 Multivariate Statistical Process Control (MSPC). The principle of Multivariate Statistical Process Control (MSPC) is to utilize the strength of analyzing many process parameters simultaneously [3].

It is known that applying individual SPC limits for many parameters will not detect outliers in a multivariate context, i.e., one will fail to detect process anomalies in some scenarios. The use of multivariate methods also produces "maps" of both the sample and variable space that give information about the similarities of the samples and (partial) correlation of the variables. In addition, critical limits for detecting outliers can be estimated, both within the model space and with respect to residuals. The most applied method for MSPC is Principal Component Analysis (PCA). A short description of PCA is given in Appendix A.1.

2.7.3 Sensor fusion and multiblock modeling. As mentioned in the Introduction, various approaches to sensor fusion and subsequent multiblock modeling can be relevant, depending on the structure of the input data and the target objective. In this study, we will not consider various alternatives for finding the unique and common information in the two blocks. The main objectives are to present options for how to handle the different number of variables in the blocks and how to represent the time series for each extruder cycle. Of special interest is the extrusion stem pressure trajectory. As mentioned above, one could simply keep 220 points of time, however, having many more variables for the press trajectory than all the individual variables combined will need some kind of block-weighting. The common approach is to give all blocks of variables the same possible impact in the model by weighting the blocks by the square root of the number of variables. However, there are other aspects to consider, e.g., if the various sets of variables should be scaled to unit variance or not. The extrusion curves variables have the same unit and might not need scaling but only to be mean-centered, whereas the variables representing the billet temperature are given both in the original unit and the delta temperature, thus scaling to unit variance is the best option.

3 RESULTS AND DISCUSSION

The PCA model for the samples defined as normal based on the extrusion curves gives scores and correlation loadings as shown in Figures 4 and 5. Based on the cross-validated explained variance and interpretation of the correlation loadings, a total of six principal components were needed to capture the information in the data. This summed up to 78% of the total variance. Interpretation of the individual PCs revealed that the various subsets of variables span different underlying dimensions. E.g., the derived variables from the extrusion curves do not correlate with the billet heating temperature variables. Adding meta-information to the data is important for the interpretation of the model and increased understanding of the process. This is typical qualitative information such as batch number, sensor, production line, and equipment that can be represented as categorical information. However, one should be careful in adding these variables as input for modeling for example by one-hot decoding them into 0/1 variables.

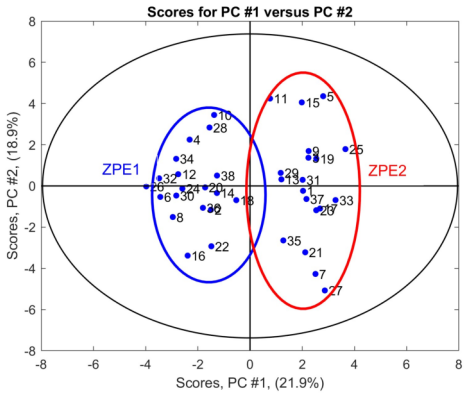

Figure 4: Score plot for PC1 vs. PC2

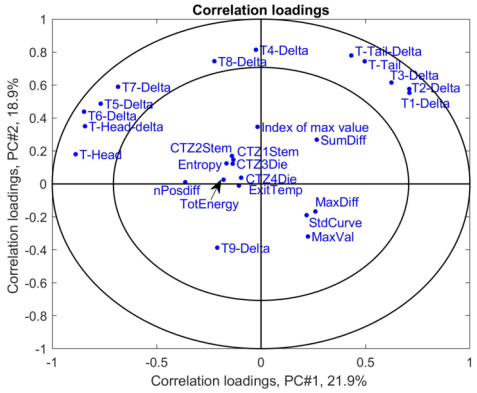

Figure 5: Correlation loadings plot for PC1 vs. PC2

The reason is that one cannot include e.g., a 0/1 numerical variable for production line and make the model "better" if line A performs better than line B. However, concluding that there is a difference might lead to increased knowledge, troubleshooting, and optimization of this product line. An example of this is given in Figure 4, where the samples are grouped according to which magnetic heater (ZPE) was in use. As can be seen, there is a systematic difference. The variables that represent these groups can be identified in the corresponding correlation loading plot. The variables pertaining to the billet temperature are grouped, which shows that the ZPEs are not heating the billets uniformly.

To evaluate the detection capability of the model, the 23 samples defined as a test set were projected onto the model to illustrate a real-time performance as well as outlier detection and for drilling down to identify anomalies. A combined plot of Hotelling's T^2 and the residual statistics with critical limits at the 99% level is shown in Figure 6. Samples inside the lower left rectangle lie inside the multivariate critical limits on both criteria. As can be seen, four samples lie outside one or two of the critical limits. An alternative visualization of the two-dimensional plot in Figure 6 is a plot over time for the two outlier statistics as shown in Figure 7, which is relevant for time-dependent processes. This visualization can

supersede individual control charts while maintaining the chosen significance level due to the multivariate approach [3].

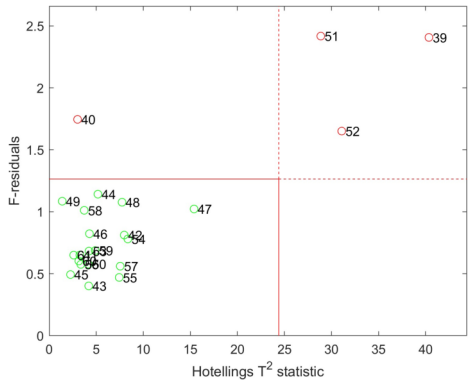

Figure 6: Influence plot for new samples

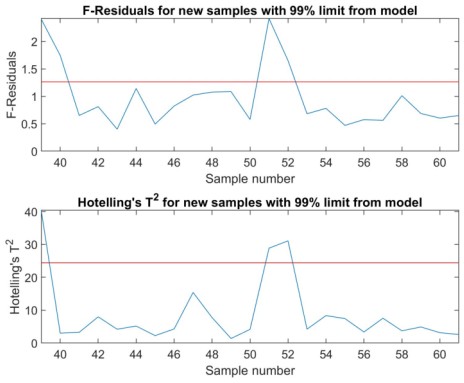

Figure 7: New samples and critical limits at the 99% level

The next step in the visualization of outlying samples is shown in the so-called contribution plot, Figure 8, see section A.3 in the Appendix for details. This plot, shown for sample 39, identifies the variables contributing to the Hotelling's T^2 statistic for a particular sample. As can be seen, the variables extracted as features from the extrusion curves are the ones of interest. This corresponds well with the visual interpretation in Figure 3. A plot of the individual residuals is shown in Figure 9. In addition to that the pattern for some of the extracted features deviates from the samples in the training set, and the value for the total energy during heating is higher than expected. This was confirmed by inspecting the raw data.

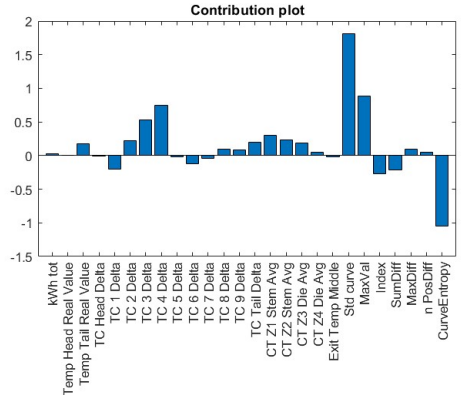

Figure 8: Contribution plot for new sample 39

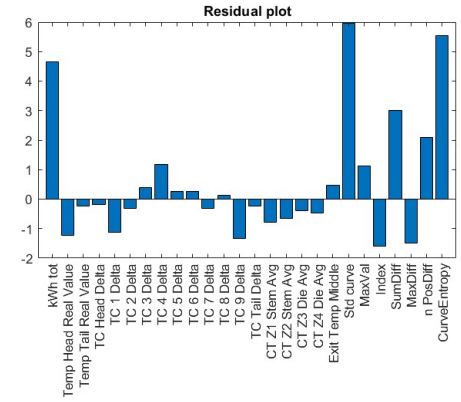

Figure 9: Residuals for new sample 39

The approach in this article can be generalized to most multi-step processes, where there are different time- and position-dependent variables that need to be aligned to enable robust process control or prediction.

The presented approach for aligning data based on time stamps does not take into account outliers or transients in the time series. Thus, some filtering of the raw data might be beneficial for making this unsupervised procedure more robust. There is a wide range of potential methods for sensor fusion and process monitoring, both within and outside of multivariate methods. The choice of methods depends heavily on the properties of the data and the domain challenges. Most processes can be handled using simple linear methods, such as PCA [3] or ICA [8]. In some settings, non-linear kernels or deep neural networks can improve performance. However, the authors of this article would strongly advocate for methods that enable detailed interpretation and identification of important variables for anomaly detection, and that allow real-time evaluation of model fit.

4 CONCLUSION

Ensuring good data quality is essential as a basis for process modeling, monitoring, classification, prediction, and closed-loop control.

Pre-processing the data involves several steps and requires process knowledge at the sensor and modeling level. Ideally, the person responsible for gathering the data should also be involved in the analysis and interpretation.

This case study presents a procedure of how to pre-process data prior to Multivariate Statistical Process Control for real-time process monitoring. Once the data structure has been defined and detailed knowledge about the sensors, time lags, and the data structure is acquired, the procedure can be applied unsupervised in real-time for in-depth analysis of new samples down to the individual variables. Feature extraction can be applied as a compressed way of representing trajectories in batch processes as an alternative to multi-block models. The outlier detection in both the model and residual space enables a single plot for troubleshooting regardless of the number of variables in the model. Furthermore, drilling down into plots of individual variables gives detailed information on why one specific sample lies outside the normal operating conditions (NOC).

The main contribution of this paper is to outline important considerations for sensor fusion and process analysis in ZDM and process monitoring use cases. The paper addresses holistic modeling of a multi-step extrusion process using transparent and inspectable machine-learning methods that allow for robust detection results.

ACKNOWLEDGEMENTS

The research described in this paper is supported by funding from the DAT4.ZERO (grant no. 958352). The authors would like to thank the European Commission for the support that made it possible to carry out this study. We also want to thank Benteler Automotive for the use of their data.

A APPENDICES

A.1 Outlier Statistics in the Model Space for PCA

The Hotelling's T^2 statistic is a multivariate generalization of the Student t-test. The form of the Hotelling's T^2 statistic for methods that are based on principal components is as follows.

$$Hotelling's\ T_i^2 = (\sum_{a=1}^{A} t_{ia}(\mathbf{t}_a^T\mathbf{t}_a)^{-1}t_{ia}^T)/(I-1) \tag{3}$$

, where:

t_a is the score vector for principal component a
t_{ia} is the score value for sample i for principal component a

The Hotelling's T^2 statistic is approximately F-distributed:

$$F_{A,I,\alpha} \sim T^2\frac{(I-A)}{A(I-1)} \tag{4}$$

where I is the number of training samples

A.2 Outliers in the Residual Space for PCA

While the Hotelling's T^2 statistic detects if samples are extreme within the model space, the X-residual statistic detects samples that have a deviating pattern for the variables. The residuals are calculated as follows:

$$e_{ik,A} = x_{ik} - \sum_{a=1}^{A} t_{ia}\hat{p}_{ka}^T \tag{5}$$

The sample residuals, $F_{i,A}$, are the mean of the squared residuals of all variables for each sample after A principal components:

$$F_{i,A} = (\sum_{i=1}^{I} {e_{ik,A}}^2)/K \tag{6}$$

The critical limits for the residuals are calculated from the standard F-distribution:

$$F_{crit} \sim (\alpha, 1, I_{training}) \tag{7}$$

A.3 Contribution Plot

Furthermore, the contribution plot gives detailed information about which variables have changed when an outlying sample has been identified[10]The contribution values per variable are the individual contributions to the Hotelling's T^2 statistic. This often gives important insight into the causality of a process. The calculation for a new sample is shown in equation 8 For more details see Jackson

$$C_{ik} = \sum_{a=1}^{A} S_{aa}^{-1}\hat{t}_{new,a}x_{new,ik}p_{a,ik} \tag{8}$$

A.4 List of variables

Table 1: List of variables and their units

Number	Full name	Abbreviation	Unit
1	Total energy in ZPE heater	TotEnergy	kWh
2	Temp Head of billet	T-Head	°C
3	Temp Tail of billet	T-Tail	°C
4	TC Head Delta	T-Head-delta	°C
5	TC 1 Delta	T1-Delta	°C
6	TC 2 Delta	T2-Delta	°C
7	TC 3 Delta	T3-Delta	°C
8	TC 4 Delta	T4-Delta	°C
9	TC 5 Delta	T5-Delta	°C
10	TC 6 Delta	T6-Delta	°C
11	TC 7 Delta	T7-Delta	°C
12	TC 8 Delta	T8-Delta	°C
13	TC 9 Delta	T9-Delta	°C
14	TC Tail Delta	T-Tail-Delta	°C
15	ContainerTempZone1StemSideTop	CTZ1Stem	°C
16	ContainerTempZone2StemSideBottom	CTZ2Stem	°C
17	ContainerTempZone3DieSideTop	CTZ3Die	°C
18	ContainerTempZone4DieSideBottom	CTZ4Die	°C
19	Extruder exit temperature	ExitTemp	°C
20	Std. deviation of extrusion curve	StdCurve	Tons
21	Maximum value of extrusion curve	MaxVal	Tons
22	Index of maximum of extrusion curve	Index of max value	N/A
23	Sum differences of extrusion curve	SumDiff	Tons
24	Maximum difference of extrusion curve	MaxDiff	Tons
25	No. of positive differences of extrusion curve	nPosdiff	N/A
26	Entropy of extrusion curve	Entropy	N/A

REFERENCES

[1] Selda Dogan Calhan, Ebru Derici Eker, and Nefise Ozlen Sahin. 2017. Quality by design (QbD) and process analytical technology (PAT) applications in pharmaceutical industry. *European Journal of Chemistry* 8, 4 (dec 2017), 430–433. https://doi.org/10.5155/eurjchem.8.4.430-433.1667

[2] Giuseppe Fragapane, Ragnhild Eleftheriadis, Daryl Powell, and Jiju Antony. 2023. A global survey on the current state of practice in Zero Defect Manufacturing and its impact on production performance. *Computers in Industry* 148 (2023), 103879. https://doi.org/10.1016/j.compind.2023.103879

[3] J Edward Jackson. 2005. *A user's guide to principal components*. John Wiley & Sons.

[4] Philip R.C. Nelson, Paul A. Taylor, and John F. MacGregor. 1996. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems* 35, 1 (1996), 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

[5] Sagar Sen, Erik Johannes Husom, Arda Goknil, Dimitra Politaki, Simeon Tverdal, Phu Nguyen, and Nicolas Jourdan. 2023. Virtual sensors for erroneous data repair in manufacturing a machine learning pipeline. *Computers in Industry* 149 (2023), 103917. https://doi.org/10.1016/j.compind.2023.103917

[6] Thomas Skov, Davide Ballabio, and Rasmus Bro. 2008. Multiblock variance partitioning: a new approach for comparing variation in multiple data blocks. *Analytica chimica acta* 615 1 (2008), 18–29.

[7] Chudong Tong and Xue feng Yan. 2017. A Novel Decentralized Process Monitoring Scheme Using a Modified Multiblock PCA Algorithm. *IEEE Transactions on Automation Science and Engineering* 14 (2017), 1129–1138.

[8] Chudong Tong, Ahmet Palazoglu, and Xuefeng Yan. 2014. Improved ICA for process monitoring based on ensemble learning and Bayesian inference. *Chemometrics and Intelligent Laboratory Systems* 135 (2014), 141–149. https://doi.org/10.1016/j.chemolab.2014.04.012

[9] Frank Westad, Lars Gidskehaug, Brad Swarbrick, and Geir Rune Flåten. 2015. Assumption free modeling and monitoring of batch processes. *Chemometrics and Intelligent Laboratory Systems* 149 (2015), 66–72. https://doi.org/10.1016/j.chemolab.2015.08.022

[10] Johan A. Westerhuis, Theodora Kourti, and John F. Macgregor. 1998. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics* 12 (1998).

Received 2023-08-04; accepted 2023-08-24