Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 104 (2021) 1334–1338

www.elsevier.com/locate/procedia

54th CIRP Conference on Manufacturing Systems

An approach to data structuring and predictive analysis in discrete manufacturing

Christian Dalheim Øien[a,]*, Sebastian Dransfeld[a]

[a]*Sintef Manufacturing, Enggata 40, 2830 Raufoss, Norway*

* Corresponding author. Tel.: +47 977 36 238. *E-mail address:* christian.dalheim.oien@sintef.no

Abstract

In discrete manufacturing the variation in process parameters and duration is often large. Common data storage and analytics systems primarily store data in univariate time series, and when analysing machine components of strongly varying lifetime and behaviour this causes a challenge. This paper presents a data structure and an analysis method for outlier detection which intends to deal with this challenge, as an alternative to predictive maintenance which often requires more data with higher quality than what is available. A case study in aluminium extrusion billet manufacturing is used to demonstrate the approach, predominantly detecting anomalies at the end of a critical component's lifetime.

Keywords: Anomaly Detection, Predictive Maintenance; Discrete Manufacturing; Big Data Analytics; Adaptive Self-learning Systems

1. Introduction

Today, production areas within a manufacturing company may well continuously collect thousands of data variables from a large variety of machines performing incoherent processes where faults and breakdowns occur sporadically. The availability of various sensors is increasing, and lifecycle management of machinery is increasingly offered by equipment suppliers [1]. Based on this level of data availability predictive maintenance (PdM) has been a popular research field for the last few decades with the goal to ensure machine function and avoid breakdowns by making predictions of features based on those variables [2]. However, such predictions are difficult and require a large amount of precise data, as well as an accurate model describing the relevant process and possible failure modes which is often laborious to obtain. The same is valid for estimations of remaining useful life for an asset or component [3]. For certain applications and machinery, like continuously running pumps with stable load and other boundary conditions, typically found in process manufacturing, such approaches are indeed viable. In addition to boundary condition stability, the number of possible failure modes and mechanisms should be low, a high amount of historical sensor data should be available with adequate quality, along plenty of failure logs well documented by detailed predefined forms [4, 5]. For more complex processes, however, typically involving discrete manufacturing and where perhaps none of the previously mentioned requirements are met, PdM is not likely to be the best way to ensure machine function and avoid failure. The multitude of relevant physical interactions of a complex manufacturing process is too big and predominantly unknown Therefore, based on the amount of data available in a given use case, it may not be possible to obtain an overview of possible failure modes and mechanisms necessary to define relevant features to predict. Furthermore, it may even not be feasible or possible to measure additional variables that those features would need to be based on. Similarly, simulation models coupled with the actual system, as a digital twin, can be a viable approach [6], but also this is difficult to achieve in many cases and should have a high threshold for application.

On the other hand, a discrete manufacturing process which seems incompatible with PdM could well be supervised by the

10.1016/j.procir.2021.11.224

application of anomaly or outlier detection in available data. In this paper we present a method based on outlier detection that can be automated and applied across a variety of discrete manufacturing processes to avoid loss of function, unwanted behaviour, and breakdowns, without ad-hoc modelling and a minimum of adaption. The method is based on univariate statistics and a data-driven description of normal behaviour, as opposed to model-based methods. The main compromise with such a supervision method is the absence of causality interpretation. The main benefits, however, are applicability without detailed historical logs or process-specific models or adaption, and transferability between a variety of processes.

The method described in this paper is intended to be applied on discrete manufacturing processes where:

- all possible failure modes and their various failure mechanisms are not fully mapped
- which specific features that can be used to predict these failure mechanisms are unknown.
- complete and detailed logs describing incidents or failures based on predefined forms are not available

1.1. State of the art

Typically, maintenance management strategies are grouped into the following three categories with various complexity [7].

Run-to-Failure (R2F) maintenance is characterized by only performing maintenance when a loss of function or failure is detected. Typically, but depending on the manufacturing processes and set-up, the cost of interventions and associated downtime after failure are usually higher with R2F than with the two following categories.

Preventive Maintenance (PvM) means performing maintenance according to a planned schedule based on time or number of process iterations. This is also referred to as scheduled maintenance. With PvM machine failures are sometimes prevented, but also unnecessary maintenance is performed.

Predictive Maintenance (PdM), also referred to as Condition-Based Maintenance [8], uses predictions or estimates of process features that can be linked to the health status of a piece of equipment [2, 9]. PdM systems enable detection of failures or loss of function ahead in time. Similarly, to PvM, this can also lead to unnecessary maintenance. Prediction tools in PdM use historical data together with ad hoc defined health models, maps of possible failure modes and mechanisms, statistical inference methods, and engineering approaches. Usually, large amounts of accurate data accompanied with detailed metadata and logs of earlier incidents are required to successfully implement PdM, and this is a limiting factor for its application.

Direct usage of anomaly or outlier detection systems, without connection to specific failure modes, is typically not regarded as part of any of the three maintenance strategies above. Anomaly detection in manufacturing systems has been a popular research field in the past decade, similarly to PdM, especially regarding usage of machine learning techniques [1, 10-13]. There seems to be a focus on applying such techniques in order to detect complex patterns on one side, as well as to cope with high dimensionality and insignificant variables. This article serves as a simpler alternative to such approaches.

2. Theory – outliers and anomalies

This article is related to detecting anomalies in manufacturing data in the form of simple statistical outliers.

Frank E. Grubbs [14] described a statistical outlier as an observation that "appears to deviate markedly from other members of the sample in which it occurs". In this perspective, it is of interest to define such observations as either "an extreme manifestation of the random variability inherent in the data", in which case the observation can be regarded as valid, or as "a result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value", in which case the observation should be disregarded since it does not represent a correct measurement of the investigated statistic.

Similarly, according to Charu C. Aggarwal [15], we can distinguish between outliers as either noise or as a "special kind of outlier that is of interest to an analyst". In the following, an outlier is regarded as a data point deviating statistically from the rest of the dataset seen as a statistical population of independent measurements. In this perspective, any attribute of the data point is not considered, such as its order or time stamp. An anomaly, on the other hand, is regarded as a data point deviating from an expectation based on a certain model of the dependency of one or more attributes of the data point, such as its time stamp.

The described differentiation between outliers and anomalies is visualised by an example in Fig. 1, showing a small data set generated by the authors for this purpose. Here, each data point represents a measured value that has an attribute value x (which could for instance be a time stamp.) In this sense, the marked anomalous data point is not an outlier as measured, but it can become an outlier in a specific context, i.e., when an expected mean from a model (in this example sin $2\pi x$) based on the attribute x is subtracted from the dataset. The red lines show the mean and ±3 standard deviations to the mean.

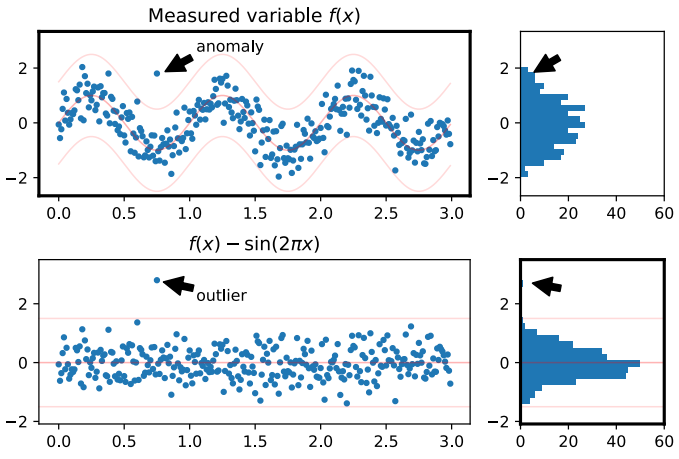

Fig. 1. An example of an anomaly in the form of a significant deviation from a model-based expectation (top left) and an outlier in the form of a statistically deviating data point with zero dimensionality (bottom right).

3. Method

The proposed method is based on converting process cycle timeseries into a feature table where one row represents a single completed process cycle, and the columns contain a set of features describing each cycle independently of its duration. This reorientation is commonly referenced to as feature extraction [16]. Furthermore, feature selection is applied in order to detect anomalies among the described process cycles as statistical outliers. The method is meant for processes or equipment where a given critical component is subject to an R2F or PvM repair or replacement strategy, and data quality is not sufficient to successfully implement PdM, by linking process cycles with the age of that critical component. In this perspective, the goal of the anomaly detection is to alert the manufacturing operations team to evaluate the process in more detail and decide if an early maintenance intervention should be carried out. The method is based on the assumption that a statistics-based description in the process cycle domain via features will yield more accurate predictions than similar models formed in the time domain.

3.1. Data integrity

The input to a feature extraction pre-processing as described above is raw data in the form of a time series. There are typically three causes of anomalies in such raw data, whereas only one of them is of interest:

- Human interaction or intervention of the production process
- Data logging system failure
- Actual process discrepancies

Naturally, only the latter type is of interest for process analysis. In order to make actual anomalies prominent, it is important to ensure integrity of the data. The best way to do this in the case of discrete manufacturing processes is to avoid univariate continuous time series of logged sensor or run-time variables [17]. Oppositely, ensuring a column-based and process cycle-ordered formatting will ensure that logged process parameter values, metadata and measured process variables are aligned with each other, and therefore unwanted anomalies due to actions such as aborted process cycles are much easier to detect. In addition to ensuring integrity, this also greatly reduces the needed time for data cleaning and pre-processing. As a minimum, a cycle counter or process parameter indicating active process should be part of the dataset, so that data stemming from active process can be correctly filtered out.

An example of raw data as a continuous time series is shown in Fig. 2, taken from the industrial use case described in section 4. The example shows data covering two process cycles, but also the time in between cycles. The largest signal values are in sequence state 1, where the process is idle, and therefore presumably irrelevant to the function of the component. When only data from active process is interesting for analysis, the presence of this process sequence parameter in the dataset is crucial for a correct pre-processing of the data.

Whether or not a laborious effort is necessary to obtain data where anomalies can be expected to be due to actual process discrepancies, the following description of the method assumes that such a set of raw data is available, constituting separate (or separable) process cycles. Additionally, replacement or maintenance actions on the critical component must be visible in the data, but the reason for the replacement or maintenance can be unknown.

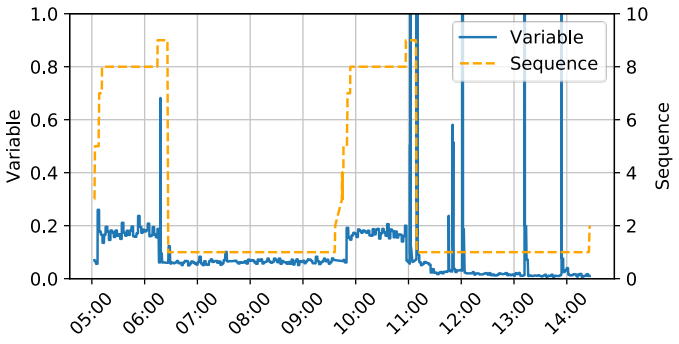

Fig. 2. An example of discrete manufacturing process data in the form of a measured variable and a process sequence parameter which reflects the sequential states of the process as an integer value.

3.2. Feature extraction

Extracting features from raw data and preparing the dataset for the proposed method can be summarized by the following steps:

- For each relevant variable or parameter in the dataset, define a list of numerical features that together would describe that variable's time series for a given process cycle. Typical features would be statistical descriptors of the values within the entire cycle or separate process sequences (if each cycle is constituted of a number of separate process sequences), analysis results based on a simple regression method over the cycle, and descriptors based on sliding windows within one or more sequences or the entire cycle.
- Be sure to align the set of variables and features with available domain knowledge, understanding of process quality, and avoid strongly interdependent features and statistical deficits such as homoscedasticity, autocorrelation, and multicollinearity.
- Now iterating through separate process cycles, and then trough relevant variables and parameters, calculate each of the defined features and store the results in a table with columns according to the list of features and one row per production cycle.
- For each process cycle, calculate the following three additional features:
  - The counter $C \in \mathbb{N}$ equal to the number of completed process cycles since the critical component was replaced or maintained.
  - The maximum value C_{max} of C for the relevant component lifetime.
  - The relative component age $c = C/C_{max}$.

This way process cycles are intended to be compared based on the age c of the critical component, as opposed to e.g., temporal proximity.

3.3. Statistical description of normal behaviour

The proposed method for detecting anomalies in the acquired feature table is based on defining normal behaviour by fitting feature values to a univariate probability distribution. Specifically, rows where $0.1 < c < 0.95$ are regarded as examples of normal behaviour. For each feature, all values from this sub-set are taken as independent measurements and fitted to a selected type of probability distribution (e.g., the Gaussian distribution). The relative component age thresholds of 0.1 and 0.95 for normal behaviour are meant to rule out slightly abnormal behaviour in the first period after replacement or maintenance of the critical component, and to rule out feature values that may change due to significant deterioration, respectively.

3.4. Feature selection and anomaly detection

Given the set of univariate probability distributions and a corresponding significance level describing expected or normal behaviour of each feature, the proposed method is to define criteria for automatic feature selection. Then, based on a selected sub-set of features, anomaly detection can be applied in the form of statistical outlier detection based on said significance level. Two types of such criteria are suggested, where features are selected

- such that the feature sub-set covers detected outliers in component lifetimes that ended in a failure, corresponding to a supervised (or semi-supervised) approach.
- based on ordering all features by the proportion of detected outliers to normal process cycles at the end of the component lifetime (e.g., where $c > 0.95$), corresponding to an unsupervised approach.

Both approaches of feature selection enable automatic adaption of the proposed method to new datasets. In the case of labelled data, the significance level can be determined e.g. by maximizing the F_β-score of the yielded prediction [18, 19]. When applied on unlabelled data, however, the significance level must be set based on relevance to process control system, domain knowledge, or simply tuned according to a desired frequency of anomalies in the dataset.

4. Results and analysis

4.1. Method applied in industrial use case

The method described has been applied on a real dataset from an aluminium extrusion billet casting process. The studied critical component, maintained using a PvM policy, is a rotating component that is part of a machine through which the liquid metal flows. The studied process and dataset are well suited for the proposed anomaly detection approach as it shows complex mechanisms and relations, has high noise intensity, and lacks a detailed log of incidents. The approach was also preferred in this industrial use case due to its transferability to other processes.

The starting point for analysis was raw data in form of continuous univariate time series split timewise in durations of one month. Based on domain knowledge, one measured variable and two process parameters were chosen for the analysis, namely measured torque, an hour counter for the critical component (indicating replacements) and a process sequence parameter (integer). The data was combined by a full outer join and forward fill, and due to the data format and missing process cycle ordering, manual work was then conducted to extract valid time series of data corresponding to actual process cycles. Finally, series of process cycles were extracted according to replacement of the critical component. Incident logs were made available for the case study, but they were inconsistent and mostly uninformative being based on free text entry. The result was a semi-labelled dataset consisting of 74 maintenance cycles, of which 2 ended with a known component failure, and covering a total of 15 317 process cycles. It is known that more than the two mentioned maintenance cycles ended with component failure, but due to insufficient log consistency and quality it is not known which of the other 72 cycles it was.

A set of 140 descriptive features were defined based on simple statistical analysis, such as *minimum*, *maximum*, *mean* and *variance*, and linear regression results like slope and regression error, based on 10 discrete windows within each sequence. Feature extraction was conducted and the three features c, C and $C_{\max}$ described in section 3.2 were added to the table. Normal behaviour was defined as described in section 3.3 by use of Gaussian distributions. Supervised feature selection was done by selecting features containing outliers within the two specific component lifetimes with known failure. This process is visualized in Fig. 3 for one of those two lifetimes. This resulted in a subset of 10 features. Furthermore, the threshold for outlier detection was set simply according to the resulting number of outliers detected based on the selected features. As a result of low normality of the data a threshold of six standard deviations was chosen, resulting in a total of 69 anomalies among the mentioned total of 15 317 process cycles.

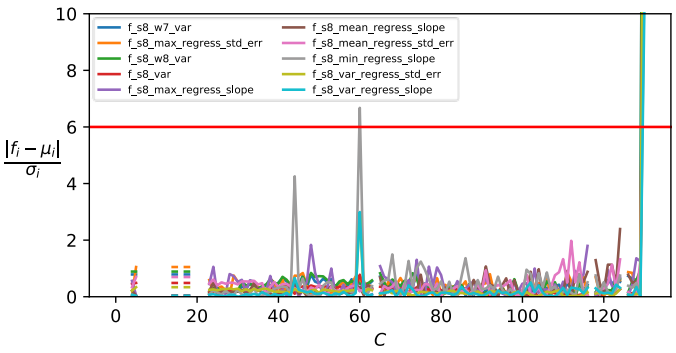

Fig. 3. A component lifetime of $C_{\max} = 131$ process cycles where 10 features f_i indicate one or more outlying cycles. The mean μ_i and standard deviation σ_i of feature i were based on occurrences where $0.1 \leq c \leq 0.95$.

4.2. Detected anomalies

Outlying feature values were detected across the entire dataset, and the originating rows were then classified as anomalous process cycles. Due to the partial labelling of the dataset, descriptors like accuracy, precision or recall cannot be computed. However, the results can be evaluated in terms of

frequency of detected anomalies as a function of component lifetime. Specifically, 36 out of 802 process cycles (4.5%) with $c > 95\%$ were classified as anomalous. These anomalies can be argued to be reasonable since a component replacement would reduce the component lifetime by less than 5% and potentially avoid failure. At the same time, 33 out of 14 515 process cycles (0.2%) with $c \leq 95\%$ were classified as anomalous. These can be regarded as false positives occurring amid the component lifetime. A histogram showing the frequency of the 69 detected anomalous process cycles as a function of c is shown in Fig. 4. These 69 process cycles cover 24 out of 74 maintenance cycles (32%).

5. Discussion

The frequency of anomalies is approximately 20 times higher for $c > 95\%$ compared to $c \leq 95\%$. This shows that the selected set of features to some extent contain information that characterize the end of the critical component's lifetime. In other words, the results above confirm that the given industrial case and dataset are applicable to the proposed method, and that it may serve as an alternative to PdM in cases where the quality of data and incident logs are insufficient. An important difference between the proposed approach and PdM would be that in the former, it is not attempted to estimate or predict the remaining useful life or health status of the component repeatedly during a portion of its lifetime, but instead it is evaluated whether each production cycle is anomalous as a possible indication that the remaining useful life may be short, and the cost of a maintenance intervention is low. By applying anomaly detection in the process cycle domain, as opposed to the domain of pure date and time, the relative age of a critical component has been considered in a simple data driven description of normal behaviour, and the resulting detected anomalies are mainly occurring at the end of the critical component's lifetime.

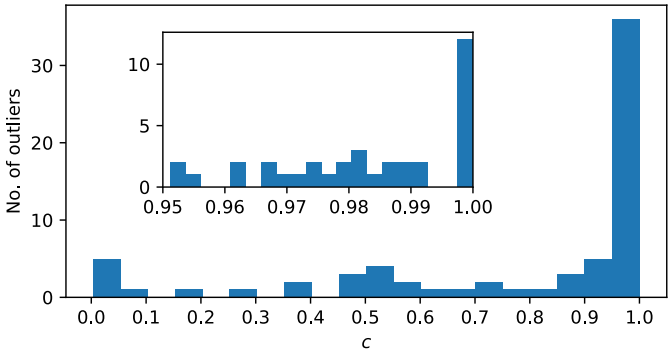

Fig. 4. An overview of the 69 detected anomalous process cycles in intervals of relative component age c.

6. Further work

Additional insight end experience with transferability of this method to different industry cases would be valuable. Specifically, the method should be tested on real datasets to see if it would detect intentionally introduced anomalies. Also, further work with the method itself would be of interest, such as evaluating alternative feature selection approaches to the ones described in section 3.4, as well as other possible methods for setting the significance level described in the same section.

Acknowledgements

We wish to thank The Research Council of Norway for their financial contribution to the project *BIA KPN CPS-Plant* which the R&D work in this article was funded by.

References

[1] D. Hendrycks, M. Mazeika, T. Dietterich, Deep Anomaly Detection with Outlier Exposure, in, 2018, pp. arXiv:1812.04606.
[2] R.K. Mobley, An introduction to predictive maintenance, Second Edition ed., Elsevier, 2002.
[3] X.-S. Si, W. Wang, C.-H. Hu, D.-H. Zhou, Remaining useful life estimation–a review on the statistical data driven approaches, European journal of operational research, 213 (2011) 1-14.
[4] C. Lee, Y. Cao, K.K.H. Ng, Big Data Analytics for Predictive Maintenance Strategies, in: Supply Chain Management in the Big Data Era, IGI Global, 2017.
[5] H. Shao, H. Jiang, Y. Lin, X. Li, A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders, Mechanical Systems and Signal Processing, 102 (2018) 278-297.
[6] P. Aivaliotis, K. Georgoulias, G. Chryssolouris, The use of Digital Twin for predictive maintenance in manufacturing, International Journal of Computer Integrated Manufacturing, 32 (2019) 1067-1080.
[7] G.A. Susto, A. Schirru, S. Pampuri, S. McLoone, A. Beghi, Machine Learning for Predictive Maintenance: A Multiple Classifier Approach, IEEE Transactions on Industrial Informatics, 11 (2015) 812-820.
[8] R. Langone, C. Alzate, B.D. Ketelaere, J. Suykens, Kernel spectral clustering for predicting maintenance of industrial machines, IEEE Symposium on Computational Intelligence, Data Mining (2013) 39-45.
[9] L. Krishnamurthy, R. Adler, P. Buonadonna, J. Chhabra, M. Flanigan, N. Kushalnagar, L. Nachman, M. Yarvis, Design and deployment of industrial sensor networks: experiences from a semiconductor plant and the north sea, in: Proceedings of the 3rd international conference on Embedded networked sensor systems, Association for Computing Machinery, San Diego, California, USA, 2005, pp. 64–75.
[10] W. Yan, L. Yu, On accurate and reliable anomaly detection for gas turbine combustors: A deep learning approach, arXiv:1908.09238v1 [cs.LG] (2019).
[11] O. Niggemann, A. Vodencarevic, A. Maier, S. Windmann, H.K. Büning, A learning anomaly detection algorithm for hybrid manufacturing systems, in: The 24th International Workshop on Principles of Diagnosis (DX-2013), Jerusalem, Israel, 2013.
[12] C.J. Kuo, K. Ting, Y. Chen, State of product detection method applicable to Industry 4.0 manufacturing models with small quantities and great variety: An example with springs, in: 2017 International Conference on Applied System Innovation (ICASI), 2017, pp. 1650-1653.
[13] T. Ergen, A.H. Mirza, S.S. Kozat, Unsupervised and semi-supervised anomaly detection with LSTM neural networks, arXiv:1710.09207v1 [eess.SP], (2017).
[14] F.E. Grubbs, Procedures for detecting outlying observations in samples, Journal of Technometrics, 11 (1969) 1-21.
[15] C.C. Aggarwal, Outlier analysis, in: Data mining, Springer, 2015, pp. 237-263.
[16] C.C. Aggarwal, Data mining: the textbook, Springer, 2015.
[17] E. Sølvsberg, C.D. Øien, S. Dransfeld, R.J. Eleftheriadis, O. Myklebust, Analysis-oriented structure for runtime data in Industry 4.0 asset administration shells, Procedia Manufacturing, 51 (2020) 1106-1110.
[18] B.C. Vickery, Reviews : van Rijsbergen, C. J. Information retrieval. , 2nd edn. ed., London, Butterworths, 1979.
[19] C. Goutte, E. Gaussier, A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation, in: 27th European conference on Advances in Information Retrieval Research, Springer, Santiago de Compostela, Spain, 2005, pp. 345-359.