Towards Novel Statistical Methods for Anomaly Detection in Industrial Processes

Simone Tonini
Department of Excellence EMbeDS, Sant'Anna School for Advanced Studies
Pisa, Italy

Fernando Barsacchi
A. Celli Group
Porcari, Lucca, Italy

Francesca Chiaromonte
Department of Excellence EMbeDS and Institute of Economics, Sant'Anna School for Advanced Studies
Pisa, Italy
Dept. of Statistics and Huck Institutes of the Life Sciences, The Pennsylvania State University
USA

Daniele Licari
Department of Excellence EMbeDS, Sant'Anna School for Advanced Studies
Pisa, Italy

Andrea Vandin
andrea.vandin@santannapisa.it
Department of Excellence EMbeDS and Institute of Economics, Sant'Anna School for Advanced Studies
Pisa, Italy
DTU Technical University of Denmark
Lyngby, Denmark

ABSTRACT

This paper presents a novel methodology based on first principles of statistics and statistical learning for anomaly detection in industrial processes and IoT environments. We present a 5-level analytical pipeline that cleans, smooths, and eliminates redundancies from the data, and identifies outliers as well as the features that contribute most to these anomalies. We show how smoothing can make our methodology less sensitive to short-lived anomalies that might be, e.g., due to sensor noise. We validate the methodology on a dataset freely available in the literature. Our results show that we can identify all anomalies in the considered dataset, with the ability of controlling the amount of false positives. This work is the result of a research project co-funded by the Tuscany Region and a company leader in the paper and nonwovens sector. Although the methodology was developed for this domain, we consider here a dataset from a different industrial sector. This shows that our methodology can be generalized to other contexts with similar constraints on limited resources, interpretability, time, and budget.

CCS CONCEPTS

• **Computing methodologies** → **Anomaly detection**.

KEYWORDS

anomaly detection, industrial processes, mahalanobis distance

ACM Reference Format:
Simone Tonini, Fernando Barsacchi, Francesca Chiaromonte, Daniele Licari, and Andrea Vandin. 2023. Towards Novel Statistical Methods for Anomaly Detection in Industrial Processes. In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering (ICPE '23 Companion), April 15–19, 2023, Coimbra, Portugal.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3578245.3585036

ICPE '23 Companion, April 15–19, 2023, Coimbra, Portugal

ACM ISBN 979-8-4007-0072-9/23/04…$15.00
https://doi.org/10.1145/3578245.3585036

1 INTRODUCTION

With the advent of Industry 4.0, industries have begun to make intensive use of sensors to collect, archive, and analyze data. The goal could be, e.g., to monitor the state of a process to identify anomalies in the behavior or performance, or to compare the performance of machines operating in different plants. From this perspective, anomaly detection has become more and more a task of interest to practitioners, but also to researchers in statistics, and computer science. Anomaly detection is the process of identifying unexpected values or events in data. This type of analysis can be useful in various fields such as quality control, medicine, finance, image analysis, and chemistry (see [16] for a selected list of papers on anomaly detection in such fields).

As highlighted in [5], we can distinguish three main kinds of anomaly: *points* anomaly, *contextual* anomaly and *collective* anomaly. Here we focus on point anomalies (hereafter anomalies), i.e., a sequence of points that abruptly deviates from the *usual* values. For anomaly detection, practitioners traditionally set upper and lower control limits, and values outside these limits are considered anomalies. It is worth noting that anomalies can appear as temporal noise, often caused by sensor errors (usually referred to as noise) or

abnormal system operation (usually referred to as system anomaly). The former are shocks of short duration, while the latter might involve longer time intervals.

This paper introduces an empirical data-driven methodology to *identify* anomalies in industrial contexts where the production process is characterized by time series with unknown distribution. In particular, this paper shows the methodology currently developed within the AutoXAI2 project, born from the cooperation between Sant'Anna School for Advanced Studies of Pisa, and the A.Celli company of Lucca, leader in the supply of machinery and advanced technologies for the paper and nonwovens market[1]. The project is co-funded by Tuscany region and the company. The company also provided the data relating to a tissue paper machine. The variables, approximately 300, were collected with a frequency of approximately one observation per second. The data does not have anomaly labels, and the objective of the study was to identify anomalies during the production process, both to monitor the progress of a specific production line and to compare different production lines to see how they differ although they should perform in a similar way. In the initial validation phases, discussions with domain experts are obviously needed to determine whether the identified anomalies are real process anomalies or just noise. In this paper we focus on the first problem, anomaly detection, validating our proposed approach on a dataset equipped with anomaly labels freely available online.

Related work. Researchers had developed various statistical tools to classify observations into regular points and outliers. One of them is the Mahalanobis distance [11], a useful way to determine the similarity of an unknown sample space to a known one, based on correlations between variables by which different patterns can be identified and analyzed. In particular, Mahalanobis-type distances are calculated and a limit value based on the distribution of these distances is used to classify observations as anomalies [12, 13]. However, most applications assume that the variables follow a Normal distribution or, in the most extreme cases, are skewed [3, 8, 10, 14, 17, 18].

In recent years, deep learning has attracted the attention of anomaly detection scholars. Deep learning techniques have the advantage of learning the complex dynamics in the data without having to make assumptions about the underlying patterns within the data. This property makes them an appreciated tool for anomaly detection even in the case of time series. A recent and exhaustive review of the state-of-the-art on deep learning-based anomaly detection approaches for time-series data is provided by [2, 5].

However, although research on deep learning techniques has reached the state of the art with respect to anomaly detection tasks, there are contexts in which they are difficult to apply. In particular, some of the limitations that most affect the use of such advanced techniques are: lack of sufficient data, lack of adequate computational resources, time constraints for development or inference, unknown distribution of data, interpretability of results, and high cost.

Contribution. In this paper we propose an agnostic 5-steps methodology to classify one or more observations as anomalies, which is based on first principles of statistical learning (variance inflation factor, Mahalanobis distance, and Chebyshev's inequality). The proposed methodology solves most of the problems listed above, namely it is easy to implement, fast to run and does not require to know the distribution of the variables.

The first three steps refine the dataset to be used in later steps to identify anomalies. In particular, the original data is cleaned of irrelevant variables and filtered with a median smoothing procedure. Next, a selection of variables is made using the variance inflation factor [6] to remove multicollinearity. The dataset defined by these three steps is then used to estimate the Mahalanobis distance. Thanks to this, we can classify the observations as anomalies if they exceed a certain threshold value of the standardized Mahalanobis distance, i.e., the Mahalanobis distance scaled to have mean 0 and variance 1. The value of such threshold, say k, can be chosen by using first principles from statistics, i.e. Chebyshev's inequality [1]. Imposing a value k as a threshold has the effect of classifying as anomalies events with a probability of occurrence less than or equal to $1/k^2$. Thus, Chebyshev's Inequality ensures that the identified anomalies have a degree of rarity determined by the tuning parameter k. We will deepen the aspects of the adopted methodology later.

To validate our methodology, we use a popular dataset from the literature [15], the Server Machine Dataset (SMD), which comes with labels to indicate whether an observation is an anomaly[2]. Therefore, in this exercise the domain expert is replaced by the labels provided in the SMD dataset, and we use them to validate our methodology and to see how it behaves based on the varying of some tuning parameters.

As discussed, the steps for constructing the analytical pipeline are well established. However, to the best of our knowledge, they have never been used in a combined and integrated manner as in our methodology, which has therefore never been addressed in the established literature.

In this paper we want to present the experiences encountered so far in developing and applying our methodology, providing a first validation of it.

2 THEORETICAL BACKGROUND

Some tools we use in this paper have critical issues that we try to address through the proposed methodology. In particular, the Mahalanobis distance is very sensitive to variations that are present in the data, which can be considered as noise and not as anomalies of the production process. We address this issue by using labeled data that allow us to validate the impact on our procedure of a smoothing step, where we consider a coarse-grained version of the data using the median of windows (or steps) of w observations, with w a parameter that can be set by the user. This can, e.g., remove very short-lived anomalies that a domain expert might consider to be noise. Obviously, the definition of process anomaly plays a key role in setting the data smoothing step and, consequently, it determines the performance of the proposed methodology. In this perspective, using the aforementioned SMD labeled dataset, we show how the estimation of anomalies can be controlled for different degrees of smoothness and compare them with the true ones.

[1] www.acelli.it

[2] https://github.com/NetManAIOps/OmniAnomaly

2.1 Technicalities and Challenges

Let $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ denote an $n \times T$ rectangular array of observations concerning n stationary time series (i.e. time series with finite variance). Given an n-dimensional data point $\{\mathbf{x}_i\} = \{(x_{1i}, x_{2i}, \ldots, x_{ni})'\}$, we focus on the basic question:

> *how distant is* $\{\mathbf{x}_i\}$ *from the center of the distribution of* $\mathbf{X}$*?*

To answer this question we consider the distance metric developed by Mahalanobis [11]. Let $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_n\}'$ be the vector of the arithmetic means of the n variables, i.e. μ_i is the mean of $\mathbf{x}_i$, and let $\boldsymbol{\Sigma}$ be the usual sample covariance matrix. The Mahalanobis distance (MD) is defined as

$$MD_i = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})'}. \tag{1}$$

MD_i tell us how far $\mathbf{x}_i$ is from the center of the data by taking into account the covariance (or correlations if the $\mathbf{x}_i$'s are re-scaled to have variance 1). It is zero for $\mathbf{x}_i$ at $\boldsymbol{\mu}$, and grows as $\mathbf{x}_i$ moves away from $\boldsymbol{\mu}$ along the principal component axis[3]. Note that, if the distribution of the n variables is exactly multivariate normal, then

$$MD_i \leq \chi^2_n(\alpha) \tag{2}$$

with probability $1 - \alpha$, where $\chi^2_n(\alpha)$ is the Chi-Squared distribution with n degrees of freedom. Here we propose to use the MD_i to identify anomalies in the production process. However, it is important to highlight that the considered data presents critical issues and challenges mitigated by the proposed methodology.

Challenges: (i) The masking effect. As for the MD, the authors of [11] stress that (1) suffers from the so called *masking effect*, according to which multiple anomalies do not necessarily have a large MD_i. This is due to the fact that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are not robust: a small cluster of outliers will attract $\boldsymbol{\mu}$ and will inflate $\boldsymbol{\Sigma}$ in its direction. We solve this issue by identifying a sufficient long portion of the dataset where production was in line with expectation, i.e. a portion of the dataset where all the variables move in an ideal range of values. Our idea is to identify first such a dataset without anomalies, let us call it X_A. This should be provided, e.g., by a domain expert during initial analysis phases. Then, we iteratively merge X_A to small pieces of new data, further portions of the dataset, that we call $X_B, X_C, \ldots$, obtaining datasets (X_A, X_B), (X_A, X_C), ... For each such recombined dataset, we evaluate whether it contains anomalies using our methodology. It is important to note that a possible alternative solution is to consider robust estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e. resistant against the influence of clusters of anomalies, giving rise to a robust Mahalanobis distance [3, 4]. However, in this paper we want to show that if a large enough anomaly-free dataset is identified, then the simple original version of the Mahalanobis distance can detect anomalies in other small samples. We do not rule out using robust versions of the Mahalanobis distance in future works.

Challenges: (ii) Multicollinear variables. As reported in (1), the Mahalanobis distance requires the covariance matrix to be invertible. However, many variables may be multicollinear and this can prevent for the covariance matrix to be invertible. In order to remove multicollinear variables we apply the variance inflation factor (VIF, [6]). The VIF consists of two steps. First, each variable is regressed on the other $n - 1$ variables. Second, the VIF for each variable x_i, $i = 1, \ldots, n$, is obtained as $1/(1 - R_i^2)$, where R_i^2 is the coefficient of determination of the regression model run at the previous step. We remove the variable with the largest VIF and run again the algorithm until the largest VIF is less than 5, a threshold value commonly used in the literature [9].

Challenges: (iii) Unknown distributions. Another issue regards the fact that the distribution of the data of a production process is typically undefined or unknown. It means that we can not use the χ^2_n distribution to define anomalies as in (2). We propose to solve this limitation by using Chebyshev's inequality. Let μ_{md} and σ^2_{md} be the mean and variance of the MD obtained on a sample (for example (X_A, X_B)), respectively. The Chebyshev's inequality says that for $Z_{md_i} = \frac{MD_i - \mu_{md}}{\sigma_{md}}$ we have

$$P\{|Z_{md_i}| \geq k\} \leq \frac{1}{k^2}. \tag{3}$$

The inequality in (3) says that, regardless of the distribution of the data, the probability of a Z_{md_i} with an absolute value grater than or equal to a value k must be less than or equal to $1/k^2$. Therefore, no matter what the distribution of Z_{md_i}, the probability of observing, for example, Z_{md_i} of 5 is no greater than 0.04 (i.e. 1/25).

3 METHODOLOGY

Let $\mathbf{X}_A$ be an $n \times T_A$ array of observations relative to the production process without anomalies. Let $\mathbf{X}_B$ be an $n \times T_B$ array of observations relative to the production process. We want to detect possible anomalies in $\mathbf{X}_B$. We recall that for the Mahalanobis distance we need $T_A \gg T_B$ i.e. X_A must be significantly larger than X_B. For the $n \times T_{tot}$ matrix $\mathbf{X}_{tot} = (\mathbf{X}_A, \mathbf{X}_B)$, with $T_{tot} = T_A + T_B$, the proposed procedure is summarized as follows.

Step 1 - Data cleaning. We start by cleaning the dataset, i.e. we remove all categorical variables that are constant in our subsample $\mathbf{X}_{tot}$, as well as the variables that are pointed out as irrelevant by the domain expert, and therefore only bring noise.

Step 2 - Smoothing. We smooth the data through the median. let us consider a window of size h, a median filter applied to each of the T_{tot}-dimensional vector $\mathbf{x}_i$'s works as follows:

$$\begin{aligned} w_{i1} &= med(x_1, \ldots, x_h), \\ w_{i2} &= med(x_2, \ldots, x_{h+1}), \\ &\vdots \\ w_{iT_{tot}-h+1} &= med(x_{T-h}, \ldots, x_{T_{tot}}). \end{aligned}$$

Therefore, given a window of size h, we replace the $n \times T_{tot}$ matrix $\mathbf{X}_{tot}$ with the $n \times T_{tot} - h + 1$ matrix $\mathbf{W}$. This step allows to tune our methodology to specific considered domains. Indeed, our methodology might be *too sensitive*, meaning that it might identify anomalies that happen for very short amounts of time, and therefore are not considered anomalies by the domain experts.

[3] Note that if the principal component axes are re-scaled to have variance 1, then MD_i correspond to the Euclidean distance.

Step 3 - VIF. As previously said, we remove the variable with the largest VIF and run again the algorithm until the largest VIF is less than 5. Thus, after this step we get the $m \times T_{tot} - h + 1$ matrix $\widetilde{\mathbf{W}}$, where $m \leq n$.

Step 4 - Anomalies detection. We calculate the Mahalanobis distance (1) relative to the dataset $\widetilde{\mathbf{W}}$, denoted as $MD(\widetilde{\mathbf{W}})$, and re-scale it to have 0 mean and variance 1, namely $Z_{md(\widetilde{\mathbf{w}})}$. An observation i is labeled as anomaly if its standardized re-scaled Mahalanobis distance, i.e. $Z_{md_i(\widetilde{\mathbf{w}})}$, is greater than k, where k is the tuning parameter which determines the upper bound on the probability of occurring according to the Chebyshev's inequality (3). Thus, we obtain a new binary variable, named Y, equal to 1 if $|Z_{md_i(\widetilde{\mathbf{w}})}| \geq k$ and 0 otherwise.

Step 5 - Variable detection. If we identify anomalies in the observations relating to the dataset $\mathbf{X}_B$, then we identify also the variable that most contributes to those anomalies as

$$max_{1 \leq j \leq m} Corr(Y, X_{Bj}). \tag{4}$$

This could be useful, e.g., to suggest to the domain expert how to interpret the anomaly, and how to address it.

Discussion. The goal of the methodology just shown is to provide an anomaly detection tool in a completely agnostic context where data are not labeled. The anomalies identification process must be supervised by a domain expert who must confirm whether they are true anomalies, by checking whether the relevant variables have exceeded the upper and lower control limits. In particular, the domain expert will determine whether some short duration shocks are true system anomalies or simple sensor noises that can be ignored. In this sense the procedure is adaptive. For example, let's consider the case where we do not apply Step 2 related to smoothing. Our expectation is that the methodology is sensitive to short duration shocks labeling them as anomalies, which however can be sensor noise. If the domain expert will evaluate these anomalies as irrelevant, then we could set a middle-high value of the filtering window h and remove the noise.

Furthermore, it is important to underline that in this agnostic context our methodology does not aim to perfectly estimate the anomalies, but simply to give an indication of where they have occurred in order to give the domain expert the possibility to: (*i*) verify if there are any regular patterns in the formation of anomalies (e.g. there are particular variables that contribute to their formation); (*ii*) being able to make a summary of the anomalies that each machine has accumulated in a certain period and thus to compare machines that should perform in the same way.

Therefore, assuming that the anomalies are of medium-long duration, we expect that, once the smoothing parameter is well set, our methodology does not lead to false positives, at the cost of having false negatives. This means that we expect to find anomalies that have a shorter duration than the true ones due to smoothing, but which occur in a time interval where a real anomaly occurred.

4 VALIDATION

The Server Machine Dataset. To validate our methodology, we use the Server Machine Dataset (SMD), a recent 5-week-long dataset collected from a large Internet company [15] and composed by 38 variables. It is one of the largest public datasets currently available for evaluating multivariate time-series anomaly detection. It contains metrics like CPU load, network usage, memory usage, etc. SMD is made up by data from 28 different machines where the observations are collected per minute, and each of them is divided into training and test set. It is important to note that there is no information related to the variables involved in the SMD dataset. In fact, the variables are labeled based on the number of the column where they are located into the data matrix. For each anomaly, these numbers are used to indicate the variables that caused it. The dataset has been validated elsewhere [15], providing it with labels to denote anomalies. The test set contains period with and without anomalies, and therefore can be used to validate the methodology presented in Section 3.

Setting. We try to identify anomalies for the first machine by focusing on the test set only. In fact, the test set contains a large first part with no anomalies (15800 timepoints) and a last part with 8 clusters of anomalies (about overall 12700 timepoints for all 8 clusters) as depicted in Figure 1. We use the first part of the test set as a sufficiently large basis for our procedure and merge it every time with a piece of the dataset containing one of the clusters of anomalies. As reported in Section 3, the initial part of the test set becomes X_A and each of the following 8 parts containing clusters of anomalies become datasets $X_B, X_C, \ldots, X_I$. The proposed methodology is thus applied to the datasets (X_A, X_B), (X_A, X_C),. . . , (X_A, X_I), separately. We set the tuning parameter relative to the degree of rarity, i.e. k, at 10, meaning that the anomalies are those events with a probability of occurring less or equal 1% according to the Chebyshev's inequality (3). Figure 2 reports the results obtained for the 9 datasets (the dataset X_A without anomalies and the other 8 datasets with anomalies). We consider 3 different values of the smoothing parameter, i.e. the size of the h window to which the median filter is applied. The values of h are 1, 10, 60. We have the first case, h=1 where the data are not smoothed, shown in the first row of Figure 2. The intermediate case (h=10) is shown in the second row, where a smoothing that we could define soft is applied. Finally, the case with stronger smoothing (h=60) is given in the third row.

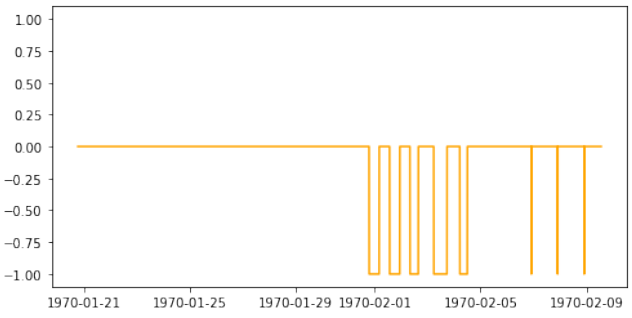

Figure 1: Anomalies in the test set of the first machine of the SMD dataset as specified in its labeling. The x-axis shows the dates on which the data were collected. Anomalies are denoted with y value, a zero-value denotes the absence of anomalies, while -1 indicates the presence of anomalies.

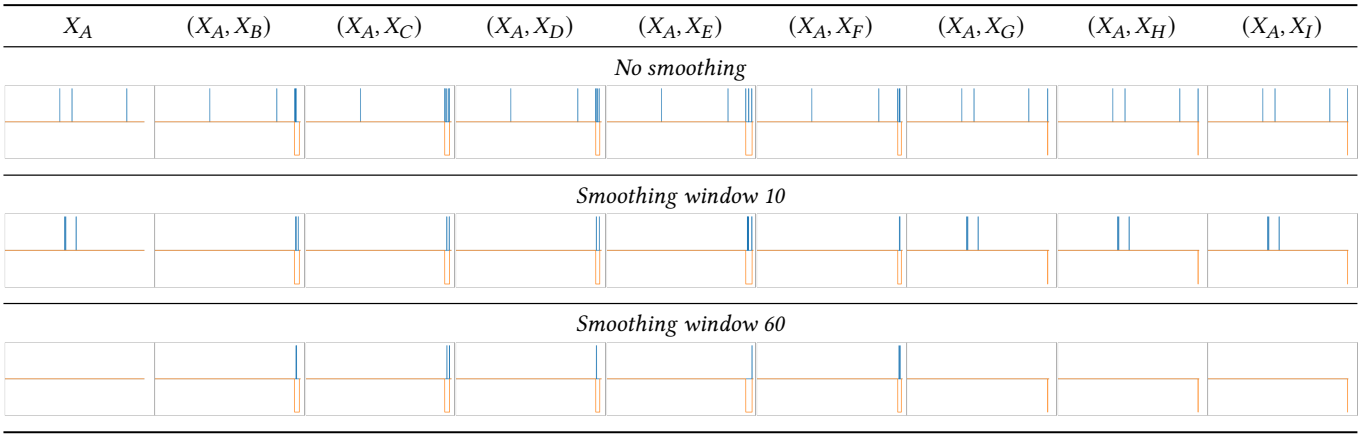

Figure 2: Our methodology for anomaly detection applied to the first machine of the SMD dataset. Column X_A considers the anomalies-free portion of the dataset. Columns (X_A, X_-) consider the dataset obtained by adding portions of the dataset with anomalies to X_A. Each row considers a different smoothing level.

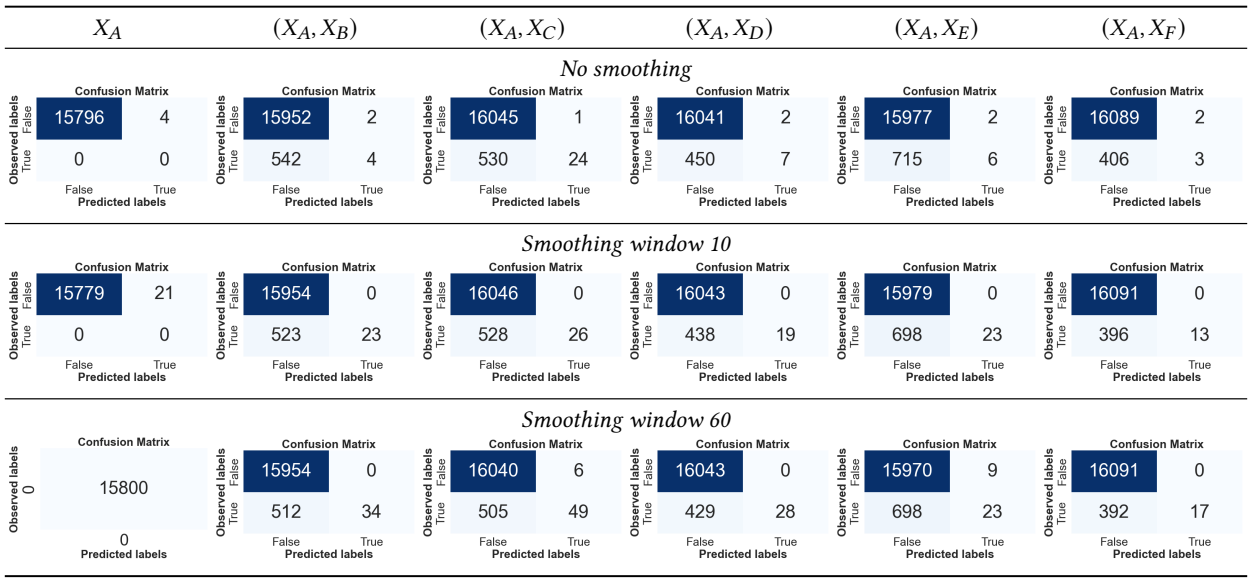

Figure 3: Confusion matrices for the imputed anomalies from data X_A to (X_A, X_E). The upper left quadrant reports the number of true negatives, while the lower right quadrant reports the number of true positives. The upper right and lower left quadrants report false positives and false negatives, respectively. Each row considers a different smoothing level.

The blue line reports the anomalies detected by our methodology and the yellow line reports the true ones from the labels of the dataset. Both lines are 0 when there are no anomalies, whereas their values diverge in the case of anomalies to better show results, namely 1 for our methodology and -1 for the true anomalies.

Hypothesis. We expect to observe that as the smoothing window increases, false positives decrease, reducing however the ability to identify anomalies of short duration, i.e. the number of false negatives increases. At the opposite side, low (or absent) smoothing could allow us to identify all anomalies, at the price of increasing the number of false positives.

Discussion/results. Observing the yellow lines we can informally divide the true anomalies into two different types. In fact, from dataset (X_A, X_B) to (X_A, X_F) the anomalies have a relatively large duration, while from the dataset (X_A, X_G) to (X_A, X_I) they affect few observations. The results of this validation study can be summarized as follows. When we do not smooth the data (first row of Figure 2) the proposed methodology always identifies the anomalies in correspondence with the true ones. However, we observe that there are also some *false positives* among the identified anomalies. These are due to the fact that without a smoothing step, the proposed procedure is sensitive to short irrelevant shocks (noises) in the data.

If we insert a slight smoothing of the data (second row), the problem is solved when the real outliers have a relatively large duration, i.e. from case (X_A, X_B) to case (X_A, X_F). When this does not happen because there are no anomalies (case X_A), or because the anomalies are too short (cases from (X_A, X_G) to (X_A, X_I)), then the methodology is strongly conditioned by small variations in the data which condition the vector $\boldsymbol{\mu}$ and the matrix $\boldsymbol{\Sigma}$, leading to false positives.

In the most extreme considered case of high smooth (third row), we observe that, on the one hand, there are no false positives in any of the 9 datasets. However, on the other hand, we see that the anomalies are selected only in the cases from (X_A, X_B) to (X_A, X_F), i.e. when they have a duration relatively large. In fact, cases (X_A, X_G) to (X_A, X_I) have too short a duration and have been removed from smoothing.

It is important to underline that in our case it is more convenient to reduce false positives to zero at the cost of increasing false negatives. This is because according to the A.Celli's domain experts, their anomalies have a common trend, i.e. they are never shocks of one or a few observations, but events that last for a prolonged period. Therefore, it is sufficient to use hard smoothing to remove sensor noise even if this increases the number of false negatives. However, this does not pose a problem, since we know that, by construction, the anomalies are short due to smoothing but refer to long-duration events. Thus, we are not interested in finding the entire interval within which the anomaly occurred, rather we are interested in finding a signal that tells us, quickly and effectively, when it occurred and which variables it involved.

Figure 3 shows the confusion matrices for the imputed anomalies from data X_A to (X_A, X_E), i.e. data without anomalies and data where the anomalies continue for a long period in the cases of $w = 1$ (first row of plots), $w = 10$ (second row of plots) and $w = 60$ (third row of plots). For each plot, the upper left quadrant (in dark blue) reports the number of true negatives, while the lower right quadrant reports the number of true positives. The upper right and lower left quadrants report false positives and false negatives, respectively. The results clearly show that a higher smoothing window eliminates false positives in dataset X_A that contains no anomalies. However, it is observed that going for $w = 60$ there are some false positives for the datasets (X_A, X_C) and (X_A, X_E), but which are not obtained for $w = 10$. This is due to the fact that for these two cases the imputed anomalies with $w = 60$ are slightly outside the range of the true anomalies. In our opinion, this does not pose a problem since the few false positives do not refer to another cluster of anomalies, but to the same one and may be due to the observations involved in the smoothing. Therefore, despite this problem, we consider $w = 60$ to be preferable to $w = 10$ since the former allows us to impute zero anomalies in the X_A dataset.

The last result we present is shown in Figure 4. It concerns the selection of the variable most correlated with the anomaly identified in the case of strong smoothing ($h = 60$). This result is obtained from step 5 of the procedure presented in Section 3. The first row contains the same information of Figure 2, while the second shows the value of the variables that has been identified according to (4). We can observe that for each anomaly the variable that is most correlated with it is stationary over the whole sample, with the exception of the final part where a strong anomaly is observed. In all 5 cases the variable that has been selected appears to be part of the set of variables that caused the anomaly according to the domain experts of the SMD dataset (as reported in [15]). Remember that, as previously pointed out, the only information relating to the relevant variables for each anomaly is their position in the dataset.

5 CONCLUSIONS

We have presented a novel preliminary methodology for identifying anomalies in time series from industrial processes. The procedure is based on well-known statistical methods such as the Variance Inflation Factor, the Mahalanobis distance and Chebyshev's inequality. We have validated our methodology on a well-studied dataset from the literature, namely the Server Machine Dataset (SMD) [15] which contains 5 weeks of data from a large Internet company.

We showed how a smoothing parameter can be used to make our method less sensitive to short anomalies, which may be uninteresting in certain domains. Overall, with slight smoothing we were able to identify all anomalies present in the considered dataset, while giving some false positives. The latter could be eliminated by increasing the smoothing, obtaining a few false negatives for real but short anomalies.

This research is part of the AutoXAI2 research project. The project is partially funded by Tuscany region and the A.Celli group. Within the project, the authors collaborate on the development of the presented methodology and its application to the company's data from the paper and nonwovens industrial sector. The aim of this paper was also to discuss the experience gained in this collaboration.

Limitations and Future works. We have told the experience developed in this industrial collaboration, and the methodology originated from solving the related problems. However, in future works, we will further develop and complete our methodology by focusing on discovering the variables that caused the identified anomalies. We will also consider further datasets, e.g. considering more machines from the SMD dataset, so to compare data across different machines. Finally, we will consider further datasets, and if possible we will present results on the data from our industrial partner.

The following is a detailed list of future works.

- *Weighted Mahalanobis distance.* Evaluate the use of a weighted version of the Mahalanobis distance [19] in order to give less weight to variables with short-term noises and greater weight to variables with short-term anomalies that are crucial for the process. Obviously, this improvement requires a phase of comparison with the domain expert to select the irrelevant variables from the potentially relevant ones.
- *Detecting critical variables.* Instead of identifying the variable that contributes most to the anomaly using maximum correlation, there are more robust approaches that can determine which variables have the greatest influence on the Mahalanobis distance [7].
- *Explaination phase.* Better develop the last step of the methodology to identify which variables caused a discovered anomaly. This will be useful to suggest the domain expert how understand and fix the methodology. In particular, the proposed procedure involves a loss of information mainly due to the VIF and the use of the Mahalanobis distance. The first

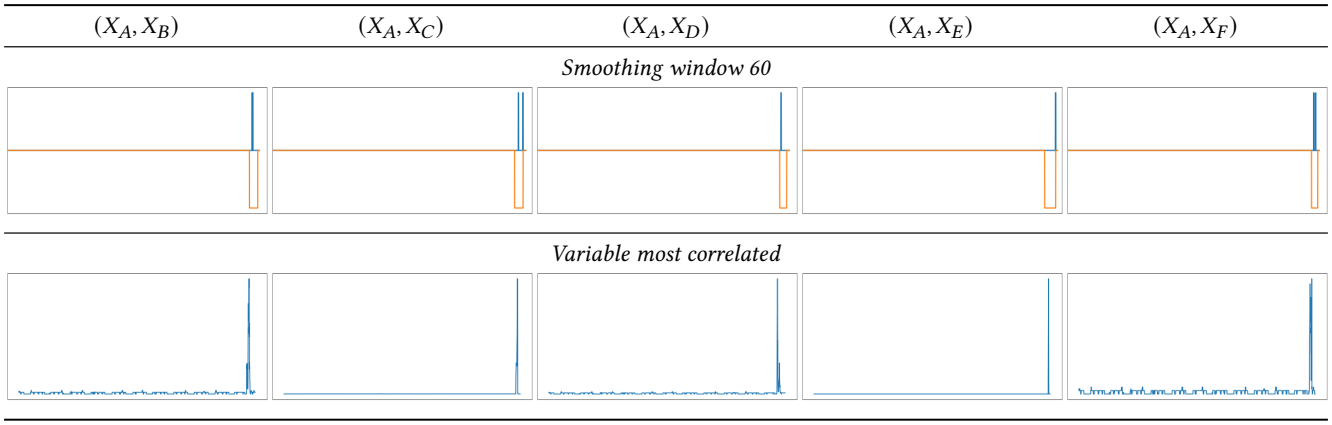

Figure 4: Variable most correlated with the anomalies obtained by smoothing data with a window of size $h = 60$.

reduces the number of variables used for the identification of anomalies, the last consists in a real transformation of the data. Due to this loss of information, not all variables that are deemed relevant by the domain expert may be present in the dataset used to identify anomalies. We could go back to all the original relevant variables by adopting a multivariate approach that starts from the relevant variables that remained in the dataset.

- *Expand application.* Apply the proposed methodology to other machines related to the SMD dataset to simulate a comparison between machines. For example, by assessing if there are machines that perform worse than others getting more anomalies.

6 ACKNOWLEDGMENTS

This work has been partly funded by the FSE regional Tuscan project AUTOXAI2 J53D21003810008.

REFERENCES

[1] Alsmeyer, G. *Chebyshev's Inequality.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 239–240.
[2] Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR) 54*, 3 (2021), 1–33.
[3] Cabana, E., Lillo, R. E., and Laniado, H. Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators. *Statistical Papers 62*, 4 (nov 2019), 1583–1609.
[4] Campbell, N. A. Robust procedures in multivariate analysis i: Robust covariance estimation. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 29*, 3 (1980), 231–237.
[5] Choi, K., Yi, J., Park, C., and Yoon, S. Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. *IEEE Access* (2021).
[6] Craney, T. A., and Surles, J. G. Model-dependent variance inflation factor cutoff values. *Quality Engineering 14*, 3 (2002), 391–403.
[7] Garthwaite, P., and Koch, I. Evaluating the contributions of individual variables to a quadratic form. *Australian & New Zealand Journal of Statistics 58* (03 2016).
[8] Hubert, M., and Van der Veeken, S. Outlier detection for skewed data. *Journal of Chemometrics 22*, 3-4 (2008), 235–246.
[9] James, G., Witten, D., Hastie, T., and Tibshirani, R. An introduction to statistical learning: with applications in r.
[10] Kamoi, R., and Kobayashi, K. Why is the mahalanobis distance effective for anomaly detection?, 2020.
[11] Mahalanobis, P. C. On the generalized distance in statistics. National Institute of Science of India.
[12] Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. *Robust statistics: theory and methods (with R).* John Wiley & Sons, 2019.
[13] Rousseeuw, P. J., and Leroy, A. M. *Robust regression and outlier detection.* John wiley & sons, 2005.
[14] Rousseeuw, P. J., and van Zomeren, B. C. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association 85*, 411 (1990), 633–639.
[15] Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., and Pei, D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & amp; Data Mining* (New York, NY, USA, 2019), KDD '19, Association for Computing Machinery, p. 2828–2837.
[16] Tao, L., Liu, H., Zhang, J., Su, X., Li, S., Hao, J., Lu, C., Suo, M., and Wang, C. Associated Fault Diagnosis of Power Supply Systems Based on Graph Matching: A Knowledge and Data Fusion Approach. *Mathematics 10*, 22 (November 2022), 1–28.
[17] Tiku, M. L., Islam, M. Q., and Qumsiyeh, S. B. Mahalanobis distance under non-normality. *Statistics 44*, 3 (2010), 275–290.
[18] Todeschini, R., Ballabio, D., Consonni, V., Sahigara, F., and Filzmoser, P. Locally centred mahalanobis distance: A new distance measure with salient features towards outlier detection. *Analytica Chimica Acta 787* (2013), 1–9.
[19] Wölfel, M., and Ekenel, H. K. Feature weighted mahalanobis distance: Improved robustness for gaussian classifiers. In *2005 13th European Signal Processing Conference* (2005), pp. 1–4.