RESISTANT STATISTICAL METHODOLOGIES FOR ANOMALY DETECTION IN GAS TURBINE DYNAMIC TIME SERIES: DEVELOPMENT AND FIELD VALIDATION

Giuseppe Fabio Ceschini[1], Nicolò Gatta[2], Mauro Venturini[2], Thomas Hubauer[1], Alin Murarasu[1]

[1] Siemens AG, Nürnberg, Germany

[2] Dipartimento di Ingegneria, Università degli Studi di Ferrara, Ferrara, Italy

ABSTRACT

The reliability of gas turbine health state monitoring and forecasting depends on the quality of sensor measurements directly taken from the unit. Outlier detection techniques have acquired a major importance, as they are capable of removing anomalous measurements and improve data quality. To this purpose, statistical parametric methodologies are widely employed thanks to the limited knowledge of the specific unit required to perform the analysis. The backward and forward moving window (BFMW) k-σ methodology proved its effectiveness in a previous study performed by the authors, to also manage dynamic time series, i.e. during a transient. However, the estimators used by the k-σ methodology are usually characterized by low statistical robustness and resistance.

This paper aims at evaluating the benefits of implementing robust statistical estimators for the BFMW framework. Three different approaches are considered in this paper. The first methodology, k-MAD, replaces mean and standard deviation of the k-σ methodology with median and mean absolute deviation (MAD), respectively. The second methodology, σ-MAD, is a novel hybrid scheme combining the k-σ and the k-MAD methodologies for the backward and the forward windows, respectively. Finally, the bi-weight methodology implements bi-weight mean and bi-weight standard deviation as location and dispersion estimators.

First, the parameters of these methodologies are tuned and the respective performance is compared by means of simulated data. Different scenarios are considered to evaluate statistical efficiency, robustness and resistance. Subsequently, the performance of these methodologies is further investigated by injecting outliers in field data sets taken on selected Siemens gas turbines.

Results prove that all the investigated methodologies are suitable for outlier identification. Advantages and drawbacks of each methodology allow the identification of different scenarios in which their application can be most effective.

NOMENCLATURE

c	parameter of the bi-weight methodology
k	acceptability threshold for the k-σ test criterion
m	outlier magnitude factor
med	median
N	number
s	noise factor
$\bar{s}$	standard deviation
$\hat{s}$	bi-weight standard deviation
u	weight
w	number of measurements in the window sample
x	measurement in the time series
$\bar{x}$	mean
$\hat{x}$	bi-weight mean

Subscripts and Superscripts

b	backward window
f	forward window

Acronyms

GT	gas turbine
FNR	false negative rate
FPR	false positive rate
MAD	median absolute deviation
SD	standard deviation
TPR	true positive rate

INTRODUCTION

The energy market demand sets high requirements to the productivity of gas turbine (GT) units, requiring high availability and efficiency levels to achieve cost effectiveness. Furthermore, the complexity of the units implies a high level of insight on the health state and condition of the turbines. For these reasons, the activity of researchers and company R&D departments focuses on the development of tools for monitoring, diagnostics and trending of the health state of GT fleets installed worldwide [1, 2, 3, 4, 5]. Data processing by means of software tools aimed at GT health analysis usually relies on direct measurements acquired by different types of sensors installed in the GT unit.

However, the environment in which sensors operate can be extremely harsh. Among other causes, high temperatures, enhanced fluid speed and unsteady operation of the units may cause hardware degradation and failures of measurement devices, thus severely affecting the quality of sensor data [6]. The input of low quality data for GT health analysis affects the reliability of results, implying misleading performance assessments and erroneous anomaly calls. False positives constitute a severe issue for O&M companies as they may lead to excessive deployment of maintenance operations, causing unnecessary downtime, extra costs and loss of reputation towards customers.

For these reasons, the application of methods at intermediate stage between data acquisition and GT health analysis is fundamental to enhance input data quality, thus posing outlier identification as a promising field of research for both GT users and manufacturers [7, 8, 9, 10]. In order to be suitable for industrial applications, anomaly identification techniques require simplicity of tuning. In fact, even if detection capabilities are somewhat related to the number of model parameters, in most cases this is seen more as a source of issues rather than a benefit. Statistics-based methodologies perfectly fit this target, with a less demanding tuning procedure than heuristics based techniques or complex thermodynamic models. Furthermore, the specific knowledge about the unit required to perform the analysis can be almost null. This feature enhances the applicability of statistical methodologies to different GT units without any substantial modification in their tuning. In fact, outliers are identified as a consequence of excessive deviation from a statistical model derived from available observations, by means of direct statistical inference [11, 12, 13, 14] or derived from autoregressive models [15, 16, 17, 18]. However, tuning complexity and poor performance achieved on large data sets reduce the attractiveness of autoregressive models techniques for industrial applications.

In this sense, the k-σ methodology proved to be particularly attractive for assessing the reliability of gas turbine sensor readings [12]. In [19], the authors derived direct and generally applicable guidelines for tuning the methodology and developed a backward and forward moving window scheme (BFMW) that also allowed the application to dynamic time series, i.e. during a transient. The methodology infers the statistical feature of a certain portion of the time series, defined by the backward and forward moving window size, by adopting sample mean and standard deviation (SD) as location and scale estimators. As these estimators present limitations, mainly in terms of statistical robustness and resistance, this paper aims at evaluating the benefit of implementing robust statistical estimators for the acceptability rule of the k-σ methodology. Therefore, different statistical methodologies for outlier detection are implemented and their performance is compared to that of the BFMW k-σ methodology.

This paper sets as a part of wider frame of research, which documents the efforts made by Siemens to continuously improve automated data processing to ultimately enhance data quality and detect failures. The implementation of automated processing techniques, without the involvement of frequent human decisions, is in fact crucial, considering that tens of gigabytes of data are collected on a daily basis from Siemens units. The research frame, in which this paper is set, includes two additional works by the authors. In [19], which is introductory to the current paper, two improved approaches are considered in addition to the standard k-σ methodology and the moving window approach is found to be the best on simulated data. In the current paper, different solutions to improve the methodology proposed in [19] are evaluated, in order to implement the best performing statistical methodology in a comprehensive tool for Detection, Classification and Integrated Diagnostics of Gas Turbine Sensors (named DCIDS). The tool, described and tested in [20], assesses the reliability of GT sensor measurements by using both single-sensor and multi-sensor analysis. Furthermore, the DCIDS tool is able to classify anomalies according to their characteristics and identify different fault scenarios. The performance of the tool is assessed by analysing different types of field measurements taken on several Siemens gas turbines.

This paper is organized as follows. The first section provides an overview of the key features of the statistical estimators considered in the analysis. Tuning guidelines of general validity for gas turbines applications are reported in the subsequent section, together with a first assessment of the capabilities of the methodologies against simulated data. Then, performance is compared by means of field data from Siemens units with outliers injected at randomly selected time points. The paper also includes a discussion about results together with best practices to fully exploit the potential of all the considered methodologies. These guidelines represent the main achievement and novel contribution of this paper.

RESISTANT STATISTICAL METHODOLOGIES FOR ANOMALY DETECTION

The performance of statistical estimators is commonly evaluated in terms of efficiency, robustness and resistance.

Efficiency is defined as the measure of sampling variability, i.e. the influence of sample elements on the estimation. It is usually expressed as a percentage referred to the sampling variability of the traditional estimator, i.e. sample mean and SD referred to a perfectly Gaussian distribution [6]. For this reason, it is commonly defined as Gaussian efficiency.

Robustness describes the ability of a model or test to effectively perform, while its variables or assumptions are altered. In fact, given a set of random observations, it is necessary to formulate assumptions regarding the characteristics of their underlying probability distribution. An estimator is “robust” if its efficiency holds despite such assumptions are completely unfulfilled.

Resistance refers to the capability of a statistical technique of keeping the estimation uninfluenced by the presence of outliers in the sample [21]. An indicator of the resistance of an estimator is the breakdown point, defined as the percentage of sample data points that can be replaced with arbitrary numbers without affecting the estimate value [22].

Sample mean and SD are well-known location and scale estimators for randomly distributed data sets. Under the assumption of homogenous, i.e. stationary, time series with Gaussian distributed samples, these estimators are known to be particularly effective and easy to compute [6]. In [19], the authors demonstrated the effectiveness of sample mean and SD for anomaly identification in dynamic time series, so proving to be suitable for processing gas turbine measurements. The methodology, named Backward and Forward Moving Window (BFMW) k-σ, implements these estimators by means of a parametric test criterion applied in a two-sided moving window time frame. According to the (BFMW) k-σ approach, an observation x_{t} is considered reliable if the following acceptability rule holds:

$$\frac{\left|x_{\mathrm{t}}-\bar{x}_{\mathrm{b}}\right|}{\bar{s}_{\mathrm{b}}}<k_{\mathrm{b}} \text{ OR } \frac{\left|x_{\mathrm{t}}-\bar{x}_{\mathrm{f}}\right|}{\bar{s}_{\mathrm{f}}}<k_{\mathrm{f}} \qquad (1)$$

The BFMW k-σ methodology (hereafter simply referred as k-σ) proved its effectiveness towards simulated dynamic time series in [19], achieving large percentages of detection combined with small percentages of false positive calls.

However, the employment of mean and SD as estimators may result ineffective in case assumptions regarding the underlying data distributions are not completely verified [6]. Despite some specific cases, the set of random observations is assumed as Gaussian distributed. In this case, sample mean and SD provide an efficient and unbiased estimate of location and scale with minimum variance [23]. However, as deviations from the Gaussian models occur in the data set, the performance of such estimators severely decreases [6, 22]. Anomalous observations, i.e. outliers, represent a perfect example of deviations from the assumed underlying distribution and their negative effect on mean and SD is particularly severe, being the breakdown point for theses estimators equal to 0%. Consequently, the performance of the k-σ methodology can be lowered, because of the scarce robustness and null resistance of the estimators implemented in its parametric test.

These considerations suggest potential benefits from the application of robust statistics location and scale estimators in the test criterion. This solution aims at the development of a statistical methodology implemented in the BFMW structure with enhanced robustness and resistance. However, resistance and efficiency are somewhat competing factors; thus a trade-off solution should be searched. Therefore, three alternatives to sample mean and SD in the test criterion are evaluated, implemented and investigated in this paper.

The first alternative to the BFMW k-σ methodology is a well-known robust scale estimator, i.e. the median absolute deviation (MAD) [22, 24], which became popular for outlier identification [24, 25] mainly because of its breakdown point, i.e. 50%. In spite of its exceptional robustness, the performance of the MAD in terms of efficiency is not so encouraging, achieving 37% for the Gaussian case [25]. Another drawback of the MAD is that MAD is equal to 0 if more than 50% of data are equal [24]. In such a case, observations are identified as outliers despite their absolute difference from the median. This phenomenon, called implosion [6], can occur for rounded data coming from sensor readings. As the MAD replaces SD in Eq. (1), the median replaces the mean. Therefore, the acceptability criterion for the k-MAD methodology can be expressed as:

$$\frac{\left|x_{\mathrm{t}} - med_{\mathrm{b}}\right|}{MAD_{\mathrm{b}}} < k_{\mathrm{b}} \ \text{ OR } \ \frac{\left|x_{\mathrm{t}} - med_{\mathrm{f}}\right|}{MAD_{\mathrm{f}}} < k_{\mathrm{f}} \qquad (2)$$

The second alternative to the BFMW k-σ methodology is a hybrid scheme. According to the k-σ methodology, observations prior to the one under assessment have already been processed and consequently they can be considered reliable. Therefore, the probability that the backward window contains outlier is rather low and the distribution is more likely to verify the Gaussian assumption. Under this condition, efficient estimators like mean and SD become more desirable than robust and resistant ones in the backward window. On the contrary, outliers can still occur among observations in the forward window, thus potentially affecting the quality of the estimate and consequently of the test criterion. In this case, robust estimators with consistent breakdown point, such as the median and the MAD, are required to prevent the reliability assessment from being excessively conditioned by outliers. The hybrid scheme consists of the application of the k-σ test criterion to the backward window and of the k-MAD test criterion to the forward window. The hybrid σ-MAD acceptability rule can be expressed as:

$$\frac{\left|x_{\mathrm{t}} - \bar{x}_{\mathrm{b}}\right|}{\bar{s}_{\mathrm{b}}} < k_{\mathrm{b}} \ \text{ OR } \ \frac{\left|x_{\mathrm{t}} - med_{\mathrm{f}}\right|}{MAD_{\mathrm{f}}} < k_{\mathrm{f}} \qquad (3)$$

This resistance towards the presence of outliers is fundamental, but at the same time the stiffness of median and MAD is likely to cause a high number of false positive calls.

For these reasons, a third alternative to the BFMW k-σ methodology, aimed at being an intermediate solution between the k-σ methodology and the k-MAD methodology, can be offered by the bi-weight mean and SD estimators, which present a more adaptive behaviour than median and MAD, while keeping a 50% breakdown point. Lanzante *et al.* [6] demonstrated that the bi-weight estimate of dispersion is more efficient than pure MAD in a sample composed of 30 elements, which is very similar to the size of the window analysed by the authors in [19] for the BFMW scheme. The estimate is performed by assigning different weights to observations according to their proximity to the centre of the distribution, which is inversely proportional to their probability of being outliers. The weights are assigned according to a bi-weight function, gradually decreasing as measurements detach from the location estimate until they drop to 0 when a certain distance is reached. This threshold value, determined by the parameter *c*, influences the measure in which each weight u_t for each observation x_t is assigned and consequently the performance of the methodology, according to Eq. (4).

$$u_t = \frac{x_t - med}{c \cdot MAD} \tag{4}$$

However, despite the interesting property of assigning less impact on measurements with high probability of being outliers, the tuning of the methodology may result complex. Directions for tuning are available in literature, but they refer to specific cases of application. Hoaglin et al. [26] suggest values of c between 6 and 9. Lanzante *et al.* adopt $c = 7.5$, while Kafadar [23] identifies 4-6 as a suitable range, concluding that the best performance is obtained with the value of 6.

In this paper, on the basis of the information available in literature, the parameter c is assumed equal to 6. The acceptability criterion for the bi-weight methodology is expressed by replacing mean and SD in Eq. (1) with their counterparts calculated by means of the previously described weighting procedure:

$$\frac{|x_t - \hat{x}_b|}{\hat{s}_b} < k_b \text{ OR } \frac{|x_t - \hat{x}_f|}{\hat{s}_f} < k_f \tag{5}$$

It can be seen that, if the value of the scale estimator in the test criteria in Equations (1), (2), (3) and (5) is zero, an undetermined solution occurs. In order to prevent the denominator from being equal to zero in the test criterion, a non-biasing infinitesimal quantity can be added to SD, MAD and bi-weight SD, respectively. The best performing tuning for k_b and k_f thresholds and an insight on capabilities are evaluated for each methodology by means of simulated data. Subsequently, the algorithms will be tested towards real datasets with injected outliers used to replace randomly chosen measurements in the time series.

TUNING OF THE RESISTANT METHODOLOGIES BY MEANS OF SIMULATED DATA

The best performing tuning parameters for the methodologies are determined by means of simulated data. The algorithm for time series generation and outlier contamination is the same employed by the authors to investigate the performance of the k-σ methodology in [19]. Data are distributed according to a Gaussian distribution with noise expressed as a percentage of the Gaussian standard deviation. Outliers are added to the time series with a user specified magnitude (±3%, ±4%, ±5% and ±7% with respect to the mean value) at randomly selected time points.

Three widely-used performance indices, derived from statistical test theory [27], are considered for the quantitative assessment of the different methodologies.

a) *True Positive Rate (TPR)* i.e. percentage ratio between the true positives identified by the algorithm and the number of total observations flagged as anomalies by the algorithm, i.e. the sum of true positives and false positives:

$$TPR = \frac{N_{\text{true positive}}}{N_{\text{true positive}} + N_{\text{false positive}}} \tag{6}$$

b) *False Negative Rate (FNR)* i.e. percentage ratio between the number of anomalous data incorrectly flagged as reliable and the number of true outliers

$$FNR = \frac{N_{\text{false negative}}}{N_{\text{true outlier}}} = 1 - \frac{N_{\text{true positive}}}{N_{\text{true outlier}}} \tag{7}$$

c) *False Positive Rate (FPR)* i.e. percentage ratio between the number of false positives and the number of true reliable data

$$FPR = \frac{N_{\text{false positive}}}{N_{\text{true reliable}}} \tag{8}$$

The simulated data scenarios employed for tuning the methodologies, as well as the field data used for their assessment, were carefully selected by the authors so that their morphology could enhance the evaluation of efficiency and resistance by means of the three performance indices *TPR*, *FNR* and *FPR*, to comprehensively evaluate the detection capabilities of each methodology.

Tuning of the methodologies. The optimal tuning identified for the methodologies is reported in Table 1. While the optimal values of w_b, k_b, w_f and k_f for the k-σ methodology were determined in [19], the parameters for the other techniques have been determined in this paper to achieve the maximum *TPR* and the minimum *FNR*, as made in [19]. The settings identified as optimal in Table 1 are adopted for the subsequent analysis performed in this paper.

Table 1 – Optimal tuning of the resistant statistical methodologies

	w_b	k_b	w_f	k_f
k-σ [19]	50	3	25	2
k-MAD	50	3	25	3
Hybrid σ-MAD	50	3	25	3
Bi-weight	50	3	25	3

The performance of the methodologies is investigated towards different simulated scenarios. As the effectiveness of the methodologies depends on the characteristics of the estimators implemented in the parametric test, it is important to investigate their capabilities towards different criteria. Namely, efficiency is evaluated by simulating the set point change maneuver already employed for the sole k-σ methodology in [19] with 1% data noise. The robustness of the methodologies is evaluated by changing the characteristics of the Gaussian distribution used for the generation of the time series. Therefore, the same maneuver used for assessing efficiency is also simulated with a significantly higher noise (i.e. 2%). Finally, resistance towards clustered outliers is evaluated by considering a restricted portion of a stationary time series with different contamination rates, i.e. percentage of outliers over total data. Similarly to [19], the results for *FPR* are not reported since *FPR* resulted lower than approximately 0.1% in all the analyzed cases.

Efficiency. The considered simulation scenario consists of a dynamic time series with outliers of different magnitudes (±3%, ±4%, ±5% and ±7% with respect to the mean value). Simulated outliers represent the 5% of data. The step magnitude during the transient is fixed at 10% and Gaussian noise at 1%. A sample of the simulated time series is presented in Figure 1.

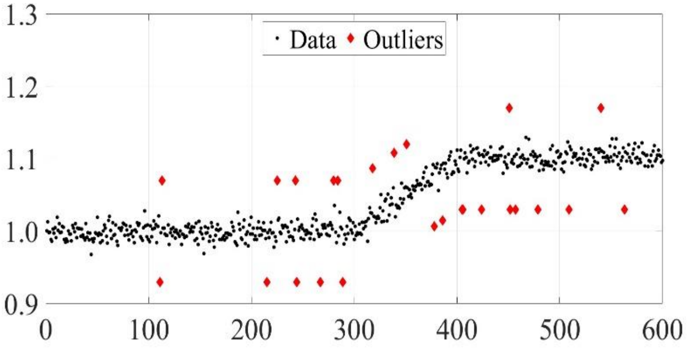

Figure 1 – Simulated time series for efficiency evaluation (1% noise; outlier magnitude equal to 7%)

Results regarding *TPR* and *FNR* are presented in Figure 2 and Figure 3. The *TPR* performance for the 7%, 5% and 4% scenarios are similar for all the considered methodologies, while differences are highlighted in the 3% magnitude scenario. In fact, the k-σ and the k-MAD methodologies achieve the highest *TPR*, which sets at 82% instead of 54% and 41% achieved by the bi-weight and the hybrid scheme, respectively.

The index *FNR* achieves satisfactorily low values for all the methodologies for outlier magnitudes greater than 4%. On the contrary, for outlier magnitude at 3%, *FNR* values considerably increase to 80%-95%. Across all the considered scenarios, the bi-weight methodology achieves the highest *FNR*, while the k-σ methodology proves the best and even allows *FNR* equal to 0 in the 5% outlier magnitude scenario.

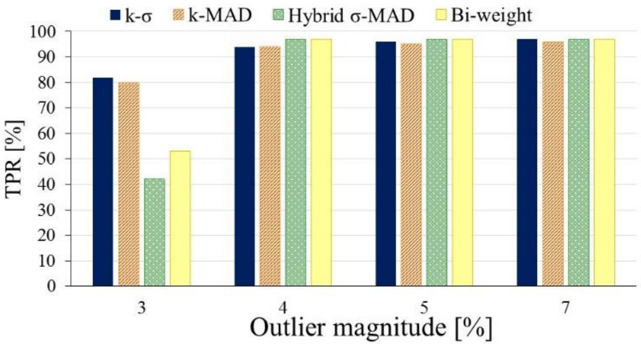

Figure 2 – *TPR* results for efficiency evaluation

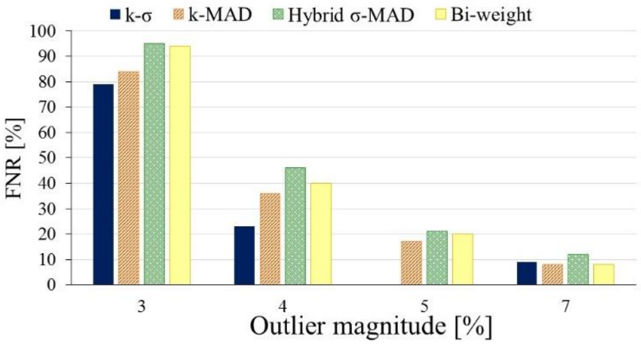

Figure 3 – *FNR* results for efficiency evaluation

Robustness. The considered simulation scenario is the same as the one considered for evaluating the efficiency, with the only difference that Gaussian noise is increased to 2%. A sample of the simulated time series is presented in Figure 4.

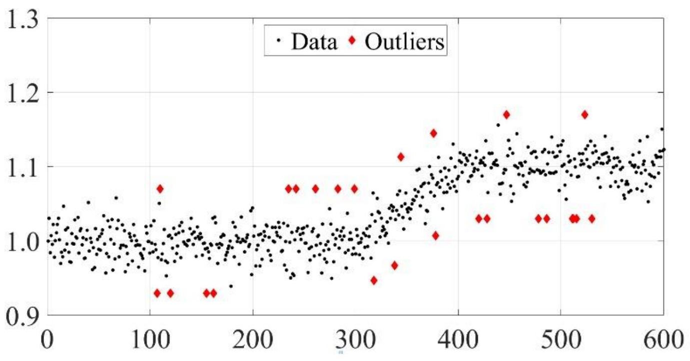

Figure 4 – Simulated time series for robustness evaluation (2% noise; outlier magnitude equal to 7%)

As it can be seen from Figure 5, the performance of all the methodologies is affected by the significant increase of measurement noise (2% in Figure 4 instead of 1% in Figure 1). In particular, the *FNR* is higher than 95% for all the considered methodologies already at 5% outlier magnitude, as shown in Figure 6. The k-σ methodology usually achieves the highest *TPR* values.

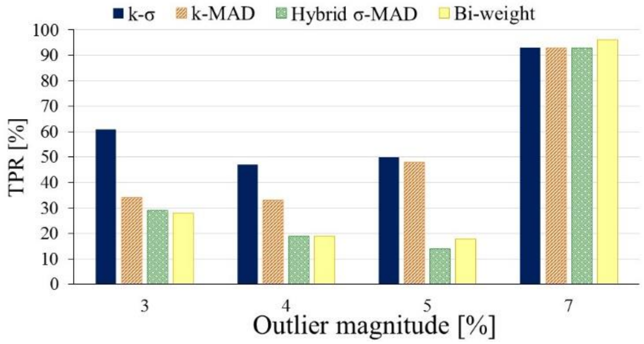

Figure 5 – *TPR* results for robustness evaluation

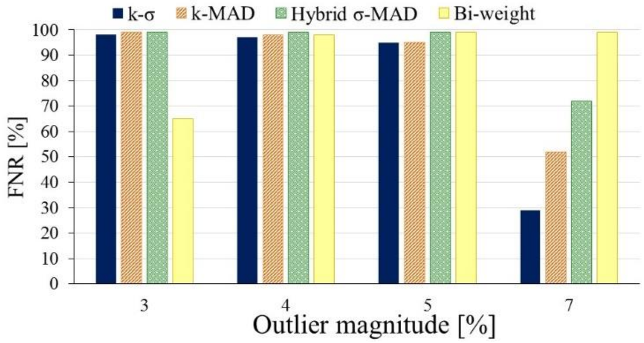

Figure 6 – *FNR* results for robustness evaluation

Resistance. The considered simulation scenario consists of a steady state time series with outlier of magnitude at 5%, imposed to be clustered in a data window of 100 observations. Different levels of contamination (5, 10, 25, 35 and 50 outliers out of 100 observations) are considered. Gaussian noise is assumed equal to 1%. A sample of the simulated time series is presented in Figure 7, with 50 imposed outliers. Results are reported in Figures 8 and 9. It can be seen that the performance of the k-σ methodology is very sensitive to the presence of clustered outliers, showing a marked decrease of *TPR* as the contamination rate increases.

Furthermore, unlike the previous cases, the k-σ methodology experiences the highest sensitivity of the *FNR* towards the contamination rate. Similar trends are observed for the hybrid scheme and the bi-weight methodology, with the latter performing better than hybrid scheme and k-σ methodology. Instead, the k-MAD methodology proves to be the most resistant towards (i.e. less affected by) clustered outliers among the evaluated methodologies, both in terms of *TPR* and *FNR*. The methodology holds its performance, achieving 95% *TPR* and 25% *FNR* until the contamination rate reaches 25%.

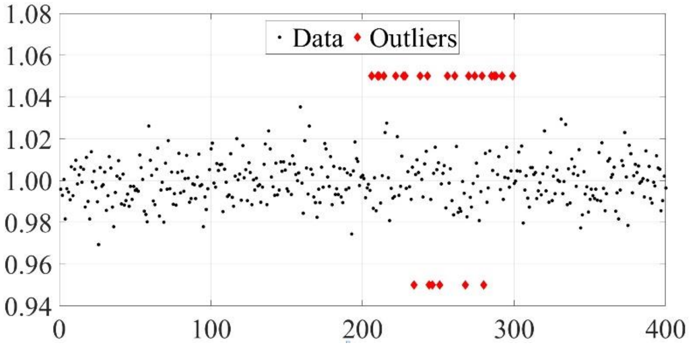

Figure 7 – Simulated time series for resistance evaluation (1% noise; outlier magnitude equal to 5%)

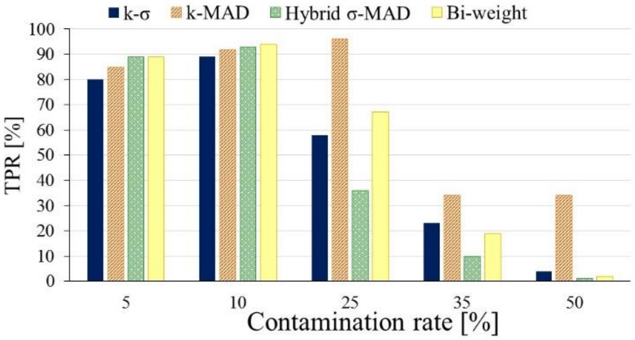

Figure 8 – *TPR* results for resistance evaluation

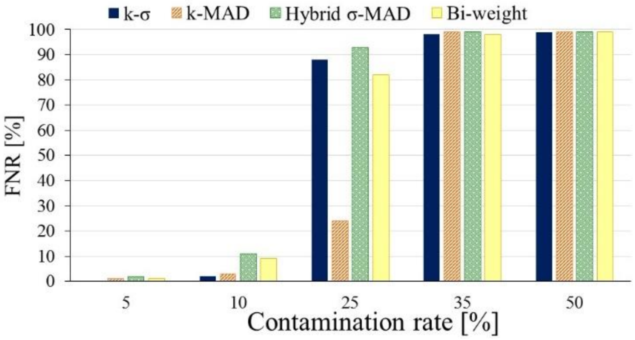

Figure 9 – *FNR* results for resistance evaluation

ASSESSMENT OF RESISTANT METHODOLOGIES BY MEANS OF FIELD DATA WITH INJECTED OUTLIERS

Even if simulated data were designed to be representative of actual operating conditions, they still represent an ideal environment, whose underlying assumptions may not be verified in field data. For this reason, the capabilities of the resistant methodologies are challenged by injecting outliers with controlled magnitude into Siemens field data sets, according to the approach adopted e.g. in [28]. In fact, as the position of anomalies is known, it is possible to evaluate *TPR, FNR* and *FPR* performance in a less controlled environment.

The magnitude of inserted outliers is expressed in terms of multiples *m* of the noise factor *s*, i.e. the SD of each steady state in the time series. Therefore, the value of the observation x_t, selected to be replaced with an outlier, is calculated according to the following expression:

$$x_{\mathrm{t\ outlier}} = x_{\mathrm{t}} + ms \qquad (9)$$

In order to establish a coherent comparison to the analysis on simulated data, field data sets are selected to reflect similar situations, i.e. set point change maneuver, relevant measurement noise and steady-state maneuver with clustered outliers. In order to avoid that injected outliers affect the quality of the analysis, the time series employed should be free from any previous anomaly. To this purpose, datasets were previously analyzed by means of visual inspection and application of engineering sense to verify that no evident outliers occurred. Moreover, it has to be pointed out that outliers are not injected during a transient, since, according to industry practice, the main goal is the evaluation of outlier identification capability at steady state conditions.

The capabilities of the methodologies are investigated in terms of efficiency and resistance by using time series with morphologies aimed at being meaningful for each analysis, in order to highlight benefits and possible drawbacks deriving from their application. Note that dynamic time series are normalized by using the mean value of the first steady state segment, while mean of all observations is used for stationary data sets.

The considered datasets and results are reported and discussed in details in the following. As can be seen, the performance of the methodologies is not strictly data dependent. In fact, even if values are different, the trends of both *TPR* and *FNR* indices are similar across the analyzed data sets, so that general considerations can be grasped from the analysis of results. The best performing methodology is the one that achieves the best balance between high *TPR* and low *FNR*. In fact, in the following, the results for *FPR* are not reported since *FPR* resulted lower than approximately 1% in all the analyzed cases.

Datasets used for efficiency analysis. The analysis is performed on the basis of four different datasets with measurements acquired from Siemens units. In compliance with the cases analyzed by using simulated data, dynamic time series with different levels of data dispersion are evaluated. Outliers are injected with different magnitudes (i.e. from 3% to 7%). Injected outliers represent the 5% of total observations.

The first time series, here referred as temperature T1, contains temperature measurements collected with a sampling frequency equal to 1 second. The time series consists of two different stationary states separated by a step change of about 7%. The SD is rather low; in fact, it is 0.005 and 0.001 for the two segments, respectively. A sample of the data with injected outliers is reported in Figure 10.

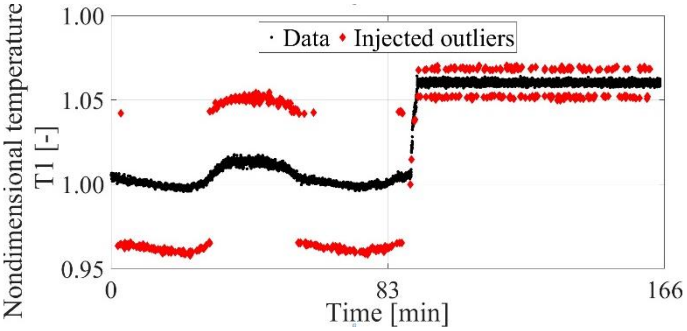

Figure 10 – Nondimensional temperature T1 dataset with 7% magnitude injected outliers

The second dataset, hereafter referred as temperature T2, contains temperature readings collected with 1 minute frequency. This dataset is considered particularly challenging for the methodologies. In fact, as it can be seen in Figure 11, four different segments can be identified, with SD 0.05, 0.02, 0.02 and 0.01 respectively. These are separated by two rapid transient maneuvers (at *t*=333 min and t=423 min) with a step change of -10% and +55%, respectively. Moreover, there is a “spike” at *t*=700 min with a 30% step.

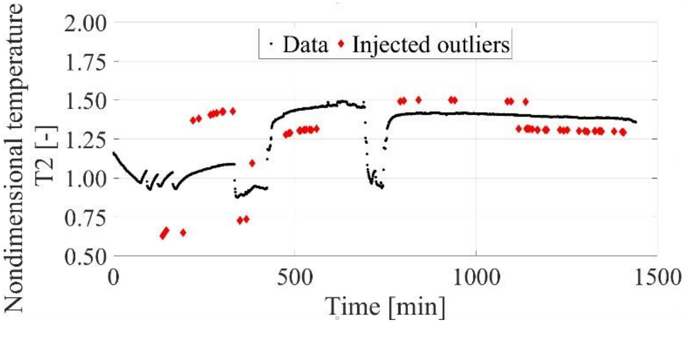

Figure 11 – Nondimensional temperature T2 dataset with 7% magnitude injected outliers

The third dataset consists of vibration measurements (i.e. displacement) collected with a 1 minute frequency. The nondimensional values of this time series, hereafter referred as vibration V1, together with injected outliers, are reported in Figure 12. Two subsequent steady states with SD 0.06 and 0.08 respectively can be identified. These are connected by a transient maneuver of +60% step change.

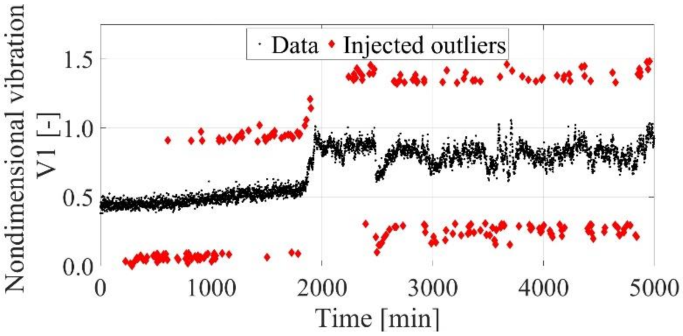

Figure 12 – Nondimensional vibration V1 dataset with 7% magnitude injected outliers

The fourth dataset, labelled as pressure P1 and reported in Figure 13, contains pressure measurements collected with a frequency of 1 minute. The time series consists of a steady state with significant noise. SD even reaches 0.6, which is one order of magnitude larger than in previous cases. For this reason, this dataset is considered particularly challenging.

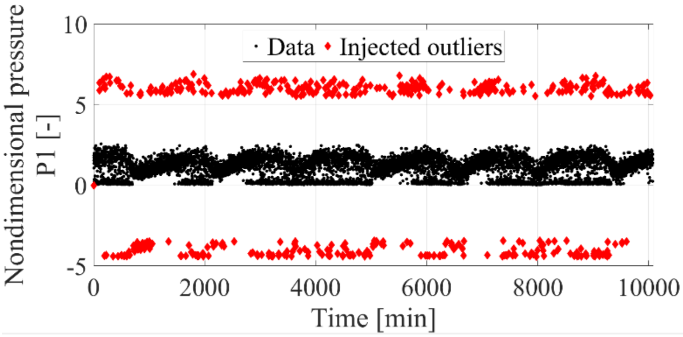

Figure 13 – Nondimensional pressure P1 dataset with 7% magnitude injected outliers

Datasets used for resistance analysis. Similarly to the approach adopted for simulated data, outliers with magnitude of 5% are injected by contaminating a portion (i.e. 25%) of the stationary time series with different rates (i.e. 5%, 10%, 25%, 35% and 50%) of outliers. Since the time series is almost stationary, particular attention is focused on data dispersion (expressed in terms of SD) to evaluate its influence on detection capabilities with clustered outliers. Four different datasets with measurements acquired from Siemens units are considered for this analysis.

The first dataset contains temperatures collected with a sampling frequency of 1 minute and is labelled as temperature T3. The SD value is 0.001. Outliers are injected in a time frame composed of 2500 time points, ranging from t = 42 min to t = 83 min. The normalized time series, together with injected outliers, is reported in Figure 14.

The second time series is composed of measurements of rotational speed, collected at the frequency of 1 second. This dataset is hereafter referred as rotational speed S1. Data noise is rather low; in fact SD is just 0.002. Outliers are injected in a 1000 time points window, between *t* = 33 min and *t* = 50 min, as shown in Figure 15.

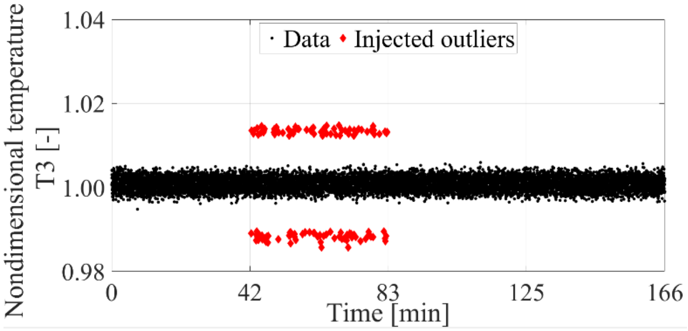

Figure 14 – Nondimensional temperature T3 dataset with 5% magnitude injected outliers

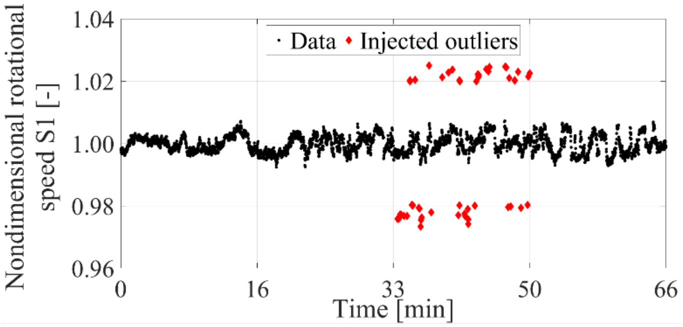

Figure 15 – Nondimensional rotational speed S1 dataset with 5% magnitude injected outliers

The third dataset, labelled as temperature T4, includes temperature measurements taken with a frequency of 1 s. Data dispersion is one order of magnitude larger than in the previous cases, as SD reaches 0.01. Outliers injection occurs in a window of 195 time points, starting at *t* = 8 min, as shown in Figure 16.

Finally, the same dataset pressure P1, previously employed for efficiency evaluation (see Figure 13), is also used here to assess statistical resistance. For this purpose, outliers are inserted in a window of 2500 time point, ranging from *t* = 4000 min to t = 6500 min. Moreover, as for all the cases considered in this section (i.e. Figures 14, 15 and 16), outliers are injected with 5% magnitude. The nondimensional time series with injected outliers is presented in Figure 17.

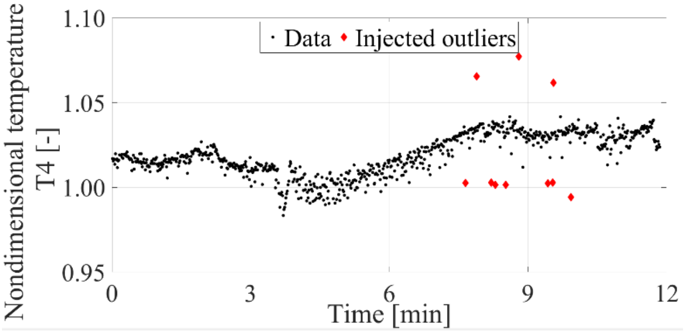

Figure 16 – Nondimensional temperature T4 dataset with 5% magnitude injected outliers

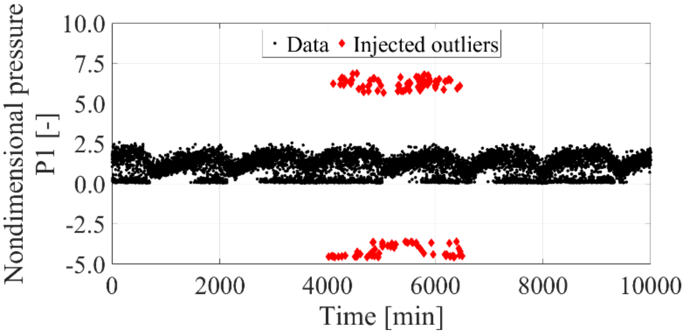

Figure 17 – Nondimensional pressure P1 dataset with 5% magnitude injected outliers

Efficiency analysis. In general, as shown in Figures 18 through 21, The k-σ methodology achieves the best *TPR* performance for each outlier magnitude scenario for all the considered datasets. Namely, the *TPR* values range from 84% to 99%, maintaining high performance levels even in the most challenging scenario, i.e. when the outlier magnitude reaches 3%. This is in fact the scenario in which the methodology most clearly overcomes the others, as in the cases of the temperature dataset T2 (Figure 19) and the pressure dataset P1 (Figure 21). For these datasets, the k-σ methodology overcomes the hybrid k-MAD methodology by 30% and 15%, respectively. The values of the *TPR* index achieved by the k-σ methodology prove its low sensitivity towards the decrease of the outlier magnitude, meaning a certain stability in detection performance over different operational scenarios. The hybrid k-MAD methodology follows the k-σ in terms of *TPR*, achieving values of this index in the range from 55% to 90%. The k-MAD and the bi-weight methodologies achieve similar performance in terms of *TPR*, but lower than that of the hybrid scheme and k-σ methodologies.

The temperature dataset T2 proves to be the most challenging. Only the k-σ methodology is able to maintain its performance, experiencing a 10% decrease of the *TPR* values with respect to less challenging time series. The same decrease sets at 30% for the hybrid methodology and at 40% for the bi-weight and k-MAD methodologies. This gap in *TPR* performance between the k-σ and the other methodologies is caused by the capability of the k-σ methodology to identify few false positive calls, of which the increase makes the value of *TPR* decrease according to Eq. (6). The low sampling frequency (1 minute), combined with the rapidness and the considerable step change of the transient in the T2 dataset, creates a consistent gap in subsequent measurements during the transient, thus increasing the probability for observations to trespass the acceptability thresholds imposed by the parametric test. The k-σ methodology can contrast this phenomenon better than the other methodologies.

The *FNR* values can be considered satisfactorily low in all datasets for all the methodologies. The highest *FNR* values are achieved by the k-σ methodology with the widest gap with respect to the other methodologies experienced in the case of the T1 and T2 datasets. In any case, the maximum value is 38% in the 3% outlier scenario, which is rather satisfactory. The other methodologies perform better than the k-σ methodology in terms of *FNR*, with values lower than 25%.

The *FNR* results, combined with the considerations on false positives calls, highlight the key features of the k-σ methodology, i.e. capability to detect true positives and tendency to avoid false positive calls. Therefore, false negative rates become higher as the outlier magnitude decreases. Even if rather far from k-σ methodology performance, the hybrid scheme methodology also proves to be efficient. Finally, the k-MAD and the bi-weight methodologies prove to be the least efficient.

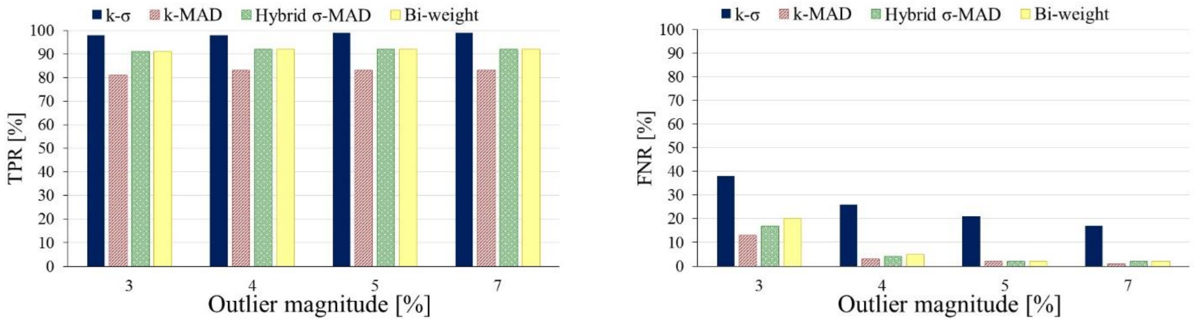

Figure 18 – *TPR* and *FNR* values as a function of outlier magnitude for the temperature T1 dataset

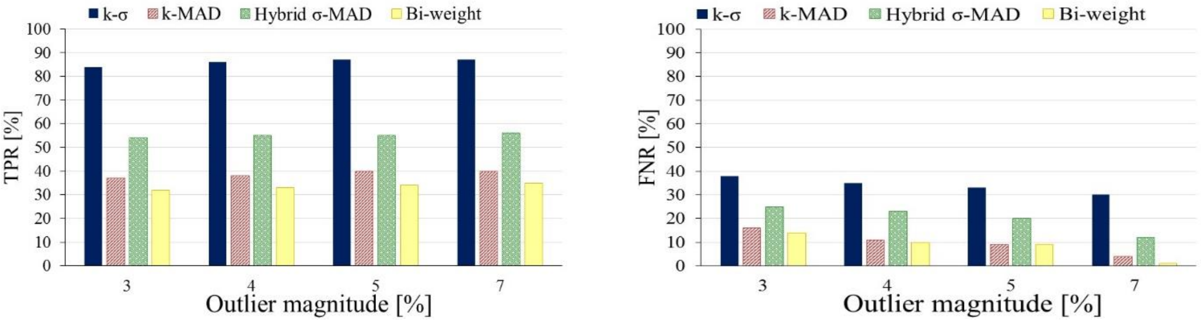

Figure 19 – *TPR* and *FNR* values as a function of outlier magnitude for the temperature T2 dataset

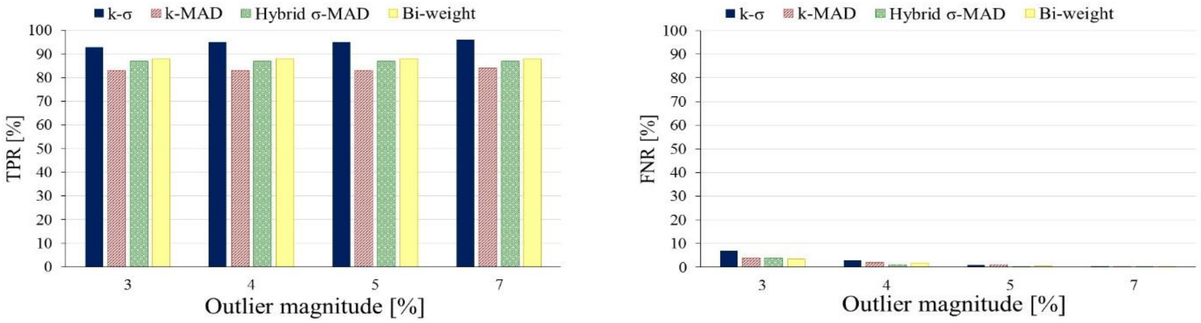

Figure 20 – *TPR* and *FNR* values as a function of outlier magnitude for the vibration V1 dataset

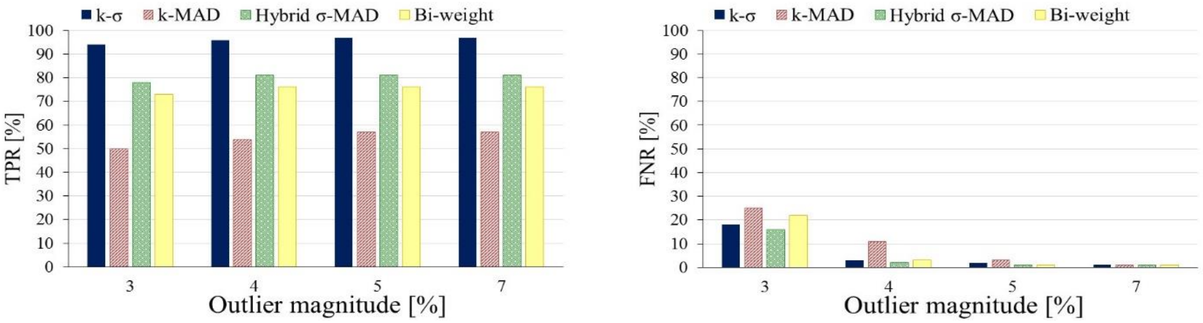

Figure 21 – *TPR* and *FNR* values as a function of outlier magnitude for the pressure P1 dataset

Resistance analysis. Since all in all the performance of the methodologies as a function of the contamination rate follows a similar pattern for all the different time series, as shown in Figures 22 through 25, it is possible to evaluate the resistance of the methodologies on the whole.

The k-σ methodology proves to be effective for outlier detection with low contaminated datasets. According to the results of the efficiency analysis, the k-σ methodology is the best methodology when the contamination rate sets at 5% for 3 out of 4 datasets, achieving the highest *TPR* (between 85% and 40%) with satisfactorily low *FNR* (between 20% and 1%). The exception is the temperature T3 dataset, in which all the methodologies achieve 98% *TPR*, but the *FNR* allowed by the k-σ methodology sets at 25%, instead of the almost null *FNR* of all the other methodologies.

However, the detection capabilities of the k-σ methodology progressively decrease as the contamination rate increases. This performance decrease can be noticed by a sensible increase of the *FNR*. In fact, the values of this index almost triplicate as the contamination rate passes from 5% to 10%, reaching very high values for the S1 and T3 datasets (60% and 70% respectively). For all the considered datasets, the detection capabilities of the k-σ methodology are almost null as the contamination rate reaches 25%, with the *FNR* at 95%. Therefore, the k-σ methodology proves to be the most affected by an increase of the contamination rate.

On the contrary, the k-MAD methodology demonstrates high effectiveness towards severely contaminated datasets. The detection capabilities achieved by the k-MAD methodology are enhanced as the contamination rate increases, reaching its performance peak in the 35% contamination rate scenario. This turns into an extremely high number of outliers identified in the field time series. At low contamination rates, e.g. 5% for all datasets and 10% for the sole case of the T4 dataset, the performance of the k-MAD methodology is sensibly lower in terms of *TPR* with respect to the other methodologies. However, the k-MAD methodology is the most suitable to allow *FNR* lower than 4%.

According to the definition of the performance indices, the lower *TPR* and the extremely low *FNR* are symptoms of a higher number of false positive calls, i.e. higher *FPR*. The k-MAD methodology remarkably overcomes the other methodologies as the contamination rate reaches 35%. Even in this very challenging scenario, the k-MAD *TPR* is always higher than 60% and reaches 99%, e.g. in the case of the T3 dataset.

Regarding *FNR* performance, the k-MAD methodology overcomes the others, while the bi-weight methodology is the second best. The performance of the bi-weight methodology in fact follows a trend in terms of *TPR*, similarly to the k-MAD methodology, but at the cost of a usually higher *FNR*. The hybrid scheme achieves *TPR* values close to, but slightly lower than, those of the bi-weight methodology, combined with higher *FNR*.

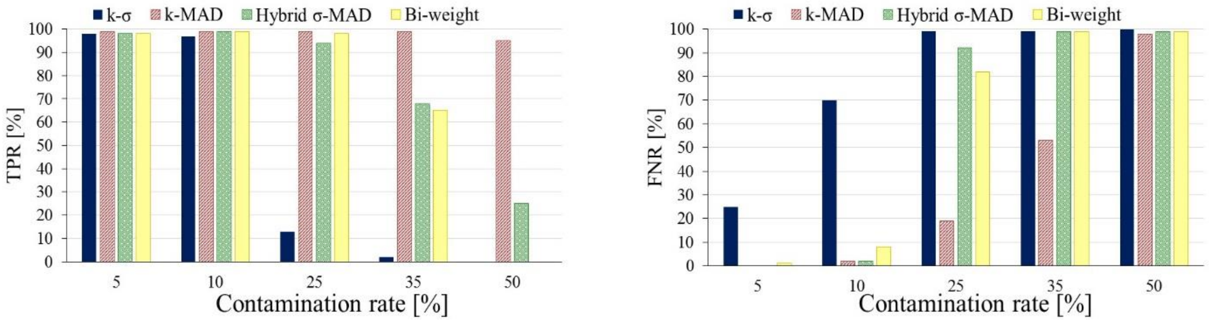

Figure 22 – *TPR* and *FNR* values as a function of outlier contamination rate for the temperature T3 dataset

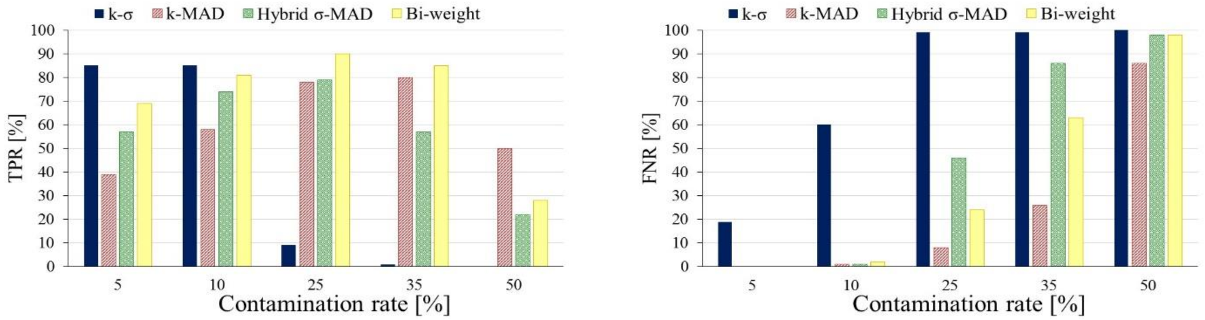

Figure 23 – *TPR* and *FNR* values as a function of the presence of outlier contamination rate for the rotational speed S1 dataset

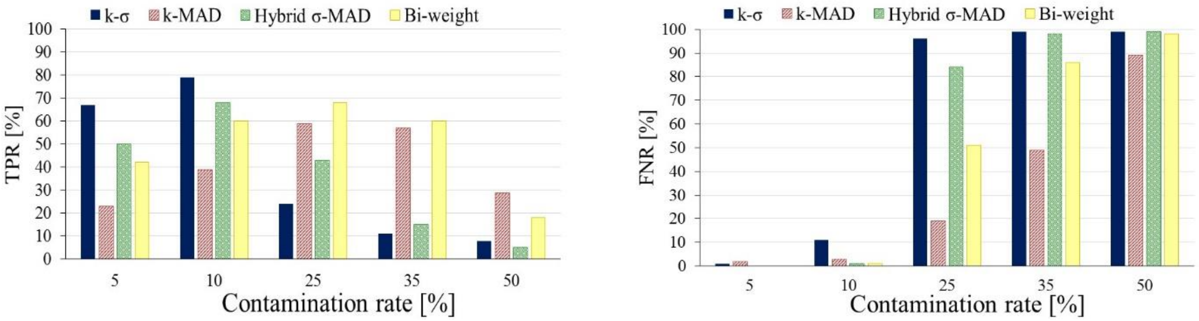

Figure 24 – *TPR* and *FNR* values as a function of the presence of outlier contamination rate for the temperature T4 dataset

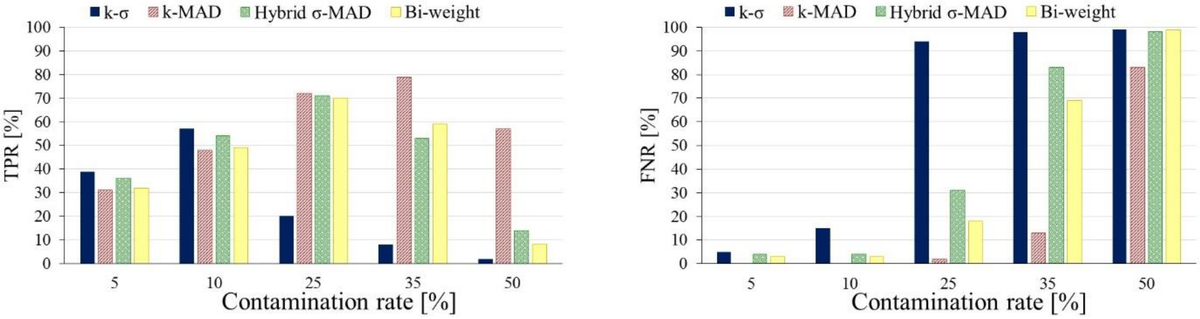

Figure 25 – *TPR* and *FNR* values as a function of the presence of outlier contamination rate for the pressure P1 dataset

Discussion. On the basis of the analysis performed in this paper on seven experimental datasets, the main advantages and drawbacks of each methodology can be summarized as in Table 2. Given the variety of considered measurements (4 temperature datasets, 1 vibration dataset, 1 pressure dataset and 1 rotational speed dataset), the indications reported in Table 2 can be seen as guidelines for implementing the considered methodologies.

The k-σ methodology is mostly suitable for datasets with low contamination rate, in which it is important not to lose information due to false positives. Instead, the k-MAD methodology can be applied to datasets with high contamination rate, in which it is important to improve data quality despite the risk of losing information due to false positives.

Finally, the hybrid σ-MAD and bi-weight methodologies can be recommended for datasets in which false detection, and its consequent loss of information, represents an issue, but a safeguard on the potential presence of clustered outliers is desired. To discriminate between these two methodologies, it can be observed that the capability of the hybrid σ-MAD methodology is closer to that of the k-σ methodology, while the capability of the bi-weight methodology is closer to that of the k-MAD methodology.

Table 2 – Summary of advantages and drawbacks of the resistant statistical methodologies

Methodology	**Advantage**	**Drawback**
k-σ	Highest statistical efficiency. Highest *TPR* and lowest *FPR* with 5% contamination rate.	Lowest statistical resistance. Detection performance severely affected by the presence of clustered outliers.
k-MAD	Highest statistical resistance. Best detection capabilities in presence of clustered outliers. Almost null *FNR* at 10% contamination rate.	Highest *FPR* with 5% contamination rate. Lower statistical efficiency than k-σ methodology.
Hybrid σ-MAD	Statistical resistance higher than k-σ methodology. Better detection capabilities in presence of clustered outliers.	Lower statistical efficiency than k-σ methodology. Lower *TPR* and higher *FPR*.
Bi-weight	Same as hybrid σ-MAD methodology.	Same as hybrid σ-MAD methodology, but requires the tuning of a parameter specific to this methodology.

CONCLUSIONS

The backward and forward moving window (BFMW) k-σ methodology usually proves to be somewhat effective in managing dynamic time series, but its estimators lack statistical resistance and robustness. Therefore, by maintaining the same BFMW structure, three different methodologies with robust statistical estimators were developed in this paper by implementing the k-σ test criterion, i.e. the k-MAD methodology, a hybrid scheme called σ-MAD methodology and the bi-weight methodology.

Results prove the effectiveness of the application of the parametric test criterion as a two-sided moving window scheme. Moreover, all the considered methodologies prove to be effective. Therefore, this paper provides a portfolio of four different methodologies and the most proficient scenario of application is identified for each of them, together with guidelines to efficiently perform anomaly identification in data time series in a comprehensive range of operational scenarios.

In particular, the k-σ methodology proves its effectiveness for outlier detection of gas turbine measurements, both in case of dynamic data sets and high noise. In fact, this methodology combines high detection capabilities with low sensitivity towards outlier magnitude.

The *TPR* values achieved by the k-σ methodology on four different field datasets, range from 99% to 84% as outlier magnitude passes from 7% to 3%. Even when its detection capabilities decrease due to the complexity of data, this methodology still overcomes the others in terms of *FPR*. Furthermore, the *FNR* results confirm that this methodology tends not to detect anomalies rather than produce false positives. This is an important feature if a reliable outlier detection method characterized by a simple tuning procedure is desired. However, the limits of the methodology to analyze severely corrupted data sets are evident. Results suggest that when 25% of data is corrupted, the detection capability of the k-σ methodology is null and many false negatives are highlighted. Therefore, when the dataset is known to be potentially severely corrupted, it is suggested to support the k-σ methodology with additional levels of detection to improve *TPR*.

The k-MAD methodology allowed a lower detection efficiency both on simulated and field data, together with sensibly higher *FPR*s. Instead, the k-MAD methodology is characterized by exceptional resistance towards outliers, i.e. this methodology is the most suitable in case severely corrupted datasets are expected. The methodology, in fact, significantly overcomes the others in terms of resistance, achieving *TPR* levels between 99% and 60% at 25% contamination rate for all the analyzed datasets. Furthermore, the k-MAD methodology is the only methodology that achieves *TPR*s higher than 50% when contamination reaches 35%. Results in terms of *FNR* improve as well together with the contamination of the dataset, maintaining values close to 0 at 10% contamination rate and in any case lower than 30% as the contamination rate reaches 25%. Therefore, the k-MAD methodology represents a desirable solution if the aim is to clean as many data as possible and provide a neater trend regardless of the possible loss of information.

The hybrid σ-MAD and bi-weight methodologies proved intermediate performance between the k-σ and the k-MAD methodologies. In particular, the characteristics of the hybrid scheme are closer to those of the k-σ methodology, while the bi-weight methodology behaves similarly to the k-MAD methodology. The hybrid σ-MAD and the bi-weight methodologies achieve higher detection efficiencies than the k-MAD methodology, combined with low *FNR* values, which can be usually achieved by the k-σ methodology. Therefore, these solutions are suitable when false detection and consequent loss of information represent an issue, but a safeguard on the potential presence of clustered outliers is desired. Among the two, tuning the hybrid σ-MAD methodology is easier to perform and more generally applicable than tuning the bi-weight methodology, of which performance is also influenced by the value of the bi-weight parameter.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Siemens for the permission to publish the results.

REFERENCES

[1] Roumeliotis, I., Aretakis, N., Alexiou, A., 2016, “Industrial Gas Turbine Health and Performance Assessment with Field Data”, Proceedings of ASME Turbo Expo 2016, GT2016-57722

[2] Tsoutsanis, E., Meskin, N., Benammar, M., Khashayar K., 2015, “Transient Gas Turbine Performance Diagnostics Through Nonlinear Adaptation of Compressor and Turbine Maps”, *J Eng Gas Turb Power* **137**(9), 091201 (Sep 01, 2015) (12 pages)

[3] Jiang , X., Foster, C., 2014, “Plant Performance Monitoring and Diagnostics – Remote, Real-Time and Automation”, Proceedings of ASME Turbo Expo 2014, GT2014- 27314

[4] Venturini, M., Therkorn, D., 2013, “Application of a Statistical Methodology for Gas Turbine Degradation Prognostics to Alstom Field Data”, *J Eng Gas Turb Power* **135**(9), 091603 (10 pages), doi:10.1115/1.4024952.

[5] Cavarzere, A., Venturini, M., 2011, “Application of Forecasting Methodologies to Predict Gas Turbine Behavior Over Time” *J Eng Gas Turb Power* **134**(1), 012401 (8 pages)

[6] Lanzante, J., 1996, "Resistant, Robust and Non-Parametric Techniques for the Analysis of Climate Data: Theory and Examples, Including Applications to Historical Radiosonde Station Data", International Journal of Climatology, **16**(11), pp. 1197-1226.

[7] Van Paridon, A., Bacic, M., Ireland, P. T., 2016, “Kalman Filter Development For Real Time Proper Orthogonal Decomposition Disc Temperature Model”, Proceedings of ASME Turbo Expo 2016, GT2016- 56330

[8] Hurst, A. M., Carter, S., Firth, D., Szary, A., Van De Weert, J., 2015, “Real-Time, Advanced Electrical Filtering for Pressure Transducer Frequency Response Correction”, Proceedings of ASME Turbo Expo 2015, GT2015-42895.

[9] Gutierrez , L. A., Pezzini, P., Tucker, D., Banta, L., 2014, “Smoothing Techniques For Real-Time Turbine Speed Sensors”, Proceedings of ASME Turbo Expo 2014, GT2014-25407

[10] Dewallef, P. and Borguet, S., 2013, "A Methodology to Improve the Robustness of Gas Turbine Engine Performance Monitoring Against Sensor Faults", *J Eng Gas Turb Power* **135**(5), p. 051601.

[11] Gomez, J., 2011, Kalman filtering, Nova Science Publishers, Hauppauge, N.Y.

[12] Pinelli, M., Venturini, M., and Burgio, M., 2003, "Statistical Methodologies for Reliability Assessment of Gas Turbine Measurements ", Proceedings of the Turbine Technical Conference and Exposition, (ASME Paper GT2003-38407), pp. 787-793.

[13] Martin, R. and Thomson, D., 1982, "Robust-resistant spectrum estimation", Proceedings of the IEEE, 70(9), pp. 1097-1115.

[14] Hampel, F., 1974, "The Influence Curve and its Role in Robust Estimation", *Journal of the American Statistical Association*, **69**(346), pp. 383-393.

[15] Bhattacharya, G., Ghosh, K., and Chowdhury, A., 2015, "Outlier detection using neighborhood rank difference", Pattern Recognition Letters, 60-61, pp. 24-31.

[16] Takahashi, T., Tomioka, R., and Yamanishi, K., 2014, "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection", *IEEE Trans. Knowl. Data Eng*., **26**(1), pp. 120-130.

[17] Takeuchi, J. and Yamanishi, K., 2006, "A unifying framework for detecting outliers and change points from time series", *IEEE Trans. Knowl. Data Eng*., **18**(4), pp. 482-492.

[18] Xu, S., Baldea, M., Edgar, T., Wojsznis, W., Blevins, T., and Nixon, M., 2014, "An improved methodology for outlier detection in dynamic datasets", *AIChE Journal*, **61**(2), pp. 419-433.

[19] Ceschini, G., Gatta, N., Venturini, M., Hubauer, T., Murarasu, A., 2017, “Optimization of Statistical Methodologies for Anomaly Detection in Gas Turbine Dynamic Time Series”, ASME Paper GT2017-63409.

[20] Ceschini, G., Gatta, N., Venturini, M., Hubauer, T., Murarasu, A., 2017, “A Comprehensive Tool for Detection, Classification and Integrated Diagnostics of Gas Turbine Sensors (DCIDS)”, ASME Paper GT2017-63411.

[21] Huber, P., 1981, Robust statistics, Wiley, New York.

[22] Rousseeuw, P. and Croux, C., 1993, "Alternatives to the Median Absolute Deviation", *Journal of the American Statistical Association*, **88**(424), p. 1273.

[23] Kafadar, K., 1983, "The Efficiency of the Biweight as a Robust Estimator of Location", *J. Res. Natl. Bur. Stan*., **88**(2), p. 105.

[24] Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L., 2013, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median", *Journal of Experimental Social Psychology*, **49**(4), pp. 764-766.

[25] Yahaya, S., Othman, A. and Keselman, H., 2004, “Testing the Equality of Location Parameters for Skewed Distributions Using S1 with High Breakdown Robust Scale Estirnators”, *Statistics for Industry and Technology*, pp.319-328.

[26] Hoaglin, D., Mosteller, F., and Tukey, J., 1983, Understanding robust and exploratory data analysis, Wiley, New York.

[27] Young, G. A. and Smith, R. L., 2005, Essential of Statistical Inference, Cambridge University Press, Cambridge, UK.

[28] Sharma, A., Golubchik, L., and Govindan, R., 2010, "Sensor faults: detection methods and prevalence in real-world datasets", *ACM Transactions on Sensor Networks*, **6**(3), pp. 1-39.

List of figure captions

List of table captions