Proceedings of the Canadian Society for Mechanical Engineering International Congress
32nd Annual Conference of the Computational Fluid Dynamics Society of Canada
Canadian Society of Rheology Symposium
CSME/CFDSC/CSR 2025
May 25–28, 2025, Montréal, Québec, Canada

Adaptive Filtering Techniques for Self-Detecting Outliers in Signals: A Case Study of Runner Strain Measurements

Quang Hung Pham[1]*, Martin Gagnon[1]

[1]Institut de recherche d'Hydro-Québec (IREQ), Varennes, Québec, Canada
*Pham.QuangHung2@hydroquebec.com

***Abstract*—**The presence of outliers in a signal measured by a physical sensor is sometimes unavoidable. This is particularly true if the sensor is installed in a harsh environment, where distinguishing between outliers and the physical stochastic phenomenon of interest becomes challenging. The problem is that we generally do not have access to the ground truth, making it impossible to determine whether there are outliers. In this paper, we propose an adaptive filter that enables self-detection of potential outliers. This filter is created by applying the Local Outlier Factor (LOF) method in a 3D-space generated from the 1D time series, using its derivatives. The methodology is verified on strain gauge measurement data from hydroelectric turbine blades.

***Keywords-component*—Adaptive Filter; Outlier; Local Outlier Factor; Runner Strain**

I. INTRODUCTION

Data obtained from physical sensors installed directly on a site of interest generally provides the most reliable information for further analysis (diagnostics, prognostics, etc.). However, this is true only if the measurements are of an acceptable quality and free from abnormal behaviours such as outliers related to unwanted noise, saturated values, aberrations, etc. Quality control for sensor data remains challenging, especially in industrial applications. For example, certain factors such as environmental noise cannot be completely removed, and human factors are sometimes unavoidable during installation, data acquisition, etc.

Consequently, a strategy to perform posteriori sensor data cleaning, such as outlier detection, is needed. Such a method would function as a filter that detects and then removes or replaces outliers before the dataset can undergo further analysis. Theoretically, outliers can be defined as data points that do not follow the typical behaviour of a dataset. In their review, Blázquez-García et al. (2021) [1] divide outliers into two main groups, based on the interest of analysis: unwanted data and events of interest. Unwanted data are abnormal points in the timeseries that are useless for further analysis. Events of interest, on the other hand, can in some cases also be abnormal points (e.g., fraud detection). This raises the question of how to identify abnormal points in a dataset when there are no clues about the phenomenon of interest behind them. In particular, for sensor data that exhibit strong stochastic behaviour such as wind speed measurements and runner strain measurements in water [2], outliers and extreme values resulting from physical phenomena may be difficult to distinguish.

In this study, we would like to highlight this point by means of a case study using hydroelectric turbine measurements and illustrate the performance of the proposed solution to filter outliers. In our case, the "true" cleaned data is unknown or outside our current knowledge. Thus, our goal is to have an adaptive filter for self-detecting outliers that is based only on the behaviour of studied data themselves. A non-supervised outlier detection technique known as the Local Outlier Factor (LOF) method is used [3]. This method detects points located far from their neighbour points in a multidimensional space. The space is constructed based on the measured values themselves, combined with their first and second derivatives. Performance is evaluated based on two criteria to assess the filtered data's reliability in both the time and frequency domains:

- ratio of detected outliers to the total number of measured values,
- distortion levels in the frequency domain between before and after the filtration.

The paper is organized as follows: Section II introduces the studied data and their stochastic behaviour. Then, the proposed filter is described in Section III. Finally, the results regarding the performance of the proposed filter are presented in Section IV.

II. STUDIED DATA

The studied dataset consists of strain signals measured at several steady-state operating conditions of a Francis turbine operated by Hydro-Québec. Strain measurements on runners are reliable data points that allow for a more realistic fatigue life estimate [4]. However, due to the harsh environment and high cost, these measurements are often available only for a limited time during the measurement campaign. The water flows around the blades generate stochastic dynamic behaviours, especially in operating conditions with partial loads that are far from the best efficiency point (BEP). This stochasticity is represented by the highly fluctuating peaks in the measurement, which can sometimes be indistinguishable from abnormal values. Figure 1 shows sample measurements: In the top figure, the variations between the high peaks could lead an observer to conclude that these represent abnormal data, unlike in the bottom figure, where doing so might be difficult. Another challenge is that we do not have any idea what the expected clean measurements may be.

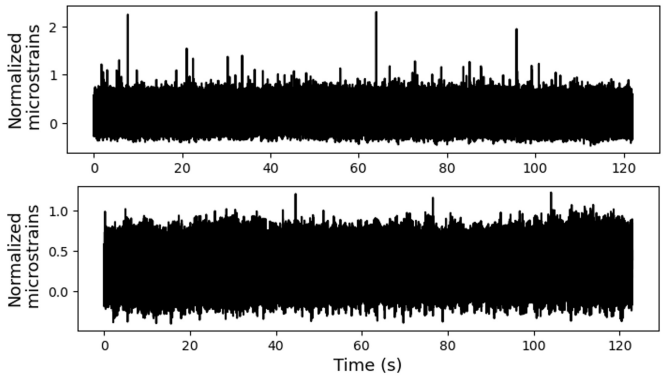

Figure. 1. Examples of runner strain signals measured by strain gauge.

III. METHODOLOGY

A. Local outlier factor method (LOF)

LOF uses unsupervised machine learning approach to assess a point's degree of isolation by comparing its local density to that of its neighbours. The idea is that points with a significantly lower local density than their neighbours are considered anomalies. This local density is calculated by determining the average distance between this point and its k nearest neighbours. This approach allows for the detection of local anomalies rather than global ones. For more details on LOF, please refer to [3].

Let $d(i,j)$ denote the distance between two points i and j. The main steps are as follows:

i. For each point, the distance $k_{dist}(i)$ to their k-nearest neighbours is calculated as the smallest radius around i that encloses at least k other points in the dataset.

ii. The reachable distance $d_{reach}(i,j)$ is determined by:

$$d_{reach}(i,j) = max\left\{k_{dist}(j), d(i,j)\right\} \tag{1}$$

By assigning $k_{dist}(j)$ for points i that are "very" close to j, this proposed reachable distance ensures that these very close points do not bias the density measurements.

iii. Local density is estimated by inverting the averaged distance:

$$d_{\text{local}}(i) = \frac{1}{\left(\frac{\sum_{o\in N_{\text{min}}(i)} d_{\text{reach}}(i,j)}{|N_{\text{min}}(i)|}\right)} \tag{2}$$

where $N_{min}(i)$ represents the set of neighbouring points of i, with a required minimum number.

iv. Finally, the outlier score lof is calculated as the ratio of the average local density of a point's neighbours to the local density of the point itself. If the score is greater than 1, the point is more likely to be an anomaly.

$$\text{lof}(i) = \frac{\sum_{o\in N_{\text{min}}} \left(\frac{d_{\text{local}}(i)}{d_{\text{local}}(j)}\right)}{|N_{\text{min}}(i)|} \tag{3}$$

However, the LOF method only provides a score for each point in a dataset. To determine whether a point is an anomaly, its outlier score must exceed a fixed threshold. This threshold is identified by trial, taking overall performance into account.

B. Proposed filter

Theoretically, the LOF method is applicable only to a multidimensional dataset (a set with at least two dimensions), while in our case we have only a unique signal. Consequently, a step is required to transform the target signal x into a multidimensional space. In this paper, we have chosen a 3D space composed of the timeseries value x, its first derivative dx, and its second derivative d^2x. The use of this space is inspired by Goring et al. (2002), who represented the signal by an ellipsoid in the (x, dx, d^2x) space and considered all points outside this ellipsoid as outliers [5]. As mentioned earlier, the actual cleaned measurements are beyond our knowledge in this case. Therefore, the only reliable factor we can rely on is the behaviour of the measurement itself, i.e., the change in value during measurement. Our hypothesis is that if there is an extremely abrupt change in a measurement point compared to its neighbours, that point is potentially an anomaly. Thus, applying LOF in a (x, dx, d^2x) space allows us to detect outliers more effectively by analyzing not only the change in values but also the rate of change.

Figure 2 illustrates the proposed filter. After transforming the signal into the 3D space (x, dx, d^2x), the LOF method is applied with a fixed threshold and a given number of neighbouring points k. In this way, the potential anomalies are identified (see Figure 2). A count of the detected outlier points is also displayed to ensure that we do not remove an excessive number of points, ensuring the smallest possible distortion in the time domain after removal/replacement. These anomalies are then eliminated and re-estimated using a linear interpolation based on the remaining signal data points.

Evaluating the performance of this filtering process is challenging due to the absence of ground truth. As mentioned above, in this case, we propose an analysis of the difference

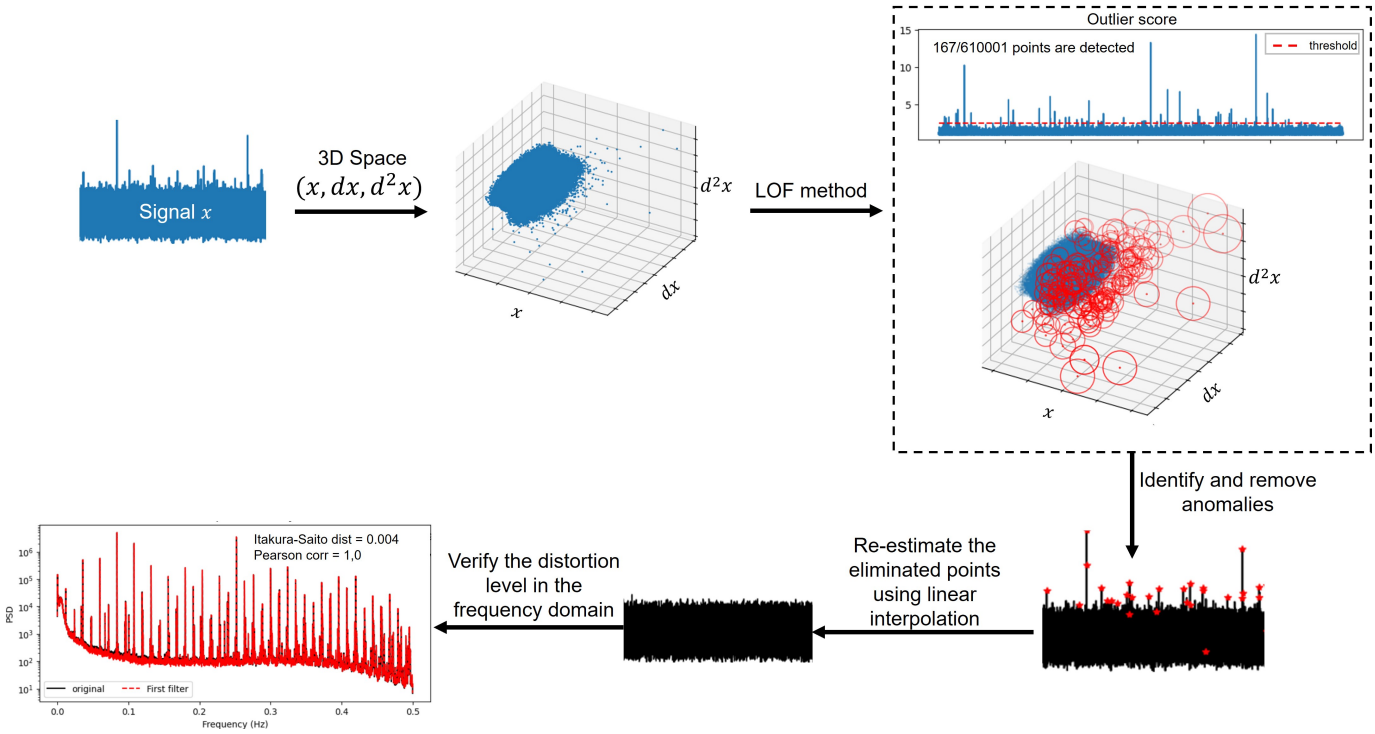

Figure. 2. Overview of methodology for the proposed filter.

between the original measurement and the filtered signal in the frequency domain. Metrics such as the Itakura-Saito distance and Pearson's correlation are used to evaluate the distortion between the power spectral densities (PSD) before and after filtering. This ensures that all important phenomena are present in the filtered signal without any major distortions.

C. Removal of outliers (optional step)

The proposed methodology allows us to detect potential anomalies that exhibit unusual behaviour in a measurement. It can also identify anomalies that do not correspond to an obvious visible peak in the time domain, which we refer to here as non-extreme anomalies. Figure 3 shows an example of this type of outlier. If the change in the amplitude of x is small, but dx and d^2x change significantly, the result is a high anomaly score. For this reason, this approach also identifies non-extreme values within the signal distribution (see Figure 3). Generally, it is not necessary to remove outliers of this type, even though they may be actual anomalies. Therefore, to minimize the number of points to be replaced after filtering, we add an optional step that cleans all anomalies within a 99% confidence interval (CI 99%). The idea is that only anomalies outside this confidence interval will be considered for removal and re-estimation (see Figure 3). Note that this CI 99% level is determined non-parametrically on empirical signal distribution without the estimated outlier in order to avoid the potential influence of abnormally high values.

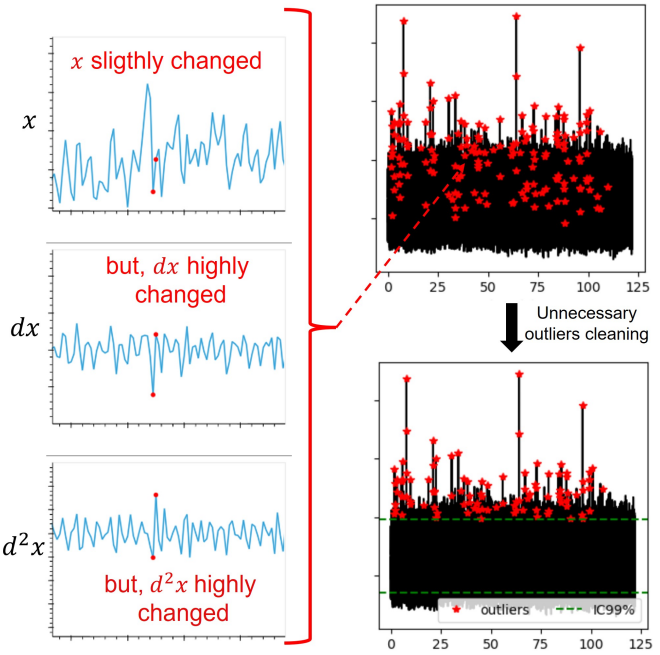

Figure. 3. Optional step to clean unnecessary filtered points. The lefthand column shows an example of a non-extreme anomaly.

IV. RESULTS AND DISCUSSIONS

In this section, several concrete examples are presented to demonstrate the performance of this filter. It should be noted that the thresholds of 2.5 and 1,000 neighbour points (for k-

distance) have been set for this case study.

Figure 4 shows filtering performance in cases of signals where high peaks are visible. Most of the high peaks are detected and re-estimated by the filter. The distortion in the frequency domain is significantly low (see the metrics in Figure 4). A large number of internal anomalies are also detected and then filtered out during the cleaning step.

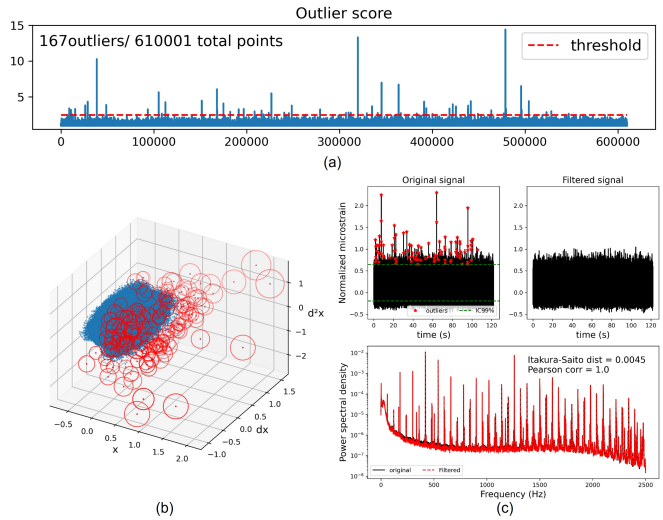

Figure. 4. Example 1 – Filtering Results: (a) Outlier scores calculated by LOF; (b) 3D-space (x, dx, d^2x) with outliers circled in red; (c) Comparison of signals and their PSDs, before and after filtering

Figures 5 and 6 present two cases of signals containing peaks that are potentially abnormal but hidden by other peaks. With this example, we aim to show that the proposed filtering does not filter out all peaks with the same amplitudes in the time domain. Instead, it effectively detects peaks that exhibit unusual behaviour in value changes.

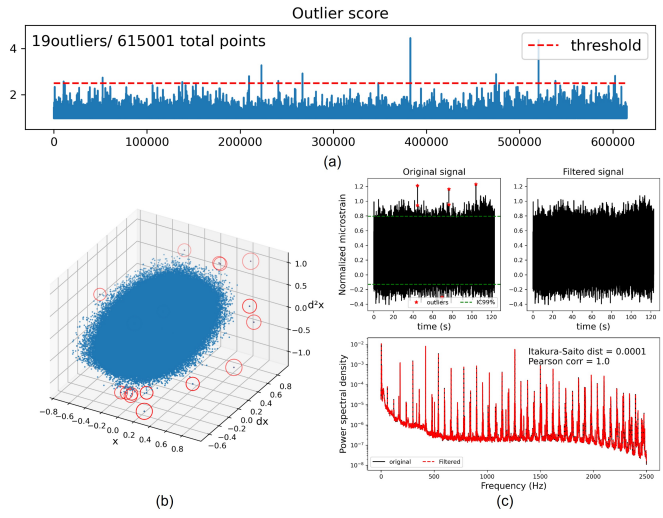

Figure. 5. Example 2 – Filtering Results: (a) Outlier scores calculated by LOF; (b) 3D-space (x, dx, d^2x) with outliers circled in red; (c) Comparison of signals and their PSDs, before and after filtering

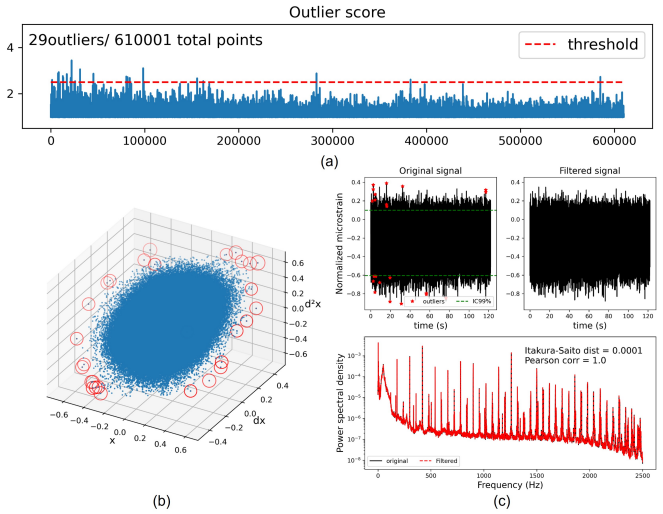

Figure. 6. Example 3 – Filtering Results: (a) Outlier scores calculated by LOF; (b) 3D-space (x, dx, d^2x) with outliers circled in red; (c) Comparison of signals and their PSDs, before and after filtering

V. CONCLUSION

The proposed filtering method is non-parametric and adaptive, as it does not require an assumption about the distribution of the targeted signal. The case study shows that this filtering technique works well for real measurements. The verification carried out using metrics in the frequency domain ensures that the essential physical phenomena in the hydroelectric turbine blade strain signal are preserved after filtering. Moreover, the number of points removed is low compared to the total length of the signal, which reduces the risk of significant distortion after filtering.

REFERENCES

[1] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," ACM Comput. Surv., vol. 54, no. 3, p. 56:1-56:33, Apr. 2021, doi: 10.1145/3444690.

[2] Q. H. Pham, J. Antoni, A. Tahan, M. Gagnon, and C. Monette, "Simulation of non-Gaussian stochastic processes with prescribed rainflow cycle count using short-time Fourier transform," Probabilistic Engineering Mechanics, vol. 68, p. 103220, Apr. 2022, doi: 10.1016/j.probengmech.2022.103220.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, New York, NY, USA, May 2000, pp. 93–104. doi: 10.1145/342009.335388.

[4] A. Presas, Y. Luo, Z. Wang, and B. Guo, "Fatigue life estimation of Francis turbines based on experimental strain measurements: Review of the actual data and future trends," Renewable and Sustainable Energy Reviews, vol. 102, pp. 96–110, Mar. 2019, doi: 10.1016/j.rser.2018.12.001.

[5] D. G. Goring and V. I. Nikora, "Despiking Acoustic Doppler Velocimeter Data," Journal of Hydraulic Engineering, vol. 128, no. 1, pp. 117–126, Jan. 2002, doi: 10.1061/(ASCE)0733-9429(2002)128:1(117).