Handling Irregularly Sampled IoT Time Series to Inform Infrastructure Asset Management

Manuel Herrera * **Manu Sasidharan** * **Jorge Merino** *
Ajith K. Parlikad *

* *Institute for Manufacturing, Dept. of Engineering, University of Cambridge, CB3 0FS, United Kingdom*
(e-mail: amh226,mp979,jm2210,aknp2@cam.ac.uk)

Abstract: Infrastructure systems in today's increasingly interconnected world employ the capabilities of the Internet of Things (IoT) technologies for their monitoring, operational control, and asset management. IoT devices can be defined as sensors (of different types) collecting, processing, and sharing time series of data. The analysis of such data often face challenges as a consequence of the high frequency of data collection and the increasing number of sensors placed on infrastructure. Power related issues, timestamp misalignment, and heterogeneous sampling designs are among the most common issues that the IoT data collection may suffer alongside the inherent complexities of large scale databases. This paper provides an overview of time series mining techniques adapted to tackle such issues in IoT data. The aim is to have a pattern recognition tool-set for developing anomaly detection algorithms. Particularly, the paper investigates how to efficiently handle large-scale time series coming from multiple sensors in a stream and following an unevenly spaced - irregular - sampling. The analysis is demonstrated through a case study of time series data mining of sensors installed for supporting the predictive maintenance of quay-cranes at the Port of Felixstowe, the largest container port in Britain.

Keywords: Time series mining, Anomaly detection, Condition monitoring, IoT, Smart ports

1. INTRODUCTION

Failure prediction and preventive action are increasingly becoming a crucial part of many infrastructure asset management strategies. The ongoing digitalisation of the critical infrastructure systems such as transport (Sasidharan et al., 2021; Jing et al., 2021), power (Pawar and Deosarkar, 2017; Cao et al., 2012) and water (Badawi, 2019) demonstrate the enormous capabilities to collect vast amount of data and information from various sources to inform effective operation and maintenance strategies. This is enabled by the use of internet of things (IoT) sensors that serve as data collectors as well as communication channels between sensors, and with the physical infrastructure that they monitor and control (Fathy et al., 2018b; Herrera et al., 2020). When integrated with robotics and smart devices, they may serve as actuators over the physical infrastructure. IoT can be used as a vehicle for optimal operation in the working context of cyber-physical systems management. The implementation of IoT in critical infrastructures is well known in practice. Their applications are disparate ranging from anomaly detection in water supply (González-Vidal et al., 2019a) to railway infrastructure maintenance (Jo et al., 2017), passing by healthcare system management (Uslu et al., 2020).

Sensors are installed on or around a component to gather a sequence of values of a variable over time, i.e. time series. From a data analysis perspective, IoT time series are large scale databases and a Big Data approach is often needed for data management and analysis. Promising alternatives focus on data preprocessing that is aimed to reduce the data size (Fathy et al., 2018a). Streaming analytics, ideally conducted in near real-time, is another challenge associated with IoT time-series data. For instance, in the literature, it is possible to find the work of Fathy et al. (2018c) who developed a near-real-time change detection algorithm for streaming IoT data. Despite the extensive research done in addressing the IoT issues, some challenges are still overlooked in the literature. This is the case of dealing with unevenly spaced or irregular time series data due to intermittent data transfer windows; the consequence of a lack of timestamp alignment in sensors, power cuts, or issues with the IoT connectivity.

This paper provides an overview of the most common ways to work with IoT time-series data, particularly on the methods for irregularly sampled time series. Additionally, the paper proposes an adaptation of time series data mining procedures that will address not only the sampling problem but also provides an efficient solution to work with large-scale database dimension and other needs associated with streaming analytics of IoT data. The paper shows the performance of IoT time series methods compared to the time series mining adaptation and discusses their advantages and disadvantages. To this end, the results of a case study on the IoT sensors installed to monitor the mechanical properties of quay cranes (QCs) at the Port of Felixstowe (PoF) is presented.

2. IOT TIME SERIES DATA ANALYSIS

The information coming from IoT devices are often large scale databases of temporal data. There are multiple challenges associated with such data analysis (De Francisci Morales et al., 2016). In addition to scale issues, a well-investigated problem is about data coming at multiple interconnected IoT devices simultaneously. One way to deal with it is by approaching the data analysis under a network traffic framework (Lopez-Martin et al., 2019). Another solution is a database transformation coupled with a feature selection process to enable working with multiple time series (González-Vidal et al., 2019b). Other challenges related to IoT data analysis are those coming from the need for near-real-time models. Particularly, it is about the model adaptation to what the IoT technology informs infrastructure operation and management (Brentan et al., 2017). In addition to all the challenges mentioned earlier, it highlights the problem of dealing with irregularly sampled time series due to issues ranging from timestamps misalignment to power cuts; and providing unevenly spaced data and large sequences of missing information. This problem attracts a significant part of current research in IoT time series and the following subsections will discuss it.

2.1 Overview of irregularly sampled time-series analysis

In the literature, the analysis of irregularly sampled time series is twofold. The first solution is re-sampling, or interpolation, of the time series for its consequent transformation into equally spaced data (and then applying existing time series methods). This solution may come with a bias; in some cases because of the underestimating seasonal components and, in others, to working with estimated data. An alternative solution is to work with transformations coming from the spectral analysis of the time series. The following bullet points provide a brief description of each method:

- Transformation to equally spaced time series:
  - Re-sampling: This strategy generally involves a down-sampling method for mapping the database to a suitable frequency in which the time series is back to be sampled at regular intervals thanks to a process of data aggregation. Other common strategies are based on up-sampling, increasing the frequency of the data collection by a process of data augmentation. A combination of both, down- and up-sampling strategies, can also be designed.
  - Interpolation: This strategy addresses the irregularly sampled time series by creating timestamps at regular intervals and filling the gaps made by missing data that come with new timestamps. Standard and time-series specific methods of interpolation can be used in this process.
- Least-squares spectral analysis (LSSA): The fundamentals of this procedure revolve around the Fourier transform. However, LSSA can handle several limitations inherent in the classical Fourier transform. For instance, working with unequally spaced values, handling missing data, and a non-constant average. There are two main developments from LSSA:
  - Lomb-Scargle periodogram (LSP): This method enables a fast computation of a Fourier-like spectrum from irregular time series data (VanderPlas, 2018).
  - Least-squares wavelet analysis (LSWA): This is an extension of the LSSA to analyse non-stationary and irregularly sampled time series. The basis of the method lies in the spectrogram analysis of the time-series frequency domain (Ghaderpour and Pagiatakis, 2017). This is done by a time series segmentation and estimation of spectral peaks of sine and cosine functions at each segment. The analysis can be enhanced by using the co-variance matrix associated with the time series.

Adding to the brief summary, it is worth mentioning the possibility of working with a feature analysis (state-space approach for time series) rather than from a regular time series perspective (Bahadori and Liu, 2012). This framework is out of the scope of the current paper, but it is of our highest interest and will be considered in our future investigations.

2.2 Signal analysis ensemble for IoT time series

Model combination and/or ensemble often provides better performance for general data mining. In this case, irregularly sampled time series can be analysed also by a combination of multiple methodologies involving those exposed above and others from the time-series mining literature.

The main process involves working with the algorithm of fast Fourier transform (FFT) that is used to transform the time series, or signal collected by the sensors, from a time domain to a frequency domain. The advantage of using FFT in this context lies in its natural reduction of the signal dimension as well as a way of removing part of the signal noise, by considering multiple signal resolution levels. FFT also facilitates a process of time-series embedding that will enable a further pattern extraction and, consequently, an anomaly detection process.

A basic FFT of the signal $x(t)$ follows the expression of Equation (1),

$$\mathcal{F}\{x(n)\} = X(\xi) = \sum_{n=-\infty}^{\infty} x(t)e^{-j2\pi\xi_n x}, \qquad (1)$$

where $\mathcal{F}$ represents the Fourier transform of the signal evolving over time, $x(t)$. The frequency of the signal is represented by ξ_n, where $\xi_n = n/T$ and T is such that the interval $[-T/2, T/2]$ is an interval in which the signal is not zero. FFT works well for modelling signals coming from vibration and acceleration data that is often monitored for mechanical components. This is because the imaginary unit, $j = \sqrt{-1}$, is embedded in Equation (1) by definition; adding the spatial-angle information to the magnitude modelling of the signal. However, working with signals related to vibration and acceleration.

The FFT also has an inverse operation enabling signal reconstruction after transformation. The FFT inverse is expressed in Equation (2),

$$x(n) = \sum_{n=-\infty}^{\infty} X(\xi) e^{j2\pi\xi_n x} \Delta\xi, \qquad (2)$$

where $\Delta\xi = \frac{n+1}{T} - \frac{n}{T} = \frac{1}{T}$.

There is a direct extension into a multidimensional FFT for the case of having a n-tuple signal, defined by $\mathbf{x} = (x_1, \ldots, x_n)$, evolving over time. The dimension of the frequencies is also extended, considering the vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$.

To the above mentioned useful properties of FFT, others such as time shifting and scaling or reversal can also be considered. As a result, the FFT is the option chosen herein for compressing the signal information and further analysis. Given the irregular sampling of the time series to analyse, the proposed methodology involves a multi-step procedure that is summarised in Figure 1 and detailed in the following phases:

(1) Time series segmentation process. This phase splits the signal, $\{t_1, \ldots, t_T\}$, into multiple time segments grouped by the sampling frequency in the data collection. Each group is, then, made of $k_1, \ldots, k_n$ elements. Timestamps at the starting and end of each time segment are added to the data group information for indexing purposes. So, for each j-th subsequence, the data is $\{x_j \ldots x_{kj}; t_j, t_{j+kj}\}$.
(2) FFT at each sub-sequence.
(3) Feature extraction in the spectral domain. This enables to follow up with an embedding procedure of such features representing the time sub-sequences.
(4) Clustering process for the sub-sequence's new embedding information.

A variation of FFT known as power spectral density function (PSD) computes the intensity of the variations in a signal as a function of the frequency, which is useful for vibration and acceleration analysis.

3. CASE STUDY: SMART PORTS

Ports are a national critical infrastructure, key for international trading, supply chain and the overall sustenance of a country's economy. The digitalisation of industrial processes brought about by Industry 4.0 has seen the port operation and management being aided by a network of smart sensors, actuators, communication devices, data centres, and decision support systems (Yang et al., 2018). The port's efficiency is often influenced by the availability and condition of the QCs that moves containers between the shore and shipping vessels. QCs experience extensive stresses and cyclic loading during their prolonged operations, making them prone to disruptions due to component deterioration. Their downtime not only paralyses the port operations but also adversely impacts the global supply chain (Kizilay and Eliiyi, 2021). To this end, it is essential to provide accurate and early fault detection and diagnosis to guide predictive maintenance of these critical assets.

A trail is ongoing at the PoF, Britain's largest container port, to monitor the critical components of the QCs using IoT sensors communicating via the port's private 5G; and to employ artificial intelligence to identify pre-incident trigger conditions. Figure 2 shows a schematic of the

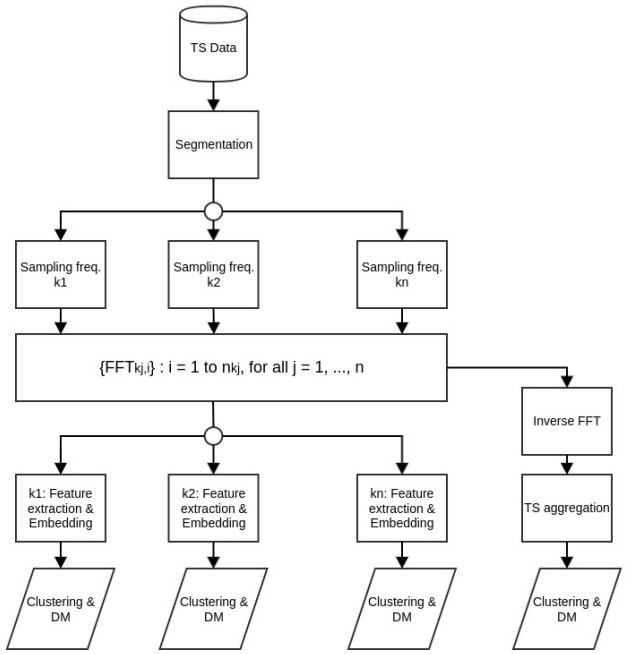


Fig. 1. Data mining process for irregularly sampled IoT time series data

physical IoT sensing set up and the data pipeline. An OASIS standard messaging protocol, MQTT, is employed to publish the data to the cloud. The IoT time series methods are applied to the data collected by the micro electro-mechanical system (MEMS) sensors that monitor the QC's hoist motor engine and gears.

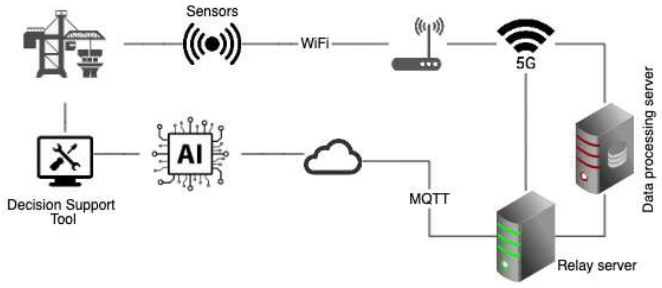


Fig. 2. Sensor setup and data pipeline at the PoF

Figure 3 shows the data collected from a sensor recording the vibration of a QC hoist engine. The sensor captures the readings every hour in batches of four minutes at 2 milliseconds frequency.

Figures 4a and 4b show the amplitude vs. frequency periodogram of the signals above, corresponding to acceleration in the edges X and Y. The edge Z has not been considered in this analysis since it shows a low variability over the records collected within this study. However, the addition of a dimension to the current analysis would be straightforward if it is required.

The extraction of statistical features over the frequency domain is based on selecting the top k peaks of amplitude and at which frequency it happened. Other descriptive statistics (quartiles, max, min, mean, standard deviation, and root mean square) are also taken into account. The

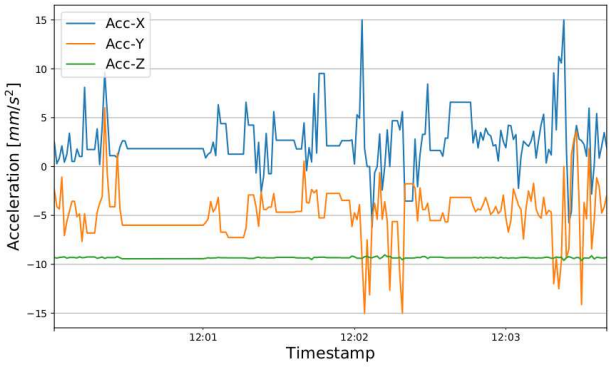
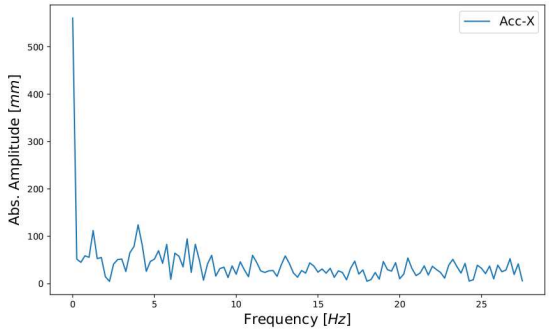
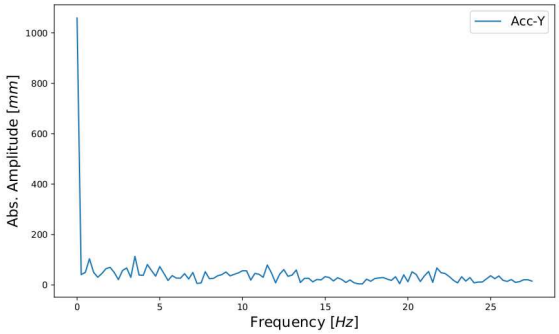


Fig. 3. Zoom into 4 minutes interval of vibration sensor



(a) Amplitude vs frequency periodogram for acceleration at edge X



(b) Amplitude vs frequency periodogram for acceleration at edge Y

Fig. 4. Amplitude vs frequency periodograms for the acceleration at the edges X and Y. Case of 4 min. sample at single vibration sensor.

action is repeated for every batch of four minutes of data observed during a week of the crane's functioning. Since the batches of data come every hour, it makes a total of $7 \times 24 = 168$ batches, each composed of 120,000 sensor readings (1 reading every 2 milliseconds for 4 minutes in each batch). We have, then, a new database of spectral features from which we can run a clustering algorithm to detect the existence of a cluster of a significantly lower number of signals than the others. In this case, we have an array of k peaks and k frequencies taken 168 times over the week.

One of the advantages of the procedure presented within this paper is the possibility to automatically classify new observations of batches of 4 minutes data into a closer cluster. This strategy will aid the development of anomaly detection processes in near real-time. Any clustering algorithm can be used here. A common option is the algorithm of k-means, given its good performance and high efficiency. Algorithms such as hierarchical agglomerative clustering can be used for visualisation purposes. Everitt et al. (2011) provides a comprehensive guide to clustering methods and analysis.

The clustering results on the spectra of the acceleration statistical features both for the edges X and Y can be summarised by the main characteristics of the amplitude of highest frequencies (see Table 1).

Table 1. Main features of the cluster of largest size

	Size (total – pct.)	Freq. (Hz)	Amp. (mm)
Edge X	100 – 60 %	2, 4	100, 120
Edge Y	120 – 71 %	2, 4	60, 80

Table 1 columns comprise the number of batches corresponding to the cluster of largest size (the most common signal configuration); together with a column with the top frequencies as well as a third column showing their amplitude.

The clustering configuration will help to detect groups representing potential anomalies based on their statistical features. Such groups are normally small in size and with a long distance to their closer neighbour. This can become into a near real-time process through an out-of-sample classification method by metering the distance of each new batch of observations taken by a sensor to the current set of cluster centroids (Rousseeuw and Hubert, 2018).

Considering the evolution of the time-series spectra, a parallel analysis has also been computed. This is by using the sequence on time of such spectra for further pattern extraction. This enables the analysis to consider the overall evolution of the sensor readings as well as to deal with multi scale data-analysis. Particularly, the spectra time-series has been successfully analysed by using the matrix profile method for computing distances based on an intensive computation of sub-sequences distances (Zhu et al., 2020; Herrera et al., 2021).

4. CONCLUSION

Infrastructure condition monitoring contributes to improving the asset management processes through timely information for near real-time decision-making. In this article, we preprocessed and visually explored FFT processed time series data from a network of IoT sensors that are monitoring a QC's hoist motor engine and gears. The spectral amplitudes are averaged by date and frequency, and time alignment of the data is performed. IoT time-series data presents certain characteristics that need to be particularly addressed instead of proceeding with a general time series analysis. Time series evolution is explored using a concatenation of time-series spectra that takes advantage of the sequential repetition of the on and off sampling phases in the case study presented herein. Over such a spectral concatenation we run a well-known algorithm mining time-series distance, such as the one used to compute matrix profile. These analyses can be considered done at a different time scale than those just comparing the spectral features of the data batches.

Additional experiments are required for the further improvement of the performance and accuracy with which

the process detect anomalies and to develop a classification of newly observed data leading to a near real-time anomaly detection process. The main limitation of the proposed approach is that it considers certain regularity in the time series data for the analysis of the evolution at multiple scales. Another limitation of this approach emerges from the need for maintenance records for labelling the data sets and the need for large amounts of quality data with information such as component breakdowns and failure. These kinds of data may not be easily available due to disparities in records. Future work will focus on exploring the possibilities of deep learning autoencoders to reconstruct the irregularly sampled time series into another equivalent, but taken at regular intervals.

ACKNOWLEDGEMENTS

This work was supported by the UK's Department for Digital, Culture, Media and Sport through the 5G Testbeds and Trials (5G Port of Felixstowe project). The authors acknowledge the support from Hutchison Ports, Three UK and Blue Mesh Solutions.

REFERENCES

Badawi, W.A. (2019). Underground pipeline water leakage monitoring based on IoT. *International Journal of MC Square Scientific Research*, 11(3), 01–08.

Bahadori, M.T. and Liu, Y. (2012). Granger causality analysis in irregular time series. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 660–671. SIAM.

Brentan, B.M., Luvizotto Jr, E., Herrera, M., Izquierdo, J., and Pérez-García, R. (2017). Hybrid regression model for near real-time urban water demand forecasting. *Journal of Computational and Applied Mathematics*, 309, 532–541.

Cao, Y., He, J., Huang, X., and Zhang, Z. (2012). Application of the internet of things technology in power transmission equipments condition monitoring. *Journal of Electric Power, Science, and Technology*, 27(3), 16–27.

De Francisci Morales, G., Bifet, A., Khan, L., Gama, J., and Fan, W. (2016). IoT big data stream mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2119–2120.

Everitt, B.S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster analysis 5th ed.* John Wiley.

Fathy, Y., Barnaghi, P., and Tafazolli, R. (2018a). An adaptive method for data reduction in the internet of things. In *2018 IEEE 4th World Forum on Internet of things (WF-IoT)*, 729–735. IEEE.

Fathy, Y., Barnaghi, P., and Tafazolli, R. (2018b). Large-scale indexing, discovery, and ranking for the internet of things (iot). *ACM Computing Surveys (CSUR)*, 51(2), 1–53.

Fathy, Y., Barnaghi, P., and Tafazolli, R. (2018c). An online adaptive algorithm for change detection in streaming sensory data. *IEEE Systems Journal*, 13(3), 2688–2699.

Ghaderpour, E. and Pagiatakis, S.D. (2017). Least-squares wavelet analysis of unequally spaced and non-stationary time series and its applications. *Mathematical Geosciences*, 49(7), 819–844.

González-Vidal, A., Cuenca-Jara, J., and Skarmeta, A.F. (2019a). IoT for water management: Towards intelligent anomaly detection. In *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, 858–863. IEEE.

González-Vidal, A., Jiménez, F., and Skarmeta, A.F. (2019b). A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. *Energy and Buildings*, 196, 71–82.

Herrera, M., Pérez-Hernández, M., Kumar Parlikad, A., and Izquierdo, J. (2020). Multi-agent systems and complex networks: Review and applications in systems engineering. *Processes*, 8(3), 312.

Herrera, M., Proselkov, Y., Pérez-Hernández, M., and Parlikad, A.K. (2021). Mining graph-fourier transform time series for anomaly detection of internet traffic at core and metro networks. *IEEE Access*, 9, 8997–9011.

Jing, G., Siahkouhi, M., Qian, K., and Wang, S. (2021). Development of a field condition monitoring system in high speed railway turnout. *Measurement*, 169, 108358.

Jo, O., Kim, Y.K., and Kim, J. (2017). Internet of things for smart railway: feasibility and applications. *IEEE Internet of Things Journal*, 5(2), 482–490.

Kizilay, D. and Eliiyi, D.T. (2021). A comprehensive review of quay crane scheduling, yard operations and integrations thereof in container terminals. *Flexible Services & Manufacturing Journal*, 33(1).

Lopez-Martin, M., Carro, B., and Sanchez-Esguevillas, A. (2019). Neural network architecture based on gradient boosting for IoT traffic prediction. *Future Generation Computer Systems*, 100, 656–673.

Pawar, R.R. and Deosarkar, S. (2017). Health condition monitoring system for distribution transformer using Internet of Things (IoT). In *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, 117–122. IEEE.

Rousseeuw, P.J. and Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2), e1236.

Sasidharan, M., Burrow, M.P.N., Ghataora, G.S., and Marathu, R. (2021). A risk-informed decision support tool for the strategic asset management of railway track infrastructure. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 09544097211038373.

Uslu, B.Ç., Okay, E., and Dursun, E. (2020). Analysis of factors affecting IoT-based smart hospital design. *Journal of Cloud Computing*, 9(1), 1–23.

VanderPlas, J.T. (2018). Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplement Series*, 236(1), 16.

Yang, Y., Zhong, M., Yao, H., Yu, F., Fu, X., and Postolache, O. (2018). Internet of things for smart ports: Technologies and challenges. *IEEE Instrumentation & Measurement Magazine*, 21(1), 34–43.

Zhu, Y., Gharghabi, S., Silva, D.F., Dau, H.A., Yeh, C.C.M., Senobari, N.S., Almaslukh, A., Kamgar, K., Zimmerman, Z., Funning, G., et al. (2020). The swiss army knife of time series data mining: ten useful things you can do with the matrix profile and ten lines of code. *Data Mining and Knowledge Discovery*, 34(4), 949–979.