Computer Networks 218 (2022) 109384

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Time-series clustering for sensor fault detection in large-scale Cyber–Physical Systems

Ahmed A. Alwan *, Allan J. Brimicombe, Mihaela Anca Ciupala *, Seyed Ali Ghorashi, Andres Baravalle, Paolo Falcarin

School of Architecture Computing and Engineering, University of East London, Docklands Campus, University Way 4, E16 2RD, London, United Kingdom

ARTICLE INFO

Keywords:
Cyber–physical systems (CPSs)
Wireless sensor networks(WSNs)
Time-series clustering
Dynamic time warping
K-shape
Characteristics based time-series clustering

ABSTRACT

Large-scale Cyber–Physical Systems (CPSs) are information systems that involve a vast network of sensor nodes and other devices that stream observations in real-time and typically are deployed in uncontrolled, broad geographical terrains. Sensor node failures are inevitable and unpredictable events in large-scale CPSs, which compromise the integrity of the sensors measurements and potentially reduce the quality of CPSs services and raise serious concerns related to CPSs safety, reliability, performance, and security. While many studies were conducted to tackle the challenge of sensor nodes failure detection using domain-specific solutions, this paper proposes a novel sensor nodes failure detection approach and empirically evaluates its validity using a real-world case study. This paper investigates time-series clustering techniques as a feasible solution to identify sensor nodes malfunctions by detecting long-segmental outliers in their observations' time series. Three different time-series clustering techniques have been investigated using real-world observations collected from two various sensor node networks, one of which consists of 275 temperature sensors distributed around London. This study demonstrates that time-series clustering effectively detects sensor node's continuous (halting/repeating) and incipient faults. It also showed that the feature-based time series clustering technique is a more efficient long-segmental outliers detection mechanism compared to shape-based time-series clustering techniques such as DTW and K-Shape, mainly when applied to shorter time-series windows.

1. Introduction

Cyber–Physical Systems (CPSs) can be seen as networks of physical components such as sensors and often actuators effectively incorporated using a computational and communication core [1,2]. Sensors collect physical environment measurements and transmit them as raw data to a computational unit. The computational unit generates feedback and sends it to the actuators that regulate the physical conditions based on the received data. This cycle ultimately achieves the self-awareness of the CPS via its ability to assess and correctly adjust its behaviour and performance in real-time [3]. Large-scale CPS applications, such as environmental monitoring systems, typically involve many low-cost sensor nodes deployed in broad geographical terrains, forming a large-scale Wireless Sensor Network (WSN) [4–6]. Ecological factors may compromise the accuracy of sensor observations, as extreme temperature [7] or humidity [4]. Failures in sensor nodes and sensor networks are inevitable events in large-scale CPS applications, which may negatively affect data quality [8], producing invalid information and potentially reducing the quality of their service [9,10]. In general, sensor nodes in WSNs have limited computing power, storage capacity and transmission radius [11,12]. Therefore, wireless sensor nodes cannot directly send observations to a remote data destination (sink). Instead, a hub device or another sensor node works as a bridge to transfer readings from other sensor nodes. Sensor nodes closer to the sink consume more power because they support other sensors to transmit their observations, and they are expected to have more power failures causing sensor node failure issues [13,14]. Thus, sensor nodes may determine the network lifetime based on their battery capacity and affect the system's quality of service [15,16].

The contribution of this research lies in the successful implementation of time-series clustering techniques as a sensor failure detection mechanisms based on detecting long-segmental outliers associated with sensors faults. It investigates the possibility of utilising time-series clustering as a sensor node failure detection mechanism, focusing on detecting long-segmental outliers associated with halting and incipient sensor failure patterns. Dynamic Time Warping (DTW), K-Shape and the Characteristics-based time-series clustering techniques are tested to

* Corresponding authors.
E-mail addresses: a.alwan@uel.ac.uk (A.A. Alwan), a.j.brimicombe@uel.ac.uk (A.J. Brimicombe), m.a.ciupala@uel.ac.uk (M.A. Ciupala), s.a.ghorashi@uel.ac.uk (S.A. Ghorashi), a.baravalle@uel.ac.uk (A. Baravalle), falcarin@uel.ac.uk (P. Falcarin).

https://doi.org/10.1016/j.comnet.2022.109384
Received 21 March 2022; Received in revised form 26 July 2022; Accepted 20 September 2022
Available online 4 October 2022
1389-1286/© 2022 Elsevier B.V. All rights reserved.

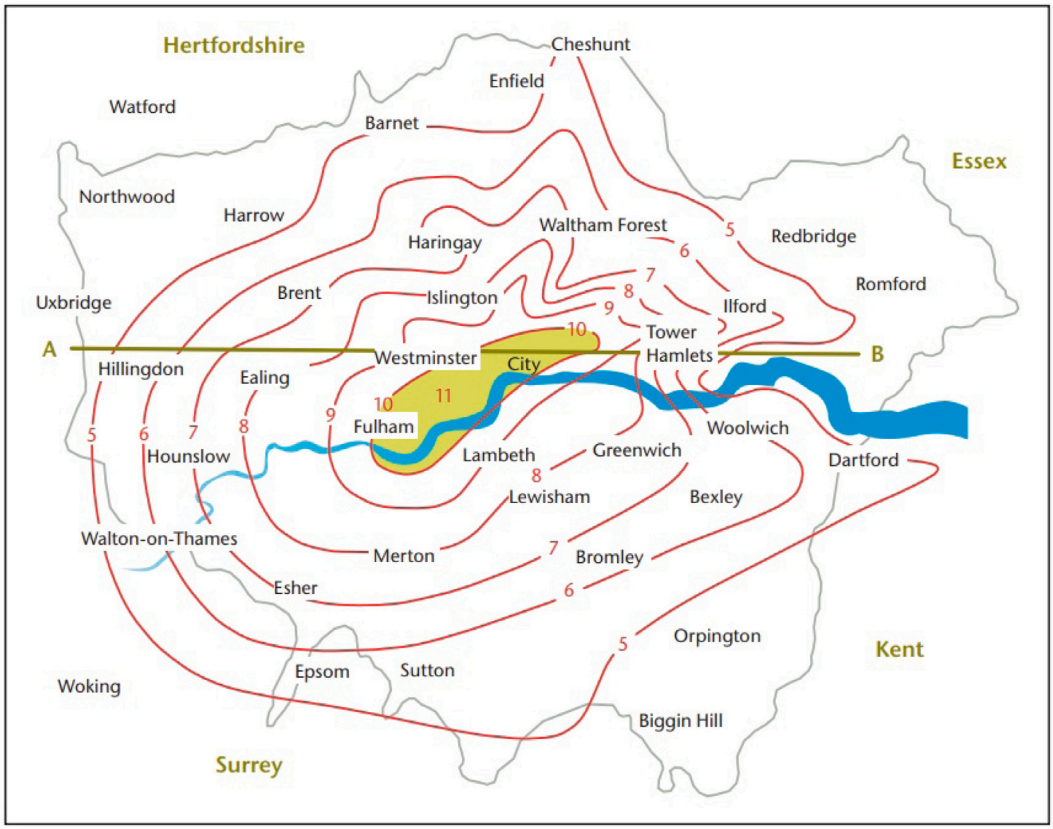

Fig. 1. The ambient temperature profile of London and the effect of the Urban Heat Islands [17].

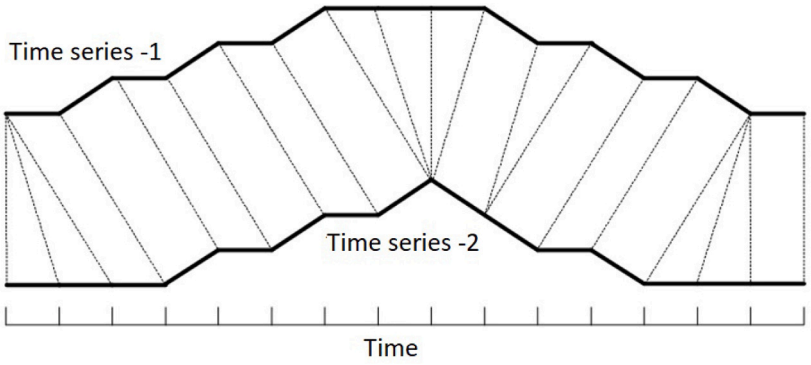

Fig. 2. An illustration of how DTW warps one time-series into another one [18].

prove the validity of the proposed approach. The accuracy of these methods in detecting the incipient sensor node failure pattern is empirically evaluated using time series collected from a local ambient temperature sensor node network deployed at the University of East London, UK. The accuracy of these methods in detecting the halting sensor node failure pattern is evaluated using time-series collected from a large-scale sensor node network comprising 275 ambient temperature sensors distributed around London.

This paper is organised as follows: Section 2 provides details of the related work, Section 3 is an introduction to time-series clustering. Section 4 describes the research data set, Section 5 presents the implementation results of the Dynamic Time Warping (DTW) and K-Shape time-series clustering techniques, and Section 6 provides the results of testing the Characteristics-based time-series clustering technique. Finally, Section 7 presents the concluding remarks.

2. Related work

Many sensor node failure detection methods were proposed in the literature, which can be mainly categorised as; technical-based and data mining-based sensor failure detection techniques.

Technical-based sensor node failure detection techniques are domain-specific solutions designed to detect sensor failures in particular applications. For example, Togneri et al. [14] utilised signal processing as sensor's failure detection mechanism for monitoring the hardware status of a large-scale weather sensor network. The Togneri et al. [14] sensor's failure detection approach cannot be adopted as a

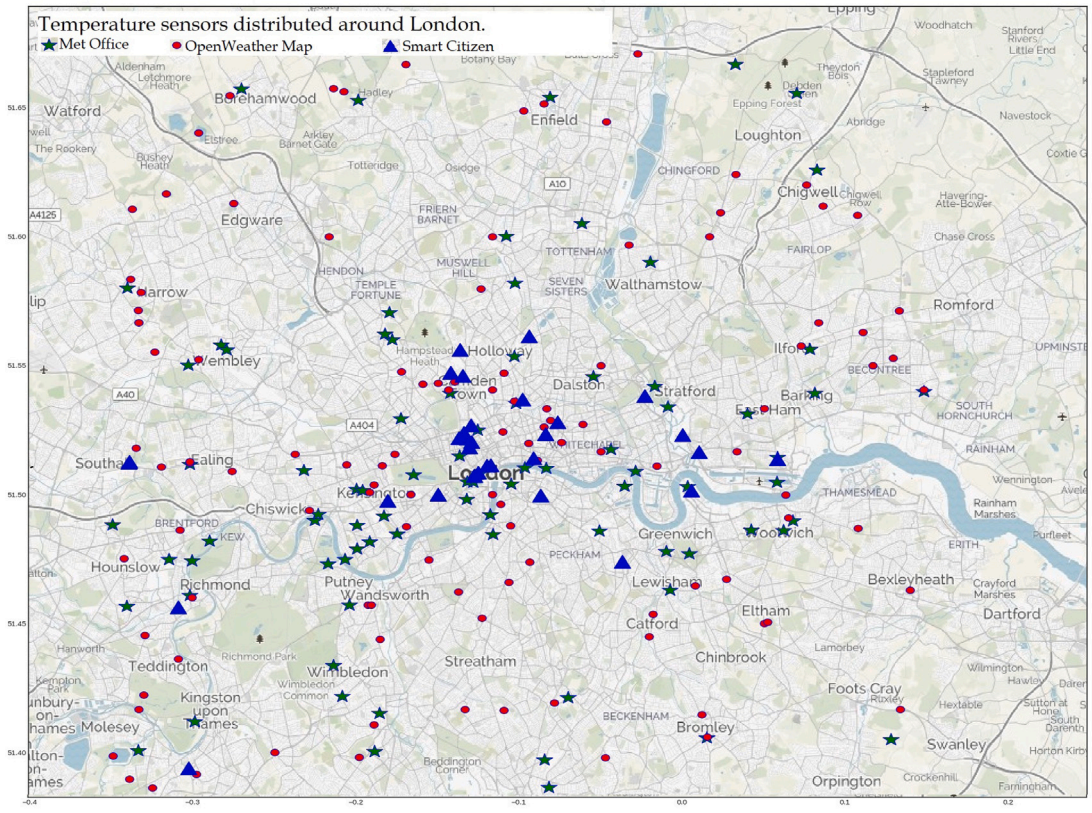

Fig. 3. The geographical distribution of the real-world sensor nodes adopted in this study.

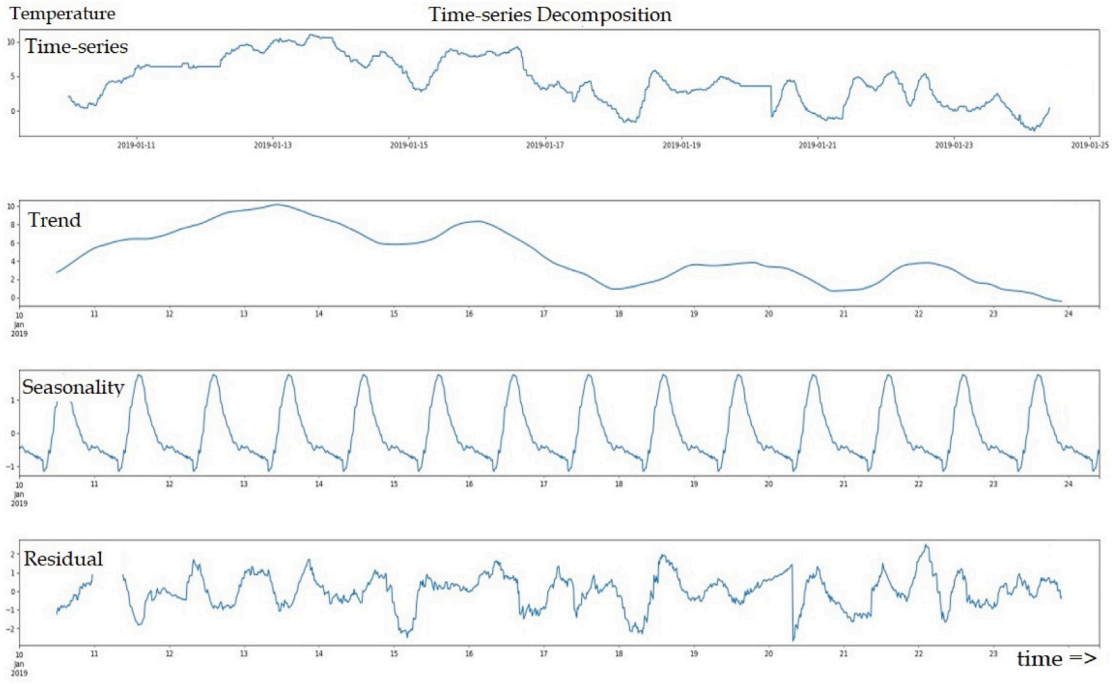

Fig. 4. Time-series decomposition of a single time-series collected from a real-world temperature sensor node of the large-scale sensor network distributed around London.

generic solution for detecting sensor failures. Thus, it does not present a systematic or generic approach for detecting sensor node failures in large-scale CPS applications. Furthermore, this approach requires direct access to the sensor management network to check their status, and such access may not be guaranteed in large-scale CPS applications.

Data mining-based sensor failure detection techniques utilise data analysis models to detect abnormal data patterns (outliers) in sensor observations associated with sensor node failures, mainly categorised into the anomaly and predictive analysis models [19,20]. Where an outlier is an extreme sensor node's measurement, it is "an observation

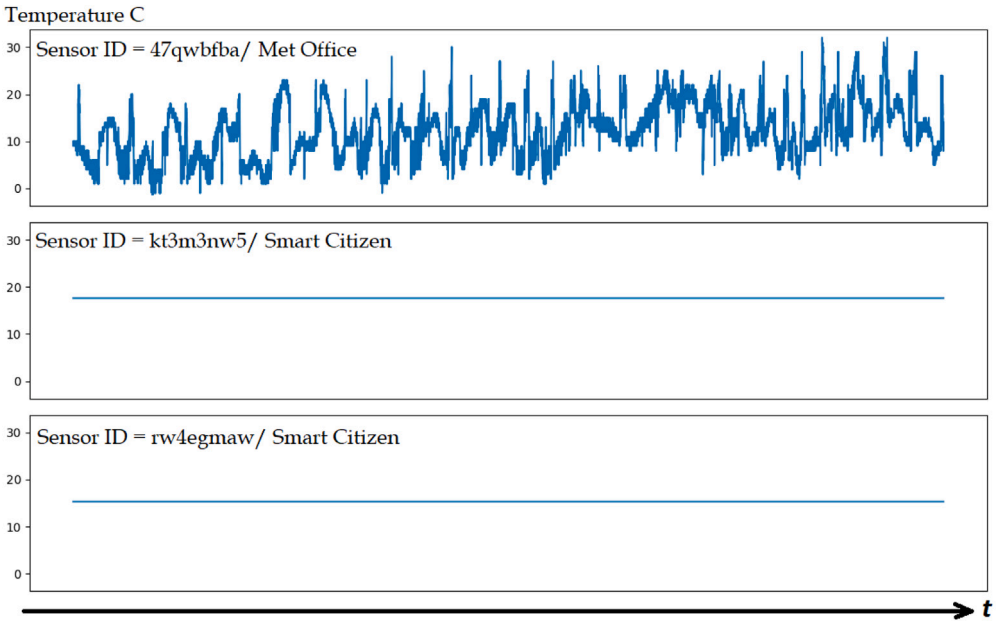

Fig. 5. An example of two temperature time-series with (halting) long-segmental outliers (Sensor ID= kt3m3nw5 and rw4egmaw/Smart Citizen) comparing with a normal time-series collected from a functional sensor node (Sensor ID=47qwbfba/Meteorological Office).

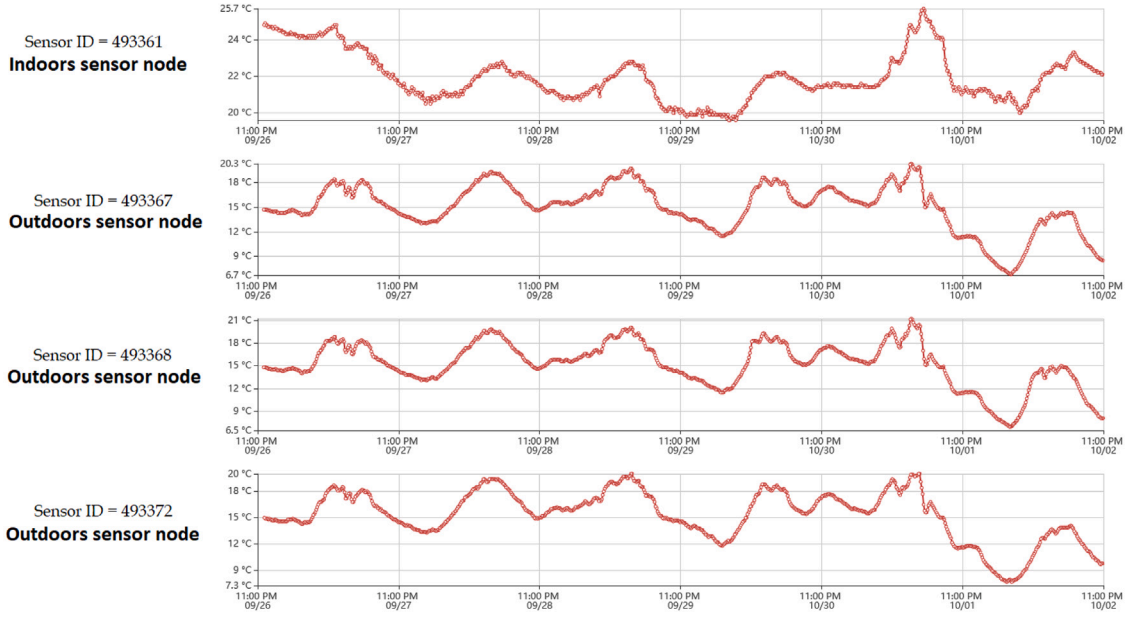

Fig. 6. Time-series of the local sensor node network do not show significant differences in the shape of their trend. However, they show differences in the value attribute, especially with the indoors sensor (Sensor ID = 493361) which streamed a time-series with a consistent offset of 10–15 C^{o} from other sensors.

which deviates so much from other observations as to arouse suspicions that a different mechanism generated it" [21].

1. **Anomaly analysis**, also known as outlier detection, identifies unusual data patterns that do not comply with well-established normal behaviour [22]. Suppose the absolute value of a sensor observation deviation is significantly diverted from observations of other neighbouring (spatially correlated) sensors at the same point in time. In that case, this observation is an outlier and potentially streamed from a faulty sensor node [19,23]. Anomaly analysis is a significant research field that is mainly investigated using statistical and machine-learning based outlier detection techniques. For example, Deep Neural Networks (DNN) [24], K-Nearest Neighbours algorithm (KNN) [24], K-means clustering algorithm are machine-learning-based outlier detection methods [25]. In contrast, standard deviation, correlation coefficient [26] and the density-based spatial clustering of applications with noise (DBSCAN) are statistical-based outlier detection methods [27–29]. Outlier detection techniques rely on the assumption that the value of sensor nodes' observations is correlated spatially, temporally, or both spatially and temporally. However, these assumptions are not necessarily always

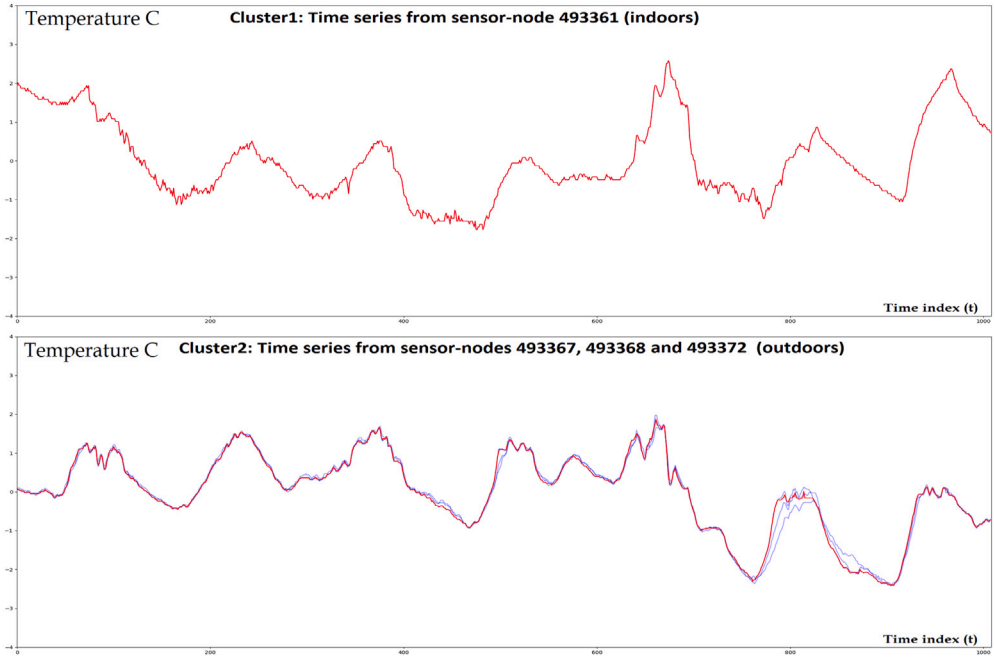

Fig. 7. DTW and K-Shape were able to successfully differentiate the incipient faults time-series of indoor sensor-493361 from the other (normal) outdoors time-series.

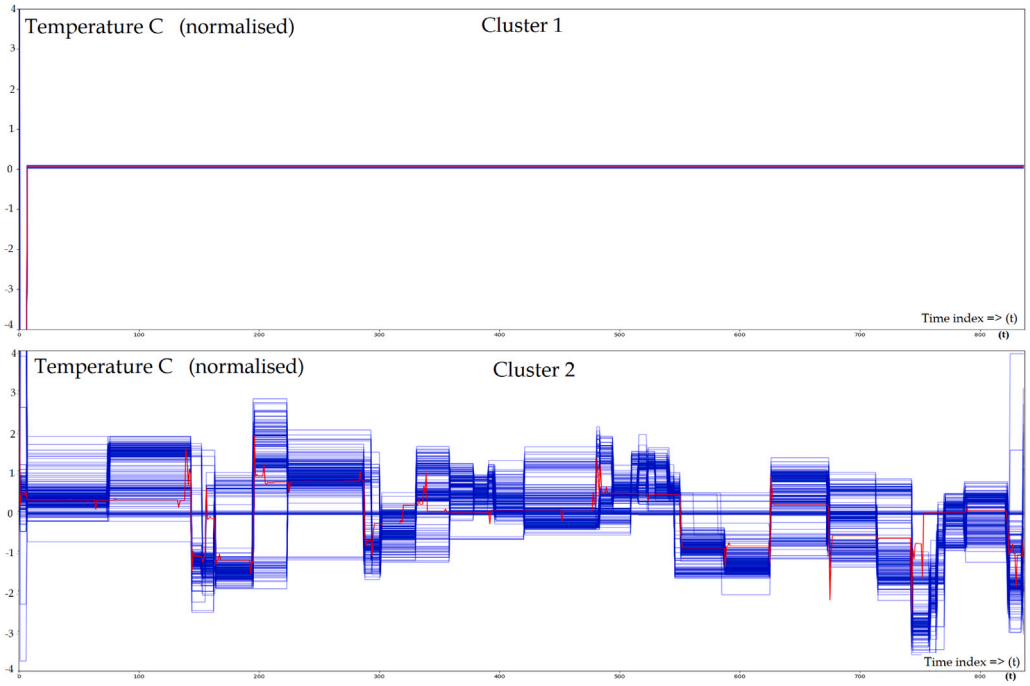

Fig. 8. DTW successfully separated time-series with the long-segmental outliers (Cluster 1) from other (typical) time-series (Cluster 2) when applied to seven-day window real-world dataset.

valid, especially in large-scale CPS applications where the correlations between sensor nodes may be violated by many external effects, such as the size of the deployment environment and the geographical distribution of sensor nodes [30]. For example, outlier detection cannot be applied directly to ambient temperature observations collected from the sensor nodes distributed around London because of a phenomenon known as the Urban Heat Islands (UHI), as shown in the heat profile map of London in Fig. 1 [17,31]. UHI causes up to 6 degrees C^0 of unexpected divergence among ambient temperature sensors observations, violating the spatial continuity constraints among these sensors and undermining the effectiveness of anomaly analysis methods in identifying unusual data patterns that do not comply with typical ambient temperature sensors observations in comparison with nearby ambient temperature sensors.

2. **Predictive analysis** is the process of mining current and historical data to identify patterns and forecast the future values of time series [32,33]. Predictive analysis can be conducted

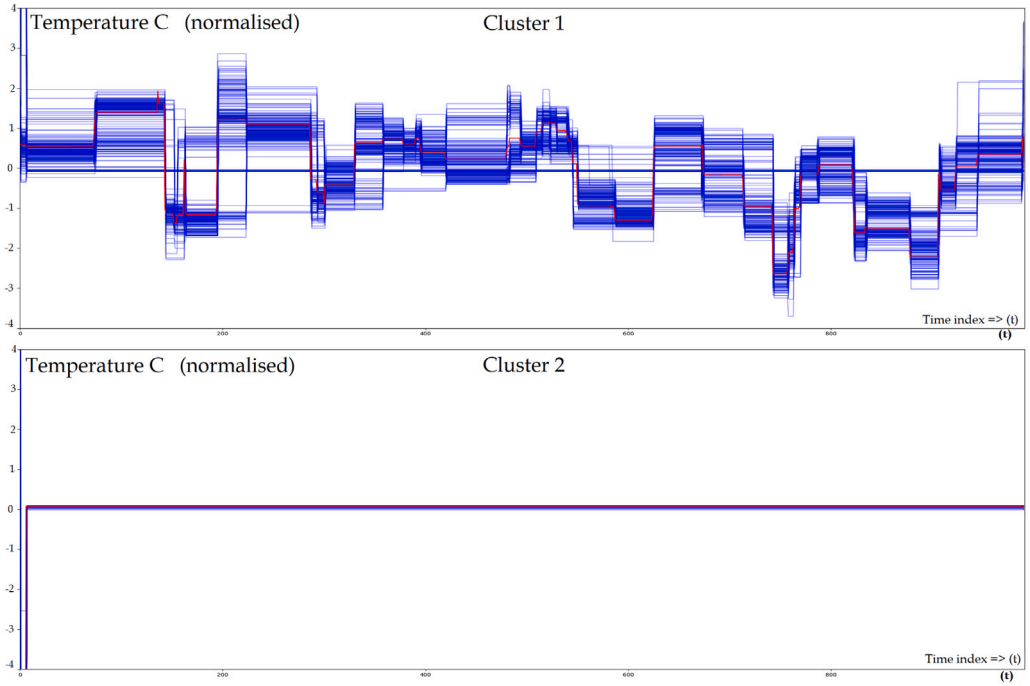

Fig. 9. K-Shape successfully separated time-series with the long-segmental outliers (Cluster 2) from the other (typical) time-series (Cluster 1) when applied to seven-day window real-world dataset.

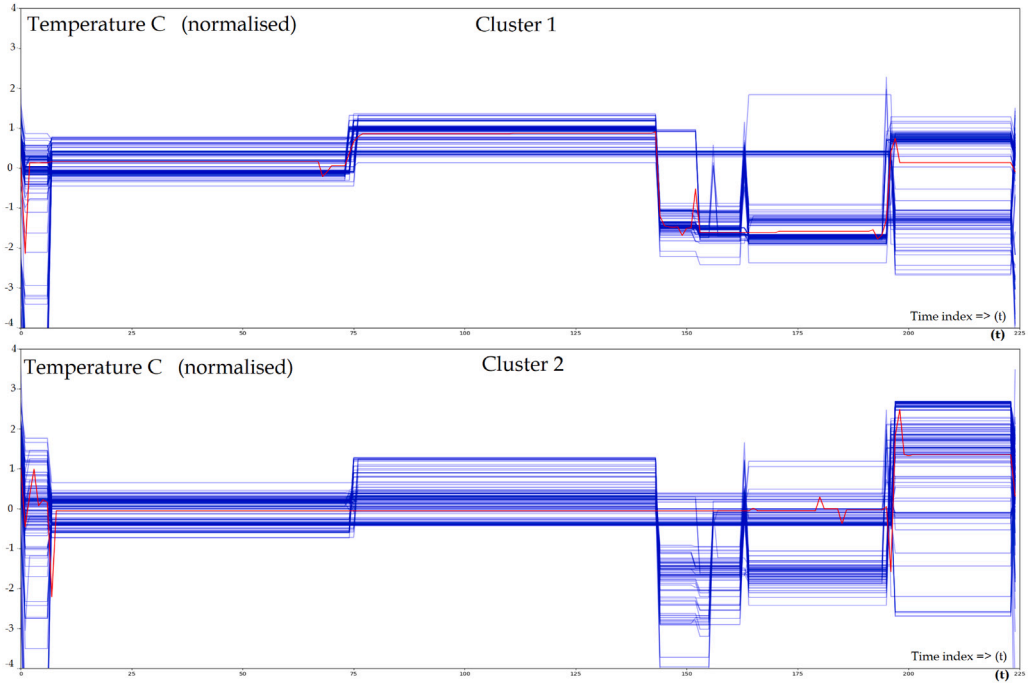

Fig. 10. DTW is not able to differentiate time-series with long-segmental outliers from other typical time-series when it was applied to a shorter two day time-window of real-world time-series.

using statistical or machine-learning-based techniques [34]. For example, a machine learning model based on the Random Forest Prediction (Random Forest Regression) technique is adopted by Farooqi et al. [35] for developing an anomaly detection mechanism for weather data. Another example is based on statistical predictive analysis, using the one step-forward approach of the Autoregressive Moving Average (ARMA) model, to tackle the inevitable challenge of sensors and sensor network failures in power terminals [9]. Some applications require a mixed-methods approach, where both machine-learning and statistical methods are used to tackle a particular data quality challenge. For example, Okafor et al. [4] investigated the use of artificial

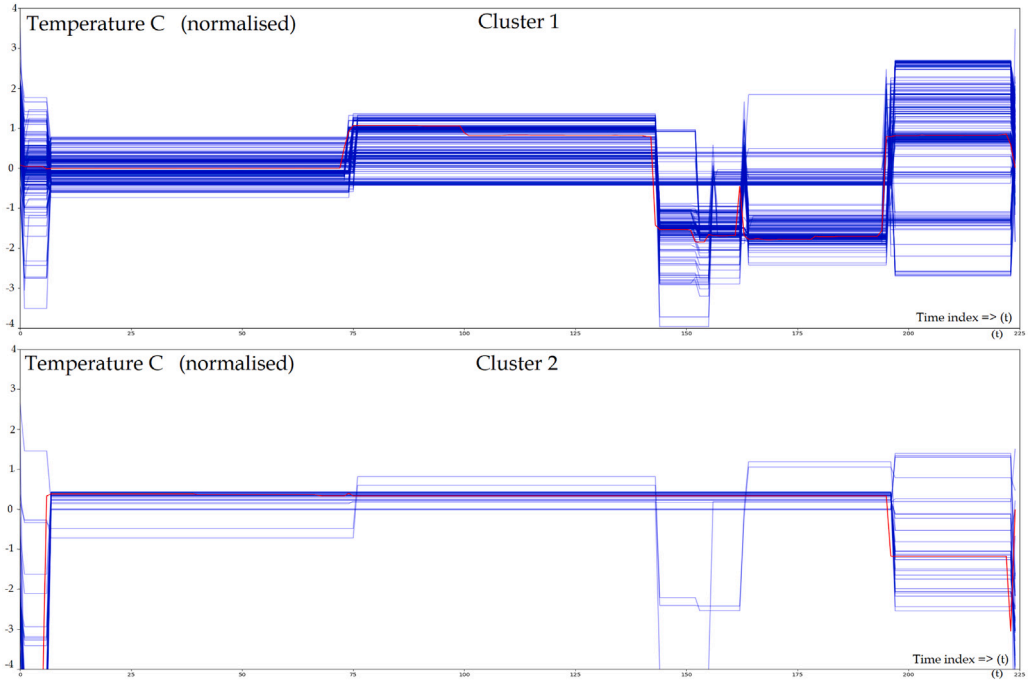

Fig. 11. K-Shape is less able to differentiate time-series with long-segmental outliers from other typical time-series when it was applied to a shorter, two day time-window, of real-world time-series.

neural networks and linear regression for calibrating low-cost environmental monitoring sensors to improve their service life by reducing the probability of their failure due to battery failure. Predictive analysis methods rely on models developed using historical data as a training data set. Therefore, predictive analysis is most suitable for detecting measurement errors that appear for a short time interval (short outliers). Measurement errors that occur for a relatively long time affect the ability of the predictive models to render accurate predictions. The pattern of time series with long outliers will be distorted to a certain extent, reflecting the wrong measurement as the standard pattern, leading to higher forecast errors and limiting the ability of predictive analysis modes to detect data accuracy issues correctly [36].

Outliers in sensor node networks are mainly categorised into short, simple and long-segmental outliers [37]. Long-segmental outliers, also known as shape outliers, are irregular observations that emerge for a relatively long time [38] and change the time-series pattern (set of observations) [39]. Long-segmental outliers occur in particular cases where a phenomenon has a long-term impact, such as forest fires or oil spills or due to sensor nodes failure [40]. Long-segmental outliers associated with sensor failures are categorised according to the behaviour of faulty sensor nodes [41] into:

- **Continuous halting faults:** long outliers that show no or minimal variation in the value attribute of their data stream for a relatively long interval of time.
- **Abrupt (emerging) and incipient faults:** a constant or linear increase offset to the measurement values that occur over a longer time interval than expected.

As long-segmental outliers occur for a relatively long time and change data patterns, they break the temporal correlation of observations after and before the anomaly and violate the possibility of using predictive outlier detection techniques to detect this type of anomalies [36].

This paper investigates time-series clustering as a novel solution that addresses the limitations of both anomaly and predictive analysis approaches in detecting long outliers associated with sensor node failures in the context of large-scale CPSs.

3. Time-series clustering

Time-series similarity measures define outliers in time-series windows by comparing them with other non-overlapping windows using a measurement metric, such as the Euclidean distance, which measures the distance between different time series [39,42,43]. Therefore, time-series similarity measures were utilised in time-series clustering methods to compare the pattern of an entire or a substantial time series window with another based on their long-term temporal correlation [42,44]. The purpose of time-series clustering is to identify faulty sensor nodes by comparing the shape or features of their time series with the time-series of other properly functioning sensor nodes. In this paper, Dynamic Time Wrapping (DTW) time-series clustering technique is tested as an anomaly detection mechanism. The DTW test has been extended to include K-Shape and Characteristic-Based Clustering techniques to find a higher performance clustering technique that can render accurate results while examining shorter time series.

3.1. Dynamic time warping

Dynamic Time Warping (DTW) is a time-series clustering technique that finds corresponding regions of similarity between time-series. DTW can stretch or shrink (warp) time-series non-linearly along its time axis to find the optimal correlation between different time-series [18], as shown in Fig. 2.

DTW has many implementations in different disciples, such as gesture recognition, robotics, and manufacturing. However, it was mainly used for data mining as a distance measure between time-series data points [18]. DTW is not sensitive to time-shifting, and it does not require the time-series to be on the same length as a condition to compare among them [45]. To compare time-series T1, T2 of lengths

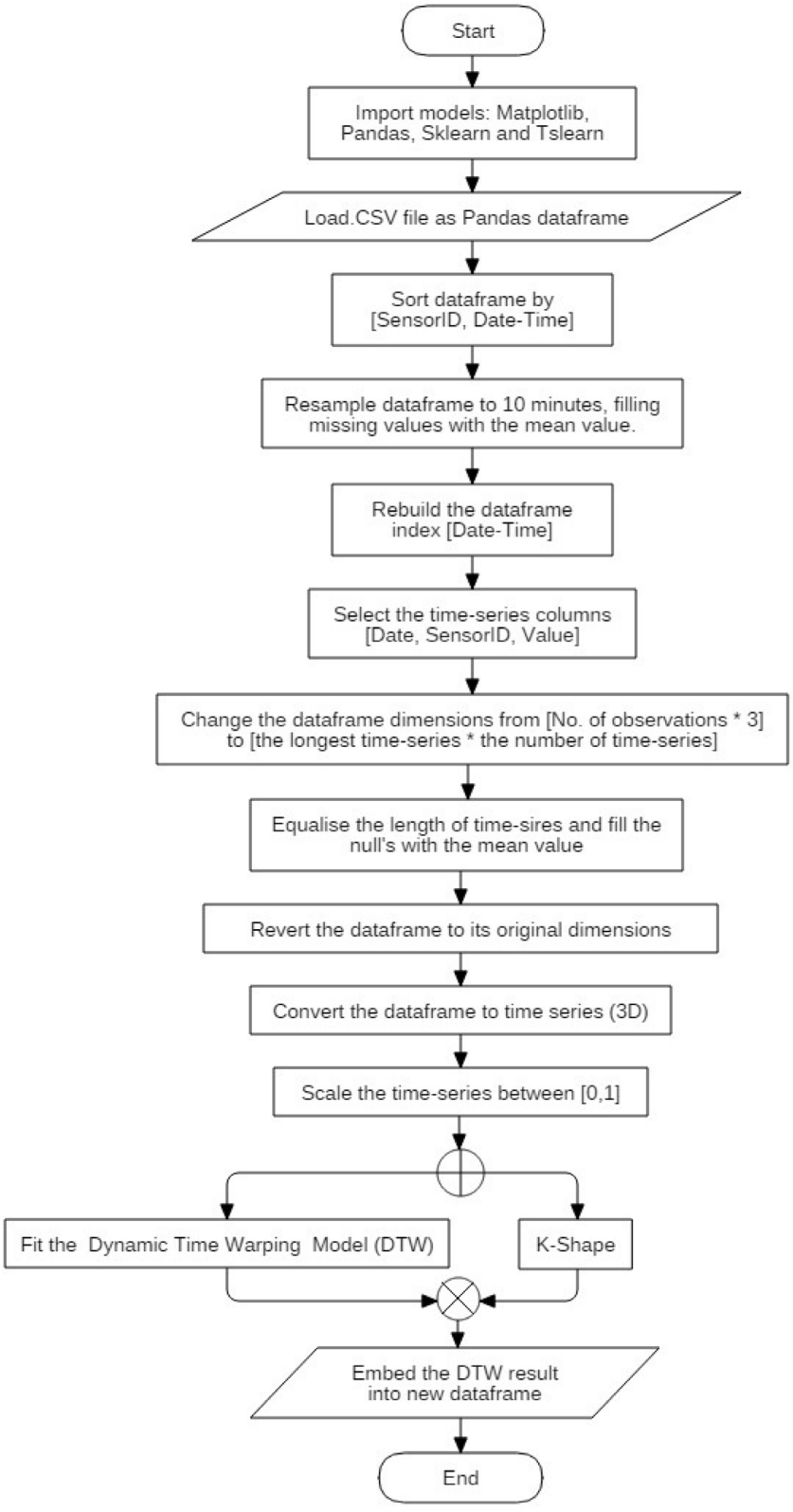

Fig. 12. The process diagram of the technical steps implemented to fit all available time-series as a 3D array into the Dynamic Time Warping (DTW) and K-Shape time-series clustering models.

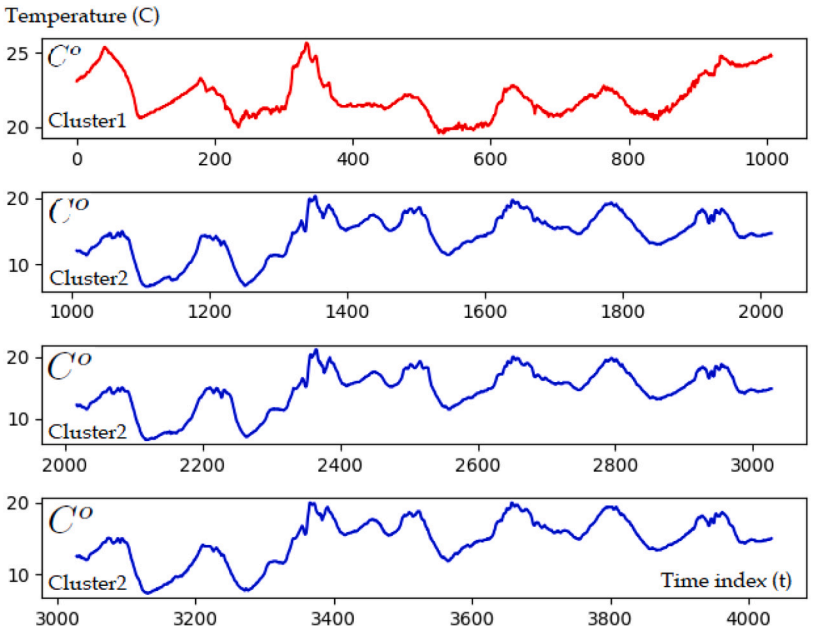

Fig. 13. The feature-based time-series clustering method differentiated the indoors time-series (Cluster 1) from the other typical time-series (Cluster 2) of the ideal dataset.

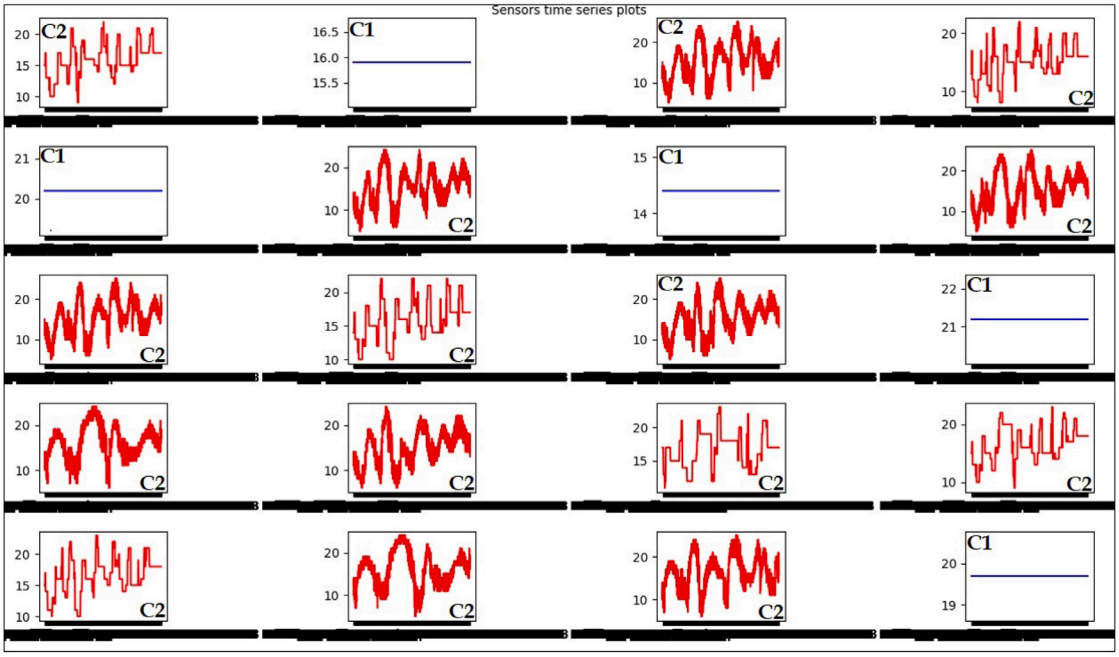

Fig. 14. The feature-based time-series clustering technique successfully differentiated time-series with long-segmental outliers (C1 = Cluster 1) from other typical time-series (C2 = Cluster 2) when applied to the real-world seven-day time window.

n and m, DTW is going to measure the distance (T1, T2) with time complexity of $(n * m)$. Thus, DTW is a computationally expensive method for simultaneously clustering long time-series or numerous time-series [46].

3.2. K-Shape

K-Shape and Dynamic Time Warping (DTW) are shape-based time-series clustering methods. K-Shape is a time-series clustering algorithm that uses cross-correlation measures to measure the distance and the centroids for time-series clusters. K-Shape analyses the shape of the time series while clustering them. The theory behind K-Shape is similar to the one used by the K-means algorithm. K-means is a distance-based clustering algorithm that divides the unlabelled dataset into several k non-overlapping subsets (clusters), each of which is represented by the mean of the distance between its data points [47]. Both K-Shape and K-means rely on the iterative refinement procedure, which scales linearly and produces equivalent and sufficiently separated clusters. K-Shape is considered a highly efficient and more domain-independent time-series clustering method than the DTW method. DTW considers the shape similarity between time series regardless of differences in amplitude

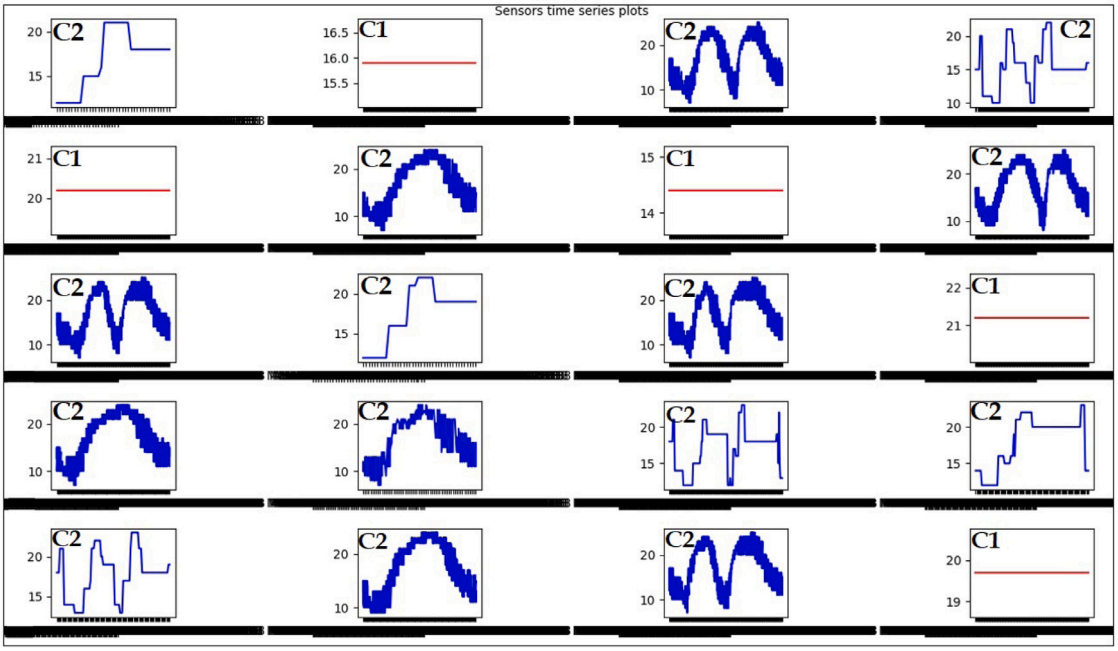

Fig. 15. The feature-based time-series clustering technique successfully differentiated time-series with long-segmental outliers (C1 = Cluster 1) from other typical time-series (C2 = Cluster 2) even when applied to relatively short time-series (two-day time window) of the real-world dataset.

and phase. At the same time, K-Shape relies on the time-series cross-correlation measures, which are significantly faster than the time-series distance measures method adopted by DTW [48].

3.3. Characteristic-based time-series clustering

Characteristic-based time-series clustering is also known as features extraction-based or statistical characteristics-based time-series clustering. Unlike the shape-based time-series clustering methods, such as DTW or K-Shape, the characteristic-based clustering does not use the distance measure or the cross-correlation measures methods. Alternatively, this method clusters time-series based on their captured global characteristics using classical statistical methods. The features extracted from each time series can be fitted into any arbitrary clustering algorithm. The extracted features describe a time series' statistical characteristics (global measures). These features can be extended to over 100 different features, such as the absolute sum of changes, autocorrelation, standard deviation and partial autocorrelation. Characteristic-based clustering reduces time-series dimensions, making it much less sensitive to the effect of missing values or noisy data. The advantage of characteristic-based clustering is its high performance, even if used to perform similarity searches or clustering amongst very long time series [49].

4. Experimental settings and London case study

Time-series clustering techniques are used in this study to detect continuous (halting), abrupt (emerging) and incipient faults using real-world datasets collected from two different sensor node networks, as follows:

4.1. Large-scale sensor node network

Large-scale sensor node network is the primary data source of this study. It consists of 275 temperature sensor nodes distributed around London and managed by different providers, such as the Meteorological Office (Met Office) [50], Open Weather Map [51] and Smart Citizen [52]. The geographical distribution of these sensors is shown in Fig. 3.

Data streams from these sensor nodes were coordinated by an Internet of Things (IoT) search engine called Thingful [53]. Thingful is owned by a U.K. based company named Umbrellium.Ltd [54] specialises in IoT projects associated with smart cities, connected vehicles, machine learning and big data analytics. These sensor nodes data streams collected through the Thingful network will be utilised to test the ability of the time-series clustering techniques to detect continuous (halting) and abrupt (emerging) long-outliers. Thus, these types of long-outliers have been detected in some time series of this large-scale dataset while focusing on temperature time series. Typically, temperature time-series show daily seasonality and a trend, as shown in Fig. 4.

Therefore, temperature time series with a constant value attribute or a very low seasonality for a relatively long-time (long-outlier) are highly likely to encompass data quality issues related to observations accuracy. This behaviour can be related to a sensor node's hardware failure that affects their detection ability, or it may indicate that these time-series were streamed from sensor nodes that are down, and the system compensates for their missing observations by repeating the last observation received from these faulty sensors. An example of long-segmental outliers is shown in Fig. 5. Fig. 5 shows the time-series of two sensors which displays constant readings (Fig. 5 Sensor ID= kt3m3nw5 and rw4egmaw/Smart Citizen) for a relatively long time, compared to another time-series (Fig. 5 Sensor ID=47qwbfba/Meteorological Office) generated by a functional sensor node managed by the Meteorological Office during the same time window.

The accuracy and performance of the three time-series clustering methods, DTW and K-Shape and the characteristics-based time-series clustering technique, examined in this study are evaluated based on their ability to identify time-series with halting or emerging long-segmental outliers and the time required to render the clustering results. The evaluation was conducted using two different time-series windows. The first is a seven-day time-series window. The second is a two-day time-series window used to evaluate the accuracy and performance of time-series clustering techniques compared to the seven-day

time window. The time series were collected from a local network of four sensor nodes and a large-scale network of 275 sensor nodes distributed around London.

4.2. Local sensor node network

This time series was used as a benchmark dataset to test the ability of the time series clustering techniques to detect incipient faults with consistent offset (long-outlier). The local sensor node network consists of four high-quality sensor nodes deployed at The University of East London. One of these network sensor nodes is installed indoors, and the other three are deployed outdoors. Since all the local sensor nodes were distributed in a relatively small geographical area, their time-series do not show significant differences in the shape of their trend. However, they show some differences in the value attribute, especially with the indoors sensor, which streamed a time series with a consistent offset of 10–15 C^o from other sensors. Thus, the indoors sensor node, in this case, represented a sensor with an incipient fault pattern, as shown in Fig. 6. Furthermore, since this time series is a high-quality dataset with no missing values or outliers, it was used as a benchmark to test and calibrate the time-series clustering techniques before applying them to the large-scale sensor node network time series.

5. Dynamic Time Warping (DTW) and K-Shape

The Dynamic Time Warping (DTW) and K-Shape time-series clustering were applied using the Python package *tslearn.clustering* provided by Scikit-learn [55]. The outcome from applying the DTW and K-Shape time-series clustering techniques to the local sensor node network dataset is shown in Fig. 7 (identical outcome).

Both DTW and K-Shape time-series clustering techniques successfully identified the time series of the indoor sensor node (the incipient fault pattern Sensor ID = 493361) from the other time series of the outdoor sensor nodes. This result is significant because both DTW and K-Shape are shape-based time series clustering techniques, and all the time series used in this test exhibit a significant similarity in the shape pattern, as shown in Fig. 6. The top graph line in Fig. 7 is the indoors sensor (Sensor ID = 493361) time series in the first cluster (incipient), while Cluster 2 (the bottom graph lines) presents the other (normal) time-series.

The time series used in the second test were collected from the large-scale sensor network. The dataset of this test is much larger than the dataset of the local sensor node network. Both DTW and K-Shape rendered identical clustering results when applied to the seven-day time series, as shown in Figs. 8 and 9.

Both DTW and K-Shape successfully separated time-series with long-segmental outliers from the other time-series that exhibit typical variation in the trend and seasonality when applied to the seven-day time window. The temperature axis (y-axes) in Figs. 8 and 9 do not reflect the actual value attribute of the observations, since all time-series were normalised using the Python package *"tslearn.preprocessing.TimeSeriesScalerMeanVariance"* [55] so that each output time series had zero mean and unit variance before being fitted to the time-series clustering models, as shown in Fig. 12. Applying DTW and K-Shape to the two-day time series showed that DTW is more sensitive to the window length of the clustered time series than the K-Shape. The ability of the DTW to differentiate the faulty time-series from other (typical) ones was more significantly affected compared to K-Shape, as shown in Figs. 10 and 11.

In general, both shape-based time-series clustering techniques require relatively long time series to enhance their clustering results, especially the DTW. Figs. 10 and 11 illustrate that K-Shape can maintain its clustering accuracy when applied to a relatively shorter time series than DTW. Both techniques were able to differentiate time-series that showed the patterns of the continuous and abrupt long-segmental outliers with 100% accurate detection ratio when applied to seven days, or longer time-series, as shown in Figs. 8 and 9. The main technical steps required to fit all available time series from all sensor nodes as a three-dimensional data array to the DTW and K-Shape models are illustrated in the process flowchart diagram shown in Fig. 12.

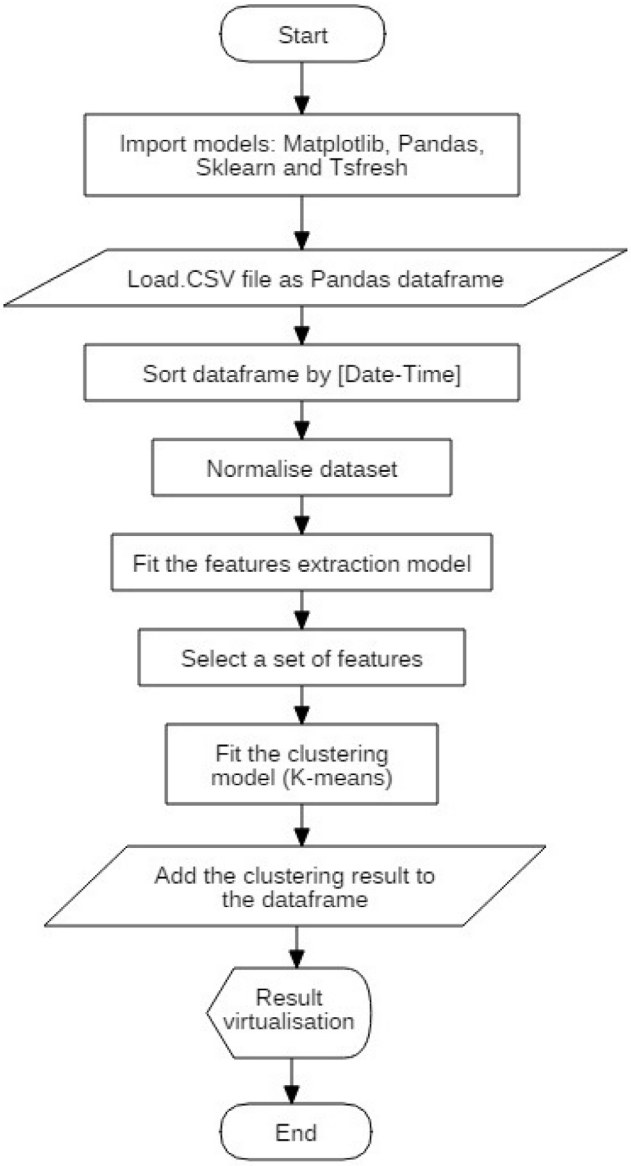

Fig. 16. The process diagram of the technical steps implemented to fit time-series to the characteristic-based time-series clustering model.

6. Characteristics-based time-series clustering

The characteristics (features) based time-series clustering technique was implemented using [56]'s Python *tsfresh* package and the time series collected from the local sensor node network as a benchmark test. The characteristic-based time-series clustering technique successfully separated the time-series of the indoors sensor node (incipient faults pattern) from the rest of the time series, as shown in Fig. 13.

The characteristics based time series clustering model relies on using arbitrary clustering algorithms, such as K-means, for clustering the set of features extracted from the examined time series. The selected features may vary from one application to another based on the characteristics of the time series chosen to be used as clustering

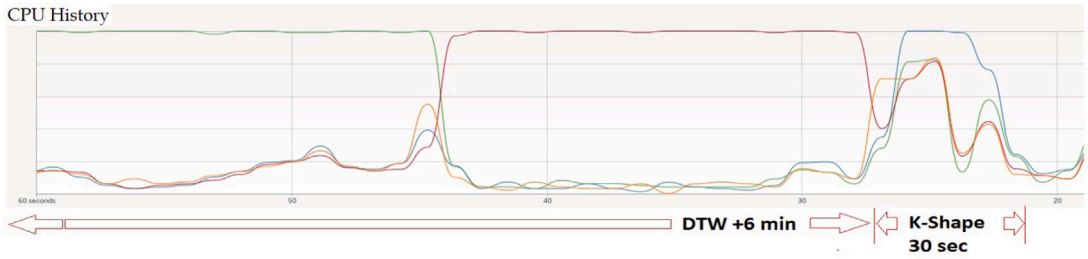

Fig. 17. The performance of the CPU cores and the time required to perform the same task by DTW comparing to K-Shape, each graph line represents the performance of a single CPU core.

reference. In this case study, the "absolute sum of changes" was the main parameter used and fitted to the K-means clustering model to detect continuous (halting/repeating) faults of sensors time series that show no or minimal variation in their observations value attributes. The outcome of applying the feature-based time-series clustering technique to the time series of the large-scale sensor node network is shown in Figs. 14 and 15.

The feature-based time series clustering technique successfully categorised time series with long-segmental outliers even when applied to the relatively short two-day time-series window, as shown in Fig. 15. The technical aspects required to fit all available time series to the characteristic-based time-series clustering models are illustrated in the flowchart diagram in Fig. 16.

Since all the used time-series clustering techniques (DTW, K-Shape and Characteristics-Based Time-Series Clustering) were applied to the same dataset, it was possible to evaluate the performance of each of these methods based on the time spent to render their clustering results. DTW required a significant amount of time to render the results, around 360 s compared to the feature-based and K-Shape time-series clustering techniques which required around 30 s when applied to the seven-day time window. It is essential to highlight that these results may vary according to the number and the type of the extracted features and the selected clustering algorithm. Although DTW demanded more time than K-Shape to render the clustering results, it seems that the K-Shape Python package managed the processing resources of the CPU cores more efficiently than DTW, as shown in Fig. 17.

Note : All tests of this study were conducted using Python 3.7 64 bits installed over a Linux (Fedora 64 bits) workstation. The processor of the workstation is an Intel(R) Core (TM) i7-7920HQ CPU @ 3.10 GHz (4 CPUs), 3.1 GHz with 32 GB of RAM.

7. Conclusion

The novelty of this research lies in the successful implementation of time-series clustering techniques as a sensors failure detection mechanism via detecting long-segmental outliers associated with time-series of faulty sensors. The purpose of this study is to test Dynamic Time Warping (DTW), K-Shape, and the Characteristics-based time series clustering technique as long-segmental outlier detection methods for sensor node fault detection in large-scale CPSs. The study focused on detecting the failures of continuous (halting/repeating) and incipient sensor nodes. The time series clustering techniques were evaluated using real-world observations collected from two real-world sensor node networks. All of the examined time series clustering techniques proved their ability to detect sensor node faults associated with long-outliers in their time series with some differences in accuracy and complexity. The feature-based time series clustering technique maintained its detection accuracy even when applied to a relatively short time series compared with the shape-based (DTW and K-Shape) time series clustering techniques. Furthermore, the empirical tests of these techniques showed that feature-based time-series clustering could be a more efficient long-segmental outlier detection mechanism than the shape-based time-series clustering techniques, such as DTW and K-Shape, mainly when applied to shorter time-series windows.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This research was supported by the University of East London (UEL) through the PhD scholarship scheme and the UEL Research Internship programme in collaboration with Umbrellium Ltd, UK, which gave access to their sensors data; in particular, we thank Usman Haque for his insight and expertise that greatly assisted the research.

References

[1] M. Törngren, F. Asplund, S. Bensalem, J. McDermid, R. Passerone, H. Pfeifer, A. Sangiovanni-Vincentelli, B. Schätz, Characterization, analysis, and recommendations for exploiting the opportunities of cyber-physical systems, Cyber-Phys. Syst. Found. Princ. Appl. (2017) 3–14, http://dx.doi.org/10.1016/b978-0-12-803801-7.00001-8, URL: https://www.sciencedirect.com/science/article/pii/B9780128038017000018.

[2] A. Platzer, Logical Foundations of Cyber-Physical Systems, Springer, 2019.

[3] S. Kounev, J.O. Kephart, A. Milenkoski, X. Zhu, S.I.P. Ag, Self-Aware Computing Systems, Cham Springer International Publishing Springer, 2018.

[4] N.U. Okafor, Y. Alghorani, D.T. Delaney, Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach, ICT Express 6 (3) (2020) 220–228, http://dx.doi.org/10.1016/j.icte.2020.06.004.

[5] G.R.C. de Aquino, C.M. de Farias, L. Pirmez, Hygieia: data quality assessment for smart sensor network, in: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Association for Computing Machinery, New York, NY, USA, 2019, http://dx.doi.org/10.1145/3297280.3297564.

[6] A. Abid, A. Kachouri, A. Ben Fradj Guiloufi, A. Mahfoudhi, N. Nasri, M. Abid, Centralized KNN anomaly detector for WSN, in: 2015 IEEE 12th International Multi-Conference on Systems, Signals Devices (SSD15), 2015, pp. 1–4, http://dx.doi.org/10.1109/SSD.2015.7348091.

[7] P.S. Banerjee, S.N. Mandal, D. De, B. Maiti, RL-sleep: Temperature adaptive sleep scheduling using reinforcement learning for sustainable connectivity in wireless sensor networks, Sustain. Comput.: Inf. Syst. 26 (2020) 100380, http://dx.doi.org/10.1016/j.suscom.2020.100380, URL: https://www.sciencedirect.com/science/article/pii/S2210537919301155.

[8] A.A. Alwan, M.A. Ciupala, A.J. Brimicombe, S.A. Ghorashi, A. Baravalle, P. Falcarin, Data quality challenges in large-scale cyber-physical systems: a systematic review, Information Systems (ISSN: 0306-4379) 105 (2022) 101951, http://dx.doi.org/10.1016/j.is.2021.101951, https://www.sciencedirect.com/science/article/pii/S0306437921001484.

[9] D. Li, L. Yan, Y. Liu, Q. Yin, S. Guo, H. Zheng, Data quality improvement method based on data correlation for power internet of things, in: 2019 12th International Symposium on Computational Intelligence and Design (ISCID), Vol. 2, 2019, pp. 259–263, http://dx.doi.org/10.1109/ISCID.2019.10142.

[10] N. Ghosh, R. Paul, S. Maity, K. Maity, S. Saha, Fault matters: Sensor data fusion for detection of faults using Dempster–Shafer theory of evidence in IoT-based applications, Expert Syst. Appl. 162 (2020) 113887, http://dx.doi.org/10.1016/j.eswa.2020.113887, URL: https://www.sciencedirect.com/science/article/pii/S0957417420306898.

[11] V.J. Lawson, L. Ramaswamy, TAU-FIVE: a multi-tiered architecture for data quality and energy-sustainability in sensor networks, in: 2016 International Conference on Distributed Computing in Sensor Systems (DCOSS), IEEE, 2016, pp. 169–176.

[12] M.Z.A. Bhuiyan, J. Wu, G. Wang, Z. Chen, J. Chen, T. Wang, Quality-guaranteed event-sensitive data collection and monitoring in vibration sensor networks, IEEE Trans. Ind. Inf. 13 (2) (2017) 572–583, http://dx.doi.org/10.1109/TII.2017.2665463.

[13] W. Liao, S. Kuai, C. Chang, Energy harvesting path planning strategy on the quality of information for wireless sensor networks, in: 2019 IEEE International Conference of Intelligent Applied Systems on Engineering (ICIASE), 2019, pp. 82–85, http://dx.doi.org/10.1109/ICIASE45644.2019.9074030.

[14] R. Togneri, G. Camponogara, J. Soininen, C. Kamienski, Foundations of data quality assurance for IoT-based smart applications, in: 2019 IEEE Latin-American Conference on Communications (LATINCOM), 2019, pp. 1–6, http://dx.doi.org/10.1109/LATINCOM48065.2019.8937930.

[15] P. Du, Q. Yang, Z. Shen, K.S. Kwak, Quality of information maximization in lifetime-constrained wireless sensor networks, IEEE Sens. J. 16 (19) (2016) 7278–7286, http://dx.doi.org/10.1109/JSEN.2016.2597439.

[16] D.-I. Curiac, C. Volosencu, Ensemble based sensing anomaly detection in wireless sensor networks, Expert Syst. Appl. 39 (10) (2012) 9087–9096, http://dx.doi.org/10.1016/j.eswa.2012.02.036, URL: https://www.sciencedirect.com/science/article/pii/S0957417412002801.

[17] MetOffice, National meteorological library and archive fact sheet 14 - microclimates, 2019, URL: https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/research/library-and-archive/library/publications/factsheets/factsheet_14-microclimates.pdf.

[18] S. Salvador, P. Chan, Toward accurate dynamic time warping in linear time and space, Intell. Data Anal. 11 (5) (2007) 561–580, http://dx.doi.org/10.5555/1367985.1367993.

[19] L. Chen, Y. Ho, H. Hsieh, S. Huang, H. Lee, S. Mahajan, ADF: An anomaly detection framework for large-scale PM2.5 sensing systems, IEEE Internet Things J. 5 (2) (2018) 559–570.

[20] A.A. Alwan, M.A. Ciupala, A. Baravalle, P. Falcarin, HADES: a hybrid anomaly detection system for large-scale cyber-physical systems, in: 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), 2020, pp. 136–142, http://dx.doi.org/10.1109/FMEC49853.2020.9144751.

[21] D.M. Hawkins, Identification of Outliers, Springer, Dordrecht, The Netherlands, 1980, http://dx.doi.org/10.1007/978-94-015-3994-4.

[22] A. Appice, A. Ciampi, F. Fumarola, D. Malerba, Sensor networks and data streams: Basics, in: Data Mining Techniques in Sensor Networks, Springer, 2014, pp. 1–8.

[23] M.-H. Lee, Y.-H. Choi, Fault detection of wireless sensor networks, Comput. Commun. 31 (14) (2008) 3469–3475, http://dx.doi.org/10.1016/j.comcom.2008.06.014, URL: http://www.sciencedirect.com/science/article/pii/S0140366408003587.

[24] Hanrong Lu, Xin Chen, Xuhui Lan, Feng Zheng, Duplicate data detection using GNN, in: 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2016, pp. 167–170, http://dx.doi.org/10.1109/ICCCBDA.2016.7529552.

[25] H. Liu, X. Wang, S. Lei, X. Zhang, W. Liu, M. Qin, A Rule Based Data Quality Assessment Architecture and Application for Electrical Data, Association for Computing Machinery, New York, NY, USA, 2019, http://dx.doi.org/10.1145/3371425.3371435.

[26] Y. Xinrui, W. Lei, L. Ruiyi, Data quality evaluation of Chinese wind profile radar network in 2018, in: 2019 International Conference on Meteorology Observations (ICMO), 2019, pp. 1–4, http://dx.doi.org/10.1109/ICMO49322.2019.9026025.

[27] M. Jayswal, M. Shukla, Consolidated study analysis of different clustering techniques for data streams, in: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 3541–3547.

[28] A. Abid, A. Kachouri, A. Mahfoudhi, Outlier detection for wireless sensor networks using density-based clustering approach, IET Wirel. Sensor Syst. 7 (4) (2017) 83–90, http://dx.doi.org/10.1049/iet-wss.2016.0044.

[29] N. Nesa, T. Ghosh, I. Banerjee, Outlier detection in sensed data using statistical learning models for IoT, in: 2018 IEEE Wireless Communications and Networking Conference (WCNC), 2018, pp. 1–6, http://dx.doi.org/10.1109/WCNC.2018.8376988.

[30] P.M. Laso, D. Brosset, J. Puentes, Analysis of quality measurements to categorize anomalies in sensor systems, in: 2017 Computing Conference, IEEE, 2017, pp. 1330–1338.

[31] T.J. Chandler, The Climate of London, Hutchinson, 1965.

[32] M. Adhikari, S. Kar, S. Banerjee, U. Biswas, Big data analysis for cyber-physical systems, 2015, Undefined. URL: https://www.semanticscholar.org/paper/Big-Data-Analysis-for-Cyber-Physical-Systems-Adhikari-Kar/2fbe376b34c56ef8a3aa12797fd111c1aa58ae4b.

[33] D.B. Rawat, J.J.P.C. Rodrigues, I. Stojmenovic, Cyber-Physical Systems: From Theory to Practice, CRC Press, Inc., USA, 2015.

[34] B. Ratner, Statistical and Machine-Learning Data Mining, Third Edition: Techniques for Better Predictive Modeling and Analysis of Big Data, third ed., Chapman & Hall/CRC, Chapman & Hall/CRC, 2017, http://dx.doi.org/10.5555/3161097.

[35] M.M. Farooqi, H. Ali Khattak, M. Imran, Data quality techniques in the internet of things: Random forest regression, in: 2018 14th International Conference on Emerging Technologies (ICET), 2018, pp. 1–4, http://dx.doi.org/10.1109/ICET.2018.8603594.

[36] K. Berk, Time series analysis, in: Modeling and Forecasting Electricity Demand: A Risk Management Perspective, Springer Fachmedien Wiesbaden, Wiesbaden, 2015, pp. 25–52, http://dx.doi.org/10.1007/978-3-658-08669-5_3.

[37] C.C. Aggarwal, Managing and Mining Sensor Data | Charu C. Aggarwal | Springer, Springer US, 2013, http://dx.doi.org/10.1007/978-1-4614-6309-2.

[38] Y. Zhuang, L. Chen, In-network outlier cleaning for data collection in sensor networks, in: CleanDB, 2006.

[39] C.C. Aggarwal, Data Mining: The Textbook, Springer, 2015.

[40] O. Ghorbel, M.W. Jmal, M. Abid, H. Snoussi, Distributed and efficient one-class outliers detection classifier in wireless sensors networks, in: International Conference on Wired/Wireless Internet Communication, Springer, 2015, pp. 259–273.

[41] F. Sailhan, T. Delot, A. Pathak, A. Puech, M. Roy, Dependable sensor networks, 2010, pp. 1–15, URL: https://hal.archives-ouvertes.fr/hal-01125818.

[42] J. Dean, Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, John Wiley & Sons, 2014.

[43] S. Aghabozorgi, A. Seyed Shirkhorshidi, T. Ying Wah, Time-series clustering – A decade review, Inf. Syst. 53 (2015) 16–38, http://dx.doi.org/10.1016/j.is.2015.04.007, URL: https://www.sciencedirect.com/science/article/pii/S0306437915000733.

[44] C.C. Aggarwal, Time series and multidimensional streaming outlier detection, in: Outlier Analysis, Springer, 2017, pp. 273–310.

[45] Z. Bankó, J. Abonyi, Correlation based dynamic time warping of multivariate time series, Expert Syst. Appl. 39 (17) (2012) 12814–12823, http://dx.doi.org/10.1016/j.eswa.2012.05.012, URL: https://www.sciencedirect.com/science/article/pii/S0957417412007105.

[46] C.C. Aggarwal, C.K. Reddy, Data clustering, in: Algorithms and Applications, Chapman&Hall/CRC Data Mining and Knowledge Discovery Series, Londra, 2014.

[47] J.A. Hartigan, M.A. Wong, Algorithm AS 136: A k-means clustering algorithm, J. Roy. Statist. Soc. Ser. C 28 (1) (1979) 100–108.

[48] J. Paparrizos, L. Gravano, K-shape: Efficient and accurate clustering of time series, SIGMOD Rec. 45 (1) (2016) 69–76, http://dx.doi.org/10.1145/2949741.2949758.

[49] X. Wang, K. Smith, R. Hyndman, Characteristic-based clustering for time series data, Data Min. Knowl. Discov. 13 (3) (2006) 335–364, http://dx.doi.org/10.1007/s10618-005-0039-x.

[50] MetOffice, Weather and climate change, 2021, URL: https://www.metoffice.gov.uk. [Online; accessed 24. May 2021].

[51] OpenWeatherMap, Current weather and forecast - OpenWeatherMap, 2021, URL: https://openweathermap.org. [Online; accessed 24. May 2021].

[52] S. Citizen, Smart citizen empowers communities to better understand their environment, 2021, URL: https://smartcitizen.me. [Online; accessed 24. May 2021].

[53] Thingful, A search engine for the internet of things, 2021, URL: https://www.thingful.net. [Online; accessed 24. May 2021].

[54] Umbrellium, Engaging cities, 2021, URL: https://umbrellium.co.uk. [Online; accessed 24. May 2021].

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[56] M. Christ, N. Braun, J. Neuffer, A.W. Kempa-Liehr, Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package), Neurocomputing 307 (2018) 72–77.

Ahmed A Alwan got his Ph.D. in computer science at the University of East London in 2021, after working for years as Database expert for the Iraqi government. His research interests are on data quality and reliability in wireless sensor networks.

Allan Brimicombe is Professor Emeritus at the University of East London where he was Head of the Centre for Geo-Information Studies for 20 years. Allan is co-currently a Director of Terra Cognita Limited engaged in independent research, publishing and bespoke consultancy. He is a Chartered Geographer, a Fellow of the Royal Statistical, Royal Geographical and Geological Societies, a Director of the British Society of Criminology, and is a Fellow of the Academy of Social Sciences. Allan has worked in both the private sector and academia and gained his higher degrees at the University of Hong Kong whilst 19 years abroad in the Far East where he pioneered spatial decision-support systems using numerical simulation techniques. His other research interests include data quality issues, spatial data mining and machine learning, quantitative and mixed methods. These have been applied to crime, health, education, natural hazards, utilities and business. Allan has been a Specialist Adviser to the House of Lords Select Committee on Olympic and Paralympic Legacy.

Dr Mihaela Anca Ciupala received the Ph.D. degree in Earthquake Engineering and Safety of Structures from the Technical University "Gh Asachi", Iasi, Romania, in 1999. She is currently a Reader in Structural Engineering, University of East London, UK. She was an EU Marie Curie research Fellow, and her main research interests span over the areas of Structural Engineering, Environmental Engineering and Data Science, focusing on the application of machine/deep learning, and computer vision for structural and environmental solutions.

Seyed Ali Ghorashi received his B.Sc. and M.Sc. degrees from the School of Electrical and Computer Engineering, the University of Tehran, Iran and his Ph.D. degree from Centre for Telecommunications Research (CTR), King's College London, UK. He has worked for Samsung Electronics (UK) Ltd, Shahid Beheshti University, Middlesex University, Goldsmiths, University of London and now is working as senior lecturer at the University of East London. He is a senior member of IEEE, holds US and international patents and has published over 120 technical papers mainly on the applications of optimisation, artificial intelligence and machine learning in Internet of Things (IoT) and Wireless Communications.

Andres Baravalle got his Ph.D. in Computer Science at Turin University in 2004. He has worked in the UK for the Open University, the University of Sheffield, and for ten years at the University of East London as a senior lecturer. His research focus on the interaction of web, security and data. In the past years, his research on the Dark Web has been featured on the first page of The Times, on the BBC and on most of main UK newspapers. He recently moved to industry as Principal Product Data Analyst.

Paolo Falcarin received his Master and Ph.D. degrees in Software Engineering from the Polytechnic of Turin (Italy) in 2004, where he worked as research fellow. Afterwards he worked as a Reader at the University of East London (UK), where he led the Secure Software Engineering research group until 2021. He is currently Associate Professor at Ca' Foscari University of Venice (Italy). He has published more than 80 papers in the areas of Software Security, Software Engineering, and Distributed Systems.