HADES: a Hybrid Anomaly Detection System for Large-Scale Cyber-Physical Systems

Ahmed Abdulhasan Alwan
School of Architecture Computer Science and Engineering
University of East London
London, United Kingdom
a.alwan@uel.ac.uk

Mihaela Anca Ciupala
School of Architecture Computer Science and Engineering
University of East London
London, United Kingdom
m.a.ciupala@uel.ac.uk

Andres Baravalle
Akamai Technologies Ltd.
London, United Kingdom
a.baravalle@akamai.com

Paolo Falcarin
School of Architecture Computer Science and Engineering
University of East London
London, United Kingdom
falcarin@uel.ac.uk

***Abstract*— Smart cities rely on large-scale heterogeneous distributed systems known as Cyber-Physical Systems (CPS). Information systems based on CPS typically analyse a massive amount of data collected from various data sources that operate under noisy and dynamic conditions. How to determine the quality and reliability of such data is an open research problem that concerns the overall system safety, reliability and security. Our research goal is to tackle the challenge of real-time data quality assessment for large-scale CPS applications with a hybrid anomaly detection system. In this paper we describe the architecture of HADES, our Hybrid Anomaly DEtection System for sensors data monitoring, storage, processing, analysis, and management. Such data will be filtered with correlation-based outlier detection techniques, and then processed by predictive analytics for anomaly detection.**

Keywords*— *Data Quality, Quality of Service, Cyber-Physical System, Smart Cities, Anomaly Detection, Predictive Analysis, Wireless Sensor Network

I. INTRODUCTION AND BACKGROUND

Cyber-Physical Systems (CPS) are advanced information systems that integrate communication, computation and control with physical processes to add new capabilities to physical systems [1]. CPS are an active area of research in many crucial domains, such as manufacturing, health care, smart power grids, transportation, and smart cities [2]. CPS are the backbone of next-generation information technology of the fourth industrial revolution (Industry 4.0) [3] and have significant importance for smart cities development [4].

In smart cities, CPS form large-scale distributed information systems. In most cases, CPS have real-time requirements where data have to be sensed and processed in real-time: traffic control systems, energy management systems, transportation systems, water resources monitoring and control systems, logistics and disaster management systems [5][6], and environment monitoring systems all typically involve a large number of sensor nodes deployed in broad geographical territories to form large-scale real-time cyber-physical systems [7]. As CPS collect and analyse data in real-time, the quality of service provided by the system relies on the quality of the collected data [8]. The quality of data in CPS is a significant concern, especially for applications which analyse a massive amount of data from various sources and operate under noisy and dynamic conditions [9]. Since CPS rely on sensor-nodes which are usually deployed in uncontrolled, remote environments, unexpected measurements may occur due to external noise, or due to issues in the sensor nodes themselves such as power failure, calibration and ageing. Thus, these unexpected measurements need to be carefully managed and interpreted based on domain knowledge and computational models [10]. For example, CPS applications such as environmental monitoring systems, typically involve a large number of sensor nodes deployed in a broad geographical territory to form a large-scale sensor-nodes network [7][11].

A sensor-nodes network may consist of few to thousands of small-size and low-cost sensor-nodes that do not have the computational power to execute complex analysis or calculations. Their only role is sensing and sending observations to another neighbour sensor-node, to the network, or directly to a central server [12]. Sensor-nodes cannot decide if the values of their observations are accurate or not: because of their limitation of resources, they cannot host built-in mechanisms for data quality assessment, that are particularly computationally intensive [13].

According to Shih et al. [14], ensuring data quality in cyber-physical systems is a challenge that has not been fully addressed yet. For example, more work is needed to tackle the issue of data lifetime, which has not been investigated enough, especially in applications in which data validity has a limited duration [14]. CPS are real-time systems and in case of data transmission delays, missing critical readings from sensors, or receiving incorrect observations, there might be life-threatening consequences, due to compromised safety constraints [15].

In addition, more research is needed to interpret data quality standards and data quality assessment methods, especially for applications which analyse a massive amount of data from various resources [16]. Although the challenge of data quality management is as old as data itself [17], it has a higher level of impact now, especially considering real-time CPS applications which might involve telecom services [18], business corporations and government agencies [19]; data management can be even more difficult when considering mobile CPS [20] that may include smartphones data and user-generated contents [21] that have short and volatile lifetime.

In this paper we describe the architecture of HADES, (Hybrid Anomaly DEtection System) a software for sensors data monitoring, storage, analysis, and management. Such data will be filtered with correlation-based outlier detection techniques, and then processed by predictive analytics for anomaly detection on large-scale Cyber-Physical Systems.

II. SMART CITIES AS LARGE-SCALE CPS

Smart cities are advanced information systems that rely on data coming from large-scale heterogeneous Cyber-Physical Systems to provide more automated and efficient services to improve the life quality of smart cities' residents [22]. CPS rely on networks of distributed wireless connected sensor-nodes known as wireless sensor networks (WSNs), which have been widely utilized, from building control to environmental monitoring systems [19].

A wireless sensor network typically consists of a group of a specialised micro-sensors usually deployed in the area of interest for monitoring physical or environmental phenomena such as temperature, humidity or seismic events [23]. Sensor-nodes are the building blocks of wireless sensor networks, a low-cost monitoring tool with one major drawback which is the reduced power capacity which may limit their service lifetime [24] [25].

WSNs have been employed successfully in many large-scale applications which involve deploying sensor-nodes in remote and uncontrolled environments, such as for environmental monitoring and agriculture monitoring [26]. Unexpected measurements may occur due to external noise or may be caused by the sensor-node itself, e.g. due to power failure, calibration, or ageing.

Large-scale CPS applications are typically based on components from different manufacturers and with different implementations; as they usually operate under noisy and dynamic conditions, their operation may involve data quality issues [9][6]. In general, data quality issues in CPS information systems can be classified into three types:

- Errors in measurements caused by precision problems within the sensors.
- Noise in the communication networks.
- Loss of precision in sampling discrete measurements of continuous variables, related to both spatial and time parameters.

The challenges of data quality assessment become greater in applications dealing with large volumes of data and having restricted requirements of data availability. Table I shows the main data quality characteristic and challenges in CPS application [27], [9].

TABLE I. MAIN DATA CHARACTERISTICS AND CHALLENGES IN CPS APPLICATIONS.

Characteristics		Challenges
Big Data	Variety	Interoperability
	Volume and Velocity	Scalability
Quality of Information	Granularity	Discrete measurements (spatial and time)
	Precision	Device calibration, accuracy and adaptive sampling
	Accuracy	Noise in communication networks
Constraints	Energy	Context-aware data collection and processing
	Connectivity	

III. DATA QUALITY AND ANOMALY DETECTION

Anomaly Detection is the process of identifying unusual patterns in data sets which do not comply with well-established normal behaviour [28]. These atypical patterns in data sets are called anomalies or outliers. An observation from a sensor-node is considered to be an anomaly or outlier if its value significantly deviates from a pre-calculated threshold value [29].

Anomaly detection techniques can be grouped in two main categories: (i) correlation-based anomaly detection, and (ii) predictive analytics-based anomaly detection.

A. Correlation-Based Anomaly Detection

Correlation-based anomaly detection models rely on the assumption that the values of sensor-nodes observations are correlated spatially, temporally, or both spatial and temporal. However, this assumption is not necessarily always valid, especially in large-scale CPS applications where the correlation between sensor-nodes observations may be affected by many parameters such as the scale of deployment area and the geographical distribution of sensor-nodes [30]. Another limitation is the challenge of calculating threshold values of observations for each sensor-node based on the temporal model in a large-scale CPS application, which needs a relatively long time-series of anomaly-free data, which cannot be guaranteed in real-world scenarios.

Correlation-based anomaly detection techniques aim at partitioning data into groups or clusters according to a chosen parameter(s) [31]. Once a cluster is defined, a centroid value (threshold value) would be calculated as representative of each cluster. Typically, clustering-based anomaly detection relies on comparing individual sensor-nodes observations with the centroid value of its cluster. It assumes that observations which belong to the same group or cluster are relatively similar at a particular point in time: this approach is justified by Tobler's law of geography, which states that "everything is related to everything else, but near things are more related than distant things" [32].

Clustering for anomaly detection 1) has no performance issues dealing with a high volume of data 2) does not need to analyse sensor-nodes time-series and 3) it can dynamically adapt to changes in the CPS network, like addition or removal of sensor-nodes [29].

Clustering-based anomaly detection is sensitive to observations with extreme values which may be caused by faults in sensor-nodes or due to external noise [6]. Moreover, calculating the value of clusters centroid value may involve a level of uncertainty because it requires determining both the optimum number of sensor-nodes in each cluster and the centroid value for each cluster in real-time [29][27]. Assuming that spatially related observations of sensor nodes are always correlated is not always valid. Moreover, the calculation of the centroid value (threshold value) of a cluster may be affected by the profile of a single sensor-node.

B. Predictive Analytics-Based Anomaly Detection

Predictive analytics is a branch of data science that involves the process of mining current and historical data to identify patterns and to forecast the future values of time-series [33], using statistical or machine learning methods [34]. Predictive analysis has been used in different industrial domains such as power management [35], transportations [36] and congestion and pollution control [37].

Predictive analysis employs historical data to train a prediction model for forecasting future values. A temporal anomaly detection models typically require more time to be

trained, which is a critical parameter, especially in real-time anomaly detection applications [6], [25].

Predictive analysis might provide an effective data quality assessment solution. However, this is not always possible, especially in large-scale CPS applications where hundreds or thousands of sensor-nodes stream data in real-time. Checking the time-series of each sensor-node in a large-scale CPS application before getting the next observation is a challenge, which needs high computational power and might not be practically possible. Typically, predictive analysis has a limited ability to adapt with the dynamically changing environment of large-scale CPS, because it requires to retrain the prediction system after any significant change in the CPS sensor-nodes network [29], [38], [39].

This research paper proposes HADES, a hybrid anomaly detection system which employs both correlation and predictive analysis as an anomaly detection technique:

1) Correlation-based anomaly detection: as a filtering layer; it summarises and reduces the number of sensor-nodes to be passed to the next unit.

2) Predictive-based anomaly detection: to perform temporal predictive analysis.

IV. THE HADES ARCHITECTURE

The HADES anomaly detection system consists of three layers, as shown in Fig. 1.

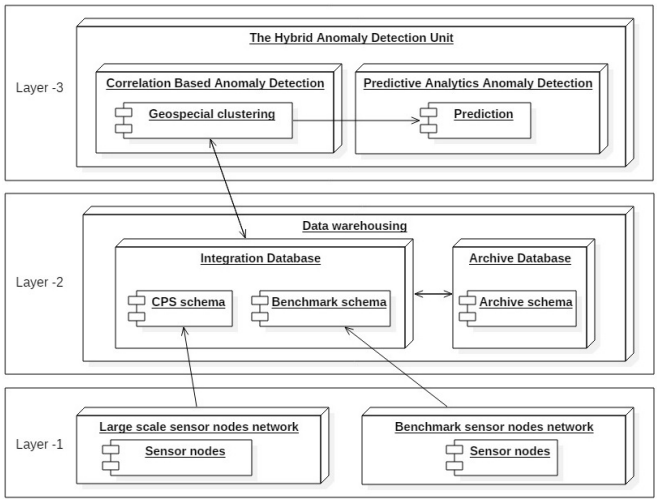

Figure 1 The UML component diagram of HADES.

HADES is able to collect sensor-nodes data streams effectively in real-time, and it collects data every T minutes. T is a dynamic parameter that may change based on the changes in the duty-cycles of the sensor nodes in the WSN, and it is smaller than the shortest duty-cycle of any sensor node in the sensor network. HADES can automatically react to changes in the sensor networks, which may lose some active sensor nodes or may have some sensor nodes added to it.

A. The Sensor-nodes layer (Layer-1)

Data related to physical processes or environmental phenomena such as temperature, humidity, air quality are collected and provided as input data to the HADES system. The data can be collected from industry, university (e.g. weather monitoring networks), or IT companies that collect observations from outsourced sensor-nodes networks.

At this level, data of different environmental parameters such as temperature, humidity, water levels and air quality were collected and provided as input data to the system. The first source is an outsourced CPS data from a large-scale sensor nodes network. The second source of data will serve as benchmark values for ensuring the quality of data of the first data source.

A benchmark data source is required to test and verify the accuracy and performance of the system before using it in real-world scenarios. The benchmark data source must be reliable and provide the same type of data parameters as the primary data source nearly at the same time and from the same geographical area.

B. The Data warehousing and integration layer (Layer-2)

HADES performs data integration processes at real-time, which, in this case, could be a challenge due to the relatively large amount of data collected in different formats and different structures. Data warehousing techniques were adopted to collected data and link them together or integrated them based on various parameters, such as observations timestamps, location and type. The data warehouse layer consists mainly of two databases (Fig. 1 Layer-2): the first database is the Integration Database, which hosts the data from two or more different CPS sensor-nodes networks.

The Archive Database has the same structure of the Integration Database but contains historical data that have been transferred automatically from the Integration Database Data which are not being used anymore for decision making will be transferred from the integration database to the archive database and will be deleted from the integration database sequentially; this approach will keep the decision-making data insulated while keeping a continues time series for all the collected data in the archive database. Fig. 2 shows the high-level sequence diagram of HADES data acquisition process.

C. The Hybrid Anomaly Detection Layer (Layer-3)

The Hybrid Anomaly Detection Layer consists of two main components: The Geospatial Clustering-Based Filtering unit and The Predictive Analytics-Based Anomaly Detection unit as shown in Fig.1 Layer-3.

1) The Geospatial Clustering-Based Filtering Unit

Observations from hundreds or thousands of sensor-nodes are divided into groups of observations which are spatially correlated to each other.

Machine learning-based clustering algorithms can be used for this purpose. The clustering process is based on a data snapshot in time from the data stream, which is the most recent reading from sensor-nodes since the last duty-cycle.

The number of the generated clusters is not relevant in this case, however, determining the optimum diameter of the clustering method automatically at real-time is a challenge especially considering that its value might vary based on the nature of the application and the minimum number of neighbour sensor-nodes (the density of the sensor-nodes in the targeted area) to form a cluster.

This component of the Layer-3 is a filtering mechanism: it compares the observation from a sensor-node with the cluster

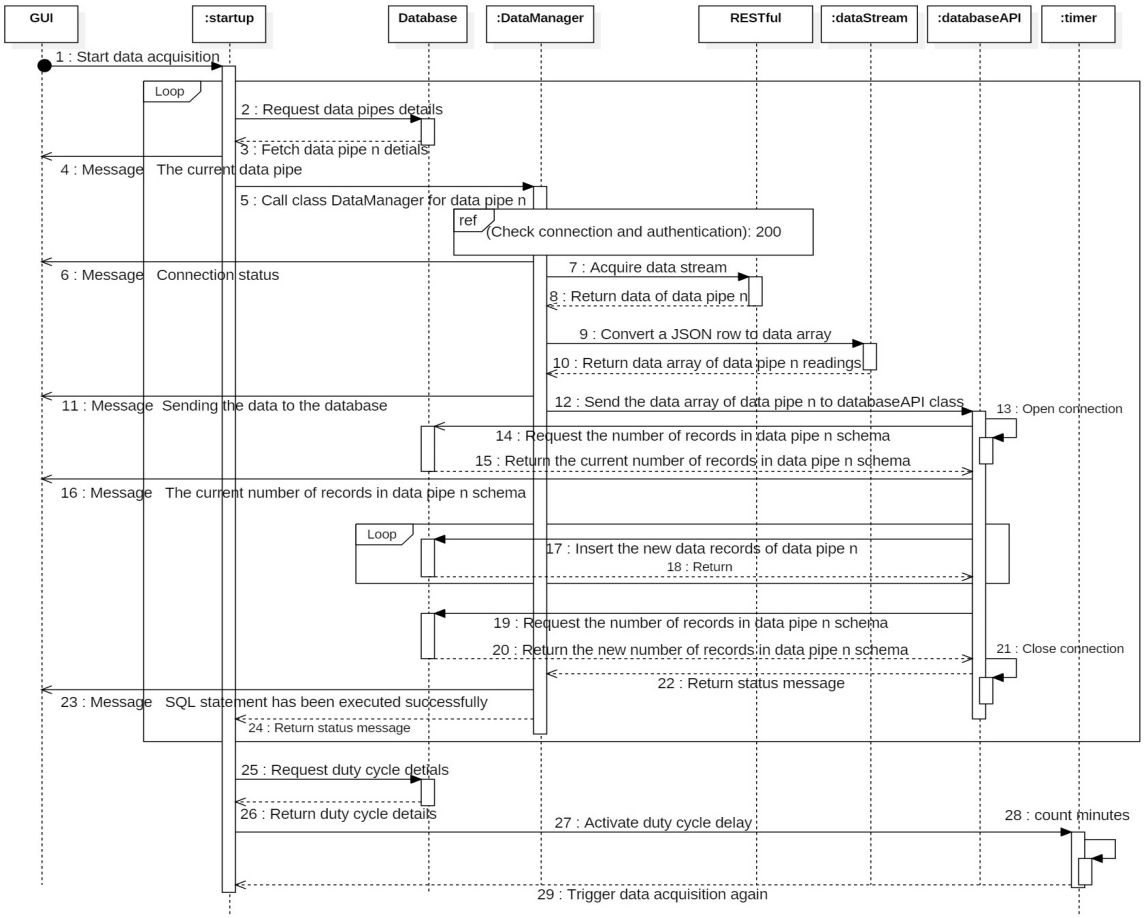

Figure 2 The high-level sequence diagram of HADES data acquisition process.

centroid value. Sensor-nodes that their observations have a significant deviation from the centroid value are considered to be potential outliers and must be verified using its time-series.

2) The Predictive Analytics-Based Anomaly Detection unit

The second component of the third layer of the proposed system examines the time-series of each sensor-node filtered by the Geospatial Clustering-Based Anomaly Detection unit and sensor-nodes that do not fit in a cluster.

This component is based on a predictive modelling approach. Statistical and machine learning-based predictive algorithms can be used to develop the predictive system. The time-series of the filtered sensor-nodes are used to train the predictive model [28]. The data which will be used to train the predictive model is theoretically considered to be an open-ended time-series because there is a relatively long sequence of observations. Using the whole time series may overfit the trained model and practically not possible. To tackle this issue, the sliding window modelling [28] will be used to select a part of the time-series only, the size of the sliding window will vary based on the application itself.

Time series decomposition technique will be applied, combined with statistical and machine learning technique such as the Seasonal Autoregressive Integrated Moving Average, (SARIMA) as shown in Fig. 7. Artificial Neural Network (ANN) [17] or the Gaussian Process Regression (GPR) [40] will be adjusted regularly based on the continuous evaluation of the prediction accuracy. The high-level activity diagram of Layer-3 is shown in Fig. 3.

V. LONDON CASE STUDY

This section is to highlight some of the preliminary results of applying data analysis to validate the functionality of the HADES hybrid anomaly detection system.

Two different data sources were used: the first is a benchmark sensor-nodes network consists of four high-quality wireless temperature sensor-nodes and a wireless Gateway

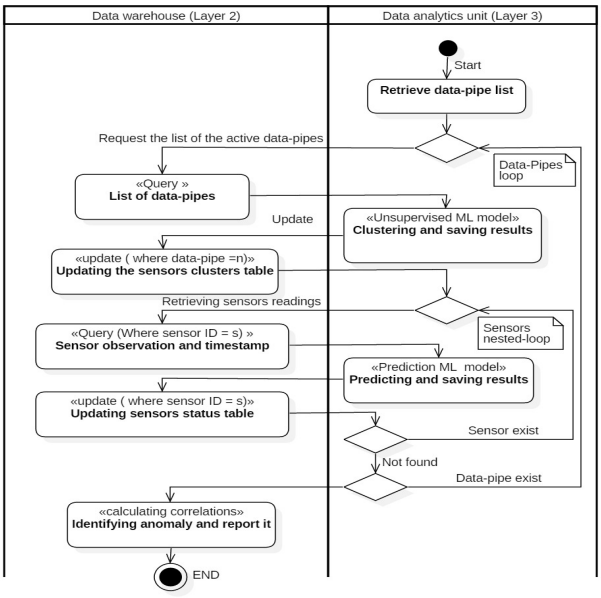

Figure 3. The UML sequence diagram of HADES data acquisition process.

deployed at the University of East London. The second data source is a large-scale network of more than 400 sensor-nodes distributed around London (see Fig.4), accessing through the Thingful API[1].

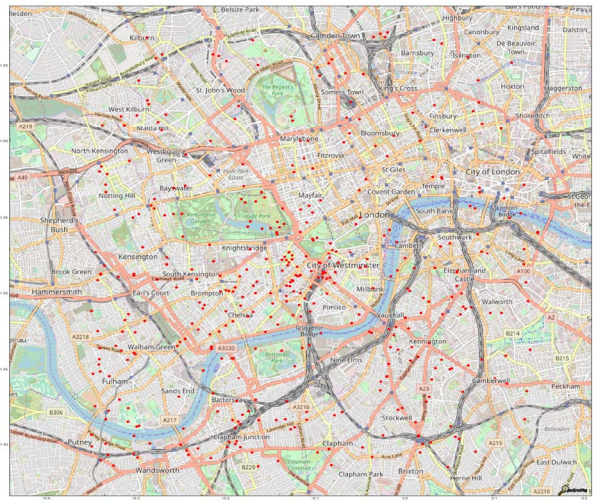

Figure 4 The geographic distribution of temperature sensor nodes (London)

HADES has been deployed on three Dell Workstations with Linux RHEL7 / Fedora 28: one for running the data acquisition process (layer-1), the second as a warehousing database server and the third for HADES anomaly detection processes (layer-3).

HADES has been fetching data since May 2018 from more than 400 sensors every 10 minutes, from 4 data pipes, collecting more than 40 million observations in its database. Each entry in the database is one observation from 5 different data pipes: Environment, Energy, Bikes, Noise, and Air Quality. Each data pipe contains common data (like latitude and longitude of the sensor, and the timestamp of the observation) and a set of observations from different types of sensors.

For example, the Environment data pipe contains: water level, water flow rate, and rainfall; while the Weather data pipe contains: sensor voltage, latitude and longitude, humidity, time-stamp, temperature, Noise level, WiFi networks count, atmospheric pressure, MSLP (Mean Sea Level Pressure), Nitrogen Dioxide (NO_2) and Carbon Monoxide (CO) chemical concentrations.

Many clustering algorithms were tested: K-Means[41], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [41] and Mean-shift [41]. All the tested clustering algorithms were able to identify centroid points of each cluster successfully as showing in Fig. 5.

Our preliminary tests showed that in real-world scenarios, sensor-nodes are not geographically uniformly distributed and the ones that are allocated in a low-density area might not fit in any cluster (Fig. 6, indicated by red arrows), while sensor-nodes in the high-density area are clustered successfully.

Thus, Geospatial Clustering-Based approach may produce less accurate results for remote sensor-nodes with no neighbour sensor-nodes around them. And support the approach of using correlation-based anomaly detection as filtering mechanism rather than a real-world anomaly detection technique.

Figure 5 Geospatial clustering of temperature sensor-nodes in London: the white dots are clusters centroids

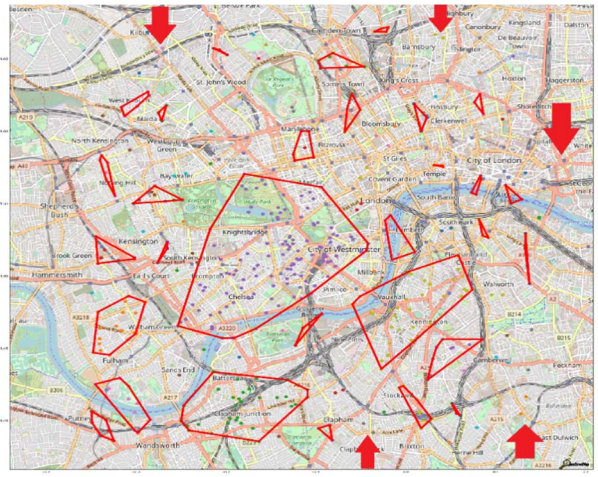

Figure 6 The density of the temperature sensor nodes in and around London (23-Jan-2020).

Time series decomposition was used to analyse the time-series of the filtered sensor-nodes by the correlation-based (clustering) algorithms. The time series decomposition showed that temperature has a trend which gradually increases or decreases over days of slow-changing and it has a clear daily seasonality where the temperature rises to the maximum during the mid-day and reduce to its minimum at the very early morning. Thus, predictive analysis can be used to forecast future observations and estimate its values to evaluate the quality of data. Holt-Winters Seasonal, ARMA, Non-Seasonal ARIMA and Seasonal ARIMA Models were tested as statistical-based forecasting methods. ARMA, Non-Seasonal ARIMA and Seasonal ARIMA were able to produce high accuracy predictions using the one-step-ahead forecasting method as shown in Fig. 7. However, none of the ARMA based forecasting models was able to provide reliable predictions for extended intervals as the forecast confidence

[1] Thingful Ltd, a search engine for the internet of things. On-line at https://www.thingful.net/

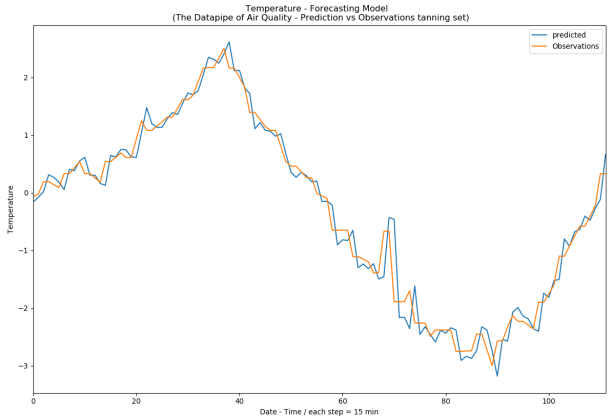

Figure 7 ARIMA one-step-ahead forecasting method.

interval overgrows with time. More tests will be conducted in the context of this research to cover more analytics prediction methods using statistical and machine-learning techniques.

VI. CONCLUSIONS AND FUTURE WORKS

The challenge of data quality assessment becomes greater in large-scale CPS applications which typically deals with large volumes of data in real-time. This research paper proposes HADES, a hybrid anomaly detection system based on combining both correlation-based and predictive-based anomaly detection techniques. The correlation-based anomaly detection acts as a filtering mechanism that identifies sensor-nodes with potential data quality issues. The predictive-based anomaly detection unit performs temporal predictive analysis on the time-series of sensor-nodes identified by the correlation-based unit or sensor-nodes which do not fit into a cluster.

With almost two years of sensors data collected, this approach could be used to develop more efficient data quality assessment systems based on time-series analysis. The next step is to validate the proposed system using machine learning and combining different prediction techniques to detect anomalies earlier.

ACKNOWLEDGEMENTS

This research was supported by the University of East London (UEL) PhD scholarship scheme 2017 and through the UEL Research Internship programme 2019 in collaboration with Thingful Ltd, UK, which gave us access to their sensors data across London; in particular we thank Usman Haque for his insight and expertise that greatly assisted the research. We also thank Professor Allan Brimicombe for assistance with time analysis techniques that helped us improving the work.

NOTE

In Greek and Roman mythology, Hades was the God of the Underworld: his name means the "The Unseen"; similarly, HADES aims at overseeing what is normally unseen, like sensors.

REFERENCES

[1] P. Marwedel, *Embedded system design embedded systems, foundations of cyber-physical systems, and the internet of things*. Springer International Publishing, 2018.

[2] A. A. Letichevsky, O. O. Letychevskyi, V. G. Skobelev, and V. A. Volkov, "Cyber-Physical Systems," *Cybern. Syst. Anal.*, vol. 53, no. 6, pp. 821–834, 2017, doi: 10.1007/s10559-017-9984-9.

[3] S. Patnaik, *New Paradigm of Industry 4.0 Internet of Things, Big Data & Cyber Physical Systems*. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer Nature Switzerland AG, 2019.

[4] Wachs D, "Industry 4.0, IoT in Digital Manufacturing, creating smart factories | BearingPoint," *www.bearingpoint.com*, 2019. .

[5] S. Jianjun, W. Xu, G. Jizhen, and C. Yangzhou, "The Analysis of Traffic Control Cyber-physical Systems," *Procedia - Soc. Behav. Sci.*, vol. 96, pp. 2487–2496, Nov. 2013, doi: 10.1016/J.SBSPRO.2013.08.278.

[6] L. A. Tang, J. Han, and G. Jiang, "Mining sensor data in cyber-physical systems," *Tsinghua Sci. Technol.*, vol. 19, no. 3, pp. 225–234, 2014, doi: 10.1109/TST.2014.6838193.

[7] D. B. Rawat, J. Rodrigues, and I. Stojmenović, *Cyber-physical systems from theory to practice*. Broken Sound Parkway NW: CRC Press Taylor & Francis Group, LLC, 2016.

[8] W. Kang, K. Kapitanova, and S. H. Son, "RDDS: A real-time data distribution service for cyber-physical systems," *IEEE Trans. Ind. Informatics*, vol. 8, no. 2, pp. 393–405, May 2012, doi: 10.1109/TII.2012.2183878.

[9] P. Barnaghi, M. Bermudez-Edo, and R. Tönjes, "Challenges for Quality of Data in Smart Cities," *J. Data Inf. Qual.*, vol. 6, no. 2–3, pp. 1–4, Jun. 2015, doi: 10.1145/2747881.

[10] Y. Zhang, W. Duan, and F. Wang, "Architecture and real-time characteristics analysis of the cyber-physical system," *2011 IEEE 3rd Int. Conf. Commun. Softw. Networks, ICCSN 2011*, pp. 317–320, May 2011, doi: 10.1109/ICCSN.2011.6013602.

[11] A. Ordonez, V. Alcázar, J. C. Corrales, and P. Falcarin, "Automated context aware composition of Advanced Telecom Services for environmental early warnings," *Expert Syst. Appl.*, vol. 41, no. 13, 2014, doi: 10.1016/j.eswa.2014.03.045.

[12] F. Hu and Q. Hao, *Intelligent Sensor Networks*. 2012.

[13] Q. Huang, C. Lu, and K. Chen, *Big Data Analytics for Sensor-Network Collected Intelligence*. 2017.

[14] C.-S. Shih, J.-J. Chou, N. Reijers, and T.-W. Kuo, "Designing CPS/IoT applications for smart buildings and cities," *IET Cyber-Physical Syst. Theory Appl.*, vol. 1, no. 1, pp. 3–12, Dec. 2016, doi: 10.1049/iet-cps.2016.0025.

[15] W. Grega and A. J. Kornecki, "Real-Time Cyber-Physical Systems: Transatlantic Engineering Curricula Framework," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2015, pp. 755–762, doi: 10.15439/2015F45.

[16] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Sci. J.*, vol. 14, no. 0, p. 2, 2015, doi: 10.5334/dsj-2015-002.

[17] A. Brimicombe, *GIS, environmental modeling and engineering*. CRC Press, 2010.

[18] P. Falcarin, M. Valla, J. Yu, C. A. Licciardi, C. Frà, and L. Lamorte, "Context data management: an architectural framework for context-aware services," *Serv. Oriented Comput. Appl.*, pp. 1–33, Jul. 2012, doi: 10 1007/s11761-012-0115-1.

[19] S. Sadiq, *Handbook of Data Quality*. 2013.

[20] Y. Guo, X. Hu, B. Hu, J. Cheng, M. Zhou, and R. Y. K. Kwok, "Mobile Cyber Physical Systems: Current Challenges and Future Networking Applications," *IEEE Access*, vol. 6, pp. 12360–12368, 2018, doi: 10.1109/ACCESS 2017.2782881.

[21] C. Baladron *et al.*, "Integrating User-Generated Content and Pervasive Communications," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 58–61, Oct. 2008, doi: 10.1109/MPRV.2008.76.

[22] H. Song, D. B. Rawat, S. Jeschke, and C. Brecher, *Cyber-Physical Systems Foundations, Principles and Applications*. 2016.

[23] H. I. Kobo, A. M. Abu-Mahfouz, and G. P. Hancke, "A Survey on Software-Defined Wireless Sensor Networks: Challenges and Design Requirements," *IEEE Access*, vol. 5, pp. 1872–1899, 2017, doi: 10.1109/ACCESS.2017.2666200.

[24] S. Zeadall and N. Jabeur, *Cyber-Physical system design with sensor networking technologies*, vol. 53, no. 9. London: Institution of Engineering and Technology, 2016.

[25] S. Chen, P. Sinha, N. B. Shroff, and C. Joo, *Rechargeable Sensor Networks*, vol. 22, no. 4. 2014.

[26] T. Addabbo, A. Fort, M. Mugnaini, E. Panzardi, A. Pozzebon, and V. Vignoli, "A city-scale IoT architecture for monumental structures monitoring," *Meas. J. Int. Meas. Confed.*, vol. 131, pp. 349–357, Jan. 2019, doi: 10.1016/j.measurement.2018.08.058.

[27] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: a survey," *Digit. Commun. Networks*, vol. 4, no. 3, pp. 161–175, Aug. 2018, doi: 10.1016/j.dcan.2017.10.002.

[28] A. Appice, A. Ciampi, F. Fumarola, and D. Malerba, *Data Mining Techniques in Sensor Networks*. 2014.

[29] L.-J. Chen, Y.-H. Ho, H.-H. Hsieh, S.-T. Huang, H.-C. Lee, and S. Mahajan, "ADF: An Anomaly Detection Framework for Large-Scale PM2.5 Sensing Systems," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 559–570, Apr. 2018, doi: 10.1109/JIOT.2017.2766085.

[30] P. M. Laso, D. Brosset, and J. Puentes, "Analysis of quality measurements to categorize anomalies in sensor systems," in *2017 Computing Conference*, 2017, pp. 1330–1338, doi: 10.1109/SAI.2017.8252263.

[31] J. Huang, Z. L. Yu, and Z. Gu, "A clustering method based on extreme learning machine," *Neurocomputing*, vol. 277, pp. 108–119, Feb. 2018, doi: 10.1016/J.NEUCOM.2017.02.100.

[32] H. J. Miller, "Tobler's First Law and Spatial Analysis," *Ann. Assoc. Am. Geogr.*, vol. 94, no. 2, pp. 284–289, Jun. 2004, doi: 10.1111/j.1467-8306.2004.09402005.x.

[33] A. Kumar, *Learning predictive analytics with python gain practical insights into predictive modelling by implementing Predictive Analytics algorithms on public datasets with Python*. Packt Publishing, 2016.

[34] B. Ratner, *Statistical and Machine-Learning Data Mining*, Third Edit. 2017.

[35] G. Sideratos, A. Ikonomopoulos, and N. Hatziargyriou, "A Committee of Machine Learning Techniques for Load Forecasting in a Smart Grid Environment," *Int. J. Energy Power*, vol. 4, no. 0, p. 98, 2015, doi: 10.14355/ijep.2015.04.016.

[36] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 1352–1372, Oct. 2015, doi: 10.1016/J.RSER.2015.04.065.

[37] C. Goves, R. North, R. Johnston, and G. Fletcher, "Short Term Traffic Prediction on the UK Motorway Network Using Neural Networks," *Transp. Res. Procedia*, vol. 13, pp. 184–195, Jan. 2016, doi: 10.1016/J.TRPRO.2016.05.019.

[38] A. Akbar, A. Khan, F. Carrez, and K. Moessner, "Predictive analytics for complex IoT data streams," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1571–1582, 2017, doi: 10.1109/JIOT.2017.2712672.

[39] J. Murphree, "Machine learning anomaly detection in large systems," *AUTOTESTCON (Proceedings)*, vol. 2016-Octob, pp. 1–9, 2016, doi: 10.1109/AUTEST.2016.7589589.

[40] C. Williams and C. E. Rasmussen, "Gaussian processes for regression," *Adv. Neural Inf. Process. Syst. 8*, vol. 8, no. August, pp. 514–520, 1996, doi: 10.1145/1102351.1102369.

[41] P. P. Angelov and X. Gu, *Empirical Approach to Machine Learning*, vol. 800. Cham: Springer International Publishing, 2019.