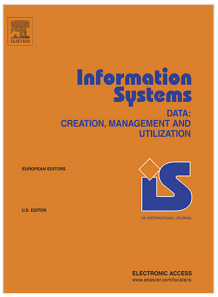
Journal Pre-proof



Data quality challenges in large-scale cyber-physical systems: A systematic review

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

PII: S0306-4379(21)00148-4
DOI: https://doi.org/10.1016/j.is.2021.101951
Reference: IS 101951

To appear in: *Information Systems*

Received date : 18 March 2021
Revised date : 4 October 2021
Accepted date : 11 November 2021

Please cite this article as: A.A. Alwan, M.A. Ciupala, A.J. Brimicombe et al., Data quality challenges in large-scale cyber-physical systems: A systematic review, *Information Systems* (2021), doi: https://doi.org/10.1016/j.is.2021.101951.

Data Quality Challenges in Large-Scale Cyber-Physical Systems: A Systematic Review

Ahmed Abdulhasan Alwan[a], Mihaela Anca Ciupala[a], Allan J. Brimicombe[a], Seyed Ali Ghorashi[a], Andres Baravalle[b], Paolo Falcarin[a,*]

[a] *University of East London, ACE School of Architecture Computing and Engineering, Docklands Campus, University Way, London, United Kingdom, E16 2RD.*
[b] *Akamai Technologies Ltd.,London, United Kingdom.*

Abstract

Cyber-physical systems (CPSs) are integrated systems engineered to combine computational control algorithms and physical components such as sensors and actuators, effectively using an embedded communication core. Smart cities can be viewed as large-scale, heterogeneous CPSs that utilise technologies like the Internet of Things (IoT), surveillance, social media, and others to make informed decisions and drive the innovations of automation in urban areas. Such systems incorporate multiple layers and complex structure of hardware, software, analytical algorithms, business knowledge and communication networks, and operate under noisy and dynamic conditions. Thus, large-scale CPSs are vulnerable to enormous technical and operational challenges that may compromise the quality of data of their applications and accordingly reduce the quality of their services. This paper presents a systematic literature review to investigate data quality challenges in smart-cities large-scale CPSs and to identify the most common techniques used to address these challenges. This systematic literature review showed that significant work had been conducted to address data quality management challenges in smart cities, large-scale CPS applications. However, still, more is required to provide a practical, comprehensive data quality management solution to detect errors in sensor nodes' measurements associated with the main data quality dimensions of accuracy, timeliness, completeness, and consistency. No systematic or generic approach was demonstrated for detecting sensor nodes and sensor node networks failures in large-scale CPS applications. Moreover, further research is required to address the challenges of ensuring the quality of the spatial and temporal contextual attributes of sensor nodes' observations.

Keywords: , Cyber-physical systems (CPS), Wireless Sensor networks(WSN), Data quality management, Data quality dimensions, Smart cities, Quality of observations

'Declarations of interest: none.

*Corresponding author
Email addresses: a.alwan@uel.ac.uk (Ahmed Abdulhasan Alwan), m.a.ciupala@uel.ac.uk (Mihaela Anca Ciupala), a.j.brimicombe@uel.ac.uk (Allan J. Brimicombe), s.a.ghorashi@uel.ac.uk (Seyed Ali Ghorashi), a.baravalle@akamai.com (Andres Baravalle), falcarin@uel.ac.uk (Paolo Falcarin)

1. Introduction

Cyber-physical systems (CPSs) are designed as a network of computational elements that combine physical input and output mechanisms to interact with the surrounding environment [1]. CPSs can be seen as the new generation of engineering systems with high computation and communication capabilities that perform dedicated functions, typically, according to strict real-time constraints [2] [3]. Data typically circulate continuously among the different CPSs components in real-time [4]. CPSs rely on data acquisition from sensor nodes, data processing in the control (computing) unit(s) and data communication with the actuators to regulate the physical environment. This data cycle is necessary for the CPSs to meet their operational requirement and ultimately enables the system's self-control and awareness, especially in real-time applications [5], [6], [7]. Therefore, data has a crucial role in the successful operation of CPSs [5], especially considering that CPSs may cause severe consequences in the case of providing decisions based on low-quality data [8], [9], [10].

CPSs might compromise safety constraints and might have life-threatening consequences in cases of receiving incorrect data, missing time deadlines or missing critical readings from sensors in real-time [11]. Ensuring the quality of data is an open challenge in large-scale CPSs applications [12], [13], [14], [15], [16], mainly because of the large amount of data that these systems exchange at (near) real-time, the vast geographical area, and the dynamic and noisy conditions where these systems are usually deployed [17].

In response, many studies have been published to address existing data-quality challenges in large-scale, smart-cities' CPS applications. Some examples are as listed in the second section of Table 1, where most of these studies focused on the methods or algorithms used to address a domain-specific data quality challenge. Despite the importance of these efforts, there is still a lack of review studies that analyse data quality challenges in large-scale CPSs in a more multivariate perspective.

Table 1: Cyber-Physical Systems Cross-Domain Applications in the Context of Smart Cities.

CPS applications/systems	Smart Environment	Smart Transportation	Smart Healthcare	Human activity/ Smart spaces	Smart Governance
Smart utility management	[18]	[19]	[20]	[21]	[22]
Traffic and road management	[23]	[24]	/	[25]	[23]
Sensors and sensing technology	[26]	[27]	[28]	[29]	[30]
Energy management	[31]	[32]	/	[33]	[34]
Common challenges of large-scale CPS in smart cities					
Big Data management	[35]	[36]	[37]	[38]	[39]
Data quality management	[40]	[41]	[42]	[43]	[44]

This paper offers a review of data quality challenges in large-scale, smart cities' CPSs and the most popular data quality assessment/management methods or techniques adopted to address these challenges. Furthermore, this review analyses the existing literature to draw meaningful conclusions related to data quality challenges in large-scale CPSs to reveal existing research gaps in the field of data quality management and point towards potential research directions.

The rest of this paper is organised as follows: Section 2 is an introduction to smart cities as large-scale CPSs, Section 4 provides details of the review process and methodology, Section 5, describes the review

conduct and primary studies selection processes. Section 6 specifies the leading data quality challenges in large-scale CPS applications, Section 7 presents data mining and data quality management in large-scale CPSs, and Section 8 presents the unaddressed data quality management challenges in large-scale CPSs. Finally, Section 9 presents the concluding remarks.

2. Smart Cities as Large-Scale CPSs

CPSs are the next generation information systems that integrate communication, computation, and control to achieve higher performing buildings and better public services with more energy-efficient operations and a higher level of automation [45]. CPSs are an active area of research [46], [47], with a significant importance to the future of smart cities [48]. CPSs are multidisciplinary cross-domain information systems which bring together different sectors of smart cities' public services, such as smart transportation management, smart utility management, smart buildings [49], smart environment management [50] and smart governance, Where data sensing, knowledge extraction, and higher automation are critical elements in the future of these services [51]. The future cities can provide smarter services by utilising IoT solutions that relay on an extended number of sensors and can provide scalable and interactive functionalities, e.g., the city scale traffic management solution illustrated in Figure 1, [52]. The smart city traffic management solution detects traffic

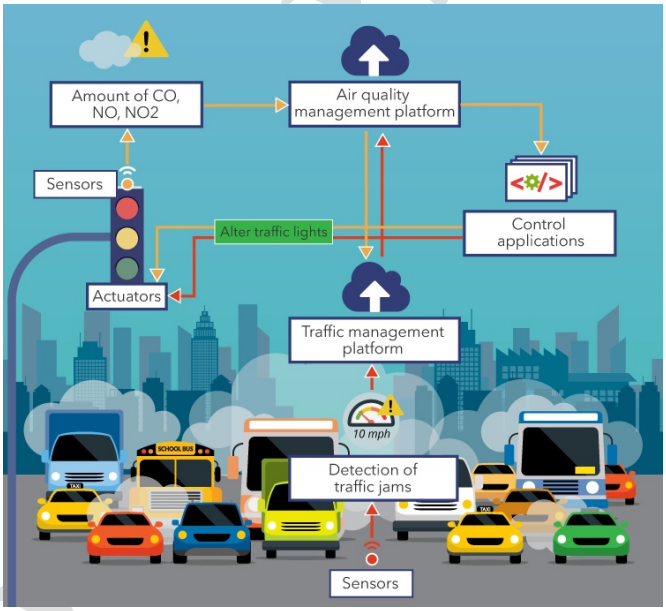


Figure 1: A smart city large-scale IoT solution for traffic management [52].

jam in real-time. It controls the traffic while monitoring the air quality to ensure that the traffic jam does not affect the environment. Such a system relies on air quality monitoring sensors to measure the ratio of harmful gases in the air and utilises traffic lights as actuators to reduce the traffic in low air quality areas [52].

Most of these large-scale CPSs are heterogeneous and multi-model information systems that analyse massive amount of data collected from various devices provided by different manufacturers [17]. Typically, smart cities' large-scale CPS applications are designed to sense, process and react to real-time changes [53]. These systems rely on hundreds of sensor nodes and other devices which continuously sense and stream readings of various parameters rendering large volumes of data, which is known as Big Data [54]. The term Big Data describes a massive volume of complex and different types of structured and unstructured data that accumulate in a relatively high velocity [55]. Mining and analysing big data has a significant role in providing a rich source of information about smart cities' utilities and citizens' activities, providing more efficient management, better services and sustainable development [56].

Ensuring the quality of data in large-scale CPS applications is a critical requirement to guarantee that their analytical core will make more reliable decisions [57], [58]. The quality of data of CPS applications is mainly affected by inaccurate observations that do not represent the actual value of measured phenomena [59]. Data quality issues may occur in large-scale CPS applications because of many reasons such as sensor nodes malfunctions [60], calibration issues, poor sensor nodes quality, environmental effects, external noise [61], networks or communication errors, and real-time scheduling problems [8], [58]. Furthermore, limitations in communication channels may cause observations' overlook in sensor networks which usually occurs during data transmission or aggregation processes [17], [62].

The challenge of data quality becomes greater in large-scale CPS applications, e.g. in environmental and noise monitoring systems, which rely on various sensors and other devices connected by extended networks and usually operate under noisy and dynamic conditions [44], [63], [60]. Such applications have enormous technical challenges because of their multiple layers and complex structure that combines hardware, software, analytical algorithms, business knowledge and communication infrastructure [57]. CPSs implementations in different sectors of smart cities, public services are listed in Table 1, which also highlights big data and data quality management as common challenges across all of these large-scale CPSs applications.

3. Data Quality Concepts and Terminology

According to the International Standardization Organisation (ISO), quality, in general, is the "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs" [64]. In comparison, Data Quality is defined as data fitness for the purpose of the intended use [65, 66] or its conformance to requirements [67, 68]. This definition outlines that data which very well meet some predefined expectations, specifications, or standards are considered to be high-quality data that fit for use in a particular application. The concept of fitness to use associated with data quality also covers how effectively the data describe any events, observations or measurements it was created to represent the characteristics of the data that circulating in the system [69].

Data quality can be quantified, measured and monitored using a set of context-dependent parameters or indicators known as Data Quality Characteristics or Dimensions [70, 69]. More than 200 data quality

dimensions have been introduced since the eighties [71]. However, these dimensions can be categorised into four core data quality dimensions: accuracy, completeness, timeliness and consistency [69, 71, 72, 73], which are ,typically, associated with data quality requirements and mapped to define data quality assessment criteria [74].

4. Review Process and Methodology

This systematic literature review (SLR) was conducted based on the guidelines proposed by [75], which provides an organised and repeatable procedure to perform the SLR based on three primary stages: planning, conducting, and reporting the review results. The overall review of the processes adopted to conduct this SLR are illustrated in Figure 2.

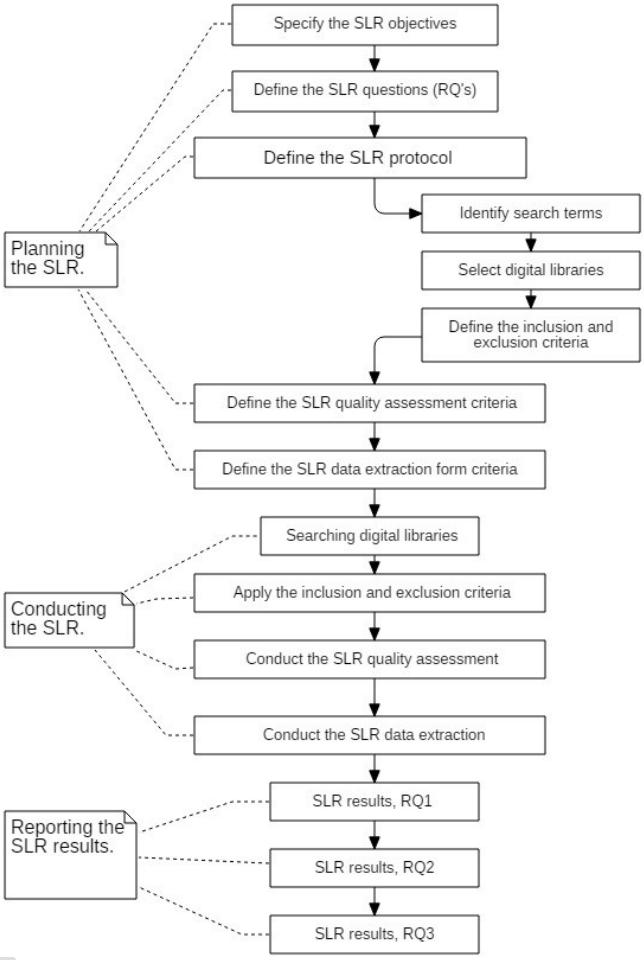


Figure 2: A holistic overview of the processes adopted in this systematic literature review.

4.1. SLR Questions and Objectives

The SLR review questions (RQ) have a significant role in driving the review methodology and identifying the primary studies. Thus, the analysis and synthesis process of the primary studies must extract the data in a way that answers the review questions [75]. The SLR review questions are listed in Table 2.

4.2. SLR Protocol (Strategy)

The Review protocol is the strategy of implementing a set of specified steps to undertake the SLR. The purpose of the SLR protocol is to narrow down the possibility of researcher bias by pre-defining the review processes and procedures of selecting and analysing the primary studies that will address the research questions. The review protocol involves specifying the research terms (keywords), digital libraries, refinement terms (synonyms for the main search terms), the quality questionnaire and the data extraction forms [75], [76], as follows:

Table 2: SLR Review Questions and Objectives.

RQ#	SLR Research Question / Objectives
RQ1	What are the most common data quality challenges associated with large-scale CPS applications?
RQ2	Which solutions/methods were adopted to address data quality challenges in large-scale CPS?
RQ3	What is the overall effectiveness of the current solutions/methods used to address data quality challenges in large-scale CPS?

4.2.1. Identifying Search Terms

Digital libraries must be searched using search terms and keywords to identify the primary studied that will address the review questions. The search terms typically extracted from the search questions, including any possible alternative terms or synonyms as shown in Table 3. The search method is based on incorporating the keywords and terms from Table 3 using Boolean expressions (OR, AND, NOT...etc) to form Boolean search string, which used to search the pre-selected digital libraries.

4.2.2. Selecting Digital Libraries

Selecting digital libraries is an essential step for identifying relevant primary studies that will address the research questions. It is critical to include many digital libraries in the search process since no single source

Table 3: The SLR Search Terms and Keywords.

Category	Search Terms	Level of Abstraction
Primary	Quality of data	Title and abstract
	Data quality	
	Quality of information	
	Information quality	
Secondary	Cyber-physical system	
	Internet of things	
	Wireless sensors network	
Exclusion	Social Media	Fully reviewing the study
	Smart-Wearables	
	Vehicle Services	
	Social Sensing	

can comprehensively provide all relevant primary studies and to ensure resource-dependent search to cover the search topic. Table 4 shows the list of specialised digital libraries selected as the literature source for identifying relevant primary studies to the research topic.

4.2.3. Defining Inclusion and Exclusion Criteria

The purpose of the inclusion and exclusion section is to define the criteria of selecting which primary studies will be approved for further analysis, while excluding other studies that do not satisfy these criteria. The inclusion and exclusion criteria of this SLR are listed in Table 5.

4.3. SLR Quality Assessment

The purpose of the SLR quality assessment is to evaluate the relevance of primary studies that already met the inclusion criteria to the review topic. SLR quality assessment is crucial because it is a further measure to limit the possibility of researcher bias [77], it presents a repeatable guideline for interpreting the results, and it provides a quantitative numeric mean to determine how strongly the selected primary studies are associated with the SLR objectives via a quality score. Typically, the SLR quality assessment can be implemented by scoring individual primary studies using a quality questionnaire form and based on assessment criteria [76].

Table 4: The list of digital libraries used for identifying the SLR primary studies.

ID	Digital Library	Online Search Interface
1	IEEE Xplore	https://ieeexplore.ieee.org
2	ACM Digital Library	https://dl.acm.org
3	IET Digital Library	https://digital-library.theiet.org
4	Science Direct	https://www.sciencedirect.com

Table 5: The SLR Inclusion and Exclusion Criteria.

	Inclusion criteria
-	The study is categorised as a peer-reviewed journal and conference paper relevant to the SLR topic and addresses one or more of its review questions.
-	The study is relevant to large-scale CPS or IoT applications.
	Exclusion criteria
-	The study is an editorial, tutorial, magazine, book, course, poster or it is not a peer-reviewed journal.
-	The focus of the study is mobile CPS or IoT.
-	The study is written in a different language other than English.
-	The full version of the study is not available.
-	The study is published before 2014.
-	Duplicated studies.

The SLR primary studies were scored based on the quality assessment questions listed in Table 6.

4.4. SLR Data Extraction Form

The data extraction form summarises and extracts information from the primary studies to answer the review questions. It specifies which primary study addresses which of the SLR review questions, analyses the

results and identifies the primary study strengths and weaknesses. The structure of the data extraction form used in this SLR is as follows:

- References details.
- Study purpose/application.
- Dataset types/details.
- Targeted data quality dimensions.
- Addressed data quality challenges.
- Proposed solutions/methods.

Table 6: The SLR Quality Assessment Questions (Matrix).

Q#	SLR Quality Assessment Questions	Q. score		
		Y	P	N
Q1	A review or an empirical study?	1	n/a	0.5
Q2	Does the study combine multi-methods techniques to address data quality challenges?	1	0.5	n/a
Q3	Does the study justify the use of these different methods/techniques?	1	n/a	0
Q4	Is there any comparative analysis of the different used methods/techniques?	1	n/a	0
Q5	How many data quality issues associated with the four core data quality dimensions are the study addressing?	4	3 - 1	n/a

5. Review Conduct and Primary Studies Selection

The SLR review was conducted using the pre-defined structure highlighted in the review process and methodology section and based on the three key steps: selecting, evaluating, and summarising the primary studies as follows:

5.1. Searching Digital Libraries

The first step to implement the SLR processes and methodology was to identify relevant primary studies by searching the digital libraries listed in Table 4 using search strings developed based on the keywords and terms specified in Table 3 as shown in Table 7[1].

5.2. Applying the Inclusion and Exclusion Criteria

The next step is to determine whether the identified primary studies satisfy all of the pre-defined inclusion and exclusion criteria, listed in Table 5. The number of the primary studies included in the SLR after applying the inclusion and exclusion criteria is shown in Table 8.

[1] There are some slight differences among the Boolean search strings used to search different digital libraries, these differences are related to the design of the search interface of the digital libraries and the availability of the primary studies.

Table 7: The SLR digital libraries, final search strings and the number of identified primary studies.

ID	Digital Library	Action	Boolean Search Strings (10/10/2020)	No. of Papers
1	IEEE Xplore	Search string	(("Document Title":"cyber physical system" OR Internet of things" OR "wireless sensors network") AND "Document Title":"data quality" OR "quality of data" OR "quality of information")	376
		Filter	Publication Type (Conferences, Journals), Publication Topics (learning (artificial intelligence) Internet of Things data analysis data mining wireless sensor networks Big Data decision making pattern classification data handling optimisation pattern clustering information systems quality of service statistical analysis) Published between (2014 and 2020)	
2	ACM	Search string	[Publication Title: "data quality"] OR [Publication Title: "quality of data"] OR [Publication Title: "quality of information"] AND [Publication Date: (01/01/2014 TO 12/31/2020)]	91
		Filter	Published between (2014 and 2020)	
3	IET	Search string	("data quality" OR "quality of data" OR "quality of information") AND ("cyber physical system" OR Internet of things")	52
		Filter	Published between (2014 and 2020)	
4	Science Direct	Search string	Articles with these terms ("cyber physical system" OR Internet of things" OR "wireless sensors network") and Title ("data quality" OR "quality of data" OR "quality of information")	23
		Filter	Review articles, Research articles, published between (2014 and 2020)	

Table 8: The final number of primary studies included in the SLR after applying the inclusion and exclusion criteria and fully reviewing all studies.

Activity \ Digital Library	IEEE	ACM	IET	Science Direct	Total
Searching digital libraries and applying filters	376	91	52	23	542
Reviewing titles and abstracts	78	26	6	8	118
Fully reviewing all studies	40	13	4	3	**60**

5.3. Conducting the SLR Quality Assessment

The quality assessment (as highlighted in Section 4.3) is a crucial step to evaluate the relevance of primary studies and scoring them according to the assessment matrix specified in Table 6 where the relationship among the quality assessment questions is shown in Equation 1.

$$Quality\ Score\ =\ Q1 + Q2 + Q3 + Q4 + Q5 \tag{1}$$

The primary studies referencing details and their overall quality assessment score are listed in Appendix Appendix A, Table A.9. Although this approach does not answer the review questions, it provides an opportunity to find the trend of most recent studies, evaluating their impact[2] and the geographical distribution of interest in data quality management of large-scale CPSs . The final number of the primary studies included in the SLR after applying the inclusion and exclusion criteria and after fully reviewing all studies is detailed in Table 8. Figure 3 shows the trend of the number of SLR primary studies associated with data quality challenges in large-scale CPSs by the year of publication.

Figure 4 shows the number of SLR studies associated with data quality challenges in large-scale CPSs by the country of publication.

[2] Using the quality score as a quantitative reference.

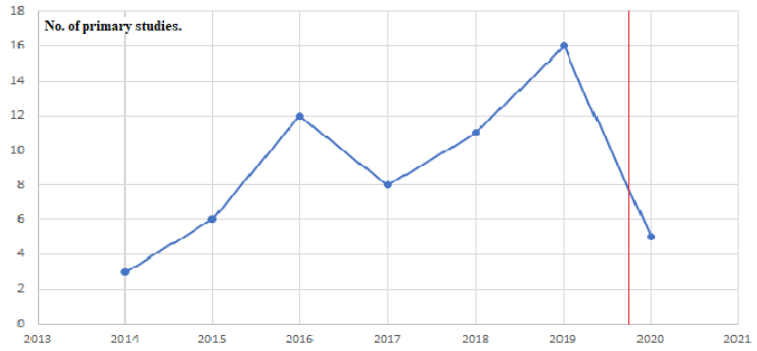


Figure 3: The number of SLR primary studies associated with data quality challenges in large-scale CPSs by the year of publication, (October 2020).

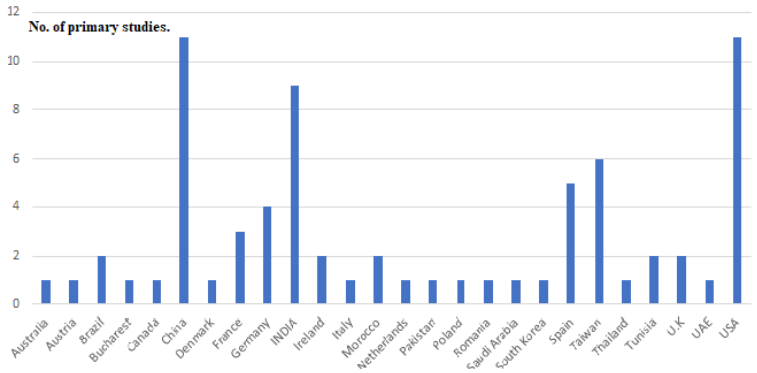


Figure 4: The number of SLR primary studies associated with data quality challenges in large-scale CPSs by the country of publication.

5.4. SLR Data Extraction

The purpose of the data extraction process (as highlighted in section 4.4) is to quantitatively summarise the information from the primary studies to answer the SLR review questions. Table A.10 shows the results of the data extraction process.

6. RQ1: Data Quality Challenges in Large-Scale CPS Applications.

This section is to answer the first SLR review question (RQ1) listed in Table 2.

Cyber-Physical Systems are designed as a network of computational elements that combine physical input and output mechanisms to interact with the surrounding environment [1]. CPSs are getting more popular in the context of large-scale, smart cities applications which produce a significant amount of data from numerous devices raising quality of service concerns mainly related to real-time big data analysis and data quality management [78], [79], [80]. The quality of data in CPS applications is mainly related to inaccurate observations that do not represent the actual value of the measured phenomena [59].

Data quality issues may occur in large-scale CPSs because of many reasons such as sensor nodes malfunctions [60], calibration issues, poor sensor nodes quality, environmental effects, external noise [61], networks or communication errors, and real-time scheduling problems [8], [58]. Furthermore, limitations in communication channels may cause observations' overlooking in sensor networks during data transmission or aggregation processes [17], [62]. The challenges of data quality management becomes greater in large-scale CPSs, e.g. in environmental and noise monitoring systems, which rely on various sensors and other devices connected by extended networks and usually operate under noisy and dynamic conditions [44], [63], [60]. Large-scale CPS applications, such as environmental monitoring systems, typically involve a large number of low-cost sensor nodes deployed in broad geographical terrains forming a large-scale Wireless Sensors Network (WSN) [61], [62], [81]. Failures in sensor nodes and sensor networks are an inevitable events in large-scale CPS applications, which may cause severe data missing, produce invalid information and potentially reduce the quality of their service [82]. In general, sensor nodes in WSN's have limited computing power, limited storage capacity and limited transmission radius [44], [83]. Therefore, wireless sensor nodes can not send observations to a remote data destination (the sink) directly. Alternatively, a hub device or other sensor nodes works as a bridge to transfer other sensor node's observations. Sensor nodes that are closer to the sink consume more power because they support other sensors to transmit their observations and are expected to have more power failures causing data quality issues [84], [57]. Therefore, sensor nodes may determine the network lifetime based on their battery capacity and affect the system's quality of information [85]. Typically, wireless sensors-nodes of WSN's are distributed according to a spatial or geographical logic over the targeted environment or area of interest [86]. Large-scale applications which exchange geographic information may face spatial data quality challenges mainly because of the amount of the delivered data from remote sensing devices which may directly affect the correctness of related spatial analysis and spatial decision making [87]. Thus, data quality challenges are not only related to observations value attributes but also into mismatches in sensor nodes temporal and spatial contextual attributes [57], [17].

Based on the SLR data extraction process presented in Table A.10, it is possible to link all of the data quality challenges in large-scale CPS applications in to the following categories:

- Errors in sensor nodes measurements.
- Hardware failures in sensor nodes or communication networks.
- Mismatches in sensor nodes spatial and temporal contextual attributes.

Figure 5 shows the main data quality dimensions defined by the SLR data extraction Table A.10 according to the ratio of the primary studies addressing data quality challenges associated with these dimensions[3].

Figure 6 shows a holistic view of the main data quality management methods, data quality dimensions and the main unaddressed data quality challenges in large-scale CPSs.

[3]Descriptive studies were excluded.

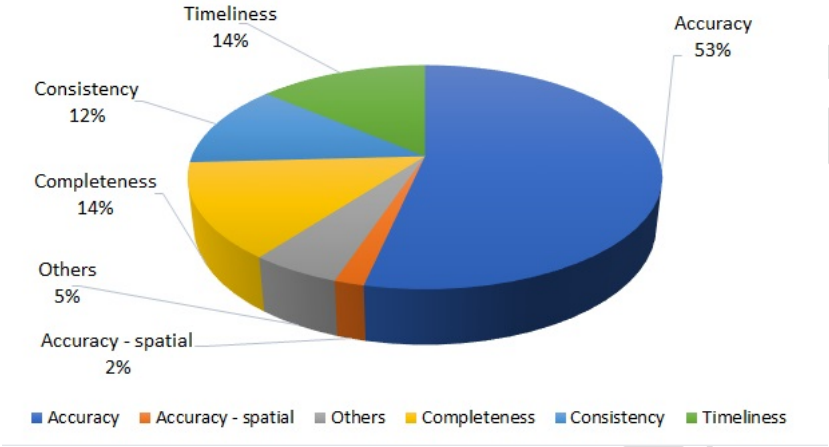


Figure 5: The ratio of the data quality dimensions as addressed by the SLR primary studies.

7. RQ2: Data Mining and Data Quality Management in Large-Scale CPSs.

This section is to answer the second SLR review question (RQ2) listed in Table 2 based on the results of the SLR.

Data quality assessment in large-scale CPS applications using traditional methods is no longer efficient because of the heterogeneous large volume of data that these systems typically exchange [57]. Thus, such systems, usually rely on numerous sensor nodes that stream large volume of data in real-time which requires a high-performance, scalable and flexible tools to effectively provide insight real-time data processing and analysing mechanisms [80], [44], [59], [88]. Based on the results of the SLR data extraction process illustrated in Table A.10, many statistical, technical and machine-learning models were proposed, tested and evaluated mostly for identifying data quality issues, decreasing their occurrence probability and overcoming their impact on the system. Most of these proposed solutions, methods, or models were developed to enhance the reliability of a particular system by improving its data quality based on prior knowledge extracted from the data itself, a process known as Data Mining. Considering the SLR empirical studies only, it is possible to categorise all the adopted data quality assessment/management methods, techniques or solutions into three primary groups:

- Data mining.
- Technical solutions/ models.
- Mathematical models.

Figure 7 shows the usage ratio of the methods of each of the above groups, indicating that data mining methods are the most widely used compared to other technical or mathematical techniques.

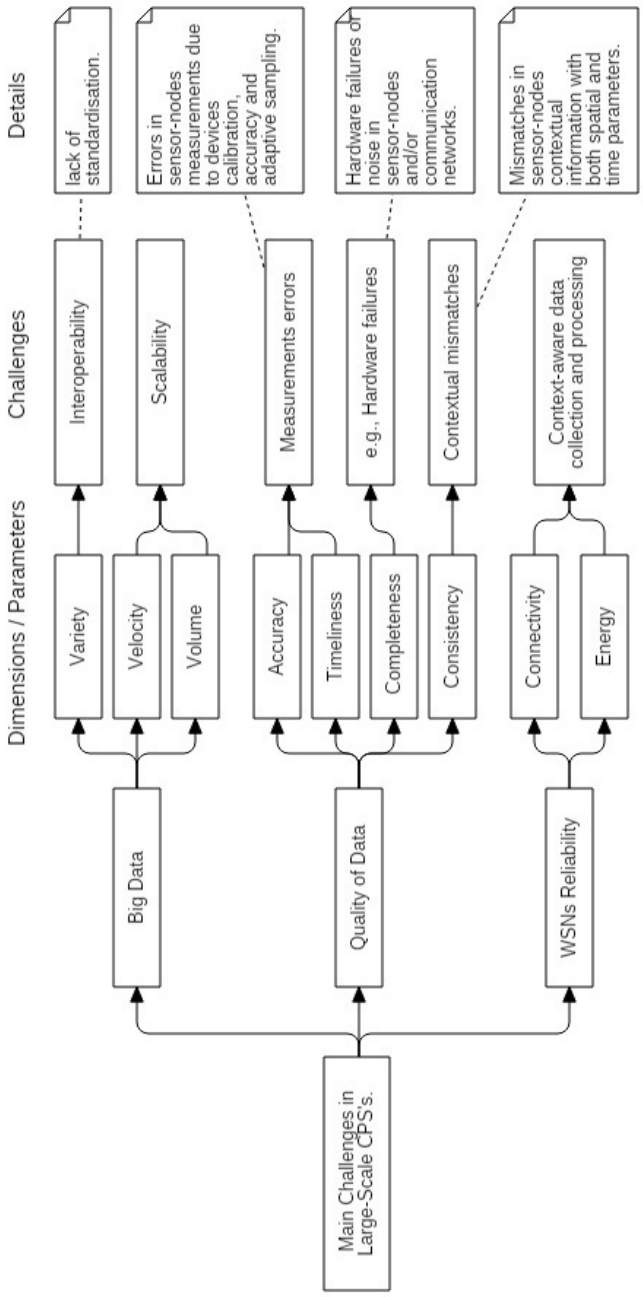


Figure 6: A holistic view of the main data quality management methods, data quality dimensions and the main unaddressed data quality challenges in large-scale CPSs.

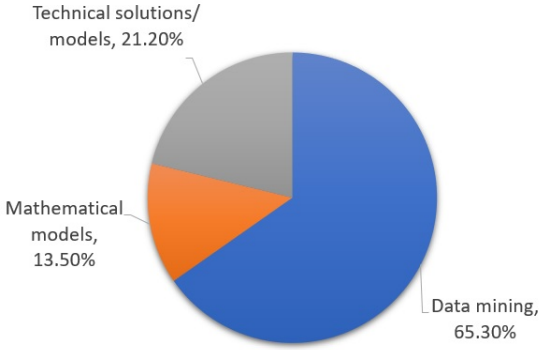


Figure 7: The most popular data quality assessment/management methods or techniques in large-scale CPS applications based the number of the SLR studies.

Data mining is the process of auto-discovering knowledge, patterns or models from large volumes of data using advance data analysis methods [89]. Data mining techniques are essential for data analysis in large-scale CPS applications which relay on sensor node networks that, typically, stream a continuous flow of spatiotemporal[4] data at a relatively high-speed and dynamicity [90]. Focusing on the SLR primary studies that adopted data mining techniques for tackling data quality challenges in large-scale CPS applications reveals that these methods are mainly divided into statistical and machine-learning based methods. Furthermore, it reveals that most popular data mining techniques used for data mining in large-scale CPS applications are anomaly analysis, predictive analysis and clustering analysis, as shown in Figure 8. Moreover, these three leading data mining techniques were applied to address various data quality issues associated with the main data quality dimensions, as shown in Figure 9.

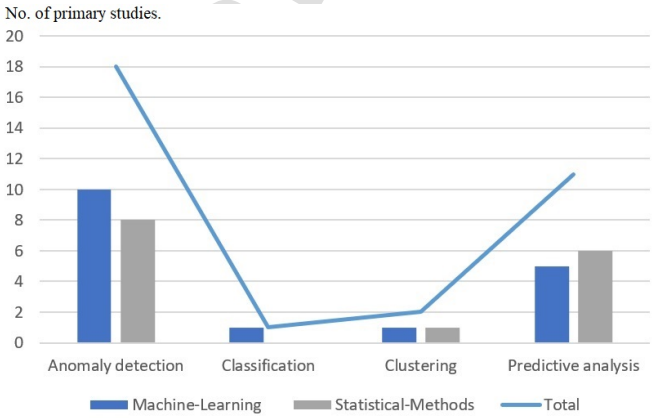


Figure 8: The most popular data mining techniques in large-scale CPS applications based on the No. of the SLR studies.

Figure 10 shows a holistic diagram of the main data quality management/assessment techniques and data

[4]Spatiotemporal data are sensor nodes observations of events that occur in a given place at a particular time.

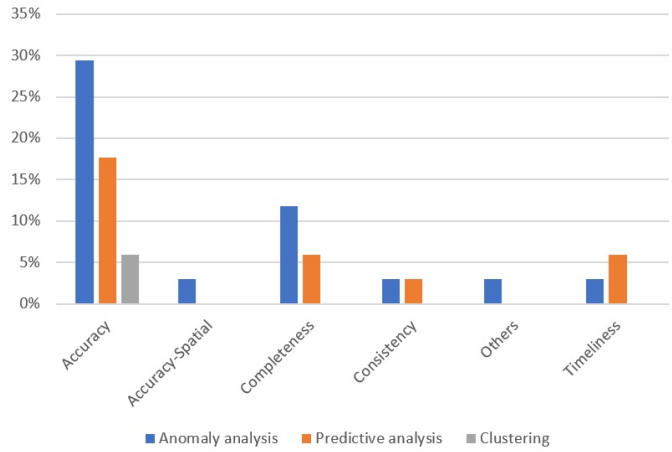


Figure 9: Data mining techniques and the main data quality dimensions in large-scale CPS applications.

quality dimensions based on the SLR results.

7.1. Anomaly Analysis for Data Quality Management

Anomaly analysis, also called outlier detection, is the process of identifying unusual patterns in datasets which do not comply with well-established normal behaviour [90]. If the absolute value of the deviation degree of a sensor node's observation is higher than a pre-calculated threshold value, then this observation is an outlier [91]. As shown in Figures 8 and 9, anomaly analysis is a significant research field in the context of data quality assessment in large-scale CPSs, which mainly investigated using statistical and machine-learning based outlier detection techniques. e.g., Deep Neural Networks (DNN) [92], K-Nearest Neighbours algorithm (KNN) [92], K-means clustering algorithm [93] as machine-learning based outlier detection methods and, standard deviation, correlation coefficient [94] and DBSCAN [88], [95], [96] as statistical outlier detection methods.

Outlier detection relies on the assumption that the values of sensor nodes' observations are correlated spatially, temporally or both spatially and temporally. However, these assumptions are not necessarily always valid, especially in large-scale CPS applications where the correlations between sensor nodes may be affected by many parameters such as the size of the deployment environment and the geographical distribution of sensor nodes [97]. For example, the approach of spatial continuity cannot be applied directly to the real-world temperature observations collected from the temperature sensor nodes distributed around London because of a phenomenon known as the Urban Heat Islands [5]. According to the Meteorological Office (Met Office), the phenomenon of heat islands is caused by many associated factors, such as the heat released from industrial, domestic facilities, concrete and other building material which observe sun heat during the day and release

[5]https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/research/library-and-archive/library/publications/factsheets/factsheet_14-microclimates.pdf

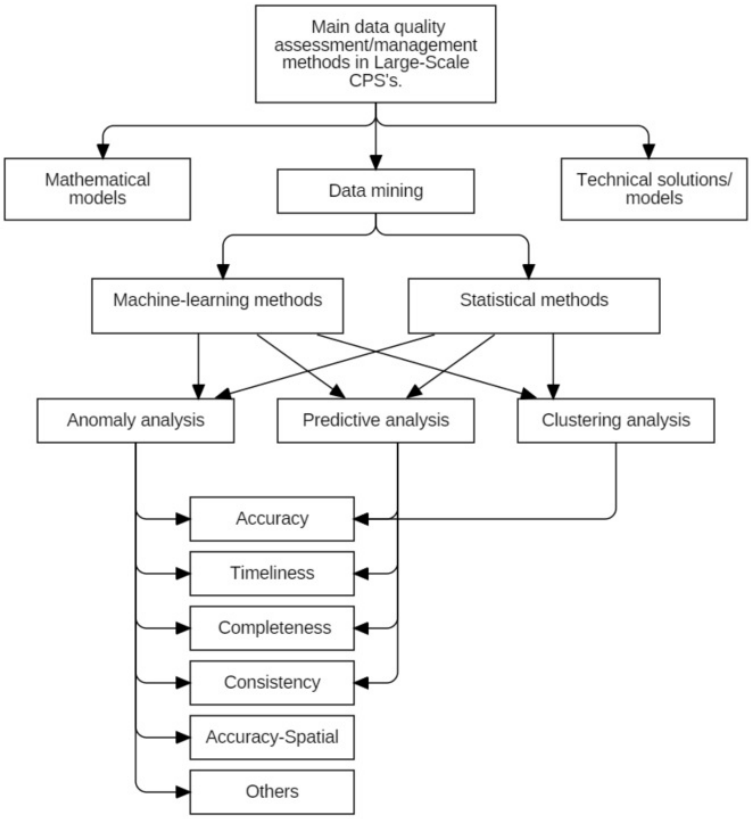


Figure 10: A holistic diagram of the main data quality management/assessment methods/techniques and data quality dimensions based on the SLR results.

it back during the night. The phenomenon of urban heat islands may cause up to 5 degrees (unexpected) deviation among sensor nodes observations at the same point in time, which violates the spatial continuity constrains [98] among sensor nodes observations. The heat profile map of London is shown in Figure 11, where the temperatures in central London may reach 11 C^o while dropped by 6 degrees C^o in the suburbs at the same point in time [98], [99], as shown in Figure 11.

7.1.1. Clustering-Based Outlier Detection

Clustering-based outlier detection relies on comparing individual correlated sensors' observations with the centroid value of their clusters. Therefore, it needs no prior knowledge of the sensor node historical data. Clustering-based outlier detection can be e.g., implemented using DBSCAN clustering algorithm for detecting errors, noise and failures outliers in high-speed, non-stationary, large volume WSN's data, [96], [88]. However, according to [91], clustering can not be considered as a reliable anomaly detection technique in real-world scenarios. It can be used as an outlier filtering mechanism due to challenges in determining both clusters' optimum number of sensor nodes and determining their centroid value in each cluster.

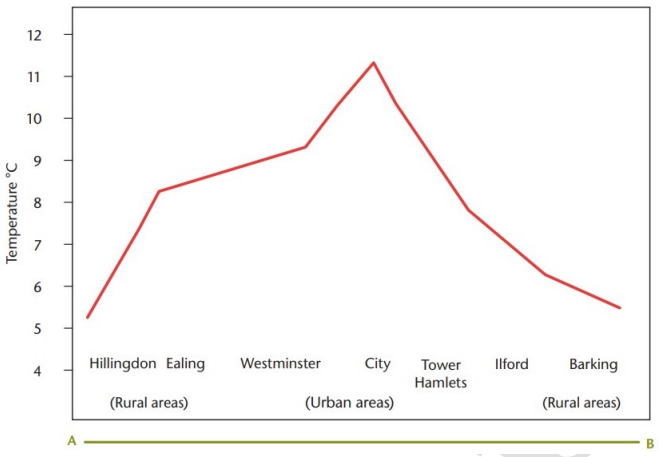


Figure 11: The heat profile map of London highlighting the impact of urban heat islands, [98].

7.1.2. Predictive Analysis Based Outlier Detection

Predictive analysis is the process of mining current and historical data to identify patterns and to forecast the future values of time series [100], [5]. Predictive analysis might be conducted using statistical or machine learning based techniques [101]. For example, machine learning model based on the Random Forest Prediction (Random Forest Regression) method was adopted by [12] for developing an automated data quality control mechanism for weather data. Another example based on statistical predictive analysis using the one step-forward approach, autoregressive moving average (ARMA) model for tackling the inevitable challenge of sensors and sensor networks failure in power terminals, [82]. Furthermore, some applications required a mixed-methods approach, where both machine-learning and statistical methods were adopted to tackle a particular data quality challenge. For example, [61] investigated the use of artificial neural network and linear regression for calibrating low-cost environmental monitoring sensors to improve the accuracy of their observations. Predictive analysis methods rely on predictive models developed using historical data as a training data set. Therefore, using predictive analysis in real-time (online mode) applications raises performance concerns because of the complexity and volume of the required training data set [78], [102]. Using predictive analysis is a challenge in real-time large-scale CPS applications; thus it may require analysing hundreds of sensor nodes data streams in a relatively short time [103], [104]. Furthermore, the training process for predictive analysis requires relatively long and valid (anomaly-free) time series, which cannot be guaranteed in real-world scenarios [91].

8. RQ3: Unaddressed Data Quality Management Challenges in Large-Scale CPSs and Research Gap.

This section is to answer the third SLR review question (RQ3) listed in Table 2.

Data is the bridge between the real physical and the digital worlds where data are used to make intelligent

decisions in CPS applications [105], [12]. Large-scale CPSs relay on the data gathered by sensors and other devices to make intelligent decisions, low-quality data may affect these decisions' reliability, and compromise these systems' quality of services. Ensuring the quality of data in large-scale CPS applications is a challenge due to the following:

- The heterogeneous nature of their data structures, the scale of data that these systems exchange and due to their real-time requirements [12], [63].
- CPS applications are vulnerable to several external and internal factors such as communication network errors, sensor nodes failures which interrupt data transferring process, compromise the integrity of data and reduce the performance and reliability of these applications [106]. Failures in wireless sensors and sensor networks are inevitable events in large-scale CPS applications, and unusually such failures are unpredictable [82], [107]. Furthermore, there is a high possibility of getting erroneous data from sensor node networks due to the limitation in their computing power, storage capacity and communication capabilities [84], [81], [108].
- There are no standard criteria to define high-quality data which typically diverse in measure attributes and requirements from application to another [109]. Data quality is a subjective concept that varies by the purpose or the intended use of the data. Therefore, data quality standards have not been fully identified or applied successfully in large-scale CPS applications [16].

The SLR data extraction Table A.10 illustrates many attempts to tackle data quality issues associated with large-scale CPSs while revealing further emerging data quality challenges in which very little or no work has been done. Addressing data quality in large-scale CPS applications is still an open challenge that is not fully enclosed yet [13], [16], [12], [14], [15], [110] which offers new research opportunities and higher possibilities for having more attention in the future. Following is a list of these data quality management issues that the SLR did not resolve:

8.1. Sensor Nodes' Measurement Errors Detection

SLR primary studies which adopted prediction analysis models as data accuracy assessment techniques are sharing the following limitations:

1. All of the proposed prediction analysis models were based on an assumption that data accuracy issues occur for a short interval of time (point outliers). None of the SLR primary studies proposed a solution to address data accuracy issues associated with long outliers. Long outliers change the time-series' pattern, so the inaccurate observations appear as the standard. In case, a time-series with long outliers is used as the predictive model training dataset. It will compromise the modes' ability to detect data accuracy issues correctly.
2. No systematic method or approach was demonstrated by any of the SLR empirical primary studies on how it was possible to ensure the quality of real-world dataset used to train or calibrate the predictive analysis model.

3. None of the SLR primary studies provided a comparison or a justification for why a particular predictive analysis technique was chosen over another. For example, it is not clear when to apply deep learning neural networks as a predictive technique [111] instead of linear regression [61].
4. SLR primary studies that investigated anomaly analysis as a solution to evaluate the accuracy of sensor nodes measurements by comparing their observations with different sensor nodes or to a pre-calculated threshold value were based on the assumption that these sensor nodes are spatially correlated. However, this assumption is not necessarily always valid in large-scale CPS applications. The spatial continuity among sensor nodes in large-scale CPS applications might be compromised because of the vast distance separating these devices or other factors that disrupt the spatial continuity constraints, as detailed in section 7.1.

8.2. Sensor Nodes' and Sensors Networks' Failures Detection

The SLR primary studies provided no systematic method or a generic approach for detecting sensor nodes and sensor node networks hardware failures in large-scale CPS applications. All proposed failure detection mechanisms were mainly domain-specific solutions. For example, signal processing techniques were explicitly utilised for monitoring the hardware status of a Chinese network of weather radars by [57] which can not be applied as a generic solution for hardware failures detection in sensor node networks of large-scale CPSs.

8.3. Ensures the Quality of Observations' Spatial and Temporal Contextual Attributes

The SLR primary studies revealed that further research is required to address the challenge of ensuring the quality of sensor nodes' spatial and temporal contextual attributes. Spatial data quality issues (sensor nodes location) may affect the validity of any related spatial analysis. Furthermore, very limited or no research has practically investigated the possibility of using observations timestamp analysis techniques as a potential solution to improve the quality of sensor nodes' spatial contextual attributes.

9. Conclusion

CPSs are multidisciplinary cross-domain information systems that bring together different smart cities' public services. Data sensing, knowledge extraction, and higher automation are critical elements in these services' future. Ensuring data quality in smart cities large-scale CPS applications is a critical requirement to guarantee their service quality. Although a significant effort was conducted to address data quality management challenges large-scale CPS applications using advanced statistical and machine-learning techniques, data quality management in large-scale CPSs is still an open challenge. This study concluded that more work is required to provide a practical, comprehensive data quality management solution to detect sensor nodes measurement errors associated with the main data quality dimensions of accuracy, timeliness, completeness, and consistency. No systematic or generic approach was demonstrated for sensor nodes and sensor node networks hardware failure detection in large-scale CPS applications. Moreover, further research is required

to address the challenge of ensuring the quality of the spatial and temporal contextual attributes of sensor nodes observations.

10. Acknowledgements

This study was supported by the University of East London, UK, through the PhD scholarship scheme. The financial support is greatly acknowledged.

Appendix A. Primary Studies Referencing Details

Table A.9: Primary Studies Referencing Details and their Overall Quality Assessment Score.

Ref.	Study Identifier	Year	Research type	Approach	Q1	Q2	Q3	Q4	Q5	Score
S1 - [12]	IEEE Conferences	2018	Solution proposal	Framework	1	0.5	1		1	3.5
S2 - [82]	IEEE Conferences	2019	Solution proposal	Method	1	0.5	1	1	1	4.5
S3 - [105]	IEEE Conferences	2015	Review	Guideline	0.5	0.5			1	2
S4 - [13]	IEEE Conferences	2019	Solution proposal	Framework	1	0.5			1	2.5
S5 - [109]	IEEE Conferences	2016	Solution proposal	Framework	1				2	3
S6 - [80]	IEEE Journals	2019	Solution proposal	Model	0.5					0.5
S7 - [84]	IEEE Conferences	2019	Solution proposal	Method	1				1	2
S8 - [16]	IEEE Conferences	2018	Solution proposal	Framework	0.5					0.5
S9 - [107]	IEEE Conferences	2014	Solution proposal	Model	0.5	0.5	1		3	5
S10 - [85]	IEEE Journals	2016	Solution proposal	Algorithm	1				1	2
S11 - [102]	IEEE Journals	2018	Solution proposal	Framework	1				2	3
S12 - [44]	IEEE Conferences	2016	Solution proposal	Framework	1	0.5			4	5.5
S13 - [63]	IEEE Journals	2014	Solution proposal	Framework	1	0.5			2	3.5
S14 - [58]	IEEE Conferences	2016	Solution proposal	Model	1	0.5			1	2.5
S15 - [57]	IEEE Conferences	2019	Solution proposal	Framework	1	0.5	1		1	3.5
S16 - [112]	IEEE Conferences	2019	Solution proposal	Tool	1	0.5	1		1	3.5
S17 - [113]	IEEE Conferences	2017	Review	Guideline	0.5				1	1.5
S18 - [104]	IEEE Conferences	2016	Solution proposal	Tool	1	1				2
S19 - [114]	IEEE Journals	2016	Solution proposal	Model	1	0.5			1	2.5
S20 - [87]	IEEE Conferences	2015	Solution proposal	Framework	0.5				1	1.5
S21 - [115]	IEEE Conferences	2017	Review	Guideline	0.5					0.5
S22 - [14]	IEEE Conferences	2016	Review	Guideline	0.5					0.5
S23 - [116]	IEEE Journals	2020	Solution proposal	Model	1				1	2
S24 - [83]	IEEE Journals	2017	Solution proposal	Method	1	0.5			1	2.5
S25 - [108]	IEEE Conferences	2019	Review	Survey	0.5	1				1.5
S26 - [111]	IEEE Conferences	2018	Solution proposal	Tool	1	0.5			3	4.5
S27 - [106]	IEEE Conferences	2019	Solution proposal	Algorithm	1	0.5	1		1	3.5
S28 - [94]	IEEE Conferences	2019	Solution proposal	System	1	0.5			2	3.5
S29 - [88]	IEEE Conferences	2016	Review	Guideline	0.5	0.5	1			2
S30 - [86]	IEEE Conferences	2017	Solution proposal	Model	0.5					0.5
S31 - [103]	IEEE Conferences	2019	Solution proposal	Tool	1	0.5			1	2.5
S32 - [117]	IEEE Conferences	2018	Evaluation research	Guideline	1	0.5			1	2.5
S33 - [81]	IEEE Conferences	2015	Solution proposal	Tool	1	0.5			1	2.5
S34 - [118]	IEEE Conferences	2015	Solution proposal	Method	1	1	1	1	1	5
S35 - [119]	IEEE Conferences	2018	Solution proposal	Model	1				1	2
S36 - [120]	IEEE Conferences	2019	Solution proposal	Tool	1				1	2
S37 - [92]	IEEE Conferences	2016	Solution proposal	Method	1	0.5	1		1	3.5
S38 - [121]	IEEE Journals	2016	Solution proposal	Framework	1					1
S39 - [122]	IEEE Conferences	2018	Solution proposal	Method	1				1	2
S40 - [95]	IET Journals	2017	Solution proposal	Method	1	0.5			1	2.5
S41 - [78]	IEEE Conferences	2019	Solution proposal	Model	0.5					0.5
S42 - [96]	IEEE Conferences	2018	Solution proposal	Framework	1	0.5			1	2.5
S43 - [17]	ACM Journals	2015	Review	Guideline	0.5				3	3.5
S44 - [123]	ACM journals	2018	Solution proposal	System	1	0.5			3	4.5
S45 - [8]	ACM Journals	2015	Review	Guideline	0.5					0.5
S46 - [60]	ACM Journals	2017	Review	Guideline	0.5					0.5
S47 - [124]	ACM Journals	2014	Review	Guideline	0.5					0.5

Table A.9 continued from previous page

Ref.	Study Identifier	Year	Research type	Approach	Q1	Q2	Q3	Q4	Q5	Score
S48 - [59]	ACM Journals	2016	Solution proposal	Framework	1	0.5				1.5
S49 - [125]	ACM Conference	2020	Review	Survey	0.5					0.5
S50 - [126]	ACM Conference	2019	Solution proposal	Method	1					1
S51 - [127]	ACM Conference	2018	Solution proposal	Model	1				1	2
S52 - [93]	ACM Conference	2019	Solution proposal	Model	1	0.5			2	3.5
S53 - [128]	ACM Conference	2019	Solution proposal	Method	1	0.5			1	2.5
S54 - [62]	ACM Journals	2019	Solution proposal	Method	1					1
S55 - [15]	IET Journals	2016	Review	Guideline	1	0.5			1	2.5
S56 - [129]	IET Journals	2017	Review	Guideline	0.5				2	2.5
S57 - [130]	IET Journals	2020	Solution proposal	Method	1	0.5		1	1	3.5
S58 - [79]	SD Journals	2020	Solution proposal	Method	1				1	2
S59 - [131]	SD Journals	2017	Solution proposal	Model	1					1
S60 - [61]	SD Journals	2020	Solution proposal	Model	1	1	1	1	1	5

Table A.10: Results of the SLR data extraction process addressing data quality main challenges and the proposed solutions in large-scale CPS.

Ref.	Purpose/application	Dimensions	Data quality challenges	Proposed solutions/methods
S1 - [12]	Weather data quality control.	Accuracy	Automatic verification of data quality, data integrity and scalability in weather data.	Improving the accuracy of data using machine learning models based on the Random Forest Prediction method (Random Forest Regression), which reduces overfitting without increasing the ratio of error.
S2 - [82]	Data quality enhancement in power terminals.	Completeness	Sensors and sensor networks failures are inevitable events in power IoT systems, which may cause severe data missing.	A one-step forward forecasting model based on the autoregressive–moving-average (ARMA) algorithm was implemented for detecting and mitigating the impact of missing data.
S3 - [105]	An overview of data outliers detection process.	Accuracy	Improving data quality, focusing on data accuracy in the context of IoT applications.	Outlier detection for enhancing the quality of data more efficiently and systematically in IoT environments.
S4 - [13]	An anomaly analysis platform to monitor the quality of data in ubiquitous power IoT.	Accuracy	Monitoring the quality of data of ubiquitous power IoT platform considering its high data exchanging rate, diversity of components and the absence of any effective data management mechanism.	Anomaly detection based on the isolated forests integrated unsupervised machine-learning algorithm. For training the ML model, the historical data was reconstructed to form a time series using the sliding time window model.
S5 - [109]	A data quality reporting framework using graphical editors and models.	Accuracy, Completeness	Data quality is a subjective concept that varies by the purpose or the intended use of the data. There are no standard criteria to define high-quality data which typically diverse in measure attributes and requirements.	A Model-Driven Architecture (MDA) framework developed by Object Management Group (OMG) for software development. It initially developed for data quality management in the context of web applications.
S6 - [80]	A process-centric framework to improve the quality of streamed sensors data.	-	Improving the quality of data in IoT applications which rely on real-time data streaming sensors and have different data structures.	A proposed data quality management framework based on the Process Reference Model (PRM) which only suitable for offline applications with a well-defined process.
S7 - [84]	A mechanism to optimise data collection process in WSN while maintaining the level of the quality of information (IoT).	Timeliness	Improving the quality of information by reducing observations delay and enhance the data lifetime in WSN networks. Improving the reliability of WSN and extending its lifetime by reducing its power consumption rates.	A proposed data transmission path planning mechanism named the Energy Harvesting Path Planning Strategy. It manages observations travel path from sensor nodes to the network sink.
S8 - [16]	A framework for managing data quality in smart connected product (SCP) / IoT environments.	-	The open challenges in SCP/ IoT applications are: data quality standardisation, data quality management especially for applications that collect a significant volume of data from different sources.	A guideline for improving data quality management in SCP environments aligned with ISO/IEC 25012 characteristics and proposed an IoT model based on ISO 8000–62 including the processes of part 8000–61.
S9 - [107]	A computational model for clinical data quality assessment.	Accuracy, Timeliness, Consistency (Dependability)	Improving telemedicine systems technological context to become data quality-aware systems.	A computational model to assess the quality of context data based on optimising the end-to-end resource configuration chain.
S10 - [85]	An algorithm to improve the QoI in WSNs	Accuracy	Improving the lifetime of WSNs, enhancing its data transmission rate while maintaining the quality of information (QoI).	Using the proximal optimisation approach (algorithms), which enhances the performance metrics of WSNs.
S11 - [102]	A QoI framework for WSNs, focusing on completeness and timeliness.	Completeness , Timeliness	Scalability and performance prediction in WSNs concerning the QoI requirements.	Top-K algorithm was adopted for evaluating data completeness metric. Top-k is an image selection algorithm which was implemented to address the non-linear relationship of data completeness with throughput.

Table A.10 continued from previous page

Ref.	Purpose/application	Dimensions	Data quality challenges	Proposed solutions/methods
S12 - [44]	A cloud-service framework for optimising the quality of data streams in real-time WSNs.	Accuracy, Timeliness, Completeness, Consistency	Investigating the quality of data of remote environmental sensors data streams in relation to energy efficiency in WSNs.	A cloud-service framework for optimising the quality of data streams in WSNs while assessing their energy efficiency in real-time. The proposed framework dynamically modify and regulate sensors to maintain data quality and energy-efficient operation in WSN.
S13 - [63]	Energy management of environmental sensors while maintaining the QoI constraints.	Accuracy, Timeliness (latency)	Efficient energy management of environmental monitoring sensors while maintaining the quality-of-information (QoI) in a multitask-oriented environment.	An energy management service compatible with sensors lower layer protocols and over-arching applications, based on signal propagation and processing latency modelling.
S14 - [58]	Enhancing the quality of the information in real-time decisions-based IoT.	Accuracy (value)	To enhance the quality of the information in real-time decisions-based IoT applications which bring many safety and security challenges related to real-time scheduling problems comparing to traditional applications, especially in data processing and smart devices management.	A scheduling model was proposed to enhance the quality of the information in applications that need multiple data items to make decisions based on quality adjustment algorithms and scheduling policies.
S15 - [57]	Data quality assurance in IoT applications	Accuracy (value)	Providing higher data quality assurance in regards to data completeness (availability) and consistency(integrity) of IoT sensors data, which usually affected by sensors failures.	Anomaly detection using the Local Outlier Factor algorithm to identify sensors failures and mismatch in sensors spatial contextual information.
S16 - [112]	Data quality advisor solution for large-scale IoT.	Accuracy (value)	Developing an interactive, large-scale sensors data quality advisor for large-scale, IoT Applications.	A data quality framework that automatically performs data validations. The core of the framework is based on the Direct Acyclic Graph(DAG) model for data quality checks and Scalable Execution Engine (SEE) for executing the validation function.
S17 - [113]	A data quality assessment framework for heterogeneous data resources.	Accuracy	Meeting the expectations of data accuracy and reliability in large sensor networks is a significant challenge due to the heterogeneous nature of engineering data.	Data quality of sensors observations which form long time series can be examined using outlier detection and trend analysis. However, this approach does not address the challenges of checking and analysing a system of sensors network or a realm of heterogeneous time series simultaneously.
S18 - [104]	QoI assessment as a service platform for smart cities applications.	Accuracy, Timeliness	Developing an autonomic, collaborative, extensible and configurable solution to cope with the challenge of QoI assessment within smart cities sensing platforms.	The study proposes an Information Quality Assessment solution as a Service (iQAS) based on measuring data attributes such as accuracy and timeliness using filtering and prediction mechanisms for a given application.
S19 - [114]	Energy efficiency and data quality improvement in large-scale WSNs.	Accuracy	Increasing energy efficiency in WSNs without sacrificing the quality of data.	The study proposes a model for enhancing energy-efficiency in large-scale WSNs by controlling the number of sensors transmissions using the second-order data coupled clustering (SODCC) and the compressive projections principal component analysis (CPPCA) algorithms.
S20 - [87]	Addressing spatial data quality concerns.	Consistency	Addressing spatial geometric inconsistency and topological inconsistencies in geographic information systems.	A proposed framework for correcting the inconsistency in spatial data based on the Triangular Pyramid Framework for spatial analysis.
S21 - [115]	An overview of the challenges of sensor streams in large-scale IoT applications.	-	Addressing the Quality of Observation (QoO) challenges between IoT sensors and their observations destination.	The study proposes a cloud-based IoT platform for collecting, processing and delivering sensors observations.
S22 - [14]	A review of data quality issues in WSNs.	-	The study specified four data quality challenges in WSNs; synchronisation issues, inefficient testing of algorithms, energy management and the lack of novel mathematical modelling.	The study discussed the existing data quality and fault tolerance techniques in WSNs.

Table A.10 continued from previous page

Ref.	Purpose/application	Dimensions	Data quality challenges	Proposed solutions/methods
S23 - [116]	Sensors data trust in IoT applications using temporal correlation.	Accuracy	Assessing the trust of sensors data in large-scale IoT applications.	A model for assessing trust in large-scale IoT sensors data using a temporal correlation-based approach and adopting Deep Neural Networks (DNN).
S24 - [83]	Data quality of event-sensitive monitoring in vibration sensor networks.	Accuracy (value)	Ensuring the quality of data in vibration data-intensive monitoring applications which must deliver high-resolution observations accurately and continuously to the system processing core.	A decentralised control and data reduction algorithm utilising the Goetzel algorithm to address data quality challenges in event sensitive WVSN applications.
S25 - [108]	A review of outlier detection techniques in WSNs in IoT frameworks.	-	Addressing data quality checking techniques in wireless sensors networks.	The quality of data in large-scale IoT frameworks can be examined using machine learningtechniques such as neural Networks, clustering and classification for being powerful methods to detect outliers in sensors data.
S26 - [111]	IoT reliability in sensors networks systems.	Accuracy, Consistency, Reliability	Optimising sensors coverage and reducing energy consumption in IoT sensing networks.	The study proposes a model which uses the minimum set cover theorem for identifying reliable sensor nodes with more extended sensing sequence of observations, higher accuracy rate and consistency per sensing region to facilitate optimal coverage.
S27 - [106]	Addressing the issue of missing data in medical IoT applications.	Accuracy	Developing a prediction model for imputing missing data in IoT applications.	The study proposed a prediction model for detecting and estimating missing data in IoT applications using deep learning neural networks.
S28 - [94]	Data quality control of the Chinese wind radars' network.	Accuracy, Completeness	Ensuring data accuracy and conventional functionality of a large-scale wind radars' network.	The data quality evaluation and detection mechanisms are mainly based on statistical techniques such as standard deviation, correlation coefficient and data acquisition rate of the observation collected from the wind profile radar network.
S29 - [88]	Reviewing different clustering techniques for detecting outliers in data streams.	Accuracy	Outliers detection in streamed data due to its high-speed, non-stationary, large volume, and attributes diversity comparing to static data sets.	The study concluded that clustering has a fundamental role in data streams mining possess for outliers detection, especially the basic density-based clustering (DBSCAN) algorithms.
S30 - [86]	Improving the quality of data of WSNs semantic information.	-	Improving the quality of semantic row data in WSNs, and improving sensors spatial and temporal ontology.	A model for providing semantic sensor data through a Semantic Sensor Web (SSW) services to enhance the quality of sensors semantic data using data integration and fusion techniques.
S31 - [103]	A big data accuracy assessment tool.	Accuracy	Developing a big data quality assessment tool.	The study proposes a data accuracy assessment tool based on machine learning (K-Nearest Neighbors, Logistic Regression and Decision Trees) and Apache Spark for handling large-scale datasets.
S32 - [117]	Inconsistency analysis in large-scale, non-stationary and inconsistent time series.	Consistency	Interpolation of missing/insufficient data in real-world, large time-series.	The study outlined four different time-series interpolation/ predictions methods for short-term statistical time-series analysis using the one step ahead prediction and the moving data window approach.
S33 - [81]	Anomaly detection based on spatial distribution data in WSNs.	Accuracy of spatial attributes	Detecting abnormalities based on spatial distribution data of sensor nodes and using numerical data outlier detectors in WSNs.	K-nearest neighbours algorithm (KNN) and Euclidian distance were adopted to detect abnormalities from the spatial distribution of data and depending on WSNs Low Energy Adaptive Clustering Hierarchy protocol (LEACH).
S34 - [118]	Controlling the quality of data in large-scale water-level monitoring system.	Accuracy	Developing a solution to replace (DBSCAN) (Density-Based Spatial Clustering of Applications with Noise) with a more efficient higher performance clustering algorithm.	A linear-clustering algorithm was developed to replace DBSCAN for data quality control in a large-scale, water-level monitoring system. The experimental results indicated that the performance of the proposed domain-specific outlier detection algorithm is higher than DBSCAN.
S35 - [119]	Missing data estimation/replication in industrial WSNs.	Availability	Enhancing data availability in the presence of sensor nodes failures in industrial WSNs (IWSNs).	The proposed solution is based on utilising sensor nodes memory space to save measurements from their neighbouring nodes and carry the last observation forward to estimate missing data. This approach is limited to time series with stable trend.

Table A.10 continued from previous page

Ref.	Purpose/application	Dimensions	Data quality challenges	Proposed solutions/methods
S36 - [120]	A monitoring system for large-scale sensors networks.	-	Automatically monitoring the infrastructure of large-scale sensors networks (124 stations) deployed over vast geographical terrain (20 sq. km).	The proposed solution is based on a rule engine which reads the system's parameters and compares them against pre-calculated threshold values.
S37 - [92]	Duplicate records detection in real-world applications.	Duplication	Detecting and cleaning duplicated records to ensures the quality of data and maintains applications performance.	A genetic neural network-based approach for detecting duplicated records.
S38 - [121]	QoI framework for smart cities applications.	-	Meeting information quality requirements for smart cities scale data analysis applications.	The study proposes a large-scale data analysis framework to provide near real-time machine-interpretable data for smart cities applications. The proposed framework considered many quality measures and fault recovery techniques to enable quality-aware and up-to-date smart city applications.
S39 - [122]	Evaluating the QoI in IoT as a service in a smart cities scale applications.	Accuracy	Enhancing public information assets using advanced methods to support public administrations services.	The study proposes a quality of information evaluation strategy based on Multiple Criteria Decision Making (MCDM) methods in the context of evaluating the quality of public data and related metadata in the scale of smart cities applications.
S40 - [95]	Outliers detection in WSNs.	Accuracy	Ensuring the quality of data through outlier detection for identifying intrusion, errors and noise in wireless sensor networks applications.	Density-based outlier detection technique was evaluated using (DBSCAN) as outlier detection technique for systems with expected normal behaviour.
S41 - [78]	Data quality evaluation in a large-scale transportation system.	-	Addressing the problem of real-time data analysis and handling imperfections in sensors data of smart cities IoT applications.	Proposing a data integration platform from different sources to interpret the information certainty level using an evidential database based on the evidence theory.
S42 - [96]	Density-based clustering for outlier detection in WSNs.	Accuracy	Improving the quality of information via outlier measurements, mainly by detecting errors, noise and failures in wireless sensor networks.	A modified density-based spatial clustering of applications with noise (DBSCAN)-OD algorithm was developed based DBSCAN algorithm in order to detect computing and spatial-temporal parameters to identify outliers from standard sensors.
S43 - [17]	A guideline for data quality challenges in smart cities.	-	Identifying the main data quality challenges in smart cities applications especially issues related to wireless networks energy restrictions, sensors bandwidth or connectivity limitations or for challenges associated with the large data volumes, high data velocity, dynamicity or diversity of types and structures.	The study classified data quality issues in smart cities scale applications into three main types: measurements or precision errors in sensor nodes, external noise or network communication errors and integrity of sensors observations in both spatial and temporal dimensions.
S44 - [123]	Automating large-scale data quality verification.	Accuracy, Consistency, Completeness	Verifying the quality of data against missing or incorrect information.	The study proposes an automated data quality verifications system. The proposed system adopts a declarative API to combine standard data quality constraints with user pre-defined validation rules and leverages machine learning for anomaly detection using data predictability approach based on historical time series.
S45 - [8]	A guideline to the main data quality challenges in CPSs.	-	The most significant challenge in CPSs is identifying and filtering faulty data.	The study highlights the need for developing novel algorithms and protocols that can effectively detect and filter erroneous data in CPS applications such as faulty data and information loss models, localized algorithm and the lightweight secure data storage and transmission protocols.
S46 - [60]	An introduction to dynamic data quality challenges.	Accuracy	Ensuring the quality of dynamic data in IoT applications which typically generated by multi-vendors devices, micro services, automated processes and different types of sensors.	The study highlights that maintaining the quality of dynamic data in IoT applications is an open challenge which provides new research opportunities.
S47 - [124]	A review of process-driven data quality management.	-	Developing a broadly applicable process-based model for improving and sustaining the quality of data.	The study concludes that further representational analysis is required to enhanced process modelling language for process-driven data quality management (PDDQM) modelling.

Table A.10 continued from previous page

Ref.	Purpose/application	Dimensions	Data quality challenges	Proposed solutions/methods
S48 - [59]	Data quality management for data streams.	-	Maintaining the quality of the data stream without affecting the real-time performance of the system.	The study proposes an ontology-based data quality monitoring framework based on the characteristics of relational data stream management to observe data quality values and take counteractions to balance the performance.
S49 - [125]	A survey about the importance of high-quality data for machine learning.	-	Enhancing the performance and accuracy of machine learning models by ensuring the quality of their training dataset.	The study concluded that researches were focusing on improving the quality of machine learning models. In contrast, insufficient works were conducted to improve the quality of data in the context of its value for machine learning applications.
S50 - [126]	Distributed data mining for identifying data quality issues.	-	Facilitating data quality analysis of data in their distributed state.	The study proposed a data quality issues identifier based on the knowledge extracted from pre-clustering data in its distributed status. The experimental results showed comparable results with those conducted on the integrated warehoused data.
S51 - [127]	Data quality challenges in large industrial environment.	Consistency	Addressing the challenge of data inconsistency to enhance the quality of data in large industrial data environment.	A proposed mathematical data quality assessment and monitoring model based on data cleaning, duplicated records detection and traditional data sorting and merging methods.
S52 - [93]	Data quality assessment for electrical data.	Accuracy, Completeness	To address the challenge of data quality assessment for electricity consumption big data.	The study proposes a model that addresses six data quality assessment indexes including accuracy, completeness and comprehensiveness using time-relevant k-means to detect outliers in voltage curves.
S53 - [128]	Data quality control for weather data.	Completeness (missing values)	Improving the accuracy of weather data which can be degraded by missing sensors readings.	The study evaluated five strategies for detecting failed sensors and statistically identifying anomalies; Mean imputation, MAP imputation, Reduction, Marginalization and Proportional distribution and concluded that missing values handling algorithms can significantly enhance the reliability of weather systems.
S54 - [62]	Data quality assessment in smart sensor networks.	-	Smart Sensor Networks (SSNs) rely on sensors with limited resources and usually deployed in remote and harsh environments which impose data quality challenges in IoT applications.	The study proposes a mechanism to reduce memory and network communication overhead and to impose networks delay.
S55 - [15]	A review of the challenges associated with designing large-scale CPS/IoT applications.	Timeliness	Five main challenges oppose the development of CPS/IoT applications in smart cities applications; middleware development, computation models, fault tolerance, data quality management, and a virtual run-time environment.	The study examined a correlation model among sensors using readings from different sensors to calibrate or verify another sensor's observations when the data are missing.
S56 - [129]	A survey about the quality of observations within sensors web systems.	Accuracy, Timeliness	Addressing the challenge of ensuring the quality of observations in sensors webs which represent the middleware layer between sensors and applications.	The study identified essential requirements for developing the future adaptive quality of observations aware sensor web solutions including standardisation, the need for a layer-based architecture, mediation, adaptation and reconfiguration.
S57 - [130]	Enhancing situation awareness in renewable power systems.	Accuracy	Developing situation awareness system for power systems, that can accurately detect anomalies and robust against multiple data corruptions.	This study tackles two primary challenges faced by conventional situation awareness in power systems: 1) accurately detect anomalies using aggregation of random matrix and long short-term memory network. 2)To be robust against multiple data corruptions using a dedicated workflow designed to decrease the impact of data corruptions.
S58 - [79]	Faulty data detection in cyber-physical systems.	Timeliness	Detecting and filtering faulty data efficiently to improve the quality of the collected data from a system's perspective.	The study proposes an automatic reliability improvement framework of three data quality assessment stages performed on the system input, output and feedback data using machine learning, and operator in the loop approach for detecting faulty-data and improving the reliability of the system.

Table A.10 continued from previous page

Ref.	Purpose/application	Dimensions	Data quality challenges	Proposed solutions/methods
S59 - [131]	Data quality management for manufacturing cyber-physical systems.	-	Developing effective managerial policies for controlling the quality of the data generated by improper operations of physical and cyber components of a service-oriented manufacturing CPS.	The study proposes a two-stage optimization model for data quality management of service-oriented manufacturing CPS (SMCPS). Formal semantics of workflow nets (WF-nets) algorithm together with a two-stage optimization model were used to find the optimal policies that balance the system's objectives.
S60 - [61]	Improving the quality of data of low-cost IoT environmental monitoring networks.	Accuracy	Identifying the main factors that affect the data quality (accuracy) of low-cost IoT sensors in environmental monitoring networks.	The study investigated the use of artificial neural network and linear regression for calibrating low-cost environmental monitoring sensors to improve the accuracy of their readings. These devices are vulnerable to environmental factors such as temperature and humidity; therefore, it is necessary to take these parameters into account when developing the calibration model. The results demonstrated the importance of feature selection process in optimising multi-parameter calibration models.

References

[1] D. E. Robbins, M. M. Tanik, Cyber-Physical Ecosystems: App-Centric Software Ecosystems in Cyber-Physical Environments, in: Applied Cyber-Physical Systems, Springer, New York, NY, 2013, pp. 141–147. doi:10.1007/978-1-4614-7336-7_12.

[2] D. P. F. Möller, Systems and Software Engineering, in: Guide to Computing Fundamentals in Cyber-Physical Systems: Concepts, Design Methods, and Applications, Springer, Cham, 2016, pp. 235–305. doi:10.1007/978-3-319-25178-3_6.

[3] A. A. Jahromi, D. Kundur, Fundamentals of Cyber-Physical Systems, in: Cyber-Physical Systems in the Built Environment, Springer, Cham, 2020, pp. 1–13. doi:10.1007/978-3-030-41560-0_1.

[4] F. Tao, Q. Qi, A. Liu, A. Kusiak, Data-driven smart manufacturing, J. Manuf. Syst. 48 (2018) 157–169. doi:10.1016/j.jmsy.2018.01.006.

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

[6] L. Zhang, Multi-view, multi-domain, and multi-paradigm approaches for specification and modeling of big data driven cyber-physical systems (2015).
URL https://api.semanticscholar.org/CorpusID:63432037

[7] M. Pan, J. Wang, S. M. Errapotu, X. Zhang, J. Ding, Z. Han, Big Data Privacy Preservation for Cyber-Physical Systems, Springer, 2019.

[8] K. Sha, S. Zeadally, Data Quality Challenges in Cyber-Physical Systems, J. Data and Information Quality 6 (2-3) (2015) 1–4. doi:10.1145/2740965.

[9] D. Williams, H. Tang, Data Quality Management for Industry 4.0: A Survey - ProQuest, [Online; accessed 13. Sep. 2020] (Mar 2020).
URL https://search.proquest.com/docview/2386939130?pq-origsite=gscholar&fromopenview=true

[10] S. Vaidya, P. Ambad, S. Bhosle, Industry 4.0 – A Glimpse, Procedia Manuf. 20 (2018) 233–238. doi:10.1016/j.promfg.2018.02.034.

[11] W. Grega, A. J. Kornecki, Real-time cyber-physical systems transatlantic engineering curricula framework, in: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, 2015, pp. 755–762.

[12] 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. doi:10.1109/ICET.2018.8603594.

[13] B. Peng, F. Shang, Y. Wang, G. Chen, Z. Zhou, L. He, Research on data quality detection technology based on ubiquitous state grid internet of things platform, in: 2019 IEEE 3rd International Electrical and Energy Conference (CIEEC), 2019, pp. 1018–1023. doi:10.1109/CIEEC47146.2019.CIEEC-2019384.

[14] B. Prathiba, K. J. Sankar, V. Sumalatha, Enhancing the data quality in wireless sensor networks — a review, in: 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), 2016, pp. 448–454. doi:10.1109/ICACDOT.2016.7877626.

[15] C.-S. Shih, J.-J. Chou, N. Reijers, T.-W. Kuo, Designing CPS/IoT applications for smart buildings and cities, IET Cyber-Phys. Syst.: Theor. Appl. 1 (1) (2016) 3–12. doi:10.1049/iet-cps.2016.0025.

[16] R. Perez-Castillo, A. G. Carretero, M. Rodriguez, I. Caballero, M. Piattini, A. Mate, S. Kim, D. Lee, Data quality best practices in iot environments, in: 2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC), 2018, pp. 272–275. doi:10.1109/QUATIC.2018.00048.

[17] P. Barnaghi, M. Bermudez-Edo, R. Tönjes, Challenges for Quality of Data in Smart Cities, Data and Information Quality 6 (2-3) (2015) 1–4. doi:10.1145/2747881.

[18] J. Včelák, A. Vodička, M. Maška, J. Mrňa, Smart building monitoring from structure to indoor environment, in: 2017 Smart City Symposium Prague (SCSP), IEEE, 2017, pp. 1–5.

[19] I. Mahmood, J. A. Zubairi, Efficient waste transportation and recycling: Enabling technologies for smart cities using the internet of things, IEEE Electrification Magazine 7 (3) (2019) 33–43.

[20] M. Goldberg, Z. Zhang, A cyber-physical system framework towards smart city and urban computing to aid people with disabilities, in: 2018 27th Wireless and Optical Communication Conference (WOCC), IEEE, 2018, pp. 1–5.

[21] E. Kim, Smart city service platform associated with smart home, in: 2017 International Conference on Information Networking (ICOIN), IEEE, 2017, pp. 608–610.

[22] Q. Zhang, R. Duan, J. Wang, Y. Cui, Smart building environment monitoring based on gaussian process, in: 2019 International Conference on Control, Automation and Information Sciences (ICCAIS), IEEE, 2019, pp. 1–6.

[23] D. R. Naik, L. B. Das, T. Bindiya, Wireless sensor networks with zigbee and wifi for environment monitoring, traffic management and vehicle monitoring in smart cities, in: 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), IEEE, 2018, pp. 46–50.

[24] A. A. Brincat, F. Pacifici, S. Martinaglia, F. Mazzola, The internet of things for intelligent transportation systems in real smart cities scenarios, in: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), IEEE, 2019, pp. 128–132.

[25] C. Lin, G. Han, J. Du, T. Xu, L. Shu, Z. Lv, Spatio-temporal congestion-aware path planning towards intelligent transportation systems in software-defined smart city iot, IEEE Internet of Things Journal.

[26] L. Liu, W. Chen, A. Solanas, A. He, Knowledge, attitude, and practice about internet of things for healthcare, in: 2017 International Smart Cities Conference (ISC2), IEEE, 2017, pp. 1–4.

[27] L. F. Herrera-Quintero, J. C. Vega-Alfonso, K. B. A. Banse, E. C. Zambrano, Smart its sensor for the transportation planning based on iot approaches using serverless and microservices architecture, IEEE Intelligent Transportation Systems Magazine 10 (2) (2018) 17–27.

[28] S. Bose, N. Mukherjee, S. Mistry, Environment monitoring in smart cities using virtual sensors, in: 2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud), IEEE, 2016, pp. 399–404.

[29] F. Bonafini, D. F. Carvalho, A. Depari, P. Ferrari, A. Flammini, M. Pasetti, S. Rinaldi, E. Sisinni, Evaluating indoor and outdoor localization services for lorawan in smart city applications, in: 2019 II Workshop on Metrology for Industry 4.0 and IoT (MetroInd4. 0&IoT), IEEE, 2019, pp. 300–305.

[30] J. Santos, T. Wauters, B. Volckaert, F. De Turck, Resource provisioning for iot application services in smart cities, in: 2017 13th International Conference on Network and Service Management (CNSM), IEEE, 2017, pp. 1–9.

[31] M. Bisadi, A. Akrami, S. Teimourzadeh, F. Aminifar, M. Kargahi, M. Shahidehpour, IoT-Enabled Humans in the Loop for Energy Management Systems: Promoting Building Occupants' Participation in Optimizing Energy Consumption, IEEE Electrif. Mag. 6 (2) (2018) 64–72. doi:10.1109/MELE.2018.2816844.

[32] A. R. Patel, S. Azadi, M. H. Babaee, N. Mollaei, K. L. Patel, D. R. Mehta, Significance of robotics in manufacturing, energy, goods and transport sector in internet of things (iot) paradigm, in: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–4.

[33] J. Walia, A. Walia, C. Lund, A. Arefi, The characteristics of smart energy information management systems for built environments, in: 2019 IEEE 10th International Workshop on Applied Measurements for Power Systems (AMPS), 2019, pp. 1–6.

[34] D. Minoli, K. Sohraby, B. Occhiogrosso, IoT Considerations, Requirements, and Architectures for Smart Buildings—Energy Optimization and Next-Generation Building Management Systems, IEEE IoT J. 4 (1) (2017) 269–283. doi:10.1109/JIOT.2017.2647881.

[35] G. R. C. Andrés, Cleanwifi: The wireless network for air quality monitoring, community internet access and environmental education in smart cities, in: 2016 ITU Kaleidoscope: ICTs for a Sustainable World (ITU WT), 2016, pp. 1–6.

[36] M. M. Rathore, A. Ahmad, A. Paul, G. Jeon, Efficient graph-oriented smart transportation using internet of things generated big data, in: 2015 11th International Conference on Signal-Image Technology Internet-Based Systems (SITIS), 2015, pp. 512–519.

[37] M. Chen, J. Yang, L. Hu, M. S. Hossain, G. Muhammad, Urban Healthcare Big Data System Based on Crowdsourced and Cloud-Based Air Quality Indicators, IEEE Commun. Mag. 56 (11) (2018) 14–20. doi:10.1109/MCOM.2018.1700571.

[38] R. Lee, R. Jang, M. Park, G. Jeon, J. Kim, S. Lee, Making iot data ready for smart city applications, in: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 2020, pp. 605–608.

[39] X. Liu, X. L. Yu, T. Fei, Research on Building Data Acquisition Methods in Smart City, 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (2012) 144–147doi:10.1109/ICITBS49701.2020.00038.

[40] X. Fang, Improving data quality for low-cost environmental sensors, Ph.D. thesis, University of York (Aug 2018).
URL http://etheses.whiterose.ac.uk/21259

[41] S. Shukla, Balachandran K, Sumitha V S, A framework for smart transportation using big data, in: 2016 International Conference on ICT in Business Industry Government (ICTBIG), 2016, pp. 1–3.

[42] N. Larburu, R. Bults, M. van Sinderen, H. Hermens, Quality-of-Data Management for Telemedicine Systems, Procedia Comput. Sci. 63 (2015) 451–458. doi:10.1016/j.procs.2015.08.367.

[43] T. Luo, J. Huang, S. S. Kanhere, J. Zhang, S. K. Das, Improving IoT Data Quality in Mobile Crowd Sensing: A Cross Validation Approach, IEEE IoT J. 6 (3) (2019) 5651–5664.
URL https://ieeexplore.ieee.org/document/8666717

[44] 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.

[45] M. Foehr, J. Vollmar, A. Calà, P. Leitão, S. Karnouskos, A. W. Colombo, Engineering of next generation cyber-physical automation system architectures, Multi-Disciplinary Engineering for Cyber-Physical Production Systems (2017) 185–206doi:10.1007/978-3-319-56345-9_8.
URL https://link.springer.com/chapter/10.1007/978-3-319-56345-9_8

[46] Y. Ashibani, Q. H. Mahmoud, Cyber physical systems security: Analysis, challenges and solutions, Computers and Security 68 (2017) 81–97. doi:10.1016/j.cose.2017.04.005.
URL https://www.sciencedirect.com/science/article/pii/S0167404817300809

[47] M. Lohstroh, P. Derler, M. Sirjani, Principles of Modeling - Essays Dedicated to Edward A. Lee on the Occasion of His 60th Birthday | Marten Lohstroh | Springer, Springer International Publishing, 2018. doi:10.1007/978-3-319-95246-8.

[48] M. Haseeb, H. I. Hussain, B. Ślusarczyk, K. Jermsittiparsert, Industry 4.0: A solution towards technology challenges of sustainable business performance, Social Sciences 8 (2019) 154. doi:10.3390/socsci8050154.
URL https://www.mdpi.com/2076-0760/8/5/154/htm

[49] A. A. Alwan, A. Baravalle, M. A. Ciupala, P. Falcarin, An open source software architecture for smart buildings, in: B. R. Arai K., Kapoor S. (Ed.), Advances in Intelligent Systems and Computing. IntelliSys 2018, Springer, 2019. doi:10.1007/978-3-030-01057-7_14.

[50] A. Ordonez, V. Alcázar, J. C. Corrales, P. Falcarin, Automated context aware composition of advanced telecom services for environmental early warnings, Expert Systems with Applications 41 (13) (2014) 5907–5916. doi:https://doi.org/10.1016/j.eswa.2014.03.045.
URL https://www.sciencedirect.com/science/article/pii/S0957417414001833

[51] F.-J. Wu, T. Luo, H. P. Tan, Case studies of wsn-cps applications, Cyber-Physical System Design with Sensor Networking Technologies 2 (2016) 269.

[52] A. Grizhnevich, Iot for smart cities: Use cases and implementation strategies, Science Soft.

[53] A. Hakiri, A. Gokhale, Work-in-Progress: Towards Real-Time Smart City Communications using Software Defined Wireless Mesh Networking, 2018 IEEE Real-Time Systems Symposium (RTSS) (2014) 177–180doi:10.1109/RTSS.2018.00034.

[54] E. Badidi, N. E. Neyadi, M. Al Saeedi, F. Al Kaabi, M. Maheswaran, Building a Data Pipeline for the Management and Processing of Urban Data Streams, Springer, Cham, 2018. doi:10.1007/978-3-319-97271-8_15.

[55] S. Kale, H. Tamakuwala, V. Vijayakumar, L. Yang, B. S. R. Kshatriya, Big Data in Healthcare: Challenges and Promise, Springer, Singapore, 2019. doi:10.1007/978-981-32-9889-7_1.

[56] S. E. Bibri, Introduction: The Rise of Sustainability, ICT, and Urbanization and the Materialization of Smart Sustainable Cities, Springer, Cham, 2018. doi:10.1007/978-3-319-73981-6_1.

[57] 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. doi:10.1109/LATINCOM48065.2019.8937930.

[58] J. Kim, T. Abdelzaher, L. Sha, A. Bar-Noy, R. Hobbs, W. Dron, On maximizing quality of information for the internet of things: A real-time scheduling perspective (invited paper), in: 2016 IEEE 22nd International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), 2016, pp. 202–211. doi:10.1109/RTCSA.2016.47.

[59] S. Geisler, C. Quix, S. Weber, M. Jarke, Ontology-Based Data Quality Management for Data Streams, J. Data and Information Quality 7 (4) (2016) 1–34. doi:10.1145/2968332.

[60] A. G. Labouseur, C. C. Matheus, An Introduction to Dynamic Data Quality Challenges, J. Data and Information Quality 8 (2) (2017) 1–3. doi:10.1145/2998575.

[61] 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. doi:10.1016/j.icte.2020.06.004.

[62] G. R. C. de Aquino, C. M. de Farias, L. Pirmez, Hygieia: data quality assessment for smart sensor network, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3297280.3297564.

[63] C. H. Liu, J. Fan, J. W. Branch, K. K. Leung, Toward qoi and energy-efficiency in internet-of-things sensory environments, IEEE Transactions on Emerging Topics in Computing 2 (4) (2014) 473–487. doi:10.1109/TETC.2014.2364915.

[64] I. ISO 8402, Iso 8402: Quality management and quality assurance — vocabulary (1994).
URL https://www.iso.org/standard/20115.html

[65] J. M. Juran, F. M. Gryna, et al., Juran's quality control handbook, Vol. 4, McGraw-Hill New York, 1988.

[66] A. Maydanchik, Data quality assessment, Technics publications, 2007.

[67] P. B. Crosby, Quality is free: The art of making quality certain, Vol. 94, McGraw-hill New York, 1979.

[68] C. Batini, M. Scannapieco, Data Quality Dimensions, Springer International Publishing, Cham, 2016, pp. 21–51. doi:10.1007/978-3-319-24106-7_2.
URL https://doi.org/10.1007/978-3-319-24106-7_2

[69] L. Sebastian-Coleman, Measuring Data Quality for Ongoing Improvement, Morgan Kaufmann, 2013. doi:10.1016/C2011-0-07321-0.

[70] R. Y. Wang, H. B. Kon, S. E. Madnick, Data quality requirements analysis and modeling, in: Proceedings of IEEE 9th International Conference on Data Engineering, IEEE, 1993, pp. 670–677.

[71] F. Guillet, H. J. Hamilton, Quality measures in data mining, Vol. 43, Springer, 2007.

[72] M. Scannapieco, P. Missier, C. Batini, Data quality at a glance., Datenbank-Spektrum 14 (January) (2005) 6–14.

[73] Y. Wand, R. Y. Wang, Anchoring data quality dimensions in ontological foundations, Commun. ACM 39 (11) (1996) 86–95. doi:10.1145/240455.240479.

[74] C. Fürber, M. Hepp, Using semantic web resources for data quality management, in: P. Cimiano, H. S. Pinto (Eds.), Knowledge Engineering and Management by the Masses, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 211–225.

[75] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, EBSE Technical Report EBSE-2007-01 / Keele University - UK.

[76] R. Malhotra, Empirical research in software engineering: concepts, analysis, and applications, CRC Press, 2015.

[77] B. A. Kitchenham, D. Budgen, P. Brereton, Evidence-based software engineering and systematic reviews, Vol. 4, CRC press, 2015.

[78] H. B. Sta, Strategy for evaluation the data in the context of smart cities: Case study of transport system, in: 2019 IEEE International Smart Cities Conference (ISC2), 2019, pp. 611–618. doi:10.1109/ISC246665.2019.9071775.

[79] G. R., B. R., K. P., Faulty-data detection and data quality measure in cyber–physical systems through Weibull distribution, Comput. Commun. 150 (2020) 262–268. doi:10.1016/j.comcom.2019.11.036.

[80] S. Kim, R. P. D. Castillo, I. Caballero, J. Lee, C. Lee, D. Lee, S. Lee, A. Mate, Extending data quality management for smart connected product operations, IEEE Access 7 (2019) 144663–144678. doi:10.1109/ACCESS.2019.2945124.

[81] 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. doi:10.1109/SSD.2015.7348091.

[82] 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. doi:10.1109/ISCID.2019.10142.

[83] 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 Transactions on Industrial Informatics 13 (2) (2017) 572–583. doi:10.1109/TII.2017.2665463.

[84] 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. doi:10.1109/ICIASE45644.2019.9074030.

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

[86] L. B. Bhajantri, R. Pundalik, Data processing in semantic sensor web: A survey, in: 2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2017, pp. 166–170. doi:10.1109/ICATCCT.2017.8389126.

[87] B. Bahl, Inconsistency quality concerns for spatial database, in: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 1328–1334.

[88] 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.

[89] K. Black, Business statistics: for contemporary decision making, John Wiley & Sons, 2019.

[90] 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.

[91] 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 of Things Journal 5 (2) (2018) 559–570.

[92] 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. doi:10.1109/ICCCBDA.2016.7529552.

[93] 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. doi:10.1145/3371425.3371435.

[94] 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. doi:10.1109/ICMO49322.2019.9026025.

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

[96] 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. doi:10.1109/WCNC.2018.8376988.

[97] 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.

[98] 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

[99] T. J. Chandler, The climate of London, Hutchinson, 1965.

[100] M. Adhikari, S. Kar, S. Banerjee, U. Biswas, Big Data Analysis for Cyber-Physical Systems, undefined.
URL https://www.semanticscholar.org/paper/Big-Data-Analysis-for-Cyber-Physical-\Systems-Adhikari-Kar/2fbe376b34c56ef8a3aa12797fd111c1aa58ae4b

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

[102] S. T. Rager, E. N. Ciftcioglu, R. Ramanathan, T. F. La Porta, R. Govindan, Scalability and satisfiability of quality-of-information in wireless networks, IEEE/ACM Transactions on Networking 26 (1) (2018) 398–411. doi:10.1109/TNET.2017.2781202.

[103] G. Mylavarapu, J. P. Thomas, K. A. Viswanathan, An automated big data accuracy assessment tool, in: 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), 2019, pp. 193–197. doi:10.1109/ICBDA.2019.8713218.

[104] A. Auger, E. Exposito, E. Lochin, iqas: An integration platform for qoi assessment as a service for smart cities, in: 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), 2016, pp. 88–93. doi:10.1109/WF-IoT.2016.7845400.

[105] A. Karkouch, H. Al Moatassime, H. Mousannif, T. Noel, Data quality enhancement in internet of things environment, in: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), 2015, pp. 1–8. doi:10.1109/AICCSA.2015.7507117.

[106] N. Al-Milli, W. Almobaideen, Hybrid neural network to impute missing data for iot applications, in: 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, pp. 121–125. doi:10.1109/JEEIT.2019.8717523.

[107] N. Larburu, R. G. A. Bults, I. Widya, H. J. Hermens, Quality of data computational models and telemedicine treatment effects, in: 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom), 2014, pp. 364–369. doi:10.1109/HealthCom.2014.7001870.

[108] N. Ghosh, K. Maity, R. Paul, S. Maity, Outlier detection in sensor data using machine learning techniques for iot framework and wireless sensor networks: A brief study, in: 2019 International Conference on Applied Machine Learning (ICAML), 2019, pp. 187–190. doi:10.1109/ICAML48257.2019.00043.

[109] A. Karkouch, H. Mousannif, H. A. Moatassime, T. Noel, A model-driven architecture-based data quality management framework for the internet of things, in: 2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech), 2016, pp. 252–259. doi:10.1109/CloudTech.2016.7847707.

[110] P. Falcarin, M. Valla, J. Yu, C. A. Licciardi, C. Fra, L. Lamorte, Context data management: an architectural framework for context-aware services, Service Oriented Computing and Applications 7 (2) (2013) 151–168.

[111] M. B. Krishna, Group-based incentive and penalizing schemes for proactive participatory data sensing in iot networks, in: 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), 2018, pp. 796–801. doi:10.1109/WF-IoT.2018.8355208.

[112] S. Shrivastava, D. Patel, A. Bhamidipaty, W. M. Gifford, S. A. Siegel, V. S. Ganapavarapu, J. R. Kalagnanam, Dqa: Scalable, automated and interactive data quality advisor, in: 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 2913–2922. doi:10.1109/BigData47090.2019.9006187.

[113] N. Micic, D. Neagu, F. Campean, E. H. Zadeh, Towards a data quality framework for heterogeneous data, in: 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2017, pp. 155–162. doi:10.1109/iThings-GreenCom-CPSCom-SmartData.2017.28.

[114] M. I. Chidean, E. Morgado, M. Sanromán-Junquera, J. Ramiro-Bargueño, J. Ramos, A. J. Caamaño, Energy efficiency and quality of data reconstruction through data-coupled clustering for self-organized large-scale wsns, IEEE Sensors Journal 16 (12) (2016) 5010–5020. doi:10.1109/JSEN.2016.2551466.

[115] A. Auger, E. Exposito, E. Lochin, Sensor observation streams within cloud-based iot platforms: Challenges and directions, in: 2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN), 2017, pp. 177–184. doi:10.1109/ICIN.2017.7899407.

[116] G. C. Karmakar, R. Das, J. Kamruzzaman, Iot sensor numerical data trust model using temporal correlation, IEEE Internet of Things Journal 7 (4) (2020) 2573–2581. doi:10.1109/JIOT.2019.2957201.

[117] T. Pełech-Pilichowski, On adaptive prediction of nonstationary and inconsistent large time series data, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 1260–1265. doi:10.23919/MIPRO.2018.8400228.

[118] N. Pattanavijit, P. Vateekul, K. Sarinnapakorn, A linear-clustering algorithm for controlling quality of large scale water-level data in thailand, in: 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2015, pp. 269–274. doi:10.1109/JCSSE.2015.7219808.

[119] H. Zhou, K. Yu, M. Lee, C. Han, The application of last observation carried forward method for missing data estimation in the context of industrial wireless sensor networks, in: 2018 IEEE Asia-Pacific Conference on Antennas and Propagation (APCAP), 2018, pp. 1–2. doi:10.1109/APCAP.2018.8538147.

[120] D. Tomescu, A. Heiman, A. Badescu, An automatic remote monitoring system for large networks, in: 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2019, pp. 71–73. doi:10.1109/CSE/EUC.2019.00023.

[121] D. Puiu, P. Barnaghi, R. Tönjes, D. Kümper, M. I. Ali, A. Mileo, J. Xavier Parreira, M. Fischer, S. Kolozali, N. Farajidavar, F. Gao, T. Iggena, T. Pham, C. Nechifor, D. Puschmann, J. Fernandes, Citypulse: Large scale data analytics framework for smart cities, IEEE Access 4 (2016) 1086–1108. doi:10.1109/ACCESS.2016.2541999.

[122] M. Giacobbe, R. Di Pietro, A. Longo Minnolo, A. Puliafito, Evaluating information quality in delivering iot-as-a-service, in: 2018 IEEE International Conference on Smart Computing (SMARTCOMP), 2018, pp. 405–410. doi:10.1109/SMARTCOMP.2018.00037.

[123] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, A. Grafberger, Automating large-scale data quality verification, Proc. VLDB Endow. 11 (12) (2018) 1781–1794. doi:10.14778/3229863.3229867.
URL https://doi.org/10.14778/3229863.3229867

[124] P. Glowalla, A. Sunyaev, Process-driven data quality management: A critical review on the application of process modeling languages, J. Data and Information Quality 5 (1-2) (2014) 1–30. doi:10.1145/2629568.

[125] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, V. Munigala, Overview and Importance of Data Quality for Machine Learning Tasks, Association for Computing Machinery, New York, NY, USA, 2020. doi:10.1145/3394486.3406477.

[126] E. Januzaj, V. Januzaj, P. Mandl, An Application of Distributed Data Mining to Identify Data Quality Problems, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3366030.3366103.

[127] A. Guo, X. Liu, T. Sun, Research on Key Problems of Data Quality in Large Industrial Data Environment, Association for Computing Machinery, New York, NY, USA, 2018. doi:10.1145/3265639.3265680.

[128] T. Zemicheal, T. G. Dietterich, Anomaly detection in the presence of missing values for weather data quality control, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3314344.3332490.

[129] A. Auger, E. Exposito, E. Lochin, Survey on Quality of Observation within Sensor Web systems, IET Wireless Sens. Syst. 7 (6) (2017) 163–177. doi:10.1049/iet-wss.2017.0008.

[130] Q. Wang, S. Bu, Deep learning enhanced situation awareness for high renewable-penetrated power systems with multiple data corruptions, IET Renewable Power Gener. 14 (7) (2020) 1134–1142. doi:10.1049/iet-rpg.2019.1015.

[131] Z. Song, Y. Sun, J. Wan, P. Liang, Data quality management for service-oriented manufacturing cyber-physical systems, Comput. Electr. Eng. 64 (2017) 34–44. doi:10.1016/j.compeleceng.2016.08.010.

This paper presents a thorough survey of resarch on data quality of cyber-physical systems, with focus on smart cities.

No conflicts of interest to declare.