Computer Science Review 48 (2023) 100554

Contents lists available at ScienceDirect

Computer Science Review

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

Review article

State of the art on quality control for data streams: A systematic literature review

Mostafa Mirzaie [a], Behshid Behkamal [a,*], Mohammad Allahbakhsh [a], Samad Paydar [a], Elisa Bertino [b]

[a] Ferdowsi University of Mashhad, Mashhad, Iran
[b] Purdue University, West Lafayette, IN, USA

ARTICLE INFO

Article history:
Received 17 March 2022
Received in revised form 2 September 2022
Accepted 31 March 2023
Available online xxxx

Keywords:
Data streams
Data quality
Systematic literature review
Quality framework

ABSTRACT

These days, endless streams of data are generated by various sources such as sensors, applications, users, etc. Due to possible issues in sources, such as malfunctions in sensors, platforms, or communication, the generated data might be of low quality, and this can lead to wrong outcomes for the tasks that rely on these data streams. Therefore, controlling the quality of data streams has become increasingly significant. Many approaches have been proposed for controlling the quality of data streams, and hence, various research areas have emerged in this field. To the best of our knowledge, there is no systematic literature review of research papers within this field that comprehensively reviews approaches, classifies them, and highlights the challenges.

In this paper, we present the state of the art in the area of quality control of data streams, and characterize it along four dimensions. The first dimension represents the goal of the quality analysis, which can be either quality assessment, or quality improvement. The second dimension focuses on the quality control method, which can be online, offline, or hybrid. The third dimension focuses on the quality control technique, and finally, the fourth dimension represents whether the quality control approach uses any contextual information (inherent, system, organizational, or spatiotemporal context) or not. We compare and critically review the related approaches proposed in the last two decades along these dimensions. We also discuss the open challenges and future research directions.

Contents

* Corresponding author.
E-mail addresses: mostafa.mirzaie@mail.um.ac.ir (M. Mirzaie), behkamal@um.ac.ir (B. Behkamal), allahbakhsh@um.ac.ir (M. Allahbakhsh), s-paydar@um.ac.ir (S. Paydar), bertino@cerias.purdue.edu (E. Bertino).

https://doi.org/10.1016/j.cosrev.2023.100554
1574-0137/

1. Introduction

There is a surge of mass data produced by social networks, sensor networks, and other Internet-based platforms in various domains [1]. The data generated in such ecosystems are enormous in terms of volume, velocity of generation, and variety of sources, which justify why they are referred to as Big Data [2]. Many organizations have recently shown a growing interest in processing big data, especially the data from social communities such as Twitter and Facebook. Such processing can help companies better understand their customers' interests and behaviors, directly impacting their effectiveness and even financial income [3–7]. However, data generation speed in social platforms is very high, making it challenging and almost impossible to process the incoming data using traditional data mining techniques and methods. Hence, it is essential to use data stream processing techniques since the data are continuously generated at a high rate [8,9]. Processing data streams is a complicated and challenging issue since they come from various sources with unknown or different trust and credibility levels. In addition to the general characteristics of big data, data streams have some unique features: (i) an infinite number of data elements, (ii) a high rate of data arrival, and (iii) potential changes in the data distribution [10]. Each of these can create different forms of challenges such as dealing with the tremendous size of data, fast data processing, and real-time detection of possible changes in the data distributions.

Data quality is a multifaceted concept and several definitions for this concept have been proposed.

According to ISO 8000,[1] data quality control is generally defined as "the degree to which a set of inherent characteristics of data fulfills the requirements". Based on ISO 25012,[2] the quality of a data product is "the degree to which data satisfies the requirements defined by the product-owner organization". The requirements, which assure the quality level of data, are reflected in a data quality model through its attributes, e.g., accuracy, completeness, timeliness, etc. However, these quality attributes can be redefined or customized for a particular domain. For example, Behkamal et al. [11] customized the quality attributes defined in ISO 25012 for Linked Open Data and proposed a set of factors for measuring the inherent quality of datasets.

On the other hand, new definitions of data quality have been proposed for specific domains. For example, Florian et al. [12] defined the quality control of a crowdsourced task as "the extent to which the output meets and/or exceeds the requester's expectations", or in collaborative content generation systems, Allahbakhsh et al. [13], showed that the quality of a human-generated artifact depends on the quality of its content, quality of the contributing people and quality of the venue. However, these quality definitions and models are domain-specific and do not cover all aspects of the quality of big data. Therefore, Merino et al. [14] introduced a quality model for big data and introduced contextual adequacy, temporal adequacy, and operational adequacy as important big data quality attributes. Contextual adequacy refers to "the capability of datasets to be used within the same domain". Temporal adequacy refers to the fact that "data is within an appropriate time slot for the analysis". Finally, operational adequacy means that "there are sufficient and appropriate resources available to perform the analysis".

Since low-quality data may lead to inaccurate results and decisions [15], it is vital for organizations to employ data stream quality control approaches [16]. In other words, without quality control of data streams, a proper understanding of what happens in the market would be difficult or even impossible for a business [17]. Low data quality in data streams may cause extra costs, delays in developing systems, low credibility, and more time for data reconciliation [17]. According to a report from the California Independent System Operator (CAISO), which is responsible for monitoring the electric power systems in California, about 17% of the data received suffered from quality problems [18]. Besides, IBM reported that only one in three corporate executives trust their analytic results due to the low data quality [19]. Moreover, a recent study has shown that low-quality data costs the USA three trillion dollars per year [20]. Another study shows that two-thirds of the European and American businesses could not unlock value from the data generated at a high rate [21]. All these reports highlight the importance of controlling the data quality in the domain of data streams.

Quality control for data streams, due to its importance, has been widely investigated in the literature. Because of the wide variety of the proposed quality control approaches, a systematic review of these approaches is critical to a comprehensive understanding of the data quality problem. It also helps individuals and organizations to adopt the proper technique that fits their requirement, when needed. In the current literature, including [22], quality control for streaming data is not covered, and there are still various challenges to be addressed.

The contributions of this systematic literature review (SLR) are as follows:

- To the best of our knowledge, our work is the first to systematically and comprehensively review the literature in the area of quality control approaches for data streams.
- This review analyzes approaches to data stream quality from four different viewpoints, which are quality analysis goal, quality control method, quality control techniques, and contextual information used.
- This review addresses all quality control goals including quality assessment (techniques applied to the already-generated data items to assess their quality) and quality improvement (policies, strategies, and mechanisms used to increase the chance of obtaining high-quality data). While other surveys focus on outlier detection approaches, which is one of the quality assessment methods.

1 https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-4:v1:en.
2 https://iso25000.com/index.php/en/iso-25000-standards/iso-25012.

- The current review investigates the evolution of techniques over time, reports domain applications, and presents active conferences/journals.
- The presented review classifies challenges, discusses them and explains future research directions.

The rest of the paper is organized as follows. Section 2 highlights the importance of this study by comparing it with existing review papers. Section 3 presents background about the fundamental concepts related to our literature review. The research questions and the search methodology are presented in Section 4. Section 5 proposes a classification of the various approaches. Section 6 surveys the state-of-the-art approaches in data stream quality control. Section 7 presents the results of our systematic analysis and comparison among the surveyed approaches. Challenges and suggestions for further research are provided in Section 8. Finally, we conclude the paper in Section 9.

2. Related reviews

Several articles have been published in the form of surveys, reviews, and systematic literature reviews (SLR) in the field of data streams, most of which focus on outlier detection approaches. This section compares and highlights the importance and need of the current study. Table 1 shows this comparison from various aspects, including contribution (the novelty of each work), trend analysis (Does the review analyze the approaches based on some metrics over time?), challenge analysis (Does the review analyze the existing challenges and gaps?), systematic review (Does the review select approaches systematically by a search methodology?), time span (represents the start and end time of the publications considered in each review), covered approaches (represents the focus of each review whether quality assessment approaches are considered or quality improvement or both. For quality assessment, it is specified that the focus is on outlier detection or general), and domain.

As shown in Table 1, most surveys have focused on classifying quality assessment and outlier detection methods for data streams. Li et al. [42] explained spatial data quality in the Internet of things and the usage of contextual information. Also, Ane et al. [40] proposed a classification based on input data, outlier type, and nature of method (univariate, multivariate). Moreover, Yang et al. [23] and Ayadi et al. [30] classified the approaches based on statistics, clustering, classification, artificial intelligence, and the nearest neighbor. Furthermore, a review of outlier detection, and concept drift detection for data streams is presented by Park [32]. A different classification is presented by Gupta et al. [25], that classified papers according to the data types. This study [25] considers other data types in addition to the data stream, Hence, more approaches than our work are included. Munir et al. [37] and Pang et al. [39] discussed just deep-learning-based anomaly detection approaches for data streams and other techniques have not been included.

Further, the classification given by Chen et al. has four categories, namely statistical-based, distance-based, density-based, and clustering methods [31]. Salehi et al. [35] categorized anomaly detection approaches in evolving data into three dimensions distance-based, clustering-based and model-based. The data stream techniques in the work by Shukla et al. [27] are divided into two groups based on clustering and outlier detection algorithms. Additionally, in this work [27] and a paper by Karkouch et al. [28], the characteristics of data and data quality attributes on the Internet of things have been identified for the first time. The impact of essential factors on data quality along these dimensions has been investigated, followed by a description of data quality control methods in terms of technique, scope, data stream type, and data characteristics. Rassam et al. [24] proposed an approach by which the sensor network's anomaly detection requirements are specified; these requirements include data reduction, distributed detection, online detection, correlation exploitation, and adaptive detection. Next, the methods are classified for detecting anomalies in the sensor network based on statistics, clustering, nearest neighbor, and classification methods.

There is only one review paper [38] that uses the systematic method to search and retrieve the papers, which is different from our study in terms of scope and research questions. Firstly, the study [38] aims to find the types of errors in sensor data and how they can be addressed. Then, the authors focused only on sensor data, and all related works were classified in terms of the applied techniques. Besides, the keywords used were sensor data and data quality. Finally, a total number of 57 papers were reviewed, none of which were published in 2020. However, these studies review related approaches, have some shortcomings that are covered in our systematic literature review:

- Most of the review papers are not conducted systematically. The key focus of a systematic literature review is to identify, evaluate, and summarize the research results and findings through research questions.
- However, some of the review papers propose a classification, all focusing on the technique dimension. The current systematic literature review presents a classification along four dimensions including goal, method, technique, and context-awareness level.
- Most of these studies focus on outlier detection approaches, which is one of the quality assessment methods, while our work addresses both quality control goals including quality assessment and quality improvement.
- However, none of the review papers inspect the evolution of techniques over time; instead the current review investigates the evolution of techniques over time, displays domain applications, and presents active conferences/journals.
- Some of the review studies mention the key challenges of the field, but in the presented review, challenges are categorized into three groups, i.e., source-dependent, inherent, and technique-dependent. Thus, the details of challenges in each category are discussed and future research directions are investigated.

3. Quality in data streaming

Considering the current literature in the area, and the lack of an all-encompassing view of the quality of data streams, we developed a framework to depict our notion of data quality in the context of streaming data and show how quality is dealt with in the literature. As presented in Fig. 1, this holistic view highlights the key aspects one has to face when developing quality control mechanisms.

This framework is proposed based on our previous experiences in data quality control [11–13,43–50], as well as on an extensive literature review of related areas, discussions with colleagues, and experimentation with systems and prototypes, which allowed us to identify common building blocks for the different variations in quality control approaches. Accordingly, we proposed a quality framework with two main elements of Quality Model and Quality Control Approach, each of which is described as follows.

3.1. Quality model

There are some definitions and classifications of data quality models. One of the most well-known among these proposals is by

Table 1
Comparison of review papers [23–42].

Ref.	Year	Contribution	Trend Analysis	Challenge Analysis	Systematic Review	Time Span	Covered Approaches	Domain
[23]	2010	Provides a technique-based categorization to classify existing outlier detection approaches	No	No	No	1980-2007	Quality Assessment: Outlier Detection	WSN
[24]	2013	Presents the challenges of outlier detection in WSNs and explains the requirements to design effective detection models	No	No	No	1990-2013	Quality Assessment: Outlier Detection	WSN
[25]	2014	Provides a technique-based classification for outlier detection of temporal data	No	No	No	1978-2013	Quality Assessment: Outlier Detection	General
[26]	2014	Identifies gaps and challenges of data stream mining	No	Yes	No	1993-2014	Quality Assessment: General	General
[27]	2015	Reviews different techniques of outlier detection for the data stream and their issues	No	No	No	2001-2014	Quality Assessment: Outlier Detection	General
[28]	2016	Surveys data quality approaches in the Internet of Things	No	Yes	No	2012-2016	Quality Assessment: General	IoT
[29]	2016	Discusses various approaches for outlier detection on data streams	No	No	No	1998-2015	Quality Assessment: Outlier Detection	General
[30]	2017	Describes an overview of existing outlier detection techniques specifically used for wireless sensor networks.	No	No	No	2006-2016	Quality Assessment: Outlier Detection	WSN
[31]	2017	Provides issues and challenges of detecting outliers over big data stream	No	No	No	1998-2016	Quality Assessment:	General
							Outlier Detection	
[32]	2018	Focuses on unusual patterns and behavior in the streaming data	No	No	No	1999-2016	Quality Assessment: Outlier Detection	General
[33]	2018	Focuses on unusual patterns and behavior in the streaming data	No	No	No	1999-2018	Quality Assessment: Outlier Detection	General
[34]	2018	Represents issues of outlier detection in trajectory streams	No	Yes	No	2005-2017	Quality Assessment: Outlier Detection	WSN
[35]	2018	Proposes a technique-based classification in evolving data	No	No	No	1996-2018	Quality Assessment: Outlier Detection	General
[36]	2019	Reviews challenges and approaches of outlier detection in the Internet of Things	No	No	No	2008-2018	Quality Assessment: Outlier Detection	IoT
[37]	2019	Analyses traditional and deep learning-based outlier detection approaches for streaming data	No	No	No	2001-2019	Quality Assessment: Outlier Detection	General
[38]	2020	Systematically reviews data quality approaches in the sensor data domain	No	No	Yes	1982-2019	Quality Assessment: General	WSN
[39]	2021	Reviews deep learning-based outlier detection approaches for streaming data	No	No	No	2002-2020	Quality Assessment: Outlier Detection	General
[40]	2021	Proposes a classification based on input data, outlier type, and nature of method for outlier detection approaches of time-series data	No	No	No	1999-2019	Quality Assessment: Outlier Detection	General
[41]	2021	Proposes solutions for trajectory outliers	No	No	No	1987-2020	Quality Assessment: Outlier Detection	IoT
[42]	2022	Focuses on spatial data quality in the Internet of Things	No	Yes	No	1970-2021	Quality Assessment: General – Quality Improvement	IoT
Our study	2022	Proposes a classification based on goal, method, technique, and context-awareness level used for classifying quality control approaches of data streams. Classifies new challenges and states open future directions	Yes	Yes	Yes	2000-2022	Quality Assessment: General – Quality Improvement	General

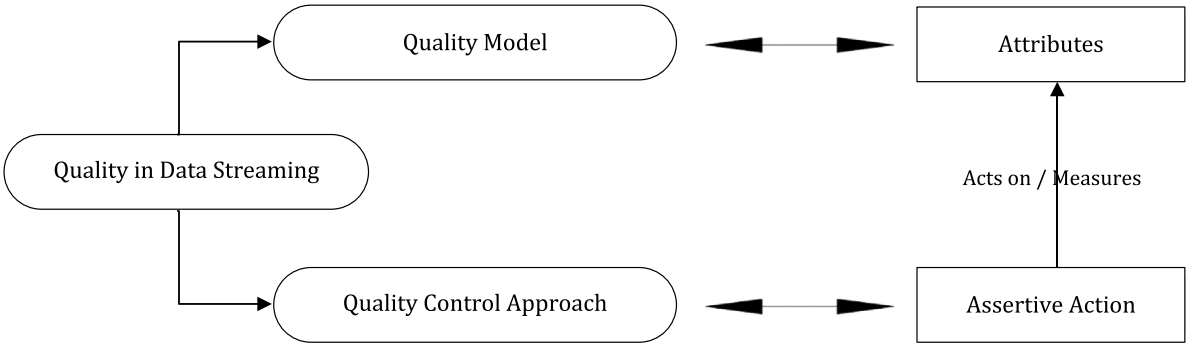

Fig. 1. Quality framework for steaming data.

ISO 25012. Based on ISO 25012 definition, "the data quality model represents the grounds where the system for assessing the quality of data products is built on. In a data quality model, the main data quality attributes that must be taken into account when assessing the properties of the intended data product are established".

The quality model captures which quality attributes have been identified so far in the literature. Quality attributes (also known as quality characteristics and quality factors) characterize properties (qualities) of the data, such as accuracy, or timeliness. Attributes are concrete if they are measurable; they are abstract if they are not directly measurable and their values are derived from concrete attributes (e.g., aggregations).

A broad spectrum of research has focused on identifying and proposing data quality attributes. The related literature [14,51–55], as well as the ISO standard namely 25012 and 8000, have identified a list of quality attributes, classified into inherent, and system dependent. These categories are not disjoint, i.e., some attributes fit into two categories. Table 2 shows the attributes, their definition, classification, and references that consider each of the attributes. All descriptions are taken from ISO 25012.

According to ISO 25012, the Inherent dimension refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions; while system-dependent quality aspect is reached and preserved within a computer system when data is used. As shown in Table 2, the quality attributes presented only in the inherent dimension, are those that have also been contemplated in the literature. In other words, the four inherent quality attributes that have attracted more attention, are more significant in the context of streaming data. These quality attributes are accuracy, completeness, validity, and timeliness.

3.2. Quality control approaches

As shown in Fig. 1, quality control approaches are actions that act on or measure quality attributes, to make sure that they conform to quality requirements. In other words, actions can be deployed that can measure the quality of generated data, based on selected quality attributes, against predefined thresholds to check if they meet the requirements or not. These approaches are called quality assessment approaches. The approaches by Rezvani et al. [65], Sadik et al. [66], and Geisler et al. [63] are examples of such approaches.

Also, there are strategies designed to maximize the chance of receiving high-quality data. These approaches are based on the analysis and understanding of system dynamics and include decisions taken at the design time of the system. These approaches are generally called quality improvement approaches. The approaches by Wang et al. [67], Hu et al. [68], and Su et al. [69] are general examples of improvement approaches.

The main focus of this study is to analyze and classify quality control approaches. So, we discuss them in more detail and from various perspectives in the following sections.

4. Research questions and search methodology

An essential step in a systematic literature review is to identify the scope and the research questions. The main research questions for our survey are presented in Table 3.

Our study was conducted on the papers available by October 2022. This study's search methodology includes two main phases: search and analysis, which are fully explained in Appendix. The systematic review's scope is first defined in the search phase, and the appropriate terms are identified. Then, the search queries are specified, and the related venues are selected. Finally, the eligibility criteria are specified to plan the data extraction. The search then returned 399 papers, 15 of which are theses. The retrieved papers are stored in a local dataset using Mendeley software, as the paper pool.

In the analysis phase, all papers are carefully examined. The abstract, the keywords, and the citations for each of the papers are re-evaluated. In other words, after reviewing the abstract and the keywords as well as applying the inclusion and exclusion criteria, a paper is selected for further review when it is considered relevant and appropriate. This is also done for citations, and if the papers' title is related to the defined scope, the papers are added to the paper pool. 58 papers are added after examining citations, leading to the inclusion of 207 papers in the paper pool for further analysis.

Furthermore, the references and authors of the papers are separately inspected. The references are selected if they are related to the research scope. Further, if other authors' publications are available on academic or social networks, the relevant papers are examined, and added to the final pool, if related. Fifteen papers were retrieved by examining the references and authors. As a result, we retrieved 92 papers that are thoroughly investigated and compared.

5. Classification

This section addresses the first research question (i.e. RQ1 in Table 3) that is related to the classification of approaches for quality control of data streams. The proposed classification is based on a comprehensive view of the problem of quality control in data streams. In order to design the classification, we first inspected all papers selected by the systematic search methodology explained in Appendix. We then extracted the key features of each approach. These features are Keywords (the keywords of each paper mentioned by the authors), Data Generation Strategy (batch or stream), Data Type (unstructured, semi-structured, or structured), Quality Attributes Focused (accuracy, completeness, etc.) Processing Method (online, offline), Architecture (centralized or distributed), Goal (detection, cleansing), Technique (distance-based, etc.), Processing Level (user-level or

Table 2
Quality attributes considered in the literature.

#	Attribute	Description	Classification		Cited in data stream field
			Inherent	System dependent	
1	Accuracy	The level to which data value has features that correctly represent the true value of the considered attribute of a concept or event in a specific context of use.	✔		[56–62]
2	Completeness	The level to which data associated with an entity is not null and has values for all regarded features and related entity instances in a specific context of use.	✔		[56–64]
3	Consistency	The level to which data has features that are free from conflict and are coherent with other data values in a specific context of use.	✔		
4	Validity	The level to which data has features that are expected as true and acceptable by users in a specific context of use.	✔		[59,63,64]
5	Timeliness	The level to which data has features that are of the right age in a specific context of use.	✔		[56–64]
6	Accessibility	The level to which data can be accessed in a specific context of use.	✔	✔	
7	Compliance	The level to which data has features that adhere to standards in force and similar rules relating to data quality in a specific context of use.	✔	✔	
8	Confidentiality	The level to which data has features that ensure that it is only accessible by authorized users in a specific context of use.	✔	✔	
9	Efficiency	The level to which data has features that can be processed and supply the expected levels of performance in a specific context of use.	✔	✔	
10	Traceability	The level to which data has features can track the access to the data and of any changes made to the data in a specific context of use.	✔	✔	
11	Precision	The level to which data has features that are exact or that provide discrimination in a specific context of use.	✔	✔	
12	Understandability	The level to which data has features that enable it to be explained by users, and are expressed in appropriate languages, symbols, and units in a specific context of use.	✔	✔	
13	Availability	The level to which data has features that enable it to be retrieved by permitted users and/or applications in a specific context of use.		✔	
14	Portability	The level to which data has features that enable it to be set up, replaced, or moved from one device to another preserving the existing quality in a specific context of use.		✔	
15	Recoverability	The level to which data has features that enable it to keep and preserve a specified level of operations and quality, even in the failure time, in a specific context of use.		✔	

Table 3
Research questions.

#	Question	Rationale	Where will be answered?
RQ1	How can approaches for data quality control in data streams be categorized?	The answer will help understanding the classification of proposed methods for the data stream quality.	Section 5
RQ2	How has the trend of quality control related activities evolved over time?	The answer will help researchers and practitioners to have a comprehensive understanding of the shifts in the state of the art over the past two decades.	Sections 6 and 7
RQ3	What are the main research gaps and challenges in the domain of quality control in data streams?	The answer will highlight the research gaps, and help researchers to identify possible future research directions.	Section 8

source-level), Dataset used, and Domain applications. After integrating the data values of each feature, we compared, and then classified them into a category. For example, distance-based, density-based, and similarity-based approaches demonstrate a unified concept and can be grouped into a single category called distance-based techniques. In some cases, after a deeper analysis, the feature values were updated. For instance, some approaches use hybrid (both offline and online) methods which were not in the first classification. We redesigned the classification with the features by replacement again and again. To have a legible and understandable classification, we followed the idea behind the feature subset selection method which is a well-known method in data mining tasks [70]. There are three basic heuristic methods in feature subset selection including forward selection, backward elimination, and decision tree induction. In forward selection, the best features are added one by one to the final selection pool, while in backward elimination all features are included in the selection and in each step, one irrelevant feature is removed.

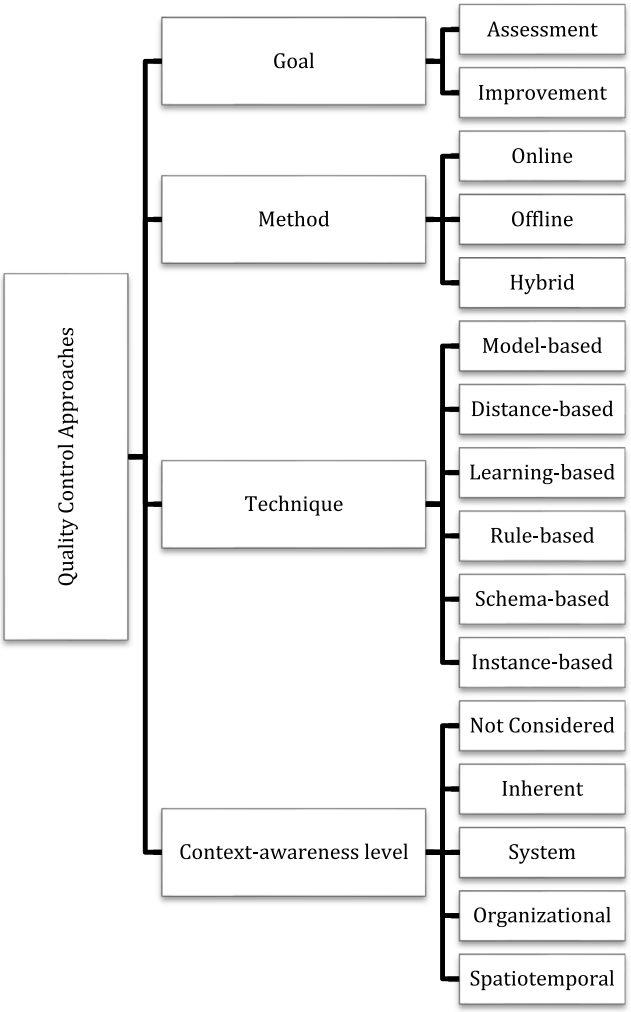

Fig. 2. Classification of approaches for the quality control of data streams.

In decision tree induction, one of the decision tree algorithms is used to construct a tree based on the most significant and distinctive features. We followed the forward selection method, and selected those features that were most significant and distinctive. Finally, the extracted classification was compared to the quality control studies that provide a classification in other domains, such as the approach proposed by Batini et al. [71] in terms of naming and structure to have a unified classification. For instance, after comparing the naming of each dimension with similar studies, the data values of the Goal dimension are replaced by assessment and improvement.

In what follows, we explain each dimension in more detail.

5.1. Goal

As Fig. 2 shows, the first dimension of the classification is the goal of quality analysis, which can be classified into quality assessment, and improvement, according to Batini et al. [72] and Florian et al. [12].

- ***Quality Assessment*** includes the techniques applied to the already-generated data items to assess their quality [72]. Once the data is generated, the quality assessment process detects the data with low quality. Most approaches (66 out of 92) have quality assessment as their goal [56–66,73–127]. Some approaches assess the quality of the data streams using specific quality factors (e.g., accuracy, completeness, timeliness) to report the data quality to the data consumer [57–59,61,63,64]. Regardless of the quality factors, other approaches use specific techniques such as clustering methods to monitor data"- streams and detect data with low quality.
- ***Quality Improvement*** refers to the policies, strategies, and techniques used to increase the chance of obtaining high-quality data [71]. Selecting credible and reliable sources is the first step to obtain high-quality data. Then cleaning techniques can be considered to improve the quality of data streams. For example, selecting highly reliable and rechargeable sensors in a sensor network can result in receiving data with higher quality levels. However, such sensors might be expensive or inaccessible for a sensor network application. Hence, other quality control strategies as well as cleaning techniques should be adopted. Fewer approaches (26 out of 92) are assigned to this category than the assessment goal [67–69,128–150].

5.2. Method

As shown in the classification, the second dimension represents the data processing method which can be either online, offline, or hybrid.

- ***Online Processing*** refers to real-time quality control of the data streams. Twenty-six approaches use online processing to control data stream quality. Data streams are generated at different and high arrival rates. Therefore, assessing the quality of data in such scenarios, in which it is impossible to store entire data, is arduous. To address such a challenge, online processing methods typically consider a window of data and perform real-time quality control on this window. On the other hand, not having access to all the data can decrease the accuracy of the assessment, but this seems to be the only viable solution when dealing with a tremendous volume of real-time data.
- ***Offline Processing*** refers to methods that control the streaming data quality after data is stored in secondary storage and has reached the desired volume. In these scenarios, due to the fact that data is entirely stored, as well as historical data is used, a comprehensive quality control process can be performed. Therefore, the evaluation's accuracy is increased, the right decisions are made, and more precise results are obtained.
- ***Hybrid Processing*** refers to the methods that obtain necessary information from the stored data offline and control the streaming data quality online. This method is in fact a combination of both online and offline methods. On the one hand, it uses historical data to gain a comprehensive view and then uses this view to apply a more precise control on the quality of real-time streams.

5.3. Technique

The third dimension of our proposed classification refers to the techniques applied by each approach. The techniques are classified into seven major following groups.

- ***Model-based techniques*** build a statistical model using previously observed data to predict the data quality. These techniques use historical data. Thus, such approaches' processing method is either offline or hybrid. Consider the pair $(\boldsymbol{S}, \boldsymbol{D})$, where $\boldsymbol{S}$ is the sample space, and $\boldsymbol{D}$ denotes the data distribution on sample space $\boldsymbol{S}$. Also, let $\boldsymbol{S}(\boldsymbol{x}_{t-1})$ indicate the data value $\boldsymbol{x}$ observed at time $\boldsymbol{t-1}$, and let $\boldsymbol{D}$

(y_t) refer to the data value y obtained at time t from the same data distribution, i.e. D. To detect low-quality data, x_{t-1} and y_t are compared. If their values differ more than a given threshold **T**, y_t is considered a low-quality data item. Although model-based techniques are the first option for identifying anomalies, it should be noted that most of these methods can only detect anomalies locally, unless they use contextual information.

- ***Distance-based techniques*** (also known as nearest neighbor-based approaches) use distance measures, such as the Euclidean distance, to measure the distance between different data values to identify low-quality data. Let f be a feature of interest; consider $f(x_{t-1})$ and $f(x_t)$ where x_{t-1} and x_t represent the data value of f respectively at time $t-1$ and time t. In order to assess the quality of recent data values in such methods, the distance between two recent data values, i.e. x_t and x_{t-1}, is calculated. If the distance is greater than a given threshold **T**, x_t is considered a low-quality data item. In some approaches, after calculating the distance, the data value's density may be estimated for detecting the outliers, which are also included in the distance-based methods. The basic idea of this method is that data with more neighbors is considered normal, whereas data with fewer neighbors is considered anomalous.
- ***Learning-based techniques*** include clustering, classification, neural networks, and machine-learning-based algorithms to evaluate or improve the quality of the data streams. For example, in the clustering methods, similar data are located in the same cluster, and a sparse cluster is considered an outlier. Assume that C_1 and C_2 are predefined clusters with cluster centers c_1 and c_2, correspondingly. Moreover, x_t denotes the data value at time t. In order to assign x_t to the correct cluster, the distance between x_t and both c_1 and c_2 is calculated, and x_t is assigned to the cluster whose distance of cluster center with x_t is the least (e.g., C_1). Finally, if the number of the data values in C_1 is less than the threshold **T**, all data values, including x_t are considered outliers; otherwise, they are deemed to be normal.
- ***Rule-based techniques*** rely on a set of rules to find low-quality data items. These rules are generally defined by domain experts. Assume that x_t is the data value at time t and R_1, and R_2 are the defined rules. R_1 says "If x is less than 10, then delete it" and R_2 says "if x is between 10 and 15, then replace it with the mean" (mean value is the mean of all observed data values at the time t) and otherwise, store the value. Therefore, decisions can be made based on existing rules when new data are observed.
- ***Schema-based techniques*** leverage the constraints defined in the data schema to control data quality (e.g. data value should not be negative). Let f be a feature of interest. Consider $f(x_t)$ where x_t denotes the data value x of f at time t and $Cons\ f$ = {type: integer and range between 0 and 100}. Consequently, if x_t is equal to 200, it can be concluded that an outlier is detected.
- ***Instance-based techniques*** are used to obtain constraints when the data schema is not available. The constraints in these methods can be obtained by analyzing the available data. Let f be a feature of interest. Consider $f(x_t)$ where x_t denotes the data value x of f at time t and $Cons_t\ f$ is the schema constraint of the f at time t and $Cons_t\ f$ = {type: integer and range between 0 and 20}. As a result, it can be deduced that low-quality data are detected for such cases that x_t is negative (e.g. -10). $Cons_t\ f$ can also be updated after a while with strategies adopted by researchers.

5.4. Context-awareness level

The last dimension of our classification refers to the context in which data is collected. Any relevant information which can be extracted from the environment in both active and passive ways is called contextual information. The benefits of embedding contextual information into an application have been already evidenced in different applications [151,152]. Here, we define four types of contextual information affecting the quality of data streams as follows.

- ***Inherent Context*** includes all information about data values and the intrinsic characteristics of the data, such as data schema and constraints. Consider for example a number of sensors monitoring the vital signs of a patient such as blood pressure. To assess the quality of the received data, information items such as the thresholds for minimum and maximum blood pressure can be retrieved from the data schema or other similar sources.
- ***System Context*** describes all information about data sources and relevant infrastructure. This includes features and limitations of hardware or software such as power consumption, device type, and memory capacity. Knowing these features and limitations of the system can help in assessing data quality. Reliability of data could be measured, for example, by the level of accuracy, precision, or security in using the technology (e.g., biometric authentication) to extract or process contextual information [152].
- ***Organizational Context*** defines features, limitations, specifications, and any information that is provided directly by the relevant organizations, or indirectly, by other sources such as websites, ontologies, datasets, or organizational regulations. Sometimes organizational constraints may impact the quality of data. For instance, poorly designed regulations or a shortage in the budgets can lead to limitations in the data assessment process, and consequently, can result in low-quality data. Moreover, information about employee roles and skills also can affect how the data is interpreted, as another form of contextual information.
- ***Spatiotemporal Context*** points to any temporal or spatial information related to the data value. There are attributes such as time and location that determine when and where the data has been generated. These attributes can be used to get other useful information, such as weather conditions, nearby resources, temperature, and humidity. These data can be extracted from external sources like ontologies, websites, and other online services. Consider a set of sensors are responsible for monitoring the temperature of a forest, and a specific sensor reports a high temperature. Having access to an additional data source, such as other sensor reports, weather conditions, or local services that provide incidents reported by citizens, can be highly beneficial in assessing the accuracy of the data reported by that sensor.

6. State of the art

In this section, we first analyze the related literature based on our proposed classification. The results of this analysis are presented in Tables 4 to 7. Each approach is analyzed and the goal of the approach, the method, technique, and context-awareness level are identified. We next review all approaches listed in Tables 4 to 7 in detail in chronological order.

Moreover, we study how the quality control approaches have evolved over time during these two decades. In what follows, we have divided the past 20 years into four periods based on the emergence of new techniques or applications and studied the related works accordingly.

Table 4
Comparison of approaches (2000–2011).

No	Ref.	Year	Paper type	Goal		Method			Technique	Context-awareness level
				Assessment	Improvement	Offline	Online	Hybrid		
1	[114]	2011	Conference	✔			✔		Learning-based	System
2	[63]	2011	Conference	✔			✔		Instance-based	Not considered
3	[66]	2011	Conference	✔				✔	Distance-based	Not considered
4	[74]	2010	Journal	✔			✔		Distance-based	Not considered
5	[82]	2010	Thesis	✔			✔		Distance-based	Not considered
6	[58]	2009	Journal	✔			✔		Instance-based	Not considered
7	[56]	2007	Conference	✔			✔		Instance-based	Not considered
8	[57]	2007	Conference	✔			✔		Instance-based	Not considered
9	[141]	2007	Journal		✔		✔		Distance-based	Not considered
10	[128]	2006	Conference		✔			✔	Learning-based	Not considered
11	[101]	2006	Conference	✔		✔			Learning-based	Not considered
12	[102]	2005	Conference	✔		✔			Distance-based	Not considered

Table 5
Comparison of approaches (2012–2015).

No	Ref.	Year	Paper type	Goal		Method			Technique	Context-awareness level
				Assessment	Improvement	Offline	Online	Hybrid		
1	[60]	2015	Journal	✔				✔	Model-based	Not considered
2	[65]	2015	Conference	✔				✔	Model-based	Not considered
3	[85]	2015	Conference	✔			✔		Distance-based	Not considered
4	[86]	2015	Journal	✔				✔	Learning-based	Spatiotemporal
5	[131]	2015	Journal		✔			✔	Model-based	Not considered
6	[136]	2015	Conference		✔	✔			Model-based	Not considered
7	[142]	2015	Conference		✔	✔			Distance-based	Not considered
8	[150]	2014	Chapter		✔		✔		Distance-based	Inherent
9	[87]	2014	Conference	✔				✔	Learning-based	Spatiotemporal
10	[76]	2014	Journal	✔			✔		Learning-based	Spatiotemporal
11	[89]	2013	Thesis	✔			✔		Distance-based	Spatiotemporal
12	[88]	2013	Thesis	✔			✔		Instance-based	Spatiotemporal
13	[77]	2013	Journal	✔				✔	Model-based	Not considered
14	[84]	2013	Thesis	✔			✔		Distance-based	Not considered
15	[73]	2012	Journal	✔			✔		Distance-based	Spatiotemporal
16	[130]	2012	Conference		✔			✔	Model-based	Not considered
17	[115]	2012	Journal	✔			✔		Distance-based	Spatiotemporal

Table 6
Comparison of approaches (2016–2017).

No	Ref.	Year	Paper type	Goal		Method			Technique	Context-awareness level
				Assessment	Improvement	Offline	Online	Hybrid		
1	[148]	2017	Journal		✔			✔	Learning-based	Not considered
2	[78]	2017	Conference	✔				✔	Learning-based	Not considered
3	[79]	2017	Conference	✔				✔	Learning-based	Not considered
4	[80]	2017	Journal	✔				✔	Learning-based	Not considered
5	[132]	2017	Conference		✔			✔	Rule-based	Not considered
6	[133]	2017	Conference		✔			✔	Model-based	Not considered
7	[137]	2017	Conference		✔			✔	Model-based	Not considered
8	[138]	2017	Conference		✔	✔			Model-based	Not considered
9	[140]	2017	Conference		✔	✔			Learning-based	Not considered
10	[143]	2017	Journal		✔			✔	Learning-based	Not considered
11	[134]	2017	Conference				✔		Model-based	Not considered
12	[64]	2016	Journal	✔			✔		Instance-based	Not considered
13	[59]	2016	Conference	✔		✔			Instance-based	Not considered
14	[90]	2016	Thesis	✔				✔	Schema-based	Not considered
15	[149]	2016	Journal					✔	Model-based	Not considered
16	[81]	2016	Thesis	✔				✔	Learning-based	Not considered
17	[83]	2016	Conference	✔			✔		Distance-based	Not considered
18	[135]	2016	Conference		✔	✔			Model-based	Not considered
19	[129]	2016	Journal		✔	✔			Rule-based	Not considered
20	[116]	2016	Journal	✔			✔		Learning-based	Not considered
21	[117]	2016	Conference	✔			✔		Distance-based	Not considered

6.1. 2000 to 2011: Distance-based, learning-based, and instance-based techniques

In the first decade of the century, learning-based, distance-based, and instance-based approaches, which are listed in Table 4, gained much attention.

The first approach, which is an offline distance-based method, was proposed by Du et al. [102]. The authors try to determine the degree of localization of data values after all data values are received and stored. The computed localization degree of each data value is compared with a predefined threshold to find low-quality data values in the data stream. The same research group then proposed a hybrid learning-based method [128], in which each sensor performs clustering on its data and then sends the result to its cluster head in the sensor network. The cluster head then aggregates the data received from the sensors and sends

Table 7
Comparison of approaches (2018-Present)..

No	Ref.	Year	Paper type	Goal		Method			Technique	Context-awareness level
				Assessment	Improvement	Offline	Online	Hybrid		
1	[118]	2022	Journal	✔			✔		Learning-based	Not considered
2	[68]	2022	Journal		✔		✔		Model-based	Not considered
3	[119]	2022	Conference	✔			✔		Learning-based	Not considered
4	[120]	2021	Conference	✔			✔		Distance-based	Not considered
5	[121]	2021	Conference	✔			✔		Distance-based	Not considered
6	[111]	2021	Journal	✔			✔		Distance-based	Not considered
7	[112]	2021	Journal	✔			✔		Distance-based	Not considered
8	[113]	2021	Journal	✔			✔		Distance-based	Not considered
9	[122]	2021	Conference	✔				✔	Learning-based	Not considered
10	[123]	2021	Journal	✔			✔		Learning-based	Not considered
11	[124]	2021	Conference	✔			✔		Learning-based	Not considered
12	[144]	2020	Journal		✔			✔	Distance-based	Not considered
13	[67]	2020	Journal		✔			✔	Model-based	Not considered
14	[145]	2020	Journal		✔	✔			Distance-based	Spatiotemporal
15	[98]	2020	Journal	✔			✔		Distance-based	Not considered
16	[99]	2020	Journal	✔				✔	Model-based	Not considered
17	[91]	2019	Journal	✔			✔		Distance-based	Spatiotemporal
18	[92]	2019	Journal	✔				✔	Learning-based	Not considered
19	[94]	2019	Journal	✔		✔			Learning-based	Not considered
20	[95]	2019	Conference	✔		✔			Model-based	Not considered
21	[97]	2019	Conference	✔			✔		Distance-based	Not considered
22	[69]	2019	Journal					✔	Learning-based	Not considered
23	[106]	2019	Journal	✔		✔			Distance-based	Not considered
24	[107]	2019	Conference		✔			✔	Model-based	Not considered
25	[108]	2019	Journal		✔		✔		Distance-based	Not considered
26	[109]	2019	Journal		✔	✔			Learning-based	Not considered
27	[110]	2019	Journal		✔	✔			Learning-based	Not considered
28	[62]	2019	Conference		✔	✔			Instance-based	Not considered
29	[125]	2019	Conference		✔		✔		Learning-based	Not considered
30	[126]	2019	Conference		✔		✔		Learning-based	Not considered
31	[127]	2019	Conference		✔		✔		Learning-based	Not considered
32	[146]	2018	Journal	✔				✔	Model-based	Not considered
33	[103]	2018	Journal		✔	✔			Learning-based	Not considered
34	[104]	2018	Conference		✔			✔	Learning-based	Not considered
35	[105]	2018	Chapter		✔		✔		Distance-based	Spatiotemporal
36	[139]	2018	Journal	✔				✔	Model-based	Not considered
37	[61]	2018	Journal		✔		✔		Instance-based	Not considered
38	[75]	2018	Journal		✔		✔		Learning-based	Spatiotemporal
39	[147]	2018	Journal	✔			✔		Distance-based	Not considered
40	[93]	2018	Journal		✔	✔			Distance-based	Spatiotemporal
41	[96]	2018	Conference		✔		✔		Distance-based	Spatiotemporal
42	[100]	2018	Journal		✔		✔		Distance-based	Not considered

the resulting cluster information back to its cluster sensors so that the global model provides each sensor with a global view of the clusters. Finally, each cluster that has an intra-distance, i.e., the distance between its members, higher than a predefined threshold is considered an outlier, and the sensors with which the cluster is associated are labeled as bad sensors. Another learning-based technique assesses the streaming data quality in an offline manner [101]. For the training phase, Bayesian belief networks are used to classify data. If the data value is in the class range specified in the test phase, it is considered normal. Otherwise, it is considered an outlier.

The need of improving the data stream quality was first pointed out by Basu and Meckesheimer [141]. Similar to the approach by Du et al. this approach also uses a distance-based technique to improve the quality of data streams. Firstly, the median of the data values received in the data stream is calculated. Secondly, the new incoming data is compared with the median. If its distance from the median is higher than a specific threshold, the new data is considered an outlier. Two median values are computed in this approach: one side median and two side median. The one-side median is calculated separately from each k latest data value in the data stream. The two-side median is computed over a window of $2k$ latest data values from the two sides of the data stream. Finally, the two medians are summed up, and the sum is considered the final median. Dealing with k outliers in a row is the reason for computing two median values.

Klein introduced an approach based on the quality window concept, which implies that data quality evaluation must be performed at the window level [56]. Furthermore, considering accuracy, completeness, timeliness, and confidence as the data quality attributes, they propose an instance-based technique. First, each data item in the current window is evaluated based on these three dimensions. Then all the results are aggregated based on timestamp and data attributes. Finally, the measured values are used to analyze the data and compare it with the corresponding thresholds to determine whether the data has the minimum required quality. Also, Klein et al. extended the approach and proposed the notion of the quality window in another work [57]. They argued that data quality analysis based on a single window is not sufficient, and multiple windows need to be simultaneously included in the quality evaluation.

Klein and Lehner answered the problem of suitable quality window size by proposing to reduce the window size as soon as low-quality data is observed [58]. In this paper, four functions are introduced to identify the suspicious data values and to decide whether to decrease or increase the window size by comparing the data value with the given threshold. Geisler et al. proposed an ontology-based framework for evaluating the quality of data streams [63,64]. The main idea is to use a threefold concept (Query-based DQ, Content-based DQ, and Application-based DQ) that enables flexible data quality control. In Query-based data quality, query expressions are responsible for assessing data

quality factors. In the Content-based notion, the quality of data streams is evaluated by semantic rules, and finally, in Application-based DQ, the user can submit their own function via a user interface to assess the quality of data streams.

Hill and Minsker proposed a method that aims to detect and improve anomalous data by the nearest cluster, single-layer linear network, or multilayer perceptron methods [74]. In these algorithms, the specific volume of data is used for constructing a model and predicting new data. Then, as soon as the new data is received, it is compared with the predicted value and the threshold. If it is outside this range, data is considered to be anomalous, otherwise normal. Next, the data value must be updated to enable further data prediction. The authors discuss two strategies for dealing with anomalous data. In the first strategy, called Anomaly Detection (AD), the identified anomalous data is labeled and added to the dataset. In the second strategy, called Anomaly Detection and Mitigation (ADAM), the anomalous data is replaced by the predicted one, and the cleaned data is added to the dataset. Then it is discussed that the ADAM strategy has higher precision than the AD strategy.

Moreover, Rogers proposed another distance-based method to detect the outliers [82]. By using the Euclidean distance, the number of data close to each data value is determined. If this number is less than a certain threshold, that data is classified as an outlier. Another distance-based method is proposed by Sadik and Gruenwald [66], in which two concepts of local deviation and global deviation are introduced. The purpose of local deviation is to check the new data value with the average of recent data items. On the other hand, the new data value is compared with the average of data values for global deviation. If the difference between the new data value and the local or global deviation is more than three times the variance, then that data is an outlier.

Junghans et al. [114] first defined quality factors and then solved the quality problem of data streams by proposing an optimization solution in order to maximize quality factors with respect to given resource constraints,

Reviewing the approaches proposed in the first decade, we observed that learning-based, instance-based, and distance-based techniques have been devoted to considerable attention. On the other hand, all studied instance-based methods were online.

6.2. 2012 to 2015: Emergence of model-based techniques

As presented in Table 5, in this period, both model-based and distance-based techniques attracted much attention. First, we review the model-based approaches, and then we study other related techniques.

A model-based technique was first introduced by Zhao and Ng in the domain of object tracking [130]. First, a baseline algorithm is introduced that constructs a model, based on the current position of the dynamic objects, and future locations are predicted. Then, it is determined that if the data is corrupted, the prediction process encounters an error. Furthermore, a radius notion is also considered to determine the radius of the object's motion. As a result, the process of predicting the next location of each object is improved. In their paper, Zhao and Ng mentioned that determining the correct radius is challenging because if a large radius is selected, the prediction of the next location is complex, and if a small radius is used, it may not be possible to make an accurate prediction.

Rassam et al. proposed an approach [75], in which initially, each sensor builds its model by running a one-class principal component classifier (OCPCC) algorithm locally. Then, the model is sent to the cluster head (CH), and the CH aggregates sensor models to obtain the global normal model (GNM). After that, it sends the GNM back to the sensors to compare the predicted value to the real one. Zhang et al. have used local and global models, where the model is based on the SVM algorithm [77]. However, unlike the former approach, the model is sent from each sensor to its neighbors. The approach proposed by Kerchove and Van Dooren [153] has been developed by Rezvani et al. [65] for online processing. In [153], the IF (Iterative Filtering) algorithm is used to model data. This algorithm predicts new data based on a data model. Since IF-based algorithms use all data values to improve prediction accuracy, they are suitable for static data. Notwithstanding, Rezvani and colleagues also applied this method to data streams [65]. In order to achieve this goal, the IF algorithm is used in the first data window to construct the model. Then, if the new data differs from the measured variance, it is an outlier. Also, they argue that if the number of outliers exceeds the number of normal data, the model is no longer valid, and it needs to be updated.

Fagúndez et al. proposed a framework in which the data sources are the sensors in body sensor networks [60]. The framework has three main components. The first component is the Monitoring component, in which the requirements and quality parameters are defined. These requirements are sent to the second component, which is the Middleware component, and managed the requirements after receiving data from different sensors. The Data Quality Manager is the third component; it is responsible for assessing data quality and analyzing the measured values by examining relevant historical data. If the measured values do not match the minimum and maximum historical values, alarms are generated.

Gill and Lee introduced another model-based approach to detect low-quality data [131]. First, the data value with a recognizable error is determined. A statistical model of the existing data is then created, which is used to identify the noisy data. Different algorithms are used to construct the model, and the most accurate one is chosen. Furthermore, these authors in another work proposed a method in which the model and the data received from each sensor include other types of information, such as wind speed and the distance of buses with specified locations [136].

Iyer in his thesis, presented a cleaning algorithm for wireless sensor networks [88]. In his model, when a sensor determines a corrupted or lost data item, a set of k sensors associated with that sensor is chosen, and the final value is obtained from the values retrieved from those sensors.

The approaches proposed by Hayes et al. in [86,87] include two main components: (i) Content Detection and (ii) Context Detection. In Content Detection, each sensor uses its historical data to generate the corresponding regression model, and then the data that is far from the model is considered a content anomaly. Context Detection has two tasks: clustering and profiling the sensors, and comparing the content anomaly with the average value of the sensor group. The k-means algorithm is used for clustering the sensors, and the clustering parameters are location, year, time, date, and weather phenomena. If the content anomaly is far from the average of each cluster, it is detected as a contextual anomaly. Zhang et al. introduced a method in which a predicted value is derived from the mean of the data produced by the sensors correlated with the corresponding sensor [150]. Then, the newly generated data is compared with the predicted value. If the difference is less than a threshold, the data is correct, otherwise, the predicted value is considered to be the improved value.

In some attempts, principal component analysis (PCA)-based methods are used in order to detect the outliers of each sensor [75,76]. In the first approach [75], initially, each sensor builds its model by running a PCA algorithm locally. Then, the model is sent to the cluster head, and the cluster head combines local models to build a global model. After that, it sends the model back to

the sensors, so each sensor updates its model, and when the new data comes, if the distance of the data from the predicted value is higher than the threshold, that data is considered an outlier. In the second approach [76], only a local model is provided, and no global and distributed approach is employed.

A spatiotemporal-based approach is proposed by Zhang et al. [73]. First, a temporal-based method (TOD) is designed for detecting outliers, in which each sensor analyzes its time-series data, and predicts the next value. If the newly generated data is far from the predicted value, it is considered an outlier. Then, the data is analyzed for spatial analysis (SOD), where the neighbors' sensor data is examined, a model is constructed, and eventually, the new value is predicted. Next, in the third step, in which the two previous methods are combined (TSOD), each sensor's data is initially investigated by each sensor by mean of a temporal analysis. Subsequently, by receiving data from its neighbors, each sensor performs the spatial analysis, and ultimately, the outlier is detected.

Another spatiotemporal distance-based approach is by Alessia et al. [115]. The idea relies on a rough set theoretic representation of the anomaly set, in which a rough set approach is described to detect outliers in a spatiotemporal dataset.

In another thesis conducted by Pumpichet [89], a method for cleaning the data stream of mobility sensors is proposed. As far as we know, this is the first time that the concept of the virtual sensor was introduced. Since sensors are moving and the location cannot be accurately predicted, virtual static sensors clean the actual sensors' missing values. In order to do so, the weighted average of actual sensor values is calculated, and the data value of each virtual sensor is updated so that the weight of the data produced is higher by the sensors near the virtual sensors. Then, with a predictive model, the new virtual sensor data is predicted. If the base station (stations that act as a gateway between sensor nodes and the end user) determines that the actual sensor's data has an error, based on that sensor's timestamp and spatial range, the corresponding virtual sensor's predicted value is considered the cleaned data.

An algorithm called Orion is proposed by Sadik et al. The algorithm first calculates the similarity between data values and then computes the stream density and k-distance criteria for each data [83,84]. The stream density is the number of stream neighbors that is obtained by the k-nearest neighbor method. Next, Orion uses a clustering algorithm to cluster the data. The clusters are divided into three groups: small, average, and large. Data that belong to the small density and large k-distance clusters are considered an outlier. Xiang et al. introduced another distance-based approach in which the data space is split into several grids [85]. The information of each grid, including the number of data items, the mean, the variance, and the time, is updated. After the update process, each grid's density is evaluated, and if the density is less than a threshold, then the grid is said to be an abnormal grid, and the data contained in an abnormal grid is considered an outlier.

Furthermore, in another distance-based approach, new data is compared to all previous data stored in a buffer [142]. If the distance is greater than a defined threshold, the value is classified as an outlier and is passed to the density calculation stage. At this stage, the data density is computed with the local outlier factor (LOF) formulas, and if it exceeds the threshold, the data value will be deleted from the buffer.

6.3. 2016 to 2017: Dominance of learning-based techniques

Learning-based techniques gained much attention in this period, due to the emergence and evolution of big data processing tools. These approaches are listed in Table 6.

Xie and Chen proposed a method to detect bad sensors; i.e. sensors that produce outliers more frequently [148]. Firstly, eigenvectors and eigenvalues are obtained by the PCA algorithm, and then data with a low eigenvalue is considered an outlier. Next, by employing the Bayesian Network, the relationship between the sensors is identified and bad sensors are identified. In another learning-based method proposed in [81], first, an adaptive algorithm called SOStream is presented. Then, an online spatiotemporal clustering algorithm, called ASMM, is proposed, which is able to dynamically adjust its clustering structure based on the incoming data. In this clustering algorithm, when new data is observed, its Euclidean distance from the centroids of the existing clusters is computed, and the closest one is chosen. If the selected centroids' distance is less than the desired threshold, the data is transferred to that cluster and otherwise returned to the input buffer. If the number of data transfers to the buffer exceeds the threshold, then that data is considered an outlier.

Furthermore, a clustering approach is proposed by Zaarour et al. which has six components [78]. The first component, called System Adapter, is responsible for coordinating the other components. The Event Extractor is the second component and is responsible for reading incoming events, setting a timestamp for each data, and adding metadata. The third component, called Central Splitter, distributes the data to the existing nodes to enhance scalability. Next, the k-means clustering algorithm is used for grouping the data and finding the cluster center. The Markov model is then used to detect outliers, where the data outside of each cluster is considered an outlier. Finally, the Ordering Operator component merges the collected data and sends it to the base station.

Similar to the work by Zaarour et al. in [78], the technique proposed by Jankov et al. uses the k-means clustering and Markov model to detect anomalies [79]. In this approach, the input data is in RDF (Resource Description Framework) triple format, and must first be parsed and then fed to the window. Finally, the output must also be transformed into a triple format. Another clustering-based approach is presented in [80], where the received signal is first converted into a data format and then, the k-means algorithm clusters the data, and the HMM and PSO algorithms are used to optimize the parameters, and ultimately, new data is predicted.

Liu et al. presented a learning-based cleaning method that consists of three parts: k-means clustering, anomaly detection, and cleaning [140]. First, the k-means algorithm is used to cluster data in distinct clusters. In each cluster, the max and min values are introduced as the cluster boundary. Then, the data outside this range is classified as an outlier. Finally, the outlier data is replaced by the minimum or average values. Besides, Pullabhotla and Supreethi proposed a hybrid method for data preprocessing, in which the k-nearest neighbor algorithm runs on historical data and is applied to a window of a data stream [143]. After cleaning the data stream by the Multivariate Singular Spectrum Analysis (MSSA) algorithm, the model is updated adaptively with cleaned data to make the next predictions more accurate.

Furthermore, Kumar et al. [116] proposed an online visual assessment that uses a cluster heat map to visualize anomalies in evolving data streams. To do so, in each data window, a clustering algorithm determines which data points are normal m which one is far from the main cluster.

Lei et al. [117] proposed a framework for efficient shared processing of a huge number of distance-based anomaly detection requests over sliding window streams. In this approach, a multi-query outlier detection workload is transformed into a single-query problem. Hence, in addition to decreasing parameter settings over the sliding window, less computational load is employed.

A model-based approach for assessing the quality of data in the health domain is proposed by Serhani [59]. In their approach, after data collection, data quality is evaluated before and after the preprocessing stage. Quality attributes such as accuracy, validity, timeliness, and completeness are evaluated before the pre-processing phase. Then, the transformation, and filtering techniques are applied, and again, the accuracy, correctness, and completeness of the data are re-evaluated to determine the degree of quality improvement.

Another model-based approach is by Zhang et al. [134] in which time-series data points with truth labels are modeled using auto-regression and auto-aggressive models. Then changes in data can be detected by the models and candidate bad data are evaluated and final outliers are repaired. Another quality improvement approach is proposed by Zhang et al. [135]. In this paper, the authors model the likelihood of a repair by tracking its speed changes. Under the principle that changes speed should not change greatly in a time period, the approach detects and cleans minimum changes.

Another method for detecting quality issues and then improving quality in sensor data is proposed by Lei et al. [149]. Firstly, each sensor analyzes time series data, and one that is not similar to the majority is marked as an outlier. On the other hand, outlier data may be identified as normal data when compared with other sensors' data. Each sensor compares its neighbors' data with its own to detect the event outliers, and if a conflict is observed, the data is tagged. Next, the data is sent to the base station, where after analyzing the correlation between all sensor data, a smoothing process is performed. Finally, the cleaned data is transmitted to the server for further processing phases.

Moreover, Liu et al. have provided a framework for assessing and cleaning Chinese electricity data [133]. The framework has three main modules: (i) a data collection module where data is accumulated from sensors, power grids, and databases; (ii) a storage module that is designed to store different data types, and (iii) a calculation module in which historical and real-time data are processed. In addition to data processing, this module is also responsible for assessing the quality of the data. Furthermore, the approach proposed by Sibai et al. in [137] is also a model-based approach in which an abstract framework is provided to examine wireless sensor data quality and its cleaning. In this framework, after collecting data from sensors, outliers are first detected, and then the new value is generated and replaced by the smoothing models.

Quality improvement is also considered in some other approaches introduced by Yu et al. [138] and Hao et al. [139]. Since the data is stored in a matrix, the low-rank matrix method is used in both approaches. The problem is translated into an optimization problem. If the given and the optimal matrices are close to each other, the given matrix's rank is smaller and, therefore, is complete. The Singular Value Thresholding (SVT) algorithm is used to recover the matrix. This algorithm continues until it reaches the optimal answer and finally delivers the complete matrix.

Moreover, Dai proposed a method to assess the quality of data streams [90]. Firstly, the author uses three algorithms: a reference algorithm, a frequency algorithm, and an entropy algorithm to identify essential data in the database. Secondly, a quality assessment model based on statistical information is employed to evaluate the quality of data streams.

Furthermore, a rule-based method is proposed by Tian in [132] with two basic modules: (i) a detection module and (ii) a cleaning module for this purpose, he employs two types of worker processes: detect worker and router worker. First, experts define a set of rules, and each detected worker is responsible for investigating the violation of the corresponding rule. The router sends the input data to the corresponding detect worker. In each worker, historical data is also available to compare the new data with, and if there is a violation, it is identified. Then, the noisy data enters to the cleaning module. In this module, the equivalence class algorithm is used in order to clean data. In this algorithm, the same data is placed in a group, and the majority value is introduced as new data.

Licong et al. [129] proposed another rule-based approach for quality improvement of biomedical data using ontology. The authors claimed that merging ontologies for quality improvement of biomedical data takes much time. Hence, to speed up the process, scalable cloud computing environment such as map-reduce and Hadoop is used.

6.4. 2018 to present: A diversity of techniques

In the last four years, a wide variety of techniques is observed which are listed in Table 7. We first discuss the distance-based approaches.

Similar context-aware distance-based approaches are presented in [93,105,145], where spatiotemporal information is used to detect outliers in the sensor networks. Aleman et al. introduced a method to clean data in mobile sensor networks in which devices' location is considered [96]. Using location helps to find similar trajectories behavior and to detect outliers. Another context-aware method is proposed by Arfaoui et al. [91] to assess data quality in a body sensor network. First, outlier detection is performed locally by each sensor, where each sensor compares new data with the mean of its recent data values. If the new data exceeds the threshold, it is labeled as a candidate outlier. Finally, the processing unit considers other information such as blood pressure and heart rate, and detects the final outliers using the distance criterion.

The method proposed by Dai and Ding [144] is an approach that aims to make sensors data highly reliable (quality improvement goal). In this approach, each sensor detects its own outlier data by comparison with its historical data. Then, the detected outliers are compared with other sensors' observations using a distance measure to detect abnormal data values. On the other hand, the approach by Yessembayev et al. in [147], which is similar to the approach by Dai and Ding [144], aims to detect unreliable sensors. In this approach, the local outlier factor is used to detect outliers, and then by clustering data values, bad sensors are detected. Another method uses the local outlier factor, which combines the k-nearest neighbor, reverses the nearest neighbor, and shares the nearest neighbor to assess the quality of data streams [100]. Yu et al. introduced a density-based method in which a probability density function reflects the most recent data distribution [97]. Then, the obtained density is compared with a threshold to detect outliers. A KNN-based method is proposed by Zhu et al. to assess the quality of data streams [98]. After identifying the neighbors of each data value with the distance criterion, if the number of neighbors is less than a number k, the data value is considered low-quality. Similar methods using the distance measure to detect low-quality data have been proposed [106,108,111–113,115,120,121].

Peng et al. proposed a framework to distribute computing and perform data quality assessment by profiling techniques [62]. Similarly, a quality assessment architecture and an instance-based method are proposed by Karkouch et al. in [61]; in this architecture, there is a system by which the data consumer defines its requirements, and the system automatically and accordingly stores valid data in the database. In this system, which is accessible through a graphical user interface, the data consumer can choose the quality attributes and the specific data

attributes. The system transforms these requirements into executable code and executes this code to extract high-quality data.

Gil et al. proposed to detect abnormal data streams with a combination of PCA and SVM algorithms [92], while Zhang et al. used an artificial neural network to determine whether the data values measured by the sensors are outliers [94]. Moreover, Cejnek and Bukovsky introduced learning-based algorithms such as learning entropy and learning-based novelty detection [103] for detecting abnormal events. Besides, Su et al. in [69] proposed an approach in which, first, all similar sensors are clustered, and then based on data correlation, which is one of the data stream characteristics, bad sensors are detected. Another offline method proposed in [109] tries to improve the performance and accuracy of new values prediction by combining a model-based algorithm with a deep learning method. Finally, a neural network-based algorithm is proposed by Venskus et al. in [110] to classify data objects as normal or abnormal. The model is trained three times on a training dataset to enhance precision and sensitivity.

A hybrid learning-based approach was proposed by Chen [122] to detect anomalies online based on the adversarial generated time series. The main idea is to train an auto-encoder to learn the normal pattern of multivariate data points, and then use the reconstruction error to find abnormal data. A similar idea is proposed by Zhang et al. [123], Campos et al. [124], and Zhang et al. [125] to detect anomalies of data streams using temporal information, in which first, a deep convolutional auto-encoder is built and then using a bidirectional LSTM with attention, temporal dependencies are captured from data streams to distinguish between normal and abnormal data points. Furthermore, Su et al. [126] proposed a stochastic recurrent neural network for anomaly detection of data streams. The idea is to capture normal patterns by learning robust representations with a stochastic recurrent neural network and reconstruct incoming data points by the representations, and use the reconstruction probabilities to determine anomalies.

Another learning-based approach is by Ren et al. [127]. In this paper, Spectral Residual (SR) and Convolutional Neural Network (CNN) are employed to detect time-series anomalies at Microsoft services. The authors claimed that this is the first attempt to borrow the SR model from the visual saliency detection domain to time-series anomaly detection.

Moreover, Han et al. [119] proposed a deep probabilistic-based time-dependent approach for anomaly detection which employs deep learning-based methods to detect outliers from trajectory data streams. Bhatia et al. [118] proposed another learning-based approach to find an anomaly in graph edge. The approach focuses on combining statistical (chi-squared test) and algorithmic (count-min sketch) to find streaming micro-cluster anomalies.

A model-based approach that aims to detect abnormal sensors in automated vehicle sensors is proposed by Wang et al. [67]. In such an approach, model-based signal filtering is combined with an extended Kalman filter to smooth sensor readings. Moreover, using a car-following model and previous observations, abnormal sensors are detected. Furthermore, Widanage et al. in [107] and Wang et al. in [146] proposed Hierarchical Temporal Memory (HTM) algorithms to model data and predict the new data in real time. Also, Poornima and Paramasivan proposed another model-based method that benefits from locally weighted projection regression to model data values and predict candidate values [99]. If the difference between the incoming value and predicted data value exceeds a predefined threshold, the incoming value is considered an outlier. Hu et al. [68] proposed an online approach to preprocess high dimensional data streams, in which a small-sized dictionary captures the data patterns and updates itself by incoming data. Then the objective function is formulated to detect outliers as well as faulty readings over a long period of time.

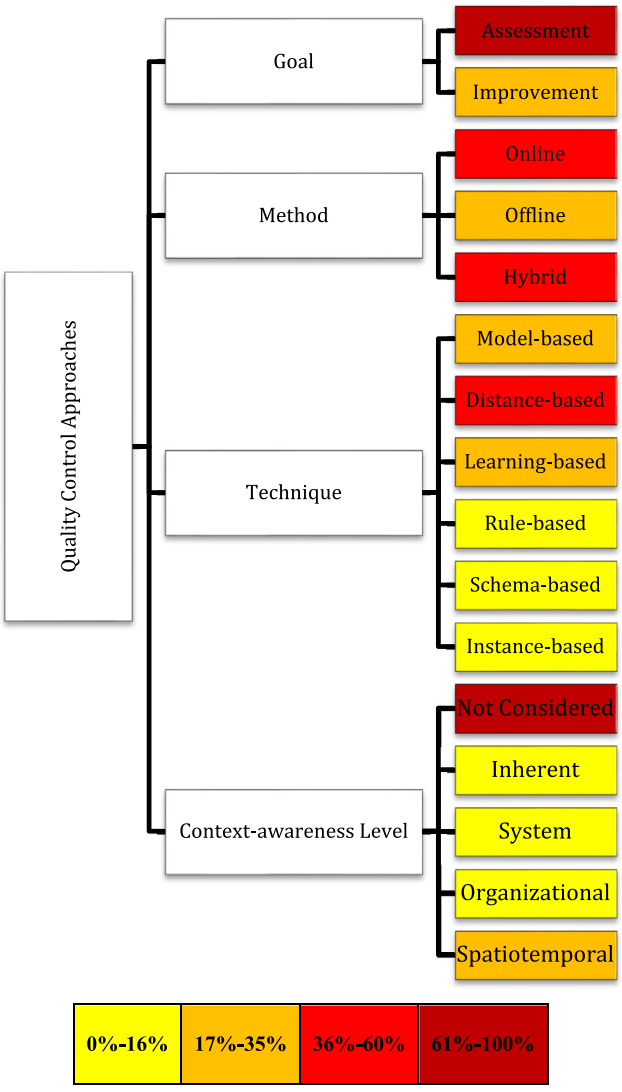

Fig. 3. Heat map of the extracted approaches. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7. Results and discussion

In this section, the second research question' answer (i.e. RQ2 in Table 3) addresses the research volume's inquiry in the target area. First, we use our proposed classification to provide a heat map to show the density and focus of studies in different dimensions of the classification. Fig. 3 provides a quantitative summary of the extracted approaches in the form of a heat map with values ranging from 0% (light yellow) to 100% (dark red).

As shown in Fig. 3, the quality assessment goal has received special attention amongst others. Additionally, hybrid processing has been gaining great attention than other methods, due to its advantages such as using a combination of offline processing (based on historical data) and online processing. Moreover, distance-based techniques have been proposed more frequently than other techniques. Furthermore, most approaches barely use any contextual information to deal with global changes. In a few cases in which context is considered, spatiotemporal context is used far more than other context types. The evolution of the techniques mentioned in Fig. 2 during the last two decades is one of the related analyses shown in Fig. 4. As demonstrated

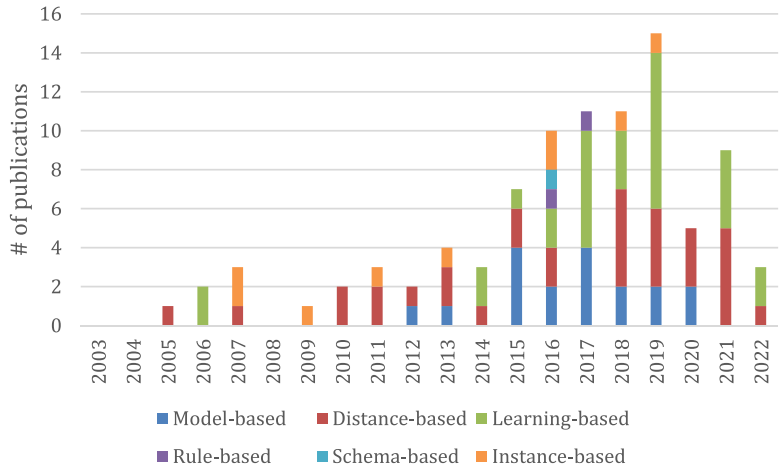

Fig. 4. Evolution of techniques in quality control of data streams over time.

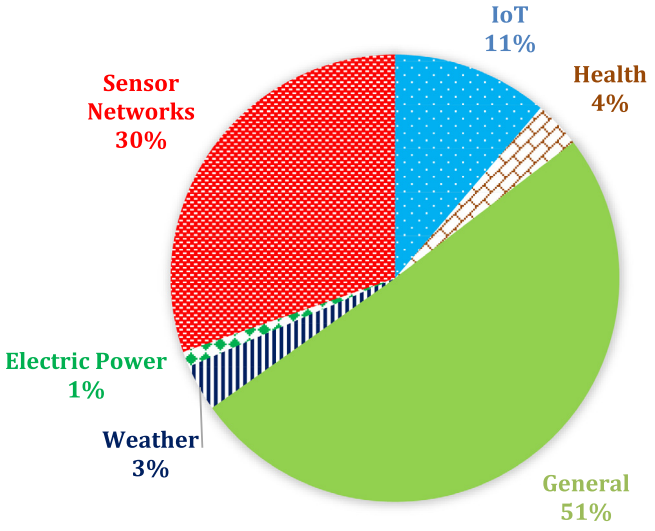

Fig. 5. Application domains of the proposed approaches.

Table 8
Distribution of study type.

No.	Research type	Number of papers
1	Books/Book chapters	2
2	Conference papers	40
3	Journal papers	44
4	Theses	6

in the figure, distance-based, learning-based, and model-based techniques have been gaining much attention during these years.

Table 8 presents the distribution of papers based on research type. The high number of journal papers emphasizes the importance of the topic.

Fig. 5 shows the percentage of approaches in terms of their application domain. As we can see from this figure, 51% of the papers do not have a specific domain. The sensor networks domain has gained more attention than others with 30% of the papers. Furthermore, less consideration is given to the Internet of things, weather, health, and electric power domains.

Fig. 6 shows the active journals/conferences, which help the researchers in the quality control of data streams community in following and searching their interesting topics.

Fig. 7 shows the geographical distribution of publications in the quality control of the data stream field. "Others" is also shown in the figure and refers to the sum of the frequencies of all countries with less than 3 percent. As we can see from the figure, the USA and China have the highest share of the total number of publications. India, Germany, France, and Canada stand in the next rank.

8. Challenges and future work

In this section, the third research question (i.e. RQ3 in Table 3) is answered. The issues not properly addressed in previous approaches are reported in Section 8.1 and directions for future research are presented in Section 8.2.

8.1. Challenges

Generally, the challenges can be divided into three categories related to source, inherent, and technique, respectively. First, Section 8.1.1 discusses the source dependent challenges. Then Section 8.1.2 discussed the inherent challenges and finally, Section 8.1.3 describes the technique dependent challenges.

8.1.1. Source dependent challenges

We identify the following source dependent challenges:

- ***Resource constraints:*** The quality control process often requires high computational capacity and time [56–58,61,65,75,77,88,130,133,149]. In some data production environments, such as sensor networks and the Internet of things, where sensors are responsible for generating data, there are constraints in terms of computational power, communication, and memory that make the assessment process more challenging. For example, if a sensor generates data, the sensor's constraints, such as its lack of required precision, can affect the data quality. Hence, sensor precision is a source dependent challenge.
- ***Source heterogeneity:*** Source heterogeneity is another challenge for quality assessment [76,148]. Addressing this challenge needs integration and evaluation of multiple data streams from different sources with various structures.
- ***Scalability:*** Another critical challenge is scalability, which is discussed in [78,86,87,133,140,142]. Evaluation methods cannot be implemented in a scalable manner in environments where it is not possible or cost-effective to distribute data or employ a distributed algorithm.

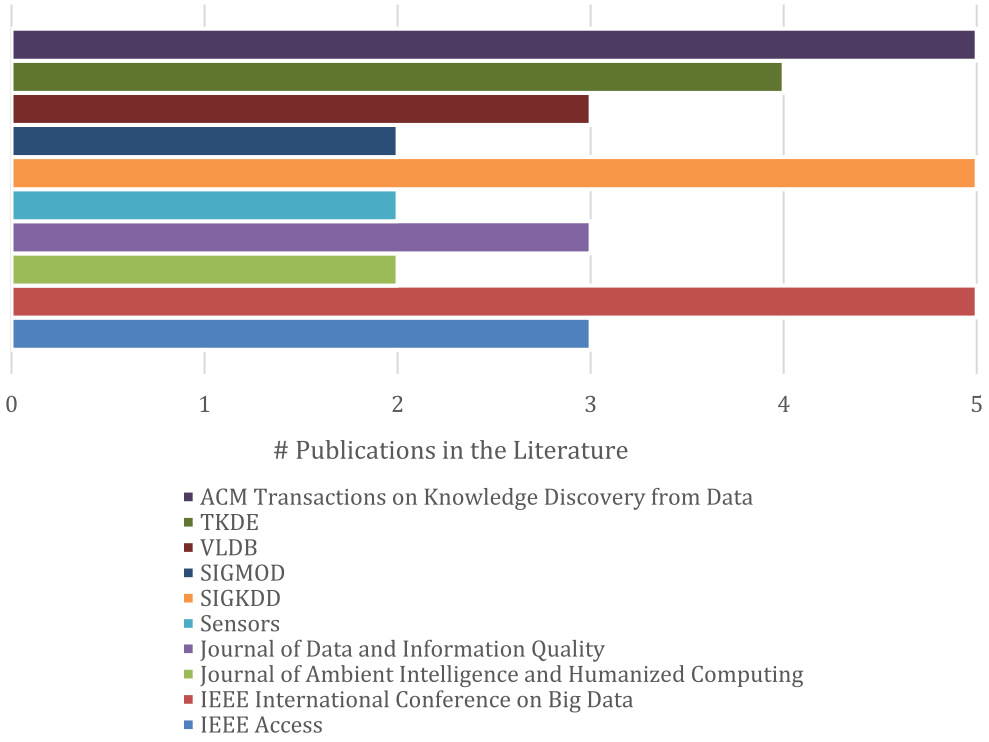

Fig. 6. Active Journals/Conferences in the quality control of data streams field.

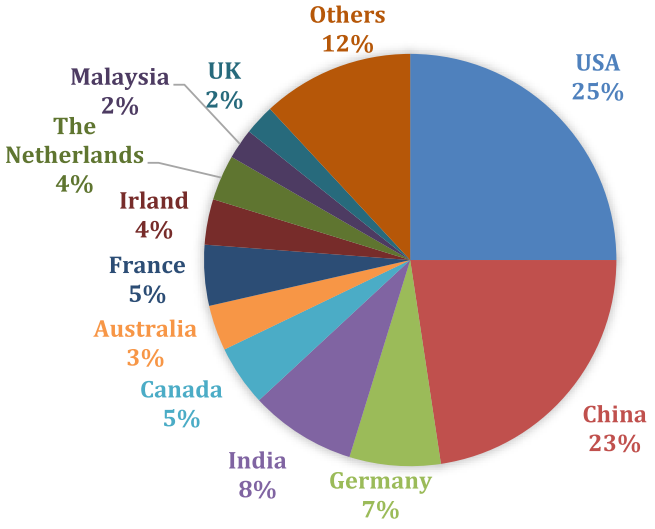

Fig. 7. Geographical distribution of publication.

8.1.2. Inherent challenges

Some challenges are inherent to the nature of data streams. The quality control approach should address these challenges:

- ***Variety of arrival rates:*** Data values arrive at different rates; therefore, the quality control approach should provide a mechanism for controlling the quality of the existing data before the arrival of new incoming data. Some previous papers discussed this problem [27,29].
- ***Infinite:*** As Tamboli and Shukla explained [29], in a data stream, the data is continuously received, and the quality control approach must be performed online without interrupting the primary retrieval process.
- ***Transient:*** In a data stream, transient data represents a significant challenge [27,29]. Data expires after a relatively short period of time and loses its credibility. Hence, data processing should be performed in a limited time window.
- ***Concept drift:*** Data distribution may change after a while. If the evaluation algorithm fails to find the new distribution, it cannot perform accurately. Paying attention to the distribution of data is considered one of the significant challenges in designing quality assessment/improvement algorithms [66,83,131].
- ***Heterogeneous schema:*** This challenge arises when different types of schemas, which may be represented by using different data models, are provided [84,148]. One of the complexities of data preprocessing is providing an efficient solution to deal with this issue.
- ***Distributed data points:*** Data may be received from different sources. It is necessary to integrate this data in order to obtain useful information. This challenge was discussed in [75,85,131,133].
- ***Appropriate dataset:*** The use of an appropriate dataset for testing proposed algorithms is essential. The lack of appropriate datasets for specific domains is one of the most critical data quality challenges. Synthetic datasets are often used for data quality.

8.1.3. Technique dependent challenges

Some challenges are related to the techniques used for quality control. Each technique (see Fig. 2) has its advantages and disadvantages, briefly explained as follows.

- ***Learning-based:*** Clustering and classification algorithms challenges were discussed in [23,24,31]. Clustering algorithms can be adapted to complex data types and used to obtain contextual information. However, the number of clusters is one of the crucial parameters for most clustering algorithms. Since there is no access to all data in the data stream, it is almost impossible to specify the number of

clusters, which leads to the production of arbitrary cluster shapes and difficulty in analyzing the clusters. Classification algorithms do not need to set the parameter. In the batch, the data can be tested and classified with reasonable accuracy by constructing the model. However, specifying the size of test data and updating the model are major challenges for data streams. Furthermore, the computational complexity of classification methods is higher than clustering algorithms.

- ***Model-based:*** Model-based methods use temporal correlation to build the model, and any changes in data distribution result in low-quality data. The simplicity of these methods justifies the prevalence of their usage for analyzing data streams. On the other hand, updating the model is one of the challenges of these techniques. Specifically, these methods cannot adapt themselves to the new data distribution when concept drifts occur.
- ***Distance-based:*** As explained in [23–25,27,30,31], these techniques (also known as nearest neighbor-based) do not assume any data distribution. Applying these techniques for different data types is simple since one only needs to define the appropriate distance criterion. However, it is challenging to define a suitable distance criterion for complex data. The computation cost for multivariate data is prohibitive, and those techniques are thus not suitable for high-dimensional data. Also, the scalability of these methods is a significant concern. A threshold value is typically used to detect outliers in these methods. An inappropriate threshold setting may result in poor results.
- ***Rule-based:*** The rules are identified by a domain expert in this technique, hence, making the evaluation's precision higher than other methods. Since these rules are obtained by examining the historical data, the corresponding quality control technique can be performed in a batch or a hybrid manner. One of the disadvantages of this approach is human intervention, which may take longer to define the rules [24,28].
- ***Instance and schema-based:*** Constraints for each attribute are obtained by profiling data items when the schema is available. Otherwise, they can obtain this by analyzing and inspecting the data values [56–58]. The schema, which often exists in batch data, can improve the profiling results' accuracy. On the other hand, in data streams with no data schema, the evaluation process becomes onerous, and the quality of data is usually evaluated by identifying a set of quality attributes [56–58,61,63,64]. Finding appropriate data quality attributes suitable for a target dataset is another challenge in data assessment.

8.2. Future research directions

Despite the advances in quality control of data streams, some problems are still unsolved. In what follows, a list of these problems and possible future directions are presented.

- ***Using hybrid methods to achieve higher accuracy:*** In some techniques such as learning-based approaches, data volume is necessary for reaching an acceptable accuracy level. In contrast, this is not a challenge for offline approaches, due to the availability of an appropriate volume of data. Hybrid methods, because of their reliance on historical data, are more suitable for evaluating data in real-time, than online approaches [132,143].
- ***Distributed architecture, more efficient but less used:*** The centralized architecture has been adopted in many approaches. However, researchers should use a distributed architecture due to the high volume of data and the computational constraints and provide an appropriate solution to the fault-tolerance problem. Distributed architecture can be a good solution to the high communication cost [76,77,79,86,87,131–133,136,140,142,149].
- ***Self-adaptive approaches as a way to recognize the concept drift:*** Some approaches use an adaptive solution for quality control of data streams [24,66,76,77,130,143]. As discussed before, data distribution may change over time. In this situation, the algorithm should adapt itself to the changes in the data distribution. The proposed approach should be adaptable when the method fails to predict and adapt to the data distribution, and it is difficult to detect outliers or evaluate the data.
- ***Context-awareness — wider view, more accurate detection:*** Reliance on local data is not sufficient for increasing the assessment's accuracy, which implies that global data or metadata should be considered. For example, the low-quality data produced by a specific sensor in a sensor network could be due to the sensor's failure. If the assessment algorithm only considers each sensor's data individually, it becomes hard to reach the desired outcome. However, the algorithm can have better performance if it has access to the data of other sensors. Due to the small number of context-aware approaches [86–89] and the importance of the topic as well as its higher accuracy, approaches should be designed and assessed that use this technique to evaluate the quality of data streams.
- ***Human intervention, better accuracy in batch, and less velocity in the stream:*** In some approaches, experts have to intervene to adjust some of the functional parameters for improving the accuracy of methods [24,28]. The expert's intervention is more applicable and suitable for batch processing, while it may be difficult for online processing. It is expected that the proposed approach either adopts a mechanism to minimize this intervention or takes this process in the shortest possible time in order not to slow down data processing.
- ***Quality characteristics as broad variety and choice difficulty:*** Given that many different quality models have been proposed, the proper selection of quality attributes should be domain-specific, since an inappropriate selection and evaluation may waste time and decrease precision.
- ***Lack of suitable tools:*** Providing a proper tool for controlling the data stream quality in real time is another problem yet to be solved. As indicated by Gao et al. in [154], the most commonly used tools only assess the quality of batch data without considering data streams, while real-time data stream evaluation is critical for many applications. Moreover, these tools profile the data, only consider the accuracy and completeness dimensions, and do not evaluate other quality attributes such as timeliness and consistency.

9. Conclusion

Data quality assessment plays a significant role in extracting meaningful data, especially those from online data sources. With this paper, we systematically and comprehensively studied the approaches proposed in the literature for quality control of data streams to address the three research questions listed in Table 3.

In the paper, we first provide a classification that categorizes approaches based on four dimensions, including goal (quality assessment, or improvement), method (involving online, offline, or hybrid), technique (comprising model-based, distance-based, learning-based, rule-based, instance-based, and schema-based), and context-awareness level (inherent, system, organizational, and spatiotemporal). We then review and compare all the identified approaches based on classification dimensions.

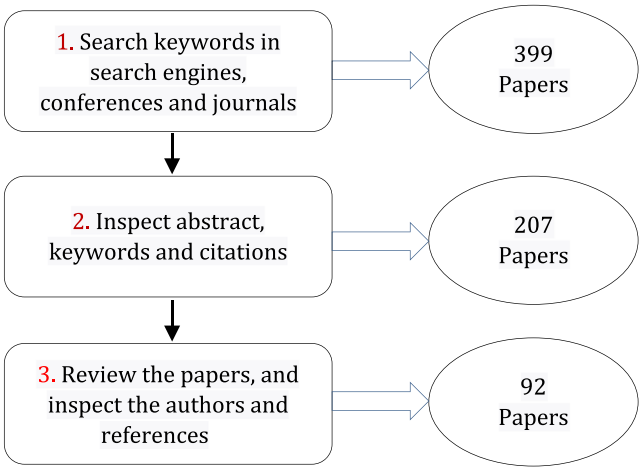

Fig. 8. Inclusion process and results.

Next, we graphically present the results based on the classification dimensions. We observe that quality assessment has received more attention than quality improvement. Moreover, due to the fact that calculating distance in a data window is less cost-effective than when data is completely stored, the distance-based techniques have gained greater consideration in the data stream with online processing. Additionally, hybrid processing has been gaining more attention than other methods due to its advantages, such as using a combination of offline processing (based on historical data) and online processing. Finally, we identify challenges, findings, and future directions. We divide the identified challenges into three categories: source, inherent, and technique-dependent, which are discussed together with future directions in Section 8. According to the analysis of the evolution of the research reported in the scientific literature over time, it can be concluded that quality control of data streams has attracted great attention in the last two decades, but the main problems, i.e., increasing the accuracy of detecting and improving low-quality data when dealing with tremendous volumes of real-time data, are yet to be solved.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Appendix. Search methodology

Search Methodology

We have conducted our systematic review using the guidelines proposed in [155,156], and the procedure described in [155, 156]. The systematic review process includes two main steps of planning and conducting, as described below.

Planning Phase

In this phase, we first define the scope of our systematic review and identify appropriate terms. Then, we specify the search engines and select related venues. Finally, we specify the eligibility criteria in order to plan the data extraction. These steps are explained as follows.

Defining the scope and terms. The paper's primary purpose is to collect broad research topics, summarize and review presented techniques, and express challenges and future directions. After defining the "Quality of data streams" as the target scope, we have identified a set of keywords, which were used as follows:

(Real-time *OR* distributed *OR* context-aware *OR* ") *AND* (data stream *OR* Internet of things (IoT) *OR* Big data) *AND* (quality *OR* assessment *OR* evaluation *OR* methodology *OR* improvement *OR* preprocessing *OR* cleaning *OR* outlier detection *OR* anomaly detection)

Specifying the search engines. In this step, we select a list of search engines to access the papers. These engines include Google Scholar,[3] ACM Digital Library,[4] IEEE Xplore Digital Library,[5] Springer Link,[6] Science Direct,[7] DBLP,[8] ProQuest,[9] OATD,[10] and California State University Library.[11]

Specifying related conferences and journals. To reduce the risk of missing some related works, we have selected a list of related venues, i.e., conferences and journals, as provided in Table 9.

Specifying the eligibility criteria. A list of inclusion and exclusion criteria was obtained as listed in Table 10.

Planning the data extraction. To extract the data from the selected papers, we have prepared a list of required items. Table 11 lists these items and the reason for their selection. Regarding the systematic review, if there is a paper in the pool of papers, it should be thoroughly reviewed and, if confirmed, the required data is extracted based on the items mentioned in this table.

Conducting Phase

After defining the search strategy, 396 papers have been retrieved, fifteen of which are theses. All papers have been carefully examined in three steps in this stage, as shown in Fig. 8.

In the first phase, the defined keywords are searched both in the search engines and the related conferences and journals. Papers are filtered, and then the paper's title is inspected, and if the paper has a related title, it is added to Mendeley software. At this phase, 399 papers are obtained, of which 15 are theses.

In the second phase, the abstract, the keywords, and the citations for each of the 399 papers are re-evaluated. After reviewing the abstract and the keyword and applying the inclusion and exclusion criteria, we conclude that the study is appropriate and is selected for reviewing thoroughly. Also, the first conducting phase is applied to citations, and if the title of the paper is appropriate, the paper will be added to Mendeley in the current phase. The number of 58 papers was added to the Mendeley after inspecting each citation's title, and finally, The number of 207 papers was included in the papers pool to be adequately studied.

In the final phase, all papers in the papers pool are thoroughly studied. References and authors of the papers are also reviewed separately. If the reference is appropriate for the study, it will be selected. Also, if other authors' publications are on social networks like ResearchGate, Google Scholar, and DBLP, their relevant

[3] https://scholar.google.com.
[4] https://dl.acm.org/.
[5] https://ieeexplore.ieee.org/Xplore/home.jsp.
[6] https://link.springer.com/.
[7] https://www.sciencedirect.com/.
[8] https://dblp.uni-trier.de/.
[9] https://pqdtopen.proquest.com/search.html.
[10] https://oatd.org/.
[11] https://csulb.libguides.com/dissertations.

Table 9
List of related conferences and journals.

Conferences	International Conference on Information Quality (ICIQ) CDOIQ Symposium The Data Governance and Information Quality Conference IEEE International Conference on Big Data IEEE International Conference on Data Mining IEEE Big Data Congress International Congress on Internet Of Things (ICIOT) ICDM: IEEE International Conference on Data Mining Quality of Information and Communications Technology (QUATIC) Quality Aspects in Big Data Systems (QABiD) International Conference On Big Data, IoT, And Data Science International Conference on Machine Learning and Big Data (ICMLB) Workshop on Quality of Open Data (QOD) International Conference on KDD (SIGKDD) ACM SIGMOD
Journals	Journal of Healthcare Quality Journal of Data and Information Quality (JDIQ) The International Journal on Very Large Data Bases (VLDB) International Journal of Sensor Networks International Journal of Information Quality IEEE Sensors Journal ACM Transactions on Sensor Networks IEEE Transactions on Knowledge and Data Engineering Data Mining and Knowledge Discovery IEEE Network Advances in Data Analysis and Classification ACM Transactions on Knowledge Discovery from Data Computer Communications Computer Networks Wireless Networks Wireless Personal Communications Ad Hoc Networks Journal of Network and Computer Applications Big Data Research

Table 10
Eligibility criteria.

Inclusion criteria	Exclusion criteria
Articles published in English between 2000 and 2020.	Papers shorter than four pages.
Papers where the search terms were found in the title or abstract.	Papers that did not propose any methodology or framework.
Papers where the full text is available.	Papers that were not peer-reviewed.

Table 11
Extracted items and usages.

Item	Usage
Title, DOI, Volume, Number, Month, Pages, Publisher, Venue Title	To describe the paper
Method (Online, Offline, and Hybrid), Goal (Assurance, Assessment, and Improvement), Technique (such as learning-based), and Context	To answer RQ1 in Table 3
Authors name, Countries, Affiliation, Domain, Year, Paper type (conference, journal, theses, chapter)	To answer RQ2 in Table 3
Challenges addressed in each paper	To answer RQ3 in Table 3
Keywords and Citations	To do conducting phase

papers are selected and reviewed, and if it is appropriate, they will be added to the final pool. A number of 15 papers have been obtained by inspecting the references and authors. As a result of this phase, we retrieved 92 papers fully explained and compared in Section 6. A total of 207 papers were thoroughly studied, and data from 92 papers have been extracted; the results are discussed in the next section.

Threats to Validity

Our study's validity has some threats, including the review process, primary paper selection, and data extraction, which are explained as follows.

Review process. The review process is the first threat, and one of the available guidelines should be used to deal with it. Various guidelines and references were provided, including the review and search process listed in [156–158], which [156] was selected as the primary source of guidance.

Primary paper selection. In order to prevent selection bias, papers were searched in Google Scholar, ACM Digital Library, IEEE Xplore Digital Library, Springer Link, Science Direct, and DBLP. To reduce the risk of missing some related works due to relying on specified search engines, we have selected a list of related conferences and journals presented in Table 9 and have accessed their published papers through their websites. Also, to find out more precisely the theses, several relevant search engines were listed: ProQuest, OATD, and California State University Library. Searching for papers in multiple databases can reduce the effect of paper selection threats.

Data extraction. To prevent data extraction bias, as Kitchenham et al. [156] stated, different authors should conduct data extraction independently. The results from the researchers should be compared to reach a consensus. In this paper, the authors have also analyzed the results in various sessions to deal with the threat.

References

[1] H.-N. Dai, R.C.-W. Wong, H. Wang, Z. Zheng, A.V. Vasilakos, Big data analytics for large-scale wireless networks, ACM Comput. Surv. 52 (2019) 1–36, http://dx.doi.org/10.1145/3337065.

[2] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, S. Guo, Protection of big data privacy, IEEE Access 4 (2016) 1821–1834, http://dx.doi.org/10.1109/ACCESS.2016.2558446.

[3] B. Mantha, Five guiding principles for realizing the promise of big data, Bus. Intell. J. (2014) 8, http://connection.ebscohost.com/c/articles/95066192/five-guiding-principles-realizing-promise-big-data (accessed March 4, 2019).

[4] A. Immonen, P. Paakkonen, E. Ovaska, Evaluating the quality of social media data in big data architecture, IEEE Access 3 (2015) 2028–2043, http://dx.doi.org/10.1109/ACCESS.2015.2490723.

[5] S. Bhatia, J. Li, W. Peng, T. Sun, Monitoring and analyzing customer feedback through social media platforms for identifying and remedying customer problems, in: Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. - ASONAM '13, ACM Press, New York, New York, USA, 2013, pp. 1147–1154, http://dx.doi.org/10.1145/2492517.2500287.

[6] F. Antunes, J.P. Costa, Integrating decision support and social networks, Adv. Hum.-Comput. Interact. 2012 (2012) 1–10, http://dx.doi.org/10.1155/2012/574276.

[7] A. Fabijan, H.H. Olsson, J. Bosch, Customer Feedback and Data Collection Techniques in Software R & D: A Literature Review, Springer, Cham, 2015, pp. 139–153, http://dx.doi.org/10.1007/978-3-319-19593-3_12.

[8] Z. Qian, Y. He, C. Su, Z. Wu, H. Zhu, T. Zhang, L. Zhou, Y. Yu, Z. Zhang, TimeStream: Reliable stream computation in the cloud, in: Proc. 8th ACM Eur. Conf. Comput. Syst. EuroSys 2013, 2013, pp. 1–14, http://dx.doi.org/10.1145/2465351.2465353.

[9] C.C. Aggarwal, Data mining: The textbook, 1981, http://dx.doi.org/10.1016/0304-3835(81)90152-X.

[10] L. Rutkowski, M. Jaworski, P. Duda, Stream data mining: Algorithms and their probabilistic properties | SpringerLink, 2020, https://link.springer.com/book/10.1007/978-3-030-13962-9 (accessed July 23, 2020).

[11] B. Behkamal, M. Kahani, E. Bagheri, Z. Jeremic, A metrics-driven approach for quality assessment of linked open data, J. Theor. Appl. Electron. Commer. Res. 9 (2014) 64–79, http://dx.doi.org/10.4067/S0718-18762014000200006.

[12] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, M. Allahbakhsh, Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions, ACM Comput. Surv. 51 (2018) http://dx.doi.org/10.1145/3148148.

[13] M. Allahbakhsh, H. Amintoosi, B. Behkamal, A. Beheshti, E. Bertino, SCiMet: Stable, scalable and reliable metric-based framework for quality assessment in collaborative content generation systems, J. Informetr. 15 (2021) 1–19, http://dx.doi.org/10.1016/j.joi.2020.101127.

[14] J. Merino, I. Caballero, B. Rivas, M. Serrano, M. Piattini, A data quality in use model for big data, Futur. Gener. Comput. Syst. 63 (2016) 123–130, http://dx.doi.org/10.1016/J.FUTURE.2015.11.024.

[15] X. Dong, H. He, C. Li, Y. Liu, H. Xiong, Scene-Based Big Data Quality Management Framework, Springer, Singapore, 2018, pp. 122–139, http://dx.doi.org/10.1007/978-981-13-2203-7_10.

[16] A.F. Haryadi, J. Hulstijn, A. Wahyudi, H. van der Voort, M. Janssen, Antecedents of big data quality: An empirical examination in financial service organizations, in: 2016 IEEE Int. Conf. Big Data (Big Data), IEEE, 2016, pp. 116–121, http://dx.doi.org/10.1109/BigData.2016.7840595.

[17] N. Abdullah, S.A. Ismail, S. Sophiayati, S.M. Sam, Data quality in big data: A review, Int. J. Adv. Soft Comput. Appl. 7 (2015) 16–27.

[18] C. ISO, Five year synchrophasor plan, 2011.

[19] J. Taylor, IBM big data and information management, 2011.

[20] Thomas C. Redman, Getting in front of data: The who does what, 2016.

[21] P. Reid, C. Petley, R. McClean, J. Jones, K. Ruck, Seizing the information advantage: How organizations can unlock value and insight from the information they hold, 2015, www.pwc.co.uk (accessed March 4, 2019).

[22] C. Batini, M. Scannapieco, Data and information quality, 2016, http://dx.doi.org/10.1007/978-3-319-24106-7.

[23] Yang Zhang, N. Meratnia, P. Havinga, Outlier detection techniques for wireless sensor networks: A survey, IEEE Commun. Surv. Tutorials 12 (2010) 159–170, http://dx.doi.org/10.1109/SURV.2010.021510.00088.

[24] M. Rassam, A. Zainal, M. Maarof, Advancements of data anomaly detection research in wireless sensor networks: A survey and open issues, Sensors 13 (2013) 10087–10122, http://dx.doi.org/10.3390/s130810087.

[25] M. Gupta, J. Gao, C.C. Aggarwal, J. Han, Outlier detection for temporal data: A survey, IEEE Trans. Knowl. Data Eng. 26 (2014) 2250–2267, http://dx.doi.org/10.1109/TKDE.2013.184.

[26] G. Krempl, I. Žliobaite, D. Brzeziński, E. Hüllermeier, M. Last, V. Lemaire, T. Noack, A. Shaker, S. Sievi, M. Spiliopoulou, J. Stefanowski, Open challenges for data stream mining research, ACM SIGKDD Explor. Newsl. 16 (2014) 1–10, http://dx.doi.org/10.1145/2674026.2674028.

[27] M. Shukla, Y.P. Kosta, P. Chauhan, Analysis and evaluation of outlier detection algorithms in data streams, in: 2015 Int. Conf. Comput. Commun. Control, IEEE, 2015, pp. 1–8, http://dx.doi.org/10.1109/IC4.2015.7375696.

[28] A. Karkouch, H. Mousannif, H. Al Moatassime, T. Noel, Data quality in internet of things: A state-of-the-art survey, J. Netw. Comput. Appl. 73 (2016) 57–81, http://dx.doi.org/10.1016/J.JNCA.2016.08.002.

[29] J. Tamboli, M. Shukla, A survey of outlier detection algorithms for data streams, in: 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev., IEEE, 2016.

[30] A. Ayadi, O. Ghorbel, A.M. Obeid, M. Abid, Outlier detection approaches for wireless sensor networks: A survey, Comput. Netw. 129 (2017) 319–333, http://dx.doi.org/10.1016/J.COMNET.2017.10.007.

[31] L. Chen, S. Gao, X. Cao, Research on real-time outlier detection over big data streams, Int. J. Comput. Appl. (2017) 1–9, http://dx.doi.org/10.1080/1206212X.2017.1397388.

[32] C.H. Park, Anomaly pattern detection on data streams, in: 2018 IEEE Int. Conf. Big Data Smart Comput., IEEE, 2018, pp. 689–692, http://dx.doi.org/10.1109/BigComp.2018.00127.

[33] C.H. Park, Outlier and anomaly pattern detection on data streams, J. Supercomput. 75 (2019) 6118–6128, http://dx.doi.org/10.1007/s11227-018-2674-1.

[34] E. Leal, L. Gruenwald, Research issues of outlier detection in trajectory streams using GPUs, ACM SIGKDD Explor. Newsl. 20 (2018) 13–20, http://dx.doi.org/10.1145/3299986.3299989.

[35] M. Salehi, L. Rashidi, A survey on anomaly detection in evolving data, ACM SIGKDD Explor. Newsl. 20 (2018) 13–23, http://dx.doi.org/10.1145/3229329.3229332.

[36] A.A. Cook, G. Misirli, Z. Fan, Anomaly detection for IoT time-series data: A survey, IEEE Internet Things J. 7 (2020) 6481–6494, http://dx.doi.org/10.1109/JIOT.2019.2958185.

[37] M. Munir, M.A. Chattha, A. Dengel, S. Ahmed, A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data, in: Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019, 2019, pp. 561–566, http://dx.doi.org/10.1109/ICMLA.2019.00105.

[38] H.Y. Teh, A.W. Kempa-Liehr, K.I.K. Wang, Sensor data quality: a systematic review, J. Big Data 7 (2020) 1–49, http://dx.doi.org/10.1186/s40537-020-0285-1.

[39] G. Pang, C. Shen, L. Cao, A. Van Den Hengel, Deep learning for anomaly detection: A review, ACM Comput. Surv. 54 (2021) http://dx.doi.org/10.1145/3439950.

[40] A. Blázquez-García, A. Conde, U. Mori, J.A. Lozano, A review on outlier/anomaly detection in time series data, ACM Comput. Surv. 54 (2021) http://dx.doi.org/10.1145/3444690.

[41] Y. Djenouri, D. Djenouri, J.C.W. Lin, Trajectory outlier detection: New problems and solutions for smart cities, ACM Trans. Knowl. Discov. Data 15 (2021) http://dx.doi.org/10.1145/3425867.

[42] H. Li, H. Lu, C.S. Jensen, B. Tang, M.A. Cheema, Spatial data quality in the internet of things: Management, exploitation, and prospects, ACM Comput. Surv. 55 (2022) 1–41, http://dx.doi.org/10.1145/3498338.

[43] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H.R. Motahari-Nezhad, E. Bertino, S. Dustdar, Quality control in crowdsourcing systems: Issues and directions, IEEE Internet Comput. 17 (2013) 76–81, http://dx.doi.org/10.1109/MIC.2013.20.

[44] E. Bertino, M.R. Jahanshahi, Adaptive and cost-effective collection of high-quality data for critical infrastructure and emergency management in smart cities—Framework and challenges, J. Data Inf. Qual. 10 (2018) 1–6, http://dx.doi.org/10.1145/3190579.

[45] F. Abedinzadeh, H. Amintoosi, M. Allahbakhsh, An iterative model for quality assessment in collaborative content generation systems, 2022, pp. 125–138, http://dx.doi.org/10.1007/978-3-031-14135-5_10.

[46] M. Allahbakhsh, H. Amintoosi, B. Behkamal, S.S. Kanhere, E. Bertino, AQA: An adaptive quality assessment framework for online review systems, IEEE Trans. Serv. Comput. (2020) http://dx.doi.org/10.1109/TSC.2020.2997737.

[47] S. Ghalibafan, B. Behkamal, M. Kahani, M. Allahbakhsh, An ontology-based method for improving the quality of process event logs using database bin logs, Int. J. Metadata Semant. Ontol. 14 (2020) 279–289, http://dx.doi.org/10.1504/IJMSO.2020.115436.

[48] B. Behkamal, M. Kahani, E. Bagheri, Quality metrics for linked open data, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Vol. 9261, 2015, pp. 144–152, http://dx.doi.org/10.1007/978-3-319-22849-5_11/COVER.

[49] B. Behkamal, Metrics-driven framework for LOD quality assessment, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), in: LNCS, vol. 8465, 2014, pp. 806–816, http://dx.doi.org/10.1007/978-3-319-07443-6_54/COVER.

[50] B. Behkamal, M. Kahani, E. Bagheri, M. Sazvar, A metric suite for systematic quality assessment of linked open data, Int. J. Inf. Commun. Technol. Res. 8 (2016) 27–45, http://ijict.itrc.ac.ir/article-1-60-en.html (accessed September 2, 2022).

[51] C. Cappiello, W. Samá, M. Vitali, Quality awareness for a successful big data exploitation, in: Proc. 22nd Int. Database Eng. Appl. Symp. - IDEAS 2018, ACM Press, New York, New York, USA, 2018, pp. 37–44, http://dx.doi.org/10.1145/3216122.3216124.
[52] I. Taleb, M.A. Serhani, R. Dssouli, Big data quality: A survey, in: 2018 IEEE Int. Congr. Big Data (BigData Congr.), IEEE, 2018, pp. 166–173, http://dx.doi.org/10.1109/BigDataCongress.2018.00029.
[53] L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the big data era, Data Sci. J. 14 (2015) 2, http://dx.doi.org/10.5334/dsj-2015-002.
[54] M. Klas, W. Putz, T. Lutz, Quality evaluation for big data: A scalable assessment approach and first evaluation results, in: 2016 Jt. Conf. Int. Work. Softw. Meas. Int. Conf. Softw. Process Prod. Meas., IEEE, 2016, pp. 115–124, http://dx.doi.org/10.1109/IWSM-Mensura.2016.026.
[55] S. Juddoo, Overview of data quality challenges in the context of big data, in: 2015 Int. Conf. Comput. Commun. Secur., IEEE, 2015, pp. 1–9, http://dx.doi.org/10.1109/CCCS.2015.7374131.
[56] A. Klein, Anja, Incorporating quality aspects in sensor data streams, in: Proc. ACM First Ph.D. Work. CIKM - PIKM '07, ACM Press, New York, New York, USA, 2007, p. 77, http://dx.doi.org/10.1145/1316874.1316888.
[57] A. Klein, H.-H. Do, G. Hackenbroich, M. Karnstedt, W. Lehner, Representing data quality for streaming and static data, in: 2007 IEEE 23rd Int. Conf. Data Eng. Work., IEEE, 2007, pp. 3–10, http://dx.doi.org/10.1109/ICDEW.2007.4400967.
[58] A. Klein, W. Lehner, Representing data quality in sensor data streaming environments, J. Data Inf. Qual. 1 (2009) 1–28, http://dx.doi.org/10.1145/1577840.1577845.
[59] M.A. Serhani, H.T. El Kassabi, I. Taleb, A. Nujum, An hybrid approach to quality evaluation across big data value chain, in: 2016 IEEE Int. Congr. Big Data (BigData Congr.), IEEE, 2016, pp. 418–425, http://dx.doi.org/10.1109/BigDataCongress.2016.65.
[60] S. Fagúndez, J. Fleitas, A. Marotta, Data stream quality evaluation for the generation of alarms in the health domain, J. Intell. Syst. 24 (2015) 361–369, http://dx.doi.org/10.1515/jisys-2014-0166.
[61] A. Karkouch, H. Mousannif, H. Al Moatassime, T. Noel, A model-driven framework for data quality management in the internet of things, J. Ambient Intell. Humaniz. Comput. 9 (2018) 977–998, http://dx.doi.org/10.1007/s12652-017-0498-0.
[62] 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: Proc. 2019 IEEE 3rd Int. Electr. Energy Conf. CIEEC 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 1018–1023, http://dx.doi.org/10.1109/CIEEC47146.2019.CIEEC-2019384.
[63] S. Geisler, S. Weber, C. Quix, Ontology-Based Data Quality Framework for Data Stream Applications, ICIQ, 2011.
[64] S. Geisler, C. Quix, S. Weber, M. Jarke, Ontology-based data quality management for data streams, J. Data Inf. Qual. 7 (2016) 1–34, http://dx.doi.org/10.1145/2968332.
[65] M. Rezvani, A. Ignjatovic, E. Bertino, S. Jha, A trust assessment framework for streaming data in WSNs using iterative filtering, in: 2015 IEEE Tenth Int. Conf. Intell. Sensors, Sens. Networks Inf. Process., IEEE, 2015, pp. 1–6, http://dx.doi.org/10.1109/ISSNIP.2015.7106935.
[66] S. Sadik, L. Gruenwald, Online outlier detection for data streams, in: Proc. 15th Symp. Int. Database Eng. Appl. - IDEAS '11, ACM Press, New York, New York, USA, 2011, p. 88, http://dx.doi.org/10.1145/2076623.2076635.
[67] Y. Wang, N. Masoud, A. Khojandi, Real-time sensor anomaly detection and recovery in connected automated vehicle sensors, IEEE Trans. Intell. Transp. Syst. (2020) 1–11, http://dx.doi.org/10.1109/tits.2020.2970295.
[68] Y. Hu, A. Qu, Y. Wang, D.B. Work, Streaming data preprocessing via online tensor recovery for large environmental sensor networks, ACM Trans. Knowl. Discov. Data (2022) http://dx.doi.org/10.1145/3532189.
[69] S. Su, Y. Sun, X. Gao, J. Qiu, Z. Tian, A correlation-change based feature selection method for IoT equipment anomaly detection, Appl. Sci. 9 (2019) 437, http://dx.doi.org/10.3390/app9030437.
[70] J. Han, M. Kamber, J. Pei, Data mining: Concepts and techniques, 2012.
[71] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Comput. Surv. 41 (2009) http://dx.doi.org/10.1145/1541880.1541883.
[72] C. Batini, M. Scannapieco, Data quality: concepts, methodologies, and techniques, 2006, http://dx.doi.org/10.1007/978-1-4020-4749-5_4.
[73] Y. Zhang, N.A.S. Hamm, N. Meratnia, A. Stein, M. van de Voort, P.J.M. Havinga, Statistics-based outlier detection for wireless sensor networks, Int. J. Geogr. Inf. Sci. 26 (2012) 1373–1392, http://dx.doi.org/10.1080/13658816.2012.654493.
[74] D.J. Hill, B.S. Minsker, Anomaly detection in streaming environmental sensor data: A data-driven modeling approach, Environ. Model. Softw. 25 (2010) 1014–1022, http://dx.doi.org/10.1016/J.ENVSOFT.2009.08.010.
[75] M.A. Rassam, M.A. Maarof, A. Zainal, A distributed anomaly detection model for wireless sensor networks based on the one-class principal component classifier, Int. J. Sens. Netw. 27 (2018) 200, http://dx.doi.org/10.1504/IJSNET.2018.093126.
[76] M.A. Rassam, M.A. Maarof, A. Zainal, Adaptive and online data anomaly detection for wireless sensor systems, Knowl-Based Syst. 60 (2014) 44–57, http://dx.doi.org/10.1016/J.KNOSYS.2014.01.003.
[77] Y. Zhang, N. Meratnia, P.J.M. Havinga, Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine, Ad Hoc Netw. 11 (2013) 1062–1074, http://dx.doi.org/10.1016/J.ADHOC.2012.11.001.
[78] T. Zaarour, N. Pavlopoulou, S. Hasan, U. ul Hassan, E. Curry, Automatic anomaly detection over sliding windows, in: Proc. 11th ACM Int. Conf. Distrib. Event-Based Syst. - DEBS '17, ACM Press, New York, New York, USA, 2017, pp. 310–314, http://dx.doi.org/10.1145/3093742.3095105.
[79] D. Jankov, S. Sikdar, R. Mukherjee, K. Teymourian, C. Jermaine, Real-time high performance anomaly detection over data streams, in: Proc. 11th ACM Int. Conf. Distrib. Event-Based Syst. - DEBS '17, ACM Press, New York, New York, USA, 2017, pp. 292–297, http://dx.doi.org/10.1145/3093742.3095102.
[80] X. Xu, Z. Zhang, Y. Chen, L. Li, HMM-based predictive model for enhancing data quality in WSN, Int. J. Comput. Appl. (2017) 1–9, http://dx.doi.org/10.1080/1206212X.2017.1395133.
[81] C. Isaksson, New outlier detection techniques for data streams, 2016.
[82] J.P. Rogers, Detection of outliers in spatial–temporal data, 2010.
[83] S. Sadik, L. Gruenwald, E. Leal, In pursuit of outliers in multi-dimensional data streams, in: 2016 IEEE Int. Conf. Big Data (Big Data), IEEE, 2016, pp. 512–521, http://dx.doi.org/10.1109/BigData.2016.7840642.
[84] Sadik M.S., Online detection of outliers for data streams, 2013.
[85] Y. Xiang, L. Guohua, X. Xiandong, L. Liandong, A data stream outlier detection algorithm based on grid, in: 27th Chinese Control Decis. Conf. (2015 CCDC), IEEE, 2015, pp. 4136–4141, http://dx.doi.org/10.1109/CCDC.2015.7162657.
[86] M.A. Hayes, M.A. Capretz, Contextual anomaly detection framework for big sensor data, J. Big Data 2 (2015) 2, http://dx.doi.org/10.1186/s40537-014-0011-y.
[87] M.A. Hayes, M.A.M. Capretz, Contextual anomaly detection in big sensor data, in: 2014 IEEE Int. Congr. Big Data, 2014, pp. 64–71, http://dx.doi.org/10.1109/BigData.Congress.2014.19.
[88] V. Iyer, Ensemble stream model for data-cleaning in sensor networks, 2013.
[89] S. Pumpichet, Novel online data cleaning protocols for data streams in trajectory, wireless sensor networks a dissertation submitted in partial fulfillment of the requirements for the degree of doctor of philosophy in electrical engineering by sitthapo, 2013.
[90] W. Dai, Stream data quality assessment based on distributed computing platforms a thesis submitted to the graduate school of university of arkansas at little rock in partial fulfillment of requirements for degree of master of science in information quality in the, 2016.
[91] A. Arfaoui, A. Kribeche, S.M. Senouci, M. Hamdi, Game-based adaptive anomaly detection in wireless body area networks, Comput. Netw. 163 (2019) 106870, http://dx.doi.org/10.1016/j.comnet.2019.106870.
[92] P. Gil, H. Martins, F. Januário, Outliers detection methods in wireless sensor networks, Artif. Intell. Rev. 52 (2019) 2411–2436, http://dx.doi.org/10.1007/s10462-018-9618-2.
[93] V.M. van Zoest, A. Stein, G. Hoek, Outlier detection in urban air quality sensor networks, Water. Air. Soil Pollut. 229 (2018) 1–13, http://dx.doi.org/10.1007/s11270-018-3756-7.
[94] K. Zhang, K. Yang, S. Li, D. Jing, H.B. Chen, ANN-based outlier detection for wireless sensor networks in smart buildings, IEEE Access 7 (2019) 95987–95997, http://dx.doi.org/10.1109/ACCESS.2019.2929550.
[95] S.K. Nagdeo, J. Mahapatro, Wireless Body Area network sensor faults and anomalous data detection and classification using machine learning, in: 2019 IEEE Bombay Sect. Signat. Conf. IBSSC 2019, Institute of Electrical and Electronics Engineers Inc., 2019, http://dx.doi.org/10.1109/IBSSC47189.2019.8973004.
[96] C.S. Aleman, N. Pissinou, S. Alemany, K. Boroojeni, J. Miller, Z. Ding, Context-aware data cleaning for mobile wireless sensor networks: A diversified trust approach, in: 2018 Int. Conf. Comput. Netw. Commun. ICNC 2018, 2018, pp. 226–230, http://dx.doi.org/10.1109/ICCNC.2018.8390320.
[97] K. Yu, W. Shi, N. Santoro, X. Ma, Real-time outlier detection over streaming data, in: Proc. - 2019 IEEE SmartWorld, Ubiquitous Intell. Comput. Adv. Trust. Comput. Scalable Comput. Commun. Internet People Smart City Innov. SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 125–132, http://dx.doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00063.
[98] R. Zhu, X. Ji, D. Yu, Z. Tan, L. Zhao, J. Li, X. Xia, KNN-based approximate outlier detection algorithm over IoT streaming data, IEEE Access 8 (2020) 42749–42759, http://dx.doi.org/10.1109/ACCESS.2020.2977114.
[99] I.G.A. Poornima, B. Paramasivan, Anomaly detection in wireless sensor network using machine learning algorithm, Comput. Commun. 151 (2020) 331–337, http://dx.doi.org/10.1016/j.comcom.2020.01.005.
[100] H. Yao, X. Fu, Y. Yang, O. Postolache, An incremental local outlier detection method in the data stream, Appl. Sci. 8 (2018) 1248, http://dx.doi.org/10.3390/app8081248.

[101] D. Janakiram, A.M. Reddy V, A.V.U.P. Kumar, Outlier detection in wireless sensor networks using bayesian belief networks, in: First Int. Conf. Commun. Syst. Softw. Middleware, Comsware 2006, 2006, http://dx.doi.org/10.1109/comswa.2006.1665221.

[102] W. Du, L. Fang, P. Ning, LAD: Localization anomaly detection for wireless sensor networks, in: Proc. - 19th IEEE Int. Parallel Distrib. Process. Symp. IPDPS 2005, 2005, http://dx.doi.org/10.1109/IPDPS.2005.267.

[103] M. Cejnek, I. Bukovsky, Concept drift robust adaptive novelty detection for data streams, Neurocomputing 309 (2018) 46–53, http://dx.doi.org/10.1016/j.neucom.2018.04.069.

[104] S. Mahfuz, H. Isah, F. Zulkernine, P. Nicholls, Detecting irregular patterns in IoT streaming data for fall detection, in: 2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2018, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 588–594, http://dx.doi.org/10.1109/IEMCON.2018.8614822.

[105] M. Chenaghlou, M. Moshtaghi, C. Leckie, M. Salehi, Online clustering for evolving data streams with online anomaly detection, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer Verlag, 2018, pp. 508–521, http://dx.doi.org/10.1007/978-3-319-93037-4_40.

[106] Y. Zhong, S. Fong, S. Hu, R. Wong, W. Lin, A novel sensor data pre-processing methodology for the internet of things using anomaly detection and transfer-by-subspace-similarity transformation, Sensors 19 (2019) 4536, http://dx.doi.org/10.3390/s19204536.

[107] C. Widanage, J. Li, S. Tyagi, R. Teja, B. Peng, S. Kamburugamuve, D. Baum, D. Smith, J. Qiu, J. Koskey, Anomaly detection over streaming data: Indy500 case study, in: IEEE Int. Conf. Cloud Comput. CLOUD, IEEE Computer Society, 2019, pp. 9–16, http://dx.doi.org/10.1109/CLOUD.2019.00015.

[108] R.A. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, M.A. Amanullah, I. Abaker Targio Hashem, E. Ahmed, M. Imran, Clustering-based real-time anomaly detection—A breakthrough in big data technologies, Trans. Emerg. Telecommun. Technol. (2019) e3647, http://dx.doi.org/10.1002/ett.3647.

[109] M. Munir, S.A. Siddiqui, M.A. Chattha, A. Dengel, S. Ahmed, FuseAD: Unsupervised anomaly detection in streaming sensors data by fusing statistical and deep learning models, Sensors 19 (2019) 2451, http://dx.doi.org/10.3390/s19112451.

[110] J. Venskus, P. Treigys, J. Bernatavičienė, G. Tamulevičius, V. Medvedev, Real-time maritime traffic anomaly detection based on sensors and history data embedding, Sensors 19 (2019) 3782, http://dx.doi.org/10.3390/s19173782.

[111] U. Gupta, V. Bhattacharjee, P.S. Bishnu, Outlier detection in wireless sensor networks based on neighbourhood, Wirel. Pers. Commun. 116 (2021) 443–454, http://dx.doi.org/10.1007/S11277-020-07722-3, 2020 1161.

[112] L. Chen, G. Li, G. Huang, A hypergrid based adaptive learning method for detecting data faults in wireless sensor networks, Inf. Sci. (NY) 553 (2021) 49–65, http://dx.doi.org/10.1016/J.INS.2020.12.011.

[113] M. Safaei, M. Driss, W. Boulila, E.A. Sundararajan, M. Safaei, Global outliers detection in wireless sensor networks: A novel approach integrating time-series analysis, entropy, and random forest-based classification, Softw. Pract. Exp. 52 (2021) 277–295, http://dx.doi.org/10.1002/SPE.3020.

[114] C. Junghans, M. Karnstedt, M. Gertz, Quality-driven resource-adaptive data stream mining? ACM SIGKDD Explor. Newsl. 13 (2011) 72–82, http://dx.doi.org/10.1145/2031331.2031342.

[115] A. Albanese, S.K. Pal, A. Petrosino, Rough sets, kernel set, and spatiotemporal outlier detection, IEEE Trans. Knowl. Data Eng. 26 (2012) 194–207, http://dx.doi.org/10.1109/TKDE.2012.234.

[116] D. Kumar, J.C. Bezdek, S. Rajasegarar, M. Palaniswami, C. Leckie, J. Chan, J. Gubbi, Adaptive cluster tendency visualization and anomaly detection for streaming data, ACM Trans. Knowl. Discov. Data 11 (2016) http://dx.doi.org/10.1145/2997656.

[117] L. Cao, J. Wang, E.A. Rundensteiner, Sharing-aware outlier analytics over high-volume data streams, in: Proc. 2016 Int. Conf. Manag. Data - SIGMOD '16, ACM Press, New York, New York, USA, 2016, pp. 527–540, http://dx.doi.org/10.1145/2882903.2882920.

[118] S. Bhatia, R. Liu, B. Hooi, M. Yoon, K. Shin, C. Faloutsos, Real-time anomaly detection in edge streams, ACM Trans. Knowl. Discov. Data 16 (2022) 1–22, http://dx.doi.org/10.1145/3494564.

[119] X. Han, R. Cheng, C. Ma, T. Grubenmann, DeepTEA: Effective and efficient online time-dependent trajectory outlier detection, Proc. VLDB Endow. 15 (2022) 1493–1505, http://dx.doi.org/10.14778/3523210.3523225.

[120] P. Boniol, J. Paparrizos, T. Palpanas, M.J. Franklin, SAND: Streaming subsequence anomaly detection, Proc. VLDB Endow. 14 (2021) 1717–1729, http://dx.doi.org/10.14778/3467861.3467863.

[121] S. Yoon, Y. Shin, J.G. Lee, B.S. Lee, Multiple dynamic outlier-detection from a data stream by exploiting duality of data and queries, in: Proc. ACM SIGMOD Int. Conf. Manag. Data., 2021, pp. 2063–2075, http://dx.doi.org/10.1145/3448016.3452810.

[122] X. Chen, L. Deng, F. Huang, C. Zhang, Z. Zhang, Y. Zhao, K. Zheng, DAEMON: Unsupervised anomaly detection and interpretation for multivariate time series, in: Proc. - Int. Conf. Data Eng. 2021–April, 2021, pp. 2225–2230, http://dx.doi.org/10.1109/ICDE51399.2021.00228.

[123] Y. Zhang, Y. Chen, J. Wang, Z. Pan, Unsupervised deep anomaly detection for multi-sensor time-series signals, IEEE Trans. Knowl. Data Eng. (2021) 1–14, http://dx.doi.org/10.1109/TKDE.2021.3102110.

[124] D. Campos, T. Kieu, C. Guo, F. Huang, K. Zheng, B. Yang, C.S. Jensen, Unsupervised time series outlier detection with diversity-driven convolutional ensembles, Proc. VLDB Endow. 15 (2021) 611–623, http://dx.doi.org/10.14778/3494124.3494142.

[125] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, N.V. Chawla, A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data chuxu, in: 33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019, 2019, pp. 1409–1416.

[126] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, D. Pei, Robust anomaly detection for multivariate time series through stochastic recurrent neural network, 1485 (2019) 2828–2837. http://dx.doi.org/10.1145/3292500.3330672.

[127] H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, Q. Zhang, Time-series anomaly detection service at microsoft, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Vol. 3330680, 2019, pp. 3009–3017, http://dx.doi.org/10.1145/3292500.3330680.

[128] S. Rajasegarar, C. Leckie, M. Palaniswami, J.C. Bezdek, Distributed anomaly detection in wireless sensor networks, in: 2006 IEEE Singapore Int. Conf. Commun. Syst. ICCS 2006, 2006, http://dx.doi.org/10.1109/ICCS.2006.301508.

[129] L. Cui, S. Tao, G.Q. Zhang, Biomedical ontology quality assurance using a big data approach, ACM Trans. Knowl. Discov. Data 10 (2016) http://dx.doi.org/10.1145/2768830.

[130] Z. Zhao, W. Ng, A model-based approach for RFID data stream cleansing, in: Proc. 21st ACM Int. Conf. Inf. Knowl. Manag. - CIKM '12, ACM Press, New York, New York, USA, 2012, p. 862, http://dx.doi.org/10.1145/2396761.2396871.

[131] S. Gill, B. Lee, A framework for distributed cleaning of data streams, Procedia Comput. Sci. 52 (2015) 1186–1191, http://dx.doi.org/10.1016/J.PROCS.2015.05.156.

[132] Y. Tian, P. Michiardi, M. Vukolic, Bleach: A distributed stream data cleaning system, in: 2017 IEEE Int. Congr. Big Data (BigData Congr.), IEEE, 2017, pp. 113–120, http://dx.doi.org/10.1109/BigDataCongress.2017.24.

[133] H. Liu, F. Huang, H. Li, W. Liu, T. Wang, A big data framework for electric power data quality assessment, in: 2017 14th Web Inf. Syst. Appl. Conf., IEEE, 2017, pp. 289–292, http://dx.doi.org/10.1109/WISA.2017.29.

[134] A. Zhang, S. Song, J. Wang, P.S. Yu, Time series data cleaning: From anomaly detection to anomaly repairing, Proc. VLDB Endow. 10 (2017) 1046–1057, http://dx.doi.org/10.14778/3115404.3115410.

[135] A. Zhang, S. Song, J. Wang, Sequential data cleaning: A statistical approach, 2016, pp. 909–924, http://dx.doi.org/10.1145/2882903.2915233.

[136] S. Gill, B. Lee, E. Neto, Context aware model-based cleaning of data streams, in: 2015 26th Irish Signals Syst. Conf., IEEE, 2015, pp. 1–6, http://dx.doi.org/10.1109/ISSC.2015.7163762.

[137] R. El Sibai, Y. Chabchoub, R. Chiky, J. Demerjian, K. Barbar, Assessing and Improving Sensors Data Quality in Streaming Context, Springer, Cham, 2017, pp. 590–599, http://dx.doi.org/10.1007/978-3-319-67077-5_57.

[138] Y. Yu, J.J.Q. Yu, V.O.K. Li, J.C.K. Lam, Low-rank singular value thresholding for recovering missing air quality data, in: 2017 IEEE Int. Conf. Big Data (Big Data), IEEE, 2017, pp. 508–513, http://dx.doi.org/10.1109/BigData.2017.8257965.

[139] Y. Hao, M. Wang, J.H. Chow, E. Farantatos, M. Patel, Model-less data quality improvement of streaming synchrophasor measurements by exploiting the low-rank Hankel structure, IEEE Trans. Power Syst. (2018) 1, http://dx.doi.org/10.1109/TPWRS.2018.2850708.

[140] H. Liu, J. Chen, F. Huang, H. Li, An electric power sensor data oriented data cleaning solution, in: 2017 14th Int. Symp. Pervasive Syst. Algorithms Networks 2017 11th Int. Conf. Front. Comput. Sci. Technol. 2017 Third Int. Symp. Creat. Comput., IEEE, 2017, pp. 430–435, http://dx.doi.org/10.1109/ISPAN-FCST-ISCC.2017.29.

[141] S. Basu, M. Meckesheimer, Automatic outlier detection for time series: an application to sensor data, Knowl. Inf. Syst. 11 (2007) 137–154, http://dx.doi.org/10.1007/s10115-006-0026-6.

[142] Y. Diao, K.-Y. Liu, X. Meng, X. Ye, K. He, A big data online cleaning algorithm based on dynamic outlier detection, in: 2015 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov., IEEE, 2015, pp. 230–234, http://dx.doi.org/10.1109/CyberC.2015.68.

[143] V. Pullabhotla, K.P. Supreethi, Adaptive pre-processing and regression of weather data, 2017, pp. 9–13, http://dx.doi.org/10.1007/978-981-10-3818-1_2.

[144] T. Dai, Z. Ding, Online distributed distance-based outlier clearance approaches for wireless sensor networks, Perv. Mob. Comput. 63 (2020) 101130, http://dx.doi.org/10.1016/j.pmcj.2020.101130.
[145] S. Bharti, K.K. Pattanaik, A. Pandey, Contextual outlier detection for wireless sensor networks, J. Ambient Intell. Humaniz. Comput. 11 (2020) 1511–1530, http://dx.doi.org/10.1007/s12652-019-01194-5.
[146] C. Wang, Z. Zhao, L. Gong, L. Zhu, Z. Liu, X. Cheng, A distributed anomaly detection system for in-vehicle network using HTM, IEEE Access 6 (2018) 9091–9098, http://dx.doi.org/10.1109/ACCESS.2018.2799210.
[147] A. Yessembayev, D. Sarkar, F. Sikder, Detection of good and bad sensor nodes in the presence of malicious attacks and its application to data aggregation, IEEE Trans. Signal Inf. Process. Netw. 4 (2018) 549–563, http://dx.doi.org/10.1109/TSIPN.2018.2790164.
[148] S. Xie, Z. Chen, Anomaly detection and redundancy elimination of big sensor data in internet of things, 2017, http://arxiv.org/abs/1703.03225 (accessed August 10, 2018).
[149] J. Lei, H. Bi, Y. Xia, J. Huang, H. Bae, An in-network data cleaning approach for wireless sensor networks, Intell. Autom. Soft Comput. 22 (2016) 599–604, http://dx.doi.org/10.1080/10798587.2016.1152769.
[150] Y. Zhang, C. Szabo, Q.Z. Sheng, Cleaning environmental sensing data streams based on individual sensor reliability, 2014, pp. 405–414, http://dx.doi.org/10.1007/978-3-319-11746-1_29.
[151] W.M.P. Van Der Aalst, S. Dustdar, Process mining put into context, IEEE Internet Comput. 16 (2012) 82–86, http://dx.doi.org/10.1109/MIC.2012.12.
[152] O. Oluwatimi, D. Midi, E. Bertino, A context-aware system to secure enterprise content, in: Proc. ACM Symp. Access Control Model. Technol. SACMAT. 06–08–June, 2016, pp. 63–72, http://dx.doi.org/10.1145/2914642.2914648.
[153] C. de Kerchove, P. Van Dooren, Iterative filtering in reputation systems, SIAM J. Matrix Anal. Appl. 31 (2010) 1812–1834, http://dx.doi.org/10.1137/090748196.
[154] J. Gao, C. Xie, C. Tao, Big data validation and quality assurance – issuses, challenges, and needs, in: 2016 IEEE Symp. Serv. Syst. Eng., IEEE, 2016, pp. 433–441, http://dx.doi.org/10.1109/SOSE.2016.63.
[155] A. Rula, A. Maurino, S. Auer, A. Zaveri, J. Lehmann, R. Pietrobon, Quality assessment for linked data: A survey, Semant. Web. 7 (2015) 63–93, http://dx.doi.org/10.3233/sw-150175.
[156] B.A. Kitchenham, I.C. Society, S.L. Pfleeger, L.M. Pickard, P.W. Jones, D.C. Hoaglin, K. El Emam, I.C. Society, J. Rosenberg, I.C. Society, Preliminary guidelines for empirical research in software engineering, IEEE Trans. Softw. Eng. 28 (2002) 721–734.
[157] A. Jedlitschka, M. Ciolkowski, D. Pfahl, Reporting experiments in software engineering, in: Guid. to Adv. Empir. Softw. Eng., Springer London, London, 2008, pp. 201–228, http://dx.doi.org/10.1007/978-1-84800-044-5_8.
[158] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empir. Softw. Eng. 14 (2009) 131–164, http://dx.doi.org/10.1007/s10664-008-9102-8.