

Acronyms

DQ	Data Quality
WSN	Wireless Sensor Network
IoT	Internet of Things
FFT	Fast Fourier Transform
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
RFR	Random Forest Regressor
SVM	Support Vector Machine
k-NN	k-Nearest Neighbors
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
OPTICS	Ordering Points To Identify the Clustering Structure
LOF	Local Outlier Factor

1 Introduction

The Internet of Things (IoT) refers to a network of interconnected devices, sensors, and actuators that can collect, exchange, and process data over the internet, enabling smart and autonomous systems ([Atzori et al. 2010](#)). IoT encompasses a wide range of applications, from industrial automation and smart homes to healthcare and urban management. Wireless Sensor Networks (WSNs) are a key component of IoT consisting of spatially distributed autonomous sensors that cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion, or pollutants ([Akyildiz et al. 2002](#)). These sensor nodes collect and transmit data wirelessly to a central gateway or base station for further processing and analysis, playing a crucial role in enabling real-time monitoring and analysis of urban environments ([Gubbi et al. 2013](#)).

The importance of data quality in WSNs for smart cities cannot be overstated. Cities are complex systems that exhibit emergent properties ([Bocchi & Facchini 2016](#)) allowing them—in theory—to be adaptable and resilient to changes—much like a living organism ([Langton 1990](#)). However, unlike living organisms, cities currently lack the equivalent of a central nervous system to detect environmental changes and coordinate responses autonomously ([Tortora & Derrickson 2018](#)). High-quality data is crucial for positive decision-making in cities, just as accurate sensory information is vital for living organisms to flourish.

Currently, cities have limited decision-making functionality due to slow, incomplete, and largely disconnected information transmission systems. To build a city-scale digital twin, rapid, interconnected, and ubiquitous sensory information is needed (Mohammadi & Taylor 2017). WSN can provide the necessary sensing and communication capabilities to gather and transmit data from the physical world to the internet (Ma et al. 2004). WSN and accompanying storage and processing infrastructure have begun to emerge in the form of urban observatories in cities around the world (Smith & Turner 2019, Rusli et al. 2023). The scale of WSN needed to support a digital twin leave no room for manual calibration of sensors meaning data quality assessments must be carried out autonomously.

WSN data, particularly in real-time streaming scenarios, is inherently prone to errors and inconsistencies which can lead to suboptimal or incorrect decisions in automated decision systems (Klein & Lehner 2009). For example, a sensor counting pedestrians might fail to detect an overcrowding event due to malfunction, potentially leading to a delayed emergency response. To address this critical challenge, there is a pressing need for automated systems that are 'quality-aware' and capable of assessing and managing data quality in real-time (Bisdikian et al. 2009, Karkouch et al. 2016). These systems must integrate quality control mechanisms along all real-time data pipelines, ensuring continuous monitoring and maintenance of data quality, and provide end-users with transparent insights into data provenance (Elkhodr & Alsinglawi 2020).

2 Research Objectives

2.1 Project Objectives

This research aims to address some of the challenges in reducing human involvement by developing a library that enables the creation of scalable and quality-aware pedestrian data stream pipelines. The library will provide a comprehensive framework for automated data quality assessment, cleaning, and monitoring, empowering end-users to make informed decisions based on reliable and trustworthy sensor data. By ensuring data quality throughout the pipeline, this research seeks to enhance the accuracy, reliability, and effectiveness of automated decision systems in pedestrian activity monitoring.

2.2 Review Objectives

The literature review explores the relationship between data quality and WSNs in the context of smart cities, focusing on methodologies for assessing data quality and developing a taxonomy suitable for pedestrian monitoring. The objectives of the literature review are to:

- Adopt/adapt an existing taxonomy for data quality dimensions in object-detecting wireless sensor networks.
- Establish the key themes from the literature for real-time data quality management and monitoring frameworks.
- Discuss existing methods for quantifying data quality in IoT networks.

- Present existing case studies that have implemented urban data quality management and monitoring pipelines.

3 Literature Review

The review will be structured as follows. The first section will explore what is meant by data quality in the context of wireless sensor networks (WSN) and examine the established taxonomies for quantifying data quality. The second section will investigate common methodologies used to assess data quality in WSN. The third section, which will be the main focus of the forthcoming research, will investigate methods for detecting and monitoring data quality in real-time. The fourth section will briefly investigate the methods used to manage and improve data quality such as automated methods for missing data-imputation and denoising. The fifth section, also brief, will explore the methods used to improve the design of WSN architecture so as to improve future data quality from the source. The final section will explore the challenges and future directions in building quality-aware platforms for smart cities.

3.1 Data Quality Dimensions and Metrics

3.2 Data Quality Assessment

3.3 Data Quality Detection and Monitoring

3.4 Data Quality Management and Improvement

3.5 Data Quality Prediction and Proactive Approaches

3.6 Challenges and Future Directions

3.1 Data Quality Dimensions and Metrics

Defining data quality can be challenging due to its multifaceted nature and the number of ways it can be assessed. Establishing a clear understanding of what constitutes data quality is crucial, particularly in the context of sensed pedestrian data. Identifying the most relevant aspects of data quality for pedestrian counting WSNs enables a targeted approach to developing quality-aware systems.

3.1.1 Data Quality Taxonomy

DQ taxonomies have developed significantly over the last few decades as technology has developed and the age of 'big data' and IoT has emerged. Although it is difficult to identify a single "seminal" paper on DQ, one of the most influential and widely cited papers on DQ dimensions is [Wang & Strong \(1996\)](#). The authors describe DQ as 'data that are fit for use by data consumers'. Their taxonomy presented in [Figure 1](#) describes four main categories of DQ dimensions: intrinsic, contextual, representational, accessibility.

NOTE: Taxonomies, as described by the methodology produced by [Nickerson et al. \(2013\)](#), consist of a set of dimensions that in turn consist of mutually exclusive characteristics.

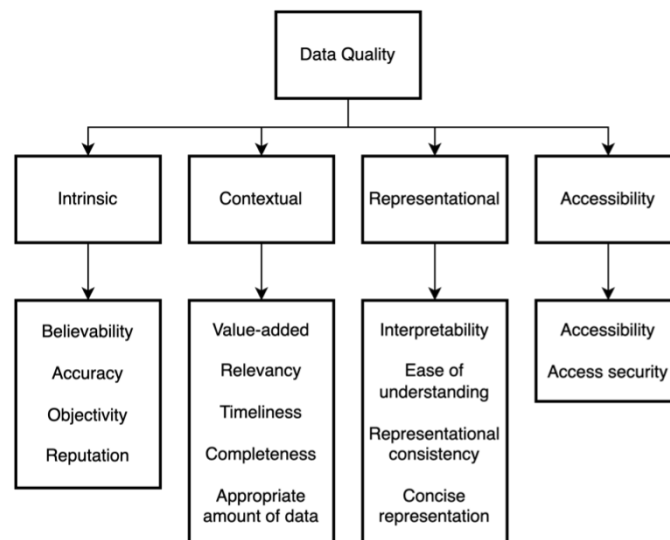


Figure 1: 'Seminal' DQ taxonomy from Wang & Strong (1996)

3.1.2 Internet of Things Data Quality Taxonomy

IoT data is highly structured (follows a schema) and can therefore use a narrower taxonomy with more precise definitions of dimensions. Karkouch et al. (2016) point out that many DQ dimensions have inconsistent definitions depending on the source. For example timeliness can (among other definitions) be defined as "currency" (when the data was last updated) (Dasu & Johnson 2003), or the average age of data in a source (Naumann 2002). The categories of intrinsic, contextual, representational and accessibility have remained consistent throughout taxonomies since Wang & Strong (1996). A simplified DQ taxonomy for IoT is presented in Figure 2.

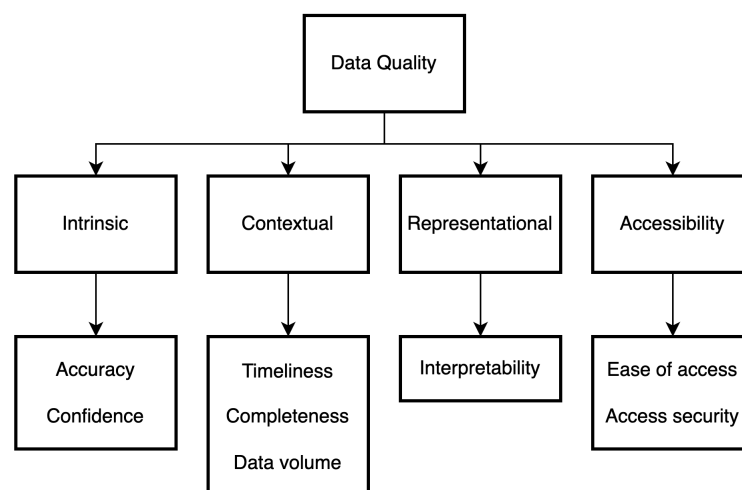


Figure 2: IoT DQ taxonomy based on Karkouch et al. (2016)

Data quality taxonomies form a hierarchy—the top level in figure 3 encompass all data quality issues (DQ), internet-of-things data quality (IoT DQ) issues form a subset (Batini & Scannapieco 2016) followed by wireless sensor network data quality issues (WSN DQ).

Existing taxonomies for WSN DQ are most relevant to this research and will be used as a foundation.

However, it is important to recognise that some DQ issues outlined in these taxonomies may not be applicable to this specific research, while other issues not covered in the taxonomies may arise, particularly those related to the performance of object detection algorithms (OD DQ). Whilst there are no taxonomies focussed on data from object detecting WSN, there are taxonomies focussed on deep-learning based object counting methodologies, which to some extent, cover data quality issues ([Heinrich et al. 2019](#)).

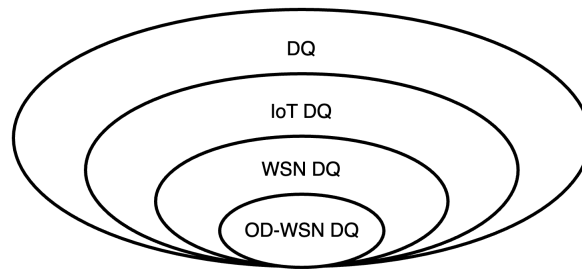


Figure 3: Subsets of taxonomies for data quality

3.1.3 Adopted Dimensions for Data Quality

[Mansouri et al. \(2023\)](#) provides a reliable and recent review of IoT data quality literature. Whilst there are limited significant changes to the dimensions proposed by [Karkouch et al. \(2016\)](#), the authors map key issues to the core dimensions from the [Karkouch et al. \(2016\)](#) framework, whilst highlighting how these issues arise from specific set of problems. The authors highlight the relationship between dimensions, problems and issues—how multiple issues can impact a single dimension and how a single problem can result in multiple data quality issues. Table 1 shows the key WSN-DQ dimensions identified by the authors that will be adopted for this research:

Data Quality Dimension	Sub-Dimension	Description	Factors to consider
Accuracy	Preciseness	How close the measured values are to the true values.	Measured values, sensor precision, measurement units, and granularity.
Accuracy	Certainty	The confidence or probability that a measured value is true.	Sensor calibration, environmental conditions, and measurement techniques, and statistical confidence intervals.
Accuracy	Confidence	Represents the reliability and credibility of the data source.	Sensor provider's reputation, sensor's conformance to specifications, certification, and independent testing results.
Timeliness	Freshness	The degree to which data is recent and not obsolete.	Time of last reading, data expiration policies, and data lifecycle management.
Timeliness	Frequency	How often the data is collected or updated.	Expected measurement frequency, data collection intervals, and synchronisation between data sources.
Completeness	Availability	Percentage of data values actually recorded compared to the expected number.	Existing data, historic expected measurement frequency, data gaps, and reasons for missing data.
Completeness	Coverage	The degree to which the recorded data covers the potential measurement space.	Spatio-temporal distribution of the data points, sensor placement, and measurement area or volume.
Consistency	Uniqueness	No duplication of records measuring the same thing.	Duplicated records, data deduplication techniques, and record identifiers.
Consistency	Integrity	Data values respect specified constraints and rules.	Data type constraints, range constraints, format consistency, and cross-field validation rules.
Usability	Interpretability	Presence of metadata to help understand encoded values.	Metadata standards, data dictionaries, measurement units, and data provenance.
Usability	Ease of manipulation	Suitability of the data format and semantics for aggregation and analysis.	Data format standards, data schema, data transformation requirements, and compatibility with analysis tools.

Table 1: DQ dimensions and factors for WSN

3.1.4 Causes of Data Quality Issues

IoT and sensor data streams present unique challenges for data quality management. [Karkouch et al. \(2016\)](#) highlight several factors that contribute to data quality issues in IoT, including deployment scale, resource constraints, network limitations, sensor inaccuracies, environmental conditions, and data stream processing to name but a few. These factors can lead to errors, inconsistencies, and data quality degradation. In the case of pedestrian sensors for example, heavy snow might affect the number of pedestrians picked up on the computer vision algorithm; Wi-Fi interference might result in a backlog of edge processed readings waiting to be sent to the database; or the sensor might malfunction resulting in no readings for a period of time. According to the [Karkouch et al. \(2016\)](#) framework, these examples would affect the dimensions of accuracy, timeliness, and completeness respectively.

3.1.5 Data Quality and Data Security

There exists another component to IoT challenges, covered in [Sicari et al. \(2016\)](#) which is that of data security. The intersection of data quality and data security issues can be summarised as follows: solutions are needed that can scale to the massive number of devices; resource constraints of devices limit applicability of existing techniques and call for lightweight approaches; and enforcing policies (cross-industry standardisation) is important. The key differences between data quality and data security are that the former focuses more on protecting against active threats and attacks, while data quality looks at accuracy and "fit for use". Privacy is a major concern from a security perspective but less central to data quality discussions. Provenance and data integrity overlap with trust issues but data quality considers many other dimensions beyond trustworthiness of sources (Figure 4). It is provenance from the perspective of transparency that is most relevant to objectives of this research, but there is potential for developing tools that address both data quality and security concerns through the same mechanisms.

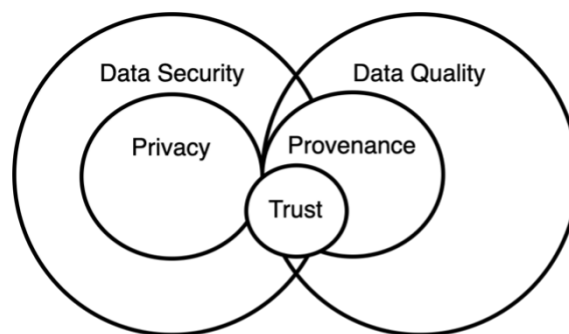


Figure 4: Data security vs quality

3.1.6 Data Quality Standards and Best Practice

The ISO/DIS 8000-210 standard outlines best practices for managing sensor data quality, emphasising the importance of data accuracy, consistency, completeness, timeliness, validity, anomaly detection, and maintenance ([ISO 2023](#)). The standard offers a comprehensive framework for data quality but has limitations due to its generalised nature, complexity in implementation, and lack of flexibility that is necessary for research projects. The dimensions of data quality outlined in the standard are consistent with those identified in the literature, but the standard does not provide specific guidelines for addressing data quality issues in IoT environments. [Perez-Castillo et al. \(2018\)](#) present twenty-three best practices for data quality in IoT environments. Although the focus is largely on the implementation of the network and the data management system, some of practices are applicable to the development of a quality aware pedestrian counting platform, such as documenting quality procedures; retaining versions of input data, workflows, programs, and models used (data provenance); and performing slope and persistence checks.

3.2 Data Quality Assessment

Immonen et al. (2015) make a useful distinction between quality assessment and quality evaluation.

- *Quality assessment*: assessing of the quality of raw data as such, without considering the context or the intended use of data.
- *Quality evaluation*: evaluating the quality of information, considering the context and the intended use of information.

This research project will mainly focus on quality assessment as the first step in the data quality management process. The dimensions of data quality identified in the previous section will be used as a foundation for assessing the quality of pedestrian data.

3.2.1 Data Quality Assessment Methodologies

DQ assessment is the first step in building an automated DQ pipeline. Batini et al. (2009) recognise that there are many methodologies to carry out such an assessment, but identify three common phases (figure 5). The first involves collecting contextual information about organisational processes, management procedures etc. The second involves measuring the value of a set of data quality dimensions and assessing the measurements against reference values. The final step concerns the selection of steps, strategies and techniques for reaching new DQ targets. As stated above, the methods investigated in this research focus mostly on the second stage, assessment and measurement of DQ.

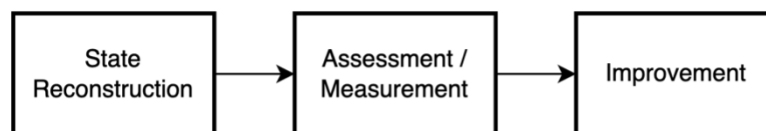


Figure 5: General stages of DQ methodology (Batini et al. 2009)

3.2.2 Event Detection

Assessing data quality in WSN (time-series data) is heavily embedded in event detection. Quantifying many of the dimensions mentioned above first require understanding the ‘normal behaviour of the data’. Events are occurrences or patterns in the data that are of significant interest or importance within a specific domain or application context (Benabbas et al. 2023). Event detection and data quality are closely linked (Kumar et al. 2023), because all data quality issues are ‘events’ (patterns in the data that are of significant interest). The following sub-dimensions of data quality would need to be assessed through event detection methods when applied to real-time WSN data:

- **Accuracy**: identifying data points that deviate significantly from the true or expected values, indicating potential accuracy issues.
- **Completeness**: detecting missing or incomplete data events, (simple) event detection techniques can assess the completeness of the incoming data.

- **Consistency:** identify inconsistencies and contradictions in the data by comparing data points across different sources or time periods.
- **Timeliness:** detecting events related to data delays, staleness, or latency can help assess the timeliness of the data.
- **Validity:** detection techniques can identify data points that violate predefined rules, constraints, or formats, indicating validity issues.

There are multiple approaches to event detection presented in the scientific literature. Broadly speaking, these are, rule-based methods, machine learning-based methods, deep learning-based methods, statistical and probabilistic methods, and hybrid methods [Kumar et al. \(2023\)](#). Table 2 gives a breakdown of what each of these methods entail.

Method	Examples	Description
Rule-based	Threshold-based detection, finite state machines, and expert systems.	Simple to implement but may struggle with complex event patterns and adaptability.
Machine learning	Supervised learning: decision trees, support vector machines, naive Bayes, etc. Unsupervised learning: clustering, anomaly detection, etc.	Can adapt to complex event patterns but require sufficient labelled data for training.
Deep learning	Convolutional neural networks (CNNs), recurrent neural networks (RNNs), Graph neural networks (GNNs).	Can handle high-dimensional and unstructured data but require large amounts of training data and computational resources.
Statistical and probabilistic	Hidden Markov models (HMMs), Bayesian networks, and Gaussian mixture models (GMMs).	Can capture the uncertainties and dependencies in event occurrences but may require prior knowledge of event distributions.
Hybrid	Combining rule-based and machine learning methods or integrating deep learning with statistical models.	Can provide more robust and accurate event detection but may increase the complexity of the system.

Table 2: Event detection methods for data quality

In the space of all domains and application contexts, every data point could, in theory, be considered an event. For the purpose of this research, the term *event of interest* means a domain and application has been applied and therefore it can be considered a collapsed point appearing somewhere on Figure 6, it could be either a novelty or an anomaly, a sub-category of one or both, or neither. Figure 6 is based on a number of research papers that seek to categorise types of events ([Chandola et al. 2009](#), [Pimentel et al. 2014](#), [Ahmad et al. 2017](#), [Aminikhanghahi & Cook 2017](#)). A brief summary of these categories is as follows:

- **Event** - generally refers to some occurrence of interest in the system being monitored. Unusual or rare events can be considered anomalies or novelties.
- **Anomalies** - data points or patterns that deviate from the expected or normal behaviour of the system. They are often considered "abnormal" or "unusual" in the context of the system's typical operation. Anomalies can be caused by errors, faults, or malicious activities in the system. The key characteristic of an anomaly is that it deviates from the expected behaviour, regardless of whether it has been observed before.

- Novelty – data points or patterns that are new or previously unseen in the context of the system's operation. They represent a new behaviour or state of the system that has not been observed in the training data or during the system's normal operation. Novelties may or may not be anomalous. A novelty could represent a new normal behaviour (e.g., due to a system update or change in environment) or a new type of anomaly. The key characteristic of a novelty is its newness or previously unseen nature, regardless of whether it is normal or abnormal.
- Outlier - a data sequence, sub-sequence or point that lies far from the rest of the data based on some measure of distance.
- Discord - refers to the most unusual subsequence or sub-series within a time series.
- Change-point - a point in time where the behaviour of the system being monitored changes abruptly.
- Drift - Drift or concept drift refers to a gradual change in the behaviour of the system over time.

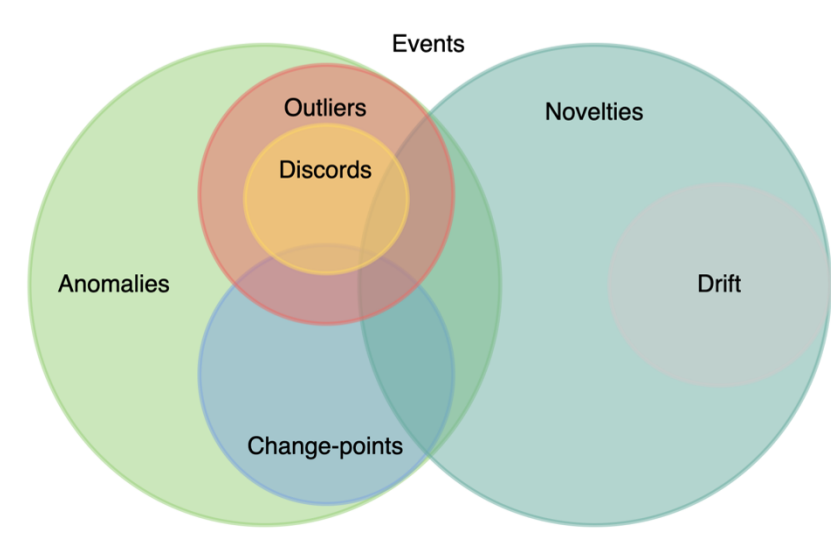


Figure 6: Theoretical event terminology

These distinctions are useful because when event detection is carried out, it is often specifically outliers, discords, change-points, or drift that being directly measured though a chosen method. Once measured, these events will then be classified as either anomalies, novelties, or anomalous novelties to inform decision making ([Abu Alsheikh et al. 2015](#)). For example, a detected anomaly, once verified as reliable, might trigger a physical response—in the context of pedestrian data this might mean dispatching an emergency services unit for crowd control. A novelty, on the other hand, might trigger the retraining of AI models as this could indicate that the factors driving pedestrian behaviour have drifted. A recent example of this is the change in pedestrian behaviour as a result of the COVID-19 pandemic ([Chen et al. 2021](#)). As pointed out by [Carreño et al. \(2020\)](#), a novel anomaly would be difficult to classify autonomously as they would not be present in any training data but trigger an alert for manual analysis or model retraining.

Having identified the dimensions critical to assessing data quality in sensor data, and establishing the different groups of methods that can be applied to the problem of assessing data quality through measurement, we can begin to look at what an integrated data quality management platform for pedestrian data might look like. Three themes will be investigated in the following sections, each of these forms a process in the data management feedback loop (Immonen et al. 2015, Ehrlinger & Wöb 2017). These are: detection and monitoring (Woodall et al. 2013, Ehrlinger & Wöb 2017); management and improvement (Khatri & Brown 2010, Batini & Scannapieco 2016); and prediction and proactive approaches (Li et al. 2012, Ardagna et al. 2018).

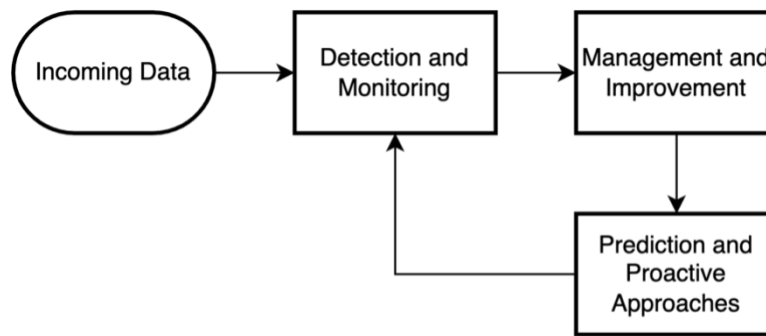


Figure 7: Data quality feedback loop

3.3 Data Quality Detection and Monitoring

This research is primarily interested in anomaly detection (defined in figure 6) which forms a binary classification problem with imbalanced classes—the objective is to categorise each instance in a dataset as belonging to one of two possible classes: “normal” or “anomalous”—shown in figure 9—(Gorshenin et al. 2024).

NOTE: Monitoring is the continuous observation and collection of data, while detection is the process of identifying specific events, anomalies, or patterns of interest within that monitored data. Monitoring provides visibility and context, while detection pinpoints specific occurrences or conditions that require attention or action (Tuychiev 2023).

3.3.1 Monitoring Methods

There are a number of DQ dimensions that can be calculated without the need for a complicated model. Fizza et al. (2022) develops a model for assessing age of data using parking sensors as a case study. The model uses a time-based calculation with spatial clustering. The authors used a heuristic approach to determine the confidence level in the state of the sensor based on the age of the data.

$$Confidence = 1 - \frac{avg.Age\ of\ the\ cluster}{max.Age} \quad (6)$$

The authors showed that this simple method could enable both better parking recommendations and highlight issues with sensor connectivity.

3.3.2 Pattern-Based Detection Methods

Pattern-based detection methods refer to techniques used in anomaly detection to identify specific patterns within data that correspond to known behaviours, signatures, or anomalies. These methods rely on predefined rules or models that describe what constitutes a normal or abnormal pattern in the dataset (Cai et al. 2023).

Detection requires identifying events. There are a number of approaches to this in the literature following the approaches shown in table 2 above. Klein & Lehner (2009) present a methodology based on sliding windows that follow a rule-based approach utilising fast Fourier transform (FFT) to assess signal behaviour (table 3). One downside to this approach is that FFT requires complete time-series data windows to work, meaning some sort of missing data imputation would need to be carried out first. However, there are methods like the Lomb-Scargle periodogram, that can create similar outputs to FFT on incomplete data (VanderPlas 2018).

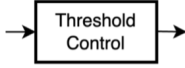
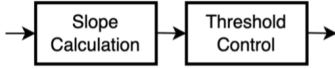
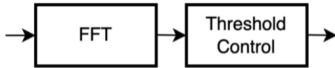
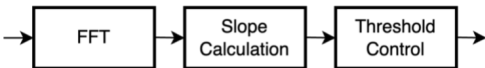
Example function	Interest indicator	DQ control operator pattern
<i>current_value()</i>	Extraordinary value ranges	
<i>sliding_slope()</i>	Extraordinary value alterations	
<i>fft_slope()</i>	Changing periodicity	
<i>fft()</i>	Unsteadiness	

Table 3: Interest indicators and operators (Klein and Lehner 2009)

3.3.3 Real-Time Monitoring and Event-Driven Architecture

System architecture is an important component in detection and monitoring as it requires a number of different processes running together. Mehmood et al. (2024) develop a portable hybrid architecture for smart cities based on device edge and cloud computing. The hybrid system combines LSTMs, PageHinkley test, adaptive windowing, and Kolmogorov-Smirnov windowing. Şimsek et al. (2024) integrate Transformers, CNN-LSTM, GRU, and RFR models into their hybrid deep-learning detection system that uses fog computing, complex event processing, and virtualisation to do event detection. Whilst these systems are not directly applicable to the research, they provide a good foundation for understanding the architecture of a real-time monitoring system.

3.3.4 Outlier Detection

If a sensor frequently produces outliers, it may indicate issues with sensor calibration, environmental conditions, or measurement techniques. By detecting and analysing pat-

terns in outliers, we can assess the reliability of the sensor data and make adjustments to improve certainty. Common outlier detection methods from literature include:

- **Statistical methods:** These methods assume that the data follows a particular distribution (e.g., Gaussian) and identify outliers based on statistical measures, such as mean, median, standard deviation, or interquartile range (Iglewicz & Hoaglin 1993, Aggarwal 2017).
- **Distance-based methods:** These methods identify outliers based on their distance from other data points. Examples include k-nearest neighbours (k-NN) and Local Outlier Factor (LOF) (Ramaswamy et al. 2000, Breunig et al. 2000)
- **Density-based methods:** These methods detect outliers based on the density of data points in their neighbourhood. Examples include DBSCAN and OPTICS (Ester et al. 1996, Ankerst et al. 1999).
- **Machine learning methods:** These methods employ machine learning algorithms, such as support vector machines (SVM) or deep learning, to learn patterns in the data and identify outliers based on their deviations from these patterns (Blázquez-García et al. 2021).

Shukla & Sengupta (2020) propose a scalable outlier detector using hierarchical clustering and LSTM neural networks, demonstrating high accuracy and outlier sensitivity tuning capabilities. Nguyen et al. (2021) combine LSTM auto-encoders with one-class SVM for anomaly detection, achieving promising results on benchmark datasets and fashion retail data. There are a number of classifying types of outlier for WSN, Zhang et al. (2010) divide these into events detection, malicious attack detection, and noise and error detection (Figure 8).

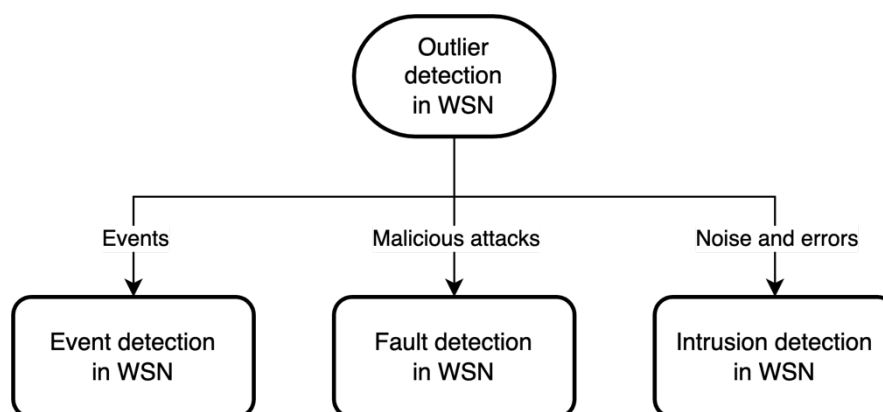


Figure 8: Outlier detection categories based on Zhang et al. (2010)

Ding et al. (2005) stress the importance of identifying the boundaries of important events in wireless sensor networks rather than just the regions where these events occur. This focus is due to the unreliability of sensor measurements. They explain the main differences between detecting important events and outlier detection. These differences include:

- Prior knowledge of trigger conditions: Knowing what triggers an event is crucial for detecting important events, while outlier detection does not require this knowledge.
- Goals: Event detection aims to specify events of interest, whereas outlier detection focuses on identifying unusual readings.
- Misclassification: In outlier detection, it's important to avoid classifying normal data as outliers. In event detection, the goal is to prevent incorrect data from affecting the reliability of detecting important events.

Despite these differences, both techniques use the spatial and temporal correlations among sensor data from nearby nodes to distinguish between actual events and errors. Event measurements are usually spatially related, while noisy measurements and sensor faults are not [Teh et al. \(2020\)](#). Although outlier detection techniques might work for event detection, they are not commonly used in this field because not all outliers need to be identified for event detection purposes. In the context of figure 8, an event of interest is one that has led to erroneous data and triggers data cleaning.

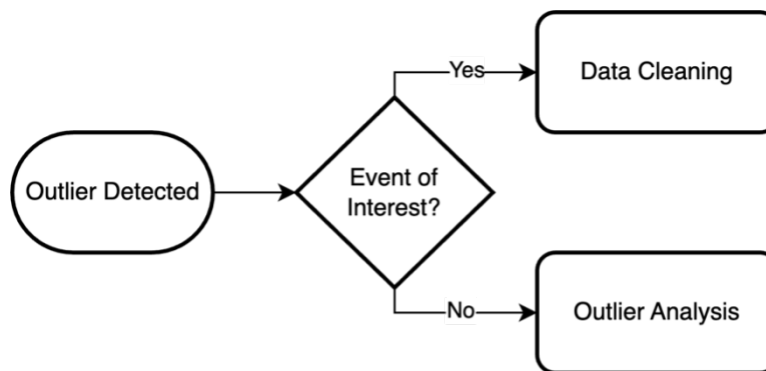


Figure 9: Decision making as a result of outlier analysis based on [Blázquez-García et al. \(2021\)](#)

3.3.5 Abnormal Node Detection in WSN

Abnormal node detection in wireless sensor networks refers to the process of identifying nodes that exhibit unusual or anomalous behaviour compared to the expected normal behaviour of the nodes in the network ([Li & Parker 2014](#)). In the context of urban mobility data such as pedestrian counts, abnormal node detection is needed to differentiate between real events such as overcrowding of a metro station and erroneous events arising from detection faults.

[Chen \(2021\)](#) take a traditional approach using multi-step methodology using spatio-temporal correlation to detect abnormal nodes. They first calculate the offset between a historic window and the current window to evaluate temporal consistency. They then use a nearest neighbour algorithm cluster the data. They combine the result of spatial and temporal offsets and repeat for all nodes in the network. Whilst the results show that this method outperforms others, the authors acknowledge that this method is computationally expensive and may not be suitable for large networks or real-time applications.

Zhang et al. (2022) propose an anomaly detection method for multimodal WSN data flow via a dynamic graph neural network. The authors inject four types of anomalies into the testing data:

- Anomaly type 1 simulates slow mode changes in the environment detected by a node that deviates from its conventional state (e.g. the observed value of the gas content of a sensor node gradually increases due to the leakage of toxic gas in the chemical plant).
- Anomaly type 2 simulates fast mode changes in the environment detected by a node (e.g. the temperature observation of a sensor node rises rapidly in a short period of time due to a fire).
- Anomaly type 3 simulates extreme changes in the external environment detected by a node (e.g. the sudden invasion of high-temperature sources leads to sudden changes in the temperature observations).
- Anomaly type 4 simulates the failure of a sensor node due to external interference, and the observed value of the mode becomes 0.

The authors use a GRU as it is less computationally expensive than an LSTM which shows promising results for real-time applications.

NOTE: In this experiment the GRU performed best a window length of 60 data points and an output layer size of 32. The authors also found that using min-max normalisation compared to z-score normalisation resulted in high misjudgement rate.

3.4 Data Quality Management and Improvement

Although data quality improvement is beyond the scope of this research, it is important to be cognisant of these methods to inform development decisions for the quality-aware system. There are a number of different methods for achieving this.

3.4.1 Correcting Errors

Teh et al. (2020) identify two categories of error correcting methods in WSN data, *missing data imputation* which attempts to correct estimated sensor measurement values that are missing and *de-noising* which aims to remove the noise associated with the measurement signal. The authors reference a number of different methods for each of these categories, including: association rule mining (Gruenwald et al. 2007), clustering (Tang et al. 2015) where the authors use a hybrid model for missing traffic volume estimation, k-nearest neighbour (Li & Parker 2014), and single-value decomposition (Xu et al. 2017) which the authors demonstrate on real-world air quality datasets.

3.4.2 Data Management Platform Architecture: Case Studies

Management frameworks for IoT/WSN follow a few main themes: edge computing; data integration techniques; cloud computing; and data analytics. Badidi et al. (2018) identify key features of an urban data stream management and processing pipeline as: facilitating real-time event detection; notification of alerts; mining the opinions of citizens regarding

the governance of their city; and building monitoring dashboards. The authors implement a prototype of the using the Kafka messaging platform. Whilst there are a number of systems that have been developed for managing data streams, there are few that focus on data quality management that have been fully implemented. [Ehrlinger et al. \(2019\)](#) present a data quality management methodology that uses machine learning algorithms to detect and correct data quality issues in real-time for industrial IoT applications, which is the closest to a functioning management system for WSN data in the literature.

3.5 Data Quality Reliability and Proactive Approaches

This sections looks at how future data quality can be improved through predictive modelling for sensor optimisation and proactive data quality management strategies. This includes understanding optimal sensor placement ([Kim et al. 2024](#)) and improving checks on incoming data ([Mohammadi & Taylor 2017](#)). This section differs from the previous, in that rather than trying to detect and correct errors, it is concerned with optimising the WSN to prevent errors from occurring in the first place.

3.5.1 Fault Tolerance

Fault tolerance in wireless sensor networks (WSNs) refers to the network's ability to continue functioning correctly even in the presence of failures or errors within some of its components. This includes the capability to handle node failures, communication errors, or environmental disruptions without significant loss of data or performance ([Guravaiah et al. 2020](#)). There are a number of methods involved in building fault tolerant WSNs, including: redundancy, error detection and correction, and self-healing mechanisms ([Guravaiah et al. 2020](#)). [El Khediri \(2022\)](#) present a comprehensive review of fault tolerance in WSNs, highlighting the importance of clustering protocols to ensure that data is not lost in the event of a node failure.

3.5.2 Preventative Measures

[Cheng et al. \(2018\)](#) explore the relationship between different data quality evaluation indicators to carry out a variety of data cleaning strategies. The results showed that the proposed strategies can improve data availability and reduce cleaning costs (computational overhead and time) for errors such as data loss, sample jitter, and gross error. The order of cleaning strategies was found to be important—starting with completeness cleaning (to repair lost data) followed by time-related cleaning (to eliminate sampling jitter) and finally correctness cleaning (to correct abnormal data) is found to be the most effective sequence. [Klein & Lehner \(2009\)](#) suggest a strategy of carefully surveying data quality restrictions and propagating quality information through data processing pipelines. They introduce jumping data quality windows to reduce data overhead and propose methods for data quality recording and storage. However, the experiments were tested on synthetic data and the authors acknowledge that further research is needed to validate the results on real-time, real-world data.

3.6 Challenges and Future Directions

3.6.1 Scalability

Scalability is a critical consideration in the development of data quality management systems for wireless sensor networks (WSNs), especially in the context of pedestrian monitoring in smart cities. As WSNs grow in size and complexity, several challenges arise that must be addressed to ensure effective and efficient data collection, processing, and analysis. Increasing the number of sensor nodes to cover larger urban areas can lead to higher data traffic, causing congestion and potential data loss. Dense networks also pose challenges in terms of interference and overlapping signals, which can degrade data quality [Akyildiz et al. \(2002\)](#). The volume of data generated by a large number of sensors means automated management processes are required, necessitating significant computational resources and efficient algorithms for real-time processing and storage [Gubbi et al. \(2013\)](#).

Ensuring data quality across a large and diverse network is complex, with issues such as data inconsistency, missing data, and noise becoming more pronounced as the network scales [Karkouch et al. \(2016\)](#). Real-time data processing is essential for applications like pedestrian monitoring, where timely information is critical, increasing the demand for real-time processing capabilities [Ahmad et al. \(2017\)](#). As networks scale, the likelihood of node failures and communication errors increases, making fault tolerance and reliable data transmission more challenging [Younis & Akkaya \(2008\)](#). Addressing these scalability challenges requires a combination of advanced data management techniques, energy-efficient protocols, and robust fault-tolerance strategies to ensure that WSNs can provide reliable and high-quality data for pedestrian monitoring in smart cities.

3.6.2 Integration

With the increasing proliferation of platforms such as the one proposed here it becomes increasingly important for them to be able to integrate with other systems. This requires centralised governance and standardisation. This exists at a high level with the Gemini principles ([Walters 2019](#)) and the National Digital Twin project ([National Digital Twin Programme \(NDTP\) 2024](#)) which deal with ethics and governance and are primarily driven by purpose (what do we want from digital twins). A more low-level data-focussed standards can be found in the Open Geospatial Consortium's CityGML and Moving Features standards ([OGC® Moving Features 2024](#), [CityGML 2024](#)), and the INSPIRE directive ([INSPIRE Knowledge Base - European Commission 2024](#)) from the European Commission are more data-focussed and policy focussed respectively. These standards are important for ensuring interoperability and scalability in smart city applications, and are essential for the development of a robust and scalable system for managing pedestrian data quality.

3.7 Summary

This literature review has explored the critical aspects of data quality in wireless sensor networks (WSNs) with a focus on pedestrian monitoring. Key dimensions such as accuracy, completeness, consistency, timeliness, and validity have been examined, highlighting the importance of maintaining high data quality for reliable pedestrian data.

Methodologies for data quality assessment, including statistical measures, machine learning approaches, and event detection techniques, have been reviewed. These methodologies are essential for real-time detection and monitoring of data anomalies, ensuring the reliability of pedestrian counts. The review also covered strategies for managing and improving data quality, such as automated imputation for missing data and denoising techniques, which are crucial for maintaining data integrity. Enhancements in WSN architecture aimed at improving data quality from the source were discussed, providing insights into building a scalable pedestrian data quality management system. Challenges and future directions emphasise the need for centralised governance and standardisation when building such platforms, ensuring interoperability and scalability in smart city applications. This research contributes to the development of a robust and scalable system for managing pedestrian data quality, facilitating more accurate and reliable urban mobility monitoring.

In summary, maintaining high data quality in WSNs for pedestrian monitoring requires a comprehensive approach, integrating advanced data quality assessment methodologies, real-time monitoring, and proactive management. This ongoing research aims to address current challenges and support the creation of effective, scalable data quality management systems for smart cities.

References

- Abu Alsheikh, M., Hoang, D. T., Niyato, D., Tan, H.-P. & Lin, S. (2015), 'Markov Decision Processes With Applications in Wireless Sensor Networks: A Survey', *IEEE Communications Surveys & Tutorials* **17**(3), 1239–1267.
- Aggarwal, C. C. (2017), *Outlier Ensembles*, Springer International Publishing, Cham, pp. 185–218.
- Ahmad, S., Lavin, A., Purdy, S. & Agha, Z. (2017), 'Unsupervised real-time anomaly detection for streaming data', *Neurocomputing* **262**, 134–147.
- Akyildiz, I. F., Su, W., Sankarasubramaniam, Y. & Cayirci, E. (2002), 'Wireless sensor networks: A survey', *Computer networks* **38**(4), 393–422.
- Aminikhanghahi, S. & Cook, D. J. (2017), 'A survey of methods for time series change point detection', *Knowledge and information systems* **51**(2), 339–367.
- Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. (1999), 'OPTICS: Ordering points to identify the clustering structure', *ACM SIGMOD Record* **28**(2), 49–60.
- Ardagna, D., Cappiello, C., Samá, W. & Vitali, M. (2018), 'Context-aware data quality assessment for big data', *Future Generation Computer Systems* **89**, 548–562.
- Atzori, L., Iera, A. & Morabito, G. (2010), 'The internet of things: A survey', *Computer networks* **54**(15), 2787–2805.
- Badidi, E., El Neyadi, N., Al Saeedi, M., Al Kaabi, F. & Maheswaran, M. (2018), 'Building a data pipeline for the management and processing of urban data streams', *Handbook of Smart Cities: Software Services and Cyber Infrastructure* pp. 379–395.
- Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009), 'Methodologies for data quality assessment and improvement', *ACM Computing Surveys (CSUR)* **41**(3), 1–52.
- Batini, C. & Scannapieco, M. (2016), 'Data and information quality', *Cham, Switzerland: Springer International Publishing*.
- Benabbas, A., Grawunder, M. & Nicklas, D. (2023), Context-aware Outlier Detection for Sensor Data Stream Processing, in '2023 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)', IEEE, pp. 540–545.
- Bisdikian, C., Branch, J., Leung, K. K. & Young, R. I. (2009), A letter soup for the quality of information in sensor networks, in '2009 IEEE International Conference on Pervasive Computing and Communications', IEEE, pp. 1–6.
- Blázquez-García, A., Conde, A., Mori, U. & Lozano, J. A. (2021), 'A review on outlier/anomaly detection in time series data', *ACM Computing Surveys (CSUR)* **54**(3), 1–33.

- Bocchi, G. & Facchini, A. (2016), Living at the edge of chaos: A complex systems view of cities, in 'Cities in the 21st Century', Routledge, pp. 129–138.
- Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. (2000), LOF: Identifying density-based local outliers, in 'Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data', SIGMOD '00, Association for Computing Machinery, New York, NY, USA, pp. 93–104.
- Cai, S., Chen, J., Chen, H., Zhang, C., Li, Q., Shi, D. & Lin, W. (2023), 'Minimal Rare Pattern-Based Outlier Detection Approach For Uncertain Data Streams Under Monotonic Constraints', *The Computer Journal* **66**(1), 16–34.
- Carreño, A., Inza, I. & Lozano, J. A. (2020), 'Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework', *Artificial Intelligence Review* **53**(5), 3575–3594.
- Chandola, V., Banerjee, A. & Kumar, V. (2009), 'Anomaly detection: A survey', *ACM Computing Surveys (CSUR)* **41**(3), 1–58.
- Chen, L., Grimstead, I., Bell, D., Karanka, J., Dimond, L., James, P., Smith, L. & Edwardes, A. (2021), 'Estimating Vehicle and Pedestrian Activity from Town and City Traffic Cameras', *Sensors (Basel)* **21**(13), 4564.
- Chen, X. (2021), 'Fault Detection Method and Simulation Based on Abnormal Data Analysis in Wireless Sensor Networks', *Journal of Sensors* **2021**(1), 6155630.
- Cheng, H., Feng, D., Shi, X. & Chen, C. (2018), 'Data quality analysis and cleaning strategy for wireless sensor networks', *EURASIP Journal on Wireless Communications and Networking* **2018**(1), 61.
- CityGML (2024), <https://www.ogc.org/standard/citygml/>.
- Dasu, T. & Johnson, T. (2003), *Exploratory Data Mining and Data Cleaning*, John Wiley & Sons.
- Ding, M., Chen, D., Xing, K. & Cheng, X. (2005), Localized fault-tolerant event boundary detection in sensor networks, in 'Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.', Vol. 2, IEEE, pp. 902–913.
- Ehrlinger, L., Haunschmid, V., Palazzini, D. & Lettner, C. (2019), A DaQL to Monitor Data Quality in Machine Learning Applications, in S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. M. Tjoa & I. Khalil, eds, 'Database and Expert Systems Applications', Springer International Publishing, Cham, pp. 227–237.
- Ehrlinger, L. & Wöb, W. (2017), Automated Data Quality Monitoring, in 'ICIQ'.
- El Khediri, S. (2022), 'Wireless sensor networks: A survey, categorization, main issues, and future orientations for clustering protocols', *Computing* **104**(8), 1775–1837.

- Elkhodr, M. & Alsinglawi, B. (2020), 'Data provenance and trust establishment in the Internet of Things', *Security and Privacy* **3**(3), e99.
- Ester, M., Kriegel, H.-P. & Xu, X. (1996), 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'.
- Fizza, K., Jayaraman, P. P., Banerjee, A., Georgakopoulos, D. & Ranjan, R. (2022), Age of Data Aware Internet of Things Applications, in '2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)', pp. 399–404.
- Gorshenin, A., Kozlovskaya, A., Gorbunov, S. & Kochetkova, I. (2024), 'Mobile network traffic analysis based on probability-informed machine learning approach', *Computer Networks* **247**.
- Gruenwald, L., Chok, H. & Aboukhamis, M. (2007), Using Data Mining to Estimate Missing Sensor Data, in 'Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)', pp. 207–212.
- Gubbi, J., Buyya, R., Marusic, S. & Palaniswami, M. (2013), 'Internet of Things (IoT): A vision, architectural elements, and future directions', *Future Generation Computer Systems* **29**(7), 1645–1660.
- Guravaiah, K., Kavitha, A., Velusamy, R. L., Guravaiah, K., Kavitha, A. & Velusamy, R. L. (2020), Data Collection Protocols in Wireless Sensor Networks, in 'Wireless Sensor Networks - Design, Deployment and Applications', IntechOpen.
- Heinrich, K., Roth, A. & Zschech, P. (2019), Everything counts: A Taxonomy of Deep Learning Approaches for Object Counting, in 'ECIS'.
- Iglewicz, B. & Hoaglin, D. C. (1993), *Volume 16: How to Detect and Handle Outliers*, Quality Press.
- Immonen, A., Pääkkönen, P. & Ovaska, E. (2015), 'Evaluating the quality of social media data in big data architecture', *IEEE Access* **3**, 2028–2043.
- INSPIRE Knowledge Base - European Commission (2024), https://knowledge-base.inspire.ec.europa.eu/index_en.
- ISO (2023), 'ISO/DIS 8000-210(en), Data quality — Part 210: Sensor data: Data quality characteristics', <https://www.iso.org/obp/ui/ru/#iso:std:84274:en>.
- Karkouch, A., Mousannif, H., Al Moatassime, H. & Noel, T. (2016), 'Data quality in internet of things: A state-of-the-art survey', *Journal of Network and Computer Applications* **73**, 57–81.
- Khatri, V. & Brown, C. V. (2010), 'Designing data governance', *Communications of the ACM* **53**(1), 148–152.

- Kim, T., Kim, J., Kim, J. & Oh, S. (2024), 'Optimization of number of wireless temperature sensors using clustering algorithm for deep learning algorithm-based Kimchi quality prediction', *Journal of Food Engineering* **367**.
- Klein, A. & Lehner, W. (2009), 'Representing data quality in sensor data streaming environments', *Journal of Data and Information Quality (JDIQ)* **1**(2), 1–28.
- Kumar, M., Singh, P. K., Maurya, M. K. & Shivhare, A. (2023), 'A survey on event detection approaches for sensor based IoT', *Internet of Things* **22**, 100720.
- Langton, C. G. (1990), 'Computation at the edge of chaos: Phase transitions and emergent computation', *Physica D: Nonlinear Phenomena* **42**(1-3), 12–37.
- Li, F., Nastic, S. & Dustdar, S. (2012), Data quality observation in pervasive environments, in '2012 IEEE 15th International Conference on Computational Science and Engineering', IEEE, pp. 602–609.
- Li, Y. & Parker, L. E. (2014), 'Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks', *Information Fusion* **15**, 64–79.
- Ma, X., Yang, D., Tang, S., Luo, Q., Zhang, D. & Li, S. (2004), Online mining in sensor networks, in 'Network and Parallel Computing: IFIP International Conference, NPC 2004, Wuhan, China, October 18-20, 2004. Proceedings', Springer, pp. 544–550.
- Mansouri, T., Sadeghi Moghadam, M. R., Monshizadeh, F. & Zararavasan, A. (2023), 'IoT data quality issues and potential solutions: A literature review', *The Computer Journal* **66**(3), 615–625.
- Mehmood, H., Khalid, A., Kostakos, P., Gilman, E. & Pirttikangas, S. (2024), 'A novel Edge architecture and solution for detecting concept drift in smart environments', *Future Generation Computer Systems* **150**, 127–143.
- Mohammadi, N. & Taylor, J. E. (2017), Smart city digital twins, in '2017 IEEE Symposium Series on Computational Intelligence (SSCI)', IEEE, pp. 1–5.
- National Digital Twin Programme (NDTP) (2024), <https://www.gov.uk/government/collections/the-national-digital-twin-programme-ndtp>.
- Naumann, F. (2002), *Quality-Driven Query Answering for Integrated Information Systems*, Springer.
- Nguyen, H. D., Tran, K. P., Thomassey, S. & Hamad, M. (2021), 'Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management', *International Journal of Information Management* **57**, 102282.
- Nickerson, R. C., Varshney, U. & Muntermann, J. (2013), 'A method for taxonomy development and its application in information systems', *European Journal of Information Systems* **22**, 336–359.

OGC® *Moving Features* (2024), <https://www.ogc.org/standard/movingfeatures/>.

Perez-Castillo, R., Carretero, A. G., Rodriguez, M., Caballero, I., Piattini, M., Mate, A., Kim, S. & Lee, D. (2018), Data quality best practices in IoT environments, in '2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)', IEEE, pp. 272–275.

Pimentel, M. A., Clifton, D. A., Clifton, L. & Tarassenko, L. (2014), 'A review of novelty detection', *Signal processing* **99**, 215–249.

Ramaswamy, S., Rastogi, R. & Shim, K. (2000), Efficient algorithms for mining outliers from large data sets, in 'Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data', SIGMOD '00, Association for Computing Machinery, New York, NY, USA, pp. 427–438.

Rusli, N., Ling, G. H. T., Hussain, M. H. M., Salib, N. S. M., Bakar, S. Z. A. & Othman, M. H. (2023), 'A review on worldwide urban observatory systems' data analytics themes: Lessons learned for Malaysia Urban Observatory (MUO)', *Journal of Urban Management*.

Shukla, R. M. & Sengupta, S. (2020), 'Scalable and robust outlier detector using hierarchical clustering and long short-term memory (Lstm) neural network for the internet of things', *Internet of Things* **9**, 100167.

Sicari, S., Cappiello, C., De Pellegrini, F., Miorandi, D. & Coen-Porisini, A. (2016), 'A security-and quality-aware system architecture for Internet of Things', *Information Systems Frontiers* **18**, 665–677.

Şimsek, M., Kök, İ. & Özdemir, S. (2024), 'DeepFogAQ: A fog-assisted decentralized air quality prediction and event detection system', *Expert Systems with Applications* **251**.

Smith, L. & Turner, M. (2019), Building the Urban Observatory: Engineering the largest set of publicly available real-time environmental urban data in the UK, in 'Geophysical Research Abstracts', Vol. 21.

Tang, J., Zhang, G., Wang, Y., Wang, H. & Liu, F. (2015), 'A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation', *Transportation Research Part C: Emerging Technologies* **51**, 29–40.

Teh, H. Y., Kempa-Liehr, A. W. & Wang, K. I.-K. (2020), 'Sensor data quality: A systematic review', *Journal of Big Data* **7**(1), 11.

Tortora, G. J. & Derrickson, B. H. (2018), *Principles of Anatomy and Physiology*, John Wiley & Sons.

Tuychiev, B. (2023), 'A Comprehensive Introduction to Anomaly Detection', <https://www.datacamp.com/tutorial/introduction-to-anomaly-detection>.

- VanderPlas, J. T. (2018), 'Understanding the lomb–scargle periodogram', *The Astrophysical Journal Supplement Series* **236**(1), 16.
- Walters, A. (2019), 'Gemini Principles', <https://www.cdbb.cam.ac.uk/DFTG/GeminiPrinciples>.
- Wang, R. Y. & Strong, D. M. (1996), 'Beyond accuracy: What data quality means to data consumers', *Journal of management information systems* **12**(4), 5–33.
- Woodall, P., Borek, A. & Parlikad, A. K. (2013), 'Data quality assessment: The hybrid approach', *Information & management* **50**(7), 369–382.
- Xu, P., Ruan, W., Sheng, Q. Z., Gu, T. & Yao, L. (2017), Interpolating the Missing Values for Multi-Dimensional Spatial-Temporal Sensor Data: A Tensor SVD Approach, in 'Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services', MobiQuitous 2017, Association for Computing Machinery, New York, NY, USA, pp. 442–451.
- Younis, M. & Akkaya, K. (2008), 'Strategies and techniques for node placement in wireless sensor networks: A survey', *Ad Hoc Networks* **6**(4), 621–655.
- Zhang, Q., Ye, M. & Deng, X. (2022), 'A novel anomaly detection method for multimodal WSN data flow via a dynamic graph neural network', *Connection Science* **34**(1), 1609–1637.
- Zhang, Y., Meratnia, N. & Havinga, P. (2010), 'Outlier detection techniques for wireless sensor networks: A survey', *IEEE communications surveys & tutorials* **12**(2), 159–170.