Continued building in data sharing platform

Ganesh college of Engineering

Submitted by: C.Nandhini
T.Swathi
S.Sevvanthi
M.Arivumathi

strategies to mitigate pollution. The goal of these systems is to monitor air quality variables, in order to provide authorities and citizens with important information about the current level of gases and particles in different areas of the city. This information can be used to take decisions aimed at preventing the negative impact of these pollutants on human health. With the aim of utilizing the monitoring data to assess, predict and reduce the pollutant levels, the environmental agencies have developed regulations that include key aspects, such as data quality objectives and indicators that must be accomplished (UNION et al., 2008; EPA, 2017).

Traditionally, air quality monitoring systems consist of a set of expensive robust stations that require on-site calibration and maintenance. Due to the high cost, the number of monitoring stations is usually low, leading to a low spatial resolution of the data (Röösli et al., 2000; Lin et al., 2020b). In the past few years, however, the development of the monitoring systems has come hand in hand with the development of new technological paradigms such as the Internet of Things (IoT), thus allowing the deployment of a larger number of air quality monitoring systems (Múnera et al., 2021). IoT is a paradigm of systems that allows connectivity and information exchange between heterogeneous objects (uniquely identifiable), in order to capture and process information ubiquitously for decision-making and action on a given context (Atzori et al., 2017). Hence, IoT-based air quality monitoring systems use low-cost sensors, thus enabling the massive development of sensor systems with lower associated costs, and allowing permanent and real-time access to the gathered data.

The data generated by IoT systems, however, has been considered unreliable for two main reasons. On one side, they utilize low-cost sensors, which lack the accuracy and precision of the robust stations. The second reason stems from the fact that these systems are exposed to many endangering factors, since their applications usually involve wide deployments and open platforms (Karkouch et al., 2016; Liu et al., 2019). These conditions have led to a significant concern regarding the data reliability and trustworthiness of the IoT-based monitoring systems. Particularly, in the context of air-quality monitoring, several researchers argue that the use of low-cost sensors is generating unreliable data (Kumar et al., 2015; Castell et al., 2017; Manikonda et al., 2016). This situation poses a new challenge in the context of smart cities and IoT: it is necessary to assess and improve the quality of the data obtained through the IoT systems, in order to establish their reliability and provide useful information to decision makers.

The study of Data Quality (DQ) emerged from the field of information systems, where large amount of data are needed to be stored in databases and managed by such information systems. Authors in Wang (1996) proposed a set of dimensions that were more important for data consumers in this field. Because of the importance of data, this concept has been adopted by other applications and fields. Specifically, in the context of IoT systems, the analysis of DQ has become relevant in order to guarantee the reliability of the data to the decision makers. Authors Liu et al. (2019) and Karkouch et al. (2016) have both conducted a systematic literature review and a state-of-the-art review of DQ in IoT, and have discussed how DQ has been addressed in IoT applications. They have also identified the challenges and most prominent research sub-fields of DQ in IoT, which include the most commonly used dimensions, endangering factors, and enhancing methods.

DQ analysis in the field of air quality monitoring systems is a fairly new topic, since massive low-cost systems have become popular in the past few years. Even though there are specific definitions for the DQ expected out of these systems, provided by the EPA (2017) and the EU (UNION et al., 2008), the studies on this topic are limited to the dimensions addressed by the deployed solutions, and do not consider the relationship between the DQ dimension and the indicator suggested by the standardization entities. In this context, this study aims at providing an overview of how DQ has been addressed in the implementation of IoT-based air quality monitoring systems. Moreover, the goal is also to find the relationship between the Data Quality Indicators (DQI) and Data Quality Objectives (DQO) defined by EPA, and the DQ dimensions traditionally used. Our contributions are hence summarized as follows.

- We review and analyze the existing guidelines for assessing DQ in air quality monitoring systems for proposing a mapping between the DQ indi-

cators (from guidelines) and the DQ dimensions (from DQ field).
- We analyze the main DQ enhancement techniques and identify how these techniques affected the DQ dimensions.
- We develop a systematic mapping study to determine the state of the evaluation of DQ in IoT-based air quality monitoring systems. We use our proposed mapping between DQI, enhancement techniques and DQ dimensions to answer the research questions that guide our systematic mapping study.
- We highlight some challenges that must be addressed in order to improve data quality in IoT-based air quality monitoring systems.

The document is organized as follows. Section 2 describes data quality principles. Section 3 presents data quality in the context of air quality, where a description of data quality indicators and objectives is discussed. Section 4 highlights the most common DQ enhancement techniques used in IoT-based air quality monitoring systems. Section 5 shows the steps for the systematic review process. Section 6 describes the results found in the systematic mapping study. Finally, Sections 7 and 8 present the discussion and conclusions, respectively.

2 Data Quality Principles

It is common to find a definition of DQ from the consumer's point of view, where this trend is based on the treatment of data as a product. In Wang (1996), it is defined as "data that are fit for use by data consumers"; similar definitions are found in Karkouch et al. (2016) and Liu et al. (2019). According to Karkouch et al. (2016), the data consumer requires data to fulfill certain criteria that are essential for the tasks at hand. Being data a product, DQ is a multi-faceted concept since users have different expectations out of it. Thus, the DQ analysis has been divided into dimensions, where each dimension stands for an attribute that is important to the data consumer, or the application. After studying the term DQ in the field of IoT, we have identified several dimensions that can be relevant to the analysis of DQ.

Tables 1, 2, 3, 4, 5, and 6 present the most relevant DQ dimensions as well as their definitions and proposed evaluation metrics. In these tables, the first column is dimension name, and the second column includes a short definition of the dimension, which is a result of the review of several sources. It can be evidenced how a dimension can take several names, but it will have the same definition over different sources. Finally, the third column shows a formula or metric to

Table 1 DQ dimensions related to data values

Dimension	Definition	Metric
Accuracy	"The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of us" (Liu et al., 2019; ISO 25000 Portal, 2019).	$\alpha = \frac{\lvert v_m - v \rvert}{v}$, $DQ_{\text{accu}} = \max(0, 1-\alpha)$, where v_m is the measured value, and v is the value accepted as true.
Precision	"The Precision is degree to which further measurements of the same phenomenon in a close time instant provides the same or similar results" (Sicari et al., 2018). It can be represented as the standard deviation of the measurement.	$DQ_{\text{prec}} = 1 - \frac{\sqrt{\frac{\sum_{i=1}^{n}(v_m - \bar{v}_m)^2}{n-1}}}{\bar{v}_m}$, where $\bar{v}_m$ is the mean of the measurement over n observations. The coefficient of variation is used to obtain a relative value.
Confidence	"The statistical error ε such that $[v-\varepsilon, v+\varepsilon]$ contains the real value with a confidence probability of p" (Klein & Lehner, 2009; Karkouch et al., 2016). It represents the statistical measurement error due to random environmental interference such as vibrations or shocks.	$\varepsilon = z \cdot \frac{\sigma}{\sqrt{n}}, n \geq 30$, $DQ_{\text{conf}} = 1 - \frac{\varepsilon}{\bar{v}_m}$, where z is the statistical value of the Z-distribution for a given confidence interval, while σ is the standard deviation of the population. Dividing by the mean give the relative margin of error.

Table 2 DQ dimensions related to the amount of data

Dimension	Definition	Metric
Completeness	"The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use" (Liu et al., 2019; ISO 25000 Portal, 2019). "The extent to which all expected data is provided by IoT services."	$DQ_{\text{comp}} = \frac{\#\text{ValidCollectedValues}}{\#\text{ExpectedValues}}$
Data volume	"The number of raw data items (values) available for use to compute a result data item (in a stream query or sub-query)" (Karkouch et al., 2016). We can define it as the number of collected values retrieved at a time instant t.	$DQ_{\text{dvol}} = \#\text{CollectedValues}(t)$
Redundancy	Data redundancy or repeated data is accounted as the amount of data items that have the same timestamp. This might be caused by abnormal network transmission that makes data to be transmitted or received multiple times (Guo & Liu, 2015).	$DQ_{\text{dupl}} = 1 - \frac{\#\text{RepeatedTimestamps}}{\#\text{CollectedValues}}$

evaluate each DQ dimension ($DQ_{\text{dimension}}$ value), and each of them has been adapted such that every value is in the range between 0 and 1 (0 for low quality and 1 for high quality).

These dimensions can be classified according to different categories. Table 1 shows the dimensions related to the specific value of the data (and its error). These dimensions include as follows: *Precision*, *Accuracy*, and *Confidence*. A second category of DQ dimensions is presented in Table 2, where the amount of data is considered. This category includes the *Data volume*, *Completeness*, and *Redundancy* dimensions. The third category gathers the time-related DQ dimensions as presented in Table 3. This category includes the *Timeliness* and *Accessibility* dimensions. Table 4 shows the dimensions that take into consideration the relationship among the data, such as *Concordance*, *Artificiality*, and *Interpretability*. Finally,

Dimension	Definition	Metric
Timeliness	"The degree to which data has attributes that are of the right age in a specific context of use" (Liu et al., 2019; ISO 25000 Portal, 2019). "The extent to which an observation for the object is updated at a desired time of interest" and it is related to terms like currentness, currency, volatility, latency, freshness, data rate, delay, frequency, or promptness.	$\text{Currency} = \text{time} - \text{Timestamp}(v_{ni})$, $DQ_{\text{time}} = \max\left(0, \frac{\text{Currency}}{\text{Volatility}}\right)$ As proposed in Batini et al. (2009), the timeliness can be calculated in terms of the currency and volatility, the latter defined as the time during which data remain valid.
Accessibility	Different from utility, Karkouch et al. (2016) define ease of access as the availability and easiness of retrieving data, while the accessibility is regarded as a category of the DQ dimensions, and defined as "how accessible data are for data consumers."	In Batini et al. (2009) it is calculated based on the time it takes to provide a result for a query: $DQ_{\text{acce}} = \max\left(0, \frac{\text{DelivTime} - \text{ReqTime}}{\text{DeadTime} - \text{ReqTime}}\right)$ where ReqTime is the request time of the query, DelivTime is the respons time, and DeadTime is the deadline time.

Table 4 DQ dimensions relation between data and context

Dimension	Definition	Calculation
Concordance	"The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use" (Liu et al., 2019). It is related to concepts like consistency.	We propose to calculate concordance as the absolute value of the Pearson's correlation coefficient between variables x_0 and x_i: $DQ_{\text{conc}} = \lvert\rho_{x_0 x_i}\rvert$
Interpretability	The interpretability tells whether data is clear in meaning and format (Karkouch et al., 2016), it can be improved by using annotations. According to Batini and Scannapieca (2006), it concerns to the documentation and metadata that are available to correctly interpret the meaning and properties of data sources.	$DQ_{\text{inte}} = \begin{cases} 1, & \text{dataset is annotated,} \\ & \text{there is metadata} \\ & \text{or there is documentation.} \\ 0, & \text{otherwise.} \end{cases}$
Artificiality	Data artificiality is proposed by Kuemper et al. (2018) and it determines whether data originates directly from a hardware sensor or whether data is estimated after the application of techniques such as interpolation, aggregation, and fusion.	$DQ_{\text{arti}} = \begin{cases} 1, \text{real sensor data.} \\ 0, \text{artificial data.} \end{cases}$

Table 5 DQ Dimensions related to the system

Dimension	Definition	Calculation
Utility	"The degree to which data can be accessed in a specific context of use" (Liu et al., 2019), which is related to the data accessibility dimension. To calculate the utility, it is necessary to keep track of user's or application's interactions with data in the form of queries or visualizations.	$DQ_{\text{util}} = \begin{cases} 1, & \text{if data was accessed at} \\ & \text{least once or it is} \\ & \text{provided in push mode.} \\ 0, & \text{otherwise.} \end{cases}$
Trust	"The probability by which data are suitable to be included in a specific process providing value" (Sicari et al., 2016), it is associated with source reputation and reliability. The source reputation is the sum of two factors, the content reputation and the owner reputation, where the former depends on the number of times the source fails to provide a good answer, while the latter depends on the history of the organization that owns the data, e.g., a node can be given a reputation based on the quality of its provided data.	Our proposal to compute the trustworthiness depends on two variables. The first is the source reputation (given by the user or the IoT system) and takes two values $\{0, 1\}$, where 0 is bad reputation of the source and 1 is a good reputation of it. The second variable is based on the correctness, i.e., the validity of data provided by the source. μ is a weight that can be adjusted to give more importance to the reputation or to the validity. $DQ_{\text{trus}} = \mu \cdot \text{Reputation} + (1 - \mu) \cdot \frac{\sum_{i=1}^{n} DQ_{\text{vali}}(i)}{n}$ (1)
Access security	Authors in Karkouch et al. (2016) define it as to secure data to protect its privacy and confidentiality. In Sicari et al. (2018) and Sicari et al. (2016) it is found to be also related to authentication and integrity. In essence, data (specially sensitive data) should remain confidential and private from its generation at the source to its storage in a database. Data authentication is related to data integrity and source authentication, where the IoT system can (and should) verify the origin of the data and confirm its integrity.	If mechanisms to preserve access security exists, these attributes can be evaluated as 1, or 0 otherwise. An example is the use of cryptographic protocols such as Transport Layer Security (TLS). Another approach, proposed by Sicari et al. (2016) to evaluate access security is based on identifying attacks and countermeasures for these attacks.

Table 6 DQ dimensions relation between data and context

DQ dimension	Definition	Calculation
Validity	In Li et al. (2012), validity is defined as a metric to evaluate the correctness of an observation. It can be seen as a set of rules or constraints that data should comply to be correct. Authors in Kuemper et al. (2018) define plausibility as whether a received data source information makes sense regarding the probabilistic knowledge about what it is measuring. The following can be validity rules related to the correctness of the data: (VR_1) data is within allowed range, (VR_2) data consistency is greater than 90%, (VR_3) data accuracy is greater than 90%, (VR_4) data precision is greater than 60%.	As proposed by Li et al. (2012), the validity can be calculated as a series of *and* operations over the rules, where m is the number of rules. $DQ_{\text{vali}} = \bigwedge_{i=1}^{m} VR_i(o)$

Table 5 presents the last category, which considers the dimensions that are related to the system, specifically *Utility*, *Trust*, and *Access security*.

As stated earlier, the relevance of each dimension depends on the specific application of the system and on how the data is going to be utilized. In that sense, a unique DQ value has not been defined in order to decide whether a data should be used. The validity dimension, however, aims at providing the system with the flexibility to define which dimensions are relevant to the DQ of the specific context (see Table 6).

3 Data Quality in the Context of Air Quality Estimation

In the context of air quality monitoring systems, *The European Parliament And The Council* has established Data Quality Objectives and Data Quality Indicators in the *DIRECTIVE 2008/50/EC* (UNION et al., 2008) guideline, while the *Environmental Protection Agency (EPA)* in the USA proposed the *Quality Assurance Handbook for Air Pollution Measurement Systems* (EPA, 2017) guideline. These documents define Data Quality Objectives (DQO) as the level of accepted threshold of the Data Quality Indicator (DQI), i.e., attributes of data quality. A close examination of these guidelines can lead to identify and match some of these indicators to the DQ dimensions previously discussed. We present below each DQI and its relation with the DQ dimensions.

- **Uncertainty**: According to JCGM (2008), it is "a parameter associated with the result of a measurement that characterizes the dispersion of the values that could be reasonably attributed to the measurand." The authors also state that uncertainty is a generic term used to describe the sum of all sources of error associated with an environmental data operation. Uncertainty has two components, namely population uncertainty and measurement uncertainty. The former is related to the representativeness of the sample, while the latter is related to the precision, bias, and detection limit (EPA, 2017).
  Regarding the DQO for particulate matter pollutants, the maximum allowed uncertainty for fixed measurements (i.e., robust monitoring stations) is 25%, while for indicative measurements (e.g., low-cost sensors measurements) is 50% (UNION et al., 2008). Based on this definition, this indicator is related to accuracy and confidence dimensions.
- **Minimum data capture**: It has a limit of 90%, which means that the maximum number of missing values within one measurement period is 10% of the expected values (UNION et al., 2008). This indicator is related to completeness dimension.
- **Minimum time coverage**: This indicator for measurements of pollutants such as particulate matter (PM10/PM2.5) has a limit of 14% (1-day measurement per week at random, evenly distributed over the year, which would result on roughly 52 1-day measurements per year, or 8 weeks evenly distributed over the year, which would result on roughly 56 1-day measurements per year) (UNION et al., 2008). This indicator is related to timeliness and completeness dimensions.

- **Minimum number of sampling points**: This indicator is defined in UNION et al. (2008) for fixed measurements, and it depends on the population of the specific area. For instance, a zone such as the Aburra Valley in Antioquia-Colombia, with about 4 million inhabitants in 2020 (Proantioquia et al., 2020), requires a minimum number of sampling points of 11. This indicator is related to data volume dimension.
- **Precision**: It represents the random component of error and is a measure of agreement among repeated measurements of the same property, under identical or very similar conditions (EPA, 2017). It is usually estimated as a derivation of the standard deviation. This indicator is part of the uncertainty components and matches the precision DQ dimension.
- **Bias**: This indicator is a component of the uncertainty and represents the systematic distortion of a measurement process that causes error in one direction. It is determined by the estimation of positive and negative deviation from the true value (EPA, 2017). This definition matches the accuracy DQ dimension.
- **Detection limit**: It is the minimum concentration of a pollutant that can be distinguished from zero (absence of the pollutant) by a single measurement at a stated level of probability (EPA, 2017). This indicator can be sorted within the validity DQ dimension.
- **Accuracy**: It is defined as data quality indicator in EPA (2017) as "measure of the overall agreement of a measurement to a known value and includes a combination of random error (precision) and systematic error (bias) components of both sampling and analytical operations." The guide recommends to use bias and precision when possible, otherwise, use accuracy as the measurement uncertainty. This indicator matches the dimension of the same name.
- **Representativeness:** In handbook (EPA, 2017), it is defined as a measurement of the population component of uncertainty and refers to "the degree to which data accurately and precisely represents the frequency distribution of a specific variable in the population". According to the guide, it does not matter how precise or unbiased the measurement values are, whether a site is unrepresentative of the population that is presumed to represent. Representativeness depends on factors such as the amount of sampling points (network size), frequency of sampling, and sampling schedule. Thus, this indicator can match timeliness and data volume DQ dimensions, as well as the "minimum number of sampling points" and "minimum time coverage", which are discussed in the guide (UNION et al., 2008).
- **Comparability:** In the EPA handbook (EPA, 2017), this indicator is defined as "a measure of the confidence with which one dataset or method can be compared to another, considering the units of measurement and applicability to standard statistical techniques". For example, if there are two datasets retrieved from monitoring stations and low-costs sensors, it is expected that both of them are comparable. This indicator can match the concordance DQ dimension.
- **Completeness:** This indicator (from EPA (2017)) directly matches definition of the data completeness DQ dimension as the ratio of obtained valid data to the expected data. EPA requires 75% data to be complete.

4 DQ Enhancement Techniques

This section describes the most used data quality enhancement techniques in IoT-based air quality monitoring systems. We found four main categories, namely Data calibration, Data Interpolation, Data aggregation/fusion, and Outlier Detection as described to follow.

4.1 Data Calibration

Low-cost sensor calibration is essential due to collected data can be affected by noise and abnormalities. However, sensor manufacturers do not often provide direct means of sensor calibration, since it is not intended for low measurements, and is under specific humidity and temperature settings (Hasenfratz et al., 2012). Moreover, a calibrated sensor can suffer of sensor drift due to it can last several years after deployment (Barcelo-Ordinas et al., 2018). Hence, automatic or additional calibration is a needed in order to

overcome the mentioned limitations. Common calibration approaches for low-cost air quality sensors are done in laboratory with artificial pollutants, as well as in field, where the sensors are located close to fixed reliable stations. Field calibration has the disadvantage of dependency on weather conditions. Therefore, different reference measurements with several weather conditions (e.g., temperature and humidity settings) are needed for a more accurate calibration process (Hasenfratz et al., 2012).

Existing works have proposed new approaches based on traditional calibration. A node-to-node calibration approach was proposed in Kizel et al. (2018). It consists in calibrating only one sensor in a chain, by using reference measurements. Then, the rest of sensors are calibrated sequentially one against the other. This approach is suitable for distributed sensor networks. Other work uses Simple Linear Regression (SLR), Multi Linear Regression (MLR), and Artificial Neural Network (ANN) for calibration (Okafor et al., 2020). One feature (i.e., measurements from one sensor) is used in a SLR model, where each sensor is calibrated individually to adjust the bias. On the other hand, MLR and ANN models use all available features and a subset of features found by an Exhaustive Feature Selection method. Another approach is to place the sensor to be calibrated and the reference sensor in a hardboard box as in Rajasegarar et al. (2014b). The authors performed cubic polynomial fit with minimum error. A similar procedure was performed in Carratu et al. (2020), where a particle generator was used, and the sensors were previously synchronized. The authors also used the cubic polynomial fitting for each sensor.

4.2 Data Interpolation

Data interpolation can be understood as the process to generate new data with the aim to improve spatial or temporal resolution of a variable under supervision. Air-quality monitoring at local scale requires spatio-temporal integration to interpolate data. Urban environments can have large variations at small scale, where traditional interpolation methods fail to obtain reliable data. A solution is the use of high-density networks, by using low-cost sensors in order to monitor variable data at local scale (Alavi-Shoshtari et al., 2013). Low-cost sensors offer finer resolution of spatio-temporal data, which can complement existing air-quality monitoring stations. However, in order to address data quality from low-cost sensors, several interpolation methods have been proposed. Spatial interpolation is a common method used to predict spatio-temporal distributions in outdoors. Spatial interpolation relates air-quality measurements to their locations in order to predict pointwise data. It increases data availability across space and time. Existing spatial interpolation algorithms include nearest neighbor, spatial averaging, inverse distance weighting, and Kriging. The most used is the Kriging method, which produces best linear unbiased estimation of air-quality data (Li et al., 2018).

4.3 Data Agregation/Fusion

Data generated by several low-cost sensors can have uncertainties since various sensors have different technical performance. Data from only one sensor cannot satisfy the needs in terms of resolution and accuracy. Hence, accurate measurements can be obtained when data from different sensors (i.e., a multisensor system) is fused (Lin et al., 2020a). Data fusion was defined by Joint Directors of Laboratories in 1991 as "a the process of dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats as well as their significance" (White, 1991).

Data fusion systems have the advantage of expanding coverage in terms of space and time, as well as to improve performance, and spatial-temporal resolution (Lin et al., 2020a). Calibration errors can be reduced by considering measurements from several sensors and multivariate regression. It helps reducing uncertainty of calibration parameters (Barcelo-Ordinas et al., 2018). Existing works have proposed data fusion methods. For example, a data-fusion framework based on Optimum Linear Data Fusion theory (based on the least squares method) and Kriging method (to estimate the spatial-temporal data) was proposed in Lin et al. (2020a). Another approach merge sensor data with environmental factors in a calibration equation by using linear regression and artificial neural networks (Okafor et al., 2020).

4.4 Outlier Detection

Outliers refer to those data points that are far from the expected pattern in the data, and cannot be explained by a model. They are not consistent in space and time with the remaining set of observations (Jain and Shah, 2017), and must be distinguished from noise and missing values (Chung & Kim, 2020). Outliers are also known as anomalies, and can be caused by interference and sensors malfunction. Increasing IoT applications has imposed the need to develop accurate outlier detection methods since there is an increase in the collected data. In addition, several IoT applications use low-cost sensors, which are more prone to outliers with a high variability. Hence, it has been reported that low-cost sensor data suffers of low accuracy and precision, as well as low correlation with the reference (Fang & Bate, 2017b). Removing outliers from datasets at preprocessing stages improves the performance of machine learning algorithms (Chung & Kim, 2020). Several approaches have addressed outlier detection by using threshold values, where if a data point is greater than that value, it is considered an outlier. However, the threshold selection is an issue since it can be very subjective (Chen et al., 2018). Several solutions for outlier detection have been studied such as density-based, distance-based, and neural network-based methods (Huang et al., 2020b).

4.5 Relation Between DQ Enhancement Techniques and Dimensions

Figure 1 depicts the relationship between air quality indicators, data quality dimensions, and data quality enhancing techniques. This mapping helps to identify which air-quality sensor data attributes are related to the IoT data quality ones, as well as which techniques are commonly used to improve those attributes. For example, the accuracy can be improved by sensor calibration, data interpolation, outlier detection, and data fusion techniques. As evidenced, not all dimensions have a mapping on the indicators side, and not all dimensions can be improved by the DQ enhancing techniques explored in the literature. A description of how the enhancement techniques are related to the DQ dimensions is explained as follows.

Data calibration involves correcting the taken measurements to improve the accuracy of the variable. Thus, the accuracy dimension is directly affected as the calibration process seeks to improve this dimen-

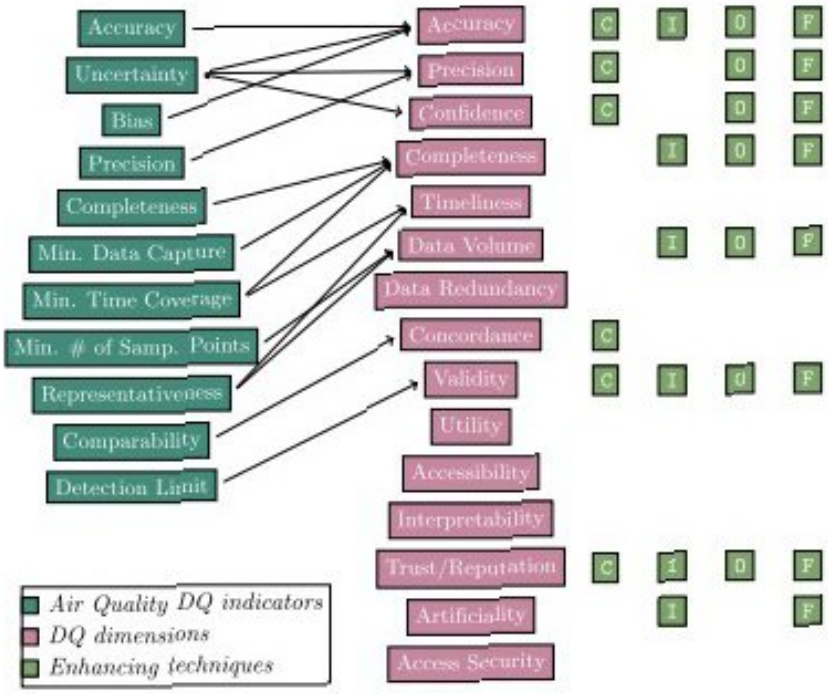


Fig. 1 Mapping air quality DQ indicators, DQ dimensions, and DQ enhancing techniques, where C is the calibration, I is the interpolation, O is the outlier detection, and F is the data fusion/aggregation

sion. When an air quality monitoring system implements a calibration process, it enhances the confidence of the system, thus affecting the confidence dimension. If the calibration process takes into account the variability of the measurement, the precision dimension is also improved. On the other hand, if the calibration process involves an additional reference measurement (e.g., robust calibrated sensors, particle generators), the concordance dimension is affected since the sensor, which is being calibrated, is compared to a reference for applying a correction mechanism.

Data interpolation creates new data points in order to fill spatial or temporal gaps, by improving the completeness of the original samples as well as the volume of captured data. Normally, interpolated data is created using mathematical or machine learning models, thus increasing the artificiality. Interpolated data is usually compared with at least one reference in order to estimate the error, and hence the accuracy dimension is altered. Precision dimension is affected if a computation of variability or standard deviation of the interpolated data is evaluated. Data interpolation is also related to the confidence dimension if confidence intervals are calculated and the interpolated data is within them. When a computation of correlation between interpolated data and near (spatial or temporal) real data is performed, the concordance dimension is considered.

Data aggregation/fusion techniques are related to several data quality dimensions. Accuracy of measurements is improved when data with poor quality are fused or aggregated with good quality data. The fused or aggregated data can have a different variability from the sources depending on the used technique, thus affecting the precision and increasing data artificiality. Completeness dimension is altered if new data is included in the fused or aggregated process. Also, new data contributes to data volume dimension, where incomplete datasets can be merged in order to obtain a more complete fused dataset. Moreover, the redundancy dimension is changed if the aggregation technique uses redundant data and the confidence dimension is affected if an error is estimated with a specific confidence interval. Additionally, concordance dimension is modified if the techniques include the correlation of multiple measurements. Finally, the validity dimension is altered if fused or aggregated data is contrasted with ground true data.

As outlier detection techniques aim to identify data that is not consistent with other observations, it can reduce the error and variability of data, by improving its accuracy, precision, and confidence, while increasing its reliability. Furthermore, if outlier detection involves removal of anomaly data, it will impact directly on the completeness of the dataset and will also reduce its volume. Detecting whether anomalous data are related to errors or important events can also be achieved by using concordance metrics; hence, this dimension is related to the technique. Having these DQ concepts in mind, we present below the design and the results of the systematic mapping proposed in this work.

5 Systematic Mapping Method

A systematic mapping study is a well organized, and a frequently used methodology to synthesize the state of the art around a particular research area. This type of studies looks for the "big picture" of some particular research topic, showing the branches and challenges associated with it James et al. (2016). This approach has been mainly used in software engineering; however, its application in the IoT field has been modest.

In this document, a systematic mapping study is developed based on the guidelines proposed by Petersen et al. (2008). Some steps were established to identify and analyze the studies about Data Quality on IoT-based air quality monitoring systems. We define the following steps for developing the systematic mapping study:

1. **Research questions:** In this step, the research questions are defined. These questions are expected to be solved when the systematic mapping process is completed.
2. **Search strategy:** This step defines the methodology of the research, starting by defining the "search chain" which will be applied to relevant academic databases.
3. **Selection criteria:** Inclusion and exclusion criteria are defined in this step. These criteria are used to filter the studies found in previous step.

4. **Data extraction:** Once the Search Strategy and Selection Criteria are applied, relevant information about the Research Questions is extracted from the selected articles.
5. **Analysis:** In this step, we analyze the results obtained for drawing conclusion about the mapping study.

5.1 Research Questions

We develop this study to identify the state-of-the-art on how data quality is applied in IoT-based air quality monitoring systems. Hence, we define five research questions (RQs) which help us to guide the review of the literature in this field.

- **RQ#1:** Which are the most relevant DQ dimensions related to IoT-based air quality monitoring systems?
- **RQ#2:** What are the most used strategies to mitigate data quality problems in IoT-based air quality monitoring systems?
- **RQ#3**: What are the system's features that threaten data quality in IoT-based air quality monitoring systems?
- **RQ#4**: How is data quality estimated for IoT-based air quality monitoring systems?
- **RQ#5**: How is degradation of data quality identified in IoT-based air quality monitoring systems?

5.2 Search Strategy

The research questions are used to identify the four main keywords in our search: "Air Quality," "Monitoring," "Data Quality," and "Internet of Things." Then, we assemble the search query including new terms from variations of these keywords. Table 7 presents the search query, for each main keyword we define a corresponding query that contains all the variants. We used the AND logical operator to connect the resulting keyword groups.

Table 8 Included and excluded publications

Item	Number of publications
Initial search	162
Snowballing	+40
Total papers found	202
Applying selection criteria	-131
Total papers selected	71

According to the analysis developed in Chen et al. (2010), we select five of the most relevant academic databases: IEEE, Web of science, Scopus, ACM, and Science Direct. We performed the search in March 2022 using the query described in Table 7 in the title, abstract, and keywords of the published works. We found a total of 162 publications, after removing duplicates.

We also developed a snowballing process from the review articles found in the initial search. The idea is to look for potential papers to include in our study by reviewing the references of these review articles. We identified 40 papers in this snowballing process.

5.3 Selection Criteria

For this study, we define one inclusion criterion and four exclusion criteria. The inclusion criterion defined for this study is "the study includes publications that propose, compare or implement methods to measure or analyze the quality of data gathered by IoT systems in the context of air pollution."

The exclusion criteria for this mapping study are the following: (1) The study excludes papers that do not propose, compare, or implement methods to measure or analyze the quality of data gathered by IoT systems in the context of air pollution. (2) The study excludes papers that are not written in proper English language. (3) The study excludes papers that are duplicated or are a previous version of a more

Table 7 Search query used in the mapping study

Main keyword	Query
"Air quality"	"air quality" OR "air pollut*" OR "atmospheric pollution"
"Monitoring"	monitoring OR detecti* OR sensing
"Data quality"	"data quality" OR "anomaly detection" OR "data anomaly"
"Internet of Things"	"internet of things" OR "IoT" OR "sensor networks"

complete study about the same research. (4) The study excludes papers such as systematic reviews, mapping studies, editorials, prefaces, article summaries, interviews, news, correspondence, discussions, comments, readers letters, tutorial summaries, panel discussions, opinion articles, poster sessions, classes, abstracts, and presentations.

We apply the inclusion/exclusion criteria to the papers retrieved in the previous step, by ensuring that each paper is analyzed by all members of the team. We develop meetings to resolve the conflicts arisen from the application of these criteria. The Rayyan web application is used for managing this process (Ouzzani et al., 2016). As a result of this step, 71 papers were selected (see Table 8).

5.4 Data Extraction

In this step, we deeply review the selected papers with the aim to extract relevant information for answering the research questions. As in Petersen et al. (2015), we divide the selected papers into five sets of 14–15 papers. Each team member extracts information from the papers in her/his set and then reviews the extraction of another team member. Following this process, we ensure that each paper is review by two team members. Then, a weekly meeting is carried out to resolve any conflict and reach a common agreement.

6 Results

This section presents the results of the mapping study developed to answer the research questions stated above. Before discussing the main results, we present a general overview of the papers under scope. The analysis of the main topics are presented around three aspects such as DQ dimensions and enhancement techniques, endangering factors, and DQ estimation and degradation.

One of the first highlighting points is that the analysis of data quality in the context of IoT-based air quality monitoring systems is a topic with rising interest in the research community, especially in the last 7 years, with an average number of near 9 papers per year, as shown in Fig. 2. Even though, there are some early approaches, such as by Harkat et al. (2006), the interest in DQ can be linked to the development and deployment of low-cost monitoring systems.

Figure 3 illustrates the venues in which the analyzed works were published. Most of the papers (57.7%) were published in high quality journals (Q1 or Q2 according to the Scimago ranking). Almost a third of the papers analyzed in this study (31%) were published in conferences. Figure 4 shows the deployment location of the AQ monitoring systems for which DQ is analyzed. These systems have been deployed in 16 different countries, being USA (with 10 AQ systems), China and Taiwan (with 7 systems each one), and Switzerland (with 4 systems) the countries with more number of deployments reported.

Figure 5 presents some details regarding the IoT-based AQ systems. Most of the systems are specifically created for outdoor monitoring (53 out 71), while 5 works are created for indoor scenarios, and 7 for both indoor and outdoor. We also analyzed the portability of these systems finding 44 implementations in fixed locations, 12 mobile system, and 4 works that can be used in both fixed and mobile.

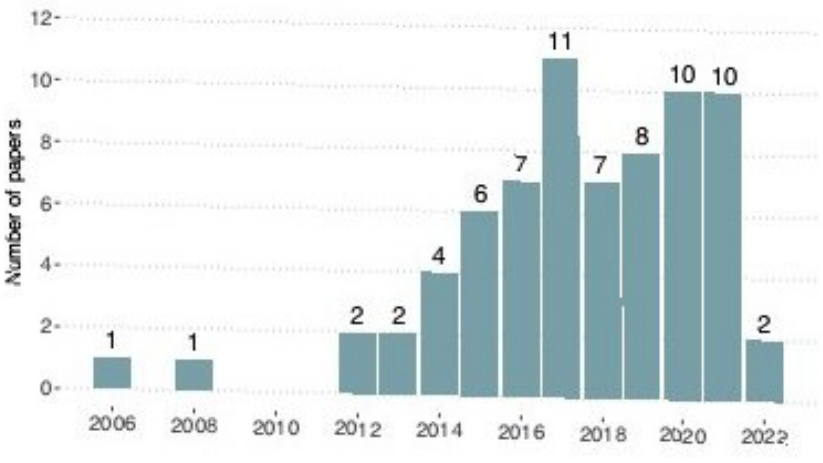
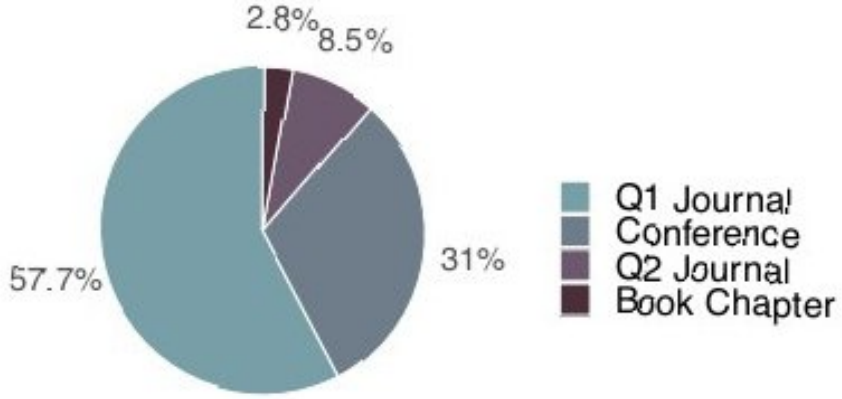


Fig. 2 Histogram of paper publications in the context of data quality in IoT-based air quality monitoring systems per year

Fig. 3 Venue of the publication



Regarding the variables of interest in the AQ monitoring system, Fig. 6 presents a histogram of the environmental variables identified in our study. The PM2.5 variable is the most frequently analyzed followed by the ozone (O3) and the nitrogen oxides (NOx). This result is in agreement with the expectations, since low-cost PM and gas sensors are more prone to low-quality measurements as mentioned before.

6.1 DQ Dimensions and Enhancement Techniques

This section aims at providing answer to research questions RQ#1 and RQ#2. Regarding RQ#1, "Which are the most relevant DQ dimensions related to IoT-based air quality monitoring systems?," most of the works analyzed usually do not refer directly to DQ dimensions, as defined in Section 2 We consider this lack of use of technical DQ concepts in IoT systems is caused by the disconnection between the IoT field and the Data Quality theory.

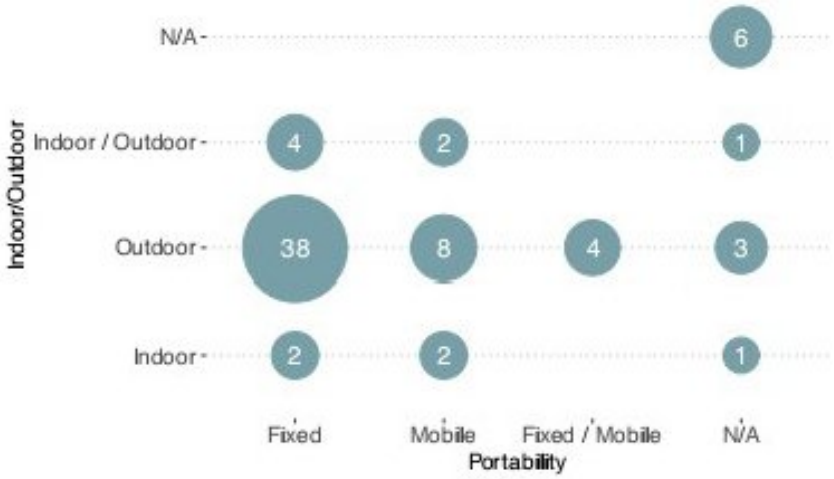
We identify the DQ dimension used in IoT systems by looking for the DQ enhancement techniques implemented in those systems, thus answering the RQ#2. Figure 7 presents the DQ enhancement techniques implemented in the analyzed works. Calibration (C) is the most used technique being implemented in about 50% of the works. Most of them implement calibration techniques on-site and at run-time as depicted in Fig. 8. Data interpolation (I) and outlier detection (O) are also frequently used in air-quality monitoring systems, being implemented in 18 works each. Finally, data aggregation and fusion are less frequently implemented, found in only four works.

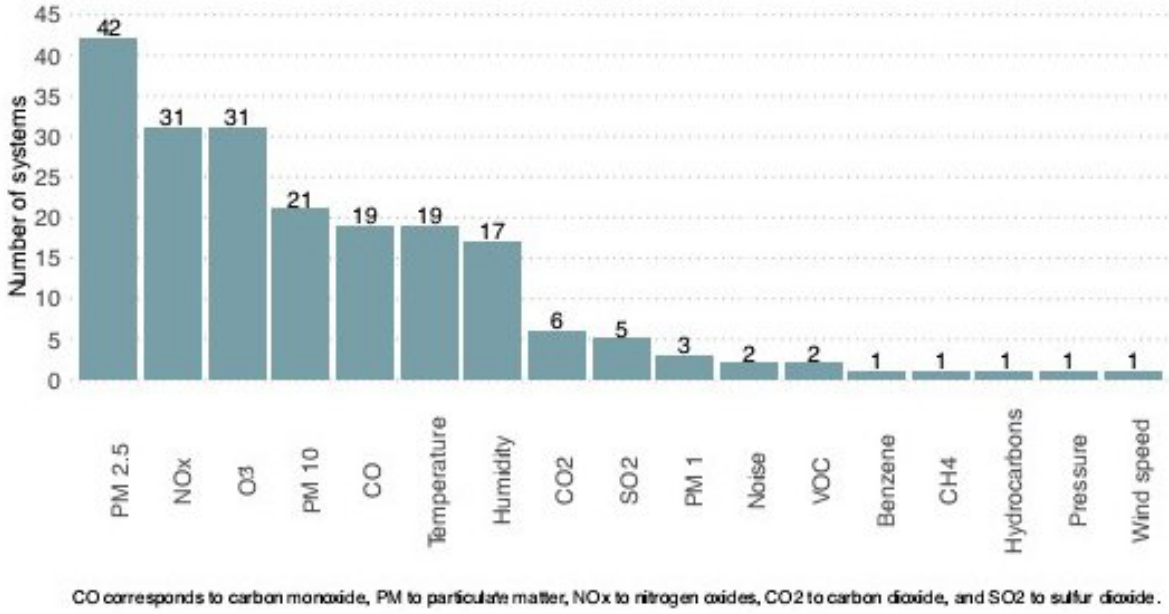
According to the discussion we develop in Section 4.5, the calibration technique is directly related to the accuracy, confidence, precision, and concordance dimensions. Furthermore, the data interpolation is related to the completeness, artificiality, accuracy, precision, confidence, and concordance dimensions. Outlier detection is associated to the following dimensions, accuracy, precision, confidence, completeness, and concordance. Finally, the data aggregation and fusion techniques are linked to the accuracy, precision. artificiality, completeness, data volume, data redundancy, confidence, concordance, and validity dimensions.

Figure 9 presents the percentage of relative importance of DQ dimensions in IoT-based air quality

Fig. 4 Location of the deployment

Fig. 5 System type

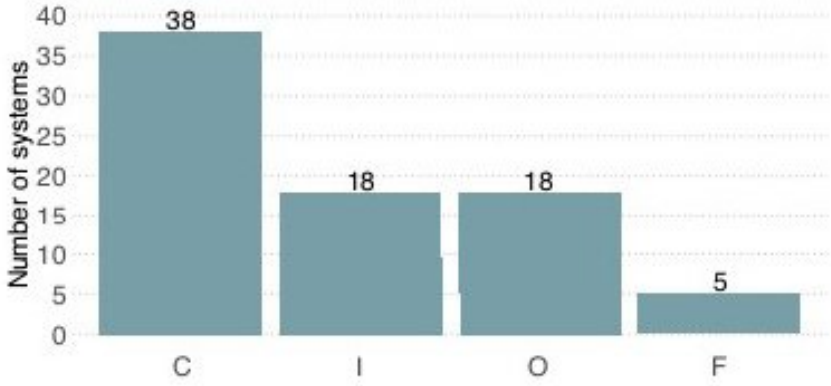




CO corresponds to carbon monoxide, PM to particulate matter, NOx to nitrogen oxides, CO2 to carbon dioxide, and SO2 to sulfur dioxide.

Fig. 6 Measured air-quality variables

Fig. 7 DQ enhancement technique used in AQ monitoring systems (C, calibration; I, data interpolation; O, outlier detection; F, data aggregation/fusion)



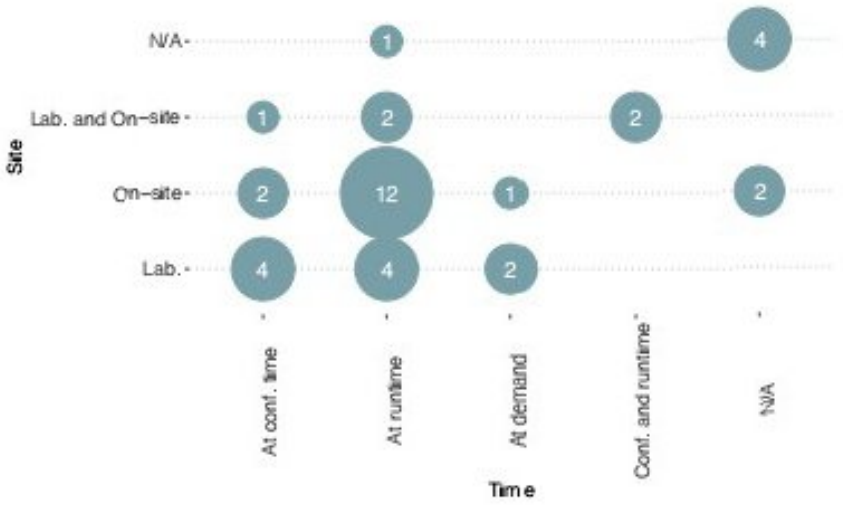


Fig. 8 Calibration time and site

monitoring systems. We define a score for representing the relative importance of a dimension, which varies from 0 to 100, where 0 means the dimension is not important, and 100 means the dimension is very important. This score is computed as the percentage of appearances of each dimension with respect to the total number of times an enhancement technique is implemented. According to this score, the Precision, Confidence, Concordance, and Accuracy dimensions, with a score of 100, are been considered the most important DQ dimensions for the IoT-based air quality monitoring systems. Then, Completeness and Artificiality dimensions have a lower importance, obtaining scores of 52 and 29. Finally, the least important dimensions are Validity, Data Volume, and Data Redundancy, with a score of 6.3 each one.

6.2 Endangering Factors

IoT-based air quality monitoring systems have been gaining popularity and are being included in a lot of new applications. Features like portability, small size, lightweight, low-cost, and first-hand data generation have motivated the creation of enthusiastic projects related to this topic. For this reasons, the trend shows that this approach will continue growing in the next decade. Figure 10 shows that most of the reviewed works (85.9%) are using low-cost sensors.

New technological approaches around air quality measurement have brought new challenges related to the degree of trust of these systems. IoT-based air quality monitoring is somehow contrary to classic, expensive, robust and certified air quality monitor-

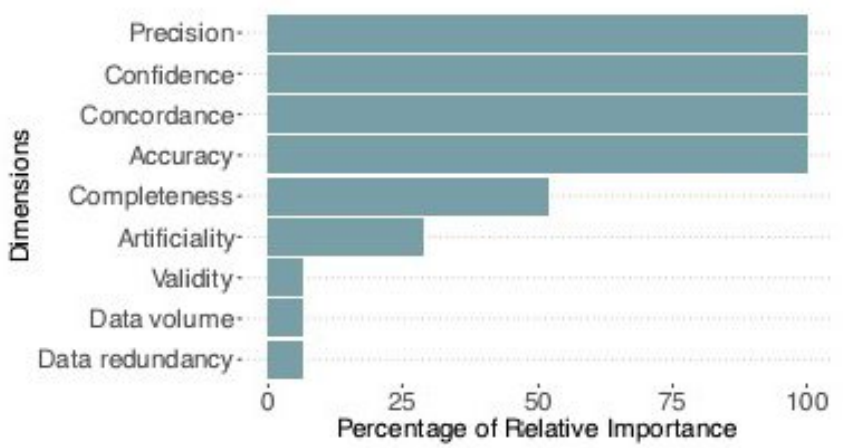


Fig. 9 Relative importance of DQ dimension in IoT-based air quality monitoring systems

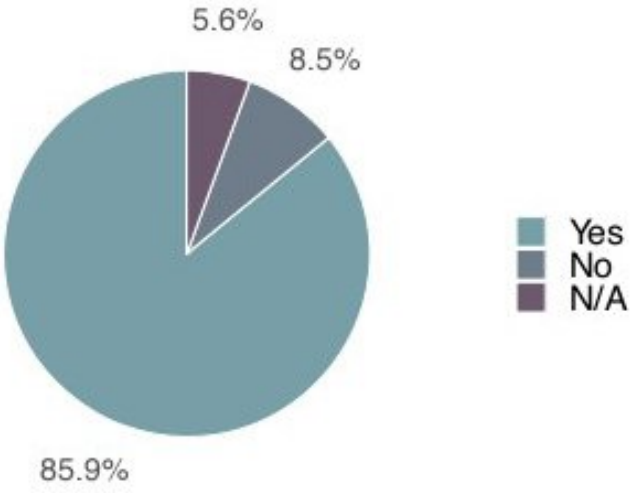


Fig. 10 Low-cost sensor usage

ing stations, which have been used as normative to determine the risks associated with air quality in overpopulated places around the world.

Moreover, some weaknesses are related to low-cost stations and their application to large-scale air quality monitoring. In particular, DQ can be seriously compromised in low-cost approaches, due to some degree of data degradation mentioned in the RQ#3. Among the weaknesses of low-cost sensors, we could identify: method of measurement, sensor aging, lack of redundancy, limited lifetime, and data error in storing/communication. The details of each identified weakness are given below.

- **Method of measurement.** Most of low-cost portable AQ sensors are related to PM2.5/PM10 and gas concentration (Budde & Riedel, 2018; Lin et al., 2018) measurements. In less proportion, other kind of gas sensors are used in IoT air-quality monitoring applications (Fig. 10). These sensors implement a widely used technique called "laser detection." In this technique, a flow of air is pumped by a fan into a chamber. The chamber has a laser light which generates shadows on a light detector when a particle is present. Constant air flow is essential to have an accurate reading. An embedded computer attached to the sensor estimates the PM value based on the light detector. Problems associated with this method include as follows: miss-computations in the detector, low or high air flow speed in the chamber, high or low environmental temperatures, and high humidity in the air around the sensor (Liu et al., 2017; Penza, 2020).
- **Sensor aging.** A high concentration of dirt and dust degrades the sensor response, generating data that can be far from the reality (Liu et al., 2017; Manikonda et al., 2016). Periodical maintenance has to be applied in order to avoid this issue. Regarding outdoor sensors, they suffer from case degradation. A lot of malfunctions in such sensors are related to electronic damage due to water leaking inside the sensor hardware. The sensor's precision decreases along time, environmental factors as humidity and temperature can seriously degrade the data. A periodical calibration strategy must be applied to improve this issue.
- **Lack of redundancy.** Redundancy is an easy way to determine when a particular sensor is showing malfunctions. In IoT AQ monitoring systems, sensor redundancy comprises including two or more sensors for comparing the deviation of their readings. Moreover, some applications can show serious problems applying redundancy of nodes due to their size and battery restrictions. Another kind of redundancy can be achieved by analyzing spatio-temporal data from near sensors (Feinberg et al., 2019; Li et al., 2018; Lin et al., 2020a). This technique is a computational demanding task that usually has to be applied in a gateway node and, again, can be hard to accomplish in some scenarios.
- **Limited lifetime.** Battery powered systems are widely used in low-cost electronic solutions, and AQ measurement system is not an exception. The typical AQ device includes a PM sensor, an on board computer, and communication and storage interfaces. With all these components demanding power from the battery. In a poorly planed AQ IoT solution, the battery lifetime can be very short due to system inefficiency (Penza, 2020). Power management design techniques as low power communications, extensive use of low power modes in the processors, time/event driven software applications development, among others, have to be applied in order to extend the lifetime of the batteries (Kendrick et al., 2019).
- **Data error in storing/communication.** It is common in electronic systems to have errors in communication and storage processes. Those errors can be caused by the electromagnetic noise in

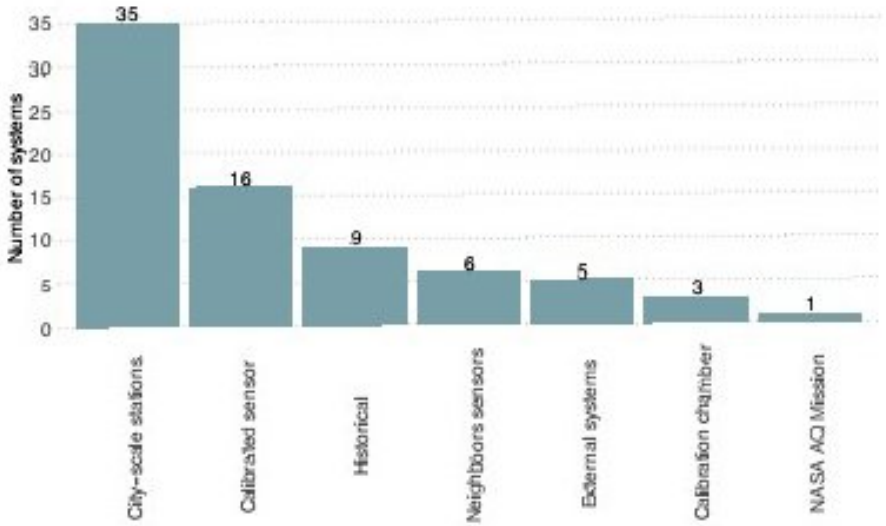


Fig. 11 Distribution of articles that compare low-cost sensors measurements to a reference

7 Discussion

An increasing interest in analyzing the DQ topic is depicted in Fig. 2, which can be interpreted as the result of the number of low-cost IoT systems deployed for AQ monitoring (see Fig. 10). However, it was found in the reviewed papers that not many authors make an explicit mention of the DQ dimensions addressed in their work. What they do is to mention terms derived from "Data Quality," where DQ information is diffuse. We believe that the main reason for this phenomenon is the language of data quality, which has not been used in a proper formal way inside the IoT and air-quality monitoring applications yet. This is considered a serious issue to confront AQ measurement under the DQ definitions here presented.

Accuracy was the most DQ indicator mentioned by different authors to measure "quality" inside AQ systems. Nevertheless, the introduction of other indicators will provide a more reliable and realistic approach inside the IoT AQ measurement. The minimum DQ dimensions and indicators that should be provided by a low-cost AQ system is a challenge that has to be established by different actors such as environmental agencies, enthusiastic developers, and technological industries around the world. We consider that environmental agencies have shown resistance in the implementation of portable and low-cost AQ supervision systems due to factors such as method of measurement, sensor aging, lack of redundancy, limited lifetime, and data error in storing/communication. Although these problems are serious and unresolved, low-cost AQ supervision will not be taken into account as a real alternative to determine the AQ in large-scale applications. On the other hand, low-cost sensors in the context of AQ applications have being growing as an alternative to empower citizens around the world. This tendency offers a lot of challenges and opportunities, which remarks the importance of an adequate DQ definitions in those applications.

Using DQI or DQ dimensions as a way to evaluate the status of an air-quality monitoring system can be a proper approach, since it will consider the attributes that are really important for the users within a context. This approach can provide a complete view of the system's DQ status, and also allow to check on specific degraded features that can be improved by using DQ enhancing techniques (see Fig. 1). Also, by identifying the endangering factors, it can be targeted improvements on the system's infrastructure to mitigate their impact on the overall DQ of the system.

Therefore, this work found that dimensions or indicators are not mentioned explicitly by the authors due to the lack of proper usage of DQ dimensions and indicators definitions, as well as the fact that most of the authors do not stick to guidelines, which standardize topics like the air-quality monitoring. In order to mitigate this issue, as a future work, we propose the development of a tool that can be used to identify and sort the dimensions and indicators for IoT-based AQ monitoring systems.

8 Conclusion and Future Work

In this paper, we studied the data quality analysis on IoT-based air quality monitoring systems. First, we identified a general overview of data-quality dimensions within an IoT context. Then, data quality indicators and objectives in air-quality monitoring systems are reported, according to the guidelines by regulatory entities. Also, we propose a mapping from indicators to dimensions to determine the relation between these concepts. In order to establish the state of data quality in IoT-based AQ systems, we developed a systematic mapping study about this field. The results showed an increasing number of studies that take into account terms related to DQ within IoT-based air quality monitoring systems in the last few years; however, there is a lack of DQ terminology adoption and a rigorous application of DQ metrics. For instance, we had to identify the most relevant DQ dimensions related to IoT-based air quality system indirectly by analyzing the used enhancement techniques. To this end, we created a mapping between the enhancement techniques and the DQ dimensions.

In general, we found authors do not use the terminology of the DQ field. We suppose this is due to two different factors. First, there is an absence of regulations that take into account indicative measurements (like low-cost sensor measurements) in the evaluation of air quality. Second, authors ignore the existing guidelines because they are not required to follow them. The primary objectives of their research are to evaluate technological alternatives or data processing techniques.

It is understandable why low-cost sensor measurements are not fully considered by agencies in charge of environmental monitoring, because of their data is prone to have more errors than a robust station. However, to avoid such distrust on low-cost sensors, an air quality monitoring system can be implemented to be DQ-aware and also to include techniques to improve the quality of its data. In addition, many low-cost sensors can complement few robust stations to improve the resolution of the system, using the robust stations directly as references or sources of data to build reference models that help to improve DQ in low-cost air pollution sensors, for example, to be used in calibration processes.

Funding Open Access funding provided by Colombia Consortium. This research was supported by the project 1135 funded by the Science and Technology Vice-Rector's office at the Universidad de Medellín in Medellín, Colombia.

Data Availability The dataset generated and analyzed during the current study are available in the figshare repository, https://figshare.com/s/d72577f85291cb52356a.

Compliance with Ethical Standards

Conflict of Interest The authors declare no competing interests.

References

Alavi-Shoshtari, M., Williams, D., Salmond, J., & et al. (2013). Detection of malfunctions in sensor networks. *Environmetrics*, *24*(4), 227–236. https://doi.org/10.1002/env.2206.

Alavi-Shoshtari, M., Salmond, J., Giurcăneanu, C., & et al. (2018). Automated data scanning for dense networks of low-cost air quality instruments. Detection and differentiation of instrumental error and local to regional scale environmental abnormalities. *Environmental Modelling and Software*, *101*, 34–50. https://doi.org/10.1016/j.envsoft.2017.12.002.

Alvarado, M., Gonzalez, F., Fletcher, A., & et al. (2015). Towards the development of a low cost airborne sensing system to monitor dust particles after blasting at open-pit mine sites. *Sensors (Switzerland)*, *15*(8), 19,667–19,687. https://doi.org/10.3390/s150819667.

Atzori, L., Iera, A., & Morabito, G. (2017). Understanding the Internet of Things: Definition, potentials, and societal role of a fast evolving paradigm. *Ad Hoc Networks*, *56*, 122–140.

Barcelo-Ordinas, J., Garcia-Vidal, J., Doudou, M., & et al. (2018). Calibrating low-cost air quality sensors using multiple arrays of sensors. pp. 1–6, https://doi.org/10.1109/WCNC.2018.8377051.

Step 1: Import all the modules required

Python3

```
# import module
import requests
from bs4 import BeautifulSoup
```

Step 2: Create a URL get function

Python3

```
# link to extract html data

def getdata(url):
    r=requests.get(url)
    return r.text
```

Step 3: Now pass the URL into the getdata function and convert that data into HTML code. The URL used here is **“https://weather.com/en-IN/forecast/air-quality/l/3dbed5c769584b3604a70d40a**

Python3

```
htmldata = getdata(# write the U
soup = BeautifulSoup(htmldata, '
result =  soup.find_all(class_="
result
```

Output:

[<div class=”styles__primaryPollutantGraphNumber__2WgP9" classname=”styles__primaryPollutantGraphNumber__2WgP9" >67 ,

class=”styles__primaryPollutantGraphNumber__2WgP9" classname=”styles__primaryPollutantGraphNumber__2WgP9" >22 ,
<div class=”styles__primaryPollutantGraphNumber__2WgP9" classname=”styles__primaryPollutantGraphNumber__2WgP9" >13 ,
<div class=”styles_N_primaryPolluta ntGraphNumber__2WgP9" classname=”styles__primaryPollutantGraphNumber__2WgP9" >30 ,
<div class=”styles__primaryPollutan

tGraphNumber__2WgP9" classname="styles__primaryPollutantGraphNumber__2WgP9" >479]

Step 4: Filter your data and Check your Air Quality according to the given data :

Python3

```
# Traverse the air quality
res_quality = soup.find(class_="

# traverse the content
air_data = soup.find_all(class_=
air_data=[data.text for data in
print("Air Quality :", res_data)
print("O3 level :", air_data[0])
print("NO2 level :", air_data[1]
print("SO2 level :", air_data[2]
print("PM2.5 level :", air_data[
print("PM10 level :", air_data[4
print("co level :", air_data[5])
```

Output:

```
Air Quality : 85
O3 level : 67
NO2 level : 22
SO2 level : 13
PM2.5 level : 30
PM10 level : 45
co level : 479
```

Step 5: Now Analyze the Air Quality with the given data:

AQI	Remark	Possible Health Impacts
0-50	Good	Minimal impact
51-100	Satisfactory	Minor breathing discomfort to sensitive people
101-200	Moderate	Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor	Breathing discomfort to most people on prolonged exposure
301-400	Very Poor	Respiratory illness on prolonged exposure
401-500	Severe	Affects healthy people and seriously impacts those with existing diseases

THANK YOU