Chapter

Quality of Information within Internet of Things Data

Tomás Alcañiz, Aurora González-Vidal,
Alfonso P. Ramallo and Antonio F. Skarmeta

Abstract

Due to the increasing number of IoT devices, the amount of data gathered nowadays is rather large and continuously growing. The availability of new sensors presented in IoT devices and open data platforms provides new possibilities for innovative applications and use-cases. However, the dependence on data for the provision of services creates the necessity of assuring the quality of data to ensure the viability of the services. In order to support the evaluation of the valuable information, this chapter shows the development of a series of metrics that have been defined as indicators of the quality of data in a quantifiable, fast, reliable, and human-understandable way. The metrics are based on sound statistical indicators. Statistical analysis, machine learning algorithms, and contextual information are some of the methods to create quality indicators. The developed framework is also suitable for deciding between different datasets that hold similar information, since until now with no way of rapidly discovering which one is best in terms of quality had been developed. These metrics have been applied to real scenarios which have been smart parking and environmental sensing for smart buildings, and in both cases, the methods have been representative for the quality of the data.

Keywords: IoT, QoI, outliers, interpolation, data quality, data integrity

1. Introduction

The emergence of Internet of Things (IoT) deployments has allowed millions of connected, communicating, and exchanging objects to be embedded seamlessly around the world, generating large amounts of data through sensor monitoring on a timely basis.

The data flow between the physical and the digital world through artificial intelligence can expand the computer's awareness of the surrounding environment, thereby obtaining the ability to act on behalf of humans through ubiquitous services.

In this IoT-based environment, the basis for making wise decisions and providing services is the data collected by sensors and actuators. If the data quality is poor, these automated decisions may be incorrect, ranging from sensor failure to deliberately providing false information with malicious intent. Data quality (DQ) is therefore needed to attract users to participate and accept IoT paradigms and services.

Data Quality refers to how well data meet the requirements of data consumers [1]. In a similar manner, Quality of Information (QoI) relates to the ability to judge whether information is adequate for a particular purpose [2, 3].

From such a well-known and accepted definition, we understand that it refers to a perception or an evaluation of the suitability of the data to fulfill its purpose in a given context, subject to the requirements of the consumer. On the literature, the quality of the data is determined by factors such as availability, usability, reliability accuracy, completeness, relevance, and novelty [4].

According to [5], ensuring data quality is crucial when deploying and leveraging devices, given that:

- Decision-making is only possible if the data available are correct and appropriate.

- Serious problems are practically unapproachable without an adequate data source.

The way to tackle this problem is through the use of so-called data quality metrics which are calculated in order to validate the Quality of the Information (QoI).

The aim of this chapter is to define some metrics for DQ and calculate them in IoT scenarios in order to test their viability.

2. Data integrity

Data integrity refers to the accuracy and reliability of data. The data must be complete, without variations or compromises from the original, which is considered reliable and accurate. Therefore, this term is closely related to the quality of the data and this in turn to the quality metrics [6].

There are several types of data integrity [7]:

- Physical integrity is the protection of the integrity and accuracy of data as they are stored and extracted. That is, it is related to the physical layer of the systems. In the context of IoT, a physical integrity problem comes from the physical degradation of the sensors, whether due to a breakdown or sabotage.

- Logical integrity preserves the data without any change, since it is used differently in a relational database. Logical integrity protects data from human errors and also from hackers, but in a very different way than physical integrity.

- The integrity of the entity is based on the creation of primary keys, or unique values, that identify data to ensure that it is not listed more than once and that there is no field in a table considered null. It is a feature of relational systems that store data in tables that can be linked and used in very different ways. In an IoT scenario, an entity integrity problem can arise in case of a sensor failure which produces redundant measurements or by a human failure in which two different sensors are assigned the same identifier, which produces redundancies in databases.

- Referential integrity is a series of processes that ensure that data is stored and used consistently. The rules built into the database structure about how foreign keys are used to ensure that only appropriate data changes, additions, or deletions occur.

- Domain integrity is the set of processes that guarantee the veracity of each data in a domain. In this context, a domain is a set of acceptable values that a column can contain. You can incorporate restrictions and other measures that limit the format, type, and amount of data entered. Due to an error in the IoT devices, one of them could be entering data that does not correspond to the correct type in a column of a database, such as saving a number when a date should be saving or a date in a format that is not adequate.

- User-defined integrity comprises the rules and constraints created by the user to suit their particular needs. Sometimes entity, referential, and domain integrity are not enough to safeguard data. Often times, specific corporate rules need to be considered and incorporated into measures regarding data integrity. In an IoT scenario, a sensor may be giving acceptable values, that is, that they respect the rest of the integrity criteria, however, it may not be meeting a necessary criterion for the correct functioning of the system, such as a sensor that collects percentage values and that you are receiving a value greater than 100.

In this section, Data Integrity has been defined, however, it is necessary to note what is the difference between this term and the term Data Quality. Data quality is related to the reliability of the information, which is necessary for planning and decision making for a specific operation. Whereas, the integrity of the data guarantees the reliability of the data in physical and logical terms.

3. Data quality metrics

In this section, we describe the metrics that have been defined to calculate and annotate the QoI for IoT data. Those were previously described on [8].

3.1 QoI basic metrics

The first set of metrics is based on a descriptive analysis. This approach was also used on the IoTCrawler framework [9]. It proposes to integrate quality measures and analysis modules to rate data sources to identify the best fitting data sources to get the needed information. The first step before implementing some quality analysis modules is to identify quality measures, which can be used to rate data sources and the delivered/produced data for their Quality of Information. To measure the QoI, we propose to use the so-called QoI Vector, which is defined in Eq. (1) and gathers the information belonging to all the metrics proposed in this framework

$$\vec{Q} = \left\langle q_{cmp}, q_{tim}, q_{pla}, q_{art}, q_{con} \right\rangle \tag{1}$$

The elements of the vector are defined as follows:

- Completeness (q_{cmp}): it represents the percentage of missing or the unusable data.

$$q_{cmp} = 1 - \frac{M_{miss}}{M_{exp}} \tag{2}$$

where M_{miss} is the sum of missing values and M_{exp} is the sum of expected values of an incoming dataset.

- Timeliness (q_{tim}): refers to the expected time of accessibility and availability of information. In other words, it represents how long is the time difference between the data capture and the reality event happening. It is crucial in critical IoT applications such as traffic safety. Its definition is:

$$q_{tim} = 1 - \frac{T_{age}}{W} \tag{3}$$

where T_{age} is the difference between the expected time and the time taken by the sensor ($T_{real} - T_{exp}$), and W is the proper time of the system, which is chosen arbitrarily.

- Plausibility (q_{pla}): shows if received data is coherent according to the probabilistic knowledge of the variables that are being measured. Sensor annotations or meta-data are used to determine an expected value range of an incoming measurement.

$$q_{pla} = \prod P_{Annotations}(\nu) \tag{4}$$

The range of Plausibility value is defined between 0 and 1.

- Artificiality (q_{art}): this metric determines the inverse degree of the used sensor fusion techniques and defines if this is a direct measurement of a singular sensor, an aggregated sensor value of multiple sources or an artificially interpolated value.

- Concordance (q_{conc}): describes the agreement between information of the data source and the information of other independent data sources, which report correlating effects. The Concordance analysis takes any given sensor x_0 and computes the individual concordances, $C(x_0, x_i)$, with a finite set of n sensors ($i = 1, \dots, n$).

$$q_{con}(x_0) = \sum_{i=1}^{n} \lambda_i(x_0) c(x_0, x_i) \tag{5}$$

with λ as a weight-function

$$\lambda_i(x_0) = \frac{1}{d(x_0, x_i)} \tag{6}$$

And $d(x_a, x_b)$ propagation and infrastructure-based distance function between sensor location x_a and x_b or sensors a and b.

All the metrics exposed in this section take values between 0 and 1, with the value 1 being the ideal case in which the quality of the data is maximum and 0 the opposite case.

These metrics represent the simplest ones that can be calculated in this kind of IoT scenarios. However, it is possible to go further and compute some metrics that give us a deeper knowledge of the IoT system.

3.2 Oultier-based metrics using heuristics

Since these metrics provide us basic information, it is possible to go further and obtain a series of metrics that can be useful. These new metrics come from the hand of Machine Learning (ML), in this case the search for outliers.

In machine learning, an outlier is an observation that diverges from an overall pattern. The number of outliers in an indicator of data quality.

In the literature, there are usually considered 4 types of basic outliers for time series: additive outliers, level shifts, temporary changes and innovational outliers, see [10, 11] for a complete description.

A metric similar to the case of q_{cmp} can be defined, only taking into account the values that are considered outliers instead of the missing ones. The percentage of outliers in the studied sensor is named q_{out} (see Eq. (7)). In order to obtain which of these values are considered outliers it is useful an Autoregressive Integrated Moving Average (ARIMA) based framework [12]. It can also determine if the oulier is innovational, additive, level shift, temporary changes or seasonal level shifts.

$$q_{out} = 1 - \frac{M_{out}}{M_{total}} \tag{7}$$

where M_{out} is the sum of outlier values on the features of the sensor and M_{total} is the sum of total features.

As important as determining whether an instance is an outlier or not is knowing how much it deviates from what would be the expected value corresponding to the normal behavior of the time series. For that purpose it is necessary to impute the data of the time series that are considered outliers as if they were missing values, in order to know what this expected value would be. Then the difference between the value and the imputation is another metric that has been computed by dividing the difference of each sensors value by the mean, median or mode of the values and then calculate their mean, median or mode ($q_{mean}, q_{median}, q_{mode}$).

$$q_{mean} = mean(|\hat{x}_i - x_i|) \tag{8}$$

$$q_{median} = median(|\hat{x}_i - x_i|) \tag{9}$$

$$q_{mode} = mode(|\hat{x}_i - x_i|) \tag{10}$$

where i corresponds to those indices of the features that present an anomalous behavior, while $\hat{x}_i$ and x_i represent the imputed value that follows the expected behavior and the value of the outlier respectively. This metric takes values between 0 and 1, with 1 being the ideal case.

Unsupervised methods are also adequate for oultier detection, so we propose q_{prob}. This metric corresponds to the probability of belonging to a certain cluster that has been computed using Gaussian Mixture Models (GMM), which consists of representing in the most faithful way possible the data points by adding some Gaussian distributions. It informs quantitatively of the anomalous values. The number of clusters or Gaussians distributions is an hyperparameter and it could be chosen in different ways. In the experiments we used silhouette coefficient.

$$q_{prob} = \sum_{i=1}^{k} G_i(x_j) \quad (11)$$

where k is the number of clusters or distributions used, G_i corresponds to distribution i and x_j is the vector taken by the sensor. Because this metric is probabilistic, it takes values between 0 and 1, in such a way that the closer the value is to 1, the more quality the instance has.

Another way to determine if the data series exhibits anomalous behavior is by using so-called AutoEncoders. An AutoEncoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The objective of these autoencoders is to learn a representation of the data to be studied, with the aim of eliminating noise, however it is possible to use this tool to detect anomalous values. AE are a specific type of feedforward neural networks where the input is the same as the output. They compress the input into a lower-dimensional code and then reconstruct the output from this representation.

The metric based on AE [13] informs us about how the correlations between the different variables of the system behave. Given that, the metric q_{rec} is based on the difference between the input and the output value of the AE, in such a way that the greater the reconstruction error, the less concordance there will be between the variables [14].

$$q_{rec} = \sum_{i=1}^{N} |x'_i - x_i| \quad (12)$$

where x_i correspond to the features of the data taken by the sensors, N is the number of total features. On the other hand x'_i is the value of the vector of variables reconstructed by the AE. Sometimes $|x'_i - x_i|$ is known as a reconstruction error and is represented as E_{rec}. Since this metric is based on a difference between two values, it can take any real value greater than 0, in such a way 0 is the value with the highest quality.

3.3 Geospatial-based metrics

Considering sensors' location is also highly relevant for knowledge extraction. In this sense, we also provide two metrics that use interpolation methods for assessing how well a sensor is coordinated and correlated with its peers according to their distance. The used models are Inverse Distance Weighting (IDW) [15] and Bayesian Maximum Entropy (BME) [16]. IDW is a deterministic estimation method in which, assuming that the near sensors are more similar, a weighted average of available values at known points is used to calculate unknown data points. BME is a knowledge-based probabilistic modeling framework for spatial and temporal information. It allows various knowledge bases to be used as prior information, and the determination rules for hard (high precision) and soft (low precision) data are logically incorporated into the modeling. Like previously, we calculate the difference between the interpolated and the real measure, and the average value will become the metric, named:

- q_{inv_mean} and q_{inv_med} for IDW.
- q_{BME_mean} and q_{BME_med} for BME.

4. Examples of implementation

In this section, 3 different IoT scenarios are introduced, in which the previous metrics are computed and highlight the possible drawbacks.

4.1 Parking data

This data was collected from 5 private parking sensors located in the city of Murcia[1], Spain.

First, the variables that are useful for our goal had to be chosen: the timestamp and the parking occupation measurements and aggregated the data in 10 minutes intervals.

This aggregation can generate redundancies on the timestamps, so the result has been averaged. Storing information about this aggregation process will be useful for the Artificiality metric.

NA (not available) instances have been kept since due to their importance in obtaining some quality metrics (Completeness). Given that the data is not measured periodically, a lot of missing values are generated at this point. For illustrative purposes, a new variable called real_time was computed, which adds a random delay to the timestamps, simulating that the data needs some time to be stored. These are some highlights:

- Completeness: it consists on counting instance by instance the percentage of non-absent values there are.
- Timeliness: the random time lag that is included in the data (T_{age}) is used, so when it is divided by the arbitrary aggregation time W (600 seconds, in this case) it shows the time that data takes to be available, as follows

$$q_{tim} = 1 - \frac{T_{age}}{W} \tag{13}$$

- Plausibility: if the data of each parking lot belongs to the interval $[0, C_i]$, this measure will be said to be plausible and will receive a value of 1. The values of C_i are: 330, 312, 305, 162 and 220 respectively.
- Artificiality: due to aggregation over time, the number of instances used for computing the mean and therefore the aggregated value were considered. Thus, if a data was obtained by means of two data-points taken in the same time frame, its metric of artificiality will be $\frac{1}{2}$.
- Concordance: the geostatistical metrics have been used for covering this concept.
- Outliers: given the amount of missing data, the ARIMA framework could not be used for detecting outliers in this dataset.

A subset of the quality metrics and data values are shown in **Table 1**. Where Park101, … , Park105 are the parkings' ids, as we can see there are many instances

[1] Their locations are stored in the following web address http://mapamurcia.inf.um.es/

that cannot be correct, that information is condensed in the quality metrics. Whereas **Figure 1** shows the histograms of all basics metrics that could be computed for the parking dataset. In **Figure 2** the histogram of outlier-based metrics is shown.

Parking data geospatial-based metric's histograms are shown in **Figure 3**. As it was said above, the calculation of these metrics replace the calculation of the concordance metric, because they provide information about the correlation of the different sensors, in this case, the lower the value of the metric, the better.

4.2 Luminosity data

In this section, the monitored luminosity from 4 sensors located in the Pleiades building of the University of Murcia was studied.

timestamp	Park$_{101}$	Park$_{102}$	Park$_{103}$	Park$_{104}$	Park$_{105}$
11:50:00	NA	163.33	NA	NA	117.5
12:00:00	NA	10000	NA	NA	116.5
12:10:00	NA	163.00	NA	10000	116.5
12:20:00	NA	165.00	NA	NA	118.0
12:30:00	NA	166.00	NA	NA	120.0
12:40:00	−1	166.50	NA	NA	119.0

Table 1.
Parking observations (number of cars) subset.

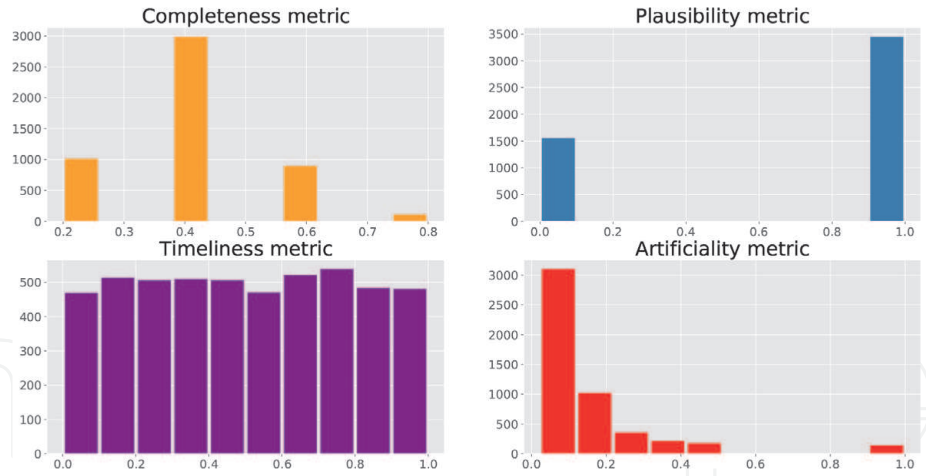

Figure 1.
Parking basic metric's histograms.

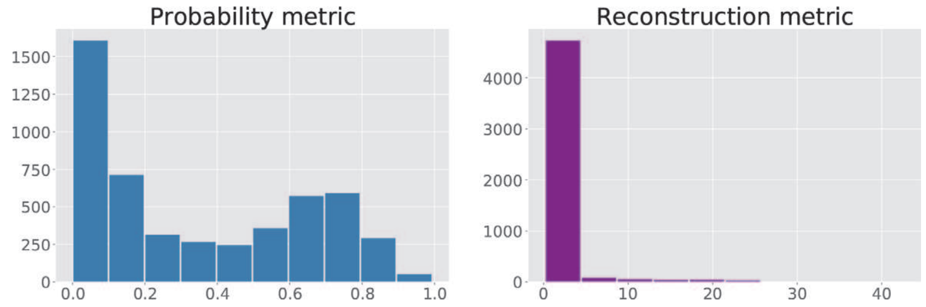

Figure 2.
Parking outlier-based metric's histograms.

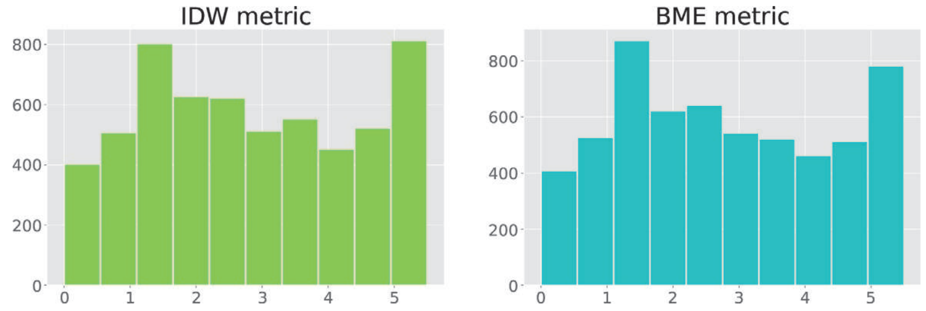

Figure 3.
Parking data geospatial-based metric's histogram.

Time	S_1	S_2	S_3	S_4
18:00:00	20	55	10	80
18:10:00	25	70	20	40
18:20:00	NA	70	10	NA
18:40:00	20	95	10	65
18:50:00	30	30	20	60
19:10:00	20	75	10	280

Table 2.
Luminosity (lumens) data subset.

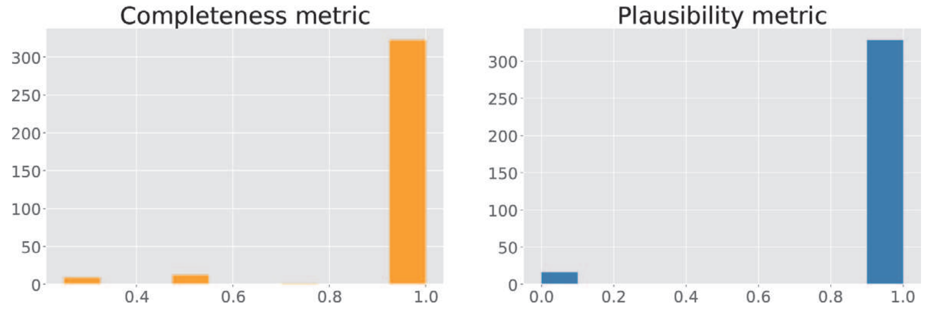

Figure 4.
Luminosity basic metric's histograms.

First, the data is aggregated using the timestamp as in the previous section, choosing a 10 minutes aggregation time. **Table 2** shows the aggregated values and also some of the computed metrics.

Figure 4 shows the histograms of all metrics that could be computed for the luminosity dataset together with basic statistics. The timeliness metric could not be calculated, since there are no signs of any lag in the data's storage. Also, the artificiality value always takes the value of 1 because the timestamps of the data are far apart. The rest of metrics are included in **Figure 5**.

By last, in **Figure 6** the geospatial luminosity's metrics can be seen. As in the case of parking, these metrics replace the concordance metric.

4.3 Pollution data

Given that the only way to calculate concordance on previous datasets has been through spatial interpolation due to poor dataset quality, a dataset of high quality has been used to compare the values that this metric takes in this situation and when they are added to it some imperfections.

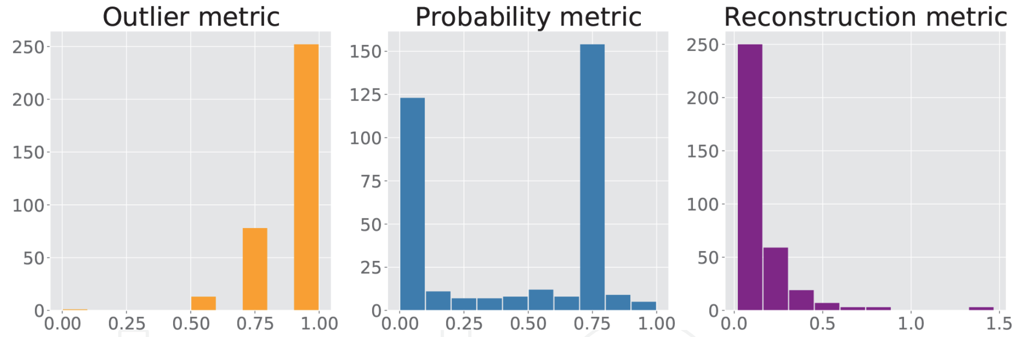

Figure 5.
Luminosity outlier-based metric's histograms.

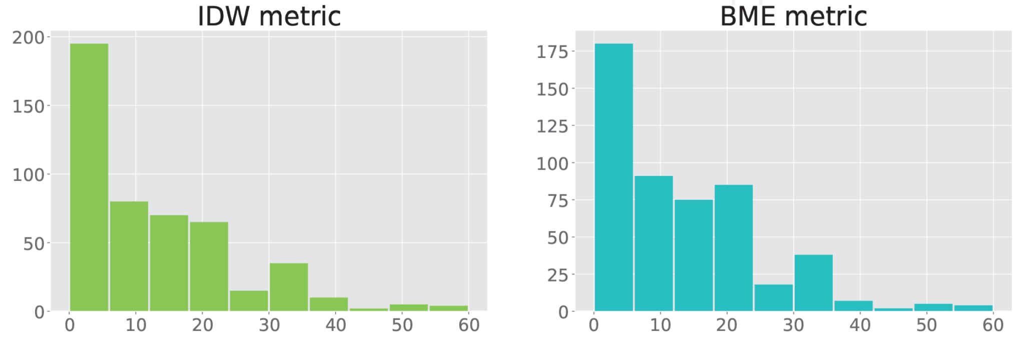

Figure 6.
Luminosity data geospatial-based metric's histogram.

Ozone	Particulate matter	Carbon monoxide	Sulfur dioxide	Nitrogen dioxide
0.18	−0.23	−1.03	−1.73	−1.66
0.29	−0.17	−1.05	−1.67	−1.68
0.31	−0.21	−1.03	−1.77	−1.70
0.22	−0.30	−0.99	−1.73	−1.66
0.27	−0.23	−1.03	−1.83	−1.77

Table 3.
Pollution data subset.

As can be seen in **Table 3**, this dataset has five variables that inform on the pollution of the atmosphere every five minutes, the data values are scaled.

Now the data are given, one way to calculate the concordance metric is to calculate the correlation between a value and the previous one, in such a way that if when the data is taken properly this value will be very close to 1, while if the data suffers any problem, this value will move away from 1. This is shown in **Figure 7**, in which we have the original dataset on the left side and the same dataset on the right side to which anomalous values have been added randomized, as it can be seen, the agreement values change significantly.

It should be noted that if the rest of the metrics are calculated in the case of the unaltered dataset, they will take perfect values, that is, they will always indicate a high quality of the dataset.

For this dataset, the rest of the metrics have been calculated, however, the results have not been added, since the dataset presents a high quality and therefore the results are not of great interest since the histograms of the metrics take the ideal behavior.

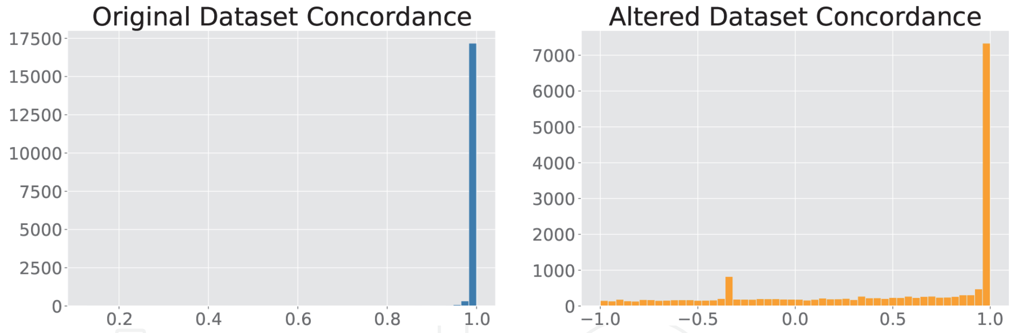

Figure 7.
Pollution dataset concordance comparison.

V1	V2	V3	V4	V5
−0.606470	−0.143913	−0.348654	2.468968	0.199896
−0.543575	0.073679	−0.223220	−0.594037	0.223077
−0.543575	−0.143913	0.090365	−0.594037	0.223077
−0.543575	−0.216444	0.090365	−0.733265	0.199896
−0.606470	−0.796689	0.090365	−0.872493	0.176716

Table 4.
Data without context subset.

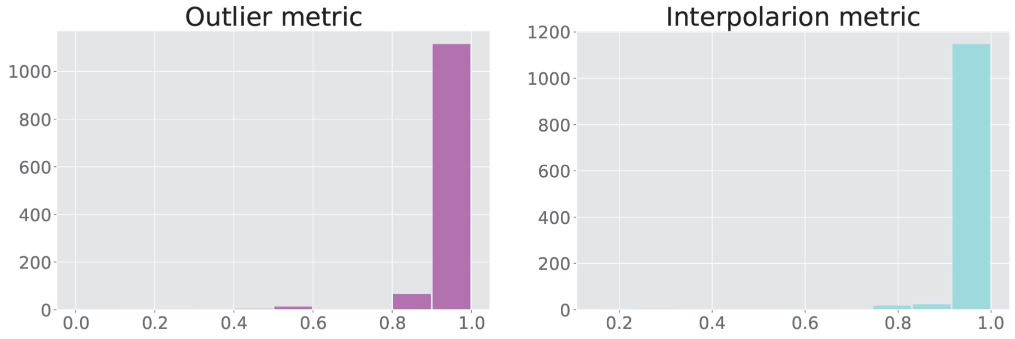

Figure 8.
Data without context outlier-based metric's histograms (I).

4.4 Data without context

For demonstration purposes, we propose to compute the quality metrics in a dataset whose context, origin and meaning are unknown. It is a dataset in which we have no knowledge about what the columns represent, how the data was collected and the timestamp of the observations. In such scenario, the only basic metric that can be computed is completeness. However, outlier-based metrics are very useful, since they consider the variables as plain time series without taking into account their physical meaning. **Table 4** shows a subset of the dataset, that presents 5 unknown variables, with 1200 instances.

Figure 8 shows the histograms of the outlier-based metrics, while **Table 5** shows the value taken by the metrics for a small data subset.

Similarly, the probabilistic and reconstruction metrics can be calculated here, since they do not assume any kind of knowledge of the data. In **Figure 9** the histogram of both metric is shown.

q_{outlier}	q_{inter}
1.00	1.00
0.90	0.80
0.85	0.80
1.00	1.00
1.00	1.00

Table 5.
Data without context subset.

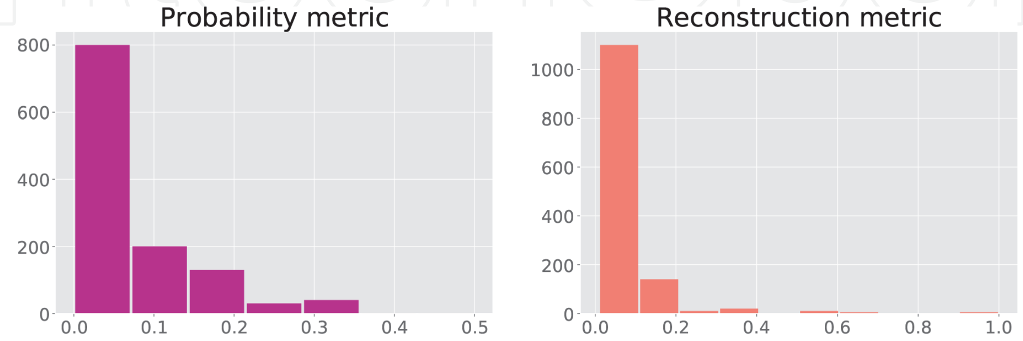

Figure 9.
Data without context outlier-based metric's histograms (II).

5. Conclusions

The proliferation of datasets thanks to the new paradigm of the Internet of Things, is populating repositories and open data platforms with data that could be of great use for the scientific community and the technologists to catalyze the growth of scientific knowledge and to make proliferate the creation of new technological solutions. Although all data has value, a point has been reached in which it is necessary to rapidly recognize the quality of a dataset, or a data stream, ideally on an only manner.

In this chapter, several concepts have been combined in order to measure the quality of data from IoT-based real-time streams (tested on real-world) sensor systems.

Three sets of quality assurance methods, descriptive, analytic and geometrical have been developed that can be used as levels of a given evaluation, or independently depending on the nature of the datasets to be evaluated.

It has been shown that the metrics can be an standard on the calculation of data quality and the majority can be applied independently on the problem context. At the same time, basic concepts that must be present in any system in which the quality of the data is to be guaranteed have been reviewed. Furthermore, it has been shown how it is possible to obtain quality metrics when knowledge about the data is limited.

The applications of this technology are linked to the proliferation of open data portals. There exist many initiatives and organizations that are working towards publishing data as open. The main funding body for engineering and physical sciences research in the UK, the Engineering and Physical Sciences Research Council (EPSRC) is supporting the management and provision of access to research data. They claim that *publicly funded research data should generally be made as widely and* *freely available as possible in a timely and responsible manner*[2]. Other initiatives are the EU Open Data Portal[3] at European level or the national-level ones such as Open Data Aarhus[4]. In that sense, the selection of data sources becomes more complicated given the great amount of data that researchers and practitioners have access to. Our system provides an easy, understandable and quick way to make an informed decision for choosing between several data sources based on data quality.

As future work, we are considering several technologies in order to make our metrics available to researchers and businesses. We consider that they have the potential to become a standard for measuring data quality.

Acknowledgements

This work has been sponsored by MINECO through the PERSEIDES project (ref. TIN2017-86885-R), by ERDF funds of project UMU-CAMPUS LIVING LAB EQC2019-006176-P by the European Comission through the H2020 IoTCrawler (contract 779852), and DEMETER (grant agreement 857202) EU Projects. It was also co-financed by the European Social Fund (ESF) and the Youth European Initiative (YEI) under the Spanish Seneca Foundation (CARM).

[2] https://epsrc.ukri.org/about/standards/researchdata/

[3] https://data.europa.eu/euodp/en/home

[4] www.opendata.dk/city-of-aarhus

Author details

Tomás Alcañiz, Aurora González-Vidal*, Alfonso P. Ramallo
and Antonio F. Skarmeta
Department of Information and Communication Engineering, Faculty of Computer Science, University of Murcia, Murcia, Spain

*Address all correspondence to: aurora.gonzalez2@um.es

References

[1] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," Communications of the ACM, vol. 40, no. 5, pp. 103–110, 1997.

[2] C. Bisdikian, L. M. Kaplan, M. B. Srivastava, D. J. Thornley, D. Verma, and R. I. Young, "Building principles for a quality of information specification for sensor information," in *2009 12th International Conference on Information Fusion*. IEEE, 2009, pp. 1370–1377.

[3] C. H. Liu, J. Fan, J. W. Branch, and K. K. Leung, "Toward qoi and energy-efficiency in internet-of-things sensory environments," IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 4, pp. 473–487, 2014.

[4] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data science journal*, vol. 14, 2015.

[5] C. Liu, P. Nitschke, S. Williams, and D. Zowghi, "Data quality and the internet of things," Computing, vol. 102, 02 2020.

[6] C. Lagoze, "Big data, data integrity, and the fracturing of the control zone," Big Data & Society, vol. 1, 11 2014.

[7] M. Celma, J. C. Casamayor, and L. Mota, *Bases de datos relacionales*. Alhambra, 2003, ch. 1, pp. 1–16.

[8] A. González-Vidal, T. Alcañiz, T. Iggena, E. Bin Illyas, and A. F. Skarmeta, "Domain agnostic quality of information metrics in iot-based smart environments," in *Intelligent Environments 2020: Workshop Proceedings of the 16th International Conference on Intelligent Environments*, vol. 28. IOS Press, 2020, p. 343.

[9] D. Kuemper, T. Iggena, R. Toenjes, and E. Pulvermueller, "Valid. iot: a framework for sensor data quality analysis and interpolation," in *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 2018, pp. 294–303.

[10] C. Chen and L.-M. Liu, "Joint estimation of model parameters and outlier effects in time series," Journal of the American Statistical Association, vol. 88, no. 421, pp. 284–297, 1993.

[11] Javier López-de-Lacalle. Detection of Outliers in Time Series. 2019. R package version 0.6-8. https://CRAN.R-project.org/package=tsoutliers

[12] A. González-Vidal, J. Cuenca-Jara, and A. F. Skarmeta, "Iot for water management: Towards intelligent anomaly detection," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*. IEEE, 2019, pp. 858–863.

[13] M. Zolotukhin, T. Hmlinen, T. Kokkonen, and J. Siltanen, "Increasing web service availability by detecting application-layer ddos attacks in encrypted traffic," in *23rd International Conference on Telecommunications (ICT)*, 2016.

[14] N. García, T. Alcañiz, A. González-Vidal, J. B. Bernabé, D. Rivera, and A. Skarmeta, "Distributed real-time slowdos attacks detection over encrypted traffic using artificial intelligence," *Journal of Network and Computer Applications*, vol. 173, p. 102871, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804520303362

[15] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong, "An experimental comparison of ordinary and universal kriging and inverse distance weighting," Mathematical Geology, vol. 31, no. 4, pp. 375–390, 1999.

[16] A. González-Vidal, P. Rathore, A. S. Rao, J. Mendoza-Bernal, M. Palaniswami, and A. F. Skarmeta-Gómez, "Missing data imputation with bayesian maximum entropy for internet of things applications," *IEEE Internet of Things Journal*, 2020.