Intelligent Environments 2020
C.A. Iglesias et al. (Eds.)

doi:10.3233/AISE200059

Domain Agnostic Quality of Information Metrics in IoT-Based Smart Environments

AURORA GONZÁLEZ-VIDAL [a,1], TOMÁS ALCAÑIZ [a], THORBEN IGGENA [b], EUSHAY BIN ILYAS [b], and ANTONIO F. SKARMETA [a].

[a] *Department of Information and Communication Engineering, University of Murcia, Spain*

[b] *Lab for Mobile Communications, Faculty of Engineering and Computer Science, University of Applied Sciences Osnabrück, Germany*

Abstract. Thanks to the proliferation of IoT devices that are interconnected, huge amounts of data are being gathered nowadays. The availability of all these new sensors, data sources and open data platforms offers new possibilities for innovative applications and use-cases that are many times dynamic. However, if we plan to depend on data for the optimal provision of services, it is of utmost importance to ensure the quality of data and the quality of information that we are handling in an online manner. Furthermore, geolocalised data provides a richer context in which the quality of information can be measured and in which services are more advanced. In order to support the process of finding the right information, we have defined several metrics in single-sensor and multi-sensor scenarios that are based on statistical analysis, machine learning algorithms and contextual information. We have applied them in two scenarios: smart parking and environmental sensing for smart buildings.

Keywords. IoT, Quality of Information, Missing data imputation, Bayesian Maximum Entropy, Autoencoders

1. Introduction

With the increasing number of devices in the Internet of Things, the availability of data has massively increased. Smart cities, industry 4.0, social networks, or even agriculture have created dozens of new data sources. This allows the development of new applications and use cases[2]. Due to the heterogeneity and the sheer amount of data sources, there is a need to ensure high quality of the used data to avoid wrong decisions and to increase the user experience for smart applications.

A major requirement for these scenarios is quality analysis metrics for data sources and their monitored data. False or misleading information might cause problems during the processing and usage of the information. This problem reaches from simple misconfigured sensors, which deliver wrong information, to intentionally provided false information with malicious intent, which leads to malfunctioning systems and applications.

[1] Corresponding Author. E-mail: aurora.gonzalez2@um.es

[2] http://www.ict-citypulse.eu/scenarios/

To approach these problems, we integrate quality measures and analysis modules to rate data sources to identify the best fitting data sources to get the needed information.

There are many definitions of Data Quality (DQ). The two predominant ones are:

- Data is of high quality if the data is fit for the intended purpose of use
- Data is of high quality if the data correctly represent the real-world construct that the data describe

In many studies, complex algorithms claim to preserve data quality [1, 2], however, they lack straightforward definitions of what quality is. In this work, we define metrics for DQ and compute them in several IoT scenarios for checking their viability.

2. Related work and background

Research on data quality became popular starting in the context of databases containing data from multiple data sources. Strong et al. came to the point, that faulty data cost billions of dollars. As a solution, they came up with the term Quality of Information (QoI), which they defined in four categories and defined measurable metrics for each of them [3]. During the next years, several frameworks to address the QoI have been developed. In [4], a general QoI framework has been designed that allows creating measurement models for specific settings and domains. Further development in the context of QoI frameworks has been taken by Bisdikian et al. [5], who described context-independent quality measurements. To do so, they split the data quality into the terms QoI and Value of Information. QoI is also a topic of research in diverse IoT frameworks designed for innovative applications in Smart Cities [6] or in IoT search [7]. A deeper review of data quality for IoT scenarios is done in [8].

Nowadays, IoT data is present in many scenarios and contexts and there exist some domain-specific approaches for estimating DQ such as health [9] or energy consumption in smart grid [10] amongst others. These works define a set of tests as queries that are written to check the properties specified by domain experts using mathematical formulas or natural language, which cannot be easily updated or applied to other problems.

Data quality tools typically address data cleansing, data integration, master data management, and metadata management. Challenges of interest have an impact when choosing a tool: Incorrect data, duplicate data, missing data and other data integrity issues can significantly impact — and undermine — the success of an initiative.

In addition to research work, several business initiatives are proposed, such as Cloudingo[3] for removing duplicates using a graphical interface, Data Ladder[4] for matching and cleaning using templates or Informatica[5], that is a data quality validation tool that provides a set of data quality checks as queries to validate the syntactic properties of the target data, such as data type and not-null constraint checks. The tool allows users to specify semantic properties to be verified by the tests.

Although these tools are general enough for use in any project, they require domain knowledge to define and update the domain-specific properties. There is a lack of open tools that automatically generate the properties to be checked by the data quality tests.

[3] https://cloudingo.com/

[4] https://dataladder.com/

[5] https://www.informatica.com/products/master-data-management.html

In machine learning, clustering is an unsupervised technique that uses intrinsic properties of the data for dividing the studied subjects into several categories. In that division, subjects within a group are similar to each other and dissimilar to the subjects assigned to other groups according to a certain metric and without labelled data.

Clustering techniques have been used in the literature for IoT-based scenarios [11]. In particular, many clustering works have focused on the use of clustering for anomaly detection. Those techniques include, but are not restricted to, K-Means and Hierarchical Clustering [12], Expectation Maximization (EM) [13], Hyperellipsoidal Clustering algorithm for Resource-Constrained Environments (HyCARCE) [14] and Gaussian Mixed Models (GMM) [15]. Autoencoders are another representation learning approach. An autoencoder investigates an efficient encoding from the data in an unsupervised manner. They have also been used for anomaly detection [16]. One-class SVM [17] is an outlier detection algorithm in which the whole training data is assumed to belong to the normal class and any data point outside of this data region is considered as an outlier.

3. Data quality metrics

In this section, we describe the metrics that have been defined to calculate and annotate the QoI for IoT data.

3.1. QoI basic metrics

The first set of metrics are based on the ones developed for the IoTCrawler framework [7].

- Completeness (q_{cmp}): missing or unusable data instances are represented with this metric. It computes the percentage of the unusable data.
- Timeliness (q_{tim}): refers to the time expectation for accessibility and availability of information. In other words, expresses how long the time difference between data capture and the real world event being captured is. In critical IoT applications such as traffic safety and control and managing power systems knowing your timeliness requirements is fundamental.
- Plausibility (q_{pla}): this metric shows if received data is coherent according to the probabilistic knowledge of the variables that are being measured.
- 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 artificial spatiotemporally interpolated value.
- Concordance (q_{conc}): This metric is used to describe the agreement between information of the data source and the information of further 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 ($i = 1,...,n$).

3.2. Outlier-based metrics

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

We have used an Autoregressive Integrated Moving Average (ARIMA) based framework [18] in order to find innovational outliers, additive outliers, level shifts, temporary changes and seasonal level shifts [19]. The percentage of outliers in the studied sensor is named q_{out}. We have also studied how much the outlier deviates from what could be considered a normal observation. The outliers are imputed with missForest [20], an iterative Random Forest-based imputation method. 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}).

Another way to detect outliers is by using unsupervised methods. q_{prob} is the probability of belonging to a certain cluster that has been computed using Gaussian Mixture Models (GMM). It informs quantitatively of the anomalous values. The number of clusters is chosen using the silhouette coefficient.

We consider that AutoEncoders (AE) are also appropriate for this task because they learn the normal relationships inherent in the data and, therefore, when looking for anomalies they have a huge potential [21]. The metric based on AE informs us about how the correlations between the different variables of the system behave. 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. 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.

To sum up, we have defined 6 new metrics: q_{out}, q_{mean}, q_{median}, q_{mode} , q_{prob} and q_{rec}.

3.3. Geospatial-based metrics

The exact location of measuring the physical world through IoT is highly relevant to extract all insights. In that sense, we have also provided two metrics that are based on the interpolation of all datapoints as if they were missing by using geostatistical models. Those models are Inverse Distance Weighting (IDW) [22] and Bayesian Maximum Entropy (BME) [23]. IDW is a deterministic estimation method where the unknown data points are calculated with a weighted average of the values available at the known points, assuming that sensors that are close are more alike. BME is a knowledge-based probabilistic modeling framework for spatial and spatiotemporal information. It allows various knowledge bases to be incorporated in a logical manner as definite rules for prior information, hard (high-precision) and soft (low-precision) data into modeling.

As defined before, we have computed the difference between the interpolated value and the real one and its mean and median are going to be our metrics, named as:

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

4. Real scenarios and implementation

In this section, we introduce 3 different IoT scenarios in which the previous metrics are computed and highlight the possible drawbacks. We also introduce how we have used OpenCPU to implement our calculations as a service that is available at any time.

4.1. Parking Data.

This data was collected from 5 private parking sensors located in the city of Murcia [6]. Fig. 1 shows the location of these parking areas.

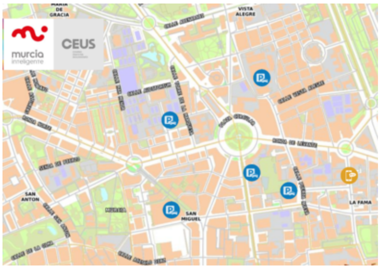


Figure 1. Parking zones

First, we have selected those variables that are useful for our goal: the timestamp and the parking occupation measurements and aggregated the data in 10 minutes intervals. This aggregation can generate redundancies on the timestamps, so we have averaged the result. Storing information about this aggregation process will be useful for the Artificiality metric. We have kept *NA* (not available) instances since they are important for 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, we have created a new variable called real_time 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: we use the random time lag that is included in the data (T_{age}), so if we divide it by the arbitrary aggregation time W (600 seconds, in this case) we get the time that data takes to be available, as follows: $q_{tim} = 1 - \frac{T_{age}}{W}$.
- 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: Since we performed 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: we used the geostatistical metrics for covering this concept.
- Outliers: given the amount of missing data, we could not use the ARIMA framework 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 and T_{age} is the time lag. Fig. 2 shows the histograms of all metrics that could be computed for the parking dataset together with basic statistics.

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

timestamp	$Park_{101}$	$Park_{102}$	$Park_{103}$	$Park_{104}$	$Park_{105}$	q_{comp}	T_{age}	q_{tim}	q_{pla}	q_{art}
2018-03-28 11:50:00	NA	163.33	NA	NA	117.5	0.4	574	0.04	1	0.33
2018-03-28 12:00:00	NA	10000	NA	NA	116.5	0.4	596	0.01	0	0.50
2018-03-28 12:10:00	NA	163.00	NA	10000	116.5	0.6	11	0.98	0	0.33
2018-03-28 12:20:00	NA	165.00	NA	NA	118.0	0.4	299	0.50	1	1.00
2018-03-28 12:30:00	NA	166.00	NA	NA	120.0	0.4	226	0.62	1	1.00
2018-03-28 12:40:00	-1	166.50	NA	NA	119.0	0.6	468	0.22	0	0.50

Table 1. Parking observations (number of cars) and quality metrics subset

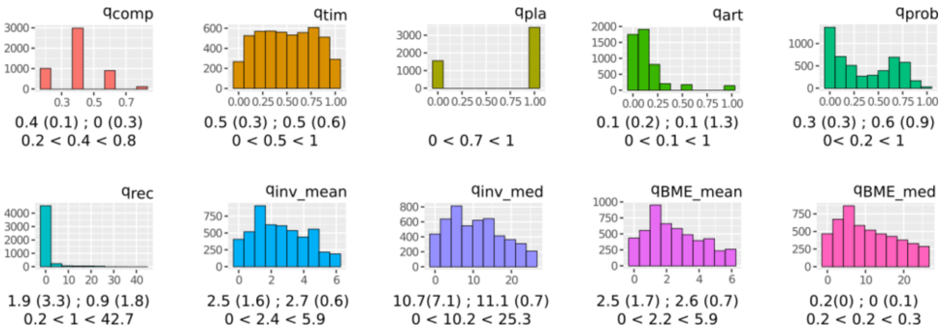


Figure 2. Parking metric's histograms and statistics. Statistics are: Mean (sd); IQR (CV) and min < mean < max

4.2. Luminosity Data

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

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.

time	S_1	S_2	S_3	S_4	q_{cmp}	q_{pla}	q_{art}	q_{prob}	q_{rec}
2020-01-21 18:00:00	20	55	10	80	1.00	1	1	0.001	0.54
2020-01-21 18:10:00	25	70	20	40	1.00	1	1	0.111	0.48
2020-01-21 18:20:00	NA	70	10	NA	0.50	1	1	0.827	0.33
2020-01-21 18:40:00	20	95	10	65	1.00	1	1	0.701	0.32
2020-01-21 18:50:00	30	30	20	60	1.00	1	1	0.110	0.29
2020-01-21 19:10:00	20	75	10	280	1.00	0	1	0.021	0.29

Table 2. Luminosity (lumens) metrics subset: completeness, plausibility and artificiality, anomaly probability and reconstruction metric

Fig. 3 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 we are not aware of any lag in the storage of the data. Also, the artificiality value always takes the value of 1 because the timestamps of the data are far apart. Therefore, it was

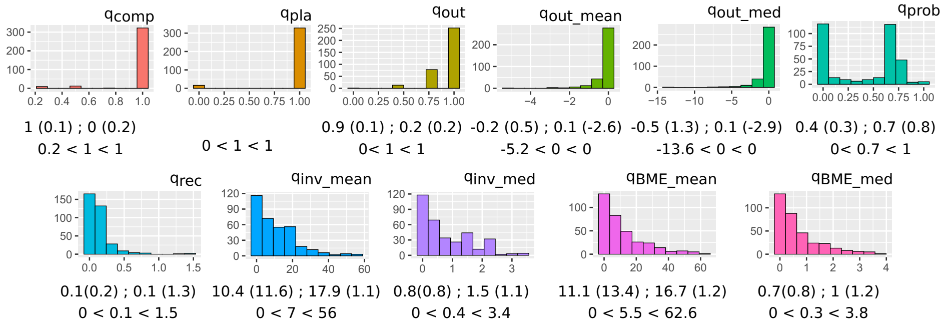


Figure 3. Luminosity metric's histograms and statistics. Statistics are: Mean (sd); IQR (CV) and min < mean < max

not included in Fig. 3. We have also computed the outlier metrics for this dataset using the ARIMA framework and the metrics created in subsection 3.2.

4.3. *Temperature (^oC) and humidity (%) data*

Finally, we have a series of temperature and humidity datasets collected by sensors located in the Pleiades building at the University of Murcia.

time	T_1	T_2	T_3	T_4	T_5	T_6	T_7	H_1	H_2	H_3	H_4	H_5	H_6
06:10:00	17.26	18.53	22.63	3.14e-193	23.40	16.74	23.18	34.90	37.05	36.09	40.02	8.88e-159	29.61
06:30:00	17.26	18.53	22.67	3.14e-193	23.51	16.74	23.18	34.90	36.90	35.87	39.92	8.88e-159	29.22
06:40:00	17.26	18.53	22.69	3.14e-193	23.41	16.74	23.18	36.20	36.35	35.53	39.82	8.88e-159	29.43
07:00:00	17.26	18.84	22.86	3.14e-193	23.39	16.74	23.18	35.65	35.71	35.23	39.61	8.88e-159	29.36
07:10:00	17.26	19.04	22.78	3.14e-193	23.31	16.74	23.18	35.40	35.63	34.89	39.97	8.88e-159	29.53
07:40:00	17.26	18.22	22.74	3.14e-193	23.42	16.74	23.18	35.38	35.24	35.65	39.74	8.88e-159	29.18

Table 3. Temperature (°C) and humidity (%) subset

Table 3 shows the temperature and humidity values for the different timestamps.

This dataset is in the ideal conditions to apply all our metrics, and the process goes as follows:

- The timestamp of the measurement is rounded and the average is taken.
- All datasets are joined, using a time union and the mean of the values is calculated.
- For the Plausibility metric, we chose as normal for temperature the interval $[10, 45]$ and for humidity, the interval $(0, 100)$. The metric takes the value 0 for the measurements out of these intervals. These values have been taken arbitrarily.
- For Artificiality it is enough to count the number of values that were used in the aggregation step. Because the frequency of data collection is very low and the time rounding has been taken from 10 minutes, all values take 1 in the Artificiality metric.
- Next, the outliers of the different time series are detected and the difference of the expected value between the outlier is calculated. This will constitute the different metrics previously calculated, that is, obtain the mean between these values, the mean weighted by the means of each of the sensor series, the similar case weighted by the median and by the mode.

- We perform similarly with the rest of the metrics

Fig. 4 shows the histograms of all metrics that could be computed for the temperature and humidity dataset together with basic statistics. Given that there are sensors that always present too low values, the Plausibility metric is always 0 and we have not included it in Fig. 4. Timeliness was again not possible to compute.

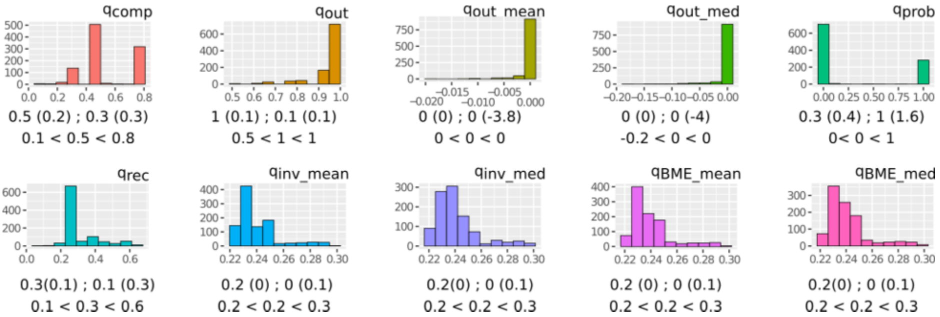
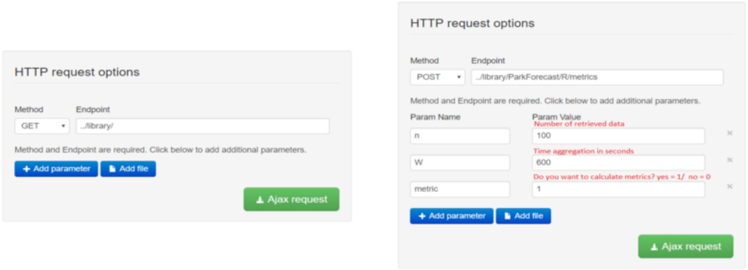


Figure 4. Temperature and Humidity metric's histograms and statistics. Statistics are: Mean (sd); IQR (CV) and min < mean < max

4.4. *OpenCPU implementation*

OpenCPU is a framework for embedded scientific computing and reproducible research. The OpenCPU server provides a reliable and interoperable HTTP API for data analysis based on R. We used OpenCPU so that our metrics can be computed by any user, since all state in OpenCPU is managed by controlling objects in sessions on a server [24]. For this purpose, an R package was constructed that performs all these calculations and supports a set of URLs in which the data is stored. The package's functions support the following parameters:

- n: number of data the function will return.
- W: proper time of the system in which the aggregation of times will be realized.
- metric: dummy variable to which a value is assigned depending on whether you want to show the quality metrics or not.



HTTP request options

Method: GET
Endpoint: ../library/

Method and Endpoint are required. Click below to add additional parameters.

HTTP request options

Method: POST
Endpoint: ../library/ParkForecast/R/metrics

Method and Endpoint are required. Click below to add additional parameters.

Param Name	Param Value
n	Number of retrieved data 100
W	Time aggregation in seconds 600
metric	Do you want to calculate metrics? yes = 1/ no = 0 1

Figure 5. OpenCPU interface

When we access the IP where the OpenCPU service is hosted, athe interface in Figure 5 (left) appears. In order to compute the metrics a POST method shold be stablished and the endpoint should point the appropriate package and functions, in this case *Park-Forecast* and *metrics*.

5. Conclusions and Future Work

In this paper we use a combination of concepts for the calculation of Quality of Information for real-time IoT-based sensor systems and shown its application to a more dataset based approach. The inclusion of an outlier detection framework allows for a more descriptive analysis of the sensor and its data. As future work we plan to incorporate our metrics into real-time systems and see how they influence the quality of results in further analysis and service provision.

Acknowledgements

This work has been sponsored by MINECO through the PERSEIDES project (ref. TIN2017-86885-R), 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).

References

[1] Y. Fathy and P. Barnaghi, "Quality-based and energy-efficient data communication for the internet of things networks," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 318–10 331, 2019.

[2] A. Gonzalez-Vidal, P. Barnaghi, and A. F. Skarmeta, "Beats: Blocks of eigenvalues algorithm for time series segmentation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 11, pp. 2051–2064, 2018.

[3] R. Y. Wang, D. M. Strong, and L. M. Guarascio, "Beyond accuracy: What data quality means to data consumers," *J. of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[4] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith, "A framework for information quality assessment," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1720–1733, 2007.

[5] C. Bisdikian, L. M. Kaplan, and M. B. Srivastava, "On the quality and value of information in sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 9, no. 4, p. 48, 2013.

[6] D. Puiu, P. Barnaghi, R. Toenjes, D. Kümper, M. I. Ali, A. Mileo, J. X. Parreira, M. Fischer, S. Kolozali, N. Farajidavar *et al.*, "Citypulse: Large scale data analytics framework for smart cities," *IEEE Access*, vol. 4, pp. 1086–1108, 2016.

[7] 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.

[8] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, "Data quality in internet of things: A state-of-the-art survey," *Journal of Network and Computer Applications*, vol. 73, pp. 57–81, 2016.

[9] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *Egems*, vol. 4, no. 1, 2016.

[10] W. Chen, K. Zhou, S. Yang, and C. Wu, "Data quality of electricity consumption data in a smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 98–105, 2017.

[11] D. Puschmann, P. Barnaghi, and R. Tafazolli, "Adaptive clustering for dynamic iot data streams," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 64–74, 2016.

[12] M. S. Parwez, D. B. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2058–2065, 2017.

[13] M. Goldstein, "Fastlof: An expectation-maximization based local outlier detection algorithm," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012).* IEEE, 2012, pp. 2282–2285.

[14] P. Rathore, A. S. Rao, S. Rajasegarar, E. Vanz, J. Gubbi, and M. Palaniswami, "Real-time urban microclimate analysis using internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 500–511, 2017.

[15] J. Diaz-Rozo, C. Bielza, and P. Larrañaga, "Clustering of data streams with dynamic gaussian mixture models: an iot application in industrial processes," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3533–3547, 2018.

[16] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 665–674.

[17] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[18] 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.

[19] 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.

[20] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[21] Y. Ikeda, K. Ishibashi, Y. Nakano, K. Watanabe, and R. Kawahara, "Anomaly detection and interpretation using multimodal autoencoder and sparse optimization," *arXiv preprint arXiv:1812.07136*, 2018.

[22] 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.

[23] 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.

[24] A. Zafeiropoulos, E. Fotopoulou, A. González-Vidal, and A. Skarmeta, "Detaching the design, development and execution of big data analysis processes: A case study based on energy and behavioral analytics," in *2018 Global Internet of Things Summit (GIoTS).* IEEE, 2018, pp. 1–6.