Data Quality Estimation in a Smart City's Air Quality Monitoring IoT Application

Julio H. Buelvas P.
Faculty of Engineering
Universidad de Antioquia
Medellin, Colombia
julioh.buelvas@udea.edu.co

Fernando E. Avila B.
Faculty of Engineering
Universidad de Antioquia
Medellin, Colombia
fernando.avila@udea.edu.co

Natalia Gaviria G.
Faculty of Engineering
Universidad de Antioquia
Medellin, Colombia
natalia.gaviria@udea.edu.co

Danny A. Munera R.
Faculty of Engineering
Universidad de Antioquia
Medellin, Colombia
danny.munera@udea.edu.co

2021 2nd Sustainable Cities Latin America Conference (SCLA) | 978-1-6654-3768-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/SCLA53004.2021.9540154

Abstract—**With the upcoming growth of the Internet of Things (IoT), which is translated into millions of interconnected devices reporting a high volume of data coming from heterogeneous sources (sensors), it is necessary to assess the confidence of data in order to provide the system with trustable information that can be used to get real insights from the physical world and thus take proper decisions or actions over it. Having in mind that ensuring data quality is key to ease user engagement, acceptance of IoT services and large scale deployments [1], a new critical issue arises which is related to the quality of the data in IoT. Some applications might have a different definitions and indicators for data quality (DQ) and thus different threshold for acceptance of the data. In this work, we explore a smart city application in the field of environmental monitoring and identify the related DQ indicators that apply within this context.**

Our approach is evaluated over a real dataset retrieved from SIATA's citizen scientist low-cost sensor network, an air quality monitoring system that can be encompassed within the IoT paradigm and that is composed by more than 200 nodes deployed all over the Aburra Valley in Antioquia, Colombia. The results show that feasibility assessing data quality and importance data quality awareness for an IoT application, as a tool for it to take proper actions on the real world.

Index Terms—**Smart City, Internet of Things, Air Quality, Environmental Monitoring, Low-cost Sensor, Data Quality.**

I. Introduction

Smart city is one of the most popular and targeted applications that can be leveraged by IoT, with the goal of improving the quality of citizen's lives by using a distributed information infrastructure to ensure the sustainability and the competitiveness of several functions through the integration of different dimensions of urban development and investments [2]. Within the context of smart cities, IoT helps to integrate different subsystems to improve the quality of services provided to citizens. This integration aims to achieve the best (most efficient) use of public resources in cities [3], including smart subsystems, such as the smart grid, mobility/transportation, health, governance, and smart environment.

Smart environmental applications for smart cities aim at monitoring hydrological, meteorological and air quality variables, to feed the government and citizens with important information about the current level of gases and particles in different areas of the city. In a smart city, this information can be used to take decision in terms of public policies to prevent the negative impact of this pollutants on people's health.

The data generated by IoT systems, however, is exposed to many endangering factors, since these applications usually involve wide deployments and open platforms [1], [4]. This fact has lead to a significant concern regarding the data reliability and trustworthiness. Particularly, in the context of environmental monitoring, several researchers argue that the use of low-cost sensors is generating unreliable data in IoT applications. This has led to the deployment of robust air quality monitoring networks, with very few nodes due to the cost of each station. In the past few years, a new trend has emerged in which several low-cost sensors are deployed in order to increase the spatio-temporal resolution of the measurements. However, since poor data quality leads to bad decisions, the data gathered by the low-cost sensors is rarely taking into consideration. Authors in [2] show that imperfect information can have an adverse effect over the performance of urban services and decision making. This situation poses a new challenge in the context of smart cities and IoT: it is necessary to assess the quality of data in order to improve the efficiency of smart cities, optimizing the use of limited resources, and providing reliable information to make adequate decisions.

In this article, we study the data quality of an air quality monitoring application, by analyzing the data gathered by two types of system in the city of Medellín, Colombia. The paper starts by presenting a definition of Data Quality in the context of IoT in Section II. The application and the specific data quality indicators are introduced in Section III. This is followed by Section IV, where we present the specific layout of the air quality monitoring system deployed by SIATA (a government project in the city of Medellín), which we use as our case of study. Sections V and VI describe and discuss the implementation and the results. Finally, the conclusions and future work ideas are presented in section VII.

II. Data Quality in the IoT Context

The development of the communication and social networks, has shown us that data can be treated as a product [5], [6], that can be very valuable. In that sense, the consumers of this product must assess the data quality (DQ) in order to decide whether it fits the needs or purposes for which it

978-1-6654-3768-4/21/$31.00 ©2021 IEEE

has been gathered. Roughly speaking, the term quality has been defined both as ”fitness for use” or as ”conformance to requirements” [4], [5], [7]. The meaning of this concept within the context of IoT can be viewed as the need for the application or end user to decide whether the gathered data complies with the requirements. This evaluation, however, is not always straightforward and may require further processing of the information. To evaluate DQ from a customer perspective, the work presented in [5] defines a set of attributes (hereafter dimensions) that are important to the data consumer.

The original definition of DQ comes from the context of information systems and databases. Authors in [1] and [4] have identified a set of DQ dimensions that are related to IoT. Some of the most important dimensions that can be found in the literature, including these from [1], [4], are: accuracy, precision, confidence, completeness, timeliness, data volume, data redundancy, concordance, validity, accessibility, interpretability, trust, artificiality and access security.

Since each IoT application is unique, it is then expected that it will have a specific set of DQ dimensions, that will be defined according to its ”fitness for purpose”. In the context of air quality monitoring systems, the Data Quality Objectives (DQO) are the levels of accepted thresholds for Data Quality Indicators (DQI). Worldwide, there are two main entities that have defined such indicators and requirements, namely *The European Parliament And The Council* in the European Union (*DIRECTIVE 2008/50/EC* [8]), and the *Environmental Protection Agency (EPA)* in the United States (*Quality Assurance Handbook for Air Pollution Measurement Systems* [9]). Using these guidelines, we have identified the most important DQ indicators, such as accuracy, uncertainty, bias, precision, completeness, minimum data capture, minimum time coverage, minimum number of sampling points, representativeness, comparability and detection limit. In order to assess the DQ within the context of air pollution monitoring in smart city, we have matched these indicators to the previously mentioned DQ dimensions, as shown in Figure 1. For instance, the minimum time coverage and the minimum data capture are DQ indicators defined by the EPA. Since neither of these have been defined as a DQ dimension, our mapping matched these quantities to the completeness of the data, as defined by [1], [4].

III. DQ Indicators in Air Quality Monitoring

This section introduces the definition of the most relevant DQ indicators, which are analyzed with the data gathered by the system sensors.

A. Uncertainty

According to [10], uncertainty is “a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could be reasonably be attributed to the measurand”. Also, [10] states 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, which is related to the representativeness of the sample and measurement uncertainty, which is related to the precision, bias and detection limit [9].

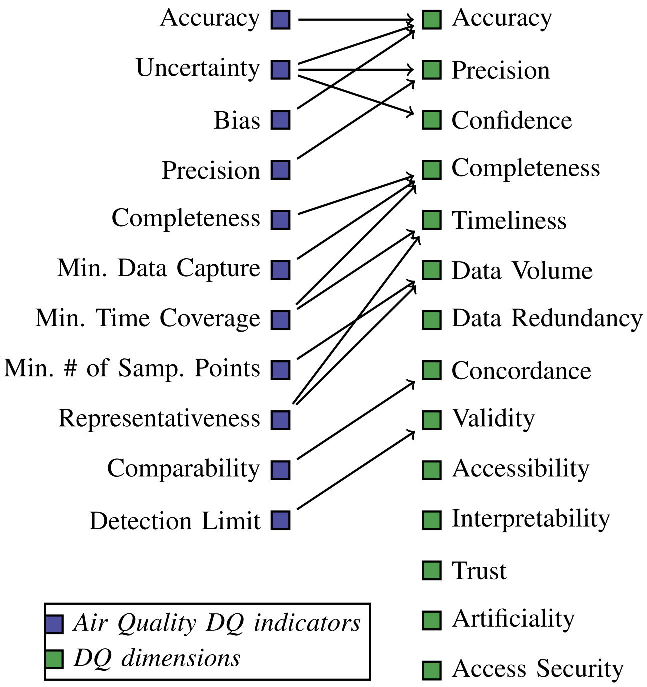

Fig. 1. Mapping Air Quality DQ indicators to DQ dimensions.

Regarding the DQO for particulate matter pollutants, the maximum allowed uncertainty for fixed measurements (robust monitoring stations) is 25%, while for indicative measurements (e.g., low cost sensors measurements) is 50% [8]. This indicator is related to data accuracy, precision and confidence.

B. Minimum Number of Sampling Points

This indicator is defined in [8] for fixed measurements and it depends on the population in a specific zone. It establishes the amount of monitoring stations regarding the population of an area. This indicator is related to data volume.

C. Precision

Precision 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 [9]. 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.

D. Accuracy

In [4], [11], it is defined as “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 use”. [4] also define it as “the extent to which an observation for the object truly reflects its real-world situation”. The accuracy as a data quality indicator is defined in [9] 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 match the dimension of the same name.

E. Comparability

In the EPA handbook [9], 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". E.g., 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 dimension.

F. Completeness

This indicator (from [9]) directly matches the definition of the data completeness DQ dimension as the ratio of valid obtained data to the expected data. EPA requires 75% of the data to consider it complete.

IV. Use case: Air Quality Monitoring

SIATA is the early warning system of Medellín and the Aburrá Valley, a science and technology project of the metropolitan area of the Aburrá Valley and the major's office of Medellín. Its objective is to identify and forecast the occurrence of natural and anthropic phenomena that alter the environmental conditions of the area, or that may generate risks to the population. To accomplish this goal a real-time monitoring system of hydrological and meteorological variables has been deployed, to timely delivery information to citizens, which added to educational processes and the development of community early warning systems, enable the protection of life and the environment in the region [12].

Regarding the air quality monitoring system and specifically for particulate matter (PM) measurements, the SIATA network counts with more than 20 robust manual and automatic monitoring stations and about 230 nodes composed of low-cost nodes (called *scientific citizens network*), both deployed over an area larger than 382 km². These networks measure air quality variables, such as PM2.5, PM1, PM10, temperature and relative humidity, among others. SIATA system gathers data and publishes it online, assigning a data quality flag to each data point. However, no further processing is applied. The DQ flags assigned to the measurements taken by the robust stations is the result of taking into account aspects like national standards, equipment providers' recommendations, validity range according to the historical data and international guidelines (e.g., Quality Assurance Handbook for Air Pollution Measurement [9]).

Robust SIATA stations report a measurement every hour, while scientific citizens low-cost nodes report measurements every minute. Scientific citizens nodes are implemented with a Davis 6830 sensor for measuring relative humidity and temperature, and two sensors for particle matter (PM1, PM2.5 and PM10) measurements, the HK-A5 Laser and the Nova - SDS011, called DF and Nova sensors, respectively.

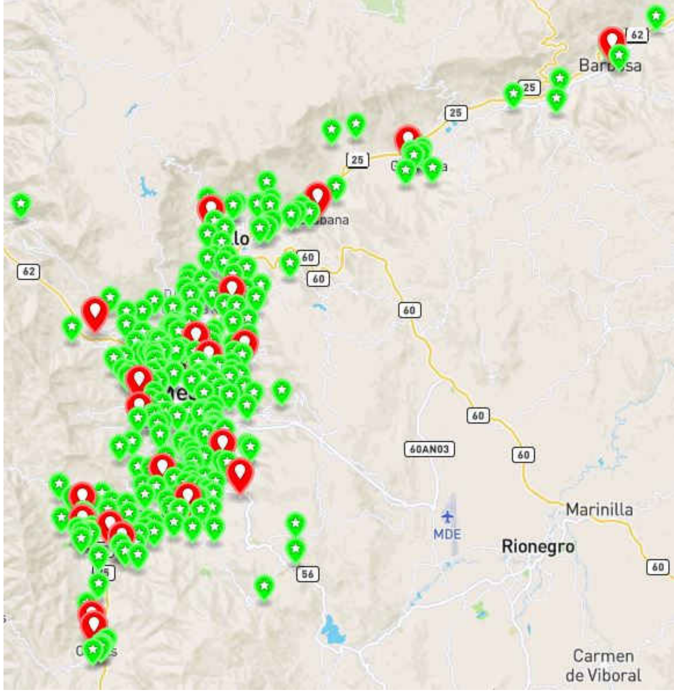

Fig. 2. Air quality monitoring stations in the Aburrá Valley (note that red spots are for robust stations and green spots are for low-cost nodes).

Regarding the number of sampling points, for zones like the Aburrá Valley in Antioquia-Colombia, with about 4 million inhabitants in 2020 [13], the minimum number is 11 if the maximum concentrations exceeds the upper threshold (the worst case) [8]. This zone has 22 fixed stations and about 230 nodes providing indicative measurements, thus complying with the DQO. Figure 2 displays the distribution of fixed and indicative stations and nodes in the mentioned region.

In this paper we assess the data quality of the scientific citizens network, which presents interesting challenges in data quality. On the one hand, as mentioned above, it is implemented an IoT monitoring system with low-cost sensors, which presumes a threat regarding the quality of the measures (calibration and adjustment needed). On the other hand, the nodes are physically located at the premises of volunteer citizens, thus nodes are exposed to endangering factors specific to their location (intermittent connection, bad placement, . . .). All these considerations make this dataset an interesting case to evaluate DQ.

V. Implementation

To assess the DQ in terms of completeness, accuracy (relative error), precision (one minus coefficient of variation), uncertainty (between sampler/instrument) and comparability (correlation, concordance), we developed an algorithm in python that implements the equations 1 to 5 and we focused our analysis in the variable 2.5μm-sized particulate matter (PM2.5).

$$completeness = \frac{\#ValidCollectedValues}{\#ExpectedValues} \quad (1)$$

Data out of range were deemed as invalid, as well as those above the maximum value in the box and whisker plot. The completeness was evaluated over the whole period with equation 1.

$$accuracy = \frac{|v_m - v|}{v} \quad (2)$$

For accuracy, equation 2 was used, where v_m is the observed or measured value by the DF or Nova sensor and v is the value accepted as true from the robust station. The station to which the node measurements are compared was chosen as the nearest one and the height is ignored, thus a bias is expected in this measurement. An hourly average was calculated for the two PM2.5 data of each node to get a one-hour measurement in order to compare it to the robust station data.

$$precision = 1 - \frac{\sigma}{\bar{v}_m} \quad (3)$$

Equation 3 is used for calculating the precision, where σ is the standard deviation from both the DF or Nova sensors, and $\bar{v}_m$ is the mean of the DF or Nova sensors measurement over the n number of observations. n was selected as 60 in order to measure the precision every hour.

$$uncertainty = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,DF} - y_{i,Nova})^2}{2n\bar{y}^2}} \quad (4)$$

Uncertainty is calculated by using equation 4, where $y_{i,DF}$ and $y_{i,Nova}$ are the readings from DF and Nova sensors respectively, n is the number of readings, and $\bar{y}$ is the average of all measurement results of the candidate method, i.e., both DF and Nova sensor readings [14]. n was selected as 60 in order to measure the uncertainty every hour. It is worth noting that there are several sources of uncertainty and in this study we only analyze the between sampler/instrument uncertainty.

$$comparability = |\rho_{x_0 x_i}| \quad (5)$$

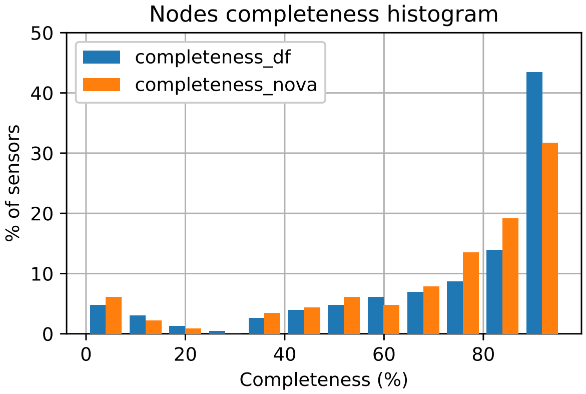

Fig. 3. Histogram for the completeness of DF and NOVA sensors

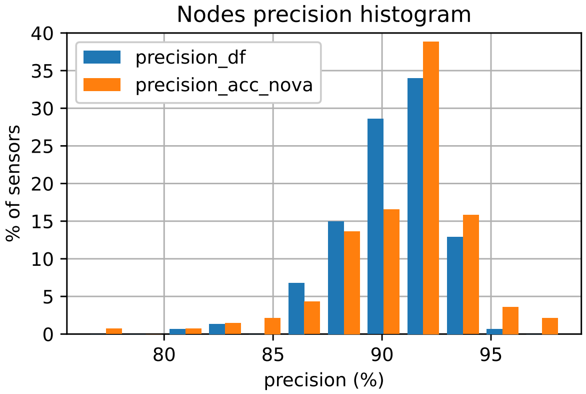

Fig. 4. Histogram for the precision of the measurements of DF and NOVA sensors

In order to evaluate the comparability, we propose the use of the correlation value as the DQ metric, since it indicates the level of agreement or concordance between variables. For our purposes, it is calculated as the absolute value of the Pearson's correlation coefficient between variables x_0 and x_i, as shown in equation 5.

VI. Results and Discussion

After processing the dataset, and specifically for the analysis of data quality of one month measurements (from 01/02/2020 to 29/02/2020) of the PM2.5 variable, the following results were obtained.

Figure 3 displays the histogram of the proportion of DF and Nova sensors regarding the completeness of their data. This figure shows that the majority of the sensors exhibits a good completeness in the considered period. If taking into account the 75% DQO, there are 147 out of 230 and 130 out of 230, DF and Nova sensors respectively, whose data is complete. Missing values can be caused by the initial cleaning of data out of range, the fact that sensors rely on user's power grid and internet service, both of which are exposed to outage, sensors out of services, malfunctioning, maintenance, etc. For the analysis of other indicators, we removed those sensors whose completeness did not comply with the DQO.

The next indicator is the precision, which is based on the standard deviation and shows the variability of PM2.5 sensor measurements. The coefficient of variation was first calculated, and the and subtracted from 100%. Figure 4 shows that the precision is greater than 75%, for both PM2.5 DF and Nova sensors. Statistically speaking, a coefficient of variation lower than 10% (in our case a precision greater than 90%) is desirable. It also tells that PM2.5 concentrations did not vary too much within one hour periods.

The histogram of the uncertainty is shown in Figure 5 (left). More than 70% of the nodes exhibit an uncertainty lower than the 20%. However, other sources of uncertainty are ignored and it could increase. If taking into account the DQO of 50%, most of the nodes are found to be compliant. Those nodes with a high uncertainty level exhibit such behavior because there is a difference between the measurements of DF and

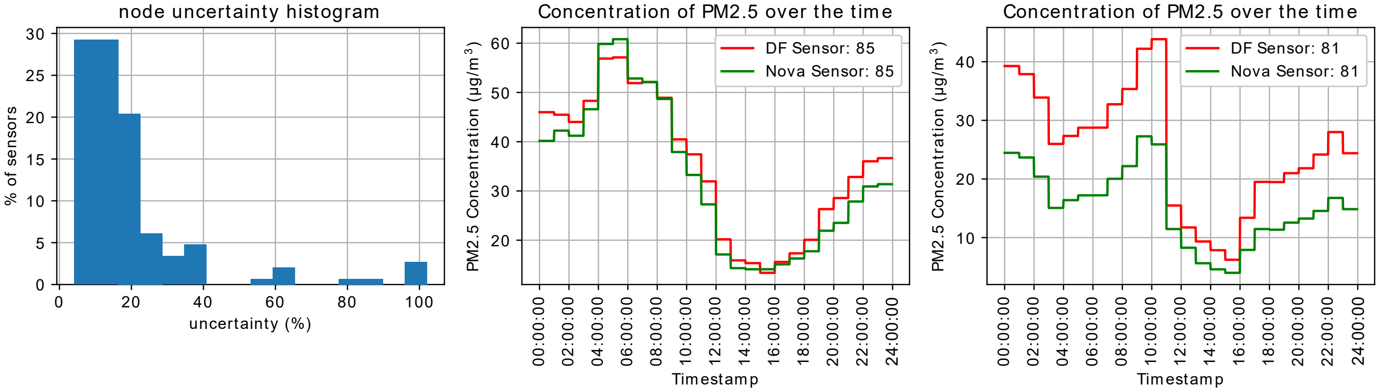

Fig. 5. Uncertainty. Left: histogram. Center: example of one node with low uncertainty. Right: example of one node with high uncertainty.

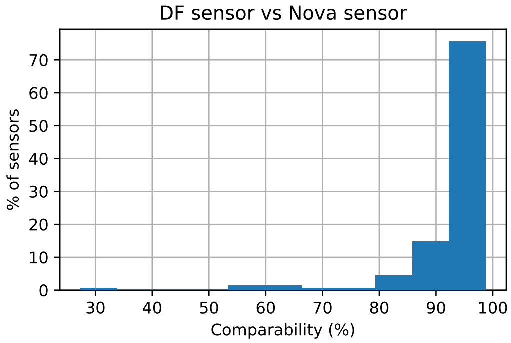

Fig. 7. Comparability between PM2.5 measurements of DF and Nova sensors

Nova sensors. This difference can be explained by lack of maintenance, loss of calibration and sensor aging. Using this indicator, one can easily identify sensors with low (Figure 5 center) and high (Figure 5 right) uncertainty. It is important to notice that 2,7% of the nodes (4 nodes) exhibit an uncertainty greater than 100%. This can be explained by a large difference between the two measurements of both sensors.

In order to compare the nodes PM2.5 measurements to an external reference, we used data from fixed stations. The comparison was performed in terms of the accuracy and the metric used is the relative error (the lower the better). Figure 6 (left) evidences that roughly 80% of nodes are within a range of 2km to the closest SIATA station that has measurements in the evaluated period. This distance will lead to a bias in the measurement of the true value, also for a region with such a topography as the Aburrá Valley, that measurement can change a lot, even more if not taking into account the height difference between the nodes and the SIATA stations. Figure 6 (right) shows that the average accuracy on more than half of the nodes is less than 40% which might be good taking into account previous considerations.

As for the comparability, we can verify in figure 7 that both sensors' measurements are highly correlated. Around 95% of the nodes exhibit a correlation larger than 80% between the PM2.5 measurement of the DF and Nova sensors. This result is confirmed by the uncertainty discussed above. In addition, the comparability to other variables such as SIATA PM2.5 measurement is low since for almost all the nodes is below 50%, not enough correlated. A reason for such value is tightly related to the distance as discussed for the accuracy results. Regarding comparisons to variables such as humidity and temperature, the results are better as shown in figure 8, but not conclusive because even if it is known that these variables impact on the measurements of PM2.5, we ignore their connection and influence in sensors' behavior.

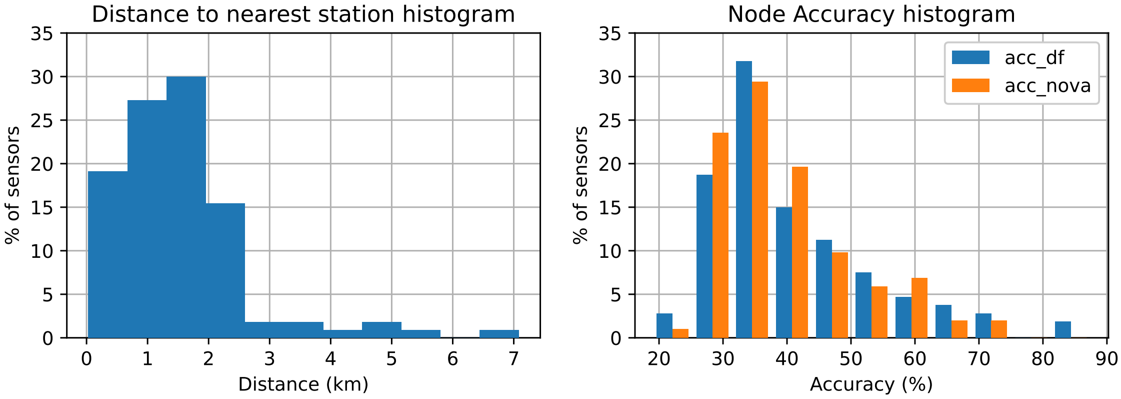

Fig. 6. Accuracy. Left: histogram for distance to reference station. Right: histogram for accuracy of the measurement of DF and NOVA sensors.

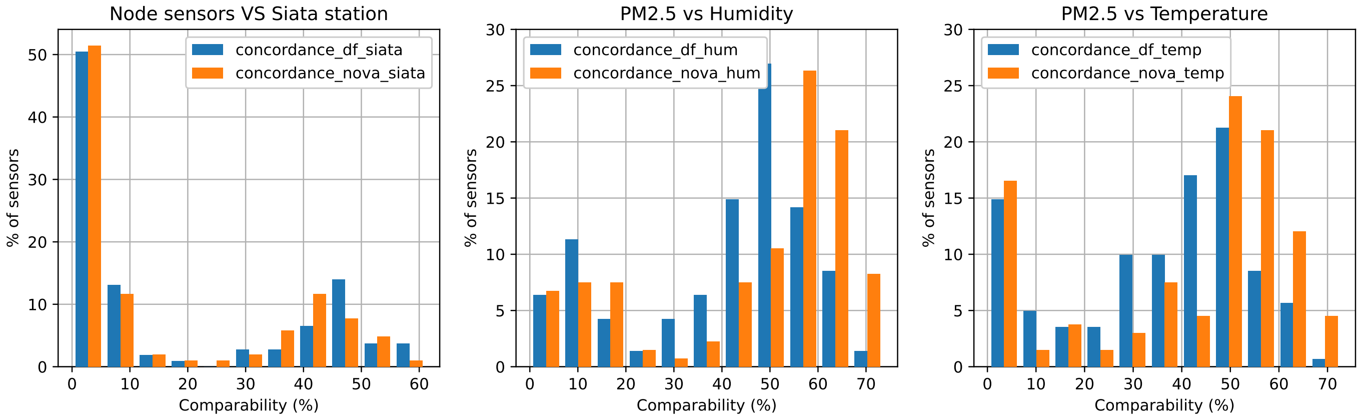

Fig. 8. Comparability. Left: between low-cost sensors and robust station (SIATA). Center: between PM2.5 measurements and humidity. Right: between PM2.5 measurements and temperature.

VII. Conclusions and Future Work

In this study, we have shown the feasibility of evaluating DQ in terms of single data attributes of an application, which provides information about the overall status and characterization of the system, as well as insights of underlying problems. This information can be used to perform maintenance and calibration of nodes to enhance their performance, it can also be used for node isolation and selection to provide trustable data to the data customer. Based on the results given and discussed in the previous section, it can be concluded that the overall DQ of the low-cost sensor is good and that the low-cost sensor network is complying with the cited DQO, evidencing the viability of their usage in conjunction to robust stations to increase the time and space monitoring resolution in the smart city.

This first DQ exploration in the context of air quality monitoring, opens the door to several lines of future work, some of them are:

- Build a model to estimate the true value at low-cost sensors' locations, based on robust stations measurements.
- Expand the analysis to other dimensions and indicators.
- As only one month data was analysed, it would be interesting to evaluate the evolution of DQ indicators through time and at different spans like hourly, weekly or even monthly.
- Develop a strategy for node and sensor selection based on their DQ indicators.
- Develop a strategy for sensor calibration and maintenance based on their DQ indicators.
- Research about the influence of sensor age, wind direction, temperature and humidity in the DQ.

Acknowledgment

The authors of this article kindly thank SIATA and its associates for their openness and willingness to share data and solve doubts about their air quality monitoring network.

References

[1] 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. [Online]. Available: http://dx.doi.org/10.1016/j.jnca.2016.08.002

[2] H. B. Sta, "Quality and the efficiency of data in "smart-cities"," *Future Generation Computer Systems*, vol. 74, pp. 409–416, 2017.

[3] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, 2017.

[4] C. Liu, P. Nitschke, S. P. Williams, and D. Zowghi, "Data quality and the Internet of Things," *Computing*, 2019. [Online]. Available: https://doi.org/10.1007/s00607-019-00746-z

[5] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[6] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, no. 2, pp. 58–65, 1998.

[7] M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications*. Springer, 2006.

[8] E. UNION *et al.*, "Directive 2008/50/ec of the european parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for europe," *Official Journal of the European Union*, 2008.

[9] U. E. P. A. EPA, *Quality Assurance Handbook for Air Pollution Measurement Systems*, 2017, vol. 2.

[10] JCGM, "Evaluation of measurement data — Guide to the expression of uncertainty in measurement," *International Organization for Standardization Geneva ISBN*, vol. 50, no. September, p. 134, 2008. [Online]. Available: http://www.bipm.org/en/publications/guides/gum.html

[11] ISO 25000 Portal, "ISO/IEC 25012," 2019. [Online]. Available: https://iso25000.com/index.php/en/iso-25000-standards/iso-25012?start=0

[12] SIATA, "Sistema de Alerta Temprana de Medellín y el Valle de Aburrá," 2021. [Online]. Available: https://www.siata.gov.co/sitio_web/index.php/home

[13] PROANTIOQUIA, Universidad EAFIT, Fundación Corona, Comfama, Comfenalco Antioquia, Cámara de comercio de Medellín para Antioquia, El Colombiano, Cámara de comercio de Bogotá, and El Tiempo, "Medellín cómo vamos," 2020. [Online]. Available: https://www.medellincomovamos.org/node/18687

[14] E. W. Group, "Guide to the demonstration of equivalence of ambient air monitoring methods," 2010.