SN Computer Science (2022) 3:216
https://doi.org/10.1007/s42979-022-01104-1

ORIGINAL RESEARCH

Completeness Assessment and Improvement in Mobile Crowd-Sensing Environments

Souheir Mehanna[1,2] · **Zoubida Kedad**[1] · **Mohamed Chachoua**[2]

Received: 20 July 2021 / Accepted: 19 March 2022 / Published online: 10 April 2022

Abstract
Mobile sensors are increasingly used to monitor air quality to accurately quantify human exposure to air pollution. These sensors are subject to various issues (misuse, malfunctions, battery problems, etc) that are likely to cause data quality problems. These quality problems may have a considerable impact on the reliability of analytical studies. In this work, we address the problem of data quality evaluation and improvement in mobile crowd-sensing environments. Our work is focused on the data completeness quality dimension. We introduce a multi-dimensional model to represent the data coming from the sensors in this context, and then present the different facets of data completeness inspired by the model. We propose quality indicators capturing different facets of completeness along with their corresponding quality metrics. We also propose an approach to improve data completeness by extending two existing data imputation techniques, SVDImpute and KNNImpute, with information about the sensor quality. Our experiments show that our quality-aware imputation approach improves the accuracy of the imputation achieved by the original techniques.

Keywords Data quality · Data completeness · Sensor data · Mobile sensors · Air quality · Data completeness improvement

Introduction

Air pollution is a global concern because of its major environmental risk and its negative effects on health. According to several World Health Organization[1] (WHO) reports, air pollution is a factor in the deterioration and worsening of people's health. It is responsible for an increasing number of deaths and a myriad of damages to ecological and economic systems, especially in dense urban cities. Air quality is often described by the WHO as an invisible killer which has been the main driver for many research projects in the recent years aiming at the quantification of human exposure to pollution [13, 26, 38, 40]. The goal is to better assess air pollution and its impact on health. This is the context of the Polluscope[2] research project [5]. The main objective of this project is to employ the emerging technologies of micro-sensors and the development of an innovative infrastructure for the acquisition and exploitation of data to assess air pollution on very fine scales. This multi-disciplinary project aims to characterize the effects of air pollutants on health, both in indoor and outdoor environments in the region of Île-de-France.

One of the main problems that arise in the Polluscope project is the reliability of the chain of acquisition and processing of spatio-temporal data. Mobile sensors and micro-detection units are well known to be less robust and more sensitive to various events including points-of-failure. By the time issues are fixed, the sensors may lose significant chunks of data. Data analysis based on poor quality data leads to ill-defined indicators and bad decisions. Hence, it is crucial to monitor data quality along the entire data workflow to provide accurate air quality indicators. This raises

This article is part of the topical collection "Web Information Systems and Technologies 2021" guest edited by Joaquim Filipe, Francisco Domínguez Mayo and Massimo Marchiori.

✉ Souheir Mehanna
souheir.mehanna@eivp-paris.fr; souheir.mehanna@uvsq.fr

Zoubida Kedad
zoubida.kedad@uvsq.fr

Mohamed Chachoua
chachoua@eivp-paris.fr

[1] DAVID Laboratory, University of Versailles Paris Saclay, Versailles, France

[2] LASTIG Laboratory, EIVP, University Gustave Eiffel, Paris, France

[1] For more information, see https://www.who.int/home.

[2] http://polluscope.uvsq.fr.

timestamp	value_from_sensor	latitude	longitude
2019-06-16 22:34:00+00	9	48.8561134338	2.37139344215
2019-06-16 22:35:00+00	9	48.8560180664	2.3714621067
2019-06-16 22:36:00+00	10	48.8560180685	2.37128782272
2019-06-16 22:37:00+00	12	48.8561134338	2.37139344215
2019-06-16 22:38:00+00	13	48.874317	2.302189
2019-06-16 22:39:00+00	13	48.87352	2.303082
2019-06-16 22:40:00+00	8	48.859813	2.299612
2019-06-16 22:41:00+00	10	48.859231	2.300347
2019-06-16 22:42:00+00	13	48.858699	2.300915
2019-06-16 22:43:00+00	11	48.8588929716686	2.30260951056391

Fig. 1 Snapshot of the data captured by sensors

the question of how credible the knowledge induced by the measurements generated by these micro-sensors is. Which leads to other questions such as: how to ensure the quality of the data from micro-sensors? How to manage the imperfections of these data? How to deal with missing data?

This work is a contribution towards data quality monitoring and improving in mobile crowd-sensing environments (MCS). We focus on completeness issues raised in this context. We first propose a multi-dimensional model representing pollution measurement data along with the relevant analysis dimensions. We then discuss the use of this model to capture the different understandings of the completeness of data coming from mobile sensors. We introduce and define three completeness facets: sensor completeness, spatial completeness, and temporal completeness. We also propose metrics for the evaluation of the aforementioned completeness facets. These contributions were first introduced in [22]. In this paper, we extend our completeness evaluation work. We propose an approach to improve data completeness by extending existing data imputation techniques with information about the quality of the measuring sensors. This approach aims to show that the quality of the measuring sensor is a crucial indicator that helps improve processing data coming from sensors. We first define sensor quality in the context of mobile crowdsensing and propose suitable evaluation metrics. Then, we propose a quality-based data imputation approach that enriches existing techniques with sensor quality. We extend SVDImpute and KNNImpute [36] by taking into account the quality of the sensors during data imputation. We compare the results of our extended techniques with their original approaches to show the usefulness of our approach.

The rest of this paper is organized as follows. "Motivating Example" presents a motivating example. "Mutli-Dimensional Data Model" introduces the proposed multi-dimensional model to represent data in MCS environments. "Sensor Completeness" introduces the sensor completeness indicator and proposes an evaluation metric. "Spatial Completeness" and "Temporal Completeness" present the spatial completeness indicator and the temporal completeness indicator respectively. "Improving Completeness: A Quality-Based Approach" describes data imputation techniques and the proposed quality-based approach for the improvement of data completeness. "Evaluations" presents our experiments and the results achieved by the proposed approaches. "Related Works" discusses the related works, and finally, "Conclusion" concludes the paper.

Motivating Example

In this paper, we focus on completeness issues in MCS environments. According to [2], data completeness has been defined as "the extent to which data are of sufficient breadth, depth and scope for the task at hand". The authors propose several metrics to evaluate data completeness in the context of relational databases. One of them is the presence of null values in a given table or column. Another metric is the comparison of the tuples present in the database to an existing set of reference tuples. In our view, such metrics are not suitable for evaluating completeness in MCS environments.

To illustrate our claim, consider the following example. The table in Fig. 1 shows a sample of the measurements from one sensor. It contains the timestamp at which the measurement was taken, the value of the pollutant, and the longitude and latitude indicating the location of the sensor at that time. If we consider that data completeness is evaluated as the proportion of Null values in the table, then we can see from Fig. 1 that there are no such values for any of the records in the table, and we can, therefore, say that our data are complete.

However, after we plot these data measurements on a map as shown in Fig. 2, we can see that these measurements cover only three cells in the studied area. We also notice that there are no measurements recorded in the remaining cells of the grid. Ideally, the measurements should have been distributed over all the cells in the considered area. Assume that we want to compute the average level of a given pollutant in this area. It is important to be aware that this characterizes only a small portion of this area, not the area as a whole.

Consider another example, and let us assume that the rate of measurement of the sensor is 1 measurement/s. This

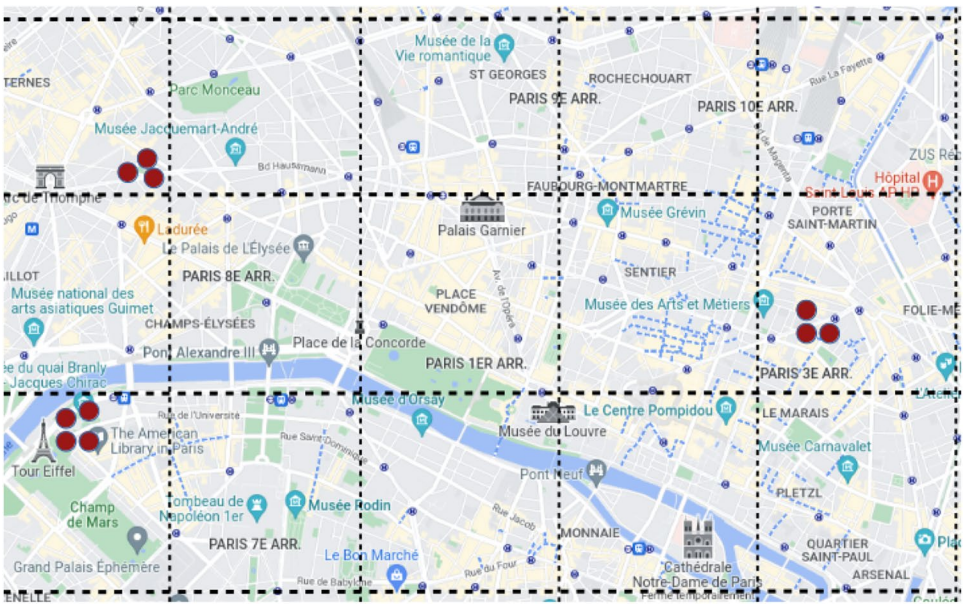

Fig. 2 Map showing the spread of the pollution measurements over the grids of a given area

means that we expect 10 min × 60 measurements/min = 600 measurements during this 10 min period of our study. Even though the data look complete in Fig. 1 with the absence of null values, there are 590 missing measurements in that table. Hence, the data in our table is in fact incomplete.

The examples presented above show that the existing completeness definitions and associated metrics are not appropriate to capture all the facets of completeness in MCS environments. In the following section, we will present a multi-dimensional model for storing pollution measurement data in the Polluscope project, and we will discuss the different facets of completeness in this context.

Mutli-dimensional Data Model

In this section, we introduce the multi-dimensional model which represents the pollution measurements in a MCS environment and the relevant analysis dimensions. We use the multi-dimensional views exposed by the model to illustrate the different facets of completeness.

In the Polluscope project, different pollution data acquisition campaigns are planned, each one having a start and an end date. Volunteering participants in the campaign are provided with a kit of sensors, which they will be expected to carry for around 7–10 days during the campaign. Each kit may consist of different sensors providing measurements of distinct pollutants such as particulate matter (PM10, PM2.5 and PM1.0), NO_2 or black carbon (BC). Each measurement is associated with a timestamp and a location. Figure 3 depicts our multi-dimensional model. A single sensor reading is represented in the fact table Measurement by the attribute measurementValue which represents the quantity of a pollutant in the air. There are six dimensions in the model. For a given measurement, the Sensor dimension represents the sensor that took the measurement, described by a sensor id, a type and a name. Location and Time dimensions give information about the spatial coordinates where the measurement was taken and the associated time. The Campaign dimension represents the campaign details during which the measurement was taken. The User dimension identifies the participant who was carrying the sensor that took this measurement; user-identity information are not saved for privacy reasons; the gender and the age are recorded for analysis purposes. The PollutantType dimension provides information about the name of the pollutant associated to the measurement value.

We leverage the different dimensions demonstrated in this model to explain the various understandings of completeness in this context. Completeness in mobile crowdsensing environments has different facets, and there are several understandings of how completeness can be perceived and represented. The multi-dimensional model in Fig. 3 helps us analyze the different facets and perspectives of completeness, we present five of them in the following:

- Completeness over a campaign, which expresses the overall completeness of a campaign. It represents the extent to which the measurements expected during this campaign from all the sensors in use and all the participants are actually stored.
- Completeness for one participant in a campaign, which expresses the completeness of the measurements from all sensors carried by this participant during their volunteering period in the campaign. Such completeness

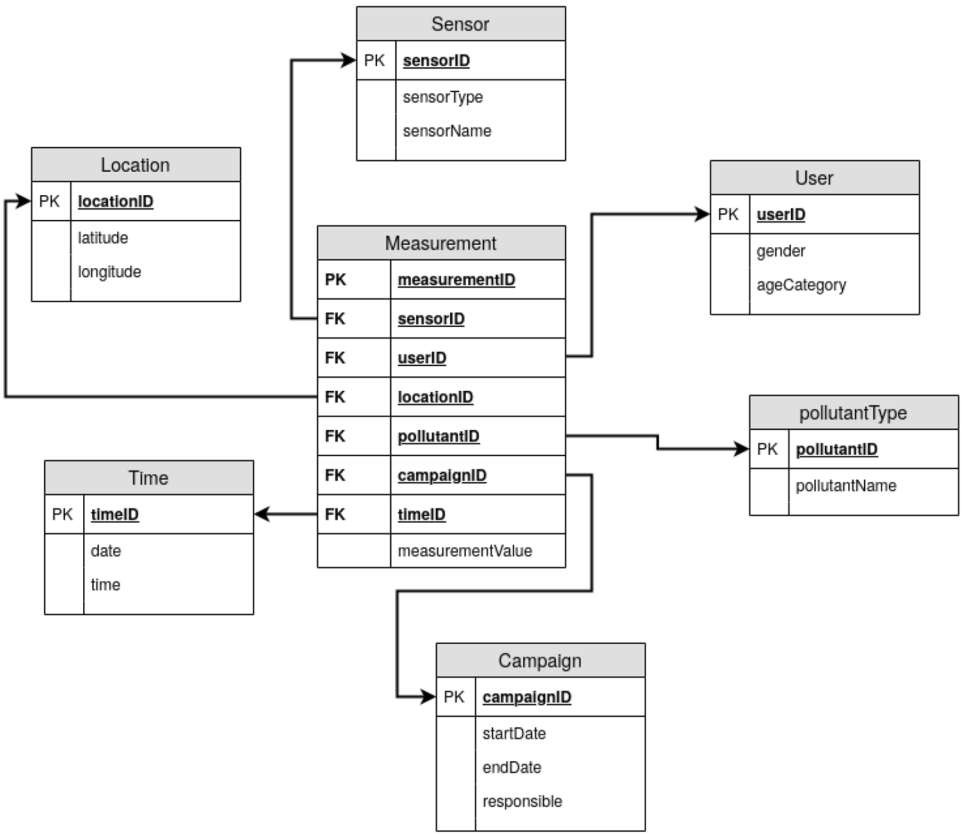

Fig. 3 A multi-dimensional model for pollution data

indicator allows for better exposure quantification to air pollutants for this participant.

- Completeness for a spatial area in one campaign is another facet which represents the spatial coverage of a designated area. It indicates the spatial dispersion of the measurements over this area. The goal is to understand the way measurements are distributed in the considered area of study, and whether the measurements are focused in a limited part of the designated area, or if they cover all of it.
- Temporal completeness characterizes the way a given period of time is covered by the collected measurements. These measurements may have been collected at regular intervals throughout the period, or taken in specific chunks of time, leaving other chunks without any measurement. Assessing such completeness assumes that the rate at which the sensors are supposed to provide their measurements is known.
- Sensor completeness which is an indicator that reflects the completeness of one specific sensor throughout the duration of the campaign. As one sensor could be used by different participants at different times during one campaign, the study of sensor completeness over a campaign shows the extent to which this sensor has provided the expected measurements regardless of the participant carrying this sensor.

In the following sections, we will present the definitions and metrics for three of the completeness facets presented above: sensor, spatial and temporal completeness.

Sensor Completeness

Sensor completeness is a quality facet that captures the extent to which the measurements of a given sensor are complete over a certain time period. It shows the completeness of the data captured and sent by this specific sensor during this time period. The nature of the sensors can be faulty and prone to many points of failures. Studying their completeness can show how reliable these sensors are by giving information about the completeness of the data captured and sent by each one.

A sensor unit can be used multiple times for different users. A sensor usage is performed by a single user. To study the completeness of one sensor over a certain period of time T, we have to study its completeness every time it was used in T. Hence, if a sensor has been used 4 times during a period, we have to study its completeness for each of these 4 times.

We evaluate the completeness of a specific sensor S_i as follows:

- Identifying all the usages of sensor S_i in the specified time period.
- Evaluating sensor completeness for each usage of S_i separately.
- Aggregating the computed evaluations of each usage to calculate the completeness of sensor S_i.

The completeness for a single sensor S_i in one campaign is evaluated as follows:

$$\text{Sen}C_{S_i} = \frac{\text{AM}_{Si}}{\text{RM}_{S_i}}, \tag{1}$$

where AM_{S_i} is the actual number of measurements sensor S_i has taken during all its usages in the specified period of time, and RM_{S_i} is the expected number of measurements from sensor S_i during its usages in this time period.

The required number of measurements RM_{S_i} for a sensor S_i throughout a specific time period is defined as:

$$\text{RM}_{S_i} = \Sigma_{j=1}^{K} n_{S_{ij}}, \tag{2}$$

where K is the number of usages of sensor S_i over the specified time period, and $n_{S_{ij}}$ is the number of required measurements for sensor S_i in usage j.

For every single usage denoted j including the sensor S_i, $n_{S_{ij}}$ is computed as follows:

$$n_{S_{ij}} = f_{S_i} * D_{C_j}, \tag{3}$$

where f_{S_i} is the sampling rate of the sensor S_i and D_{C_j} is the duration of the usage j of the sensor.

Spatial Completeness

Spatial completeness is the extent to which data sufficiently represents a specific spatial area, and it characterizes the coverage of this area considering the available measurements. In other words, spatial completeness indicates how sufficient and comprehensive the current measurements are for a particular area. This notion is similar to the concept of data skewness [3].

Completeness of measurements does not necessarily mean the more the better. It only means that we have enough measurements that ideally cover the area of study such that we can say it is spatially complete. Spatial completeness is quantifying that distance between the ideal and the actual situation of the measurements. The specification of an ideal spatial coverage, however, is to be determined. Several interpretations exist for an ideal required number of measurements for a specific area. For instance, one interpretation could be to divide the designated area into equal grid cells and aim for an equal number of measurements in each grid cell. This means that the measurements taken by the sensor are not clustered in few portions of the area but are rather evenly spread all over it. Hence, we propose a spatial completeness metric that compares the actual distribution of the measurements to a uniform distribution of the measurements over the area of study. For this particular interpretation, the evaluation steps of spatial completeness over the designated area are performed as follows:

- Dividing the area of study into equal-sized grid cells
- Computing the required number of measurements for each grid cell, and evaluating the spatial completeness of each grid cell
- Aggregating the computed evaluations of each grid cell to compute the spatial completeness of the area of study

Spatial Completeness of a Cell C_i

After dividing the designated area of study into equal sized grid cells, we compute the spatial completeness for each cell in the grid. Spatial completeness of a grid cell C_i, denoted SC_i, is computed as follows:

$$SC_i = \frac{\text{AM}_{C_i}}{\text{RM}_{C_i}}, \tag{4}$$

where AM_{C_i} is the actual number of measurements in a grid cell C_i, and RM_{C_i} is the required number of measurements in a grid cell C_i.

Different assumptions could be made to estimate RM_{C_i}, the required number of measurements in a given cell. Two of them are presented hereafter:

- **Hypothesis 1:** We consider as a reference, a uniform distribution of the measurements over the area of study A. This means that the number of measurements should be evenly distributed over the cells in the grid. Hence, the required number of measurements is:

$$\text{RM}_{C_i} = \frac{\text{AM}}{|A|}, \tag{5}$$

  where AM is the actual number of available measurements for the whole grid, and $|A|$ is the number of grid cells in the area A.
- **Hypothesis 2:** We consider as a reference, a distribution of the measurements that takes into account the variation of pollutant levels in the different cells of the area of study A. Pollutant variability could be learned from existing data obtained from previous campaigns. If for a given cell, the data show that there is a low variation of pollutant levels in all the spatial area represented by this cell, then the number of required measurements for this cell can be low without a loss of coverage. Conversely, if there is a high variability

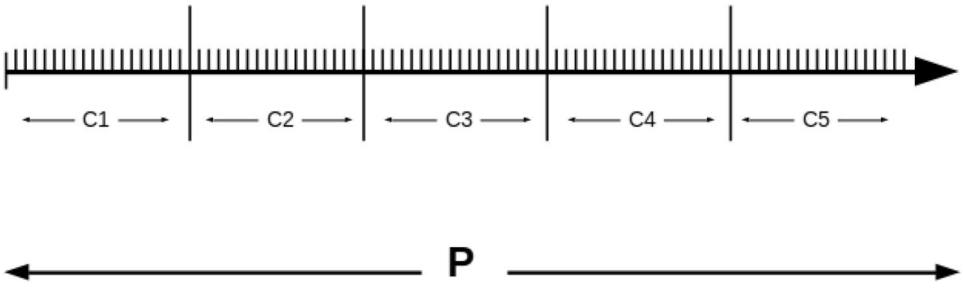

Fig. 4 Period P divided into chunks D_i

in a given cell, then the required number of measurements should be higher to better represent this cell.

The value of SC_i ranges from 0 to 1. A value of 1 meaning that the available measurements are uniformly distributed over the considered area. A low value represents the fact that the measurements are unevenly distributed over the area. Note that a high spatial completeness value does not represent the fact that a high number of measurements is available, but that the available measurements, regardless of the quantity, are more evenly distributed.

Spatial Completeness of an Area *A*

After computing spatial completeness for each cell in the grid separately, the overall spatial completeness for the whole area of study A is computed by aggregating the spatial completeness of all the cells. This could be done in different ways, for example using the average, the median, the minimum or the maximum functions.

We propose two quality metrics to compute the overall spatial completeness:

- **Cumulative average**, which computes the average of all cells' spatial completeness, as shown in the formula below:

$$SC(A) = \frac{\sum_{i=1}^{|A|} SC_{Ci}}{|A|}, \tag{6}$$

  where SC_{Ci} is the spatial completeness of one grid cell C_i, and $|A|$ is the number of cells in the grid covering area A.
- **Spatial completeness above a certain threshold**, which computes the proportion of cells having their spatial completeness above a given threshold t, as shown in the formula below:

$$SC_{(A)} = \frac{\sum_{i=1}^{|A|} \alpha_i}{|A|}, \quad \text{where} \begin{cases} \alpha_i = 1 & \text{if } SC_{Ci} \geq t \\ \alpha_i = 0 & \text{if } SC_{Ci} < t. \end{cases} \tag{7}$$

Temporal Completeness

Temporal completeness is another facet of data completeness which can be relevant in the context of MCS environments. It expresses the extent to which a considered period of time is covered by the available measurements. On one hand, sensors capturing measurements at a very high frequency may at some point add redundancy to the data, but on the other hand, a very low frequency will result in missing data. Therefore, we need a clear characterization of temporal completeness.

A high temporal completeness indicates a high coverage of the acquired measurements over a time period P. To assess temporal completeness, we divide a period P into n equal chunks and then compare the number of acquired measurements during time period P to a reference number of measurements defined for each chunk denoted by RM_{Di}. The reference number of measurements can be computed in several ways. In our work, we assume a uniform distribution of the acquired measurements in period P over the time chunks D_i where $i\epsilon\{i = 0, 1, 2..., n\}$. A high number of measurements does not necessarily mean high temporal completeness, we have to study their distribution over time as well.

The evaluation of temporal completeness for a specified period of study is done as follows:

- First, dividing the period of study P into equal-sized chunks of time as it is shown in Fig. 4
- Then computing the required number of measurements and evaluating the temporal completeness for each chunk.
- Aggregating the computed evaluations to calculate the overall temporal completeness of period P.

Temporal Completeness of a Specified Chunk of Time D_i

Different assumptions could be made to estimate the temporal completeness for a single time chunk D_i. Two of them are presented hereafter:

- **Hypothesis 1:** We consider as a reference, a uniform distribution of the measurements over time. The temporal completeness for a single time chunk D_i is:

$$TC_i = \frac{\mathrm{AM}_{Di}}{\mathrm{RM}_{Di}}, \tag{8}$$

  where AM_{Di} is the actual number of measurements in a chunk of time D_i and RM_{Di} is the required number of measurements in the chunk of time D_i. RM_{D_i} is defined for a chunk of time D_i as:

$$\mathrm{RM}_{D_i} = \Sigma_{j=1}^{K} n_{sj}, \tag{9}$$

  where K is the number of sensors, n_{sj} is the number of required measurements for sensor s_j during the time chunk D_i. For a sensor s_j, the number of required measurements during a chunk of time D_i is computed as:

$$n_{sj} = f_{sj} * |D_i|, \tag{10}$$

  where f_{sj} is the sampling rate of the sensor s_j expressed in number of measurements per minute, and $|D_i|$ is the size of the chunk D_i expressed in minutes.
- **Hypothesis 2:** We consider that the measurements are distributed considering the variation of pollutant levels at different times of the day, month or year. Pollutant measurements are highly affected by time (e.g., rush hours pollutant readings are higher than other times of the day). A possible approach would be to analyze the available data to detect variation patterns. The number of required measurements can then be set using these patterns in order to compute the temporal completeness.

Temporal Completeness of a Period P

The temporal completeness of a period of time P provides information about the way the available measurements are distributed over P, and how well P is covered by these measurements. It is computed by aggregating the temporal completeness values computed for all the time chunks in P.

Temporal completeness for a time period P can be computed as the average of all the temporal completeness values of its chunks, as shown below:

$$TC_P = \frac{\Sigma_{i=1}^{|P|} TC_i}{|P|}, \tag{11}$$

where $|P|$ is the number of chunks in a period of time P, and TC_i the temporal completeness of chunk D_i.

Improving Completeness: A Quality-Based Approach

In "Sensor Completeness", "Spatial Completeness", "Temporal Completeness" we have presented different quality factors related to the completeness dimension along with their corresponding evaluation metrics. In this section, we address the problem of improving data completeness in mobile crowdsensing environments. To this end, we will rely on existing data imputation techniques to generate missing values. Our proposal is to revisit existing techniques by introducing quality aspects during the data imputation process.

According to [16], there are three families of data imputation techniques for time series, two of which are very commonly used: pattern-based and matrix based techniques. Pattern-based approaches assume a high similarity in the patterns of the series. To generate missing data in series X_1, approaches in this family study the patterns of the reference time series at the missing block's timestamp t_j, and then look for candidate replacement patterns in other timestamps of the reference series. When a similarity in the reference series is found at some timestamp t_y, the missing block in X_1 at timestamp t_j is replaced with the pattern at timestamp t_y in X_1. Matrix-based approaches use dimensionality reduction methods to detect linear correlations across multiple series which will later be used to recover missing values and then compensate the lost accuracy with iterations over computations until a set error metric threshold is reached. Principal Component Analysis (PCA) [11] and Singular Value Decomposition (SVD) [32], among several others, are the most commonly used dimensionality reduction techniques. SoftImpute [21] and CDRec [14, 15] are matrix-based data imputation approaches based on SVD [32] and Centroid Decomposition (CD) [6], respectively. ST-MVL [39] and DynaMMo [19] are pattern-based techniques that learn hidden patterns and exploit series smoothness and neighboring sensors to generate missing values.

Existing data imputation techniques mainly rely on existing measurements either from the same sensor or other sensors or both. Given the erroneous nature of the low-cost mobile sensors, imputation techniques can sometimes perform poorly due to the usage of low-quality sensors in the imputation process. We propose an improvement to the existing techniques by providing them with sensor quality information. Our idea is to use the quality of the measuring sensor during the imputation process. A quality-aware data imputation technique prioritizes high-quality sensors. Consider the example of a sensor s_1 having a missing value at timestamp t_j and two other sensors s_2 and s_3 having a quality

of 0.2 and 0.85, respectively. Assume that the measurements provided by these two sensors at timestamp t_j are 4 and 8, respectively. Because s_3 has a higher quality than s_2, the measurement of sensor s_3 should be given a higher weight than the one of sensor s_2 when imputing the value of s_1 at timestamp t_j.

To improve the completeness of the data coming from the sensors, we propose an extension to the existing data imputation techniques to make them aware of the quality of the sensors providing the data used for the imputation. Quality of a sensor is a broad concept that has several facets. In our extension, we use the definitions of sensor completeness proposed in "Sensor Completeness" to assess the quality of a sensing unit. Provided with other quality factors characterizing sensors, we could integrate more quality information to further improve the imputation. In this section, we will first discuss sensor quality, then we will present the considered data imputation techniques and finally we will describe the proposed quality-based extensions of these techniques.

Sensor Quality

Sensor quality is a general indicator of the performance of a sensor unit expressing the extent to which we can trust the data coming from it. Some sensors perform better than others. This could be due to a variety of reasons, such as the characteristics of the device itself, like the measurement acquisition rate or the technologies at the heart of this device. In addition, the performance of sensor units of the same type and coming from the same manufacturer may vary depending on the meteorological context in which the measurements were taken. For example, the indoor air quality measuring sensor, NETATMO[3] works only for an external temperature between 0 and 50 °C with a humidity level between 0% and 100%. We can therefore deduce that the data provided by this sensor when the temperature is negative is of poor quality. The quality of the measurements can also be impacted by the way the sensor is used by its carrier. Indeed, a sensor whose battery is discharged, or even turned off by its user, will provide data of lower quality than a sensor operating continuously with a correct battery level. Hence, the data coming from each sensor unit can be disparate in terms of quality and all of the aforementioned factors can be used to determine the level of quality of a sensor. We can also evaluate the quality of a sensor using some measurements provided by reference devices if available.

There are different perspectives on how to define sensor quality and many quality factors could be added, normalized and aggregated to represent sensor quality. In this work, as stated earlier, we will use the definition of data completeness and the evaluation metrics proposed in "Sensor Completeness" to represent the quality of the sensors, and we will rely on these to extend the existing data imputation techniques.

Data Imputation Algorithms

There are many data imputation algorithms in the literature using different techniques to infer missing values. In our work, we have used two existing approaches, KNNimpute and SVDImpute proposed by [36]. To generate a missing value for a given sensor, KNNImpute uses data from k neighboring sensors. SVDImpute uses SVD to extract linear correlations across data series of several existing sensors. We will describe hereafter each of these two approaches.

KNNimpute

KNNimpute is based on k-nearest neighbors applied on the data matrix containing data measurements of k sensors at timestamp t. KNNImpute generates a missing value at timestamp t based on the measurements from neighboring sensors at timestamp t. The value of the measurement from each sensor is weighted according to its similarity with the target sensor s_t. The similarity metric is the Euclidean distance between a sensor s_i and the target sensor s_t with the missing measurement. Finally, to impute a missing value, a weighted average of the k-nearest sensors is computed using similarity as the weight. The missing value inferred using KNNImpute is defined by:

$$\hat{v}_j = \frac{\sum_{l=1}^{k} v_{lj} * d_{s_l,s_t}}{\sum_{l=1}^{k} d_{s_l,s_t}}, \tag{12}$$

where $\hat{v}_j$ is the imputed value of target sensor s_t at timestamp t_j, k is the number of neighboring sensors, v_{lj} is the measurement value of sensor s_l at timestamp t_j, s_t is the target sensor, and d_{s_l,s_t} is the Euclidean distance between sensor s_l and target sensor s_t.

SVDImpute

SVDImpute is a matrix-based technique that relies on the Singular Value Decomposition (SVD) matrix factorization technique [32] to extract the linear combinations in the data which will be later used in the regression to finally recover missing values as shown in Fig. 5. SVD is employed to obtain the principle components of the matrix containing the sensors and their corresponding values at every timestamp within a period of time P.

Let us consider a set of sensors $S = \{s_1, s_2, ..., s_n\}$. We denote by A, the matrix containing all the measurements

[3] https://www.netatmo.com/fr-fr/aircare/homecoach/specifications.

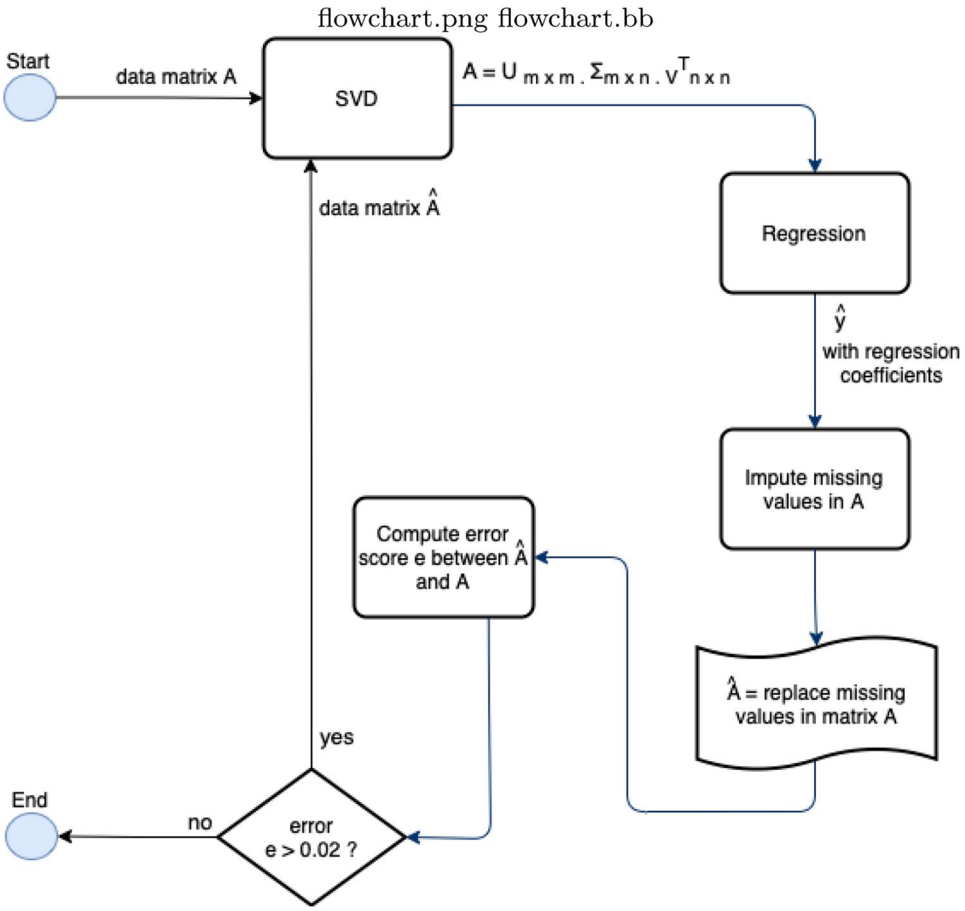

Fig. 5 Iterative SVDImpute Algorithm

from the set of sensors S. SVD factorizes the matrix A, into 3 singular matrices as follows:

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^{\mathrm{T}}_{n\times n}, \tag{13}$$

where U is an $m \times m$ unitary matrix, Σ is an $m \times n$ diagonal matrix, V^{T} is the transpose of an $n \times n$ unitary matrix that contains eigenvectors, that are quantified by their corresponding eigenvalues on the diagonal of matrix Σ. After the principle components are computed, the k most significant eigenvectors are selected from V^{T}. Then, we estimate a missing value v_{ij} in sensor s_i by first regressing this sensor against the k eigenvectors to get the coefficients $(\beta_0, \beta_1, ..., \beta_n)$ in the regression equation below.

$$\hat{v}_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon, \tag{14}$$

where $\hat{v}_{ij}$ is the imputed value of sensor s_i at timestamp t_j, β_0 is the intercept, $(\beta_1, ..., \beta_n)$ are the regression coefficients, $x_1, ..., x_n$ are the corresponding sensor values, and ϵ is the residuals.

The coefficients of the regression are used to reconstruct a missing value v_{ij} from a linear combination of the k eigenvectors. The imputation of all missing values in the data matrix A using the aforementioned technique constructs the new matrix $\hat{A}$. Finally, the difference between A and $\hat{A}$ is computed. If the difference is greater than a predefined threshold δ, we iterate and repeat the process considering $\hat{A}$ as the matrix to factorize in Eq. (13). Iterative SVDImpute is described in Fig. 5.

Quality-Based Data Imputation

In this section, we present our approach for extending existing data imputation techniques with insights about the quality of the sensors acquiring the measurements. In our approach each sensor is associated to a quality score that ranges between 0 and 1. The considered quality dimension is sensor completeness, and the completeness score is assessed using the metrics introduced in "Sensor Completeness".

The existing data imputation techniques do not make any assumption on the quality of their input data. In MCS environment, there may be important variations of the quality of the data provided by a sensor. To improve the quality of the imputation, we propose to enrich the data imputation techniques with the quality of the measuring sensor. A quality-aware data imputation technique would not consider all the sensors in the same way, but instead give higher weights

to measurements coming from sensors with higher quality scores.

We propose three different ways to extend the existing techniques with sensor quality:

1. Taking into account sensors with a quality score above a predefined threshold γ, where $0 \geq \gamma \leq 1$
2. Taking into account the data of a percentage p of the sensors having the highest quality score.
3. Considering all the sensors, but weighting the measurements by the quality score of the measuring sensor.

Our goal in this section is to extend an approach A with sensor quality denoted by q_s, and propose an extended approach A′ such that the imputed value using A′ is closer to the actual value than the imputed value using A.

Extending KNNImpute Using Sensor Completeness

KNNImpute chooses the k-nearest sensors to be used for the imputation of the missing value.

To extend KNNImpute, we use the third type of extension presented earlier, which consists in weighting the measurements with the quality score of the measuring sensors. This way, sensor quality will be considered as a weight that gives more importance to the measurements coming from good-quality sensors over those from poor-quality ones.

Consider a set of sensors $S = \{s_1,..., s_n\}$ where each sensor s_i has a quality score q_{s_i}. Assume sensor s_t has a missing measurement v_j at timestamp t_j. The imputed value $\hat{v}_j$ generated by the extended version of KNNImpute is defined by:

$$\hat{v}_j = \frac{\sum_{l=1}^{k} v_{lj} * d_{s_l,s_t} * q_{s_l}}{\sum_{l=1}^{k} d_{s_l,s_t} * q_{s_l}}, \tag{15}$$

where k is the number of neighboring sensors, d_{s_l,s_t} is the Euclidean distance between sensor s_l and target sensor s_t, and v_{lj} is the measurement value of sensor s_l at timestamp t_j.

Extending SVDImpute Using Sensor Completeness

We propose two possible extensions to SVDImpute to take into account sensor quality. One considers only a percentage p of sensors having the highest quality score in the data matrix. The second takes sensors having a quality score above a predefined threshold γ.

The first proposed extension takes place at the very first phase of the algorithm, after the data from the sensors required for the imputation is retrieved. In this data set, we will consider only the data of the percentage p of the sensors having the highest quality score. For example, if we set $p = 70\%$ and if we have data from 10 sensors, SVDImpute will consider only the data from the 7 sensors with the higher quality scores. The data from the 3 remaining sensors with lower quality scores will be discarded.

The second proposed extension is similar to the first one, but instead of considering the percentage p of sensors having the highest quality score, it considers sensors having a quality score above a threshold γ, where γ ranges between 0 and 1. In this proposition, we are only interested in the data from the sensors that have a quality score above a predefined threshold. This threshold is determined empirically. If we consider the previous example, and assume our threshold is 70% and only 4 sensors out of 10 have a quality score above 70%. In this case, we only consider the data from these 4 sensors into our data matrix and discard the remaining ones.

For both extensions, we filter out the data of the unwanted sensors by multiplying the data matrix $A_{m,n}$ by a filter $K_{m,m}$ that nullifies the rows of the unwanted sensors. Hence, the resulting matrix $A'_{m,n}$ only contains the rows of data from the selected sensors. For instance, if matrix A is of size 4×5, representing data from 4 sensors over 5 timestamps, and suppose only sensors s_2 and s_3 have a quality score above the considered threshold. Then, the matrix $A'_{m,n}$ will be computed as follows:

$$A'_{m,n} = K_{m,m} \times A_{m,n}, \tag{16}$$

$$K\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \times \begin{matrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}\begin{pmatrix} t_1 & t_2 & t_3 & t_4 & t_5 \\ a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Evaluations

In this section, we present some experiments on the approaches and metrics discussed in this paper. Our experiments are done on the real data collected in the context of the Polluscope[4] research project [5] over two campaigns. In this section, we present preliminary evaluations of the spatial, temporal and sensor completeness of the collected data using the metrics defined in this paper.

For completeness improvement, we illustrate our experiments with the proposed extensions of existing data

[4] http://polluscope.uvsq.fr.

imputation techniques from the literature. We show the improvement of the imputations by the extensions proposed to each of KNNimpute and SVDimpute.

Context of the Experiments

During the first phase of the Polluscope project, studies and experiments on pollutants and sensors were performed. The measured pollutants are PM1.0, PM2.5, PM10 (particulate matter of diameters 1.0, 2.5 and 10 respectively), NO_2 and BC (black carbon). Multiple sensors were selected to measure different pollutants.

For data acquisition, volunteers carry kits containing sensor units with them during their daily life routines (both indoor and outdoor) without any preset routes or destinations. A kit may contain either a single sensor, or multiple sensors, each measuring a different pollutant. Each acquired measurement is associated with its timestamp and its spatial coordinates.

In our completeness improvement experiments, we evaluate the inference of missing values on a subset of PM2.5 measurements selected to ensure that they have neighboring sensors within a 30 m diameter at the same time. This is because some of the studied data imputation algorithms rely on neighboring sensors for the imputation of a missing value. To conduct our experiments, we assume that each measurement in the selected set is missing, and we generate the imputed value both with the original imputation approaches and with the extended ones. Finally, we evaluate our extensions using the metrics defined in "Completeness Improvement Indicators". We extend KNNimpute by weighting every data measurement by the quality score of the measuring sensor. We have tested the two proposed extensions for SVDImpute, which consist, respectively, in considering a percentage p of the sensors having the highest quality score, and considering only data from sensors having a quality score above a predefined threshold.

Setup

We conducted our experiments on a Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz machine with 16GB System Memory and clock 100MHz. The data are stored on Postgres in a docker container on the cloud. We have used Python on Jupyter Notebook to establish a connection with the server containing the data and to be able to access the data for our evaluations. The sensor sampling rate is 1 measurement per minute.

For the completeness improvement experiments, the initial data set contains 7700 measurements. We will generate the missing values of a subset of this data set that contains 500 measurements. As discussed earlier, this subset is selected so as to ensure that for each measurement, there are other measurements taken by sensors within a 30 m distance. During our experiments, we evaluate our extensions by first assuming that every measurement in the subset is missing. We use the sensor completeness metric to compute the sensor quality score. We have set k to 5 for KNNImpute.

Completeness Improvement Indicators

To assess the proposed extensions, we compute the proportion of improved, unchanged and worsened measurements. These indicators show the accuracy of our extensions compared to that of the original versions of the imputation techniques.

We also compute the RMSE metric to quantify the error between the generated series and the actual ones. RMSE is computed between the data generated by the original approaches and the actual values. It is also computed between the values generated by the extended approaches and the actual values. Finally, we compare the two obtained results to infer whether our extensions minimized the error between the generated values and the actual ones or not.

RMSE Error Metric

RMSE measures the quality of an estimator. To get the error of the estimation models, we compute the root mean squared error RMSE error metric between a vector of estimated/predicted values $\hat{vij}$ and a vector of the actual/observed values v_{ij}. The RMSE could be computed for the estimation of the original approaches as well as the estimation from the extended approaches proposed in this paper. We can compare the error metric RMSE between the estimations from the original approaches (SVDImpute and KNNImpute) to the values generated using the extended versions of the aforementioned approaches.

Proportion of Improved Measurements

The proportion of improved measurements shows the number of measurements improved using the extension of the techniques with sensor quality, compared to the original approach. This means that the data imputation techniques generate more accurate estimations of the missing measurements with the extended versions than the original approaches. Therefore, the imputation of a missing measurement is said to be improved, if the absolute value of the difference between the actual measurement and the imputed value of the extended approach is smaller than the absolute value of the difference between the actual measurement and the imputed value of the original approach as shown in the following Eq. (17):

$$|v_{ij} - \hat{v}'_{ij}| < |v_{ij} - \hat{v}_{ij}|, \quad (17)$$

where v_{ij} is the actual measurement, $\hat{v}_{ij}$ is the imputed measurement generated by the original approach, and $\hat{v}'_{ij}$ is the imputed measurement generated by the extended approach.

Proportion of Worsened Measurements

Similarly, the proportion of worsened measurements shows the number of measurements worsened using the extension of the techniques with sensor quality, compared to the original approach. This means that the original approach provided a more accurate estimation of the missing measurements than the extended approach. The imputation of a missing measurement has been worsened using our extension if the absolute value of the difference between the actual measurement and the imputed value of the extended approach is greater than the absolute value of the difference between the actual measurement and the imputed value of the original approach as shown in the following Eq. (18).

$$|v_{ij} - \hat{v}'_{ij}| > |v_{ij} - \hat{v}_{ij}|, \quad (18)$$

where v_{ij} is the actual measurement, $\hat{v}_{ij}$ is the imputed measurement generated by the original approach, and $\hat{v}'_{ij}$ is the imputed measurement generated by the extended approach.

Proportion of Unchanged Measurements

Likewise, the proportion of unchanged measurements shows the percentage of measurements for which both the original and the extended versions of the data imputation technique generated the same value. This means that there was no improvement using sensor quality, and that both the original and the extended versions of the technique performed similarly. In such case, the absolute value of the difference between the actual measurement and the imputed value of the extended approach is equal to the absolute value of the difference between the actual measurement and the imputed value of the original approach as shown in the following Eq. (19).

$$|v_{ij} - \hat{v}'_{ij}| = |v_{ij} - \hat{v}_{ij}|, \quad (19)$$

where v_{ij} is the actual measurement, $\hat{v}_{ij}$ is the imputed measurement generated by the original approach, and $\hat{v}'_{ij}$ is the imputed measurement generated by the extended approach.

Results

Our completeness experiments are performed on the NO_2 pollutant measurements obtained in both campaigns 1 and 2. We have considered 21,398 measurements from campaign 1, and 38,834 measurements in campaign 2 for our sensor completeness experiments. The sensor completeness value was 58.66 and 59.92% for campaigns 1 and 2 respectively. Table 1 shows the sensor completeness of the selected sensor in all its usages during campaign 2.

Table 1 Completeness of a sensor measuring NO_2 for all its usages during campaign 2

Kit Nb	Sen-Comp
1	37.65%
2	77.3%
3	70.20%
4	28.55%
5	87.89%

Table 2 Computed spatial completeness of all pollutants during campaigns 1 and 2

Pollutants	SC Campaign 1	SC Campaign 2
PM1.0	15.10%	33.02%
PM2.5	15.10%	33.02%
PM10	15.10%	33.02%
NO_2	18.17%	35.15%
BC	20.38%	34.99%
Temperature	15.10%	32.91%
Humidity	15.10%	32.91%
Pressure	15.10%	33.024%

Table 3 Aggregated total average of temporal completeness of all pollutants during each sensing campaign

Pollutants	TC Campaign 1	TC Campaign 2
PM2.5	7.75%	42.23%
NO_2	60.91%	63.66%
BC	68.53%	59.49%

Over the two campaigns 1 and 2, we evaluated spatial completeness for each of the following pollutants: PM1.0, PM2.5, PM10, NO_2, BC. The evaluations are done over a manually selected area in Paris. The spatial completeness experiments were done on a total number of measurements 1,627,487 in campaign 1, and 4,229,053 measurements in campaign 2 (Table 2).

Campaign 1 has less collected data than campaign 2. We first compute spatial completeness as described in "Spatial Completeness" for every pollutant and for each of the kits in this campaign, and then we compute an average of all the kits to get the total spatial completeness. The fact that campaign 1 has less collected data than campaign 2 should be taken into consideration when analyzing spatial completeness because it means that with more data in campaign 2, there is the possibility of a wider spatial

Table 4 Evaluating quality-based data imputation approaches

Extension	Improved	Worsened	Unchanged	RMSE
KNNimpute: measurements weighted by quality	69%	11%	20%	5.4
SVDimpute top 40%	71%	21%	8%	2.7
SVDimpute quality above threshold: 0.45	62%	24%	14%	5.2

coverage. Table 3 shows the spatial completeness values computed for campaigns 1 and 2.

Over the two campaigns 1 and 2, we have also evaluated temporal completeness for each of the pollutants PM2.5, NO_2 and BC. To evaluate temporal completeness, we have extracted 582,506 measurements of the 3 selected pollutants in campaign 1 and 1,378,497 measurements in campaign 2. Table 3 shows the temporal completeness of the three pollutants PM2.5, NO_2 and BC over campaigns 1 and 2.

Finally, we evaluate the extended imputation approaches using RMSE error metric and proportion of data measurements improved, worsened, and unchanged. In Table 4, we present the values of the these metrics for the extensions of both SVDimpute and KNNimpute.

The analysis of these results leads to the following observations:

1. The three extensions show promising results. The extension of SVDImpute that considers data from the 40% of sensors having the highest quality score shows the highest percentage of improvement (71%).
2. KNNimpute shows very good results as well with 69% of the measurements imputation improved and only 11% of them worsened. 20% of the measurements remain unchanged. The RMSE is also relatively low with a value of 5.4.
3. Both extensions of SVDimpute show good results. However, considering the 40% of sensors having the highest quality score shows better results than considering only sensors with a quality score above the 0.45 threshold. This latter has to be determined empirically by testing different percentages and thresholds.
4. Taking the 40% of sensors having the highest quality score shows 71% of improved measurements in the imputation of the missing values, 21% of worsened measurements, and 8% unchanged measurements. The RMSE is also low with a value of 2.7.
5. The extension of SVDimpute that considers only sensors with a quality score above the 0.45 threshold, shows that 62% of the measurements were improved. But the percentage of worsened measurement is 24%, which is not very low. Besides, 14% of the measurements remained unchanged. The RMSE value of 5.2 is acceptable.

Discussion

From the experiments, we have seen that *Sensor Completeness* may significantly vary from one usage of the selected NO_2 sensor to another in a campaign. During the usages of the sensor in the two kits 1 and 4, sensor completeness was relatively low; whereas for the other kits, the sensor completeness value scored more than 70%. One possible reason could be that sensors used to measure NO_2 may sometimes lose their data if they run out of battery. However, overall, the sensor completeness results were relatively high for the selected NO_2 sensor.

As for the evaluations of *Spatial Completeness*, the results of campaign 2 are generally better than those of campaign 1. However, the spatial completeness achieved in both campaigns is not high and this could mean that the participants did not change their locations a lot during their participation periods. This can make sense if we think of the amount of time people spend in their homes and workplaces. The spatial completeness results are almost in the same range for both campaigns as the sensors measuring the studied pollutants were grouped in kits and carried together. The spatial areas they cover are, therefore, the same. Besides, the rates of sensor measurements in the setup of the experiments were the same for all the sensors.

The evaluation of *Temporal Completeness* scores for sensors measuring PM2.5 and NO_2 were better in campaign 2 than in campaign 1. However, the temporal completeness in sensors measuring BC was slightly better in campaign 1 than in 2. One possible reason for the very low temporal completeness for the sensor measuring PM2.5 in campaign 1 could be that during campaign 1, the sensors were unstable which caused the loss of many chunks of data. Therefore, the values of campaign 2 are more reliable for that sensor.

The preliminary results of the extended data imputation experiments with *Sensor Quality* are promising. The indicators showed approximately 70% of improvement of the performance of the existing data imputation techniques when extended with sensor quality represented in our experiments by *Sensor Completeness*. The extension of SVDImpute that considers the 40% of sensors having the highest quality score shows the best improvement result. The SVDImpute extension with quality above the 0.45 threshold shows the highest percentage of worsened measurements. The KNNImpute extension shows the highest percentage of unchanged measurements.

Related Works

Many research works have addressed the issues related to data quality. Some of them have studied quality dimensions and their evaluation metrics [2, 20, 24, 31]. Integrity assessment of maritime messages has been evaluated in [28] through both message-based and signal-based analysis. To support decision making on whether or not allocate a sensing task, [37] assessed the data quality of the inferred unsensed cells in a crowdsensing environment using re-sampling methods such as *leave-one-out* and *Bootstrap*.

A data quality assessment framework has been proposed in [1]. Dasu et al. [7] proposed two types of data quality checks, the first one monitors the data gathering process and checks the incoming data while the second monitors the quality of the content of the data streams and studies data quality according to four defined types of constraints on the data. The work presented in [27] proposes a supervised classification approach to assess the quality of sensor data. Using graph convolutional networks, the approach described in [30] defines local variation and a data quality level.

Although there are several works data quality evaluation, these proposals do not take into account the specific characteristics of the data in MCS environments. In our work, we specifically assessed one quality dimension, data completeness, with its different understandings, as one of the main issues introduced by mobile sensors is the loss of data. Some works have also addressed quality evaluation at the sensor level such as [9] who proposed a toolkit for the evaluation of micro-sensing units explaining all the factors and their metrics. Languille et al. [18] used the SET tool proposed by [9] to evaluate the performance of air quality sensors, and to justify the selection of some sensors rather than others.

Another set of works are more focused on representing and characterizing data quality in data storage systems and extending traditional existing tools to allow the association of quality indicators to data. Han et al. [10] identified two different types of sensor applications and their respective requirements, and proposed strategies for both the satisfaction and the optimization of either a single requirement or multi-dimensional quality requirements. Mustapha et al. [23] proposed a multivariate spatial time series representation model and used functional data representation for storing, aggregating, transforming and retrieving sensor data. Klein et al. [17] presented a metadata model extension for a relational database schema to store quality information along with data values, and have also extended conventional data stream systems to propagate data quality indicators.

Given the polysemous nature of the concept of data quality, some authors try to define the meaning of this concept according to the specific field and context. Han et al. [10] characterizes two types of data requirements under which they categorized each quality dimension. Rodríguez and Servigne [29] defines the quality dimensions for environmental monitoring systems and [25] defines the quality dimensions for spatial data. In addition, [8] defines and illustrates the data dimensions that are useful for the context of mobile sensing. While these works aim at discussing the application of all data dimensions to mobile sensing environments, we focus in our work on one of these dimensions, namely data completeness, we provide a characterization of the different facets of this dimension and we propose some suitable evaluation metrics. *Accuracy* and *completeness* are the most commonly described and evaluated dimensions for mobile sensing in the existing works. One of these works has specifically addressed completeness assessment [4], and the authors have developed a quality model to assess data completeness for sensor data by translating data rates to completeness values measured over a period of time. They have considered a specific *"smart home"* application context to demonstrate how completeness can be calculated. Similarly to this work, we also use the sampling/data rate to evaluate completeness, but we also introduce and discuss the different facets of completeness for the context of mobile crowd-sensing and assess completeness spatially, temporally and for a specific sensor.

There are several works on the evaluation of low-cost mobile sensors performance on different aspects and levels. In [12], the authors have reviewed the existing works related to these type of sensors and they have introduced a comparison of the existing studies and an evaluation of the agreement between low-cost sensors and reference machines. Another systematic review presented in [35] studies the quantification and detection of the missing data and outliers as these two are the most common issues in sensor data, and then illustrates the most common techniques existing for resolving them. To the best of our knowledge, there are no works that propose to extend existing imputation techniques with sensor quality information. However, there are several works that try to enrich data imputation techniques, such as in [34], which proposes the limited rule-based imputation techniques by proposing an extensive similarity neighbors extension that can generate missing data that are not revealed by the limited imputation techniques based on exact equality neighbors. In [33], the same authors extend similarity rule-based techniques to handle multiple incomplete attributes instead of one attribute and improve imputation accuracy by considering similarity neighbors under the constraints of similarity rules. However, this work requires an important amount of knowledge on the data to be able to set the similarity rules.

Conclusion

This paper is a first attempt towards characterizing and monitoring data quality in mobile crowd-sensing environments. We have first introduced a multi-dimensional data model to represent sensor data in this context. Then we have focused on data completeness and presented its different facets. We have provided the definitions and the evaluation metrics for three of these facets: sensor completeness, spatial completeness and temporal completeness. In order to improve data completeness, we have also proposed to extend two existing data imputation techniques with sensor quality, SVDImpute and KNNImpute. We have performed some evaluations on the proposed completeness evaluation metrics as well as the two extended imputation approaches and presented some experiments on real mobile sensor data provided by a mobile crowdsensing environment dedicated to air pollution measurement. The results on completeness evaluations show the benefits of studying this quality dimension from different and complementary perspectives. The results obtained using the extended imputation approach have illustrated the usefulness of considering sensor quality in the imputation process.

Beyond data completeness evaluation and improvement, our future works will address other quality dimensions such as data accuracy. We will also study some important quality-related problems in the context of MCS environments, such as anomaly detection.

Funding This work was funded by the Polluscope project (Grant ANR-15-CE22-0018 of the French National Research Agency) and the Qualiscope Impulsion project (I-SITE FUTURE, Gustave Eiffel University).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aquino GRCD, de Farias CM, Pirmez L. Hygieia: data quality assessment for smart sensor network. In: Hung C, Papadopoulos GA. (eds.) Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019, Limassol, Cyprus, 2019. ACM; 2019. p. 889–91. https://doi.org/10.1145/3297280.3297564.
2. Batini C, Scannapieco M. Data and information quality—dimensions, principles and techniques. Data-centric systems and applications. Berlin: Springer; 2016. https://doi.org/10.1007/978-3-319-24106-7.
3. Belussi A, Migliorini S, Eldawy A. Detecting skewness of big spatial data in SpatialHadoop. GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems 2018. p. 432–5. https://doi.org/10.1145/3274895.3274923.
4. Biswas J, Naumann F, Qiu Q. Assessing the completeness of sensor data. In: M. Lee, K. Tan, V. Wuwongse (eds.) Database Systems for Advanced Applications, 11th International Conference, DASFAA 2006, Singapore, April 12–15, 2006, Proceedings, Lecture Notes in Computer Science, vol. 3882. Springer; 2006. p. 717–32. https://doi.org/10.1007/11733836_50.
5. Brahem M, el Hafyani H, Mehanna S, Zeitouni K, Yeh L, Taher Y, Kedad Z, Ktaish A, Chachoua M, Ray C. Data perspective on environmental mobile crowd sensing. 2021. p. 269–88. https://doi.org/10.1016/B978-0-12-819671-7.00012-9.
6. Chu MT, Funderlic R. The centroid decomposition: Relationships between discrete variational decompositions and SVDS. SIAM J Matrix Anal Appl. 2002;23(4):1025–44. https://doi.org/10.1137/S0895479800382555.
7. Dasu T, Duan R, Srivastava D. Data Quality for Temporal Streams. Tech. rep. 2016.
8. Ferreira E, Ferreira D. Towards altruistic data quality assessment for mobile sensing. In: Lee SC, Takayama L, Truong KN. (eds.) Adjunct Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, UbiComp/ISWC 2017, Maui, HI, USA, 11–15, 2017. ACM; 2017. p. 464–9. https://doi.org/10.1145/3123024.3124439.
9. Fishbain B, Lerner U, Castell N, Cole-Hunter T, Popoola O, Broday D, Iñiguez T, Nieuwenhuijsen M, Jovašević-Stojanović M, Topalovic D, Jones R, Galea K, Etzion Y, Kizel F, Golumbic Y, Baram Tsabari A, Yacobi T, Drahler D, Robinson J, Bartonova A. An evaluation tool kit of air quality micro-sensing units. Sci Total Environ. 2017. https://doi.org/10.17863/CAM.9573.
10. Han Q, Hakkarinen D, Boonma P, Suzuki J. Quality-aware sensor data collection. Int J Sensor Netw. 2010;7(3):127. https://doi.org/10.1504/IJSNET.2010.033115.
11. Jolliffe IT. Principal component analysis. In: Lovric M, editor. International encyclopedia of statistical science. Springer: Berlin; 2011. p. 1094–6. https://doi.org/10.1007/978-3-642-04898-2_455.
12. Karagulian F, Barbiere M, Kotsev A, Spinelle L, Gerboles M, Lagler F, Redon N, Crunaire S, Borowiak A. Review of the performance of low-cost sensors for air quality monitoring. Atmosphere. 2019. https://doi.org/10.3390/atmos10090506.
13. Kaur A, Singla S, Bansal D. Quantifying personal exposure to spatio-temporally distributed air pollutants using mobile sensors. In: M.R. Eskicioglu (ed.) Proceedings of the First ACM Workshop on Mobile Crowdsensing Systems and Applications, CrowdSenSys@SenSys 2017, Delft, The Netherlands, November 5; 2017. pp. 1–6. ACM. https://doi.org/10.1145/3139243.3139248.
14. Khayati M, Böhlen MH, Gamper J. Memory-efficient centroid decomposition for long time series. In: Cruz IF, Ferrari E, Tao Y, Bertino E, Trajcevski G. (eds.) IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, 2014. IEEE Computer Society; 2014. p. 100–11. https://doi.org/10.1109/ICDE.2014.6816643.
15. Khayati M, Cudré-Mauroux P, Böhlen MH. Scalable recovery of missing blocks in time series with high and low cross-correlations. Knowl Inf Syst. 2020;62(6):2257–80. https://doi.org/10.1007/s10115-019-01421-7.
16. Khayati M, Lerner A, Tymchenko Z, Cudré-Mauroux P. Mind the gap: an experimental evaluation of imputation of missing values techniques in time series. Proc VLDB Endow. 2020;13(5), 768–782. http://www.vldb.org/pvldb/vol13/p768-khayati.pdf.
17. Klein A, Do HH, Hackenbroich G, Karnstedt M, Lehner W. Representing data quality for streaming and static data. In: Proceedings—International Conference on Data Engineering (2014), 2007. p. 3–10. https://doi.org/10.1109/ICDEW.2007.4400967.

18. Languille B, Gros V, Bonnaire N, Pommier C, Honoré C, Debert C, Gauvin L, Srairi S, Annesi-Maesano I, Chaix B, Zeitouni K. A methodology for the characterization of portable sensors for air quality measure with the goal of deployment in citizen science. Sci Total Environ. 2020;708:134698. https://doi.org/10.1016/j.scitotenv.2019.134698.
19. Li L, McCann J, Pollard NS, Faloutsos C. Dynammo: mining and summarization of coevolving sequences with missing values. In: IV JFE, Fogelman-Soulié F, Flach PA, Zaki MJ. (eds.) Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009. ACM; 2009. p. 507–16. https://doi.org/10.1145/1557019.1557078.
20. Liu C, Nitschke P, Williams S, Zowghi D. Data quality and the internet of things. Computing. 2019. https://doi.org/10.1007/s00607-019-00746-z.
21. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res. 2010;11:2287–2322. http://portal.acm.org/citation.cfm?id=1859931.
22. Mehanna S, Kedad Z, Chachoua M. Completeness issues in mobile crowd-sensing environments. In: Marchiori M, Mayo FJD, Filipe J. (eds.) Proceedings of the 16th International Conference on Web Information Systems and Technologies, WEBIST 2020, Budapest, Hungary, 2020. SCITEPRESS. p. 129–38. https://doi.org/10.5220/0010136201290138.
23. Mustapha A, Zeitouni K, Taher Y. Towards rich sensor data representation—functional data analysis framework for opportunistic mobile monitoring. In: Grueau C, Laurini R, Ragia L. (eds.) Proceedings of the 4th International Conference on Geographical Information Systems Theory, Applications and Management, GISTAM 2018, Funchal, Madeira, Portugal, 2018. SciTePress. p. 290–5. https://doi.org/10.5220/0006788502900295.
24. Nemani RR, Konda R. A framework for data quality in data warehousing. In: Yang J, Ginige A, Mayr HC, Kutsche R. (eds.) Information systems: modeling, development, and integration, third international united information systems conference, UNISCON 2009, Sydney, Australia, 2009. Proceedings, Lecture Notes in Business Information Processing, vol. 20. Springer; 2009. p. 292–7. https://doi.org/10.1007/978-3-642-01112-2_30.
25. Östman A. The specification and evaluation of spatial data quality. In: Proceedings of the 18st International Cartographic Conference, 1997. p. 836–47.
26. Qin X, Platisa L, Huu T, Tsiligianni E, Hofman J, Manna V, Deligiannis N, Philips W. Context-based analysis of urban air quality using an opportunistic mobile sensor network. Lecture notes of the institute for computer sciences, social-informatics and telecommunications engineering, LNICST 323 LNICST, pp. 285–300 (2020). https://doi.org/10.1007/978-3-030-51005-3_24.
27. Rahman A, Smith DV, Timms G. A novel machine learning approach toward quality assessment of sensor data. IEEE Sens J. 2014;14(4):1035–47. https://doi.org/10.1109/JSEN.2013.2291855.
28. Ray C. Data variety and integrity assessment for maritime anomaly detection. CEUR Workshop Proc. 2018;2343:4–7.
29. Rodríguez CCG, Servigne S. Managing sensor data uncertainty. Int J Agric Environ Inf Syst. 2013;4(1):35–54. https://doi.org/10.4018/jaeis.2013010103.
30. Seo S, Mohegh A, Ban-Weiss G, Liu Y. Automatically inferring data quality for spatiotemporal forecasting. In: International Conference on Learning Representations, 2018. https://openreview.net/forum?id=ByJIWUnpW.
31. Sidi F, Panah PHS, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. Data quality: A survey of data quality dimensions. In: Mahmod R, Abdullah R, Abdullah LN, Sembok TMT, Smeaton AF, Crestani F, Doraisamy S, Kadir RA, Ismail M. (eds.) 2012 International Conference on Information Retrieval and Knowledge Management, Kuala Lumpur, Malaysia, 2012. IEEE. p. 300–4. https://doi.org/10.1109/InfRKM.2012.6204995.
32. Skillicorn D. Understanding complex datasets: data mining with matrix decompositions. In: Chapman & Hall/CRC data mining and knowledge discovery series. CRC Press; 2007. https://books.google.fr/books?id=9SzLDi8jUnAC.
33. Song S, Sun Y, Zhang A, Chen L, Wang J. Enriching data imputation under similarity rule constraints. IEEE Trans Knowl Data Eng. 2020;32(2):275–87. https://doi.org/10.1109/TKDE.2018.2883103.
34. Song S, Zhang A, Chen L, Wang J. Enriching data imputation with extensive similarity neighbors. Proc VLDB Endow. 2015;8(11):1286–97. https://doi.org/10.14778/2809974.2809989.
35. Teh HY, Kempa-Liehr AW, Wang KI. Sensor data quality: a systematic review. J Big Data. 2020;7(1):11. https://doi.org/10.1186/s40537-020-0285-1.
36. Troyanskaya OG, Cantor MN, Sherlock G, Brown PO, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5. https://doi.org/10.1093/bioinformatics/17.6.520.
37. Wang L, Zhang D, Wang Y, Chen C, Han X, M'Hamed A. Sparse mobile crowdsensing: challenges and opportunities. IEEE Commun Mag. 2016;54(7):161–7. https://doi.org/10.1109/MCOM.2016.7509395.
38. Wonohardjo E, Kusuma Negara IGP. Air pollution mapping using mobile sensor based on internet of things. Procedia Comput Sci. 2019;157:638–45. https://doi.org/10.1016/j.procs.2019.08.224.
39. Yi X, Zheng Y, Zhang J, Li T. ST-MVL: filling missing values in geo-sensory time series data. In: S. Kambhampati (ed.) Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15. 2016. IJCAI/AAAI Press; 2016. p. 2704–10. http://www.ijcai.org/Abstract/16/384.
40. Zappatore M, Loglisci C, Longo A, Bochicchio M, Vaira L, Malerba D. Trustworthiness of context-aware urban pollution data in mobile crowd sensing. IEEE Access. 2019. https://doi.org/10.1109/ACCESS.2019.2948757.