Trust Assessment on Data Stream Imputation in IoT Environments

Tao Peng, Sana Sellami, Omar Boucelma, Richard Chbeir

Trust Assessment on Data Stream Imputation in IoT Environments

Tao Peng, Sana Sellami, Omar Boucelma[1] and Richard Chbeir[2]

[1] Aix Marseille Univ, CNRS, LIS, Marseille, France
tao.peng@univ-amu.fr, sana.sellami@univ-amu.fr, omar.boucelma@univ-amu.fr
[2] Univ Pau & Pays Adour, E2S-UPPA, LIUPPA, EA3000, Anglet, France
richard.chbeir@univ-pau.fr

Abstract. In the era of internet of Things, stream data emitted by sensors may rise quality issues such as incompletenes caused mainly by sensors failure or transmission problems. It is therefore necessary to recover missing data because missing values can impact decision making. Within this landscape, trust on data imputation is a key issue for helping stakeholders involved in such process. In this paper, we address the problem related to the trustworthiness on imputed data streams in IoT environments. We propose here a method called CSIV (Confidence Score for Imputed Values) to assess trust by assigning a confidence score to imputed data. CSIV considers both trust score of non-missing values and neighboring sensors. We have evaluated CSIV on real datasets using accuracy and trustworthiness as evaluation metrics. Experiments show that CSIV is able to assign correctly a trust score to the imputed values.

Keywords: Internet of Things · Stream data · Imputation · Trust

1 Introduction

Since the advent of IoT, access to sensor data has become commonplace in many fields: environmental monitoring, road traffic monitoring, or e-health to cite a few.

Data emitted by sensors at real time are aggregated as data streams than may be ingested by different IoT applications or services. However, those streams may rise quality issues such as inaccuracy or incompleteness, due to issues such as sensor failure or network issues and leading to loss of precision and difficulties in a decision making process.

Missing value repairing can be performed in adopting different strategies such as [15]: 1) Delete incomplete observations; 2) Manual repair; 3) Substitute by a constant/last-observation/mean; and 4) Estimate the most probable value. The first three strategies are not suitable for IoT data streams. The last strategy, also called imputation, does not need human intervention and is much more efficient than manual ones. Data imputation estimates the most probable value by using as much information as possible from the gathered observations to repair missing values [15].

In this paper, we address the problem of assessing trustworthiness on imputed data streams in IoT environments. Most of the data imputation works [8, 5] are based on the assumption that data imputation can be evaluated according to the difference between the reference values and the simulated missing ones. However, in the real scenarios, data is unavailable. Here, we propose a method called CSIV, standing for Confidence Score for Imputed Values, to assess trust by assigning a confidence score to imputed data, which extends our previous work on data trust assessment [12]. We adopt the same definition provided in [1] stating the *"Data Trustworthiness in IoT Networks is the subjective probability that data observed by a user is consistent with the data at the source"*, and consequently define imputed data trustworthiness as the subjective probability that imputed data is close to the expected value. CSIV is based on: (1) trust score of non-missing data, and (2) trustworthy neighboring sensors. Experiments conducted on real datasets demonstrate the efficiency of CSIV while assessing imputed data accuracy, hence ensuring trustworthiness of the values being imputed.

This paper is organized as follows. We present a literature review in Section 2 and describe our approach in Section 3. Section 4 presents the experiments and validation setting. Finally, Section 5 concludes the paper and pin down several future directions.

2 Related Works

In this section, we review related works on data Trustworthiness and the evaluation of imputation accuracy.

The work described in [9] proposed a cyclic trust computation framework for data streams: (a) the more trusted data reported by the sensor, the higher is the (provider's) reputation; (b) data trust depends on all of data similarity, provenance similarity and sensor reputation. This approach is based on the hypothesis that the sensor data is independent of one another and follows the same Gaussian Distribution $N\left(\mu, \sigma^2\right)$. However, data streams are non-stationary which means that they do not have the same distribution and are time-dependent.

Other recent works [1, 10, 4] suppose that the residual ($r_{f,t} = \hat{d}_{f,t} - d_{f,t}$), where $\hat{d}_{f,t}$ is the estimated value and $d_{f,t}$ is the emitted value by a sensor f, follows a Gaussian Distribution. If the prediction is not biased, then the expected value μ is close to zero. A prediction is trustworthy if the residual is within the confidence interval of 95% (i.e. $[\mu \pm 1.98 * \delta]$, where δ is the standard deviation). In [1, 10], a trust score ($s_{f,t}$ in the Equation 1), is proposed. Cumulative Distribution function takes the residual as input and outputs a trust score: if this score exceeds a threshold, the received data is trusted. However, confidence interval and trust score $s_{f,t}$ can not be applied to assess trust on missing values because they are based on the hypothesis that missing values are simulated and then available, which is not the case in real scenarios.

$$s_{f,t} = \begin{cases} \frac{2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{r_{f,t}} EXP\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x, \text{if } r_{f,t} < \mu \\ \frac{2}{\sigma\sqrt{2\pi}} \int_{r_{f,t}}^{+\infty} EXP\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x, \text{if } r_{f,t} \geq \mu \end{cases} \quad (1)$$

According to [6], evaluation methods for imputed values can be classified into two groups: indirect and direct. Indirect methods consist in: 1) training a classifier with the set of imputed data, 2) using a classifier on a set of test data without missing values, and 3) using accuracy of better classifier to assess the accuracy of imputation. However, indirect methods can not be done in real time because they need to access to all imputed data. Direct methods are based on the difference between imputed and missing values. They can be applied on the simulated missing values, but in real scenario missing values are often unavailable [14]. In [8, 5], the authors proposed imputation methods on data streams and used direct method based on the root mean squared error $RMSE_{SMV}$ (Equation 2) to evaluate the accuracy of imputed simulated missing values (SMV).

$$RMSE_{SMV} = \sqrt{\frac{1}{|SMV|} \sum_{d_{f,t} \in SMV} \left(\hat{d}_{f,t} - d_{f,t}\right)^2} \tag{2}$$

In [2], the authors used RMSE of the sliding window (w) (Equation 3) to assess the accuracy of the regression model in data streams. Equation 3 takes into account accuracy changes due to the non stationary data but assumes that $d_{f,t}$ exists which is not always true since it is a missing value.

$$RMSE_t = \sqrt{\frac{1}{w} \sum_{i=t-w}^{t} \left(\hat{d}_{f,t} - d_{f,t}\right)^2} \tag{3}$$

In a nutshell, most of the evaluation methods in data streams are applied in the case of simulated data and consider the difference between the estimated value and the simulated one to determine the accuracy of the prediction. In our work, we take advantage of the spatial and temporal characteristics of data streams and assess trust of imputed data based on the confidence score of non-missing values and the trustworthiness of neighboring sensors.

3 CSIV method

In this section, we describe our method CSIV (Confidence Scores of Imputed Values) for assessing trustworthiness in imputed values. Notations used in the followng subsections are detailed in Table 1.

3.1 Method description

As shown in Fig. 1, CSIV consists of three steps: 1) trust assessment of data, 2) trust evaluation of sensors, and 3) trust assessment of imputed data. Due to the non-stationary data, the accuracy of imputed values can change over time [5] which leads to a concept drift. In view of the concept drift and in order to represent the trust score at one point, a sliding window is used for the next steps.

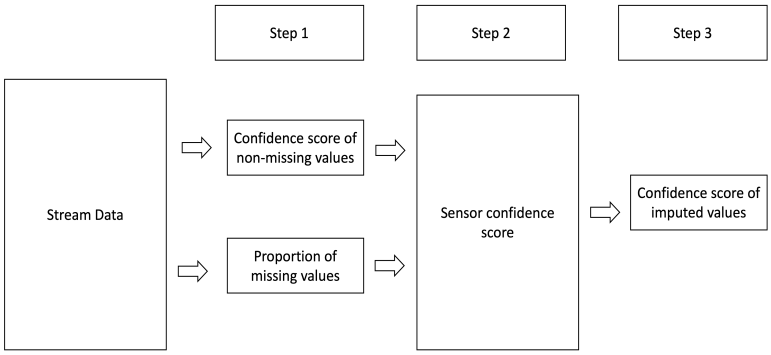

Fig. 1. CSIV method processes

Phase 1: Data Trustworthiness: For each sensor $f \in F$ at time t, if $d_{f,t}$ is not a missing value, we calculate a trust score of $d_{f,t}$ according to Equation 1. If the error is low, then $d_{f,t}$ will have a higher trust score. Indeed, the more the errors are higher, the low will be the trust score. Also, we determine, for each sensor $f \in F$ in the sliding window, the proportion of missing values in order to evaluate its trust score.

Phase 2: Sensors Trustworthiness: For each sensor $f \in F$ at time t, we consider that the more trusted data reported by the sensor, the higher is the (provider's) reputation in a sliding window (i.e., $[t-w, t]$) [3]. Moreover, if the sensor rarely generates missing values in the sliding window, then its trust score will be higher at time t [7].

Phase 3: Imputed values Trustworthiness: Given an imputed value $\hat{d}_{f,t}$, $\{f\} \bigcup K_f$ is the set of relevant sensors. The trust score of imputed value $\hat{d}_{f,t}$ is determined based on the trust score $\Re_{f,t} = \{sc_{f',t} \mid f' \in K_f\} \bigcup \{sc_{f,t}\}$ of relevant sensors which generate trustworthy data. For each imputed value, if the trust score of relevant sensors is higher, then the trust score of imputed value will be higher.

3.2 Algorithm

Algorithm 1 illustrates the pseudo-code description of CSIV. At time t, for all the sensors $f \in K$ (Algorithm 1 lines 2, 3), the trust score of a sensor $sc_{f,t}$ is determined by $sc_{f,t} \leftarrow (\frac{1}{w}\sum_{i=t-w}^{t-1} s_{f,i}) * (1 - RatioMV_{f,t})$. $(\frac{1}{w}\sum_{i=t-w}^{t-1} s_{f,i})$ is the average trust scores of the sensor values f in the sliding window w at time t. If $d_{f,t}$ is a missing value, the residual, being the difference between the predicted value $\hat{d}_{f,t}$ and the real value $d_{f,t}$, is denoted $r_{f,t} = \hat{d}_{f,t} - d_{f,t}$ (line 7). Assuming that the residual follows a Gaussian Distribution (Expected value μ and standard deviation δ), the trust score $s_{f,t}$ of $d_{f,t}$ is calculated at line 8. For all sensors $f \in K$ at time t, if $d_{f,t}$ is a missing value, then $\hat{d}_{f,t}$ is the estimated

value by an imputation algorithm such as ISTM (Incremental Space-Time-based model) [11] which provides an estimation for the missing value based on nearly historical data and the observations of neighboring sensors of the default one. The trust score of relevant sensors (line 10) of $d_{f,t}$ depends on the trust score of f and its neighbors K_f. The trust sore of $\hat{d}_{f,t}$ (denoted $\hat{s}_{f,t}$) is calculated by a function G that takes as input $\Re_{f,t}$ (line 11) and is defined as follows:

Notation	Explanation
F	is a set of sensors
f	is a sensor, $f \in K$
K_f	is the neighbors set of f where $K_f \in F$
$d_{f,t}$	is the real value generated by a sensor f at time t
$\hat{d}_{f,t}$	is the predicted value of $d_{f,t}$
$r_{f,t}$	is the error which is the difference between the predicted value and the real value
μ	is the expected value of $r_{f,t}$
δ	is the standard deviation of $r_{f,t}$
$s_{f,t}$	is the trust score of $d_{f,t}$
$\hat{s}_{f,t}$	is the trust score of $\hat{d}_{f,t}$
$sc_{f,t}$	is the trust score of a sensor f at time t
w	is the length of a sliding window
$RatioMV_{f,t}$	is the proportion of missing values in the sliding window for a sensor f at time t
$f \bigcup K_f$	are the relevant sensors of $\hat{d}_{f,t}$
$\Re_{f,t}$	$= \{sc_{f',t} \mid f' \in K_f\} \bigcup \{sc_{f,t}\}$, are trust scores of relevant sensors of $\hat{d}_{f,t}$
CSIV	is the proposed method
G	is a function that takes as input $\Re_{f,t}$ ad gives as output the trust score of $\hat{d}_{f,t}$ (denoted $\hat{s}_{f,t}$).
$G_{avg}(\Re_{f,t})$	$= average(\Re_{f,t})$.
$G_{min}(\Re_{f,t})$	$= \min(\Re_{f,t})$.
$CSIV_{min}$	is CSIV , where G is G_{min}
$CSIV_{avg}$	is CSIV , where G is G_{avg}

Table 1. CSIV algorithm notations

- $G_{min}(\Re_{f,t}) = \min(\Re_{f,t})$ which means that the trust score of the imputed value is equal to the minimum trust score values of the relevant sensors.
- $G_{avg}(\Re_{f,t}) = average(\Re_{f,t})$ which is the average of $\Re_{f,t}$.

4 Experiments

We present here the experiments that have been conducted on two real-world datasets in order to assess the accuracy of CSIV and trustworthiness of imputed data. Also, it is worthy to note that the experiments have been conducted using a MAC mini 2014, Core i5 chip, 8GB RAM, with Python 3.7.

Algorithm 1 Trust Assessment on imputed values

Require: F: set of sensors
$d_{f,t}$: a value emitted by a sensor f at time t;
$\hat{d}_{f,t}$: the estimation of $d_{f,t}$
K_f: a neighbor set on a sensor $f \in F$.
$N\left(\mu, \sigma^2\right)$: Gaussian Distribution of $\hat{d}_{f,t} - d_{f,t}$
w: the length of sliding window
$RatioMV_{f,t}$: the proportion of missing values in the sliding window with a width of w for each sensor f at time t

Ensure: $\hat{s}_{f,t}$: Trust score estimation for imputed values

for t = n,n+1,.... **do**
  for each sensor $f \in F$ **do**
    $sc_{f,t} \leftarrow (\frac{1}{w}\sum_{i=t-w}^{t-1} s_{f,i}) * (1 - RatioMV_{f,t})$
  end for
  for each sensor $f \in F$ **do**
    if $d_{f,t}$ is a non missing value **then**
      $r_{f,t} \leftarrow \hat{d}_{f,t} - d_{f,t}$
      $s_{f,t} = \begin{cases} \frac{2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{r_{f,t}} EXP\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x, \text{si } r_{f,t} < \mu \\ \frac{2}{\sigma\sqrt{2\pi}} \int_{r_{f,t}}^{+\infty} EXP\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x, \text{si } r_{f,t} \geq \mu \end{cases}$
    else
      $\Re_{f,t} \leftarrow \{sc_{f',t} \mid f' \in K_f\} \bigcup \{sc_{f,t}\}$
      $\hat{s}_{f,t} \leftarrow G(\Re_{f,t})$
      $s_{f,t} \leftarrow \hat{s}_{f,t}$
    end if
  end for
end for

4.1 Description of the datasets

Our experiments are performed on two datasets: CityPulse [3] and Appliances energy prediction Data Set from UCI Machine Learning Repository [4].

- CityPulse dataset covers seven different domains: Road Traffic, Parking, Pollution, Weather, Cultural, Social and Library Events Data of Aarhus, Denmark and Brasov, Romania for years 2014 and 2015. Among all these parts, Road Traffic Data is of greatest importance and represents data about travel information of Aarhus (Danmark) during the following periods: "2/2014 - 6/2014","8/2014 - 9/2014", "10/2014 - 11/2014"," 07/2015 - 10/2015". There is a total of 449 monitors (assuming that one sensor was installed in one area). The volume of the data in format CSV is 747.2 MB. Traffic Data is collected by many sensors installed on the road. Every 5 minutes, each sensor will send a bunch of information (one line of table Traffic Data) to a central computer center. Every 5 minutes, the center receives 29,940 Bytes (0.029MB). All the sensors located within 1km are considered as neighbors.

[3] http://www.ict-citypulse.eu/
[4] https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

There are on average 21 neighbors per sensor. The minimum value is 5 and the maximum value is 68.
– AEP data consists of the following attributes: energy assumption, humidity and temperature. Data is averaged for 10 minutes period and gathered during 4.5 months (from 11/01/2016 to 05/27/2016) resulting in a total 12 MB CSV file with 19735 instances. One humidity sensor and one temperature sensor are installed in each room and outside the building (18 sensors in total and all of them are regarded as neighbors).

Real Missing Value: For CityPulse data, the total missing value rate is close to 9%. The dataset AEP does not have missing values.

Simulated Missing Value (SMV): We simulate some values randomly (Missing completely at random or MCAR) for datasets: we randomly select a percentage of data (5%, 10%, 15%, 20%, 25%, 30%) according to the discrete uniform distribution and mark them as Simulated Missing Values (SMV). Thanks to SMV and its ground truth, we can measure the effectiveness of the reparation. The percentage rate of simulated missing values is borrowed from work [13] where the percentage of missing value or simulated missing value vary from 5% to 30%. There are both real missing values and simulated missing value in our test data.

Fig. 2 shows CityPulse data distribution (SMV in red and original one in blue) for 10 sensors over a period of 5 hours.

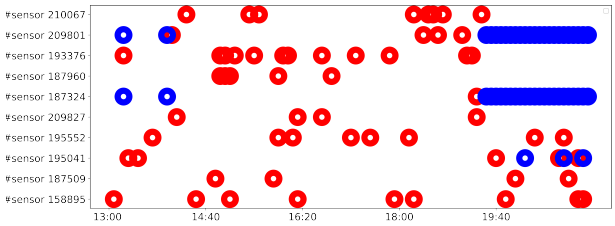

Fig. 2. 5% of SMV (red circle) and real missing value (blue circle) in CityPluse

Figure 3 shows the distribution of the error of non missing values in AEP and CityPulse. We note that the residual does not follow a Gaussian Distribution. After Kolmogorov–Smirnov test (KS test), all the PValue are higher than 0.05.

4.2 Evaluation metrics

For evaluation purposes, we used *Accuracy* and *trustworthiness* metrics:

– Prediction **Accuracy** is evaluated in terms of the root mean squared error (RMSE) of all variables (Equation 4).

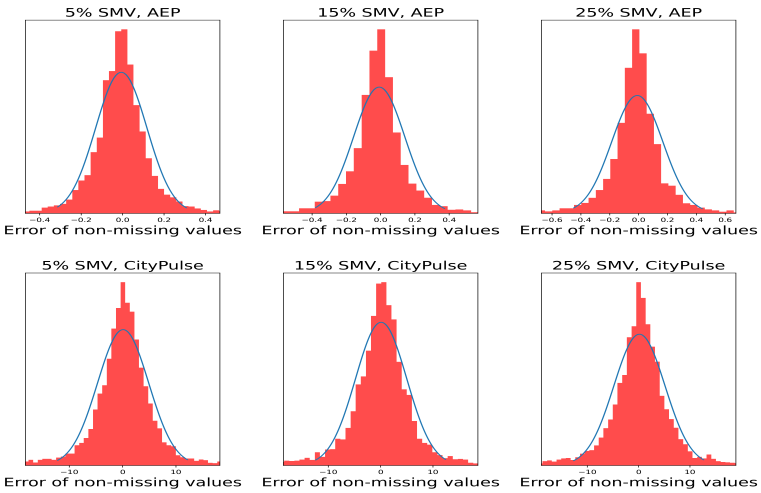

Fig. 3. Error distribution of the estimation of non-missing values, ISTM

$$RMSE = \sqrt{\frac{1}{\mid SMV \mid} \sum_{d_{f',t'} \in SMV} \left(\hat{d}_{f',t'} - d_{f',t'}\right)^2} \tag{4}$$

– Trust **Assessment**: To measure trustworthiness in a predicted value, we use Confidence Score as defined in Equation 5.

$$RMSE_{trust} = \sqrt{\frac{1}{\mid SMV \mid} \sum_{d_{f',t'} \in SMV} \left(\hat{s}_{f',t'} - s_{f',t'}\right)^2} \tag{5}$$

4.3 Configurations

In order to evaluate the trustworthiness of predicted values, we use ISTM [11] model which is an Incremental Spatio-Temporal regression method for repairing missing values in IoT data streams. The configuration of ISTM is as follows: 30% of instances are taken for initiation. For CityPulse dataset, given one sensor, neighbors' sensors within 1 km around are considered as its neighbors. The size of reference dataset g is set to 6 which represents the value where the precision of ISTM becomes stable (Table 2). For AEP dataset, all sensors are neighbors to each other and $g = 4$.

g	1	2	3	4	5	6	7	8	9
RMSE (CityPulse)	7.84	7.82	7.74	7.72	7.73	7.70	7.70	7.71	7.70
RMSE (AEP)	0.62	0.558	0.555	0.554	0.556	0.557	0.556	0.557	0.556

Table 2. Precisions of ISTM by varying g, with 25% of SMV in CityPulse and AEP datasets

4.4 Results

In this section, we highlight the obtained results along the line of two metrics: accuracy and trustworthiness.

MV (%)	5	10	15	20	25	5	10	15	20	25
	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{avg}$	$CSIV_{avg}$	$CSIV_{avg}$	$CSIV_{avg}$	$CSIV_{avg}$
w =5	0.28	0.32	0.37	0.4	0.45	0.09	0.09	0.09	0.11	0.12
w =10	0.22	0.25	0.29	0.32	0.37	0.08	0.09	0.09	0.11	0.12
w =30	0.16	0.18	0.21	0.24	0.27	0.09	0.09	0.1	0.11	0.13
w =60	0.15	0.16	0.18	0.21	0.23	0.09	0.09	0.1	0.12	0.13
w =90	0.14	0.15	0.17	0.2	0.22	0.09	0.09	0.1	0.12	0.13
w =120	0.13	0.15	0.17	0.19	0.22	0.09	0.1	0.1	0.12	0.14
w =150	0.13	0.15	0.17	0.19	0.21	0.1	0.1	0.11	0.12	0.14
w =180	0.13	0.15	0.16	0.18	0.21	0.1	0.1	0.11	0.12	0.14
w =200	0.13	0.15	0.17	0.19	0.21	0.1	0.1	0.11	0.13	0.14
w optimal	120	90	180	180	150	10	10	10	10	10
RMSE optimal	0.13	0.15	0.16	0.18	0.21	0.08	0.09	0.09	0.11	0.12

Table 3. RMSE with different sliding window length (w), AEP

Sliding Window analysis : We evaluated the accuracy of $CSIV_{avg}$ and $CSIV_{min}$ by varying sliding window size. The results (Tables 3 and 4) show that: 1) For $CSIV_{avg}$, the best length value of window is 10 and RMSE tends to increase when the length of window increases, and 2) For $CSIV_{min}$, the optimum value of sliding window is between 90 and 180 and RMSE varies slightly when the length of window increases.

Impact of SMV : We analyzed the impact of SMV on the accuracy of trust scores. Fig. 4 (a) and Fig. 4 (b) show that the RMSE of $CSIV_{min}$ and $CSIV_{avg}$ is higher when the proportion of SMV increases. In addition, we note that the accuracy of $CSIV_{avg}$ is better than $CSIV_{min}$ for both dataset. Indeed, the lower the RMSE value of $CSIV_{avg}/CSIV_{min}$, the more accurately their trust scores are estimated. Moreover, $CSIV_{min}$ underestimates trust score in comparison to $CSIV_{avg}$.

Trust score of sensors : Trust score of a sensor relies on the trustworthiness of his neighbors and the accuracy of the non missing values. Fig. 5 illustrates the trust score of a sensor 158895 and its neighbors over a two days period. We can note that the trust score of sensors is positively correlated to the accuracy of predicted

MV (%)	5	10	15	20	25	5	10	15	20	25
	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{min}$	$CSIV_{avg}$	$CSIV_{avg}$	$CSIV_{avg}$	$CSIV_{avg}$	$CSIV_{avg}$
w =5	0.3	0.34	0.38	0.42	0.45	0.09	0.1	0.11	0.12	0.14
w =10	0.25	0.28	0.31	0.35	0.38	0.09	0.1	0.11	0.12	0.14
w =30	0.21	0.23	0.26	0.28	0.31	0.1	0.1	0.12	0.13	0.15
w =60	0.2	0.22	0.24	0.26	0.29	0.1	0.11	0.12	0.13	0.15
w =90	0.19	0.21	0.23	0.25	0.28	0.1	0.11	0.12	0.14	0.16
w =120	0.18	0.2	0.22	0.24	0.27	0.1	0.11	0.12	0.14	0.16
w =150	0.17	0.19	0.2	0.24	0.26	0.1	0.11	0.12	0.14	0.16
w =180	0.17	0.19	0.21	0.23	0.26	0.1	0.11	0.12	0.14	0.16
w =200	0.17	0.19	0.2	0.23	0.26	0.1	0.11	0.12	0.14	0.16
w optimal	150	150	150	180	150	10	15	10	10	10
RMSE optimal	0.17	0.19	0.2	0.23	0.26	0.09	0.1	0.11	0.12	0.14

Table 4. RMSE with different sliding window length (w), CityPulse

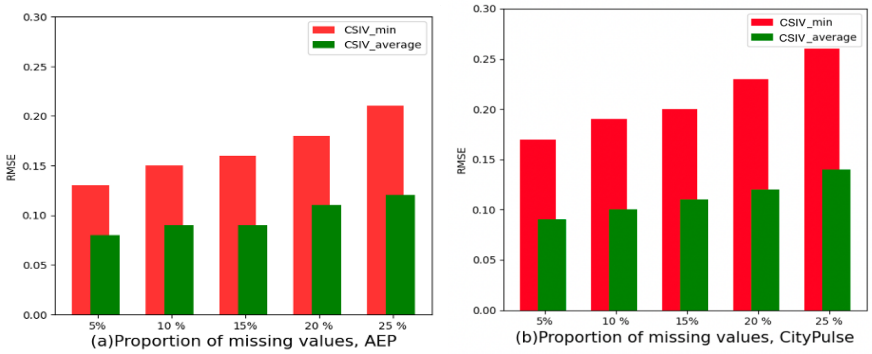

Fig. 4. RMSE of confidence score with varying proportions of missing data (sliding window lengths are optimized), ISTM.

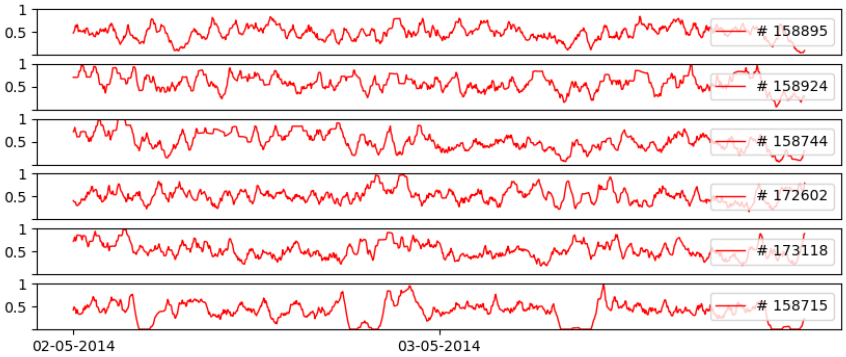

Fig. 5. Trust score of sensor 158895 and his neighbors over a two day period, CityPulse, 15% SMV, $w = 10$, ISTM.

values and ISTM is able to ensure continuous accurate in the streaming data because it is updated incrementally. Moreover, when the proportion of missing values increases (this is the case of CityPulse data which have original missing data), the trust score of the sensor is close to 0.

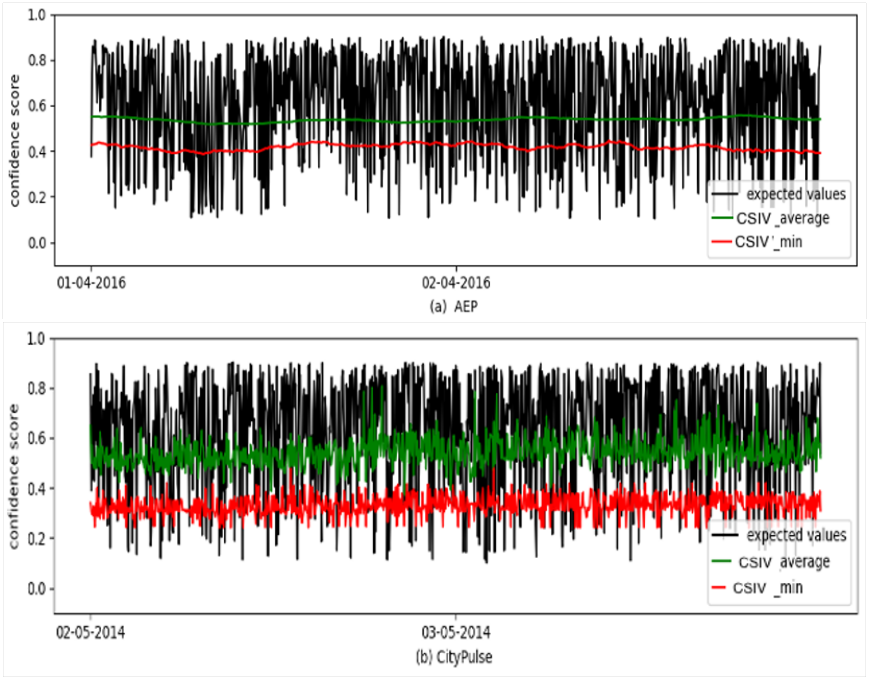

Fig. 6. Trust score of imputed values (15% SMV) of G_{min} and G_{avg} over a two day period, ISTM

Trustworthiness evaluation : We assess trustworthiness (according to Equation 5) of $CSIV_{avg}$ and $CSIV_{min}$. Fig. 6 (a) and Fig. 6 (b) show that the scores of $CSIV_{avg}$ (green curve), when the ratio of missing values is equal to 15%, are closer to the expected values (black curve) than $CSIV_{min}$ (red curve). This can be explained by the fact that $CSIV_{avg}$ combines trust scores of several sensors. Then, according to $CSIV_{avg}$, CSIV is able to assign correctly a trust score to the imputed values.

5 Conclusion and Future Work

In this paper, we described CSIV, a method that assigns a Confidence Score to Imputed Values.

For validation purposes, we conducted several experiments with imputation model on real datasets using *accuracy* and *trustworthiness* as evaluation metrics.

For the future, we intend to optimize the confidence score by taking into account only sensors which produce small proportion of missing data in order to avoid the risk of underestimation. Moreover, we aim to apply our method on other imputation methods such as deep learning models in order to be able to explain the imputation of missing values within a dataset [16].

References

1. Adams, S., Beling, P.A., Greenspan, S., Velez-Rojas, M., Mankovski, S.: Model-based trust assessment for internet of things networks. In: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). pp. 1838–1843. IEEE (2018)
2. Barddal, J.P.: Vertical and horizontal partitioning in data stream regression ensembles. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
3. Bertino, E.: Data trustworthiness—approaches and research challenges. In: Data privacy management, autonomous spontaneous security, and security assurance, pp. 17–25. Springer (2014)
4. Chhabra, G., Vashisht, V., Ranjan, J.: A comparison of multiple imputation methods for data with missing values. Indian Journal of Science and Technology **10**(19), 1–7 (2017)
5. Dong, W., Gao, S., Yang, X., Yu, H.: An exploration of online missing value imputation in non-stationary data stream. SN Computer Science **2**(2), 1–11 (2021)
6. Hasan, M.K., Alam, M.A., Roy, S., Dutta, A., Jawad, M.T., Das, S.: Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). Informatics in Medicine Unlocked **27**, 100799 (2021)
7. Junior, F.M.R., Kamienski, C.A.: A survey on trustworthiness for the internet of things. IEEE Access **9**, 42493–42514 (2021)
8. Lee, M., An, J., Lee, Y.: Missing-value imputation of continuous missing based on deep imputation network using correlations among multiple iot data streams in a smart space. IEICE TRANSACTIONS on Information and Systems **102**(2), 289–298 (2019)
9. Lim, H.S., Moon, Y.S., Bertino, E.: Provenance-based trustworthiness assessment in sensor networks. In: Proceedings of the Seventh International Workshop on Data Management for Sensor Networks. pp. 2–7 (2010)
10. Liu, J., Adams, S., Beling, P.A.: An ensemble trust scoring method for internet of things sensor networks. In: 2020 IEEE 6th World Forum on Internet of Things (WF-IoT). pp. 1–6. IEEE (2020)
11. Peng, T., Sellami, S., Boucelma, O.: Iot data imputation with incremental multiple linear regression. Open J. Internet Things **5**(1), 69–79 (2019)
12. Peng, T., Sellami, S., Boucelma, O.: Trust assessment on streaming data: A real time predictive approach. In: Lemaire, V., Malinowski, S., Bagnall, A.J., Guyet, T., Tavenard, R., Ifrim, G. (eds.) Advanced Analytics and Learning on Temporal Data - 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers. Lecture Notes in Computer Science, vol. 12588, pp. 204–219. Springer (2020)

13. Puiu, D., Barnaghi, P., Tönjes, R., Kümper, D., Ali, M.I., Mileo, A., Parreira, J.X., Fischer, M., Kolozali, S., Farajidavar, N., et al.: Citypulse: Large scale data analytics framework for smart cities. IEEE Access **4**, 1086–1108 (2016)
14. Ramirez-Gallego, S., Krawczyk, B., Garcia, S., Wozniak, M., Herrera, F.: A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing **239**, 39–57 (2017)
15. Somasundaram, R., Nedunchezhian, R.: Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications, Vol21 **21**(10) (2011)
16. Vu, M.A., Nguyen, T., Do, T.T., Phan, N., Halvorsen, P., Riegler, M.A., Nguyen, B.T.: Conditional expectation for missing data imputation. CoRR **abs/2302.00911** (2023)