REMIAN: Real-Time and Error-Tolerant Missing Value Imputation

QIAN MA, Dalian Maritime University
YU GU, Northeastern University
WANG-CHIEN LEE, The Pennsylvania State University
GE YU, Northeastern University
HONGBO LIU, Dalian Maritime University
XINDONG WU, University of Louisiana at Lafayette

Missing value (MV) imputation is a critical preprocessing means for data mining. Nevertheless, existing MV imputation methods are mostly designed for batch processing, and thus are not applicable to streaming data, especially those with poor quality. In this article, we propose a framework, called *Real-time and Error-tolerant Missing vAlue ImputatioN* (REMAIN), to impute MVs in poor-quality streaming data. Instead of imputing MVs based on *all* the observed data, REMAIN first initializes the MV imputation model based on *a-RANSAC* which is capable of detecting and rejecting anomalies in an efficient manner, and then incrementally updates the model parameters upon the arrival of new data to support real-time MV imputation. As the correlations among attributes of the data may change over time in unforseenable ways, we devise a *deterioration detection* mechanism to capture the deterioration of the imputation model to further improve the imputation accuracy. Finally, we conduct an extensive evaluation on the proposed algorithms using real-world and synthetic datasets. Experimental results demonstrate that REMAIN achieves significantly higher imputation accuracy over existing solutions. Meanwhile, REMAIN improves up to one order of magnitude in time cost compared with existing approaches.

CCS Concepts: • **Information systems → Data cleaning**; *Data stream mining*; • **Mathematics of computing → Time series analysis**; *Regression analysis*; • **Computing methodologies → Anomaly detection**;

Additional Key Words and Phrases: Missing value, poor-quality streaming data, real-time imputation

This work is partly supported by the China Postdoctoral Science Foundation (Grant No. 2019M661077), the National Natural Science Foundation of China (Grant Nos. 61772102, 61751205, and 61872070), National Science Foundation (Grant No. IIS-1717084), Liaoning Collaborative Fund (Grant No. 2020-HYLH- 17), and Liaoning Revitalization Talents Program (Grant No. XLYC1807158).
Authors' addresses: Q. Ma, Dalian Maritime University, Dalian, China, 116026; email: maqian@dlmu.edu.cn; Y. Gu (corresponding author), Northeastern University, Shenyang, China, 110819; email: guyu@mail.neu.edu.cn; W.-C. Lee, The Pennsylvania State University, State College; email: wlee@cse.psu.edu; G. Yu, Northeastern University, Shenyang, China, 110819; email: yuge@mail.neu.edu.cn; H. Liu, Dalian Maritime University, Dalian, China, 116026; email: lhb@dlmu.edu.cn; X. Wu, University of Louisiana at Lafayette, Lafayette; email: xwu@louisiana.edu.

1556-4681/2020/09-ART77 $15.00
https://doi.org/10.1145/3412364

ACM Reference format:
Qian Ma, Yu Gu, Wang-chien Lee, Ge Yu, Hongbo Liu, and Xindong Wu. 2020. REMIAN: Real-Time and Error-Tolerant Missing Value Imputation. *ACM Trans. Knowl. Discov. Data* 14, 6, Article 77 (September 2020), 38 pages.
https://doi.org/10.1145/3412364

1 INTRODUCTION

Due to various uncontrollable factors, e.g., hardware failure, unconscious malfunction, and participants refusal, the issue of *missing values* (MVs) is ubiquitous in many real-world datasets. The task of *MV imputation*, aiming to replace the MVs with some plausible ones, is critical because many data mining and analytics applications, e.g., machine learning [15, 29] and pattern mining [14, 31], do not handle the datasets with MVs well. Over years, many MV imputation algorithms [1, 25, 34, 35, 40, 45] have been developed. Most algorithms assume that the observed data as accurate ground truths and use *all* of them to impute the MVs directly. However, in many real-world applications, data are collected with no guarantee on data quality and credibility. For example, based on [28], 7% of Stock data in *Yahoo! Finance* and 5% of Flight data in *Travelocity* are inaccurate, even though those sources of Stock and Flight information are considered as highly reliable. On the other hand, many kinds of real-world data, e.g., sensory data, network traffic data, and web clicks data, usually arrives sequentially and continuously as a high-speed stream [2, 4, 5], which requires real-time processing. We consider the continuously arriving data containing MVs and anomalies as *poor-quality streaming data*. Based on our analysis, the poor-quality streaming data has the following characteristics: (1) The data arrive continuously and in real time. (2) Besides MVs, anomalies may be embedded in the data. (3) Both values and correlations among attributes of the data dynamically evolve over time.

Challenges. Owing to the aforementioned characteristics, MV imputation for poor-quality streaming data is challenging in the following aspects:

(1) **Real-time imputation.** In contrast to batch processing, the streaming data are collected sequentially and to be processed in an online fashion. Undoubtedly, the MVs in the data must be imputed in real time. On the other hand, the imputation model needs to be updated incrementally for real-time MV imputation as the observed values of the data change over time. However, most existing MV imputation algorithms, e.g., GBKII [40], CMI [45], ERACER [21], OSICM [22], and IIM [38], are designed for batch processing which does not support real-time imputation for streaming data.

(2) **Error tolerance.** Since the anomalies may mask the real distribution or the correlations among attributes of the data, the imputation model learned using all the observed data (including anomalies) may be inaccurate. Thus how to learn an effective imputation model from the poor-quality streaming data becomes another challenge. However, most existing approaches do not consider the impact of anomalies in the data. In conventional online MV imputation methods (where most of them are based on autoregression (AR) model [16]), the MVs at time t are typically imputed using all the historical data collected from a time window of previous p time points, by exploring the temporal dependencies in the data. Intuitively, if the historical data contain anomalies, the MV imputation results at time t are likely to be inaccurate. Thus, for an existing online MV imputation algorithm to perform well on poor-quality streaming data, the anomalies not only need to be *detected* but also *repaired*. Note that real-time anomaly repair [39] is a non-trivial research problem. Existing anomaly repair methods also explore the temporal dependencies in the data. Thus, the ideas behind anomaly repair and MV imputation are not independent from each other, making the MV imputation for poor-quality streaming data more complex, leading to unsatisfactory imputation results.

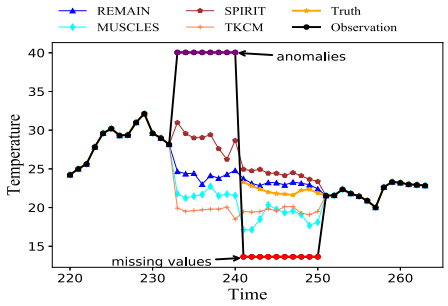


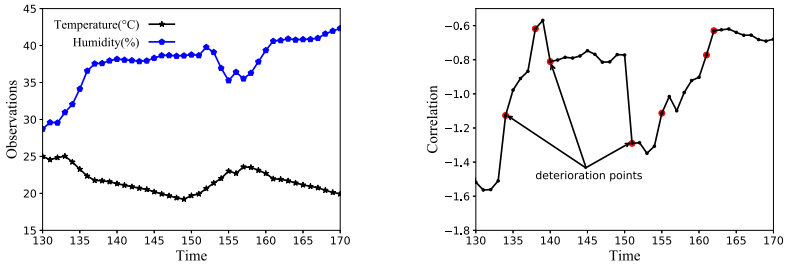
Fig. 1. A segment of temperature values of sensor 1.

Example 1. Consider a real-world dataset, IDL,[1] which contains data sampled from 54 sensors deployed in the Intel Berkeley Research Lab. Each sensor captures the temperature and humidity of the lab every 31 seconds. Figure 1 presents a segment of the observed temperature values of Sensor 18 evolving over time, denoted by black line. In this segment of data, the real anomalies occur in the period from time 233 to 240. Notably, the real observed values of anomalies are greater than 100. To avoid the differences among the imputation results based on various methods not obvious due to the large span of longitudinal, we denote the anomalies by the maximal value 40 and color them in purple in Figure 1. Moreover, in this dataset, we only label the observation "anomalous" or not manually (introduced in Section 6.1 in detail), i.e., the ground truths of anomalies are unknown. Thus, we do not show the ground truths of the anomalies in the figure. Since the ground truths of real MVs in the IDL dataset are unknown, we cannot evaluate the imputation results of various algorithms by directly using the real MVs. Moreover, there are no MVs in the adopted segment of data. Therefore, we assume that the MVs occur in the period from 241 to 250 (denoted by the minimal value 13.6 and colored in red). We implement and apply several state-of-the-art online MV imputation methods, including MUSCLES [35], SPIRIT [25], and TKCM [34] for demonstrating their deficiencies. As they are not designed to handle anomalies, we repair the anomalies by AR [16] model before applying these MV imputation methods.

As shown, existing temporal-dependency-based MV imputation (MUSCLES, SPIRIT, and TKCM) could not effectively impute MVs even though the anomalies are repaired before MV imputation. The reason is that the repaired anomalies (and imputed MVs) are used for later MV imputation (and anomaly repair), and thereby the inaccuracies of anomaly repair (and MV imputation) are propagated over the time series. Compared with temporal-dependency-based MV imputation, this case study demonstrates that the proposed REMAIN obtains the imputation results closest to the ground truths.

(3) **Deterioration detection.** In addition to observed data values, correlations among attributes of the data are also expected to evolve over time, especially in dynamic environments. In particular, at some time points, the change may be abrupt, causing the imputation model to deteriorate quickly. We term such time points as *deterioration points* (as illustrated in Example 2 and defined formally in Section 4). To achieve a satisfactory imputation result, the parameters of the imputation model need to be re-estimated at deterioration points. However, techniques for handling the deterioration points typically rely on the feedback of MV imputation performance which requires the ground truth of MVs to measure. Since the ground truths of MVs are unknown, the deterioration detection is challenging.

[1] http://db.csail.mit.edu/labdata/labdata.html.



(a) Observed values (b) Correlations

Fig. 2. The change of example data over time.

Example 2. To illustrate the dynamic nature of streaming data, based on the IDL dataset introduced in Example 1, Figure 2 (a) presents a selected segment of the observed Temperature and Humidity attributes of Sensor 1 over time during which there are no MVs and anomalies. As shown, the values of temperature and humidity are dynamically evolving over time. In addition, to illustrate the dynamic change of correlations amongst attributes of the data,[2] we plot the evolution of the correlation between attributes Temperature and Humidity in the IDL data, as shown in Figure 2(b). The evolution of the correlation is quite smooth for some periods, e.g., time points from 142 to 150. Under this scenario, it is natural to incrementally maintain the existing imputation model using newly arrival data. However, at some specific time points, the correlation changes abruptly, e.g., time points 134, 138, and 140. Note that the newly arriving data at these time points do not fit the existing model any more, i.e., at these time points (termed deterioration points and denoted by red dots in Figure 2(b)), the existing imputation model deteriorated. Thus, there is a need to re-estimate the parameters of the imputation model based only on the newly arriving data.

Since existing online MV imputation methods may not work well in imputing poor-quality streaming data (as illustrated in Example 1), we devise a *Real-time and Error-tolerant Missing vAlue ImputatioN* (REMAIN) framework. To the best of our knowledge, this is the first study on MV imputation for poor-quality streaming data. The proposed REMAIN imputes the MVs in poor-quality streaming data with polynomial time and constant space. Our major contributions are summarized as follows.

- —We formulate the MV imputation problem for poor-quality streaming data and propose a novel framework, namely, REMAIN, for real-time and error-tolerant MV imputation.
- —In REMAIN, we propose a-RANSAC (an accelerated variant of RANSAC[3]) to initialize (and re-estimate) the model parameters (in Section 3.2.1). Based on a-RANSAC, the efficiency of parameter initialization and parameter re-estimation in REMAIN is significantly better than using RANSAC. Moreover, we propose an incremental approach for updating parameters of the imputation model (in Section 3.3) to accommodate the streaming applications.
- —Considering the scenario where the correlations among attributes of the data change abruptly, we define the notion of *deterioration point*. Accordingly, we devise a deterioration detection mechanism (in Section 4) by estimating the variance of imputation error at each time point.
- —Finally, we conduct an extensive experimental evaluation using both real and synthetic datasets. The results demonstrate that the proposed REMAIN achieves significantly higher

[2]Note that the correlations among attributes are often exploited for MV imputation.
[3]RANSAC estimates the parameters of a given model from a dataset with presence of anomalies, which is not competent for streaming data.

Table 1. Notations

Notation	Definition
o_i^t	the observation of the ith object at time t
$\mathbf{x}_i^t$	the vector of values on d complete attribute in o_i^t
y_i^t	the value on the incomplete attribute in o_i^t
$\widehat{y}_i^t$	the prediction of y_i^t
O^t	the poor-quality dataset at time t
O_m^t	the incomplete set of O^t at time t
O_a^t	the anomalous set of O^t at time t
O_c^t	the consistent set of O^t at time t
δ^t	the anomaly ratio at time t
τ	the anomaly threshold
p	the probability that the estimated model is correct

imputation accuracy than existing works. Moreover, compared with the state-of-the-art existing MV imputation methods, REMAIN obtains up to one order of magnitude improvement in scalability.

2 PRELIMINARIES

In this section, we first introduce the problem of MV imputation for poor-quality streaming data. Next, we review some prior works relevant to our research.

2.1 Problem Statement

For ease of discussion on the research problem, we first formally introduce some terms below and summarize the notations used throughout the article in Table 1.

Definition 1. Given a set of objects $\{o_1, o_2, \ldots, o_n\}$, an observation, which consists of d attributes, is a data record describing an object at a time point, where o_i^t denotes the observation of ith object at time t.

In this article, an attribute containing MVs is termed as *incomplete attribute*, otherwise the attribute is a *complete attribute*. Given a dataset, assume there are d_m incomplete attributes and d_c complete attributes, i.e., $d_m + d_c = d$. Since we impute the MVs by exploring the correlations between d_m incomplete attributes and d_c complete attributes (will be introduced in detail later), the imputation for an incomplete attribute does not introduce violations in other incomplete attributes, i.e., the imputation for multiple incomplete attributes are independent. Therefore, for ease of illustration, we focus on imputation for one single incomplete attribute, i.e., $d_m = 1$, and denote an observation by $o_i^t = (y_i^t, \mathbf{x}_i^t)$, where y_i^t is the value on the incomplete attribute and $\mathbf{x}_i^t = (x_{i,1}^t, x_{i,2}^t, \ldots, x_{i,d_c}^t)$ is the vector of values on d_c complete attributes. It is worth noting that the proposed REMAIN is also capable of the imputation for multiple incomplete attributes. During the imputation for multiple incomplete attributes, the variant $y_i^t \in o_i^t$ becomes a vector $\mathbf{y}_i^t$, and the computation of parameter estimation becomes multiple matrix calculations from one matrix calculation.

Definition 2. The observation $o_i^t = (y_i^t, \mathbf{x}_i^t)$ is an incomplete observation if y_i^t is missing, otherwise o_i^t is a complete observation.

On the other hand, a complete observation is either an *anomalous observation* or a *consistent observation*, which is define as follows.

Definition 3. The observation $o_i^t = (y_i^t, \mathbf{x}_i^t)$ is an anomalous observation if (1) it is a complete observation and (2) the predicted value $\widehat{y}_i^t$ significantly differs from the original observed value y_i^t, i.e., $|\widehat{y}_i^t - y_i^t| > \tau$, where $\widehat{y}_i^t$ is computed by the MV imputation model (introduced in Section 3.2 in detail) and τ is a predefined anomaly threshold. Otherwise, o_i^t is a consistent observation.

The above definition of anomalous observation is widely adopted in existing anomaly detection works [3, 20, 27, 39]. At an abstract level, an anomaly is defined as a pattern that does not conform to expected normal behavior [9]. Given a set of observations, due to the difficulty of defining a *normal region* that encompasses possible normal behavior, most of existing works approximate the normal region by learning a (prediction) model which fits the majority observations in the data. If an observation can be fitted to the model (i.e., the difference between the predicted value (by the learned model) and the observed (real) value is smaller than a given threshold τ), it is determined as a normal observation. Otherwise, it is determined as an anomalous observation. In other words, the intuition behind the definition is that a farther distance between the observed value and its prediction indicates a higher probability of being an anomaly. Moreover, in this article, we assume that anomalies exist in incomplete attributes. The main reasons are two-fold: (1) In many real applications, noisy/erroneous values (e.g., MVs and anomalies) often occur simultaneously. For example, in applications of air monitoring and traffic jam detection, MVs and anomalies may both be introduced due to the power shortage of physical measurement sensors, mobile devices, outdoor and indoor cameras. Therefore, the existence of MVs indicates a high probability of anomalies occurrence, i.e., the anomalies are more likely to exist in the incomplete attributes with MVs simultaneously. (2) The data on the complete attributes are mainly used for learning an MV imputation model which fills the MVs by exploring the correlations between the incomplete attributes and complete attributes. For a complete attribute containing plenty of anomalies (even though the probability of this case occurrence in practice is low as analyzed above), if we do not handle (reject or repair) the anomalies on the complete attribute, the learned imputation model may be inaccurate. However, the anomaly detection and repair for the complete attribute incurs extra computation and time costs. To balance the MV imputation accuracy and efficiency, it is reasonable to reject such complete attribute with significant anomalies since our main goal is MV imputation rather than anomaly detection.

Definition 4. At time t, a poor-quality dataset $O^t = \{o_1^t, o_2^t, \ldots, o_n^t\}$ is a collection of observations which consists of three subsets: incomplete set O_m^t, anomalous set O_a^t and consistent set O_c^t, containing n_m incomplete observations, n_a anomalous observations and n_c consistent observations, respectively, such that $n_m + n_a + n_c = n$. Additionally, the anomaly ratio (i.e., the percentage of anomalous observations) at time t, denoted by δ^t, is computed as $\delta^t = \frac{|O_a^t|}{|O_a^t|+|O_c^t|} = \frac{n_a}{n_a+n_c}$.

It is notable that the anomalous set $O_a^t \subseteq O^t$ is non-deterministic because whether the observed values are anomalous or not is unknown and thus needs to be detected. Consequently, the consistent set O_c^t is also non-deterministic, and the anomaly ratio δ^t needs to be computed accordingly at each time point t. In addition, for each observation $o_i^t \in O^t$, the presence of MVs does not depend on other observations, i.e., data are missing at random.

Example 3. Table 2 shows an example segment of poor-quality streaming data sampled from IDL dataset (see Example 2). Each sensor is considered as an object and obtains the temperature and humidity of the lab continuously. For example, the observation of Sensor 1 at time t_0 is $(19.447, 39.525)$, corresponding to temperature and humidity. For illustration, in Table 2, "–" indicates MVs, and the numbers with red color indicate anomalies. Moreover, in IDL dataset, the MVs and anomalies are concentrated in the attribute Temperature. As shown, at time t_0, the temperature of Sensor 4 is missing, i.e., $O_m^{t_0} = \{o_4^{t_0}\}$, while the observed temperature of Sensor 3 is

Table 2. Example of Poor-quality Streaming Data

sensor id	observations			
	t_0	t_1	t_2	$\cdots$
1	(19.447, 39.525)	(12.346, 39.484)	(– , 39.710)	
2	(18.843, 40.917)	(19.518, 41.010)	(18.146, 41.207)	$\cdots$
3	(22.936, 45.276)	(15.215, 45.470)	(15.307, 47.296)	
4	(– , 43.585)	(– , 43.738)	(– , 44.212)	
5	(18.535, 39.622)	(24.387, 39.788)	(23.894, 40.181)	$\cdots$
6	(19.154, 39.377)	(17.743, 39.462)	(18.343, 39.507)	
7	(17.762, 42.856)	(18.392, 42.906)	(– , 42.998)	
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

anomalous, i.e., $O_a^{t_0} = \{o_3^{t_0}\}$. Accordingly, the consistent set at time t_0 is $O_c^{t_0} = \{o_1^{t_0}, o_2^{t_0}, o_5^{t_0}, o_6^{t_0}, o_7^{t_0}\}$, and the anomaly ratio at time t_0 is $\delta^{t_0} = \frac{1}{6}$.

In this article, we propose a framework, namely, REMAIN, for MV imputation of poor quality streaming data at time t ($t = 1, 2, \ldots$). Given a poor-quality dataset $O^t = \{o_1^t, o_2^t, \ldots, o_n^t\}$ as the input, REMAIN aims to tackle the following issues:

—Building an MV imputation model for poor-quality streaming data.
—Designing effective and efficient algorithms for parameter estimation, including parameter initialization and incremental parameter update, of the imputation model.
—Exploring a smart mechanism of deterioration detection that adapts quickly to the change of correlations amongst attributes of the data.
—Imputing the MVs contained in the dataset based on the learned imputation model in an online fashion.

Finally, REMAIN returns the imputed dataset O^{t*}.

2.2 Related Work

The idea of imputing MVs based on the intrinsic relationships in the underlying data is widely adopted in solving the MV imputation problem. The traditional MV imputation approaches can mainly be classified into nearest-neighbor-based imputation (NNI) [1, 42, 48], kernel-based imputation (KI) [43, 47] and regression-based imputation (RI) [21, 37, 44]. Given an incomplete observation, the NNI methods search neighbors of the incomplete observation and then take a distance-weighted mean of the k neighbors for imputation, e.g., GBKII [40], CMI [45], and OSICM [22]. KI methods employ various kernel functions/models [7] applicable to different data types to build imputation models, e.g., NIIA [43] and DIM [46]. RI methods predict the MV on an incomplete attribute by employing a regression model using the observed values on complete attributes in the same observation, e.g., ERACER [21] and IIM [38]. However, all algorithms introduced above do not consider the negative effect of anomalies. Moreover, almost all of them are proposed to work on static datasets and inefficient for dynamic streaming data.

For online MV imputation, MUSCLES [35], SPIRIT [25], and TKCM [34] have been proposed. They impute the MVs based on the temporal dependencies of the data, where the MVs at time t are estimated based on the segment of historical data in a predefined time window. However, since the intuitions of MUSCLES and SPIRIT for MV imputation are similar to that of AR model, they suffer the same anomaly repair problem introduced in Section 1. TKCM cannot also avoid the impact of anomalies during MV imputation, as the MVs in a time series s are imputed by the historical values

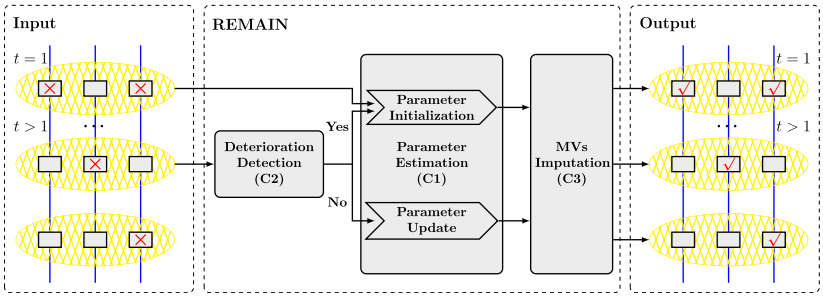


Fig. 3. An overview of REMAIN.

of s at the anchor points of k most similar patterns which are detected based on a set of reference time series of s. If there is no anomaly in the reference time series but unfortunately the historical values at anchor points in s (used for MV imputation) contain anomalies, the imputation result may be inaccurate. On the other hand, MUSCLES does not scale well to a large number of objects since it requires at least quadratic space and time. Even though the SPIRIT requires less memory and time than MUSCLES by compressing the original data based on Principle Component Analysis (PCA), its space complexity grows linearly with respect to the number of objects. The newest online MV imputation method, TKCM, is proposed with the assumption that time series often exhibit repeating patterns, which limits its applications. Additionally, a few prediction models over time series based on neural networks are proposed for forecasting, e.g., [36] based on Deep LSTM. However, they are not directly applicable to MV imputation for poor-quality streaming data due to the big data samples requirement and large time occupation for model training.

3 REMAIN FRAMEWORK AND PARAMETER ESTIMATION

In this section, we first give an overview of the proposed REMAIN framework. Then we introduce the parameter estimation which is the first component of the proposed REMAIN.

3.1 REMAIN– an Overview

First of all, in REMAIN, we adopt the *multi-variates linear regression* (MLR) as the basic MV imputation model for the following three reasons: (1) MLR model is very powerful and easy to interpret, which can even approximate the non-linear correlations amongst attributes [41]. (2) MLR model is highly efficient. Its parameters can be computed based on the Least Square method easily. (3) MLR model has good scalability and does not suffer the inaccuracy propagation. It does not require to buffer any historical data. Moreover, for a new incomplete observation, it can be imputed directly based on the learned MLR model without other computation.

Next, as shown in Figure 3, corresponding to the core issues to be tackled introduced earlier, the proposed REMAIN consists of three components:

(C1) Parameter estimation learns the parameters of a given MLR model effectively and efficiently. First, the model parameters are initialized based on the poor-quality dataset O^t at the first time point (i.e., $t = 1$) by employing a-RANSAC (*Parameter Initialization*). Afterwards, the model parameters are incrementally updated at each time point to accommodate the streaming applications (*Parameter Update*).

(C2) Deterioration detection determines if the current time t is a deterioration point. Since the correlations among attributes of the data change in unforeseen ways, the imputation result based on the (incrementally) updated model may deteriorate. When deterioration occurs, the data arriving at the current time do not fit the existing imputation model anymore. Therefore, the model

parameters need to be re-estimated just based on the poor-quality dataset O^t (via the same procedure of parameter initialization, i.e., resetting $t = 1$ basically).

(C3) MV imputation fills the MVs in the poor-quality dataset O^t based on the learned model and derives the imputed dataset O^{t*}. Moreover, the anomaly ratio δ^t at time t ($t > 1$) is effectively estimated simultaneously.

Algorithm 1 provides the pseudo-code of REMAIN. Given a poor-quality dataset O^t at time t ($t = 1, 2, \ldots$), if $t = 1$ (i.e., at the first time point), we initialize the model parameters based on a-RANSAC (Lines 4-7). If $t > 1$, we first determine whether it is a deterioration point (Line 10). If a deterioration point is detected, the model parameters are re-estimated where the procedure is the same as parameter initialization (Line 13). Otherwise, the model parameters are expected to update incrementally (Line 16). Finally, the incomplete observations in O^t are imputed based on the learned model (Line 18). Moreover, δ_{max} and δ_{min} used for deterioration detection and parameter initialization are expected to update based on the estimated anomaly ratio δ^t ($t > 1$) (Lines 19–22). It is notable that the proposed REMAIN supports both numerical and categorical data. For the categorical data, we adopt the multivariate logistic regression model as the basic imputation model which is similar to the MLR model. Additionally, in some specific applications, the number of observations arriving at a time point may be small, e.g., 54 observations from 54 sensors at each time point in IDL dataset. Under this scenario, we can collect observations from several successive time points as a logical time point t. With sufficient complete observations at a logical time point, the parameters of the imputation model are initialized/updated effectively. Next, we will introduce each component in detail in the following sections.

3.2 Parameter Estimation

As introduced earlier, we adopt the MLR model to impute MVs in poor-quality streaming data. For an incomplete observation $o_i^t = (y_i^t, \mathbf{x}_i^t)$, the MV y_i^t is estimated as a linear combination of the values in $\mathbf{x}_i^t = (x_{i,1}^t, x_{i,2}^t, \ldots, x_{i,d_c}^t)$ where $x_{i,j}^t$ is the observed value on the j-th complete attribute of o_i^t. The prediction function is shown below:

$$\widehat{y}_i^t = w_0^t + \sum_{j=1}^{d_c} w_j^t x_{i,j}^t + \varepsilon_i^t, \tag{1}$$

where $\widehat{y}_i^t$ is the prediction of y_i^t, $\mathbf{w}^t = (w_0^t, w_1^t, \ldots, w_{d_c}^t)$ is the *parameter vector* consists of all parameters in MLR model and ε_i^t is a white noise generated according to the Gaussian distribution with mean 0 and variance $(\sigma^t)^2$. If y_i^t is missing, $\widehat{y}_i^t$ is the imputation result of y_i^t. If y_i^t is observed but significantly differs from its prediction $\widehat{y}_i^t$, i.e., $|y_i^t - \widehat{y}_i^t| > \tau$, then y_i^t is considered as an anomaly.

Intuitively, the learning process of an MLR model is to determine its parameter vector $\mathbf{w}^t$. Given a poor-quality dataset O^t at time t, suppose there are n_c consistent observations in $O_c^t \subseteq O^t$. We assume that all the consistent observations at the same time point are well captured by the MLR model. Thus, based on Yule-Walker equation [12], $\mathbf{w}^t$ can be estimated from the observations as follows:

$$\mathbf{w}^t = ((\mathbf{X}^t)^T(\mathbf{X}^t))^{-1}((\mathbf{X}^t)^T\mathbf{y}^t), \tag{2}$$

where

$$\mathbf{X}^t = \begin{bmatrix} 1, \mathbf{x}_1^t \\ 1, \mathbf{x}_2^t \\ \cdots \\ 1, \mathbf{x}_{n_c}^t \end{bmatrix} = \begin{bmatrix} 1, x_{1,1}^t, x_{1,2}^t, \ldots, x_{1,d_c}^t \\ 1, x_{2,1}^t, x_{2,2}^t, \ldots, x_{2,d_c}^t \\ \cdots \\ 1, x_{n_c,1}^t, x_{n_c,2}^t, \ldots, x_{n_c,d_c}^t \end{bmatrix} \quad \mathbf{y}^t = \begin{bmatrix} y_1^t \\ y_2^t \\ \cdots \\ y_{n_c}^t \end{bmatrix}.$$

ALGORITHM 1: REMAIN(O^t, δ^1, p, s, τ)

Input: the poor-quality dataset O^t at time t ($t = 1, 2, \ldots$),
the anomaly ratio δ^1 at the first time point, i.e., $t = 1$,
the probability p,
the minimal number of observations required to estimate model parameters s,
and the anomaly threshold τ
Output: the imputed dataset O^{t*}

```
1  w^{t-1} = φ, G^{t-1} = φ
2  δ_max = δ^1, δ_min = δ^1
3  foreach time point t do
4      if t == 1 then
5          //parameter initialization
6          w^1, G^1 = ParameterInitialization(O^1, δ_max, τ, p, s)
7          w^{t-1} = w^1, G^{t-1} = G^1
8      else
9          // deterioration detection, de is a boolean variate
10         de = DeteriorationDection(O^t, w^{t-1}, δ_max, δ_min, τ, p)
11         if de then
12             //re-estimate the parameters of the imputation model
13             w^t, G^t = ParameterInitialization(O^t, δ_max, τ, p, s)
14         else
15             // incrementally update the parameters of the model
16             w^t, G^t = ParameterUpdate(O^t, w^{t-1}, G^{t-1}, τ)
17         w^{t-1} = w^t, G^{t-1} = G^t
18     O^{t*}, δ^t = MVImputation(O^t, w^t, τ)
19     if δ^t > δ_max then
20         δ_max = δ^t
21     if δ^t < δ_min then
22         δ_min = δ^t
23 return O^{t*} (t = 1, 2, ...)
```

In matrix $\mathbf{X}^t$, the 0th column is a constant vector with element 1, which is corresponding to the *intercept* w_0^t. The jth ($1 \le j \le d_c$) column is the vector of values $x_{i,j}^t (1 \le i \le n_c)$ on the jth complete attribute of observations in O_c^t, which is corresponding to the parameter w_j^t. Accordingly, $\mathbf{y}^t$ is a vector of desired values $y_i^t (1 \le i \le n_c)$ of observations in O_c^t.

3.2.1 Parameter Initialization. As introduced in the Introduction, the anomalies embedded in the data may incur inaccurate MLR model used for MV imputation. One possible solution is to employ the RANSAC (RANdom SAmple Consensus) [11] paradigm which is able to estimate the parameters of a given model from a dataset with outliers. Here, we briefly review the traditional RANSAC paradigm and analyze its pitfalls in handling streaming data.

RANSAC paradigm. Generally, RANSAC separates the observations of a dataset from *inliers* (observations explainable by the model) and *outliers* (observations not explainable by the model), assuming that the model parameters can be optimally estimated by a (usually small) set of inliers. Given a poor-quality dataset O^t, RANSAC estimates the parameters of a model by the following steps:

1. Selecting s complete observations from O^t randomly, where s is the minimal number of observations to determine the model parameters. The selected observations compose an initial consistent set $O_c^t = \{o_1^t, o_2^t, \ldots, o_s^t\}$.

2. Estimating the model parameters based on O_c^t.

3. For $\forall o_i^t \in O^t \wedge o_i^t \notin \{O_c^t \cup O_m^t\}$, it is considered as an outlier (anomalous observation) and added into O_a^t if $|y_i^t - \widehat{y}_i^t| > \tau$, where the prediction $\widehat{y}_i^t$ of y_i^t is computed based on the estimated model in Step 2. Otherwise, the observation o_i^t is considered as an inlier (consistent observation) and added into O_c^t.

4. The estimated model is reasonably correct if there are sufficient inliers, i.e., $\frac{|O_c^t|}{|O_a^t|+|O_c^t|} > C(1 - \delta^t)$ where C $(0 \leq C \leq 1)$ is a constant.

5. Re-estimating the model by using all members of the consistent set O_c^t if the estimated model in Step 4 is accepted as a correct model.

6. Repeating the above five steps for k times. The model estimated in each iteration is either rejected because the size of consistent set O_c^t is too small, or be saved if the size of O_c^t is larger than the previously saved model.

The anomalous observations and consistent observations in the dataset are considered as outliers and inliers in RANSAC, respectively. To estimate the parameters of the MLR model based on the Least Square method, let $s = d_c + 2$ because $d_c + 1$ parameters need to be learned. Given a certain probability p that the correct model can be obtained, the number of iteration k can be computed as $k = \frac{log(1-p)}{log(1-q^s)}$, where q is the probability that a selected observation o_i^t in initial consistent set O_c^t is an inlier, i.e., $q = 1 - \delta^t$. In practice, the number of iteration k is set two or three times of the expected (theoretical) number to obtain a reasonable parameter estimation [11]. In this article, we set k as two times of the expected number to trade off the effectiveness and efficiency. Moreover, we set $C(1 - \delta^t) = 1 - \delta_{max}$ to determine if the estimated model is correct (Step 4). The reason is that with a relative small ratio threshold, i.e., $C(1 - \delta^t)$, more estimated models have chances to be improved by re-estimation using all observations of the consensus set, and thereby it is more likely to find the correct model.

Parameter Initialization based on a-RANSAC. As introduced above, different from classical parameter estimation techniques which estimate the parameters of a model by optimizing the fit of the model to *all* of the observations, RANSAC has an internal mechanism for detecting and rejecting outliers. Thus it is capable of estimating the model parameters for poor-quality dataset with a high degree of accuracy. However, RANSAC is not suitable for streaming data due to the iterative estimation of model parameters.

Given a poor-quality dataset O^t at a certain time point, we have the following proposition for time and space complexity of RANSAC.

Proposition 1. *Given a poor-quality dataset O^t at time t, the parameter vector of the MLR model is estimated based on RANSAC in $O(kn_cd^2 + kd^3)$ time with $O(d^2)$ space, where n_c is the size of consistent set $O_c^t \subseteq O^t$, k is the iteration number of RANSAC, and d is the number of attributes in an observation $o_i^t \in O_c^t$.*

Proof. We need $O(d_c + 1)$ space to store the parameter vector $\mathbf{w}^t$ and $O((d_c + 1)^2)$ space to store the intermediate result $\mathbf{G}_i^t$ (to be introduced later) used for incremental parameter update. Since $\mathbf{X}^t$ and $\mathbf{y}^t$ are composed of the consistent observations at time t, rather than the historical observations, there is no extra space cost for $\mathbf{X}^t$ and $\mathbf{y}^t$. On the other hand, the computation costs for matrix multiplication and matrix inversion in Equation (2) are $O(n_c(d_c + 1)^2)$ and $O((d_c + 1)^3)$, respectively. Additionally, the parameter estimation is repeated k times, and thus the parameter initialization based on RANSAC runs in $O(k(n_cd^2 + d^3))$ time. □

Note that based on Definition 1, we have $d = d_c + d_m$. Moreover, in many real-world datasets, the number of attributes d in an observation is usually significantly smaller than the datasize n, i.e., $d \ll n$. Thus we consider d as a constant.[4] For conciseness of expression, we use d to replace d_c in space and time complexity analysis throughout the article unless noted specially. Based on Proposition 1, RANSAC is time costly when the dataset is large and/or the computation requires many iterations. To tackle this issue, we propose a variant of RANSAC, named a-RANSAC (*accelerated RANSAC*), to trade off the effectiveness and efficiency.

In Proposition 1, the time cost of parameter initialization based on RANSAC is mainly affected by (i) the number of iterations k and (ii) the size of consistent set n_c. To obtain a near-optimal parameter estimation with probability p correctness guarantee, the number of iteration k could not be easily reduced. Therefore, in a-RANSAC, our intuition is to improve the efficiency of RANSAC by reducing the size of consistent set n_c. In detail, given a poor-quality dataset O^t at time t, a-RANSAC first selects a subset of complete observations $\widehat{O}^t_{c'}$ from the set of complete observations $O^t_{c'} = O^t_a \bigcup O^t_c$ such that the distribution of anomalous observations in $\widehat{O}^t_{c'}$ is close to that of $O^t_{c'}$. Next, a-RANSAC estimates the parameters of a given model based on RANSAC by using $\widehat{O}^t_{c'}$.

The rationale behind the algorithm is that based on $\widehat{O}^t_{c'}$, we can also obtain a correct MLR model with probability p. Since the distribution of anomalous observation in $\widehat{O}^t_{c'}$ follows that in $O^t_{c'}$, it is reasonable to assume that the anomaly ratio of $\widehat{O}^t_{c'}$ is also δ^t. Based on the computation for iteration k discussed earlier, a correct MLR model with probability p can be obtained when $k \geq \frac{log(1-p)}{log(1-q^s)}$. Therefore, if there are enough complete observations to support k times non-repetitive initial consistent set selection (Step 1 of RANSAC), the learned MLR model based on $\widehat{O}^t_{c'}$ is correct with probability p. Next, we discuss how to determine the size of $\widehat{O}^t_{c'}$.

Let the size of $\widehat{O}^t_{c'}$ be $\widehat{n}_{c'} = n_{c'} + n_{a'}$ where $n_{c'}$ and $n_{a'}$ are the numbers of consistent observations and anomalous observations in $\widehat{O}^t_{c'}$, respectively. Then there are $\binom{\widehat{n}_{c'}}{s}$ possible initial consistent sets. As discussed above, $\binom{\widehat{n}_{c'}}{s}$ should satisfy $\binom{\widehat{n}_{c'}}{s} \geq k$. Based on the definition of *binomial coefficient* [24], we have

$$\prod_{i=0}^{s-1} (\widehat{n}_{c'} - i) \geq ks!, \tag{3}$$

where k and s are known. Here we aim to estimate $\widehat{n}_{c'}$. Intuitively, the exact solution for Equation (3) is difficult to obtain. We need to try $\widehat{n}_{c'}$ from 1 to $ks!$ to determine the minimal $\widehat{n}_{c'}$. To find the solution of Equation (3), we simplify Equation (3) as Equation (4):

$$\prod_{i=0}^{s-1} (\widehat{n}_{c'} - i) \geq \left(\widehat{n}^t_{c'} - s\right)^s \geq ks!. \tag{4}$$

Based on Equation (4), we have $\widehat{n}_{c'} \geq \sqrt[s]{ks!} + s$. In addition, after finding a reasonable MLR model based on the initial consistent set (Step 4 of RANSAC), the RANSAC re-tunes the model parameters based on all of the consistent observations by using least square estimation (Step 5 of RANSAC). Referring to the central limit theorem, the number of samples used empirically for least square estimation should be at least 30 or greater [18]. Thus we propose that the size of subset is $\widehat{n}_{c'} = \max\{\lceil \sqrt[s]{ks!} \rceil + s, 30\}$.

In addition, to derive a subset $\widehat{O}^t_{c'} \subseteq O^t_{c'}$ such that the distribution of anomalous observations is close to that in $O^t_{c'}$, we sample the complete observations from $O^t_{c'}$ based on the statistical distribu-

[4] As a promising future direction, it is interesting to extend the REMAIN to high-dimensional streaming data.

tion of anomalous observations at historical time points. For example, in Wireless Sensor Networks (WSNs), the anomalous observations may be generated by some sensors due to the little energy remaining. Under this scenario, the distribution of anomalous observations may be stable for a period of time. Thus we can derive $\widehat{O}^t_{c'}$ by stratified sampling based on the historical information.

Finally, Algorithm 2 shows the pseudo-code of the parameter initialization. Given a poor-quality dataset O^t at time t ($t = 1$), we first employ a-RANSAC to derive the parameter vector $\mathbf{w}^t$ with the correctness probability p (Lines 2-9). Next, the parameter vector $\mathbf{w}^t$ is incrementally updated based on the parameter update algorithm (to be introduced in Section 3.3) by using the remaining complete observations $\widetilde{O}^t_{c'} = O^t_{c'} \setminus \widehat{O}^t_{c'}$. It is notable that to solve the cold start problem (i.e., there is no available historical data), the anomaly ratio at the first time point, i.e., δ^1, requires to be provided to initialize the δ_{max} and δ_{min}. Generally, δ^1 can be given by domain experts or decided by observing the distribution of the data arriving at the first time point, i.e., O^1, based on existing visual identification approaches [8].

ALGORITHM 2: ParameterInitialization(O^t, δ_{max}, p, s, τ)

Input: the poor-quality dataset O^t,
the maximum anomaly ratio δ_{max},
the probability p,
and the minimal number of observations required to estimate model parameters s,
and the anomaly threshold τ
Output: the parameter matrix $\mathbf{W}^t$, the intermediate result for parameter update $\mathbf{G}^t$

1 $O^t_{c'} = O^t \setminus O^t_m$, $w = 1 - \delta^t$
2 // calculate the number of iteration k
3 $k = 2 \times \frac{log(1-p)}{log(1-w^s)}$
4 // determine the size of $\widehat{O}^t_{c'}$
5 $\widehat{n}_{c'} = \max\{\lceil \sqrt[s]{ks!} \rceil + s, 30\}$
6 // select $\widehat{n}_{c'}$ complete observations from $O^t_{c'}$
7 $\widehat{O}^t_{c'} = \left\{o^t_1, o^t_2, \ldots, o^t_{\widehat{n}_{c'}}\right\}$
8 // parameter estimation based on $\widehat{O}^t_{c'}$
9 $\mathbf{W}^t, \mathbf{G}^t$ = **RANSAC**($\widehat{O}^t_{c'}$, δ^t, p, s)
10 // parameter update by using the remaining complete observations
11 $\mathbf{W}^{t-1} = \mathbf{W}^t$, $\mathbf{G}^{t-1} = \mathbf{G}^t$
12 $\widetilde{O}^t_{c'} = O^t_{c'} \setminus \widehat{O}^t_{c'}$
13 $\mathbf{W}^t, \mathbf{G}^t$ = **ParameterUpdate**($\widetilde{O}^t_{c'}$, $\mathbf{W}^{t-1}$, $\mathbf{G}^{t-1}$, τ)
14 **return** $\mathbf{W}^t, \mathbf{G}^t$

Proposition 2. *Given a poor-quality stream dataset O^t at time t, for parameter initialization based on a-RANSAC, the space complexity is $O(d^2)$ and the time complexity is $O(k(n_{c'}d^2 + d^3) + n_c d^2)$.*

Proof. Referring to the Proposition 1, the time complexity of parameter estimation based on a-RANSAC by using $\widehat{O}^t_{c'}$ is $O(n_{c'}kd^2 + kd^3)$. The time complexity of parameter update by using the remaining complete observations is $O((n_c - n_{c'})d^2)$ (referring to the Proposition 5 to be introduced in Section 3.3). □

Compared with the time complexity of parameter initialization based on RANSAC, i.e., $O(n_c kd^2 + kd^3)$ (see Proposition 1), the time complexity of parameter initialization based on

a-RANSAC is linear with the increase of both k and n_c, as the size of consistent observations in $\widehat{O}_{c'}^t$, i.e., $n_{c'}$, is usually much smaller than n_c. Moreover, the proposed parameter initialization algorithm (based on a-RANSAC) improves efficiency without losing much imputation accuracy (as shown in Section 6).

3.3 Parameter Update

Note that the algorithm of parameter initialization introduced above only utilizes the poor-quality dataset at one time point. Statistically, the MLR model is more accurate if there are more data samples (i.e., consistent observations) to be used for model learning. Thus existing MV imputation approaches mostly learn the imputation model using the entire data (i.e., in batch processing). At time t ($t > 1$), if the model parameters are also estimated based on a-RANSAC using the entire dataset (the collection of data from time 1 to t), the size of the matrix $\mathbf{X}^t$ is increased to $tn_c \times d$. Thus the time complexity of parameter estimation at time t is $O(tkn_{c'}d^2 + kd^3 + tn_cd^2)$. Since t is not fixed and can grow indefinitely, parameter estimation based on a-RANSAC using the entire data is not competent for streaming data as time passes.

Based on the above analysis, online incremental parameter update is necessary. For streaming data, the observations arrive sequentially and continuously. Let $\mathcal{X}_i^t$ and $\mathcal{Y}_i^t$ be the sample matrixes constituted by all of the consistent observations having arrived thus far (including the consistent observations that arrived at historical time points $t = 1, 2, \ldots, t-1$).

Proposition 3. *Let* $\mathbf{G}_i^t = ((\mathcal{X}_i^t)^T(\mathcal{X}_i^t))^{-1}$ *be an intermediate variable. Then* $\mathbf{G}_i^t$ *could be recursively computed from* $\mathbf{G}_{i-1}^t$ *as follows:*

$$\mathbf{G}_i^t = \mathbf{G}_{i-1}^t - \left(\mathbf{G}_{i-1}^t \left(\mathbf{x}_i^t\right)^T\right)\left(\mathbf{x}_i^t\mathbf{G}_{i-1}^t\left(\mathbf{x}_i^t\right)^T + \mathbf{I}\right)^{-1}\left(\mathbf{x}_i^t\mathbf{G}_{i-1}^t\right). \tag{5}$$

Proof. Based on matrix inverse lemma [30], we have

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1}\right)^{-1}\mathbf{DA}^{-1},$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are the invertible matrices. On the other hand, $\mathbf{G}_i^t$ can be denoted as follows:

$$\begin{aligned}\mathbf{G}_i^t &= \left(\left(\mathcal{X}_i^t\right)^T\left(\mathcal{X}_i^t\right)\right)^{-1}\\ &= \left(\left(\mathcal{X}_{i-1}^t\right)^T\left(\mathcal{X}_{i-1}^t\right) + \left(\mathbf{x}_i^t\right)^T\left(\mathbf{x}_i^t\right)\right)^{-1}\\ &= \left(\left(\mathbf{G}_{i-1}^t\right)^{-1} + \left(\mathbf{x}_i^t\right)^T\mathbf{I}\left(\mathbf{x}_i^t\right)\right)^{-1}.\end{aligned}$$

Accordingly, let $\mathbf{A} = \left(\mathbf{G}_{i-1}^t\right)^{-1}$, $\mathbf{B} = \left(\mathbf{x}_i^t\right)^T$, $\mathbf{C} = \mathbf{I}$ and $\mathbf{D} = \mathbf{x}_i^t$, the conclusion is proved. □

Proposition 4. *Based on* $\mathbf{G}_i^t$, *the parameter* $\mathbf{w}_i^t$ *could be recursively computed from* $\mathbf{w}_{i-1}^t$, *as shown in Equation (6) below:*

$$\mathbf{w}_i^t = \mathbf{w}_{i-1}^t - \mathbf{G}_i^t\left(\mathbf{x}_i^t\right)^T\mathbf{x}_i^t\mathbf{w}_{i-1}^t + \mathbf{G}_i^t\left(\mathbf{x}_i^t\right)^T y_i^t. \tag{6}$$

Proof. Based on the Yule-Walker equation, we have

$$\begin{aligned}\mathbf{w}_i^t &= \left(\left(\mathcal{X}_i^t\right)^T\left(\mathcal{X}_i^t\right)\right)^{-1}\left(\left(\mathcal{X}_i^t\right)^T\mathcal{Y}_i^t\right)\\ &= \mathbf{G}_i^t\left[\left(\mathcal{X}_{i-1}^t\right)^T\mathcal{Y}_{i-1}^t + \left(\mathbf{x}_i^t\right)^T y_i^t\right]\\ &= \mathbf{G}_i^t\left(\mathcal{X}_{i-1}^t\right)^T\mathcal{Y}_{i-1}^t + \mathbf{G}_i^t\left(\mathbf{x}_i^t\right)^T y_i^t.\end{aligned}$$

Furthermore, combining $(\mathcal{X}_{i-1}^t)^T \mathcal{Y}_{i-1}^t = (\mathbf{G}_{i-1}^t)^{-1}\mathbf{w}_{i-1}^t$ and $(\mathbf{G}_{i-1}^t)^{-1} = (\mathcal{X}_{i-1}^t)^T \mathcal{X}_{i-1}^t$ into above equation, we have

$$\begin{aligned}\mathbf{w}_i^t &= \mathbf{G}_i^t \left[\left(\mathcal{X}_{i-1}^t\right)^T \mathcal{X}_{i-1}^t \right] \mathbf{w}_{i-1}^t + \mathbf{G}_i^t \left(\mathbf{x}_i^t\right)^T y_i^t \\ &= \mathbf{G}_i^t \left[\left(\mathcal{X}_i^t\right)^T \mathcal{X}_i^t - \left(\mathbf{x}_i^t\right)^T \mathbf{x}_i^t \right] \mathbf{w}_{i-1}^t + \mathbf{G}_i^t \left(\mathbf{x}_i^t\right)^T y_i^t \\ &= \mathbf{G}_i^t \left(\left(\mathcal{X}_i^t\right)^T \mathcal{X}_i^t \right) \mathbf{w}^{t-1} - \mathbf{G}_i^t \left(\mathbf{x}_i^t\right)^T \mathbf{x}_i^t \mathbf{w}_{i-1}^t + \mathbf{G}_i^t \left(\mathbf{x}_i^t\right)^T y_i^t.\end{aligned}$$

Since $\mathbf{G}_i^t((\mathcal{X}_i^t)^T \mathcal{X}_i^t) = \mathbf{I}$, the conclusion is proved. □

Proposition 5. *At time t, the space and time complexity of the parameter update based on Proposition 4 is $O(d^2)$ and $O(n_c d^2)$, respectively.*

Proof. To compute $\mathbf{W}_i^t$ incrementally, we need to store the matrix $\mathbf{G}_i^t$ of size $(d_c + 1) \times (d_c + 1)$. Thus the space complexity is $O(d^2)$. Since the size of matrix $(\mathbf{x}_i^t \mathbf{G}_{i-1}^t (\mathbf{x}_i^t)^T + \mathbf{I})$ in Equation (5) becomes a scalar, the time complexity of matrix inversion is $O(1)$. Accordingly, the time complexity of matrix multiplication in Equations (5) and (6) is $O(d^2)$. Since the parameters are updated based on proposition 4 when each consistent observation in O^t arrives, the time complexity of parameter update at time t is $O(n_c d^2)$. □

Algorithm 3 shows the pseudo-code of parameter update. Given a poor-quality dataset O^t at time t ($t > 1$), for each observation $o_i^t \in O^t$, it is omitted for the parameter update if $y_i^t \in o_i^t$ is missing (i.e., o_i^t is an incomplete observation) or $|y_i^t - \widehat{y}_i^t| > \tau$ (i.e., o_i^t is an anomalous observation) (Lines 4–9). Note that the predicted value $\widehat{y}_i^t$ is computed based on the learned model at time $t - 1$. If $o_i^t \in O^t$ is considered as a consistent observation, it is used for refining the parameter vector $\mathbf{w}_i^t$ and the intermediate matrix $\mathbf{G}_i^t$ (Lines 10–12). Finally, the updated $\mathbf{w}^t$ and $\mathbf{G}^t$ are returned.

4 DETERIORATION DETECTION

As introduced in the Introduction, the correlations among attributes of the data evolve over time in unforeseen ways, which may yield the phenomenon of *concept drift* [13]. The concept drift means that the correlations between the input variables (complete attributes) and the target variables (incomplete attributes) change over time. For the poor-quality streaming data, we target on in the article, both the existence of anomalies and the smooth/abrupt changes of streaming data may cause the concept drift. However, from the MV imputation perspective, we only concern the concept drift resulting in unsatisfactory imputation accuracy. By detecting and rejecting anomalies to learn the MV imputation model, the concept drift caused by anomalies does not affect the imputation accuracy significantly. Additionally, by updating the imputation model based on the newly arriving data incrementally, the concept drift caused by the smooth change of the correlations amongst attributes does also not introduce the degradation of the imputation accuracy. In this article, we mainly concern the concept drift caused by the abrupt change of the correlations among attributes of the data, since the imputation accuracy may deteriorate quickly if such concept drifts are not detected as early as possible. We specifically define the time points when such concept drifts occur as the *deterioration points*. Since most existing concept drift detection/fault diagnosis approaches [6, 13, 33] are proposed for complete data, and do not aims at effective MV imputation, we specifically devise an effective deterioration detection mechanism to achieve high imputation accuracy.

Given a poor-quality dataset O^t at time t, let $e^t = \{e_1^t, e_2^t, \ldots, e_{n_m}^t\}$ be the MV imputation errors of the incomplete set $O_m^t \subseteq O^t$.

ALGORITHM 3: ParameterUpdate(O^t, $\mathbf{w}^{t-1}$, $\mathbf{G}^{t-1}$, τ)

Input: the poor-quality dataset O^t at time t ($t > 1$),
the parameter vector $\mathbf{w}^{t-1}$ derived at last time point,
the intermediate result $\mathbf{G}^{t-1}$ at last time point,
and the anomaly threshold τ
Output: the updated parameter vector $\mathbf{w}^t$, the intermediate result for parameter update $\mathbf{G}^t$

$\mathbf{w}_{i-1}^t = \mathbf{w}^{t-1}, \mathbf{G}_{i-1}^t = \mathbf{G}^{t-1}$
foreach *observation* $o_i^t \in O^t$ **do**
  // o_i^t is an incomplete observation
  if y_i^t *is missing* **then**
    Continue
  $\widehat{y}_i^t = w_0^{t-1} + \sum_{j=1}^{d} w_j^{t-1} x_{i,j}^t + \varepsilon_i^t$
  // o_i^t is an anomalous observation
  if $\left|y_i^t - \widehat{y}_i^t\right| > \tau$ **then**
    Continue
  $\mathbf{G}_i^t = \mathbf{G}_{i-1}^t - \left(\mathbf{G}_{i-1}^t \left(\mathbf{x}_i^t\right)^T\right)\left(\mathbf{x}_i^t \mathbf{G}_{i-1}^t \left(\mathbf{x}_i^t\right)^T + \mathbf{I}\right)^{-1}\left(\mathbf{x}_i^t \mathbf{G}_{i-1}^t\right)$
  $\mathbf{w}_i^t = \mathbf{w}_{i-1}^t - \mathbf{G}_i^t \left(\mathbf{x}_i^t\right)^T \mathbf{x}_i^t \mathbf{w}_{i-1}^t + \mathbf{G}_i^t \left(\mathbf{x}_i^t\right)^T y_i^t$
  $\mathbf{G}_{i-1}^t = \mathbf{G}_i^t, \mathbf{w}_{i-1}^t = \mathbf{w}_i^t$
$\mathbf{w}^t = \mathbf{w}_i^t, \mathbf{G}^t = \mathbf{G}_i^t$
return $\mathbf{w}^t, \mathbf{G}^t$

Definition 5. Time point t is a deterioration point if the mean absolute deviation of e^t, denoted by $MAD(e^t)$, satisfies $MAD(e^t) > \epsilon$, where ϵ is an error tolerance threshold. Specifically, $MAD(e^t) = \frac{1}{n_m} \sum_{o_i^t \in O_m^t} |\widetilde{y}_i^t - \widehat{y}_i^t|$, where n_m is the size of O_m^t, $\widetilde{y}_i^t$ and $\widehat{y}_i^t$ are the ground truth and prediction of y_i^t respectively.

Note that the estimation of imputation error e^t is challenging because the ground truths of MVs are unknown. Hence, we cannot compute e^t directly. In this article, we explore the variance of imputation errors and adopt the Gaussian distribution which is widely used in many fields to capture the imputation errors. Suppose the imputation errors follow Gaussian distribution, i.e., $e^t \sim N(0, (\sigma^t)^2)$, where the mean is 0 as we believe that the methods do not make errors intentionally and the reliability degree of the imputation results is reflected by the variance $(\sigma^t)^2$. If the imputation results of O^t are unreliable, the distribution of imputation errors e^t has a wide spectrum and vice versa. Thus, we rewrite the definition of deterioration point in terms of variance as follows.

Definition 6. Time point t is a deterioration point if the variance $(\sigma^t)^2$ of imputation errors e^t satisfies $(\sigma^t)^2 > \varsigma^2$, where ς^2 is a variance threshold.

Mathematically, given a set of complete observations O_f^t (with size n_f) which fit the imputation model, since the mean of imputation errors is supposed to be 0, the maximum likelihood (ML) estimation of the variance $(\sigma^t)^2$ can be computed below:

$$\left(\sigma^t\right)^2 = \frac{1}{n_f}\left(\mathbf{y}_f^t - \mathbf{X}_f^t \mathbf{w}^{t-1}\right)^T \left(\mathbf{y}_f^t - \mathbf{X}_f^t \mathbf{w}^{t-1}\right). \tag{7}$$

As introduced in Section 3.2.1, the learned near-optimal MLR model is correct with probability p, i.e., there is $1 - p$ probability that the learned MLR model is incorrect. For the imputation error

estimation problem, given a poor-quality dataset O^t, we consider the extreme scenario where the consistent observations in $O_c^t \subseteq O^t$ fit the correct model well, while the anomalous observations in $O_a^t \subseteq O^t$ fit the incorrect model well. Let $(\sigma_c^t)^2$ and $(\sigma_{ic}^t)^2$ be the variances of imputation errors e^t based on correct and incorrect MLR model, respectively. Then the variance $(\sigma^t)^2$ of e^t is estimated as follows:

$$\left(\sigma^t\right)^2 = p\left(\sigma_c^t\right)^2 + (1-p)\left(\sigma_{ic}^t\right)^2. \tag{8}$$

Example 4. For the observations in Table 2 at time t_1, suppose the model parameters learned at time t_0 is $\mathbf{w}^{t_0} = (32.328, -0.348)$. At time t_1, the predicted temperature values based on $\mathbf{w}^{t_0}$ are $\widehat{\mathbf{y}}^{t_1} = (18.587, 18.057, 16.504, -, 18.482, 18.595, 17.396)$, while the observed values are $\mathbf{y}^{t_1} = (12.346, 19.518, 15.215, -, 24.387, 17.743, 18.392)$ (we do not predict the values of MVs as our goal is to detect anomalous observations in this step). The difference between $\widehat{\mathbf{y}}^{t_1}$ and $\mathbf{y}^{t_1}$ (i.e., the estimated imputation error) is $\widehat{\mathbf{y}}^{t_1} - \mathbf{y}^{t_1} = (6.241, -1.461, 1.289, -, -5.905, 0.852, -0.996)$. Thus the detected anomalous set is $O_a^{t_1} = \{o_1^{t_1}, o_5^{t_1}\}$ where their absolute differences between the observed values and predicted values are greater than the anomaly threshold $\tau = 5$. Accordingly, the consistent set is $O_c^{t_1} = \{o_2^{t_1}, o_3^{t_1}, o_6^{t_1}, o_7^{t_1}\}$. As introduced earlier, we suppose that the consistent observations and anomalous observations fit the learned correct and incorrect model well, respectively. Therefore, based on the Equation (8), the variance $(\sigma_1^{t_1})^2$ is estimated as $(\sigma_1^{t_1})^2 = 0.99 \times \frac{1}{4} \times [(-1.461)^2 + 1.289^2 + 0.852^2 + (-0.996)^2] + 0.01 \times \frac{1}{2} \times [6.241^2 + (-5.905)^2] = 1.734$ (suppose $p = 0.99$).

Next, we discuss the bound of variance $\left(\sigma^t\right)^2$. As introduced earlier, the anomalous set O_a^t is undetermined and we need to detect whether the complete observations are anomalous or not. Given a poor-quality dataset O^t, suppose it contains a_1 anomalous observations in which a_2 anomalous observations are detected.

(i) If $a_2 = a_1$, i.e., all anomalous observations are detected exactly, then every complete observation $o_i^t \in O_f^t$ is a consistent observation. We have $\forall o_i^t \in O_f^t$, $\widetilde{y}_i^t = y_i^t$, i.e., the observed value on incomplete attribute of $o_i^t \in O_f^t$ is the ground truth. Thus $\forall o_i^t \in O_f^t$, we have $0 \le \left|e_i^t\right| \le \tau$, thereby $0 \le \left(\sigma^t\right)^2 \le \tau^2$.

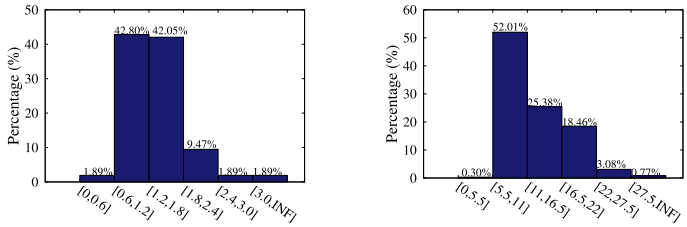
(ii) If $0 \le a_2 < a_1$, $(a_1 - a_2)$ anomalous observations are not detected and included into O_f^t incorrectly. In other words, there are $(a_1 - a_2)$ anomalous observations which do not fit the imputation model, but they are misused to estimate the variance of imputation. For $\forall o_i^t \in O_f^t$, if o_i^t is a consistent observation, we have $\widetilde{y}_i^t = y_i^t$ and $0 \le \left|e_i^t\right| \le \tau$. If o_i^t is an anomalous observation, there should be $\tau \le \left|e_i^t\right| \le +\infty$ because the anomalous values may be any values. Therefore, we have $\frac{a_1-a_2}{n_f}\tau^2 \le \left(\sigma^t\right)^2 \le +\infty$.

Finally, the variance of imputation error e^t of poor-quality dataset O^t satisfies $\frac{a_1-a_2}{n_f}\tau^2 \le \left(\sigma^t\right)^2 \le +\infty$.

Proposition 6. *Theoretically, the upper bound of the minimal variance* $\left(\sigma^t\right)^2$ *is* $\left(\sigma^t\right)^2_{min\_ub} = p\delta_{max}\tau^2 + (1-p)(1-\delta_{min})\tau^2$.

Proof. Based on the above analysis, the minimal variance $\frac{a_1-a_2}{n_f}\tau^2$ achieves the greatest value when $a_2 = 0$, i.e., there is no anomalous observation detected. Thus the upper bound of minimal variance $\frac{a_1-a_2}{n_f}\tau^2 = \frac{a_1}{n_f}\tau^2$.

When the imputation model is correct, we have $O_f^t = O_c^t$, i.e., $\frac{a_1}{n_f}\tau^2 = \delta^t\tau^2 \le \delta_{max}\tau^2$. In contrast, when the imputation model is incorrect, we have $O_f^t = O_a^t$, i.e., $\frac{a_1}{n_f}\tau^2 = (1-\delta^t)\tau^2 \le$



(a) IDL (b) Air Quality

Fig. 4. The Distribution of Estimated Variance Over Real Datasets.

$(1-\delta_{min})\tau^2$. Combined with Equation (8), we have $\frac{a_1}{n_f}\tau^2 = p\delta^t\tau^2 + (1-p)(1-\delta^t)\tau^2 \leq p\delta_{max}\tau^2 + (1-p)(1-\delta_{min})\tau^2$. Thus the proposition is proved. □

Note that the upper bound of the minimal variance $(\sigma^t)^2$ is a theoretical threshold. When the upper bound is achieved, no anomalous observation is detected, i.e., the imputation model has deteriorated (refer to the proof of Porposition 6). However, in practice, the estimated variance $(\sigma^t)^2$ is generally small. Intuitively, if an anomalous observation o_i^t is considered as a consistent observation incorrectly, we have $0 \leq |y_i^t - \widehat{y}_i^t| \leq \tau$. Moreover, since we consider the observed y_i^t of a consistent observation to be the ground truth $\widetilde{y}_i^t$, the imputation error e_i^t of the anomalous observation o_i^t is incorrectly computed as $|e_i^t| = |\widetilde{y}_i^t - \widehat{y}_i^t| = |y_i^t - \widehat{y}_i^t|$ which is less than τ, so the estimated variance is less than τ^2, rather than greater than τ^2 in theoretical analysis. Therefore, we adopt the upper bound of the minimal variance $p\delta_{max}\tau^2 + (1-p)(1-\delta_{min})\tau^2$ as the variance threshold ς^2. If the estimated variance $(\sigma^t)^2 > p\delta_{max}\tau^2 + (1-p)(1-\delta_{min})\tau^2$, we consider a deterioration is detected at time t. Algorithm 4 presents the procedure of deterioration detection. First, we determine the anomalous set O_a^t and consistent set O_c^t (Lines 2–10). Then, the variance of imputation error e^t of poor-quality dataset O^t, i.e., $(\sigma^t)^2$, is computed based on Equations (7) and (8) (Lines 11–18). Finally, by comparing the estimated variance $(\sigma^t)^2$ and the bound $p\delta_{max}\tau^2 + (1-p)(1-\delta_{min})\tau^2$, we determine whether time point t is a deterioration point or not (Lines 19–20).

Example 5. For two real-world datasets: IDL (introduced in Example 2) and Air Quality (to be introduced in Section 6), suppose the anomaly threshold for the two datasets are 5 and 15, respectively. Additionally, for IDL dataset, the $\delta_{max} = 0.1$ and $\delta_{min} = 0.028$ based on our statistics as the anomalies naturally exist, while for Air Quality dataset, the $\delta_{max} = 0.1$ and $\delta_{min} = 0.1$ as no anomalies naturally exist and we inject anomalies manually. Based on Proposition 6, the upper bounds of minimal variance $(\sigma^t)^2$ for IDL and Air Quality are 2.6 and 24.3 respectively (with $p = 0.99$).

On the other hand, for the above two datasets, we plot the real distributions of estimated variance in Figure 4. As shown, 96.22% and 96.15% of the estimated variances fall in the relative small ranges, e.g., $[0, 2.4]$ and $[0, 22]$, respectively. It confirms our analysis that the estimated variance is generally smaller than $(\sigma^t)^2_{min\_ub}$ in practice, and thus it is reasonable to adopt the upper bound of the minimal variance as the variance threshold to detect the deterioration.

5 MISSING VALUE IMPUTATION

Finally, the MVs in the incomplete observations are imputed based on the learned MLR model. Given an incomplete (anomalous) observation o_i^t, the imputed (repaired) result of $y_i^t \in o_i^t$ is:

$$\widehat{y}_i^t = \begin{cases} y_i^t & if \left|y_i^t - \widehat{y}_i^t\right| < \tau \\ w_0^t + \sum_{j=1}^{d_c} w_j^t x_{i,j}^t + \varepsilon_i^t & otherwise \end{cases} \tag{9}$$

ALGORITHM 4: DeteriorationDection(O^t, $\mathbf{w}^{t-1}$, δ_{max}, δ_{min}, p, τ)

Input: the poor-quality dataset O^t at time t ($t > 1$),
the model parameter $\mathbf{w}^{t-1}$,
the maximum anomaly ratio δ_{max},
the minimum anomaly ratio δ_{min},
the probability p,
and the anomaly threshold τ
Output: a bool variate de to indicate the deterioration

$de = False$, $O_c^t = \phi$, $O_a^t = \phi$
foreach $o_i^t \in O^t$ **do**
  if $y_i^t \in o_i^t$ *is missing* **then**
    continue //o_i^t is an incomplete observation
  else
    $\widehat{y}_i^t = w_0^{t-1} + \sum_{j=1}^{d_c} w_j^{t-1} x_{i,j}^t + \varepsilon_i^t$ // $w_0^t, w_j^{t-1} \in \mathbf{w}^{t-1}$
    if $\left|y_i^t - \widehat{y}_i^t\right| < \tau$ **then**
      $O_c^t \leftarrow \left\{o_i^t\right\}$ //o_i^t is a consistent observation
    else
      $O_a^t \leftarrow \left\{o_i^t\right\}$ //o_i^t is an anomalous observation
// When the imputation model is correct
$O_f^t = O_c^t$
$\left(\sigma_c^t\right)^2 = \frac{1}{n_f}\left(\mathbf{y}_f^t - \mathbf{X}_f^t \mathbf{w}^{t-1}\right)^T \left(\mathbf{y}_f^t - \mathbf{X}_f^t \mathbf{w}^{t-1}\right)$
// When the imputation model is incorrect
$O_f^t = O_a^t$
$\left(\sigma_{ic}^t\right)^2 = \frac{1}{n_f}\left(\mathbf{y}_f^t - \mathbf{X}_f^t \mathbf{w}^{t-1}\right)^T \left(\mathbf{y}_f^t - \mathbf{X}_f^t \mathbf{w}^{t-1}\right)$
// Finally computing the $\left(\sigma^t\right)^2$
$\left(\sigma^t\right)^2 = p\left(\sigma_c^t\right)^2 + (1-p)\left(\sigma_{ic}^t\right)^2$
if $\left(\sigma^t\right)^2 > p\delta_{max}\tau^2 + (1-p)(1-\delta_{min})\tau^2$ **then**
  $de = True$
return de

Algorithm 5 shows the pseudo-code. It is worth noting that, the anomalous observations in O_a^t can be detected and repaired at the same time. As a result, the anomaly ratio δ^t can be estimated based on the learned model. Moreover, δ_{max} and δ_{min} can be updated accordingly. In this way, there is no need to employ specialized anomaly detection approaches for anomaly ratio estimation, and thus both the time complexity and space complexity are reduced. Moreover, with the δ_{max} and δ_{min} estimated based on the learned MLR model in REMAIN, the best performance is achieved compared with existing anomaly detection/repair approaches (as shown in Appendix A).

Example 6. By the proposed REMAIN, the imputed streaming data in Example 3 is shown in Table 3, where the MVs are imputed with high-accuracy (RMSE=0.5). In addition, the anomalies can be detected and repaired effectively, thereby the anomaly ratio can be estimated accurately.

Table 3. The Imputation Result of Data in Example 3

	observations			
sensor id	t_0	t_1	t_2	$\cdots$
1	(19.447, 39.525)	(**18.585**, 39.484)	(**18.509**, 39.710)	
2	(18.843, 40.917)	(19.518, 41.010)	(18.146, 41.207)	$\cdots$
3	(**16.572**, 45.276)	(15.215, 45.470)	(15.307, 47.296)	
4	(**17.160**, 43.585)	(**17.107**, 43.738)	(**16.942**, 44.212)	
5	(18.535, 39.622)	(**18.482**, 39.788)	(**18.345**, 40.181)	$\cdots$
6	(19.154, 39.377)	(17.743, 39.462)	(18.343, 39.507)	
7	(17.762, 42.856)	(18.392, 42.906)	(**17.365**, 42.998)	
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

ALGORITHM 5: MVImputation(O^t, $\mathbf{w}^t$, τ)

Input: the poor-quality dataset O^t at time t ($t = 1, 2, \ldots$),
the model parameters $\mathbf{w}^t$,
and the anomaly threshold τ
Output: the imputed dataset O^{t*}

$O^{t*} = O^t$
$\left|O_a^t\right| = 0, \left|O_c^t\right| = 0$
foreach $o_i^t \in O^{t*}$ **do**
  $\widehat{y}_i^t = w_0^t + \sum_{j=1}^{d_c} w_j^t x_{i,j}^t + \varepsilon_i^t$ // $w_0^t, w_j^t \in \mathbf{w}^t$
  if $y_i^t \in o_i^t$ *is missing* **then**
    $y_i^t = \widehat{y}_i^t$ //o_i^t is an incomplete observation
  else
    if $\left|y_i^t - \widehat{y}_i^t\right| < \tau$ **then**
      $\left|O_c^t\right| += 1$
      continue //o_i^t is a consistent observation
    else
      $\left|O_a^t\right| += 1$
      ${y_i}^t = \widehat{y}_i^t$ //o_i^t is an anomaly observation
  $\delta^t = \frac{|O_a^t|}{|O_a^t|+|O_c^t|}$
return O^{t*}, δ^t

Proposition 7. *The space complexity of REMAIN is $O(d^2)$ which can be regarded as a constant due to $d \ll n$, while the time complexity of REMAIN is $O(n_c d^2 + n_m)$ which increases linearly with the number of observations in O^t.*

Proof. The proposed REMAIN imputes MVs based on a learned MLR model which is a global model, i.e., we only need to maintain an MLR model with $O(d)$ space for the entire streaming data. In addition, at each time point, the model parameters are updated or re-estimated just based on the poor-quality dataset at current time t without any historical data. Based on Propositions 5 and 2, the space costs of both parameter update and parameter re-estimation are $O(d^2)$. Therefore, the space cost of our REMAIN is $O(d^2)$, which can be regarded as constant due to $d \ll n$. In the same

line, the time costs of parameter update and parameter re-estimation are linearly dependent on the number of consistent complete observations (i.e., n_c) in the data. Moreover, the time complexity of MV imputation based on the learned MLR model is $O(1)$ for an incomplete observation. As a result, the time complexity of MV imputation for O^t with n_m incomplete observations is $O(n_m)$. Thus the time complexity of REMAIN is $O(n_c d^2 + n_m)$. The conclusion is proved. □

Note that in deterioration detection, we estimate the imputation error based on a widely adopted principle in data cleaning—humans or systems always try to minimize mistakes in practice [39]. The variance of the imputation errors is then leveraged to reflect the reliability of the imputation result. However, there is no guarantee for the imputation result to be always aligned with the ground truth. Thus similar to other MV imputation algorithms, the accuracy of the imputation result is unlikely to have theoretical guarantee. For this reason, we can only evaluate the accuracy of our imputation result by comparing with the ground truth in experiments, which is the common practice in MV imputation studies.

6 EXPERIMENTS

In this section, we report the result of an experimental study on REMAIN along with a number of the state-of-the-art methods. All the programs are implemented in Python and the experiments are performed on a PC with 3.4 GHz CPU and 16 GB RAM.

6.1 Experimental Settings

Datasets. To evaluate the performance of REMAIN, we adopt two real-world datasets as follows.

(1) *Air Quality (AQ) data with real MVs and synthetic anomalies.* The dataset contains observations which record concentrations for CO, Non Metanic Hydrocarbons (NMHC), Nitrogen Oxides (NO_x), Nitrogen Dioxide (NO_2), O_3, Temperature (T), Relative Humidity (RH) and Absolute Humidity (AH) collected by an Air Quality Chemical Multi-sensor device deployed in a significantly polluted area of Italy [32]. In this dataset, the real MVs (mainly appear in attributes NO_2 and NO_x) naturally exist and the corresponding ground truths are known. Therefore, we impute the inherent MVs in the dataset. Moreover, we aim to verify that REMAIN is competent for imputing MVs in multiple attributes by adopting NO_2 and NO_x as the incomplete attributes. By collecting data from 72 successive time points as the observations of a logical time point, the dataset contains 130 logical time points. In addition, we synthetically inject anomalies into the data by randomly replacing observed values on the incomplete attributes of some complete observations. Following the same line of evaluation for stream data cleaning [28], the value of each replaced observation is substituted by a random value between the minimum and maximum values in the dataset.

(2) *IDL data with real anomalies and synthetic MVs.* As introduced in Example 2, the IDL dataset contains measurements of temperature and humidity taken from 54 sensors for every 31 seconds. We adopt the data from 51 sensors as the data collected by the remain three sensors are too sparse. In this dataset, both MVs and anomalies naturally exist and mainly appear in the Temperature attribute. However, as the ground truths of MVs are unknown, we cannot evaluate the effectiveness of the proposed algorithms by directly using the real MVs contained in the dataset. Instead, we remove the observations with inherent MVs from the dataset and synthetically generate the incomplete observations by marking off certain percentage of values on the attribute Temperature in the remaining data. On the other hand, we first employ RANSAC to detect the anomalies naturally existing in the data, then we double check and label the anomalies manually. Finally, by taking three successive time points as a logical time point, a dataset with 88 logical time points where 153 observations arrive at each logical time point is obtained.

In addition, for ease of evaluation, we manually inject the same specific ratio of anomalies and MVs for the data arriving at each time point t in AQ and IDL dataset, respectively. Moreover, we normalize the data based on min-max normalization [19], and repeat 10 times for each test as the anomalies (in AQ) and MVs (in IDL) are randomly introduced. Finally, the averages are reported to obtain reliable experimental results.

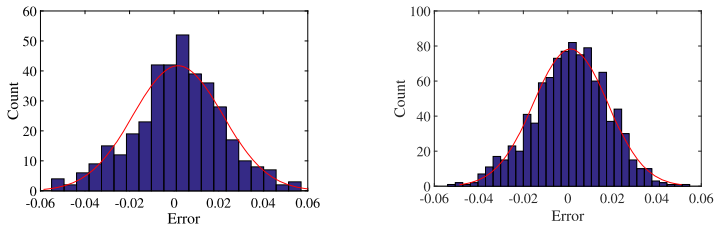
Algorithms for comparison. We experimentally compare the proposed REMAIN with four state-of-the-art online MV imputation algorithms, including (1) the multivariate linear regression method MLR [21], (2) the top-k case matching method TKCM [34], (3) the online time-series mining methods MUSCLES [35], and (4) SPIRIT [25]. In addition, as the anomalies in the data have a negative influence on MV imputation, we implement the state-of-the-art anomaly repair algorithms, including AR [16], ARX [23], IMR [39], and EWMA [17] to detect and repair the anomalies before MV imputation over the examined online MV imputation algorithms. Note that the parameters in AR model can be estimated based on either Yule–Walker equation or Bayesian approach (named BVAR model in [26]). Since the Bayesian estimation, which requires a predefined Bayesian prior, is difficult for incremental computation, we implement the AR model based on Yule–Walker equation in this article. In addition to REMAIN (which initializes and re-estimates model parameters based on a-RANSAC), we also implement REMAIN-RSAC (which initializes and re-estimates model parameters based on RANSAC) and REMAIN-NDD (REMAIN without deterioration detection).

Parameter Settings. In REMAIN, besides the poor-quality dataset O^t arriving at each time point t, four parameters, i.e., δ^1, p, s, and τ, need to be provided. As introduced in Section 3.2.1, the anomaly ratio at the first time point, i.e., δ^1, can be given by domain experts or estimated based on the arrival data O^1 by exploring the existing visual identification approaches. Intuitively, the impact of δ^1 on the final imputation accuracy is slight, as δ^1 is only used to initialize the δ_{max} and δ_{min} which are expected to update with time passes. We set $\delta^1 = 0.1$ for both AQ and IDL datasets in experiments. The parameters p and s are the input of RANSAC (introduced in Section 3.2.1), where p is the probability that a correct model can be obtained and s is the minimal number of observations required to compute the model parameters. Intuitively, a large p indicates a high probability of correctness for the learned MV imputation model while a high time cost (as shown in Figure 6). As recommended in [10], we adopt $p = 0.99$ for the proposed REMAIN. Additionally, as introduced earlier, we set $s = d_c + 2$ since $d_c + 1$ parameters need to be learned where d_c is the number of complete attributes in an observation. Finally, the parameter τ is the anomaly threshold used to identify the anomalies and needs to be predefined. It is notable that τ is also required by existing anomaly repair approaches, i.e., AR, ARX, IMR, and EWMA. In the existing works mentioned above, τ is usually decided by observing the statistical distributions of distances between the observed values and predicted values based on a segment of historical data [39]. In other words, the parameter τ is usually pre-learned based on a segment of historical data. In this article, the parameter τ can be decided based on the poor-quality dataset O^1 at the first time point with the same line of the existing works. In our experiments, we evaluate the performances of the proposed REMAIN and the compared baselines over the real-world datasets (see the experimental results in Appendix A), and set $\tau = 0.2$ as the default value since almost all approaches achieve the best performance when $\tau = 0.2$.

For existing methods adopted as baselines, their parameters are required to be provided by pre-learning based on a testing dataset (which is constituted by historical data). In addition, the existing methods may not achieve the best performance with the default parameters recommended by their authors, as the datasets adopted in this article are different from that adopted in existing works. For fair comparison, we conduct a set of experiments over the two real-world datasets used in this article (see details in Appendix B), and adopt the values with which the best performance is achieved as the default parameters. The default parameter settings are shown in Table 4. Note that

Table 4. Default Parameter Settings

Methods	Parameter Settings
REMAIN	$s = d_c + 2$, $p = 0.99$, $\delta^1 = 0.1$, $\tau = 0.2$
MUSCLES	$p' = 6$
SPIRIT	$p' = 6$, $h_s = 1$
TKCM	$L = 50$, $k' = 2$, $d' = 2$, $l = 3$ (AQ), $l = 9$ (IDL)
AR/ARX/IMR	$p' = 6$, $\tau = 0.2$
EWMA	$\lambda = 0.2$ (AQ), $\lambda = 0.3$ (IDL), $\tau = 0.2$



(a) Air Quality (b) IDL

Fig. 5. Error distribution approximation.

some notations in existing works are the same as that in REMAIN but with different meanings. To avoid ambiguity, we add the apostrophes for the repeat notations of existing works, e.g., p of AR is represented by p' in this article.

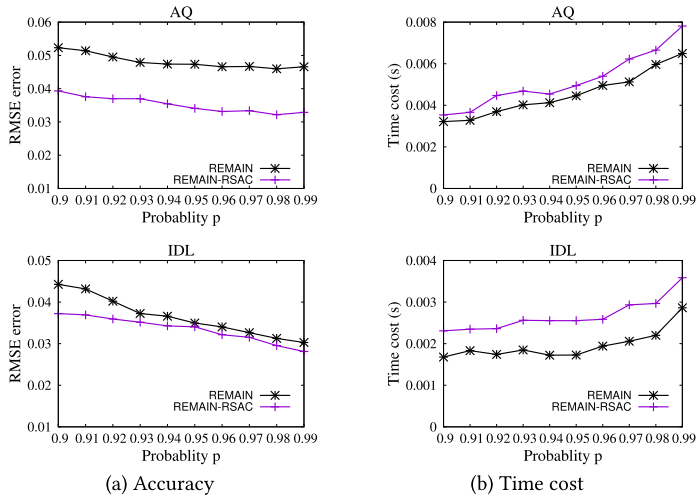
Performance metric. We adopt Root Mean Square Error (RMSE) to evaluate the imputation result. The lower RMSE is, the imputation result is closer to the ground truth and thus the imputation has better performance.

6.2 Error Distribution Evaluation

As we assume that the imputation errors follow the Gaussian distribution in deterioration detection (Section 4), here we verify this assumption using the two real-world datasets. Figure 5 shows the error distribution of imputation results from our datasets. It can be observed that Gaussian distributions are fitted and the means are approximate 0, showing consistency with our assumption.

6.3 Performance Evaluation by Varying p

Recall that the parameter p is the probability that a correct model can be obtained by RANSAC (a-RANSAC). It impacts the performances of the proposed REMAIN and REMAIN-RSAC. Thus, we evaluate the RMSE and time cost of REMAIN and REMAIN-RSAC by varying p on the two real-world datasets. With a large p, the imputation model learned by RANSAC (a-RANSAC) is more likely to be correct. However, with the increase of p, the number of iteration $k = \frac{log(1-p)}{log(1-q^s)}$ in RANSAC increase, which incurs the high time cost. Moreover, the variance threshold $\varsigma^2 = p\delta_{max}\tau^2 + (1-p)(1-\delta_{min})\tau^2$ used for deterioration detection is low when the p is large. The time costs of both REMAIN and REMAIN-RSAC is higher with more deterioration detection points being detected. Therefore, as shown in Figure 6, the RMSE of REMAIN (REMAIN-RSAC) decreases as p increases, while the time cost correspondingly increases.



(a) Accuracy (b) Time cost

Fig. 6. Performance by varying the probability p with $\tau = 0.2$, $\delta^t = 0.1$, and $s = 10$ for AQ dataset and $\tau = 0.2$, $\gamma^t = 0.1$, and $s = 3$ for IDL dataset, respectively.

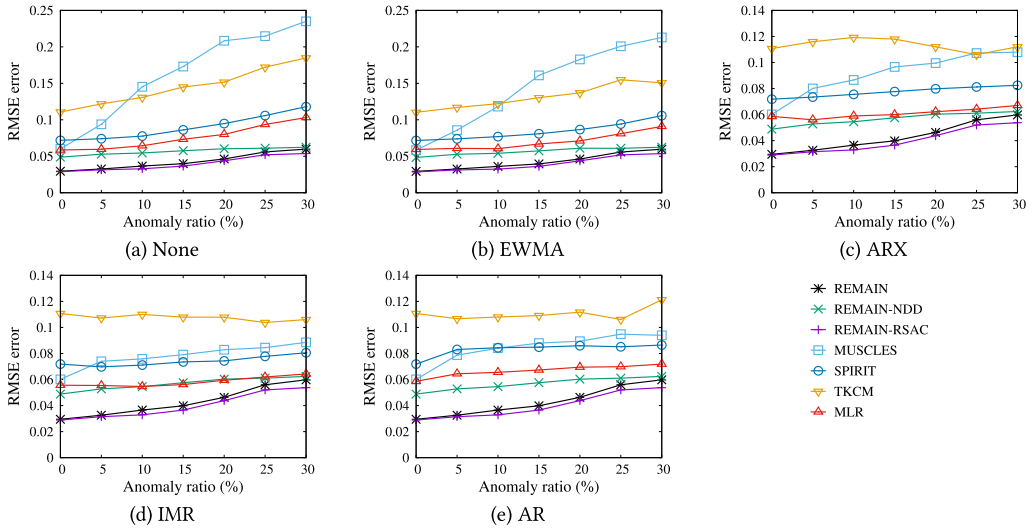


Fig. 7. Imputation accuracy by varying anomaly ratio δ^t, over Air Quality with $\tau = 0.2$, $p = 0.99$, and $s = 10$.

6.4 Evaluation on Air Quality

The experiments on real MVs and synthetic anomalies over AQ dataset mainly consider two factors: anomaly ratio δ^t and anomaly threshold τ.

6.4.1 Varying the Anomaly Ratio δ^t. First, REMAIN imputes the MVs by using data without anomalies, while existing solutions impute the MVs by employing repaired anomalies as introduced in Section 1. Therefore, as shown in Figure 7, REMAIN always achieves the lowest RMSE. In particular, when there is no anomaly in the data, i.e., the anomaly ratio is 0%, the performance of REMAIN is also the best. One reason is that existing works suffer the problem of inaccuracy propagation by imputing MVs using the historical data based on temporal dependencies, while

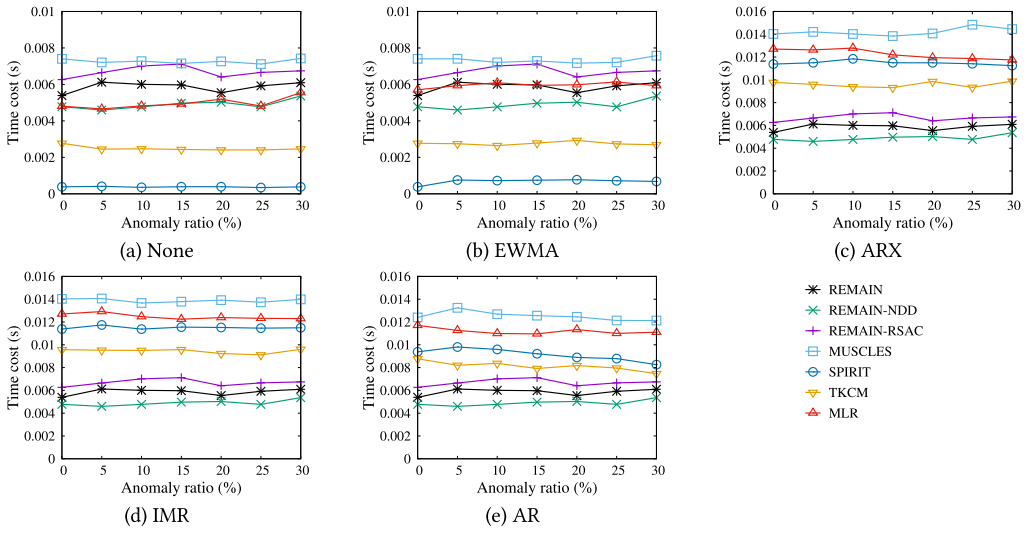


Fig. 8. Time cost by varying anomaly ratio δ^t, over Air Quality with $\tau = 0.2$, $p = 0.99$, and $s = 10$.

REMAIN imputes MVs based only on the data at the current time without being affected by historical imputation errors. Another reason is that the effective deterioration detection mechanism in REMAIN improves the imputation accuracy.

In addition, with anomaly repair before MV imputation, existing online MV imputation methods show better performance than no anomaly repair. For example, when anomaly ratio $\delta^t = 10\%$, the RMSE errors of MUSCLES with anomaly repair based on EWMA (Figure 7(b)), ARX (Figure 7(c)), IMR (Figure 7(d)), and AR (Figure 7(e)) are 0.1182, 0.0866, 0.0759, and 0.0841, respectively, which are lower than 0.1452 without anomaly repair (Figure 7(a)). Nevertheless, their performances are worse than REMAIN.

On the other hand, among REMAIN, REMAIN-RSAC, and REMAIN-NDD, the RMSE of REMAIN is slightly higher than that of REMAIN-RSAC, because in REMAIN, the parameters of the imputation model are re-estimated based on a small subset of consistent observations, which reduces the accuracy of imputation model. However, the time cost of REMAIN is lower than that of REMAIN-RSAC (can be observed in Figures 8 and 10), which verifies that REMAIN is more competent for streaming data. Comparing with REMAIN and ERMAIN-RSAC, the RMSE error of REMAIN-NDD is higher, which verifies that the deterioration detection indeed contributes to the imputation accuracy improvement.

Second, Figure 8 presents the time cost by varying anomaly ratio *delta*. Since there is no deterioration detection in REMAIN-NDD, its time cost is lower than that of REMAIN and REMAIN-RSAC. As shown in Figure 8(a), the time costs of MLR, SPIRIT, and TKCM are lower than that of REMAIN. The reason for MLR is that there is no anomaly detection and deterioration detection. For SPIRIT, it only needs to maintain h_s ($h_s = 1$ in our experiments, which is usually small) AR models by compressing n objects into h_s hidden variables. For TKCM, its time cost is low because there are a small number of MVs in the AQ dataset. It is not surprising that MUSLCES incurs higher time cost since it maintains n AR models for n objects. Similar results can be observed in Figure 8(b) with anomaly repair based on EWMA. Since the parameters of EWMA model are stable, the time cost of anomaly repair based on EWMA is much lower than that of MV imputation. As shown in Figure 8(c)–(e), the time costs for existing online MV imputation methods are significantly increased due to the high time cost of anomaly repair, while REMAIN achieves the lowest time cost.

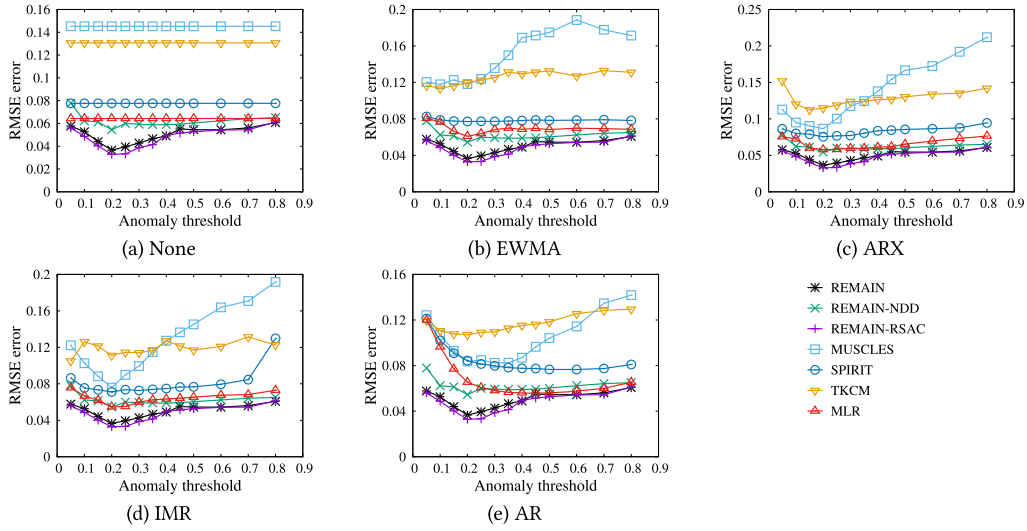


Fig. 9. Imputation accuracy by varying anomaly threshold τ, over Air Quality with $\delta^t = 0.1$, $p = 0.99$, and $s = 10$.

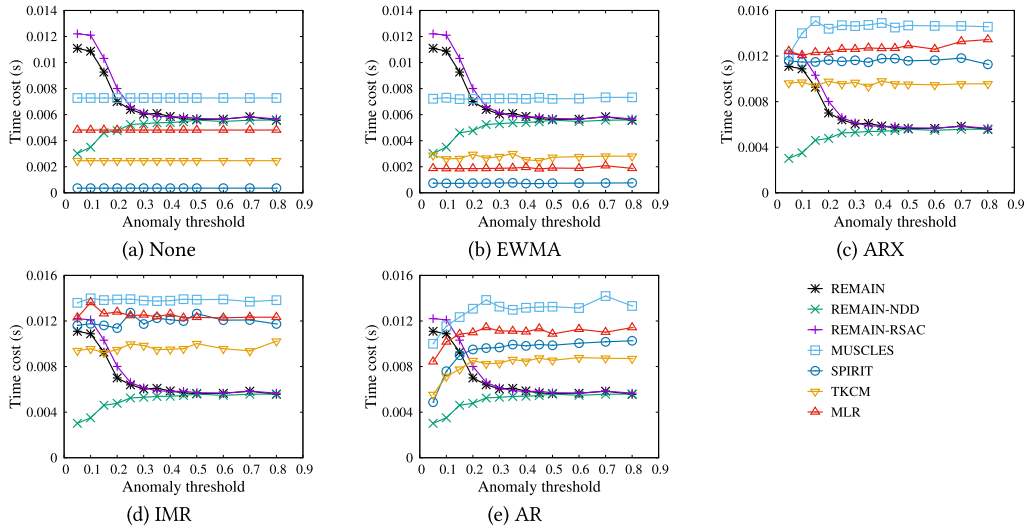


Fig. 10. Time cost by varying anomaly threshold τ, over Air Quality with $\delta^t = 0.1$, $p = 0.99$, and $s = 10$.

6.4.2 Varying the Anomaly Threshold τ. Figures 9 and 10 show the MV imputation accuracy and time cost under increased anomaly threshold τ. As shown in Figure 9, the RMSE error of REMAIN first decreases and then increases. With a small anomaly threshold, some consistent observations are detected as anomalous ones. As a result, the imputation model updated based on insufficient consistent observations does not perform well. On the contrary, with a too large an anomaly threshold, some anomalous observations are not detected, i.e., they are mistaken as consistent observations. As a result, the learned imputation model also does not perform well. Therefore, the RMSE error of REMAIN shows the tendency of deceasing first and then increasing.

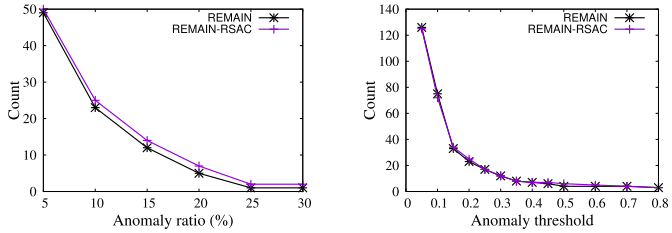


Fig. 11. Number of deterioration points over AQ dataset with $\tau = 0.2$, $\delta^t = 0.1$, $p = 0.99$, and $s = 10$.

In the same way, existing works with anomaly repair based on EWMA, AR, ARX, and IMR models show the same trend.

On the other hand, as shown in Figure 10, the time cost of REMAIN significantly decreases when the anomaly threshold is smaller than 0.3, and then remains stable with the increase of anomaly threshold. With a small anomaly threshold, the variance threshold is correspondingly small, which causes many deterioration points to be detected. At deterioration points, since the model parameters need to be re-estimated which is time consuming, the time cost of REMAIN is high when the anomaly threshold is small. Nevertheless, when the anomaly threshold reaches a certain value, e.g., 0.3, there is no anomalous observation to be detected, and the main time cost is caused by model parameter update. Thus the time cost of REMAIN tends to be stable. We adopt the default anomaly threshold $\tau = 0.2$.

Additionally, we show the number of detected deterioration points in Figure 11 by varying the anomaly ratio δ^t and the anomaly threshold τ over the AQ dataset. Since the variance threshold ς^2 is proportional to both anomaly ratio and anomaly threshold, it is not surprising that less deterioration points are detected with the increase of anomaly ratio and anomaly threshold. Moreover, the number of deterioration points detected by REMAIN is almost the same as that of REMAIN-RSAC, which further verifies that the proposed a-RANSAC (in REMAIN) does not trade much effectiveness off for the improved efficiency.

6.5 Evaluation on IDL Dataset

The experiments on real anomalies and synthetic MVs over IDL dataset focus on the evaluation by varying the missing ratio which controls the percentage of incomplete observations and the number of consecutive MVs which decide how long (in terms of consecutive time points) each incomplete observation lasts.

6.5.1 Varying the Missing Ratio. As shown in Figure 12, the imputation accuracy of REMAIN is stable with the increase of missing ratio. In addition, REMAIN always shows the lowest RMSE error compared with other online MV imputation methods under evaluation. This result demonstrates again that the proposed REMAIN works well in MV imputation for poor-quality streaming data.

Next, we compare the efficiency of various approaches in terms of time cost w.r.t. missing ratio. Intuitively, with the increase of missing ratio, more incomplete observations need to be imputed, while fewer consistent observations can be used to update or re-estimate parameters of the imputation model. Moreover, since the anomaly ratio is defined as the proportion of anomalous observations in all complete observations, with the increase of missing ratio, the number of complete observations decreases, and thereby the anomaly ratio increases. Accordingly, the variance threshold increases, which causes fewer deterioration points to be detected. Under the mutual impact of factors introduced above, Figure 13 shows that the time costs of REMAIN and REMAIN-RSAC decrease with the increase of missing ratio. Likewise, other approaches except for TKCM show

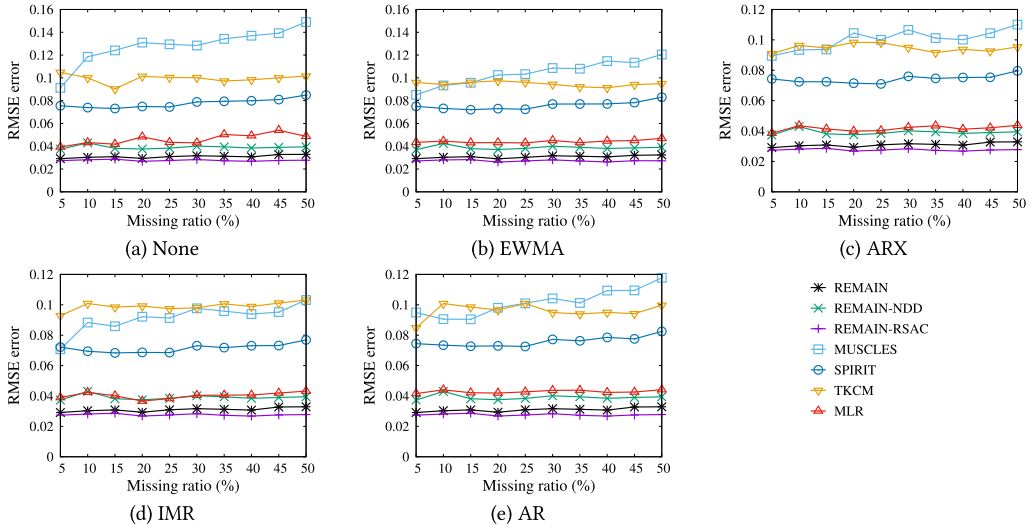


Fig. 12. Imputation accuracy by varying missing ratio γ^t, over IDL with $\tau = 0.2$, $p = 0.99$, and $s = 3$.

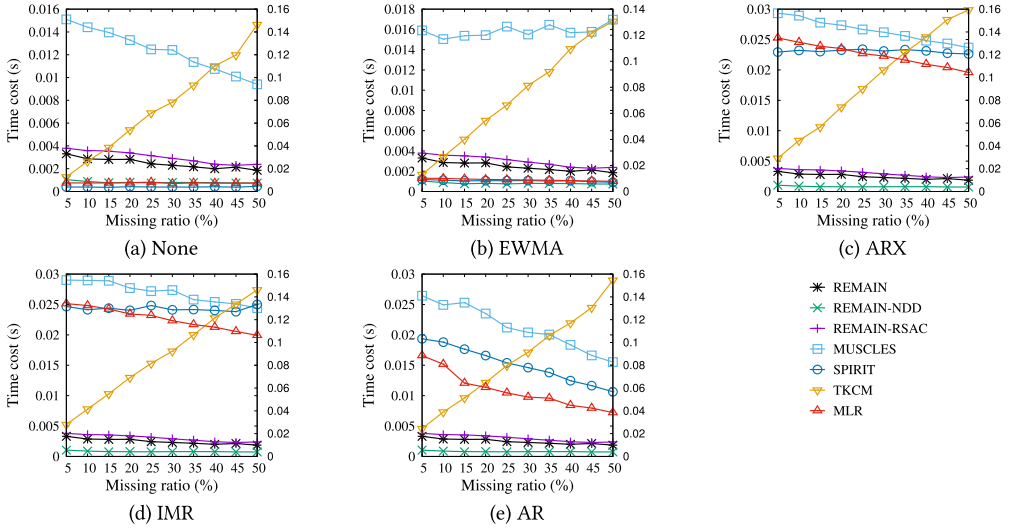


Fig. 13. Time cost by varying missing ratio γ^t, over IDL with $\tau = 0.2$, $p = 0.99$, and $s = 3$.

the similar trend, mainly because the time cost of parameter update is greater than that of MV imputation for temporal-dependency-based existing works. Since in TKCM, the MV in an incomplete observation is imputed based on the anchor points detected from the reference time series, which is time costly, its time cost shows an up-growing trend with the increase of missing ratio. Moreover, due to the high time consumption, we present the time cost of TKCM by using another ordinate (shown as right ordinate) in Figure 13.

6.5.2 Varying the Number of Consecutive MVs. In Figure 14, we study the imputation accuracy w.r.t. the number of consecutive MVs. As shown, the RMSE error of REMAIN remains stable as the number of consecutive MVs increases from 1 to 10, which verifies that our REMAIN is resilient to

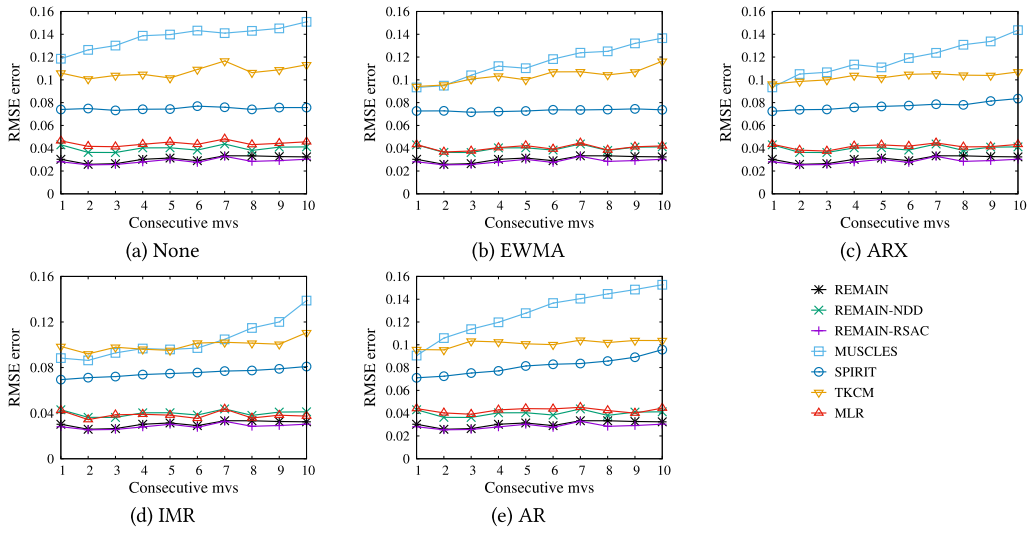


Fig. 14. Imputation accuracy by varying the number of consecutive MVs, over IDL with $\gamma^t = 0.1$, $\tau = 0.2$, $p = 0.99$, and $s = 3$.

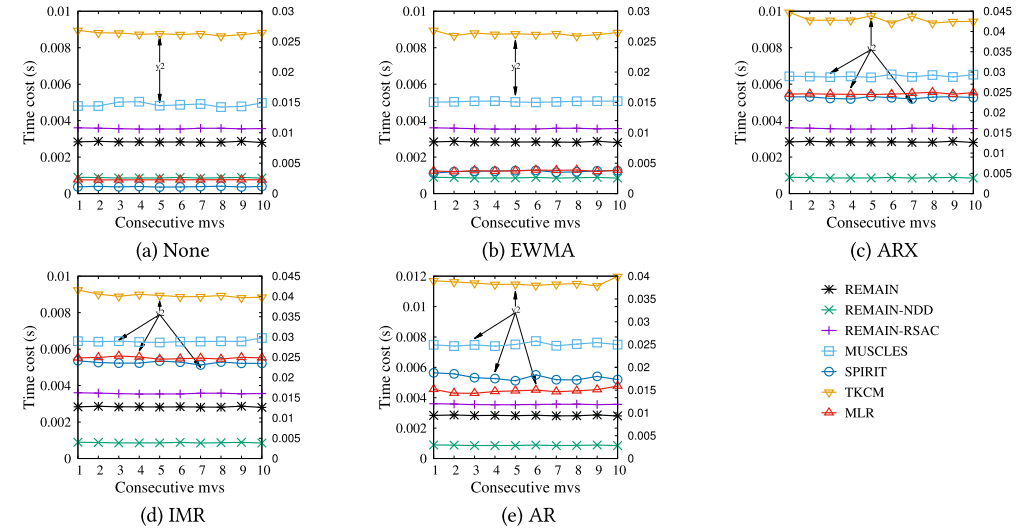


Fig. 15. Time cost by varying the number of consecutive MVs, over IDL with $\gamma^t = 0.1$, $\tau = 0.2$, $p = 0.99$, and $s = 3$.

consecutive MVs. Since MUSCLES and SPIRIT suffer from the error propagation during MV imputation, their performances become worse when they face a large number of consecutive MVs. The time cost results are reported in Figure 15. It is notable that to make the time costs of various approaches more clear, we present the results of some approaches by using another ordinate (shown as the right ordinate and annotated by "y2") in the figure. As shown, since both missing ratio and anomaly ratio are stable, the time costs of various methods are stable. In addition, REMAIN achieves the best performance in terms of time cost in most cases.

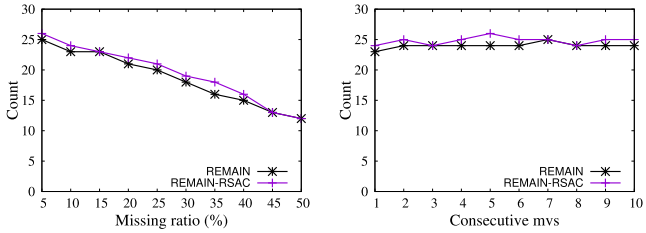


Fig. 16. Number of deterioration points over IDL dataset with $\gamma^t = 0.1$, $\tau = 0.2$, $p = 0.99$, and $s = 3$.

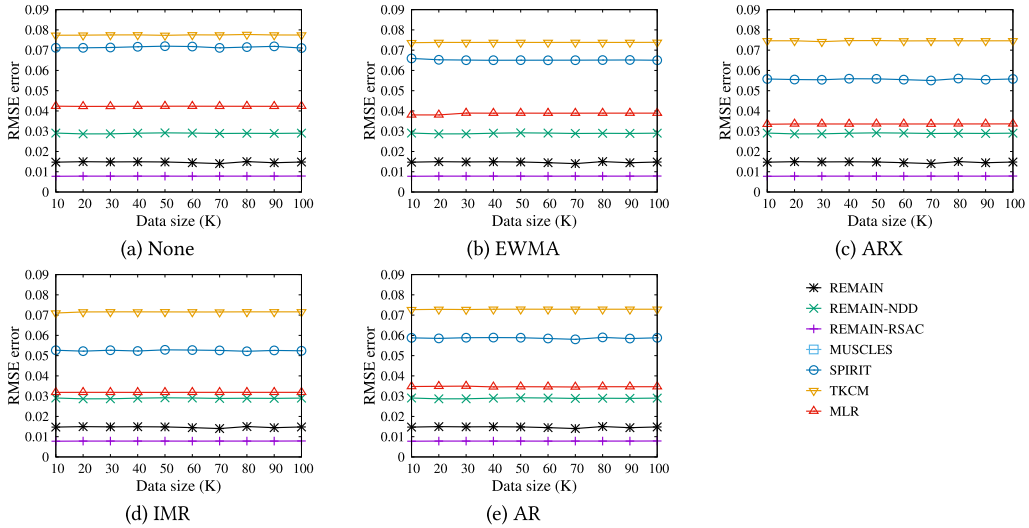


Fig. 17. Imputation accuracy by varying the data size.

Following the same line of AQ dataset, Figure 16 shows the number of detected deterioration points of REMAIN and REMAIN-RSAC over the IDL dataset by varying the missing ratio and the number of consecutive MVs. As introduced in Section 6.5.1, the increase of missing ratio incurs a large anomaly ratio, thus the number of detected points also decreases with the increase of missing ratio. Since the number of consecutive MVs do not impact the variance threshold ς^2 and both REMAIN and REMIAN-RSAC are competent for handling consecutive MVs, the number of detected deterioration points retains stable.

6.6 Evaluation on the Sythetic Dataset

To evaluate the scalability of REMAIN, we generate synthetic datasets based on the real-world dataset Air Quality. In the synthetic datasets, both anomaly ratio and missing ratio are 10%, and the anomaly threshold $\tau = 0.2$. Since the time cost of MUSCLES increases heavily with the increase of data size, we omit its results in this section.

Figure 17 shows that REMAIN has the best performance in imputation accuracy compared with other online MV imputation methods. Even though the imputation accuracy of REMAIN is slightly worse than that of REMAIN-RSAC due to the efficient parameter re-estimation, the efficiency of REMAIN is better than that of REMAIN-RSAC. Without deterioration detection, REMAIN-NDD shows the worst imputation accuracy compared with REMAIN and REMAIN-RSAC. With the anomaly repair (i.e., the anomalies are repaired before MV imputation), the improvement of MV

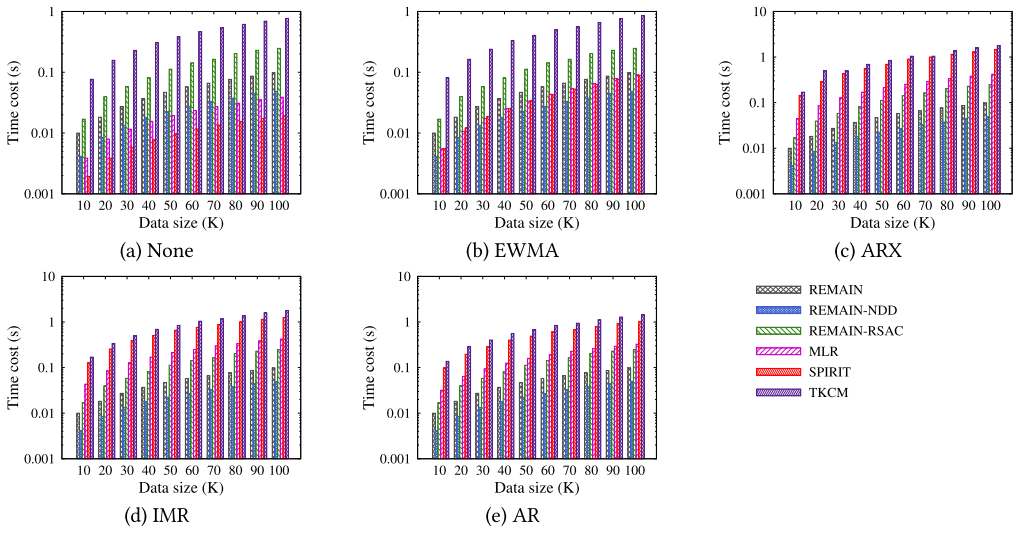


Fig. 18. Time cost by varying the data size.

imputation accuracy based on existing methods is limited, as the repaired anomalies have errors propagated into MV imputation easily.

As shown in Figure 18, it is not surprising that the time costs rise with the increase of data size for all online MV imputation approaches. Similar to the results shown in Figure 8, the time cost of SPIRIT is lower than that of REMAIN in Figure 18(a), since SPIRIT only needs to maintain a small number of AR models for hidden variables. Nevertheless, the SPIRIT has a higher imputation error. Likewise, since the anomalies are repaired based on a stable model in EWMA, similar results are observed in Figure 18(b). With anomaly repair based on AR, ARX, and IMR models, the time costs of existing works significantly increase due to the extra time cost for anomaly repair. Consequently, REMAIN achieves the best performance in time cost. Specifically, REMAIN yields one order of magnitude performance gain in time cost compared with existing online MV imputation approaches. In addition, it can be observed that the time cost of REMAIN is obviously lower than that of REMAIN-RSAC in Figure 18, which supports that REMAIN is more competent for streaming data than REMAIN-RSAC.

6.7 Summary of Experiments.

We summarize the experimental findings as follows: (1) The MV imputation accuracy is significantly improved by the proposed REMAIN compared with existing woks, e.g., MLR, MUSCLES, SPIRIT, and TKCM, even though the anomaly repair is conducted before MV imputation. (2) REMAIN shows comparable imputation accuracy to the REMAIN-RSAC, while the efficiency is significantly improved. (3) Without deterioration detection, REMAIN-NDD shows the lowest time cost while the imputation accuracy is worse than REMAIN and REMAIN-RSAC. (4) With constant space complexity, in time cost, REMAIN shows the best performance in most scenarios and shows up to one order of magnitude improvement compared with existing online MV imputation approaches.

7 CONCLUSION

In this article, we have studied the problem of MV imputation for poor-quality streaming data which has the following characteristics: (1) arriving continuously; (2) containing MVs and

anomalies simultaneously; and (3) changing dynamically for data values and the correlations among attributes. Specifically, these characteristics lead to major challenges for existing imputation methods. To address these challenges, we propose a novel framework of real-time and error-tolerant MV imputation, called REMAIN. To the best of our knowledge, this is the first study on MV imputation for poor-quality streaming data. Instead of employing MV imputation based on all of the complete observations, REMAIN learns and incrementally updates an effective imputation model by detecting and eliminating anomalous observations from the whole dataset. To make the imputation model adapt to the dynamical change of data, REMAIN detects the deterioration points through monitoring the MV imputation errors at each time point. Experimental results on two real-world and synthetic datasets demonstrate that REMAIN achieves higher imputation accuracy and scales up well compared to the state-of-the-art MV imputation methods.

APPENDICES

A ESTIMATION OF ANOMALY RATIO

Given a poor-quality dataset O^t at time t ($t > 1$), since the anomalous set $O_a^t \subseteq O^t$ is non-deterministic, the anomaly ratio δ^t needs to be estimated based on the detected anomalous observations. Moreover, the accuracy of the estimated anomaly ratio δ^t affects the accuracy of δ_{max} and δ_{min} which are used for deterioration detection and parameter initialization, and thereby affects the final imputation accuracy. In the proposed REMAIN, at each time point t, we estimate the anomaly ratio δ^t based on the learned MLR model and update δ_{max} and δ_{min} accordingly. To evaluate the effectiveness of anomaly detection based on the learned MLR model in REMAIN, we compare the proposed the REMAIN with (1) AR+REMAIN, (2) ARX+REMAIN, (3) IMR+REMAIN, and (4) EWMA+REMAIN, where the MVs are imputed based on the learned MLR model in REMAIN, while the anomalies are detected/repaired based on existing works AR, ARX, IMR, and EWMA, respectively. Moreover, we implement the REMAIN with the real anomaly ratio δ^t (denoted by "Real" in Figure 19).

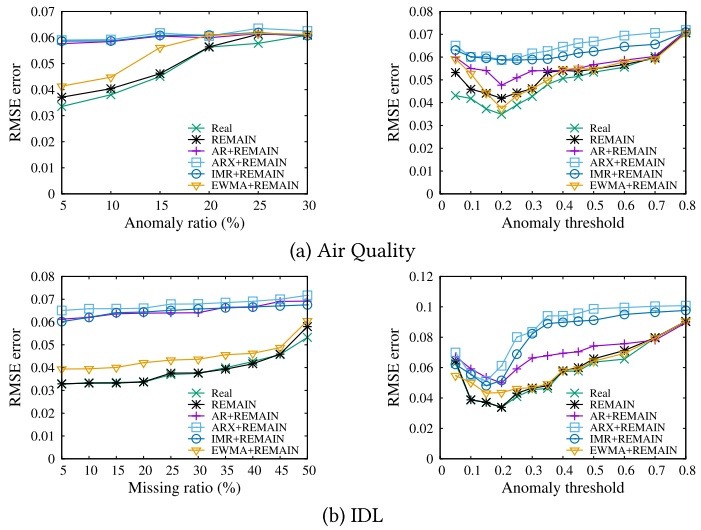


Fig. 19. Performance comparison on imputation accuracy with $p = 0.995$.

First, Figure 19(a) shows the MV imputation accuracies of various approaches by varying anomaly ratio and anomaly threshold over the AQ dataset, repectively. As introduced earlier, the learned MLR model (used for MV imputation and anomaly detection simultaneously) in REMAIN explores

the non-linear correlations among attributes in the same observation, while AR, ARX, IMR, and EWMA explore the temporal correlations of the data to detect/repair anomalies based on a segment of historical observations. Since the inaccuracies of imputed MVs and repaired anomalies in the historical data segment are to be propagated to later anomaly detection, the estimated anomaly ratio based on AR, ARX, IMR, and EWMA are likely to be inaccurate. By comparison, REMAIN does not suffer the problem of inaccuracy propagation with no requirements for historical data during model learning. Moreover, the mechanism of deterioration detection in REMAIN further improves the effectiveness of the learned MLR model by considering the abrupt change of the data. Therefore, as shown in Figure 19(a), the estimated anomaly ratio based on the learned MLR model in REMAIN is most close to the real anomaly ratio, and thereby the estimated δ_{max} and δ_{min} are the most accurate. Consequently, REMAIN achieves the lowest RMSE compared with AR+REMAIN, ARX+REMAIN, IMR+REMAIN, and EWMA+REMAIN.

Similar results can be observed in Figure 19(b) which shows the MV imputation accuracies of various approaches by varying the missing ratio and anomaly threshold over IDL dataset, respectively. Since anomalies naturally exist in the IDL dataset, i.e., at each time point, the size of anomalous set O_a^t is deterministic, the size of consistent set O_c^t is reduced with the increase of missing ratio, and thereby the anomaly ratio δ^t increases accordingly. Therefore, the change trend of MV imputation accuracy in Figure 19(b) is similar to that of in Figure 19(a). On the other hand, with the increase of anomaly threshold τ, the RMSE errors of all approaches decrease first and increase then (the reasons are elaborated in Section 6.4.2). Moreover, almost all approaches achieve the best performances when $\tau = 0.2$. Thus, we adopt 0.2 as the default value of the anomaly threshold τ for the proposed REMAIN and the existing anomaly detection/repair works, i.e., AR, ARX, IMR, and EWMA.

B DEFAULT PARAMETER SETTINGS

In existing methods, the parameters are recommended for their adopted datasets rather than the two real-world datasets adopted in this work. Thus, through extensive experimental evaluation, we adopt the parameters with which the best performance is achieved over the datasets adopted in this work.

B.1 Parameter Settings for Existing Anomaly Repair Approaches

For existing anomaly repair approaches, they are mainly used for anomaly detection before MV imputation over the examined online MV imputation algorithms. We use the $F-measure$ to evaluate the accuracy of the detected anomalies. Let $truth$ be the set of real anomalies and $found$ be the set of detected anomalies. Then $F-measure = 2 \cdot \frac{Recall \cdot Precision}{Recall+Precision}$ where $Recall = \frac{|truth \cap found|}{|truth|}$ and $Precision = \frac{|truth \cap found|}{|found|}$. The higher $F-measure$ is, the closer the detected anomalies are to the ground truths, and thereby the imputation accuracy is higher.

As shown in Figure 20(a)–(c), since the rationales of anomaly repair based on existing AR, ARX, and IMR are similar, the change trends of them are also similar to each other. With various anomaly ratios, their performances are stable when p' is in range of [4, 9], while with various missing ratios, most of them achieve the best performance when $p' = 6$. Thus, for AR, ARX and IMR, we set the default value of p' as 6. For EWMA, based on its performance shown in Figure 20(d), we set the default $\lambda = 0.2$ for Air Quality dataset and $\lambda = 0.3$ for IDL dataset, respectively.

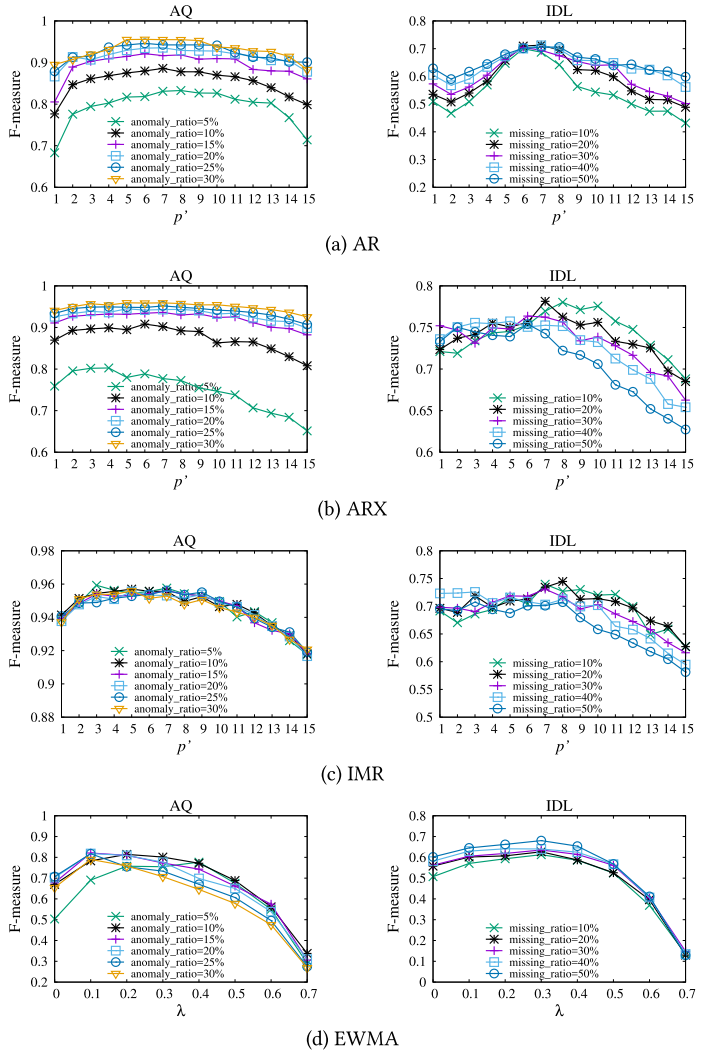


Fig. 20. Performance of existing anomaly repair approaches with various parameters.

B.2 Parameter Settings for Existing MV Imputation Algorithms

For existing online MV imputation approaches, we adopt the RMSE introduced in Section 6.1 to evaluate their imputation accuracy.

First, Figure 21 presents the MV imputation accuracy of MUSCLES by varying the parameter p' with various anomaly ratios and missing ratios over AQ and IDL datasets, respectively. We adopt the default $p' = 6$ for the following two reasons: (i) As shown in Figure 21, the RMSE error of MUSCLES shows an up-growing trend with the increase of p'. But the increase of RMSE error is slight especially for the IDL dataset, i.e., the difference of RMSE error between $p' = 6$ and $p' = 1$ is not obvious. (ii) In several existing works [25, 34, 35], all of them adopt $p' = 6$ as the default value with different datasets. Without loss of generality, we also adopt $p' = 6$ as the default value.

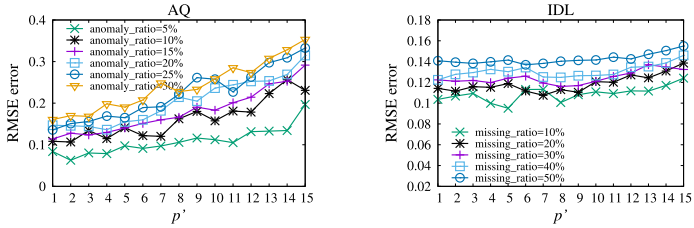
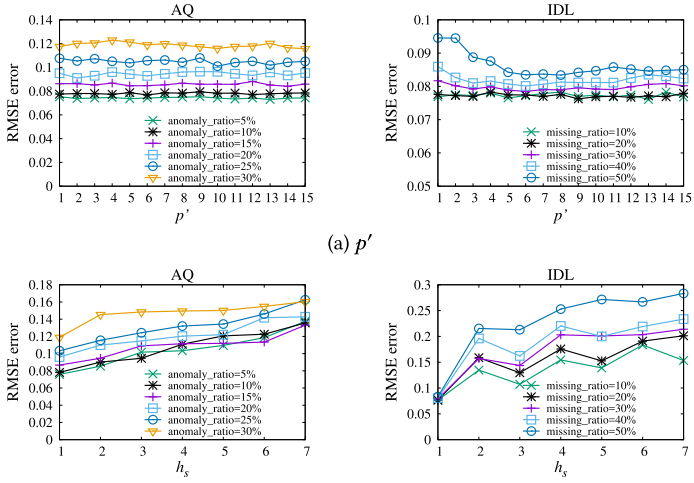


Fig. 21. Performance of MUSLES by varying p'.



(a) p'

(b) h_s

Fig. 22. Performance of SPIRIT by varying p' and h_s, respectively.

Second, in Figure 22, we study the accuracy of SPIRIT in terms of its parameters, i.e., p' and h_s. As shown in Figure 22(a), the RMSE errors are stable with the increase of p' on both AQ and IDL datasets, and thus we adopt $p' = 6$ as the default value which is consistent with MUSCLES. On the other hand, we adopt the default value of h_s as 1 because the lowest RMSE is achieved by SPIRIT when $h_s = 1$ (as shown in Figure 22(b)).

Finally, for TKCM, there are four parameters, i.e., L, d', l, and k', and Figure 23 illustrates the imputation accuracies of TKCM by varying these parameters with various anomaly ratios and missing ratios over AQ and IDL datasets, respectively. Based on the experimental results, we adopt the default values of the parameters as $L = 50$, $d' = 2$ and $k' = 2$ for both AQ and IDL datasets. Moreover, we set $l = 3$ and $l = 9$ for AQ and IDL dataset, respectively.

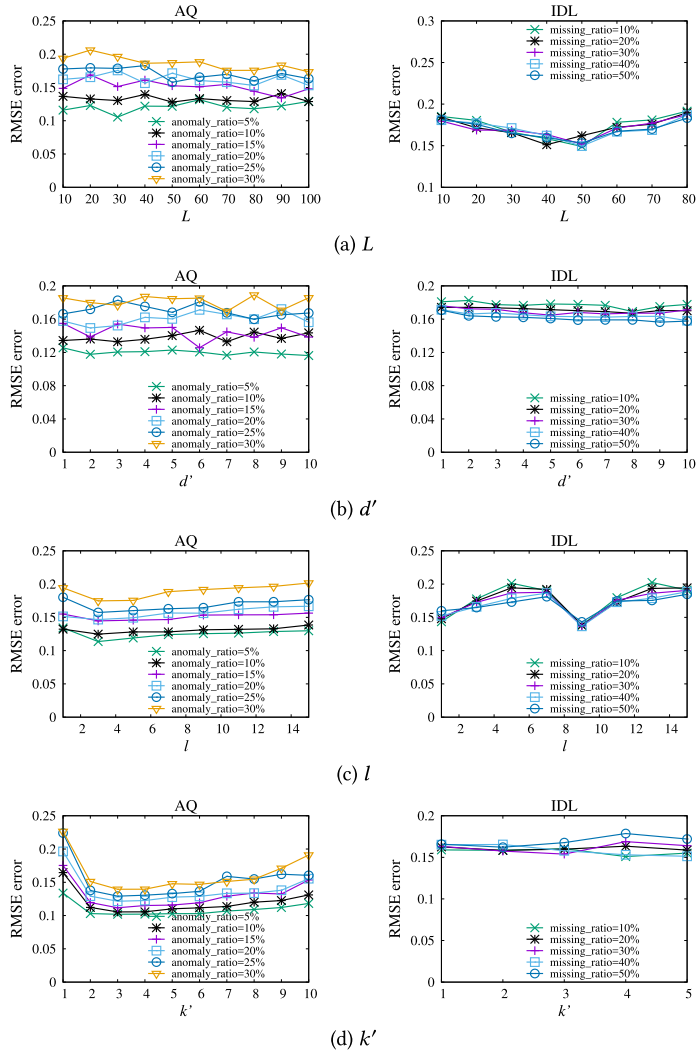


Fig. 23. Performance of TKCM by varying L, d', l, and k', respectively.

REFERENCES

[1] A. Tero. 2010. Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Briefings in Bioinformatics* 11, 2 (2010), 253–264.

[2] R. B. Hamed and C. Fazli. 2018. GOOWE: Geometrically optimum and online-weighted ensemble classifier for evolving data streams. *ACM Transactions on Knowledge Discovery from Data* 12, 2 (2018), 25:1–25:33.

[3] B. Jyoti. 2007. Time series anomaly detection using multiple statistical models. US Patent 7,310,590.

[4] C. Beidi and S. Anshumali. 2018. Densified winner take all (WTA) hashing for sparse datasets. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. 906–916.

[5] C. Carmela, F. Agostino, and P. Clara. 2019. Bursty event detection in twitter streams. *ACM Transactions on Knowledge Discovery from Data* 13, 4 (2019), 41:1–41:28.

[6] C. Huanhuan, T. Peter, R. Ali, and Y. Xin. 2013. Learning in the model space for cognitive fault diagnosis. *IEEE Transactions on Neural Networks and Learning Systems* 25, 1 (2013), 124–136.

[7] C. Huanhuan, T. Peter, R. Ali, and Y. Xin. 2013. Model-based kernel for efficient time series analysis. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 392–400.

[8] C. Nan, L. Chaoguang, Z. Qiuhan, L. Yu-Ru, Tand Xian, and W. Xidao. 2017. Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 23–33.
[9] C. Varun, B. Arindam, and K. Vipin. 2009. Anomaly detection: A survey. *ACM Computing Surveys* 41, 3 (2009), 1–58.
[10] K. G. Derpanis. 2010. Overview of the RANSAC algorithm. *Image Rochester NY* 4, 1 (2010), 2–3.
[11] A. F. Martin and C. B. Robert. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24, 6 (1981), 381–395.
[12] G. Eshel. 2003. The yule walker equations for the AR coefficients. *Internet Resource* 2 (2003), 68–73.
[13] G. João, Z. Indre, B. Albert, P. Mykola, and B. Abdelhamid. 2014. A survey on concept drift adaptation. *ACM Computing Surveys* 46, 4 (2014), 44:1–44:37.
[14] G. Wensheng, C. W. L. Jerry, F. V. Philippe, C. Han-Chieh, and S. Y. Philip. 2019. A survey of parallel sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data* 13, 3 (2019), 25:1–25:34.
[15] G. Zhabiz, Z. Xingquan, H. Arthur, and C. Michael. 2020. Deep learning for user interest and response prediction in online display advertising. *Data Science and Engineering* 5, 1 (2020), 12–26.
[16] J. H. David and S. M. Barbara. 2010. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modeling and Software* 25, 9 (2010), 1014–1022.
[17] M. H. Joseph. 2008. Quantitative data cleaning for large databases. Technical report, United Nations Economic Commission for Europe, 25.
[18] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman. 2010. *Probability and Statistical Inference*. Pearson/Prentice Hall, Upper saddle River, NJ, USA.
[19] Y. J. Kumar and K. B. Santosh. 2011. Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology* 2, 8 (2011), 45–50.
[20] L. Nikolay, A. Saeed, and F. Lan. 2015. Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1939–1947.
[21] M. Chris, N. Jennifer, and P. Sunil. 2010. ERACER: A database approach for statistical inference and data cleaning. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. 75–86.
[22] M. Qian, G. Yu, L. Wang-Chien, and Y. Ge. 2019. Order-sensitive imputation for clustered missing values. *IEEE Transactions on Knowledge and Data Engineering* 31, 1 (2019), 166–180.
[23] P. Gyuhae, C. R. Amanda, S. Hoon, and R. F. Charles. 2005. An outlier analysis framework for impedance-based structural health monitoring. *Journal of Sound and Vibration* 286, 1–2 (2005), 229–250.
[24] P. Peter and S. Markus. 1995. A mathematica version of zeilberger's algorithm for proving binomial coefficient identities. *Journal of Symbolic Computation* 20, 5–6 (1995), 673–698.
[25] P. Spiros, S. Jimeng, and F. Christos. 2005. Streaming pattern discovery in multiple time-series. In *Proceedings of the 31st International Conference on Very Large Data Bases*. 697–708.
[26] F. F. Ribeiro Ramos. 2003. Forecasts of market shares from VAR and BVAR models: A comparison of their accuracy. *International Journal of Forecasting* 19, 1 (2003), 95–110.
[27] T. S. Dominique, M. G. Jason, M. P. Paolo, and T. E. Stephen. 2017. Time series anomaly detection: Detection of anomalous drops with limited features and sparse examples in noisy highly periodic data. *Corr*, abs/1708.03665, http://arxiv.org/abs/1708.03665.
[28] S. Shaoxu, Z. Aoqian, W. Jianmin, and S. Y. Philip. 2015. SCREEN: Stream data cleaning under speed constraints. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 827–841.
[29] S. Xiaoyuan, G. Russell, M. K. Taghi, and N. Amri. 2011. Using classifier-based nominal imputation to improve machine learning. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 124–135.
[30] J. T. Daniel and R. L. S. Guy. 1986. Generalization of the matrix inversion lemma. *Proceedings of the IEEE* 74, 7 (1986), 1050–1052.
[31] T. Jin, J. Bo, Z. Aihua, and L. Bin. 2012. Graph matching based on spectral embedding with missing value. *Pattern Recognition* 45, 10 (2012), 3768–3779.
[32] D. V. Saverio, M. Ettore, P. Marco, M. Luca, and D. F. Girolamo. 2008. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical* 129, 2 (2008), 750–757.
[33] W. Heng and A. Zubin. 2015. Concept drift detection for streaming data. In *Proceedings of the 2015 International Joint Conference on Neural Networks*. 1–9.
[34] W. Kevin, H. B. Michael, D. Anton, G. Johann, and M. Hannes. 2017. Continuous imputation of missing values in streams of pattern-determining time series. In *Proceedings of the 20th International Conference on Extending Database Technology*. 330–341.
[35] Y. Byoung-Kee, S. D. Nikolaos, J. Theodore, H. V. Jagadish, F. Christos, and B. Alexandros. 2000. Online data mining for co-evolving time sequences. In *Proceedings of 16th International Conference on Data Engineering*. 13–22.

[36] Y. Rose, L. Yaguang, S. Cyrus, D. Ugur, and L. Yan. 2017. Deep learning: A generic approach for extreme condition traffic forecasting. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. 777–785.
[37] C. Y. Yang. 2010. Multiple imputation for missing data: Concepts and new development (Version 9.0). SAS Institute Inc, Rockville, MD, 49, 1–11 (2010), 12.
[38] Z. Aoqian, S. Shaoxu, S. Yu, and W. Jianmin. 2019. Learning individual models for imputation. In *Proceedings of 2019 International Conference on Data Engineering*. 160–171.
[39] Z. Aoqian, S. Shaoxu, W. Jianmin, and S. Y. Philip. 2017. Time series data cleaning: From anomaly detection to anomaly repairing. *Proceedings of the VLDB Endowment* 10, 10 (2017), 1046–1057.
[40] Z. Chengqi, Z. Xiaofeng, Z. Jilian, Q. Yongsong, and Z. Shichao. 2007. GBKII: An imputation method for missing values. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 1080–1087.
[41] Z. Indre and H. Jaakko. 2015. Optimizing regression models for data streams with missing values. *Machine Learning* 99, 1 (2015), 47–73.
[42] Z. Shichao. 2011. Shell-neighbor method and its application in missing data imputation. *Applied Intelligence* 35, 1 (2011), 123–133.
[43] Z. Shichao, J. Zhi, and Z. Xiaofeng. 2011. Missing data imputation by utilizing information within incomplete instances. *Journal of Systems and Software* 84, 3 (2011), 452–459.
[44] Z. Shichao, Q. Zhenxing, X. L. Charles, and S. Shengli. 2005. "Missing is useful": Missing values in cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering* 17, 12 (2005), 1689–1693.
[45] Z. Shichao, Z. Jilian, Z. Xiaofeng, Q. Yongsong, and Z. Chengqi. 2008. Missing value imputation based on data clustering. *Transactions on Computational Science* 1 (2008), 128–138.
[46] Z. Xiaofeng, Y. Jianye, Z. Chengyuan, and Z. Shichao. 2019. Efficient utilization of missing data in cost-sensitive learning. *IEEE Transactions on Knowledge and Data Engineering*, Early Access (2019), 1–1.
[47] Z. Xiaofeng, Z. Shichao, J. Zhi, Z. Zili, and X. Zhuoming. 2010. Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering* 23, 1 (2010), 110–121.
[48] Z. Xiaofeng, Z. Shichao, J. Zhi, Z. Zili, and X. Zhuoming. 2011. Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering* 23, 1 (2011), 110–121.

Received December 2019; revised May 2020; accepted July 2020