Chinese Journal of Chemical Engineering 53 (2023) 201–210

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

journal homepage: www.elsevier.com/locate/CJChE

Full Length Article

Univariate imputation method for recovering missing data in wastewater treatment process

Honggui Han*, Meiting Sun, Huayun Han, Xiaolong Wu, Junfei Qiao

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China
Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124, China

ARTICLE INFO

Article history:
Received 13 September 2021
Received in revised form 5 January 2022
Accepted 5 January 2022
Available online 22 April 2022

Keywords:
Univariate
Self-similarity
Waste water
Algorithm
Integration

ABSTRACT

High-quality data play a paramount role in monitoring, control, and prediction of wastewater treatment process (WWTP) and can effectively ensure the efficient and stable operation of system. Missing values seriously degrade the accuracy, reliability and completeness of the data quality due to network collapses, connection errors and data transformation failures. In these cases, it is infeasible to recover missing data depending on the correlation with other variables. To tackle this issue, a univariate imputation method (UIM) is proposed for WWTP integrating decomposition method and imputation algorithms. First, the seasonal-trend decomposition based on loess method is utilized to decompose the original time series into the seasonal, trend and remainder components to deal with the nonstationary characteristics of WWTP data. Second, the support vector regression is used to approximate the nonlinearity of the trend and remainder components respectively to provide estimates of its missing values. A self-similarity decomposition is conducted to fill the seasonal component based on its periodic pattern. Third, all the imputed results are merged to obtain the imputation result. Finally, six time series of WWTP are used to evaluate the imputation performance of the proposed UIM by comparing with existing seven methods based on two indicators. The experimental results illustrate that the proposed UIM is effective for WWTP time series under different missing ratios. Therefore, the proposed UIM is a promising method to impute WWTP time series.

1. Introduction

As water resources dwindle and freshwater demand increases in urban centers, wastewater treatment process (WWTP) plays a vital role in removing the organic wastes and producing the renewable water source [1–3]. For WWTP, various data collected from sensors are expected to provide complete and reliable state information for monitoring, control, and prediction [4–6]. Generally speaking, the data available for analysis requires both accuracy and integrity. Unfortunately, missing values are inevitable due to common mode failures, for example, network collapses, communication errors and transmission failures [7,8]. In these cases, the corresponding sensors are unable to collect data at the same time. If missing values cannot be dealt with in an appropriate manner, they will affect the effectiveness of treatment process, even lead to faulty conclusions [9]. Besides, the variables of WWTP have typical periodic variation trends, for example, influent flow rate, concentrations and temperature [10]. In addition to time-related trends, WWTP is also disturbed by heavy rains, snow melt and different wastewater composition, which leads to the uncertainty of WWTP data [11]. Due to the nonlinear and nonstationary characteristics of WWTP data, it causes more difficulty in handling missing values. Therefore, missing values processing for WWTP is an important and necessary pre-processing step before applying the data analysis algorithms.

In the past few decades, the missing data problem has attracted the attention of researchers. The typical processing methods mainly are deletion and imputation. Discarding incomplete data is arguably the most common approach due to its simplicity. However, simply discarding missing data can lead to information loss which may reduce computational efficiency [12]. In addition, it can also lead to biased and unreliable results of data-driven models due to the reduced sample size [13]. Rather than discarding missing data, imputation techniques keep the full sample size by

* Corresponding author at: Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China. Fax +86 10 67391631.
E-mail address: Rechardhan@sina.com (H. Han).

https://doi.org/10.1016/j.cjche.2022.01.033
1004-9541/

replacing missing data with estimated suitable values based on available information [14,15]. Imputation is a well-known problem in numerical methods and has been investigated using different approaches.

Recent years have witnessed a rapid development of imputation methods. Rustum and Adeloye [16] used multi-layer perception to handle missing values of activated sludge data, and the results confirmed that this method was effective. In Ref. [17], Tabari and Talaee employed artificial neural networks to successfully recover missing values of 13 water-quality parameters at five monitoring stations in the South of Iran. The work in Ref. [18] built an unsupervised deep autoencoder framework to impute patient records. Shang *et al.* [19] proposed a hybrid method for missing traffic data imputation. This method chose fuzzy c-means as the basic algorithm which can be optimized by a combination of particle swarm optimization and support vector regression. Li *et al.* [20] addressed the building fault detection and diagnosis with incomplete data by the adjacent information recovery filter. In Ref. [21], a novel fuzzy clustering algorithm integrating the grey theory concept was designed. Compared with five existing imputation approaches, the fuzzy clustering method was superior in terms of three evaluation criteria in most cases. These above methods are based on the correlations between variables, and estimate missing values based on the observed values of non-missing variables. However, when common mode failures occur, the related variables are incapable to provide values at the same time. In this case, these multivariate imputation methods are infeasible due to the unavailability of the relationship of variables.

To tackle the above mentioned issues, plenty of research efforts have been undertaken to impute the incomplete values of a variable by only relying on non-missing values of the variable being analyzed [22–24]. Moritze *et al.* [25] pointed out that univariate imputation is a particularly challenging task. In Refs. [26,27], researchers proposed to use the mean, the median or the last observation carried forward of available values to substitute each missing values. Despite their simplicity, these algorithms can lead to bias because they fill missing value with the same statistical results without considering the non-stationarity of the missing values. Autoregressive integrated moving average model (ARIMA) [28] and weighted moving average (MA) are popular imputation techniques. However, the imputation performance may be poor if ARIMA and MA are applied to time series with high nonlinearity. Bashir and Wei [29] developed an improved vector autoregressive imputation method combining an expectation minimization algorithm with the prediction error minimization method. However, stationarity is one of the main limitations of the algorithm. In addition, other imputation techniques such as linear interpolation [30] and spline interpolation [31] are inaccurate for nonlinear data. Bokde *et al.* [32] extended the pattern sequence-based forecasting algorithm by characterizing the repetitive patterns of existing observations. In Ref. [33], an STL based seasonal moving window algorithm (SWMA) was presented to search the most similar window. Phan *et al.* [34] applied dynamic time warping (DTW) to elastic match between two subsequences which could find the similar pattern. In Ref. [35], Chiewchanwattan *et al.* presented a pattern characterization approach to fill one missing point by sliding window to retrieve the closest value. Liu *et al.* [36] proposed an iterative framework based on seasonal and trend decomposition using loess (STL), which can impute the most appropriate value to ensure the integrity of monitoring sensor data. The experiment results demonstrated that the proposed method gave more accurate results than comparative methods when dealing with large gaps. Although the above methods achieved reliable results, the common weaknesses of these methods are that they require long running time to search the historical data, and most of these methods are based on a prior assumption that original data is recurrent. However, the actual natures of WWTP data are not completely known in advance. Thus, the nonlinear and nonstationary characteristics of WWTP time series should be considered simultaneously.

According to the review of the existing methods, most of these mentioned methods have limited success in WWTP due to the inability of considering both common mode failures and the characteristics of data. For time series, it is promising to combine decomposition model and imputation model. Cleveland *et al.* [37] proposed STL which can be used to decompose time series into the seasonal, trend, and remainder components. Taking into account the characteristics of STL, it has been extensively applied in the data analysis in different fields [38–40]. According to the previous studies, the most important merit of STL is that STL can produce robust sub-series due to its strong resilience to outliers in time series [41]. In addition, STL can decompose time series with seasonal frequency greater than one [42]. Meanwhile, the modeling characteristic based on numerical method makes STL easy to be applied [43].

Encouraged by the analysis above, a novel univariate imputation method is proposed by combining decomposition method and imputation algorithms. In order to fully consider the nonstationary characteristics of WWTP data, STL is chosen as the decomposition method. The three components of time series can be extracted by the decomposition step and easily be imputed based on their respective characteristics. The trend component represents the long-term change of variables influenced by the comprehensive effect of contaminant transformation and the initial quality of the water [44]. The seasonal component takes into account periodic characteristics influenced by diurnal and seasonal trends. It means the seasonal component has a regular repeated patterns within the time series. Furthermore, the remainder component reflects the random fluctuations of time series when WWTP is subjected to strong interferences such as weather changes, flow of influent water and composition of the influent. Considering the time-varying effects and influencing factors, support vector regression (SVR) has excellent fitting capabilities for the nonlinear trend and remainder components. Self-similarity decomposition (SSD) is designed to recover the seasonal component based on the intra-interval fluctuations. The main contributions and novelties of the paper are three-fold:

(1) A univariate imputation method (UIM) for WWTP imputation is provided. It is a "decomposition-single imputation-summation" procedure. In UIM, STL is treated as the decomposition procedure to process nonstationary characteristics, so as to effectively extract three robust components.
(2) Appropriate imputation operations are designed based on different characteristics of the seasonal, trend and remainder components to reduce computation time and ensure imputation accuracy.
(3) The proposed UIM is tested with six different time series of WWTP. The experimental results show that the proposed UIM is not only effective for time series with different characteristics, but also can achieve a trade-off between imputation accuracy and running time.

This article is organed as follows. The data source is briefly introduced in Section 2 along with patterns of missing WWTP data. In Section 3, an overall schematic of the imputation model is first presented, then the detailed execution steps are introduced. The experimental protocol for the imputation demonstration is given in Section 4. Finally, experimental results and discussions are provided in Section 5 and some conclusions are drawn in Section 6.

2. Preliminaries

In this section, the data imputation backgrounds of WWTP are presented, including data source and patterns of missing WWTP data.

2.1. Data source

It is well known that WWTP has been widely used to guarantee water quality for environmental protection. The activated sludge process (ASP) is the mostly used technology for WWTP. The ASP uses activated sludge including aerobic bacteria to remove nitrogen and phosphorus from wastewater [45]. The treatment process includes a primary clarifier, activated sludge reactors and a secondary clarifier. The wastewater is first preliminarily prescreened and sedimented through the primary clarifier. After the primary clarifier, the first two activated sludge tanks are respectively anaerobic and anoxic, which mainly mix the activated sludge and wastewater using mechanical mixers. The main function of the last three aerobic tanks is to degrade pollutants in wastewater through bacterial biomass. Finally, the mixture is sent to the secondary clarifier where the sludge is pumped into the activated sludge reactors and the treated wastewater is discharged into rivers. A general WWTP based on activated sludge is shown in Fig. 1.

In the activated sludge WWTP, collection of the data is implemented through deploying sensors on all the key infrastructures. The dataset includes a large volume of monitored variables such as temperature, pH, ammonia nitrogen (NH_4), dissolve oxygen, potential of hydrogen, activated sludge concentrations and so on [46]. Then, the collected data are transmitted to the information processing system through various transmission technologies such as Bluetooth, Wi-Fi and so on. Finally, the processing data are delivered and stored to database. However, the problem of missing data is often encountered during data acquisition, transmission, storage and processing. Missing values make the treatment process unmonitored and uncontrollable, and even decrease the efficiency and stability of WWTP. Hence, treating missing data is of paramount importance to ensure the accuracy and integrity of WWTP data.

2.2. Patterns of missing WWTP data

To develop an effective imputation method for WWTP, the first step is to identify the causes for the missing data. However, it is a difficult task to find the missing reasons because WWTP time series have a complex distribution. Therefore, understanding the generation process of missing data is necessary. As proposed by Rubin *et al.* [47], missing data can be regarded as in a probabilistic framework on missing data relations. Here are three common patterns of missing WWTP data as follows:

(1) Missing completely at random (MCAR): The MCAR mechanism assumes that the missing values are completely independent of each other. Therefore, data loss usually occurs when the treatment process runs into network collapses, connection errors, transformation failures and saving failures.
(2) Missing at random (MAR): The MAR mechanism supposes the missing values are only related to the observed data. For instance, when the sensor fail of influent flow rate takes place, the recycle flow rate is more likely to be missing or possible cause gross failure without regular properly adjustment.
(3) Not missing at random (NMAR): The NMAR mechanism supposes the missing values depend on both the other observed variables and missing variable itself.

In view of the above analysis, this study only focuses on missing data imputation under the MCAR type.

3. Methodology

In this section, an overall summary of the proposed UIM is given firstly. Then, the specific steps of the proposed UIM are presented in detail.

3.1. The proposed UIM

Consider a univariate time series $X(t)=\{x_t \mid t = 1, 2, \cdots, N\}$, where x_t is the observation value at time t and N is the total number of observation values in time occurring in uniform intervals. When some observation values x_t for $1 \leq t \leq N$ are missing, the time series $X(t)$ is referred to as an incomplete time series $X_G(t)$. Given that G represents the set of missing value indexes and M represents the number of gaps in $X_G(t)$. As shown in Fig. 2, the proposed UIM consists of three steps as follows:

Step 1: Decomposition. As STL requires a complete time series, linear interpolation is performed first [30]. After linear interpolation initialization, the time series $X_G'(t)$ is decomposed into the trend $T(t)$, remainder $R(t)$ and seasonal $S(t)$ components using the STL technique $X_G'(t) = T(t) + S(t) + R(t)$, $t = 1, 2, \cdots, N$. After decomposition, three missing values sub-series in the regions corresponding to the missing value indexes G are represented as $T_G(t)$, $S_G(t)$, and $R_G(t)$, respectively.

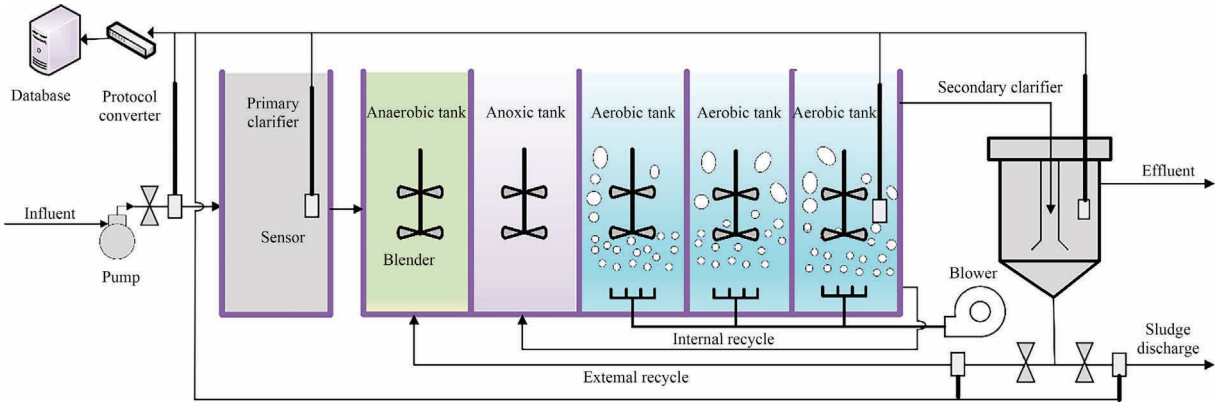

Fig. 1. The activated sludge WWTP.

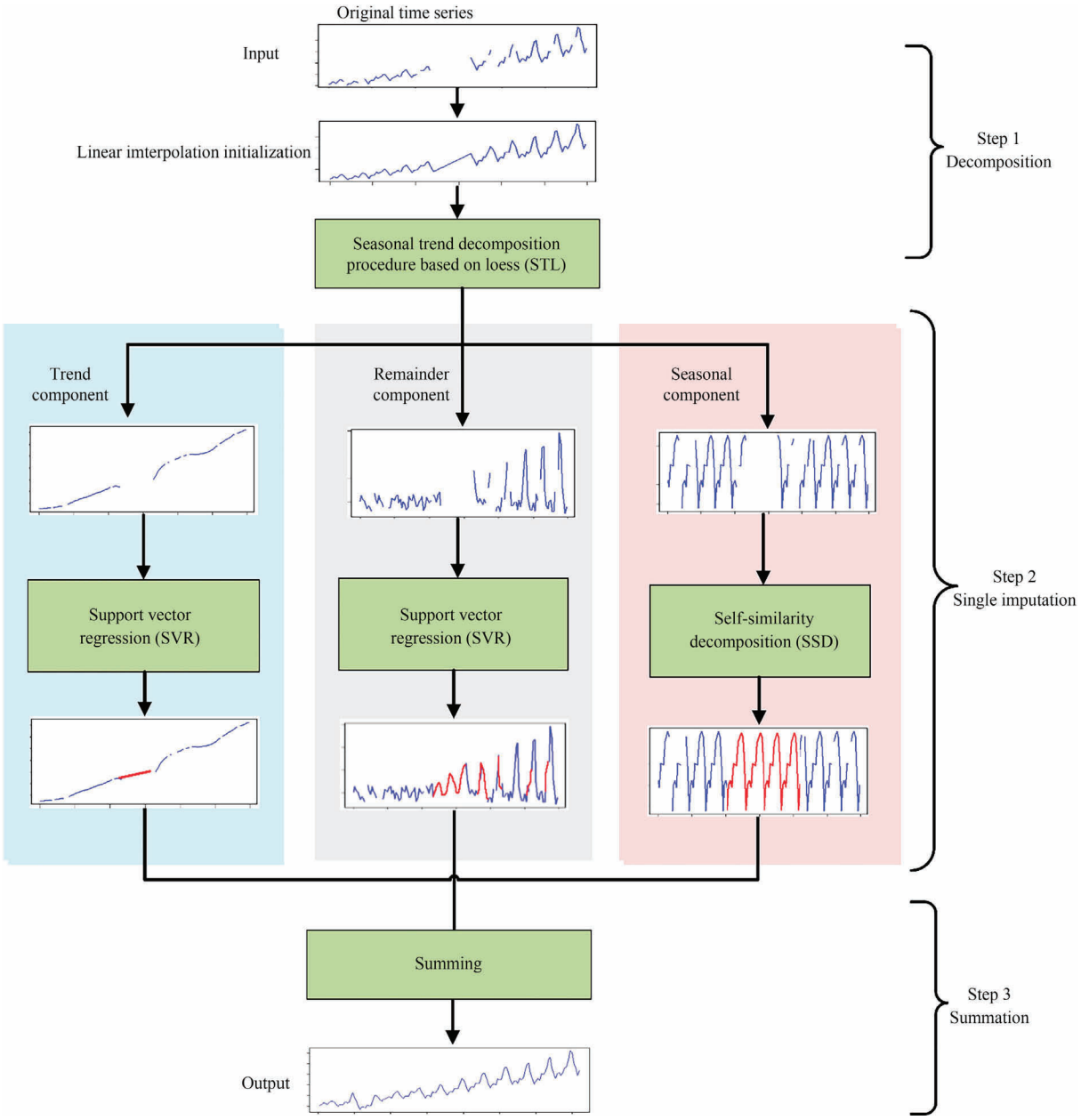

Fig. 2. Schematic of the UIM scheme.

Step 2: Single imputation. SVR is used to impute the trend component $T_G(t)$ and the remainder component $R_G(t)$. SSD is utilized to fill the seasonal component $S_G(t)$.

Step 3: Summation. The imputation results of the trend component $\widehat{T}_G(t)$, the remainder component $\widehat{R}_G(t)$, and the seasonal component $\widehat{S}_G(t)$ are summed to produce a final imputed result of the missing time series as $\widehat{X}_G(t)=\widehat{T}_G(t)+\widehat{R}_G(t)+\widehat{S}_G(t)$.

3.2. Decomposition of original time series

In WWTP, the influent flow rate, concentrations and temperature (*e.g.*, weekdays and weekends, dry weather influent and storm weather influent) are varied, which leads to a stable and stochastic influent demand, respectively. To separate useful component information, STL is adopted in this paper. STL can be beneficial to handle with WWTP data especially when missing data and noisy observations are inevitable. Also, the original time series can be decomposed into robust additive components through STL. Therefore, the imputation task is transformed into the imputation of different components.

Given a simply initialized WWTP time series $X_G'(t)$, $X_G'(t)$ can be decomposed by STL into the three additive components including the seasonal component $S(t)$, trend component $T(t)$, and remainder component $R(t)$ with: $X_G'(t) = S(t) + T(t) + R(t)$. The decomposition process is performed iteratively by inner loop and outer loop. At each iteration, each pass of the inner loop mainly applies seasonal smoothing to update the seasonal component, and uses trend smoothing to update the trend component. An iteration of the outer loop is composed of one iteration of the inner loop, which is used to adjust the robustness weight and calculates the remainder component. The robustness weights are employed in the next operation of the inner loop to reduce the effect of transient and aberrant behavior on the seasonal and trend components.

As for the imputation of the three components, three sub-series in the regions corresponding to the missing value indexes G are

represented as $T_G(t)$, $S_G(t)$ and $R_G(t)$, respectively. To increase the accuracy of imputation results, it is essential to select appropriate imputation methods based on the characteristics of each component.

3.3. Imputation of the seasonal component

Given the three sub-series obtained by STL, it is a vital issue to select a suitable imputation method for each component. Taking into full consideration of intra-interval fluctuations, the seasonal component has a regular and repeated pattern of peaks and valleys. Based on this fact, self-similar decomposition (SSD) is designed for imputing the seasonal component $S_G(t)$. SSD can detect and locate repeated subsequences which is conducted as primitives for imputation. To implement SSD, it is assumed that reference subsequence $S_C(t)$ is the longest consecutive non-missing subsequence in $S_G(t)$. Then, the main steps of SSD can be summarized as follows:

Step 1: Segment reference subsequence. Decomposing reference subsequence $S_C(t)$ is the first step of SSD. All the maxima x_t^{up} and minima in the $S_C(t)$ are identified to define the borders of segments. Then, $S_C(t)$ can be partitioned into s segments based on maxima x_t^{up} and minima x_t^{low}. The segments are denoted as $S_C(t)=\{v_i(t)|\ i = 1,2,\cdots,s\}$, where $v_i(t)$ is the ith segment. To be more concrete, every segment is enclosed between one maximum x_t^{up} and minimum x_t^{low} or *vice versa*. By this means, the segmentation consists of breaking down $S_C(t)$ into non-overlapping parts. As a conclusion of the statements above, the segment can be represented as:

$$v_i(t) = \left\{x_t^{i\text{up}}, x_t^{i2}, \cdots, x_t^{i\text{low}}\right\} \text{ or } v_i(t) = \left\{x_t^{i\text{low}}, x_t^{i2}, \cdots, x_t^{i\text{up}}\right\}, \tag{1}$$

where $x_t^{i\text{up}}$ is the maximum value of the ith segment, $x_t^{i\text{low}}$ denotes the minimum value of the ith segment. All segments preserve their position in the $S_C(t)$. After obtaining the initial set of segments, similarity measure is used to compare the segments quantitatively in order to find similar segments.

Step 2: Similarity measure. Segments are clustered based on the self-similarity of their statistics. Using the Pearson correlation coefficient, the similarity of segment can be calculated by:

$$R\left(v_i(t), v_j(t)\right) = \frac{n\sum_{k=1}^{n} v_{ik}(t)\,v_{jk}(t) - \left(\sum_{k=1}^{n} v_{ik}(t)\right)\left(\sum_{k=1}^{n} v_{jk}(t)\right)}{\sqrt{\left[n\sum_{k=1}^{n} v_{ik}(t)^2 - \left(\sum_{k=1}^{n} v_{ik}(t)\right)^2\right]\left[n\sum_{k=1}^{n} v_{jk}(t)^2 - \left(\sum_{k=1}^{n} v_{jk}(t)\right)^2\right]}}, \tag{2}$$

where $v_i(t)$ is the ith segment, n is the length of segment. A higher Pearson correlation coefficient ($R \in [-1,1]$) highlights positive linear correlation.

Step 3: Segments aggregation. To detect the primitive, segments are aggregated after similarity measures. The aggregation process starts with the first segment $v_1(t)$. When $R(v_1(t),\ v_j(t)) \geq 0.98$, $v_1(t)$ and $v_j(t)$ are marked as similar segments. Otherwise, segments are not similar. The recursion goes s depth for K_1 until no more connected similar segments are detected. If the number of similar segments is greater than 1, all segments are marked as K_1. Next, the segment $v_2(t)$ is searched. If segment $v_2(t)$ is not similar to the previous segments, the segment $v_2(t)$ is considered as a new original shape. The recursive procedure of segment $v_2(t)$ repeats the same procedure of segment $v_1(t)$ and is marked as K_2. The processing is executed until all segments are marked with similar shape index. Finally, the adjacent but dissimilar segments are aggregated which produces the most basic shape of $S_C(t)$. We denote the most basic subsequence as H operator. the H operator is used to impute missing values of the seasonal component.

Step 4: Impute missing values. The missing gap length L_i, where $i \leq M$, is compared with the H operator length l, and the missing gaps are filled in different ways. For smaller missing gaps with the missing gap length L_i being less than $l/3$, mean interpolation is performed. SSD is performed when the length of missing gap L_i is greater than $l/3$. Finally, the seasonal component $S_G(t)$ can be imputed with.

$$\widehat{S}_G(t) = \begin{cases} \bar{X}_{-G}(t)\ , & \text{if } L_i < \frac{l}{3} \\ \text{rep}(H\ \text{ operator},\ \lceil \frac{L}{l} \rceil\ \ ), & \text{if } L_i \geqslant \frac{l}{3} \end{cases}, \tag{3}$$

where $X_{-G}(t)$ refers as all elements in $X(t)$ except the missing value indexes G, l is the length of the H operator, L_i is the length of the ith missing gap, $\lceil\cdot\rceil$ is integer ceiling function, rep$(\cdot)$ is the function used to replicate the H operator.

3.4. Imputation of the trend and remainder components

Support vector regression (SVR) model, a non-parametric modelling technique, employs the principle of structural risk minimization to replace the empirical risk of traditional neural network [48]. Moreover, SVR has a good performance in dealing with nonlinear regression problems. Taking the nonlinear characteristics of trend and remainder components into account, SVR is applied to impute missing values in $T_G(t)$ and $R_G(t)$.

Since there are M missing gaps in $R_G(t)$, the $R_G(t)$ can be segmented into M + 1 sub-series. Consider each missing gap, an autoregressive model is constructed to impute the remainder component $R_G(t)$ since it is a univariate time series. Thus, SVR uses the previous L-order observation values to impute the missing values:

$$x_t = \theta_1 x_{t-1} + \theta_2 x_{t-2} + \cdots + \theta x_{t-L} + e_t, \tag{4}$$

where x_{t-L} is the value before missing gap of the remainder component $R_G(t)$, e_t is the error term at tth point.

The training set of each sub-series is constructed with the corresponding L-order variables of the elements in the label set.

$$\boldsymbol{T} = \begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ x_2 & x_3 & \cdots & x_{L+1} \\ \vdots & \vdots & & \vdots \\ x_{t-L-1} & x_{t-L} & \cdots & x_{t-2} \end{bmatrix},\ \boldsymbol{I} = \begin{bmatrix} x_{L+1} \\ x_{L+2} \\ \vdots \\ x_{t-1} \end{bmatrix}, \tag{5}$$

where $\boldsymbol{T}$ is the training set, $\boldsymbol{I}$ is the label set, and x_t is a value before the missing gap of the remainder component.

By using autoregressive model to construct the imputation function, and the optimization model is given as follows:

$$\min \frac{1}{2}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right) + C\sum_{t=L}^{N-1} \left|x_t - f(x_{t-1}, \cdots, x_{t-L})\right|_{\varepsilon}, \tag{6}$$

where $\boldsymbol{w}$ is a regression coefficient vector, C is a positive constant that determines the degree of penalized loss when a training error occurs, ε is an error tolerance, N is the number of non-missing values in each sub-series. In the situation that Eq. (6) is infeasible, the slack variables can be introduced to obtain Eq. (7).

$$\min \frac{1}{2}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right) + C\left(\sum_{t=L}^{N-1} \left(\varepsilon_t + \varepsilon_t^*\right)\right), \tag{7}$$

$$\text{s.t.} \begin{cases} x_t - f(x_{t-1}, \cdots, x_{t-L}) \leqslant \varepsilon + \xi_t \\ f(x_{t-1}, \cdots, x_{t-L}) - x_t \leqslant \varepsilon + \xi_t^*, \\ \xi_t, \xi_t^* \geqslant 0 \end{cases} \tag{8}$$

where ξ_t and ξ_t^* are slack variables that deal with the corresponding positive and negative errors when subject to an error tolerance ε at the tth point, respectively.

The remainder component $R_G(t)$ of WWTP time series is nonlinear. In experiments, we adopt the widely used the redial basis function as nonlinear function. The decision function of missing values is as follows:

$$\hat{R}_G(t) = f(x_{t-1}, \cdots, x_{t-L}) = \sum_{t=1}^{p} \boldsymbol{w}_t \cdot \exp\left(-\gamma \cdot x_t - x^2\right) + b, \tag{9}$$

where γ is a kernel-specific parameter, $\boldsymbol{w}_t$ is the coefficient of imputation function and b is a bias term.

The same imputation procedures are performed for the trend component $T_G(t)$ using SVR. Upon the finish of SVR, the imputed result of $T_G(t)$ can be expressed as $\widehat{T}_G(t)$. Therefore, three imputed results have so far been calculated by SSD and SVR. Finally, the overall results will be merged to achieve the data imputation.

4. Experiment Protocol

In this section, the data description, the algorithms used for comparison, the performance indicators and the detailed experimental process are presented, respectively.

4.1. Data description

Two public datasets of WWTP are used to conduct all experiments. One dataset is collected from the wastewater treatment plant in Barcelona, Spain. The dataset is publicly available in the University of California at Irvine machine learning repository (UCI). Another dataset is NH_4 concentration sampled from 30.11.2010-16:10 to 01.01.2011-6:40 every 10 min. It can be obtained from imputeTS package in R software, which is open source and freely available. Usually, the experiment results are more significant if more different datasets are compared. The statistical features are described in Table 1.

The standard deviation of input sediments to primary settler (SED.P) and output sediments (SED.S) is large, indicating that these time series have large dispersion. The stationary of each time series is confirmed by an augmented dickey-fuller (ADF) test. The results of input pH to plant (PH.E), input sediments to primary settler (SED.P), and performance input biological demand of oxygen in primary settler (RD.DBO.P) are all failed to reject the unit-root null hypothesis when the significance level is defined as 0.05, which means that they are nonstationary.

Considering that the above time series have mixed and inconsistent feature units, data normalization is indispensable before missing imputation. In this experiment, six time series are normalized using min-max normalization which can reduce data to its canonical form. The min-max normalization is expressed as follows:

$$x(t)_{\text{normalized}} = \frac{x_t - x_{\min}}{x_{\max} - x_{\min}}, \tag{10}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum of the time series, respectively. According to Eq. (10), all the time series can be transformed into values between zero and one. In practice, the final output can be de-normalized again for application.

4.2. Univariate imputation algorithms

To validate the imputation ability of the propose UIM, seven existing methods from imputeTS package [49] and DTWBI package [34] are introduced as the benchmarks for comparison. All imputation methods are written in the R programming language. The seven comparison methods are described as follows:

(1) na.interpolation: Missing values are replaced by values of linear, spline or stineman interpolation. In our experiment, we use linear interpolation (na.linear) [30] and spline interpolation (na.spline) [31]. The default parameter values of na.linear and na.spline are used.
(2) na.ma: This method refers to the weight moving average. In this experiment, we use semi-adaptive window size to replace missing values [28]. The order of moving average smoother is set to 4.
(3) na.seadec: This method is in fact an STL approach, which recovers missing values based on seasonally decomposed missing value imputation [49]. In na.seadec, the weighted moving average is utilized after decomposition.
(4) na.kalman: It uses Kalman smoothing on structural time series models for imputation [50]. A structural model fitted by maximum likelihood is used to recover missing values.
(5) na.smwa: Missing values are filled by the best cyclic pattern from the past available data, which is known as SMWA. When implementing na.smwa, the head size and tail size are set to be $w/10$ as suggested in Ref. [33], where w is the length for searching the best fitting window.
(6) DTWBI: It applies dynamic time warping to search the most similar subsequence, and uses the subsequence to complete missing values [34].

4.3. Performance criteria

For the purpose of evaluating the imputation performance of the proposed UIM quantitatively, two indicators are selecting, *i.e.*, root mean square error (RMSE) and similarity (Sim). The RMSE and Sim can be calculated as follows:

$$\text{RMSE}(\widehat{x}_t, x_t) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\widehat{x}_t - x_t\right)^2}, \tag{11}$$

$$\text{Sim}(\widehat{x}_t, x_t) = \frac{1}{N}\sum_{t=1}^{N}\frac{\max(x) - \min(x)}{(\max(x) - \min(x)) + \left|\widehat{x}_t - x_t\right|}, \tag{12}$$

where N is the number of missing values, $\widehat{x}_t$ is the imputed value, and x_t is the real value. The lower RMSE and the higher Sim highlight the better imputation performance.

4.4. Experimental process

To comprehensively evaluate the imputation methods, the experimental process consists of four steps:

Table 1
Dataset description

No.	Variable name	Length	Maximum	Minimum	Mean	Standard deviation	ADF test
1	Input pH to plant (PH.E)	200	8.5	7.3	7.83	0.24	0.07
2	Input sediments to primary settler (SED.P)	200	46.0	1.0	5.12	3.58	0.09
3	Output suspended solids (SS.S)	200	131.0	8.0	20.91	11.55	0.01
4	Output sediments (SED.S)	200	0.0	3.5	0.04	0.20	0.01
5	Performance input biological demand of oxygen in primary settler (RD.DBO.P)	200	79.1	0.6	39.13	14.93	0.07
6	NH_4 concentration (NH_4)	4552	48.2	0.7	12.4	7.7	0.01

Table 2
Average RMSE and average Sim of eight imputation methods on six time series

Missing ratio	Method	Q.E		SED.P		SS.S		SED.S		RD.DBO.P		NH_4	
		RMSE	Sim	RMSE	Sim	RMSE	Sim	RMSE	Sim	RMSE	Sim	RMSE	Sim
5%	UIM	**0.131**	**0.819**	**0.049**	**0.793**	**0.058**	0.763	**0.012**	0.759	0.159	0.802	**0.080**	**0.913**
	na.linear [30]	0.143	0.805	0.063	0.766	0.064	**0.773**	0.012	0.761	0.174	0.783	0.133	0.852
	na.spline [31]	0.241	0.710	0.147	0.566	0.234	0.464	0.022	0.598	0.660	0.521	0.444	0.676
	na.ma [28]	0.145	0.802	0.055	0.792	0.067	0.753	0.012	**0.772**	**0.151**	**0.805**	0.145	0.842
	na.seadec [49]	0.146	0.794	0.059	0.783	0.075	0.746	0.024	0.647	0.170	0.795	0.084	0.907
	na.kalman [50]	0.137	0.812	0.052	0.789	0.063	0.725	0.014	0.663	0.157	0.807	0.134	0.849
	na.smwa [33]	0.200	0.729	0.101	0.733	0.095	0.697	0.018	0.743	0.257	0.702	0.143	0.843
	DTWBI [34]	0.235	0.722	0.066	0.748	0.164	0.630	0.065	0.658	0.276	0.679	0.170	0.820
10%	UIM	**0.146**	**0.859**	**0.050**	**0.825**	**0.080**	**0.827**	**0.010**	**0.822**	**0.188**	**0.831**	**0.102**	**0.892**
	na.linear [30]	0.158	**0.859**	0.059	0.795	0.090	0.807	0.010	0.816	0.211	0.815	0.117	0.870
	na.spline [31]	0.277	0.743	0.291	0.484	0.396	0.529	0.043	0.458	0.797	0.575	0.695	0.644
	na.ma [28]	0.165	0.829	0.064	0.785	0.100	0.797	0.011	0.788	0.192	0.826	0.143	0.853
	na.seadec [49]	0.160	0.836	0.065	0.780	0.096	0.806	0.022	0.748	0.201	0.821	0.110	0.874
	na.kalman [50]	0.159	0.828	0.054	0.815	0.085	0.811	0.012	0.697	0.192	0.829	0.160	0.838
	na.smwa [33]	0.198	0.799	0.078	0.780	0.133	0.789	0.035	0.760	0.256	0.794	0.164	0.835
	DTWBI [34]	0.206	0.801	0.136	0.720	0.134	0.769	0.060	0.708	0.266	0.784	0.186	0.816
15%	UIM	**0.162**	**0.858**	**0.048**	**0.846**	**0.084**	0.845	**0.011**	**0.770**	**0.189**	**0.833**	**0.102**	**0.899**
	na.linear [30]	0.163	0.857	0.052	0.828	0.088	0.832	0.012	0.755	0.194	0.828	0.126	0.873
	na.spline [31]	0.611	0.680	0.273	0.568	0.412	0.500	0.049	0.414	1.050	0.521	3.080	0.385
	na.ma [28]	0.165	0.853	0.054	0.828	0.092	0.838	0.012	0.763	0.214	0.810	0.141	0.855
	na.seadec [49]	0.165	0.854	0.054	0.826	0.096	0.832	0.025	0.679	0.226	0.804	0.127	0.871
	na.kalman [50]	0.163	0.857	0.049	0.838	0.085	**0.847**	0.013	0.678	0.196	0.825	0.148	0.840
	na.smwa [33]	0.200	0.835	0.064	0.820	0.120	0.797	0.067	0.689	0.246	0.796	0.144	0.858
	DTWBI [34]	0.207	0.829	0.122	0.750	0.144	0.782	0.048	0.737	0.275	0.775	0.164	0.837
20%	UIM	**0.177**	**0.862**	**0.051**	**0.843**	**0.082**	**0.841**	**0.011**	**0.827**	**0.212**	**0.822**	**0.132**	**0.851**
	na.linear [30]	0.186	0.856	0.056	0.831	0.096	0.829	0.012	0.812	0.226	0.815	0.152	0.826
	na.spline [31]	0.598	0.678	0.494	0.495	1.26	0.421	0.187	0.384	3.330	0.457	1.810	0.398
	na.ma [28]	0.189	0.855	0.058	0.822	0.128	0.798	0.013	0.826	0.220	0.818	0.164	0.826
	na.seadec [49]	0.191	0.853	0.057	0.823	0.123	0.801	0.024	0.793	0.229	0.814	0.150	0.839
	na.kalman [50]	0.179	0.860	0.052	0.838	0.092	0.831	0.014	0.680	0.223	0.816	0.157	0.823
	na.smwa [33]	0.215	0.840	0.093	0.808	0.137	0.789	0.064	0.758	0.219	0.812	0.161	0.829
	DTWBI [34]	0.215	0.838	0.120	0.776	0.145	0.777	0.054	0.796	0.276	0.788	0.181	0.807

(1) Step 1: Take a complete time series.
(2) Step 2: Get incomplete time series by creating missing values randomly.
(3) Step 3: Apply the imputation methods to complete missing data.
(4) Step 4: Evaluate those methods by comparing the differences between two complete time series.

On each complete time series mentioned above, missing values are simulated using the gapCreation function in the DTWBI package by varying the amount of missing ratios from 5% to 20% at increasing intervals of 5%. The gapCreation function can randomly create a continuous missing gap within a univariate time series, and the created missing gap starts at a random location. Missing ratio is the total number of missing values over the total number of values in a time series. By comparing different missing ratios, the influence of missing ratios on imputation methods can be demonstrated. To eliminate the uncertainty of imputation results, the experiments are repeated 10 times by randomly selecting the missing locations for each running and performing 35 iterations for each time series.

5. Results and Discussion

Using the experiment protocol mentioned in Section 4, the imputation experiments for WWTP time series are conducted in this section including the results of quantitative accuracy performance, the visual shape performance, the computation time and maximum tolerable missing ratio.

5.1. Quantitative accuracy comparison

To verify the effectiveness of the presented UIM, the quantitative comparison experiments between the proposed UIM and traditional imputation methods are conducted. The average RMSE and average Sim for six time series are showed in Table 2. For each missing ratio, the best results have been highlighted in bold.

In Table 2, it is clear that the proposed UIM is effective for WWTP, relative to the other seven competitors listed in this study. In addition, the DTWBI produces less satisfactory results for all of the WWTP series. The reason behind the inferiority of DTWBI is that they need recurrent data and thus cannot match the most similar subsequence in WWTP time series. Apart from the proposed UIM, na.spline and DTWBI, other comparation models present some mixed results, which are analyzed in terms of different characteristics.

First, regrading the Q.E, SED.P, and RD.DBO.P time series which are nonstationary, UIM has the lowest average RMSE and the highest average Sim though there are a few exceptions, followed by na.karman, na.linear, na.ma, and na.seadec. In general, Kalman filter imputation performs well in data with nonstationary characteristics. The reason why na.kalman is superior to the six models is that the Kalman filter is used to handle stochastic uncertainty of the WWTP time series. Compared with na.kalman, UIM tries to fill the large gap by combining periodic information of the seasonality component and uncertainty information of the remainder component. In addition, na.ma ranks the first position when missing ratio is less than 5% on the RD.DBO.P time series. It is conceivable that the reason is that na.ma is a class of typical linear model. Thus, it can obtain better imputation results when a time series is stationary. However, na.ma cannot capture nonlinear patterns hidden in the RD.DBO.P time series when missing ratio is greater than 5%.

Second, the SED.S time series exhibits both stationary and high standard deviation. The na.linear ranks second after UIM, followed by na.ma, na.linear, and na.smwa in term of average RMSE and average Sim. Therefore, the performance of UIM and na.linear are highly convincing in terms of RMSE with large intervals of missing data.

Third, in the case of the SS.S and NH_4 time series, some similar results can be drawn. The proposed UIM performs better than the other existing methods, except for several outliers, followed by na.seadec, na.ma, na.linear, and na.kalman. Compared with na.seadec, UIM tries to fill missing values by combining the advantages of SSD and SVR.

In summary, several important conclusions can be drawn based on the analysis of the experimental results as: (1) The proposed UIM is the best-performance model, relative to the other methods listed in this study, for different WWTP series in terms of RMSE and Sim. (2) In terms of the comparison between the UIM and na.seadec, the UIM is better. This indicates that using different imputation methods (*i.e.*, SSD and SVR) after decomposition techniques can effectively improve the imputation performance. In UIM, SSD obtains periodic pattern by identifying the most basic shape of the seasonality component. And SVR considers the stochastic uncertainty of the remainder component and the nonlinearity of the trend component. (3) In view of the non-stationarity of WWTP time series, the STL procedure removes the trend and seasonal pattern in the WWTP time series to get result in stationary time series. Therefore, STL is suitable for decomposing time series to obtain robust components.

5.2. Visual shape comparison

To further verify the proposed method, Figs. 3 and 4 give the visual comparisons of different algorithms on the NH_4 time series. Due to space limitations, we cannot present all imputation methods in Fig. 4. Thus, we only show three comparative methods

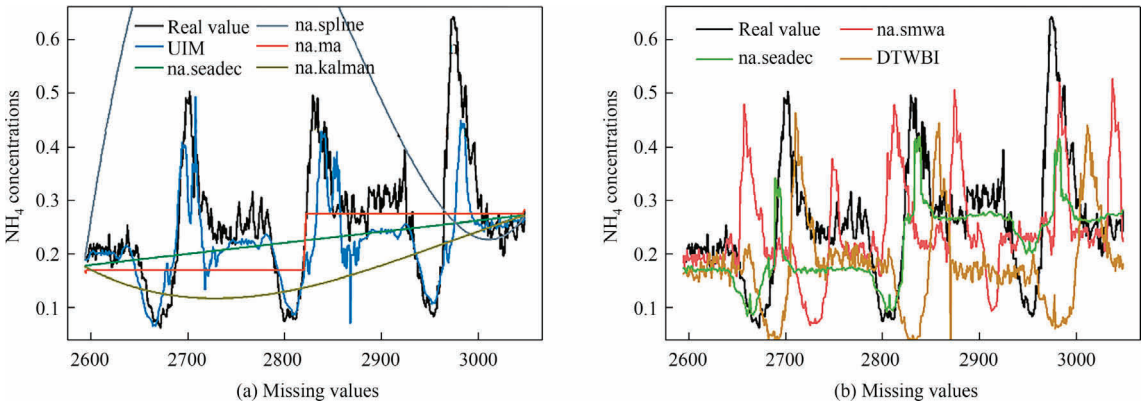

Fig. 3. Visual comparison on the NH_4 concentrations at missing ratio of 10%.

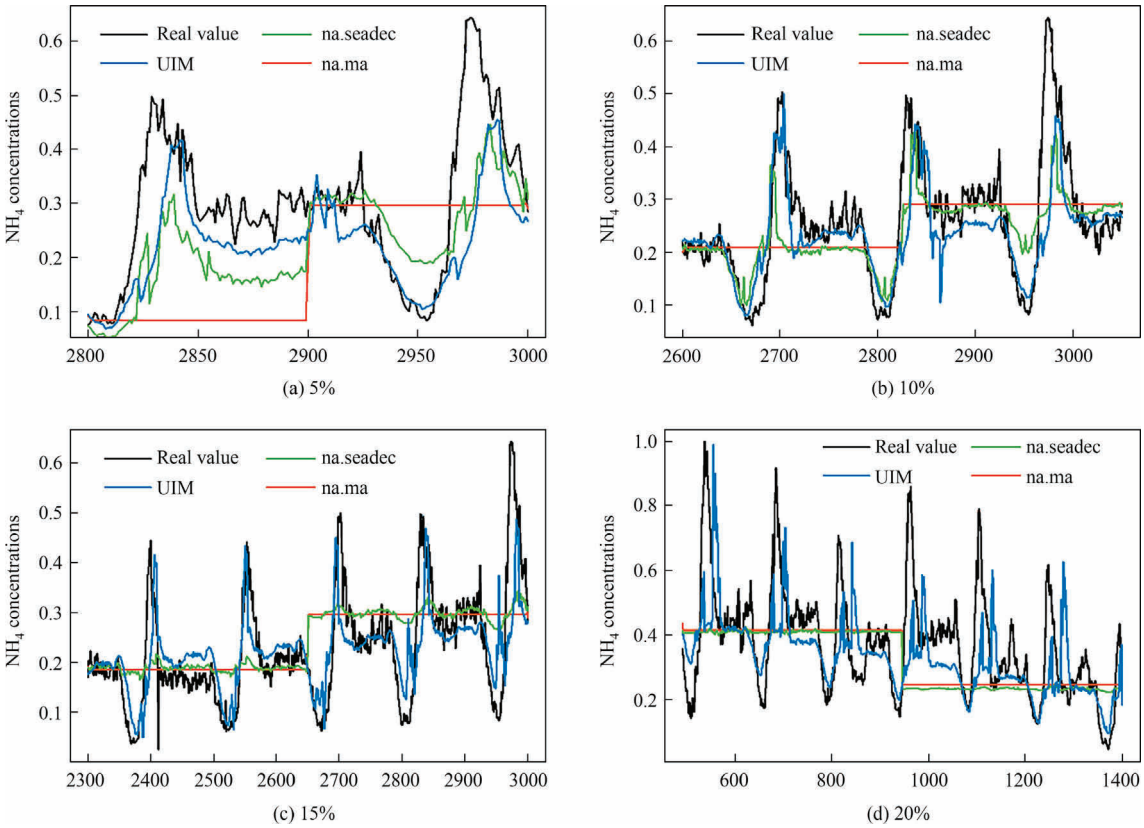

Fig. 4. Visual comparison on the NH_4 concentrations under different missing ratios.

Table 3
Computation time of different models on the NH_4 concentrations in second

Missing ratio	UIM	na.linear[30]	na.spline[31]	na.ma[28]	na.seadec[49]	na.kalman[50]	na.smwa[33]	DTWBI[34]
5%	0.53	0	0	0.03	0.03	5132.78	0.71	1.09
10%	1.76	0	0	0.80	0.73	7772.34	1.97	2.12
15%	1.92	0	0	1.04	0.95	8362.11	2.24	3.09
20%	1.97	0	0	1.09	1.35	9568.14	2.40	2.98

according to Table 2. The three comparative methods are the best methods under different missing ratios.

Fig. 3 displays that na.spline has the worst result. It can be found that the shape of na.spline is far from the shape of real values. The NH_4 series imputed by na.smwa and DTWBI models have too much fluctuation, which shows a large deviation from the real values. Conversely, the results of na.ma, na.linear, and na.kalman models are so smooth that they cannot fit fluctuation in the actual time series well. As shown in Fig. 4, the imputation values yielded by UIM are almost identical to the real values on the NH_4 time series. Taken together, the visual results confirm the effectiveness of the proposed UIM's efficacy for imputing WWTP time series.

5.3. Computation time comparison

The running time of imputation method has to be satisfactory under high accuracy. Imputation methods are conducted 30 times by randomly selecting the missing positions for each run on NH_4 series. In this case, the computational time of each method on the NH_4 series are given, as shown in Table 3.

Table 3 indicates that UIM consumes the shortest running time compared with na.kalman, na.smwa and DTWBI. Furthermore, considering the quantitative and visual performance of UIM (Table 2, Figs. 3 and 4), the required time of the proposed method is fully acceptable. Note that SSD obtains periodic pattern directly from the seasonal component, which isn't required to duplicate search of historical data.

5.4. Maximum tolerable missing ratio

To further verify the effectiveness of the propose UIM, Fig. 5 gives the average RMSE of the proposed UIM on the NH_4 concentrations under different missing ratios.

The proposed UIM can tolerate a maximum missing ratio of 45%. When missing ratio is greater than 45%, a time series does not provide enough information for the proposed UIM. Thus, the proposed UIM cannot compute an appropriate result. The DTWBI function in R package prints missing gap is too large to compute an appropriate imputation result when missing ratio is more than 30%. Compared with DTWBI, the results confirm the effectiveness of UIM's efficacy for imputing WWTP time series. Fig. 5 gives the average RMSE of the proposed UIM on the NH_4 concentrations under different missing ratios.

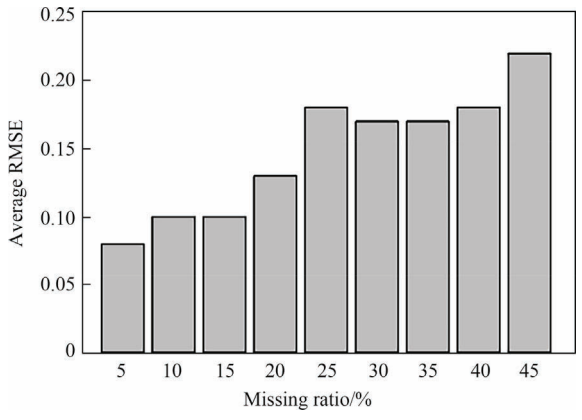

Fig. 5. The average RMSE of the proposed UIM on the NH_4 concentrations under different missing ratios.

6. Conclusions

In this paper, a univariate imputation method (UIM) with well-selected decomposition method and imputation algorithms was proposed to impute missing values of WWTP time series. UIM aims at first decomposing a time series into the seasonal, trend and remainder components, then separately deals with each component with specific imputation methods. The SSD method is designed to deal with missing values in the seasonal component by locating the primitive with repeating patterns. And, the SVR model is used to impute the missing values of the trend and remainder components. Finally, UIM integrates the processed data and outputs the imputation results. The proposed UIM has been tested using six time series of WWTP and compared with seven existing methods on accuracy and running time. The experiment results demonstrated that UIM can achieve better accuracy than the existing methods and acceptable running time. This means that the proposed UIM is effective for time series with nonlinear and nonstationary characteristics. Furthermore, the designed UIM has high practical application potentials in WWTP. In the future work, multivariate time series imputation will be investigated.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are thankful to the National Key Research and Development Project (No. 2018YFC1900800-5), the National Natural Science Foundation of China (Nos. 61890930-5, 61903010, 6202100), the Beijing Outstanding Young Scientist Program (No. BJJWZYJH01201910005020) and the Beijing Natural Science Foundation (No. KZ202110005009).

References

[1] M.A. Shannon, P.W. Bohn, M. Elimelech, J.G. Georgiadis, B.J. Mariñas, A.M. Mayes, Science and technology for water purification in the coming decades, *Nature* 452 (7185) (2008) 301–310.
[2] H.G. Han, X.L. Wu, L.M. Ge, J.F. Qiao, A sludge volume index (SVI) model based on the multivariate local quadratic polynomial regression method, *Chin. J. Chem. Eng.* 26 (5) (2018) 1071–1077.
[3] W. Wei, P.F. Xia, Z.W. Liu, M. Zuo, A modified active disturbance rejection control for a wastewater treatment process, *Chin. J. Chem. Eng.* 28 (10) (2020) 2607–2619.
[4] H.G. Han, S.G. Zhu, J.F. Qiao, M. Guo, Data-driven intelligent monitoring system for key variables in wastewater treatment process, *Chin. J. Chem. Eng.* 26 (10) (2018) 2093–2101.
[5] I. Lizarralde, T. Fernández-Arévalo, A. Manas, E. Ayesa, P. Grau, Model-based opti mization of phosphorus management strategies in Sur WWTP, Madrid, *Water Res.* 153 (2019) 39–52.

[6] H.G. Han, Z. Liu, Y. Hou, J.F. Qiao, Data-driven multiobjective predictive control for wastewater treatment process, *IEEE Trans. Ind. Inform.* 16 (4) (2020) 2767–2775.
[7] W.D. Tian, Z.J. Liu, L.N. Li, S.F. Zhang, C.K. Li, Identification of abnormal conditions in high-dimensional chemical process based on feature selection and deep learning, *Chin. J. Chem. Eng.* 28 (7) (2020) 1875–1883.
[8] P. Baraldi, F.D. Maio, D. Genini, E. Zio, Reconstruction of missing data in multidimensional time series by fuzzy similarity, *Appl. Soft Comput.* 26 (2015) 1–9.
[9] L. Rieger, I. Takács, K. Villez, H. Siegrist, P. Lessard, P.A. Vanrolleghem, Y. Comeau, Data reconciliation for wastewater treatment plant simulation studies-planning for high-quality data and typical sources of errors, *Water Environ. Res.* 82 (5) (2010) 426–433.
[10] G. Olsson, H. Aspegren, M.K. Nielsen, Operation and control of wastewater treatment—A Scandinavian perspective over 20 years, *Water Sci. Technol.* 37 (12) (1998) 1–13.
[11] H.R. Haimi, M. Mulas, F. Corona, R. Vahala, Data-derived soft-sensors for biological wastewater treatment plants: An overview, *Environ. Model. Softw.* 47 (2013) 88–107.
[12] N.M. Noor, M.M. Al Bakri Abdullah, A.S. Yahaya, N.A. Ramli, Comparison of linear interpolation method and mean method to replace the missing values in environmental data set, *Mater. Sci. Forum* 803 (2014) 278–281.
[13] G. Hawthorne, P. Elliott, Imputing cross-sectional missing data: Comparison of common techniques, *Aust. N. Z. J. Psychiatry* 39 (7) (2005) 583–590.
[14] Z.W. Wang, L. Wang, Y.Y. Tan, J.F. Yuan, Fault detection based on Bayesian network and missing data imputation for building energy systems, *Appl. Therm. Eng.* 182 (2021) 116051.
[15] Y. Deng, C. Chang, M.S. Ido, Q. Long, Multiple imputation for general missing data patterns in the presence of high-dimensional data, *Sci. Reports* 6 (2016) 21689.
[16] R. Rustum, A.J. Adeloye, Replacing outliers and missing values from activated sludge data using kohonen self-organizing map, *J. Environ. Eng.* 133 (9) (2007) 909–916.
[17] H. Tabari, P. Hosseinzadeh Talaee, Reconstruction of river water quality missing data using artificial neural networks, *Water Qual. Res. J.* 50 (4) (2015) 326–335.
[18] J.Q. da Xu, P.J.H. Sheng, T.S. Hu, C.C. Huang, Hsu, A deep learning–based unsupervised method to impute missing values in patient records for improved management of cardiovascular patients, *IEEE J. Biomed. Heal. Inform.* 25 (6) (2021) 2260–2272.
[19] Q. Shang, Z.S. Yang, S. Gao, D.R. Tan, An imputation method for missing traffic data based on FCM optimized by PSO-SVR, *J. Adv. Transp.* 2018 (2018) 2935248.
[20] D. Li, Y.X. Zhou, G.Q. Hu, C.J. Spanos, Handling incomplete sensor measurements in fault detection and diagnosis for building HVAC systems, *IEEE Trans. Autom. Sci. Eng.* 17 (2) (2020) 833–846.
[21] A.M. Sefidian, N. Daneshpour, Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model, *Expert Syst. Appl.* 115 (2019) 68–94.
[22] N. Shaadan, N.A.M. Rahim, Imputation analysis for time series air quality ($\{PM\}_{10}$) data set: A comparison of several methods, *J. Phys. Conf. Ser.* 1366 (1) (2019) 012107.
[23] A. Chaudhry, W. Li, A. Basri, F. Patenaude, A method for improving imputation and prediction accuracy of highly seasonal univariate data with large periods of missingness, *Wirel. Commun. Mob. Comput.* 2019 (2019) 4039758.
[24] S.J. Hadeed, M.K. O'Rourke, J.L. Burgess, R.B. Harris, R.A. Canales, Imputation methods for addressing missing data in short-term monitoring of air pollutants, *Sci. Total. Environ.* 730 (2020) 139140.
[25] S. Moritz, A. Sard'a, T. Bartz-Beielstein, M. Zaefferer, J. Stork, Comparison of different methods for univariate time series imputation in R, *ArXiv* 15 (2015) 3924–3944.
[26] P.D. Allison, Missing data: Quantitative applications in the social sciences, *Br. J. Math. Stat. Psychol.* 55 (1) (2002) 193–196.
[27] V. Audigier, F. Husson, J. Josse, Multiple imputation for continuous variables using a Bayesian principal component analysis, *J. Stat. Comput. Simul.* 86 (11) (2016) 2140–2156.
[28] W.O. Yodah, J.M. Kihoro, K.H.O. Arhiany, H.W. Kibunja, Imputation of incomplete non-stationary seasonal time series data, *Math. Theory Model.* 3 (12) (2013) 142–154.
[29] F. Bashir, H.L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm, *Neurocomputing* 276 (2018) 23–30.
[30] L. Bigaignon, R. Fieuzal, C. Delon, T. Tallec, Combination of two methodologies, artificial neural network and linear interpolation, to gap-fill daily nitrous oxide flux measurements, *Agric. For. Meteorol.* 291 (2020) 108037.
[31] L. Greco, M. Cuomo, B-Spline interpolation of Kirchhoff-Love space rods, *Comput. Methods Appl. Mech. Eng.* 256 (2013) 251–269.
[32] N. Bokde, F.M. Álvarez, M.W. Beck, K. Kulat, A novel imputation methodology for time series based on pattern sequence forecasting, *Pattern Recognit. Lett.* 116 (2018) 88–96.
[33] S. Chandrasekaran, S. Moritz, M. Zaefferer, J. Stork, T. Bartz-Beilelstein, Data preprocessing: A new algorithm for univariate imputation designed specifically for industrial needs, in: Proceedings of the Workshop Computational Intelligence, Dortmund, Germany, 2016, 1–12.
[34] T.T.H. Phan, É. Poisson Caillault, A. Lefebvre, A. Bigand, Dynamic time warping-based imputation for univariate time series data, *Pattern Recognit. Lett.* 139 (2020) 139–147.
[35] S. Chiewchanwattana, C. Lursinsap, C.H. Henry Chu, Imputing incomplete time-series data based on varied-window similarity measure of data sequences, *Pattern Recognit. Lett.* 28 (9) (2007) 1091–1103.
[36] Y.H. Liu, T. Dillon, W.J. Yu, W. Rahayu, F. Mostafa, Missing value imputation for industrial IoT sensor data with large gaps, *IEEE Internet Things J.* 7 (8) (2020) 6855–6867.
[37] R.B. Cleveland, W.S. Cleveland, STL: A seasonal-trend decomposition procedure based on Loess (with discussion), *J. Off. Stat.* 6 (1) (1990) 3–33.
[38] H.T. He, S.C. Gao, T. Jin, S. Sato, X.Y. Zhang, A seasonal-trend decomposition-based dendritic neuron model for financial time series prediction, *Appl. Soft Comput.* 108 (2021) 107488.
[39] F.F. Li, Z.Y. Wang, X. Zhao, E. Xie, J. Qiu, Decomposition-ANN methods for long-term discharge prediction based on fisher's ordered clustering with MESA, *Water Resour. Manage.* 33 (9) (2019) 3095–3110.
[40] T.H. Sun, T.Y. Zhang, Y. Teng, Z. Chen, J.K. Fang, Monthly electricity consumption forecasting method based on X12 and STL decomposition model in an integrated energy system, *Math. Probl. Eng.* 2019 (2019) 9012543.
[41] S. Polwiang, The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003–2017), *BMC Infect. Dis.* 20 (1) (2020) 208.
[42] M. Theodosiou, Forecasting monthly and quarterly time series using STL decomposition, *Int. J. Forecast.* 27 (4) (2011) 1178–1195.
[43] L. Qin, W.D. Li, S.J. Li, Effective passenger flow forecasting using STL and ESN based on two improvement strategies, *Neurocomputing* 356 (2019) 244–256.
[44] K.B. Newhart, R.W. Holloway, A.S. Hering, T.Y. Cath, Data-driven performance analyses of wastewater treatment plants: A review, *Water Res.* 157 (2019) 498–513.
[45] A.D. Kotzapetros, P.A. Paraskevas, A.S. Stasinakis, Design of a modern automatic control system for the activated sludge process in wastewater treatment, *Chin. J. Chem. Eng.* 23 (8) (2015) 1340–1349.
[46] H.G. Han, H.J. Zhang, Z. Liu, J.F. Qiao, Data-driven decision-making for wastewater treatment process, *Control Eng. Pract.* 96 (2020) 104305.
[47] D.B. Rubin, Inference and missing data, *Biometrika* 63 (3) (1976) 581–592.
[48] P. Wang, C.H. Yang, X.M. Tian, D.X. Huang, Adaptive nonlinear model predictive control using an on-line support vector regression updating strategy, *Chin. J. Chem. Eng.* 22 (7) (2014) 774–781.
[49] S. Moritz, T. Bartz-Beielstein, imputeTS: Time series missing value imputation in R, *R J.* 9 (1) (2017) 207.
[50] N. Alavi, J.S. Warland, A.A. Berg, Filling gaps in evapotranspiration measurements for water budget studies: Evaluation of a Kalman filtering approach, *Agric. For. Meteorol.* 141 (1) (2006) 57–66.