A New Fuzzy Logic-Based Similarity Measure Applied to Large Gap Imputation for Uncorrelated Multivariate Time Series

Thi-Thu-Hong Phan, André Bigand, Émilie Poisson Caillault

Received: 00 Month 0000 | Revised: 00 Month 0000 | Accepted: 00 Month 0000

DOI: xxx/xxxx

ARTICLE TYPE

A New Fuzzy Logic-based Similarity Measure applied to Large Gap Imputation for Uncorrelated Multivariate Time Series

Thi-Thu-Hong PHAN*[1,2] | André BIGAND*[1] | Émilie POISSON CAILLAULT[1]

[1]Univ. Littoral Côte d'Opale, EA 4491-LISIC, F-62228 Calais, France
[2]Vietnam National University of Agriculture, Department of Computer Science, Hanoi, Vietnam

Correspondence
*Thi-Thu-Hong PHAN and André BIGAND. Email: ptthong@vnua.edu.vn and bigand@univ-littoral.fr

Abstract

The completion of missing values is a prevalent problem in many domains of pattern recognition and signal processing. Analyzing data with incompleteness may lead to a loss of power and unreliable results, especially for large missing sub-sequence(s). Therefore, the aim of this paper is to introduce a new approach for filling successive missing values in low/un-correlated multivariate time series, that is to manage a high level of uncertainty. In this way, we propose to use a novel fuzzy weighting-based similarity measure. The proposed method consists of two main steps. Firstly, for each incomplete signal, the data before a gap and the data after this gap are considered as two separated reference time series with their respective query windows Qb and Qa. We then find the most similar sub-sequence (Qbs) to the sub-sequence before this gap Qb and the most similar one (Qas) to the sub-sequence after the gap Qa. To find these similar windows, we build a new similarity measure based on fuzzy grades of basic similarity measures and on fuzzy logic rules. Finally, we fill in the gap with average values of the window following Qbs and the one preceding Qas. The experimental results have demonstrated that the proposed approach outperforms the state-of-the-art methods in case of multivariate time series having low/non-correlated data but effective information on each signal.

KEYWORDS:
Fuzzy inference system, Uncorrelated multivariate time series, Incompleteness, Similarity measure, Imputation.

1 | INTRODUCTION

Nowadays huge time series can now be considered due to the availability of effective low-cost sensors, the wide deployment of remote sensing systems, internet based measure networks,.... However, collected data are often incomplete for various reasons such as sensor errors, transmission problems, incorrect measurements, bad weather conditions (outdoor sensors), for manual maintenance, ... This is particularly the case for marine samples[1] that we consider in this paper. For example, the MAREL-Carnot database characterizes sea water in the eastern English Channel, in France[2]. The data contain nineteen time series that are measured by sensors every 20 minutes as nitrate, fluorescence, phosphate, pH, ... The analysis of these data (it is important to note the important size and shape of this dataset) allows sea biologists to reveal events such as algal blooms and thus better understand phytoplankton processes[3], sea pollution,.... But these data have a lot of missing values: 62.2% for phosphate, 59.9% for nitrate, 27.22% for pH, ...and the size of missing data varies from one-third hour to several months. Most proposed models

for multivariate time series analysis often have difficulties to process incomplete datasets, despite their powerful techniques. They usually require completed data. Then the question is: how can missing values be dealt with? Ignoring or deleting them is a simple way to solve this drawback. But serious problems regularly arise when applying this solution, particularly for time series data where the considered values depend on the previous ones. Furthermore, an analysis with systematic differences between observed and unobserved data leads to biased and unreliable results[4]. Thus, it is important to propose a new technique to estimate the missing values. The imputation technique is a conventional method to handle incompleteness problems[5].

Considering imputation methods, one way to deal with incompleteness for multivariate time series usually take advantage of the correlations between variables to predict lacking data[6,7,8,9,10,11]. This means that correlations permit to use the values of available features to estimate the missing values of other features. However, considering multivariate datasets having low/non-correlations (for instance the MAREL-Carnot dataset), the observed values of full variables cannot be used to complete attributes containing missing values. One way to deal with missing data in this case is to use the observed values of the unique variable with the missing data to compute the incomplete values.

Thus the proposed method has to manage the high level of uncertainty of this kind of signal using a new similarity measure.

Particularly, imperfect time series can be modelled using fuzzy sets. The fuzzy approach makes it possible to deal with incomplete, vague and imprecise circumstances[12], which provide a high uncertainty environment to make decision. The successful use of fuzzy-based similarity measure in pattern recognition[13], in retrieval systems[12] and in recommendation systems[14] leads us to study its ability to complete missing values in uncorrelated multivariate time series. Wang *Wang2015*[15] proposed to use information granules and fuzzy clustering for time series long-term forecasting with success. But according to our knowledge, there is no application devoted to complete large gap(s) in uncorrelated multivariate time series using a fuzzy-weighted similarity measure.

Thus, this paper aims to propose a new approach, named FSMUMI, to fill large missing values in low/un-correlated multivariate time series by developing a new similarity measure based on fuzzy logic. However, estimating the distribution of missing values and whole signals is very difficult, so our approach makes an assumption of effective patterns (or recurrent data) on each signal.

The rest of this paper is organized as follows. In Section 2, related works to imputation methods and fuzzy similarity measure are reviewed. Section 3 introduces our approach for completing large missing sub-sequences in low/un-correlated multivariate time series. Next, Section 4 demonstrates our experimental protocol for the imputation task. Section 5 presents results and discussion. Conclusions are drawn and future work is presented in the last section.

2 | RELATED WORKS

This section presents - first, related work about multivariate imputation methods, followed by a review on the fuzzy similarity measure and its applications.

2.1 | Classical multivariate imputation methods

Up to now, numerous successful researches have been devoted to complete missing data in multivariate time series imputation such as[16,17,18,19,20,21,22,23,24,10,11]. Imputation techniques can be categorized in different perspectives: model-based, or machine learning-based and clustering-based imputation techniques.

In view of the model-based imputation, two main methods were proposed. The first method was introduced by Schafer[16]. With the hypothesis that all variables follow a multivariate normal distribution, this approach is based on the multivariate normal (MVN) model to determine completion values. And, the second method, namely MICE, was developed by van Buuren et al.[17] and Raghunathan et al.[18]. This method uses chained equations to fill in incomplete data: for each variable with missing values, MICE computes the imputation data by exploiting the relations between all other variables.

According to the concept of machine learning-based imputation, many studies focus on completion of missing data in multivariate time series. Stekhoven and Bühlmann[6] implemented missForest based on the Random Forest (RF) method for multivariate imputation. P.Bonissone et al.[25] proposed a fuzzy version of RF that they named fuzzy random forest FRF. At the moment FRF is only devoted to classification and in our case FRF may be only interesting to separate correlated and uncorrelated variables in multivariate time series if necessary. In[21], Shah et al. investigated a variant of MICE which fills in each variable using the estimation generated from RF. The results showed that the combination of MICE and RF was more efficient

than original methods for multivariate imputation. K-Nearest Neighbors (k-NN)-based imputation is also a popular method for completing missing values such as[22,23,26,27,28,11]. This approach identifies the k most similar patterns in the space of available features to impute missing data.

Besides these principal techniques, clustering-based imputation approaches are considered as power tools for completing missing values thanks to their ability to detect similar patterns. The objective of these techniques is to separate the data into several clusters when satisfying the following conditions: maximizing the intercluster similarity and minimizing intracluster dissimilarity. Li et al.[29] proposed the k-means clustering imputation technique that estimates missing values using the final cluster information. The fuzzy c-means (FcM) clustering is a common extension of k-means. The squared-norm is applied to measure the similarity between cluster centers and data points. Different applications based on FcM are investigated for the imputation task as[30,31,32,33,34,8,7,9]. Wang et al.[35] used FcM based on DTW to successfully predict time series in long-term forecasting.

In general, most of the imputation algorithms for multivariate time series take advantage of dependencies between attributes to predict missing values.

2.2 | Methods based on fuzzy similarity measure

Indeed similarity-based approaches are a promising tool for time series analysis. However, many of these techniques rely on parameter tuning, and they may have shortcomings due to dependencies between variables. The objective of this study is to fill large missing values in *uncorrelated multivariate time series*. Thus, we have to deal with a high level of uncertainty. Mikalsen et al.[36] proposed to use GMM (Gaussian mixture models) and cluster kernel to deal with uncertainty. Their method needs ensemble learning with numerous learning datasets that are not available in our case at the moment (marine data). So we have chosen to model this global uncertainty using fuzzy sets (FS) introduced by Zadeh[37]. These techniques consider that measurements have inherent vagueness rather than randomness.

Uncertainty is classically presented using three conceptually distinctive characteristics: fuzziness, randomness and incompleteness. This classification is interesting for many applications, like sensor management (image processing, speech processing, time series processing) and practical decision making. This paper focuses on (sensor) measurements treatment, but is also relevant for other applications. This global uncertainty is commonly modeled using fuzzy sets (FS) introduced by Zadeh[37]. These techniques consider that measurements have inherent vagueness rather than randomness.

Incompleteness often affects time series prediction (time series obtained from marine data such as salinity, temperature, ...). So it seems natural to use fuzzy similarity between sub-sequences of time series to deal with these three kinds of uncertainties (fuzziness, randomness and incompleteness). Fuzzy sets are now well-known and we only need to remind the basic definition of "FS". Considering the universe X, a fuzzy set $A \in X$ is characterized using a fuzzy membership function μ_A:

$$\mu_A : X \to [0, 1], \tag{1}$$

where $\mu_A(x)$ represents the membership of x to A and is associated to the uncertainty of x. In our case, we will consider similarity values between the sub-sequences as defined in the following. One solution to deal with uncertainty brought by multivariate time series is to use the concept of fuzzy time series[38]. In this framework, the variable observations are considered as fuzzy numbers instead of real numbers. In our case the same modelling is used considering distance measures between sub-sequences and then we compute the similarity between these fuzzy numbers to impute the missing data in observations.

Fuzzy similarity is a generalization of the classical concept of equivalence and defines the resemblance between two objects (here sub-sequences of time series). Similarity measures of fuzzy values have been compared in[39] and have been extended in[40]. In[39], Pappis and Karacapilidis presented three main kinds of similarity measures of fuzzy values, including:

- measures based on the operations of union and intersection,
- measures based on the maximum difference,
- measures based on the difference and the sum of membership grades.

In[41,42], the authors used these definitions to propose a distance metric for a space of linguistic summaries based on fuzzy protoforms. Almeida et al. extended this work to put forward linguistic summaries of categorical time series[43]. The introduced similarity measure takes into account not only the linguistic meaning of the summaries, but also the numerical characteristic attached to them. In the same way, Gupta et al.[12] introduced this approach to create an hybrid similarity measure based on fuzzy

logic. The approach is used to retrieve relevant documents. In the other research, Al-shamri and Al-Ashwal presented fuzzy weightings of popular similarity measures for memory-based collaborative recommend systems [14].

Concerning the similarity between two sub-sequences of time series, we can use the DTW cost as a similarity measure. However, to deal with the high level of uncertainty of the processed signals, numerous similarity measures can be used to compute similarity like the cosine similarity, Euclidean distance, Pearson correlation coefficient, and so on. Moreover, a fuzzy-weighted combination of scores generated from different similarity measures could comparatively achieve better retrieval results than the use of a single similarity measure [12,14].

Based on the same concepts, we propose to use a fuzzy rules interpolation scheme between grades of membership of fuzzy values. This method makes it possible to build a new hybrid similarity measure for finding similar values between sub-sequences of time series.

3 | PROPOSED APPROACH

The proposed imputation method is based on the retrieval and the similarity comparison of available sub-sequences. In order to compare the sub-sequences, we create a new similarity measure applying a multiple fuzzy rules interpolation. This section is divided into two parts. Firstly, we focus on the way to compute a new similarity measure between sub-sequences. Then, we provide details of the proposed approach (namely Fuzzy Similarity Measure-based Uncorrelated Multivariate Imputation, FSMUMI) to impute the successive missing values of low/un-correlated multivariate time series.

3.1 | Fuzzy weighted similarity measure between sub-sequences

To introduce a new similarity measure using multiple fuzzy rules interpolation to solve the missing problem, we have to define an information granule, as introduced by Pedrycz [44]. The principle of justifiable granularity of experimental data is based on two conditions: (i) the numeric evidence accumulated within the bounds of numeric data has to be as high as possible and (ii) at the same time, the information granule should be as specific as possible [15].

To answer the first condition, we take into account 3 different distance measures between two sub-sequences Q ($Q = \{q_i, i = 1, \cdots, T\}$) and R ($R = \{r_i, i = 1, \cdots, T\}$) including: Cosine distance, Euclidean distance (these two measures are widely used in the literature) and Similarity distance (this one was presented in our previous study [45]). These three measures are defined as follows:

- Cosine distance is computed by eq 2. This coefficient presents the cosine of the angle between Q and R

$$Cosine(Q, R) = \frac{\sum_{i=1}^{T} q_i . r_i}{\sum_{i=1}^{T} (q_i)^2 . \sum_{i=1}^{T} (r_i)^2} \tag{2}$$

- Euclidean distance is calculated by eq 3

$$ED^*(Q, R) = \sqrt{\sum_{i=1}^{T} (q_i - r_i)^2} \tag{3}$$

  To satisfy the input condition of fuzzy logic rules, we normalize this distance to $[0, 1]$ by this function $ED = 1/(1 + ED^*(q, r))$.

- Similarity measure is defined by the function 4. This measure indicates the similarity percentage between Q and R

$$Sim(Q, R) = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{1 + \frac{|q_i - r_i|}{max(Q) - min(Q)}} \tag{4}$$

To answer the second condition, we use these 3 distance measures (or attributes) to generate 4 fuzzy similarities (see figure 2), then applied to a fuzzy inference system (see figure 1) using the cylindrical extension of the 3 attributes which provides 3 coefficients to calculate a new similarity measure. The universe of discourse of each distance measure is normalized to the value 1.

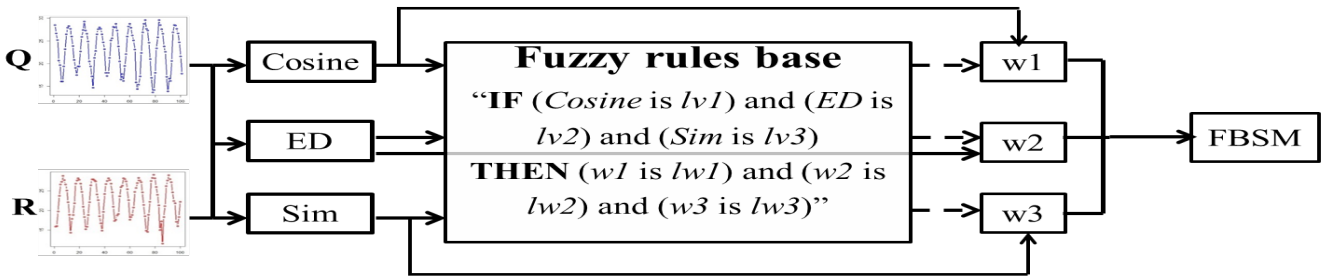

FIGURE 1 Computing scheme of the new similarity measure

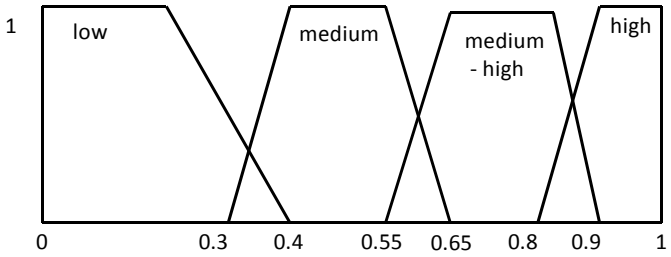

FIGURE 2 Membership function of fuzzy similarity values

And finally, the new similarity measure is determined by eq 5:

$$FBSM = w1 * Cosine(Q, R) + w2 * ED(Q, R) + w3 * Sim(Q, R) \tag{5}$$

where $w1$, $w2$, $w3$ are the weights of the Cosine, ED and Sim measures respectively. Thus uncertainty modelled using FS is kept during the similarity computation and makes it possible to deal with a high level of uncertainty as shown in the sequel. The coefficients wi are generated from the fuzzy interpolation system (figure 1). We use FuzzyR R-package[46] to develop this system. All input and output variables are expressed by 4 linguistic terms as low, medium, medium-high and high. A trapezoidal membership function is handled in this case to match input and output spaces to a degree of membership (figure 2). The multiple rules interpolation is applied to create the fuzzy rules base. So, 64 fuzzy rules are introduced. Each fuzzy rule is presented in the following form:

Rule R: **IF** ($Cosine$ is $lv1$) and (ED is $lv2$) and (Sim is $lv3$) **THEN** ($w1$ is $lw1$) and ($w2$ is $lw2$) and ($w3$ is $lw3$)
in which $lvi, lwi \in$ {low, medium, medium-high, high}, and $i = 1, 2, 3$.

3.2 | FSMUBI Approach

Let us consider some notations about multivariate time series and the concept of large gap. A multivariate time series is represented as a matrix $X_{N\times M}$ with M collected signals of *size* N. $x(t, i)$ is the value of the i-th signal at time t. $x_t = \{x(t, i), i = 1, \cdots, M\}$ is the feature vector at the t-th observation of all variables. X is called an incomplete time series when it contains missing values. We define the term gap of T-size at position t as a portion of X where at least one signal of X between t and $t + T - 1$ containing consecutive missing values ($\exists i | \forall t \in [t, t + T - 1], x(t, i) = NA$).

Here, we deal with large missing values in low/un-correlated multivariate time series. For isolated missing values ($T = 1$) or small T-gap, conventional techniques can be applied such as the mean or the median of available values[47,48]. A T-gap is large when the duration T is longer than known change process. For instance, in phytoplankton study, T is equal to one hour to characterize Langmuir cells and one day for algal bloom processes[49]. For small time series ($N < 10,000$) without prior knowledge of an application and its change process, we set a large gap when $T \geq 5\%N$.

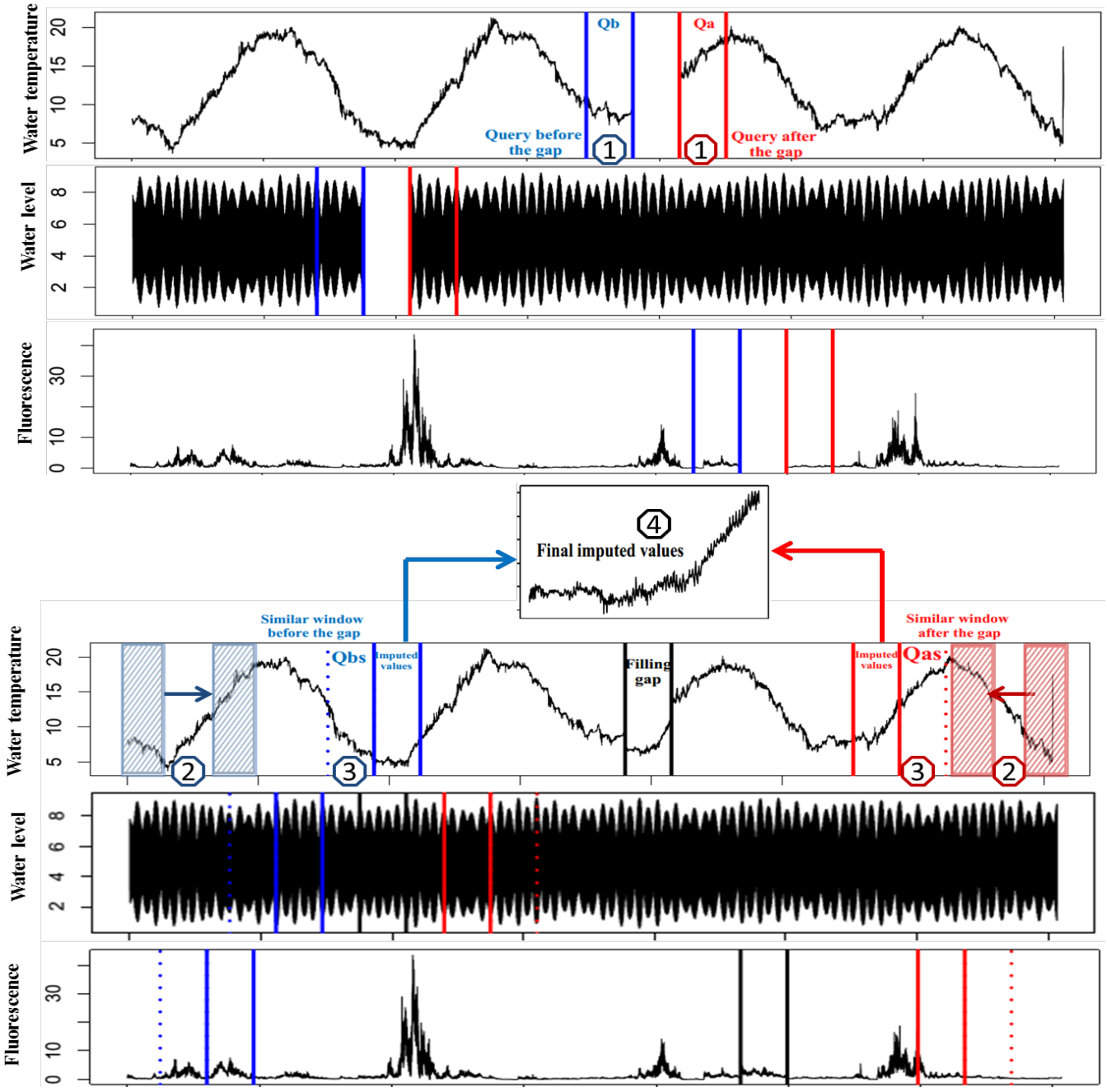

FIGURE 3 Scheme of the completion process: 1- Building queries, 2-Comparing sliding windows, 3-Selecting the most similar windows, 4- Completing gap.

The mechanism of FSMUMI approach is demonstrated in figure 3 . It includes the stage of building queries, the stage of finding similar windows and the stage of completing gap. This method concentrates to fill missing values in low/un-correlated multivariate time series. For this type of data, we can not take advantage of the relations between features to estimate missing values. So we must base our approach on observed values on each signal to complete missing data on itself. This means that we can complete missing data on each variable, one by one. And each incomplete signal will be processed as two separated time series: one time series before the considered gap and one time series after this gap. The proposed model is described in Algorithm 1 and is mainly divided into two phases:

- The first phase

Algorithm 1 FSMUMI algorithm

Input: $X = \{x^1, x^2, \dots, x^M\}$: incomplete uncorrelated multivariate time series
N: size of time seres
t: index of a gap (position of the first missing of the gap)
T: size of the gap
$step\_threshold$: increment for finding a threshold
$step\_sim\_win$: increment for finding a similar window
Output: Y - completed (imputed) time series

for each incomplete signal $x^j \in X$ **do**
  for each gap at t index in x^j **do**
    Divide x^j into two separated time series Da, Db: $Da = x^j[t+T : N]$, $Db = x^j[1 : t-1]$
    Completing all lines containing missing parameter on Da, Db by a max trapezoid function
    Construct queries Qa, Qb-temporal windows after and before the gap $Qa = Da[1 : T]$, $Qb = Db[t-T+1 : t-1]$
    for Db data **do**
      Step a: Find the threshold in the Db database
      $i \leftarrow 1$; $FSM \leftarrow NULL$
      while $i \leq length(Db)$ **do**
        $k \leftarrow i + T - 1$
        Create a reference window: $R(i) = Db[i : k]$
        Calculate a fuzzy-based similarity measure between Qb and $R(i)$: $fbsm$
        Save the $fbsm$ to FMS
        $i \leftarrow i + step\_threshold$
      end while
      return $threshold = max\{FBMS\}$
      Step b: Find similar windows in the Db database
      $i \leftarrow 1$; $Lopb \leftarrow NULL$
      while $i \leq length(Db)$ **do**
        $k \leftarrow i + T - 1$
        Create a reference window: $R(i) = Db[i : k]$
        Calculate a fuzzy-based similarity measure between Qb and $R(i)$: $fbsm$
        if $fbsm \geq threshold$ **then**
          Save position of $R(i)$ to $Lopb$
        end if
        $i \leftarrow i + step\_sim\_win$
      end while
      return position of Qbs - the most similar window to Qb having the maximum fuzzy similarity measure in the $Lopb$ list.
    end for
    for Da data **do**
      Perform **Step a** and **Step b** for Da data
      return position of Qas - the most similar window to Qa
    end for
    Replace the missing values at the position t by average vector of the window after Qbs and the one previous Qas
  end for
end for
return Y - imputed time series

For each incomplete signal and each T-gap, two referenced databases are extracted from the original time series and two query windows are built to retrieve similar windows. We noted Qb is the sub-sequence before the gap and Qa is the respective sub-sequence after the gap. These query windows have the same size T as the gap. The data before the gap (noted Db) and the data after this gap (denoted Da) are considered as two separated time series.

Then for the Db database, we build sliding reference windows (noted R) of size T. From these R windows, we retrieve the most similar window (Qbs) to the Qb query using the new similarity measure $fbsm$ as previously defined in subsection 3.1. Details are in the following:

We first find the threshold, which allows to consider two windows to be similar. For each increment $step\_threshold$, we compute a $fbsm$ similarity measure between a sliding window R and the query Qb. The $threshold$ is the maximum value obtained from the all $fbsm$ calculated (**Step a:** in Algorithm 1). The second task is to find the most similar window to the query Qb. For each increment similar window $step\_sim\_win$, a $fbsm$ of a R sliding reference and the query Qb is estimated. We then compare this $fbsm$ to the $threshold$ to determine if this R reference is similar to the query Qb. We finally choose the most similar window Qbs with the maximum $fbsm$ of all the similar windows (**Step b:** in Algorithm 1).

The same process is performed to find the most similar window Qas in Da data.

In the proposed approach, the dynamics and the shape of data before and after a gap are a key-point of our method. This means we take into account both queries Qa (after the gap) and Qb (before the gap). This makes it possible to find out windows that have the most similar dynamics and shape to the queries.

- The second phase

  When results from both referenced time series are available, we fill in the gap by averaging values of the window preceding Qas and the one following Qbs. The average values are used in our approach because model averaging makes the final results more stable and unbiased[50].

4 | EXPERIMENT PROTOCOL

The experiments are performed on three multivariate time series with the same experiment process and the same gaps, described in detail below.

4.1 | Datasets description

For the assessment of the proposed approach and the comparison of its performance to several published algorithms, we use 3 multivariate time series, one from UCI Machine Learning repository, one simulated dataset (this allows us to handle the correlations between variables and percentage of missing values) and finally a real time series hourly sampled by IFREMER (France) in the eastern English Channel.

- **Synthetic dataset**[51]: The data are synthetic time series, including 10 features, 100,000 sampled points. All data points are in the range -0.5 to +0.5. The data appear highly periodic, but never exactly repeat. They have structure at different resolutions. Each of the 10 features is generated by independent invocations of the function:

$$y = \sum_{i=3}^{7} \frac{1}{2^i} sin(2\pi(2^{2+i} + rand(2^i))t); 0 \leq t \leq 1 \tag{6}$$

  where $rand(x)$ produces a random integer between 0 and x.

  These data are very large so we choose only a subset of 3 signals for performing experiments.

- **Simulated dataset**: In the second experiment, a simulated dataset including 3 signals is produced as follows: for the first variable, we use 5 sine functions that have different frequencies and amplitudes $F = \{f_1, f_2, f_3, f_4, f_5\}$. Next, 3 various noise levels are added to data F, $S = \{F, F + noise1, F + noise2, F + noise3\}$. We then repeat S 4 times (this dataset has 32,000 sampled points). In this study, we treat with missing data in low/un-correlated multivariate time series. So to satisfy this condition, the two remaining signals are generated based on the first signal with the correlations between these signals are low ($\leq 0.1\%$). We apply the Corgen function of ecodist R-package[52] to create the second and the third variables.

- **MAREL-Carnot dataset**[2]: The third experiment is conducted on MAREL-Carnot dataset. This dataset consists of nineteen series such as phosphate, salinity, turbidity, water temperature, fluorescence, water level,... that characterize sea water.

These signals were collected from the 1^{st} January 2005 to the 9^{th} February 2009 at a 20 minute frequency. Here they were hourly sampled, so they have 35,334 time samples. But the data include many missing values, the size of missing data varying on each signal. To assess the performance of the proposed method and compare it with other approaches, we choose a subgroup including fluorescence, water level, and water temperature (the water level and the fluorescence signals are completed data, while water temperature contains isolated missing values and many gaps). We selected these signals because their correlations are low.

After completing missing values, completion data will be compared with the actual values in the completed series to evaluate the ability of different imputation methods. Therefore, it is necessary to fill missing values in the water temperature. To ensure the fairness of all algorithms, filling in the water temperature series is performed by using the na.interp method ([53]).

4.2 | Multivariate imputation approaches

In the present study, we perform a comparison of the proposed algorithm with 7 other approaches (comprising Amelia II, FcM, MI, MICE, missForest, na.approx, and DTWUMI) for the imputation of multivariate time series. We use R language to execute all these algorithms.

1. **Amelia II** (Amelia II R-package)[54]: The algorithm uses the familiar expectation-maximization algorithm on multiple bootstrapped samples of the original incomplete data to draw values of the complete data parameters. The algorithm then draws imputed values from each set of bootstrapped parameters, replacing the missing values with the drawn values.
2. **FcM-Fuzzy *c*-means based imputation**: This approach involves 2 steps. The first step is to group the whole data into k clusters using fuzzy-c means technique. A cluster membership for each sample and a cluster center are generated for each feature. The second step is to fill in the incomplete data by using the membership degree and the center centroids[29]. We base on the principles of[29] and use the c-means function[55] to develop this approach.
3. **MI - Multiple Imputation** (MI R-package)[56]: This method uses predictive mean matching to estimate missing values of continuous variables. For each missing value, its imputation value is randomly selected from a set of observed values that are the closest predicted mean to the variable with the missing value.
4. **MICE - Multivariate Imputation via Chained Equations** (MICE R-package)[57]: For each incomplete variable under the assumption of MAR (missing at random), the algorithm performs a completion by full conditional specification of predictive models. The same process is implemented with other variables having missing data.
5. **missForest** (missForest R-package)[6]: This algorithm uses random forest method to complete missing values. For each variable containing missing data, missForest builds a random forest model on the available data. To estimate missing data this model is applied in the variable. Repeating the procedure until meets a stopping condition.
6. **Linear interpolation - na.approx** (zoo R-package)[58]: This method is based on a interpolation function to predict each missing point.
7. **DTWUMI** [59]: For each gap, this approach finds the most similar window to the sub-sequence after (resp. before) the gap based on the combination of shape-features extraction and Dynamic Time Warping algorithms. Then, the previous (resp. following) window of the most similar one in the incomplete signal is used to complete the gap.

4.3 | Imputation performance measurements

In order to estimate the quantitative performance of imputation approaches, six usual criteria in the literature are used as follows:

1. Similarity: evaluates the similar percent between the estimated values (y) and the respective real values (x). This index is defined by:

$$Sim(y,x) = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{1+\frac{|y_i - x_i|}{max(x)-min(x)}} \tag{7}$$

Where T is the number of missing values. The similarity tends to 1 when the two curves are identical and tends to 0 when the amplitudes are strongly different.

2. R^2 score: is determined as the square of correlation coefficient between two variables y and x. This indicator makes it possible to assess the quality of an imputation model. A method presents better performance when its score is higher ($R^2 \in [0, 1]$)

3. RMSE (Root Mean Square Error): is computed as the average squared difference between y and x. This is an appreciate coefficient to measure global ability of a completion method. In general, a lower RMSE highlights a better imputation performance.

$$RMSE(y, x) = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i - x_i)^2} \tag{8}$$

4. FSD (Fraction of Standard Deviation): is defined as eq. 9

$$FSD(y, x) = 2 * \frac{|SD(y) - SD(x)|}{SD(y) + SD(x)} \tag{9}$$

This fraction points out whether a method is acceptable or not. Applying to the imputation task, when FSD value approaches 0, an imputation method is impeccable.

5. FB - Fractional Bias: determines the rate of predicted values y are overestimated or underestimated relative to observed values x. This indicator is given by eq. 10. An imputation model is considered ideal as its FB equals to 0.

$$FB(y, x) = 2 * |\frac{mean(y) - mean(x)}{mean(y) + mean(x)}| \tag{10}$$

6. FA2: defines the percentage of outlier between two variables y and x. It is described by eq. 11:

$$FA2(y, x) = \frac{length(0.5 \leq \frac{y}{x} \leq 2)}{length(x)} \tag{11}$$

When FA2 value is close to 1, a model is considered perfect.

4.4 | Experimental process

Indeed, evaluating the ability of imputation methods can not be done because the actual values are lacking. So we must produce artificial missing data on completed time series in order to compare the performance of imputation approaches. We use a technique based on three steps to assess the results detailed in the following:

- *The first step*: Generate simulated missing values by removing data values from full time series.
- *The second step*: Apply the imputation methods to fill in missing data.
- *The third step*: Evaluate the ability of proposed approach and compare with state-of-the-art methods using different performance indices above-mentioned.

In this paper, we perform experiments with seven missing data levels on three large datasets. On each signal, we create simulated gaps with different rates ranging from 1%, 2%, 3%, 4%, 5%, 7.5% and 10% of the data in the complete signal (here the biggest gap of MAREL-Carnot data is 3,533 missing values corresponding to 5 months of hourly sampled). For every missing ratio, the approaches are run 5 times by randomly choosing the positions of missing in the data. We then perform 35 iterations for each dataset.

5 | RESULTS AND DISCUSSION

This section provides experiment results obtained from the proposed approach and compares its ability with the seven published approaches. Results are discussed in three parts, i.e quantitative performance, visual performance and execution times.

5.1 | Quantitative performance comparison

Tables 1 , 2 , 3 illustrate the average ability of various imputation methods for synthetic, simulated and MAREL-Carnot time series using 6 measurements as previously defined. For each missing level, the best results are highlighted in bold. These results demonstrate the improved performance of FSMUMI to complete missing data in low/uncorrelated multivariate time series.

Synthetic dataset: Table 1 presents a comparison of 8 imputation methods on synthetic dataset that contains 7 missing data levels (1-10%). The results clearly show that when a gap size is greater than 2%, the proposed method yields the highest similarity, R^2, FA2 and the lowest RMSE, FB. With this dataset, na.approx gives the best performance at the smallest missing data level for all indices and is ranked second for other ratios of missing values (2-5%) for similarity and FA2, RMSE (2-4%), and R^2 (the 1^{st} rank at 2% missing rate, the 2^{nd} at 3%, 5%). The results can explain that the synthetic data are generated by a function (eq. 6). na.approx method which applies the interpolation function to estimate missing values. So it is easy to find a function to generate values that are approximate real values when missing data rates are small. But this work is more difficult when the missing sample size rises, that is why the ability of na.approx decreases as missing data levels increase, especially at 7.5% and 10% rates. Although this dataset never exactly repeats itself and our approach is proposed under the assumption of recurrent data but the FSMUMI approach proves its performance for the imputation task even if the missing size increases.

Among the considered methods, the FcM-based approach is less accurate at lower missing rates but it provides better results at larger missing ratios as regards the accuracy indices.

TABLE 1 Average imputation performance indices of various imputation algorithms on synthetic dataset (100,000 collected points)

Gap size	Method	Accuracy indices			Shape indices		
		1-Sim	1-R^2	RMSE	FSD	FB	1-FA2
1%	FSMUMI	0.136	0.261	0.051	0.358	3.253	0.364
	Amelia	0.275	0.999	0.143	0.409	2.252	0.773
	FcM	0.231	0.722	0.096	1.889	2.208	0.996
	MI	0.275	0.999	0.142	0.421	2.091	0.773
	MICE	0.258	0.944	0.13	0.406	2.452	0.72
	missForest	0.248	0.915	0.122	0.389	3.976	0.744
	na.approx	**0.052**	**0.066**	**0.019**	**0.054**	**0.29**	**0.074**
	DTWUMI	0.257	0.713	0.88	0.725	0.405	0.69
2%	FSMUMI	**0.1**	0.295	**0.046**	**0.155**	**0.395**	**0.337**
	Amelia	0.259	0.998	0.147	0.275	2.005	0.803
	FcM	0.208	0.686	0.104	1.863	2.289	0.987
	MI	0.259	0.998	0.147	0.268	2.11	0.81
	MICE	0.244	0.968	0.14	0.255	7.616	0.759
	missForest	0.239	0.968	0.133	0.279	3.156	0.792
	na.approx	0.104	**0.278**	0.047	0.224	0.398	0.347
	DTWUMI	0.237	0.775	0.867	0.509	8.449	0.646
3%	FSMUMI	**0.113**	**0.341**	**0.056**	0.219	**0.852**	**0.322**
	Amelia	0.218	0.911	0.127	**0.133**	6.128	0.76
	FcM	0.214	0.601	0.1	1.832	1.759	0.989
	MI	0.253	0.993	0.141	0.236	2.295	0.775
	MICE	0.21	0.873	0.118	0.208	5.118	0.703
	missForest	0.188	0.796	0.102	0.215	1.846	0.627
	na.approx	0.148	0.43	0.072	0.372	2.382	0.577
	DTWUMI	0.231	0.799	0.874	0.332	27.952	0.69
4%	FSMUMI	**0.06**	**0.146**	**0.037**	**0.099**	**0.738**	**0.299**
	Amelia	0.208	1	0.14	0.213	2.171	0.807
	FcM	0.155	0.759	0.095	1.85	2.09	0.986
	MI	0.208	0.999	0.14	0.196	2.302	0.807
	MICE	0.209	0.987	0.138	0.22	3.748	0.801
	missForest	0.196	0.968	0.127	0.216	3.94	0.827
	na.approx	0.145	0.721	0.092	0.252	5.251	0.689
	DTWUMI	0.148	0.586	0.918	0.185	12.688	0.719
5%	FSMUMI	**0.055**	**0.132**	**0.032**	**0.058**	**0.098**	**0.201**
	Amelia	0.214	0.997	0.15	0.147	2.238	0.79
	FcM	0.179	0.715	0.108	1.818	2.194	0.993
	MI	0.231	0.996	0.167	0.206	3.094	0.808
	MICE	0.221	0.968	0.152	0.222	2.3	0.79
	missForest	0.212	0.944	0.143	0.315	4.547	0.819
	na.approx	0.16	0.8	0.118	0.352	18.217	0.622
	DTWUMI	0.186	0.885	0.88	0.213	0.723	0.694

7.5%	FSMUMI	**0.049**	**0.071**	**0.027**	0.069	**0.505**	**0.184**
	Amelia	0.197	0.998	0.147	0.045	1.305	0.792
	FcM	0.158	0.809	0.104	1.813	1.866	0.991
	MI	0.2	0.992	0.15	**0.038**	1.645	0.797
	MICE	0.205	0.988	0.15	0.057	10.744	0.799
	missForest	0.188	0.97	0.136	0.284	4.396	0.812
	na.approx	0.192	0.971	0.142	0.669	2.163	0.712
	DTWUMI	0.133	0.653	0.908	0.064	1.113	0.571
10%	FSMUMI	**0.061**	**0.181**	**0.043**	0.114	**0.511**	**0.26**
	Amelia	0.202	0.999	0.147	0.034	4.062	0.788
	FcM	0.164	0.872	0.104	1.837	2.201	0.992
	MI	0.21	0.997	0.155	0.12	2.954	0.785
	MICE	0.209	0.996	0.15	0.055	3.994	0.779
	missForest	0.194	0.97	0.135	0.308	3.024	0.811
	na.approx	0.183	0.997	0.129	0.372	1.455	0.719
	DTWUMI	0.155	0.782	0.893	**0.026**	1.182	0.626

Simulated dataset: Table 2 illustrates the evaluation results of various imputation algorithms on the simulated dataset. The best values for each missing level are highlighted in bold. Our proposed method outperforms other methods for the imputation task on accuracy indices: the highest similarity, R^2 and the lowest RMSE at every missing ratio. However, when considering other indices such as FA2, FSD and FB, FSMUMI no longer shows its performance. It gains only at a 4% rate for the FB index and at 10% ratio for FA2. In contrast to FSMUMI, DTWUMI provides the best results for FSD indicator at all missing levels and FA2 at the first 5 missing ratios (from 1% to 5%).

Different from the synthetic dataset, on the simulated dataset, the FcM-based method is always ranked the third at all missing rates for similarity and RMSE indicators. Following FcM is missForest algorithm for the both indices.

Although, in the second experiment, data are built by various functions but they are quite complex so that na.approx does not provide good results.

TABLE 2 Average imputation performance indices of various imputation algorithms on simulated dataset (32,000 collected points)

Gap size	Method	Accuracy indices			Shape indices		
		1-Sim	1-R^2	RMSE	FSD	FB	1-FA2
1%	FSMUMI	**0.083**	**0.515**	**1.033**	0.159	2.51	0.574
	Amelia	0.157	1	2.206	0.232	3.619	0.794
	FcM	0.118	0.998	1.483	1.98	2.015	0.998
	MI	0.16	0.999	2.241	0.2	**0.915**	0.799
	MICE	0.159	0.998	2.201	0.214	1.449	0.801
	missForest	0.127	0.998	1.608	0.836	12.034	0.861
	na.approx	0.146	0.992	1.901	0.393	18.997	0.777
	DTWUMI	0.09	0.552	1.156	**0.007**	6.022	**0.562**
2%	FSMUMI	**0.068**	**0.487**	**1.166**	0.194	1.971	0.611
	Amelia	0.12	0.998	2.312	0.107	2.191	0.794
	FcM	0.093	0.999	1.672	1.985	**1.96**	0.998
	MI	0.12	1	2.307	0.123	3.949	0.789
	MICE	0.119	0.999	2.282	0.114	8.881	0.789
	missForest	0.096	1	1.769	0.941	2.777	0.858
	na.approx	0.118	1	2.261	0.721	2.059	0.786
	DTWUMI	0.074	0.523	1.545	**0.008**	3.686	**0.583**
3%	FSMUMI	**0.068**	**0.453**	**1.053**	0.076	10.649	0.582
	Amelia	0.13	0.999	2.212	0.062	3.779	0.794
	FcM	0.098	0.999	1.526	1.984	2.22	0.997
	MI	0.13	0.999	2.197	0.078	9.374	0.795
	MICE	0.129	1	2.19	0.067	**1.938**	0.792
	missForest	0.102	0.999	1.626	0.855	2.407	0.851
	na.approx	0.116	0.997	1.938	0.518	1.974	0.818
	DTWUMI	0.073	0.526	1.189	**0.01**	8.725	**0.567**
4%	FSMUMI	**0.064**	**0.412**	**1.067**	0.061	**1.374**	0.568
	Amelia	0.122	1	2.305	0.032	2.446	0.764
	FcM	0.096	1	1.607	1.982	2.325	0.997
	MI	0.125	1	2.261	0.043	2.391	0.792
	MICE	0.124	0.999	2.233	0.045	42.495	0.791
	missForest	0.101	1	1.726	0.876	2.901	0.854
	na.approx	0.109	1	1.99	0.475	1.94	0.811
	DTWUMI	0.066	0.465	1.172	**0.004**	2.079	**0.547**

5%	FSMUMI	**0.063**	**0.404**	**1.062**	0.062	4.508	0.577
	Amelia	0.122	1	2.273	0.028	4.109	0.798
	FcM	0.092	1	1.619	1.984	2.192	0.998
	MI	0.123	1	2.287	0.024	5.582	0.797
	MICE	0.121	1	2.267	0.044	2.326	0.792
	missForest	0.097	0.999	1.731	0.923	2.473	0.859
	na.approx	0.114	1	1.988	0.567	2.247	0.809
	DTWUMI	0.063	0.454	1.166	**0.003**	**1.594**	**0.545**
7.5%	FSMUMI	**0.06**	**0.408**	**1.063**	0.049	4.843	**0.566**
	Amelia	0.117	1	2.232	0.034	3.306	0.792
	FcM	0.09	1	1.605	1.981	3.562	0.998
	MI	0.119	0.999	2.259	0.025	1.946	0.793
	MICE	0.118	1	2.238	0.032	9.359	0.794
	missForest	0.094	0.999	1.695	0.907	**1.259**	0.858
	na.approx	0.108	1	1.958	0.461	3.089	0.816
	DTWUMI	0.065	0.477	1.19	**0.004**	3.851	0.566
10%	FSMUMI	**0.061**	**0.4226**	**1.086**	0.051	5.558	**0.572**
	Amelia	0.117	1	2.269	0.021	3.074	0.793
	FcM	0.089	1	1.607	1.981	2.683	0.997
	MI	0.118	0.9996	2.233	0.02	2.05	0.793
	MICE	0.118	0.9998	2.254	0.018	3.424	0.793
	missForest	0.094	0.9999	1.702	0.909	**1.87**	0.857
	na.approx	0.11	1	1.958	0.541	2.006	0.798
	DTWUMI	0.067	0.5371	1.293	**0.012**	3.093	0.577

MAREL Carnot dataset: Once again, as reported in table 3 , our algorithm demonstrates its capability for the imputation task. FSMUMI method generates the best results as regarding accuracy indices for almost missing ratios (excluding at 2% missing level on all indices, and at 5% missing rate on R^2 score). But when considering shape indicators, FSMUMI only provides the highest FA2 values at several missing levels (3%, 5%-10%). In particular, our method illustrates the ability to fill in incomplete data with large missing rates (7.5% and 10%): the highest similarity, R^2, FA2 and the lowest RMSE, FSD (excluding at 7.5%), and FB. These gaps correspond to 110.4 and 147.2 days sampled at hourly frequency.

In contrast to the two datasets above, on the MAREL-Carnot data, na.approx indicates quite good results: the permanent second or third rank for the accuracy indices (the 1^{st} order at 5% missing rate on R^2 score), the lowest FSD (from 3% to 5% missing rates) and FB at some other levels of missing data. But when looking at the shape of imputation values generated from this method, it absolutely gives the worst results (figure 6).

Other approaches (including FcM-based imputation, MI, MICE, Amelia, missForest) exploit the relations between attributes to estimate missing values. However, three considered datasets have low correlations between variables (roundly 0.2 for MAREL-Carnot data, ≤ 0.1 for simulated and synthetic datasets). So these methods do not demonstrate their performance for completing missing values in low/un-correlated multivariate time series. Otherwise, our algorithm shows its ability and stability when applying to the imputation task for this kind of data.

DTWUMI approach was proposed to fill large missing values in low/un-correlated multivariate time series. However, this method is not as powerful as the FSMUMI method. DTWUMI only produces the best results at 2% missing level on the MAREL-Carnot dataset, and is always at the second or the third rank at all the remaining missing rates on the MAREL-Carnot and the simulated datasets. That is because the DTWUMI method only finds the most similar window to a query either before a gap or after this gap, and it uses only one similarity measure, the DTW cost, to retrieve the most similar window. In addition, another reason may be that DTWUMI has directly used data from the window following or preceding the most similar window to completing the gap.

TABLE 3 Average imputation performance indices of various imputation algorithms on MAREL-Carnot dataset (35,334 collected points)

Gap size	Method	Accuracy indices			Shape indices		
		1-Sim	1-R^2	RMSE	FSD	FB	1-FA2
1%	FSMUMI	**0.051**	**0.156**	**1.532**	**0.044**	0.081	0.191
	Amelia	0.187	0.544	5.132	0.378	0.354	0.482
	FcM	0.156	0.342	4.037	0.4	0.347	0.338
	MI	0.192	0.561	5.282	0.396	0.365	0.497
	MICE	0.166	0.608	5.596	0.423	0.35	0.436
	missForest	0.165	0.472	4.422	0.385	0.355	0.381

	na.approx	0.061	0.171	1.748	0.067	**0.06**	**0.161**
	DTWUMI	0.084	0.181	2.466	0.214	0.149	0.198
2%	FSMUMI	0.045	0.037	1.446	0.053	0.083	0.182
	Amelia	0.146	0.369	4.743	0.211	0.222	0.429
	FcM	0.116	0.06	3.418	0.415	0.237	0.231
	MI	0.146	0.364	4.72	0.218	0.228	0.435
	MICE	0.129	0.369	4.711	0.197	0.21	0.413
	missForest	0.116	0.155	3.575	0.33	0.193	0.258
	na.approx	0.06	0.07	2.012	0.045	0.094	0.214
	DTWUMI	**0.042**	**0.018**	**1.095**	**0.029**	**0.066**	**0.154**
3%	FSMUMI	**0.053**	**0.11**	**1.294**	0.134	0.08	**0.166**
	Amelia	0.176	0.503	4.694	0.426	0.224	0.478
	FcM	0.139	0.251	3.35	0.441	0.237	0.314
	MI	0.17	0.531	4.474	0.354	0.221	0.476
	MICE	0.157	0.552	4.905	0.34	0.184	0.429
	missForest	0.139	0.345	3.556	0.422	0.184	0.346
	na.approx	0.068	0.224	1.79	**0.062**	**0.056**	0.169
	DTWUMI	0.096	0.216	2.587	0.329	0.136	0.223
4%	FSMUMI	**0.059**	**0.058**	**1.466**	0.094	**0.101**	0.183
	Amelia	0.171	0.44	4.389	0.287	0.2	0.456
	FcM	0.126	0.152	2.779	0.285	0.203	0.727
	MI	0.166	0.41	4.234	0.277	0.204	0.444
	MICE	0.15	0.379	4.15	0.268	0.19	0.411
	missForest	0.129	0.234	3.134	0.23	0.187	0.303
	na.approx	0.077	0.13	2.006	**0.068**	0.135	0.268
	DTWUMI	0.07	0.105	1.77	0.15	0.12	**0.138**
5%	FSMUMI	**0.051**	0.22	**2.025**	0.227	0.152	**0.167**
	Amelia	0.151	0.551	4.924	0.303	0.189	0.461
	FcM	0.113	0.337	3.606	0.301	0.199	0.254
	MI	0.143	0.567	4.612	0.249	0.123	0.448
	MICE	0.131	0.523	4.75	0.274	0.188	0.419
	missForest	0.104	0.371	3.443	0.229	0.147	0.274
	na.approx	0.065	**0.213**	2.071	**0.175**	**0.038**	0.233
	DTWUMI	0.067	0.275	2.363	0.22	0.157	0.242
7.5%	FSMUMI	**0.043**	**0.056**	**1.52**	0.075	**0.039**	**0.189**
	Amelia	0.14	0.42	4.546	0.191	0.197	0.437
	FcM	0.104	0.123	3.12	0.328	0.198	0.23
	MI	0.142	0.427	4.624	0.222	0.222	0.443
	MICE	0.126	0.38	4.375	0.206	0.208	0.437
	missForest	0.112	0.202	3.587	0.329	0.228	0.288
	na.approx	0.073	0.081	2.043	0.092	0.107	0.243
	DTWUMI	0.06	0.102	1.999	**0.071**	0.074	0.215
10%	FSMUMI	**0.053**	**0.098**	**1.642**	**0.083**	**0.055**	**0.191**
	Amelia	0.14	0.3	4.294	0.24	0.142	0.442
	FcM	0.1	0.098	3.68	0.136	0.101	0.303
	MI	0.14	0.112	4.294	0.24	0.142	0.442
	MICE	0.12	0.42	4.066	0.152	0.077	0.383
	missForest	0.097	0.461	3.049	0.104	0.117	0.255
	na.approx	0.071	0.529	1.873	0.098	0.094	0.253
	DTWUMI	0.081	0.381	3.293	0.119	0.124	0.224

5.2 | Visual performance comparison

In this paper, we also compare the visualization performance of completion values yielded by various algorithms. Figure 4 and figure 5 illustrate the form of imputed values generated from different approaches on the synthetic series at two missing ratios 1% and 5%.

At a 1% missing rate, the shape of imputation values produced by na.approx method is closer to the one of true values than the form of completion values given by our approach. However, at a 5% level of missing data, this method no longer shows the performance (figure 5). In this case, the proposed method proves its relevance for the imputation task. The shape of FSMUMI's imputation data is almost similar to the form of true values (figure 5).

Looking at figure 6 , FSMUMI one more time proves its capability for uncorrelated multivariate time series imputation: completion values yielded by FSMUMI are virtually identical to the real data on the MAREL-Carnot dataset. When comparing DTWUMI with FSMUMI, it is clear that FSMUMI gives improved results (figure 4 , 5 and 6).

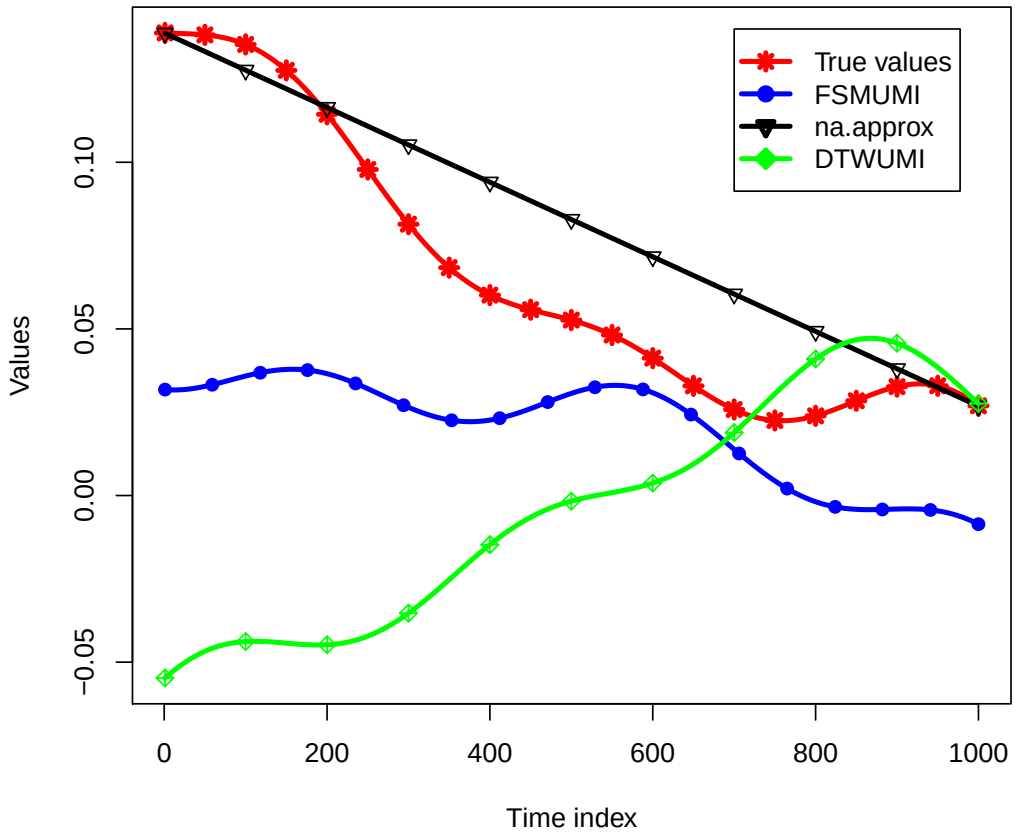

FIGURE 4 Visual comparison of completion data of different imputation approaches with real data on the 1^{st} signal of synthetic series with the gap size of 1000

5.3 | Computation time

Besides, we perform a comparison of the computational time of each method on the synthetic series (in second - s). Table 4 indicates that na.approx method requires the shortest running time and DTWUMI approach takes the longest computing time. The proposed method, FSMUMI, demands more execution time as missing rates increase. However, considering the quantitative and visual performance of FSMUMI for the imputation task (table 1 , figue 5 and figue 6), the required time of the proposed approach is fully acceptable.

TABLE 4 Computational time of different methods on the synthetic series in second (s)

Method	Gaps size						
	1%	2%	3%	4%	5%	7.5%	10%
FSMUMI	353.9	427.5	701.9	1037.8	1423.6	2525.5	3556.8
Amelia	3.2	3.4	5.2	3.2	3.2	3.2	3.2
FcM	40.9	39.8	40.0	41.1	41.2	46.7	45.6
MI	844.1	714.0	739.1	723.3	724.5	719.7	726.5
MICE	7021.1	9187.7	21909.6	13041.9	14833.9	19417.7	23812.6
missForest	26833.8	24143.8	22969.9	32056.6	36485.8	42424.1	28521.1
na.approx	**0.11**	**0.089**	**0.167**	**0.09**	**0.088**	**0.088**	**0.094**
DTWUMI	5002.67	15714.8	37645.82	64669.71	86435.38	180887.78	273879

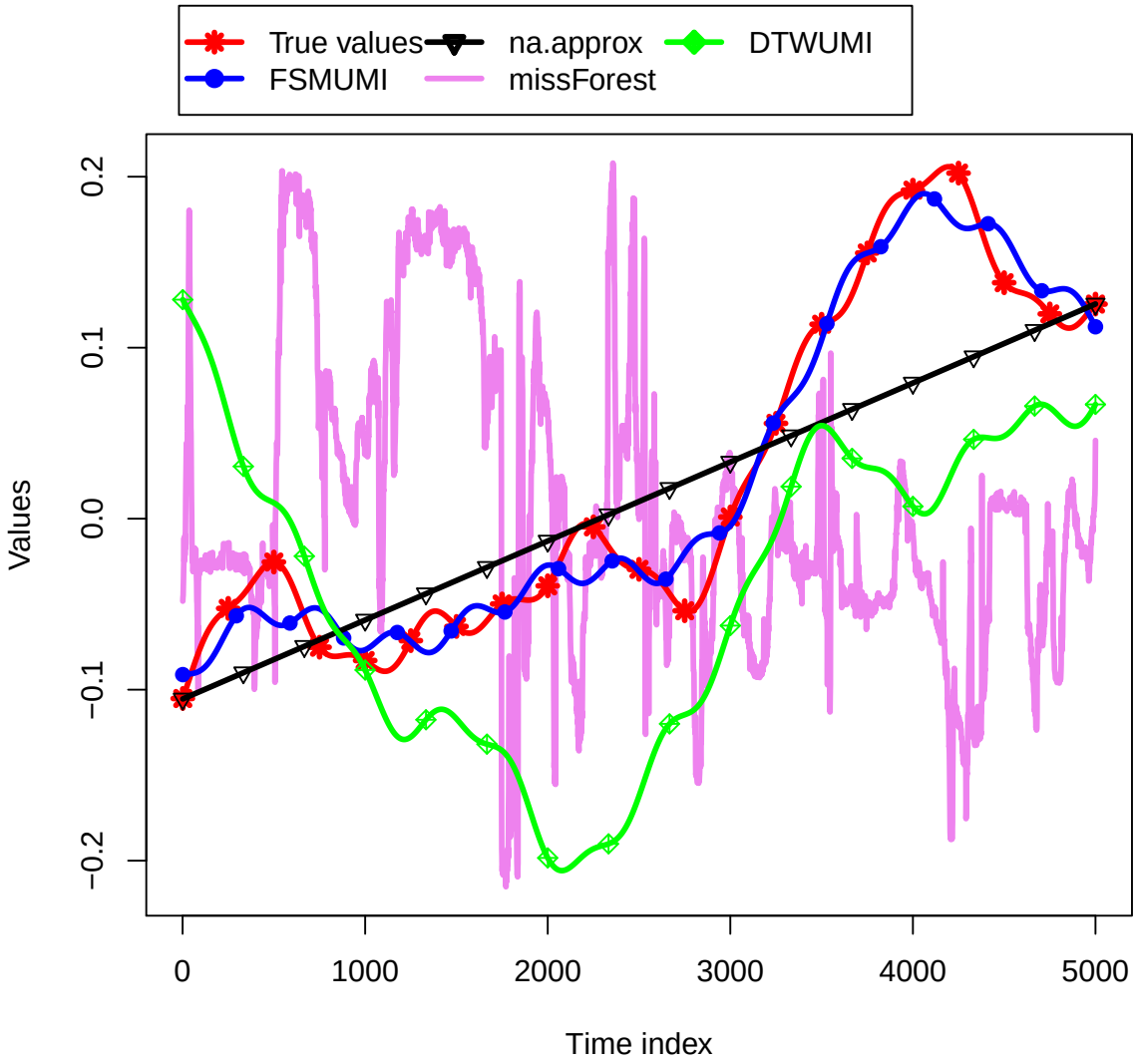

FIGURE 5 Visual comparison of completion data of different imputation approaches with real data on the 1^{st} signal of synthetic series with the gap size of 5000

6 | CONCLUSION

This paper proposes a novel approach for uncorrelated multivariate time series imputation using a fuzzy logic-based similarity measure, namely FSMUMI. This method makes it possible to manage uncertainty with the comprehensibility of linguistic variables. FSMUMI has been tested on different datasets and compared with published algorithms (Amelia II, FcM, MI, MICE, missForest, na.approx, and DTWUMI) on accuracy and shape criteria. The visual ability of these approaches is also investigated. The experimental results definitely highlight that the proposed approach yielded improved performance in accuracy over previous methods in the case of multivariate time series having large gaps and low or non-correlation between variables. However, the proposed algorithm is necessary to make an assumption of recurrent data and sufficiently large dataset.

In future work, we plan to (i) combine FSMUMI method with other algorithms such as Random Forest or Deep learning in order to efficiently fill incomplete values in any type of multivariate time series; (ii) investigate this approach applied to short-term/long-term forecasts in multivariate time series. We could also investigate complex fuzzy sets ([60]) instead of ordinary FS that have given good results using an adaptive scheme in the case of the bi-variate time series with small dataset.

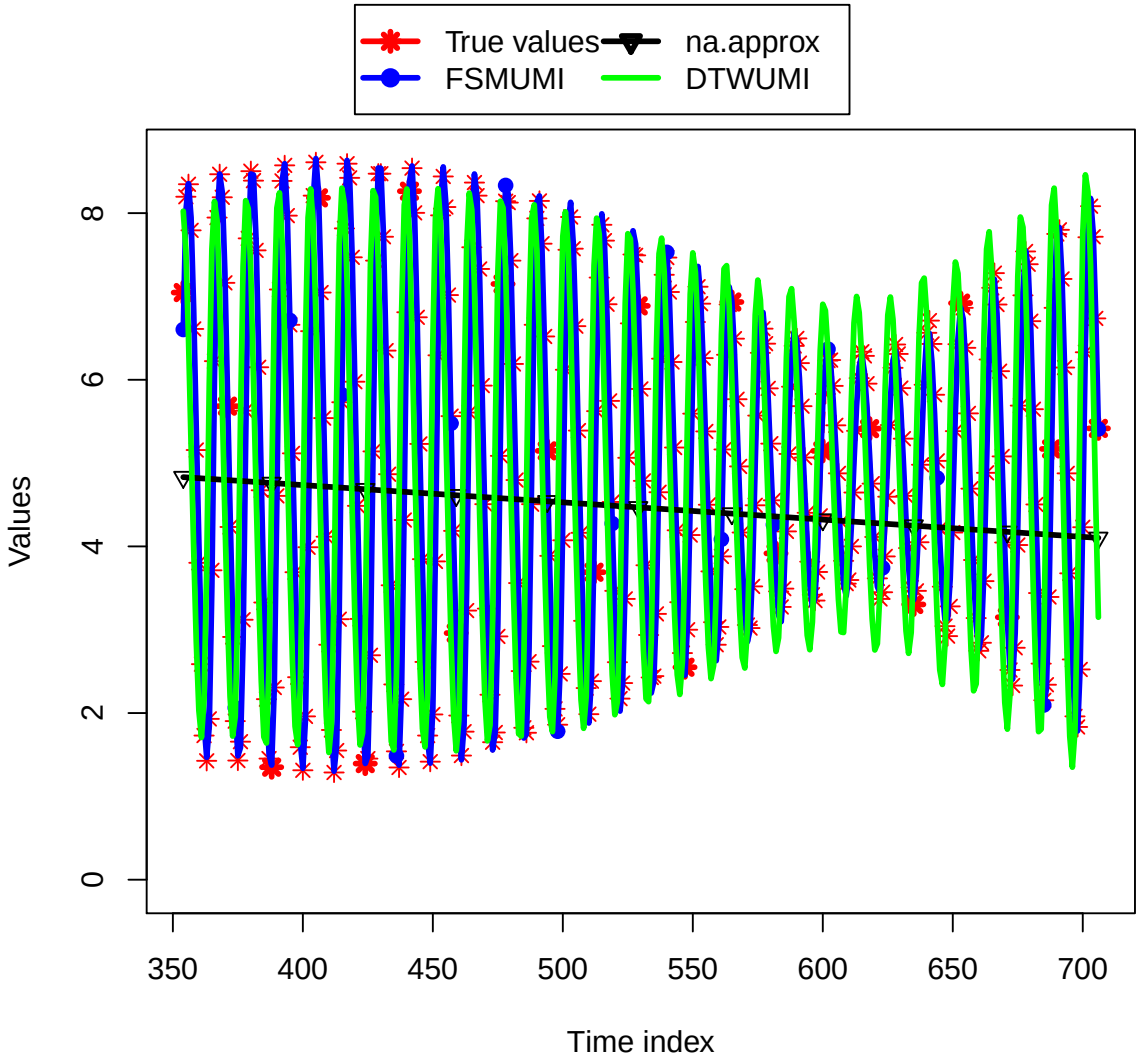

FIGURE 6 Visual comparison of completion data of different imputation approaches with real data on the 2^{nd} signal of MAREL Carnot dataset with the gap size of 353

ACKNOWLEDGMENTS

This work was kindly supported by the Ministry of Education and Training Vietnam International Education Development, the French government and FEDER, the region Hauts-de-France (CPER 2014-2020 MARCO). The experiments were carried out using the CALCULCO computing platform, supported by SCoSI/ULCO (Univ. Littoral)

References

1. Ceong H.T., Kim H.J., Park J.S.. Discovery of and Recovery from Failure in a Coastal Marine USN Service. *Journal of Information and Communication Convergence Engineering*. 2012;10(1):11–20.

2. Lefebvre Alain. *MAREL Carnot Data and Metadata from Coriolis Data Centre. SEANOE. http://doi.org/10.17882/39754.* 2015.

3. Rousseeuw K., Poisson Caillault E., Lefebvre A., Hamad D.. Hybrid Hidden Markov Model for Marine Environment Monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.* 2015;8(1):204-213.

4. Hawthorne Graeme, Hawthorne Graeme, Elliott Peter. Imputing Cross-Sectional Missing Data: Comparison of Common Techniques. *Australian and New Zealand Journal of Psychiatry.* 2005;39(7):583–590.

5. Junninen Heikki, Niska Harri, Tuppurainen Kari, Ruuskanen Juhani, Kolehmainen Mikko. Methods for Imputation of Missing Values in Air Quality Data Sets. *Atmospheric Environment.* 2004;38(18):2895–2907.

6. Stekhoven Daniel J., Bühlmann Peter. MissForest—non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics.* 2012;28(1):112–118.

7. Ichihashi Hidetomo, Honda Katsuhiro, Notsu Akira, Yagi Takafumi. Fuzzy C-Means Classifier with Deterministic Initialization and Missing Value Imputation. In: :214–221IEEE; 2007.

8. Saravanan P., Sailakshmi P.. Missing value imputation using fuzzy possibilistic c means optimized with support vector regression and genetic algorithm. *Journal of Theoretical and Applied Information Technology.* 2015;72(1):34–39.

9. Furukawa Takashi, Ohnishi Shin-ichi, Yamanoi Takahiro. Missing Categorical Data Imputation for FCM Clusterings of Mixed Incomplete Data. In: ; 2014.

10. Deng Yi, Chang Changgee, Ido Moges Seyoum, Long Qi. Multiple Imputation for General Missing Data Patterns in the Presence of High-Dimensional Data. *Scientific Reports.* 2016;6:21689.

11. Oehmcke Stefan, Zielinski Oliver, Kramer Oliver. kNN Ensembles with Penalized DTW for Multivariate Time Series Imputation. In: :2774–2781IEEE; 2016.

12. Gupta Y., Saini A., Saxena A.. Fuzzy Logic-Based Approach to Develop Hybrid Similarity Measure for Efficient Information Retrieval. *Journal of Information Science.* 2014;40(6):846–857.

13. Shahmoradi Sina, Bagheri Shouraki Saeed. Evaluation of a novel fuzzy sequential pattern recognition tool (fuzzy elastic matching machine) and its applications in speech and handwriting recognition. *Applied Soft Computing.* 2018;62:315–327.

14. Al-Shamri Mohammad Yahya H., Al-Ashwal Nagi H.. Fuzzy-Weighted Similarity Measures for Memory-Based Collaborative Recommender Systems. *Journal of Intelligent Learning Systems and Applications.* 2014;06(01):1–10.

15. Wang Weina, Pedrycz Witold, Liu Xiaodong. Time series long-term forecasting model based on information granules and fuzzy clustering. *Eng. Appl. of AI.* 2015;41:17–24.

16. Schafer J. L.. *Analysis of Incomplete Multivariate Data*. CRC Press; 1997.

17. Van Buuren Stef, Boshuizen Hendriek C., Knook Dick L., others . Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in medicine.* 1999;18(6):681–694.

18. Raghunathan Trivellore E., Lepkowski James M., Van Hoewyk John, Solenberger Peter. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey methodology.* 2001;27(1):85–96.

19. Engels J, Diehr Paula. Imputation of Missing Longitudinal Data: A Comparison of Methods. *Journal of Clinical Epidemiology.* 2003;56(10):968–976.

20. Royston Patrick. Multiple Imputation of Missing Values: Further Update of Ice, with an Emphasis on Interval Censoring. *Stata Journal.* 2007;7(4):445–464.

21. Shah Anoop D., Bartlett Jonathan W., Carpenter James, Nicholas Owen, Hemingway Harry. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology.* 2014;179(6):764–774.

22. Liao Serena G., Lin Yan, Kang Dongwan D., et al. Missing Value Imputation in High-Dimensional Phenomemic Data: Imputable or Not, and How?. *BMC Bioinformatics.* 2014;15:346.

23. Rahman Shah Atiqur, Huang Yuxiao, Claassen Jan, Heintzman Nathaniel, Kleinberg Samantha. Combining Fourier and Lagged k -Nearest Neighbor Imputation for Biomedical Time Series Data. *Journal of Biomedical Informatics.* 2015;58:198–207.

24. Gelman Andrew, Hill Jennifer, Su Yu-Sung, et al. *Mi: Missing Data Imputation and Model Checking.* 2015.

25. Bonissone Piero, Cadenas José M., Carmen Garrido M., Andrés Díaz-Valladares R.. A fuzzy random forest. *International Journal of Approximate Reasoning.* 2010;51(7):729–747.

26. Hsu Hui-Huang, Yang Andy C., Lu Ming-Da. KNN-DTW Based Missing Value Imputation for Microarray Time Series Data. *JCP.* 2011;6(3):418–425.

27. Yang Andy C., Hsu Hui-Huang, Lu Ming-Da. Missing Value Imputation in Microarray Gene Expression Data. In: :300–307; 2009; Kinmen, Taiwan.

28. Kostadinova Elena, Boeva Veselka, Boneva Liliana, Tsiporkova Elena. An Integrative DTW-Based Imputation Method for Gene Expression Time Series Data. In: :258–263IEEE; 2012.

29. Li Dan, Deogun Jitender, Spaulding William, Shuart Bill. Towards Missing Data Imputation: A Study of Fuzzy k-Means Clustering Method. In: :573–579Springer; 2004.

30. Tang Jinjun, Zhang Guohui, Wang Yinhai, Wang Hua, Liu Fang. A Hybrid Approach to Integrate Fuzzy C-Means Based Imputation Method with Genetic Algorithm for Missing Traffic Volume Data Estimation. *Transportation Research Part C: Emerging Technologies.* 2015;51:29–40.

31. Aydilek Ibrahim Berkan, Arslan Ahmet. A Hybrid Method for Imputation of Missing Values Using Optimized Fuzzy C-Means with Support Vector Regression and a Genetic Algorithm. *Information Sciences.* 2013;233:25–35.

32. Furukawa Takashi, Ohnishi Shin-ichi, Yamanoi Takahiro. On C-Means Algorithm for Mixed Incomplete Data Using Partial Distance and Imputation. In: ; 2014.

33. Azim S., Aggarwal S.. Hybrid model for data imputation: Using fuzzy c means and multi layer perceptron. 2014;:1281–1285.

34. Tang Jinjun, Wang Yinhai, Zhang Shen, Wang Hua, Liu Fang, Yu Shaowei. On Missing Traffic Data Imputation Based on Fuzzy C -Means Method by Considering Spatial–Temporal Correlation. *Transportation Research Record: Journal of the Transportation Research Board.* 2015;2528:86–95.

35. Wang Weina, Pedrycz Witold, Liu Xiaodong. Time series long-term forecasting model based on information granules and fuzzy clustering. *Engineering Applications of Artificial Intelligence.* 2015;41:17–24.

36. Mikalsen Karl Øyvind, Bianchi Filippo Maria, Soguero-Ruiz Cristina, Jenssen Robert. Time series cluster kernel for learning similarities between multivariate time series with missing data. *Pattern Recognition.* 2018;76:569–581.

37. Zadeh L.A.. Fuzzy sets. *Inform. and Control.* 1965;8:338-353.

38. H.J.Sadael , F.G.Guimaraes , Silva C.José, M.H.Lee , T.Eslami . Short-term load forecasting method based on fuzzy time series, seasonality and long memory process. *Int. Journal of Approximate Reasoning.* 2017;83:196–217.

39. Pappis Costas P., Karacapilidis Nikos I.. A Comparative Assessment of Measures of Similarity of Fuzzy Values. *Fuzzy Sets and Systems.* 1993;56(2):171–174.

40. Shyi-Ming Chen , Ming-Shiow Yeh , Pei-Yung Hsiao . A Comparison of Similarity Measures of Fuzzy Values. *Fuzzy Sets and Systems.* 1995;72(1):79–89.

41. Wilbik Anna, Keller James M.. A Distance Metric for a Space of Linguistic Summaries. *Fuzzy Sets and Systems.* 2012;208:79–94.

42. Wilbik Anna, M.Keller James. A fuzzy measure similarity between sets of linguistic summaries. *IEEE Trans.on Fuzzy Systems vol.21 (1) (2012).* 2012;.

43. Almeida Rui Jorge, Lesot Marie-Jeanne, Bouchon-Meunier Bernadette, Kaymak Uzay, Moyse Gilles. Linguistic Summaries of Categorical Time Series for Septic Shock Patient Data. In: :1–8IEEE; 2013.

44. W. Pedrycz F. Gomide. *Fuzzy Systems Engineering: Toward Human-Centric Computing*. John Wiley, Hoboken, NJ; 2007.

45. Phan Thi-Thu-Hong, Poisson Caillault Emilie, Lefebvre Alain, Bigand Andre. Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters (2017).* 2017;.

46. Garibaldi Jon, Chen Chao, Razak Tajul. FuzzyR: Fuzzy Logic Toolkit for R2017. R package version 2.1.

47. Allison Paul D.. *Missing Data* Quantitative Applications in the Social Sciences, vol. 136: . Sage Publication; 2001.

48. Bishop Christopher M.. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2006.

49. Dickey Tommy D.. Emerging ocean observations for interdisciplinary data assimilation systems. *Journal of Marine Systems.* 2003;40:5 - 48.

50. Schomaker Michael, Heumann Christian. Model Selection and Model Averaging After Multiple Imputation. *Comput. Stat. Data Anal.*. 2014;71:758–770.

51. Keogh Eamonn J., Pazzani Michael J.. An Indexing Scheme for Fast Similarity Search in Large Time Series Databases. In: :56–67IEEE; 1999.

52. Goslee Sarah C., Urban Dean L., others . The Ecodist Package for Dissimilarity-Based Analysis of Ecological Data. *Journal of Statistical Software.* 2007;22(7):1–19.

53. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R, used package in 2016. *Journal of Statistical Software.* 2008;:1-22.

54. Honaker James, King Gary, Blackwell Matthew. Amelia II: A Program for Missing Data. 2011;45,(2017).

55. Meyer David, Dimitriadou Evgenia, Hornik Kurt, Weingessel Andreas, Leisch Friedrich. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien2015. R package version 1.6-7.

56. Su Yu-Sung, Gelman Andrew, Hill Jennifer, Yajima Masanao, others . Multiple Imputation with Diagnostics (Mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software.* 2011;45(2):1–31.

57. Buuren Stef, Groothuis-Oudshoorn Karin. Mice: Multivariate Imputation by Chained Equations in R. *Journal of statistical software.* 2011;45(2011).

58. Zeileis Achim, Grothendieck Gabor, Ryan Jeffrey A., Andrews Felix. *Zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations).* 2016.

59. Phan Thi-Thu-Hong, Poisson Caillault Emilie, Bigand Andre, Lefebvre Alain. DTW-Approach for Uncorrelated Multivariate Time Series Imputation. In: ; 2017; Tokyo, Japan.

60. Yazdanbakhsh Omolbanin, Dick Scott. A systematic review of complex fuzzy sets and logic. *Fuzzy Sets and Systems.* 2018;338:1–22.

AUTHOR BIOGRAPHY

Thi-Thu-Hong PHAN received her M.Sc in information and web technology from Claude Bernard Lyon 1 University, Lyon, France. Currently, she is a Ph.D candidate in signal processing, University of Littoral, Calais, France. Her research interests include time series imputation, signal processing, computer vision, and machine learning.

André Bigand (IEEE Member) received the Ph.D. degree in 1993 from the University Paris 6 and the "HDR" degree in 2001 from University of Littoral in Calais (ULCO, France). He is currently senior associate professor in ULCO since 1993. His current research interests include uncertainty modeling and machine learning with applications to image processing and synthesis (particularly noise modeling and filtering). He is currently with the LISIC Laboratory (ULCO).

Emilie Poisson Caillault is assistant professor in the LISIC Laboratory in the University Littoral in Calais (France). She received the Ph.D degree in 2005 from the University of Nantes. Her research interests include clustering and machine learning, time series similarity and imputation. She is involved in H2020 Jerico-Next project dedicated in the monitoring of water quality and climate change.