Which DTW Method Applied to Marine Univariate Time Series Imputation

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

Which DTW Method Applied to Marine Univariate Time Series Imputation

Thi-Thu-Hong PHAN *†
†Department of Computer Science
VNUA, Hanoi, Vietnam

Émilie POISSON CAILLAULT*‡
*LISIC - ‡IFREMER

Alain LEFEBVRE ‡
‡ IFREMER, LER BL
F-62321 Boulogne-sur-mer, France

André BIGAND*
* Univ. Littoral Côte d'Opale, EA 4491-LISIC
F-62228 Calais, France

***Abstract*—Missing data are ubiquitous in any domains of applied sciences. Processing datasets containing missing values can lead to a loss of efficiency and unreliable results, especially for large missing sub-sequence(s). Therefore, the aim of this paper is to build a framework for filling missing values in univariate time series and to perform a comparison of different similarity metrics used for the imputation task. This allows to suggest the most suitable methods for the imputation of marine univariate time series. In the first step, the missing data are completed on various mono-dimensional time series. To fill a missing sub-sequence (gap) in a time series, we first find the most similar sub-sequence to the sub-sequence before (resp. after) this gap according a Dynamic Time Warping (DTW)-cost. Then we complete the gap by the next (resp. previous) sub-sequence of the most similar one. Through experiments results on 5 different datasets we conclude that i) DTW gives the best results when considering the accuracy of imputation values and ii) Adaptive Feature Based DTW (AFBDTW) metric yields very similar shape of imputation values similar to the one of true values.**

***Keywords*—Univariate time series; Missing data; Dynamic Time Warping (DTW); Derivative DTW (DDTW); Dynamic Time Warping-D (DTW-D); Adaptive Feature Based DTW (AFBDTW); Similarity measures.**

I. Introduction

Recent advances in monitoring systems, communication and information technology, storage capacity and remote sensing systems make it possible to consider huge spatial and/or time series databases. However, collected data are usually incomplete due to sensor failures, communication/transmission problems or a lack of human measures. This is particularly the case for marine samples [1], [2]. Ignoring missing data is a simple solution to deal with this drawback. But this solution may lead to serious problems, especially for time series data (the considered values would depend on the past values). First, there is a loss of information which could lose efficiency and unreliable results [3]. Second, more seriously, systematic differences between observed and unobserved data that leads to biased and unreliable results [4].

Imputation techniques are one prospective approach to solve problems with missing data [5]. In the litterature, many studies have been devoted to the imputation task of multivariate time series, such as [6]–[18] and fewer for univariate time series with missing data [5], [19]–[22]. Correlation-based method [23], and machine learning approach [24] are often used for multivariate series: missing data is filled by the value of its representative computed from others variables. Dynamic Time Warping (DTW) [25] approach is used when no information are available, the idea is to find a similar shape in a database to fill the missing values. Related works to DTW are cited below, rare works deal with large gaps in the series.

Hsu et al. [26] used k-Nearest Neighbors (k-NN) and DTW algorithms for completing DNA data. They also performed comparing different versions of DTW algorithm for better prediction and computation performance. Nevertheless, the authors did not mention to complete long missing subsequences. In [18] a weighted k-NN version is combined with DTW to compare multiple points in time simultaneously. DTW-cost is used as distance metric instead of pointwise distance measurements. Kostadinova et al. [27] proposed an Integrative DTW-Based Imputation algorithm that is particularly suited for the estimation of missing values in gene expression time series data using multiple related information in datasets. This algorithm identifies an appropriate set of estimation matrices by using DTW-cost distance in order to measure similarities between gene expression matrices. Yang et al. [28] also developed a method to impute missing values in microarray time-series data based on the combination of k-NN and DTW. In these three last cited works, the authors applied DTW method for completing missing values in multivariate data. Imputation for consecutively missing values in univariate data is not considered.

According to our knowledge, there is no application for surveying imputation algorithms with large gap(s) size using directly DTW in case of univariate time series. A gap is considering large when the process could have significant changes during this missing period. In addition, for handing missing data within univariate time series, we must only rely on the available values of this unique variable to estimate the incomplete values. Moritz et al. [22] showed that imputing univariate time series data is a particularly challenging task.

Therefore, the objective of this paper is to build a framework for filling missing values in univariate time series and to perform a comparison of different similarity DTW metrics used for the imputation task. This allows to suggest the most

suitable methods for the imputation of marine univariate time series ensuring that results are reliable and high quality.

This paper is organized as follows. Section 2 introduces materials and methods, including: Data presentation, Elastic similarity measures, Imputation based on DTW metrics, Imputation performance indicator, and Experiment protocol. Next, Section 3 demonstrates our experimental results and discussion for series with large missing subsequence. Conclusions and future works are drawn in Section 4.

II. Materials and Methods

A. Data presentation

Five datasets are used for evaluating the performance of different DTW versions, including: Cua Ong temperature, Gas online, Chlorophyll-*a*, fluorescence, and water level. The last three datasets are collected by IFREMER (France) in the eastern English Channel [29].

- Cua Ong temperature in ^{o}C - daily mean air temperature at the Cua Ong meteorological station in Vietnam from 1/1/1973 to 31/12/1999.
- Gas online - weekly data on US finished motor gasoline products supplied (in thousands of barrels per day) from 8/2/1991 to 4/11/2016 [30].
- Chlorophyll-*a* (Chla) in µg/L - weekly Chlorophyll-*a* time series from 01/1/1989 to 24/12/2014, Ifremer IGA-Gravelines monitoring [31].
- Water level in m - sampling frequency 20 minutes of water level from 01/1/2015 to 31/12/2009 [29].
- Fluorescence in FFU - sampling frequency 20 minutes of fluorescence from 1/1/2005 to 9/2/2009 [29].

In order to obtain useful information from the dataset and makes the data set easily exploitable, we analyzed these series. Table I summarizes characteristics of the datasets.

TABLE I
DATA CHARACTERISTICS BY DATASET: NUMBER OF THE DATASET, ITS NAME, THE NUMBER OF TIME SAMPLES, PRESENCE (Y=YES ELSE N=NO) OF TREND, PRESENCE OF SEASONAL CYCLE AND SAMPLING FREQUENCY

N0	Dataset name	N0 of instants	Trend (Y/N)	Seasonal (Y/N)	Frequency
1	Cua Ong temperature	9859	N	Y	Daily
2	Gas online	1344	Y	Y	Weekly
3	*Chlorophyll-a*	1352	N	N	Weekly
4	Fluorescence	106000	N	Y	20 minutes
5	Water level	131472	N	Y	20 minutes

B. Elastic similarity measures

Different DTW versions used for univariate time series imputation (namely, DTW [25], Derivative DTW (DDTW) [32], DTW-D [33], and AFBDTW (Adaptive Feature Based DTW) [34]) are studied in this paper.

Dymanic Time Warping algorithm - In general, DTW calculates an optimal match between two given sequences (*e.g.* time series) with certain restrictions. The sequences are non-linearly "warped" in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. DTW was initially introduced to recognize spoken words [25], but it has since been applied to a wide range of information retrieval and database problems. The description following is restricted to the most important steps of the algorithm (see [25], [35] for detailed explanation).

Given two time series X and Y of length n and m respectively, where: $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$.

The first step is the computation of an n-by-m distance matrix where each (i^{th}, j^{th}) element corresponds to the distance measured between x_i and y_j. Dynamic programming formulation (1) is used first to determine the cost matrix $Dist(i,j)$ by a direct pass and then to find the warping path P by a back pass $P = (p_1, p_2, \ldots, p_K)$ with $p_l = (p_i, p_j) \in [1:n]$ x $[1:m]$ for $l \in [1:K]$; $max(n,m) \leq K \leq n+m-1$. The alignment cost $DTW(x,y)$ is the last computed cost $DTW(x,y) = Dist(x_n, y_m)$.

$$Dist(i,j) = d(x_i, y_j) + min\{Dist(i-1, j-1), Dist(i-1, j), Dist(i, j-1)\} \quad (1)$$

Derivative Dymanic Time Warping algorithm (DDTW) - Keogh and Pazzani [32] replaced the signal X by a new vector $X^{'} = \{x_1, \ldots, x_i^{'}, \ldots, x_n\}$ according to the equation

$$x_i^{'} = \frac{(x_i - x_{i-1}) + \frac{x_{i+1} - x_{i-1}}{2}}{2}, 1 < i < m \quad (2)$$

This method takes into account the form of the time series and the first derivative of the sequences. So, it estimates local derivatives of the data to find the correct wrapping.

Dymanic Time Warping-D algorithm (DTW-D) - Chen et al. [33] proposed an other version of DTW by replacing DTW distances by DTW-D ones devoted to applications of semi-supervised learning, with

$$DTW - D(X,Y) = DTW(X,Y)/ED(X,Y) \quad (3)$$

Adaptive Feature Based Dymanic Time Warping algorithm (AFBDTW) - Xie and Wiltgen considered both the local and the global features of the series instead of the value itself or its derivative [34]. For each point in a sequence, a global feature and a local feature are calculated as following:

- The local feature of a data point x_i is defined as a vector of two components:

$$f_{local}(x_i) = (x_i - x_{i-1}, x_i - x_{i+1}) \quad (4)$$

- Global feature of a data point x_i is defined as a vector of two components::

$$f_{global}(x_i) = (x_i - \sum_{k=1}^{i-1} \frac{x_k}{i-1}, x_i - \sum_{k=i+1}^{n} \frac{x_k}{i-1}) \quad (5)$$

For evaluate of the distance between x_i and y_j, instead of using Euclidian distance d the authors proposed to use distances $d^{'}$ following:

$$d^{'}(x_i, y_j) = w_1\, d^{'}_{local}(x_i, y_j) + w_2\, d^{'}_{global}(x_i, y_j) \quad (6)$$

with w_1 and w_2 are weights used to adjust the percentage influence of local and global criterion. In this paper, equal influence are considered ($w_1 = w_2 = 0.5$).

$$\begin{aligned} d^{'}_{local}(x_i, y_j) = |(f_{local}(x_i))_1 - (f_{local}(y_j))_1| \\ +|(f_{local}(x_i))_2 - (f_{local}(y_j))_2| \end{aligned} \tag{7}$$

$$\begin{aligned} d^{'}_{global}(x_i, y_j) = |(f_{global}(x_i))_1 - (f_{global}(y_j))_1| \\ +|(f_{global}(x_i))_2 - (f_{global}(y_j))_2| \end{aligned} \tag{8}$$

C. Imputation based on DTW metrics

In this part, we present our method for imputing missing values of univariate time series data based on DTW metrics.

The approach consists in finding the most similar sub-sequence (Qs) to a query (Q), with Q is the sub-sequence before a gap of T size at position t ($Q = X[t - T : t - 1]$). Then, we complete this gap by the following sub-sequence of the Qs. The mechanism is illustrated on the figure 1.

To obtain the Qs similar sub-sequence, we used different versions of DTWs (as above mentioned). The dynamics and the shape of data before (resp. after) a gap are key-point of this technique. The elastic matching is used to find similar windows to the Q query of T size in the search database. Once the most similar window is identified, the following window will be copied to the location of missing values.

In order to decrease the computation time, firstly we deployed the shape-features extraction algorithm ([36]) and then applied various DTW algorithms to find imputation values. We only calculated DTW cost between the query and a reference window when the correlation between the shape-features of this window and the ones of the query is very high. The shape-features extraction algorithm is used because it better maintains the shape and dynamics of series through 9 global features (see [36] for more detail) and it is really important in our framework. In this paper, we just present the finding of similar windows before the gap. In case of finding similar windows after the gap, the method just needs to shift the corresponding index.

D. Imputation performance indicator

To assess accuracy and shape indexes of theses imputation methods, 6 indicators are computed as following:

1) Similarity - defines the similar percentage between the imputed value (Y) and the respective true values (X). It is calculated by:

$$Sim(Y, X) = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{1 + \frac{|y_i - x_i|}{max(X) - min(X)}} \tag{9}$$

Where T is the number of missing values. A higher similarity ($\in [0, 1]$) highlights a better ability to complete missing values.

2) NMAE, the Normalized Mean Absolute Error between the imputed value Y and the respective true value time series X is computed as:

$$NMAE(Y, X) = \frac{1}{T} \sum_{i=1}^{T} \frac{|y_i - x_i|}{V_{max} - V_{min}} \tag{10}$$

Where V_{max}, V_{min} are the maximum and the minimum value of original time series. A lower NMAE means better performance method for the imputation task.

3) RMSE: The Root Mean Square Error is defined as the average squared difference between the imputed value Y and the respective true value time series X. This indicator is very useful for measuring overall precision or accuracy. In general, the more effective method would have a lower RMSE.

$$RMSE(Y, X) = \sqrt{\frac{1}{T} \sum_{i=1}^{T} (y_i - x_i)^2} \tag{11}$$

4) FSD: Fraction of Standard Deviation of the imputed value Y and the respective true value time series X is defined as follows:

$$FSD(Y, X) = 2 * \frac{|SD(Y) - SD(X)|}{SD(Y) + SD(X)} \tag{12}$$

This fraction indicates whether a method is acceptable or not (here SD stands for Standard Deviation). For the imputation task, if FSD is closer to 0, the imputation values are closer to the real values.

5) FA2 - represents the fraction of data points that satisfied smoothing amplitude cover. It is calculated as:

$$FA2(Y, X) = \frac{length(0.5 \leq \frac{Y}{X} \leq 2)}{length(X)} \tag{13}$$

A model is considered perfect when its FA2 is equal to 1.

6) FB - determines whether the predicted values are overestimated or underestimated relative to those observed values.

$$FSD(Y, X) = 2 * \frac{|mean(Y) - mean(X)|}{mean(Y) + mean(X)} \tag{14}$$

A model is considered perfect when its FB tends to 0.

E. Experiment protocol

For assessing the results of imputation algorithms, we use a technique based on three steps. In the first step, we create artificial missing data by deleting data values from completed time series. The second step consists in applying the imputation algorithms to complete missing data. Finally, the third step compares the performance of different DTW metrics on various indicators as previously defined. In the present study, 5 missing data levels are considered on 5 datasets. Gaps are built at rates 0.6%, 0.75%, 1%, 1.25%, and 1.5% of the data set size (here missing sequences of the water level time series correspond to around 10 days (789 NAs) to 1 month (1972

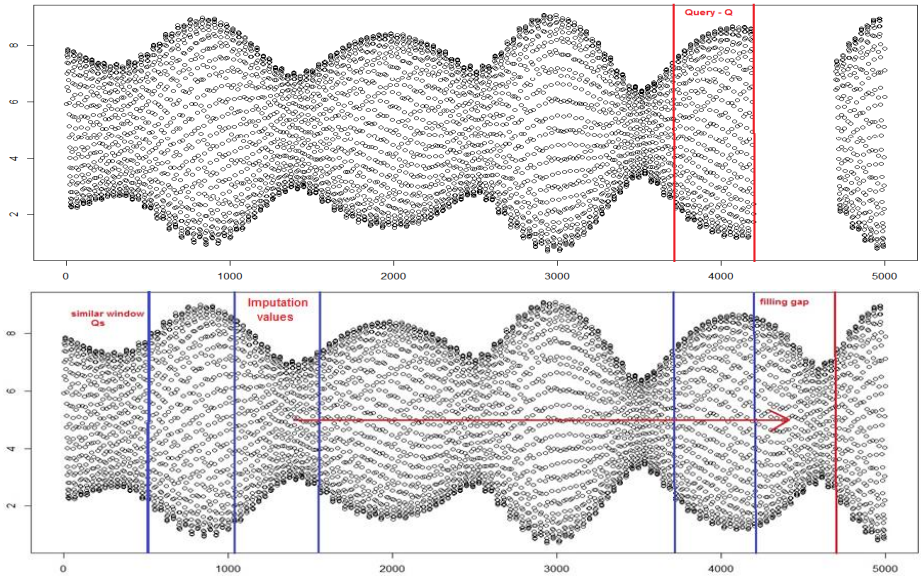

Fig. 1. Illustration of the DTW-completion process: query building (Q) and similar sequence finding (Qs), gap filling.

NAs)). For each gap, the algorithms are conducted 10 times by randomly selecting the missing positions on the data. We then run 50 iterations for each data set.

III. Experiment and Discussion

Table II, III, V, IV, and VI show average results on 6 indicators (including similarity, NAME, RMSE, FSD, FA2, and FB) using different DTW versions for completing missing data applied on 5 time series.

From the results of these tables, we find that DTW metric provides the best results on the accuracy indexes: the highest similarity and the lowest NMAE and RMSE at every missing level for all datasets. However, when considering on other indexes such as FSD, FA2 and FB (we call shape indexes), DTW no longer performs well as on the accuracy indicators.

With Cua Ong temperature (table II) and Gas online (table III) series, DTW still proves its ability on the FB index at all missing rate. For the remaining datasets (Fluorescence, water level, Chla datasets), DTW only highlights its performance at small missing rates.

According to Keogh et Pazzani [32], DDTW method presents better performance than the original DTW by minimizing the number of duplicate points. However, DDTW is not suitable for handing the imputation task, it does not prove its ability here.

AFBDTW is proposed in 2010 by Xie and Wiltgen [34]. This method takes into account both the local and global features of the series for correspondences points instead of the value itself or its derivative. That is the reason why AFBDTW proves the strength for the imputation task at large missing rates, specially in large datasets.

DTW-D method is proposed for semi-supervisor classification. Therefore, when we applied this method to complete

TABLE II
Average imputation performance indexes of various similarity metrics on Cua Ong temperature series

Gap size	Metric	Accuracy indexes			Shape indexes		
		1-Sim	NMAE	RMSE	FSD	1-FA2	FB
0.6%	DTW	**0.209**	**0.118**	**37.001**	0.269	0.005	**0.083**
	DDTW	0.232	0.138	43.003	0.333	0.008	0.118
	DTW-D	0.273	0.160	48.372	0.307	0.005	0.152
	AFBDTW	0.228	0.126	39.099	**0.252**	**0.000**	0.103
0.75%	DTW	**0.212**	**0.122**	**38.033**	**0.168**	0.014	**0.090**
	DDTW	0.237	0.145	44.627	0.200	**0.008**	0.141
	DTW-D	0.270	0.184	53.756	0.267	0.064	0.175
	AFBDTW	0.224	0.142	44.297	0.188	0.030	0.131
1%	**DTW**	**0.164**	**0.099**	**31.952**	0.159	**0**	**0.013**
	DDTW	0.171	0.106	33.561	0.176	0.008	0.060
	DTW-D	0.188	0.123	39.209	0.228	0.010	0.078
	AFBDTW	0.173	0.104	33.537	**0.125**	0.005	0.043
1.25%	DTW	**0.150**	**0.108**	**34.315**	**0.151**	0.003	**0.036**
	DDTW	0.166	0.124	39.871	0.298	**0.002**	0.076
	DTW-D	0.160	0.119	37.711	0.228	0.008	0.074
	AFBDTW	0.155	0.113	36.699	0.181	0.003	0.072
1.5%	DTW	**0.141**	**0.110**	**35.649**	0.124	0.011	**0.035**
	DDTW	0.191	0.164	51.600	0.159	0.020	0.136
	DTW-D	0.147	0.115	36.399	**0.088**	**0.005**	0.060
	AFBDTW	0.142	0.111	36.656	0.102	0.009	0.048

TABLE III
AVERAGE IMPUTATION PERFORMANCE INDEXES OF VARIOUS SIMILARITY METRICS ON GAS ONLINE SERIES

Gap size	Metric	Accuracy indexes			Shape indexes		
		1-Sim	NMAE	RMSE	FSD	1-FA2	FB
0.6%	DTW	**0.293**	**0.094**	**392.806**	0.385	0	**0.031**
	DDTW	0.303	0.100	413.314	**0.355**	0	0.033
	DTW-D	0.336	0.113	457.966	0.438	0	0.031
	AFBDTW	0.453	0.237	894.008	0.460	**0**	0.094
0.75%	DTW	**0.287**	**0.106**	**452.470**	**0.328**	0	**0.031**
	DDTW	0.330	0.137	560.240	0.484	0	0.051
	DTW-D	0.330	0.131	533.966	0.440	0	0.047
	AFBDTW	0.455	0.237	891.465	0.351	**0**	0.095
1%	**DTW**	**0.276**	**0.115**	**476.098**	0.203	**0**	**0.039**
	DDTW	0.328	0.146	575.640	0.311	0	0.053
	DTW-D	0.315	0.131	545.698	**0.174**	0	0.046
	AFBDTW	0.384	0.227	859.176	0.304	0	0.084
1.25%	DTW	**0.288**	**0.102**	**433.679**	0.266	0	**0.028**
	DDTW	0.299	0.116	473.552	0.325	**0**	0.036
	DTW-D	0.313	0.118	482.555	**0.241**	0	0.036
	AFBDTW	0.300	0.113	457.787	0.341	0	0.037
1.5%	DTW	**0.234**	**0.131**	**549.911**	0.201	0	**0.047**
	DDTW	0.277	0.168	655.410	0.238	0	0.066
	DTW-D	0.266	0.149	598.538	**0.121**	**0**	0.048
	AFBDTW	0.346	0.216	820.442	0.280	0	**0.084**

TABLE IV
AVERAGE IMPUTATION PERFORMANCE INDEXES OF VARIOUS SIMILARITY METRICS ON FLUORESCENCE SERIES

Gap size	Metric	Accuracy indexes			Shape indexes		
		1-Sim	NMAE	RMSE	FSD	1-FA2	FB
0.6%	DTW	**0.160**	**0.028**	**1.569**	**0.531**	**0.462**	**0.423**
	DDTW	0.189	0.032	1.767	1.120	0.662	0.871
	DTW-D	0.327	0.067	3.732	0.950	0.740	1.060
	AFBDTW	0.198	0.035	1.991	0.853	0.545	0.685
0.75%	DTW	**0.187**	**0.032**	**1.800**	**0.616**	**0.512**	**0.505**
	DDTW	0.190	0.034	1.883	1.364	0.731	0.974
	DTW-D	0.378	0.101	5.272	1.175	0.802	1.219
	AFBDTW	0.212	0.036	2.068	0.654	0.576	0.724
1%	DTW	**0.150**	**0.027**	**1.579**	0.838	**0.550**	**0.711**
	DDTW	0.172	0.035	1.963	1.411	0.854	1.236
	DTW-D	0.295	0.070	3.749	1.122	0.778	1.141
	AFBDTW	0.157	0.027	1.606	**0.782**	0.606	0.800
1.25%	DTW	**0.157**	**0.027**	**1.655**	0.913	0.630	0.794
	DDTW	0.175	0.034	1.925	1.415	0.825	1.132
	DTW-D	0.362	0.104	5.740	1.218	0.834	1.302
	AFBDTW	0.160	0.030	1.756	**0.778**	**0.629**	**0.744**
1.50%	DTW	**0.119**	**0.028**	**1.689**	1.033	0.659	0.790
	DDTW	0.123	0.031	1.820	1.270	0.813	0.957
	DTW-D	0.259	0.083	4.690	1.042	0.811	1.145
	AFBDTW	0.142	0.038	2.319	**0.791**	**0.622**	**0.656**

TABLE V
AVERAGE IMPUTATION PERFORMANCE INDEXES OF VARIOUS SIMILARITY METRICS ON CHLA SERIES

Metric	Accuracy indexes			Shape indexes		
	1-Sim	NMAE	RMSE	FSD	1-FA2	FB
DTW	**0.308**	**0.069**	**4.609**	**0.597**	**0.413**	**0.381**
DDTW	0.339	0.091	5.707	0.692	0.463	0.476
DTW-D	0.356	0.090	5.915	0.831	0.450	0.543
AFBDTW	0.386	0.089	5.962	0.759	0.463	0.641
DTW	**0.243**	**0.076**	**5.136**	**0.525**	**0.360**	**0.311**
DDTW	0.254	0.076	5.171	0.582	0.400	0.355
DTW-D	0.303	0.094	6.481	0.897	0.480	0.492
AFBDTW	0.281	0.086	5.876	0.646	0.460	0.535
DTW	**0.185**	**0.071**	**4.990**	**0.444**	**0.393**	**0.394**
DDTW	0.205	0.088	6.207	0.501	0.443	0.468
DTW-D	0.236	0.093	6.557	0.642	0.486	0.637
AFBDTW	0.198	0.086	6.046	0.545	0.450	0.486
DTW	**0.187**	**0.089**	**6.488**	0.812	**0.429**	0.526
DDTW	0.203	0.103	7.076	0.687	0.500	0.475
DTW-D	0.216	0.105	7.352	0.775	0.518	0.409
AFBDTW	0.222	0.104	7.136	**0.686**	0.512	**0.404**
DTW	**0.205**	**0.090**	**6.226**	0.435	0.545	0.408
DDTW	0.216	0.097	6.772	**0.407**	**0.515**	0.460
DTW-D	0.218	0.097	6.865	0.655	0.550	0.463
AFBDTW	0.217	0.098	6.721	0.510	0.525	**0.376**

TABLE VI
AVERAGE IMPUTATION PERFORMANCE INDEXES OF VARIOUS SIMILARITY METRICS ON WATER LEVEL SERIES

Gap size	Metric	Accuracy indexes			Shape indexes		
		1-Sim	NMAE	RMSE	FSD	1-FA2	FB
0.6%	DTW	**0.042**	**0.037**	**0.401**	**0.045**	**0**	**0.019**
	DDTW	0.042	0.037	0.402	0.045	0	0.022
	DTW-D	0.139	0.141	1.434	0.103	0.059	0.005
	AFBDTW	0.079	0.074	0.765	0.051	0.002	0.019
0.75%	DTW	**0.037**	**0.033**	**0.355**	**0.017**	**0**	**0.009**
	DDTW	0.042	0.038	0.401	0.019	0	0.010
	DTW-D	0.154	0.162	1.624	0.075	0.082	0.010
	AFBDTW	0.076	0.073	0.750	0.039	0.008	0.022
1%	**DTW**	**0.033**	**0.030**	**0.333**	0.026	**0**	**0.012**
	DDTW	0.034	0.030	0.333	0.027	0	0.014
	DTW-D	0.107	0.108	1.141	0.047	0.034	0.013
	AFBDTW	0.082	0.080	0.828	**0.025**	0.009	0.017
1.25%	DTW	**0.039**	**0.035**	**0.373**	0.025	**0**	0.009
	DDTW	0.039	0.035	0.373	0.025	0	0.009
	DTW-D	0.086	0.086	0.965	0.034	0.019	0.019
	AFBDTW	0.047	0.044	0.471	**0.018**	0.001	**0.009**
1.5%	DTW	**0.045**	**0.042**	**0.442**	0.030	**0**	0.022
	DDTW	0.045	0.043	0.450	0.032	0	0.025
	DTW-D	0.073	0.073	0.841	0.021	0.012	0.008
	AFBDTW	0.061	0.060	0.635	**0.020**	0.009	**0.015**

missing values, DTW-D does not work well in all datasets at every missing level. Nevertheless, when looking at FSD indicator in the table III, DTW-D gives the best results at large gaps ($\geq 1\%$). The reason may be that Gas online series has both trend and seasonality component.

Besides, the shape of imputation values generated from methods using various DTW metrics (DTW, DDTW, DTW-D, AFBDTW) are also analyzed. Fig. 2 presents the form of imputed values yielded by methods using different similarity metrics with the true values at position 444, the gap size of 14 (approximate 3 months of missing values) on the Chlorophyll-*a*. DTW metric proves again its capability to deal with missing subsequence. The shape of the imputation values generated from the method using DTW and the one of true values are

very close.

After the comparison of quantitative and visual performance of different DTW versions, we carry out examining computational time of each metric. Table VII shows that for large datasets or large gaps, AFBDTW requires the longest computational time and DTW has at least computing time.

In this paper, we also calculated Cross-Correlation (CC) coefficients between the query and each reference window and the maximum coefficient is extracted. CC demonstrates that a pattern (here that is the query) exists or not in the database. High CC value means that there exists one or more recurrence of the pattern in the database, that means: it is easy to find similar patterns. In Table VIII, we see that only for water level series, CC values are very high (approximate 1), this explains why the similarity values are very high and the error index is very low.

IV. Conclusion

This paper proposes a visual and quantitative comparison of performance between different DTW versions for univariate time series imputation. The obtained results shows that when considering the accuracy of imputation values, DTW is the best robust and when regarding the shape of completed values for the large gaps and datasets, AFBDTW is more suitable. The paper highlights two mains contributions. Firstly, we perform completing large missing subsequences in time series data. Secondly, we provide a quantitative and visual comparison of different DTW algorithms applied to various datasets. The present work will allow to measure the imputation values of multivariate time series in the future.

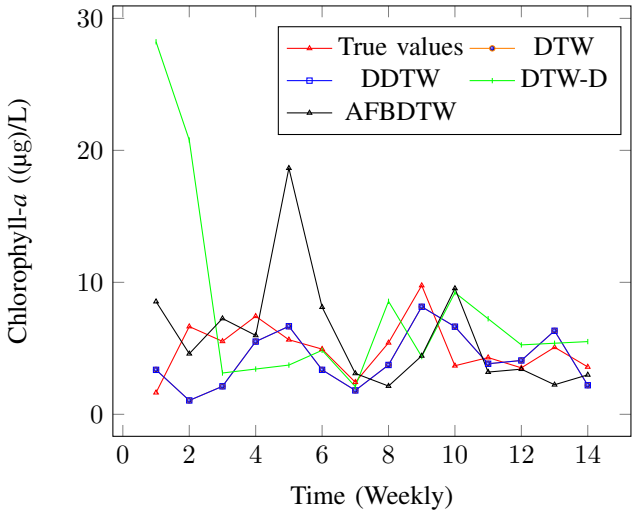

Fig. 2. Visual comparison of imputed values using different DTW metrics with true values on *Chla* series at position 444 at missing rate 1% (correspond to 14 weeks missing).

Acknowledgments

This work has been funded by the Ministry of Education and Training Vietnam International Education Development, the French government and the region Hauts-de-France in the framework of the project CPER 2014-2020 MARCO.

TABLE VII
COMPUTATIONAL TIME OF METHODS USING DIFFERENT DTW METRICS AT MISSING RATE 0.6% ON VARIOUS SERIES

Method	Cua Ong temperature	Gas online	Chla	Fluores-cence	Water level
DTW	12.459	1.670	1.08	774.718	2081.388
DDTW	13.112	1.700	1.07	786.543	2126.847
DTW-D	12.543	1.671	1.10	761.831	2088.375
AFBDTW	62.602	1.539	1.07	14219.51	49095.888

TABLE VIII
THE MAXIMUM OF CROSS-CORRELATION BETWEEN THE QUERY AND REFERENCE WINDOWS.

Gap size	Cua Ong temperature	Gas online	Chla	Fluo	Water level
0.6%	0.751	0.921	0.93	0.657	0.997
0.75%	0.762	0.889	0.92	0.694	0.996
1%	0.780	0.819	0.86	0.710	0.996
1.25%	0.789	0.788	0.86	0.753	0.996
1.50%	0.825	0.778	0.87	0.731	0.996

References

[1] K. Rousseeuw, E. P. Caillault, A. Lefebvre, and D. Hamad, "Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling," in *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS*. IEEE, 2013, pp. 3962–3965.

[2] H.-T. Ceong, H.-J. Kim, and J.-S. Park, "Discovery of and recovery from failure in a costal marine usn service," *Journal of Information and Communication Convergence Engineering*, vol. 1, no. 1, Mar 2012. [Online]. Available: http://dx.doi.org/10.6109/jicce.2012.10.1.011

[3] N. M. Noor, M. M. Al Bakri Abdullah, A. S. Yahaya, and N. A. Ramli, "Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set," *Materials Science Forum*, vol. 803, pp. 278–281, Aug. 2014.

[4] G. Hawthorne, G. Hawthorne, and P. Elliott, "Imputing cross-sectional missing data: Comparison of common techniques," *Australian and New Zealand Journal of Psychiatry*, vol. 39, no. 7, pp. 583–590, 2005.

[5] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, no. 18, pp. 2895–2907, Jun. 2004.

[6] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. CRC Press, Aug. 1997.

[7] S. Van Buuren, H. C. Boshuizen, D. L. Knook, and others, "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in medicine*, vol. 18, no. 6, pp. 681–694, 1999.

[8] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, vol. 27, no. 1, pp. 85–96, 2001.

[9] J. Engels and P. Diehr, "Imputation of missing longitudinal data: A comparison of methods," *Journal of Clinical Epidemiology*, vol. 56, no. 10, pp. 968–976, Oct. 2003.

[10] P. Royston, "Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring," *Stata Journal*, vol. 7, no. 4, pp. 445–464, 2007.

[11] E. A. Stuart, M. Azur, C. Frangakis, and P. Leaf, "Multiple Imputation With Large Data Sets: A Case Study of the Children's Mental Health Initiative," *American Journal of Epidemiology*, vol. 169, no. 9, pp. 1133–1139, May 2009.

[12] K. J. Lee and J. B. Carlin, "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation," *American Journal of Epidemiology*, vol. 171, no. 5, pp. 624–632, Mar. 2010.

[13] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study," *American Journal of Epidemiology*, vol. 179, no. 6, pp. 764–774, Mar. 2014.

[14] S. G. Liao, Y. Lin, D. D. Kang, D. Chandra, J. Bon, N. Kaminski, F. C. Sciurba, and G. C. Tseng, "Missing value imputation in high-dimensional phenomic data: Imputable or not, and how?" *BMC Bioinformatics*, vol. 15, p. 346, 2014.

[15] S. A. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg, "Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data," *Journal of Biomedical Informatics*, vol. 58, pp. 198–207, Dec. 2015.

[16] A. Gelman, J. Hill, Y.-S. Su, M. Yajima, M. Pittau, B. Goodrich, Y. Si, and J. Kropko, "Mi: Missing Data Imputation and Model Checking," Apr. 2015.

[17] Y. Deng, C. Chang, M. S. Ido, and Q. Long, "Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data," *Scientific Reports*, vol. 6, p. 21689, Feb. 2016.

[18] S. Oehmcke, O. Zielinski, and O. Kramer, "kNN ensembles with penalized DTW for multivariate time series imputation," in *Neural Networks (IJCNN), 2016 International Joint Conference On*. IEEE, 2016, pp. 2774–2781.

[19] P. D. Allison, *Missing Data | SAGE Publications Ltd*. SAGE, 2001.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[21] S. Chiewchanwattana, C. Lursinsap, and C.-H. Henry Chu, "Imputing incomplete time-series data based on varied-window similarity measure of data sequences," *Pattern Recognition Letters*, vol. 28, no. 9, pp. 1091–1103, Jul. 2007.

[22] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, and J. Stork, "Comparison of different Methods for Univariate Time Series Imputation in R," *arXiv preprint arXiv:1510.03924*, 2015.

[23] S. Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *Journal of statistical software*, vol. 45, no. 3, 2011.

[24] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012.

[25] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, vol. 16, pp. 43–49, 1978.

[26] H.-H. Hsu, A. C. Yang, and M.-D. Lu, "KNN-DTW Based Missing Value Imputation for Microarray Time Series Data," *Journal of Computers*, vol. 6, no. 3, Mar. 2011.

[27] E. Kostadinova, V. Boeva, L. Boneva, and E. Tsiporkova, "An Integrative DTW-based imputation method for gene expression time series data," in *Intelligent Systems (IS), 2012 6th IEEE International Conference*. IEEE, 2012, pp. 258–263.

[28] A. C. Yang, H.-H. Hsu, and M.-D. Lu, "Missing Value Imputation in Microarray Gene Expression Data," in *Conference on Information Technology and Applications in Outlying Islands*, 2009.

[29] A. Lefebvre, "MAREL Carnot data and metadata from Coriolis Data Centre. SEANOE," http://doi.org/10.17882/39754, 2015.

[30] US Energy Information Administration, "Product Supplied for Finished Gasoline," http://www.eia.gov/dnav/pet/PET_SUM_SNDW_A_EPM0F_VPP_MBBLPD_W.htm, 2016, gas_online_product_2016.

[31] D. Woehrling, A. Lefebvre, and R. Le Fèvre-Lehoërff, G.and Delesmont, "Seasonal and longer term trends in sea temperature along the french north sea coast, 1975 to 2002." *Journal of the Marine Biological Association U.K., 85 (1)*, pp. 39–48, 2005.

[32] E. J. Keogh and M. J. Pazzani, "Derivative Dynamic Time Warping." in *Sdm*, vol. 1. SIAM, 2001, pp. 5–7.

[33] Y. Chen, B. Hu, E. Keogh, and G. E. Batista, "DTW-D: Time series semi-supervised learning from a single example," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2013, pp. 383–391.

[34] "Adaptive Feature Based Dynamic Time Warping," vol. 10, pp. 264–273, Jan. 2010, oCLC: 255558624.

[35] F. Hermans and E. Tsiporkova, "Merging microarray cell synchronization experiments through curve alignment," *Bioinformatics*, vol. 23, no. 2, pp. e64–e70, Jan. 2007.

[36] T. T. H. Phan, E. P. Caillault, and A. Bigand, "Comparative study on supervised learning methods for identifying phytoplankton species," in *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE)*. IEEE, Jul. 2016, pp. 283–288.