HAL
open science

DTW-APPROACH FOR UNCORRELATED MULTIVARIATE TIME SERIES IMPUTATION

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

To cite this version:

Thi-Thu-Hong Phan, Alain Lefebvre, Émilie Poisson Caillault, André Bigand. DTW-APPROACH FOR UNCORRELATED MULTIVARIATE TIME SERIES IMPUTATION. 27th International Workshop on Machine Learning for Signal Processing (MLSP 2017), Sep 2017, Tokyo, Japan. pp.1-6, 10.1109/MLSP.2017.8168165 . hal-01609267

HAL Id: hal-01609267

https://hal.science/hal-01609267v1

Submitted on 3 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DTW-APPROACH FOR UNCORRELATED MULTIVARIATE TIME SERIES IMPUTATION

Thi-Thu-Hong PHAN[1,2], *Émilie POISSON CAILLAULT*[1], *André BIGAND*[1], *Alain LEFEBVRE*[3]

[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
[3] IFREMER, LER BL, F-62321 Boulogne-sur-mer, France

ABSTRACT

Missing data are inevitable in almost domains of applied sciences. Data analysis with missing values can lead to a loss of efficiency and unreliable results, especially for large missing sub-sequence(s). Some well-known methods for multivariate time series imputation require high correlations between series or their features. In this paper, we propose an approach based on the shape-behaviour relation in low/un-correlated multivariate time series under an assumption of recurrent data. This method involves two main steps. Firstly, we find the most similar sub-sequence to the sub-sequence before (resp. after) a gap based on the shape-features extraction and Dynamic Time Warping algorithms. Secondly, we fill in the gap by the next (resp. previous) sub-sequence of the most similar one on the signal containing missing values. Experimental results show that our approach performs better than several related methods in case of multivariate time series having low/non-correlations and effective information on each signal.

Index Terms— Imputation; Uncorrelated multivariate time series; Missing data; Dynamic Time Warping; Similarity measures.

1. INTRODUCTION

Recent advances in monitoring systems, the availability of effective low-cost sensors, and the deployment of remote sensing systems make it possible to consider huge time series and/or spatial databases. Most proposed methods for analysis of multivariate time series require complete data but collected data are usually incomplete due to sensor failures, communication/transmission problems or bad weather conditions for manual measures or maintenance. This is particularly the case for marine samples [1, 2].

For instance, let consider the Marel Carnot station - a marine water monitoring platform in the eastern English Channel, France [3], 19 large time series are collected every 20 minutes as fluorescence, turbidity, oxygen saturation, water level, water temperature signals, ... This data contains missing values. The size of consecutive missing data are various from one hour to few months (too large) and the phytoplankton bloom dynamics change too fast to use linear or spline imputation.

A simple rule to deal with missing values is to ignore missing values. But serious problems often occur using this solution, especially in the case of time series data (the considered values would depend on the past ones). An analysis of systematic differences between observed and unobserved data may lead to biased and unreliable results [4]. Imputation techniques are one prospective approach to solving missing data problems [5].

In the litterature, many successful studies have been devoted to multivariate time series imputation such as [6–16]. There are 2 methods of model-based imputation: the first method based on the multivariate normal (MVN), originally developed by Schafer [6]. MVN applied Markov chain Monte Carlo algorithm to compute imputed values under the assumption that all variables follow a multivariate normal distribution. And, the second method used a chained equations to complete missing data (MICE) implemented by van Buuren et al. [7, 17] and Raghunathan et al. [8]. For each variable containing missing values, MICE exploited the conditional distribution of all the other variables to estimate imputed values.

Besides these two main techniques, machine learning techniques like Random Forest or K-nearest Neighbors are used to predict missing values of one series according to the observed series. In this way, Stekhoven and Bühlmann [18] investigated a random forest-based method (called missForest). In the [11], Shah et al. implemented a new version of MICE which imputes each variable using the random forest method to perform better than the original multiple imputation methods.

K-Nearest Neighbors (k-NN) imputation fills in missing values by a function built from the average K similar patterns in the space of the available features. Four modification versions of k-NN for high-dimensional imputation were suggested by Liao et al. [12]. In other work, Rahman et al. [13] combined a lagged k-NN with Fourier methods for the imputation of biomedical time series data. Other studies combined k-NN algorithms and Dynamic Time Warping (DTW) to consider a temporal windows around the missing value or interval: DNA data imputation [19], microarray time-series data imputation [20], gene expression imputation [21]. Recently, in [16] a weighted k-NN version was combined with DTW to compare multiple points in time simultaneously. DTW-cost was used as distance metric instead of pointwise distance measurements. And, the authors also pointed out that datasets having high correlation between features are required to outperform Random Forest imputation approach [16].

Almost above imputation algorithms for multivariate time series usually exploit the correlations between features to estimate missing values. These correlations make it possible to use the values of observed variables to predict the others containing missing data. However, it is not efficient for multivariate series having low-or uncorrelated features (case of Marel Carnot dataset). For handling missing values or intervals in this case, we must only rely on the available values of the unique variable containing missing data to estimate the incomplete values. Furthermore, it is important to assure an acceptable similarity for each signal within the time series in the same temporal window.

Therefore, in this paper, we propose an efficient method for filling missing values in low/un-correlated multivariate time series under an only assumption of effective patterns (here a pattern corresponds to the sub-sequence before (resp. after) a gap).

This paper is organized as follows. Section 2 details the proposed approach and some available multivariate time series imputation algorithms. Section 3 explains experimental protocol including data presentation, imputation performance criteria and experimental process. Next, Section 4 demonstrates results and discussion for the completion of large missing sub-sequences. Conclusions and future work are drawn in Section 5.

2. IMPUTATION METHODS

2.1. DTWUMI - Proposed approach

In this part, we present our method for imputing missing intervals of low/un-correlated multivariate time series data based on DTW metric, named DTWUMI.

A multivariate time series is represented as $N \times M$ matrix X 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 vector at the t-th observation of all variables.

X is referred as incomplete time series when it contains missing values (or values are Not Available-NA). 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 ($\exists i | \forall t \in [t, t+T-1], x(t,i) = NA$).

A gap is considered large when $T \geq 5\%N$ for small time series ($N < 10,000$) or when T is larger than known process change (this depends on each application).

Figure 1 illustrates the mechanism of DTWUMI method. The approach consists in finding the most similar sub-sequence (Qs) to a query (Q), with Q is the sub-sequence after (resp. before) a gap. We then complete this gap by the previous (resp. following) sub-sequence of the Qs.

To obtain the Qs similar sub-sequence, we apply the principles of Dynamic Time Warping [22]. The dynamics and the shape of data after (resp. before) a gap are key point of this technique. Besides, conserving the same temporal window on all variables is another important factor of our algorithm. This means we create the query on all variables Q of size T (see figure 1) and look for the similar windows in the search database based on the elastic matching of multidimensional signals. Once the most similar window is identified, the previous window on the incomplete signal will be copied to the location of missing values.

In addition, the DTW algorithm requires long computational time. In order to decrease the computation time, before using DTW method to estimate imputation values, we deployed the shape-features extraction algorithm [23]. We only calculate DTW cost of 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 utilized because it better maintains the shape and dynamics of series through 9 global features (see [23] for more details).

2.2. Multivariate time series imputation algorithms

We compare our method with several commonly multivariate time series imputation approaches used state-of-the-art (including MI, MICE, na.approx, missForest). R language is applied to implement all these methods.

MI- Multiple Imputation [24]: For each observation in a variable containing missing values, this method predicts imputed value by finding an observation (from available values) with the closest predictive mean to that variable. Bayesian models and weakly informative prior distributions are used to construct more stable estimates of imputation models; multiple chains are run and convergence is assessed after a pre-specified number of iterations within each chain.

MICE - Multivariate Imputation via Chained Equations [17]: This method is based on the conditional (on all of other variables) distribution for each variable containing missing values to estimate imputed ones under the assumption that the missing data are missing at random (that means a missing value depends only on available values and can be estimated based on them).

Linear interpolation - na.approx (zoo package) [25]: This algorithm uses a generic function with interpolated values to estimate each missing data.

missForest [18]: This approach is based on random forest algorithm for filling in missing data. For each variable missForest builds a random forest model on the observed part. Then this model is used to predict missing values in the variable. The algorithm continues to repeat these two steps until a stopping criterion is met or the user specified maximum of iterations is reached. For further details see [18].

3. EXPERIMENT PROTOCOL

To validate our approach and compare with other methods, we conduct experiments on 3 different datasets with the same protocol and gaps detailed as follows.

3.1. Data presentation

Three multivariate time series are handled in this paper. We choose one from KEEL repository, one simulated dataset (this permits to control the criterion of correlations and the amount of missing data) and one real dataset hourly collected by IFREMER (France) in the eastern English Channel.

NNGC1_F1_V1_003 (NNGC) dataset [26]: This time series contains transportation data (4 attributes and 1745 instants) including highway traffic, traffic data of cars in tunnels, traffic at automatic payment systems on highways, traffic of individuals on subway systems, domestic aircraft flights, shipping imports, border crossings, pipeline flows and rail transportation. The data contains a time series of hourly frequency.

Simulated dataset: We have created a simulated dataset as follows: the first signal, 5 sine functions with various frequencies and amplitudes are generated $F = \{f_1, f_2, f_3, f_4, f_5\}$. Next, we add 3 different noise levels to F data $S = \{F, F + noise1, F + noise2, F + noise3\}$. Then S data is repeated 4 times (the size of this dataset is 32,000). The second and third signals are constructed based on the first signal to satisfy that the correlations between these series are very low ($\leq 0.1\%$). Corgen function of ecodist R-package [27] is used for generating the last two signals.

Marel Carnot dataset [3]: These real data contain a set of various signals such as nitrate, silicate, oxygen saturation, pH, water temperature, fluorescence, water level,... that characterize sea water. They are collected from 1/1/2005 to 9/2/2009 at hourly frequency (and consist of 35,334 time samples). However, this dataset contains a lot of missing values, the number of missing ones is different from each signal. In order to evaluate the performance of the proposed method and to compare with other methods, we selected a subset containing water level, fluorescence, and water temperature (the water level and the fluorescence series are full data, while water temperature signal has a few missing values). Also, these series have low correlations.

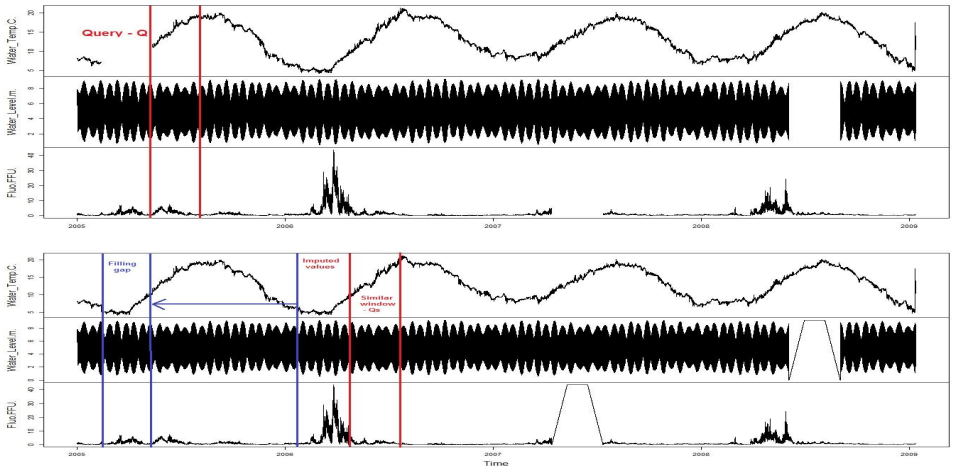

Fig. 1. Illustration of the DTW-completion process: query building and similar sequence research, gap filling.

After filling in missing data, imputation values will be compared with observed ones in the full dataset so we need to complete the water temperature signal. For guaranteeing "fair-play" to all methods, we utilized interpolation algorithm to complete these missing data in the water temperature variable.

3.2. Imputation performance analysis

To assess theses imputation methods, 6 indicators are computed including Similarity, R^2 score, RMSE (for evaluating the accuracy) and FSD, FA2, FB (for evaluating the shape).

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

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

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

$\boldsymbol{R^2}$ **score** is calculated as the square of coefficient of correlation between Y and X. A high score implies that imputation values are very closer to true values.

RMSE - Root Mean Square Error is defined as the average squared difference between Y and X (eq. 2). This indicator is very useful for measuring overall precision or accuracy. In general, the more effective method would have a lower RMSE.

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

FSD - Fraction of Standard Deviation between Y and X) defined by eq. 3 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.

$$FSD(Y,X) = 2 * \frac{|SD(Y) - SD(X)|}{SD(Y) + SD(X)} \quad (3)$$

FA2 represents the fraction of data points that satisfied smoothing amplitude cover (eq. 4). A model is considered impeccable when its FA2 is equal to 1.

$$FA2(Y,X) = \frac{length(0.5 \leq \frac{Y}{X} \leq 2)}{length(X)} \quad (4)$$

FB - Fractional Bias is defined by eq. 5. When the FB tends to 0, a model is considered perfect.

$$FB(Y,X) = 2 * \left|\frac{mean(Y) - mean(X)}{mean(Y) + mean(X)}\right| \quad (5)$$

3.3. Experiment process

We apply a technique based on three steps to evaluate the results in the following:

- *The* 1st *step*: Create artificial missing data by deleting data values from completed time series.
- *The* 2nd *step*: Use the imputation algorithms to complete missing data.
- *The* 3rd *step*: Assess the performance of proposed method and compare with published algorithms using the various indicators previously defined.

In the present study, 7 missing data levels are considered on 3 datasets. Gaps are built at rates 1%, 2%, 3%, 4%, 5%, 7.5% and 10% of the dataset size on every signal (here missing sequences on each variable of the Marel Carnot series correspond to around 15 days (353 consecutive missing) to 5 months (3533 NAs)). For each missing ratio, the algorithms are conducted 5 times by randomly selecting the missing positions on the data. We then run 35 iterations for each data set.

4. RESULTS AND DISCUSSION

Table 1 presents the average performance evaluation of different imputation algorithms for NNGC, simulated and Marel Carnot time series for the 6 indicators. The best results for each missing rate are highlighted in bold. These results confirm the good ability of DTWUMI for filling missing values in uncorrelated multivariate time series.

NNGC dataset: Table 1 shows a comparison of five imputation methods on NNGC dataset that has 7 missing ratios (1-10% missing values). We clearly find that missForest gives the highest similarity, R^2, FA2 and the lowest RMSE at every missing level. MICE is following the missForest method on these indicators. However, when considering on other indexes such as FSD and FB, missForest only proves its performance at small missing rates ($\leq$ 3%). At larger missing levels (4%-7.5%), MICE provides the smallest FB indicator. And at 5%-10% missing rates MI gives best FSD. A lower value indicates better performance. The results can explain that NNGC dataset has high correlations between variables (approximate 0.79). MICE and missForest estimate missing data based on other observed variables. That is why these algorithms have better results and our algorithm does not prove its performance when completing datasets having high correlations. MI is also based on observed values for filling in missing data but under an assumption that all variables follow a multivariate normal distribution. So with this dataset, this method does not give good performance as MICE or missForest.

Simulated and Marel Carnot datasets: From the results of table 1, it is clear that missForest, MI, and MICE do not demonstrate their performance for completing missing data on these two datasets. For all missing rates, MissForest is ranked the second as considering similarity and RMSE indexes (the simulation data) and the third or below for all indicators (Marel Carnot series). Because these two datasets have very low correlations between variables, especially for the simulated series which is an almost uncorrelated dataset. That explains why, DTWUMI illustrates the best ability for imputation task: the highest similarity, R^2, FA2 and the lowest RMSE, FSD for all missing ratios (table 1 - Simulated dataset). Regarding Marel Carnot series, this dataset has low correlations (around 0.2), so that our approach, DTWUMI, does not show the capability to fill in missing values as it does in the simulated dataset (table 1 - Marel Carnot dataset). However, this method definitely indicates its imputation performance when considering similarity, R^2, FA2, RMSE indicators at every missing level. In particular, our method further proves the ability to fill in incomplete data with large missing rates (7.5% and 10% on Marel Carnot dataset). These gaps correspond to 110.4 and 147.2 days sampled at hourly frequency.

With the NNGC series (table 1), the na.approx method always produces the worst result for every indicator. On the simulated and Marel Carnot datasets, this method gives quite good results when comparing the quantitative performance: the lowest FB and/or FSD at some missing rates (simulated series), the second rank on similarity, R^2, FA2 for all missing ratios (Marel Carnot dataset). However, when looking at the shape of imputation values generated from of this method, it absolutely shows the worst shape (figure 2, 3).

In this study, we also carry out comparing the visualization performance of imputation values generated from different methods. Figure 2 presents the shape of imputed values yielded by five different methods on the NNGC series. The missForest approach proves again the capability to deal with the successive missing of a correlated dataset. The form of imputation values produced from missForest method is very close to the form of true values. However, with low-correlated dataset as Marel Carnot data, missForest no longer demonstrates its ability (figure 3). In this case, our approach confirms its performance for the imputation task. The shape of DTWUMI's imputed values is almost identical to the form of true values (figure 3).

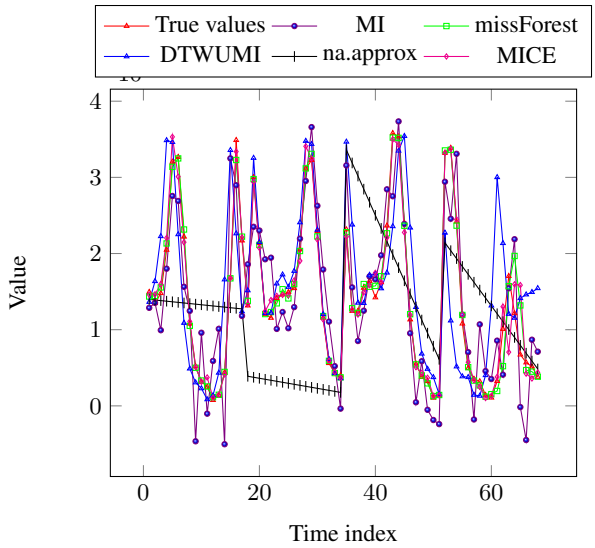

Fig. 2. Visual comparison of imputed values of different imputation methods with true values on NNGC series with the gap size of 17 on each signal.

5. CONCLUSION

In this paper, we propose an effective approach for uncorrelated multivariate time series imputation, namely DTWUMI. We have performed several experiments on artificial and real datasets to demonstrate the capability of our approach and compared it with published algorithms (na.approx, MI, MICE, and missForest) on quantitative and shape indicators. The visual performance of these methods is also evaluated. The obtained results clearly show that our approach provides better performance than the other existing methods in case of time series having low or non-correlations between variables and large gap(s). However, the proposed algorithm is restricted to applications with the necessary assumption of recurring data and sufficient large datasets size. The present work will allow combining DTWUMI with other algorithms (as Random Forest or Deep Learning) to efficiently complete missing data not only on uncorrelated datasets but also on any type of multivariate time series in the future.

6. REFERENCES

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

[2] Hee-Taek Ceong, Hae-Jin Kim, and Jeong-Seon 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.

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

[4] Graeme Hawthorne, Graeme Hawthorne, and Peter 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.

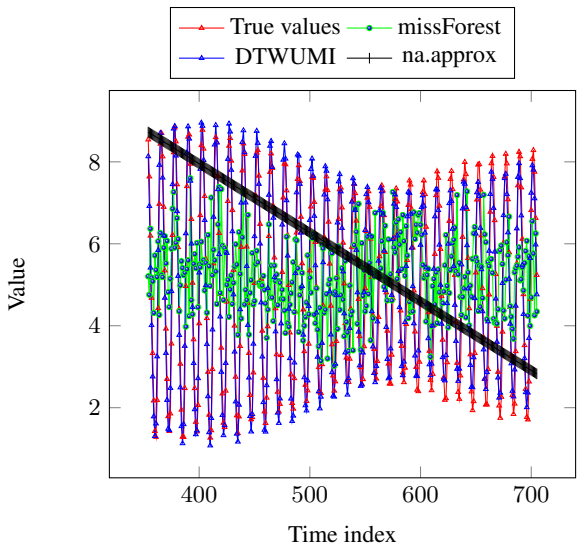

Fig. 3. Visual comparison of imputed values of different imputation methods with true values on Marel Carnot dataset with the gap size of 353 on the 2nd signal.

[5] Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, and Mikko Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, no. 18, pp. 2895–2907, June 2004.

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

[7] Stef Van Buuren, Hendriek C. Boshuizen, Dick 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] Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, and Peter 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] Elizabeth A. Stuart, Melissa Azur, Constantine Frangakis, and Philip 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.

[10] Katherine J. Lee and John 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.

[11] Anoop D. Shah, Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry 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.

[12] Serena G. Liao, Yan Lin, Dongwan D. Kang, Divay Chandra, Jessica Bon, Naftali Kaminski, Frank C. Sciurba, and George C. Tseng, "Missing value imputation in high-dimensional phenomic data: Imputable or not, and how?," *BMC Bioinformatics*, vol. 15, pp. 346, 2014.

[13] Shah Atiqur Rahman, Yuxiao Huang, Jan Claassen, Nathaniel Heintzman, and Samantha 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.

[14] Andrew Gelman, Jennifer Hill, Yu-Sung Su, Masanao Yajima, Maria Pittau, Ben Goodrich, Yajuan Si, and Jon Kropko, "Mi: Missing Data Imputation and Model Checking," Apr. 2015.

[15] Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long, "Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data," *Scientific Reports*, vol. 6, pp. 21689, Feb. 2016.

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

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

[18] Daniel J. Stekhoven and Peter Bühlmann, "MissForest—nonparametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012.

[19] Hui-Huang Hsu, Andy C. Yang, and Ming-Da Lu, "KNN-DTW Based Missing Value Imputation for Microarray Time Series Data," *Journal of Computers*, vol. 6, no. 3, Mar. 2011.

[20] Andy C. Yang, Hui-Huang Hsu, and Ming-Da Lu, "Missing Value Imputation in Microarray Gene Expression Data," 2009.

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

[22] Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 16, pp. 43–49, 1978.

[23] 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)*. July 2016, pp. 283–288, IEEE.

[24] Yu-Sung Su, Andrew Gelman, Jennifer Hill, Masanao Yajima, and others, "Multiple imputation with diagnostics (mi) in R: Opening windows into the black box," *Journal of Statistical Software*, vol. 45, no. 2, pp. 1–31, 2011.

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

[26] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2010.

[27] Sarah C. Goslee, Dean L. Urban, and others, "The ecodist package for dissimilarity-based analysis of ecological data," *Journal of Statistical Software*, vol. 22, no. 7, pp. 1–19, 2007.

Table 1. Average imputation performance indexes of various imputation algorithms on NNGC (1745 collected points), simulated (32,000 collected points) and Marel Carnot (33,344 collected points) datasets **(best results in bold)**

Gap size	Method	NNGC dataset						Simulated dataset						Marel Carnot dataset					
		Accuracy indexes			Shape indexes			Accuracy indexes			Shape indexes			Accuracy indexes			Shape indexes		
		1-Sim	1-R^2	RMSE	FSD	FB	1-FA2	1-Sim	1-R^2	RMSE	FSD	FB	1-FA2	1-Sim	1-R^2	RMSE	FSD	FB	1-FA2
1%	na.approx	0.2	0.99	11786	0.41	0.19	0.52	0.126	0.994	1.99	0.52	**1.86**	0.81	0.068	0.15	1.62	**0.07**	**0.03**	0.21
	MI	0.1	0.32	5774	0.02	0.01	0.26	0.14	0.999	2.22	0.12	1.89	0.79	0.19	0.44	4.48	0.42	0.24	0.48
	MICE	0.03	0.06	2382	0.03	0.01	0.05	0.14	0.997	2.23	0.13	2.39	0.79	0.16	0.46	4.51	0.37	0.2	0.39
	missForest	**0.02**	**0.02**	**1286**	**0.01**	**0.01**	**0.01**	0.11	0.996	1.69	0.89	5.49	0.85	0.15	0.26	3.2	0.35	0.18	0.32
	DTWUBI	0.12	0.51	7591	0.03	0.12	0.3	**0.085**	**0.51**	**1.22**	**0.01**	5.86	**0.58**	**0.056**	**0.04**	**1.02**	0.11	0.05	**0.15**
2%	na.approx	0.18	0.96	11456	0.36	0.22	0.52	0.11	0.998	1.99	0.48	2.41	0.8	0.07	0.13	1.73	**0.06**	0.12	0.18
	MI	0.1	0.33	5644	0.04	0.05	0.31	0.13	0.997	2.31	0.06	7.12	0.8	0.17	0.41	3.81	0.23	0.12	0.43
	MICE	0.04	0.11	3025	0.02	0.01	0.05	0.12	0.999	2.25	0.08	3.75	0.8	0.16	0.44	4.05	0.28	0.14	0.37
	missForest	**0.02**	**0.02**	**1210**	**0.01**	**0.01**	**0.01**	0.1	0.998	1.7	0.94	2.48	0.86	0.13	0.24	2.76	0.24	0.14	0.26
	DTWUBI	0.12	0.51	7591	0.1	0.08	0.3	**0.064**	**0.45**	**1.17**	**0.01**	**0.79**	**0.55**	**0.06**	**0.04**	**1.07**	0.1	**0.03**	**0.16**
3%	na.approx	0.18	0.99	11329	0.66	0.29	0.55	0.11	0.998	1.88	0.69	2.08	0.81	0.08	0.17	1.8	0.09	0.07	0.19
	MI	0.1	0.29	5317	0.04	0.02	0.24	0.13	1	2.27	0.03	2.63	0.8	0.21	0.49	4.53	0.41	0.33	0.47
	MICE	0.03	0.11	3112	0.02	0.02	0.05	0.13	1	2.27	0.03	2.63	0.8	0.19	0.53	5.17	0.49	0.36	0.41
	missForest	**0.02**	**0.02**	**1375**	**0.02**	**0.01**	**0.01**	0.1	1	1.71	0.91	2.49	0.85	0.18	0.37	4.09	0.39	0.37	0.36
	DTWUBI	0.05	0.19	4219	0.05	0.08	0.08	**0.064**	**0.45**	**1.16**	**0.01**	**1.72**	**0.54**	**0.056**	**0.06**	**1.07**	**0.09**	**0.02**	**0.12**
4%	na.approx	0.18	0.99	11298	0.35	0.12	0.53	0.11	0.999	2.14	0.42	**2.08**	0.79	0.057	0.09	1.68	0.06	0.07	0.22
	MI	0.11	0.43	6647	0.05	0.06	0.3	0.12	1	2.3	0.03	5.66	0.8	0.15	0.41	4.51	0.31	0.2	0.47
	MICE	0.04	0.17	3730	**0.02**	**0.01**	0.08	0.12	0.999	2.26	0.04	10.07	0.8	0.135	0.44	4.73	0.29	0.2	0.43
	missForest	**0.03**	**0.09**	**2405**	0.06	0.03	**0.05**	0.09	1	1.73	0.94	3.81	0.86	0.12	0.22	3.46	0.31	0.18	0.34
	DTWUBI	0.1	0.48	6935	0.05	0.06	0.22	**0.065**	**0.46**	**1.19**	**0.01**	4	**0.56**	**0.048**	**0.05**	**1.27**	**0.06**	**0.05**	**0.19**
5%	na.approx	0.17	0.99	10848	0.73	0.26	0.54	0.12	1	2.12	0.66	**2.09**	0.79	0.064	0.11	1.81	**0.06**	**0.06**	0.21
	MI	0.11	0.43	6823	**0.02**	0.06	0.29	0.12	1	2.27	0.04	3.67	0.79	0.15	0.41	4.36	0.21	0.21	0.44
	MICE	0.04	0.14	3483	0.03	**0.02**	0.06	0.12	1	2.27	0.04	3.27	0.79	0.13	0.4	4.42	0.27	0.23	0.41
	missForest	**0.03**	**0.09**	**2710**	0.06	0.03	**0.04**	0.1	1	1.75	0.94	1.92	0.85	0.12	0.23	3.52	0.28	0.23	0.28
	DTWUBI	0.1	0.49	7116	0.05	0.04	0.22	**0.07**	**0.46**	**1.19**	**0.01**	2.55	**0.58**	**0.054**	**0.08**	**1.59**	0.12	0.09	**0.13**
7.5%	na.approx	0.19	0.99	11803	0.49	0.19	0.57	0.11	1	1.86	0.84	**2.09**	0.82	0.07	0.3	3.2	0.19	0.16	0.24
	MI	0.11	0.39	6408	**0.013**	0.05	0.28	0.12	0.999	2.24	0.03	7.95	0.79	0.14	0.54	4.76	0.26	0.17	0.48
	MICE	0.04	0.13	3375	0.02	**0.013**	0.05	0.12	1	2.23	0.02	5.54	0.79	0.13	0.6	5.06	0.28	0.21	0.43
	missForest	**0.03**	**0.07**	**2197**	0.05	0.02	**0.03**	0.1	1	1.69	0.9	2.7	0.86	0.1	0.41	3.35	0.28	0.14	0.33
	DTWUBI	0.04	0.14	3452	0.03	0.04	0.06	**0.078**	**0.58**	**1.37**	**0.01**	5.57	**0.6**	**0.061**	**0.25**	**2.11**	**0.12**	**0.08**	**0.18**
10%	na.approx	0.18	1	11419	0.62	0.25	0.56	0.11	1	2.01	0.46	**2.02**	0.79	0.083	0.23	3.09	0.15	0.16	0.27
	MI	0.1	0.35	5892	**0.008**	0.02	0.27	0.12	1	2.24	0.02	2.18	0.79	0.13	0.43	4.35	0.16	0.14	0.46
	MICE	0.04	0.13	3435	0.01	0.01	0.06	0.12	1	2.25	0.02	16.56	0.79	0.12	0.5	4.78	0.21	0.18	0.41
	missForest	**0.02**	**0.05**	**1990**	0.02	**0.004**	**0.03**	0.09	1	1.7	0.91	1.35	0.86	0.1	0.29	3.47	0.25	0.15	0.3
	DTWUBI	0.05	0.21	4402	0.02	0.04	0.08	**0.064**	**0.47**	**1.18**	**0**	4.49	**0.56**	**0.065**	**0.2**	**2.58**	**0.12**	**0.13**	**0.2**