Reconstruction of Missing Data in Synthetic Time Series Using EMD

Tatjana Sidekerskiene
Department of Applied Mathematics
Kaunas University of Technology
Kaunas, Lithuania
Email: tatjana.sidekerskiene@ktu.lt

Robertas Damasevicius
Department of Software Engineering
Kaunas University of Technology
Kaunas, Lithuania
Email: robertas.damasevicius@ktu.lt

Abstract—**The paper presents a novel method for reconstruction of missing data in time series. The method is based on the decomposition of known parts of time series into mono-components (Intrinsic Mode Functions, IMF) using Empirical Mode Decomposition (EMD), construction of prediction models for each IMF using known parts of times series and their composition using weighted average. We demonstrate the efficiency of the proposed approach using a synthetic time series data.**

I. Introduction

Knowledge-based decision-making processes (such as business decisions) are very dependent on the availability of data, from which information can be extracted. These processes often use predictive models or other computational intelligence technique that take observed data as inputs. However, in some cases due to various reasons (power failure, sensor failure, maintenance, human error, etc.) data could be lost, corrupted or recorded incompletely, which affects the quality of data negatively. Most decision-making and machine learning tools such as the commonly used Artificial Neural Networks (ANN), Support Vector Machines (SVM), Principal Component Analysis (PCA) and many other computational intelligence techniques cannot be used for decision making if data are not complete [1]. Missing values in the data can severely affect the interpretation and hinder downstream analysis such as supervised or unsupervised classification or clustering. Since the decision output should still be maintained despite the missing data, we have to deal with the problem of missing data. Therefore, in case of incomplete or missing data, the first step in data processing is to estimate the missing values.

The task, also known as missing value imputation [2] or gap filling, is an important one in cases where it is crucial to use all available data and not discard records with missing values. It is especially relevant in many real-world time series such as obtained by remote sensing observations made without direct physical contact with the observed object [13], in which raw data can be corrupted, obstructed and hindered by multiple, often unforeseen, ways. Such time series may contain multiple gaps, which must be dealt with before applying other signal processing techniques such as spectral analysis. Consequently, the quality and the completeness of these time series are essential. Previously researchers [4] have demonstrated that missing values imputation can significantly improve the overall prediction or data analysis when information about the data is incorporated into the imputation. Thus the problem of gap-filling (or data reconstruction) is a fundamental one in computational intelligence.

The applications of gap filling (aka missing data imputation) are numerous and include bioinformatics (gene microarray data) [5] and remote sensing observations [13]. The existing gap-filling techniques depend upon the size of gap series and the nature of data. If gaps are small, the solutions are simple and well researched. For example, mean imputation [4] substitutes every missing value with the mean of the observations. Because of its simplicity, mean imputation is commonly used in the social sciences. Regression imputation replaces missing values with predictions made by a regression (curve fitting, interpolation) model of the missing value on variables observed for the data vector. If there is abundance of complete data, hot-deck imputation can be used to substitute missing values according to data vectors with similar values. However, if gaps expand over several cycles of data in time series, specific methods should be developed.

The specific gap filling methods can be categorized according to the type of information used as global and local. Using global approach, the algorithms perform missing values estimation based on global correlation information obtained from the entire data matrix. An example of such algorithms includes the SVD imputation [6]. Peng and Zhu [7] used Independent Component Analysis (ICA) and Self Organizing Maps (SOM) to handle missing values. Gaussian mixture models are used in [8]. Ssali and Marwala [9] combined decision trees, Principal Component Analysis Neural Network (PCA-NN) and Auto Associative Neural Network (AANN) to estimate the missing values. The power of evolutionary computing in imputation process is emphasized in [10] by combining ANN with either GA or Particle Swarm Optimization (PSO). In [11], [12] it was used a composite neuro-wavelet reconstruction system composed by two neural networks separately trained to obtain the reconstruction. For local approach, the algorithms exploit only local similarity structure in the data sets to perform missing values imputation.

The signal decomposition methods split the signal into additive components. The time series are then reconstructed using only the components of interest, usually by removing the high frequency components considered as noise. Because the decomposition is performed over a limited temporal window, only a limited amount of information is used when filling gaps.

The Iterative Caterpillar Singular Spectrum Analysis Method (ICSSA) [13] is a modification of the CSSA [14] method developed to describe time series and fill missing data by decomposing the time series into empirical orthogonal functions (EOF). This method allows filling gaps and forecasting data at the extremities of the time series.

Empirical mode decomposition (EMD) method [15] consists in decomposing the time series into a small number of intrinsic mode functions (IMFs) derived directly from the time series itself using an adaptive iterative process based on local characteristics of data in the time domain. The first IMF, mostly affected by noise, can be discarded or thresholded to remove the high frequency fluctuations [16]. Note that the EMD method requires the time series to be continuous. EMD has been used before in for gap-filling in [17]. In this method, gap filling is based on adding artificial local extrema in the missing part of the data based on the assumptions that behavior of the missing data is similar to the neighbourhood of the gap, and for each IMF, every two consecutive extrema in the gap are equally distant. The modified EMD produces IMFs with gaps, and the gap data is reconstructed by summing up the IMFs. However, the assumption that the behavior of the data does not change in the gap may not hold true. Furthermore, the method involves the selection of extrema points manually based on the appearance of each gap-filled IMF, which is time-consuming. Kandasamy *et al.* [13] filled the missing data by linear interpolation and then applied signal decomposition by EMD. However, as linear interpolation provides generally poor performances in case of long periods without observations, the EMD-based method fails when there is a significant fraction of gaps (more than 20 %).

The aim of this paper is to propose and evaluate a novel method for reconstruction of missing data in time series. The method is based on the decomposition of known parts time series into mono-components (IMFs) using EMD, construction of prediction models for each IMF using known parts of times series and their composition using weighted average.

The structure of the remaining parts of the paper is as follows. The proposed method is described in Section II. The metrics for evaluation the accuracy of the solution are discussed in Section III. The experiment is described in Section IV. The results of the experiment are presented in Section V. Finally, conclusions and discussion of future work are presented in Section VI.

II. Description of method

A. Task

Consider a time series $X \in R^T$ with $m \leq T-2$ data points missing. The missing data are indicated by a vector $\hat{X} \in R^m$ with components $x_t = 1$ for all $t \in \{1, 2, \dots, T\}$ where x_t is defined, and $x_t = -1$ for all $t \in \{1, 2, \dots, T\}$ where x_t is not defined. Let the zero crossing rate of X be

$$zcr(X) = \sum_{t=1}^{T} I\{x_t x_{t-1} < 0\} \qquad (1)$$

here $I\{A\}$ is the Iverson operator: $I\{A\}$ is 1 if its argument A is true, and 0 otherwise.

Hereinafter we introduce the simplification and consider only such X, for which $zcr(X) = 2$ and $x_0 = 1$. That is the time series starts with known data, then follows all missing data, and the time series ends with known data. Hereinafter, we denote the known data at the beginning of the time series X before the missing data as the left series X_l, and the known data at the end of the time series X after the missing data as the right series X_r, and the missing series as X_m.

B. Outline

The proposed method consists of the following steps:

1) Perform decomposition of X_l into a set of IMFs IMF_l, and decompose X_r into a set of IMFs IMF_r using EMD.
2) Perform prediction of IMF_l into the future by m steps using the polynomial prediction model. Let predicted set of series is IMF'.
3) Perform prediction of IMF_r into the past by m steps using the polynomial prediction model. Let the predicted set of series be IMF''.
4) Combine the predicted series IMF' and IMF'' using the linearly weighted average. Let the result be IMF.
5) Sum IMF to derive the prediction of X_m. as Y.

C. EMD for decomposition

The steps comprising EMD method [15] are as follows:

1) Identify local maxima and minima of signal $x(t)$, where t is time.
2) Perform cubic spline interpolation between the maxima and minima to obtain envelopes $E_{max}(t)$ and $E_{min}(t)$.
3) Calculate the mean of the envelopes as $M(t) = (E_{max}(t) + E_{min}(t))/2$.
4) Calculate the difference between a signal and the mean of its envelopes as $C_1(t) = x(t) - M(t)$.
5) IF the number of local extrema of $C_1(t)$, is equal to or differs from the number of zero crossings by one, and the average of $C_1(t)$ is close to zero,
THEN $IMF_1 = C_1(t)$;
ELSE repeat steps 1-4 on $C_1(t)$ instead of $x(t)$, until new $C_1(t)$ satisfies the conditions of an IMF in Step 5.
6) Calculate residue $R_1(t) = x(t) - C_1(t)$.
7) If residue $R_1(t)$ is above a threshold value, then repeat steps 1-6 on $R_1(t)$ to obtain next IMF and a new residue.

As a result, n orthogonal IMFs are obtained from which the original signal may be reconstructed as follows

$$x(t) = \sum_i IMF_i(t) + R(t). \tag{2}$$

D. Prediction using polynomial model

Prediction is performed using a polynomial model that finds coefficients for a polynomial $p(x)$ of degree n that is a best fit (in a least-squares sense) for the data in y using M previous values of x. Here a linear case of the model is shown:

$$p(x_i) = a_0 x_{i-1} + a_1 x_{i-2} + \ldots + a_{M-1} x_{i-M} + a_M. \tag{3}$$

E. Weighted averaging

The predictions are averaged using the weighted average as follows:

$$X = X' w(t) + X''(1 - w(t)), \tag{4}$$

where $w(t)$ is a linear monotonous increasing function in the range $[0;1]$ from 0 to $m-1$.

III. Performance evaluation using statistical measures

Evaluation of the data gap filling results is a crucial step to demonstrate reliability and accuracy of the proposed method. In validation, the performance indices are computed between the reconstructed and known original values. The quality of the reconstructed data also can be evaluated independently of the original data using the smoothness criterion.

We evaluate the accuracy of the results using the following metrics:

Root Mean Square Error (RMSE) of a model prediction with respect to the estimated variable y is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}, \tag{5}$$

where $x_1, x_2, \ldots, x_n$ are n observed values and $y_1, y_2, \ldots, y_n$ are the corresponding values predicted.

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of forecasts, without considering their direction. It is used to measure how close forecasts or predictions are to the eventual outcomes:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|. \tag{6}$$

Mean Square Error (MSE). If y_i is a vector of n predictions, and x_i is the vector of observed values corresponding to the inputs to the function which generated the predictions, then the MSE of the predictor can be estimated by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2. \tag{7}$$

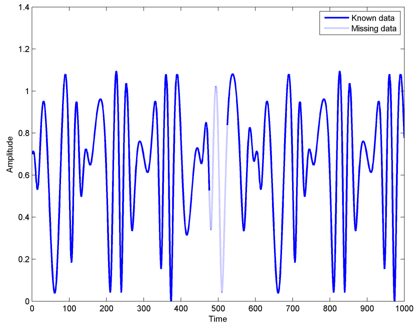

Fig. 1. A fragment of example synthetic time series.

Pearson correlation coefficient R^2 is a measure of the strength and direction of the linear relationship between two variables that is defined as:

$$R^2 = \frac{\left\{\sum(x_i - \bar{x})(y_i - \bar{y})\right\}^2}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}. \tag{8}$$

IV. Experiment

A. Data and motivating example

We use the following synthetic time series (also analyzed in [18]):

$$x(t) = \sin\left(\frac{2\pi}{300}t\right) \cdot \cos\left(\frac{2\pi}{40}t + \frac{\pi}{2}\sin\frac{2\pi}{120}t\right) \tag{9}$$

The known data to the left and right of the gap was decomposed using EMD and the IMFs were derived. The polynomial model based prediction method described in Section II.D was used to derive the forward prediction from the IMF data to left side of the gap and the backward prediction from the IMF data to the right side of the gap.

Fig. 1 shows a fragment of the original time series with missing data (gap is in the middle of the time series, gap length = 50). Fig. 2 shows the IMFs derived for known continuous fragments of data before and after the gap. Fig. 3 shows the adjusted IMFs with reconstructed IMF values for the missing data using the combination of forward and backward prediction using linearly weighted averaging. Finally, Fig. 4 shows the true data and reconstructed data (only the gap of time series is shown).

For comparison, the results of reconstruction without signal decomposition are also shown. Even by visual inspection of figures we can see that the proposed method performs better.

B. Methodology

To obtain a statistically valid evaluation of the efficiency of the method, we, first, have generated a synthetic series with a large length $(N = 10000)$. Then we have extracted $M = 20$ different subseries of selected length $(500, 1000, 2000)$ at random locations of the original large time series. The gap in the data is always located in the middle of series and has the length of $K(K = 10, 20, \ldots, 100)$. Gap filling is performed using the proposed method and accuracy is calculated. Results of computational experiments are presented in Section V.

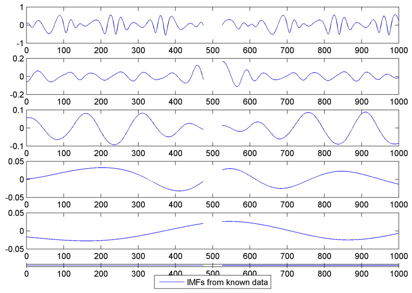

Fig. 2. IMFs from known data of synthetic time series.

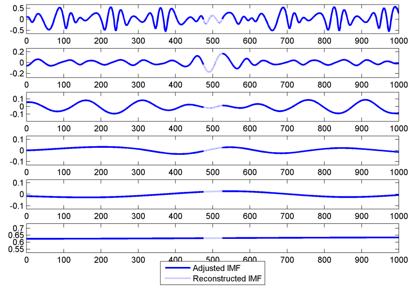

Fig. 3. Gap filling on the IMF level.

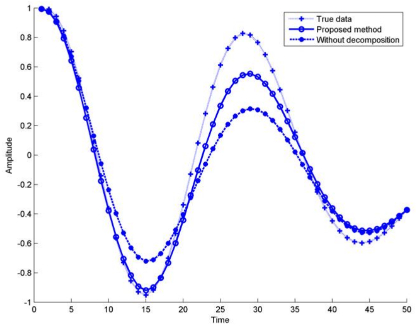

Fig. 4. Reconstructed missing data in the gap.

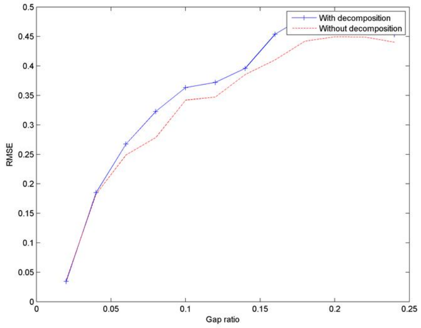

Fig. 5. Average RMSE with respect to the length of gap.

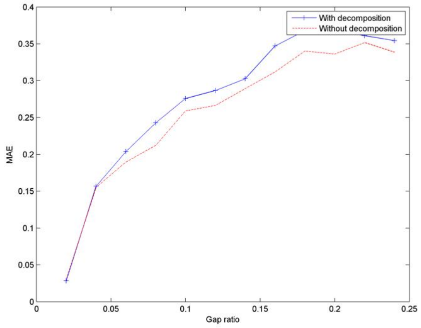

Fig. 6. Average MAE with respect to the length of gap.

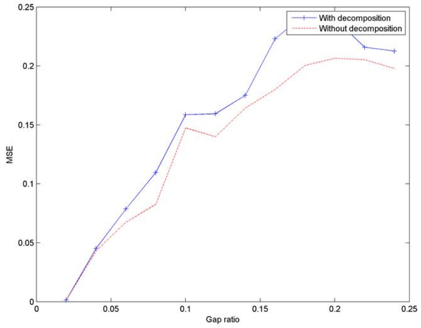

Fig. 7. Average MSE with respect to the length of gap.

V. Result

In order to evaluate the performance of the gap-filling, we have repeated the gap filling experiment 20 times with different subseries at random locations of the original time series. The numerical simulation was performed using MATLAB 8.1 (R2013a). The experimental results are presented in Figs. 5-8 for the values of RMSE, MAE, MSE and Pearson correlation values vs the gap ration (length of gap divided by the length of subseries under consideration).

We can see that the EMD decomposition-based method fails to improve the accuracy of missing data reconstruction for large gaps. Yet, when the size of gap is small (10 or less), the method allows to achieve some improvement as compared with the direct prediction of missing data from the known time series data. The results are summarized in Tables I-II.

VI. Conclusion

The paper has presented a novel method for the reconstruction of missing data in time series. The proposed method is model-free, fully data-driven, and does not impose any technical assumptions. The results of experiments with the synthetic data show that better results have not been achieved for gap filling task using the EMD decomposition if gap length is large (more than 10 missing data points). The reasons for that may be the inherent deficiencies of the EMD method

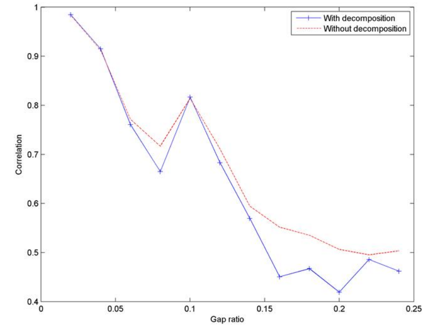

Fig. 8. Average correlation with respect to the length of gap.

TABLE I
RESULTS WITHOUT AND WITH DECOMPOSITION FOR SUBSERIES OF LENGTH 500

Gap length	Decomposition	RMSE	MAE	MSE	R^2
10	yes	0.0316	0.0262	0.0014	0.9890
	no	0.0304	0.0257	0.0014	0.9887
20	yes	0.0912	0.0764	0.0136	0.9499
	no	0.0964	0.0813	0.0148	0.9440
30	yes	0.2393	0.1823	0.0815	0.9301
	no	0.2034	0.1543	0.0561	0.9418
40	yes	0.4062	0.3085	0.1810	0.7010
	no	0.3632	0.2775	0.1514	0.7449

TABLE II
RESULTS WITHOUT AND WITH DECOMPOSITION FOR SUBSERIES OF LENGTH 1000

Gap length	Decomposition	RMSE	MAE	MSE	R^2
10	yes	0.0449	0.0382	0.0026	0.9870
	no	0.0445	0.0380	0.0025	0.9855
20	yes	0.0870	0.0714	0.0120	0.9808
	no	0.0732	0.0609	0.0093	0.9829
30	yes	0.2760	0.2166	0.1004	0.7968
	no	0.2684	0.2123	0.0974	0.7913
40	yes	0.2893	0.2220	0.0979	0.8109
	no	0.2902	0.2247	0.0985	0.7828

TABLE III
RESULTS WITHOUT AND WITH DECOMPOSITION FOR SUBSERIES OF LENGTH 2000

Gap length	Decomposition	RMSE	MAE	MSE	R^2
10	yes	0.0415	0.0354	0.0025	0.9719
	no	0.0391	0.0333	0.0023	0.9800
20	yes	0.1095	0.0920	0.0242	0.8538
	no	0.1157	0.0973	0.0250	0.8499
30	yes	0.1813	0.1409	0.0482	0.8952
	no	0.1809	0.1402	0.0466	0.9023
40	yes	0.2973	0.2278	0.1143	0.7743
	no	0.2615	0.2004	0.0922	0.8273

also noted by other researchers such as the boundary problem and the mode mixing problem, which can significantly affect the accuracy of analysis results. The correct prediction of locations of minima and maxima in data gap, could allow to better estimate the envelope of data. The main reasoning for using EMD is that IMFs would be less complex than the original data. However, the higher frequency IMFs do not always have a simple waveform, which would allow more accurate prediction of their values. Thus no gain by predicting a simpler could be achieved. Additionally, the method requires more computation, which also negatively affects the accuracy.

Future work will focus on performing more extensive experiments with synthetic time series and the application of the proposed method to real world time series and the examination of other signal decomposition and prediction methods to overcome the shortcoming of EMD and to improve the proposed method.

REFERENCES

[1] A. W. Liew and N. F. Law, H. Yan, *Missing value imputation for gene expression data: computational techniques to recover missing data from available information*, Brief Bioinform, no. 12, pp. 498-513, 2011.

[2] D. B. Rubin, *Multiple Imputation After 18+ Years*, Journal of the American Statistical Association, vol. 91, no. 434, pp. 473-489, 1996.

[3] S. Kandasamy, F. Baret, A. Verger, P. Neveux, and M. Weiss, *A comparison of methods for smoothing and gap filling time series of remote sensing observations application to MODIS LAI products*, Biogeosciences, vol. 10, pp. 4055-4071, 2013.

[4] P. D. Allison, *Missing data*, Thousand Oaks, CA: Sage Publications, Inc., 2001.

[5] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Bostein, R.B. Alman, *Missing value estimation methods for DNA microarrays*, Bioinformatics, vol. 17, pp. 520-525, 2001.

[6] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, *Imputing Missing Data for Gene Expression Arrays*, Stanford University Statistics Department Technical report, 1999.

[7] H. Peng and S. Zhu, *Handling of incomplete data sets using ICA and SOM in data mining*, Neural Computing & Applications, vol. 16, no. 2, pp. 167-172, 2007.

[8] D. Yoon, E. K. Lee, T. Park, *Robust imputation method for missing values in microarray data*, BMC Bioinformatics, vol. 8, Suppl 2, S6, 2007.

[9] G. Ssali and T. Marwala, *Computational Intelligence and Decision Trees for Missing Data Estimation*, in International Joint Conference on Neural Networks (IJCNN 2008), Hong Kong, 1-8 Jun 2008.

[10] S. M. Dhalmini, F. V. Nelwamondo, T. Marwala, *Condition Monitoring of HV Bushings in the Presence of Missing Data Using Evolutionary Computing*, WSEAS Transactions on Power Systems, vol. 1, no. 2, pp. 296-302, 2006.

[11] G. Capizzi, C. Napoli, L. Paterno, *An Innovative Hybrid Neuro-wavelet Method for Reconstruction of Missing Data in Astronomical Photometric Surveys*, Proc. of International Conference on Artificial Intelligence and Soft Computing ICAISC (1), pp. 21-29, 2012.

[12] C. Napoli and E. Tramontana, *Massively parallel WRNN reconstructors for spectrum recovery in astronomical photometrical surveys*, Neural Networks, vol. 83, pp. 4250, 2016.

[13] S. Kandasamy, P. Neveux, A. Verger, S. Buis, M. Weiss, and F. Baret, *Improving the consistency and continuity of MODIS 8 day leaf area index products*, International Journal of Electronics and Telecommunications, vol. 58, pp. 141-146, 2012.

[14] N. Golyandina and E. Osipov, *The Caterpillar-SSA method for analysis of time series with missing values*, Journal of Statistical Planning and Inference, vol. 137, pp. 2642-2653, 2007.

[15] N. E. Huang, Z. Shen, S. R. Long, *A new view of nonlinear water waves: the Hilbert spectrum*, Annual Review of Fluid Mechanics, vol. 31, pp. 417-457, 1999.

[16] R. Damasevicius, M. Vasiljevas, I. Martisius, V. Jusas, D. Birvinskas, M. Wozniak, *BoostEMD: An Extension of EMD Method and Its Application for Denoising of EMG Signals*, Electronics and Electrical Engineering, vol. 21, no. 6, pp. 57-61, 2015.

[17] A. Moghtaderi, P. Borgnat, P. Flandrin, *Gap-filling by the empirical mode decomposition*, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3821-3824, 2012.

[18] D. Kondrashov and M. Ghil, *Spatio-temporal filling of missing points in geophysical data sets* Nonlinear Processes in Geophysics, vol. 13, pp. 151-159, 2006.