GAP-FILLING BY THE EMPIRICAL MODE DECOMPOSITION

Azadeh Moghtaderi⋆*, Pierre Borgnat*† *and Patrick Flandrin*†

⋆Queen's University, Department of Mathematics and Statistics
Kingston, Ontario, Canada, K7L 3N6
†CNRS, École Normale Supérieure de Lyon, Laboratoire de Physique (UMR 5672)
46 allée d'Italie 69364 Lyon Cedex 07, France

ABSTRACT

We propose a novel gap-filling technique, based on the empirical mode decomposition (EMD). The idea is that a signal with missing data can be decomposed into a set of intrinsic mode functions (IMFs) with missing data. Filling the gaps in each IMF should be easier than filling the gaps in the original signal. This is because each IMF varies much more slowly than the original signal, and also because the IMFs are known to have useful regularity properties. We demonstrate the performance of our technique on environmental pollutant data.

***Index Terms*—** Signal reconstruction, signal restoration, interpolation, signal processing algorithms

1. INTRODUCTION

Many real-world signals often contain multiple gaps which reflect missing data. On the other hand, most of the standard signal processing tools are only applicable to signals without gaps; for instance, this is true of spectral analysis. Thus the problem of *gap-filling* (or "interpolating") is fundamental. There is a sizable literature on this problem; a survey can be found in [1]. The existing gap-filling techniques make technical assumptions which limit their applicability. For instance, the Papoulis–Gerchberg algorithm [2] assumes that the signal is stationary, band-limited and the bandwidth is known. The goal of this paper is to develop a gap-filling technique which makes minimal technical assumptions. The technique we propose is called *empirical mode decomposition gap-filling*. As its name suggests, our technique is fundamentally based on the empirical mode decomposition, or EMD. Recall [3] that the EMD decomposes a signal into a finite superposition of oscillatory modes. The method is model-free, fully data-driven, and more importantly does not impose any technical assumptions. In particular, this means that the EMD is well-suited to the analysis of non-stationary and non-linear signals.

The paper is organized as follows. After briefly reviewing the EMD in §2, we describe EMD gap-filling in §3. The performance of EMD gap-filling on a real-world data is demonstrated in §4.

2. THE EMPIRICAL MODE DECOMPOSITION

The *empirical mode decomposition* (EMD) is an algorithm which decomposes a real-valued signal into a finite additive superposition of oscillatory components [3]. Each oscillatory component is called an *intrinsic mode function* (IMF). The EMD does not rely on any technical assumptions concerning the nature of the signal. The IMFs are computed subject to two requirements: First, the number of local extrema and number of zero crossings of each IMF should vary by at most one. Second, the mean of the upper and lower envelopes of each IMF should be approximately equal to zero. The IMFs are computed by means of the so-called "sifting process." The sifting process stops when the two requirements are satisfied within a prescribed tolerance [3].

Consider a discrete-time signal $\boldsymbol{\mathcal{X}} = (X_1, X_2, \ldots, X_N)$. The EMD computes the IMFs of $\boldsymbol{\mathcal{X}}$ using the following algorithm. As an initialization step, set $i = 1$ and $\boldsymbol{\rho}^0 = \boldsymbol{\mathcal{X}}$.

1. SIFTING PROCESS

(a) Identify the local maxima and local minima of $\boldsymbol{\rho}^{i-1}$. For example, X_t is a local maximum if $X_{t-1} \leq X_t$ and $X_{t+1} \leq X_t$.

(b) Use the local maxima and local minima obtained in (1a) to compute the upper and lower envelopes of $\boldsymbol{\rho}^{i-1}$. This requires an interpolation method, for which we use cubic splines.

(c) Set $\boldsymbol{Q}^i$ to be the mean of the upper and lower envelopes of $\boldsymbol{\rho}^{i-1}$, and set $\boldsymbol{h} = \boldsymbol{\mathcal{X}} - \boldsymbol{Q}^i$.

(d) If $\boldsymbol{h}$ is not an IMF, in the sense that it does not satisfy the two requirements described above within the prescribed tolerance, then go to (1a) with $\boldsymbol{\rho}^{i-1} = \boldsymbol{h}$. If $\boldsymbol{h}$ is an IMF, then go to (2a).

2. IMF EXTRACTION

(a) The *ith intrinsic mode function* of $\boldsymbol{\mathcal{X}}$ is $\boldsymbol{\mathcal{A}}^i = \boldsymbol{h}$, and the *$i$th residual* is $\boldsymbol{\rho}^i = \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}^i$. Increment i and go to (1a).

The algorithm halts when the local maxima and local minima of the ith residual cannot be identified. It is known that this algorithm halts in a reasonably small number of steps. We denote by $\mathcal{I}$ the largest index for which $\boldsymbol{\mathcal{A}}^i$ is defined. Then the EMD of $\boldsymbol{\mathcal{X}}$ is

$$\boldsymbol{\mathcal{X}} = \sum_{i=1}^{\mathcal{I}} \boldsymbol{\mathcal{A}}^i + \boldsymbol{\rho}^{\mathcal{I}}.$$

In this decomposition, $\boldsymbol{\mathcal{A}}^1$ through $\boldsymbol{\mathcal{A}}^{\mathcal{I}}$ can be thought of as containing a "spectrum" of local oscillations in $\boldsymbol{\mathcal{X}}$, with the highest-frequency oscillations represented in $\boldsymbol{\mathcal{A}}^1$ and the lowest-frequency oscillations represented in $\boldsymbol{\mathcal{A}}^{\mathcal{I}}$.

3. EMD GAP-FILLING

In this section, $\boldsymbol{\mathcal{X}}$ is given as in §2. Consider

$$\widetilde{\boldsymbol{\mathcal{X}}} = (X_1, \ldots, X_{q-1}, X_{q+Q}, \ldots, X_N),$$

which we regard as $\boldsymbol{\mathcal{X}}$ with a single gap. We assume $q \gg 1$ and $q + Q - 1 \ll N$, so that there is information available on both sides of the gap. Our problem is to fill the gap in $\widetilde{\boldsymbol{\mathcal{X}}}$; that is, to estimate $(X_q, X_{q+1}, \ldots X_{q+Q-1})$.

3.1. Modified EMD

In this section we describe how to modify the EMD to extract IMFs from $\widetilde{\boldsymbol{\mathcal{X}}}$. This is needed for the following reasons: First, the fact that X_q and X_{q+Q-1} are missing means that X_{q-1} and X_{q+Q} cannot be deemed as local extrema. Consequently X_{q-1} and X_{q+Q} go unused. Second, the interpolation step of the sifting process suffers from overshoot near the gap. It turns out that the overshoot can be mitigated by adding "artificial local extrema" based on the information contained in X_{q-1} and X_{q+Q}. The modified EMD computes the IMFs of $\widetilde{\boldsymbol{\mathcal{X}}}$ using the following algorithm. As an initialization step, set $i = 1$, $\boldsymbol{\rho}^0 = \widetilde{\boldsymbol{\mathcal{X}}}$, and $k = 1$.

1. SIFTING PROCESS

(a) Identify the local maxima and local minima of $\boldsymbol{\rho}^{i-1}$. Let A and B be the total numbers of local minima and local maxima, respectively. Let A^ℓ and B^ℓ be the numbers of local minima and maxima, respectively, on the left side of the gap. Denote the local minima on the left and right side of the gap and local maxima on the left and right side of the gap by respectively

$$\begin{aligned} \mathcal{M}_{\min}^{\ell} &= (X_{t_1}, X_{t_2}, \ldots, X_{t_{A^\ell}}), \\ \mathcal{M}_{\min}^{r} &= (X_{t_{A^\ell+1}}, X_{t_{A^\ell+2}}, \ldots, X_{t_{A^\ell+C}}), \\ \mathcal{M}_{\max}^{\ell} &= (X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_{B^\ell}}), \quad \text{and} \\ \mathcal{M}_{\max}^{r} &= (X_{\tau_{B^\ell+1}}, X_{\tau_{B^\ell+2}}, \ldots, X_{\tau_{B^\ell+D}}). \end{aligned}$$

Let m_1 and M_1 be $\min(X_1, X_{t_1})$ and $\max(X_1, X_{\tau_1})$ at time index 1, m_{q-1} and M_{q-1} be $\min(X_{q-1}, X_{t_{A^\ell}})$ and $\max(X_{q-1}, X_{\tau_{B^\ell}})$ at time index $q-1$, m_{q+Q} and M_{q+Q} be $\min(X_{q+Q}, X_{t_{A^\ell+1}})$ and $\max(X_{q+Q}, X_{\tau_{B^\ell+1}})$ at time index $q + Q$, and finally m_N and M_N be $\min(X_N, X_{t_{A^\ell+C}})$ and $\max(X_N, X_{\tau_{B^\ell+D}})$ at time index N. We now redefine

$$\begin{aligned} \mathcal{M}_{\min}^{\ell} &= (m_1, \mathcal{M}_{\min}^{\ell}, m_{q-1}), \\ \mathcal{M}_{\min}^{r} &= (m_{q+Q}, \mathcal{M}_{\min}^{r}, m_N), \\ \mathcal{M}_{\max}^{\ell} &= (M_1, \mathcal{M}_{\max}^{\ell}, M_{q-1}), \\ \mathcal{M}_{\max}^{r} &= (M_{q+Q}, \mathcal{M}_{\max}^{r}, M_N). \end{aligned}$$

(b) Use the new local extrema obtained above together with their associated time indices to compute the upper and lower envelopes of $\boldsymbol{\rho}^{i-1}$.

(c) Set $\boldsymbol{Q}^i$ to be the mean of the upper and lower envelopes of $\boldsymbol{\rho}^{i-1}$ at time indices $1 \le t \le q-1$ and $q+Q \le t \le N$, and set $\boldsymbol{h} = \widetilde{\boldsymbol{\mathcal{X}}} - \boldsymbol{Q}^i$.

(d) Due to the gap in $\boldsymbol{h}$, we cannot determine whether or not $\boldsymbol{h}$ is an IMF. We therefore make the convention that if $k \le 100/i$, then we increment k and go to (1a) with $\boldsymbol{\rho}^{i-1} = \boldsymbol{h}$. If $k > 100/i$, then set $k = 1$ and go to (2a).

2. IMF EXTRACTION

(a) The ith IMF of $\widetilde{\boldsymbol{\mathcal{X}}}$ is $\boldsymbol{\mathcal{A}}^i = \boldsymbol{h}$, and the ith residual is $\boldsymbol{\rho}^i = \widetilde{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{A}}^i$. Increment i and go to (1a).

In the modified EMD, there are two stopping criteria: Do not allow i to exceed 10; For $i < 10$ we only deem $\boldsymbol{h}$ to be an IMF if it has at least one local minimum and one local maximum on each side of the gap. Otherwise, stop the algorithm.

3.2. Identifying local extrema and their indices in the gap

The modified EMD produces IMFs with gaps. In this section we describe how to estimate the total number of local extrema appearing in the gap of each IMF. We then describe how to estimate the indices of the local extrema.

The indices of the local minimum and local maximum nearest to the left side (resp. right side) of the gap are t_{A^ℓ} and τ_{B^ℓ} (resp. $t_{A^\ell+1}$ and $\tau_{B^\ell+1}$). Assuming that every two consecutive minimum and maximum are connected by a line, we use the equation of this line to define the "index of the zero crossing" as the a real number at which the line crosses zero. We denote the indices of the zero crossing nearest to the left and right side of the gap by κ_{z^ℓ} and κ_{z^r} respectively. Let I be an empty vector. Use the following algorithm:

- Let $I_1 = \max(t_{A^\ell}, \tau_{B^\ell}, \kappa_{z^\ell})$ and $I = [I, I_1]$;
- If $I_1 = \kappa_{z^\ell}$ then $I_2 = \max(t_{A^\ell}, \tau_{B^\ell})$ and $I = [I, I_2]$;
- Let $I_3 = \min(t_{A^\ell+1}, \tau_{B^\ell+1}, \kappa_{z^r})$ and $I = [I, I_3]$;
- If $I_3 = \kappa_{z^r}$ then $I_4 = \min(t_{A^\ell+1}, \tau_{B^\ell+1})$ and $I = [I, I_4]$.

Depending on whether the closest indices to the left and right side of the gap are associated with minima, maxima or zero crossings, we can come up with 16 different vector I. These are reported in the first column of Table 1.

Recalling the properties of IMFs such as being zero-mean and mutually orthogonal, we immediately know that every

I	A'	B'	T'
$[t_{A^\ell}, t_{A^{\ell+1}}]$, $[t_{A^\ell}, \kappa_{z^r}, t_{A^{\ell+1}}]$, $[t_{A^\ell}, \kappa_{z^\ell}, \kappa_{z^r}, t_{A^{\ell+1}}]$, $[t_{A^\ell}, \kappa_{z^\ell}, t_{A^{\ell+1}}]$	h	h+1	2h+1
$[t_{A^\ell}, \tau_{B^{\ell+1}}]$, $[t_{A^\ell}, \kappa_{z^r}, \tau_{B^{\ell+1}}]$, $[\tau_{B^\ell}, \kappa_{z^\ell}, \tau_{B^{\ell+1}}]$, $[t_{A^\ell}, \kappa_{z^\ell}, \tau_{B^{\ell+1}}]$, $[\tau_{B^\ell}, \kappa_{z^r}, t_{A^{\ell+1}}]$, $[\tau_{B^\ell}, t_{A^{\ell+1}}]$	h	h	2h
$[\tau_{B^\ell}, \tau_{B^{\ell+1}}]$, $[\tau_{B^\ell}, \kappa_{z^\ell}, \kappa_{z^r}, \tau_{B^{\ell+1}}]$, $[\tau_{B^\ell}, \kappa_{z^r}, \tau_{B^{\ell+1}}]$, $[\tau_{B^\ell}, \kappa_{z^\ell}, \tau_{B^{\ell+1}}]$	h+1	h	2h+1
$[\tau_{B^\ell}, \kappa_{z^\ell}, \kappa_{z^r}, t_{A^{\ell+1}}]$, $[t_{A^\ell}, \kappa_{z^\ell}, \kappa_{z^r}, \tau_{B^{\ell+1}}]$	h+1	h+1	2h+2

Table 1. 16 possibilities for I together with the expected number of minima denoted A' (resp. maxima denoted B') in the gap with respect to a non-negative integer h. Also we have the total number of extrema in the gap denoted $T' = A' + B'$.

minimum or maximum should be surrounded by zero crossings and that the order in which a minimum or maximum occurs alternates. For example, for the first case reported in Table 1, we have minima on both sides of the gap. In the simplest case, the minimum on the left should be followed by a zero-crossing, maximum, and zero-crossing. The next possible case is that the minimum on the left is followed by a zero-crossing, maximum, zero-crossing, minimum, zero-crossing, maximum, and zero-crossing. For the first possibility, we have one maximum in the gap and no minimum. For the second possibility, we have two maxima and one minimum in the gap. This can go on depending on how many extrema we can fit inside the gap. We therefore use a non-negative integer h to represent the total number of extrema in the gap. For the example given above, if the total number of minima is represented by h and the total number of maxima by $h + 1$, then the first possibility can be described by $h = 0$ and the second one by $h = 1$. This approach is used for all cases in Table 1. The question is how to estimate h.

In order to estimate h we first make a few assumptions. 1) The behaviour of the missing data is similar to the neighbourhood of the gap. 2) For each IMF, every two consecutive extrema in the gap are equally distant.

We now use the following algorithm to estimate h:

- Compute $d^l = |t_{A^\ell} - \tau_{B^\ell}|$ and $d^r = |t_{A^{\ell+1}} - \tau_{B^{\ell+1}}|$;
- Compute $\bar{d} = \min(d^l, d^r)$;
- Compute $D^l = q - \max(t_{A^\ell}, \tau_{B^\ell})$ and $D^r = \min(t_{A^{\ell+1}}, \tau_{B^{\ell+1}}) - (q + Q - 1)$;
- Select $0 \leq P^l \leq 1$ and $0 \leq P^r \leq 1$ and then compute $J^l = q - \lfloor P^l \times D^l \rfloor$ and $J^r = q + Q - 1 + \lfloor P^r \times D^r \rfloor$;
- Compute $\lfloor J^r - J^l / \bar{d} \rfloor - 1$ as the estimated number of extrema in the gap. This is an estimate for T' in Table 1;
- Use the h-dependent expression for T' from Table 1 to obtain an estimate for h.

We have now completed the first goal in this subsection.

In order to obtain the second goal in this subsection we proceed as follows.

- Using the estimated h, we first use the h-dependent expressions in Table 1 to estimate the total number of minima A' and the total number of maxima B' in the gap.
- Compute the increment $\Delta T = \lfloor J^r - J^l / \bar{d} \rfloor$.
- Compute the indices $J^l + k\Delta T$ where $1 \leq k \leq \widehat{T'}$.
- Using the knowledge of the estimated A' and B', we determine the type of extrema that associates with the above indices.

Remarks: The percentage points P^l and P^r must be selected for each IMF. For example, if we have five IMFs, we must select ten percentage points. We select these points manually based on the appearance of each gap-filled IMF. This is not illogical but can be time-consuming.

3.3. Final Gap-filling

The goal in this subsection is to describe the final stage of the EMD gap-filling.

For each IMF, compute the upper and lower envelopes following what we described in the modified EMD. We then assign a point on the upper envelope (resp. lower envelope) at every estimated maxima indices (resp. minima indices) in the gap. We use all the extrema in the gap which we located on the envelopes and use spline interpolation to interpolate the missing data in the gap.

Since the residual term does not contain enough oscillations (in the sense that we do not have two extrema on each side of the gap), we only use a simple spline in order to fill out the gap in the residual. We finally sum all the gap-filled IMFs to reconstruct the signal in the gap.

Remark: For simplicity, we have assumed one gap in the signal. It is often the case that we have multiple number of gaps. The method described above can be used for multiple gaps. The only necessary step is to break down the signal with multiple gaps into a few sub-signals each of which contains one of the gaps. The size of the sub-signal is selected so that we obtain on average 5 to 6 IMFs for each sub-signal.

4. EXAMPLE

In this section, we provide an example in order to demonstrate the performance of the EMD gap-filling. This example uses the Nitrogen dioxide (NO_2) pollutant signal from 1990 in Toronto, Canada taken from http://www.etc-cte.ec.gc.ca/napsdata/Default.aspx. We denote this signal by $\boldsymbol{\mathcal{X}}$. We then construct $\widetilde{\boldsymbol{\mathcal{X}}}$ by inserting three synthetic gaps into $\boldsymbol{\mathcal{X}}$ at base times 80, 166, and 310 and lengths 10, 30, and 16 respectively. Fig. 1 displays $\boldsymbol{\mathcal{X}}$ and $\widetilde{\boldsymbol{\mathcal{X}}}$. In order to extract the IMFs from $\widetilde{\boldsymbol{\mathcal{X}}}$, we construct three signals such that $\widetilde{\boldsymbol{\mathcal{X}}}_1 = \widetilde{\boldsymbol{\mathcal{X}}}(30 : 140)$, $\widetilde{\boldsymbol{\mathcal{X}}}_2 = \widetilde{\boldsymbol{\mathcal{X}}}(106 : 255)$, and $\widetilde{\boldsymbol{\mathcal{X}}}_3 = \widetilde{\boldsymbol{\mathcal{X}}}(280 : 365)$. Applying the modified EMD described in subsection 3.1, we obtained 6 IMFs for $\widetilde{\boldsymbol{\mathcal{X}}}_1$, 6 IMFs for $\widetilde{\boldsymbol{\mathcal{X}}}_2$ and finally 5 IMFs for $\widetilde{\boldsymbol{\mathcal{X}}}_3$. As an example, Fig. 2 displays all

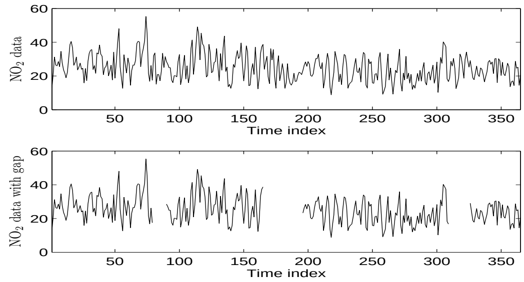

Fig. 1. Top: NO_2 pollutant data from 1990 in Toronto. Bottom: NO_2 pollutant data with three inserted gaps.

the IMFs extracted for $\widetilde{\mathcal{X}}_1$. We now use the EMD gap-filling

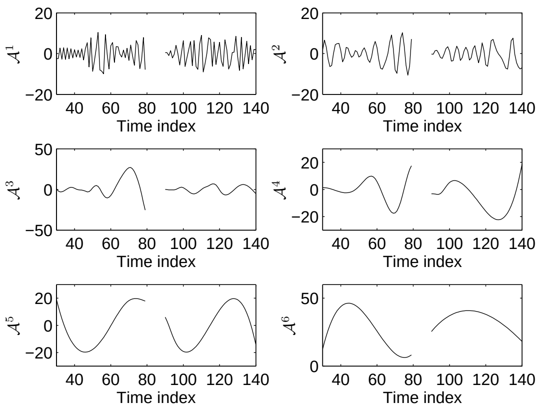

Fig. 2. 6 IMFs extracted from $\widetilde{\mathcal{X}}_1$.

to interpolate the missing data in $\widetilde{\mathcal{X}}_1$, $\widetilde{\mathcal{X}}_2$, and $\widetilde{\mathcal{X}}_3$ independently. The P^l selected for $\widetilde{\mathcal{X}}_1$, $\widetilde{\mathcal{X}}_2$, and $\widetilde{\mathcal{X}}_3$ are respectively $[1, 0.9, 0.8, 0.5, 0]$, $[0.8, 1, 1, 1, 0.5]$, and $[0.8, 0.8, 1, 1]$ and the P^r selected for $\widetilde{\mathcal{X}}_1$,$\widetilde{\mathcal{X}}_2$, and $\widetilde{\mathcal{X}}_3$ are respectively $[0.5, 1, 0.9, 0.8, 0]$, $[1, 0.5, 0, 1, 0.5]$, and $[0.8, 0.5, 0, 0]$. An example of EMD gap-filling is shown in Fig. 4 for the IMFs extracted from $\widetilde{\mathcal{X}}_1$. The gap-filled IMFs for each $\widetilde{\mathcal{X}}_1$,$\widetilde{\mathcal{X}}_2$, and $\widetilde{\mathcal{X}}_3$ are now summed in order to gap-fill $\widetilde{\mathcal{X}}$. Fig. 4 displays the gap-filled $\widetilde{\mathcal{X}}$ in comparison with $\mathcal{X}$ using small regions around each gap. In order to evaluate the performance of the EMD gap-filling, we compare it with 4 other interpolation techniques such as linear , Wiener [4], Papoulis–Gerchberg [2], and singular spectral analysis [5] with window length 100. To do so, we compute the multi-taper spectrum estimate [6] of $\mathcal{X}$ and the gap-filled $\widetilde{\mathcal{X}}$ using all 5 techniques. We then compute the log-Euclidean distance between the estimated spectrum of $\mathcal{X}$ and the estimated spectra of the gap-filled $\widetilde{\mathcal{X}}$ (see Table 2). The results indicates that the EMD gap-filling outperforms the other 4 interpolation techniques.

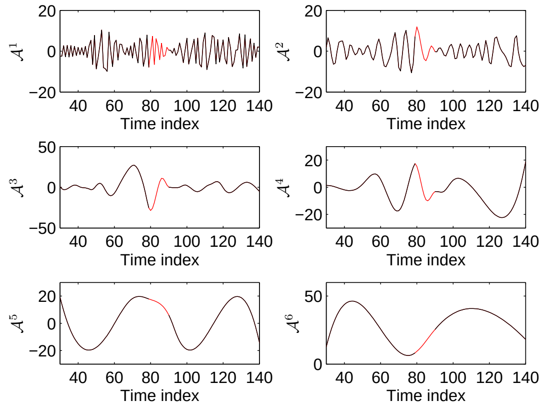

Fig. 3. Gap-filled IMFs for $\widetilde{\mathcal{X}}_1$ using EMD gap-filling.

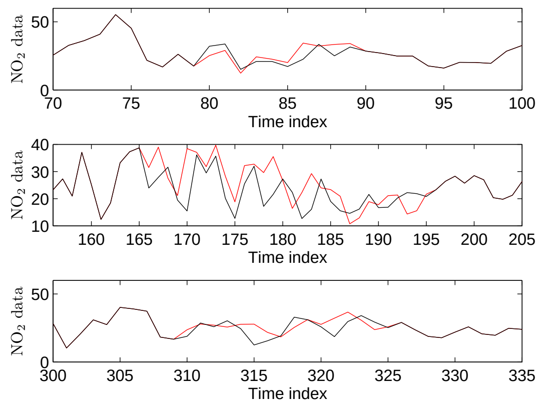

Fig. 4. Gap-filled $\widetilde{\mathcal{X}}$ (red) and $\mathcal{X}$ (black).

Interpolation	log-Euclidean distance
EMD gap-filling	49.1312
Wiener	70.7185
Linear	56.5355
Papoulis–Gerchberg	55.3522
Singular spectral analysis	63.2790

Table 2. The log-Euclidean distance between the estimated spectrum of $\mathcal{X}$ and the estimated spectra of the gap-filled $\widetilde{\mathcal{X}}$ using 5 gap-filling techniques.

5. REFERENCES

[1] E. Meijering, “A chronology of interpolation: From ancient astronomy to modern signal and image processing,” *Proceedings of the IEEE*, vol. 90, no. 4, pp. 319–342, 2002.

[2] P. Jorge and S. G. Ferreira, “Interpolation and the discrete papoulis-gerchberg algorithm,” *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2596–2606, 1994.

[3] N. E. Huang, Z. Shen, S. R. Long, M. L. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, “The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis,” *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, 1998.

[4] D. J. Thomson, L. J. Lanzerotti, and C. G. Maclennan, “Interplanetary magnetic field: Statistical properties and discrete modes,” *Journal of Geophysical Research*, vol. 106, no. A8, pp. 15941–15962, 2001.

[5] 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.

[6] D. J. Thomson, “Spectrum estimation and harmonic analysis,” *Proceedings of the IEEE*, vol. 70, pp. 1055–1096, 1982.