應用疊代式濾波法於潮汐數據修補的初步研究

The First Study upon Tide Data Repairing via Iterative Wave Decomposition Method

Yih Nen Jeng (鄭育能)
Professor of Aero. and Astro. Dept., National Cheng-Kung Univ., Taiwan
Email: z6208016@email.ncku.edu.tw
Yueh-Jiuan G. Hsu (徐月娟)
Director of Marine Meteorology Center, Central Weather Bureau, Taiwan
You-Chi Cheng (鄭又齊)
Engineer, Zu-Soft company, Taipei, Taiwan

ABSTRACT

The iterative filter using the Gaussian smoothing method is employed to decompose and repair a tide data string composed of many tidal wave components. Since the iterative filter can ignore the effect of missing data to certain extent, the tidal wave components can be successively decomposed. However, because the employed wave decomposition method cannot decompose a composite wave found by two wave components whose frequencies close to each other, three beats are found. For those wave components whose wavelengths are larger than $\sqrt{2}$ times the drop-out period, the missing data can be satisfactorily achieved by merely applying the filter. For a longer period of missing data, an iterative technique is developed to repair the data. The tide data of the Houbihu harbor in Pin-Tong at south Taiwan in the period of Jan. 1 through Dec. 31/2001 is employed to demonstrate the procedure of wave decomposition and data repairing.

Keywords: Iterative filter, wave decomposition, data repairing.

INTRODUCTION

Because of the rapid development of computer technique, people can collect many data string simultaneously now. In near future, the nano technology will further increase the number of data string exponentially. Today, before analyze a data string, people often classify, validate, and edit the data which exclude unavailable part, arrange the available part. For example, in Ref.[1], six types of random data anomalies are listed and are recommended to exclude them manually. However, as the number of data string increases to a certain level, it is impossible to do such a data qualification manually again. Therefore, the automatically data qualification procedure becomes more and more important and urgent.

In this study, the procedure of repairing the data drop-outs, which is one of the data anomalies, is examined. It is believed that, during the procedure, one can learn how to accumulate necessary fundamental insights of the automatic data qualification.

The other important issue of data analysis is how to decompose a composite wave which may or may not involving data drop outs. Since every components of the composite wave may change from time to time and involves data anomaly, we can not directly employ classical analyzing method. To the authors' knowledge, the iterative filtering method of Ref.[2-4] has the potential to treat this complicate issue. Therefore, this study employs it as a tool to do the data repairing.

For a data string without serious data anomaly, there are two available methods proposed to decompose a time-series data string consist of many waves with different and variable wavelengths: the matrix pencil method [5,6] and empirical mode decomposition methods of Hwang et al.[7,8]. The former method is successively applied in the electro-magnetic wave problem and the latter is widely used in the surface wave problems. Unfortunately, the matrix pencil method is a zero-th order method which can not take care of the problem with rapidly varying amplitude, frequency, and phase angle. The latter method may or may not suffer by the artificial numerical addition introduced by the cubic spline interpolation. Although the employed iterative filtering method has the limitation that the data string should contain a finite frequency gap around the cut-off frequency, the method has the property of Fourier serious expansion on an unstructured mesh system and can be treated as a semi-analytic method. In other words, it has the intrinsic property that a data drop out may be ignored by the method in some independent spectral range so that it can repair the isolated data drop as will be shown in the following content.

Tides around Taiwan result from shoaling effects of tidal constituents in the Pacific Ocean propagated westward to the continental shelf. According to the harmonic analyses of tide data at Houbihu harbor in southern coast of Taiwan, the Luni-solar Diurnal (K1) and the Principle Solar Diurnal (O1) are the largest diurnal tidal components, and the Principle Lunar (M2), the Principle Solar (S2), the Larger Lunar Elliptic (N2) and the Luni-solar Semidiurnal (K2) are the largest semidiurnal tidal components. Besides, the monthly, fortnightly, annual, semiannual and the long-term (18.61 years) variations of water level are also included.

ANALYSIS

Iterative Filter

Consider a set of data $(x_i, y_i), i = 0, \ 1, 2, ..., n$ to be approximated by a polynomial. The Gaussian smoothing method employs a zero-degree approximation to minimize the following error measure function,

$$I_k = \sum_j \exp[-\frac{(x_j - x_k)^2}{2\sigma^2}][y_j - \sum_{m=0}^{M} A_{m,k}(x_j - x_k)^m]^2 \qquad (1)$$

Where the point k may or may not be belong to points of $j = 0,1,...,n$. The resulting smoothing becomes the following form,

$$\begin{aligned} \bar{y}_j &= \frac{1}{k}\sum_{i=-\infty}^{\infty} e^{-(i-j)^2(\Delta x)^2/2\sigma^2} y_i \\ k &= \sum_{i=-\infty}^{\infty} e^{-(i-j)^2(\Delta x)^2/2\sigma^2} \approx \sqrt{2\pi}\sigma \big/ \Delta x \end{aligned} \qquad (2)$$

This formula can be applied to an irregular point distribution. Assume that a data string y defined on uniform spacing $x_i = i\Delta x$, it is expanded in form of

$$y = \sum_{j=1}^{J} y_j(\lambda_j), \qquad (3)$$

where $y_j(\lambda_j)$'s are wave components consist of sine or cosine functions. After employing the Gaussian smoothing method, the resulting data string can be written in the following form:

$$\begin{aligned} \bar{y}_1 &= S(y) = S(\sum_{j=1}^{J} y_j(\lambda_j)) = \sum_{j=1}^{J} S(y_j(\lambda_j)) \\ &= \sum_{j=1}^{J} a(\sigma/\lambda_j) y_j(\lambda_j) \end{aligned}, \qquad (4)$$

where $S(y)$ denotes the linear smoothing operator, $a(\sigma/\lambda_j, 0)$ is the attenuation factor reflecting the effect of S upon the j-th wave. In Ref.[2], it is shown that

$$0 \le a(\sigma/\lambda_j) \le 1 \qquad (5)$$

where

$$a(\sigma/\lambda_j) \approx \exp\left[-2\pi^2\sigma^2/\lambda_j^2\right] \le 1, \ \ \forall \sigma > 0, \qquad (6)$$

In Ref.[2], the first smoothed waveform of the data string is denoted as $\bar{y}_1$ and the residual waveform, $y_1^{'}$, is :

$$y_1^{'} = y - \bar{y}_1 = \sum_{\lambda_j}[1 - a(\sigma/\lambda_j)] y_j(\lambda_j). \qquad (7)$$

The same smoothing can be applied to the first residual waveform again to obtain the second smoothed waveform,

$$\bar{y}_2 = S(y - \bar{y}_1) = \sum_{\lambda_j}[1 - a(\sigma/\lambda_j)] a(\sigma/\lambda_j) y_j(\lambda_j), \qquad (8)$$

and the second residual wave,

$$y_2^{'} = y_1^{'} - \bar{y}_2 = \sum_{\lambda_j}[1 - a(\sigma/\lambda_j)]^2 y_j(\lambda_j) \qquad (9)$$

This procedure can be applied up to the m-th cycle and the residual and accumulated smoothed waves of the m-th times yield:

$$y_m^{'} = \sum_{\lambda_j}[1 - a(\sigma/\lambda_j)]^m y_j(\lambda_j) \qquad (10)$$

$$\bar{y}(m) = \bar{y}_1 + \bar{y}_2 + ... + \bar{y}_m = \sum_{\lambda_j}\left\{1 - [1 - a(\sigma/\lambda_j)]^m\right\} y_j(\lambda j_j),$$

respectively.

Assume that a data string composed of two wave zones separated by a finite $\Delta\lambda = \lambda_{c2} - \lambda_{c1}$, where λ_{c1} is the shortest wavelength of the high frequency part and λ_{c2} is the longest wavelength of the low frequency part. From Eqs.(3) and (6), the following equations can be constructed

$$
\begin{aligned}
b(\sigma/\lambda_{c1}, m) &= 1 - [1 - \exp\{-\frac{2\pi^2\sigma^2}{\lambda_{c1}^2}\}]^m = 0.001 \\
b(\sigma/\lambda_{c2}, m) &= 1 - [1 - \exp\{-\frac{2\pi^2\sigma^2}{\lambda_{c2}^2}\}]^m = 0.999
\end{aligned}
\tag{11}
$$

After solving these equations, one can get the necessary iteration number m and smoothing factor σ. For the sake of clarity, the iterative procedure is listed below [2].

1. Smooth the original data to get the first smoothed and high frequency parts using the Gaussian smoothing with the given σ.
2. Smooth the high frequency part to obtain the second smoothed and high frequency part.
3. Repeat the same procedure for m times.
4. The final high frequency part is the desired high frequency part and the difference between the original data and final high frequency part is the desired smoothed part.

This iterative filter has a limitation that the frequency gap should be finite and large enough. In other words, for those component waves whose wave lengths close to each other, the beat wave will be obtained. Fortunately, this beat can be further analyzed by checking its spectrum. In this study, the simple procedure of generating spectrum introduced in Ref.[10] is employed.

Data Repairing

Note that the given factor σ in the iterative procedure should be properly given to achieve a suitable filtering job and can be fixed to a certain value by iteration. If the data drop-outs are at isolated points or only but a few continuous points, the iteration can be done automatically.

For those missing data running over a long period, the following simple strategy is proposed. Although the iterative filter can give a solution for such a data drop-outs, the missing data will inevitably introduce certain error within a region whose length is approximately equal to $\sqrt{2}\sigma$ (or one wavelength) around the missing points. Beyond the region, the resulting wave component data is nearly not influenced. Consider the data drop-outs of a beat wave shown in Fig.1. Basing on the above mentioned fact, most data remote from the missing point can be employed to be a reference data string. Around the data drop-out region, the upper and lower envelopes are constructed by connecting the local maximum and minimum points via the following monotonic cubic spline interpolation [9,10], respectively.

$$
\begin{aligned}
& y(x) = c_3(x-x_i)^3 + c_2(x-x_i)^2 + c_1(x-x_i) + c_0, \\
& c_0 = y_i, c_1 = y_i^{'}, \quad s_{i+1/2} = \frac{y_{i+1}-y_i}{x_{i+1}-x_i} \\
& c_2 = \frac{3s_{i+1/2} - 2y_i^{'} - y_{i+1}^{'}}{x_{i+1}-x_i}, \quad c_3 = \frac{y_i^{'} + y_{i+1}^{'} - 2s_{i+1/2}}{(x_{i+1}-x_i)^2}, \\
& y_i^{'} = \mathrm{sgn}(t_i)\min[\frac{1}{2}\left|p_{i-1/2}^{'}(x_i) + p_{i+1/2}^{'}(x_i)\right|, \max(k|s_i|, \frac{k}{2}|t_i|)] \\
& p_{i-1/2}^{'}(x_i) = s_{i-1/2} + d_{i-1/2}(x_i - x_{i-1}), \\
& p_{i+1/2}^{'}(x_i) = s_{i+1/2} + d_{i+1/2}(x_i - x_{i+1}) \\
& t_i = \min\mathrm{mod}[p_{i-1/2}^{'}(x_i), p_{i+1/2}^{'}(x_i)] \\
& d_{i+1/2} = \min\mathrm{mod}(d_i, d_{i+1}), \quad d_i = \frac{s_{i+1/2} - s_{i-1/2}}{x_{i+1} - x_{i-1}}, \\
& s_i = \min\mathrm{mod}[s_{i-1/2}, s_{i+1/2}] \\
& y_i^{''}, y_{i+1}^{''} = 0, \quad \text{if} \quad |y_{i+1} - 2y_i + y_{i-1}| \le \varepsilon \\
& \qquad \text{or} \quad |y_{i+2} - 2y_{i+1} + y_i| \le \varepsilon
\end{aligned}
\tag{12}
$$

Next, these envelopes are employed to scale up the wave data outside the drop-out region and the scaled data is shown as thin line whose amplitude is almost constant. Third, as shown in Fig.2, a segment of scaled wave data within the range marked by two arrows is shift to the drop-out region and is shown as dotted line. Fourth, the shift data is scaled back as shown in Fig.3. Fifth, in the drop-out region, this shifted data is chosen as the repaired data component. Note that the original wave component is different from the repaired one outside the drop-out region because the original one is seriously affected by the missing data.

The same procedure is applied to all wave components except the highest frequency part which is referred as noise. For those wave components with long enough wave length, the short period data drop-out region is automatically repaired and need not to treat it. Finally, the repaired data of all wave components are summed up to replace the data drop-out points. However, as the above mentioned discussion about Fig.3, the repaired data may or may not consist with the existing data. Therefore, the above procedure should be repeatedly applied until the slopes at all end points of every drop-out region are smooth.

RESULTS AND DISCUSSIONS

The tide data of the Houbihu harbor in Pin-Tong at south Taiwan in the period of Jan.1 through Dec. 31/2001 is employed to illustrate the proposed repairing procedure. The data is the water level in milli-meter recorded at every hour. It contains many isolated data drop-out points, a three-hour missing period, a one day drop-out, and a two day drop-out region. The Original data and the result of employing $\sigma = 90$ and 30 are shown in Fig.4. At the missing point, the original data takes a zero value. It seems that the first two long waves are not affected by the data drop-outs. These two waves are principally generated by the variation of the northern Pacific Ocean current heading the Taiwan Strait that is modulated by the location of sun heating. The result of employing $\sigma = 10,5$ is shown in Fig.5, and $\sigma = 3$ in Fig.6. For these waves generated by $\sigma \geq 3$, the isolated, short period, and long period data missing are all automatically repaired. It seems that, for those waves decoupled with a σ larger than the drop-out period (or wave length larger than $\sqrt{2}$ times the period), the iterative filter can automatically repair the data. The resulting long wave of employing $\sigma = 1$ is shown in Fig.7 and some portion of those long waves employing $\sigma = 0.4,$ $0.25, 0.125$ and the final high frequency part are shown in Fig.8. The last three long waves are beats composed of two single waves whose frequency close to each other. For these high frequency waves, except the final high frequency part which is considered as the noise, the isolated and short period data are automatically repaired. Fig.9a is one part of the repaired data. For the long period data drop-out, two repairing procedures are employed twice and the results are shown in Fig.10a and 10b, respectively. Although there is not answer how the correct data is, it seems that the result of the second cycle captures the trend of neighboring data and is better than that of the first cycle.

In order to inspect the detail of these pseudo-sinusoidal wave components except the non-sinusoidal part decoupled by $\sigma = 90$, their spectrums evaluated by the method of ref.[10] are shown below. Note that the non-sinusoidal part always generates an exponentially decreasing low frequency part and is not easy be understood in the spectrum domain. All the pseudo-sinusoidal waves are carefully cut to be zero at two ends via the linear interpolation without modifying the data. Consequently, all the undesired low frequency part is eliminated.

Figure 11 shows the spectrum of the wave decoupled by $\sigma = 30, 10, 5, 3, 1, 0.4, 0.25,$ and 0.125, respectively. In Fig.11a, the wide dominate band shows that the wavelength is not constantly over the computed range. The first peak (amplitude is about 11 mm and wave length is about 26.92 days) of Fig.11b is related to the MM and MSM tidal components (the amplitudes are 8.6 and 10.1 mm and wavelengths are 27.554 and 31.82 days, estimated by the harmonic analysis in the range of 2002/08/01 through 2004/07/31) but has certain error. However, spectrums shown in Figs.11c through 11e, there is no obviously dominant frequency. It seems that this harbor's long wavelength tide is seriously influenced by the northern Pacific Ocean current and complicated weather so that all the other long wavelength tide cannot be seen.

In Fig.11f, which is corresponding to the beat of Fig.8 decoupled by $\sigma = 0.4$, there are two dominant frequencies (23.768 hrs for the larger peak and 25.596 through 25.845 hrs for the smaller peak) and close to the K1 (215.2mm and 23.934 hrs) and O1 (199mm and 25.819 hrs) tidal components with errors induced by insufficiently fine data resolution. As to the amplitude estimation, the estimated peaks are 70 and 52 mm (the maximum amplitude indicated in Fig.8 is about 600 mm which are a weighted sum of every modes) which are different from that estimated by the harmonic analysis shown before. A careful inspection upon Fig.11f reveals that both peaks are not simple impulse which reflects amplitudes change slowly and wavelengths may be not of constant values. The O1 further indicates that it is a beat formed by two wave components. The authors believe these are the reason of difference of amplitude estimation between the present method and harmonic analysis.

Figure 11g shows the semidiurnal wave beat (the amplitude and wavelength of the larger peak are 111mm and 12.425 hrs, respectively, and that of the smaller peak are 39 mm and 11.960 hrs). The related waves estimated by the harmonic analysis are M2 (259.9mm and 12.420 days), S2 (115.5mm and 12 hrs), K2 (30.5 mm and11.967 hrs), and N2 (53.9 mm and 12.568 hrs) tidal components with error causing by insufficiently fine data resolution. That shown in Fig.11h is the anther beat formed by similar tidal components. The fact that the beat of Fig.11h can be separated from the beat of Fig.11g reflects that the former is a more complicated combination than the latter and need a more robust decomposition method to decouple them.

The above discussion shows that the proposed data repairing procedure upon every wave component separately can effectively repair the data and capture the dominant tidal components very well.

CONCLUSIONS

The iterative filtering procedure can effectively decompose tidal wave components. All the dominant tidal components are captured in the form of beats. The proposed data repairing method can also effectively recover isolated data drop-outs. It seems that this procedure can be applied to many other fields provided that the beats can be further decomposed into single wave components.

ACKNOWDGEMENT

This work is supported by the National Science Council of Taiwan under the grant number NSC-93 -2212-E006-037.

REFERENCE

1. Bendat, J. S. and Piersol, A. G., “*Random Data Analysis and Measurement Procedures*,” 3rd ed., John Wiley & Sons, Singapore, 2000.
2. Jeng, Y. N., Huang, G. P. G., and Chen, H., “Wave Decomposition in Physical Space Using Iterative Moving Least Squares Methods,” 11th national conference on computational fluid dynamics Taiwan, paper no. CFD11-0107, Aug. 2004.
3. Jeng, Y. N., Cheng, Y. C., and Yang, T. M., “Wave Decomposition Across Discontinuity Using Iterative Moving Least Squares Method,” 11th national conference on computational fluid dynamics Taiwan, paper no. CFD11-0110, Aug. 2004.
4. Jeng, Y. N. and Huang, G. P. G., “Evaluation of Derivatives from Composite Wave via Iterative Filter and A Numerical Method of Solving Partial Differential Equations,” 11th national conference on computational fluid dynamics Taiwan, paper no. CFD11-0111, Aug. 2004.
5. Hua, R. and Sarkar, T. K., “Generalized Pencil-of-Function Method for Extracting Poles of an EM System from Its Transient Response,” IEEE Trans. Antennas and Propagation, vol.37, no.2, Feb. 1989, pp.229-234.
6. Ruscio, D. D., “On Subspace Identification of Extended Observability Matrix,” Proc. Of 36th Conf. on Decision & Control., San Diego, Cal. U.S.A., Dec. 1997, Paper no.TA05, pp.1841-1847.
7. Huang N. E., Z. Shen, and S. R. Long, “A New View of Nonlinear Water Waves: the Hilbert Spectrum,” Annual Reviews of Fluid Mechanics, 1999, vol.51, pp.417-457.
8. Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. and Liu, H. H., “The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis,” Proc. Royal Soc. London A, vol. 454, pp.903-995, 1998.
9. Huynh, H. T., “ Accurate Monotone Cubic Interpolation,” SIAM J. Number. Anal. vol.30, no.1, pp57-100, Feb.1993
10. Y. N. Jeng and Y. C. Cheng, “A Simple Strategy to Evaluate the Frequency Spectrum of a Time Series Data with Non-Uniform Intervals,” 10-th National Computational Fluid Dynamics Conference of Taiwan, Hua-Lien, 2003, Paper no.A-9, also Transactions of the Aeronautical and Astronautic Society of the Republic of China in press, Sept. 2004.

應用疊代式濾波法於潮汐數據修補的初步研究

鄭育能 成功大學航太系教授

徐月娟 中央氣象局海象測報中心主任

鄭又齊 儒園系統公司 工程師

摘　要

本文應用疊代式高斯濾波器拆解組合成潮汐之波形，發現可以拆解出一組全日潮和兩組半日潮的複合波，組合成全日潮之一個波的振幅可能會有緩慢的變化。本文也對各波形發展出空缺數據之修補法。對於單一和少數連續數據之空缺點，疊代式濾波器可以自動修補之。對於較長時間的空白數據點，波長大於空缺時間之波形也可以自動修補之。針對較短波長的波形，本文發展出針對各波形修補的簡易方法。本文並以屏東後壁湖漁港的水位數據為例子，說明拆解和修補過程，初步結果令人滿意。

關鍵詞：疊代型濾波器，潮汐波形拆解，空缺數據修補法。

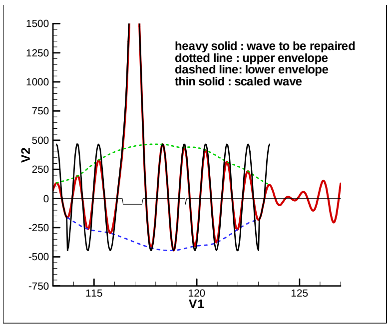


Fig.1 The schematic diagram of data repairing. For the central line: 0 denotes regular data, -50 is drop-out point; dotted lines are upper and lower envelopes; heavy line is the original decoupled wave; thin solid line is the scaled wave.

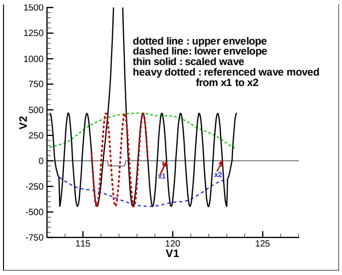


Fig.2 Wave data in the region marked by two arrows are moved to the data drop-out region and is shown as dotted line.

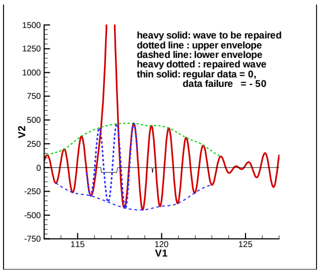


Fig.3 The shifted data string is scaled back as the dotted line.

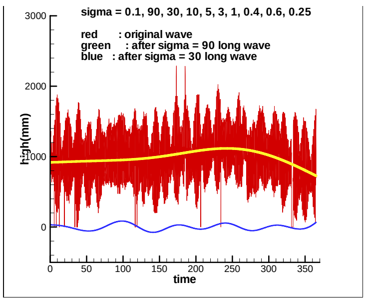


Fig.4 The original data and two decoupled waves. At the data drop-out point, the original data takes a zero value. The wave decoupled by $\sigma = 90$ is shown in heavy line around the middle region of the original data, the wave decoupled by $\sigma = 30$ is at the bottom region.

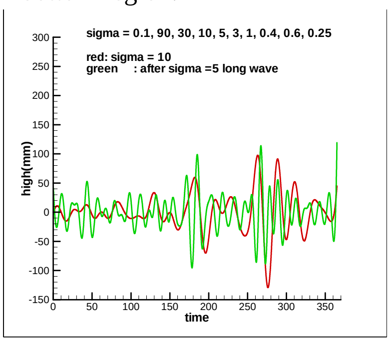


Fig.5 The wave components decoupled by $\sigma = 10, 5$.

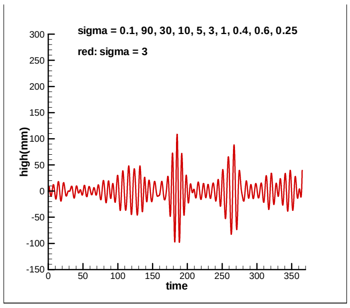


Fig.6 The wave component decoupled by $\sigma = 3$.

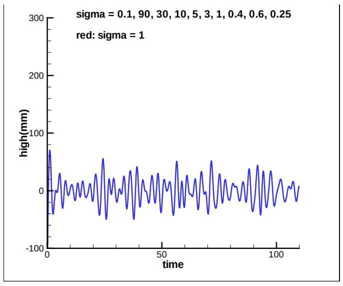


Fig.7 The wave component decoupled by $\sigma = 1$

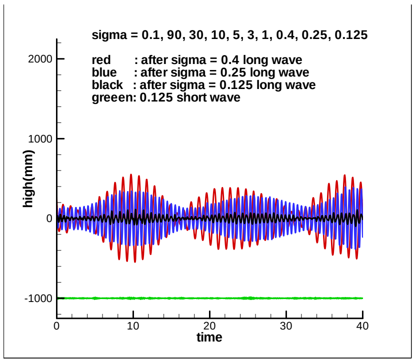


Fig.8 Four decoupled waves: long waves generated by $\sigma = 0.4, 0.25, 0.125$ and bottom wave is the final high frequency part which is considered as the noise.

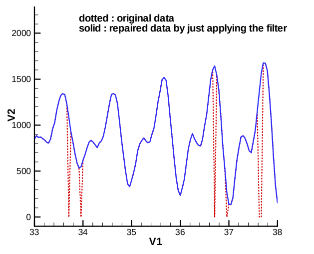


Fig.8 Some part of the original and repaired data by just applying the filter with $\sigma = 0.075$.

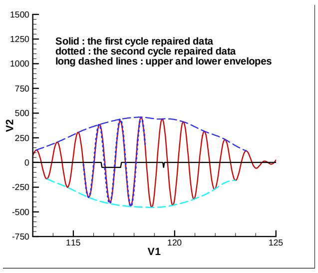
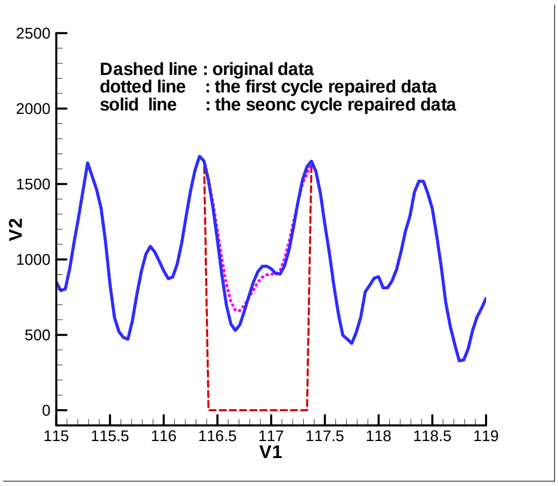


Fig.9 The first cycle and second cycle repaired data coincide to each other.

(10a)



(10b)

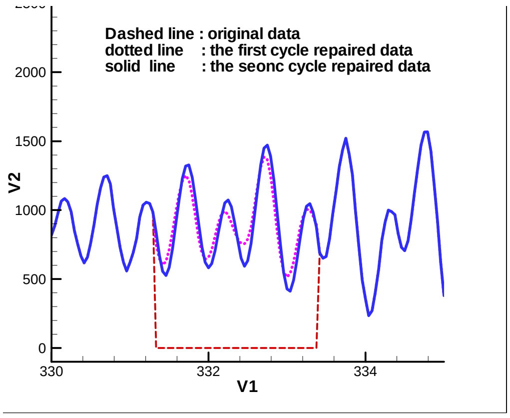
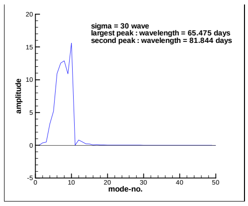
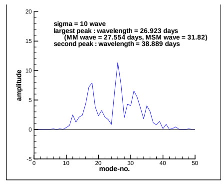


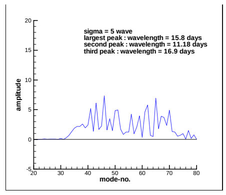
Fig.10 The original, first cycle and second cycle repaired data: (a) 1 day drop-out and (b) 2 days dropout.



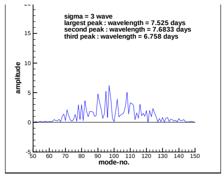
(11a)



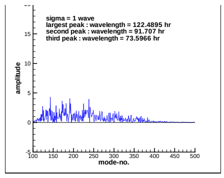
(11b)



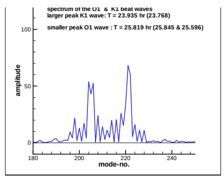
(11c)



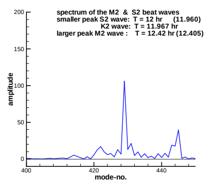
(11d)



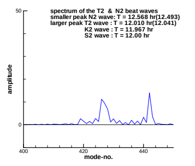
(11e)



(11f)



(11g)



(11h)

Fig.11 Spectrum of the decoupled wave components: 9a through 9h.