

PREFACE

Empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA) represent a desperate attempt to break the suffocating hold on data analysis by the twin assumptions of linearity and stationarity. To analyze the data from nonlinear and non-stationary processes, various attempts such as Spectrograms, Wavelet analysis, and the Wigner-Ville distribution have been made, but the EMD-HSA approach is unique and different from the existing methods of data analysis. The EMD-HAS is truly an adaptive time-frequency analysis. It does not require an *a priori* functional basis. Instead, the basis functions are derived adaptively from the data by the EMD sifting procedures; the instantaneous frequencies are computed from derivatives of the phase functions of the Hilbert transform of the basis functions; the final result is presented in the time-frequency space. The EMD-HSA is a magnifying glass for analyzing the data from nonlinear and non-stationary processes. The EMD-HSA results are intriguing and are no longer shackled by spurious harmonics (the artifacts of imposing a linearity property on a nonlinear system) or limited by the uncertainty principle (the consequence of Fourier transform pairs in data analysis).

EMD-HSA was originally designed in 1995 specifically to study water surface wave evolution, the phenomenon of high frequency waves with short fetch evolving into low frequency waves at long fetch. With the EMD-HSA method, it was found that the evolution of the waves was not continuous but abrupt, discrete and local. Subsequently, NEH spent two years visiting Caltech at the invitation of Professor Theodore Y. Wu. Under the guidance of Professor Wu and Professor Owen M. Phillips of the Johns Hopkins University, the EMD-HSA method was further developed and various applications explored. Professor Wu designated the method as the Hilbert-Huang Transform (HHT), a name later adopted by NASA to avoid the awkward name of EMD-HSA. It is only fair to say that the HHT would not have been developed without the encouragement and guidance of Professors Wu and Phillips.

The HHT's power and effectiveness in data analysis have been demonstrated by its successful application to many important problems covering engineering, biomedical, financial and geophysical data. The mathematical development of the HHT, however, is undergoing the same path as other significant and historical data analysis methods as in Fourier analy-

sis and wavelet analysis: Applications are leading to development, and the mathematical theories are following, since the methods were motivated by applications. Mathematicians' apparent interest in the HHT motivated our organization of an HHT mini-symposium at the joint meeting between the Society for Industrial and Applied Mathematics and the Canadian Applied and Industrial Mathematics Society in June of 2003 at Montreal.

This book contains most of the presentations made at the mini-symposium with some additions. The book contents are divided into two groups: the theoretical aspects and the applications, with the applications further grouped into geophysics, structural safety, and visualization. In the theoretical aspects, the chapters cover the attempts of mathematicians to apply rigor to the empirical method such as the representation of the IMF by B-spline functions, filter based decompositions, and the statistical characteristics of the IMFs. This book also represents a plea for help from the mathematical community. A list of outstanding mathematical problems is given in Chapter 1. The chapters on applications include the correction of satellite orbit drifting, detection of failure of highway bridges and other structures, discoveries of the patterns and anomalies in climate data, and calculation of the instantaneous frequency of water waves. The objectives of the book are to provide HHT users with a collection of successful HHT applications, to supply graduate students and researchers with an HHT tutorial, and to inform data analysis mathematicians of the outstanding mathematical problems of HHT.

This book is intended as a reference for anyone who are involved in signal analysis by processing data from nonlinear and non-stationary systems. Although each chapter is independent from the others, it is sufficiently pedagogical so that every single chapter or the entire book is suitable as a part of a graduate course on signal analysis. To use this book efficiently, the readers should have background knowledge of calculus, Fourier transform, numerical analysis and differential equations. The HHT algorithm has been patented by NASA; non-commerical users may obtain it at the website: <http://techtransfer.gsfc.nasa.gov>.

Much effort went into compiling this collection of papers into a book form. In this processes, we owe our gratitude to Dr. Dean Duffy for his skillful editing and typesetting, and without his efficient and professional work, this book would not have been possible.

Norden E. Huang and Samuel S. P. Shen,
Greenbelt, Maryland

TABLE OF CONTENTS

Preface	i
Table of Contents	iii
<i>Theoretical Aspects</i>	
1 Introduction to the Hilbert-Huang Transform and Its Related Mathematical Problems	1
<i>Norden E. Huang</i>	
1.1 Introduction	1
1.2 The Hilbert-Huang transform	3
1.2.1 The empirical mode decomposition method (the sifting process)	5
1.2.2 The Hilbert spectral analysis	13
1.3 Recent developments	16
1.3.1 Normalized Hilbert transform	16
1.3.2 Confidence limit	19
1.3.3 Statistical significance of IMFs	20
1.4 Mathematical problems related to the HHT	21
1.4.1 Adaptive data-analysis methodology	22
1.4.2 Nonlinear system identification	22
1.4.3 The prediction problem for nonstationary processes (the end effects of EMD)	24
1.4.4 Spline problems (the best spline implementation for HHT, convergence and 2-D)	24
1.4.5 The optimization problem (the best IMF selection and uniqueness mode mixing)	26
1.4.6 Approximation problems (the Hilbert transform and quadrature)	27
1.4.7 Miscellaneous statistical questions concerning HHT	28
1.5 Conclusion	28

2 B-Spline Based Empirical Mode Decomposition	33
<i>Sherman Riemenschneider, Bao Liu, Yuesheng Xu and Norden E. Huang</i>	
2.1 Introduction	33
2.2 A B-spline algorithm for empirical mode decomposition	36
2.3 Some related mathematical results	39
2.4 Performance analysis of BS-EMD	47
2.5 Application examples	53
2.6 Conclusion and future research topics	58
3 EMD Equivalent Filter Banks, from Interpretation to Applications	67
<i>Patrick Flandrin, Paulo Gonçalvès and Gabriel Rilling</i>	
3.1 Introduction	67
3.2 A stochastic perspective in the frequency domain	68
3.2.1 Model and simulations	68
3.2.2 Equivalent transfer functions	69
3.3 A deterministic perspective in the time domain	73
3.3.1 Model and simulations	74
3.3.2 Equivalent impulse responses	74
3.4 Selected applications	75
3.4.1 EMD-based estimation of scaling exponents	75
3.4.2 EMD as a data-driven spectrum analyzer	78
3.4.3 Denoising and detrending with EMD	81
3.5 Concluding remarks	84
4 HHT Sifting and Filtering	89
<i>Reginald N. Meeson</i>	
4.1 Introduction	89
4.2 Objectives of HHT sifting	91
4.2.1 Restrictions on amplitude and phase functions	92
4.2.2 End-point analysis	95
4.3 Huang's sifting algorithm	96
4.4 Incremental, real-time HHT sifting	97
4.4.1 Testing for iteration convergence	98
4.4.2 Time-warp analysis	99
4.4.3 Calculating warp filter characteristics	101
4.4.4 Separating amplitude and phase	102
4.5 Filtering in standard time	103

Table of Contents

v

4.6 Case studies	105
4.6.1 Simple reference example	105
4.6.2 Amplitude modulated example	106
4.6.3 Frequency modulated example	109
4.6.4 Amplitude step example	112
4.6.5 Frequency shift example	117
4.7 Summary and conclusions	120
4.7.1 Summary of case study findings	121
4.7.2 Research directions	121
5 Statistical Significance Test of Intrinsic Mode Functions	125
<i>Zhaohua Wu and Norden E. Huang</i>	
5.1 Introduction	125
5.2 Characteristics of Gaussian white noise in EMD	128
5.2.1 Numerical experiment	129
5.2.2 Mean periods of IMFs	129
5.2.3 The Fourier spectra of IMFs	130
5.2.4 Probability distributions of IMFs and their energy	132
5.3 Spread functions of mean energy density	135
5.4 Examples of a statistical significance test of noisy data	138
5.4.1 Testing of the IMFs of the NAOI	140
5.4.2 Testing of the IMFs of the SOI	141
5.4.3 Testing of the IMFs of the GASTA	142
5.4.4 <i>A posteriori</i> test	145
5.5 Summary and discussion	145

Applications to Geophysics

6 The Application of Hilbert-Huang Transforms to Meteorological Datasets	149
<i>Dean G. Duffy</i>	
6.1 Introduction	149
6.2 Procedure	151
6.3 Applications	156
6.3.1 Sea level heights	156
6.3.2 Solar radiation	161
6.3.3 Barographic observations	164
6.4 Conclusion	167

7 Empirical Mode Decomposition and Climate Variability	173
<i>Katie Coughlin and Ka Kit Tung</i>	
7.1 Introduction	173
7.2 Data	174
7.3 Methodology	177
7.4 Statistical tests of confidence	179
7.5 Results and physical interpretations	184
7.5.1 Annual cycle	184
7.5.2 Quasi-Biennial Oscillation (QBO)	184
7.5.3 ENSO-like mode	185
7.5.4 Solar cycle signal in the stratosphere	186
7.5.5 Fifth mode	188
7.5.6 Trends	189
7.6 Conclusions	189
8 EMD Correction of Orbital Drift Artifacts in Satellite Data Stream	193
<i>Jorge E. Pinzón, Molly E. Brown and Compton J. Tucker</i>	
8.1 Introduction	193
8.2 Processing of NDVI imagery	196
8.3 Empirical mode decomposition	199
8.4 Impact of orbital drift on NDVI and EMD-SZA filtering	200
8.5 Results and discussion	204
8.6 Extension to 8-km data	207
8.7 Integration of NOAA-16 data	211
8.8 Conclusions	212
9 HHT Analysis of the Nonlinear and Non-Stationary Annual Cycle of Daily Surface Air Temperature Data	217
<i>Samuel S. P. Shen, Tingting Shu, Norden E. Huang, Zhaohua Wu, Gerald R. North, Thomas R. Karl and David R. Easterling</i>	
9.1 Introduction	217
9.2 Analysis method and computational algorithms	222
9.3 Data	225
9.4 Time analysis	227
9.4.1 Examples of the TAC and the NAC	227
9.4.2 Temporal resolution of data	229
9.4.3 Robustness of the EMD method	231

Table of Contents

vii

9.4.3.1 EMD separation of a known signal in a synthetic dataset	231
9.4.3.2 Robustness with respect to data length	233
9.4.3.3 Robustness with respect to end conditions	233
9.5 Frequency analysis	233
9.5.1 Hilbert spectra of NAC	233
9.5.2 Variances of anomalies with respect to the NAC and TAC .	236
9.5.3 Spectral power of the anomalies with respect to the NAC and TAC .	237
9.6 Conclusions and discussion	240
10 Hilbert Spectra of Nonlinear Ocean Waves	245
<i>Paul A. Hwang, Norden E. Huang, David W. Wang, and Jame M. Kaihatu</i>	
10.1 Introduction	245
10.2 The Hilbert-Huang spectral analysis	246
10.3 Spectrum of wind-generated waves	251
10.4 Statistical properties and group structure	254
10.5 Summary	258

Applications to Structural Safety

11 EMD and Instantaneous Phase Detection of Structural Damage	263
<i>Liming W. Salvino, Darryll J. Pine, Michael Todd and Jonathan M. Nichols</i>	
11.1 Introduction to structural health monitoring	263
11.2 Instantaneous phase and EMD	266
11.2.1 Instantaneous phase	267
11.2.2 EMD and HHT	268
11.2.3 Extracting an instantaneous phase from measured data .	270
11.3 Damage detection application	272
11.3.1 One-dimensional structures	273
11.3.2 Experimental validations	276
11.3.3 Instantaneous phase detection	280
11.4 Frame structure with multiple damages	282
11.4.1 Frame experiment	282
11.4.2 Detecting damage by using the HHT spectrum	286
11.4.3 Detecting damage by using instantaneous phase features .	287

11.4.4 Auto-regressive modeling and prediction error	292
11.4.5 Chaotic-attractor-based prediction error	295
11.5 Summary and conclusions	299

12 HHT-Based Bridge Structural Health-Monitoring

Method	305
<i>Norden E. Huang, Kang Huang and Wei-Ling Chiang</i>	

12.1 Introduction	305
12.2 A review of the present state-of-the-art methods	307
12.2.1 Data-processing methods	308
12.2.2 Loading conditions	311
12.2.3 The transient load	313
12.3 The Hilbert-Huang transform	315
12.4 Damage detection criteria	316
12.5 Case study of damage detection	318
12.6 Conclusions	326

Applications to Visualization

13 Applications of HHT in Image Analysis	335
<i>Steven R. Long</i>	

13.1 Introduction	335
13.2 Overview	337
13.3 The analysis of digital slope images	338
13.3.1 The NASA laboratory	338
13.3.2 The digital camera and set-up	338
13.3.3 Acquiring experimental images	339
13.3.4 Using EMD/HHT analysis on images	342
13.3.5 The digital camera and set-up	342
13.3.5.1 Volume computations and isosurface visualization .	345
13.3.5.2 Use of EMD/HHT in image decomposition	348
13.4 Summary	352

Index	355
--------------	-----

CHAPTER 1

INTRODUCTION TO THE HILBERT-HUANG TRANSFORM AND ITS RELATED MATHEMATICAL PROBLEMS

Norden E. Huang

The Hilbert-Huang transform (HHT) is an empirically based data-analysis method. Its basis of expansion is adaptive, so that it can produce physically meaningful representations of data from nonlinear and nonstationary processes. The advantage of being adaptive has a price: the difficulty of laying a firm theoretical foundation. This chapter is an introduction to the basic method, which is followed by brief descriptions of the recent developments relating to the normalized Hilbert transform, a confidence limit for the Hilbert spectrum, and a statistical significance test for the intrinsic mode function (IMF). The mathematical problems associated with the HHT are then discussed. These problems include (i) the general method of adaptive data-analysis, (ii) the identification methods of nonlinear systems, (iii) the prediction problems in nonstationary processes, which is intimately related to the end effects in the empirical mode decomposition (EMD), (iv) the spline problems, which center on finding the best spline implementation for the HHT, the convergence of EMD, and two-dimensional EMD, (v) the optimization problem or the best IMF selection and the uniqueness of the EMD decomposition, (vi) the approximation problems involving the fidelity of the Hilbert transform and the true quadrature of the data, and (vii) a list of miscellaneous mathematical questions concerning the HHT.

1.1. Introduction

Traditional data-analysis methods are all based on linear and stationary assumptions. Only in recent years have new methods been introduced to analyze nonstationary and nonlinear data. For example, wavelet analysis and the Wagner-Ville distribution (Flandrin 1999; Gröchenig 2001) were designed for linear but nonstationary data. Additionally, various nonlinear time-series-analysis methods (see, for example, Tong 1990; Kantz and Schreiber 1997; Diks 1999) were designed for nonlinear but stationary and deterministic systems. Unfortunately, in most real systems, either natural or even man-made ones, the data are most likely to be both nonlinear

and nonstationary. Analyzing the data from such a system is a daunting problem. Even the universally accepted mathematical paradigm of data expansion in terms of an *a priori* established basis would need to be eschewed, for the convolution computation of the *a priori* basis creates more problems than solutions. A necessary condition to represent nonlinear and nonstationary data is to have an adaptive basis. An *a priori* defined function cannot be relied on as a basis, no matter how sophisticated the basis function might be. A few adaptive methods are available for signal analysis, as summarized by Windrow and Stearns (1985). However, the methods given in their book are all designed for stationary processes. For nonstationary and nonlinear data, where adaptation is absolutely necessary, no available methods can be found. How can such a basis be defined? What are the mathematical properties and problems of the basis functions? How should the general topic of an adaptive method for data analysis be approached? Being adaptive means that the definition of the basis has to be data-dependent, an *a posteriori*-defined basis, an approach totally different from the established mathematical paradigm for data analysis. Therefore, the required definition presents a great challenge to the mathematical community. Even though challenging, new methods to examine data from the real world are certainly needed. A recently developed method, the Hilbert-Huang transform (HHT), by Huang et al. (1996, 1998, 1999) seems to be able to meet some of the challenges.

The HHT consists of two parts: empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). This method is potentially viable for nonlinear and nonstationary data analysis, especially for time-frequency-energy representations. It has been tested and validated exhaustively, but only empirically. In all the cases studied, the HHT gave results much sharper than those from any of the traditional analysis methods in time-frequency-energy representations. Additionally, the HHT revealed true physical meanings in many of the data examined. Powerful as it is, the method is entirely empirical. In order to make the method more robust and rigorous, many outstanding mathematical problems related to the HHT method need to be resolved. In this section, some of the problems yet to be faced will be listed, in the hope of attracting the attention of the mathematical community to this interesting, challenging and critical research area. Some of the problems are easy and might be resolved in the next few years; others are more difficult and will probably require much more effort. In view of the history of Fourier analysis, which was invented in 1807 but not fully proven until 1933 (Plancherel 1933), it should be anticipated that significant time and

effort will be required. Before discussing the mathematical problem, a brief introduction to the methodology of the HHT will first be given. Readers interested in the complete details should consult Huang et al. (1998, 1999).

1.2. The Hilbert-Huang transform

The development of the HHT was motivated by the need to describe nonlinear distorted waves in detail, along with the variations of these signals that naturally occur in nonstationary processes. As is well known, the natural physical processes are mostly nonlinear and nonstationary, yet the data analysis methods provide very limited options for examining data from such processes. The available methods are either for linear but nonstationary, or nonlinear but stationary and statistically deterministic processes, as stated above. To examine data from real-world nonlinear, nonstationary and stochastic processes, new approaches are urgently needed, for nonlinear processes need special treatment. The past approach of imposing a linear structure on a nonlinear system is just not adequate. Other than periodicity, the detailed dynamics in the processes from the data need to be determined because one of the typical characteristics of nonlinear processes is their intra-wave frequency modulation, which indicates the instantaneous frequency changes within one oscillation cycle. As an example, a very simple nonlinear system will be examined, given by the non-dissipative Duffing equation as

$$\frac{d^2x}{dt^2} + x + \epsilon x^3 = \gamma \cos(\omega t), \quad (1.1)$$

where ϵ is a parameter not necessarily small, and γ is the amplitude of a periodic forcing function with a frequency ω . In (1.1), if the parameter ϵ were zero, the system would be linear, and the solution would be easily found. However, if ϵ were non-zero, the system would be nonlinear. In the past, any system with such a parameter could be solved by using perturbation methods, provided that $\epsilon \ll 1$. However, if ϵ is not small compared to unity, then the system becomes highly nonlinear, and new phenomena such as bifurcations and chaos will result. Then perturbation methods are no longer an option; numerical solutions must be attempted. Either way, (1.1) represents one of the simplest nonlinear systems; it also contains all the complications of nonlinearity. By rewriting the equation in a slightly different form as

$$\frac{d^2x}{dt^2} + x(1 + \epsilon x^2) = \gamma \cos(\omega t), \quad (1.2)$$

its features can be better examined. Then the quantity within the parenthesis can be regarded as a variable spring constant, or a variable pendulum length. As the frequency (or period) of the simple pendulum depends on the length, it is obvious that the system given in (1.2) should change in frequency from location to location, and time to time, even within one oscillation cycle. As Huang et al. (1998) pointed out, this intra-frequency frequency variation is the hallmark of nonlinear systems. In the past, when the analysis was based on the linear Fourier analysis, this intra-wave frequency variation could not be depicted, except by resorting to harmonics. Thus, any nonlinear distorted waveform has been referred to as “harmonic distortions.” Harmonics distortions are a mathematical artifact resulting from imposing a linear structure on a nonlinear system. They may have mathematical meaning, but not a physical meaning (Huang et al. 1999). For example, in the case of water waves, such harmonic components do not have any of the real physical characteristics of a real wave. The physically meaningful way to describe the system is in terms of the instantaneous frequency, which will reveal the intra-wave frequency modulations.

The easiest way to compute the instantaneous frequency is by using the Hilbert transform, through which the complex conjugate $y(t)$ of any real valued function $x(t)$ of L^p class can be determined (see, for example, Titchmarsh 1950) by

$$\mathcal{H}[x(t)] = \frac{1}{\pi} \text{PV} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau, \quad (1.3)$$

in which the PV indicates the principal value of the singular integral. With the Hilbert transform, the analytic signal is defined as

$$z(t) = x(t) + iy(t) = a(t)e^{i\theta(t)}, \quad (1.4)$$

where

$$a(t) = \sqrt{x^2 + y^2}, \quad \text{and} \quad \theta(t) = \arctan\left(\frac{y}{x}\right). \quad (1.5)$$

Here, $a(t)$ is the instantaneous amplitude, and θ is the phase function, and the instantaneous frequency is simply

$$\omega = \frac{d\theta}{dt}. \quad (1.6)$$

A description of the Hilbert transform with the emphasis on its many mathematical formalities can be found in Hahn (1996). Essentially, (1.3) defines the Hilbert transform as the convolution of $x(t)$ with $1/t$; therefore, (1.3)

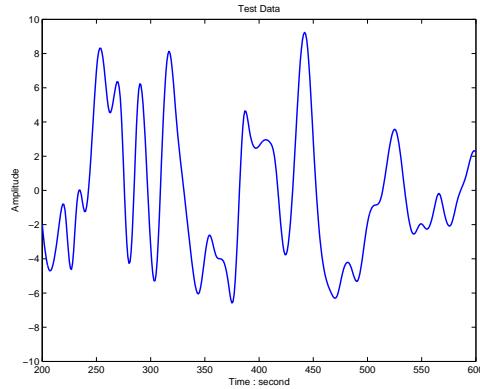


Figure 1.1: The test data.

emphasizes the local properties of $x(t)$. In (1.4), the polar coordinate expression further clarifies the local nature of this representation: it is the best local fit of an amplitude and phase-varying trigonometric function to $x(t)$. Even with the Hilbert transform, defining the instantaneous frequency still involves considerable controversy. In fact, a sensible instantaneous frequency cannot be found through this method for obtaining an arbitrary function. A straightforward application, as advocated by Hahn (1996), will only lead to the problem of having frequency values being equally likely to be positive and negative for any given dataset. As a result, the past applications of the Hilbert transform are all limited to the narrow band-passed signal, which is narrow-banded with the same number of extrema and zero-crossings. However, filtering in frequency space is a linear operation, and the filtered data will be stripped of their harmonics, and the result will be a distortion of the waveforms. The real advantage of the Hilbert transform became obvious only after Huang et al. (1998) introduced the empirical mode decomposition method.

1.2.1. *The empirical mode decomposition method (the sifting process)*

As discussed by Huang et al. (1996, 1998, 1999), the empirical mode decomposition method is necessary to deal with data from nonstationary and nonlinear processes. In contrast to almost all of the previous methods, this new method is intuitive, direct, and adaptive, with an *a posteriori*-defined basis, from the decomposition method, based on and derived from the data.

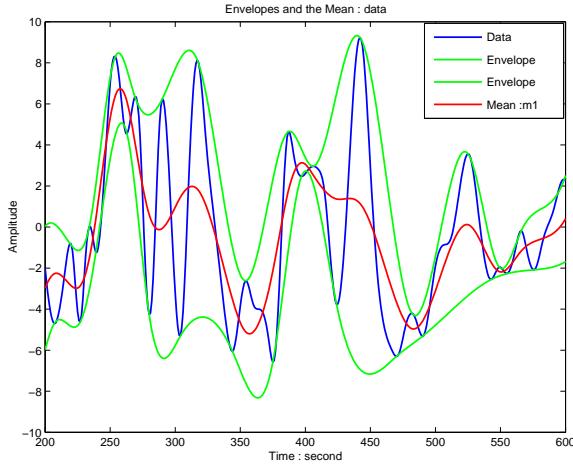


Figure 1.2: The data (blue) upper and lower envelopes (green) defined by the local maxima and minima, respectively, and the mean value of the upper and lower envelopes given in red.

The decomposition is based on the simple assumption that any data consists of different simple intrinsic modes of oscillations. Each intrinsic mode, linear or nonlinear, represents a simple oscillation, which will have the same number of extrema and zero-crossings. Furthermore, the oscillation will also be symmetric with respect to the “local mean.” At any given time, the data may have many different coexisting modes of oscillation, one superimposing on the others. The result is the final complicated data. Each of these oscillatory modes is represented by an intrinsic mode function (IMF) with the following definition:

- (1) in the whole dataset, the number of extrema and the number of zero-crossings must either equal or differ at most by one, and
- (2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

An IMF represents a simple oscillatory mode as a counterpart to the simple harmonic function, but it is much more general: instead of constant amplitude and frequency, as in a simple harmonic component, the IMF can have a variable amplitude and frequency as functions of time. With the above definition for the IMF, one can then decompose any function as follows: take the test data as given in Fig. 1.1; identify all the local extrema,

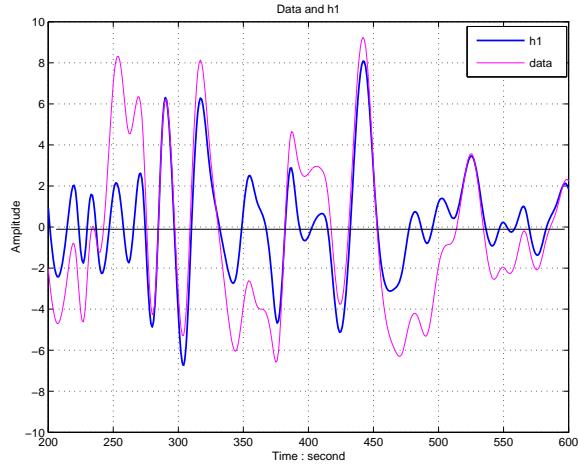


Figure 1.3: The data (red) and h_1 (blue).

then connect all the local maxima by a cubic spline line as shown in the upper envelope. Repeat the procedure for the local minima to produce the lower envelope. The upper and lower envelopes should cover all the data between them, as shown in Fig. 1.2. Their mean is designated as m_1 , also shown in Fig. 1.2, and the difference between the data and m_1 is the first component h_1 shown in Fig. 1.3; i. e.,

$$h_1 = x(t) - m_1. \quad (1.7)$$

The procedure is illustrated in Huang et al. (1998).

Ideally, h_1 should satisfy the definition of an IMF, for the construction of h_1 described above should have made it symmetric and have all maxima positive and all minima negative. However, even if the fitting is perfect, a gentle hump on a slope can be amplified to become a local extremum in changing the local zero from a rectangular to a curvilinear coordinate system. After the first round of sifting, the hump may become a local maximum. New extrema generated in this way actually reveal the proper modes lost in the initial examination. In fact, with repeated siftings, the sifting process can recover signals representing low-amplitude riding waves.

The sifting process serves two purposes: to eliminate riding waves, and to make the wave profiles more symmetric. While the first purpose must be achieved for the Hilbert transform to give a meaningful instantaneous frequency, the second purpose must also be achieved in case the neighboring wave amplitudes have too large a disparity. Toward these ends, the sifting

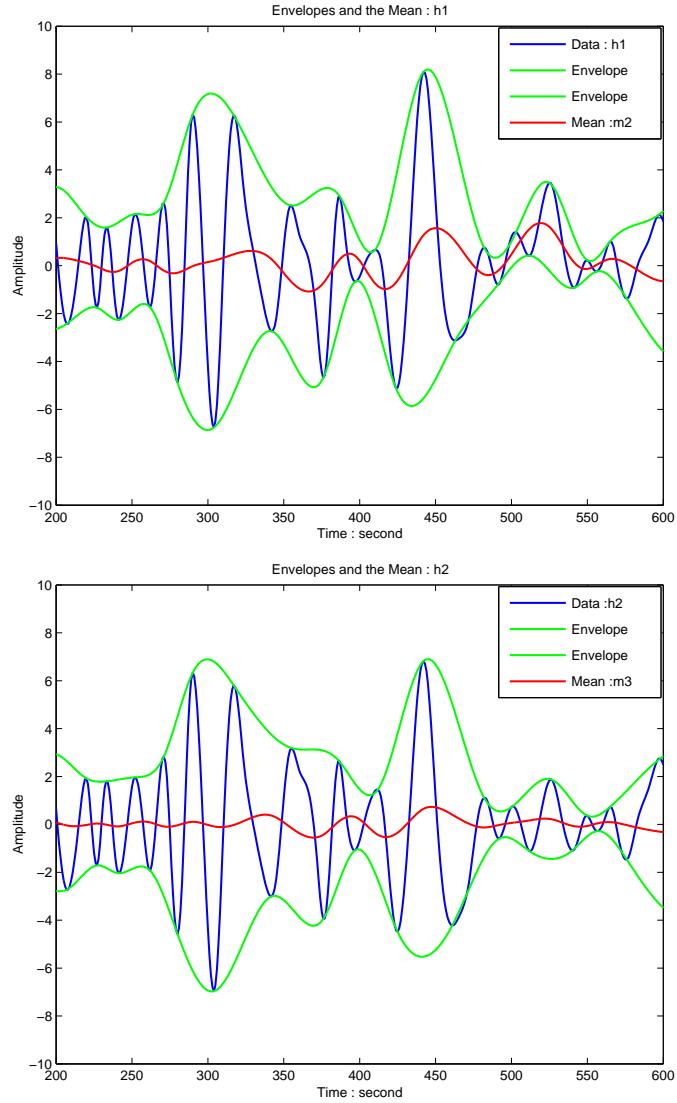


Figure 1.4: (a, top) Repeated sifting steps with h_1 and m_2 . (b, bottom) Repeated sifting steps with h_2 and m_3 .

process has to be repeated as many times as is required to reduce the extracted signal to an IMF. In the subsequent sifting processes, h_1 can be treated only as a proto-IMF. In the next step, it is treated as the data;

then,

$$h_{11} = h_1 - m_{11}. \quad (1.8)$$

After repeated siftings in this manner, shown in Fig. 1.4a,b, up to k times, h_{1k} becomes an IMF; that is,

$$h_{1k} = h_{1(k-1)} - m_{1k}; \quad (1.9)$$

then, it is designated as

$$c_1 = h_{1k}, \quad (1.10)$$

the first IMF component from the data shown in Fig. 1.5. Here, a critical decision must be made: the stoppage criterion. Historically, two different criteria have been used: The first one was used in Huang et al. (1998). This stoppage criterion is determined by using a Cauchy type of convergence test. Specifically, the test requires the normalized squared difference between two successive sifting operations defined as

$$SD_k = \frac{\sum_{t=0}^T |h_{k-1}(t) - h_k(t)|^2}{\sum_{t=0}^T h_{k-1}^2} \quad (1.11)$$

to be small. If this squared difference SD_k is smaller than a predetermined value, the sifting process will be stopped. This definition seems to be rigorous, but it is very difficult to implement in practice. Two critical questions need to be resolved: first, the question of how small is small enough needs an answer. Second, this criterion does not depend on the definition of the IMFs. The squared difference might be small, but nothing guarantees that the function will have the same numbers of zero-crossings and extrema, for example. These shortcomings prompted Huang et al. (1999, 2003) to propose a second criterion based on the agreement of the number of zero-crossings and extrema. Specifically, a S -number is pre-selected. The sifting process will stop only after S consecutive times, when the numbers of zero-crossings and extrema stay the same and are equal or differ at most by one. This second choice has its own difficulty: how to select the S number. Obviously, any selection is *ad hoc*, and a rigorous justification is needed.

In a recent study of this open-ended sifting, Huang et al. (2003) used the many possible choices of S -numbers to form an ensemble of IMF sets, from which an ensemble mean and confidence were derived. Furthermore, through comparisons of the individual sets with the mean, Huang et al. established an empirical guide. For the optimal siftings, the range of S -numbers should be set between 4 and 8. More details will be given later.

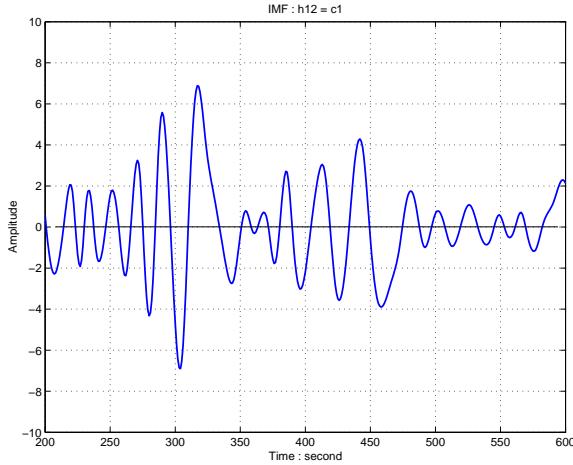


Figure 1.5: The first IMF component c_1 after 12 steps.

Now assume that a stoppage criterion was selected, and that the first IMF c_1 was found. Overall, c_1 should contain the finest scale or the shortest period component of the signal. It follows that c_1 can be separated from the rest of the data by

$$r_1 = x(t) - c_1. \quad (1.12)$$

Since the residue r_1 still contains longer period variations in the data, as shown in Fig. 1.6, it is treated as the new data and subjected to the same sifting process as described above. This procedure can be repeated with all the subsequent r_j 's, and the result is

$$\begin{aligned} r_2 &= r_1 - c_1 \\ &\vdots \\ r_n &= r_{n-1} - c_n \end{aligned} \quad (1.13)$$

The sifting process can be stopped finally by any of the following predetermined criteria: either when the component c_n or the residue r_n becomes so small that it is less than the predetermined value of substantial consequence, or when the residue r_n becomes a monotonic function from which no more IMFs can be extracted. Even for data with zero mean, the final residue still can be different from zero. If the data have a trend, the final residue should be that trend. By summing up (1.12) and (1.13), we finally

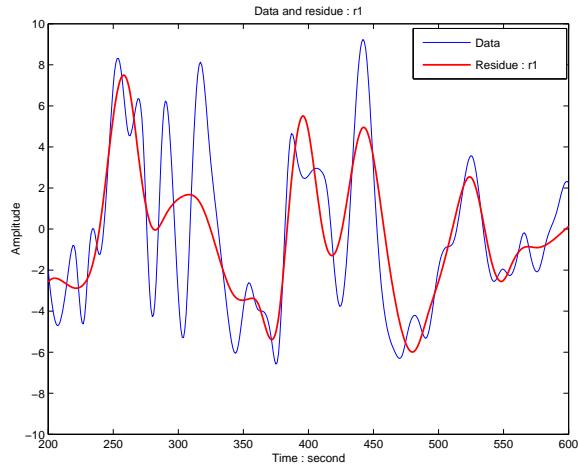


Figure 1.6: The original data (blue) and the residue r_1 .

obtain

$$x(t) = \sum_{j=1}^n c_j + r_n. \quad (1.14)$$

Thus, a decomposition of the data into n -empirical modes is achieved, and a residue r_n obtained which can be either the mean trend or a constant. As discussed here, to apply the EMD method, a mean or zero reference is not required; the EMD technique needs only the locations of the local extrema. The zero reference for each component will be generated by the sifting process. Without the need for the zero reference, EMD has the unexpected benefit of avoiding the troublesome step of removing the mean values for the large DC term in data with a non-zero mean.

The components of the EMD are usually physically meaningful, for the characteristic scales are defined by the physical data. To understand this point, consider the length-of-day data shown in Fig. 1.7, which measure the deviation of the rotational period from the fixed cycle of 24 h. The mean and the standard deviation of the IMFs, given in Fig. 1.8a,b, were obtained after using a different S -number for sifting. The sifting results are quite robust with respect to the selection of a stoppage criteria, as indicated by the low standard deviation values; thus, these IMF results are physically meaningful. The first component represents the very short period of perturbation caused by large-scale storms to the earth's rotational speed;

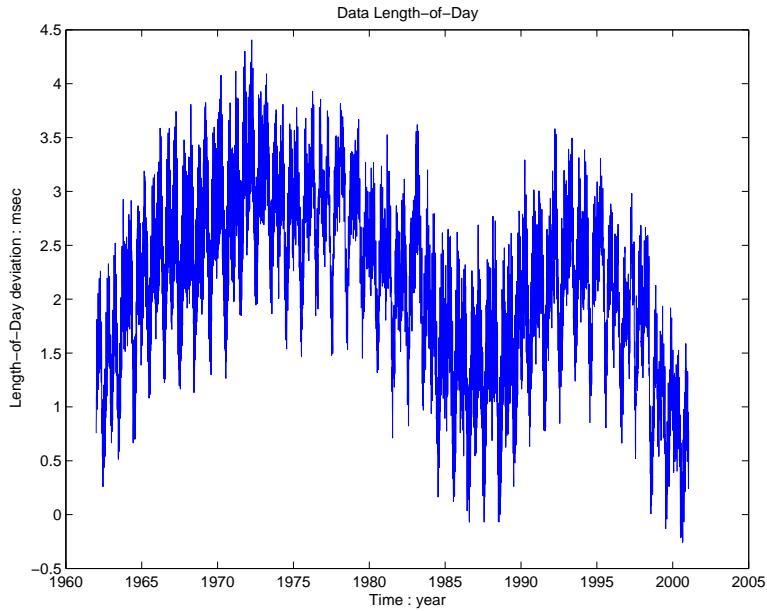


Figure 1.7: The length-of-day data.

this perturbation could be measured only after the 1990s by using many Global Positioning Satellites (GPS). The second component represents the half-monthly tides; the eighth component, the annual tidal variations. In fact, a plot of the annual variation by itself in Fig. 1.9 shows that the inter-annual variations are actually associated with the El Niño events. During an El Niño event, the equatorial water in the Pacific Ocean is warmed up, and this warming imparts more energy into the atmosphere. This result, in turn, causes the atmosphere to be more energetic. The resulting increase in angular momentum makes the rotational speed of the earth slow down. Even more surprisingly, the standard deviation values of the different siftings were unusually large during 1965 to 1970, and 1990 to 1995, the periods identified by NOAA as the anomaly periods for El Niño events.

This example established the physical meaning of the IMF components beyond any real doubt. Even more fundamental, the recent studies by Flandrin et al. (2004), Flandrin and Gonçalves (2004) and Wu and Huang (2004) further established the statistical significance of the IMF components. Thus, whether a given IMF contains significant information or just represents noise can now be tested.

1.2.2. The Hilbert spectral analysis

Having obtained the intrinsic mode function components, one will have no difficulty in applying the Hilbert transform to each IMF component, and in computing the instantaneous frequency according to (1.2)–(1.6). After performing the Hilbert transform on each IMF component, the original data can be expressed as the real part \Re in the following form:

$$x(t) = \Re \left\{ \sum_{j=1}^n a_j(t) \exp \left[i \int \omega_j(t) dt \right] \right\} \quad (1.15)$$

Here, the residue r_n has been left out on purpose, for it is either a monotonic function or a constant. Although the Hilbert transform can treat the monotonic trend as part of a longer oscillation, the energy involved in the residual trend representing a mean offset could be overpowering. In consideration of the uncertainty of the longer trend, and in the interest of obtaining the information contained in the other low-energy but clearly oscillatory components, the final non-IMF component should be left out. However, it could be included if physical considerations justify its inclusion.

Equation (1.15) gives both the amplitude and frequency of each component as functions of time. The same data expanded in a Fourier representation would be

$$x(t) = \Re \left[\sum_{j=1}^n a_j e^{i\omega_j(t)t} \right], \quad (1.16)$$

with both a_j and ω_j as constants. The contrast between (1.15) and (1.16) is clear: the IMF represents a generalized Fourier expansion. The variable amplitude and the instantaneous frequency have not only greatly improved the efficiency of the expansion, but also enabled the expansion to accommodate nonlinear and nonstationary data. With the IMF expansion, the amplitude and the frequency modulations are also clearly separated. Thus, the restriction of the constant amplitude and fixed frequency of the Fourier expansion has been overcome, with a variable amplitude and frequency representation. This frequency-time distribution of the amplitude is designated as the “Hilbert amplitude spectrum” $H(\omega, t)$, or simply “Hilbert spectrum.” If amplitude squared is the more preferred method to represent energy density, then the squared values of the amplitude can be substituted to produce the Hilbert energy spectrum just as well.

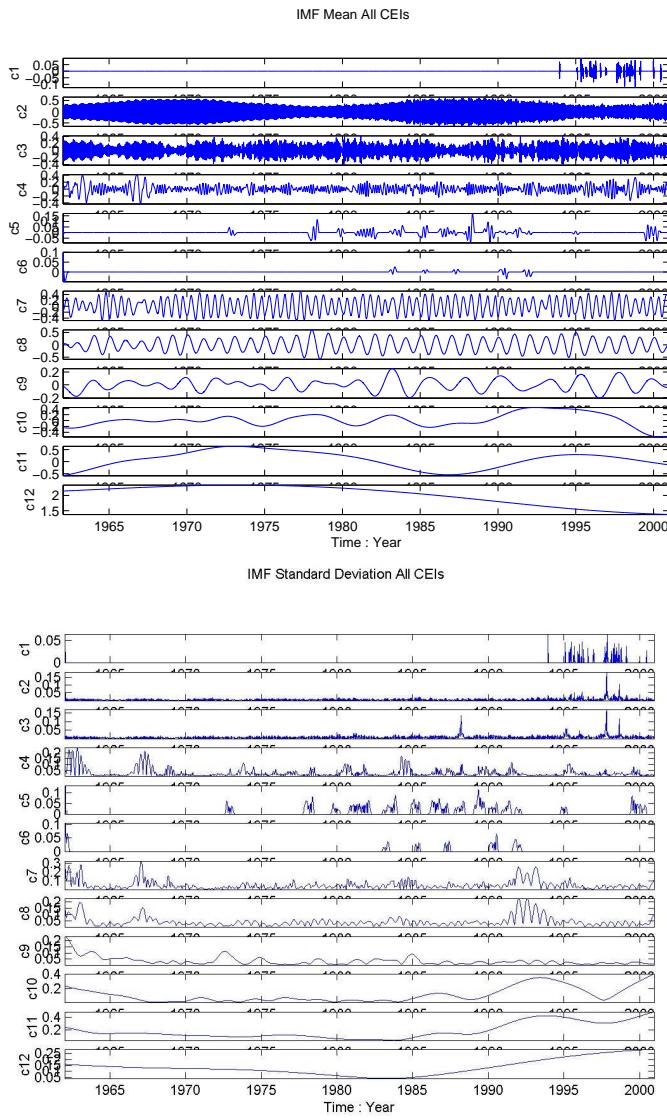


Figure 1.8: (a, top) The mean IMF for nine different siftings. (b, bottom) The standard deviation of the IMF for nine different siftings.

The skeleton Hilbert spectrum presentation is more desirable, for it gives more quantitative results. Actually, Bacry et al. (1991) and Carmona et al.

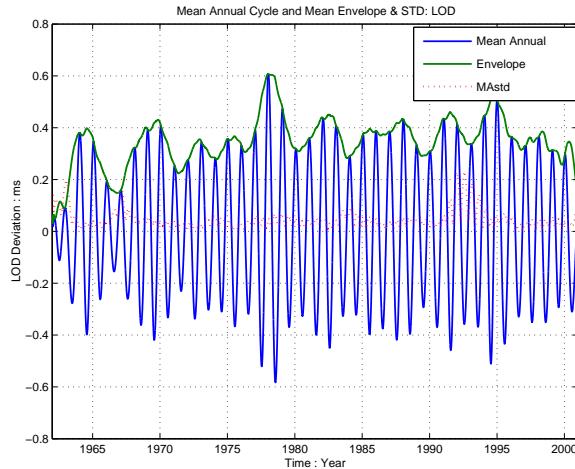


Figure 1.9: The mean annual cycle and its envelope. Each peak of the envelope coincides with an El Niño event.

(1998) have tried to extract the wavelet skeleton as the local maximum of the continuous wavelet coefficient. Even that approach is still encumbered by the harmonics. If more qualitative results are desired, a fuzzy representation can also be derived from the skeleton Hilbert spectrum presentation by using two-dimensional smoothing. The result is a smoother presentation of time-frequency distribution, but the spurious harmonics are still not needed.

With the Hilbert Spectrum defined, we can also define the marginal spectrum $h(\omega)$ as

$$h(\omega) = \int_0^T H(\omega, t) dt. \quad (1.17)$$

The marginal spectrum offers a measure of the total amplitude (or energy) contribution from each frequency value. This spectrum represents the accumulated amplitude over the entire data span in a probabilistic sense.

The combination of the empirical mode decomposition and the Hilbert spectral analysis is also known as the “Hilbert-Huang transform” (HHT) for short. Empirically, all tests indicate that HHT is a superior tool for time-frequency analysis of nonlinear and nonstationary data. It is based on an adaptive basis, and the frequency is defined through the Hilbert transform. Consequently, there is no need for the spurious harmonics to represent nonlinear waveform deformations as in any of the *a priori* basis methods,

and there is no uncertainty principle limitation on time or frequency resolution from the convolution pairs based also on *a priori* basis. A comparative summary of Fourier, wavelet and HHT analyses is given in the following table:

	Fourier	Wavelet	Hilbert
Basis	<i>a priori</i>	<i>a priori</i>	adaptive
Frequency	convolution: global uncertainty	convolution: regional uncertainty	differentiation: local, certainty
Presentation	energy- frequency	energy-time- frequency	energy-time- frequency
Nonlinear	no	no	yes
Nonstationary	no	yes	yes
Feature Extraction	no	discrete: no; continuous: yes	yes
Theoretical base	theory complete	theory complete	empirical

This table shows that the HHT is indeed a powerful method for analyzing data from nonlinear and nonstationary processes: it is based on an adaptive basis; the frequency is derived by differentiation rather than convolution; therefore, it is not limited by the uncertainty principle; it is applicable to nonlinear and nonstationary data and presents the results in time-frequency-energy space for feature extraction.

1.3. Recent developments

Some recent developments in the following areas will be discussed in some detail:

- (1) normalized Hilbert transform
- (2) confidence limit
- (3) statistical significance of the IMFs.

1.3.1. *Normalized Hilbert transform*

It is well known that although the Hilbert transform exists for any function of L^p class, the phase function of the transformed function will not always yield a physically meaningful instantaneous frequency, as discussed

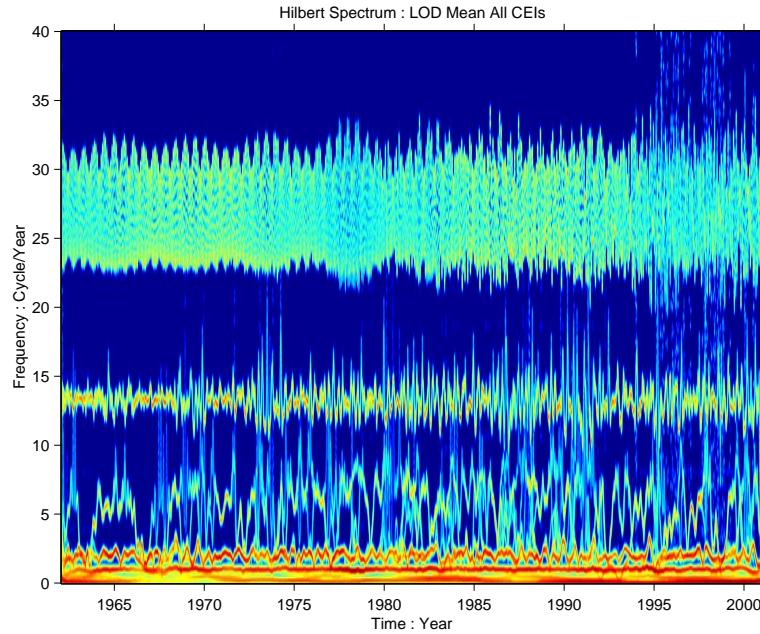


Figure 1.10: The mean Hilbert spectrum.

above. Reducing the function into IMFs has improved the chance of getting a meaningful instantaneous frequency, but obtaining IMFs satisfies only the necessary condition; additional limitations have been summarized succinctly in two additional theorems:

First, the Bedrosian theorem (1963) states that the Hilbert transform for the product of two functions $f(t)$ and $h(t)$ can be written as

$$\mathcal{H}[f(t)h(t)] = f(t)\mathcal{H}[h(t)], \quad (1.18)$$

only if the Fourier spectra for $f(t)$ and $h(t)$ are totally disjoint in frequency space, and the frequency range of the spectrum for $h(t)$ is higher than that of $f(t)$. This limitation is critical: if the instantaneous frequency is to be computed from the phase function as defined in (1.3)–(1.6), the data can be expressed in the IMF form as

$$x(t) = a(t) \cos[\theta(t)]; \quad (1.19)$$

then, the Hilbert transform will give us the conjugate part as

$$\mathcal{H}\{a(t) \cos[\theta(t)]\} = a(t)\mathcal{H}\{\cos[\theta(t)]\}. \quad (1.20)$$

However, according to the Bedrosian theorem, (1.20) can be true only if the amplitude is varying so slowly that the frequency spectra of the envelope and the carrier waves are disjoint. This condition has made the application of the Hilbert transform problematic. To satisfy this requirement, Huang and Long (2003) proposed that the IMFs be normalized as follows: start from the data that is already an IMF. First, find all the maxima of the IMFs; then, define the envelope by a spline through all the maxima, and designate the envelope as $E(t)$. Now, normalize the IMF by dividing it by $E(t)$ as

$$Co(t) = \frac{x(t)}{E(t)}, \quad (1.21)$$

where $Co(t)$ should be the carrier function with all local maxima equal to unity. The normalized function of the above example is given in Fig. 1.10d.

This construction should give an amplitude always equal to unity, but anomalies clearly exist, and complications can arise from the spline fitting, which mainly occurs at the point where the amplitude fluctuation is large. Then the spline line could go under the data momentarily and cause the normalized function to have an amplitude greater than unity. Though these conditions are rare, they can occur. Whenever they do, error will certainly occur, which will be discussed next. Even with a perfect normalization, not all of the problems have been solved. The next difficulty is given by the Nuttall theorem.

The Nuttall (1996) theorem states that the Hilbert transform of a cosine is not necessarily a simple 90° phase shift, resulting in the sine function with the same phase function for an arbitrary phase function. Nuttall gave an error bound ΔE , defined as the difference between the Hilbert transform C_h and the quadrature C_q (with a phase shift of exactly 90°) of the function as

$$\Delta E = \int_0^T |C_q(t) - C_h(t)|^2 dt = \int_{-\infty}^0 S_q(\omega) d\omega, \quad (1.22)$$

in which S_q is the Fourier spectrum of the quadrature function. Though the proof of this theorem is rigorous, the result is hardly useful: first, it is expressed in terms of the Fourier spectrum of a still unknown quadrature; and second, it gives a constant error bound over the whole data range. For a nonstationary time series, such a constant bound will not reveal the location of the error on the time axis. With the normalization, Huang and Long (2003) have proposed a variable error bound as follows: the error will

be the difference between the squared amplitude of the normalized signal and unity.

The proof of this concept is simple: if the Hilbert transform is exactly the quadrature, by definition the square amplitude will be unity, and the difference zero. If the squared amplitude is not exactly unity, then the Hilbert transform cannot be exactly the quadrature. Two possible complications can contribute to the errors: First, the normalization processes are not clean, as discussed above, so the normalized amplitude could exceed unity, and the error would not be zero. The second complication could come from a highly complicated phase function, as discussed in Huang et al. (1998); then the phase plan will not be a perfect circle. Any deviation from the circle will result in the amplitude being different from unity. Huang and Long (2003) and Huang et al. (2005) conducted detailed comparisons and found the result quite satisfactory. Alternatively, Huang et al. (2005) suggested that the phase function can be found by computing the inverse cosine of the normalized function. The results obtained in this manner were also found to be satisfactory. Two problems, however, still plagued this approach: first, any imperfect normalization will occasionally give the values of the normalized function greater than unity, as discussed above. Under that condition, the inverse cosine will break down. Second, the computation precision requirement is too high near the phase angle 0° and 180° . One can show that the problem of the normalized Hilbert transform always occurs at the location where the amplitude either changes drastically or is very low.

1.3.2. Confidence limit

In data analysis, a confidence limit is always necessary; it provides a measure of assurance about the legitimacy of the results. Therefore, the confidence limit for the Fourier spectral analysis is routinely computed, but the computation is based on ergodic theory, where the data are subdivided into N sections, with spectra from each section being computed. The confidence limit is determined from the statistical spread of the N different spectra. When all the conditions of ergodic theory are satisfied, the temporal average is treated as the ensemble average. Unfortunately, the ergodic condition is satisfied only if the processes are stationary; otherwise, averaging them will not make sense. Huang et al. (2003) have proposed a different approach by utilizing the existence of infinitely many ways to decompose one given function into difference components. Even using EMD, many different sets

of IMFs may be obtained by varying the stoppage criteria. For example, Huang et al. (2003) explored the stoppage criterion by changing the S -number. Using the length-of-day data, they varied the S -number from 1 to 20 and found the mean and the standard deviation for the Hilbert spectrum. The confidence limit so derived does not depend on the ergodic theory. If the same data length is used, the spectral resolution is not downgraded in frequency space through sub-dividing of the data into sections. Additionally, Huang et al. have also invoked the intermittence criterion and forced the number of IMFs to be the same for different S -numbers. As a result, Huang et al. were able to find the mean for the specific IMFs shown in Fig. 1.9. Of particular interest are the periods of high standard deviations, from 1965 to 1970, and 1990–1995. These periods are the anomaly periods of the El Niño phenomenon, when the sea-surface-temperature values in the equatorial region were consistently high based on observations, indicating a prolonged heating of the ocean, rather than the changes from warm to cool during the El Niño to La Niña changes.

Finally, from the confidence-limit study, an unexpected result is the determination of the optimal S -number. Huang et al. (2003) computed the difference between the individual cases and the overall mean and found that a range always exists where the differences reach a local minimum. Based on their limited experience from using different datasets, Huang et al. concluded that a S -number in the range of 4 to 8 performed well. Logic also dictates that the S -number should not be high enough to drain all the physical meaning out of the IMF, nor low enough to leave some riding waves remaining in the resulting IMFs.

1.3.3. Statistical significance of IMFs

EMD is a method for separating data into different components according to their scales. The question of the IMFs' statistical significance is always an issue. In data containing noise, how can the noise be separated confidently from the information? These questions were addressed by Flandrin et al. (2004), Flandrin and Gonçalvès (2004), and Wu and Huang (2004) through a study of noise only.

Flandrin et al. (2004) and Flandrin and Gonçalvès (2004) studied the fractional Gaussian noises and found that the EMD is a dyadic filter. These researchers also found that when one plotted the root-mean-squared (RMS) values of the IMFs as a function of the mean period derived from the fractional Gaussian noise on a log-log scale, the results formed a straight

line. The slope of the straight line for white noise is -1 ; however, the values change regularly with the different Hurst indices. Based on these results, Flandrin et al. (2004) and Flandrin and Gonçalvès (2004) suggested that the EMD results could be used to determine what kind of noise one was encountering.

Instead of fractional Gaussian noise, Wu and Huang (2004) studied the Gaussian white noise only. They also found the same relationship between the RMS values of the IMFs as a function of the mean period. Additionally, they also studied the statistical properties of the scattering of the data and found the bounds for the noise data distribution analytically. From the scattering, they deduced a 95% bound for the white noise. Therefore, they concluded that when a dataset is analyzed by using EMD, if the RMS-mean period values exist within the noise bounds, the components most likely represent noise. On the other hand, if the mean period-RMS values exceed the noise bounds, then those IMFs must represent statistically significant information. Thus, with the study of noise, Wu and Huang have found a way to discriminate noise from information. They applied this method to the Southern Oscillation Index (SOI) and concluded that the phenomena with the mean periods of 2.0, 3.1, 5.9 and 11.9 years are statistically significant signals.

1.4. Mathematical problems related to the HHT

Over the past few years, the HHT method has gained some recognition. Unfortunately, the full theoretical base has not been fully established. Up to this time, most of the progress with the HHT has been in its application, while the underlying mathematical problems have been mostly left untreated. All the results have come from case-by-case comparisons conducted empirically. The work with the HHT is presently at the stage corresponding historically to that of wavelet analysis in the earlier 1980s, producing great results but waiting for mathematical foundations on which to rest its case. The work is waiting for someone like Daubechies (1992) to lay the mathematical foundation for the HHT as was done for wavelets. The outstanding mathematical problems at the forefront at the present time are as follows:

- (1) Adaptive data analysis methodology in general
- (2) Nonlinear system identification methods
- (3) Prediction problem for nonstationary processes (end effects)
- (4) Spline problems (best spline implementation for the HHT, convergence and 2-D)

- (5) Optimization problems (the best IMF selection and uniqueness)
- (6) Approximation problems (Hilbert transform and quadrature)
- (7) Miscellaneous questions concerning the HHT.

1.4.1. Adaptive data-analysis methodology

Most data-analysis methods are not adaptive. The established approach is to define a basis (such as trigonometric functions in Fourier analysis, for example). Once the basis is determined, the analysis is reduced to a convolution computation. This well established paradigm is specious, for we have no *a priori* reason to believe that the basis selected truly represents the underlying processes. Therefore, the results produced will not be informative. This paradigm does, however, provide a definitive quantification with respect to a known metric for certain properties of the data based on the basis selected.

If one gives up this paradigm, no solid foundation remains, yet data-analysis methods need to be adaptive, for their goal is to find out the underlying processes. Only adaptive methods can let the data reveal their underlying processes without any undue influence from the basis. Unfortunately, no mathematical model or precedent exists for such an approach. Recently, adaptive data processing has gained some attention. Some adaptive methods are being developed (see Windrows and Stearns 1985). Unfortunately, most of the methods available depend on feedback; therefore, they are limited to stationary processes. Generalizing these available methods to nonstationary conditions is a difficult task.

1.4.2. Nonlinear system identification

System-identification methods are usually based on having both input and output data. For an ideally controlled system, such datasets are possible, yet for most of the cases studied, natural or man-made, no such luxury involving data is available. All that might be available is a set of measured results. The question is whether the nonlinear characteristics can be identified from the data. This problem might be ill-posed, for this is very different from the traditional input vs. output comparison. Whether the system can be identified through data only is an open question. Unfortunately, in most natural systems, control of the input is not possible. Additionally, the input and even the system itself are usually unknown. The only data available usually correspond to the output from an unknown system. Can the system be identified? Or short of identification, can anything be learned about the

system? The only thing that might be available is some general knowledge of the underlying controlling processes connected with the data. For example, the atmosphere and ocean are all controlled by the generalized equations for fluid dynamics and thermodynamics, which are nonlinear. The man-made structures, though linear under design conditions, will approach nonlinearity under extreme loading conditions. Such *a priori* knowledge could guide the search for the characteristics or the signatures of nonlinearity. The task, however, is still daunting.

So far, most of the definitions or tests for nonlinearity from any data are only necessary conditions: for example, various probability distributions, higher-order spectral analysis, harmonic analysis, and instantaneous frequency (see, for example, Bendat 1990; Priestly 1988; Tong 1990; Kantz and Schreiber 1997). Certain difficulties are involved in making such identifications from observed data only. This difficulty has made some scientists talk about only “nonlinear systems” rather than “nonlinear data.” Such reservations are understandable, but this choice of terms still does not resolve the basic problem: How to identify the system nonlinearity from its output alone. Is doing so possible? Or, is there a definite way to define a nonlinear system from the data (system output) at all? This problem is made even more difficult when the process is also stochastic and nonstationary. With a nonstationary process, the various probabilities and the Fourier-based spectral analyses are all problematic, for those methods are based on global properties, with linear and stationary assumptions.

Through the study of instantaneous frequency, intra-wave frequency modulation has been proposed as an indicator for nonlinearity. More recently, Huang (2003) identified the Teager energy operator (Kaiser 1990; Maragos et al. 1993a,b) as an extremely local and sharp test for harmonic distortions within any IMF derived from data. The combination of these local methods offers some hope for system identification, but the problem is not solved, for this approach is based on the assumption that the input is linear. Furthermore, all these local methods also depend on local harmonic distortion; they cannot distinguish a quasi-linear system from a truly nonlinear system. A test or definition for nonlinear-system identification based on only observed output is urgently needed.

1.4.3. *The prediction problem for nonstationary processes (the end effects of EMD)*

End effects have plagued data analysis from the beginning of any known method. The accepted and timid way to deal with these effects is by using various kinds of windowing, as is done routinely in Fourier analysis. Although sound in theory, such practices inevitably sacrifice some precious data near the ends. Furthermore, the use of windows becomes a serious hindrance when the data are short. In the HHT approach, the extension of data beyond the existing range is necessary, for a spline through the extrema is used to determine the IMF. Therefore, a method is needed to determine the spline curve between the last available extremum and the end of the data range. Instead of windowing, Huang et al. (1998) introduced the idea of using a “window frame,” a way to extend the data beyond the existing range in order to extract some information from all the data available.

The extension of data, or data prediction, is a risky procedure even for linear and stationary processes. The problem that must be faced is how to make predictions for nonlinear and nonstationary stochastic processes. Here the age-old cozy shelter of the linear, stationary, low-dimension and deterministic assumptions must be abandoned, and the complicated real world must be faced. The data are mostly from high-dimensional nonlinear and nonstationary stochastic systems. Are these systems predictable? What conditions must be imposed on the problem to make it predictable? How well can the accuracy of the predictions be quantified? In principle, data prediction cannot be made based on past data alone. The underlying processes have to be involved. Can the available data be used to extract enough information to make a prediction? This issue is an open question at present.

However, EMD has an advantage to assist the analysis: the whole data span need not be predicted, but only the IMF, which has a much narrower bandwidth, for all the IMFs should have the same number of extrema and zero-crossings. Furthermore, all that is needed is the value and location of the next extrema, not all the data. Such a limited goal notwithstanding, the task is still challenging.

1.4.4. *Spline problems (the best spline implementation for HHT, convergence and 2-D)*

EMD is a “Reynolds type” decomposition: it is used to extract variations from the data by separating the mean, in this case the local mean, from the

fluctuations by using spline fits. Although this approach is totally adaptive, several unresolved problems arise from this approach.

First, among all the spline methods, which one is the best? The answer to this question is critical, for it can be shown easily that all the IMFs other than the first are a summation of spline functions, for from (1.5) to (1.8), it follows that

$$c_1 = x(t) - (m_{1k} + m_{1(k-1)} + \cdots + m_{11} + m_1), \quad (1.23)$$

in which all m functions are generated by splines. Therefore, from equation (1.10),

$$r_1 = x(t) - c_1 = m_{1k} + m_{1(k-1)} + \cdots + m_{11} + m_1 \quad (1.24)$$

is totally determined by splines. Consequently, according to (1.11), all the rest of the IMFs are also totally determined by spline functions. What kind of spline is the best fit for the EMD? How can one quantify the selection of one spline vs. another? Based on experience, it was found that the higher-order spline functions needed additional subjectively determined parameters, yet the requirement violates the adaptive spirit of the approach. Furthermore, higher-order spline functions could also introduce additional length scales, and they are also more time-consuming in computations. Such shortcomings are why only the cubic spline was selected. However, the possible advantages and disadvantages of higher-order splines and even a taut spline have not been definitively established and quantified.

Finally, the convergence of the EMD method is also a critical issue: is there a guarantee that in finite steps, a function can always be reduced into a finite number of IMFs? All intuitive reasoning and experience suggest that the procedure is converging. Under rather restrictive assumptions, the convergence can even be proved rigorously. The restricted and simplified case studied involved sifting with middle-points only. With further restriction of the middle-point sifting to linearly connected extrema, the convergence proof can be established by *reductio ad absurdum*, and it can be shown that the number of extrema of the residue function has to be less than or equal to that in the original function. The case of equality exists only when the oscillation amplitudes in the data are either monotonically increasing or decreasing. In this case, the sifting may never converge and forever have the same number in the original data and the IMF extracted. The proof is not complete in another aspect: can one prove the convergence once the linear connection is replaced by the cubic spline? Therefore, this approach to the proof is not complete.

Recently, Chen et al. (2004) used a B-spline to implement the sifting. If one uses the B-spline as the base for sifting, then one can invoke the variation-diminishing property of the B-spline and show that the spline curve will have less extrema. The details of this proof still have to be established.

1.4.5. *The optimization problem (the best IMF selection and uniqueness mode mixing)*

Does the EMD generate a unique set of IMFs, or is the EMD method a tool to generate infinite sets of IMFs? From a theoretical point of view, infinitely many ways to decompose a given dataset are available. Experience indicates that the EMD process can generate many different IMF sets by varying the adjustable parameters in the sifting procedure. How are these different sets of IMF related? What is the criterion or criteria to guide the sifting? What is the statistical distribution and significance of the different IMF sets? Therefore, a critical question involves how to optimize the sifting procedure to produce the best IMF set. The difficulty is that it must not sift too many times and drain all the physical meaning out of each IMF component, and, at the same time, one must not sift too few times and fail to get clean IMFs. Recently, Huang et al. (2003) studied the problem of different sifting parameters and established a confidence limit for the resulting IMFs and Hilbert spectrum, but the study was empirical and limited to cubic splines only. Optimization of the sifting process is still an open question.

This question of the uniqueness of the IMF can be traced to this more fundamental one: how to define the IMF more rigorously? The definition given by Huang et al. (1998, 1999) is hard to quantify. Fortunately, the results are quite forgiving: even with the vague definition, the results produced are similar enough. Is it possible to give a rigorous mathematical definition and also find an algorithm that can be implemented automatically?

Finally, there is the problem of IMF mode rectifications. Straightforward implementation of the sifting procedure will produce mode mixing (Huang et al. 1999, 2003), which will introduce aliasing in the IMFs. This mode mixing can be avoided if an “intermittence” test is invoked (see Huang et al. 2003). At this time, one can implement the intermittence test only through interactive steps. An automatic mode rectification program should be able to collect all the relevant segments together and avoid the unnecessary

aliasing in the mode mixing. This step is not critical to the HHT, but would be a highly desirable feature of the method.

1.4.6. Approximation problems (the Hilbert transform and quadrature)

One of the conceptual breakthroughs involving the HHT has been the ability to define the instantaneous frequency through the Hilbert transform. Traditionally, two well-known theorems, the Bedrosian theorem (Bedrosian 1963) and the Nuttall theorem (Nuttall 1966), have considered the Hilbert transform to be unusable. The Bedrosian theorem states that the Hilbert transform for the product functions can be expressed only in terms of the product of the low-frequency function and the Hilbert transform of the high-frequency one, if the spectra of the two functions are disjointed. This condition guarantees that the Hilbert transform of $a(t) \cos[\theta(t)]$ is given by $a(t) \sin[\theta(t)]$. The Nuttall theorem (Nuttall 1966), further stipulates that the Hilbert transform of $\cos[\theta(t)]$ is not necessarily $\sin[\theta(t)]$ for an arbitrary function $\theta(t)$. In other words, a discrepancy exists between the Hilbert transform and the perfect quadrature of an arbitrary function $\theta(t)$. Unfortunately, the error bound given by Nuttall (1966) is expressed in terms of the integral of the spectrum of the quadrature, an unknown quantity. Therefore, the single valued error bound cannot be evaluated.

Through research, the restriction of the Bedrosian theorem has been overcome through the EMD process and the normalization of the resulting IMFs (Huang 2003). With this new approach, the error bound given by Nuttall has been improved by expressing the error bound as a function of time in terms of instantaneous energy. These developments are major breakthroughs for the Hilbert transform and its applications. However, the influence of the normalization procedure must be quantified. As the normalization procedure depends on a nonlinear amplification of the data, what is the influence of this amplification on the final results? Even if the normalization is accepted, for an arbitrary $\theta(t)$ function, the instantaneous frequency is only an approximation. How can this approximation be improved?

Also related to the normalization scheme, are other questions concerning the Hilbert transform: for example, what is the functional form of $\theta(t)$ for the Hilbert transform to be the perfect quadrature and also be analytic? If the quadrature is not identical to the Hilbert transform, what is the error bound in the phase function (not in terms of energy as it has been achieved now)?

One possible alternative is to abandon the Hilbert transform and to compute the phase function by using the inverse cosine of the normalized data. Two complications arise from this approach: the first one is the high precision needed for computing the phase function when its value is near $n\pi/2$. The second one is that the normalization scheme is only an approximation; therefore, the normalized functional value can occasionally exceed unity. Either way, some approximations are needed.

1.4.7. Miscellaneous statistical questions concerning HHT

The first question concerns the confidence limit of the HHT results. Traditionally, all spectral analysis results are bracketed by a confidence limit, which gives either a true or false measure of comfort. The traditional confidence limit is established from the ergodicity assumption; therefore, the processes are necessarily linear and stationary. If the ergodic assumptions are abandoned, can a confidence limit still exist without resorting to true ensemble averaging, which is practically impossible for most natural phenomena? The answer seems to be affirmative for Fourier analysis. For HHT, however, a confidence limit has been tentatively established, based on the exploitation of repeated applications of the EMD process with various adjustable parameters, which produces an ensemble of IMF sets. How representative are these different IMFs? How can the definition be made more rigorous? How can the statistical measure for such a confidence limit be quantified?

The second question concerns the degree of nonstationarity. This question has led to another conceptual breakthrough, for the qualitative definition of stationarity has been changed to a quantitative definition of the degree of nonstationarity. In Huang et al. (1998), in addition to a degree of nonstationarity, a degree of statistical nonstationarity was also given. For the degree of statistical nonstationarity, an averaging procedure is required. What is the time scale needed for the averaging?

1.5. Conclusion

Some of the problems encountered in the present state of the research have been discussed. Even though these issues have not been settled, the HHT method is still a very useful tool, but when they are settled, the HHT process will become much more rigorous, and the tool more robust. The author is using the HHT method routinely now, for as Heaviside famously said, when encountering the purist's objections on his operational calculus:

“Shall I refuse my dinner because I do not fully understand the process of digestion.” For us, the “the process of digestion” consists of fully addressing the questions that we have raised in this chapter. The path is clear; work must now begin.

Finally, the need for a unified framework for nonlinear and nonstationary data analysis is urgent and real. Currently, the field is fragmented, with partisans belonging to one camp or another. For example, researchers engaged in wavelet analysis will not mention the Wagner-Ville distribution method, as if it does not exist (see, for example, any wavelet book). On the other hand, researchers engaged with the Wagner-Ville distribution method will not mention wavelets (see, for example, Cohen 1995). Such a position is unscientific, and unhealthy for the data-analysis community. The time is right for some support from everyone to unify the field and push it forward. A concerted effort should be mounted to attack the problems of nonlinear and nonstationary time series analysis. One logical suggestion is to organize an activity group within SIAM to address all the mathematical and application problems, as well as all the scientific issues related to nonlinear and nonstationary data analysis. This task is worthy of the effort.

References

- Bacry, E., A. Arnéodo, U. Frisch, Y. Gagne, and E. Hopfinger, 1991: Wavelet analysis of fully developed turbulence data and measurement of scaling exponents. *Proc. Turbulence89: Organized Structures and Turbulence in Fluid Mechanics*, M. Lesieur, O. Métais, Eds., Kluwer, 203–215.
- Bedrosian, E., 1963: A product theorem for Hilbert transform. *Proc. IEEE*, **51**, 868–869.
- Bendat, J. S., 1990: *Nonlinear System Analysis and Identification from Random Data*. Wiley Interscience, 267 pp.
- Carmona, R., W. L. Hwang, and B. Torresani, 1998: *Practical Time-Frequency Analysis: Gabor and Wavelet Transform with an Implementation in S*. Academic Press, 490 pp.
- Chen, Q., N. E. Huang, S. Riemenschneider, and Y. Xu, 2005: A B-spline approach for empirical mode decomposition. *Adv. Comput. Math.*, in press.
- Cohen, L., 1995: *Time-Frequency Analysis*. Prentice Hall, 299 pp.
- Daubechies, I., 1992: *Ten Lectures on Wavelets*. CBMS-NSF Series in Applied Mathematics. Vol. 61, SIAM, 357 pp.

- Diks, C., 1999: *Nonlinear Time Series Analysis: Methods and Applications*. World Scientific Press, 180 pp.
- Flandrin, P., 1999: *Time-Frequency/Time-Scale Analysis*. Academic Press, 386 pp.
- Flandrin, P., G. Rilling, and P. Gonçalvès, 2004: Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.*, **11**, 112–114.
- Flandrin, P., and P. Gonçalvès, 2004: Empirical mode decompositions as data-driven wavelet-like expansions. *Int. J. Wavelets Multiresolut. Inform. Process.*, **2**, 477–496.
- Gröchenig, K., 2001: *Foundations of Time-Frequency Analysis*. Birkhäuser, 359 pp.
- Hahn, S., 1995: *Hilbert Transforms in Signal Processing*. Artech House, 442 pp.
- Huang, N. E., S. R. Long, and Z. Shen, 1996: The mechanism for frequency downshift in nonlinear wave evolution. *Adv. Appl. Mech.*, **32**, 59–111.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of water waves – The Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Huang, N. E., 2003: Empirical mode decomposition for analyzing acoustic signal. US Patent 10-073857, pending.
- Huang, N. E., and S. R. Long, 2003: Normalized Hilbert transform and instantaneous frequency. NASA Patent Pending GSC 14,673-1.
- Huang, N. E., M. C. Wu, S. R. Long, S. S. P. Shen, W. Qu, P. Gloersen, and K. L. Fan, 2003: A confidence limit for empirical mode decomposition and Hilbert spectral analysis. *Proc. R. Soc. London, Ser. A*, **459**, 2317–2345.
- Kaiser, J. F., 1990: On Teager's energy algorithm and its generalization to continuous signals. *Proc. 4th IEEE Signal Process. Workshop*, Sept. 16–19, Mohonk, NY, IEEE, 230 pp.
- Kantz, H., and T. Schreiber, 1999: *Nonlinear Time Series Analysis*. Cambridge University Press, 304 pp.
- Maragos, P., J. F. Kaiser, and T. F. Quatieri, 1993a: On amplitude and frequency demodulation using energy operators. *IEEE Trans. Signal Process.*, **41**, 1532–1550.
- Maragos, P., J. F. Kaiser, and T. F. Quatieri, 1993b: Energy separation in signal modulation with application to speech analysis. *IEEE Trans.*

- Signal Process.*, **41**, 3024–3051.
- Nuttall, A. H., 1966: On the quadrature approximation to the Hilbert transform of modulated signals. *Proc. IEEE*, **54**, 1458–1459.
- Plancherel, M., 1933: Sur les formules de réciprocité du type de Fourier. *J. London Math. Soc.*, Ser. 1, **8**, 220–226.
- Priestley, M. B., 1988: *Nonlinear and Nonstationary Time Series Analysis*. Academic Press, 237 pp.
- Titchmarsh, E. C., 1950: *Introduction to the Theory of Fourier Integral*. Oxford University Press, 394 pp.
- Tong, H., 1990: *Nonlinear Time Series Analysis*. Oxford University Press, 564 pp.
- Windrows, B., and S. D. Stearns, 1985: *Adaptive Signal Processing*. Prentice Hall, 474 pp.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.

Norden E. Huang

Goddard Institute for Data Analysis, Code 614.2, NASA/Goddard Space Flight Center, Greenbelt, MD 20771, USA
norden.e.huang@nasa.gov

CHAPTER 2

B-SPLINE BASED EMPIRICAL MODE DECOMPOSITION

Sherman Riemenschneider, Bao Liu, Yuesheng Xu and Norden E. Huang

This paper discusses some mathematical issues related to empirical mode decomposition (EMD). A B-spline EMD algorithm is introduced and developed for the convenience of mathematical studies. The numerical analysis using both simulated and practical signals and application examples from vibration analysis indicate that the B-spline algorithm has a comparable performance to that of the original EMD algorithm. It is also demonstrated that for white noise, the B-spline algorithm acts as a dyadic filter bank. Our mathematical results on EMD include Euler splines as intrinsic mode functions, the Hilbert transform of B-splines, and the necessary and sufficient conditions which ensure the validity of the Bedrosian identity of the Hilbert transform of product functions.

2.1. Introduction

Information processing is important in both pure research and practical applications. More often than not, information is embedded in data corrupted by noise. With the rapid development of science and technology, we are now inundated with a voluminous amount of data every day. The data need to be processed to extract meaningful information for various applications. Fourier transform is the most used and powerful traditional technique of data analysis. However, this technique is not effective in processing non-stationary and nonlinear signals because the basis functions it uses are not localized and cannot properly characterize the spectrum evolution in time. Time-frequency analysis is considered the major approach for overcoming the limitations of the traditional techniques. It aims at representing a signal with a joint function of both time and frequency, thereby providing a revealing picture in the time-frequency domain that allows the investigation of the non-stationary and nonlinear characteristics of a signal.

For over fifty years, researchers have put significant effort into seeking effective and efficient ways of time-frequency representation. Short-time

Fourier transform (Gabor 1946; Cohen 1995; Qian 1996), a representative achievement of this effort, is powerful in various applications. Unfortunately, it has difficulties in processing some types of signals such as those composed of small bursts plus quasi-stationary components because its time and frequency resolution is fixed. Although a number of improved methods have been developed by adapting the window size to the local composition of a signal (cf., Jones and Park 1990), these methods are either computationally expensive or effective only for certain specific applications. The Wigner-Ville distribution (Cohen 1995; Qian 1996) is another classical technique. It can well preserve the time and frequency concentration of a signal but has the drawback of cross-term interference, which often obscures the useful pattern over the time-frequency plane. To overcome this problem, various techniques (Choi and Williams 1989; Zhao et al. 1990), have been developed based on Cohen's general framework, (Cohen 1966) with the aim of reducing the cross-term interference while retaining the desirable properties of Wigner-Ville distribution. The recent advances of wavelet analysis opened a new path for time-frequency analysis. A significant breakthrough of wavelet analysis was the use of multi-scales to characterize signal events. This technique has led to the development of several wavelet-based time-frequency analysis techniques (Daubechies 1992; Mallat 1998). Some of them are adaptive, such as wavelet packets (Coifman et al. 1992), matching pursuit (Mallat and Zhang 1993), and basis pursuit (Chen et al. 2001). Despite the great success of numerous applications, all the above-mentioned techniques, however, each have their own limitations, and almost all of them are ineffective for characterizing the detailed time-frequency composition of nonlinear signals. On the other hand, most of the existing nonlinear time series analysis methods (Diks 1999) are designed only for stationary systems.

Over the past several decades, researchers have been attempting to exploit the concept of instantaneous frequency derived from analytic signals to construct a time-frequency representation. It is expected that such a representation may not have the fundamental drawbacks in the above-mentioned techniques. However, due to the existence of some paradoxes involving the instantaneous frequency derived this way and the concept of frequency in Fourier analysis or in our intuition (Cohen 1995), this effort did not have much success until Huang et al. (1998, 1999) developed the empirical mode decomposition (EMD) method. This method is now known as Hilbert-Huang transform (HHT) in the literature. It provides a powerful tool for improving the time-frequency technology. The basic approach of the EMD method is to decompose a signal into a collection of intrinsic mode

functions (IMF) that allow well-behaved Hilbert transforms for computation of physically meaningful instantaneous frequencies. This consequently makes it possible to construct a time-frequency representation, known as the “Hilbert spectrum,” by using instantaneous frequency. Another breakthrough is that, unlike wavelet analysis, which characterizes the scale of a signal event with pre-specified basis functions, HHT decomposes a signal by direct extraction of the local energy associated with the intrinsic time scales of the signal itself. The Hilbert spectrum, therefore, can depict not only the inter- but also the intra-wave time-frequency characteristics of a local event. Thus, HHT is applicable to both non-stationary and nonlinear signals.

HHT has been proved remarkably effective in various applications (Echeverria et al. 2001; Pines and Salvino 2002; Zhang et al. 2003). However, most of the underlying mathematical problems has still not been treated. As HHT finds wider and wider applications, the need for a rigorous mathematical foundation becomes more urgent. Recently, in the attempt to circumvent the mathematical difficulties, Chen et al. (2004) developed a variation of the original EMD, which appears potentially more convenient for an analytical formulation of EMD and for the study of certain related mathematical issues. This method represents the local mean of a signal by using the moving averages of the extrema as combinations of B-splines and avoids the use of envelopes, for which a proper mathematical definition is still an unsolved issue. In addition, this method overcomes the problem that, in implementation of the original EMD, the upper and lower envelopes may cross. Recently, we applied this method in an analysis of vibration signals for equipment fault diagnosis (Liu et al. 2004). The results showed that this method had a comparable performance to that of the original EMD.

The present paper reviews the work on the B-spline EMD (BS-EMD) based on Chen et al. (2004), Liu et al. (2004), and Xu and Yan (2004). In the next section, the algorithm of BS-EMD is reviewed. Then in section 2.3, some related mathematical results are presented, including Euler splines as a prototypical examples of B-spline IMF's, the properties of the Hilbert transform of B-splines, and the necessary and sufficient conditions which ensure the validity of the Bedrosian identity of the Hilbert transform of product functions. Section 2.4 addresses the properties of the BS-EMD as a filter bank and investigates the performance of the BS-EMD through comparison with the original EMD. In section 2.5, we present some application examples of the BS-EMD method and its corresponding Hilbert spectrum in vibration signal analysis. Finally, in section 2.6, the conclusion

is presented and future research problems are outlined.

2.2. A B-spline algorithm for empirical mode decomposition

The empirical mode decomposition method decomposes a signal into a finite sum of intrinsic mode functions that allow the computation of a physically meaningful instantaneous frequency defined in terms of the Hilbert transform. To describe the B-spline algorithm for empirical mode decomposition, we first recall the definition of the Hilbert transform. For a real signal $s(t)$, the Hilbert transform is defined by the principal value (PV) integral

$$\mathcal{H}s(t) := \frac{1}{\pi} \text{PV} \int_{-\infty}^{\infty} \frac{s(\tau)}{t - \tau} d\tau. \quad (2.1)$$

This yields an analytic signal

$$Z(t) = s(t) + i\mathcal{H}s(t) = a(t)e^{i\theta(t)}, \quad (2.2)$$

with

$$a(t) = \sqrt{s(t)^2 + \mathcal{H}s(t)^2}, \quad \theta(t) = \arctan\left[\frac{\mathcal{H}s(t)}{s(t)}\right], \quad (2.3)$$

where $a(t)$ and $\theta(t)$ are the amplitude and phase, respectively, of the signal at time t . The instantaneous frequency can be computed by

$$\omega(t) = \frac{d\theta(t)}{dt}. \quad (2.4)$$

Note that both the amplitude and instantaneous frequency in the above are functions of time. We naturally hope to construct a time-frequency representation by using the Hilbert transform. If this representation were correct for any arbitrary signal, we would have a time-frequency representation with a higher energy concentration than that of the short-time Fourier transform and wavelet transform, and a clearer representation pattern than that of the Wigner-Ville distribution. Unfortunately, a significant difficulty is that the instantaneous frequency obtained in this way may have frequency values which are meaningless in physics. This problem has plagued researchers for many years because the attempts to overcome it have been based on classical Fourier methods and filter theory.

The development of the EMD is a method for solving the above-mentioned problem. Huang et al. (1998) found that to define a meaningful instantaneous frequency for a function by using the Hilbert transform, the function has to satisfy the following two conditions:

- (1) The number of the extrema and the number of the zero crossings of the function must be equal or differ at most by one.
- (2) At any point of the function, the mean value of the envelopes defined by the local extrema should be zero.

Such a function is called the “intrinsic mode function” (IMF). Since the IMFs defined in this way admit computation of meaningful instantaneous frequencies, it becomes possible to construct a time-frequency representation based on the Hilbert transform.

The EMD algorithm provides a method to obtain the IMFs with the basic idea being the removal of the local mean from a signal by using a sifting process. In the original EMD, the local mean is computed as the mean value of the upper and lower envelopes. The envelopes themselves are approximated as the cubic spline interpolant of the local maxima and local minima, respectively. It can be seen that “envelopes” play a crucial role in this algorithm. However, a good mathematical description of envelopes remains an unsolved issue. For the convenience of studying the mathematical foundation of the EMD method, a more direct link from the local extrema to the mean in the sifting process is desirable. This need led to the development of the B-spline algorithm of EMD (Chen et al. 2004). It uses the moving average of the extrema as combinations of B-splines and is very amenable to mathematical study. We present the basic idea of the BS-EMD from Chen et al. (2004) in the following.

We first define the B-splines of order k for an arbitrary knot sequence (De Boor 1978). For a given increasing sequence $\tau_j, j \in Z$, the j th B-spline of order k is defined by the k th order divided difference, $[\tau_j, \dots, \tau_{j+k}]$, at the $k+1$ points $\tau_j, \dots, \tau_{j+k}$ applied to the truncated power as a function of x

$$B_{j,k,\tau}(t) := (\tau_{j+k} - \tau_j) [\tau_j, \dots, \tau_{j+k}] (x - t)_+^{k-1}, \quad t \in \mathbb{R}, \quad (2.5)$$

where $(x - t)_+^{k-1}$ is zero if $x < t$ and equals $(x - t)^{k-1}$ if $x \geq t$. These B-splines form a basis for the space of splines of order k with knots $\tau_j, j \in Z$. In other words, for any function s in this space, there exist unique scalars a_j such that

$$s(t) = \sum_{j \in Z} a_j B_{j,k,\tau}(t), \quad t \in \mathbb{R}. \quad (2.6)$$

B-splines satisfy the following recursive formula

$$B_{j,k,\tau}(t) = \frac{t - \tau_j}{\tau_{j+k-1} - \tau_j} B_{j,k-1,\tau}(t) + \frac{\tau_{j+k} - t}{\tau_{j+k} - \tau_{j+1}} B_{j+1,k-1,\tau}(t), \quad (2.7)$$

with

$$B_{j,1,\tau}(t) = \begin{cases} 1, & \tau_j \leq t < \tau_{j+1}, \\ 0, & \text{elsewhere.} \end{cases} \quad (2.8)$$

They are normalized so that

$$\sum_{j=p}^q B_{j,k,\tau}(t) = 1, \quad q \geq p+k, \quad (2.9)$$

on $[\tau_{p+k-1}, \tau_{q+1}]$.

For use in EMD, we define an operator on a given signal s as follows: The knots $\tau^s := \{\tau_j : j \in Z\}$ are taken as the extreme points of s . The knots in the support of the B-spline B_{j,k,τ^s} of order k are τ_{j+l} , $l = 0, 1, \dots, k$ with B_{j,k,τ^s} vanishing outside (τ_j, τ_{j+k}) and strictly positive inside that interval (hence, at $\tau_{j+1}, \dots, \tau_{j+k-1}$). At these extreme points strictly inside the support of B_{j,k,τ^s} , the following linear functionals are defined:

$$\lambda_{j,k,\tau^s} : s \mapsto \frac{1}{2^{k-2}} \sum_{l=1}^{k-1} \binom{k-1}{l} s(\tau_{j+l}). \quad (2.10)$$

This definition is a binomial average of the extrema contained in the support of B_{j,k,τ^s} , where more weight is given toward the center of the support. We take

$$V_{\tau^h,k} s := \sum_{j \in Z} \lambda_{j,k,\tau^s}(s) B_{j,k,\tau^s} \quad (2.11)$$

as the operator to replace the mean envelope in the original EMD algorithm.

As particular examples, when $k = 3$, we have the quadratic B-spline approximation

$$V_{\tau^s,3} s := \sum_{j \in Z} \frac{1}{2} [s(\tau_{j+1}) + s(\tau_{j+2})] B_{j,3,\tau^s}, \quad (2.12)$$

and when $k = 4$, we have the cubic B-spline approximation

$$V_{\tau^s,4} s := \sum_{j \in Z} \frac{1}{4} [s(\tau_{j+1}) + 2s(\tau_{j+2}) + s(\tau_{j+3})] B_{j,4,\tau^s}. \quad (2.13)$$

Using the operator defined above, we obtain the B-spline algorithm for the empirical mode decomposition. This algorithm extracts the first IMF of a signal s by using the following sifting process:

- (1) Find the local extrema of s .
- (2) Apply the operator $V_{\tau^h,k}$ in (2.11) to s .
- (3) Compute $h = s - V_{\tau^h,k} s$.

- (4) If h is an IMF, stop. Otherwise, treat h as the signal and iterate on h through Steps 1 to 4.

Denote by c_1 the first IMF and set $r_1 = s - c_1$, the first residue. The algorithm proceeds to select the next IMF by applying the above procedure to the first residue r_1 . This process is repeated until the last residue r_n has at most one extremum (excluding the ends) or becomes constant. The original signal then can be represented as

$$s = \sum_{j=1}^n c_j + r_n. \quad (2.14)$$

As in Huang et al. (1998), the stopping condition in Step 4 is to limit the following standard deviation from two consecutive results in the sifting process:

$$SD = \sum_{t=0}^T \frac{[h_{m-1}(t) - h_m(t)]^2}{h_{m-1}^2(t)}, \quad (2.15)$$

where T is the length of the signal, and h_m , the sifting result in the m th iteration. A typical stopping value of SD is set between 0.2 and 0.3.

Our numerical studies show that after a finite number of iterations, the sifting result of the BS-EMD will be an IMF. After applying the Hilbert transform on each IMF, we obtain an analytic signal corresponding to the signal $s(t)$

$$z(t) = \sum_{j=1}^n a_j(t) \exp\left[i \int_{-\infty}^t \omega_j(\tau) d\tau\right] + r_n(t). \quad (2.16)$$

The first term on the right side of (2.16) can be considered as a generalization of the Fourier expansion. It differs from the latter in that the components in (2.16) have a variable amplitude and frequency. This allows the amplitude and the frequency modulation to be separated and makes possible the conversion of the signal to a joint function of time and frequency. The residue r_n characterizes the trend of the signal and can be treated separately.

2.3. Some related mathematical results

In this section, we review some mathematical results related to IMFs, Hilbert transforms and the Bedrosian identity in the context of EMD. Most of the results discussed in this section come from Chen et al. (2004) and Xu and Yan (2004).

IMFs obtained from the EMD algorithm admit a physically meaningful instantaneous frequency. However, the mathematical definition and characterization of IMFs remain unsolved. We present an illuminating mathematical example of IMFs which is very insightful for further development in the mathematical characterizations of IMFs. An interesting observation is that the Euler splines of Schoenberg are an important class of spline type intrinsic mode functions which tend to the simplest IMF, the harmonic signal, when their degree tends to infinity.

We now recall the definition of Euler polynomials and Euler splines. The Euler polynomial of degree zero is defined by $p_0 = 1$, and Euler polynomials of higher degree are defined recursively by

$$p'_n(t) := np_{n-1}(t), \quad t \in \mathbb{R}, \quad n > 1 \quad (2.17)$$

under the constraint conditions $p_n(0) = -p_n(1)$. Specifically, the Euler polynomials of lower order are

$$p_1(t) = t - \frac{1}{2}, \quad p_2(t) = t^2 - t, \quad p_3(t) = t^3 - \frac{3}{2}t^2 + \frac{1}{4}, \quad (2.18)$$

$$p_4(t) = t^4 - 2t^3 + t, \quad p_5(t) = t^5 - \frac{5}{2}t^4 + \frac{5}{2}t^2 - \frac{1}{2}, \quad p_6(t) = t^6 - 3t^5 + 5t^3 - 3t. \quad (2.19)$$

The simplest Euler spline is the piecewise constant periodic function

$$E_0(t) = \text{sgn}[\sin(\pi t)]. \quad (2.20)$$

In general, by using Euler polynomials, Euler splines are generated as follows: Define splines E_n on \mathbb{R} as extensions of the Euler polynomials p_n on $[0, 1]$ to all of \mathbb{R} via the functional equation

$$E_n(t+1) = -E_n(t), \quad t \in \mathbb{R}. \quad (2.21)$$

That is,

$$E_n(t) := (-1)^{[t]} p_n(t - [t]), \quad t \in \mathbb{R}. \quad (2.22)$$

Schoenberg's Euler splines \mathcal{E}_n are defined to be

$$\mathcal{E}_n(t) := \begin{cases} E_n(t)/E_n(0), & \text{if } n \text{ is odd;} \\ E_n(t - 1/2)/E_n(1/2), & \text{if } n \text{ is even.} \end{cases} \quad (2.23)$$

Note that Euler splines are 2-periodic piecewise polynomials.

Figure 2.1 shows that Euler splines are non-stationary signals since their frequencies vary with time. However, the instantaneous frequencies of Euler splines tend rather rapidly to π when the order of the splines tends to infinity. This observation can be explained by the interesting results of

Schoenberg (1964, 1972, 1976, 1983) and Golitschek (1972), which state that,

$$\lim_{n \rightarrow \infty} \mathcal{E}_n(t) = \cos(\pi t), \text{ uniformly for } t \in \mathbb{R}, \quad (2.24)$$

and

$$\lim_{n \rightarrow \infty} \mathcal{E}_{2n-1}^{(k)}(t) = \cos^{(k)}(\pi t), \text{ uniformly for } t \in \mathbb{R}, \text{ for each } k \in \mathbb{N}. \quad (2.25)$$

In fact, from (6.10) in Schoenberg (1976), the first of these can be quantified as

$$\mathcal{E}_n(t) = \frac{\sum_{\nu=0}^{\infty} \frac{(-1)^{\nu(n+1)}}{(2\nu+1)^{n+1}} \cos[\pi(2\nu+1)t]}{\sum_{\nu=0}^{\infty} \frac{(-1)^{\nu(n+1)}}{(2\nu+1)^{n+1}}} = \cos(\pi t) + O[3^{-(n+1)}]. \quad (2.26)$$

By the absolute convergence of the series in (2.26), it also follows easily that

$$\lim_{n \rightarrow \infty} \mathcal{H}\mathcal{E}_n(t) = \sin(\pi t), \text{ uniformly for } t \in \mathbb{R}, \quad (2.27)$$

with the same error bound. Thus,

$$\lim_{n \rightarrow \infty} \left\{ [\mathcal{E}_n(t)]^2 + [\mathcal{H}\mathcal{E}_n(t)]^2 \right\} = 1, \quad t \in \mathbb{R}. \quad (2.28)$$

B-splines play a crucial role in the BS-EMD algorithm since except for the first IMF, all the others are linear combinations of B-splines. Furthermore, cubic splines used for interpolation of the envelopes in the original EMD can also be written as linear combinations of B-splines. Since time-frequency representation is obtained from the Hilbert transform of each IMF, it is desirable to consider the Hilbert transform of B-splines. Recursive formulas of the Hilbert transform of B-splines were established in Chen et al. (2004). We now review that paper's main results regarding the recursive formulas.

We recall the definition of equally spaced cardinal B-splines. Let χ_I denote the characteristic function of the interval I . We set $B_1 = \chi_{[0,1]}$ and cardinal B-splines of higher orders are defined recursively by convolution with B_1 ; that is,

$$B_n = B_{n-1} * B_1 = B_1 * B_{n-1}, \quad n = 2, 3, \dots \quad (2.29)$$

It can be seen that B_n is a spline of order n with the knots at integers $j = 0, 1, \dots, n$. Cardinal B-splines enjoy many nice properties (cf., De Boor

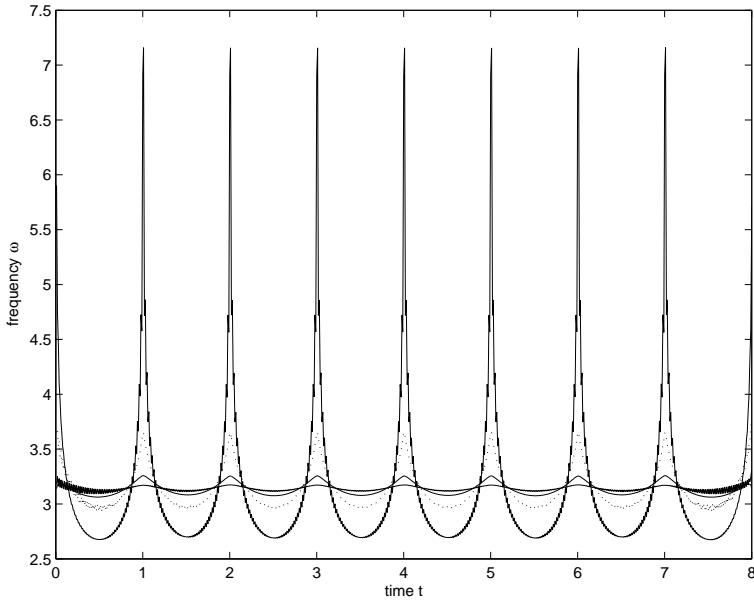


Figure 2.1: The comparisons of instantaneous frequency of Euler splines: The curve with the highest variation is the instantaneous frequency of \mathcal{E}_2 . The curve of dotted line denotes that of \mathcal{E}_3 . The remaining two curves of solid lines are the instantaneous frequency of \mathcal{E}_4 and \mathcal{E}_5 . The variation of the instantaneous frequency decreases from \mathcal{E}_2 to \mathcal{E}_5 , and the instantaneous frequency of \mathcal{E}_5 is already very close to π .

1978). If we let ∇ denote the backward difference operator defined recursively by

$$\nabla f(\cdot) = f(\cdot) - f(\cdot - 1) \quad \text{and} \quad \nabla^j = \nabla (\nabla^{j-1}), \quad (2.30)$$

the derivative of the B-splines have the form

$$\frac{d^j}{dt^j} B_n(t) = \nabla^j B_{n-j}(t), \quad t \in \mathbb{R}. \quad (2.31)$$

The B-splines satisfy the recursive formula

$$B_n(t) = \frac{t}{n-1} B_{n-1}(t) + \frac{n-t}{n-1} B_{n-1}(t-1), \quad t \in \mathbb{R}. \quad (2.32)$$

Such a recursive formula makes B-splines very convenient to use in applications. We also know that the cardinal B-splines are symmetric about the center of their support; i.e.,

$$B_n\left(\frac{n}{2} + t\right) = B_n\left(\frac{n}{2} - t\right), \quad t \in \mathbb{R}. \quad (2.33)$$

We will see that these properties carry over to the Hilbert transform of B-splines. Using the recursive definition of the cardinal B-splines and the properties of the Hilbert transform, we have that

$$\mathcal{H}B_n(t) = \mathcal{H}B_{n-1} * B_1 = \mathcal{H}B_1 * B_{n-1}, \quad (2.34)$$

where $\mathcal{H}B_1$ has a specific expression given by

$$\mathcal{H}B_1(t) = \frac{1}{\pi} \ln \left| \frac{t}{t-1} \right|. \quad (2.35)$$

For the derivative of $\mathcal{H}B_1$, we obtain the formula

$$\frac{d}{dt} \mathcal{H}B_1(t) = \frac{1}{\pi t} - \frac{1}{\pi(t-1)}. \quad (2.36)$$

This formula can be generalized to the high order derivative of the Hilbert transform of B-splines of a higher order. In the following theorem, we show that the j -th derivative of the Hilbert transform of B_n is the j -th backward difference of $\mathcal{H}B_{n-j}$.

Theorem 2.1: *Let n be a positive integer. Then, the following statements hold.*

(i) *For any positive integer j with $j \leq n$,*

$$\frac{d^j}{dt^j} (\mathcal{H}B_n)(t) = \nabla^j \mathcal{H}B_{n-j}(t). \quad (2.37)$$

(ii) *The Hilbert transform of cardinal B-splines have a recursive formula*

$$\mathcal{H}B_n(t) = \frac{t}{n-1} \mathcal{H}B_{n-1}(t) + \frac{n-t}{n-1} \mathcal{H}B_{n-1}(t-1), \quad t \in \mathbb{R}. \quad (2.38)$$

(iii) *The Hilbert transform of the cardinal B-splines B_n is anti-symmetric about the middle point $n/2$ of the support of B_n ; that is*

$$\mathcal{H}B_n\left(\frac{n}{2} + t\right) = -\mathcal{H}B_n\left(\frac{n}{2} - t\right), \quad t \in \mathbb{R}. \quad (2.39)$$

We remark that it follows from part (iii) of the last theorem that $\mathcal{H}B_n$ has a zero at $n/2$.

The analytic signal of the cardinal B-spline B_n is defined by

$$Z_n(t) := B_n(t) + i\mathcal{H}B_n(t), \quad t \in \mathbb{R}. \quad (2.40)$$

This definition immediately implies that

$$Z_1(t) = \chi_{[0,1]}(t) + \frac{i}{\pi} \ln \left| \frac{t}{t-1} \right|. \quad (2.41)$$

The properties of B-splines and the Hilbert transform of B-splines are translated to Z_n , which we present in the following theorem.

Theorem 2.2: *The following statements hold for the analytic signal Z_n of the cardinal B-spline B_n :*

(i) *The function Z_n has the convolution property*

$$Z_n = B_1 * Z_{n-1} = Z_1 * B_{n-1}; \quad (2.42)$$

(ii) *The derivative of Z_n satisfies the relation*

$$\frac{d^j}{dt^j} Z_n(t) = \nabla^j Z_{n-j}(t), \quad t \in \mathbb{R}; \quad (2.43)$$

(iii) *The function Z_n may be computed by using the recursive formula*

$$Z_n(t) = \frac{t}{n-1} Z_{n-1}(t) + \frac{n-t}{n-1} Z_{n-1}(t-1), \quad t \in \mathbb{R}. \quad (2.44)$$

In practice, the spline representation obtained from the EMD method has non-uniform knots. Hence, it is desirable to consider the Hilbert transform of B-splines with non-uniform knots. In this case, we have the recurrence relation of (2.7). It is also known that the derivative of a B-spline can be written as a combination of lower-order B-splines:

$$B'_{j,k,\tau}(t) = (k-1) \left[\frac{B_{j,k-1,\tau}(t)}{\tau_{k+j-1} - \tau_j} - \frac{B_{j+1,k-1,\tau}(t)}{\tau_{k+j} - \tau_{j+1}} \right]. \quad (2.45)$$

Chen et al. (2004) proved that the Hilbert transform of B-splines has exactly the same recursive relation as the B-splines.

Theorem 2.3: (i) *The Hilbert transforms of B-splines satisfy the recursion relations*

$$\mathcal{H}B_{j,k,\tau}(t) = \frac{t - \tau_j}{\tau_{j+k-1} - \tau_j} \mathcal{H}B_{j,k-1,\tau}(t) + \frac{\tau_{j+k} - t}{\tau_{j+k} - \tau_{j+1}} \mathcal{H}B_{j+1,k-1,\tau}(t) \quad (2.46)$$

with

$$\mathcal{H}B_{j,1,\tau}(t) = \frac{1}{\pi} \ln \left| \frac{t - \tau_j}{t - \tau_{j+1}} \right|. \quad (2.47)$$

(ii) *The derivative of the Hilbert transform of B-splines satisfies the formula*

$$\frac{d}{dt}(\mathcal{H}B_{j,k,\tau})(t) = (k-1) \left[\frac{\mathcal{H}B_{j,k-1,\tau}(t)}{\tau_{k+j-1} - \tau_j} - \frac{\mathcal{H}B_{j+1,k-1,\tau}(t)}{\tau_{k+j} - \tau_{j+1}} \right]. \quad (2.48)$$

This recursive formula can be used to compute the Hilbert transform of B-splines of a higher order from the Hilbert transform of B-splines of a lower order.

The Bedrosian type formula (cf., Bedrosian 1963) plays a crucial role in computing the Hilbert transform of a product of two functions. This formula presents that the Hilbert transform of the product of a signal with lower frequency and one with higher frequency is equal to the signal with lower frequency times the Hilbert transform of the one with higher frequency. In the rest of this section, we show the validity of the Bedrosian type formula (cf., Bedrosian 1963) of the product of a polynomial and a signal having vanishing moments. Our consideration of a product of this type is motivated by an important characteristic of IMFs. We have seen either numerically or theoretically that an IMF has vanishing moments of a certain order. Specifically, we proved in Chen et al. (2004) that if a function has vanishing moments of order k , then the Hilbert transform of the product of this function with a polynomial of degree k is equal to the product of the polynomial and the Hilbert transform of this function.

For this purpose, we define the vanishing moments in the sense of the Cauchy principal value. We say a function f has vanishing moments of order k in this sense if the principal value integrals satisfy

$$PV \int_{\mathbb{R}} t^j f(t) dt = 0, \quad 0 \leq j \leq k-1. \quad (2.49)$$

It can be readily verified that the basic harmonic signal $\cos(\omega_0 t)$ has vanishing moments of order 2 if $\omega_0 \neq 0$. Because of this property, the Bedrosian formula holds for the Hilbert transform of this function. We state this result in the next proposition.

Proposition 2.4: *Let $\omega_0 \neq 0$ and p_2 be a polynomial of degree 2. Then, the Bedrosian type formula holds*

$$\mathcal{H}[p_2(\cdot) \cos(\omega_0 \cdot)](t) = p_2(t) \sin(\omega_0 t). \quad (2.50)$$

Moreover, we have the stronger result that this property holds for a general function having vanishing moments of order k .

Theorem 2.5: *Let f be a real valued function having vanishing moments of order k in the principal value sense. Then, for any polynomial p_n with degree $n \leq k$, there holds the Bedrosian type formula*

$$\mathcal{H}(p_n f) = p_n \mathcal{H}f. \quad (2.51)$$

Recently, Xu and Yan (2004) studied the necessary and sufficient conditions which ensure the validity of the Bedrosian identity of the Hilbert transform of product functions fg . These authors presented convenient sufficient conditions, which cover the classical Bedrosian theorem and provide new additional insightful information. We now review the results from Xu and Yan (2004).

Theorem 2.6: *Let $f \in W^{1,2}(\mathbb{R})$ and $g, fg, fH(g) \in L^2(\mathbb{R})$. Then the Hilbert transform of function fg satisfies the Bedrosian identity*

$$H(fg) = fH(g) \quad (2.52)$$

if and only if

$$\int_0^1 \int_{\mathbb{R}} \frac{\omega}{t^2} e^{-2i\pi x\omega(t-1)/t} \hat{f}\left(\frac{\omega}{t}\right) \overline{\hat{g}(\omega)} d\omega dt = 0. \quad (2.53)$$

For a nonempty set $\Omega \subseteq \mathbb{R}$ and a real number t , we let

$$t\Omega := \{tx : x \in \Omega\}, \quad (2.54)$$

and for the unit interval $I := [0, 1]$, we define the product set $I \cdot \Omega$ by

$$I \cdot \Omega := \bigcup_{t \in [0, 1]} t\Omega. \quad (2.55)$$

The following result is a direct consequence of Theorem 2.6.

Proposition 2.7: *Let $f \in W^{1,2}(\mathbb{R})$ and $g, fg, fH(g) \in L^2(\mathbb{R})$. If*

$$(I \cdot \text{supp}(\hat{f})) \cap \text{supp}(\hat{g}) = \emptyset, \quad (2.56)$$

then the Hilbert transform of function fg satisfies the Bedrosian identity (2.52).

The classical Bedrosian theorem follows immediately from Proposition 2.7.

Corollary 2.8: *Let $a > 0$ and suppose that $f, g \in L^2(\mathbb{R})$ with*

$$\text{supp}(\hat{f}) \subseteq (-a, a) \text{ and } \text{supp}(\hat{g}) \subseteq (-\infty, -a) \cup (a, \infty). \quad (2.57)$$

Then, the Hilbert transform of function fg satisfies the Bedrosian identity (2.52).

Xu and Yan (2004) proved the following lemma and used it to relax the hypothesis of Proposition 2.7 on the function f .

Lemma 2.9: *If $f \in L^2(\mathbb{R})$ satisfies the condition that*

$$(I \cdot \text{supp}(f)) \cap K = \emptyset \quad (2.58)$$

for some closed set K , then for any $\varepsilon > 0$, there exists $\varphi \in \mathcal{D}(\mathbb{R})$, the space of functions of bounded supports in C^∞ , such that

$$\|f - \varphi\|_2 < \varepsilon \quad (2.59)$$

and

$$(I \cdot \text{supp}(\varphi)) \cap K = \emptyset. \quad (2.60)$$

Theorem 2.10: *Let $f \in L^2(\mathbb{R})$, $g \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and $H(g) \in L^\infty(\mathbb{R})$. If condition (2.56) holds, then the Hilbert transform of functions fg satisfies the Bedrosian identity (2.52) almost everywhere.*

2.4. Performance analysis of BS-EMD

The EMD method decomposes a signal based on its intrinsic time scales. The energy of the IMFs with different local scales is distributed in different frequency bands. Numerical analysis has shown that for white and colored stationary noise, the original EMD acts as a dyadic filter bank, similar to the dyadic wavelet transform (Wu and Huang 2004; Flandrin et al. 2003). Here we present our research results on the behavior of the BS-EMD in this aspect. Most of the results have been reported in Liu et al. (2004). Our numerical experiment was carried out as follows. First, we generated 3000 independent Gaussian white noise time series of length 1024. Each was decomposed by using BS-EMD. Corresponding to each mode of the decomposition, the IMF was windowed and Fourier transformed. Then, we estimated the power spectrum for each mode as the average of the squared absolute values of the corresponding Fourier transforms over all the realizations.

Figures 2.2a–b show the estimated power spectra corresponding to the cubic and quadratic B-spline EMD respectively, where each curve corresponds to one decomposition mode. All the curves together in each figure can be interpreted as the frequency output of an equivalent filter bank. For comparison, Figure 2.2c presents the power spectra corresponding to the original EMD obtained by using the same procedure. One can see that the BS-EMD behaves similarly to the original EMD.

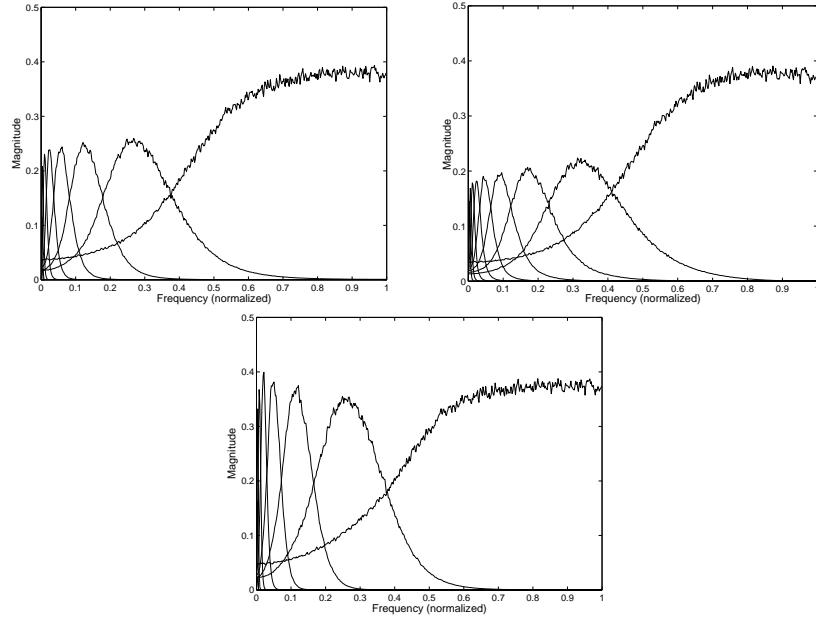


Figure 2.2: Power spectra of the IMFs obtained by using the cubic B-spline EMD (a, top left) the quadratic B-spline EMD (b, top right), and the original EMD (c, bottom). In each figure, from the right to the left, the first curve (half-bell shaped) is the power spectrum corresponding to IMF1, the second curve is that corresponding to IMF2, and so forth.

We now provide an interpretation of the properties of the BS-EMD as a filter bank. A rigorous mathematical proof requires further research. Consider a signal composed of a high frequency sinusoid riding on a low frequency sinusoid, as shown in Figure 2.3a, where the local extrema are marked by “o”. From (2.10), we know that the coefficients of the B-splines in the operator defined by (2.11) are essentially the output of a low-pass filter applied to the local extremum sequence; the low-pass filter here is the binomial average. For the quadratic and cubic B-spline operators, the filters are $\{0.5, 0.5\}$ and $\{0.25, 0.5, 0.25\}$, respectively. In other words, the coefficient sequence in (2.11) is a smoothed version of the local extremum sequence and represents the low frequency part of the latter. Furthermore, it has been proved that a function represented in the form of (2.11) does not have more sign changes than the coefficient sequence itself (De Boor 1978). We may thus consider that the component obtained by the operator

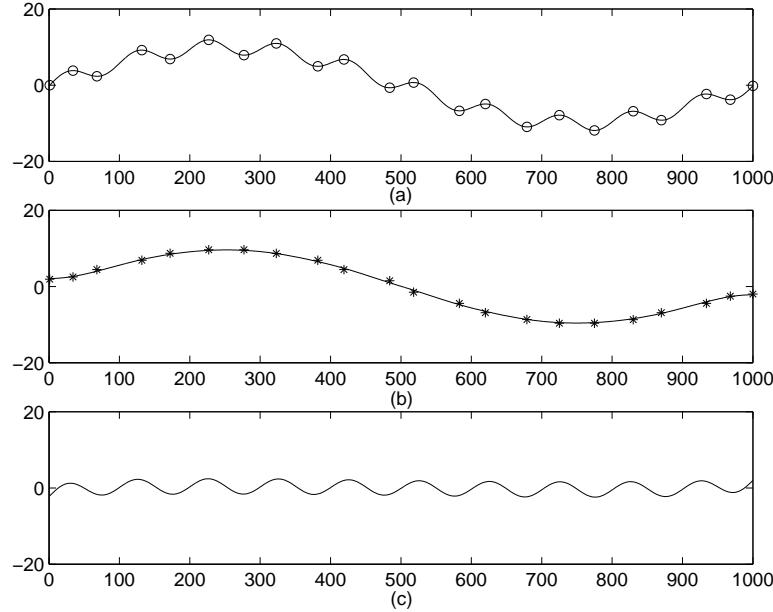


Figure 2.3: (a) Top. A signal composed of a high-frequency sinusoid riding on a low-frequency sinusoid, where the local extrema are marked by “o”. (b) Middle. The waveform obtained by applying the operator defined in (2.11) to the signal. The points “*” represent the coefficients of the operator. (c) Bottom. The difference between the signal and the waveform in (b).

of (2.11) represents the low-frequency part of the signal. Figure 2.3b shows such a component of the signal depicted in Fig. 2.3a obtained by using the cubic B-spline EMD, where the coefficients of the operator are represented by “*”. The high frequency part of this signal obtained by subtraction of the low-frequency part is shown in Fig. 2.3c. In the next iteration, a new low-frequency part is generated from the current high-frequency part. When the stoppage criterion is satisfied, we obtain a high-frequency part, i.e., the first IMF, and the difference between the signal and the IMF, which is equal to the sum of the low-frequency parts generated in all the iterations. The next IMF is obtained in the same way except that it is extracted from the current low-frequency part. One can imagine that when all the IMFs are obtained, the signal will be decomposed into a number of frequency bands similar to those in a wavelet transform.

We should point out that like the original EMD, the BS-EMD decom-

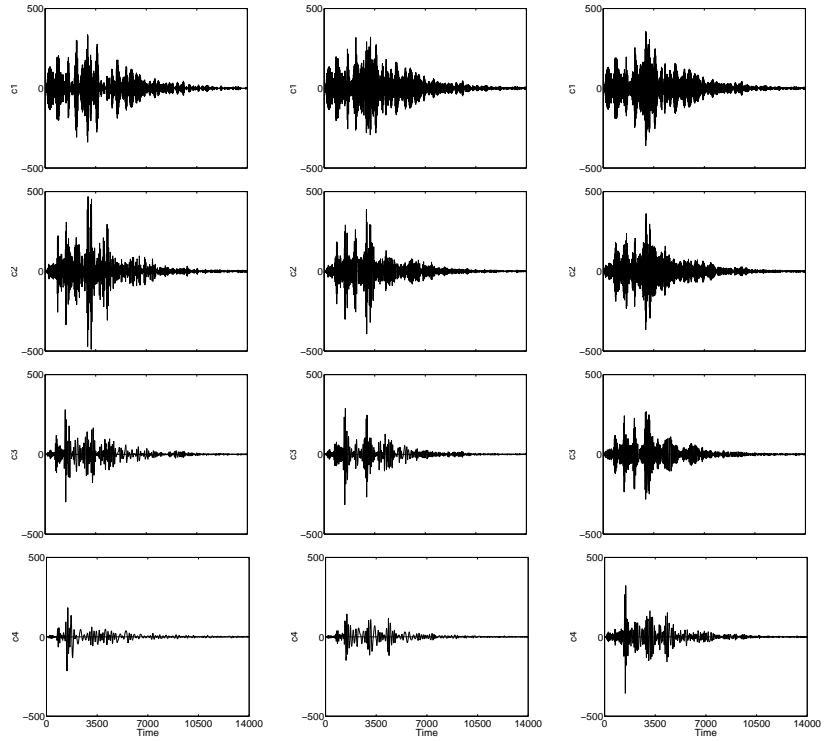


Figure 2.4: Plot of the first 4 IMFs for the earthquake data by the original EMD on the left, by the cubic B-spline EMD in the center, and by the quadratic B-spline EMD on the right.

poses a signal based on its local time scales. As such, the BS-EMD is a time-varying filter bank that is adaptive to the local time scales of a signal. This capability is different from that of a wavelet transform, in which the filter bank is predetermined. Consider a signal whose first half is a high-frequency sinusoid and second half, a low-frequency sinusoid. Differing from the frequencies and bandwidths of a wavelet transform, the central frequencies and bandwidths of the BS-EMD filter bank will change from the first to the second halves of the signal. For a white noise signal, the local scale varies randomly. The results presented in Figs. 2.2a–c thus are not applicable to a specific realization of the random process. They describe only the “average behavior” of the EMD as a filter bank.

We present another example, based on Chen et al. (2004), to further

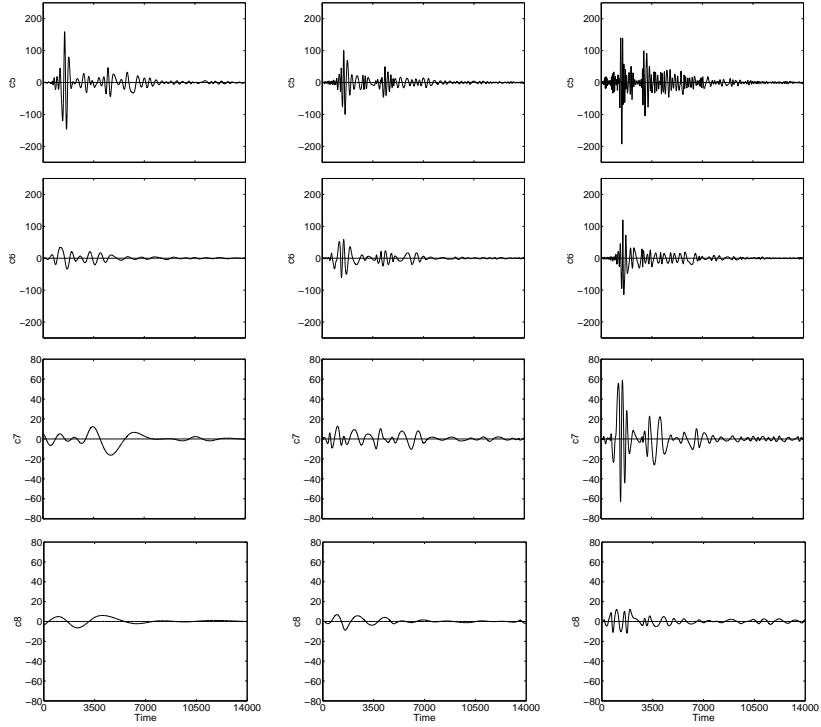


Figure 2.5: Plot of the next 4 IMFs for the earthquake data from the original EMD on the left, from the cubic B-spline EMD in the center, and from quadratic B-spline EMD on the right.

investigate the performance of the BS-EMD. The signal in the analysis is an earthquake record from the Chi-Chi event discussed originally by Huang et al. (2001). Our intention is not to decipher the meaning of the results, but merely to show that the BS-EMD can give similar results as the original EMD. In Figs. 2.4–2.6, the plots on the left are for the original envelope method, the plots in the center are for the cubic B-spline EMD, while the plots on the right are for the quadratic B-spline EMD. The residues are given separately in Fig. 2.7. As one can see, the original EMD produces only 10 IMF components, while the cubic and quadratic B-spline produce 11 and 12 IMF components, respectively. With averaging as in the B-spline, the wave groups tend to persist longer temporally and also through many different IMF components. As a result, the B-spline approaches would give a finer decomposition than the envelope approach. However, the similarity

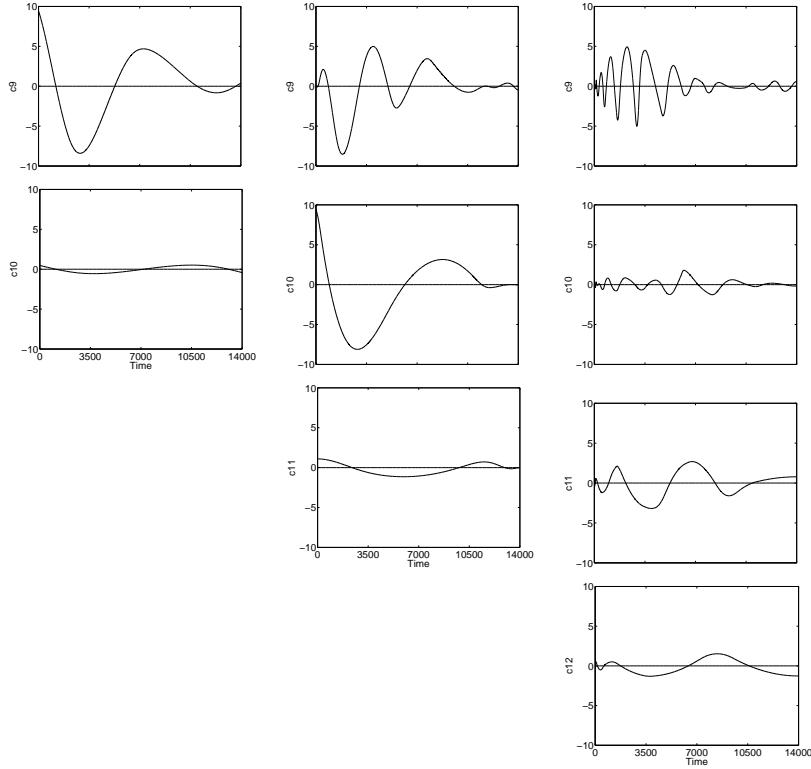


Figure 2.6: Plot of the last few IMFs for the earthquake data from the original EMD on the left, from the cubic B-spline EMD in the center, and from the quadratic B-spline EMD on the right.

between the envelope and the cubic B-spline approaches is evident. Even the last three components and the residues are qualitatively and quantitatively similar.

To evaluate the performance of the BS-EMD more accurately, we present the results of our investigation on the extent of the orthogonality among the IMFs and the energy conservation of the decompositions (Chen et al. 2004). The measures for orthogonality are defined by

$$IO_{max} := \max_{k \neq j} \frac{|\langle c_j, c_k \rangle|}{\|c_j\|_2 \|c_k\|_2}, \quad IO_{ave} = \text{Average}_{k < j} \frac{|\langle c_j, c_k \rangle|}{\|c_j\|_2 \|c_k\|_2}. \quad (2.61)$$

We compute the sum of squared values of the IMFs and divide it by the

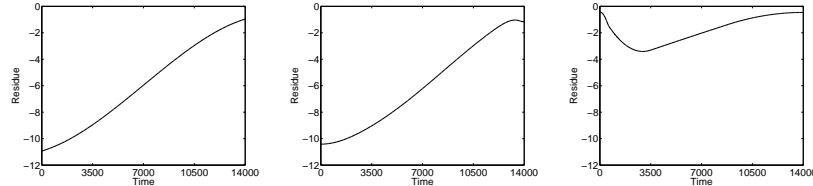


Figure 2.7: Plot of the residues for the earthquake data from the original EMD on the left, from the cubic B-spline EMD in the center, and from the quadratic B-spline EMD on the right.

squared values of the original signal minus the residue

$$IEC := \frac{\sum_t \sum_j |c_j(t)|^2}{\sum_t |s(t) - r(t)|^2} \quad (2.62)$$

to obtain an index of energy conservation. The latter reflects the fact that the sum of the IMFs represent the signal minus the trend given in the residue. For the earthquake signal, we have

Method	IO _{max}	IO _{ave}	IEC
Cubic spline	0.6111	0.0523	1.1439
Cubic B-spline	0.3334	0.0579	1.0459
Quadratic B-spline	0.3034	0.0519	1.2916

The test results summarized in the above table show that the B-spline approaches give very comparable results to those of the original envelope approach. The cubic B-spline EMD performs especially well on the energy conservation test, and this finding indicates that the total energy deviation amongst the components is the smallest. The exact choice of the order of the B-spline is an open question. Although the lower order B-spline preserves local characteristics better, it will follow the data more closely. As our goal is to find the mean through the data, the higher order B-spline should give better mean and smaller orthogonal indices in both the maximum and average values. We thus recommend the cubic B-spline EMD as the choice for further mathematical investigation and applications.

2.5. Application examples

In this section, we present two application examples of the BS-EMD in transient detection. Transients are typical non-stationary signal components. The detection of such components is of interest in various applications. The

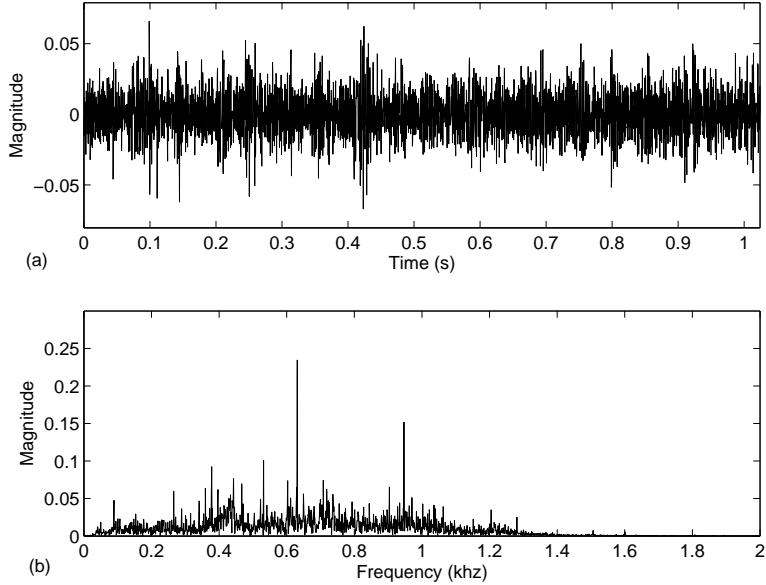


Figure 2.8: (a, top) A vibration signal collected from the gearbox with a tooth crack, and (b, bottom) its Fourier transform. The signal has been normalized as unit energy.

problem investigated here is the detection of transients in vibration signals generated in mechanical systems. Our purpose is to diagnose incipient localized failures, which are a problem of great concern to industry (Braun 1986). When localized failures such as cracks and surface spalls exist in a mechanical system, the relatively-moving machine parts often impact one another. The impacts then excite transients into the background vibration. The detection of the transients is therefore a promising way to diagnose localized failures. However, doing so has been a challenging task because compared with the background vibration, the transients excited by incipient failures are usually very small. The results presented here show that the EMD techniques provide a promising tool for dealing with this problem.

The first example comes from Liu et al. (2004). The vibration signals in this example were collected from a fatigue test of an automobile gearbox. The transmission train had four pairs of gears. At the end of the test, one tooth of the driving gear in the last gear pair, which ran at 5.9 Hz, was broken. In the early stage of development, such a fault would excite a sequence of small transients into the background vibration. Our purpose is to detect the transients from the vibration signals collected before the

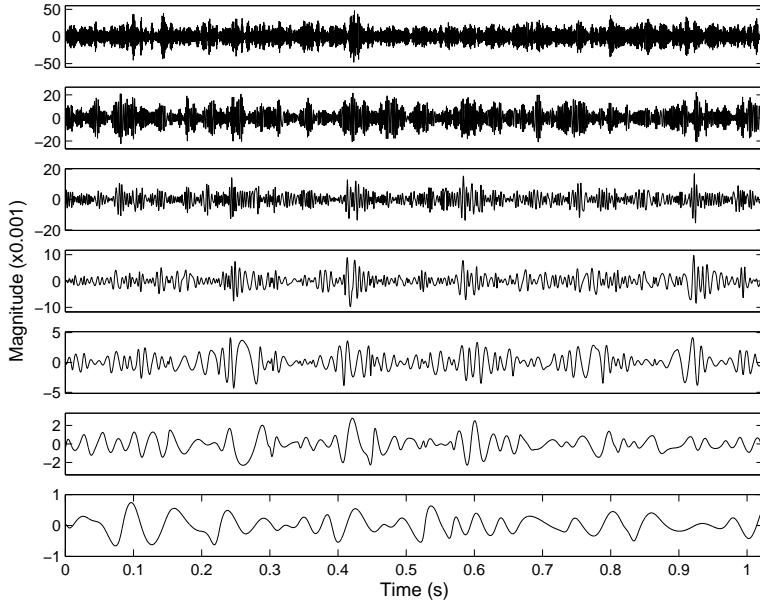


Figure 2.9: IMFs of the signal in Fig. 2.8a obtained by using the BS-EMD. The indexes of the IMFs increase from the top to the bottom.

breakage occurred. During this stage, a crack might be developing. In the test, the picked-up vibration signals were lowpass-filtered at 1.8 kHz and digitized at the sampling frequency of 4 kHz.

Figure 2.8a shows a signal collected before the tooth was broken. Since we are interested mainly in the time-frequency composition, the signal is normalized as unit energy for the sake of convenience. We also do so for the signal in the second application example presented later. The Fourier transform of the signal in Fig. 2.8a is given in Fig. 2.8b. It is difficult to understand the condition of the gearbox based directly on the Fourier transform. Figure 2.9 shows the first seven IMFs obtained by using the BS-EMD. One can see that IMFs 3–6 contain equally spaced impulses representing the transient components in the signal. It can be determined that the transients are distributed over the frequency range of these four IMFs. The Hilbert spectrum obtained from the Hilbert transform of these four IMFs is given in Fig. 2.10, where the brighter patches show the time-frequency characteristics of the transients. For better identification of the featured components, a smoothed representation of the Hilbert spectrum is used in this section (see Huang et al. 1998). From Figs. 2.9 and 2.10, we can ob-

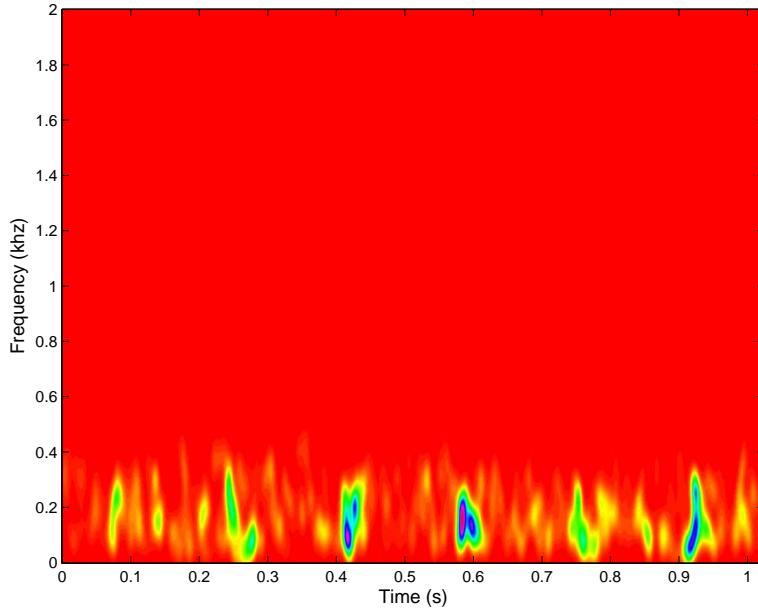


Figure 2.10: Hilbert spectrum of the third to sixth IMFs shown in Fig. 2.9.

tain the following information. Firstly, the average time spacing between the neighboring transients is about 0.17 s, corresponding to 5.9 Hz in frequency – the rotation frequency of the damaged gear. Secondly, the frequency of the impulses ranges from 0 to about 300 Hz and covers the sidebands of the meshing frequency of the last gear pair in the transmission path. This mesh frequency is about 88.4 Hz. Finally, the fifth IMF shows that a phase delay occurs in some of the impulses. Based on the general knowledge of gearbox diagnosis (Braun 1986), we know that these features indicate the existence of a localized defect on the driving gear of the last gear pair.

Figures 2.11 and 2.12 show the first seven IMFs and the Hilbert spectrum of IMFs 3–6 obtained by using the original EMD. For this specific signal, the IMFs and Hilbert spectrum obtained by using the BS-EMD appear to reveal the impulses better than those obtained by using the original EMD.

The problem that the second example involves is the localized failure diagnosis of a rolling element bearing through transient detection. The specifications of the bearing in the test were as follows: number of rolling elements, 8; diameter of the rolling elements, 15 mm; medium diameter, 65

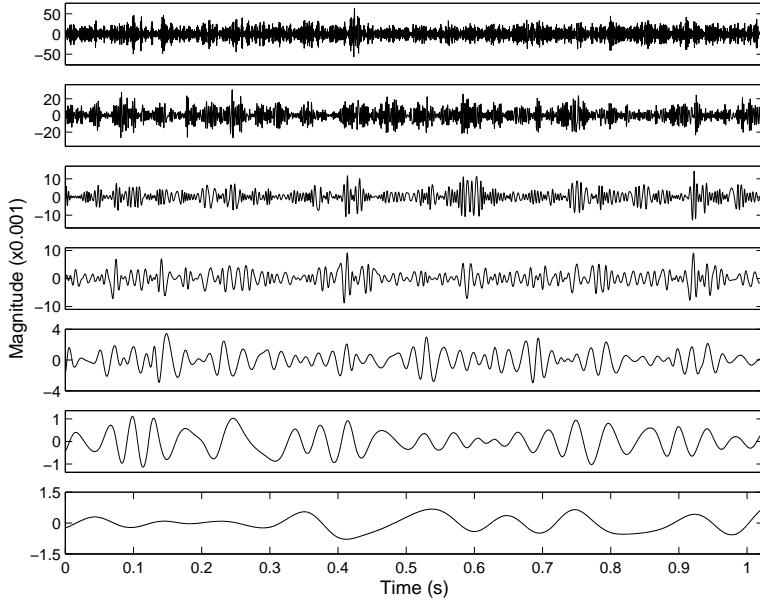


Figure 2.11: IMFs of the signal in Fig. 2.8a obtained using the original EMD. The indexes of the IMFs increase from the top to the bottom.

mm; and contact angle, 0° . The bearing carried a defect on its inner race induced in the form of a dimple measuring about 0.1 mm in depth and 1 mm in diameter to simulate an incipient surface spalling. The vibration signals were picked up at a constant inner race rotation speed of 1900 rpm. They were lowpass-filtered at 9 kHz and digitized at the sampling rate of 20 kHz. An inner race defect could generate a sequence of high-frequency vibration transients. These transients are equally spaced in time but usually have different amplitudes and frequency ranges if the inner race rotates. Under the present test conditions, it can be computed that the average time spacing between the neighboring transients is 6.4 ms (Braun 1986).

Figure 2.13 shows a typical vibration record from the bearing and its Fourier transform. The signal contains some impulses and high-frequency components. These features more or less show the presence of a localized defect in the bearing. The IMFs obtained by using the BS-EMD are presented in Fig. 2.14, of which the first three IMFs better reveal the existence of the transients than the waveform of the original signal. Figure 2.15 shows the corresponding Hilbert spectrum. It reveals the corresponding frequency range of each transient. Some transients that are difficult to identify from

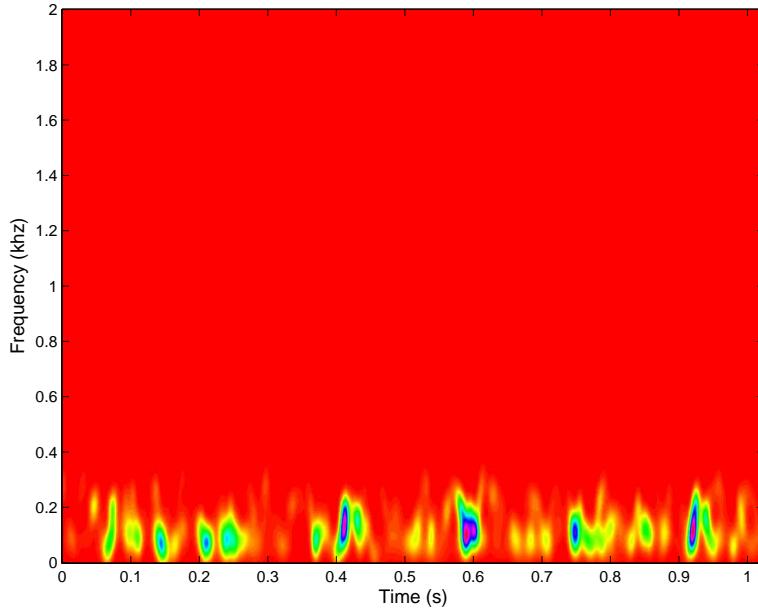


Figure 2.12: Hilbert spectrum of the third to sixth IMFs shown in Fig. 2.11.

the waveforms of the IMFs become clearer in the Hilbert spectrum, such as the transient represented by the bright patch close to the time instant of 0.01 s and in the frequency band from 7 kHz to 9 kHz. Figures 2.14 and 2.15 also reveal that the average time spacing between the neighboring transients is about 6 ms or its multiple. This value is close to the computed one mentioned above. Based on these features, the finding that the bearing carried a localized inner race defect is convincing.

Figures 2.16 and 2.17 show the IMFs obtained by using the original EMD and the Hilbert spectrum computed from the first three IMFs. They provide similar information to that in Figs. 2.14 and 2.15. We have analyzed more signals. The results show that the performances of the two methods are generally comparable.

2.6. Conclusion and future research topics

The empirical mode decomposition and Hilbert spectral analysis, comprising the Hilbert-Huang transform, has been applied with great success for nonlinear and non-stationary signal analysis in various areas. However, most of the underlying mathematical problems have been left untreated.

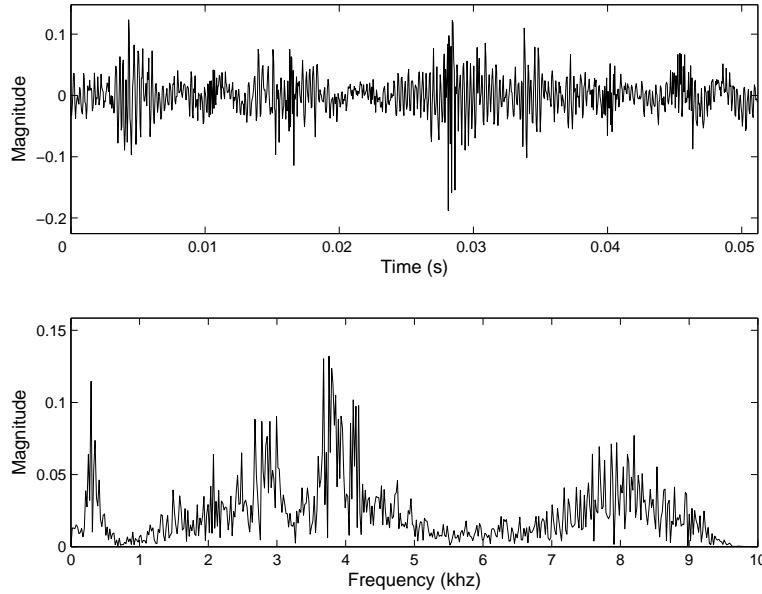


Figure 2.13: (a, top) A vibration signal collected from the bearing, and (b, bottom) its Fourier transform. The signal has been normalized as unit energy.

Given the rapid progress made in the methodology and applications, we urgently need to develop a firm mathematical foundation for the Hilbert-Huang transform. This paper reviewed the work in that direction based on Chen et al. (2004), Liu et al. (2004), and Xu and Yan (2004). We described the algorithm of our B-spline based EMD and presented the results of some related mathematical studies. The simulated and practical application examples given in the paper showed that the BS-EMD has a comparable performance to that of the original EMD. To facilitate future progress, we have identified the outstanding mathematical problems in Chen et al. (2004). We now summarize those problems closely related to the B-spline EMD.

In time-frequency analysis, traditional methods are often conducted with an *a priori* basis constructed based on certain general criteria. For non-linear and non-stationary signals, which have varying local characteristics, it is impossible to expect a fixed basis to fit all the signals without invoking spurious harmonics. A breakthrough in improving the technology is the development of the adaptive representation methods using an over-complete basis library, e.g., Coifman et al. (1992), Mallat and Zhang (1993), Chen

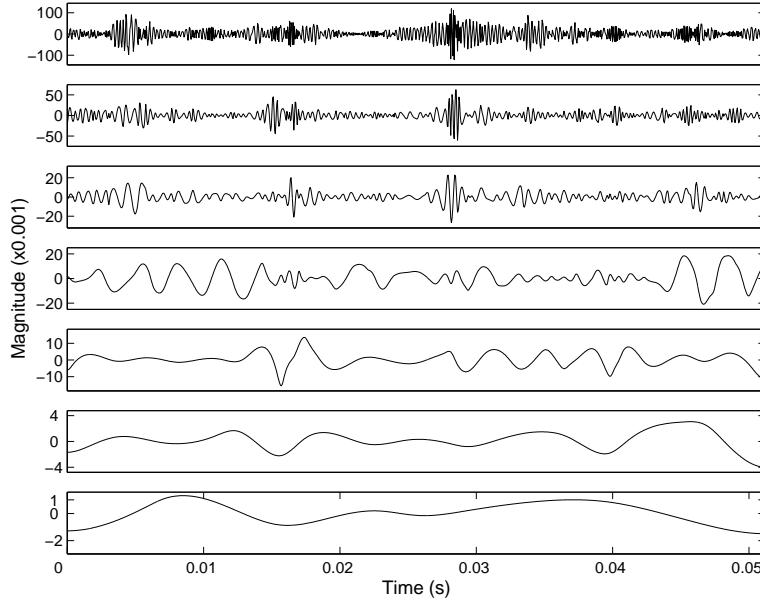


Figure 2.14: IMFs of the signal in Fig. 2.13a obtained by using the BS-EMD. The indexes of the IMFs increase from the top to the bottom.

and Donoho (1995). They select from the library a basis that is “adapted” to the analyzed signal. These methods have good mathematical grounds but are not fully adaptive, for the library itself is *a priori*, and increasing the size of the library will result in computational difficulties. In the case of the EMD approach, the representation is general and adaptive enough but not mathematically rigorous. A fundamental reason for the latter is that the intrinsic mode functions lack a mathematically rigorous definition. This problem may be addressed within the B-spline EMD method, as all the intrinsic mode functions, other than the first one, obtained from either the original or the B-spline EMD can be represented as sums of B-spline curves.

For the same reason mentioned above, the B-spline EMD could also be used to address the problem of the convergence of EMD. By “convergence,” we mean that the EMD will produce only a finite set of IMF components. Although all intuitive reasoning and numerical experiments suggest that the EMD procedure should be convergent, a general and complete proof is still wanting. In general, the cubic spline mean envelope could create extra extrema by itself. This possibility introduces more difficulties for attempts

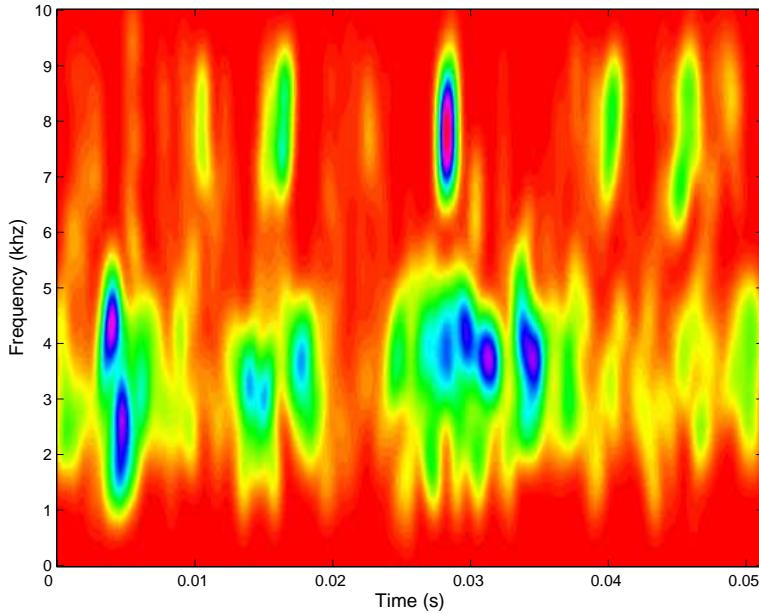


Figure 2.15: Hilbert spectrum of the first to third IMFs shown in Fig. 2.14.

to prove the convergence. An expected benefit from using B-splines is that the convergence might come as a result of the nice variation-diminishing properties of B-spline series. Certainly, there still exist challenges, such as how to deal with the hidden scales caused by the inflection points in the data.

Several implementation issues need to be settled for the B-spline approach, just as they do for the envelope approach. Firstly, the B-spline EMD also depends on the stoppage criterion. Since different stoppage criteria yield different sets of IMFs, an optimal one is naturally desired. Secondly, although we have an intuitive reason, as mentioned in section 2.4, to choose a cubic B-spline for the BS-EMD, the exact choice of the order still remains a question. Finally, since we are dealing with finite data, our algorithm must also be adjusted to use some form of boundary conditions. Although one can invoke the “clamped” end point option to fix the ends, how to select the fixed end is still a problem. Even if the local modification property of the B-spline will limit the influence to the end regions, the region of influence will become larger and larger as the scale of the IMF mode becomes larger, and the influence of the ends will propagate into the

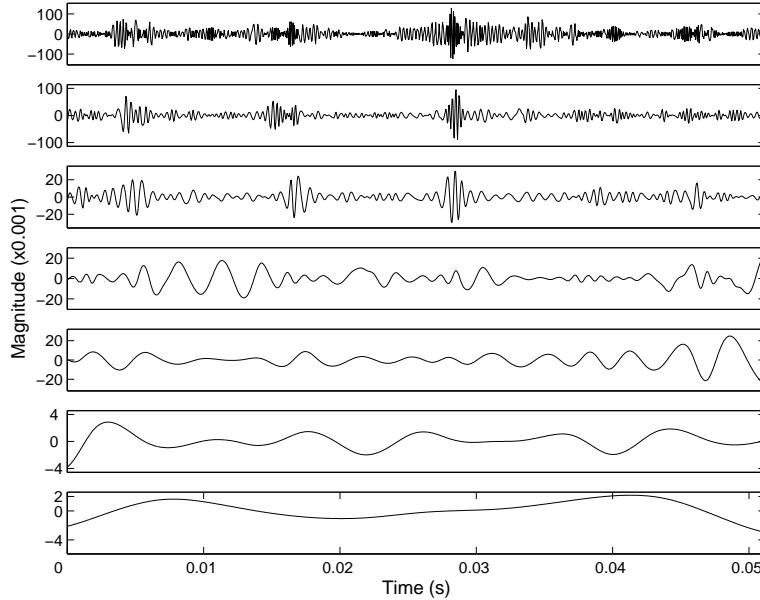


Figure 2.16: IMFs of the signal in Fig. 2.13a obtained using the original EMD. The indexes of the IMFs increase from the top to the bottom.

low-frequency data components. Clearly, if the B-spline approach is to become a viable alternative to the envelope approach, a detailed investigation of these problems is urgently needed.

Although the B-spline approach has the same problems as the original approach concerning the above-mentioned issues, with the analytical form, the B-spline approach, however, is more amenable to mathematical analysis. The variation-diminishing properties and the partition of unity properties of the B-spline series may be particularly useful. With these advantages, the B-spline approach really deserves more detailed investigation.

Acknowledgements

This work was supported in part by National Aeronautics and Space Administration under grant NAG5-5364, and National Science Foundation under grants NSF0314742 and NSF0312113.

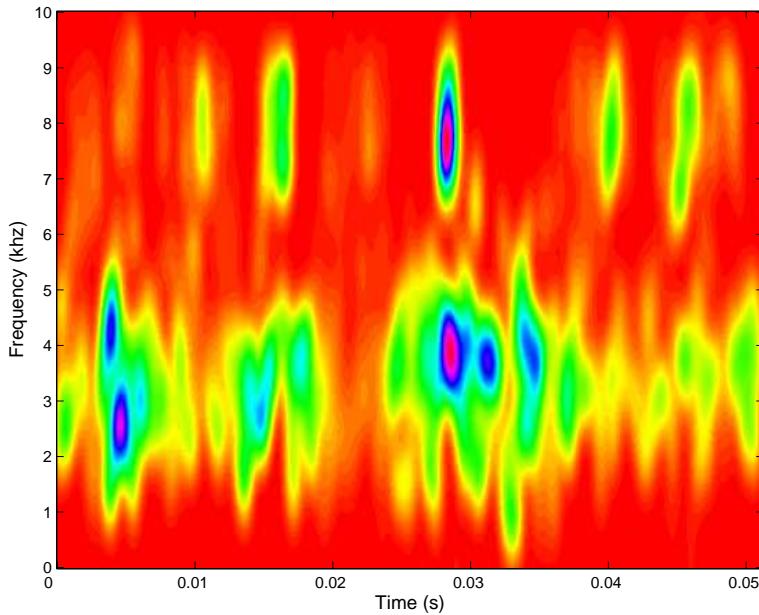


Figure 2.17: Hilbert spectrum of the third to sixth IMFs shown in Figure 2.16.

References

- Bedrosian, E., 1963: A product theorem for Hilbert transform. *Proc. IEEE*, **51**, 868–869.
- Braun, S., 1986: *Mechanical Signature Analysis*. Academic Press, 385 pp.
- Chen, S., D. Donoho, and M. Saunders, 2001: Atomic decomposition by basis pursuit. *SIAM Rev.*, **43**, 129–159.
- Chen, Q., N. E. Huang, S. Riemenschneider, and Y. Xu, 2004: A B-spline approach for empirical mode decompositions. *Adv. Comput. Math.*, in press.
- Choi, H. I., and W. J. Williams, 1989: Improved time-frequency representation of multicomponent signals using exponential kernels. *IEEE Trans. Acoust. Speech Signal Process.*, **37**, 862–871.
- Cohen, L., 1966: Generalized phase-space distribution functions. *J. Math. Phys. (Woodside, NY)*, **7**, 781–786.
- Cohen, L., 1995: *Time-Frequency Analysis*. Prentice Hall, 299 pp.
- Coifman, R. R., Y. Meyer, and M. V. Wickerhauser, 1992: Wavelet analysis and signal processing. *Wavelets and Their Applications*, M. B. Ruskai et al., Eds., Jones and Bartlett, 153–178.

- Daubechies, I., 1992: *Ten Lectures on Wavelets*. CBMS-NSF Series in Applied Mathematics. Vol. 61, SIAM, 357 pp.
- De Boor, C., 1978: *A Practical Guide to Splines*. Springer-Verlag, 392 pp.
- Diks, C., 1999: *Nonlinear Time Series Analysis*. World Scientific Press, 220 pp.
- Echeverria, J. C., J. A. Crowe, M. S. Woolfson, and B. R. Hayes-Gill, 2001: Application of empirical mode decomposition to heart rate variability analysis. *Med. Biol. Eng. Comput.*, **39**, 471–479.
- Flandrin, P., G. Rilling, and P. Gonçalvès, 2004: Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.*, **11**, 112–114.
- Gabor, D., 1946: Theory of communications. *J. IEE*, **93**, 429–457.
- Golitschek, M. V., 1972: On the convergence of interpolating periodic spline functions of high degree. *Numer. Math.*, **19**, 146–154.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of nonlinear water waves: The Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Huang, N. E., C. C. Chern, K. Huang, L. W. Salvino, S. R. Long, and K. L. Fan, 2001: A new spectral representation of earthquake data: Hilbert spectral analysis of Station TCU129, Chi-Chi, Taiwan, 21 September 1999. *Bull. Seism. Soc. Am.*, **91**, 1310–1338.
- Jones, D. L., and T. W. Parks, 1990: A high resolution data-adaptive time-frequency representation. *IEEE Trans. Acoust. Speech Signal Process.*, **38**, 2127–2135.
- Liu, B., S. Riemenschneider, and Y. Xu, 2004: Gearbox fault diagnosis using empirical mode decomposition and Hilbert spectrum. *Mech. Syst. Signal Process.*, in press.
- Mallat, S., 1998: *A Wavelet Tour of Signal Processing*. Academic Press, 637 pp.
- Mallat, S., and Z. Zhang, 1993: Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Process.*, **41**, 3397–3415.
- Pines, D., and L. Salvino, 2002: Health monitoring of one dimensional structures using empirical mode decomposition and the Hilbert-Huang transform. *Proc. SPIE*, **4701**, 127–143.
- Qian, S., and D. Chen, 1996: *Joint Time-Frequency Analysis: Methods and Applications*. Prentice Hall, 302 pp.
- Schoenberg, I. J., 1964: On interpolation by spline functions and its minimal

- properties. *On Approximation Theory (Proc. Oberwolfach Conf. 4–10 Aug. 1963)*, ISNM Vol. 5, Birkhäuser, 109–129.
- Schoenberg, I. J., 1972: Notes on spline functions I: The limits of the interpolating periodic spline functions as their degree tends to infinity. *Indag. Math.*, **34**, 412–422.
- Schoenberg, I. J., 1976: On the remainders and the convergence of cardinal spline interpolation for almost periodic functions. *Studies in Spline Functions and Approximation*, S. Karlin et al., Eds., Academic Press, 277–303.
- Schoenberg, I. J., 1983: A new approach to Euler splines. *J. Approx. Theory*, **39**, 324–337.
- Ville, J., 1948: Théorie et applications de la notion de signal analytique. *Câbles Transmissions*, **2A**, 61–74.
- Wigner, E. P., 1932: On the quantum correction for thermodynamic equilibrium. *Phys. Rev.*, **40**, 749–759.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.
- Xu, Y., and D. Yan, 2004: The Hilbert transform of product functions and the Bedrosian identity. Preprint, Department of Mathematics, Syracuse University, USA, 11 pp.
- Zhang, R. R., S. Ma, and S. Hartzell, 2003: Signatures of the seismic source in EMD-based characterization of the 1994 Northridge, California, earthquake recordings. *Bull. Seism. Soc. Amer.*, **93**, 501–518.
- Zhao, Y., L. E. Atlas, and R. J. Marks, 1990: The use of cone-shaped kernels for generalized time-frequency representations of non-stationary signals. *IEEE Trans. Acoust. Speech Signal Process.*, **38**, 1084–1091.

Sherman Riemenschneider

Department of Mathematics, West Virginia University, Morgantown, WV 26506, USA
sherm@math.wvu.edu

Bao Liu

Department of Mathematics, West Virginia University, Morgantown, WV 26506, USA
bliu@math.wvu.edu

Yuesheng Xu

*Department of Mathematics, Syracuse University, Syracuse, NY 13244,
USA*

yxu06@syr.edu

Norden E. Huang

*Goddard Institute for Data Analysis, Code 614.2, NASA/Goddard Space
Flight Center Greenbelt, MD 20771, USA*

norden.e.huang@nasa.gov

CHAPTER 3

EMD EQUIVALENT FILTER BANKS, FROM INTERPRETATION TO APPLICATIONS

Patrick Flandrin, Paulo Gonçalvès and Gabriel Rilling

Huang's data-driven technique of empirical mode decomposition (EMD) is given a filter bank interpretation from two complementary perspectives. First, a stochastic approach operating in the frequency domain shows the spontaneous emergence of an equivalent dyadic filter bank structure when EMD is applied to the versatile class of fractional Gaussian noise processes. Second, a similar structure is observed when EMD is operated in the time domain on a deterministic pulse. A detailed statistical analysis of the observed behavior is carried out involving extensive numerical simulations that suggest a number of applications. New EMD-based approaches are used to estimate the scaling exponents in the case of self-similar processes, to perform a fully data-driven spectral analysis, and to denoise-detrend signals that contain noise.

3.1. Introduction

Empirical mode decomposition (EMD) has been recently pioneered by Huang et al. (1998) for adaptively decomposing signals into a sum of “well-behaved” AM-FM components consisting of natural “intrinsic” building blocks that describe the complicated waveform. The technique has already been employed successfully in various applications (Coughlin and Tung 2004; Fournier 2002; Huang et al. 1998; Neto et al. 2004; Wu et al. 2001).

Although EMD is quite simple in principle, it still lacks a theoretical foundation. Indeed, it is presently defined only as the output of an iterative algorithm, with no analytical definition that could be used for performance evaluation. The only way to better understand this technique is to resort to extensive numerical simulations in well-controlled situations. Such an “input-output” approach is adopted here, with the objective of obtaining a detailed, yet empirical, statistical knowledge of the EMD behavior, just as we might do for some unknown “filter” in signal processing.

Because the EMD algorithm is not uniquely defined since it depends

on a number of user-controlled tunings such as the particular interpolation scheme for envelope extraction, the stopping criterion criterion used in the sifting process, and the manner in which border effects are treated, we assume that its principle and the manner in which it is implemented are known. More precisely, the algorithm used in this study was developed on the basis on algorithmic considerations described in earlier publications (see Rilling et al. 2003) and is available as a MATLAB code on the Internet (<http://perso.ens-lyon.fr/patrick.flandrin/emd.html>).

3.2. A stochastic perspective in the frequency domain

Our first characterization of EMD is carried out in the frequency domain from a stochastic perspective. The idea is to apply EMD to some broadband noise in order to understand how a full spectrum process is split into its “intrinsic mode functions” (IMF). One versatile class of full spectrum processes is provided by *scaling processes* for which wavelets (unanimously considered as a naturally fitted analysis tool; see Abry et al. 2000) can be used as a benchmark for performance evaluation.

3.2.1. Model and simulations

Fractional Gaussian noise (fGn, see Embrechts and Maejima 2002; Mandelbrot and van Ness 1968) is a generalization of ordinary white noise. It is a versatile model for a homogeneously spreading broadband noise without any dominant frequency band, is an intrinsically discrete-time process, and may be described as the increment process of fractional Brownian motion (fBm) since fBm is the only self-similar Gaussian process with stationary increments. Consequently the statistical properties of fGn are entirely determined by its second-order structure, which depends solely upon one single scalar parameter, H , its Hurst exponent. More precisely, $\{x_H[n], n = \dots, -1, 0, 1, \dots\}$ is a fGn of index H (with $0 < H < 1$) if and only if it is a zero-mean Gaussian stationary process whose autocorrelation sequence $r_H[k] := \mathbb{E}\{x_H[n]x_H[n+k]\}$ is

$$r_H[k] = \frac{\sigma^2}{2} (|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}). \quad (3.1)$$

It is well known that the special case $H = \frac{1}{2}$ reduces to (discrete-time, uncorrelated) white noise, whereas other values induce non-zero correlations, either negative when $0 < H < \frac{1}{2}$ or positive when $\frac{1}{2} < H < 1$ (long-range dependence). Taking the discrete Fourier transform of (3.1),

we readily obtain the power spectrum density of fGn, or

$$\mathcal{S}_H(f) = C \sigma^2 |e^{i2\pi f} - 1|^2 \sum_{k=-\infty}^{\infty} \frac{1}{|f + k|^{2H+1}}, \quad (3.2)$$

with $|f| \leq \frac{1}{2}$. If $H \neq \frac{1}{2}$, we have $\mathcal{S}_H(f) \sim C \sigma^2 |f|^{1-2H}$ when $f \rightarrow 0$. It therefore follows that fGn is a convenient model for power-law spectra at low frequencies. From its spectral properties, the particular value $H = \frac{1}{2}$ delineates two domains with contrasting behaviors. In the regime $0 < H < \frac{1}{2}$, we have $\mathcal{S}_H(0) = 0$, and the spectrum is high-pass (sometimes referred to as an “ultraviolet” situation). On the other hand, within the range $\frac{1}{2} < H < 1$, we have $\mathcal{S}_H(0) = \infty$ with a “ $1/f$ ”-type spectral divergence (“infrared” catastrophe). In both situations, the power-law form of the spectrum, although not exactly verified, is well approximated over most of the Nyquist frequency band. In other words, we have a quasi-linear relation in log-log coordinates,

$$\log \mathcal{S}_H(f) \approx (1 - 2H) \log |f| + C,$$

for most frequencies $-\frac{1}{2} \leq f \leq \frac{1}{2}$.

3.2.2. Equivalent transfer functions

Extensive simulations were carried out on fGn processes, with H values ranging from 0.1 to 0.9. The present study [whose results were first proposed in Flandrin et al. (2004) with further extensions in Flandrin and Gonçalvès (2004)] generalizes the study conducted independently by Wu and Huang (2004) for white noise only ($H = \frac{1}{2}$) and consistently supports their findings.

In all of our simulations, the data length was taken to be $N = 512$, and, for each value of H , $J = 5000$ independent sample paths of fGn were generated via the Wood and Chan (1994) algorithm. EMDs were computed for all sample paths $\{x_H^{(j)}[n]; n = 1, \dots, N\}$ (with $j = 1, \dots, J$), resulting in a collection of IMFs referred to as $\{d_{k,H}^{(j)}[n]; n = 1, \dots, N; k = 1, \dots, K_j\}$. Although the number K_j of IMFs varied from one realization to the other, none of the realizations generated less than 7 modes. Therefore, $K = 7$ has been taken in this study as the common number of modes for all realizations.

Given this dataset, a spectral analysis was carried out mode by mode with the estimated power spectrum density (PSD) given by

$$\hat{\mathcal{S}}_{k,H}(f) := \sum_{m=-N+1}^{N-1} \hat{r}_{k,H}[m] w[m] e^{-i2\pi f m}, \quad |f| \leq \frac{1}{2},$$

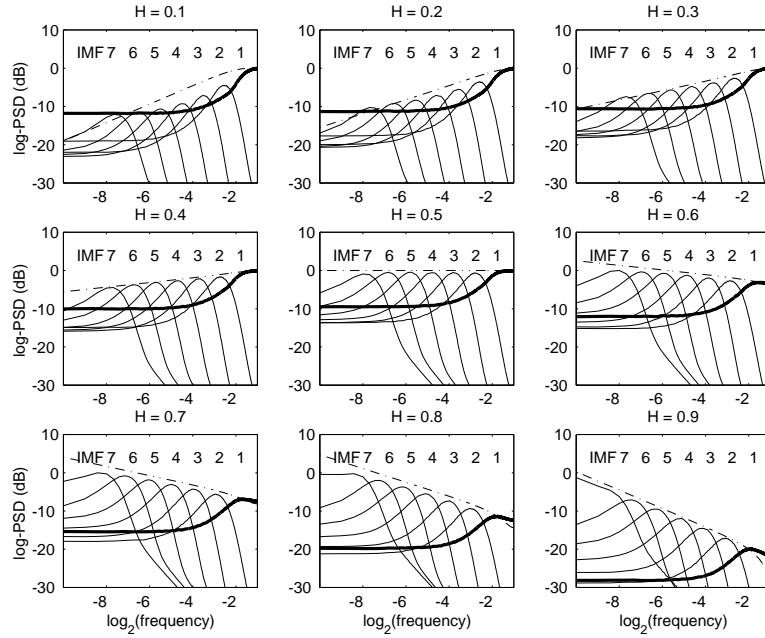


Figure 3.1: IMF power spectra in the case of fractional Gaussian noise. The logarithm of the estimated power spectrum densities (log-PSD) is plotted as a function of the logarithm of the normalized frequency for the first seven IMFs. For $H = 0.1, 0.2, \dots, 0.9$, the spectral estimates have been computed on the basis of 5000 independent sample paths of 512 data points. Theoretical PSDs of the full processes are superimposed as dashed-dotted curves. (Originally published in *Int. J. Wavelets Multiresolut. Inform. Process.*, **2**, 477–496, ©2004 World Scientific.)

where $w[n]$ is a Hamming taper, and

$$\hat{r}_{k,H}[m] = \frac{1}{J} \sum_{j=1}^J \left(\frac{1}{N} \sum_{n=1}^{N-|m|} d_{k,H}^{(j)}[n] d_{k,H}^{(j)}[n + |m|] \right), \quad |m| \leq N - 1$$

is the ensemble average (over the J realizations) of the empirical estimates of the auto-correlation function. The result of this spectral analysis is plotted in Fig. 3.1, whose graphs reveal a number of striking features:

- (1) Regardless of the value of the Hurst exponent H , the behavior of the first IMF (thick line) differs from that of the other modes. To a first approximation, it possesses the characteristics of a high-pass filter while higher order modes behave similarly to a band-pass filter. The (roughly

half-band) high-pass character of the first mode must be tempered, however, by the fact that the maximum attenuation in the stop-band is no more than 10 dB (as compared to the maximum which occurs at the Nyquist frequency $\frac{1}{2}$, and there is a non-negligible contribution in the lower half-band in “ultraviolet” situations $H < \frac{1}{2}$).

- (2) As H varies from 0.1 to 0.9, the spectrum of the last IMF ($k = 7$) progressively evolves from band-pass to increasingly low-pass, in accordance with the increasing predominance of low frequencies (“infrared catastrophe”).
- (3) In a similar, but more general manner, the energy balance among the different modes reflects the behavior of the global spectrum (superimposed dashed-dotted curve) described by (3.2), the flat spectrum when $H = \frac{1}{2}$ (the case of white noise), and the increasing (decreasing) power-law spectrum when $H < \frac{1}{2}$ ($H > \frac{1}{2}$).
- (4) For modes $k = 2$ to 6 (band-pass IMFs), all the spectra appear nearly the same, with some shifts in abscissa and ordinate, and this finding is surprisingly reminiscent of what is currently being observed in wavelet decompositions (see Flandrin 1999; Mallat 1998).

This last observation suggests that we should examine in greater detail how the different spectra are related to each other for a given H . To this end, we can use the unique structure of IMFs: all extrema appear as an alternation of local minima and maxima separated by only one zero-crossing. Finding the average number of zero-crossings in a mode is, therefore, a meaningful way of characterizing its mean frequency. The average number of zero-crossings $z_H[k]$ is plotted in Fig. 3.2a as a function of the IMF number k ; this figure suggests the functional relation

$$z_H[k] \propto \rho_H^{-k}, \quad (3.3)$$

where ρ_H very nearly equals 2.

A more precise check of (3.3) is shown in Fig. 3.2b, where the estimated scaling factor ρ_H is given by the slope from a linear fit of a semi-log diagram of $\log_2 z_H[k]$ vs. k for $k = 2$ to 6. The observed decrease in the number of zero-crossings as the order of the IMFs increases is nearly equal to 2 and

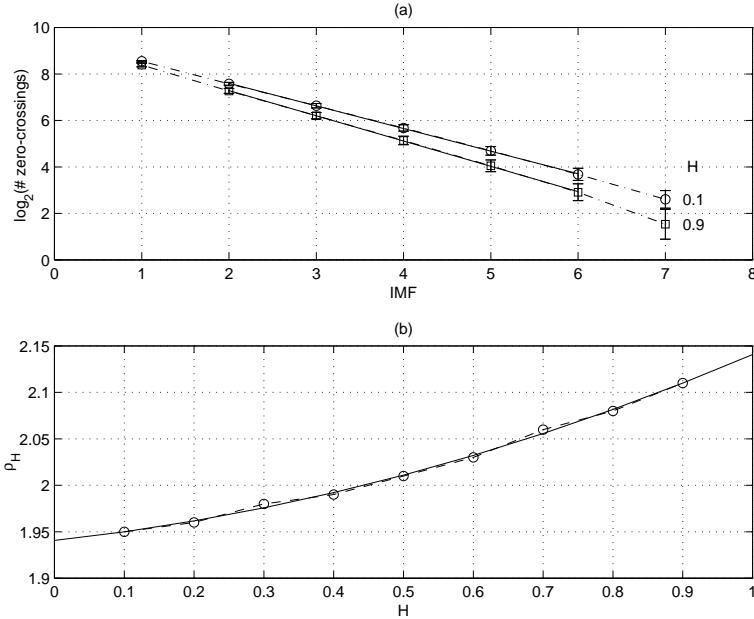


Figure 3.2: (a) IMF average number of zero-crossings *zero-crossing* in the case of fractional Gaussian noise. For clarity, only those curves corresponding to the extreme indices $H = 0.1$ (circles) and $H = 0.9$ (squares) have been plotted in the diagram; the remaining cases ($H = 0.2, 0.3, \dots, 0.8$) lead to regularly intertwined similar curves. The superimposed solid lines correspond to linear fits within the IMF range $k = 2$ to 6 . (b) Corresponding decrease rate of zero-crossings (circles and dashed line), with the least-squares quadratic fit given by (3.4) superimposed as a solid line. (Originally published in *Int. J. Wavelets Multiresolut. Inform. Process.*, **2**, 477–496, ©2004 World Scientific.)

may be approximated by the quadratic expression:^{*}

$$\rho_H \approx 2.01 + 0.2 \left(H - \frac{1}{2} \right) + 0.12 \left(H - \frac{1}{2} \right)^2. \quad (3.4)$$

Using (7.4), we can improve our search for self-similarity in the “filter bank” structure of Fig. 3.1. If we restrict ourselves to the band-pass IMFs ($k = 2$ to 6), self-similarity means that

$$\mathcal{S}_{k',H}(f) = \rho_H^{\alpha(k'-k)} \mathcal{S}_{k,H}(\rho_H^{k'-k} f) \quad (3.5)$$

^{*}The accuracy of this approximation is slightly dependent on the manner used to compute the IMFs and, in particular, on the choice of the stopping criterion used for the sifting process. Further studies will be necessary to clarify this point.

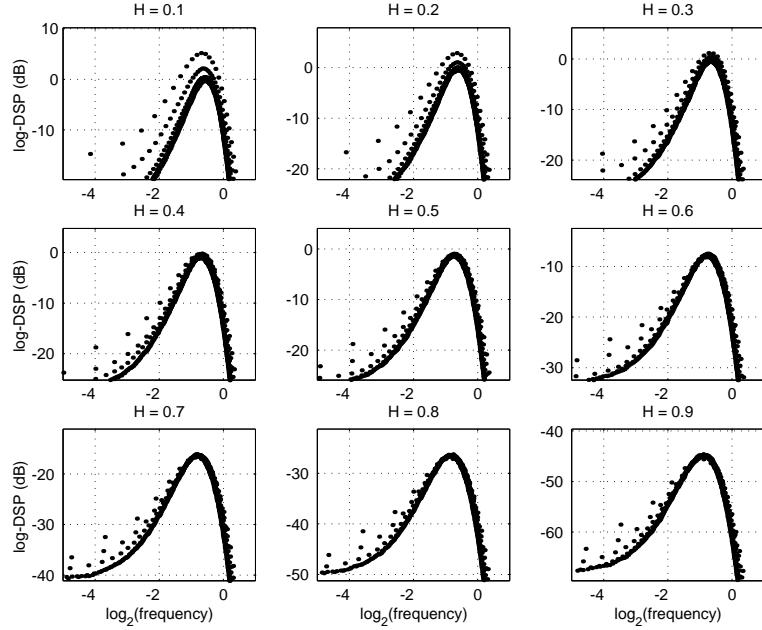


Figure 3.3: Renormalized IMF spectra in the case of fractional Gaussian noise. For each value of H , the band-pass IMFs ($k = 2$ to 6) of Figure 3.1 are plotted according to the renormalization given by (3.5) with $\alpha = 2H - 1$, and the values of ρ_H are given in Fig. 3.2b. (Originally published in *Int. J. Wavelets Multiresolut. Inform. Process.*, **2**, 477–496, ©2004 World Scientific.)

for some α and any $k' > k \geq 2$. Consequently, the power spectra of all IMFs should collapse onto a single curve when properly renormalized. Indeed, setting $\alpha = 2H - 1$ verifies this assumption as the corresponding renormalizations converge to the same template (see Fig. 3.3). Even if some low frequency discrepancies are present (especially when $H < \frac{1}{2}$), these diagrams support our claim that, to a first approximation, EMD acts on fGn as a dyadic filter bank of constant- Q band-pass filters.

3.3. A deterministic perspective in the time domain

Our second approach to the characterization of the filter-like structure of EMD is constructed deterministically and in the time domain. Therefore, we seek to obtain an equivalent impulse response of the analysis.

3.3.1. Model and simulations

Finding the impulse response of a system usually amounts to observing its output when excited with a Dirac pulse $\delta(t)$ or, in discrete-time, a function $\delta[n]$ that is zero everywhere except when $n = 0$ with $\delta[0] = 1$. Doing so is not possible here since such an input signal would not consist of enough local extrema to initiate the algorithm. An alternative is to consider an idealized pulse as the limit of a noisy pulse as the signal-to-noise ratio goes to infinity. From this point of view, we model the noisy pulse $\delta_\varepsilon[n] := \delta[n] + \varepsilon x_{1/2}[n]$, and the effective IMFs are defined as

$$d_k[n] = \lim_{\varepsilon \rightarrow 0^+} \mathbb{E}\{d_{k,\varepsilon}[n]\},$$

where $d_{k,\varepsilon}[n]$ denotes the k -th IMF of $\delta_\varepsilon[n]$.

In practice, we used in our simulations zero mean unit variance Gaussian white noise $x_{1/2}[n]$, with $\varepsilon = 0.02$ (corresponding to a signal-to-noise ratio (SNR) of 34 dB, with SNR defined as $10 \log_{10}(1/\text{var}\{\varepsilon x_{1/2}[n]\})$). The data length of each realization has been fixed to $N = 256$, and simulations have been carried out on $J = 5000$ independent realizations.

3.3.2. Equivalent impulse responses

The result of the simulation (average EMD for our slightly noisy pulse) is plotted in Fig. 3.4. Again, this figure shows a striking resemblance to what we would have obtained by using a wavelet analysis, the different averaged IMFs apparently all having the same shape for each order k .

To understand our results, we again seek a self-similar structure which would reduce all of the IMFs to one universal waveform thanks to a well-chosen renormalization. Figure 3.5 shows that this reduction is indeed possible. In constructing our waveform, we first plotted the maximum amplitude of the different IMFs as a function of their index. Doing so allowed us to identify an exponential variation of the form:

$$d_k[0] = 2^{C-pk},$$

where $p \approx 0.85$. In the second step, we found that a dilation factor $\alpha = 2^p \approx 1.80$ gave a superposition of the different waveforms (and their spectra) which we could express as

$$d_k[n] = \frac{1}{\alpha^k} \psi\left(\left\lfloor \frac{n}{\alpha^k} \right\rfloor\right),$$

where $\psi(t)$ is a reference waveform and analogous to a mother wavelet in a multiresolution analysis. For our analysis, the interpolation scheme used in

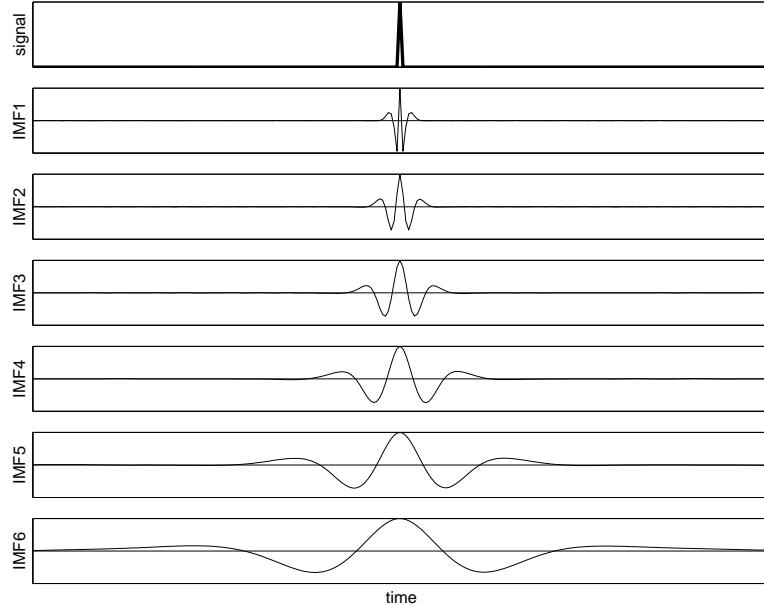


Figure 3.4: Impulse response. The EMD equivalent impulse response was obtained by averaging each IMF by using a large number of decompositions computed with a slightly noisy pulse. Here, 5 000 independent realizations from 256 data points were simulated with a signal-to-noise ratio of 34 dB. The first frame shows the average pulse and each successive frame corresponds to the ensemble average of the first six IMFs normalized in amplitude mode by mode.

the EMD was a cubic spline (see Huang et al. 1998; Rilling et al. 2003; available online at <http://perso.ens-lyon.fr/patrick.flandrin/emd.html>). Note how very similar $\psi(t)$ is to a cubic spline wavelet (third convolution power of the Haar wavelet).

3.4. Selected applications

If we accept that EMD may be characterized in some cases as a “spontaneous” filter bank, then several potentially useful applications are immediately suggested.

3.4.1. *EMD-based estimation of scaling exponents*

Our first application concerns the estimation of the Hurst exponent H for fGn based on the EMD spectral analysis described in Section 3.2. Given

the self-similar relation (3.5) for PSDs for band-pass IMFs (index $k > 1$), we can deduce how the variance should evolve as a function of k . Assuming that (3.5) holds for any $k' > k \geq 2$ and $\alpha = 2H - 1$, we have

$$\begin{aligned} V_H[k'] &:= \text{var } d_{k',H}[n] = \int_{-1/2}^{1/2} \mathcal{S}_{k',H}(f) df \\ &= \rho_H^{\alpha(k'-k)} \int_{-1/2}^{1/2} \mathcal{S}_{k,H}(\rho_H^{k'-k} f) df \\ &= \rho_H^{(\alpha-1)(k'-k)} V_H[k], \end{aligned}$$

which leads to

$$V_H[k] = C \rho_H^{2(H-1)k}. \quad (3.6)$$

The IMF variance should be an exponentially decreasing function of the IMF index with a decay rate which is a linear function of the Hurst exponent H . Experimental evidence for this behavior is given in Fig. 3.6 where a semi-log diagram (in base 2) gives the (energy-based) empirical variance estimate

$$\hat{V}_H[k] := \frac{1}{J} \sum_{j=1}^J \left[\frac{1}{N} \sum_{n=1}^N \left(d_{k,H}^{(j)}[n] \right)^2 \right] \quad (3.7)$$

as a function of the index k . From a logarithmically linearized version of (3.6), straight lines may be fitted to the different curves. The slope κ_H then gives an estimated Hurst exponent \hat{H} via

$$\hat{H} = 1 + \frac{\kappa_H}{2}. \quad (3.8)$$

Figure 3.6 shows that (3.6) holds only for IMF indices $k > 1$. Furthermore, the error increases as H becomes small (typically, the model fits the data reasonably well for $H > \frac{1}{4}$).

To better understand our ability to estimate H from the slope of a “log-energy vs. IMF index” diagram, one must consider not only the evolution of the variance as a function of the modes, but also the possible correlations which may exist within and between modes. To this end, we focus on band-pass IMFs ($k > 1$) and evaluate the two-dimensional correlation function[†]

$$D_H[k', n'] := \mathbb{E} \{ d_{k,H}[n] d_{k+k',H}[n+n'] \}$$

[†]The definition of this quantity is based on the implicit assumption that IMFs are, jointly, second-order stationary processes.

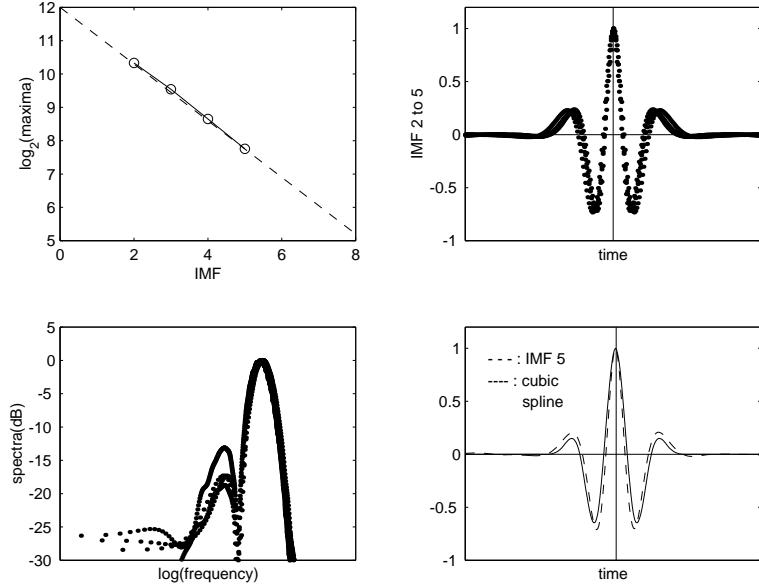


Figure 3.5: Self-similarity. The logarithm of the maximum amplitude of IMFs 2 to 5 in Fig. 3.4 is given by a linear function of the IMF index (top left diagram). Renormalizing these IMFs, either in time (top right) or in frequency (bottom left), yields an unique curve. The “impulse response” is similar to a cubic spline wavelet (bottom right).

by using the averaged empirical estimate

$$\hat{D}_H[k', n'] := \frac{1}{J} \sum_{j=1}^J \left(\frac{1}{NK} \sum_{k=2}^{K-|k'|} \sum_{n=1}^{N-|n'|} d_{k,H}^{(j)}[n] d_{k+|k'|,H}^{(j)}[n + |n'|] \right), \quad (3.9)$$

with $|n'| \leq N - 1$ and $|k'| \leq K - 2$. Here, K denotes the largest IMF index minus 1, and we have discarded the residual. This two-dimensional correlation function of the full IMF matrix is plotted in Fig. 3.7 and shows that modes with different indices are essentially uncorrelated. The only significant values of $\hat{D}_H[k', n']$ correspond to $k' = 0$, i.e., to intra-scale correlations, with a correlation decay which becomes slower as H is increased.

The effects of using our estimate of the Hurst exponent H given by (3.8) and the slope κ_H deduced from (3.7) on (3.9) are twofold. First, because of the non-zero intra-scale correlations, the variance estimate $\hat{V}_H[k]$ given by (3.7) would experience large fluctuations, especially when the Hurst exponents H and IMF indices k are large. Second, the negligible inter-scale correlations should allow for an estimate of the slope κ_H from a weighted

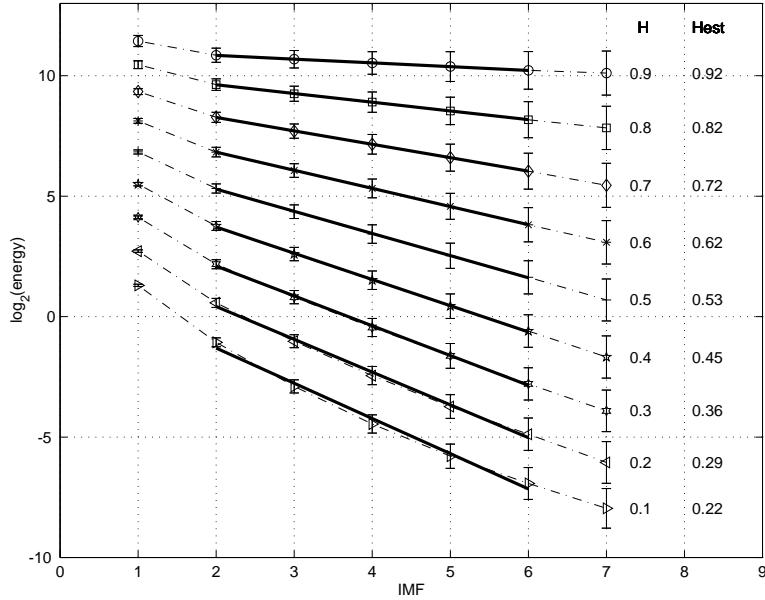


Figure 3.6: Estimated IMF \log_2 -variance in the case of fractional Gaussian noise. The values of the empirical (energy-based) variance estimates are given by dotted lines for different values of the Hurst exponent H . The error bars correspond to the standard deviations associated with the 5 000 realizations run in the study. The mean value of the estimated Hurst exponents is also given, based on a weighted linear fit within the IMF indices range $k = 2$ to 6. For clarity, all curves have been arbitrarily shifted along the vertical axis to avoid overlapping. (Originally published in *Int. J. Wavelets Multiresolut. Inform. Process.*, **2**, 477–496, ©2004 World Scientific.)

linear regression from a semi-log diagram of $\log_2 \hat{V}_H[k]$ vs. k . Further results from the effective performance of this EMD-based estimator of H (and comparisons with wavelet-based approaches) can be found in Flandrin and Gonçalvès (2003).

3.4.2. EMD as a data-driven spectrum analyzer

If one accepts that EMD behaves as a homogeneous filter bank for processes whose (full) spectrum varies monotonically, one can further investigate how this method decomposes processes with a less regular spectrum. Figure 3.8 shows the preliminary results obtained for an auto-regressive (AR) process of order 4. While EMD does achieve a filter bank-like decomposition in this

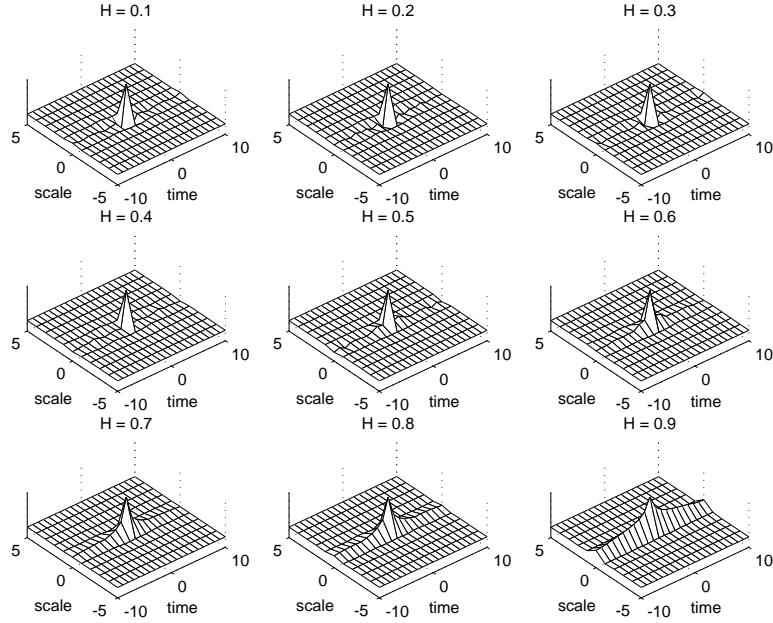


Figure 3.7: Two-dimensional correlation function of the IMF matrix in the case of fractional Gaussian noise. For each Hurst exponent H , the graph displays the quantity $|\hat{D}_H[k', n']|$ given by (3.9) as a function of time and scale (IMF index lag).

case, the interpretation of these modes requires some caution. With regard to the first IMF, the selected main frequency band is fully data-driven and automatically adapted to the highest frequency resonance. On the other hand, as noted earlier with regards to the non-negligible capture of low frequencies by IMF 1, some contributions at lower frequencies also occur. These contributions may include resonances at lower frequencies; they can also correspond to artifacts which must be compensated by IMFs of higher orders. These situations may be identified by examining intermodal correlation coefficients: The larger the correlation is, the less significant is the splitting into separate components. A quantitative evaluation of this inter-modal correlation is plotted in Fig. 3.8, where

$$\Theta[k, k'] := \left| \frac{1}{J} \sum_{j=1}^J \frac{\Theta^{(j)}[k, k']}{\sqrt{\Theta^{(j)}[k, k] \Theta^{(j)}[k', k']}} \right| \quad (3.10)$$

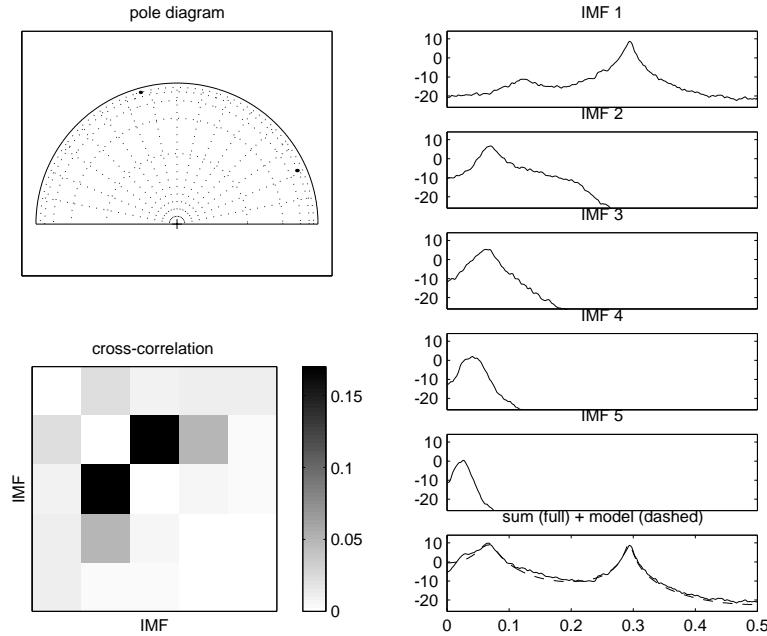


Figure 3.8: EMD as a spectrum analyzer. In the case of an AR(4) process, whose pole constellation in the upper half unit circle is plotted on the top left diagram, the ensemble averaged spectral analysis (over 50 realizations) of the first five IMFs is given in the right column of the figure. The bottom frame displays the cumulative spectrum (solid line) obtained by summing up the five spectra and compares it to the model's spectrum (dotted line). The bottom left diagram is a schematic of the normalized IMF cross-correlation (3.10) with the (unit) diagonal artificially forced to be zero so as to enhance the gray scale dynamic range (IMF indices grow from left to right and from top to bottom).

and

$$\Theta^{(j)}[k, k'] := \frac{1}{N} \sum_{n=1}^N d_{k,H}^{(j)}[n] d_{k',H}^{(j)}[n].$$

By definition, we have $0 \leq \Theta[k, k'] \leq 1$. Figure 3.8 clearly shows that the non-negligible values (as compared to 1) of $\Theta[k, k']$ correspond to index pairs (k, k') for which the IMF DSPs have a large amount of frequency overlap.

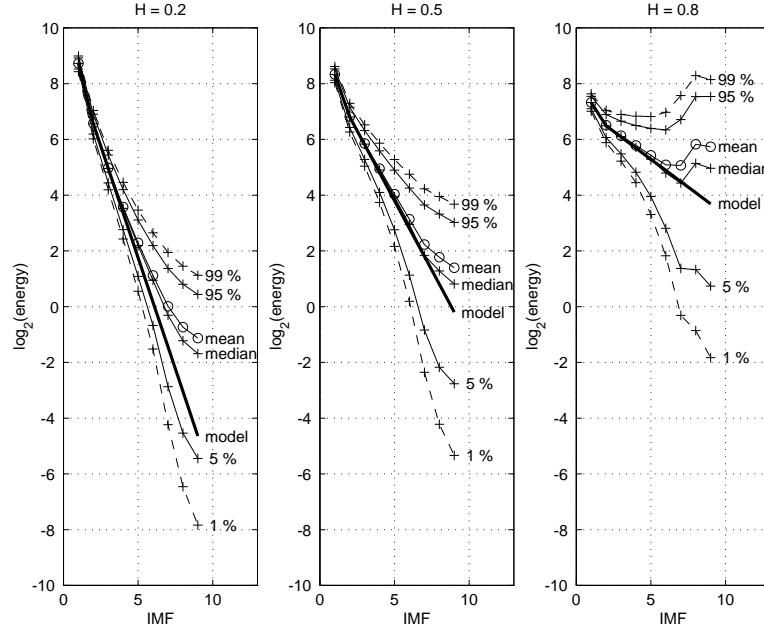


Figure 3.9: Experimental “modegrams” in the case of fractional Gaussian noise. For three different values of the Hurst exponent, the statistical characteristics (mean, median, confidence intervals) of the logarithm of the estimated EMD variance have been plotted as a function of the IMF index. The linear model (3.11) is also plotted. (Originally published in *Int. J. Wavelets Multiresolut. Inform. Process.*, **2**, 477–496, ©2004 World Scientific.)

3.4.3. Denoising and detrending with EMD

A detailed knowledge of IMF statistics in situations where noise is present can help in gauging the significance of a given mode. This idea, which has been pioneered by Wu and Huang (2004), can be used to separate a signal from noise. Two possible methods, namely *denoising* (by removing those modes identified as noise) and *detrending* (by keeping only them), can be used.

With regards to the variability of the variance estimate, Fig. 3.6 gives a rough, second-order indication based on the observed standard deviation. A greater appreciation can be gained from Fig. 3.9, in which the experimental mean, median and various confidence intervals are plotted for $H = 0.2, 0.5$

and 0.8, as well as

$$\log_2 V_H[k] = \log_2 \hat{V}_H[2] + 2(H-1)(k-2) \log_2 \rho_H \quad (3.11)$$

for $k \geq 2$, which was derived from (3.6). This series of simulations (which was carried out on 10 000 realizations of 2048 data points in each case) shows increasingly larger fluctuations for modes as the indices increase.[†] This finding agrees with (and is a generalization of) the findings reported in Wu and Huang (2004) for the case of white noise. Figure 3.10 also suggests that we may parameterize $T_H[k]$ by using the formula:

$$\log_2 (\log_2 (T_H[k]/W_H[k])) = a_H k + b_H, \quad (3.12)$$

where $W_H[k]$ denotes the H -dependent variation of the IMF energy. As noted earlier, the best linear fit occurs when the median of the IMF's energy is used to compute $W_H[k]$ over the realizations. The parameters a_H and b_H can be deduced from the simulation results in Fig. 3.9, and their values are reported in Table 3.1. In practice, $W_H[1]$ can be estimated from

$$\hat{W}_H[1] = \sum_{n=1}^N d_{1,H}^2[n], \quad (3.13)$$

and subsequent values of $W_H[k]$ are given by

$$\hat{W}_H[k] = C_H \rho_H^{-2(1-H)k}, \quad k \geq 2, \quad (3.14)$$

where $C_H = \hat{W}_H[1]/\beta_H$. The parameter β_H used to compute C_H can, in turn, be estimated from the data displayed in Fig. 3.9, and its values are given in Table 3.1.

Given these results, a possible strategy for denoising a signal corrupted by fGn (with a known H) is as follows:

Table 3.1: Confidence Interval Parameters for the Linear Model (3.12).

H	β_H	$a_H(95\%)$	$b_H(95\%)$	$a_H(99\%)$	$b_H(99\%)$
0.2	0.487	0.458	-2.435	0.452	-1.951
0.5	0.719	0.474	-2.449	0.460	-1.919
0.8	1.025	0.497	-2.331	0.495	-1.833

[†]The skewed (marginal) distribution of these “modegrams” yields better agreement if the linear model (3.12) uses the median rather than the mean of the realizations.

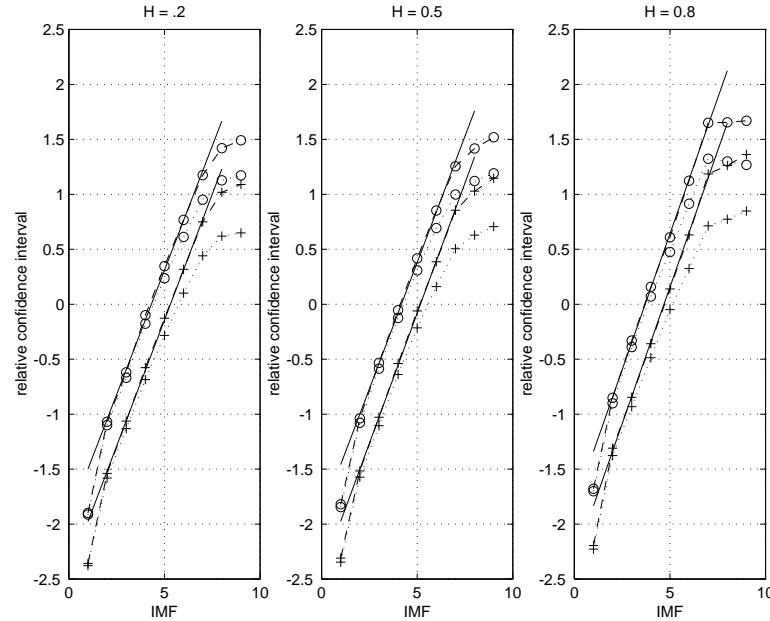


Figure 3.10: Relative confidence intervals. The (base 2) logarithm of the relative confidence intervals given in Fig. 3.9, i.e., $\log_2(\log_2(T_H[k]/W_H[k]))$, behaves essentially linearly as a function of the IMF index k , suggesting (3.12). For each of the three values of H , the crosses (circles) correspond to a confidence interval of 95% (99%), the dotted (dashed) lines refer to the cases where the reference $W[k]$ is chosen by using the mean (median) of the IMF energies over the realizations, and the solid lines indicate the corresponding best linear fit.

- (1) Assuming that the first IMF captures most of the noise, estimate the noise level in the noisy signal by computing $\hat{W}_H[1]$ from (3.13).
- (2) Estimate the “noise only” model by using (3.13) and (3.14).
- (3) Estimate the corresponding model for a chosen confidence interval from (3.12) and Table 3.1.
- (4) Compute the EMD of the noisy signal, and compare the IMF energies by using the confidence interval as a threshold.
- (5) Compute a partial reconstruction by keeping only the residual and those IMFs whose energy exceeds the threshold.

An alternative strategy for detrending fGn-type noise process consists of computing the complementary partial reconstruction based on only those IMFs whose energy is below the threshold.

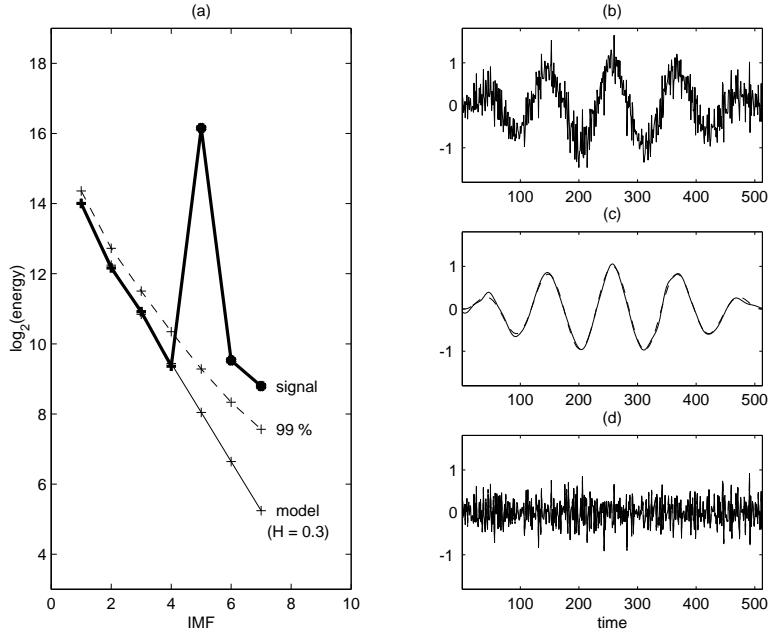


Figure 3.11: Denoising and detrending. An example of an amplitude-modulated, low-frequency oscillation embedded in fractional Gaussian noise with a Hurst exponent $H = 0.3$ is plotted in (b). The estimated energies of the seven IMFs are plotted in (a) as the thick line, together with the “noise only” model (thin line) and the 99% confidence interval (dotted line). The partial reconstruction obtained by adding the EMD residual and IMFs 5 to 7 [only those whose energies exceed the threshold in (a)] is plotted in (c) as a solid line and is superimposed on the actual signal component (dotted line). The partial reconstruction of IMFs 1 to 4 is plotted in (d).

A simple example of the EMD approach to denoising and detrending is given in Fig. 3.11, which presents the case of an oscillatory, low frequency waveform embedded in fractional Gaussian noise. A companion example containing actual data (heart-rate variability) is given in Fig. 3.12.

3.5. Concluding remarks

We have shown that EMD achieves a specific form of hierarchical filtering. This result is in agreement with the intuition associated with the EMD principle. However, because EMD still lacks a sound theoretical foundation, a careful and detailed analysis based on extensive numerical simula-

tions was necessary for asserting and quantifying this behavior. In the cases shown here, we observed the “spontaneous” emergence of an equivalent filter bank structure which has the advantage of being fully data-driven. Furthermore, because it is local in time, this structure can adapt automatically to nonstationary situations with greater flexibility than other approaches using a pre-determined decomposition scheme. Although some possible applications have been outlined, their potential usefulness will require further studies that compare EMD to alternative methods for specific tasks while endeavoring to make the theory more rigorous.

References

- Abry, P., P. Flandrin, M. Taqqu, and D. Veitch, 2000: Wavelets for the analysis, estimation, and synthesis of scaling data. *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., Wiley, 39–88.
- Coughlin, K. T., and K. K. Tung, 2004: 11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.*, **34**, 39–88.
- Embrechts, P., and M. Maejima, 2002: *Selfsimilar Processes*. Princeton Univ. Press, 111 pp.
- Flandrin, P., 1999: *Time-Frequency/Time-Scale Analysis*. Academic Press, 386 pp.
- Flandrin, P., and P. Gonçalvès, 2003: Sur la décomposition modale empirique. *Proc. 19ème Coll. GRETSI sur le Traitement du Signal et des Images*, Paris, 149–152.
- Flandrin, P., and P. Gonçalvès, 2004: Empirical mode decompositions as data-driven wavelet-like expansions. *Int. J. Wavelets Multiresolut. Inform. Process.*, **2**, 477–496.
- Flandrin, P., G. Rilling, and P. Gonçalvès, 2004: Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.*, **11**, 112–114.
- Fournier, R., 2002: *Analyse stochastique modale du signal stabilométrique. Application à l'étude de l'équilibre chez l'Homme*. Thèse de Doctorat, Univ. Paris XII Val de Marne, 263 pp.
- Huang, N. E., Z. Shen, S. R. Long, M. L. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Mandelbrot, B. B., and J. W. van Ness, 1968: Fractional Brownian motions,

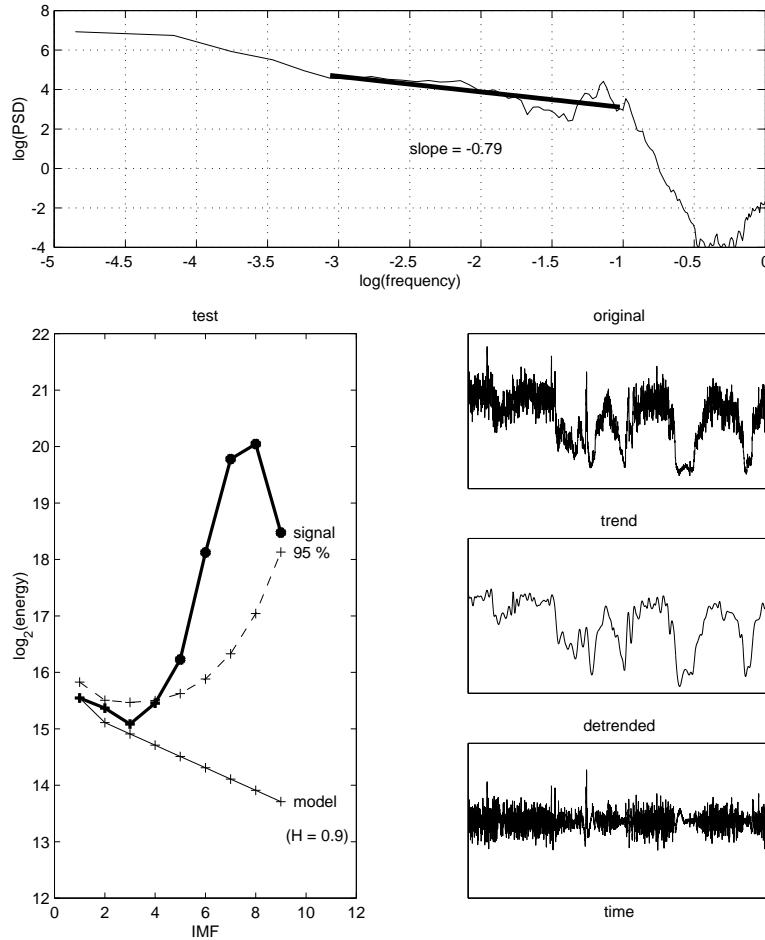


Figure 3.12: Denoising and detrending of a heart-rate signal. Top diagram: Signal spectrum in log-log coordinates (thin line) with a linear fit of slope $p \approx -0.8$ in the mid-frequency range (thick line), supporting a fGn model with Hurst exponent $H = (1 - p)/2 \approx 0.9$. Bottom diagram: Model-based detrending. Left: Estimated energy of nine IMFs, plotted as the thick line, together with the “noise only” model given by $H = 0.9$ (thin line) and the 95% confidence interval (dotted line). Top right: Original signal. Middle right: Estimated trend obtained from the partial reconstruction with IMFs 5 to 9 (only those whose energies exceed the threshold in the left diagram) and the residual. Bottom right: Detrended signal obtained from the partial reconstruction with IMFs 1 to 4.

- fractional noises and applications. *SIAM Rev.*, **10**, 422–437.
- Mallat, S., 1998: *A Wavelet Tour of Signal Processing*. Academic Press, 577 pp.
- Neto, E. P. S., M. A. Custaud, C. J. Cejka, P. Abry, J. Frutoso, C. Gharib, and P. Flandrin, 2004: Assessment of cardiovascular autonomic control by the empirical mode decomposition. *Method. Inform. Med.*, **43**, 60–65.
- Rilling, G., P. Flandrin, and P. Gonçalvès, 2003: On empirical mode decomposition and its algorithms. *IEEE-EURASIP Workshop on Nonlinear Signal Image Process. NSIP-03*, Grado, Italy.
- Wood, A. T., and G. Chan, 1994: Simulation of stationary processes in $[0, 1]^d$. *J. Comp. Graph. Stat.*, **3**, 409–432.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.
- Wu, Z., E. K. Schneider, Z. Z. Hu, and L. Cao, 2001: The impact of global warming on ENSO variability in climate records. *COLA Technical Report, CTR 110*, 25 pp.

Patrick Flandrin

Laboratoire de Physique (UMR 5672 CNRS), Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07 France
flandrin@ens-lyon.fr

Paulo Gonçalvès

INRIA Rhône-Alpes, On leave at IST-ISR, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
Paulo.Goncalves@inria.fr

Gabriel Rilling

Laboratoire de Physique (UMR 5672 CNRS), Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07 France
grilling@ens-lyon.fr

CHAPTER 4

HHT SIFTING AND FILTERING

Reginald N. Meeson, Jr.

Time-frequency analysis is the process of determining what frequencies are present in a signal, how strong they are, and how they change over time. Understanding how the frequencies in a signal change with time can explain much about the physical processes that generate or influence the signal. Better resolution of the details of frequency changes provides better insight into these underlying physical processes.

The Hilbert-Huang transform (HHT) offers higher frequency resolution and more accurate timing of transient and non-stationary signal events than conventional integral transform techniques. The HHT separates complex signals into simpler component signals, each of which has a single, well-defined, time-varying frequency. Real-time HHT algorithms enable this enhanced signal analysis capability to be used in process monitoring and control applications.

“Sifting” is the central signal separation process of the HHT algorithm. This chapter compares the component signal separations of Huang’s sifting process with those produced by filtering techniques.

Although intuition seems to suggest that filtering, with appropriate real-time adjustments to parameters, could be substituted for Huang’s sifting process, our results did not support this suggestion. Five case studies present HHT and filtering results for stationary amplitude- and frequency-modulated signals, as well as signals with more dynamic transient behavior. These examples show that, in general, HHT sifting and filtering separate signal components quite differently. Our experiments with example signals led to the discovery of aliasing in the HHT sifting algorithm.

4.1. Introduction

One way to describe a timed series of measurements, referred to as a “signal,” is in terms of the frequencies in its variations. The process of determining what frequencies are present, how strong they are, and how they change over time is called “time-frequency analysis.” Conventional time-frequency

analysis techniques use integral calculus transforms to map time-based signals into frequency-based or joint time- and frequency-based representations. Examples of these techniques include Fourier transforms, windowed Fourier or Gabor transforms, wavelet transforms, and joint time-frequency distributions [see Cohen (1995) for a thorough introduction to these techniques].

The Hilbert-Huang transform (HHT) is a new time-frequency analysis technique that offers higher frequency resolution and more accurate timing of transient and non-stationary signal events than conventional Fourier and wavelet transform techniques. This approach was introduced by Huang (1998). Conventional techniques assume signals are stationary, at least within the time window of observation. Fourier analysis assumes further that the signal is harmonic and repeats itself with a period exactly matching the width of the sampling window. These analysis techniques are employed widely even though their (theoretically necessary) enabling conditions rarely hold for signals of interest.

In addition, integral transform techniques suffer from an uncertainty problem similar, mathematically, to Heisenberg's uncertainty principle in physics. This uncertainty limits their ability to accurately measure timing and frequency at the same time. That is, after a point, higher-resolution frequency measurements cannot be achieved without sacrificing timing accuracy, and vice versa. The HHT is able to resolve frequencies accurately and time them precisely without this limiting uncertainty.

The original HHT algorithm was formulated as a "batch" computation, in which a complete dataset is collected and then processed as a whole. An incremental algorithm that transforms evolving input data streams into streams of HHT results has also been developed [see Meeson (2002)]. Modern microprocessors and signal processing chips offer sufficient performance for this incremental algorithm to be used in many real-time applications. For this study the incremental algorithm served as a bridge connecting the original HHT algorithm to the incremental filtering techniques.

"Sifting" is the central signal separation process of the HHT algorithm. In the seminal work on the HHT, Huang (1998) described sifting informally as analogous to an adaptive filtering process, but then developed a different algorithmic procedure to separate signal components. This development led us to conjecture that filters, with parameters appropriately adjusted in real time, could mimic the HHT sifting process. It seemed natural to analyze the results from filtering and to compare them with the HHT results. Huang's original HHT sifting algorithm was the starting point for this comparison.

The results from the original and incremental HHT algorithms are virtually identical for the example signals used in this analysis.

“Filtering,” for this discussion, means conventional finite impulse response (FIR) digital filtering where filter coefficients and cutoff frequencies can be adjusted on a sample-by-sample basis. Our experiments with these signal-analysis techniques revealed new insights into the mathematical properties of the HHT signal separation process and may help refine HHT-processing techniques.

In section 4.2, we describe the objectives of the HHT signal-separation process and the desired attributes of separated components. Huang’s original empirical mode decomposition algorithm, which later became known as the HHT, is described in section 4.3. Section 4.4 describes the incremental HHT algorithm and analyzes a special case where an analogy to conventional digital filtering techniques can be used. In section 4.5, we describe the shift from special-case static filtering to a general method using dynamically adjustable filters. HHT and filtering results for five example signals are compared in section 4.6. Section 4.7 concludes with a summary and some directions for future research.

4.2. Objectives of HHT sifting

The HHT sifting process separates a signal into a series of amplitude- and frequency-modulated component signals in the form

$$s(t) = \sum_i a_i(t) \cos[\varphi_i(t)],$$

where $a_i(t)$ represents the amplitude modulation and $\varphi_i(t)$ denotes the phase functions that represent the frequency modulation characteristics of each component.

Numerous possible solutions are available for this separation scheme. One familiar solution is the Fourier series, which is made up of constant amplitude and constant frequency (linear phase) functions [see, for example, Oppenheim and Shafer (1989) for a discussion of the Fourier series]. The solution that the HHT seeks is quite different. Rather than trying to represent a signal in terms of predetermined basis functions, the HHT tracks and adapts dynamically to transient, non-stationary, and nonlinear changes in component frequencies and amplitudes as the signal evolves over time.

Windowed Fourier and wavelet-signal-analysis techniques are also able to track slowly changing signal behavior; but, as described above, they suffer

from an uncertainty problem that can limit the accuracy of the frequency (scale for wavelets) and timing information they yield. The product of the frequency (scale) variance and the timing variance for the results from these techniques has a positive lower bound. Consequently, once this limit is reached, increasing the accuracy of frequency measurements can be achieved only by sacrificing timing accuracy, and vice versa [see Cohen (1995) for a discussion of time and frequency uncertainty].

Many signals of interest contain short-duration transients that are difficult to analyze because of this uncertainty limitation. With conventional analysis techniques, it is not possible to accurately time when specific frequencies were present. Transient events can be timed accurately, but accurate frequency information cannot be resolved within that narrow time window.

HHT signal separations are not subject to this limitation and provide both accurate frequency and accurate timing simultaneously. HHT has this unique advantage over conventional time-frequency analysis techniques. HHT analysis of earthquake data, as described by Huang (2001), for example, shows a very different distribution of frequencies over time than conventional Fourier analysis. This difference may prove tremendously important in analyzing the strength of buildings, bridges, and other structures.

4.2.1. Restrictions on amplitude and phase functions

In order to extract the desired amplitude and frequency information, without conflicting interpretations or paradoxical results, restrictions must be imposed on the amplitude and phase functions, $a_i(t)$ and $\varphi_i(t)$. The primary requirement for HHT components is that they be sufficiently well behaved to allow extraction of well-defined amplitude and phase functions. Such functions are called “monocomponent” functions, and we distinguish them from “multicomponent” functions, from which amplitude and phase cannot be cleanly extracted. Although there seems to be no generally accepted mathematical definition of “monocomponentness,” there is little debate over one primary criteria, which is that at any time a monocomponent signal must have a single, well-defined, positive instantaneous frequency represented by the derivative of its phase function.

The first approach suggested for finding the necessary conditions for a separated component’s “monocomponentness” was to look at the component’s analytic signal, which is given by

$$\mathcal{A}[c(t)] = c(t) + i\mathcal{H}[c(t)],$$

where $c(t) = a(t) \cos[\varphi(t)]$, $\mathcal{H}(\cdot)$ is the Hilbert transform, and $i = \sqrt{-1}$ [see Cohen (1995) for a discussion of analytic signals and the Hilbert transform]. The analytic signal is a complex function whose Fourier transform is twice that of $c(t)$ over the positive frequencies and zero over negative frequencies. The spectrum of this signal, therefore, contains only positive frequencies. This fact does not guarantee, however, that the signal's instantaneous frequency (the derivative of its phase) will always be positive. Cohen (1995) shows examples of analytic signals that have paradoxical instantaneous frequency characteristics, including some with negative instantaneous frequencies. The analytic signal, therefore, by itself, does not appear to provide sufficient criteria for separating monocomponent signals.

A second approach suggested for finding monocomponent conditions was to consider the function's quadrature model:

$$\mathcal{Q}[c(t)] = a(t)e^{i\varphi(t)} = a(t)\{\cos[\varphi(t)] + i \sin[\varphi(t)]\}.$$

By using the additional knowledge about the Hilbert transform that $\mathcal{H}\{\cos[\varphi(t)]\} = \sin[\varphi(t)]$, the quadrature model can be compared with the analytic signal. The two are the same when the amplitude function can be factored out of the signal's Hilbert transform; i.e., when

$$\mathcal{H}\{a(t) \cos[\varphi(t)]\} = a(t)\mathcal{H}\{\cos[\varphi(t)]\} = a(t) \sin[\varphi(t)].$$

The conditions under which this relationship holds were established by Bedrosian (1963) and elaborated by Nuttall (1966). The conditions are, for some positive frequency ω_0 :

- a. The spectrum of the amplitude function is restricted to frequencies below ω_0 , and
- b. The spectrum of the cosine term is restricted to frequencies above ω_0 .

An example of a function that does not satisfy these conditions is

$$s(t) = 1.25 \cos(t) - \cos(2t).$$

The analytic signal of this function is similar to one of Cohen's problematic signals, $\mathcal{A}[s(t)] = 1.25e^{it} - e^{2it}$, which cannot be expressed in the form $a(t)e^{i\varphi(t)}$ without either $a(t)$ oscillating rapidly or $\varphi'(t)$ turning negative periodically. As can be seen in the graph shown in Fig. 4.1, the real signal $s(t)$ has local minima with positive values. Such a signal cannot be expressed in the form $a(t) \cos[\varphi(t)]$ with a slowly varying amplitude and an increasing phase function. If we assume a slowly varying amplitude, to satisfy Bedrosian's first condition, then the $\cos[\varphi(t)]$ term would have to

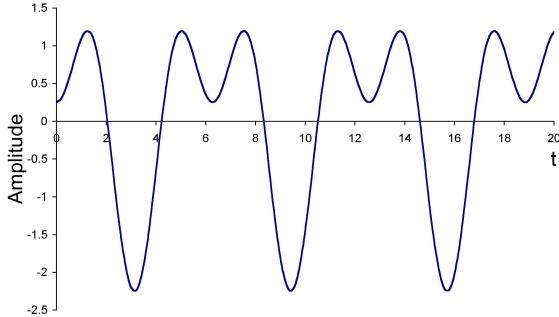


Figure 4.1: Example of a multicomponent signal.

turn and go back up again without going negative. The phase function, therefore, would have to decrease for a time, resulting in a negative instantaneous frequency. This result would violate Bedrosian's spectral separation conditions, since the amplitude function would have to have a negative upper-frequency limit. If we stipulate an increasing phase function, then the amplitude must peak near $t = (2n + 1)\pi$ and dip to a minimum near $t = 2n\pi$, giving an average frequency of $\omega = 1$, the same as the average change in phase. Either way, Bedrosian's condition is not satisfied.

Bedrosian's conditions are too restrictive for our needs, however. Purely frequency-modulated signals with constant amplitude can have spectra that extend down to zero frequency. Any amplitude modulation imposed on such a "carrier" signal would violate Bedrosian's conditions—even though the signal would make a perfectly good HHT component. The case studies below show that solutions must allow phase functions that exhibit this sort of frequency-modulated behavior.

Teager's energy operator, Ψ , was suggested by Maragos (1993) as a possible non-linear approach for restricting amplitude and phase functions for combined amplitude-modulated (AM) and frequency-modulated (FM) signals

$$\Psi(s(t), t) = [s'(t)]^2 - s(t)s''(t).$$

For component signals of the form $a(t) \cos[\varphi(t)]$, Ψ can be expanded as

$$\begin{aligned} \Psi\{a(t) \cos[\varphi(t)], t\} &= [a(t)\varphi'(t)]^2 + \frac{1}{2}a^2(t) \sin[2\varphi(t)]\phi''(t) \\ &\quad + \cos^2[\varphi(t)]\Psi[a(t), t]. \end{aligned}$$

If a signal has a dominant high-frequency component, the first term in this formula will dominate the others. Maragos (1993) describes the secondary

terms as “error” terms and shows how they can be minimized by constraining the AM and FM indexes of modulation, and the modulating signal bandwidth.

The integrals of the two terms in Teager’s Ψ operator are both equal to the signal’s total energy times its average square frequency; i.e.,

$$\int [s'(t)]^2 dt = - \int s(t)s''(t) dt = \int \omega^2 |S(\omega)|^2 d\omega = E\langle\omega^2\rangle,$$

where $S(\omega)$ is the signal’s Fourier transform, E is its total energy,

$$E = \int [s(t)]^2 dt = \int |S(\omega)|^2 d\omega,$$

and $\langle\omega^2\rangle$ is the average square frequency.

Instantaneously, though, Teager’s two terms are quite different. Ψ may not even yield positive results. For the signal in Fig. 4.1, for example, the values of Ψ are negative in the vicinity of $t = 2n\pi$ (where $s'(t) < 0$, $s(t) > 0$, and $s''(t) > 0$).

For lightly modulated signals, Ψ produces a stable output dominated by $[a(t)\varphi'(t)]^2$. Maragos (1992) showed that, as long as the “error” terms are sufficiently small, Ψ can be used to demodulate the signal and extract approximate values for $a(t)$ and $\varphi'(t)$ by applying Ψ to the signal

$$\Psi[s(t), t] = \Psi\{a(t) \cos[\varphi(t)], t\} \approx [a(t)\varphi'(t)]^2$$

and to its derivative

$$\Psi[s'(t), t] \approx a^2(t)[\varphi'(t)]^4.$$

Teager’s formula appears to offer possibilities for identifying signals that would satisfy our general notion of monocomponentness. Turning these results into algorithms for separating monocomponent signals from more complex ones, however, is still an open problem.

We proceed from this point without a concrete definition of “monocomponentness,” but we recognize that it implies constraints on phase monotonicity ($\varphi'(t) > 0$), amplitude and “carrier” signal bandwidth, and degrees of amplitude and frequency modulation.

4.2.2. End-point analysis

An area of interest to many scientists is the extraction of frequency information at the very beginning and at the very end of the data they collect. We have not pursued this problem and are skeptical about prospects for significant advances in this area. Cohen (1995) states that the frequency content

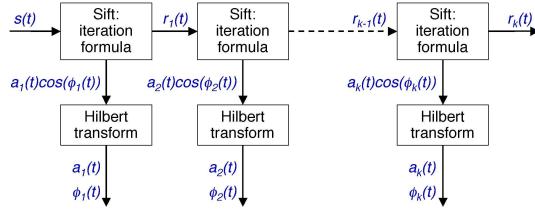


Figure 4.2: Diagram of the HHT signal separation process.

of a signal at any point in time depends entirely on the context of its behavior both before and after the time in question. In the absence of data providing this context, assumptions could be made about the signal's probable past and future behavior, but the analysis would then merely reflect these assumptions. The primary reason for our interest in time-frequency analysis techniques for transient and non-stationary signals is because their behavior is unpredictable. We conclude, therefore, that unless the necessary assumptions and predictions are strongly supported by additional knowledge about the physical processes underlying the signal, end-point analyses should be treated as at least somewhat suspect.

4.3. Huang's sifting algorithm

Huang's sifting process [see Huang (1998)] separates the highest-frequency component embedded in a multicomponent signal from all the lower-frequency components. The remaining lower-frequency components together make up the signal trend. A signal can be described in terms of its first component and residual trend functions by

$$s(t) = a_1(t) \cos[\varphi_1(t)] + r_1(t).$$

The sifting process for a single component is repeated by using the trend output from one stage as the input for the next, producing the series of $a_i(t) \cos[\varphi_i(t)]$ terms that add up to reconstruct the original signal, $s(t)$. A diagram of this process is shown in Fig. 4.2.

To determine $r(t)$, Huang fit smooth envelope curves (using cubic splines) to the local maxima of the signal and to the local minima. The average of these two envelopes provides a rough estimate of $r(t)$. (Local maxima are referred to as “positive peaks” even though the signal values at those points may be positive or negative. Local minima are similarly referred to as “negative peaks.”) Huang then applied an iteration scheme

to refine the estimated trend. The iteration scheme can be formulated as

$$r_{(n+1)}(t) = r_{(n)}(t) + \rho(c_{(n)}, t),$$

where $c_{(n)}(t) = s(t) - r_{(n)}(t)$. The function ρ represents the spline curve fitting and averaging process applied to the peaks of function $c_{(n)}(t)$. (Subscripts in parentheses indicate the iteration count.) This calculation is repeated (starting with $r_{(0)}(t) = 0$) until a fixed point is reached, and $\rho(c_{(n)}, t)$ converges to zero (within some small ϵ). Once the residual or trend function is determined, the difference between it and the input signal is the highest-frequency separated component, $c_i(t) = a_i(t) \cos[\varphi_i(t)]$.

To separate the $a_i(t)$ and $\varphi_i(t)$ functions, Huang computed the component's analytic signal by using Fourier transforms. The Fourier transform of a function's Hilbert transform satisfies the relation

$$\mathcal{F}\{\mathcal{H}[s(t)]\} = -i \operatorname{sign}(\omega) \mathcal{F}[s(t)],$$

where $\mathcal{F}(\cdot)$ is the Fourier transform and $\mathcal{H}(\cdot)$ is the Hilbert transform. The Fourier transform of a function's analytic signal can then be formulated as

$$\mathcal{F}\{\mathcal{A}[s(t)]\} = \mathcal{F}[s(t)] + \operatorname{sign}(\omega) \mathcal{F}[s(t)],$$

which is zero for all negative frequencies and double the input signal's values for all positive frequencies.

Taking a separated component's Fourier transform, zeroing its negative-frequency terms and doubling its positive-frequency terms, and then applying the inverse Fourier transform, produces the component's complex analytic signal. The magnitude of the analytic signal is $a(t)$ and the argument is $\varphi(t)$.

Huang called this separation technique “empirical mode decomposition,” and the individual component signals “intrinsic mode functions.” His colleagues later named the method “the Hilbert-Huang transform.”

Components separated by this process are well behaved, although, because the process is defined only in terms of this algorithm, the mathematical monocomponentness properties they satisfy are not easily determined. The work described below represents an attempt to link the HHT results to filtering, the mathematical properties of which are well established.

4.4. Incremental, real-time HHT sifting

In Huang's original HHT algorithm, the data passed between the processing blocks in Fig. 4.2 are arrays containing an entire time series. The incremental algorithm [see Meeson (2002)] turns these batch-processing blocks into

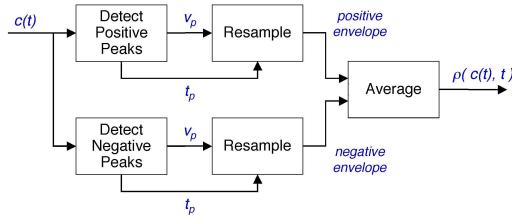


Figure 4.3: Diagram of one iteration of ρ .

pipeline processes that operate incrementally on streams of data, passing one data sample at a time.

The first step in sifting is to identify signal peaks. The calculation of peak values and times in the incremental HHT algorithm is the same as in Huang's original algorithm, except that peak value and time pairs, $\langle v_p, t_p \rangle$, are produced incrementally as the input stream evolves. The resulting stream of peak values corresponds to sampling the input signal at its peak times rather than at uniform intervals.

Spline interpolation uses global information to calculate the derivative of the positive envelope at each positive peak and for the negative envelope at each negative peak. For incremental processing, only local information is available, so we must rely on Hermite interpolation, which is nearly identical to spline interpolation but uses derivative values estimated from local signal behavior [see, for example, Kincaid (1991) for a thorough discussion of interpolation techniques].

By using the spline parameters derived for each segment of the positive peak envelope, values are calculated at points corresponding to the signal's original sample times. This resampling process produces a stream of uniformly sampled envelope values, although with some latency from the peak-detection and spline-interpolation process. The same process is applied to the negative-peak data. The two resampled envelope streams are then averaged to produce a stream of trend values. This process, diagrammed in Fig. 4.3, represents one application of Huang's ρ function. Each stage of this process is performed incrementally, so the calculation of ρ is achieved incrementally.

4.4.1. Testing for iteration convergence

The iteration process involving repeated applications of ρ does not always exhibit smooth convergence. Removing the residual or trend component

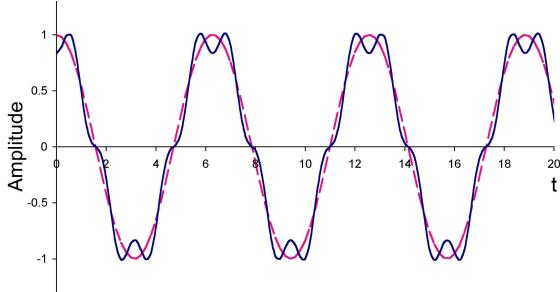


Figure 4.4: Example of a signal containing hidden peaks.

occasionally exposes new peaks that appeared only as inflections in the input signal. An example where these peaks occur is illustrated by the signal

$$s(t) = \cos(t) - 0.167 \cos(5t).$$

The graph of this function is shown in Fig. 4.4. The trend function produced by ρ , also shown in the figure (dashed line), cuts through the inflection points in the signal as it crosses the axis and produces new peaks in the next iteration $c(t)$ that were not present in the input for this iteration. These new peaks are included in all further iterations.

The discovery of new peaks introduces highly nonlinear disturbances in both the high-frequency component and the trend that may require several additional iterations to smooth out. These disturbances can occur even when the trend has nearly converged to a fixed point.

Huang devised a global test for convergence that spans the entire signal duration. This solution, however, is not in keeping with our objective of incremental processing. We have not yet found a satisfactory incremental test for convergence. In practice we have used fixed-length chains of ρ operations, making them long enough so that errors from terminating the iteration too early are rare. While not ideal for real-time performance, the results are comparable to those obtained by using the original HHT algorithm.

4.4.2. Time-warp analysis

If the peaks of an input signal are uniformly spaced, a number of simplifying assumptions can be made in the sifting process. These assumptions do not apply in general, so this approach cannot be used to process arbitrary

signals, but the analysis provides insights that can be generalized.

Disregard, for the moment, the timing information that accompanies the incremental stream of peak values described above, and assume these peak values have been sampled at some uniform rate. The distortion this sampling introduces is referred to as a “time warp,” since the actual peak times in general are not uniformly spaced. Although all of the nonlinear phase information between peaks in the original signal is lost (for the moment), the trend of the warped signal can be easily calculated by using standard low-pass digital filtering techniques.

In the warped world, one iteration of Huang’s fixed-point function, ρ , for a series of warped peak values v at time t_p , corresponds to the expression

$$\rho(v, t_p) = \frac{v_p}{2} - \frac{v_{p-3}}{32} + \frac{9v_{p-1}}{32} + \frac{9v_{p+1}}{32} - \frac{v_{p+3}}{32}.$$

This expression is the average of the two envelopes, one of which is represented by v_p , and the other is interpolated from the spline curve derived from the neighboring opposite-sign peaks (v_{p-3} , v_{p-1} , v_{p+1} and v_{p+3}) at time t_p . This expression corresponds to a simple low-pass digital filter, which has the frequency response shown in Fig. 4.5. This graph shows that the transition band for one pass through this filter crosses at approximately one-half of the warped signal’s Nyquist frequency.

If the timing of peaks does not change from iteration to iteration, multiple iterations correspond to passing the signal through this filter multiple times. (The timing of peaks may change slightly, usually in the initial iterations.) Multiple passes through a simple filter are equivalent to a single pass through a larger filter [see, for example, Parks (1987) for a discussion of filter composition]. Huang’s iteration scheme is formulated so that the high-pass filter is iterated and its iteration successively reduces the resulting pass band. The corresponding longer low-pass filters have wider pass-band regions and sharper transitions to the stop band. Examples of transfer functions for filters representing different iterations of ρ are shown in Fig. 4.5.

The original HHT algorithm uses the shrinking corrections of the iteration process to judge when it has converged. This method corresponds to choosing the characteristics of filters dynamically, based on the signal’s behavior. Figure 4.5 shows that each iteration of ρ shifts the filter transfer function to a higher-frequency cutoff point. Note also that successive iterations have less and less effect on the size of the frequency shift. Rather than iterate the simple filter corresponding to ρ , we wish to determine the filter characteristics necessary to directly satisfy the monocomponent criteria

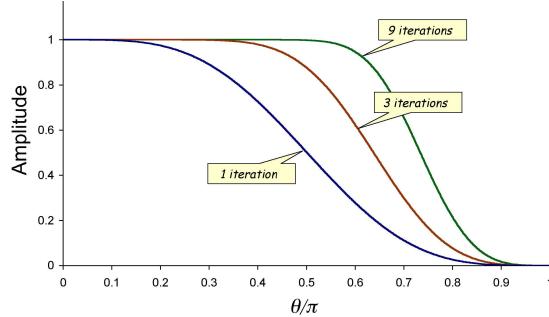


Figure 4.5: Frequency response of HHT trend-estimating process.

and separate the component from the trend in a single pass.

4.4.3. Calculating warped filter characteristics

Consider that the separated warped signal can be described by $s_p = a_p + r_p$ for all positive peaks, and $s_p = -a_p + r_p$ for all negative peaks, where a_p is the absolute value of the high-pass filter output, and r_p is the low-pass filter output for each peak. The a_p values are interpreted as approximating a warped sampling of the amplitude function, $a(t)$. The r_p values are similarly interpreted as a warped sampling of the residual function, $r(t)$.

The spectrum of the warped residual function is controlled by the low-pass filtering effects of multiple iterations of ρ . This same filtering process also controls the spectrum of the warped amplitude function. The spectrum of the series of a_p values is shifted upward by the modulating effects of the warped “carrier” signal, $\cos(p\pi)$. The spectrum captured by the high-pass filter, therefore, is that of the amplitude function shifted upward by π . If $R(\theta)$ is the low-pass filter transfer function for the r_p values, then the corresponding transfer function for the a_p values is

$$A(\theta) = 1 - R(\pi - \theta).$$

This relationship, for an idealized separation filter, is shown in Fig. 4.6. [The transfer function for the high-pass filter is shown as $C(\theta)$.] From these graphs we can see that, to satisfy Bedrosian and keep the $\cos(p\pi)$ and a_p spectra from overlapping, the stop band breakpoint for the high-pass filter must be no lower than half the warped Nyquist frequency.

Ten iterations of this filter would reduce the effective filter throughput at one-half the warped Nyquist frequency to approximately $2^{-10} (\approx 10^{-3})$,

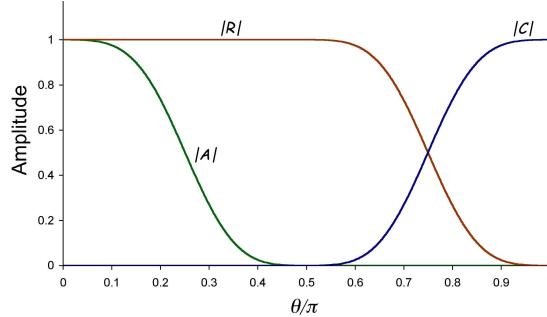


Figure 4.6: Warped filter transfer functions

which should satisfy Bedrosian's separation criteria for many practical purposes. Iterating the simple warped filter or substituting a more efficient filter, however, will not adapt to the frequency changes introduced by new peaks. In practice, we have often encountered signals that require 25 to 30 iterations of Huang's ρ operator to converge. Much of this disparity in iteration counts is attributable to the nonlinear disturbances caused by the discovery of new peaks.

4.4.4. Separating amplitude and phase

To separate the amplitude and phase functions incrementally, we substituted a Hilbert transform filter for the batch Fourier transform process described earlier for calculating analytic signals. A Hilbert transform filter has a transfer function that approximates the Fourier transform of a signal's Hilbert transform: $H(\omega) = -j\text{sign}(\omega)$ [see, for example, Parks (1987) for a discussion of Hilbert transformer filter design]. For a monocomponent signal, $a(t) \cos[\varphi(t)]$, this filter approximates

$$h(t) * \{a(t) \cos[\varphi(t)]\} = a(t) \sin[\varphi(t)],$$

where $h(t)$ represents the Hilbert transform filter coefficients and “*” represents convolution. The amplitude and phase functions are easily separated by using this result:

$$a(t) = \sqrt{\{a(t) \sin[\varphi(t)]\}^2 + \{a(t) \cos[\varphi(t)]\}^2},$$

and

$$\varphi(t) = \arctan\{\sin[\varphi(t)] / \cos[\varphi(t)]\}.$$

Once the phase function is extracted, the signal's instantaneous frequency is calculated by passing $\varphi(t)$ through a differentiating filter (after compensating for the discontinuities in the arctan results). All of these calculations are done incrementally.

The band-limiting effects of warped filtering on the amplitude envelope indicate that $a(t)$ should be relatively smooth. That is, we expected $a(t)$ to look like the smooth spline-connected envelopes calculated in the final iteration of ρ in the sifting process, with all of the high-frequency content captured by the phase function, $\varphi(t)$. Both the Hilbert transform filter and the Fourier batch technique, however, were found to introduce a high-frequency "ripple" in the amplitude results for some signals.

The explanation for this seeming anomaly is that, within certain limits, the spectral energy of a combined amplitude- and frequency-modulated signal can be freely exchanged between the amplitude and phase functions. While we expected a band-limited amplitude, the Hilbert transform appears to split the difference, sharing the high-frequency content between the amplitude and phase functions. The result, therefore, is sometimes a bit different from what we expected, but is an equivalent representation of the signal.

We experimented with a number of different possible techniques for separating the amplitude and phase, including the use of Teager's energy operator. None of these other techniques were as successful as the use of the Hilbert transform filter. Teager's operator worked well for the signal itself, but occasionally produced negative results for the derivative of the signal, spoiling Maragos's (1992) demodulation approach. Boashash (1992) provides an extensive discussion of additional techniques for extracting a signal's instantaneous frequency.

4.5. Filtering in standard time

The next objective, to test our conjecture about filtering substituting for HHT sifting, was to reproduce the effects of Huang's ρ operation in standard time, without resampling the original input signal. In the process, we wanted to avoid the unreasonable time-warp analysis assumptions about uniformly spaced peaks. The question posed was, Is there a corresponding standard-time filter that will isolate a comparable (unwarped) trend function and, if so, what are its characteristics? Any filter that approximates this response will have to change its attributes over time (possibly every few samples) to track transient and non-stationary changes in the signal.

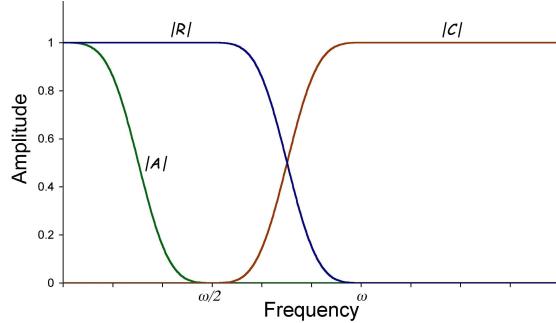


Figure 4.7: Filter transfer functions.

The transfer function for this low-pass filter is shown schematically in Fig. 4.7 as $R(\theta)$. The transfer function for the complementary high-pass filter for the $a(t) \cos[\varphi(t)]$ term is shown as $C(\theta)$. This filtering should also leave the spectrum of the amplitude function as shown by $A(\theta)$, maintaining Bedrosian's separation from the minimum frequency of the $\cos[\varphi(t)]$ term. All we have to do is determine the breakpoint frequencies, ω and $\omega/2$, for these filters and calibrate the horizontal scale.

The spectrum of the $a(t) \cos[\varphi(t)]$ term will, in general, contain both AM and FM components. Amplitude modulation of a constant-frequency “carrier” signal shifts the spectrum of the amplitude signal from the origin to the carrier frequency. If $A(\theta)$ is the spectrum of $a(t)$, then the spectrum of $a(t) \cos(\omega_c t)$ will be $A(\theta + \omega_c)$, where ω_c is the carrier frequency. Frequency modulation redistributes the spectrum of its modulating signal in much more complex ways.

In a combined AM and FM signal, the FM spectrum overlaps and mixes with the AM spectrum so that separating the two components by using a simple linear process (like conventional filtering) does not appear promising. The HHT process, however, is able to make a separation, although not always in exactly the same form as that used to formulate sample inputs. (Recall that solutions satisfying the HHT monocomponent separation criteria are not unique.)

As a first approximation for the breakpoint for the high-pass pass band, we used the minimum peak-to-peak frequency of the signal over the time span covered by the filter.[§] This frequency is marked as ω on the

[§]We note that Bedrosian's spectral separation criteria, being based on integral transform

axis in Fig. 4.7. The pass band breakpoint for the high-pass filter was set to this frequency. The stop band breakpoint, based on our experience with warped filtering, was set to one-half this frequency. As signals pass through the filter, their peak-to-peak frequencies are monitored, and the filter coefficients are adjusted to track any changes.

4.6. Case studies

In this section we present five case studies that illustrate and compare the results produced by the HHT and filtering approaches. The first example is a simple composite signal that serves as a reference for comparison with the second example. The second example is a steady-state AM signal. The third example is a steady-state FM signal. The fourth and fifth examples contain unit step changes in amplitude and frequency, respectively, and allow us to begin to explore the dynamic capabilities of the HHT and filtering mechanisms.

4.6.1. Simple reference example

The first example is a simple combination of constant amplitude sinusoids defined by

$$s(t) = \cos(t) + 0.5 \cos(t/2).$$

The graph of this function is shown in Fig. 4.8, along with the signal trend (dotted line). The maximum timing between peaks is slightly greater than π , indicating the need for high- and low-pass filters with upper breakpoint frequencies at $\omega = 0.97$. The result of filtering this signal, because of our selection of filter breakpoints, produces a nearly perfect separation of the two components, namely $c_1(t) = \cos(t)$ and $r_1(t) = 0.5 \cos(t/2)$. The HHT sifting process produces nearly identical results. One difference is that HHT sifting approximates the trend using splines, so its trend is represented by a series of cubic polynomials pieced together at the peaks. These small differences are of little concern here. Our primary interest in this simple signal is its similarity to the next example.

analysis, must hold (theoretically) for all time, not merely for the time span covered by the filter. We conjecture that this rather severe constraint can be relaxed by using more modern tight-frame analysis techniques. We have not completed the analysis to formally confirm this conjecture, however, and we proceed, taking it as an assumption.

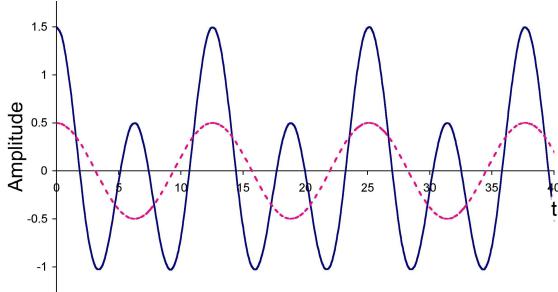


Figure 4.8: Simple two-component example signal.

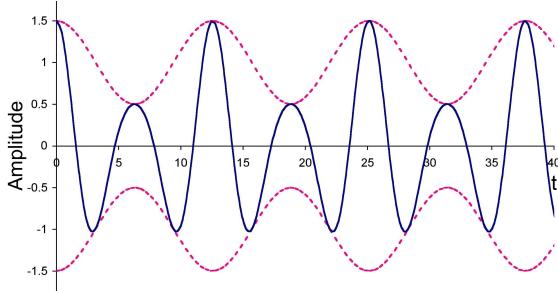


Figure 4.9: Example of an amplitude-modulated signal.

4.6.2. *Amplitude modulated example*

The second example is a stationary amplitude-modulated signal defined by

$$s(t) = [1 + 0.5 \cos(t/2)] \cos(t).$$

The graph of this function is shown in Fig. 4.9, along with this function's positive and negative envelope (dotted lines). Note that a very similar envelope could also be constructed for the previous example.

The differences between this example and the first one are that the tall positive peaks are a little narrower, and the shorter positive peaks are a little broader. The positive peaks have exactly the same values and timing. The negative peaks extend slightly lower (to -1.03), and their timing is shifted slightly toward the tall positive peaks. Another way to examine these signals is to expand this example's definition and apply a trigonometric identity for the product of two cosines:

$$\begin{aligned} s(t) &= \cos(t) + 0.5 \cos(t/2) \cos(t) \\ &= \cos(t) + 0.25 \cos(t/2) + 0.25 \cos(3t/2). \end{aligned}$$

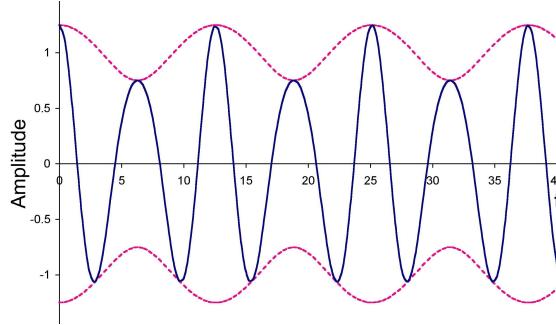


Figure 4.10: High-frequency component separated from the AM signal by filtering.

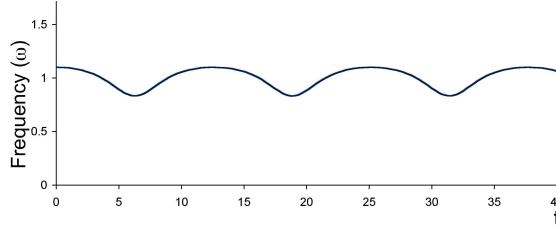


Figure 4.11: Instantaneous frequency of the AM signal component separated by filtering.

The above equations show that the difference between this example and the previous one is a smaller coefficient for the $\cos(t/2)$ term and an additional higher-frequency term, $0.25 \cos(3t/2)$.

The maximum timing between peaks is again slightly greater than π , indicating the need for filters with upper breakpoint frequencies at $\omega = 0.93$. The result of filtering this signal separates the lower-frequency $\cos(t/2)$ term from the two higher frequency components; i.e.,

$$c_{filter}(t) = \cos(t) + 0.25 \cos(3t/2)$$

and

$$r_{filter}(t) = 0.25 \cos(t/2).$$

The high-frequency component produced by filtering, $c_{filter}(t)$, is shown in Fig. 4.10, along with its amplitude envelope. The instantaneous frequency of the filtering solution ranges from approximately 0.83 to 1.10, as shown in Fig. 4.11.

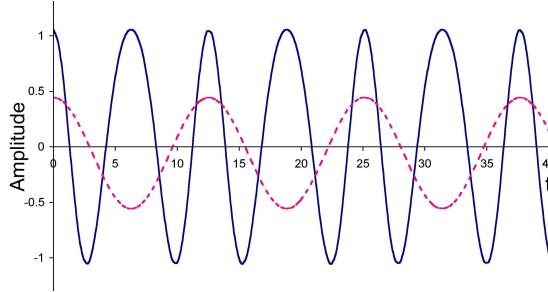


Figure 4.12: High-frequency component separated from the AM signal by the HHT sifting process.

The result produced by HHT sifting is quite different, as shown in Fig. 4.12. The HHT sifting algorithm produces a constant-amplitude, frequency-modulated component, and a trend that is the same frequency as the filter trend but twice its amplitude. The component's constant amplitude makes its frequency modulation more clearly evident.

The HHT-separated component and trend signals can be described mathematically by

$$c_{HHT}(t) = \cos(t) + 0.25 \cos(3t/2) - 0.25 \cos(t/2) + 0.0563$$

and

$$r_{HHT}(t) = 0.5 \cos(t/2) - 0.0563.$$

The small constant terms in the HHT formulas offset the frequency modulation effects that result when the three cosine terms in $c_{HHT}(t)$ are combined. These effects are discussed in the next example.

The instantaneous frequency of this signal, shown in Fig. 4.13, has a larger range than that for the filter solution. The instantaneous frequency of the HHT sifting solution ranges from approximately 0.69 to 1.19.

Both solutions produce monocomponent high-pass components and band-limited trend signals, satisfying the HHT objectives as we characterized them earlier. The filter produces a mixed AM and FM component with a smaller-amplitude trend signal. HHT sifting produces a purely FM component with larger frequency variations and a larger-amplitude trend signal.

In this example, the HHT result illustrates a classic example of signal aliasing. The HHT and warped filtering processes, being based on peak

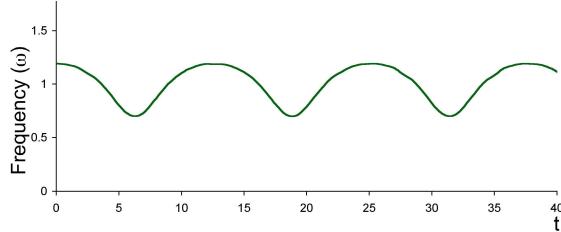


Figure 4.13: Instantaneous frequency of the AM signal component separated by HHT sifting.

values, which are sampled at a frequency of $\omega = 2$, under-sample the input signal and misinterpret the energy from the higher-frequency $\cos(3t/2)$ component and attribute it incorrectly to the lower-frequency $\cos(t/2)$ term. The extra $\cos(t/2)$ energy in both the HHT component and the trend for this signal does not accurately redistribute the $\cos(t/2)$ energy contained in the input signal. Aliasing often leads to unintended consequences, which we believe is the case here.

4.6.3. Frequency modulated example

The third example is a stationary frequency-modulated signal defined by

$$s(t) = \cos[t + 0.5 \sin(t)].$$

The amplitude of this signal is constant, but its phase increases nonlinearly. The graph of this function, shown in Fig. 4.14, shows sharpened positive peaks and rounded negative peaks, much like the solutions to Stokes's equation (although these results are not a solution to Stokes's equation). See Huang (1999) for further discussion and analysis of nonlinear wave dynamics.

HHT analysis of this signal finds evenly spaced constant-valued positive and negative peaks. The trend function is a constant zero, and the separated component captures the entire signal. The instantaneous frequency derived from the HHT results, as shown in Fig. 4.15, matches our expectations: $\varphi'(t) = 1 + 0.5 \cos(t)$.

The filtering results are a bit more complicated to explain. The coefficients of the Fourier series for a frequency-modulated signal are defined in terms of Bessel functions. [See, for example, Lathi (1965) or Schwartz (1990) for details.] If the signal is generalized to

$$s(t) = A \cos[\omega_c t + \beta \sin(\omega_m t)],$$

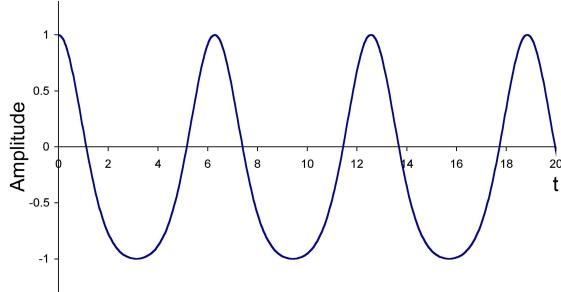


Figure 4.14: Example of a frequency-modulated signal.

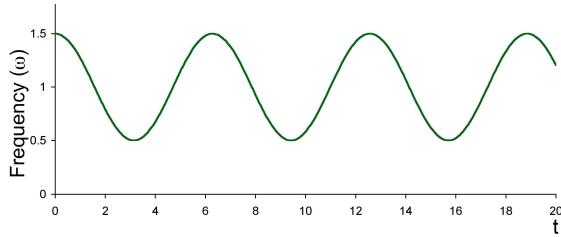


Figure 4.15: Instantaneous frequency of the FM signal component separated by HHT sifting.

where A denotes the signal's constant amplitude, ω_c is its "carrier" frequency, β is the index of modulation, and ω_m is the modulating frequency, then the equivalent Fourier series is

$$s(t) = A \sum_n J_n(\beta) \cos[(\omega_c + n\omega_m)t],$$

where $J_n(\cdot)$ is the Bessel function (first kind) of order n . The summation, theoretically, ranges over integral values of n from $-\infty$ to ∞ . Bessel function values for small values of β , however, are essentially zero for all but a few terms. An approximate Fourier series for this signal is

$$s(t) \approx -0.242 + 0.969 \cos(t) + 0.242 \cos(2t) + 0.031 \cos(3t).$$

This representation of the signal shows that its nonlinear phase gives it a constant "DC" term as well as higher-frequency harmonic components. The filter breakpoint frequencies for this signal, determined from the signal's peak-to-peak timing, were $\omega = 1$ and $\omega = \frac{1}{2}$. This filter produced the

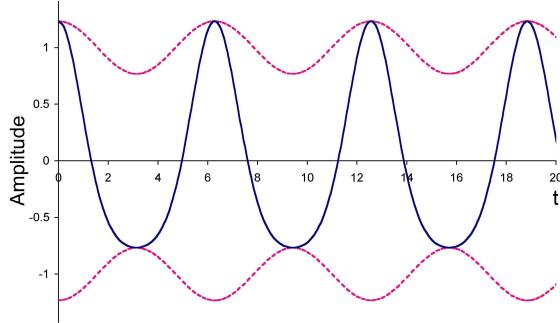


Figure 4.16: High-frequency component separated from the FM signal by filtering.

separation

$$c_{filter}(t) = 0.969 \cos(t) + 0.242 \cos(2t) + 0.031 \cos(3t),$$

and

$$r_{filter}(t) = -0.242.$$

The results for the high-pass component are shown in Fig. 4.16, along with a smooth amplitude envelope connecting the absolute values of the peaks (dashed lines).

The output from this filter differs from the monocomponent signal we started out with, although the basic shape of the input signal is preserved. The oscillating amplitude appears problematic, since the input signal contained no amplitude modulation. Furthermore, the amplitude oscillations have the same average frequency as the signal, which violates Bedrosian's spectral separation conditions. These amplitude oscillations appeared because the filtering process removes the constant term in the signal's Fourier series. Our earlier time-warp analysis showed that the amplitude envelope should be band-limited to below one-half of the signal's "carrier" frequency. The observed higher-frequency content, therefore, is an unexpected artifact that must be attributed to the filtering process.

Similar nonlinear signal behavior was encountered in the previous (AM) example. The high-frequency component separated by HHT sifting (shown in Fig. 4.12) contains alternating narrow and wide positive peaks. This nonlinear phase behavior gives this signal a constant term similar to that described here. As these examples show, any signals with nonlinear phase behavior may contain low-frequency energy that will produce similar artifacts in filter results.

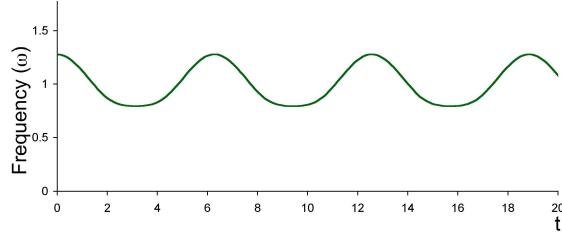


Figure 4.17: Instantaneous frequency of the FM signal component separated by filtering.

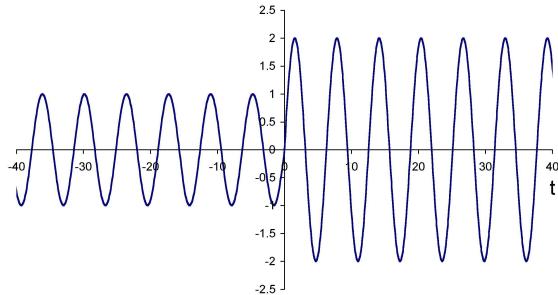


Figure 4.18: Amplitude step example signal.

The instantaneous frequency derived from the high-pass filter output is shown in Fig. 4.17. This signal has a smaller frequency range than the HHT component ($\omega = 0.79$ to 1.28), and the variations are not purely sinusoidal.

4.6.4. Amplitude step example

The preceding examples are all stationary signals that could be handled by static filtering techniques (if the frequencies are known in advance). The signal shown in Fig. 4.18 begins to exercise the dynamic capabilities of the HHT and filtering processes. This signal contains a step discontinuity in its amplitude at time $t = 0$; i.e.,

$$s(t) = \begin{cases} \sin(t) & \text{for } t \leq 0 \\ 2 \sin(t) & \text{for } t \geq 0. \end{cases}$$

Both the HHT and filtering processes are expected to smooth out this amplitude transition because of limitations on amplitude bandwidth suggested by the monocomponentness considerations. The results plotted in

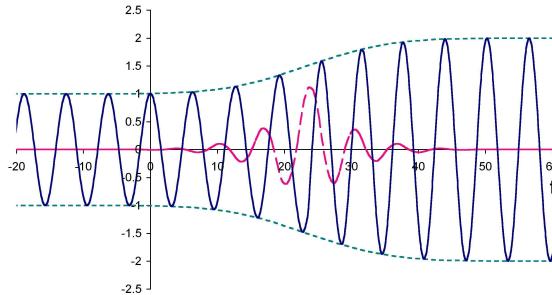


Figure 4.19: HHT component and trend results for the amplitude step signal.

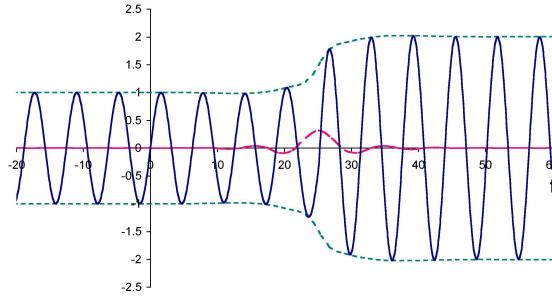


Figure 4.20: Filter high- and low-pass results for the amplitude step signal.

Figs. 4.19 and 4.20 verify this expectation. The differences in smoothing are a result of the differing filter transfer functions and, in the case of the HHT, its signal aliasing behavior. The trend signals in both cases are shaped somewhat like sampling functions. The HHT trend has a considerably higher amplitude than the filter low-pass signal.

There is also a time delay of approximately 24 time units for the incremental HHT result and 25 time units for the filtering results. These delays are necessary to collect data on the signal's future behavior, which both processes need before they can produce their results.

The instantaneous frequencies, derived numerically, for the two separated high-pass components are shown in Figs. 4.21 and 4.22. In both cases, the effect of smoothing out the amplitude step transient has created transient frequency modulations. This result suggests the presence of a “conservation of transient energy” law that allows amplitude transients to be transformed into frequency transients.

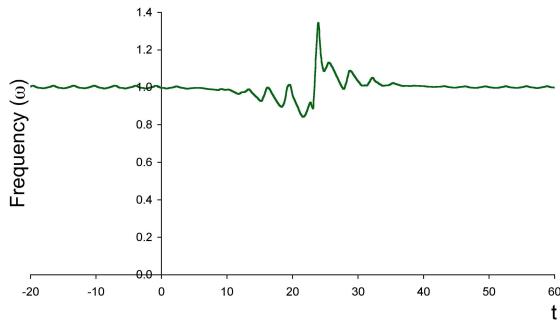


Figure 4.21: Instantaneous frequency of the amplitude step component separated by HHT sifting.

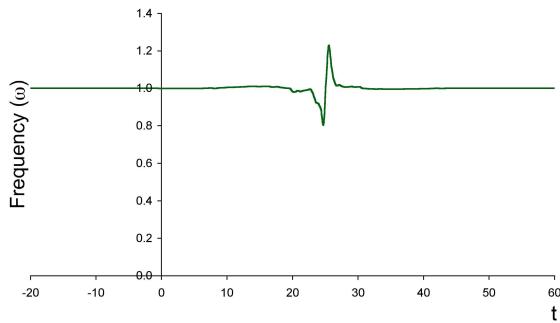


Figure 4.22: Instantaneous frequency of the amplitude step component separated by filtering.

Although our understanding of this frequency behavior is incomplete, we can explain the behavior of the two signal separation processes by using their representation in the frequency domain. The Fourier transform of the amplitude step signal is

$$S(\omega) = 3\pi i[\delta(\omega + 1) - \delta(\omega - 1)]/2 + 1/(1 - \omega^2),$$

where $\delta(\cdot)$ denotes the Dirac delta function. Figure 4.23 shows the magnitude of this transform. It has complex poles at $\omega = \pm 1$, which reflect the $\sin(t)$ term in the signal. The bandwidth contributed by the amplitude step is distributed smoothly over the entire frequency spectrum.

Figure 4.24 shows how filtering separates the amplitude step signal in the frequency domain. The low-pass (solid) curve shows the spectrum of the

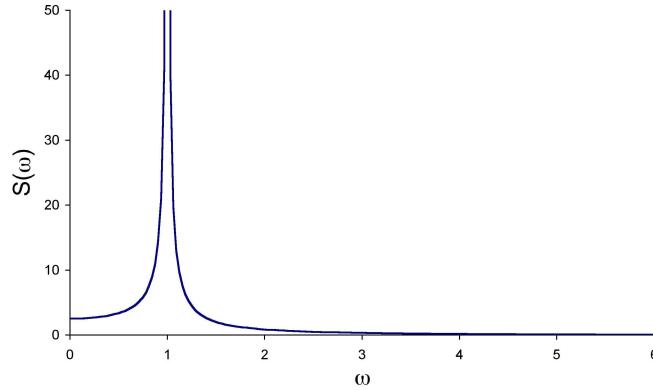


Figure 4.23: Fourier transform (magnitude) of the amplitude step signal.

signal trend and the high-pass results (dashed) curve shows the spectrum of the separated component. Inverting these transforms back into the time domain reproduces the trend and component signals shown in Fig. 4.20.[¶]

In practice, the results shown back in Fig. 4.20 are produced by direct convolution of the signal with the digital filter coefficients, not by applying transforms. The results, however, are the same by using either process.

An explanation of the HHT results requires introducing the effects of the peak curve-fitting and iteration process. Figure 4.25 shows the spectra of the signals separated by the HHT algorithm. The low-frequency “hump” (solid line) is the trend’s spectrum. The second curve (dashed line) is the spectrum of the separated high-frequency component. Transforming these spectra back into the time domain reproduces the signal trend and separated component shown in Fig. 4.19.

The third curve in Fig. 4.25 (dotted line) shows the apparent spectrum of the signal that was derived by resampling it at its peaks. This result is a direct effect of aliasing. Because the peak sampling rate is below the signal’s original sampling rate, aliasing creates overlapping replicas of the spectrum

[¶]Care must be taken with numerical Fast Fourier Transform (FFT) tools when analyzing these signals and spectra. The results presented here are for continuous infinite-integral transforms of one-time transient events. Numerical techniques that operate on finite-duration numerical representations of signals and their spectra can easily generate different results. For example, a finite representation of the signal shown in Fig. 4.18 will be presumed to repeat periodically. While the graph still looks like a one-time unit step amplitude change, the transform produced will be for a repeating square-wave modulated signal.

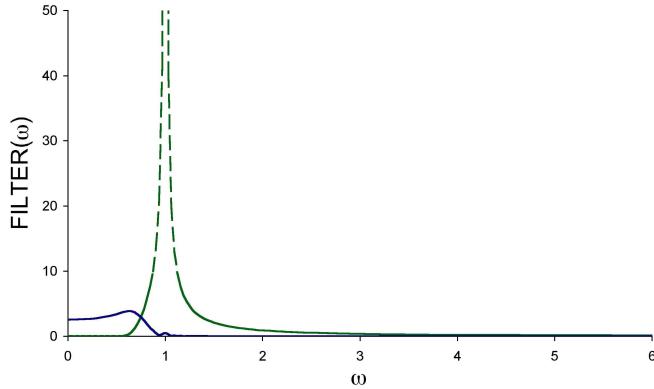


Figure 4.24: Filter high- and low-pass spectra for the amplitude step signal.

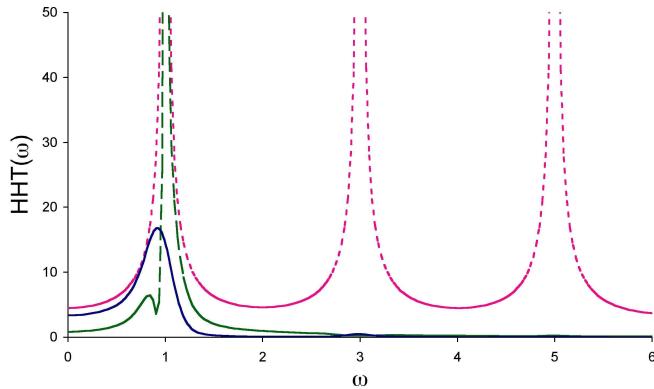


Figure 4.25: Spectra of the HHT trend and separated component for the amplitude step signal.

shown in Fig. 4.23. The results shown in Fig. 4.25 reveal that aliasing has a significant effect on the signal's apparent spectrum, causing the HHT algorithm to attribute considerable energy to the trend that is not part of the input signal. The separated high-frequency component is calculated by resampling the trend at its original sample times and subtracting that result from the original input signal. Only the trend, therefore, is directly affected by the aliased spectrum.

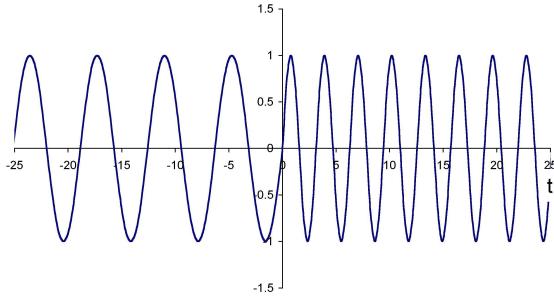


Figure 4.26: Frequency shift example signal.

4.6.5. Frequency shift example

The final example signal to be explored contains a step discontinuity in frequency at time $t = 0$, given by

$$s(t) = \begin{cases} \sin(t) & \text{for } t \leq 0 \\ \sin(2t) & \text{for } t \geq 0. \end{cases}$$

A graph of this signal is shown in Fig. 4.26. Because the signal amplitude is constant, the HHT trend remains constant (zero) through this frequency shift. The aliasing has no effect because the trend is zero. The first separated HHT component captures the entire input signal. It seems clear from this example and the earlier frequency-modulated example that the HHT will separate any constant-amplitude, monotonically-increasing phase signal as a single component.

The instantaneous frequency extracted from the signal, which is the HHT-separated component, is shown in Fig. 4.27. It tracks the signal nearly perfectly through the transition. While the HHT produced considerable smoothing of the amplitude step in the previous example, it makes no attempt to smooth out the frequency shift here.

Filtering produces quite different results, as shown in Fig. 4.28. The high-pass signal (solid line) shows a clear disturbance, although it is difficult to characterize. The low-pass signal (central dotted line) looks something like an inverted sampling function, centered at the point where the frequency shift takes place. The amplitude envelope around the high-frequency signal (upper and lower dotted lines) also reflects the disturbance.

The instantaneous frequency, derived numerically, for the high-pass component signal is shown in Fig. 4.29. We do not fully understand why the frequency behavior should take this shape.

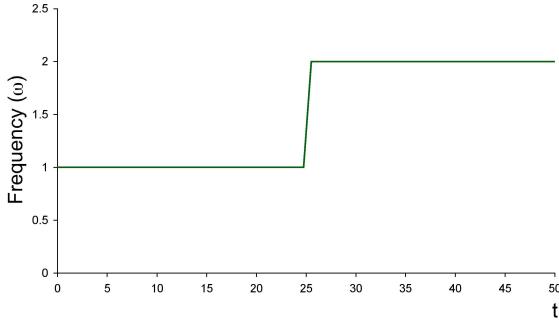


Figure 4.27: Instantaneous frequency of the frequency shift component separated by HHT sifting.

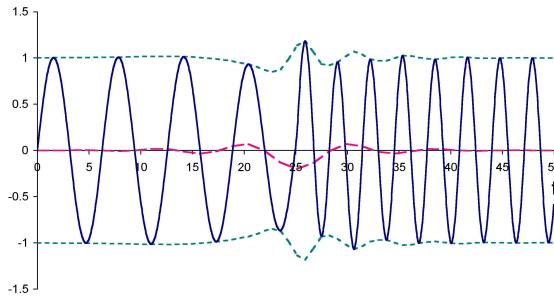


Figure 4.28: Filter high- and low-pass results for the frequency shift signal.

As with the previous example, we turn to the frequency domain to explain the behavior of the filter. The Fourier transform of the frequency step signal is

$$\begin{aligned} S(\omega) = & \pi i[\delta(\omega + 1) - \delta(\omega - 1)]/2 + \pi i[\delta(\omega + 2) - \delta(\omega - 2)]/2 \\ & -1/(1 - \omega^2) + 2/(4 - \omega^2). \end{aligned}$$

The magnitude of this transform is shown in Fig. 4.30. The complex poles at $\omega = \pm 1$ and $\omega = \pm 2$ reflect the signal's two sinusoidal frequencies. The bandwidth contributed by the frequency transition is distributed smoothly over the entire spectrum.

Figure 4.31 shows how filtering separates the frequency shift signal in the frequency domain. The low-pass curve (solid line) shows the spectrum of the signal trend. The high-pass curve (dashed line) shows the spectrum of the separated component. Inverting these transforms reconstructs the signals

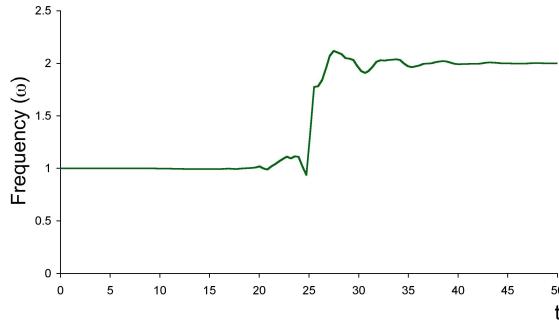


Figure 4.29: Instantaneous frequency of the frequency shift component separated by filtering.

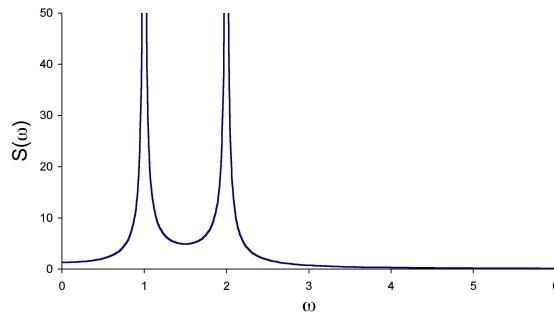


Figure 4.30: Fourier transform (magnitude) of the frequency shift signal.

shown in Fig. 4.28. The results shown in Fig. 4.28, however, were calculated by direct convolution of the signal with the digital filter coefficients, not by applying transforms.

Because the filter breakpoint frequencies are determined by the lowest peak-to-peak frequency within the span of the filter, these results are effectively the same as for static filters with breakpoints at $\omega = \frac{1}{2}$ and $\omega = 1$. Once the last low-frequency peak passes through the filter, its coefficients are adjusted to move the breakpoint frequencies to $\omega = 1$ and $\omega = 2$. This adjustment has no effect on the filter outputs because, in both cases, the signal resides completely within the high-pass pass band.

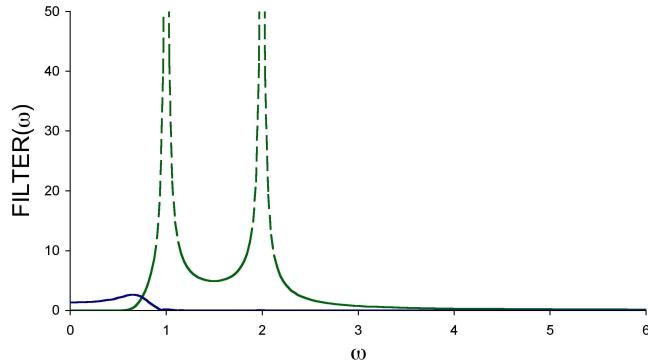


Figure 4.31: Filter high- and low-pass spectra for the amplitude shift signal.

4.7. Summary and conclusions

In this chapter, the HHT component separation, or “sifting,” process has been compared with a filtering process that was intended to mimic HHT behavior. The conjecture that conventional digital filters, by adapting dynamically to signal frequency content, could substitute for the HHT process was found to be incorrect. The results from several example signals showed that under most conditions, the two techniques produce distinct results.

The experiments we conducted to compare the HHT and filtering processes led to the discovery of aliasing in the HHT sifting algorithm. The process of sampling a signal at its peak times results in a classic example of under-sampling that leads to misinterpretation of signal frequency content. Specifically, signal content at frequencies above the peak-to-peak sampling rate is misinterpreted as lower-frequency content.

The question of whether aliasing is a problem or a “feature” in terms of HHT signal separations has not yet been completely resolved. High-frequency components separated from examples where aliasing arises appear to satisfy the requirements for “monocomponentness,” so they are expected to have well-defined instantaneous frequencies. Where aliasing arises, however, it introduces anomalous energy into both the high-frequency component and the trend, and this result can be considered a form of signal corruption. Further investigation is needed to determine if unaliased filtering results are indeed “better,” or if the aliasing is in some unusual way a necessary aspect of the HHT sifting process.

4.7.1. Summary of case study findings

For signals with a dominant highest frequency (case study #1), the HHT and filtering were found to produce equivalent separations.

For stationary amplitude-modulated signals with a dominant central “carrier” frequency (case study #2), filtering separates the lower sidebands as the trend, and the carrier and upper sidebands as the high-frequency component. The HHT, because of aliasing, misinterprets the upper sideband energy as lower-frequency energy, effectively doubling the lower sideband amplitude. This result gives the high-frequency component a nearly constant amplitude and larger variations in instantaneous frequency.

For signals with transient amplitude changes (case study #4), HHT sifting produced a broad smoothing of the amplitude transition and, because of aliasing, a large trend amplitude. Filtering also smoothed out the amplitude transition, but not as broadly as the HHT. Its trend amplitude was small compared to that of the HHT trend.

For frequency-modulated signals with monotonically increasing phase (case studies #3 and #5), the HHT high-frequency component captures the entire signal, leaving a zero-valued residual trend. The extracted phase function, $\varphi(t)$, and instantaneous frequency, $\varphi'(t)$, for these signals tracked the signal behavior very closely, even with significant transients in frequency (case study #5). Filtering had considerably more difficulty with FM signals. Signals with nonlinear phase functions often have significant low-frequency content. Conventional filtering separates the high- and low-frequency energy, disrupting the input signal’s monocomponent characteristics.

4.7.2. Research directions

Although this paper investigated a key step in separating signal components, additional aspects of the overall problem need attention. The following research areas have been identified as areas still to be explored.

Resolving the question about aliasing is of high priority. Our preference for a solution would be an algorithm that separates complex signals into components without aliasing, and without the amplitude disturbances filtering causes with FM signals.

Second on our list is finding a better way to separate amplitude and phase information from monocomponent signals. Although the Hilbert transform is the obvious theoretical solution, the finite numerical approximations we used produced anomalous results.

Episodes of signals with only very low-frequency content compared to

their sampling rate (that is, with many samples between peaks) would require excessively long filters to achieve the separations we propose. To process such signals, a method is needed for adaptively down-sampling or decimating the signal and for automatically restoring higher sampling rates when higher-frequency content returns. Static down-sampling is used extensively in wavelet transform processing [see, for example, Daubechies (1992)]. To our knowledge, the idea of a dynamic down-sampling mechanism has not been explored.

The residual trend signals that are passed to successive stages of sifting have their high-frequency content removed, resulting in signals with lower and lower frequency content. This result is a prime example of where signal down sampling is needed. Non-uniform sampling techniques may be useful here, although they appear to require more complex up-sampling procedures to restore their original sampling rates than do uniformly sampled signals [see Aldroubi (2001) for examples of possible techniques].

Real-world signals often contain components that turn on and off intermittently, like the telephone that rings while we are listening to our favorite music or are eating dinner. Huang (1999) developed a technique for dealing with such intermittent components that attempts to minimize the disturbance in analysis of more continuous “background” components. Although a clear need exists for this capability, it has not yet been addressed in our real-time algorithms.

References

- Aldroubi, A., and K. Gröchenig, 2001: Nonuniform sampling and reconstruction in shift-invariant spaces. *SIAM Review*, **43**, 585–620.
- Bedrosian, E., 1963: A product theorem for Hilbert transforms. *Proc. IEEE*, **51**, 868–869.
- Boashash, B., 1992: Estimating and interpreting the instantaneous frequency of a signal. *Proc. IEEE*, **80**, 520–568.
- Cohen, L., 1995: *Time-Frequency Analysis*. Prentice Hall, 299 pp.
- Daubechies, I., 1992: *Ten Lectures on Wavelets*. CBMS-NSF Series in Applied Mathematics, Vol. 61, SIAM, 357 pp.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of nonlinear water

- waves: The Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Huang, N. E., C. C. Chern, K. Huang, L. W. Salvino, S. R. Long, and K. K. Fan, 2001: A new spectral representation for earthquake data: Hilbert spectral analysis of station TCU129, Chi-Chi, Taiwan, 21 September 1999. *Bull. Seism. Soc. Am.*, **91**, 1310–1338.
- Kincade, D., and W. Cheney, 1991: *Numerical Analysis*. Brooks/Cole Publ., 690 pp.
- Lathi, B. P., 1965: *Signals, Systems and Communication*. John Wiley & Sons, 607 pp.
- Maragos, P., J. F. Kaiser, and T. F. Quatieri, 1992: On separating amplitude from frequency modulations using energy operators. Preprints, *Intl. Conf. on Acoust., Speech, and Signal Process.*, San Francisco, CA, IEEE, Vol. 2, 1–4.
- Maragos, P., J. F. Kaiser, and T. F. Quatieri, 1993: On amplitude and frequency demodulation using energy operators. *IEEE Trans. Signal Process.*, **41**, 1532–1550.
- Meeson, R., 2002: An Incremental, Real-Time Algorithm for the Hilbert Huang Transform. IDA Paper P-3656, Institute for Defense Analyses, 44 pp.
- Nuttall, A. H., 1966: On the quadrature approximation to the Hilbert transform of modulated signals. *Proc. IEEE*, **54**, 1458–1459.
- Oppenheim, A., and R. Schafer, 1989: *Discrete-Time Signal Processing*. Prentice Hall, 879 pp.
- Parks, T. W., and C. S. Burrus, 1987: *Digital Filter Design*. John Wiley & Sons, 342 pp.
- Schwartz, M., 1990: *Information Transmission, Modulation, and Noise*. McGraw-Hill, 742 pp.

Reginald N. Meeson, Jr.

*Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA
22311-1882, USA*

meeson@ida.org

CHAPTER 5

STATISTICAL SIGNIFICANCE TEST OF INTRINSIC MODE FUNCTIONS

Zhaohua Wu and Norden E. Huang

One of the preliminary tasks when analyzing a dataset is to determine whether it or its components contain useful information. The task is essentially a binary hypothesis testing problem in which a null hypothesis of pure noise is often pre-proposed. To test against the null hypothesis, the characteristics of noise need to be understood first, and often, these characteristics pertain to the analysis method used.

In this paper, the characteristics of Gaussian white noise are studied by using the empirical mode decomposition (EMD) method. Statistical testing methods for Gaussian white noise for the intrinsic mode functions (IMFs) are designed based on the characteristics of Gaussian white noise by using EMD. These methods are applied to well-studied geophysical datasets to demonstrate the method's validity and effectiveness.

5.1. Introduction

The word “noise” can possibly be traced back to the Latin word “nausea,” “seasickness, feeling of sickness.” In the scientific community, “noise” refers to a disturbance, especially a random and persistent disturbance that obscures or reduces the clarity of a signal. The causes of noise are numerous. In radar, noise is often caused by ambient radiation and the receiver’s electronics. In a digital communication system, the signal is usually distorted due to limited channel bandwidth and is corrupted by addictive channel noise. In nature, noise can be generated by local and intermittent instabilities, irresolvable sub-grid phenomena, some concurrent phenomena in the environment where investigations are being conducted, and by the sensors and recording systems. As a result, when we are dealing with data, we are inevitably dealing with an amalgamation of signal and noise,

$$x(t) = s(t) + n(t), \quad (5.1)$$

where $x(t)$ denotes the data, and $s(t)$ and $n(t)$ are the signal and noise, respectively.

The detection of the information content of a noisy dataset is fundamental to decision making and information extraction. Usually, the extraction of information requires a knowledge of the characteristics of both the signal and the noise. When the processes that generate the dataset are linear, and the noise in the data has distinct characteristics from those of the true signal, effective filters can be designed based on the characteristics of the signal and the noise to separate a dataset's signal from the noise. However, such cases are relatively rare since knowledge of the signal in a noisy dataset is limited prior to the analysis of the data. The problem can be more complicated if the signal is nonlinear and nonstationary, and the data is of limited size. In such cases, a short piece of pure noise data may behave like a signal, and therefore, all the possible behaviors of a short piece of noise should be considered. Under such circumstances, a less ambitious goal is often set to decide whether a noisy dataset has any signals. The latter is often a binary hypothesis testing problem associated with a null hypothesis that assumes the dataset contains only noise. Under such a hypothesis, the characteristics of noise can be used as a reference for discriminating the data (or their components) from pure noise without having any pre-knowledge of the signal. Furthermore, knowing the characteristics of the noise is an essential first step before one can attach any significance to the signal eventually extracted from the data. The characteristics of noise are usually related closely to the analysis methods used to examine the noise. For example, white noise in the temporal domain is characterized by the independence among any data points with a zero autocorrelation, whereas in the Fourier frequency domain by a flat Fourier spectra.

Many time-series-analysis methods are currently available for use. When the processes generating the data are linear, and the noises have distinct time or frequency scales different from those of the true signal, these analysis methods may have some capability of distinguishing the data from noise. However, most of these methods suffer more or less from various drawbacks even in linear and stationary cases. For example, even if the real signal and the noises have distinct fundamental frequencies, their harmonics can still mix with the noise during a Fourier spectrum analysis. This mixing of the harmonics with noise will make the Fourier spectrum analysis an ineffective noise-discriminating method. The problem could be even worse when the time series to be analyzed is both nonlinear and non-stationary. Therefore, more effective methods, as well as an understanding of the characteristics of noise pertaining to these methods, are needed so that the signal content of real data can be estimated.

In recent years, a new method, entitled empirical mode decomposition (EMD), has been developed (Huang et al. 1998; Huang et al. 1999; Huang et al. 2003) and has been applied to various fields of scientific research and industry. In this book, the method is introduced and many new applications are illustrated. EMD is an adaptive method to decompose any time series into a set of intrinsic mode function (IMF) components, which become the basis for representing the data. Because the basis is adaptive and locally determined, it usually offers a more physically meaningful representation of the underlying processes. Due to the adaptive nature of the basis, harmonics are not needed; therefore, EMD is ideally suited for analyzing data generated by nonlinear, non-stationary processes.

In this chapter, we will examine the characteristics of Gaussian white noise by using EMD and then design statistical test methods to distinguish the IMFs of real data from those of Gaussian white noise. The characteristics of uniform white noise revealed by using EMD have already been reported (Wu and Huang 2004). Some characteristics of white noise described in Wu and Huang and in this chapter can also be found in a study by Flandrin et al. (2004) and a chapter by Flandrin et al. in this book. We will show that if we know the characteristics of the noise, we can offer some measure of the information content of signals buried under data with an unknown noise level.

In this chapter, we will demonstrate that Gaussian white noise has almost identical characteristics to those of uniform white noise, which we have already reported (Wu and Huang 2004): (1) the EMD is an effective dyadic filter capable of separating white noise into IMFs having mean periods exactly twice the value of the previous one; (2) the IMFs are all normally distributed; and (3) the Fourier spectra of the IMF components are identical in shape and cover the same area on a semi-logarithmic period scale. These results are useful for determining the relationship between the product of the mean energy density of an IMF and its corresponding mean period and also the spread function of the energy density. The characteristics and the derived relationships are verified by using the Monte-Carlo method, which analyzes a large synthetically generated Gaussian white noise dataset. These quantities also provide the necessary information for us to design a statistical significance test method by using the bounds for the energy density spread function of the IMFs of Gaussian white noise. Some well-known climate time series are used to illustrate the effectiveness of the methodology of assigning statistical significance.

The chapter is arranged as follows: Section 5.2 will present the numerical

experiment and the empirical relationship between the energy density and the mean period. Section 5.3 will focus on the empirical result of normally distributed IMF components, and the energy spread function derived. Section 5.4 will discuss the statistical significance test method and illustrate its validity by applying the method to some well-known climate time series and to the series defined using climate system model outputs. A discussion and some conclusions will be presented in section 5.5.

5.2. Characteristics of Gaussian white noise in EMD

This section is on the statistical characteristics of the IMF components of the Gaussian white noise. These characteristics are derived by numerically studying a lengthy Gaussian white noise of 2^{20} data points generated by using a method described by Press et al. (1992). We are forced to use numerical methods since EMD is an algorithm, and the IMFs have no analytical expression. The empirical results presented below are not sensitive to the random number generators.

Table 5.1: The mean periods of IMFs. Each column corresponds to an IMF. The second row is the number of maxima; the third row is the mean period calculated from the number of maxima; and the fourth row is the Fourier spectrum weighted mean period [see (5.8)].

IMFs	1	2	3	4	5
number of maxima	365358	173012	86247	43152	21701
Mean Period (counting extrema)	2.870	6.061	12.16	24.30	50.65
Mean Period (Spec.-weighted)	3.467	5.405	9.841	18.61	35.45

IMFs	6	7	8	9
number of maxima	10843	5429	2717	1345
Mean Period (counting extrema)	96.71	193.1	385.9	779.6
Mean Period (Spec.-weighted)	67.81	133.7	259.4	492.4

5.2.1. Numerical experiment

The Gaussian white noise data generated are decomposed into IMFs by using the EMD method. An IMF is any function having symmetric envelopes defined by the local maxima and minima separately, and also having the same number of zero-crossing and extrema. Practically, an IMF is extracted through a sifting process which stops when a certain criterion is satisfied. In this study, a Cauchy-type stoppage criterion modified from that in Huang et al. (1998) is used; i.e.,

$$SC = \frac{\sum_{i=1}^N [h_{n,k+1}(i) - h_{n,k}(i)]^2}{\sum_{i=1}^N h_{n,k}^2(i)} < 2 \times 10^{-4}, \quad (5.2)$$

where N is the length of data being decomposed and $h_{n,k}$ is the k th sifting result for n th IMF. This modification essentially eliminates the unstable jump of the value of the traditional Cauchy-type stoppage criterion defined in Huang et al. (1998) in the sifting process and is consistent with the stoppage criterion of the repetitiveness of the number of extrema described in Huang et al. (2003). In the experiment, the number of the iteration of the sifting process for each IMF is between seven and ten.

5.2.2. Mean periods of IMFs

Based on the definition of an IMF, we can determine the mean period of an IMF by counting the number of local maxima of the function. The results of the mean periods are listed in Table 5.1. In this table, the second row is the total number of local maxima, the peaks of each IMF in 2^{20} data points. The third row is the extrema-counting mean period measured in terms of the number of data points (2^{20} divided by total number of maxima of an IMF). The fourth row is the spectrum weighed mean period of an IMF, as defined by (5.8) later. (The spectrum-weighted mean period of an IMF is smaller than the corresponding extrema-counting mean period defined. We will further discuss the issue in section 5.3.) The mean period of the n th IMF of Gaussian white noise based on extrema counting is slightly larger than that of the n th IMF of uniformly distributed white noise, which we have already reported on (Wu and Huang 2004). However, the mean period doubling property remains since EMD serves as a dyadic filter, consistent with the results obtained by Flandrin et al. (2004).

5.2.3. The Fourier spectra of IMFs

Another characteristic that we are interested in is the detailed distribution of the energy density of an IMF in terms of the Fourier spectrum as a function of its period (the inverse of frequency). The derivation here follows what is described in Wu and Huang (2004), which provides more details. Since the IMFs are nearly orthogonal to each other, we have the total energy for the data f_j for $j = 1, 2, \dots, N$, to a high degree of approximation, as

$$\sum_{j=1}^N f_j^2 = \sum_{k=1}^N |F_k|^2 = N \sum_n E_n. \quad (5.3)$$

In (5.3),

$$F_k = \sum_{j=1}^N f_j e^{-2\pi i j k / N} \quad (5.4)$$

is the Fourier transform of data f_j , and

$$E_n = \frac{1}{N} \sum_{j=1}^N [C_n(j)]^2 \quad (5.5)$$

is the energy density of the n th IMF, where $i = \sqrt{-1}$, $|F_k|$ is the norm of F_k , and $C_n(j)$ is the n th IMF. The expected Fourier spectrum of a white noise time series is a constant, indicating that the contribution to the total spectrum energy comes from each Fourier component uniformly and equally. (For a synthetically generated white noise time series of short length, however, its Fourier spectrum may be a constant superimposed on by many spikes. The spikes of the Fourier spectrum of an individual copy will be smoothed out when the Fourier spectra of many copies of white noise series of the same length are averaged, and the average of the Fourier spectra approaches a constant.) The Fourier spectra for the IMFs, however, will not yield a constant white spectrum, for the decomposition through EMD effectively subjects the data to a dyadic filter (Flandrin et al. 2004).

To illustrate the shape of the Fourier spectrum for each IMF, the Fourier spectra of 256 (2^8) independent segments of 4096 (2^{12}) data points each are calculated. The averaged Fourier spectra for nine IMFs are plotted on a semi-logarithmic scale in Fig. 5.1. It is clear that all the Fourier spectra, except the first one, have almost identical shapes in terms of the $\ln(T)$ -axis, where T is the period of a Fourier component. Figure 5.1 implies that the ratios of the periods of the neighboring spectra are almost identically

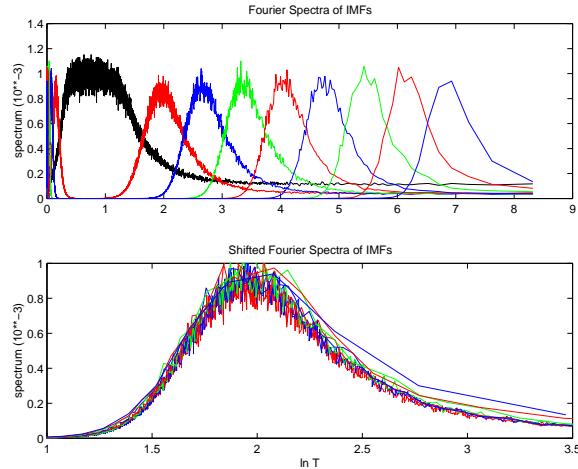


Figure 5.1: The Fourier spectra of IMFs as a function of the logarithm of their periods. In the upper panel, the black line is the Fourier spectrum of the first IMF, and the Fourier spectra of IMFs 2-9 are located away from that of the first IMF, respectively. In the lower panel, the Fourier spectra of IMFs 2-9 are redrawn and the Fourier spectrum of the n th IMF is shifted to the left by $(n - 2)\ln(2)$ for IMFs 3-9.

equal to 2, a fact consistent with the doubling of the mean periods of the neighboring IMFs.

Based on the fact of identical spectral shape and identical area coverage for each IMF, an approximate integral expression of the functional form of the Fourier spectrum for any IMF, except the first one, can be made as

$$\int S_{\ln(T),n} d[\ln(T)] = \text{constant}, \quad (5.6)$$

where $S_{\ln(T),n}$ denotes the Fourier spectrum of the n th IMF as a function of $\ln(T)$, and the subscript n denotes the n th IMF. From (5.6), one can derive for a normalized Gaussian white noise series that

$$\ln(\bar{E}_n) + \ln(\bar{T}_n) = 0, \quad (5.7)$$

where

$$\bar{T}_n = \frac{\int S_{\ln(T),n} d[\ln(T)]}{\int S_{\ln(T),n} d[\ln(T)]/T} \quad (5.8)$$

is the spectrum-weighted mean period of n th IMF as $N \rightarrow \infty$. The simple relation in (5.7) has already been stated in Wu et al. (2001).

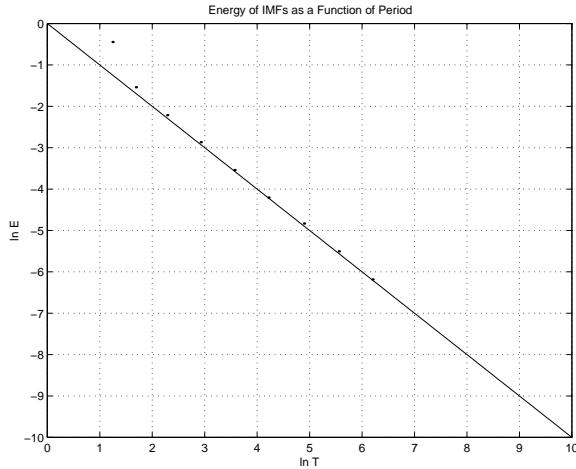


Figure 5.2: The relation between the energy density and the spectrum-weighted mean period. The black dots are the energy density as a function of the spectrum-weighted mean period for IMFs 1–9 based on the Fourier spectra displayed in Fig. 5.1. The black straight line from upper left to lower right is the theoretical line corresponding to (5.7).

The verification of (5.9) is given in Fig. 5.2, where the spectrum-weighted mean periods for IMFs 1–9 are calculated based on the averaged Fourier spectra of the corresponding IMFs displayed in Fig. 5.1. The straight black line from the upper left to the lower right is the expectation line derived from (5.9). Clearly, (5.9) offers an excellent fit to these scattered points.

5.2.4. Probability distributions of IMFs and their energy

In this subsection, we will examine the probability density functions of IMFs and their corresponding energy. Before we present the results, we re-examine the mathematical meanings of IMFs. To achieve this goal, we rewrite part of (5.3) as

$$f_j = \sum_{i=1}^n C_i(j) = \sum_{i=1}^m C_i(j) + R_m(j) = \sum_{i=1}^{m+1} C_i(j) + R_{m+1}(j), \quad (5.9)$$

where $R_m(j)$ is the remainder of f_j after m number of IMFs are extracted. It is easy to recognize that

$$R_i(j) = C_{i+1}(j) + R_{i+1}(j). \quad (5.10)$$

Since $C_{i+1}(j)$ is purely oscillatory with respect to $R_{i+1}(j)$, we can consider $R_{i+1}(j)$ as the local mean of $R_i(j)$. Similarly, $R_i(j)$ is the local mean of $R_{i-1}(j)$ and so forth, so that $R_2(j)$ is the local mean of $R_1(j)$. Therefore, $R_{i+1}(j)$ can also be considered as the local mean of f_j over a local timescale that is approximately the local period of $C_{i+1}(j)$. Since the local period of $C_{i+1}(j)$ is concentrated near the extrema-counting mean period of $C_{i+1}(j)$, to a good approximation, the distribution of $R_{i+1}(j)$ is close to the distribution of the running mean of f_j over the extrema-counting mean period of $R_i(j)$.

Two major theorems in statistics (Papoulis, 1986) can then be applied to infer the distribution of $C_{i+1}(j)$. The first one is the central limit theorem, which states that the mean over a given finite number of random samples from a distribution with finite variance results in a normal distribution. Therefore, $R_i(j)$ and $R_{i+1}(j)$ are both approximately normal distributions. The second theorem states that the linear combination of two normal distributions results in another normal distribution. Therefore, we can infer that all IMFs of Gaussian white noise are approximately normally distributed.

Figure 5.3 plots the probability distribution of each IMF for a sample of 50 000 data points. The results are consistent with the discussion we presented above. Indeed, the deviation from the normal distribution function grows as the mode number increases because in the higher modes, the IMFs contain a small number of oscillations; therefore, the number of events decrease, and the distribution becomes less smooth. When a sample of longer length is used, the IMFs of the higher modes will have more oscillations, and the distribution will converge to a normal distribution according to the central limit theorem.

According to the theory of probability density function, for a time series that has a normal distribution, its energy defined by (5.5) should have a *chi-square* distribution with the degrees of freedom of the chi-square distribution equal to the mean of the energy.

To determine the exact number of degrees of freedom for the chi-square distribution of IMFs decomposed from a white noise series of length N , we can argue as follows: we use the Fourier spectrum of a white noise series of the same length, N . For such a white noise series, its number of degrees of freedom is N , and that number is an invariant when the data are mapped into another space. The decomposition of such a white noise series in terms of Fourier components results in N Fourier components which form a complete set. Each component has a unit degree of freedom;

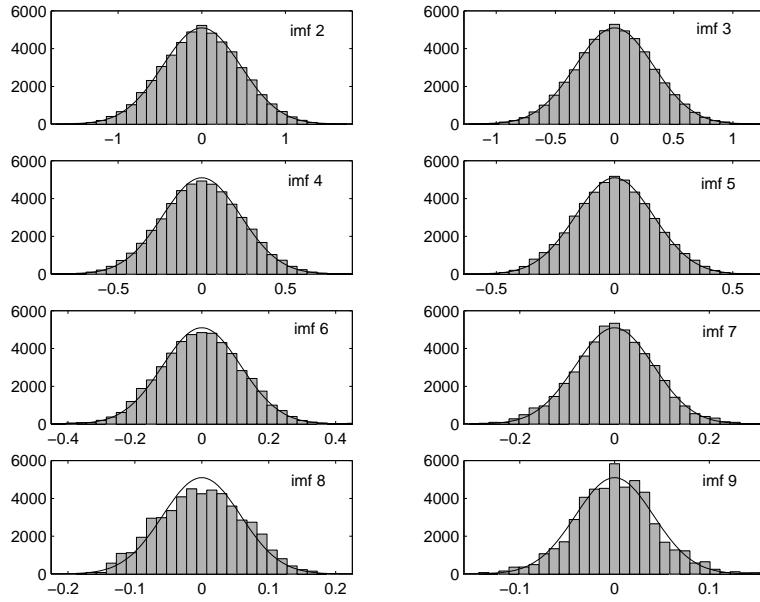


Figure 5.3: The histograms of IMFs (modes) 2–9 for a white noise sample with 50 000 data points. The superimposed black lines are the Gaussian fitting for each IMF.

therefore, the number of degrees of freedom of an IMF is essentially the sum of the Fourier components it contains. As the energy in a white noise series is evenly distributed to each Fourier components, we propose that the fraction of energy contained in an IMF is the same as the fraction of the number of degrees of freedom. For a normalized white noise time series with unit total energy, the number of degrees of freedom of the n th IMF should be the energy of that particular IMF; thus, $r_n = N\bar{E}_n$. Therefore, the probability distribution of $N\bar{E}_n$ is the chi-square distribution with $N\bar{E}_n$ degrees of freedom:

$$\rho_N(N\bar{E}_n) = (N\bar{E}_n)^{N\bar{E}_n/2-1} e^{-N\bar{E}_n/2}. \quad (5.11)$$

Therefore, the probability distribution of E_n is

$$\rho(E_n) = N(N\bar{E}_n)^{N\bar{E}_n/2-1} e^{-N\bar{E}_n/2}. \quad (5.12)$$

A Monte Carlo test confirms our conjecture. Figure 5.4 shows the histogram of the distribution of energy for each IMF for 1024 samples of white noise series each of the length of 1024 data points. The red lines are the corre-

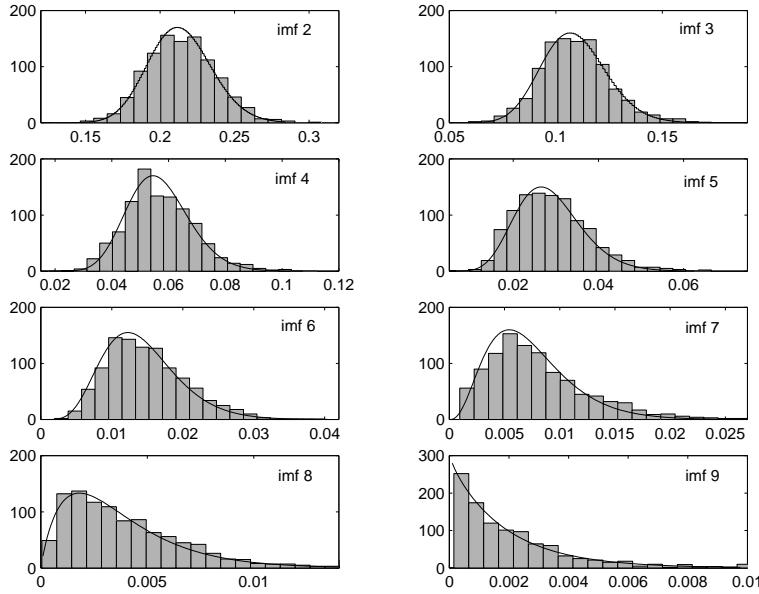


Figure 5.4: The histograms of the energy density for IMFs (modes) 2–9 for 1024 white noises samples of 1024 data points each. The superimposed black lines are the chi-square fitting for each IMF.

sponding chi-square distributions based on (5.12). Clearly, the theoretical lines and histograms are in excellent agreement with each other.

5.3. Spread functions of mean energy density

Having determined the distribution function of the energy, we are ready to derive the spread of the energy densities of the IMFs of white noise samples of certain length N . Since the characteristics of Gaussian white noise are essentially identical to those of the uniformly distributed white noise that we described in Wu and Huang (2003, 2004), the spreads of the energy density of the IMFs of Gaussian white noise are also the same as the results we obtained in Wu and Huang (2003, 2004). Here, we repeat the derivation of functions that we included in Wu and Huang (2003, 2004).

Since the probability distribution of energy density E_n is given by (5.12),

we can derive the probability distribution of a new variable $y = \ln(E)$ as

$$\begin{aligned}\rho(y) &= N(Ne^y)^{N\bar{E}/2-1} e^{-NE/2} e^y \\ &= C \exp\left(y \frac{N\bar{E}}{2} - \frac{NE}{2}\right) = C \exp\left[-\frac{N\bar{E}}{2} \left(\frac{E}{\bar{E}} - y\right)\right],\end{aligned}\quad (5.13)$$

where $C = N^{N\bar{E}/2}$. For simplicity, the subscript n is omitted in (5.13). Since

$$\frac{E}{\bar{E}} = e^{y-\bar{y}} = 1 + y - \bar{y} + \frac{(y - \bar{y})^2}{2!} + \frac{(y - \bar{y})^3}{3!} + \dots,\quad (5.14)$$

substituting (5.14) into (5.13) leads to

$$\begin{aligned}\rho(y) &= C \exp\left\{-\frac{N\bar{E}}{2} \left[1 - \bar{y} + \frac{(y - \bar{y})^2}{2!} + \frac{(y - \bar{y})^3}{3!} + \dots\right]\right\} \\ &= C' \exp\left\{-\frac{N\bar{E}}{2} \left[\frac{(y - \bar{y})^2}{2!} + \frac{(y - \bar{y})^3}{3!} + \dots\right]\right\},\end{aligned}\quad (5.15)$$

where $C' = C \exp[-N\bar{E}(1 - \bar{y})/2]$. From (5.15), one can determine the spread of the different confidence levels.

When $|y - \bar{y}| \ll 1$,

$$\rho(y) = C \exp\left\{-\frac{N\bar{E}}{2} \left[1 - \bar{y} + \frac{(y - \bar{y})^2}{2!}\right]\right\} = C' \exp\left[-\frac{N\bar{E}(y - \bar{y})^2}{4}\right],\quad (5.16)$$

Therefore, the distribution of $y = \ln(E)$ is approximately a Gaussian with a standard deviation of

$$\sigma^2 = \frac{2}{N\bar{E}} = \frac{2\bar{T}}{N}.\quad (5.17)$$

For such a case, the spread lines can be approximately defined as

$$y = -x \pm k \sqrt{\frac{2}{N}} e^{x/2},\quad (5.18)$$

where $x = \ln(\bar{T}_n)$ and k is a constant determined by percentiles of a standard normal distribution. For example, when $k = -2.326, -0.675, 0.0, 0.675$, and 2.326 , we obtain the first, 25th, 50th, 75th and 99th percentiles, respectively.

Figure 5.5 plots the spread lines for the 5th and 95th percentiles based on (5.15)–(5.16), as well as the scattered pairs of the mean energy density and spectrum-weighted mean period for the IMFs of each sample of 1024 data points. In this figure, the Fourier spectra of the IMFs of each sample are obtained through the Fourier transform of the IMFs of 1024 data points,

which is a quarter of the length of the samples used to calculate the Fourier spectra in Fig. 5.2.

In general, the Monte Carlo results displayed in Fig. 5.5 agree well with the theoretical inferences discussed above. The simplified calculations based on (5.18) agree well with the theoretical lines (bold dashed) that are based on (5.16), which provide more details of the abnormal distribution that skews toward the lower energy side. However, for slowly oscillating IMFs, the averages of the spectrum-weighted mean period deviate from the theoretical line (the bold black) significantly. For example, the scattered pairs of the mean energy density and the spectrum weighted mean period (red dots) for IMF 9, as well as their averages, are located systematically on the lower-left side of the theoretical expectation line of the mean energy density and the spectrum weighted mean period. As is evident in Fig. 5.2, such a large systemic error is not seen when the spectra are calculated by using longer samples, because the Fourier transform of a short piece of a slowly oscillating IMF with a large mean period is likely to represent artificially the low frequency components of the IMF with high-frequency components. Such drawbacks of the Fourier transform do not affect the high-frequency IMFs much since these IMFs do not contain much low-frequency information, but can lead to the severe overestimation (underestimation) of the mean frequency (spectrum-weighted mean period) of slowly oscillating IMFs.

The problem mentioned above has negative implications for designing a statistical significance test method for IMFs based on the characteristics of Gaussian white noise that we discussed previously: the spread lines of a certain percentage work only for IMFs of high frequency but not for IMFs of low frequency. To eliminate the problem, we need to search for a better estimation of the mean periods of IMFs. Fortunately, the extrema-counting and Hilbert transform lead to quite an accurate estimation of the mean period. The methods were adopted in Wu and Huang (2003, 2004). However, the latter approach also has a side effect: the expectation line of the mean period and the energy density is likely to deviate from the theoretical line.

Figure 5.6 shows the results of the latter approach. The dashed black line is the fitting of the averaged mean energy density and the averaged mean period based on extrema counting and the Hilbert transform of all 1024 samples. This fitting can be expressed as

$$\ln(\bar{E}) = 0.12 - 0.934 \ln(\bar{T}). \quad (5.19)$$

With this empirical correction and the spread function of energy density

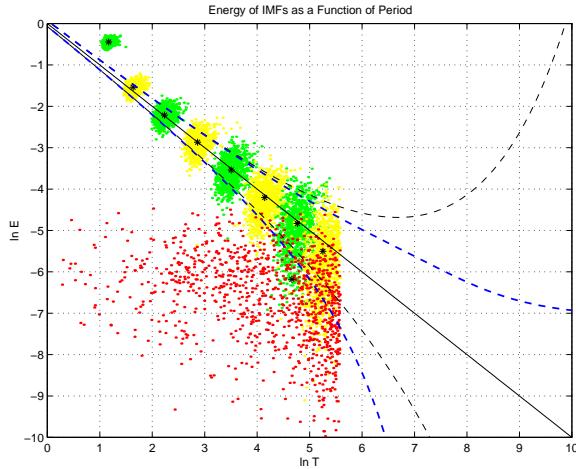


Figure 5.5: The spread function. The grouped green and yellow dots from the upper left to the lower right are the mean energy density as a function of the spectrum-weighted mean period for IMFs 1–8 for all 1024 samples with an identical length of 1024 data points. The grouped red dots are the same as the green and yellow dots except they are for IMF 9. The solid black line from the upper-left corner to the lower-right corner is the same as in Fig. 5.2. The bold dashed lines are the first and 95th percentiles calculated from (5.15). The thin dashed lines are the first and 95th percentiles calculated from (5.16). The asterisks correspond to the pairs of the averaged mean energy density and the averaged spectrum-weighted mean period of all 1024 samples for each IMF.

expressed by (5.15), we can obtain the lines for the various confidence levels. for the 5th and 95th percentiles are shown by the bold dashed lines. Clearly, these lines are consistent with the Monte Carlo test.

5.4. Examples of a statistical significance test of noisy data

The relationship between the energy density and the average period, the energy distribution, and its spread function for Gaussian white noise (which is confirmed by Monte Carlo tests) revealed in previous sections provides a unique opportunity for us to design a method to distinguish IMFs of a dataset from those of pure Gaussian white noise, so that we can determine which IMFs from a dataset contain information, and which IMFs may be only components of the pure Gaussian white noise that is contained in this dataset. The method is based on the rejection of a null hypothesis, which states that all IMFs are the components of a pure Gaussian white noise

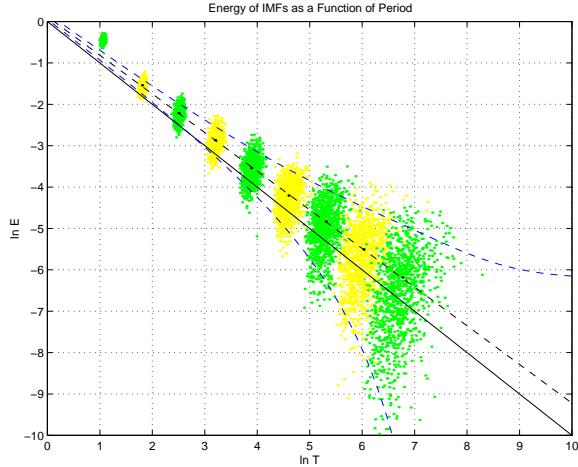


Figure 5.6: Same as Fig. 5.5 except that the abscissa is now the spectrum-weighted mean period for IMFs 1-9. There are no red dots or asterisks. The black dots correspond to the pairs of averaged mean energy density and averaged extrema-counting mean period.

dataset. Our goal is to demonstrate that the null hypothesis is an untrue statement about the origin of these IMFs. If the possibility is small (say, less than 0.05) that the IMFs decomposed from the noisy dataset could be the results of the decomposition of any Gaussian white noise of the same length as that of the targeted dataset, we reject the null hypothesis and consider that these IMFs contain signal information. With this consideration in mind, we design a significance test method utilizing the statistical characteristics of Gaussian white noise, which, by definition, contains no information. The method is then simple: first, decompose the targeted noisy dataset (normalized) into IMFs; second, use (5.15) to calculate the spread function of given percentiles for the IMFs decomposed from Gaussian white noise; third, select the confidence-limit level (e.g., 95%) and determine the upper and lower spread lines; and finally, compare the energy density for the IMFs from the noisy data with the spread functions. The IMFs with their energy located above the upper bound or below the lower bound should be considered as containing signal information at that selected confidence level. In the final step, if the targeted dataset is non-stationary and has a large trend, the trend should be excluded, and all the oscillatory IMFs should be rescaled by using the total energy of all the oscillatory IMFs.

In the following, we will use EMD to decompose three well-known time

series in climate study: (1) the North Atlantic Oscillation Index (NAOI, monthly data from January 1860 to December 1999); (2) the Southern Oscillation Index (SOI, monthly data from January 1866 to December 1997); and (3) the globally averaged surface temperature anomaly (GASTA, monthly data from January 1860 to December 1999). The decompositions of these indices use the same sifting stoppage criterion as that used in (5.2). The indices being analyzed here are all maintained by P. Jones of the Climate Research Unit, University of East Anglia. The advantages and drawbacks of our proposed significance testing method will be illustrated when the method is applied to test the IMFs of these three time series.

5.4.1. Testing of the IMFs of the NAOI

The North Atlantic Oscillation (NAO) is a well-known climate pattern that has great impact on Europe's climate. The NAO is often indexed by the difference in sea-level pressure between Iceland, representing the strength of the Icelandic climatological low, and the Azores of Lisbon (NAOI), near the central ridge of the Azores high. When the index is high, the Icelandic low is strong. This result increases the influence of the cold Arctic air masses on the northern seaboard of North America and enhances the ability of the eastward winds to carry warmer, moister air masses into western Europe during winter (Hurrell 1995). Thus, NAO anomalies are related to the downstream wintertime temperature and precipitation across Europe, Russia, and Siberia.

The NAOI being analyzed in this paper is described in Hurrell (1995) and Jones et al. (1997). As is well known among climatologists, the NAOI is a Gaussian-white-noise-like index. The auto-correlation of the NAOI with one data point lag is 0.088. Therefore, we expect that our method will generally show that most of the IMFs of the NAOI should be located between the 5% and 95% confidence levels.

The result of the testing is presented in Fig. 5.8. IMFs 2–7 and 9 are not distinguishable from the IMFs of the Gaussian white noise at the 95% confidence level. IMF 8 seems to be statistically significant at the 95% level. However, it cannot survive when the significance level is raised to 99%. The apparent significance of IMF 8 is not very surprising, since the NAOI does have a small auto-correlation. Overall, the application of our proposed testing method on a Gaussian-white-noise-like dataset is successful.

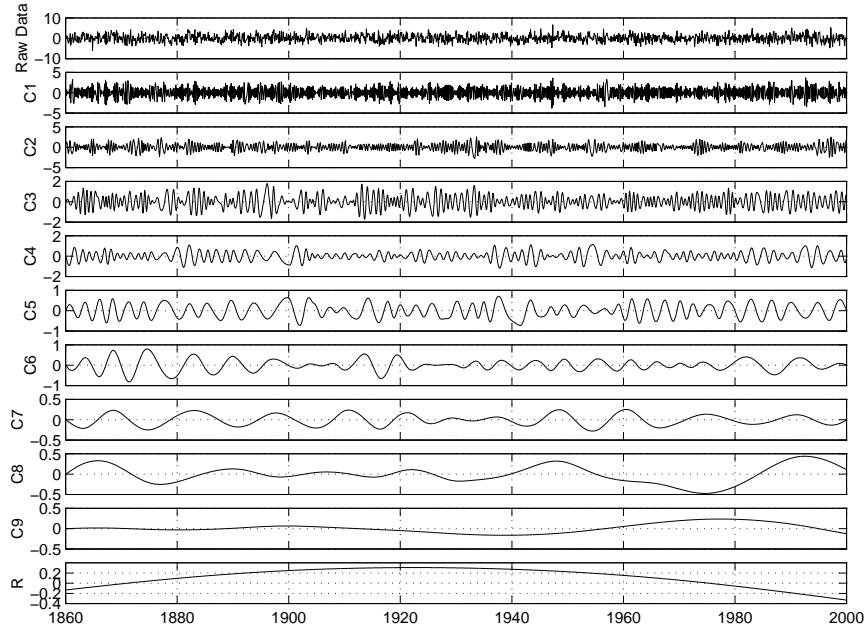


Figure 5.7: The raw NAOI (top panel), the corresponding intrinsic mode functions ($C_1 - C_9$), and the trend (R).

5.4.2. Testing of the IMFs of the SOI

The Southern Oscillation Index (SOI) is a normalized monthly sea level pressure index that reflects primarily the large-scale dynamically coupled system of atmosphere and ocean in the tropical Pacific (Trenberth 1984). A large negative (positive) peak of SOI, which often happens at two-to-seven-year intervals, corresponds to a strong El Niño (La Niña) event. With its rich statistical properties and scientific importance, the SOI is one of the most prominent time series in the geophysical research community and has been well studied. Many time-series-analysis tools have been used to analyze this time series to display their ability to reveal useful scientific information (e.g., Wu et al. 2001; Ghil et al. 2002).

The SOI that we use in this study is described in Ropelewski and Jones (1987) and Allan et al. (1991). The SOI has a lag-one auto-correlation of 0.69, indicating that it is significantly different from Gaussian white noise. The SOI and its IMFs are displayed in Fig. 5.9. The statistical significance test of the IMFs is displayed in Fig. 5.10.

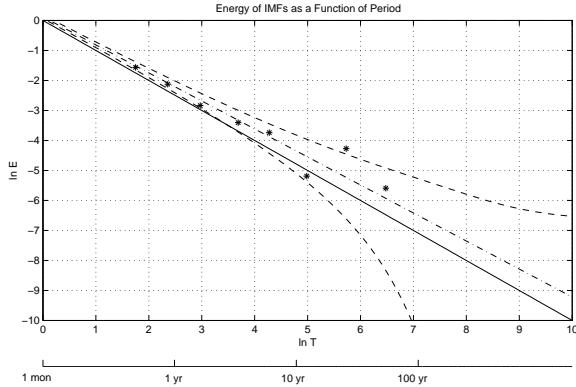


Figure 5.8: Significance test of the IMFs of the NAOI. The solid line from the upper-left corner to the lower-right corner is the same as in Fig. 5.6. The dash-dotted black line is the empirical fitting of the averaged mean energy density and the averaged mean period for Gaussian white noise. The dashed lines are the 5th and 95th percentiles calculated from (5.16). The asterisks correspond to the pairs of the averaged mean energy density and the averaged mean period of $C2 - C9$ of NAOI (from left to right).

The test shows that 5 IMFs are significant at the 95% confidence level. The averaged periods for these IMFs are 2.3 yr, 4.5 yr, 8.5 yr, 16.5 yr, and 32.9 yr, respectively. However, the latter two IMFs are not statistically significant at the 99% confidence level. This test result is consistent with the statistical test using the regular spectral analysis method, which also shows that the SOI has inter-annual peaks that are statistically significant.

5.4.3. Testing of the IMFs of the GASTA

The globally averaged surface temperature anomaly (GASTA) is one of the most popular time series in the climate, environmental, and even the social sciences. It includes the responses to the strengthening anthropogenic forces such as accumulating green house gases in the Earth's atmosphere and deforestation in many regions, in addition to the natural variability of the Earth's climate. Therefore, the GASTA is composed by two parts: a strong trend and oscillatory variations on all timescales. To examine whether an IMF comes from pure noise, we must first extract the trend part. Fortunately, the EMD method naturally separates the trend from the oscillatory variability in data, so we determine whether an IMF is significant at a cer-

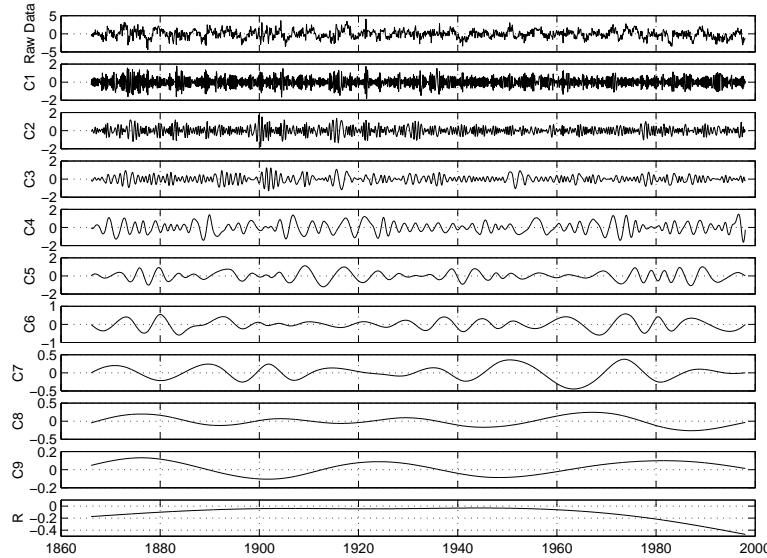


Figure 5.9: Same as Fig. 5.7, but for SOI.

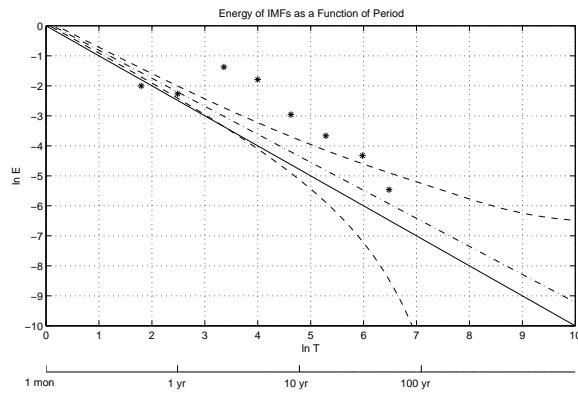


Figure 5.10: Same as Fig. 5.8, but for SOI.

tain confidence level after the IMF has been normalized by the total energy of all the oscillatory components.

The GASTA that we use in this study is introduced and studied in Jones et al. (1999), Folland et al. (2001a), and Folland et al. (2001b). The GASTA and its IMF components are displayed in Fig. 5.11. It is clear that

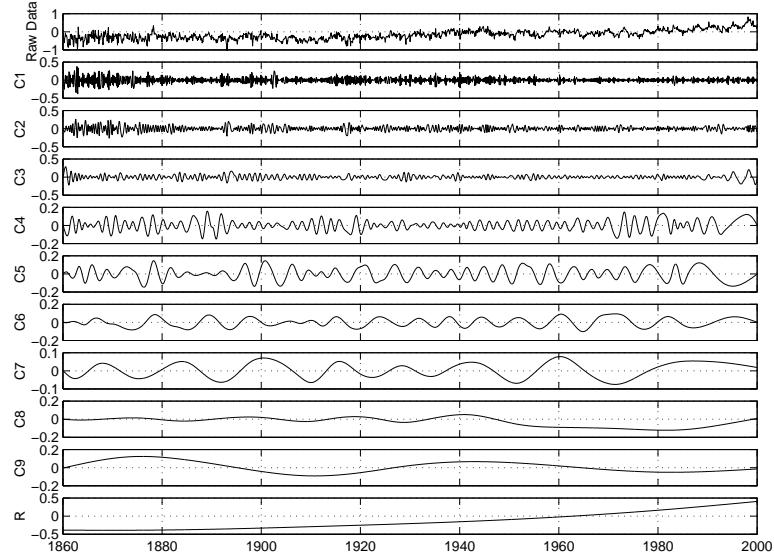


Figure 5.11: Same as Fig. 5.7, but for GASTA.

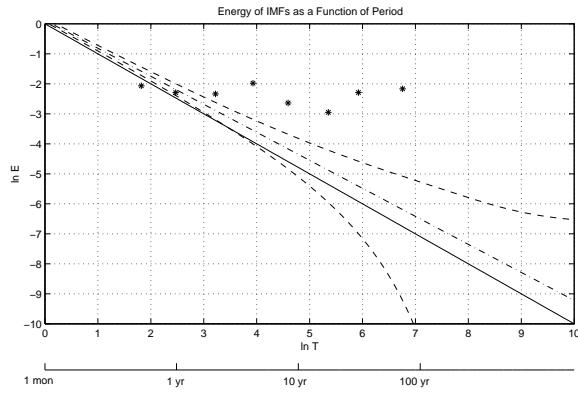


Figure 5.12: Same as Fig. 5.8, but for GASTA.

the trend is very large. The temperature increased by about 0.8°C over 140 yr from 1860 to 2000.

The significance test of IMFs 2–9 is displayed in Fig. 5.11. The results reveal that IMFs 4–9 are significant. However, this conclusion is misleading. In the discussions that we presented so far in this chapter, we have

implicitly or explicitly assumed that the noise included in data has been Gaussian white noise. However, the possibility cannot be ruled out that the noise in the data is not Gaussian white noise. If the GASTA does not contain Gaussian white noise, the statistical significance of an IMF against Gaussian white noise does not mean the IMF is physically or even statistically significant. Indeed, the oscillatory part of GASTA is more like a fractional Gaussian noise (fGn) series with its Hurst exponent close to 0.9. Therefore, we must be cautious about concluding that an IMF is statistically significant.

5.4.4. *A posteriori* test

The examples presented above are all *a priori* tests in which we do not know the noise level in the data. However, if we can ascertain that any specific IMF contains little useful information, then we can assume that the energy of that IMF comes solely from noise, and assign it on the 99% line. Then, we can use the energy level of that IMF to re-scale the rest of the IMFs. If the energy level of any IMF lies above the theoretical reference white noise line, we can assume that this IMF contains statistically significant information. If the rescaled energy level lies below the theoretical white noise line, then we can assume that the IMF contains little useful information. The latter is called a *posteriori* test. An example of a *posteriori* test is included in Wu and Huang (2003, 2004).

5.5. Summary and discussion

The properties of Gaussian white noise were studied by using the EMD method. We carried out numerical experiments to decompose uniformly distributed white noise into IMFs. The empirical findings were almost identical to those given in Wu and Huang (2003, 2004) for uniformly distributed white noise. Therefore, we followed the example of Wu and Huang (2003, 2004) when we deduced expressions for the various statistical properties of Gaussian white noise. These results were all tested by using the Monte Carlo method.

The known characteristics of Gaussian white noise were used to design statistical test methods that were applied to three well-known climate time series: (1) the North Atlantic Oscillation index (NAOI), which is close to pure Gaussian white noise; (2) the Southern Oscillation index (SOI), which contains dominant oscillation on interannual timescales; and (3) the global averaged surface temperature anomaly (GASTA), which contains a large

trend and possibly noise other than Gaussian white noise. The tests of the NAOI and the SOI obtained results consistent with those of the tests using Fourier analysis methods, indicating that our methods are valid and effective when data contain Gaussian white noise. The results of the test of the GASTA, which likely contains other types of noise, seem to exaggerate the significance of some IMFs. To eliminate this problem, the methods that can test against other types of noise should be used. Our proposed testing method is not just a trivial testing method, but is consistent with other analysis methods. The EMD method's results give us the following additional information: First, EMD identifies the significant IMFs. Because IMFs are adaptive, they can represent the underlying processes more effectively than pure sinusoids. Furthermore, the IMFs isolate physical processes of various time scales and also give the temporal variation with the processes in their entirety without resorting to the linear assumption as in the Fourier-based decomposition. Because we are free of harmonics, the IMFs can show the nonlinear distortion of the waveform locally, as was discussed by Huang et al. (1998) and Wu et al. (2001). Finally, the IMFs can be used to construct the time-frequency distribution in the form of the Hilbert spectrum, which offers minute details of the time variation of the underlying processes.

A *posteriori* test method was also discussed. The IMFs are more effective in isolating the physical processes of various timescales than the usual Fourier component.

Acknowledgements

The authors would like to thank Professor Samuel Shen of the University of Alberta for suggesting to us to study the Fourier spectra of the IMFs from the white noise. Throughout our study, he also offered valuable comments and encouragement. The authors also gratefully acknowledge the assistance of Dr. Dean G. Duffy at NASA/GSFC in preparing our manuscript. ZW is supported by NSF under grant ATM-0342104, and NEH is supported in part through a grant from NASA RTOP Oceanic Processes Program, an ONR Processes and Prediction Program grant number N00014-98-F-0412, and a NOAA Climate Center grant number NEEF4100-3-00269.

References

- Allan, R. J., N. Nicholls, P. D. Jones, and I. J. Butterworth, 1991: A further extension of the Tahiti-Darwin SOI, early ENSO events and Darwin pressure. *J. Climate*, **4**, 743–749.

- Flandrin, P., G. Rilling, and P. Gonçalvès, 2004: Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.*, **11**, 112–114.
- Folland, C. K., N. A. Rayner, S. J. Brown, T. M. Smith, S. S. P. Shen, D. E. Parker, I. Macadam, P. D. Jones, R. N. Jones, N. Nicholls, and D. M. H. Sexton, 2001a: Global temperature change and its uncertainties since 1861. *Geophys. Res. Lett.*, **28**, 2621–2624.
- Folland, C. K., T. R. Karl, J. R. Christy, R. A. Clarke, G. V. Gruza, J. Jouzel, M. E. Mann, J. Oerlemans, M. J. Salinger, and S.-W. Wang, 2001b: Observed climate variability and change. *Climate Change 2001: The Scientific Basis*, J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds., Cambridge University Press, 99–181.
- Ghil, M., M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, 2002: Advanced spectral methods for climatic time series. *Rev. Geophys.*, **40**, 10.1029/2000GR000092.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of water waves – The Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Hurrell, J.W., 1995: Decadal trends in the North Atlantic Oscillation and relationships to regional temperature and precipitation. *Science*, **269**, 676–679.
- Jones, P. D., T. Jonsson, and D. Wheeler, 1997: Extension using early instrumental pressure observations from Gibraltar and SW Iceland to the North Atlantic Oscillation. *Int. J. Climatol.*, **17**, 1433–1450.
- Jones, P. D., M. New, D. E. Parker, S. Martin, and I. G. Rigor, 1999: Surface air temperature and its variations over the last 150 years. *Rev. Geophys.*, **37**, 173–199.
- Papoulis, A., 1986: *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 576 pp.
- Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in C*. Cambridge University Press, 994 pp.
- Ropelewski, C. F., and P. D. Jones, 1987: An extension of the Tahiti-Darwin Southern Oscillation Index. *Mon. Wea. Rev.*, **115**, 2161–2165.
- Trenberth, K. E., 1984: Signal versus noise in the Southern Oscillation. *Mon. Wea. Rev.*, **112**, 326–332.

- Wu, Z., E. K. Schneider, Z.-Z. Hu, and L. Cao, 2001: The impact of global warming on ENSO variability in climate records. *COLA Technical Rep.* **110**, 24 pp.
- Wu, Z., and N. E. Huang, 2003: A study of the characteristics of white noise using the empirical mode decomposition method. *COLA Technical Rep.* **133**, 27 pp.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.

Zhaohua Wu

Center for Ocean-Land-Atmosphere Studies, 4041 Powder Mill Road, Suite 302, Calverton, MD 20705, USA
zhwu@cola.iges.org

Norden E. Huang

Goddard Institute for Data Analysis, Code 614.2, NASA/Goddard Space Flight Center, Greenbelt, MD 20771, USA
norden.e.huang@nasa.gov

CHAPTER 6

THE APPLICATION OF HILBERT-HUANG TRANSFORMS TO METEOROLOGICAL DATASETS^{||}

Dean G. Duffy

Recently a new spectral technique has been developed for the analysis of aperiodic signals from nonlinear systems – the Hilbert-Huang transform. It is shown how this transform can be used to discover synoptic and climatic features: For sea level data, the transforms capture the oceanic tides as well as variations in precipitation patterns. In the case of solar radiation, variations in the diurnal and seasonal cycles are observed. Finally, from barographic data, the Hilbert-Huang transform reveals the passage of extratropical cyclones, fronts, and troughs. Thus, this technique can detect signals on synoptic to interannual time scales.

6.1. Introduction

For generations researchers have used Fourier analysis to analyze signals. Because both the signal and its Fourier transform are important in understanding most processes, contour plots of the signal energy as functions of time and frequency (temporal-frequency analysis) have the potential of painting a more revealing picture than just the temporal signal or frequency analysis alone.

The earliest time-frequency representation (TFR) was the short-time Fourier transform (STFT; see Allen and Rabiner 1977). This scheme divides the temporal signal $f(t)$ into a series of small overlapping pieces. Each piece is then windowed and individually Fourier transformed. The STFT of a function $f(t)$ is defined by

$$F_{\text{ST}}(t, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)h(t - \tau)e^{-i\omega\tau} d\tau, \quad (6.1)$$

^{||}This paper is reprinted with the kind permission of the American Meteorological Society which holds its copyright. It originally appeared in the *Journal of Atmospheric and Oceanic Technology*, **21**, No. 4, 599–611, in an article by Dean G. Duffy entitled “The Application of Hilbert-Huang Transforms to Meteorological Datasets.”

where $h(\tau)$ is the window function. Contour plots of the energy density function $|F_{\text{ST}}(t, \omega)|$ are typically presented. This scheme is most useful when the physical process is linear, so that the superposition of sinusoidal solutions is valid and the time series is *locally* stationary, so that the Fourier coefficients are slowly changing.

One of the drawbacks of STFT is the presence of a fixed window although Czerwinski and Jones (1997) have developed a short-time Fourier analysis with an adaptively adjusting window. Wavelet analysis (see Daubechies 1992; Torrence and Compo 1998) seeks to address this defect by decomposing the time series into *local*, time-dilated, and time-translated wavelet components using time-frequency atoms or wavelets ψ . The wavelet transform of the signal $f(t)$ is then

$$F_{\text{WT}}(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-b}{a}\right) dt, \quad (6.2)$$

where a is the scale and b is the time shift. The wavelet transform represents the energy in the signal of temporal scale a at $t = b$. Wavelet analysis is attractive because 1) it is local, although higher frequencies are more localized; 2) it has uniform temporal resolution for all frequency scales; and 3) it is useful for characterizing gradual frequency changes. However, it is nonadaptive because the same basic wavelet is used for all data.

Finally, empirical orthogonal function (EOF) analysis or its Fourier transform version, the singular spectral analysis (SSA), decomposes a time series using eigenfunctions of the covariance matrix (see Ghil et al. 2002). This analysis is quite different from the short-time Fourier transform or wavelet analysis because the EOFs are derived from data. However, its distribution of eigenvalues does not yield characteristic time or frequency scales. Furthermore, the eigenfunctions themselves are not necessarily linear or stationary and therefore are not easily analyzed by spectral modes.

From our experience with short-time Fourier transforms, wavelets, and EOF analyses, an ideal scheme for the spectral analysis of signals would be complete (i.e., the sum of the modes equals the original signal), orthogonal, local and adaptive. This method would also allow us to extract local time and frequency scales. The Hilbert-Huang transform is another step toward this goal.

Hilbert transforms were originally developed to solve integral equations. Instead of re-expressing a function of time with its Fourier transform that depends on frequency, the Hilbert transform yields another temporal func-

tion that has been phase shifted by -90° via the integral definition:

$$\hat{f}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(\tau)}{t - \tau} d\tau. \quad (6.3)$$

By itself, it holds little interest for us. However, when Gabor (1946) developed his theory of communications Hilbert transforms appeared in his concept of *analytic signal*, $z(t) = f(t) + i\hat{f}(t)$. A particularly interesting case occurs if $f(t)$ is band limited. Then we can rewrite $z(t) \approx A(t)e^{i\theta(t)}$, a local time-varying wave with amplitude $A(t)$ and phase $\theta(t)$.

Most signals are not band-limited. Huang's great contribution was to devise a method, which he calls *sifting*, that decomposes a wide class of signals into a set of band-limited functions, which he calls *intrinsic mode functions* (IMFs). In their original paper, Huang et al. (1998a) tested out their analysis on simple nonlinear systems such as Stokes waves and the solutions to the Duffing and Lorenz equations. Subsequently, the Hilbert–Huang analysis has been applied to signals from pulmonary blood pressure (Huang et al. 1998b) to earthquakes (Huang et al. 2001) to the rotational residuals from the solar convection zone (Komm et al. 2001). In the atmospheric sciences, these transforms have been applied to climatic signals (Pan et al. 2003, Salisbury and Wimbush 2002, Wu et al. 1999, Xie et al. 2002). The purpose of this paper is to illustrate the advantages of applying Hilbert–Huang transforms to signals that include synoptic as well as climatic signals.

A detailed description of the scheme is provided in section 6.2. In section 6.3, this scheme is then applied to datasets of sea level heights, incoming solar radiation, and barographic data. Conclusions are presented in section 6.4.

6.2. Procedure

The process of developing time-frequency diagrams using Hilbert–Huang transforms consists of three steps. The first step decomposes or “sifts” the signal into its intrinsic mode functions. The first intrinsic mode gives the smallest *local* variations of the original signal. Using a cubic spline, two envelope curves are generated; one of which connects the maxima of the signal while the other connects the minima. From these curves the mean is computed. The difference between the signal and this mean constitutes a first guess of the first IMF. If the IMF were sinusoidal, then the number of extrema would equal the number of zero crossings, or differ by one. This is usually not the case and suggests that our first guess, while good, needs

further refinement because there may be yet smaller scales buried in the data.

Originally Huang et al. [1998a, their equation (5.5)] repeated this sifting process until a Cauchy-like integral condition was satisfied. In later papers they adopted a stopping criterion based on the number of extrema and zero-crossings. When these quantities were equal, or differed by one, for three consecutive iterations, the IMF is set to the values found in the final iteration. This is the criterion that shall be used. Although it cannot be proven mathematically, this procedure always converged for the datasets tested here.

To illustrate the decomposition process, Fig. 6.1 presents various steps in computing the first IMF. Figure 6.1a presents a small portion of the original dataset—sea level observations taken at the mouth of the Chesapeake Bay during the 1980s. In addition to the oscillations due to the oceanic tides, there is a large peak at 133 h due to the passage of Hurricane Gloria during the early hours of 27 September 1985.

Figures 6.1b and 6.1c show the first IMF after it has satisfied the convergence criteria for the first and last (third) time, respectively. Also shown are the top and bottom envelopes (the dashed lines) as well as the mean (the dotted line). Note how the sifting process has generated a mode that is symmetrical with respect to the abscissa.

It will be shown shortly that this first IMF represents the semidiurnal tides. This mode varies smoothly except in the interval 125–150 h. If the original data record in Fig. 6.1a is examined this behavior can be associated with several “kinks” in the data record. It is unclear whether these kinks are real or due to a failure of the instrument.

To examine the effects that such kinks have on the construction of the IMFs, the original data was smoothed with a single pass of a simple Shapiro filter [a digital filter developed by Ralph Shapiro (1970) that eliminates waves with a period of $2\Delta t$ but leaves longer-period waves relatively unaffected]. After sifting, Fig. 6.1d gives the first IMF of the original (solid line) and smoothed (dashed line) data. This improved behavior in the structure of the IMF might argue for smoothing the data before applying Hilbert-Huang transforms. Unfortunately, filtering is a two-edged sword: It modifies both the true signal and eliminates noise. Because Huang-Hilbert transforms were designed specifically to analyze aperiodic and nonlinear signals (signals from nonlinear systems), the use of a linear filter could alter the signal in way that might compromise the usefulness of the technique. For this reason, the presence of the noise will simply be endured.

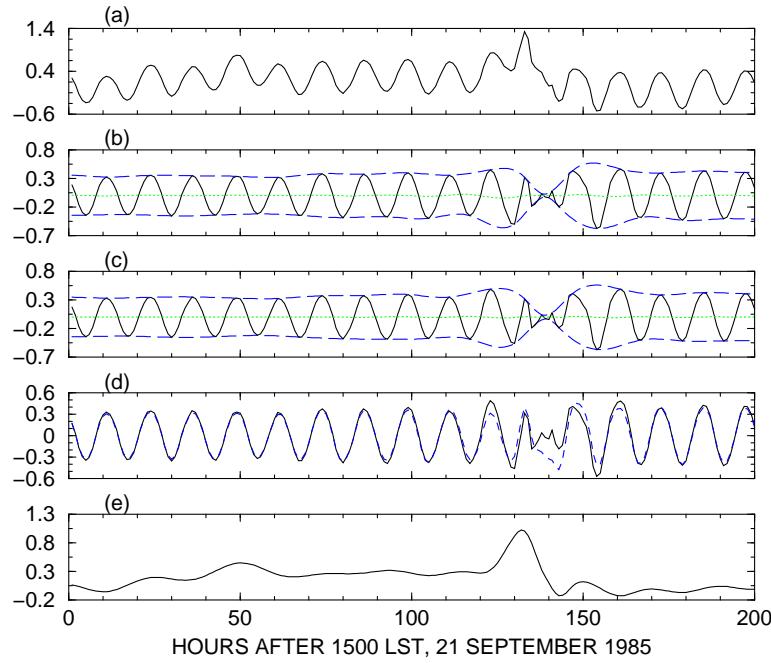


Figure 6.1: The computation of the first intrinsic mode function from a sea level dataset. The solid line in (a) shows a portion of the original data with sea level height (in m). The solid lines in (b) and (c) show the first intrinsic mode when the mode's extrema and zero crossings are equal or differ by one for the first and third times, respectively. The dashed lines give the envelopes, while the dotted line is the mean. (d) The first IMF from the original dataset (solid line) and the original data after applying a Shapiro filter (dashed line). (e) The signal after the first intrinsic mode has been removed.

Once the first IMF is found, it is subtracted from the signal, yielding a residual that is smoother and has a lower frequency than the original signal because the first IMF captures the smallest *local* variation of the signal (see Fig. 6.1e). The process then begins anew with the residual to obtain the second IMF.

Having found the second IMF, it is subtracted from the residual, leaving an even smoother residual. Further modes are found in the same manner, and the sifting process concludes when there is no longer any maxima or minima in the residual. The original signal equals the sum of the various IMFs plus this small residual trend.

One drawback in using a cubic spline to obtain the envelope curve is the possibility of large swings near the endpoints. The method of Komm et al. (2001), that adds buffer zones on each end that equal the length of the original data, has been adopted. The signal was extrapolated into these buffer zones by fitting a sine wave using the closest extrema and zero crossing to define the frequency and amplitude of the wave.

Having determined the IMFs, the second step consists of computing the Hilbert transform $\hat{f}(t)$ of each mode $f(t)$ and then the corresponding analytic signal $z(t) = f(t) + i\hat{f}(t)$. Because only numerical values are available, the conventional method of computing $\hat{f}(t)$ is to take the fast Fourier transform of the data, multiply the transform by $i \operatorname{sgn}(\omega)$, and then take the inverse Fourier transform of this product (see Čížek 1970 or Henrici 1986, p. 203). Here $\operatorname{sgn}()$ denotes the sign function.

Computing the instantaneous frequency $\omega(t)$ from the data is difficult because it is the time derivative of $\theta(t)$. Barnes (1992) tested a number of methods to compute it from $f(t)$ and $\hat{f}(t)$. He found that the best representation of the instantaneous frequency is

$$\omega(t) = \frac{1}{2\Delta t} \tan^{-1} \left[\frac{f(t - \Delta t)\hat{f}(t + \Delta t) - f(t + \Delta t)\hat{f}(t - \Delta t)}{f(t - \Delta t)f(t + \Delta t) + \hat{f}(t + \Delta t)\hat{f}(t - \Delta t)} \right], \quad (6.4)$$

where Δt denotes the time between observations. This is the method that will be used.

Figure 6.2 illustrates the instantaneous amplitude $A(t)$ and period $P(t) = 2\pi/\omega(t)$ corresponding to the first IMF shown in Fig. 6.1. The solid and dashed lines give the results for the original signal and Shapiro filtered signal, respectively. In both cases, the period equals approximately 12.5 h outside of the time interval from 125 and 150 h. Within the interval there are considerable differences between the original and smoothed data. Because analytic signals cannot have negative frequencies, the appearance of some suggests that the data are not correct here. For this reason, we will discard any amplitudes and periods when they are negative. This is acceptable because when the frequencies are negative the amplitude is small, as Fig. 6.2 shows.

Once the analytic signal for each IMF is obtained, the final task remains to display the results graphically. Although $A(t)$ and $P(t)$ could be plotted for each IMF, a better idea is to plot the square of the amplitude as a function of $P(t)$ and time, combining amplitude and period measurements of all IMF components on a single figure. This contour plot is commonly called the Hilbert spectrum, $H(t, \omega)$. Although two IMFs can have the same

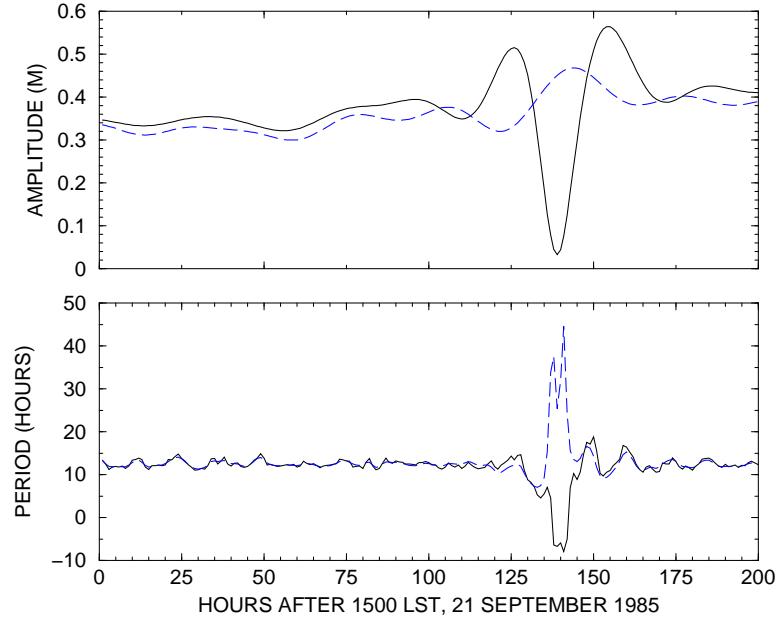


Figure 6.2: The solid line gives the amplitude and period of the first IMF shown in Fig. 6.1. The dashed line also gives the amplitude and period of the first IMF of the signal shown in Fig. 6.1 but the data have first been Shapiro filtered.

period, this will not occur at the same time and there is no ambiguity in constructing the Hilbert spectrum. Because most geophysical datasets contain a mixture of phenomena, with time scales varying from hours to years, it was found convenient to work with the logarithm of the instantaneous period rather than period itself.

One problem with displaying the results is the irregularity of the location of the frequencies associated with each IMF at a given instant. Not only are these locations irregular but they also vary with time. Although most graphical packages can handle such irregularly spaced data, the plots are too noisy. One possible solution would be to construct a smoothed field before plotting.

If the instantaneous amplitudes are viewed as “data” observed at various instantaneous periods, then regression techniques developed by statisticians for fitting a curve through data can be used (Ryan 1997, chapter 10). Because there is little a priori knowledge about the shape of the curve,

a nonparametric scheme that includes a kernel smoother is used. [Such a scheme was developed by Herrmann (1996); the FORTRAN 77 code is available online at <http://www.unizh.ch/biostat/Software/kernf77.html>.]

For large datasets, this method was modified to include time averaging. There are two possible methods. One method would group all of the instantaneous amplitudes and periods within a particular time interval and assign them a common time. The second method would find curve fits at each time and then these curves would be averaged over an appropriate time interval. Figure 6.3 shows the instantaneous amplitudes and periods at four instances during the 24-h interval from 2330 LST 24 September 1985 to 2330 LST 26 September 1985 using sea level data. The solid line shows the grouping of instantaneous amplitudes and phases to form a single time value, while the dashed line shows the averaging of the curve fits. The first method retains the character of the higher-temporal-resolution amplitudes compared to the second method.

As will be seen shortly, time-frequency plots contain a wealth of detail. For that reason, it is useful to integrate $H(t, \omega)$ over time, say from 0 to T . Because this *marginal spectrum* represents the energy of the signal, it is analogous to the power spectrum in Fourier analysis. Here a monochromatic, linear and periodic signal appears as a sharp peak in the marginal spectrum, whereas a nonlinear and nonperiodic phenomenon yields a broad peak in the spectrum. If the marginal spectrum is normalized by T , we have the *average marginal spectrum*.

6.3. Applications

Having presented the procedure for constructing a Hilbert-Huang transform, this technique is now applied to three very different data fields: sea level heights, incoming solar radiation, and barographic observations.

6.3.1. Sea level heights

As a first application of the Hilbert-Huang transform to geophysical data, observations were obtained of the sea level heights observed at the Chesapeake Bay Bridge and Tunnel (CBBT) from 0000 LST 8 April 1984 to 2300 LST 28 September 1990. These 48 000 hours of consecutive observations contain a wealth of physical phenomena—from highly predictable, astronomically forced tides to chaotic coastal storms.

Figure 6.4 shows the Fourier power spectrum and average marginal spectrum for the sea level heights. The Fourier analysis only reveals a semidiur-

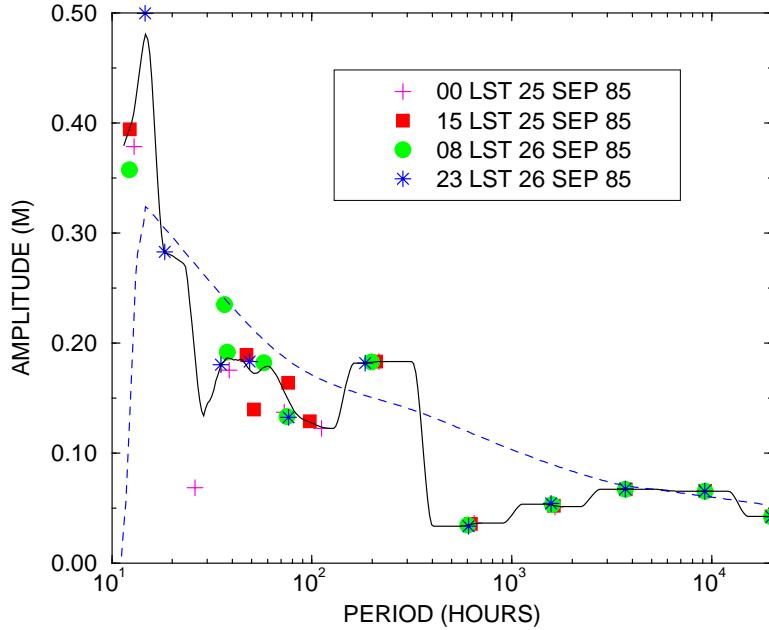


Figure 6.3: Using sea level data, the curve fit when instantaneous amplitudes and periods over a 24-h interval are grouped together to form a single Hilbert-Huang transform (solid line). The dotted line is the average of 24 curve fits to each hour's instantaneous amplitudes and periods. Some of the instantaneous amplitudes used in computing the averages are plotted as data points.

nal tide with a period of 12.42 h and a diurnal tide with a period of 23.93 h. The average marginal spectrum also reveals the semidiurnal and diurnal tides. This is consistent with a study by Susanto et al. (2000), who showed that Hilbert-Huang transforms capture the diurnal and semidiurnal tides in the Makassar Straits. Furthermore, the marginal spectrum captures a wealth of other physical phenomena. For example, there is a peak near 100 h due to baroclinic instability. (Baroclinic instability is the instability of the prevailing atmospheric westerlies that generates cyclones. Its time scale is on the order of days.) At lower periods, we see peaks with annual and semiannual periods associated with the increased precipitation during the warmer months and the melting of snow during the spring. Finally, there are peaks associated with the intraseasonal event of El Niño/Southern Oscillation (ENSO).

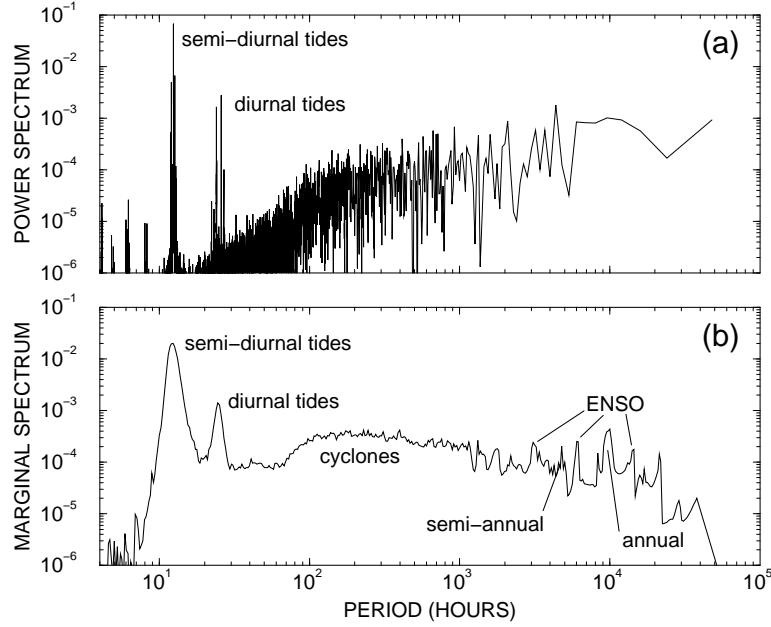


Figure 6.4: The (a) Fourier power spectrum and (b) average marginal spectrum (in m^2) derived from 48 000 h of sea level observations at the mouth of the Chesapeake Bay.

Figure 6.5 shows the Hilbert spectrum. A persistent maximum is found at approximately 12 h; this is the semidiurnal tides. A modulation in the amplitude that has a period of approximately one month is noted; this is the lunar forcing of the tides. There is also a faint green line at 24 h (denoted by the letter A), which is the diurnal tides.

At longer periods there are many local maxima. Two particularly large ones are highlighted in Fig. 6.5 by the letters B and C. Maximum B occurs during mid-November 1985 and is associated with the “Election Day Flood,” 4–7 November 1985, the second-worst river flood in Virginia during the twentieth century. During this event, heavy rain occurred in the mountainous regions of Virginia and West Virginia, and shown in Fig. 6.5 is the corresponding runoff.

February 1989 was the snowiest month (62 cm) in 50 yr (1948–1999) for Norfolk, Virginia, due in large part from its heaviest 24-h snowfall (36 cm), which occurred during 17–18 February 1989. Point C in Fig. 6.5 corresponds to the melting and runoff of this snowpack during late February and early

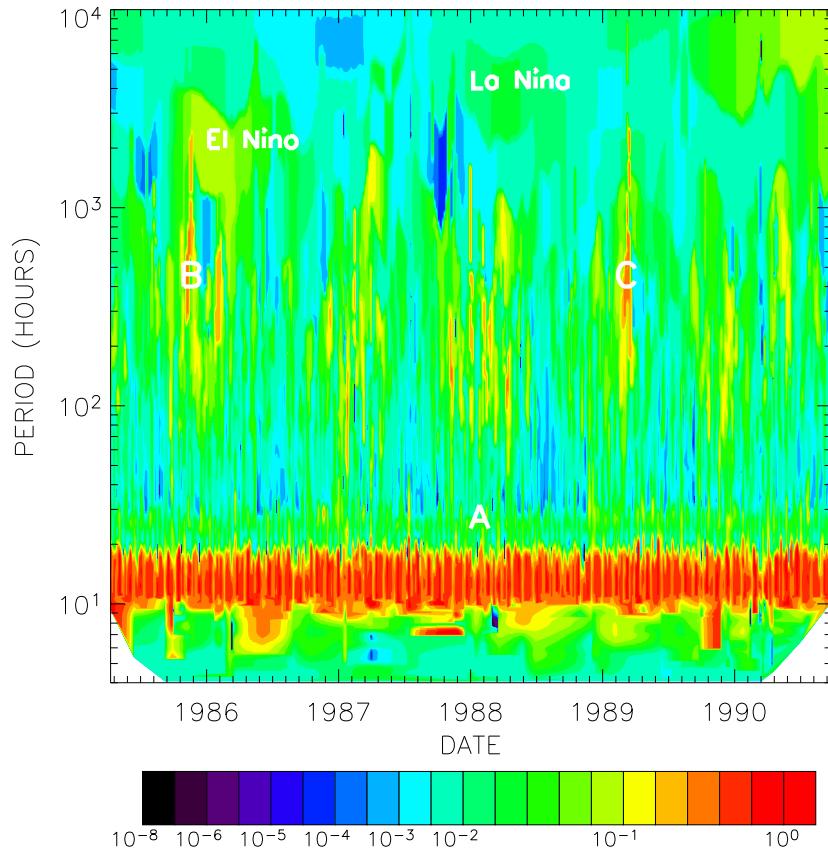


Figure 6.5: Contour plot of the Hilbert spectrum of the sea level heights (in m^2) at the mouth of the Chesapeake Bay as a function of time and period. The amplitudes were time averaged, as described in the text, over 480 h.

March 1989.

During the late 1980s, there were two significant ENSO events: an El Niño event in 1986 and a La Niña event in 1988. Because ENSO influences the precipitation and temperature along the U. S. East Coast, it is not surprising to see its signatures in the Hilbert spectrum.

The above analysis of Fig. 6.5 has shown that one of the possible uses of Hilbert-Huang transforms is the identification of significant events from the

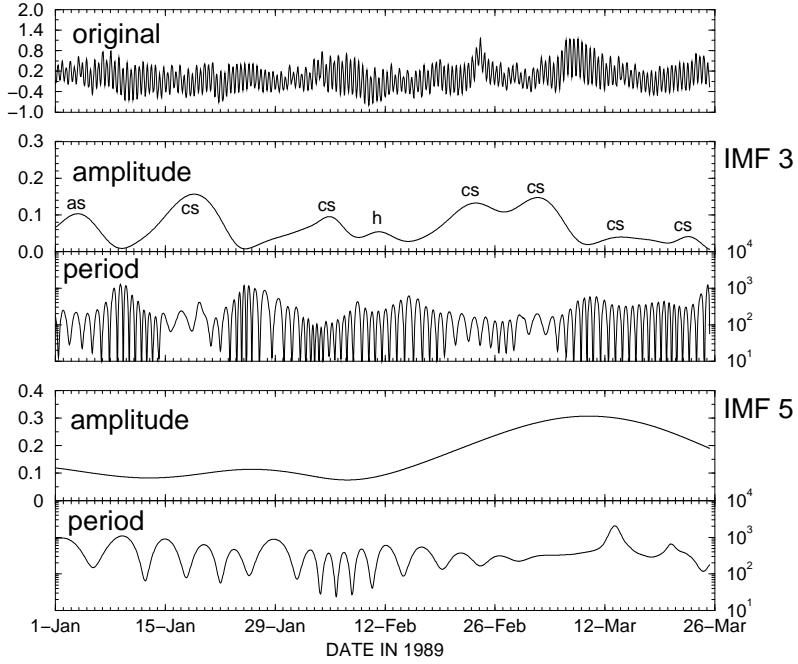


Figure 6.6: A portion of third and fifth IMFs from the Hilbert-Huang analysis of 48 000 h of sea level heights taken at the mouth of the Chesapeake Bay. For the IMFs, the scale for the amplitudes (in m) is given on the left while the period (in h) is given on the right.

Hilbert spectrum. However, one might ask whether the original data could not simply be used. To answer this question, the original sea level data from the first 2000 h of 1989 have been plotted in Fig. 6.6. Also included are the third and fifth IMFs for the same period of time. By consulting daily weather maps, each peak in the third IMF's amplitude could be traced to the presence of cyclones in the North Atlantic (labelled as), coastal storms (cs), or a strong high pressure (h) located over New England, and its strong northerly winds. It would be very difficult to detect these phenomena in the original data. The large maximum in the fifth IMF's amplitude is due to the spring melt of the snow pack in the Appalachian Mountains.

As mentioned in the introduction, Hilbert-Huang transforms are one of several temporal-frequency techniques. A particularly popular one is wavelet analysis. Figure 6.7 shows the wavelet analysis [using a IDL package by Torrence and Compo (1998)] for our sea level height data. Clearly seen in the power spectrogram is the presence of the tides, storms (includ-

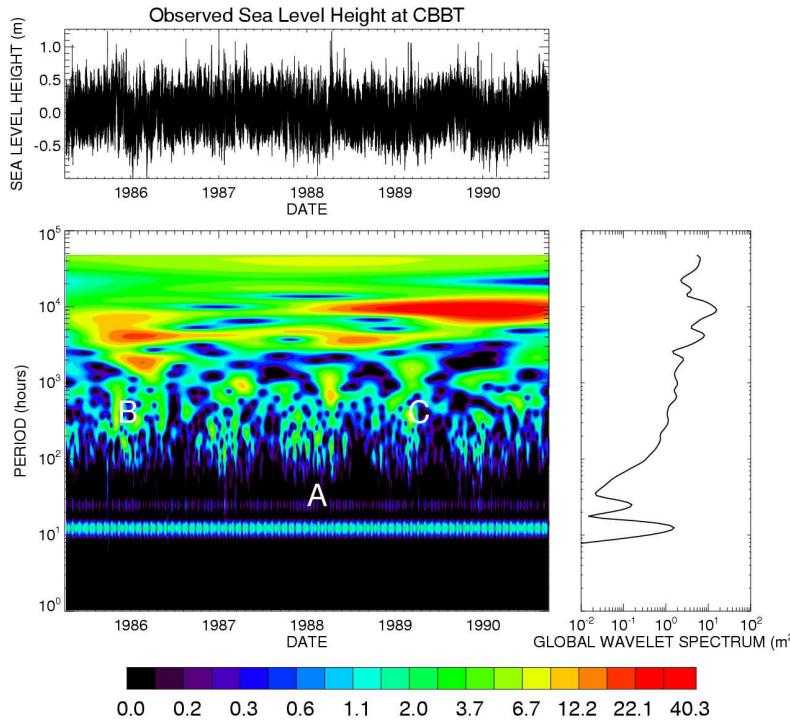


Figure 6.7: Morlet wavelet analysis of sea level heights at the mouth of the Chesapeake Bay. The letter A denotes the diurnal tides, while B and C denote the Election Day Flood of Nov 1985 and the Feb 1989 snowstorm, respectively.

ing the Election Day Flood and 1989 snowstorm), and two ENSO events. The strong signal at $\sim 10^4$ h is from variations in the seasonal cycle. The primary difference is the smoothness of the spectrogram compared to the Hilbert spectrum.

In summary, the Hilbert-Huang transform can discern persistent, periodic features such as the tides, as well as episodic events such as a snowmelt and heavy precipitation events.

6.3.2. Solar radiation

For the second dataset, the hourly diffuse horizontal radiation (the amount of solar radiation from the sky excluding the solar disk) for Des Moines, Iowa, from 1 January 1980 through 13 Decem-

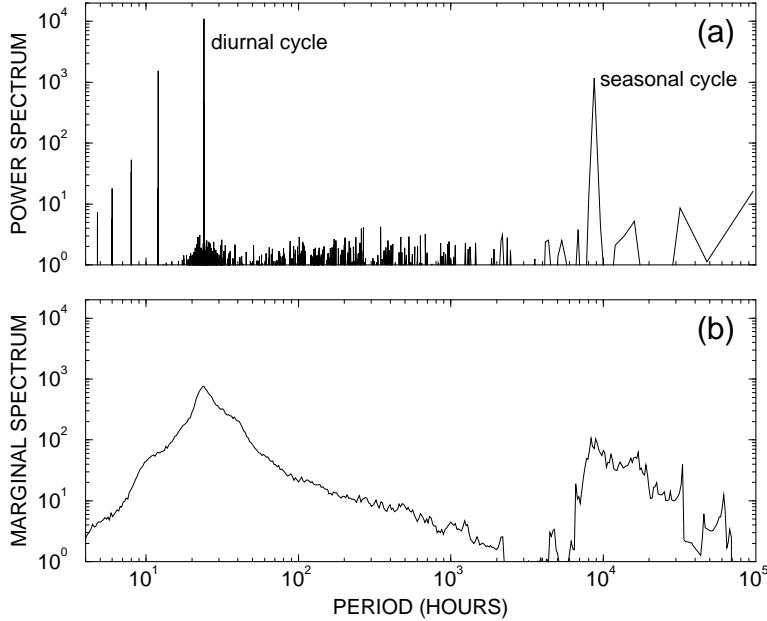


Figure 6.8: The (a) Fourier power spectrum and (b) average marginal spectrum [in $(\text{Wm}^{-2})^2$] derived from 96 000 h of diffuse solar radiation measured on a horizontal surface at ground level in Des Moines, IA.

ber 1990 is analyzed. (These observations may be found online at http://rredc.nrel.gov/solar/old_data/nsrdb/hourly.)

Figure 6.8 shows the Fourier power spectrum and average marginal spectrum. As one might expect, there are two peaks, one corresponding to the diurnal cycle and the other to the seasonal cycle.

Figure 6.9 shows the Hilbert spectrum for the signal. The maximum values occur along the line corresponding to a period of 24 h, the diurnal cycle, and during the summer months. For periods between 10^2 and 10^3 h, there are relative minima during the winter months due to the increased cloud cover that reduces the incoming radiation at ground level. Finally, for periods near 10^4 h the seasonal cycle is clearly seen. Of particular interest are the maxima that occur during 1982 and 1988. What physical phenomenon is the Hilbert-Huang transform detecting?

Recently, Koch and Mann (1996) studied the spatial and temporal variation of ^{7}Be , a natural radionuclide that is produced in the stratosphere and

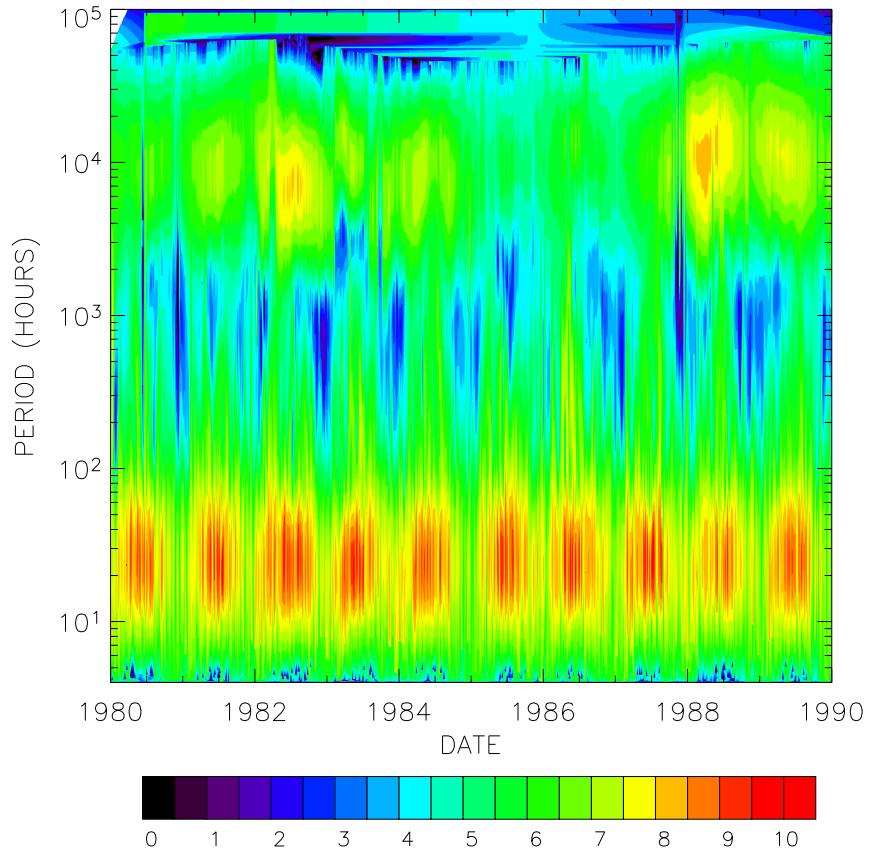


Figure 6.9: Contour plot of the (natural) logarithm of the Hilbert spectrum of diffuse solar radiation [in $(\text{Wm}^{-2})^2$] as a function of time and period. The amplitudes were time averaged, as described in the text, over 480 h.

upper troposphere and is carried on aerosols. They found that ENSO events could significantly affect the concentration of this radionuclide. They suggested that this signal was due to observed rainfall anomalies during ENSO and its effect on the aerosol concentration in the troposphere. Because the amount of sunlight scattered by aerosols depends on their concentration, our maxima correspond to the El Niño event of 1982, as well as the eruption

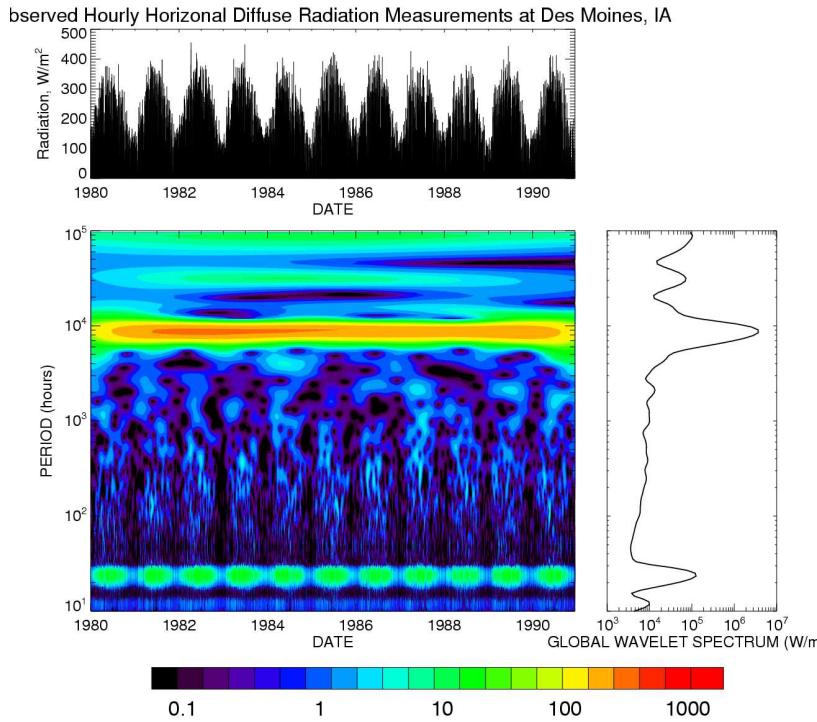


Figure 6.10: Morlet wavelet analysis of diffuse solar radiation over Des Moines, IA, during the 1980s.

of El Chichón, , and the La Niña event of 1988. Apparently the El Niño event of 1986 did not significantly affect the aerosol concentration of Des Moines.

Finally, in Fig. 6.10, the wavelet analysis of our solar data is presented. It captures the diurnal and seasonal cycle (the signal at $\sim 10^4$ h) and the El Niño event of 1982. However, there is no maxima for either the 1986 or 1988 ENSO events, and the spectrogram is much smoother than the Hilbert-Huang analysis.

6.3.3. Barographic observations

Our third test involves barographic data averaged to produce observations at 2-min intervals from a network located in central Illinois (Grivet-Talocia et al. 1999). These data were grouped by station and quarter of the year.

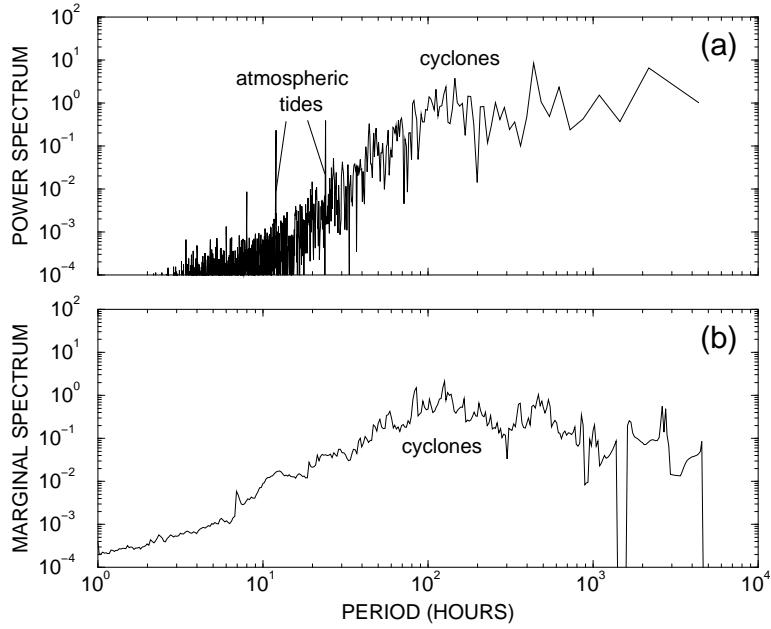


Figure 6.11: The (a) Fourier power spectrum and (b) average marginal spectrum (in hPa^2) derived from barographic observations near Champaign-Urbana, IL, during the latter half of 1995.

Figure 6.11 gives the Fourier power and average marginal spectrum for the latter half of 1995. The largest peak occurs at 100 h and is due to the dominance of baroclinic instability—the instability of the prevailing atmospheric westerlies that creates cyclones and anticyclones. It is interesting to note that conventional Fourier analysis detected the atmospheric tides.

Figure 6.12 is the Hilbert spectrum for the final quarter of 1995. Overall, the amplitudes increase with time due to the onset of winter. Using the daily weather maps from 1200 UTC published by the National Oceanic and Atmospheric Administration (NOAA), each maximum was easily correlated with the nearby passage of a cyclone and its associated front; these maxima have been labeled A–F in Fig. 6.12. For example, maximum A corresponds to the passage of the remnants of Hurricane Opal on 6 October 1995. The passage of a particular strong cold front on 11 November 1995 is highlighted with the letter G.

Because these data have a high temporal resolution, one would hope to see mesoscale signals as well as synoptic events. To this end, Fig. 6.13 shows

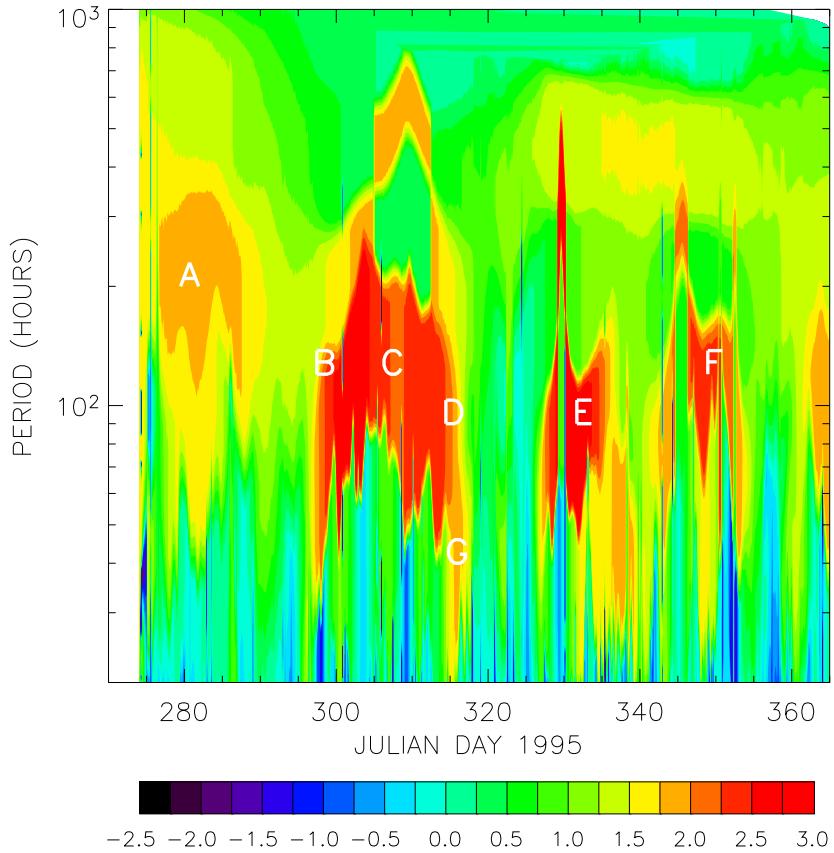


Figure 6.12: Contour plot of the (natural) logarithm of the Hilbert spectrum of surface pressure (in hPa^2) as a function of time and period. The amplitudes were time averaged, as described in the text, over 128 minutes.

the fifth and eighth IMFs as a function of time (19–29 October 1995), while the original data is plotted on top. Turning to the fifth IMF, many of the peaks could be paired with meteorological events such as the passage of a cold front (cf), a warm front (wf), and trough (t). There are probably other mesoscale events present in this analysis, but the lack of other independent data precludes their classification. The eighth IMF clearly shows the passage

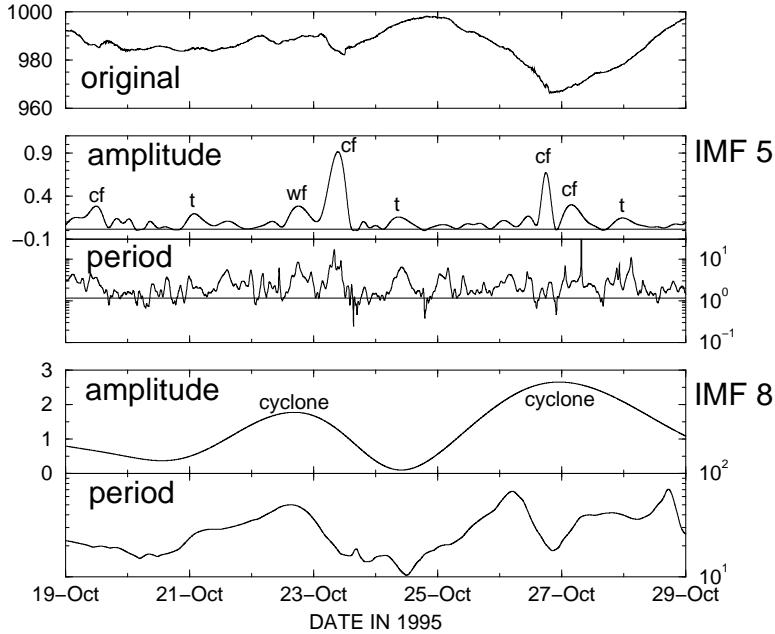


Figure 6.13: The fifth and eighth IMFs from the Hilbert-Huang analysis of surface pressure measurements over central Illinois during the 10 days from 19 to 29 Oct 1995. For the IMFs, the scale for the amplitudes (in hPa) is given on the left, while the period (in h) is given on the right.

of two major cyclones during this period. Taken together, Figs. 6.12 and 6.13 show how well the Hilbert-Huang transform can discern various mesoscale and synoptic events.

Finally, Fig. 6.14 shows the Morlet wavelet analysis of the same pressure observations used earlier. The strong signal at ~ 500 h is associated with the transition from the summer regime to winter. The wavelet analysis agrees with the Hilbert-Huang analysis except that it fails to pick out the individual cyclones and corresponding frontal passages that the Hilbert-Huang transform captures.

6.4. Conclusion

The purpose of data analysis is to discover the physical processes underlying the observations. For generations scientists have only had Fourier analysis. As our analysis has shown, this technique is quite good for periodic signals,

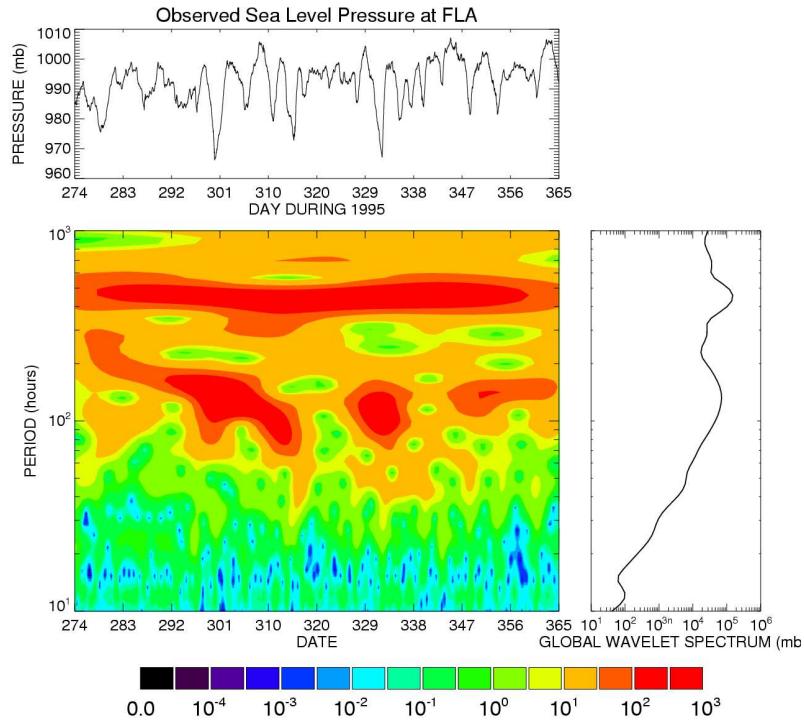


Figure 6.14: Morlet wavelet analysis derived from barographic observations near Champaign-Urbana, IL, during the last quarter of 1995.

such as atmospheric and oceanic tides. However, it is inadequate for most meteorological signals because they are nonlinear and aperiodic.

Recently several new techniques (short-time Fourier transforms, wavelets, and EOFs) have been developed to handle these aperiodic and nonlinear signals. To this list we now add the promising technique commonly known as Hilbert-Huang transforms because it is adaptive, local, complete, and nearly orthogonal in the sense of Reynolds decomposition $(f(t) - \bar{f}(t)) \cdot \bar{f}(t) = 0$, where the overbar denotes a local average [see Huang et al. 1998a, their Eq. (6.1)].

The procedure for computing a Hilbert-Huang transform consists of two steps. First, the data is sifted and decomposed into a set of intrinsic mode functions. If the data consisted of a pure sine wave, then its first (and only) IMF would be the original sine wave. The second step consists of taking

the Hilbert transform of each IMF and then computing the instantaneous amplitude and frequency. In the case of a pure sine wave, the Hilbert transform would be a negative cosine and both the instantaneous amplitude and frequency are constant.

So far this technique has only been applied to climatological datasets. Here the method was tested on datasets that also contain synoptic signals: sea level heights, incoming solar radiation, and barographic observations. These tests showed that the Hilbert-Huang transforms capture a wide variety of phenomena: the diurnal cycle, frontal passages, baroclinic instability, and the seasonal cycle. Therefore, we can use this technique to flag important weather events from floods to ENSO events.

In this paper two methods for presenting Hilbert-Huang transforms are highlighted. One method plots the instantaneous amplitude as a function of time and (instantaneous) period or frequency via the Hilbert spectrum. These plots are useful in suggesting the location and nature of significant events in the original data. Unfortunately, a significance test has not yet been developed for the Hilbert-Huang transform. The second method is the marginal spectrum. It is similar to the popular concept of power spectrum in Fourier analysis and gives a global picture of the dataset.

Acknowledgments

The author would like to gratefully acknowledge Dr. N. E. Huang for allowing the use of his technique and providing considerable insight into the mechanics of how it works. The barographic data was kindly provided by Dr. Stefano Grivet-Talocia. This paper was greatly improved by suggestions from Dr. David O.C Starr and two anonymous reviewers.

References

- Allen, J. B., and L. R. Rabiner, 1977: A unified approach to short-time Fourier spectrum analysis and synthesis. *Proc. IEEE*, **65**, 1558–1564.
- Barnes, A. E., 1992: The calculation of instantaneous frequency and instantaneous bandwidth. *Geophysics*, **57**, 1520–1524.
- Čížek, V., 1970: Discrete Hilbert transform. *IEEE Trans. Audio Electroacoust.*, **AU-18**, 340–343.
- Czerwinski, R. N., and D. L. Jones, 1997: Adaptive short-time Fourier analysis. *IEEE Signal Process. Lett.*, **4**, 42–45.
- Daubechies, I., 1992: *Ten Lectures on Wavelets*. CBMS-NSF Series in Applied Mathematics. Vol. 61, SIAM, 357 pp.

- Gabor, D., 1946: Theory of communications. *J. IEE*, **93**, 429–457.
- Ghil, M., M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, 2002: Advanced spectral methods for climatic time series. *Rev. Geophys.*, **40**(1), 1001, doi:10.1029/2000RG00092.
- Grivet-Talocia, S., F. Einaudi, W. L. Clark, R. D. Dennett, G. D. Nastrom, and T. E. VanZandt, 1999: A 4-yr climatology of pressure disturbances using a barometer network in central Illinois. *Mon. Wea. Review*, **127**, 1613–1629.
- Henrici, P., 1986: *Applied and Computational Complex Analysis. Vol. 3: Discrete Fourier Analysis-Cauchy Integrals-Construction of Conformal Maps-Univalent Functions*. John Wiley & Sons, 637 pp.
- Herrmann, E., 1996: On the convolution type kernel regression estimator. Preprints, Dept. Math., Tech. Univ. Darmstadt, 31 pp.
<http://wwwbib.mathematik.tu-darmstadt.de/Math-Net/Preprints/Listen/shadow/pp1833.html>
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998a: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, W., Z. Shen, N. E. Huang, and Y. C. Fung, 1998b: Use of intrinsic modes in biology: Examples of indicial response of pulmonary blood pressure to \pm step hypoxia. *Proc. Natl. Acad. Sci. USA*, **95**, 12766–12771.
- Huang, N. E., C. C. Chern, K. Huang, L. W. Salvino, S. R. Long, and K. L. Fan, 2001: A new spectral representation of earthquake data: Hilbert spectral analysis of Station TCU129, Chi-Chi, Taiwan, 21 September 1999. *Bull. Seism. Soc. Am.*, **91**, 1310–1338.
- Koch, D. M., and M. E. Mann, 1996: Spatial and temporal variability of ^{7}Be surface concentrations. *Tellus, Ser. B*, **48**, 387–396.
- Komm, R. W., F. Hill, and R. Howe, 2001: Empirical mode decomposition and Hilbert analysis applied to rotation residuals of the solar convection zone. *Astrophys. J.*, **558**, 428–441.
- Pan, J.-Y., X.-H. Yan, Q.-N. Zheng, W. T. Liu, and V. V. Klemas, 2003: Interpretation of scatterometer ocean surface wind vector EOFs over the Northwestern Pacific. *Remote Sens. Environ.*, **84**, 53–68.
- Ryan, T. P., 1997: *Modern Regression Methods*. John Wiley & Sons, 515 pp.
- Salisbury, J. I., and M. Wimbush, 2002: Using modern time series analysis

- techniques to predict ENSO events from the SOI time series. *Nonlinear Process. Geophys.*, **9**, 341–345.
- Shapiro, R., 1970: Smoothing, filtering, and boundary effects. *Rev. Geophys. Space Phys.*, **8**, 359–387.
- Susanto, R. D., A. L. Gordon, J. Sprintall, and B. Herunadi, 2000: Intraseasonal variability and tides in Makassar Strait. *Geophys. Res. Lett.*, **27**, 1499–1502.
- Torrence, C., and G. P. Compo, 1998: A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.*, **79**, 61–78.
- Wu, M.-L. C., S. Schubert, and N. E. Huang, 1999: The development of the South Asian summer monsoon and the intraseasonal oscillation. *J. Climate*, **12**, 2054–2075.
- Xie, L., L. J. Pietrafesa, and K. J. Wu, 2002: Interannual and decadal variability of landfalling tropical cyclones in the southeast coastal states of the United States. *Adv. Atmos. Sci.*, **19**, 677–686.

*Dean G. Duffy
Code 613.1, NASA/Goddard Space Flight Center, Greenbelt, MD 20771,
USA
dean.g.duffy@nasa.gov*

CHAPTER 7

EMPIRICAL MODE DECOMPOSITION AND CLIMATE VARIABILITY

Katie Coughlin and Ka Kit Tung

Empirical mode decomposition (EMD) is used to extract inter-annual atmospheric signals. Our climate contains components that are often distinguished from one another on temporal, rather than spatial, scales, and EMD is found to be a useful tool in extracting physically meaningful information from the data. The National Centers for Environmental Prediction has provided reanalyzed data, which are used here to demonstrate that dynamical variables in the zonally averaged troposphere and lower stratosphere contain only five oscillating modes and a trend. These oscillations contain interesting amplitude and frequency modulations which are cumbersome and difficult to interpret when viewed in the traditional Fourier frequency domain. EMD analysis provides a unique, relatively unbiased and useful way to decompose the time series. Statistical tests are developed to determine the confidence in our analysis. Physical interpretations of the significant modes are also discussed with corresponding justification. By combining the information found for the time series of each horizontal strip into a spatial matrix, valuable spatial information about these physical processes have been obtained.

7.1. Introduction

Empirical mode decomposition (EMD) can be a useful time series analysis tool, particularly for analyzing the climate records of the atmosphere beyond annual time scales. Global climate phenomena are often separated in temporal, rather than spatial, scales. Therefore, time series analysis is more appropriate for the initial decomposition of climate data than spatial methods, which have previously been used to find climate components. Furthermore, nonlinear trends, irregular frequencies and amplitude modulations, which are inherently present in these physical phenomena, make EMD especially appropriate for this problem.

Understanding the variability of our climate is important. Atmospheric variability on inter-annual time scales is investigated here with a focus on

observations in the regions which most influences us; the lower stratosphere ($\sim 12 - 30$ km) and the troposphere (0 – 12 km). The main difficulty is disentangling the atmospheric signals in these regions. In both the stratosphere and the troposphere, inter-annual signals are mixed with one another as well as with noise. Isolating the relevant signals is a daunting task. Traditional time-series techniques, based on Fourier transforms, tend to spread the signal energy into multiple frequency bins which sometimes leads to ambiguous interpretations of the data. Other methods, such as empirical orthogonal function (EOF) analysis, depend on spatial decompositions to filter the noisy data. However, as mentioned above, the signals of interest are mostly global in extent and distinguishable from one another temporally, not by their spatial scales. EOF analysis leads to a leading mode which contains climate signals of various time scales projected — unphysically — onto a spatial pattern which contains the most variance. Long period climate oscillations, which usually have a smaller variance than other oscillations, are either grossly distorted or missing altogether in the leading EOFs. There is no real reason why a spatial decomposition should isolate temporal signals. Here, we use the empirical mode decomposition (EMD) method, Huang et al. (1998) designed for time series analysis. The modes generated by this method are derived directly from the data. Here, we show that the atmospheric signals in the stratosphere and troposphere can be completely decomposed into five of these oscillating modes: an annual cycle, a Quasi-Biennial Oscillation (QBO) signal, an El Niño/Southern Oscillation (ENSO)-like mode and a solar cycle mode. The last mode is a trend which indicates cooling in the stratosphere and warming in the troposphere, as expected from the recent increase in greenhouse gases. Statistical tests, used to determine the significance of these EMD modes, will also be discussed here.

7.2. Data

Atmospheric measurements are often made from balloons and aircraft with reference to pressure levels instead of height, since the pressure can be measured more accurately in situ. The analysis shown here is plotted on pressure surfaces, measured in hecto-Pascals (hPa).^{**} The pressure decreases as the distance from the surface increases and provides a useful axis for looking at atmospheric dynamics.

^{**}A hPa is equivalent to a mbar where pressure at the surface of the earth is ~ 1 bar = 1000 mb = 1000 hPa.

The monthly averages of the NCEP Daily Global Analyses data are provided by the NOAA-CIRES Climate Diagnostics Center in Boulder, Colorado, USA, available on line at <http://www.cdc.noaa.gov/> (Kalnay et al. 1996). These averages are used here to create a time series of geopotential height from January 1958 to July 2002. A spatial average of the total geopotential height from 20°N to 90°N is performed at 17 levels, from 10 hPa down to 1000 hPa. This averaged Northern Hemisphere time series is decomposed by using the EMD method, and the results are analyzed here. To investigate the latitudinal dependence, zonally averaged latitudinal strips of the total geopotential height are decomposed and analyzed as well. A typical tropospheric (stratospheric) decomposition of the total geopotential height at 700 hPa (30 hPa), averaged from 20°N to 90°N, is shown in Fig. 7.1. The results for all 17 levels show similar decompositions of the five modes, each with a variable amplitude and period, and a trend.

The sunspot record is provided by the SIDC, RWC Belgium, World Data Center for the Sunspot Index, Royal Observatory of Belgium, available on line at <http://sidc.oma.be/>. This record consists of monthly data from January 1749 to September 2002. Only the data from 1958 to 2002 are utilized here.

The data used to characterize the QBO winds are from personal email communications with Barbara Naujokat. As one of the compilers of the daily stratospheric charts at the Stratospheric Research Group of the Free University Berlin, she tabulated the daily wind observations of selected stations near the equator since 1957. For the earlier years, the values were extracted from the Northern Hemisphere Data Tabulations. From these daily values, the monthly mean zonal wind components were calculated for the levels 70, 50, 40, 30, 20, 15, and 10 hPa, and a dataset from 1953 to the present was produced by combining the observations from the three radiosonde stations; Canton Island (closed 1967), Gan/Maledive Islands (closed 1975), and Singapore. The negative values are easterly (from the east), and the positive values are westerly. This dataset is supposed to represent the equatorial belt since all studies have shown that the longitudinal differences in the phase of the QBO are small. However, some uncertainties arose at higher levels during the early years because of the scarcity of observations. More information on the data can be found in Naujokat (1986).

The Multivariate ENSO Index (MEI) bimonthly time series from Dec/Jan 1949/1950 up to Nov/Dec 2003 is used as a robust measure of the ENSO. More information on the MEI can be found on the MEI homepage, available online at www.cdc.noaa.gov/~kew/MEI/mei.html, and in

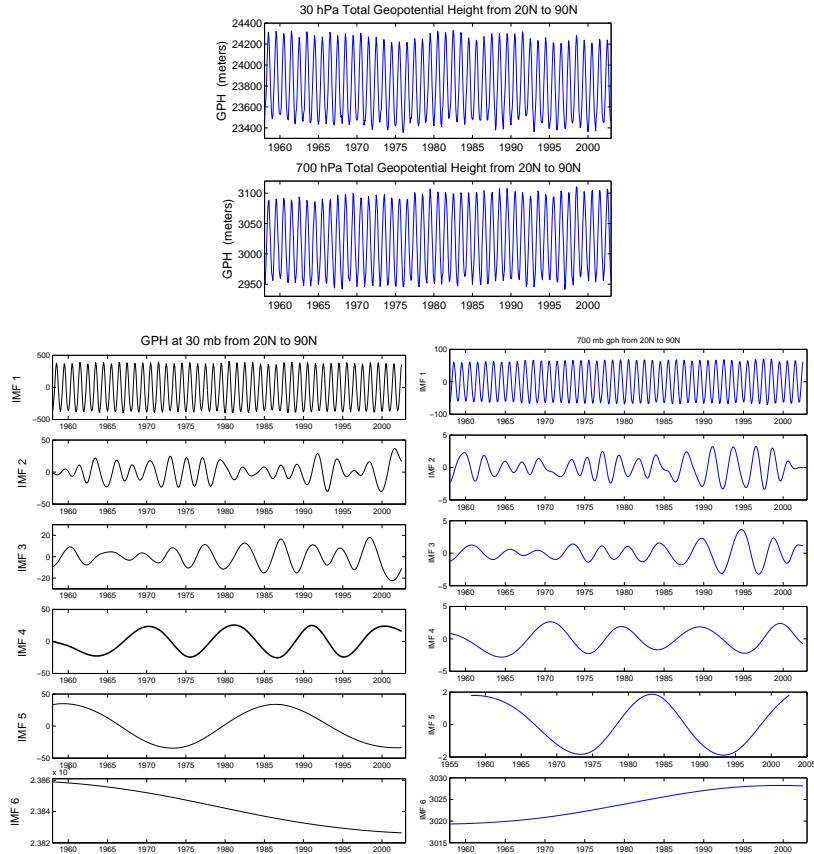


Figure 7.1: *Top.* The total geopotential height at 30 hPa and 700 hPa spatially averaged over 20°N to 90°N . *Bottom.* The decomposition of the 30 hP (left) and 700 hPa (right) geopotential height produces five modes and a trend. The first mode is the annual cycle. The second mode is the extratropical QBO, with an average period of 28 months. The third ENSO-like mode has an average period of around four years, and the fourth mode is highly correlated with the 11-yr sunspot cycle. We refrain from commenting on the 22-yr mode found since the data record contains only two periods of this oscillation. The trend in recent decades indicates cooling in the troposphere and warming in the stratosphere. This finding is consistent with the anticipated effect of the increase in greenhouse gases. Figure 7.1 was taken from Coughlin, K. T., and K. K. Tung, 2004: 11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.*, **34**, 323–329 with permission.

Wolter (1987) and Wolter and Timlin (1993,1998).

7.3. Methodology

Climate signals in the atmosphere are often lost among the noise and data-collection and assimilation problems. Although climate signals are recognized mainly by their time scales, previous analysis, such as the use of empirical orthogonal functions (EOFs), relies on spatial decompositions to remove excess noise. In some cases, this method is useful, but in general, there is no *a priori* reason to expect temporal signals to be defined in terms of spatial patterns ordered by their variances (as required by EOF analysis). In fact, time series analysis is more appropriate for the decomposition of climate signals. The traditional time series analysis tools usually rely on Fourier transforms in one way or another. However, Fourier transforms lead to inconclusive interpretations due mainly to the global nature (in the time domain) of the transforms. Even wavelet analysis, developed to deal with non-stationarity and local frequency changes, produces confusing and sometimes contradictory results when applied to climate signals (Mak 1995; Oh et al. 2003; Sonechkin and Datsenko 2000). These three papers all use similar types of wavelet analysis on global, surface climate data and obtain three different results. With wavelet analysis, it is sometimes difficult to distinctly define local frequency changes because the spectra are created by stepping through various predetermined frequencies, often producing a blurred result. Wavelets have the additional problem of shift variance. If the starting point is varied by dropping the initial point, for example, wavelet analysis can produce completely different results. In comparison, the EMD method makes no assumption about linearity or stationarity, and the intrinsic mode functions (IMFs) are usually easy to interpret and relevant to the physical system being studied (see Flandrin et al. 2004; Huang et al. 1998, 1999; Zhu et al. 1997 for further discussion of the method and comparison to other time series analysis techniques).

EMD must be used cautiously even though it is a powerful method. One difficulty encountered when using it is the sensitivity to end point treatments. The envelopes are calculated by using a cubic spline; however, splines are notoriously sensitive to end points. The end effects must not be allowed to propagate into the interior solution. As described in Coughlin and Tung (2004a), this problem is dealt with by extending both the beginning and end of the data by the addition of typical waves,

$$\text{wave extension} = A \sin(2\pi t/P) + \text{phase} + \text{local mean.} \quad (7.1)$$

The typical amplitude A and period P are determined by the nearest local extrema; i.e.,

$$A_{beginning} = \frac{1}{2} \|\max(1) - \min(1)\|, \quad (7.2)$$

$$A_{end} = \frac{1}{2} \|\max(N) - \min(N)\|, \quad (7.3)$$

$$P_{beginning} = 2 \|\text{time}(\max(1)) - \text{time}(\min(1))\|, \quad (7.4)$$

$$P_{end} = 2 \|\text{time}(\max(N)) - \text{time}(\min(N))\|, \quad (7.5)$$

where $\max(1)$ and $\min(1)$ are the first two local extrema in the time series, and $\max(N)$ and $\min(N)$ are the last two local extrema. This calculation takes place every iteration so that the additional waves are continually changing in amplitude and frequency. Because the additional waves have the same amplitude as the nearest oscillations, the addition of these waves causes the slope of the envelope to tend toward zero at the beginning and end of the time series. This technique eliminates large swings in the spline calculation that may otherwise form when the slope is artificially forced to zero. Two or three oscillations of approximately the same amplitude are needed in order to flatten the spline. Any more than this can adversely affect the low frequency modes by artificially leveling the ends of any long-term trend present in the data.

The stopping criterion is another issue that can affect the decomposition. In order to allow more natural flexibility in the modes, we stop the sifting process and define a mode when the number of extrema is equal to or one different from the number of zero crossings (our criterion is similar to the one suggested in Huang et al. 1999). We also require this condition to be consistent for three iterations in a row in order to constrain the local means to be close to zero. If the Hilbert transform is to be used on the subsequent modes, a stricter condition on the local mean may be useful since the Hilbert transform is very sensitive to changes in this mean. In our analysis, however, we compare the modes directly to physically relevant oscillations and are less concerned with a precise zero mean.

Another important component of the analysis, as described in Coughlin and Tung (2004a), is the inclusion of the annual cycle. Unlike more traditional methods, where a twelve-month climatology is subtracted from the data and only the anomalies are studied, this method decomposes the total atmospheric signal, not just the anomalies. In fact, the removal of climatology, which is a linear operation, can actually degrade the nonlinear analysis.

Here, the annual cycle is an important component and is retrieved as the first IMF in the decomposition. In some cases, a three-month running average can be applied to the time series to damp month-to-month variations and retrieve the annual cycle as our first mode. These higher month-to-month frequencies are either intermittent or appear to be intermittent. In either case, the resulting modes contain a mixture of frequencies, and these mixed modes are much more difficult to interpret than other modes. Since we are not interested in intraseasonal variations, a minimal amount of pre-smoothing is performed so that the first mode will contain the annual cycle. Since the annual cycle is relatively small near the equator, where the seasons are less distinct, we expected our results to be somewhat degraded in this region due to the pre-smoothing process that takes place there, but surprisingly, we find that the equatorial signals remain relatively robust in most cases.

The IMFs will generally be ordered from high to low frequency, and the last IMF often contains a trend. The significance of the IMFs must be determined. We should not expect all modes to be significant.

7.4. Statistical tests of confidence

The significance of the EMD modes must be calculated. Our confidence in these modes can be tested by comparing their energy spectra to the energy spectrum of decomposed noise. Our statistics are developed specifically for this climate problem (see Coughlin and Tung 2004a).

To distinguish the signals from noise in the EMD method is used, we examine the energy of all the IMFs and compare the energy in each mode to the energy distribution of red noise. Typically, we may expect the highest frequency mode to contain only noise, so the power in this mode can then be used to calibrate the noise distribution. However, in this case, the first mode contains a climatological annual cycle, which obviously does not contain pure noise. To estimate the power of the noise present in this time series, we subtract the climatology from the first mode and assume that the remainder is noise. This estimate of the noise is still very conservative. However, by using this criteria, we can normalize the red noise spectrum and perform a Monte-Carlo simulation to test the significance of the remaining IMFs. In atmospheric phenomena, the data at one time step tend to be related to the data at the previous time step. For example, the temperature on a day (or month) will be similar to the temperature on the day (or month) before. This similarity is called “persistence,” and this “redness” or

autocorrelation should be accounted for when creating Monte-Carlo simulations to test the significance of the decomposition. In our simulations Coughlin and Tung (2004a,b), we create 500 random “red” time series with the same autocorrelation as that of the difference between the data and the climatology. The variance of these time series is calculated based on the “noise” in the first mode, which is found by subtracting the climatology from the first mode, and then the original climatology is added back into each of these time series before using EMD to define their modes. In this way, the first mode contains the annual cycle, and the power of the original remaining modes can then be compared to the power in the modes created from the red noise. The red noise, or random time series, is

$$\text{rts}(t_n) = A_{\text{noise}} E(t_n) + \rho \text{rts}(t_{n-1}), \quad (7.6)$$

$$A_{\text{noise}} = \text{standard deviation (IMF1 - climatology)}, \quad (7.7)$$

and

$$\rho = \text{autocorrelation of (total geopotential height - climatology)}, \quad (7.8)$$

where E contains a uniform distribution of random numbers with values between 1 and -1, t_n is the n th time step, and rts is the random time series generated. The climatology is the geopotential height averaged over all years of the data record for each individual month and is added to the random time series before applying the EMD method.

EMD modes are generated by using these time series, and then the statistics of this simulation are calculated to create the energy distribution of the noise. The power of each IMF is then compared to this distribution to determine its significance.

In Fig. 7.2, the average power per period of modes 2, 3 and 4 in the 30-hPa and 700-hPa geopotential height is compared to that of modes 2, 3 and 4 of 500 randomly generated time series. These series are created by adding the 30-hPa, or 700-hPa, climatology to appropriately scaled red noise. For randomly generated time series, the EMD method exhibits a period doubling phenomenon so that each mode tends to have an average period of about twice that of the previous mode (as described in Flandrin et al. 2004). This result creates a clustering of the modes about certain periods in the Monte-Carlo plot (Fig. 7.2). The average power of the random modes varies inversely to that of the average period (power $\sim 1/\text{period}$) so that the modes quickly decrease in power as the average period increases. Here, the power per period is plotted, allowing one to directly compare each mode,

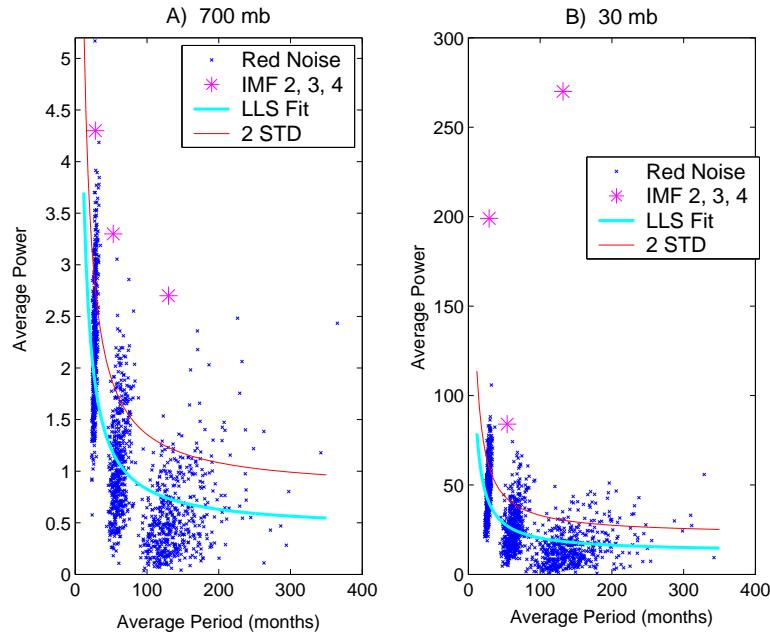


Figure 7.2: Monte-Carlo t-tests showing the significance of EMD modes. (a) 700 hPa climatology is added to 500 autocorrelated time series to generate a Monte-Carlo test. The red noise is calibrated by using the difference between the first mode and the climatology (Coughlin and Tung 2004a), and the autocorrelation of the noise is the same as the autocorrelation of the 700 hPa geopotential height anomalies, 0.54. These time series are decomposed by using the EMD method, and the average power in modes 2, 3 and 4 of each of the decompositions is plotted in this figure. The solid blue line represents the linear least squares fit of these points, and the red line is two standard deviations from the best-fit line. The average power of IMFs 2, 3 and 4 of the 700 hPa geopotential heights is plotted as stars. (b) Same as (a) except for the 30 hPa decomposition. Note that the 11-yr mode (IMF 4) is clearly above all random noise. Figure 7.2 was taken from Coughlin, K. T., and K. K. Tung, 2004: 11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.*, **34**, 323–329 with permission.

so the points actually decrease in inverse proportion to the square of the period. The thick line is the linear least squares fit of the noise, and the dashed line is one standard deviation above this best fit line. The three stars represent the average power per period of IMF 2, IMF 3 and IMF 4 of the 30-hPa, or 700 hPa, geopotential height at average periods of 28, 59 and

132 months, respectively. These modes all fall above the confidence interval and therefore are significant. This finding indicates that modes 2, 3 and 4 are real signals, different from random red noise. They make significant contributions to the observed geopotential height.

Once the significance of the mode is established, its physical relevance can be determined. Here, we use a student t-test (Hogg and Tanis 1997),

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (7.9)$$

to calculate the significance of correlations between the modes and physically relevant indices. The student t value is t , r is the correlation and $\nu = N - 2$ is the degree of freedom. Special care must be taken in estimating the degrees of freedom. Because the modes are non-stationary, typical techniques may not be applicable for finding the degree of freedom. Here, we use the following argument: In the EMD method, each IMF is successively subtracted from the original time series, leaving behind only variations with local timescales longer than those in the subtracted modes. We assume, then, that the data points in the residual time series contain timescales equivalent to or greater than the periods of the last subtracted mode. For example, in the decomposition of the 30-hPa geopotential height, the third IMF has an average period of 4.4 yr. This finding implies that the fourth IMF is composed of data which are dependent on timescales of about 4.4 yr. Conservatively, then, we estimate that the fourth IMF has independent time intervals of 5 yr. By using this estimate, nine independent measurements of the time series can be made for the 4th IMF at 30-hPa, and, therefore, $\nu = 7$ degrees of freedom for the correlation calculation. For this example, correlations with the solar flux are calculated by comparing the data every 60 months (see Fig. 7.3). These correlations average to 0.70. For comparison, the correlation coefficient, calculated by using every month in the unsmoothed sunspot numbers, is 0.72. For this mode, the correlation is very robust. If the student t-test is used, the 0.70 correlation is statistically different from the null hypothesis of a zero correlation. Although this correlation is significant at a higher level, all of our significance testing is done at a 95% confidence level.

This statistical relationships can also be verified by using a regression with auto-regressive (AR) errors. For the above example, we assume that the IMF contains the solar cycle plus an autoregressive error,

$$\text{IMF4}(t) = b[\text{SI}(t) - \mu] + e(t). \quad (7.10)$$

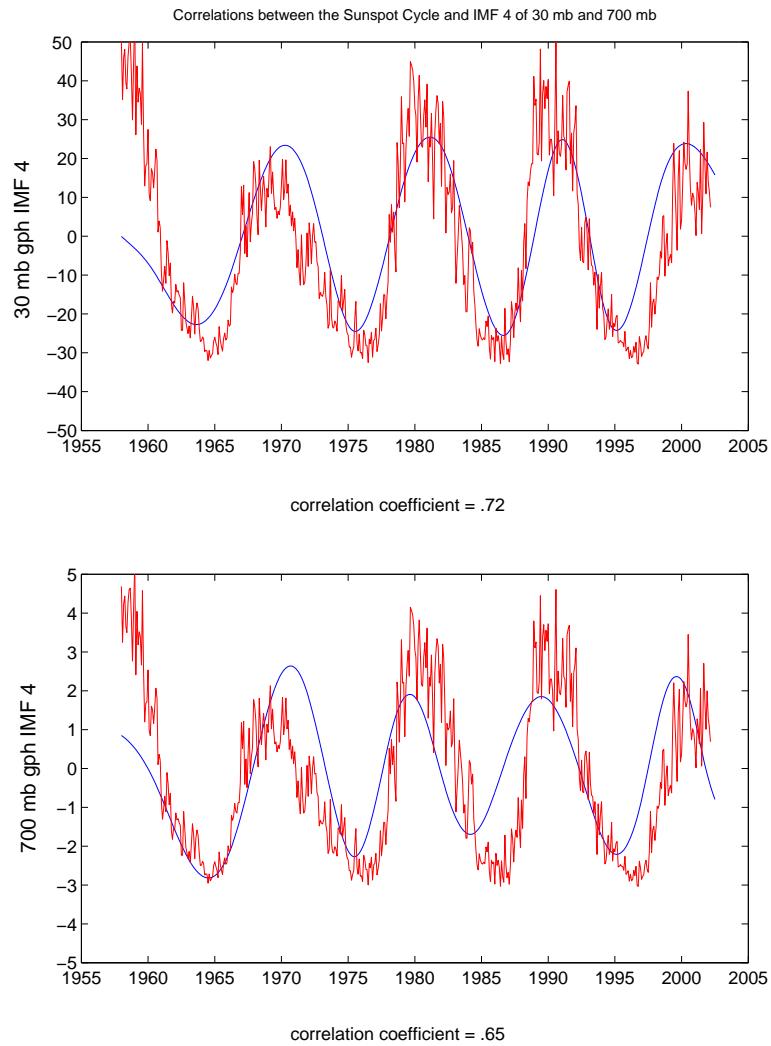


Figure 7.3: Significant correlation is 0.70 for 30 mb and 0.58 for 700 mb. Significance assumes that the mode has only 7 degrees of freedom. Figure 7.3 was taken from Coughlin, K. T., and K. K. Tung, 2004: 11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.*, **34**, 323–329 with permission.

Here, SI is a solar cycle index defined to be the sunspot number divided by 1000. The parameters, b and μ and their variances can be found by fitting the error $e(t)$ to an AR(p) process for some order p . More information about

this type of method can be found in Newton (1988). For this mode, the parameter values do not change significantly for different orders; however, we find that the best fit for the errors is an autoregressive process of order 5. This order gives the parameter values of $\mu = 0.082 \pm 0.013$ and $b = 42 \pm 7$ as the best fit for this model. Since $b = 42$ is about 6 times its standard error, we can effectively rule out the hypothesis $b = 0$. This conclusion implies that the solar cycle is directly related to the fourth IMF. The regression is highly significant, and the sunspot cycle statistically explains the majority of the fourth mode. A similar type of analysis is performed for the other modes.

7.5. Results and physical interpretations

Figure 7.1 shows the complete decomposition of the 30-hPa and 700-hPa geopotential height spatially averaged over all longitudes and latitudes from 20°N to 90°N. Although the EMD modes are empirically determined, they remain locally orthogonal to one another. The time series is separated into five modes and a trend. The first IMF contains the annual cycle. The second IMF has an average period of 28 months and is anti-correlated with the equatorial QBO. The third ENSO-like mode has an average period of four years. The fourth mode, with an average period of 11 yr, is highly correlated with the solar cycle, and the trend indicates cooling in the stratosphere over time. These trends are consistent with theories of stratospheric cooling due to increases in greenhouse gases. Amazingly, similar decompositions are found at almost all heights and latitudes.

7.5.1. Annual cycle

Although most analysis of atmospheric data begins by subtracting a climatological mean from the time series, the winter-to-summer differences are retained in this analysis. As a result, the very first mode is always a large annual cycle. Because the EMD method is non-stationary and nonlinear, we can look at the seasonal variations. Figure 7.4 shows that the annual cycle is not strictly regular in frequency or amplitude.

7.5.2. Quasi-Biennial Oscillation (QBO)

The second mode has an average period of 28 months and, in the extratropics, is anticorrelated with the equatorial QBO. Its see-saw pattern of geopotential height (see Holton and Tan 1980) with a node at 50°N is apparent in

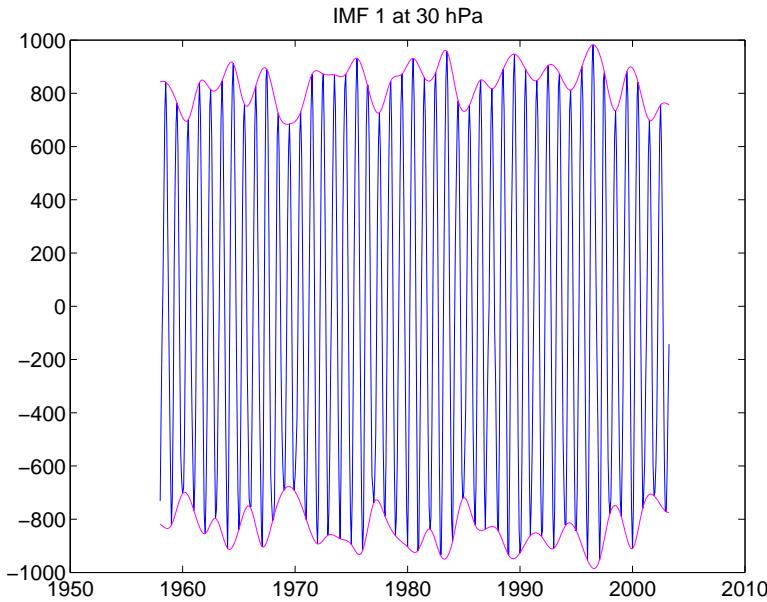


Figure 7.4: Amplitude modulations of the first mode. At 30 hPa the first IMF of the total geopotential height at the pole, from 65°N to 90°N .

Fig. 7.5. The extra-tropical manifestation of the QBO in the stratosphere is relatively well understood: During winter, the polar geopotential height is anticorrelated with the equatorial winds due to the Holton-Tan mechanism (Holton and Tan 1980, 1982; Tung and Yang 1994) and to the asymmetrical QBO direct circulation in the lower stratosphere (Jones et al. 1998; Kinnersley 1998; Kinnersley and Tung 1999). The tropospheric mechanism is not yet understood but has been observed in previous work (Coughlin and Tung 2001; Thompson et al. 2002). Here, the correlations in the troposphere are small, probably due to the short data record and the possible phase shifts associated with the influence of the equatorial QBO at various levels besides 30 hPa.

7.5.3. ENSO-like mode

The third ENSO-like mode has an interesting latitudinal and height structure. It is positively correlated with the Multivariate ENSO index in the tropical troposphere and anticorrelated in the extratropical troposphere, between 40°N and 60°N . The signal also seems to have influence higher up

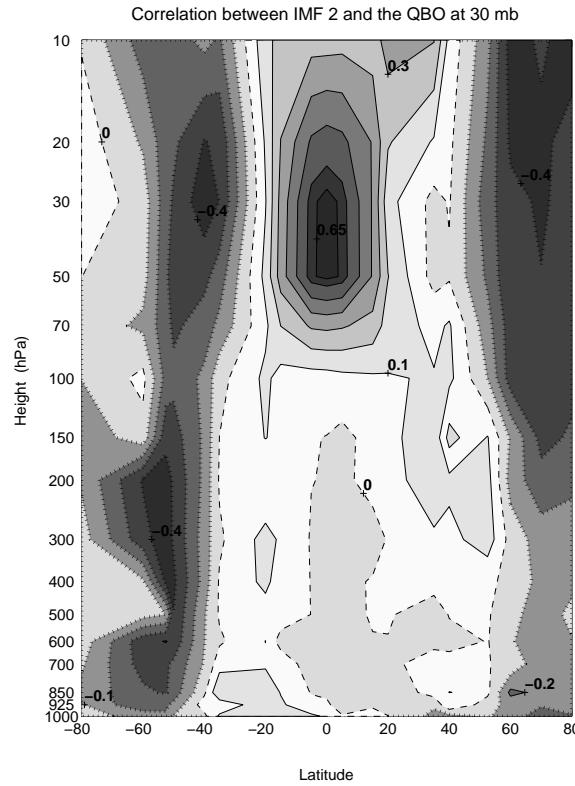


Figure 7.5: A correlation map of the second mode and the equatorial QBO. A map of correlation coefficients between the second IMF at each height latitudinal strips of 20° and the QBO Index. The coefficients are plotted at the height and mean latitude of each decomposition. The horizontal axis ranges from 40°S to 80°N .

into the atmosphere, both in the tropical and extratropical regions, but the correlations there are not statistically significant because a zonal average is taken here, which cancels out most of the positive and negative signals in the Pacific basin. Wu and Huang (2004) have also used EMD to analyze an ENSO signal. In general, the influence of this signal deserves further analysis.

7.5.4. Solar cycle signal in the stratosphere

The fourth IMF has the same frequency as that of the 11-yr solar cycle and also contains amplitudes comparable to those previous estimates of

geopotential height solar cycle variations. More in-depth discussion of this mode can be found in Coughlin and Tung (2004a,b). In our analysis, the peak-to-peak amplitude variation of the solar cycle signal is about 50 m for the geopotential height averaged over the Northern Hemisphere at 30 hPa. In Labitzke's (2001) paper, a 3-yr running mean of geopotential height at 30 hPa and at a selected location, 3°N/150°W, from 1958 to 1997 has peak-to-peak amplitude variations ranging from 40 to 90 meters.

Although many reports have been made of the 11-yr solar cycle in atmospheric data, a considerable debate is taking place about the spatial and temporal extent to which the atmosphere is influenced by the 11-yr variations in the solar irradiance and about the validity of the statistical significance of these claims. A major problem is the shortness of the data record, which prevents a straightforward extraction of an 11-yr signal in the energy spectrum of dynamical variables.

A small amplitude solar cycle signal has, nevertheless, been found in the quiescent regions of the dataset by previous authors. In the extratropical middle atmosphere, this signal has been found during summers (Labitzke 1987; Labitzke and van Loon 1988, 1989; Kodera 1991; Dunkerton and Baldwin 1992; van Loon and Labitzke 1994) — away from the time periods of most variability — over the mid-latitude Pacific (Gray and Ruth 1993; Haigh 2002; Labitzke 2001; Labitzke and van Loon 1992, 1994, 1999) — away from the spatial areas of the most variability — and during the westerly phases of the QBO (Labitzke 1987; Labitzke and van Loon 1988, 1994; Gary and Ruth 1993, Ruzmaikin and Feynman 2002) — away from the periods of the most dynamical disturbance.

During these quiescent periods of time and in quiescent regions of space, the stratosphere is observed to be positively correlated with the solar cycle. What we find (see Coughlin and Tung 2004a,b) is a solar cycle signal in a more general and global sense, without resorting to any specific grouping of the data, either by season, area or the phase of the QBO. This consideration is important in the test for the statistical significance of the signal, as the stratospheric record is short with less than five cycles of the solar signal. The spatial structure of the fourth mode is shown in Fig. 7.6. From these correlations, we see that a component of the climate is always in sync with the solar cycle. In the region between 50°S and 50°N, the coherence between the atmospheric signal and the solar flux is robust. These solar signals are constant in amplitude and have very little latitudinal or vertical variation. At the poles, the signal is less robust, perhaps because of poor data quality at the poles, or because of a real difference in this region. We reserve that

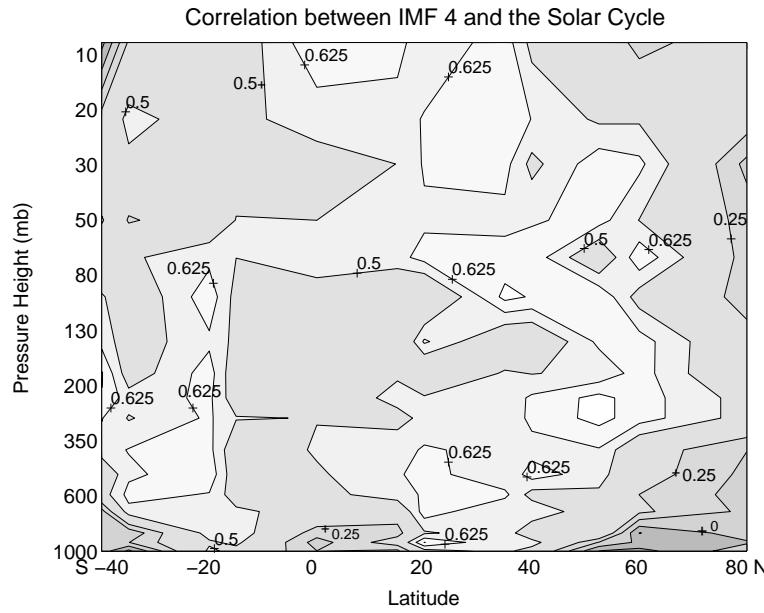


Figure 7.6: A correlation map of the fourth mode and the solar flux. A map of correlation coefficients between the fourth mode of each latitudinal strip and the solar flux. The coefficients are plotted at the pressure altitude and mean latitude of each decomposition. The horizontal axis ranges from 40°S to 80°N . Figure 7.6 was taken from Coughlin, K. T., and K. K. Tung, 2004: 11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.*, **34**, 323–329 with permission.

analysis for future investigation.

The correlation map in Fig. 7.6 shows statistically how closely these signals are related to the solar flux. The solar flux is positively correlated with the fourth mode almost everywhere in our domain (see Fig. 7.6). In contrast to mode 2 (Fig. 7.5) and mode 3, which have interesting spatial structures, mode 4 is comparatively uniform. The only minor exception is in the south polar region, and the correlations there are not statistically significant.

7.5.5. Fifth mode

The fifth IMF contains only one or two oscillations, and little is known about it. Some researchers have speculated that it may be related to a double solar period or to the Pacific Decadal Oscillation (PDO), but cor-

relations show that it probably is not. Moreover, although it is correlated with the double solar cycle, it is not statistically significant because we only have two such periods. Another theory is that this mode may actually be physically associated with the trend. In that case, it might represent further nonlinearities in the anthropogenic influences on the atmosphere. For now, however, we reserve further comment until these possibilities can be distinguished from one another.

7.5.6. Trends

The trends, as seen in Fig. 7.1, are very robust in that the stratosphere exhibits cooling, and the troposphere has experienced warming over the last few decades. This finding is consistent with the anticipated effect of recent increases in greenhouse gases.

7.6. Conclusions

Through the EMD representation, we are able to see the natural variations in the atmosphere. This method is both compact and complete. We are able to completely describe the changes in the lower atmosphere with only 5 modes and a trend. Each of these modes is empirically determined and not artificially constrained to have fixed amplitudes or frequencies. Furthermore, the statistically significant modes can be identified with known physical phenomena. Here, this ability allows us to gain a straightforward view of our climate, its components and the temporal/spatial distribution of these physical phenomena. The signals occur almost everywhere in the lower atmosphere (Figs. 7.5 and 7.6). We find that the significant inter-annual atmospheric modes include the QBO mode, the ENSO-like mode, the solar cycle, and climate trends that are consistent with the increases in greenhouse gases.

Acknowledgments

We thank Dr. Zhaohua Wu for first introducing us to the EMD method, and we thank Dr. Dean Duffy for his meticulous editing and typesetting. This research was supported by the National Science Foundation, Climate Dynamics Program, under grant ATM-0332364.

References

- Coughlin, K. T., and K. K. Tung, 2001: QBO signal found at the extratropical surface through northern annular modes. *Geophys. Res. Lett.*,

- 28, 4563–4566.
- Coughlin, K. T., and K. K. Tung, 2004a: 11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.*, **34**, 323–329.
- Coughlin, K., and K. K. Tung, 2004b: Eleven-year solar cycle signal throughout the lower atmosphere. *J. Geophys. Res.*, **109**, D21105, doi:10.1029/2004JD004873.
- Dunkerton, T. J., and M. P. Baldwin, 1992: Modes of interannual variability in the stratosphere. *Geophys. Res. Lett.*, **19**, 49–52.
- Flandrin, P., G. Rilling, and P. Gonçalvès, 2004: Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.*, **11**, 112–114.
- Gray, L. J., and S. Ruth, 1993: The modeled latitudinal distribution of the ozone quasi-biennial oscillation using observed equatorial winds. *J. Atmos. Sci.*, **50**, 1033–1046.
- Haigh, J. D., 2002: The effects of solar variability on the Earth's climate. *Philos. Trans. R. Soc. London, Ser. A*, **361**, 95–111.
- Hogg, R. V., and E. A. Tanis, 1997: *Probability and Statistical Inference*. Prentice Hall, 722 pp.
- Holton, J. R., and H. C. Tan, 1980: The influence of the equatorial quasi-biennial oscillation on the global circulation at 50 mb. *J. Atmos. Sci.*, **37**, 2200–2208.
- Holton, J. R., and H. C. Tan, 1982: The quasi-biennial oscillation in the Northern Hemisphere lower stratosphere. *J. Meteor. Soc. Japan*, **60**, 140–148.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-steady time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of water waves – The Hilbert spectrum. *Ann. Rev. Fluid Mech.*, **31**, 417–457.
- Jones, D. B. A., H. R. Schneider, and M. B. McElroy, 1998: Effects of the quasi-biennial oscillation on the zonally averaged transport of tracers. *J. Geophys. Res.*, **103**, 11235–11249.
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471.
- Kinnersley, J. S., and K. K. Tung, 1999: Mechanisms for the extratropical QBO in circulation and ozone. *J. Atmos. Sci.*, **56**, 1942–1962.
- Kodera, K., 1991: The solar and equatorial QBO influences on the stratospheric circulation during the early Northern Hemisphere winter. *Geo-*

- phys. Res. Lett.*, **18**, 1023–1026.
- Labitzke, K., 1987: Sunspots, the QBO, and the stratospheric temperature in the north polar region. *Geophys. Res. Lett.*, **14**, 535–537.
- Labitzke, K., and H. van Loon, 1988: Associations between the 11-year solar cycle, the QBO and the atmosphere: I. The troposphere and stratosphere in the Northern Hemisphere winter. *J. Atmos. Terr. Phys.*, **50**, 197–206.
- Labitzke, K., and H. van Loon, 1989: The 11-year solar cycle in the stratosphere in the northern summer. *Ann. Geophys.*, **7**, 595–597.
- Labitzke, K., and H. van Loon, 1992: On the association between the QBO and the extratropical stratosphere. *J. Atmos. Terr. Phys.*, **54**, 1453–1463.
- Labitzke, K., and H. van Loon, 1994: Trends of temperature and geopotential height between 100 and 10 hPa on the Northern Hemisphere. *J. Meteor. Soc. Japan*, **72**, 643–652.
- Labitzke, K., and H. van Loon, 1999: *The Stratosphere: Phenomena, History, and Relevance*. Springer, 179 pp.
- Labitzke, K., 2001: The global signal of the 11-year sunspot cycle in the stratosphere: Differences between solar maxima and minima. *Meteor. Z.*, **10**, 83–90.
- Mak, M., 1995: Orthogonal wavelet analysis: Interannual variability in the sea surface temperature. *Bull. Amer. Meteor. Soc.*, **76**, 2179–2186.
- Newton, H. J., 1988: *TIMESLAB: A Time Series Analysis Laboratory*. Wadsworth and Brooks/Cole Publishing Co., 625 pp.
- Oh, H.-S., C. M. Ammann, Ph. Naveau, D. Nychka, and B. L. Otto-Bliesner, 2003: Multi-resolution time series analysis applied to solar irradiance and climate reconstructions. *J. Atmos. Solar-Terr. Phys.*, **65**, 191–201.
- Naujokat, B., 1986: An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics. *J. Atmos. Sci.*, **43**, 1873–1877.
- Ruzmaikin, A., and J. Feynman, 2002: Solar influence on a major mode of atmospheric variability. *J. Geophys. Res. - Atmos.*, **107**, Art. No. 4209.
- Sonechkin, D. M., and N. M. Datsenko, 2000: Wavelet analysis of non-stationary and chaotic time series with an application to the climate change problem. *Pure Appl. Geophys.*, **157**, 653–677.
- Thompson, D. W. J., M. P. Baldwin, and J. M. Wallace, 2002: Stratospheric connection to Northern Hemisphere wintertime weather: Implications for prediction. *J. Climate*, **15**, 1421–1428.
- Tung, K. K., and H. Yang, 1994: Global QBO in circulation and ozone. Part I: Reexamination of observational evidence. *J. Atmos. Sci.*, **51**,

- 2699–2707.
- Van Loon, H., and K. Labitzke, 1994: The 10–12-year atmospheric oscillation. *Meteor. Z.*, **3**, 259–266.
- Wolter, K., 1987: The Southern Oscillation in surface circulation and climate over the Atlantic, Eastern Pacific, and Indian Oceans, as captured by cluster analysis. *J. Climate Appl. Meteor.*, **26**, 540–558.
- Wolter, K., and M. S. Timlin, 1993: Monitoring ENSO in COADS with a seasonally adjusted principal component index. *Proc. 17th Climate Diag. Workshop*, Norman, OK, pp. 52–57.
- Wolter, K., and M. S. Timlin, 1998: Measuring the strength of ENSO event: How does 1997/98 rank? *Weather*, **53**, 315–324.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.
- Zhu, X., Z. Shen, S. D. Eckermann, M. Bittner, I. Hirota, and J.-H. Yee, 1997: Gravity wave characteristics in the middle atmosphere derived from the empirical mode decomposition method. *J. Geophys. Res.*, **102**, 16545–16561.

Katie Coughlin

University of Washington, Department of Applied Mathematics, Box 352420, Seattle, WA 98195, USA
katie@amath.washington.edu

Ka Kit Tung

University of Washington, Department of Applied Mathematics, Box 352420, Seattle, WA 98195, USA
tung@amath.washington.edu

CHAPTER 8

EMD CORRECTION OF ORBITAL DRIFT ARTIFACTS IN SATELLITE DATA STREAM

Jorge E. Pinzón, Molly E. Brown and Compton J. Tucker

Orbital drift results in late equatorial crossing times for the afternoon National Oceanic and Atmospheric Administration (NOAA) polar satellites and in changes of illumination that affect measurements made by the advanced very high resolution radiometer (AVHRR). Processing and correcting for calibration variation(s) and atmospheric effects have improved, but one of the standard AVHRR products, the normalized difference vegetation index (*NDVI*), may still contain variations due to orbital drift or changes in sun-target-sensor geometry. In this study, the solar zenith angle (SZA) trends associated with orbital drift are identified and analyzed with respect to their effects on the *NDVI*. The adaptive empirical mode decomposition (EMD) method is used to identify and remove the induced artifacts from the *NDVI* time series. The EMD is based on the local characteristic time scale of the data and is used to identify embedded nonlinear and nonstationary variation. Trend artifacts associated with drift were uncoupled from the surface signal, and their contributions were quantified at all latitudes. The approach was tested on 1 degree and 8-km *NDVI* global datasets, and showed that it is very suitable for addressing the long-standing issues of orbital shifting or the inconsistencies of the AVHRR data among sensors. The results showed that the interference of satellite drift artifacts with the surface signal was (i) large in a tropical forest, (ii) moderate in the tropics for less-densely vegetated areas, and (iii) lowest at higher northern and lower southern latitudes.

8.1. Introduction

The use of the advanced very high resolution radiometer (AVHRR) data provided by the National Oceanic and Atmospheric Administration (NOAA) polar satellite series has demonstrated the potential of remote sensing for monitoring land-surface variables at a regional to a global scale. This archive of global 4-km AVHRR data was compiled from five different AVHRR instruments on five different NOAA polar-orbiting meteorologi-

cal satellites (Cracknell 1997). Although recent sensors provide improved coarse resolution global land satellite data, the longest AVHRR record, which now extends over more than 23 years, constitutes an invaluable and irreplaceable archive of historical land information (Cracknell 1997, 2001).

Spectral vegetation indices are by far the most used of any of the products derived from the AVHRR instruments (Cracknell 2001, Los 1998, Tucker 1996), a development not anticipated by their designers (Cracknell 2001). These indices are composed of red and near infrared radiances or reflectance, sometimes with additional channels included (Tucker 1979). For the AVHRR instruments, only two bands can be used. Consequently, most of these applications are based on the normalized difference vegetation index (*NDVI*), which is the ratio of the near infrared (NIR) and the red visible (VIS) radiances ,

$$NDVI = \frac{NIR - VIS}{NIR + VIS}. \quad (8.1)$$

This ratio yields a measure of photosynthetic capacity such that the higher the value of the ratio, the more photosynthetically active the cover type (Sellers 1985). Thus, the NDVI product provides the longest existing record of spatial coverage needed for monitoring global vegetation dynamics .

However, collecting a consistent time series record and linking historical observations to current and future measurements is a difficult task. Known inaccuracies for AVHRR arise from orbital drifts,^{††} coarse resolution, lack of on-board calibration capability for the visible and near infrared channels, and their wide spectral bandwidths. These effects increase the probability of sub-pixel clouds that interfere with the surface signal, and increase the variation in the sun-target-view geometries convolved with surface bidirectional reflectance properties of the land surface, topography, and soil background reflectance (Privette et al. 1995). Overall, these inaccuracies result in both gradual and abrupt changes (artifacts) in the data with each successive satellite.

In order to reduce these inaccuracies, vicarious calibration, atmospheric correction and compositing techniques are commonly included in all *NDVI*-processing systems (Cihlar 1994, Los et al. 1994, Los 1998). Among these techniques, the maximum *NDVI* compositing technique has been used in nearly all operational AVHRR processing chains (Cracknell

^{††}The afternoon (PM) platforms drift about 20 to 40 minutes per year at later times of the day with a discontinuity drop back to earlier times of the day when the satellite changes (Price 1991, Los 1998).

2001). This procedure requires the processing of a series of multitemporal georeferenced satellite data into *NDVI* images. On a pixel-by-pixel basis, each *NDVI* value is examined, and only the highest value is retained for each pixel location. Holben (1986) showed that maximum *NDVI* imagery is highly related to green vegetation dynamics, and the aforementioned problems have been minimized. Moreover, Pinzón et al. (2001) showed that of the most commonly used compositing technique, the maximum *NDVI* provides an image with the highest spatial coherence.

The combined techniques used in most *NDVI* processing systems, although effective in reducing atmospheric and sun-target-sensor effects, which tend to cancel each other out (Holben 1986), are still confounded by variations of sun-target-sensor geometry and cloud contamination in some regions (Cihlar et al. 1998, Gutman 1999, Privette et al. 1995). Removing these artifacts from the *NDVI* time series remains a challenge, since an analytical or numerical solution to the system would be prohibitive with current real time processing, given the huge amount of data to process both spatially and temporally. Recent advances in atmospheric correction (El Saleous et al. 2000), bidirectional reflectance distribution functions (BRDF) (Schaaf 2002), and new capabilities for calibration as a result of crossovers with other satellites carrying more recent sensors have opened the way for significant improvements in data processing for the AVHRR sensor (El Saleous et al. 2000). However, this series of improvements has yet to be applied to the full AVHRR record and validated. Meanwhile, the *NDVI* may still contain variations due to orbital drift and changes in sun-target-sensor geometry, since the closure of such a system involves more unknowns than the measurable quantities (Myneni et al. 1993, Privette et al. 1995, Sellers et al. 1994).

Parametric approaches from model simulations to curve fitting have been used in the past, with a heavy set of simplifying assumptions (Gutman 1999, Los et al. 1994, Privette et al. 1995, Sellers et al. 1994). One main assumption is that any spurious trend in the *NDVI* over stable targets, e.g., deserts, can be linearly extrapolated and corrected in all surfaces. However, the complex interaction between vegetation responses, land cover, cloud contamination, bidirectional reflectance properties, and other *NDVI* dependencies prevent this generalization. Here, one can turn to the problem of inferring the underlying governing equations from measured data, or of the so-called observational approach (Gershenson 1999). The empirical mode decomposition (EMD) is proposed as a means of separating the *NDVI* signal into components and identifying those that interfere with the

surface signal, thus producing a consistent long record of remote sensing data for climate and vegetation studies. The EMD is based on the local characteristic time scale of the data and provides sharp identifications of embedded structures of nonlinear and nonstationary processes [see next section and, for more detail, Huang et al. (1998) for definitions]. To be precise, the *NDVI* and SZA signals are decomposed in terms of intrinsic mode functions (IMFs) by using the EMD to simplify and describe accurately the *NDVI*-SZA relationship. That is, one can tune the EMD to uncouple the seasonal and interannual components of the *NDVI* signals from SZA-drift induced structures.

This chapter is organized as follows. The standard processing of *NDVI* imagery is briefly described; this application of the EMD technique is presented with examples of trend detection in SZA latitudinal profiles to measure and correct orbital drift impact; a correction approach is tested and evaluated within the framework of EMD-filtering on 1 degree *NDVI* datasets; and finally, the extension to 8-km datasets is investigated, and the necessary pre-processing steps are described.

8.2. Processing of NDVI imagery

The *NDVI* data used in the analysis were the standard 8-km bimonthly continental maximum *NDVI* product of the Global Inventory Mapping and Monitoring System (GIMMS) group, from 1981 to the present (hereafter referred to as “the GIMMS dataset”). This dataset was mapped to an Albers equal-area projection, calibrated by using the method of Vermote and Kaufman (1995) and corrected for sensor degradation by using a technique based on stable desert targets (Los 1998). The dataset also includes corrections for stratospheric volcanic aerosols for the April 1982 El Chichon eruption (applied to the data from April 1982 to December 1984) and the June 1991 Mt. Pinatubo eruption (applied to the data from June 1991 to December 1993). The technique uses maximum *NDVI* as a compositing technique with cloud screening based on an AVHRR channel 5 thermal threshold value over Africa (283K) and South America (273K). Pixels having scan angles greater than 45 degrees from nadir view and pixels associated with unrealistic reflectance values were also screened prior to compositing (see Mahoney et al. 2001 for more details of the GIMMS processing system).

The reliability of a long *NDVI* time series can be checked by examining the trends over regions of known and unchanging vegetation cover, such as

deserts. From the other sources of *NDVI* variation other than land-surface vegetation, the solar zenith angle and, possibly, volcanic aerosols are known to cause a systematic trend in satellite datasets (Cihlar et al. 1998, Gutman 1999, Kaufmann et al. 2000). Independently, Los (1998), Kaufmann et al. (2000), and Slayback et al. (2003) showed that the effects of the solar zenith angle and volcanic aerosol on the *NDVI* over desert regions are negligible. Because these factors are small, residual errors in calibration are best detected from the desert *NDVI*. Three features of these desert time series are useful for evaluation: long-term trends, features at satellite transitions, and the variance. Slayback et al. (2003) reported that no trends, either for individual satellites or across the entire time period, were significant at the 95 percent level for different desert areas in the GIMMS dataset. However, Slayback et al. (2003) found statistically significant low-latitude trends from the 5° to 25° latitude bands, which are contaminated by substantial solar zenith angle effects and do not reflect real vegetative trends. Importantly, these researchers also reported that smooth trends were generally maintained across satellite transitions in desert signals, indicating the success of the calibration and sensor degradation schemes in intercalibrating the data from the various satellites to compose the GIMMS dataset (Slayback et al. 2003).

Figure 8.1 shows $1^\circ \times 1^\circ$ uncorrected *NDVI* data at various latitude and longitude coordinates. Vegetated tropical areas (Fig. 8.1b–f), and especially tropical forests (Fig. 8.1c–f), present an evident trend connected to the orbital drift or SZA variation. Desert areas (Fig. 8.1a) and northern latitudes (Fig. 8.1g–i) appear less affected by changes in SZA due to orbital drift. These results are consistent with Slayback et al.'s (2003) observations. Previous studies of growing season patterns in northern latitudes (greater than 45°N) by using the *NDVI* from the AVHRR (Kaufmann et al. 2000, Myneni et al. 1997, Slayback et al. 2003, Tucker et al. 2001) found no statistically meaningful relation between *NDVI* trends and trends connected with the orbital drift or SZA variation. Qualitatively, this finding suggests that the magnitude of SZA trends at those higher latitudes is very small compared with the high seasonal variation in the *NDVI* of vegetated areas. Thus, the orbital drift's contribution would be minimal to the *NDVI* signal at those latitudes. However, explicit quantification of these contributions has not been reported. Moreover, Gutman (1999), using a different AVHRR dataset, doubted the reliability of *NDVI* data for inferring the northern latitude greening trend since potential artifacts related to SZA and inter-sensor changes could remain in the data. In the next section, it is explained

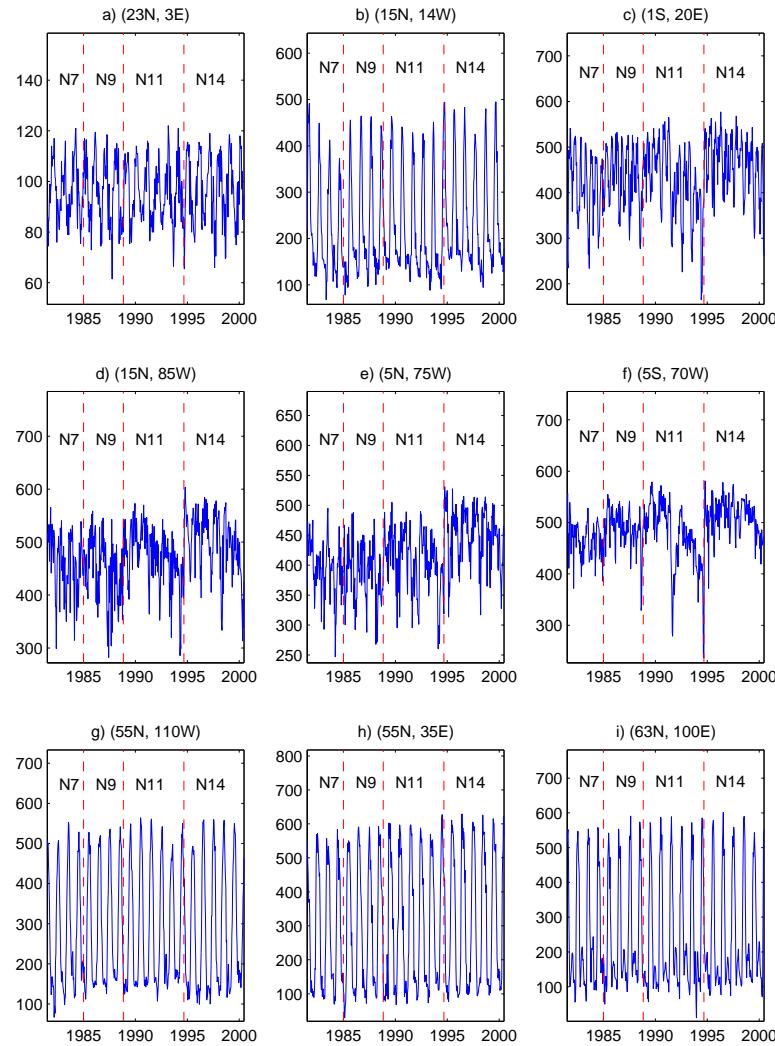


Figure 8.1: The $1^\circ \times 1^\circ$ uncorrected $NDVI \times 1000$ data at nine different locations. Note that the vertical scales are a function of land cover: vegetated tropical areas (b-f), and especially tropical forests (c-f), present an evident trend connected to the SZA variation; desert areas (a), and northern latitudes (g-i) appear less affected by changes in SZA.

how to extract interannual trends from the $NDVI$ signal, measure the contributions and significance of the SZA trends in the $NDVI$ variability at

all latitudes, and remove them from the *NDVI* signal accordingly.

8.3. Empirical mode decomposition

Huang et al. (1998) introduced empirical mode decomposition (EMD) as method for the representation of nonlinear and nonstationary data that shows clearly a physical scale or frequency content. As Huang et al. (1998, 1999) explained, the EMD method, contrary to almost all others previously proposed, is empirical, intuitive, direct, *a posteriori*, and adaptive, with the basis functions based on, and derived from, the data in question.

The decomposition uses the simple assumption that any data consist of different simple intrinsic modes of oscillations. Each mode may or may not be linear and will have the same number of extrema and zero-crossings. Furthermore, the oscillation will also be symmetric with respect to the “local mean.” At any given time, the data $X(t)$ may have many different coexisting modes of oscillation, each one superimposed on the others. The result of combining these modes is the final complicated data. In the EMD, each of these oscillatory modes is represented by an intrinsic mode function (IMF) with the following definition: (a) in the whole dataset, the number of extrema and the number of zero-crossings must either be equal or differ at most by one, and (b) at any point, the mean value defined by the envelopes of local maxima and local minima is zero.

With this definition, one can decompose any function or data as follows: derive the “local mean” m_1 by computing the mean between the upper envelope and lower envelope of the signal. The upper (lower) envelope is found by connecting all the local maxima (minima) by a cubic spline line. The difference, $h_1 = X(t) - m_1$, ideally should be an IMF, for the construction of h_1 described above should satisfy all the requirements of an IMF. However, even if the fitting is perfect, a gentle hump on a slope can be amplified to become a local extremum, and a new iteration over h_1 will be needed.

While the first condition is absolutely necessary for separating the intrinsic modes and for defining a meaningful instantaneous frequency, the second condition is also necessary in case the neighboring wave amplitudes have too large a disparity. In this way, one can actually recover the proper modes lost in an initial examination. Huang et al. (1998) named this process “sifting.” In fact, the sifting process can recover signals representing low amplitude riding waves by iteration. The procedure is illustrated in Huang et al. (1998,1999). The sifting process serves two purposes: to eliminate riding waves and to make the wave profiles more symmetric. Toward

these ends, the sifting process has to be repeated as many times as are required to reduce the extracted signal to an IMF. In the subsequent sifting process, h_1 is treated as the data, then $h_1 - m_{11} = h_{11}$. After repeated sifting, up to k times, h_{1k} becomes an IMF, $h_{1(k-1)} - m_{1k} = h_{1k}$; then, the first IMF component from the data is designated as $c_1 = h_{1k}$.

Overall, c_1 should contain the finest scale or the shortest period component of the signal. We can separate c_1 from the rest of the data by

$$X(t) - c_1 = r_1 . \quad (8.2)$$

Since the residue r_1 may still contain longer period components, it is treated as the new data and subjected to the same sifting process as described above. This procedure can be repeated to all the subsequent r_j 's, and the result is

$$r_1 - c_2 = r_2 , \dots , r_{n-1} - c_n = r_n . \quad (8.3)$$

The sifting process can be stopped by any of the following predetermined criteria: either when the component c_n or the residue r_n becomes so small that it is less than the predetermined value of substantial consequence, or when the residue r_n becomes a monotonic function from which no more IMFs can be extracted. Even for data with zero mean, the final residue still can be different from zero. If the data have a trend, the final residue should be that trend. By summing up Eqs. (8.2) and (8.3), we finally obtain

$$X(t) = \sum_{j=1}^n c_j + r_n . \quad (8.4)$$

Thus, we achieve a decomposition of the data into n -implicit modes, with a residue r_n , which can be either the mean trend or a constant. The trends are obtained from the last IMFs of the EMD approach. In this application of the EMD, the SZA and NDVI signals are overextended at the boundaries with seasonal profiles with no trends to reduce known boundary problems in the EMD decomposition that might also affect the correction.

8.4. Impact of orbital drift on NDVI and EMD-SZA filtering

This section presents EMD as a viable alternative to minimize *NDVI* SZA-inaccuracies and to characterize uncertainties more robustly than was possible previously. As discussed, the EMD method adjusts itself to local extrema and generates, by the sifting process, zero references and trends. Figure 8.2

shows the effects of the satellite drift on the SZA at different 10° latitude bands. An increasing trend is observed in each satellite due to its delay in the equatorial crossing time. This trend, superimposed on each plot, is accurately extracted and associated with the r_n component in the EMD iteration. The drift effects are more pronounced at lower latitudes (higher slopes), whereas seasonal variations dominate at higher latitudes in the two hemispheres. Note also that equatorial SZA plots (10°S to 10°N) have an extra oscillation due to the solar nadir moving past the target latitude, causing an increase in the SZA at six-month intervals rather than at yearly intervals (Privette et al. 1995).

Similarly, the EMD is used to extract the *NDVI* trends that may be associated with satellite orbital drift, reducing the interference of many other components with the surface signal, for example.

The SZA values in general are different in the same latitudinal bins in the two hemispheres because of the difference in the local time of observation within the same orbit; i.e., on a given day, a given southern latitude is observed at a later local time than the same northern latitude. A remarkable extra oscillation is observed in the 60° – 75°N bin. This oscillation is an indication of the data-screening scheme used in the GIMMS dataset for high SZA angles. Furthermore, this oscillation is an artifact from the averaging over a wide range of latitudes when the SZA is greater than 85° during the winter and is reporting an arbitrary missing value. In the *NDVI* data, this artifact will be associated with terminator effects and screened out. As expected, this artifact is not observed in the highest southern bin since land-vegetation data goes only from 40°S to 57°S, and the SZA angles are less than 85° during the winter. The components of the EMD, as Fig. 8.2 suggests, are usually physically meaningful, since each scale is defined by the physical data themselves.

Similarly, the EMD is used to extract *NDVI* trends that may be associated with satellite orbital drift, reducing the interference of many other components with the surface signal, e.g., the BRDF, or the cloud contamination. Thus, *NDVI* time series are decomposed into two components for each satellite: a wave-seasonal and trend components. The hypothesis in this work is that if the *NDVI* is affected by orbital drift, its trend component is the one that should show this effect. Correlating the *NDVI*-trend and SZA-trend components, we can account for a large percentage (> 90%) of the solar zenith angle affect upon the *NDVI* (hereafter referred to as t_{szac}), remove these effects from the *NDVI* data, leaving intact the phenologic components in the wave-seasonal components, and reconstruct the

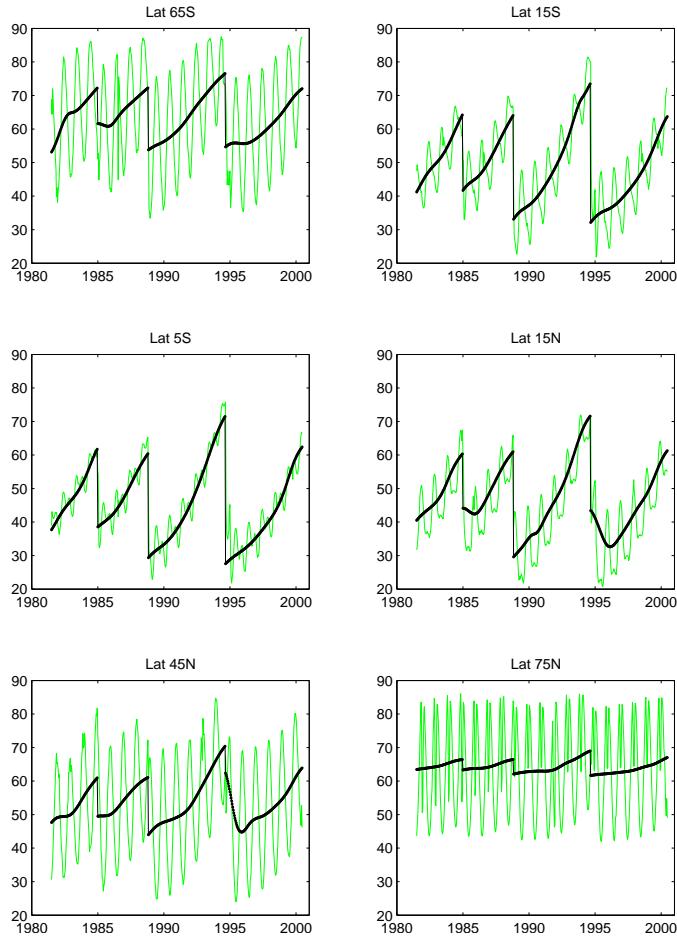


Figure 8.2: Solar zenith angle drift by 10° latitude bands. Drift effects are more pronounced at lower latitudes (higher slopes), whereas seasonal variations dominate at higher latitudes in the two hemispheres. An extra oscillation is observed in equatorial plots (10°S to 10°N) due to the solar nadir moving past the target latitude. A remarkable extra oscillation in the $60^\circ - 75^\circ\text{N}$ bin is observed. This oscillation is an indication of the data screening scheme used in the GIMMS dataset for very high solar zenith angles.

NDVI signal without the solar zenith angle variation.

In fact, as Huang et al. (1998) suggested, we can also use the IMF components as time-space filters. Traditionally, filtering is carried out in frequency space only. Using IMF, however, we can devise a time-space filter-

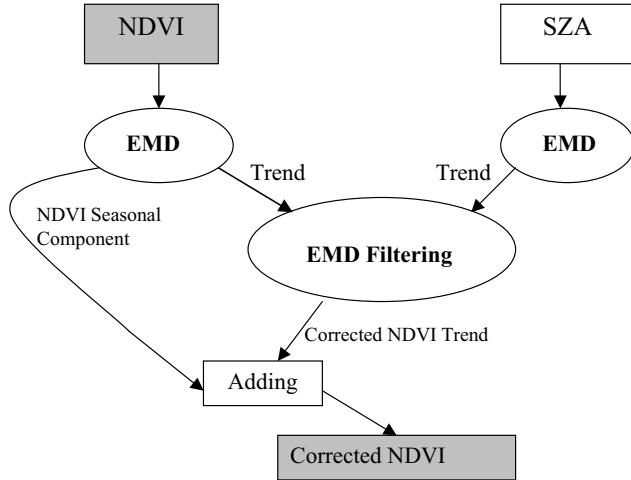


Figure 8.3: Flowchart of the solar zenith angle time-space EMD-filtering.

ing. For example, the result of low, high, and band time-space EMD-filters in a signal having n -IMF components can be simply expressed as

$$\begin{aligned}
 X_{lk}(t) &= \sum_{j=k_1}^n c_j + r_n, \\
 X_{hk}(t) &= \sum_{j=1}^{k_2} c_j,
 \end{aligned} \tag{8.5}$$

and

$$X_{bk}(t) = \sum_{j=k_1}^{k_2} c_j,$$

respectively, where $1 < k_1 < k_2 < n$. The advantage of this time-space filtering is that the results preserve full meaning, nonlinearity and nonstationarity in physical space. The EMD-filtering strategy just described is summarized in Fig. 8.3.

To quantify more precisely the contribution of drift interference with the surface signals and their spatial distribution, a canonical correlation analysis was performed. This analysis is a multivariate statistical technique used to examine and describe the strength of a linear association between

two sets of random variables (Bretherton et al. 1992, Cherry 1996, Wallace et al. 1992). The canonical correlation analysis is based on the singular value decomposition (SVD) of the paired-mode of the correlation matrix between the EMD-trend components of the *NDVI* and the SZA: $[U, S, V] = \text{SVD}(A)$, where A is the correlation matrix. The singular value decomposition of a matrix A produces a diagonal matrix S and a set of matrices U , and V , with orthonormal column vectors, such that $A = U \times S \times V^t$. The column vectors of the matrices U and V , referred to as the canonical factors, constitute an optimal orthonormal basis for the EMD-trend components of the *NDVI* and the SZA, respectively. The contribution of each canonical factor to the correlation matrix A is a function of the values in the diagonal matrix S , which are sorted in descending order. For more details about the singular value decomposition, see Golub and Van Loan (1989) and Trefethen and Bau (1997).

8.5. Results and discussion

In this section, we focus our EMD analysis on the global 1° degree *NDVI* time series starting in July 1981 (NOAA-7) and ending in November 2000 (NOAA-14). Table 8.1 shows the results of the SZA EMD-filtering approach for selected 1° latitude-longitude pixels from regions representing different biomes. It was found that the percentage variance of the *NDVI* variability explained by the filtered t_{szac} component decreased as the latitude increased (k-p) and as vegetation biomass decreased (a,b). This finding implies that t_{szac} variability is very small at those northern latitude regions and that the the *NDVI* is minimally sensitive as the vegetation biomass decreases. This result is confirmed by the coefficient of determination between t_{szac} and the correspondent 10° latitude SZA trends. In most of those regions, r^2 is less than 75, but not in tropical areas (c-h), especially tropical forests (c-g), where the percentage variance explained is greater than 10 and the r^2 above 85. The coefficient of determination between t_{szac} and the uncorrected (UNDVI) and corrected (CNDVI) trends indicates that the EMD-filtering efficiently removed those artifacts related to the solar zenith angle and intersensor changes at all latitudes and biomes.

Each panel of Fig. 8.4 shows the uncorrected *NDVI* and its associated trend (U-time series), the solar zenith angle component t_{szac} removed (S in the figure) and the corresponding corrected *NDVI* and trend (C-time series), respectively. Each panel represents the same one degree time series at the latitude-longitude pixels described in Table 8.1. This figure

Table 8.1: Percentage variance of $NDVI$ explained by the filtered t_{szac} component, coefficient of determination of orbital drift SZA trends and extracted $NDVI$ t_{szac} , and coefficient of determination of extracted t_{szac} and other $NDVI$ trends: UNDVI (uncorrected), CNDVI (corrected). All coefficients are significant with $p < 0.00001$. Column (a) gives SZA and t_{szac} , column (b) reports the t_{szac} and UNDVI, and column (c) states the t_{szac} and CNDVI, respectively.

Location	(Lat, Lon)	% (a)	$r^2 \times 100$		
			(b)	(c)	
a) Sahara	(23°N, 3°E)	2	72	88	37
b) Sahel	(15°N, 14°W)	3	84	100	25
c) Central Forest	(1°S, 20°E)	14	87	100	27
d) Upper Guinea	(8°N, 5°W)	10	79	99	7
e) Center America	(15°N, 85°W)	18	86	99	27
f) Colombia	(5°N, 75°W)	29	89	99	25
g) Amazon	(5°S, 70°W)	23	87	100	37
h) NE-Brasil	(5°S, 38°W)	11	86	99	36
i) Indochina	(15°S, 105°E)	9	74	99	12
j) Australia	(25°S, 135°E)	26	92	91	56
k) Mexico	(27°N, 105°W)	8	71	83	26
l) Great plains	(45°N, 105°W)	2	80	97	37
m) Canada	(55°N, 110°W)	1	68	95	77
n) Labrador	(55°N, 65°W)	1	52	94	43
o) Central Russia	(55°N, 35°E)	1	68	96	60
p) Siberia	(63°N, 100°E)	1	51	98	58

confirms the main conclusion of Table 8.1: (1) regions in the tropics, especially tropical forests, are the most SZA-affected; (2) regions with a low vegetation biomass and at the high northern or low southern latitudes are least contaminated; and (3) all remaining components in the C-time series are found to be statistically independent of the SZA. Therefore, the EMD-filtering approach presented here constituted a sound SZA correction technique, especially in tropical forests where the $NDVI$ signal was shown to be most affected. Notice in Fig. 8.4 that the SZA EMD-filtering keeps the known salient features found with the $NDVI$ time series, like the drop in the Sahel in 1984–1985 (b) due to a serious drought (Nicholson et al. 1998); the aforementioned long-term stability of deserts (a), and most tropical

forests (c-f) (Los 1998); and Mt. Pinatubo's volcanic eruption's effects on high-vegetated tropical areas (f). Furthermore, an increasing trend is observed in the northern latitude panels (g-h). The corrected trends still reveal an expected discontinuity around September of 1994. This discontinuity is due to the loss of NOAA-11 and the lack of new satellites that could take its place and continue the collection of data. The only option available to preserve the continuity of the time series was the old NOAA-9 now passing over the equator in the descending node. In late January of 1995, NOAA-14 was operational and took NOAA-9's place. Lotsch et al. (2003), applying independent component analysis (ICA) on the seasonal NDVI anomalies of the corrected data (i.e., the seasonal cycle was removed from the data before performing temporal and spatial ICA), showed that the time series is a consistent record suitable for studying long-term vegetation dynamics (Lotsch et al. 2003).

Figure 8.5 shows similar time series as in Fig. 8.4 but with 10° latitude bands. Although the mixture of biomes by latitude is high, these results reinforce the conclusions of Fig. 8.4 and Table 8.1, and show explicitly (LAT75N) the terminator effects upon the *NDVI* at a high solar zenith angle during winter. These results also show that the corrected trend presents a steady increase at latitudes higher than 35° , extending previous studies that reported only northern latitude greening for latitudes higher than 45° (Myneni et al. 1997, Slayback et al. 2003, Tucker et al. 2001). A thorough analysis of these trends is not within the scope of this present chapter, but should provide the quantitative confirmation of the plant growth associated with a longer active growing season at those latitudes, a confirmation that is lacking in previous studies.

Figure 8.6 shows the contribution of drift interference with the surface signal and their spatial distribution according to the canonical correlation analysis. The EMD-filtering removes *NDVI* trends that are more than 80% (red areas) correlated to the SZA trends. Areas with trends that have a lower correlation and their contribution lower than the *NDVI* uncertainty (usually less than 50%, yellow areas) were not altered. In summary, the results show that (1) regions in the tropics are the most SZA-affected due to their SZA trend magnitudes, and that (2) the high northern latitudes and regions with low vegetation biomass are less contaminated since the SZA component represents a small part of the *NDVI* signal.

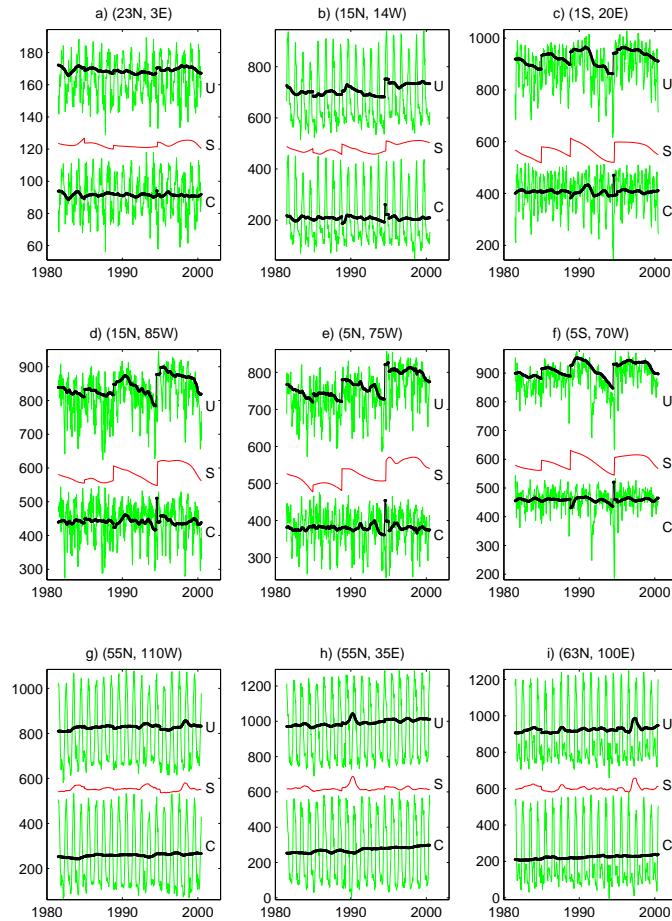


Figure 8.4: Each panel shows the $NDVI \times 1000$ at different stages of the EMD-filtering scheme: the uncorrected $NDVI$ and its associated trend (U-time series), the EMD-filtered component t_{szac} (S-time series) and the reconstructed $NDVI$ and its associated corrected trend (C-time series) at different 3° by 3° latitude-longitude regions. U- and S-time series are shifted by the range of variation of the particular C-time series of each panel. Note that vertical scales are a function of land cover. In panel (g), the drop in 1991 is due to Pinatubo effects.

8.6. Extension to 8-km data

Since 1° datasets are smoother than 8-km data due to spatial averaging, the extension of the approach to 8-km data was straightforward with a

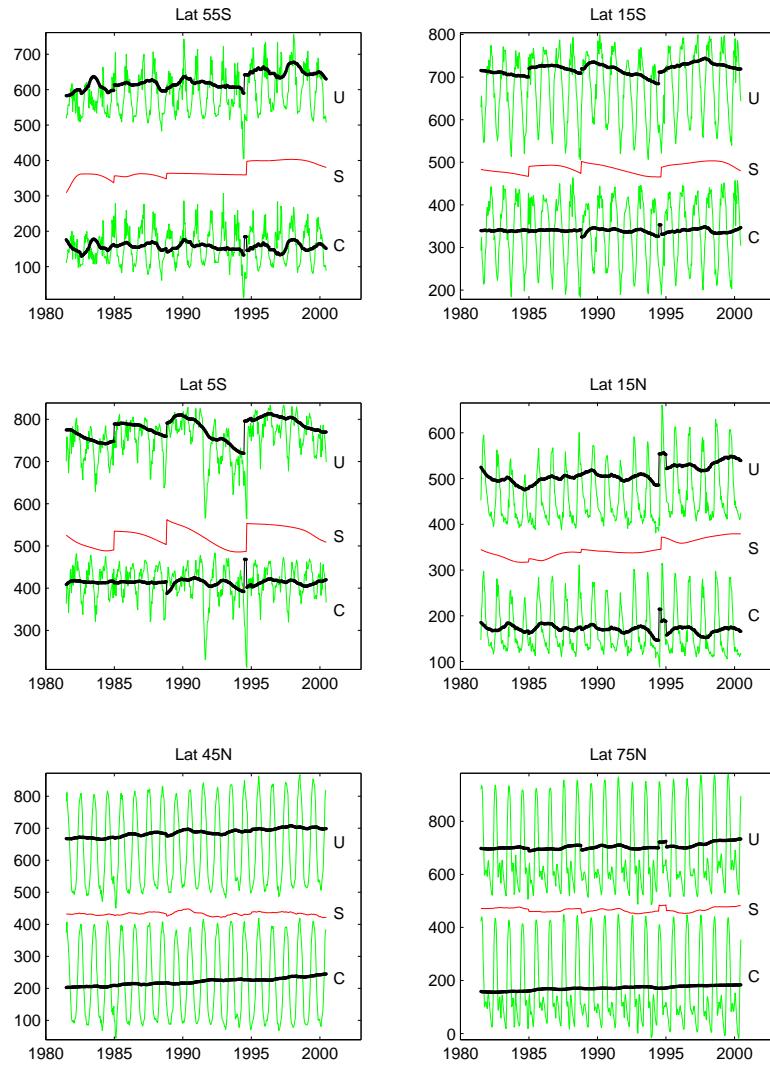


Figure 8.5: Each panel shows different stages of the EMD-filtering scheme, as in Fig. 8.4, of the average of $NDVI \times 1000$ at 10° latitude bins. As before, the U- and S-time series are shifted by the range of variation of the particular corrected $NDVI$ of each panel.

pre-processing step of a two-month temporal median nonlinear low-pass filter applied to each pixel and a kriging interpolation for missing values,

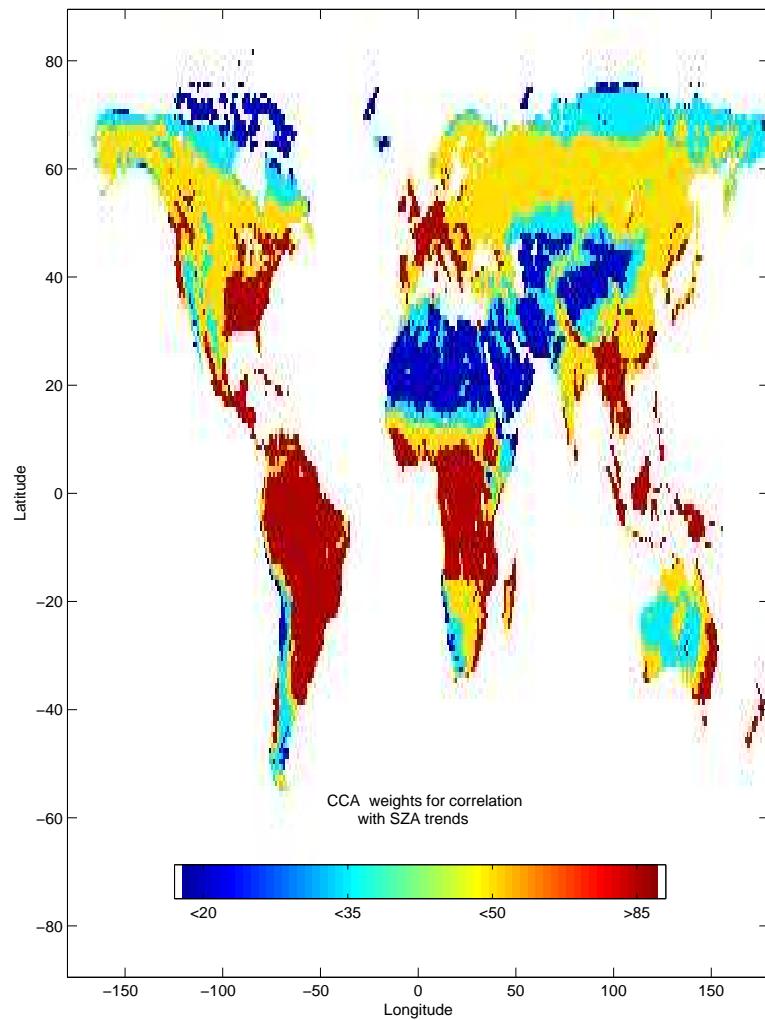


Figure 8.6: Contribution of drift interference with the surface signals and their spatial distribution according to the canonical correlation analysis. Red areas represent regions where the orbital drift affected the NDVI signal and the EMD correction was applied; yellow areas have a low correlation between the drift and NDVI trends and its contribution is lower than NDVI uncertainty; and blue areas are regions where calibration already accounts for possible drift artifacts.

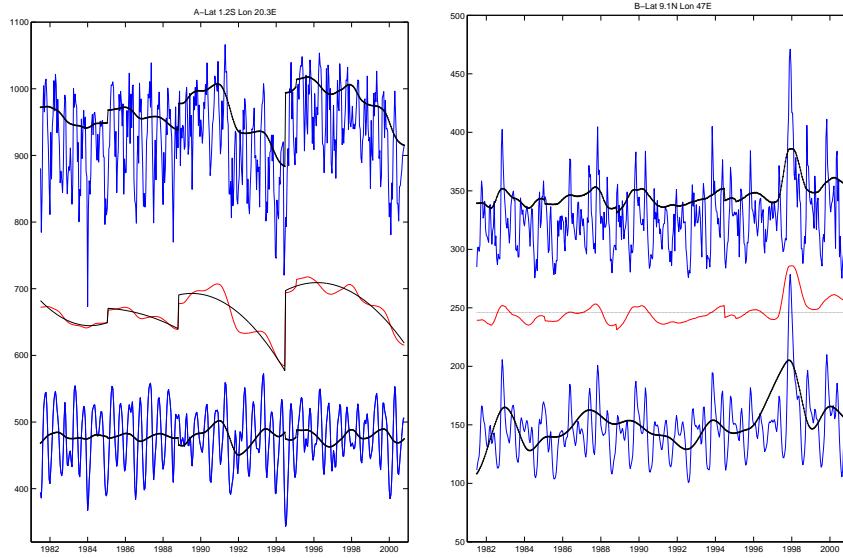


Figure 8.7: EMD decomposition and removal of SZA-correlated trends. (a) Tropical forest signal before satellite drift correction (series at top), trend removed (middle) and resulting series (bottom). (b) Low vegetated signal that is not corrected due to low correlation with the SZA trends. The uncorrected and trend signals are shifted by the range of variation of the particular corrected NDVI $\times 1000$. Note that the vertical scales are a function of land cover.

especially in winter months in northern latitudes. This step increases the signal-to-noise ratio of the 8-km *NDVI* time series, allowing for the use of the time-space EMD-filtering technique without affecting trends. We applied the time-space EMD-filtering to those 8-km *NDVI* pixels whose corresponding 1° *NDVI*-trend has a contribution of more than 50% from the SZA trends. This contribution was quantified by applying a canonical correlation analysis between *NDVI* trends and latitudinal SZA trends (Fig. 8.6). Notably, as in the 1° case, there are 8-km *NDVI* pixels with large or small solar zenith angle interference. Figure 8.7 shows both cases: trend removal from the *NDVI* signal due to high correlation with the satellite drift, and trends that have not been altered because of low contribution to the *NDVI* signal from the satellite drift.

8.7. Integration of NOAA-16 data

Figure 8.8 shows two EMD features that are very useful in classification problems. The *NDVI* seasonal variability of the Sahel is characterized by the standard deviation of the EMD-wave seasonal component. A simple classification based on uniform quantization of the median value of the EMD-trend component provides the amount of vegetation biomass expected in the region. While the classification shows the different characteristic latitudinal gradients observed in the Sahel (Nicholson et al. 1998), the standard deviation of the seasonal component identifies regions with low to high phenologic activity.

For purposes of comparison and validation, we have added a small box to the median value image that corresponds to the 1° box, labeled as (b), used in Figs. 8.1 and 8.4 and Table 8.1. According to the classification image, the average of the trend on this box is around 200, which coincides with the average value at 1° resolution (Fig. 8.4b).

These features appear as invariants and were thoroughly exploited as a result of crossovers with other satellites carrying more recent sensors. This method provides a way to account for spectral differences and, thus, to obtain significant improvements in the NDVI-AVHRR. This procedure is used as an intercalibration of the single-gain sensors aboard NOAA-7-14 (historical data) satellites and the dual-gain sensors aboard NOAA-16-17 satellites with data from SPOT-Vegetation global data. SPOT data were averaged from 1-km to 8-km globally and then decomposed by using the EMD method. The interannual EMD-trend of NDVI-SPOT signals were computed from 40 months of data, 20 months overlapping with NOAA-14 and 20 months with NOAA-16. This trend was determined to be invariant through time over the period examined. A similar trend was extracted from the same period of NOAA-14 and NOAA-16. A non-linear regression using seasonal variability and the median trend from each sensor was performed to establish the coefficients that transform the historical AVHRR record into the range of variation of the current and upcoming suite of visible and NIR sensors, such as MODIS, SPOT-Vegetation and others. A similar procedure was performed for the NOAA-16 data. Once the trends from the historical and NOAA-16 data were transformed into the common range, the data were reconstructed, and a consistent time series was established. A krigging interpolation was applied on the integrated data to reduce noise and attenuate the effect of cloudy and missing pixels. This data will be available via ftp from the Global Land Cover Facility at the University of

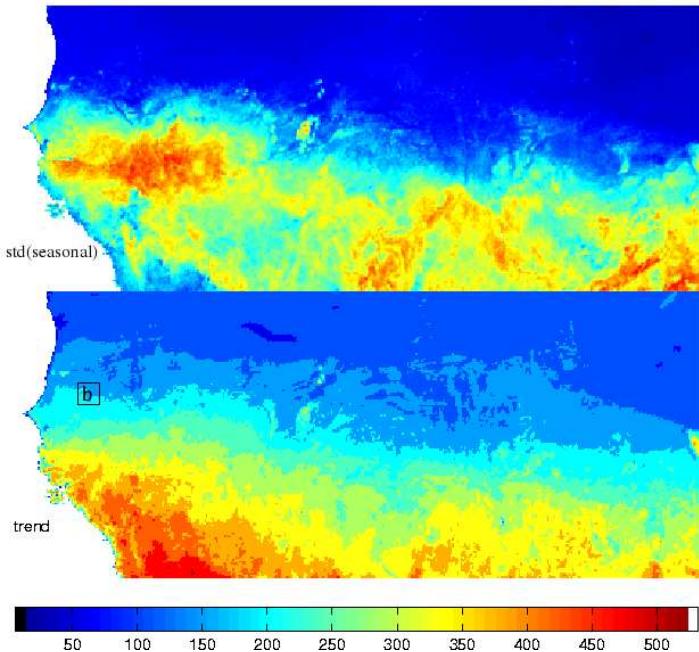


Figure 8.8: (a) Seasonal $NDVI$ variability and classification by uniform quantization of the $NDVI$ trend in the Sahel. The labeled box (b) represents the 1° box of Figs. 8.1b and 8.4b with a quantized median value close to 200.

Maryland (available online at www.glc.umd.edu).

8.8. Conclusions

We have presented an EMD approach for improving the $NDVI$ -AVHRR time series by removing spurious SZA trends induced by satellite drift and for intercalibrating the $NDVI$ -AVHRR and $NDVI$ from sensors with narrow spectral bands. By using this decomposition, we found a mean trend in the $NDVI$ that can be explained almost entirely by SZA trends. The remaining IMF components of the $NDVI$ were found to be statistically independent of the SZA. Although the correction eliminates up to 30 percent of the variability in $NDVI$ signals, it keeps all known vegetation features captured by the $NDVI$ time series. Therefore, we have shown that the EMD-filtering approach for obtaining the associated mean SZA trend from the $NDVI$ time series constituted a sound SZA correction technique.

In particular, we have removed SZA trends from tropical areas, especially forests, where the *NDVI* signal was shown to be more affected. As an additional gain, we have shown that the corrected *NDVI* and associated IMFs features can be used in concert for better spatial characterization and time series analysis. However, a thorough comparison with other measurements should be conducted in order to independently validate these results. Fortunately, in a few years time, the data archive from the new generation of satellites (SeaWiFS, MODIS, SPOT VEGETATION) may be long enough to allow the useful comparisons with current orbiting AVHRR sensors aboard NOAA-16 and NOAA-17. A comparison of ENSO-related interannual variations in land-surface *NDVI* from SeaWiFS and AVHRR over a 2-year period does show a high degree of consistency (Behrenfeld et al. 2001).

References

- Behrenfeld, M. J., J.T. Randerson, C. R. McClain, G. C. Feldman, S. O. Los, C. J. Tucker, P. G. Falkowski, C. B. Field, R. Frouin, W. E. Esaias, D. D. Kolber, and N. H. Pollack, 2001: Biospheric primary production during an ENSO transition. *Science*, **291**, 2594–2597.
- Betherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **6**, 541–560.
- Cherry, S., 1996: Singular value decomposition analysis and canonical correlation analysis. *J. Climate*, **9**, 2003–2009.
- Cihlar, J., J. M. Chen, Z. Li, F. Huang, R. Latifovic, and R. Dixon, 1998: Can interannual land surface signal be discerned in composite AVHRR data? *J. Geophys. Res.*, **103**, 23163–23172.
- Cihlar, J., D. Manak, and M. D'Iorio, 1994: Evaluation of compositing algorithms for AVHRR data over land. *IEEE Trans. Geosci. Remote Sens.*, **32**, 427–437.
- Cracknell, A. P., 1997: *The Advanced Very High Resolution Radiometer (AVHRR)*. Taylor & Francis, 300 pp.
- Cracknell, A. P., 2001: The exciting and totally unanticipated success of the AVHRR in applications for which it was never intended. *Adv. Space Res.*, **28**, 233–240.
- El Saleous, N. Z., E. F. Vermote, C. O. Justice, J. R. G. Townshend, C. J. Tucker, and S. N. Goward, 2000: Improvements in the global biospheric record from the Advanced Very High Resolution Radiometer (AVHRR).

- Int. J. Remote Sens.*, **21**, 1251–1277.
- Gershenson, N., 1999: *The Nature of Mathematical Modeling*. Cambridge University Press, 344 pp.
- Golub, G. H., and C. F. Van Loan, 1989: *Matrix Computations*. John Hopkins University Press, 664 pp.
- Gutman, G. G., 1999: On the use of long-term global data of land reflectances and vegetation indices derived from the advanced very high resolution radiometer. *J. Geophys. Res.*, **104**, 6241–6255.
- Holben, B. N., 1986: Characteristics of maximum-value composite images from temporal AVHRR data. *Int. J. Remote Sens.*, **7**, 1417–1434.
- Huang, N.E., Z. Shen, and S. R. Long, 1999: A new view of nonlinear water waves: the Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Kaufmann, R. K., L. Zhou, Y. Knyazikhin, N. V. Shabanov, R. B. Myneni, and C. J. Tucker, 2000: Effect of orbital drift and sensor changes on the time series of AVHRR vegetation index data. *IEEE Trans. Geosci. Remote Sens.*, **38**, 2584–2597.
- Los, S., 1998: *Linkages Between Global Vegetation and Climate*. PhD thesis, Vrije Universiteit and NASA/GSFC, 179 pp.
- Los, S., C. O. Justice, and C. J. Tucker, 1994: A global 1° by 1° NDVI data set for climate studies derived from GIMMS continental NDVI data. *Int. J. Remote Sens.*, **15**, 3493–3518.
- Lotsch, A., A. F. Mark, and J. E. Pinzón, 2003: Spatio-temporal deconvolution of NDVI image sequences using independent component analysis. *IEEE Trans. Geosci. Remote Sens.*, **41**, 2938–2942.
- Mahoney, R. L., C. J. Tucker, A. Anyamba, M. Brown, D. Slayback, S. Los, J. Pinzon, J. Kendall, E. Pak, Z. Brodner, D. Grant, M. Paris, and A. Morahan, 2001: Global remote sensing of vegetation from space by the NASA/GSFC GIMMS group. *International Workshop on Global Change*, J. Kudoh and K. Yamada, eds., Sendai Kiodo Printing Co. Ltd., 17–29.
- Myneni, R. B., I. Impens, and G. Asrar, 1993: Simulation of space measurements of vegetation canopy bidirectional reflectance factors. *Remote Sens. Rev.*, **7**, 19–41.
- Myneni, R. B., C. D. Keeling, C. J. Tucker, G. Asrar, and R. R. Nemani, 1997: Increased plant growth in the northern high latitudes from 1981–

1991. *Nature*, **386**, 698–702.
- Nicholson, S. E., C. J. Tucker, and M. B. Ba, 1998: Desertification, drought, and surface vegetation: An example from the west African Sahel. *Bull. Amer. Meteor. Soc.*, **79**, 815–829.
- Pinzón, J. E., J. Pierce, C. J. Tucker, and M. Brown, 2001: Evaluating coherence of natural images by smoothness membership in besov spaces, *IEEE Trans. Geosci. Remote Sens.*, **39**, 1879–1889.
- Price, J. C., 1991: Timing of NOAA afternoon passes. *Int. J. Remote Sens.*, **12**, 193–198.
- Privette, J. L., C. Fowler, G. A. Wick, D. Baldwin, and W. J. Emery, 1995: Effects of orbital drift on advanced very high resolution radiometer products: Normalized difference vegetation index and sea surface temperature. *Remote Sens. Environ.*, **53**, 164–171.
- Schaaf, C. B., F. Gao, A. H. Strahler, W. Lucht, X. Li, T. Tsang, N. Strugnell, X. Zhang, Y. Jin, J. P. Muller, P. Lewis, M. Barnsley, P. Hobson, M. Disney, G. Roberts, R. P. Dunderdale, R. P. d'Entremont, B. Hu, S. Liang, J. Privette, and D. Roy, 2002: First operational BRDF albedo and nadir reflectance products from MODIS. *Remote Sens. Environ.*, **83**, 135–148.
- Sellers, P. J., 1985: Canopy reflectance, photosynthesis and transpiration. *Int. J. Remote Sens.*, **6**, 1335–1372.
- Sellers, P. J., S. O. Los, C. J. Tucker, G. J. Collatz, C. O. Justice, D. A. Dazlich, and D. A. Randall, 1994: A global 1° by 1° NDVI data set for climate studies. Part 2: The generation of global fields of terrestrial biophysical parameters from the NDVI. *Int. J. Remote Sens.*, **15**, 3519–3545.
- Slayback, D. A., J. E. Pinzón, S. O. Los, and C. J. Tucker, 2003: Northern Hemisphere photosynthetic trends 1982–99. *Glob. Change Biol.*, **9**, 1–15.
- Trefethen, L. N., and D. Bau, 1997: *Numerical Linear Algebra*. SIAM, 361 pp.
- Tucker, C. J., 1979: Red and photographic infrared linear combinations monitoring vegetation. *Remote Sens. Environ.*, **8**, 127–150.
- Tucker, C. J., 1996: History of the use of AVHRR data for land applications. *Advances in the Use of NOAA AVHRR Data for Land Applications*, G. D'Souza, A. S. Belward, and J. P. Malingreau, eds., Brussels, 1–19.
- Tucker, C. J., D. Slayback, J. E. Pinzon, S. Los, R. B. Myneni, and M. G. Taylor, 2001: Higher northern latitude normalized difference vegetation index and growing season trends from 1982 to 1999. *Int. J. Biometeor.*,

- 45, 184–190.
- Vermote, E. F., and Y. J. Kaufman, 1995: Absolute calibration of AVHRR visible and near infrared channels using ocean and cloud views, *Int. J. Remote Sens.*, **16**, 2317–2340.
- Wallace, J. M., C. Smith, and C. S. Betherton, 1992: Singular value decomposition of winter time sea surface temperature and 500-mb height anomalies. *J. Climate*, **5**, 561–576.

Jorge E. Pinzón

Science Systems and Applications, Inc., Biospheric Sciences Branch, Code 614.4 NASA/Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

Jorge_Pinzon@ssaihq.com

Molly E. Brown

Science Systems and Applications, Inc., Biospheric Sciences Branch, Code 614.4, NASA/Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

mebrown@pop900.gsfc.nasa.gov

Compton J. Tucker

Biospheric Sciences Branch, Code 614.4, NASA/Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

compton@ltpmail.gsfc.nasa.gov

CHAPTER 9

HHT ANALYSIS OF THE NONLINEAR AND NON-STATIONARY ANNUAL CYCLE OF DAILY SURFACE AIR TEMPERATURE DATA

Samuel S. P. Shen, Tingting Shu, Norden E. Huang, Zhaohua Wu,
Gerald R. North, Thomas R. Karl and David R. Easterling

The empirical mode decomposition (EMD) method is used to analyze the nonlinear and non-stationary annual cycle (NAC) in climate data. The NAC is defined as an intrinsic mode function generated in the EMD process and has a mean period equal to a year. Both Hilbert and Fourier spectra of the NAC are examined to validate the power density at the frequency of one cycle per year. The NAC differs from the strictly periodic thirty-year-mean annual cycle (TAC), which is now commonly used in climate anomaly analysis. It is shown that the NAC has a stronger signal of the annual cycle than that of the TAC. Thus, the anomalies derived from the NAC have less spectral power at the frequency of one cycle per year than those from the TAC. The marginal Hilbert spectra of the maximum daily surface air temperature of ten North America stations demonstrate the expected characteristics of an annual cycle: the NAC's strength is proportional to the latitude of the station location and inversely proportional to the heat capacity of the Earth's surface materials around a station. The NAC of the Niño3.4 sea surface temperature analysis indicates that El Niño events correspond to a weaker annual cycle, as suggested by wavelet analysis.

9.1. Introduction

Conventionally, when studying climate change via surface air temperature (SAT) data, one analyzes anomaly data, which are the departures of the observed SAT from its normal value called “climate normal” or “climatology.” The climatology of a station is often defined as the thirty-year mean of the station’s data for each month. The 30-year period progresses as time passes. Currently, 1961–1990 is commonly used as the climatology period, whereas 1951–1980 was used in the 1980s and the early 1990s. Thus, the January climatology is the simple mean of the January temperatures from 1961–1990. This climatology is the expected monthly temperature through-

out a year and is the so-called “annual cycle,” or “seasonal cycle.” Thus, the annual cycle of climate data describes the expected climate’s regular oscillation from the winter cold to the summer heat and back to the winter cold. This fixed cycle, marked by the spring, summer, fall, and winter seasons, reflects the relative positions of the sun and the Earth, and the heat capacity of the Earth’s surface (land or ocean), and some topographic characteristics. The commonly used thirty-year-mean annual cycle, referred to as the TAC hereafter, has a temporal resolution of a month. Apparently, the thirty-year mean of the SAT of a fixed day of a year, say October 6, can be very different from each other when one considers the 30-year means of 1951–1980 and 1961–1990. Thus, one normally does not consider an annual cycle of daily resolution when using the 30-year-mean approach for calculations. The monthly TAC also varies when using different periods of 30 years, but the variation is much smaller compared to that of the daily TAC. Figure 9.1 shows the daily and monthly TAC of the maximum daily surface air temperature of the Victoria station, Canada, and the results are calculated by using three different climatology periods: 1951–1980, 1961–1990, and 1971–2000. The figure clearly indicates a large variation of the daily climatology and a relatively smaller variation of the monthly climatology. However, the variation of the monthly climatology is still noticeable, particularly in the months of January, March and April.

Among the periodic or quasi-periodic signals of the climate, the annual cycle is often the strongest one in the temporal climate data, particularly for SAT and precipitation in the mid or high latitudes. This fact can be intuitively understood but may not be always easy to quantify. The Fourier spectrum of the data of a station for SAT shows a distinctive peak at one cycle per year (Fig. 9.2). The other peaks, such as those at the bi-annual and interannual cycles, are several orders smaller than that of the annual cycle in Victoria, Canada. However, due to the motions of the atmosphere and ocean, the annual cycle is far from being this simple. The TAC definition of the monthly annual cycle has become a standard way to derive climate anomalies in climate data processing, but the daily climatology is not well defined in terms of the 30-year-mean. The daily resolution data are useful in assessing climatic change due to the nonlinear interaction in the Earth’s climate system. This interaction can result in energy transfer from high frequency oscillations (daily oscillations) to low frequency oscillations (inter-decadal oscillations). Thus, the daily anomalies should be considered, just as they are in climate modeling studies. To obtain the daily anomalies, one has to use a well-defined climatology. Due to the large

Annual Cycle of Daily Surface Air Temperature Data

219

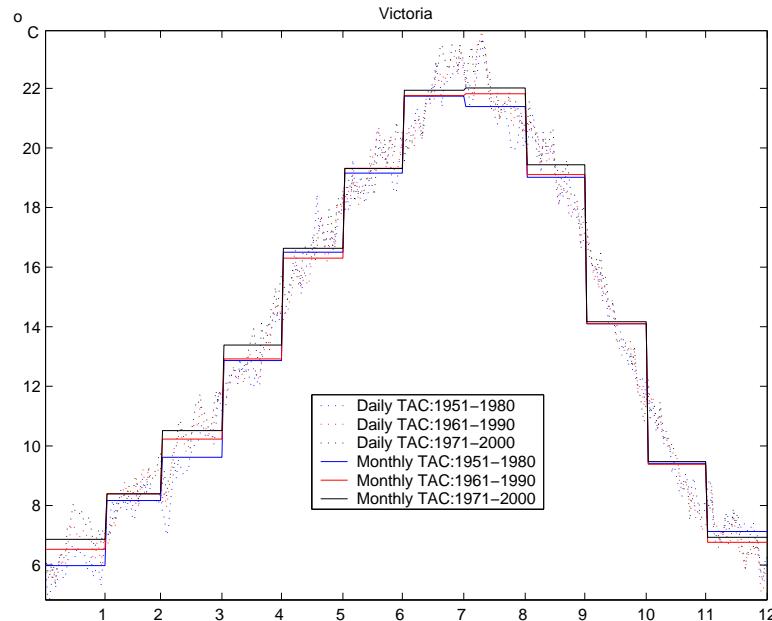


Figure 9.1: Daily and monthly TAC derived from the daily maximum surface air temperature data at the Victoria station, Canada.

variation of the 30-yr mean of daily data and due to the step-function behavior of the monthly TAC, alternative methods are required to define the daily climatology, or, in general, a method is needed to define the climatology at the resolution of a given dataset, whether daily, weekly or monthly. Time-frequency analysis appears to be the right tool for this purpose. Thus, the window-Fourier-transform, the wavelet analysis, and the Hilbert-Huang transform (HHT) can be used. The empirical mode decomposition (EMD) in the HHT procedure can lead to an intrinsic mode function that has a mean period of a year and carries some nonlinear and non-stationary signals of a climate system (Gloersen and Huang, 2003). The nonlinear and non-stationary annual cycle, referred to as the NAC, is thoroughly analyzed in both time and frequency domains, and is compared with the TAC whenever appropriate.

In a TAC, the nonlinear interactions of the different components of a climate system are largely ignored, and the anomalies derived from the TAC climatology still contain some apparent signal of the annual cycle because the TAC cannot completely filter out the annual signal from the

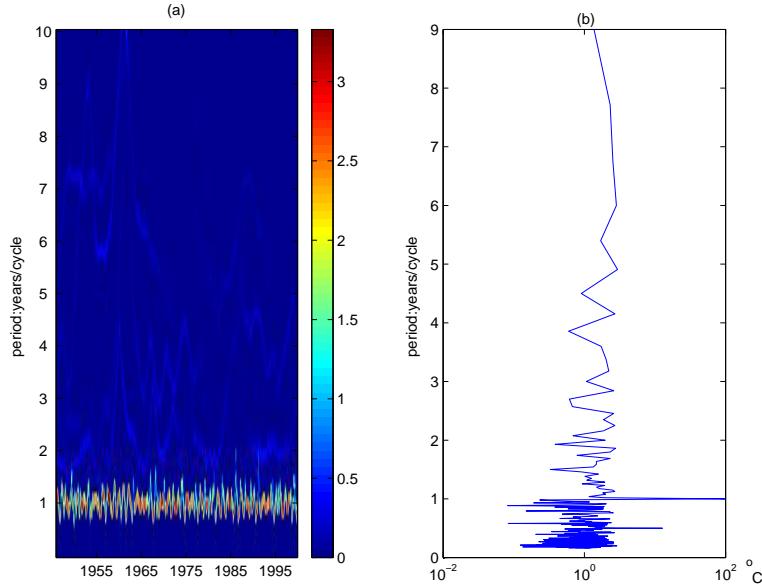


Figure 9.2: (a) Hilbert spectrum of the monthly maximum SAT data at Victoria station, Canada. The unit of the color scales is $^{\circ}\text{C}$. (b) Fourier spectrum of the same data.

climate data. Due to the nonlinear nature of the climate system, the spectral power of the TAC is not monotone at the one cycle per year; rather, it is distributed in a frequency interval that contains the annual cycle of one cycle per year. Thus, when the annual cycle in the residue of data minus the TAC is strong, the calculations of the linear trends of the seasonal or monthly climate changes can be affected because the non-monotonic “annual cycles” in the anomaly data are not synchronized to have peaks and valleys at the same month or season. Thus, the anomalies with respect to the TAC can lead to much uncertainty in a linear trend that is being used to measure the local, monthly climatic changes.

In contrast, the NAC reflects not only the periodic variation of the solar irradiance but also the nonlinear and non-stationary processes of atmospheric motions. Since the NAC is obtained through a sifting process, the sum of the last few IMFs can be regarded as the nonlinear and non-stationary trend. The nonlinear trend assessment of the monthly, local climatic changes is more robust and useful for attribution studies of the climatic changes.

Why do we want to remove the annual cycle? Although removing it is

not important to the linear trend when a data stream is sufficiently long, the removal of this cycle becomes necessary when the data stream is shorter than 30 or 20 years. For a short stream of data, if we want to get a long-term linear trend for the regional or global average SAT, the seasonal variation will influence the trend, because only the departures from the “climate normal” in a cold area, such as Alberta, and a hot area, such as Arizona, can be compared in the spatial average. The annual cycle defined by the TAC assumes that the climate system can settle down into a “normal” pattern in a sufficiently long time period, say, 30 years. This assumption is a good approximation for many applications, yet when nonlinearity and non-stationarity have to be carefully considered, the assumption becomes problematic, since even in a low dynamic nonlinear chaotic system, such as the Lorenz attractor, the signal will not ever repeat itself. When the annual cycle appears to be an intrinsic property of a climate system, an adaptive approach, rather than *a priori* functional method, should be used to find the annual cycle. Furthermore, a successful removal-operation of annual cycles should satisfy the following criteria: (1) the anomaly data should have no spectral power peaks around the period of a year, and (2) the annual cycle should carry some physically meaningful signal, such as El Niño. Criterion (1) reflects the variation of the solar irradiance, and criterion (2) reflects some major interactions between the dynamic components of a climate which is appropriate for a linear system, cannot satisfy even criterion (1), much less (2). The El Niño signal is left in the anomaly data. Thus, the EMD approach is a natural way to define the annual cycle, that is the NAC. However, the NAC is not immune from uncertainties due to different sifting parameters, while the TAC’s uncertainties are caused by different period of 30 years. The contents of this chapter are arranged as follows: Section 9.2 describes the method used in this research, regarding the end-point conditions, spectral representation, intermittence test, and stop criterion. Section 9.3 describes the daily and monthly data used in the analysis. Section 9.4 performs the time analysis of the NAC and analyzes the NAC’s robustness with respect to the length of the data stream, sifting parameters, and end-point algorithms. Section 9.5 performs the frequency analysis of both the NAC and the TAC. Section 9.6 contains conclusions and a discussion.

9.2. Analysis method and computational algorithms

The analysis method used here follows mainly that of Huang et al. (1998), a time domain analysis using EMD, and a frequency domain analysis using Hilbert and Fourier spectra. However, both the time and frequency analyses of Huang et al. (1998, 1999, 2003) have flexibilities in the computing details. Thus, we will specify our computing details to help others repeat our results.

The key problem in time-frequency analysis is to find the instantaneous frequency of a time series. Several ways are available to find instantaneous frequencies, including the wavelet transform (WT) and Hilbert transform (HT) (Mallat 1998, pp. 91–110; Boashash 1992a). The popular WT spectrum does not have a good resolution for nonlinear processes and hence cannot effectively discern the nonlinear signals. The EMD pre-processed signals allow high-resolution Hilbert-spectra even for nonlinear signals. However, the HT often cannot be directly applied to a time series. For example, the HT was applied to the analysis of nonlinear vibration motions (Feldman 1997), but the instantaneous frequency could not be found from the original time series. The possible problems were with the phase function, which are non-differentiable, or have unbounded derivatives, or are non-physical. H98 showed examples of these. The problem, thus, becomes how to find a differentiable and physically meaningful phase function so that the instantaneous frequencies are well defined, and their implications can be explained for the system where the data come from. Huang and his colleagues established their EMD-HAS (empirical mode decomposition-Hilbert spectral analysis), i.e., the HHT, based upon the belief that a physically meaningful mode should be self-coherent, namely, either quasi-periodic and quasi-symmetric, or quasi-monotone. Mathematically, phase functions and their derivatives are well defined for these types of functions by using HT. The quasi-periodic and quasi-symmetric functions, or quasi-monotone functions are the IMFs that satisfy the following two conditions:

- (1) In the temporal domain of data, the number of extrema and the number of zero-crossings must either be equal or differ at most by one, and
- (2) At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

With these requirements, an IMF oscillates in a narrow frequency band, a reflection of quasi-periodicity and nonlinearity. Of course, the non-constant frequency means non-stationarity. When an IMF $c(t)$ is found, its HT,

$\mathcal{H}[c(t)]$, can be found. Then

$$c(t) + i\mathcal{H}[c(t)] = a(t) \exp[i\omega(t)] \quad (9.1)$$

is well defined. For a given t , this equation yields a frequency $\omega(t)$ and an amplitude $a(t)$. The triplet (t, ω, a) forms the Hilbert spectral power described in Huang et al. (1998). The procedures for finding an IMF $c(t)$ can be found in Chapter 1 of this book. However, the conditions of the end points, stopping criterion and intermittence are still flexible up to certain subjective decisions. The details of our algorithm are described below.

The envelopes of local maxima and minima require the extrapolation of the original time series outside the two temporal end points: the first point and the last point of the time series. However, usually no physical laws can be easily used to make the extrapolation. Various spline extrapolation methods have been developed. We used two extrapolation methods to deal with the end effect: the signal extension approach by Wu and Huang (2004) and the extrema-prediction approach. The signal-extension approach includes two steps: the signal extension step and the signal-damping step. In the first step, the targeted time series, f_i , $i = 1, \dots, N$, is extended by using the anti-symmetric extension. In the second step, the extended signal is damped at the two ends. One can also use a one-step signal extension approach: a simple reflection extension. Apparently, many other signal-extension methods can be used. The physical nature of a problem is helpful in determining which methods are appropriate for the problem.

Due to the use of cubic splines, each EMD sifting step requires two extrema points outside of the data's temporal domain (possibly including the first end data point) and before the first extrema point. Similarly, each step requires two extrema points after the last extreme point. These four extrema must be predicted.

Many possible approaches can be used for predicting these extrema. For example, Coughlin and Tung (2003) extended the signal by adding a characteristic wave at each end of the sifted data stream. Our approach is as follows: Let us denote the first data point by $(0, x_0)$, the first two extrema by (t_1, x_1) and (t_2, x_2) , where $0 < t_1 < t_2$, and the predicted extrema by (t_{p1}, x_{p1}) and (t_{p2}, x_{p2}) . Setting $t_{p1} = \min(0, 2t_1 - t_2)$ and $t_{p2} = t_{p1} - (t_2 - t_1)$, then $x_{p1} = x_2$ if $t_{p1} < 0$; otherwise, $x_{p1} = x_0$. On the other hand, x_{p2} always equals x_1 . The two extrema after the last data point are predicted in a similar manner.

The stopping criterion, determining when the EMD sifting stops for an IMF mode, is another important component for the sifting process. A

stopping criterion through the standard deviation (SD) is proposed in H98. The sifting process for an IMF is stopped if this SD value is between 0.2 and 0.3. However, this SD value depends on the length of the time series. An alternative stop criterion is that the number of zero crossings and extrema remains the same for N successive sifting steps (Huang et al. 1999,2003). The choice of N is subjective, and many trials have to be made to find the best N suitable for a certain time series. The stopping criterion used here is that the sifting process stops when

$$\text{SD} = \frac{\sum_{i=1}^N m_i^2}{\sum_{i=1}^N h_i^2} < 10^{-5}, \quad (9.2)$$

where h_i , $i = 1, \dots, N$, are the data of the step of being stopped, and m_i , $i = 1, \dots, N$, are the mean of the envelope of maxima and that of the minima of the data h_i , $i = 1, \dots, N$.

Mode mixture, a phenomenon of different time scales mixed in a single IMF component, often occurs in a practical sifting process. The intermittency test is used to separate the intrinsic mode in the sifting process to prevent the modes from mixing. The intermittence criterion adopted in this study is a number M selected as the limit to the distances between each pair of the three successive maxima. If these distances are both greater than M , the points between the two zero crossings among the three extrema are assigned as zero. The M value ranges between 120 and 150 for the station SAT anomaly data in this research.

Our HT is computed by using the Fourier transform (Marple 1999). The signal $x(t)$ and its HT forms an analytic signal $z(t) = x(t) + i\mathcal{H}[x(t)]$. The N -point spectrum $Z(m)$ of the discrete analytic signal $z(t)$ is computed by

$$Z(m) = \begin{cases} 2X(m), & 0 < m < N/2, \\ X(m), & m = 0, N/2, \\ 0, & N/2 < m < N, \end{cases} \quad (9.3)$$

where $X(m)$ is the Fourier transform of $x(t)$. The discrete-time analytic signal $z[n]$ is then computed by using the N -point inverse discrete Fourier transform, and the imaginary part of $z[n]$ is $\mathcal{H}[x(t)]$.

The amplitude and phase of $z(t)$ are $a(t) = \sqrt{[x(t)]^2 + [\mathcal{H}[x(t)]]^2}$ and $\theta(t) = \arctan\{\mathcal{H}[x(t)]/x(t)\}$. The instantaneous frequency of a continuous signal is defined as (Boashash 1992a)

$$\omega(t) = \frac{1}{2\pi} \frac{d\theta(t)}{dt}. \quad (9.4)$$

Computationally, for a discrete-time analytic signal $z[n]$, we can compute the amplitude $a[n]$ and the angle $\theta[n]$. The discrete-time instantaneous frequency $\omega[n]$ is computed by using the central difference scheme

$$\omega[n] = \frac{1}{2\pi} \frac{\theta[n+1] - \theta[n-1]}{2T}, \quad (9.5)$$

where T is the time interval (Boashash 1992b). Thus, a given time n corresponds to a frequency $\omega[n]$ and an amplitude $a[n]$. Thus, on the (n, ω) -plane, each point corresponds to an amplitude that is a function of both time n and frequency ω , but the time n and frequency ω are not independent; rather, they are related by a function $\omega[n]$. The triplet $(n, \omega[n], a[n])$ determines a point in the three-dimensional space (n, ω, a) . For a given n , find a point $\omega[n]$, hence a point on the (n, ω) -plane. From this point, find the corresponding amplitude $a[n]$. One can find this $a[n]$ for all IMFs and hence for many amplitudes on the (n, ω) -plane. These amplitudes form the discrete Hilbert-spectra. They are then smoothed by a 9-point grid, moving average on the (n, ω) -plane to yield a series of smooth ridges in the three-dimensional space (n, ω, a) , and each ridge corresponds to an IMF (see Figs. 9.2a and 9.4b for two examples of Hilbert spectra). For more examples, see Huang et al. (1998).

9.3. Data

The dataset used in this study is the daily maximum surface air temperature data at 10 land stations of different latitude in USA and Canada in the time interval of 1 January 1946 to 31 December 2000. The data are from the Global Daily Climatology Network (GDCN) dataset of the US National Climatic Data Center (NCDC). The GDCN (v1.0) is a global dataset and contains different records of 32 857 locations around the world but covers mainly the northern hemisphere. The earliest and the latest dates of observation are 1 March 1840 and 30 November 2001, respectively, and most stations have missing records. All of the GDCN data have gone through an extensive set of quality-control procedures viz. simple datum checks and the statistical analysis of sets of observations to locate and identify potential outliers and/or erroneous data. However, the station data have not been homogenized. To demonstrate the NAC properties, we selected 10 stations in the United States and Canada that have few missing data and are distributed at four different latitude levels and from a coast to an area inland. Among the ten stations, three are distributed around the latitude zone 32.47°N , three around 41.15°N , three around 48.65°N , and one on

62.47°N. Among these stations, San Diego, Medford, and Victoria are close to the Pacific. They have few missing data in 55 years: from 1 January 1946 to 31 December 2000. Since the number of missing data is small with respect to the length of the entire data, the missing data are interpolated by using cubic spline fitting. The inventory of the ten stations is shown in Table 9.1.

The land stations at different latitudes were chosen to show the increase of the strength of the annual cycle as the latitude gets higher. It is also important to determine the annual cycle for the equatorial area where the incoming solar irradiance does not vary according to seasons. While no clear annual cycle is apparent on the equator, the climatology and climate anomalies are routinely computed for the SST data in both climate research and forecasting. To investigate the annual cycle of the equatorial area, we processed the SST data, not the SST anomaly data, over the Niño 3.4 region (5°N–5°S, 120°W–170°W) from January 1950 to March 2003. The data are from the National Centers for Environmental Prediction (NCEP) data repository. The El Niño and La Niña events defined by Trenberth (1997) are used to compare the ENSO signal carried by the NAC.

Table 9.1: Inventory of the ten stations used in this study.

Station name	Location	Number of missing data
Roscoe, Texas	32.45°N, 100.53°W	38
Shreveport, Louisiana	32.47°N, 93.82°W	1
San Diego, California	32.73°N, 117.17°W	43
Elko, Nevada	40.83°N, 115.78°W	7
Cheyenne, Wyoming	41.15°N, 104.80°W	4
Medford, Oregon	42.38°N, 122.87°W	12
Priest River, Idaho	48.35°N, 116.83°W	3
Victoria, B.C.	48.65°N, 123.43°W	4
Regina, Saskatchewan	50.43°N, 104.67°W	2
Yellowknife, NWT	62.47°N, 114.43°W	7

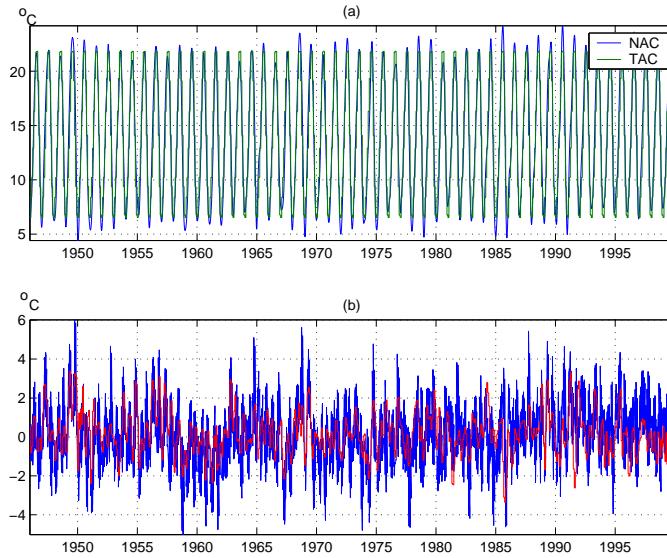


Figure 9.3: (a) The daily NAC (blue) and TAC (green) cycles of the Victoria station, Canada, and (b) the difference of the daily NAC minus the TAC (blue) and the red is the difference computed from the monthly data.

9.4. Time analysis

9.4.1. Examples of the TAC and the NAC

Figure 9.3 shows an example of the NAC and TAC of the daily maximum temperature of the Victoria station. The TAC is a strictly periodic cycle and computed for every month in a year. To compare the TAC with the NAC, the former is decomposed into daily resolution: all the days in a month are given the same climatology value. The NAC, although quasi-periodic, is automatically synchronized with the TAC by using the EMD method. Their highs and lows appear at the same time. This result is quite amazing since the EMD itself does not force a periodicity.

The difference of the NAC minus the TAC is shown in Fig. 9.3b (blue line). As shown in Fig. 9.1, the decomposed daily TAC has jumps from the last day of the previous month to the first day of the current month. These jumps cause a large difference. When using the monthly data to find the NAC, the difference of the NAC minus the monthly TAC is smaller (the red line in Fig. 9.3b). In both cases, there is no apparent drift of the differences away from zero.

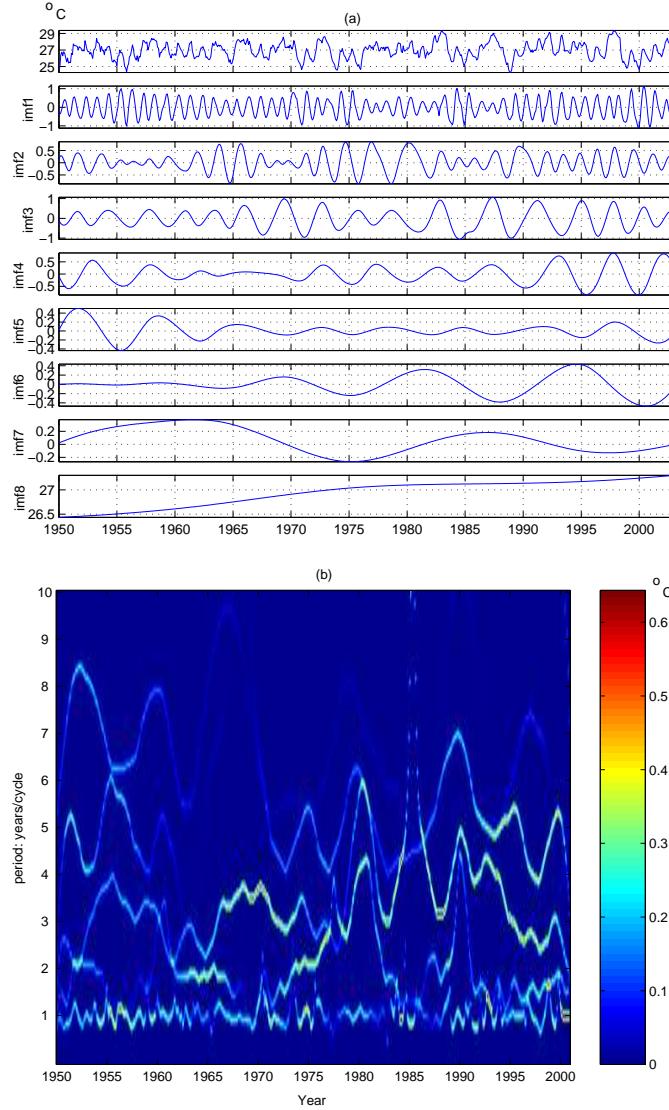


Figure 9.4: (a) Top. The Niño 3.4 data and their IMFs. The first frame is the Niño 3.4 SST data, the second frame is the NAC, and the remaining frames are the higher IMFs. (b) Bottom. Hilbert spectrum of the Niño 3.4 data according to all the IMFs.

The IMFs of the Niño 3.4 data are shown in Fig. 9.4. The first IMF is the NAC, and the second IMF corresponds to the biennial cycle. The third to fifth IMFs are the three- to seven-year cycles. The sixth IMF is the decadal cycle. The seventh IMF is the multi-decadal cycle. The last one is a warming trend. The strength of the NAC is found to have been enhanced during three strong La Niña events in 1955–1956, 1970–1971, and 1974–1975, and weakened during most El Niño events before 1990. Similar results have been found by using the wavelet transform (Wang 1994, Wang and Wang 1996). The influence of the ENSO on the annual cycle appears to have changed after 1990. The prolonged El Niño events from 1991 to 1995 did not change the amplitude of the NAC, but the NAC still appeared to be strong during the latest La Niña period from 1998 to 2000. The smaller amplitude of the NAC in the second panel of Fig. 9.4a corresponds to the smaller amplitude of the lowest smoothed bright line around the frequency of one cycle yr^{-1} in Fig. 9.4b. To clearly explain the Niño 3.4 SST data, their NAC, its NAC and TAC anomalies, we display them in Fig. 9.5. It is quite surprising that the EMD-derived NAC is quasi-periodic and oscillates in the frequency of one cycle per year, because the incoming solar radiation does not have an apparent annual cycle. Of course, the NAC's amplitude is small with a high-low difference within 2 degrees. We also calculated the TAC, and the anomalies with respect to the NAC and TAC. The anomalies are almost the same. The TAC anomalies have been commonly used to define El Niño events (Trenberth 1997). The comparison of the NAC's and the TAC's anomalies seems to suggest that the NAC anomalies can also be used to define El Niño. If so, the definition of El Niño is more natural and the nonlinearity of the ENSO dynamics is explicitly reflected in the statistical computing of the ENSO index. However, this conclusion needs further justification, and the EMD method for describing the ENSO dynamics requires a thorough investigation.

9.4.2. Temporal resolution of data

The NAC is in an annual time scale and is expected to be independent of the daily or monthly data resolution. This expectation was confirmed by analyzing the daily and monthly data of the ten stations. We performed sifting processing on the daily and monthly SAT data. The resulting daily NAC is averaged in each month, and the difference between the mean daily NAC and the monthly NAC is small. Figure 9.6a shows the IMFs derived from the daily maximal temperature of the Victoria station. The IMF11 is

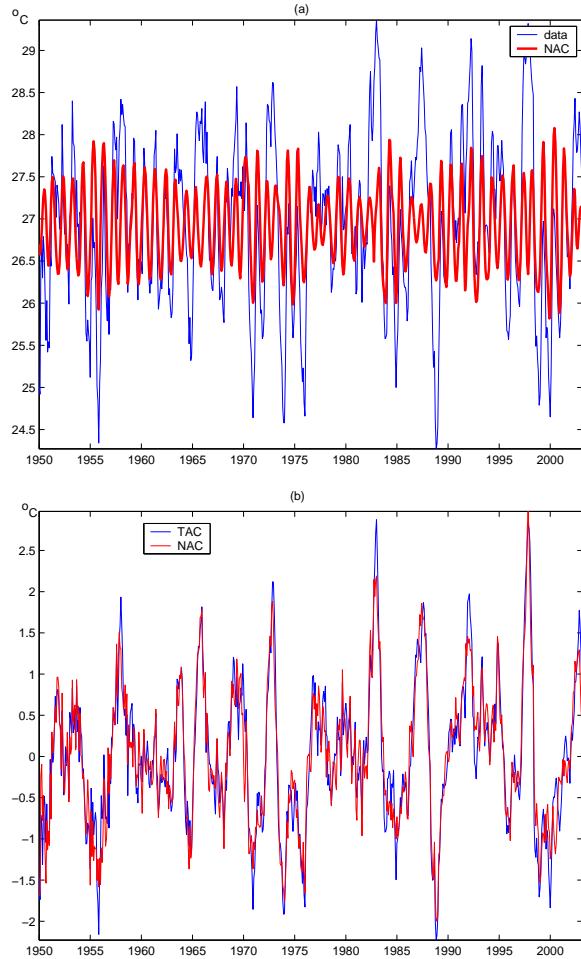


Figure 9.5: (a) Top. The Niño 3.4 SST data (blue) and the NAC (red line). Here, the mean of the Niño 3.4 SST data is added to the NAC to compare with the Niño 3.4 SST data. (b) Bottom. The anomaly of the TAC (blue) and that of the NAC (red). Here, the mean of the Niño 3.4 SST data is subtracted from the NAC anomaly to compare with the TAC anomaly.

the NAC, which is compared with the NAC derived from the monthly data (imf1 in Fig. 9.6b). The difference, shown in Fig. 9.6c, oscillates around zero and has no apparent drift to the positive or negative side. Occasionally, the magnitude of the difference reaches 2 degrees. This large difference may be

caused by some abnormal nonlinear interaction of the climate components. However, the exact cause is yet to be understood.

The above result does not mean that the daily resolution data are not important. On contrary, they are very important in reflecting the nonlinear interactions of the climate components in different time scales, but the interactions are reflected in the anomalies derived from the NAC.

9.4.3. Robustness of the EMD method

Another important aspect concerning our confidence in the NAC results is the robustness of the IMFs. Two questions are involved: (1) can the EMD lead to the separation of a given signal in a dataset? and (2) are the EMD-derived IMFs insensitive to the length of the data streams, the perturbation of the end conditions, the intermittency test, and the stop criterion? Huang et al. (2003) addressed some aspects of this problem. More examples are given below.

9.4.3.1. EMD separation of a known signal in a synthetic dataset

A dataset of 7 300($= 365 \times 20$) data points, uniformly spaced in time t , has three signal components and a noise component and is expressed by the following:

$$s[n] = -5 \cos\left(\frac{2\pi n}{365}\right) + \cos\left(\frac{2\pi n}{365 \times 5}\right) + (10^{-6}n)^2 + 10^{-4}n + w[n], \quad (9.6)$$

where $w[n]$ is a sequence of white noise with mean zero and variance $\sigma_w^2 = (1.5)^2$ and $n = 1, \dots, 20 \times 365$ (Fig. 9.7a). The dataset mimics a 20-yr climate data of two periodic components at the frequency of one cycle yr^{-1} and $\frac{1}{5}$ cycles yr^{-1} , a quadratic trend, and white noise. The annual cycle is much stronger than the 5-yr cycle. The nonlinear trend is very weak. This signal is contaminated by white noise. The signal-extension approach was applied to mitigate the end effect. The sifting results show the IMFs representing dyadically-filtered white noise (Flandrin et al. 2003, Wu and Huang 2004), the annual cycle component, and the low-frequency components, and the last two IMFs are trends (Fig. 9.7b). We used the sum of the low-frequency components to reconstruct the 5-yr cycle component in the signal and the sum of the last two IMFs to reconstruct the trend component. The results show that, as one expects, the EMD can accurately recover the strong annual component. The recovery of the weak 5-yr cycle

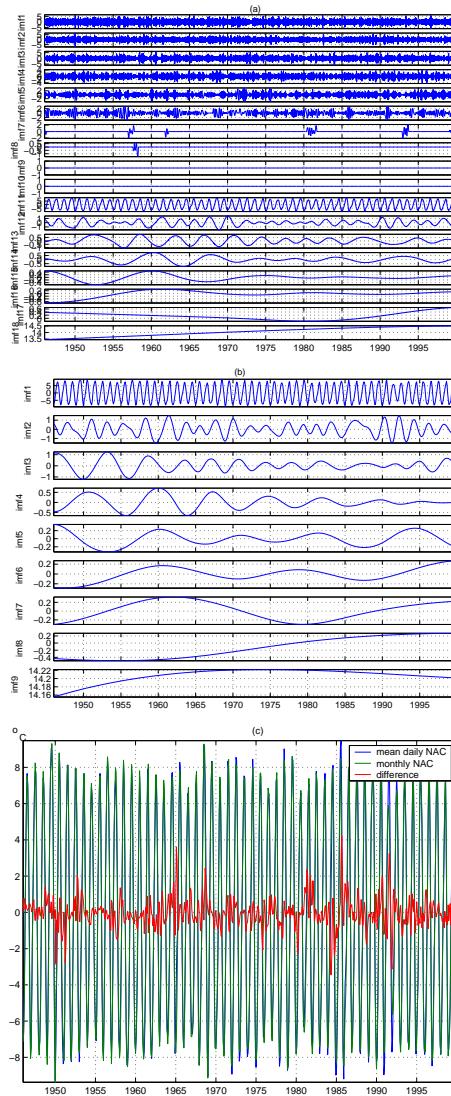


Figure 9.6: (a) Top. IMFs from Victoria station obtained by using the daily data. (b) Middle. IMFs from Victoria station obtained by using the monthly data. (c) Bottom. The monthly mean NAC from the daily data (blue line), the NAC from the monthly data (green line), and the difference (red line) of the mean NAC from the daily data and the NAC from the monthly data.

is also very good except in the neighborhoods of the end points. The end-point effects are carried to the trend recovery. The EMD-derived trends from 0 to 2300 and from 4500 to 7300 are apparently deviating away from the original trend. See Fig. 9.7c. These results call for further revisions of the end-point conditions.

9.4.3.2. Robustness with respect to data length

The full dataset of the daily maximum temperature data at Victoria station is from 1 January 1946 to 31 December 2000. In order to test the IMF's robustness with respect to the data length, we took a partial set of the data from January 1, 1946 to December 31, 1999 and calculated the IMFs. The NACs derived from the full dataset and the partial dataset are basically the same (Fig. 9.8).

9.4.3.3. Robustness with respect to end conditions

We used three end conditions in the EMD method, i.e., the signal extension approach using anti-symmetric extension, the signal extension approach using reflection extension, and the extrema-prediction approach. The results shown in Fig. 9.9 imply that the NAC does change much with respect to end-point conditions, and hence the NAC is robust with respect to these end-condition perturbations. Of course, the end conditions have to be reasonable. One can expect very different NACs if the end conditions are arbitrarily set.

9.5. Frequency analysis

9.5.1. Hilbert spectra of NAC

After performing the sifting process and intermittency test, the NAC is obtained for each station's data series. The NAC for the daily maximum temperature contains no mode mixture. The intermittency parameter M is 150–180 d in the sifting process of the daily temperature of the ten stations. The daily data have high-frequency oscillations before the NAC. After the NAC IMF, the rest of the IMFs manifest the biennial cycle with periods from 1.5 to 2.5 yr, the inter-annual cycle with periods from 4 to 6 yr, the decadal cycle with periods from 7 to 10 yr, the multi-decadal cycle with a period greater than 15 yr, and a trend, respectively.

To assess the effect of solar radiation and the heat capacity of the Earth's surface materials on the strength of the NAC, the marginal Hilbert spectra

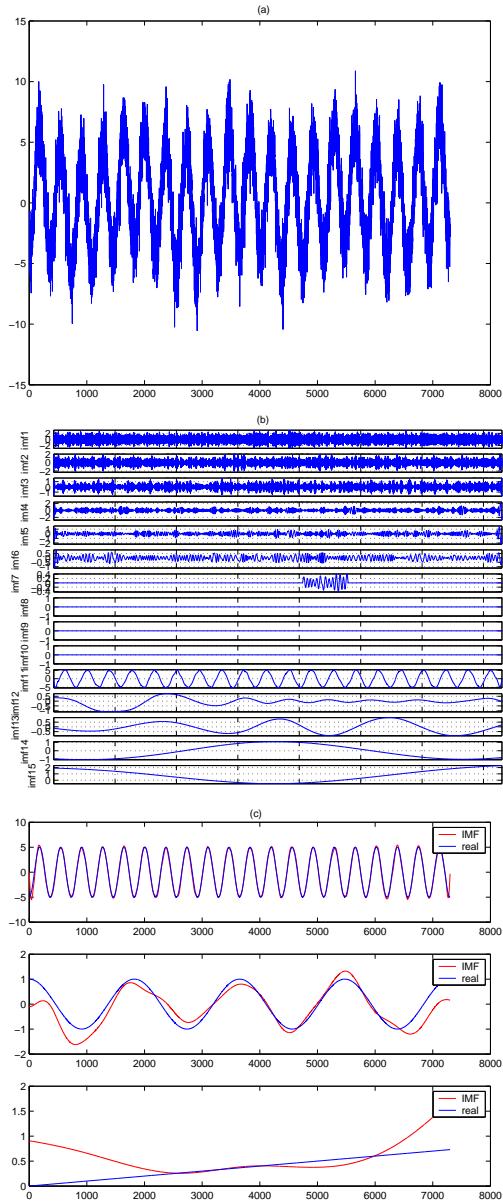


Figure 9.7: (a) Top. Synthesized data $s[n]$. (b) Middle. IMFs of $s[n]$ from the sifting process. (c) Bottom. Comparisons of the IMFs with the three signal components.

Annual Cycle of Daily Surface Air Temperature Data

235

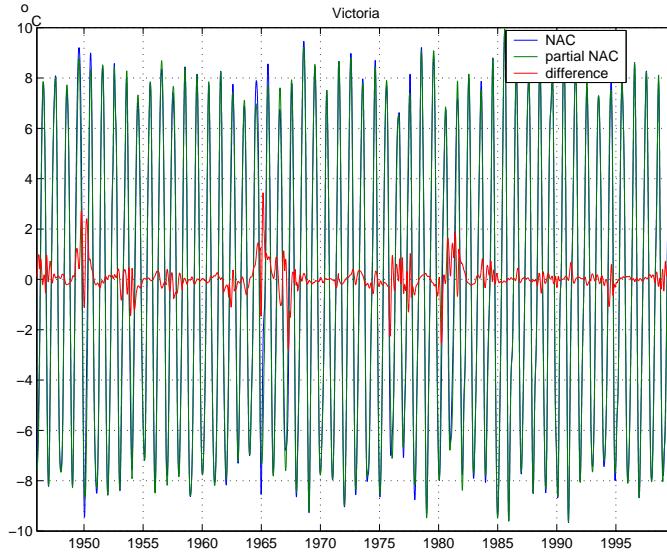


Figure 9.8: The NAC of the daily maximum SAT from 1 January 1946 to 31 December 2000 (blue line), and that of the SAT from 1 January 1946 to 31 December 1999 at the Victoria Station (green line), and the difference between the two NACs (red line).

(i.e., the temporal integration of the Hilbert spectra) for the data of the ten stations are compared in four groups: stations located at the similar latitudes but with variable distances from large water bodies, and stations located inland and at very different latitudes (Fig. 9.10). The results in Fig. 9.10 indicate the following. First, at the same latitude, the NAC strengths are comparable, but the San Diego station, Medford station, and Victoria station have less NAC spectrum power than the other stations located at about the same latitude (Fig. 9.10a–c). This finding reflects the effect of the heat capacity of the large water bodies: the larger the heat capacity, the weaker the annual cycle. Second, at different latitudes, the NAC's strength increases with latitude (Fig. 9.10d) because the higher the latitudes of the stations, the larger the seasonal variation in receiving the energy from solar radiation.

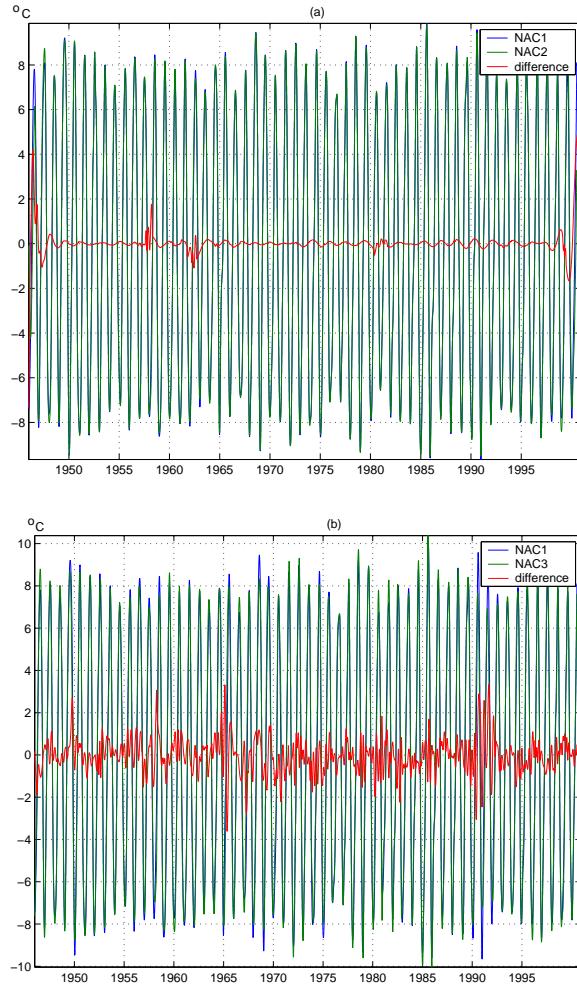


Figure 9.9: (a) Top. The NAC from the signal extension approach by using the reflection extension as end conditions (blue line), the NAC from the signal extension approach by using anti-symmetric extension (green line), and their difference (red line). (b) Bottom. The NAC from the signal extension approach by using reflection extension (blue line), the NAC from the extrema-prediction approach (green line), and their difference (red line).

9.5.2. Variances of anomalies with respect to the NAC and TAC

We compared the variances of anomalies with respect to the NAC and TAC at all stations. The results show that the mean monthly variances of the

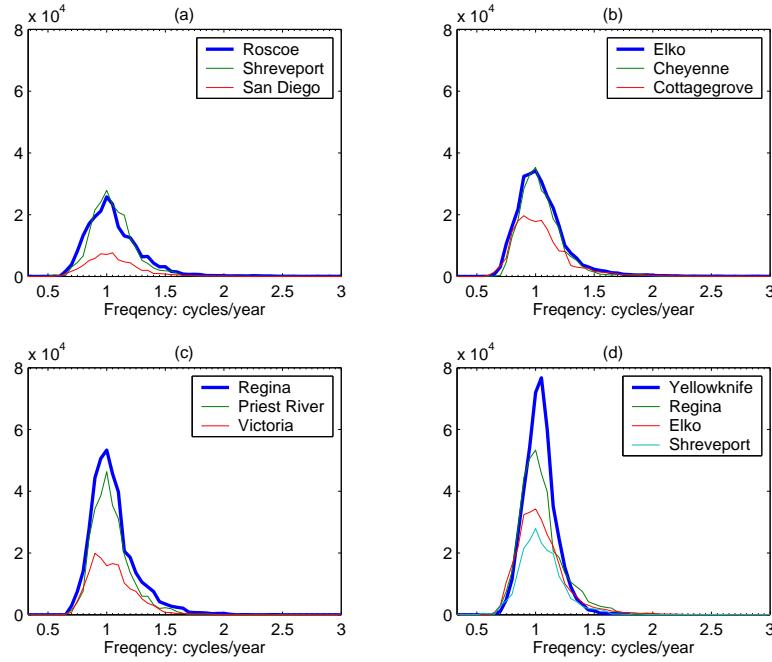


Figure 9.10: The marginal Hilbert spectra of the NAC for the daily maximum temperature at ten North America land stations. (a) Roscoe and Shreveport, San Diego. These three stations are located around the latitude zone 32.47°N . (b) Elko, Cheyenne, and Medford. These three stations are located around the latitude zone 41.15°N . (c) Regina, Priest River, and Victoria. These three stations are located around the latitude zone 48.65°N . (d) Yellowknife (the north-most), Regina, Cheyenne and Roscoe (the south-most). The ordinate's unit is $[\text{ }^\circ\text{C}][\text{yr}/365]$.

NAC anomalies are smaller than those of the TAC in most months. The overall variances of the NAC anomalies are smaller than those of the TAC anomalies (Fig. 9.11). These results seem to suggest that the EMD can separate the intrinsic modes of the SAT data better than the TAC filtering.

9.5.3. Spectral power of the anomalies with respect to the NAC and TAC

After the removal of the TAC, one expects that the annual cycle will be filtered out from a station's temperature data since the data's Fourier transform shows a strong peak at the frequency of one cycle yr^{-1} . However, one

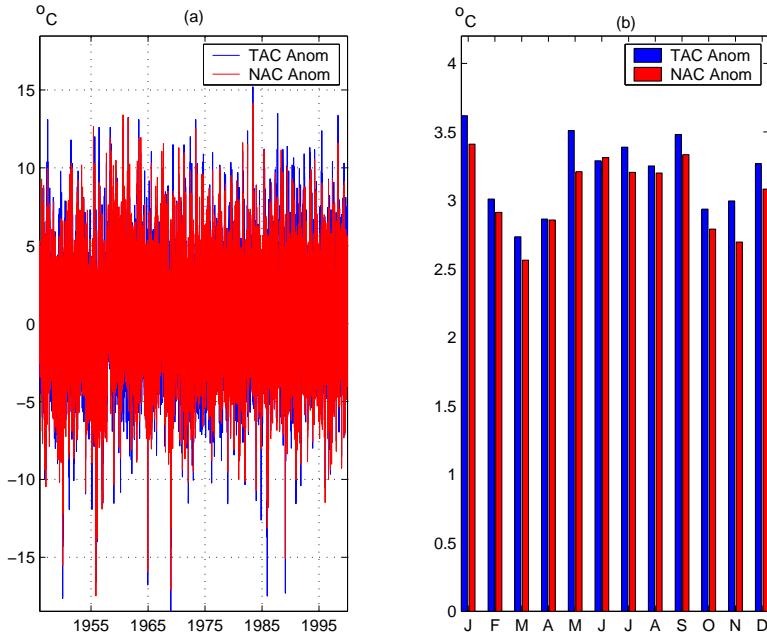


Figure 9.11: The anomalies of the TAC and NAC of the daily maximum SAT data at Victoria station (left) and the mean deviation of the anomalies for each month (right).

can detect some non-trivial spectral power in the frequency neighborhood of one cycle yr^{-1} . A good filter for the annual cycle should remove not only the monotonic spectral power of one cycle yr^{-1} , but also the spectral power in this neighborhood. The NAC can filter out the annual cycle more cleanly than the TAC, mainly because of the nonlinearity in the climate system, whose spectra are not purely discrete, but continuous. Thus, the anomalies about the TAC contain more spectral power around the annual cycle than those about the NAC. Figure 9.12 shows the Fourier spectra of the daily maximal SAT's NAC and TAC and the corresponding anomalies for four stations. Each station has two panels: the first one is for the TAC and NAC, and the second is for the NAC and TAC anomalies. The spectral power of the TAC anomalies is always larger than that of the NAC anomalies around the frequency of one cycle yr^{-1} . Thus, empirically speaking, the NAC is a better filter for the annual cycle than the TAC. Of course, the ordinates of the first and second panels are of different sizes. The spectral power in the anomalies is much smaller than that of the original data around the

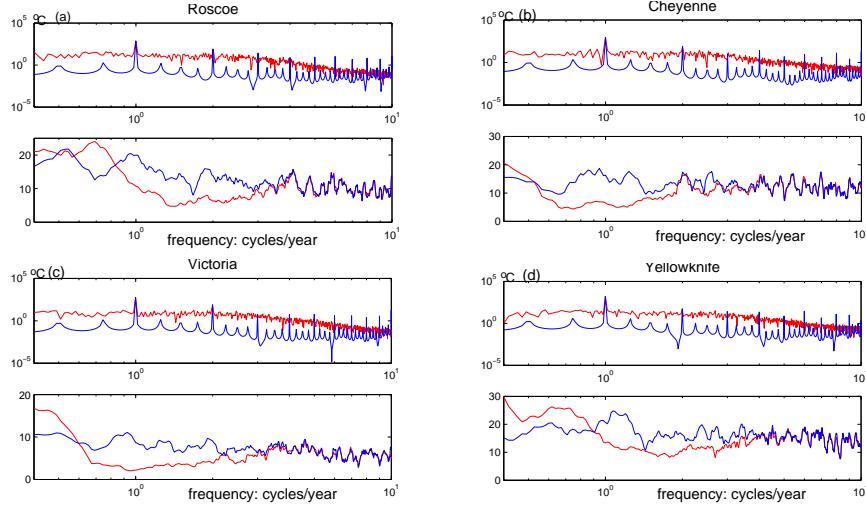


Figure 9.12: Fourier spectra of the NAC (red) and the TAC (upper frame), and those of the anomaly data of the NAC and the TAC (lower frame) at stations (a, top, left) Roscoe, (b, top, right) Cheyenne, (c, bottom, left) Victoria , and (d, bottom, right) Yellowknife. The Fourier spectra of the anomaly data are smoothed by a 10-point moving average.

frequency of one cycle yr^{-1} . However, due to the climate system's nonlinearity, the anomalies are very important and can transform non-negligible energy that produces an abnormal climate, such as El Niño.

The second panels of Figs. 9.12a–d indicate the relevant positions of the NAC and TAC anomalies' spectral lines: the TAC anomaly's spectra are above the NAC anomaly's spectra. To demonstrate the relative position quantitatively, we integrated the spectral power of both anomalies around the frequency of one cycle yr^{-1} . The results for the interval $[0.9, 1.1]$ cycles yr^{-1} are shown in Table 9.2. In general, the spectral power of the NAC anomalies in this interval is less than $\frac{1}{2}$ of that of the TAC anomalies. The table also includes the ratios of the TAC spectral power to the NAC spectral power. It was expected that these ratios would be less than 1. Among the ten stations considered, nine fulfill this expectation. The exception is the Medford station ($42.38^\circ\text{N}, 122.87^\circ\text{W}$), Oregon, the United States, where the ratio is 1.02. A nearby station, Cottage Grove ($42.72^\circ\text{N}, 123.05^\circ\text{W}$), Oregon was considered to check this abnormal case. The ratio is 0.97. Thus, in general, we can still conclude that the ratios of the TAC spectral power to the NAC spectral power around the frequency of one cycle yr^{-1} are less

than 1.0.

Table 9.2: The ratio of the energy around one cycle yr^{-1} of anomaly data with respect to the TAC and the NAC and that of the annual cycles at all stations.

Station name	$\frac{\int_{0.9}^{1.1} Anom_{TAC} ^2 d\omega}{\int_{0.9}^{1.1} Anom_{NAC} ^2 d\omega}$	$\frac{\int_{0.9}^{1.1} TAC ^2 d\omega}{\int_{0.9}^{1.1} NAC ^2 d\omega}$
Roscoe	3.6	0.93
Shreveport	6.2	0.96
San Diego	4.1	0.95
Elko	8.0	0.99
Cheyenne	5.6	0.98
Medford	1.4	1.02
Priest River	8.5	0.99
Victoria	18.7	0.94
Regina	17.3	0.96
Yellowknife	2.3	0.98

9.6. Conclusions and discussion

We have used the EMD method to derive the nonlinear non-stationary annual cycle for both the SAT at land stations in North America and the SST in the Niño 3.4 region. The nonlinear and non-stationary annual cycle compared with the commonly used thirty-year-mean annual cycle (TAC). The NAC allows daily resolution and is robust with respect to the length of a data stream and the end conditions of the EMD method. The EMD procedure used in the paper can accurately separate the annual cycle in a synthetic dataset composed of multiple harmonics, a nonlinear trend, and white noise. Comparison of the NAC and TAC in spectral space indicates that the NAC is a cleaner filter for the annual cycle than the TAC.

Many problems involving annual cycles are worth investigation. Three are listed here:

- (1) Can one use the NAC anomalies of the Niño 3.4 SST data or the buoy data in the same area to define the El Niño events?
- (2) If the NAC anomalies are calculated from the $5^\circ \times 5^\circ$ grid boxes for the NCAR/NCEP Reanalysis data and the EOFs are computed from these anomalies, then are the ENSO patterns distributed in the same way as those calculated from the commonly used TAC anomalies?
- (3) Discrete wavelet analysis can also yield modes of different scales. Is a systematic method available for comparing the modes derived from the EMD, wavelet analysis, and even some dynamical models?

Acknowledgements

This work was supported by the NOAA Office of Global Programs. Shen also thanks the US National Research Council for the Associateship award, MITACS (Mathematics of Information Technology and Complex Systems) for a research grant, and the Chinese Academy of Sciences for an Overseas Assessor's research grant and for the Well-Known Overseas Chinese Scholar award.

References

- Boashash, B., 1992a: Estimating and interpreting the instantaneous frequency of a signal. Part 1: Fundamentals. *Proc. IEEE*, **80**, 520–538.
- Boashash, B., 1992b: Estimating and interpreting the instantaneous frequency of a signal. Part 2: Algorithms and applications. *Proc. IEEE*, **80**, 539–568.
- Coughlin, K. T., and K. K. Tung, 2003: 11-year solar cycle in the lower stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.*, **34**, 323–329.
- Feldman, M., 1997: Non-linear free vibration identification via the Hilbert transform. *J. Sound Vib.*, **208**, 475–489.
- Flandrin, P., G. Rilling, and P. Gonçalvès, 2004: Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.*, **11**, 112–114.
- Gloersen, P., and N. E. Huang, 2003: Comparison of interannual intrinsic modes in hemispheric sea ice covers and other geophysical parameters. *IEEE Trans. Geosci. Remote Sens.*, **41**, 1062–1074.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of nonlinear water waves: The Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decompo-

- sition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., M. C. Wu, S. R. Long, S. S. P. Shen, N. H. Hsu, D. Xiong, W. Qu, and P. Gloersen, 2003: On the establishment of a confidence limit for the empirical mode decomposition and Hilbert spectral analysis, *Proc. R. Soc. London, Ser. A*, **459**, 2317–2345.
- Mallat, S., 1998: *A Wavelet Tour of Signal Processing*. Academic Press, 637 pp.
- Marple, S. L., 1999: Computing the discrete-time analytic signal via FFT. *IEEE Trans. Signal Process.*, **47**, 2600–2603.
- Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.*, **78**, 2771–2777.
- Wang, B., and Y. Wang, 1996: Temporal structure of the Southern Oscillation as revealed by waveform and wavelet analysis. *J. Climate*, **9**, 1586–1598.
- Wang, B., 1994: On the annual cycle in the tropical eastern-central Pacific. *J. Climate*, **7**, 1926–1942.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.

Samuel S. P. Shen

*Department of Mathematical and Statistical Sciences, University of Alberta,
Edmonton, Alberta T6G 2G1, Canada*
shen@ualberta.ca

Tingting Shu

*Department of Mathematical and Statistical Sciences, University of Alberta,
Edmonton, Alberta T6G 2G1, Canada*
ttshu@ualberta.ca

Norden E. Huang

*Goddard Institute for Data Analysis, Code 971, NASA/Goddard Space
Flight Center, Greenbelt, MD 20771, USA*
norden.e.huang@nasa.gov

Zhaohua Wu

*Center for Ocean-Land-Atmosphere Studies, 4041 Powder Mill Rd., Suite
302, Calverton, MD 20705-3106, USA*
zhwu@cola.iges.org

Gerald R. North

Department of Atmospheric Science, Texas A&M University, College Station, TX 77843, USA
north@csrp.tamu.edu

Thomas R. Karl

NOAA/National Climatic Data Center, Asheville, NC 28801, USA
tkarl@ncdc.noaa.gov

David R. Easterling

NOAA/National Climatic Data Center, Asheville, NC 28801, USA
deasterl@ncdc.noaa.gov

CHAPTER 10

HILBERT SPECTA OF NONLINEAR OCEAN WAVES

Paul A. Hwang, Norden E. Huang, David W. Wang and James M. Kaihatu

The Hilbert-Huang transform (HHT) analysis interprets wave nonlinearity in terms of frequency modulation instead of harmonic generation. The resulting spectrum contains much higher spectral energy at low frequency and sharper drop off at high frequency in comparison with the spectra derived from Fourier-based analysis methods. The high energy level in the low-frequency components of the Hilbert spectrum seems to be consistent with the rich group structure apparent in typical ocean-wave records. For wind-generated waves, the spectral level of the Fourier spectrum is about two orders of magnitude smaller than that of the Hilbert spectrum at the first subharmonic of the peak frequency. The mean frequency of the Fourier spectrum is 20% higher than that of the Hilbert spectrum. Furthermore, the frequency of the wave groups is also more likely to be properly identified in the Hilbert spectrum than in the Fourier spectrum. The implications for ocean engineering and air-sea interaction are discussed.

10.1. Introduction

Fourier-based spectral analysis methods have been widely used for studying random waves. One major weakness of these methods is the assumption of linear superposition of wave components. As a result, the energy of a nonlinear wave spreads into many harmonics, which are phase-coupled via the nonlinear dynamics inherent in ocean waves. In addition to the nonlinearity issue, Fourier spectral analysis should, strictly speaking, be used for periodic and stationary processes only, but wave propagation in the ocean is certainly neither stationary nor periodic.

Recently, Huang and his colleagues developed a new analysis technique, the HHT. Through analytical examples, they demonstrated the superior frequency and temporal resolutions of the HHT for analyzing nonstationary and nonlinear signals (e.g., Huang et al. 1998, 1999). A brief description

of the HHT analysis technique is presented in section 10.2. Using this analysis, the physical interpretation of nonlinearity is frequency modulation, which is fundamentally different from the commonly accepted concept associating nonlinearity with harmonic generation. Huang et al. argued that harmonic generation is caused by the perturbation method used in solving the nonlinear equation governing the physical processes; thus, the harmonics are produced by the mathematical tools used for the solution rather than being a true physical phenomenon.

In section 10.3, we investigate the HHT technique for ocean-wave analysis. The spectrum of wind-generated waves is presented as a case study. The Hilbert spectrum is compared with that obtained by using the Fourier-based techniques [wavelet and fast Fourier transform (FFT) algorithms]. The wavelet technique is based on Fourier spectral analysis but with adjustable frequency-dependent window functions, the mother wavelets, to provide temporal/spatial resolution for nonstationary signals (e.g., Shen and Mei 1993; Shen et al. 1994; Liu 2000; Massel 2001). As expected, the Fourier-based analysis interprets wave nonlinearity in terms of harmonic generation; thus, the spectral energy leaks into the higher-frequency components. The HHT interprets wave nonlinearity as frequency modulation, and the spectral energy remains near the base frequencies. As a result, these two sets of spectra have significantly different spectral characteristics.

The implications of these differences for ocean-wave spectra for ocean engineering applications and air-sea interaction processes, such as the characteristic frequency of the forcing waves and other statistical properties and the group structure of a wave field, are discussed in section 10.4. A summary is given in section 10.5.

10.2. The Hilbert-Huang spectral analysis

The Hilbert transformation was first used for water-wave analysis in the 1980s (e.g., Melville 1983; Bitner-Gregersen and Gran 1983; Hwang et al. 1989). A main application of the Hilbert analysis is to derive the local wavenumber in a spatial series or the instantaneous frequency in a time series. To use the Hilbert transformation, the proper preprocessing of the signal is critical. Large errors in the computed local frequency or wavenumber can occur when small wavelets are riding on longer waves or when a sharp change in the frequency occurs in the wave signal. A quantitative illustration of the riding wave problem has been discussed in greater detail by Huang et al. (1998) and sharp changes in the frequencies of oscillations

by Guillaume (2002) and will not be repeated here. The common approach in the past to alleviate such problems was to apply a low-pass filter to the signal prior to the Hilbert transformation. The determination of the low-pass frequency is somewhat subjective, and the signals removed may contain the information of nonlinearity, which is frequently the feature to be studied. Furthermore, a simple low-pass operation may not eliminate the riding wave problem.

The key ingredient in the HHT is empirical mode decomposition (EMD) designed to reposition the riding waves at the mean water level. Huang et al. (1998, 1999) have extensively discussed EMD. The main idea is to find the trend that can represent the mean local average so that riding waves can be identified. The EMD method uses the point-by-point average of the signal envelopes for the local mean. The difference between the original signal and the local mean represents a mode of the signal. The local mean may also contain riding waves, and the mode decomposition process continues until no riding waves exist in the local mean signal. The process is called “sifting” by Huang et al. (1998). From experience, even for very complicated random signals, a time series can usually be decomposed into a relatively small number of modes, $M < \log_2(N)$, where M denotes the number of modes and N is the number of data points. Each mode is free of riding waves; thus, the Hilbert transformation yields the accurate local frequency of the mode. The spectrum of the original signal can be obtained by the sum of the Hilbert spectra of all modes. Extensive tests have been carried out by Huang et al. (1998, 1999). Here, we present three cases to illustrate the superior resolutions of the Hilbert spectrum.

Case 1 is an example of the ideal sinusoidal oscillations of constant amplitude. The frequency of the first half of the signal is twice that of the second half (Fig. 10.1a). The spectra computed by the HHT and wavelet techniques are displayed in Figs. 10.1b and 10.1c, respectively. The Hilbert spectrum yields very precise frequency resolution and also high temporal resolution in identifying the sudden change of signal frequency at about the half-point of the time series. In comparison, the wavelet spectrum has only a mediocre temporal resolution of the frequency change. A serious leakage problem also occurs and the spectral energy of the simple oscillations spreads over a broad frequency range. Unless specified otherwise, the spectral contours plotted in the figures presented in this chapter are 3-dB (a factor of two) apart and cover a 30-dB range. For the example given in Fig. 10.1c, the 3-dB contour near the spectral peak extends to between 0.8 and 1.2 times of the spectral peak frequency for the wavelet spectrum. In

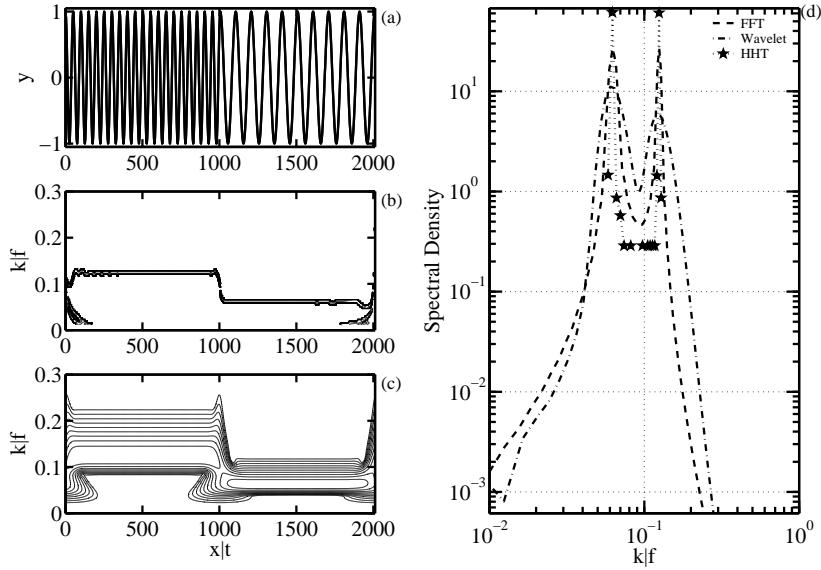


Figure 10.1: (a) Simple sinusoidal oscillations with the frequency or wavenumber of the first half double that of the second half, (b) the computed Hilbert spectrum, (c) the computed wavelet spectrum, and (d) a comparison of the spatially or temporally averaged spectra computed by the FFT, wavelet and HHT methods. The frequency (wavenumber) is normalized by the Nyquist value.

contrast, the Hilbert spectral energy is pretty much contained at the two spectral peak frequencies, and the spectral density of the next frequency bin is at least 10 dB down, as estimated by averaging the marginal spectrum over the whole time sequence (Fig. 10.1d). The spectral peak discrimination power can be quantified by the ratio between the spectral peaks and the neighboring spectral valleys. For the HHT, this number is 23 dB, the FFT analysis gives 16 dB, and wavelet 8 dB. Also noticeable in Fig. 10.1d is the unequal spectral densities at the two peaks of the wavelet spectrum, caused by the application of frequency-dependent windows in the wavelet analysis. The spectral density at the second-frequency component is only about 60 % of the spectral density at the first-frequency component.

Case 2 is a single cycle sinusoidal oscillation occurring at the middle of the otherwise quiescent signal stream (Fig. 10.2a). The period of a single cycle is 32 s. The precise temporal resolution of the HHT method is clearly demonstrated by the sharp rise and fall of the Hilbert spectrum coincident

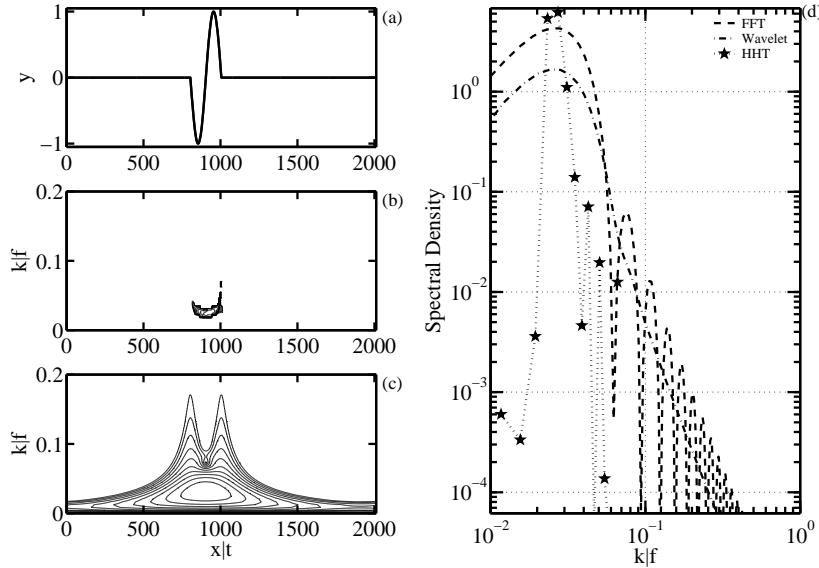


Figure 10.2: Same as Fig. 10.1 but for a transient sinusoidal wave of one cycle.

with the transient signal, as illustrated in Fig. 10.2b. By comparison, the wavelet spectrum is much more smeared both in the frequency and temporal resolutions (Fig. 10.2c). The marginal spectrum derived from the analysis shows a much sharper frequency definition of the single oscillating cycle as compared to that of the wavelet and Fourier spectra (Fig. 10.2d).

Case 3 is a sinusoidal function with periodically oscillating frequencies (Fig. 10.3a)

$$y(t) = a \sin[\omega t + \epsilon \sin(\omega t)]. \quad (10.1)$$

This equation is the exact solution for the nonlinear differential equation (Huang et al. 1998)

$$\frac{d^2y}{dt^2} + [\omega + \epsilon \omega \cos(\omega t)]^2 y - \sqrt{1 - y^2} \epsilon \omega^2 \sin(\omega t) = 0. \quad (10.2)$$

If the perturbation method is used to solve (10.2), the solution to the first order of ϵ is

$$y_1(t) = \cos(\omega t) - \epsilon \sin^2(\omega t) = \cos(\omega t) - \epsilon \left\{ \frac{1}{2} [1 - \cos(2\omega t)] \right\}. \quad (10.3)$$

The Hilbert spectrum (Fig. 10.3b) correctly reveals the nature of the oscillatory frequencies of the exact solution (10.1). In contrast, the wavelet spectrum (Fig. 10.3c) shows a dominant component at the base frequency and

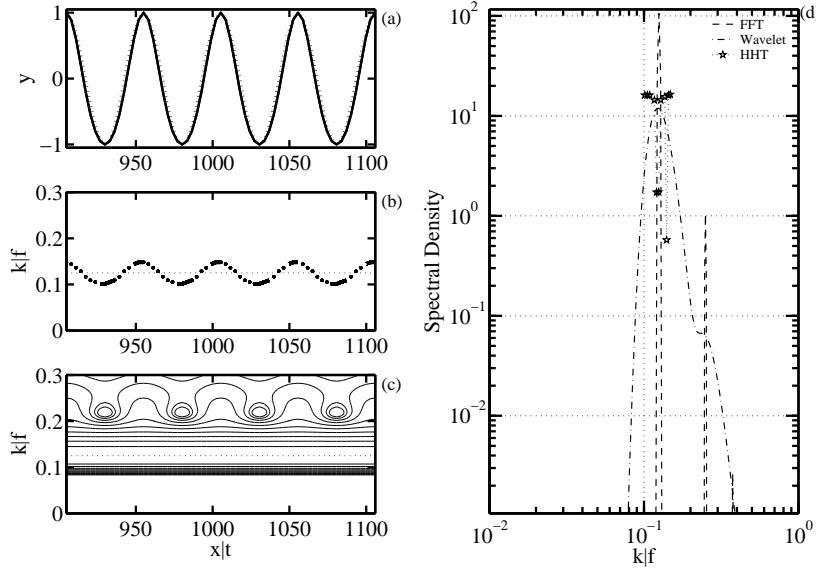


Figure 10.3: Same as Fig. 10.1 but for a signal with modulated frequency, $y(t) = a \cos[\omega t + \epsilon \sin(\omega t)]$. The unmodulated mean frequency is shown by the dotted lines in (b) and (c) for reference.

periodic oscillations of the second-harmonic component. In the marginal spectrum (Fig. 10.3d), the HHT analysis shows that the spectral energy is confined in the narrow frequency band surrounding the base frequency, which reflects the nature of the frequency modulation of the nonlinear system (10.2). Both FFT and wavelet analyses spread the spectral energy into higher frequencies as a result of harmonic generation by the Fourier decomposition of a nonlinear signal. The Fourier decomposition turns out to be a perfect match for representing the perturbation solutions such as (10.3). In this example, we have chosen $\epsilon = \frac{1}{5}$, so the spectral density of the second harmonic is $\frac{1}{100}$ of the primary component, which is accurately reproduced by the Fourier spectrum. The wavelet spectrum under-predicts the magnitude of the second harmonic by about 40%, similar to the results in Case 1 (Fig. 10.1d).

The three examples shown above illustrate the excellent temporal (spatial) and frequency (wavenumber) resolution of the HHT method for processing nonlinear and nonstationary signals. Many more demonstration cases are presented by Huang et al. (1998, 1999).

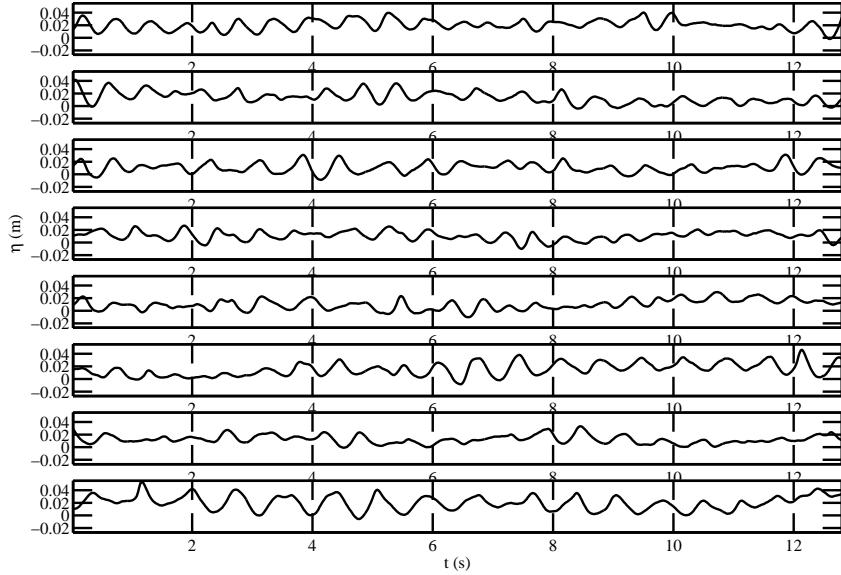


Figure 10.4: Examples of the time series of wind-generated waves used for spectral comparison. The average wind speed is about 5 m s^{-1} . The data are measured by a fast response wire gauge sampled at 50 Hz.

10.3. Spectrum of wind-generated waves

Here, we investigate the impact on the wind wave spectral functions by using different the spectral analysis techniques described in section 10.2. The wave record is acquired by using a fast-response wire gauge (Chapman and Monaldo 1991; Hwang and Wang 2004) during a test deployment in a canal (approximately 100 m wide and 400 m long). The data are sampled at 50 Hz, and the wind condition is light and variable with a range between 0 and 5 m s^{-1} . Figure 10.4 displays examples of the wave record showing the typical quasi-random time series of wind waves rich with group structure. The peak wave period is somewhat longer than 0.6 s, and one expects considerably lower frequency energy associated with the wave groups. Most groups have between 3 and 10 carrier waves.

Wave spectra are calculated by using the three methods using 20 segments of the wave record. Each segment contains 640 data points (12.8 s). For the Fourier spectrum, the mean of the 20 raw spectra is further running-averaged across nine frequency bins, resulting in the final spectrum with 360 degrees of freedom. For the Hilbert and wavelet spectra, each data

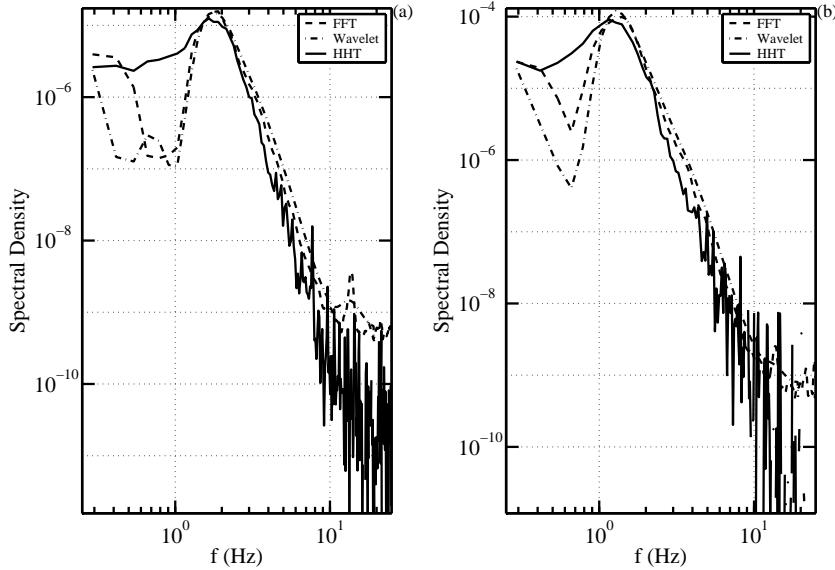


Figure 10.5: Spectra of wind-generated waves. The average wind speed is (a) 2 m s^{-1} and (b) 5 m s^{-1} .

segment produces a temporal variation of the wave spectrum. The average over time gives the marginal (one-dimensional) spectrum. The procedure is repeated for the 20 segments to obtain the final average Hilbert and wavelet frequency spectra. Figure 10.5 compares the spectra derived from these three different processing procedures. The similarities and differences of the spectral properties are described below.

The peak frequencies of the three spectra are close to 1.9 Hz for average wind speed $U = 2 \text{ m s}^{-1}$ (Fig. 10.5a) and 1.4 Hz at $U = 5 \text{ m s}^{-1}$ (Fig. 10.5b). A secondary peak near the frequency component with minimum phase speed, $f_m = 13.6 \text{ Hz}$, is very prominent in the Fourier spectrum. The secondary peak is still discernable in the wavelet spectrum, but is buried in the noise of the Hilbert spectrum.

Overall, the wavelet spectrum represents a smoothed version of the Fourier spectrum. The two Fourier-based spectra produce essentially similar results. The differences between the two spectra can be attributed to the degree of freedom, which is considerably higher in the wavelet analysis through the multiple windowing procedure.

The Hilbert spectrum differs from the other two Fourier-based spectra

in two main areas. The spectral density at the low-frequency portion is considerably higher in the Hilbert spectra, but near the peak and at higher frequencies, the reverse is true. As we have emphasized in the last section, this result is expected because of the HHT's and the Fourier-based methods' different interpretations of wave nonlinearity. Fourier-based techniques always decompose a nonlinear wave into its base frequency and higher harmonics; therefore, some spectral energy in the higher frequencies is leaked from their lower frequency subharmonics. There are higher-order spectral-processing methods, such as bispectrum and trispectrum, designed to restore the nonlinearity-contributed high-frequency spectral energy to the base frequency. It is fair to say that Fourier-based methods always overestimate the spectral level at frequencies higher than the spectral peak. The HHT interprets wave nonlinearity in terms of frequency modulation, and the spectral energy of a nonlinear wave remains at the neighborhood of the base frequency (see also Fig. 10.3d).

Figure 10.6a shows the ratio of the spectral densities, S_F/S_H and S_w/S_H , where subscripts F , H , and w denote the FFT, HHT, and wavelet, respectively. If we use the Hilbert spectrum as a reference, the spectral density derived from Fourier-based analysis is in general much lower (by a factor of about 80 at its minimal point) at low frequencies and much higher (by about a factor of 10) at high frequencies. We also processed the spectral difference normalized by the peak spectral density, $(S_F - S_H)/S_H(f_p)$ and $(S_w - S_H)/S_H(f_p)$. The results are shown in Fig. 10.6b. Significant differences in the spectral properties are obvious in the frequency region lower than the second harmonic of the peak frequency.

These differences in the frequency distribution of wave energy certainly impact ocean engineering designs. For example, the mean frequency as defined by the normalized first moment of the wave spectrum,

$$f_1 = \int f S(f) df / \int S(f) df, \quad (10.4)$$

is 13% lower in the Hilbert spectrum than that in the Fourier spectrum, and 21% lower than that in the wavelet spectrum for the examples shown in Fig. 10.5. These results clearly show that the Hilbert spectral level is considerably higher than the Fourier-based spectra in the lower frequency region. In the higher frequency portion, the Hilbert spectrum shows a steeper dropoff than the Fourier-based spectra. These differences in the wave spectral properties affect many engineering applications such as the frequency response of marine structures. Based on the interpretation of nonlinearity as frequency

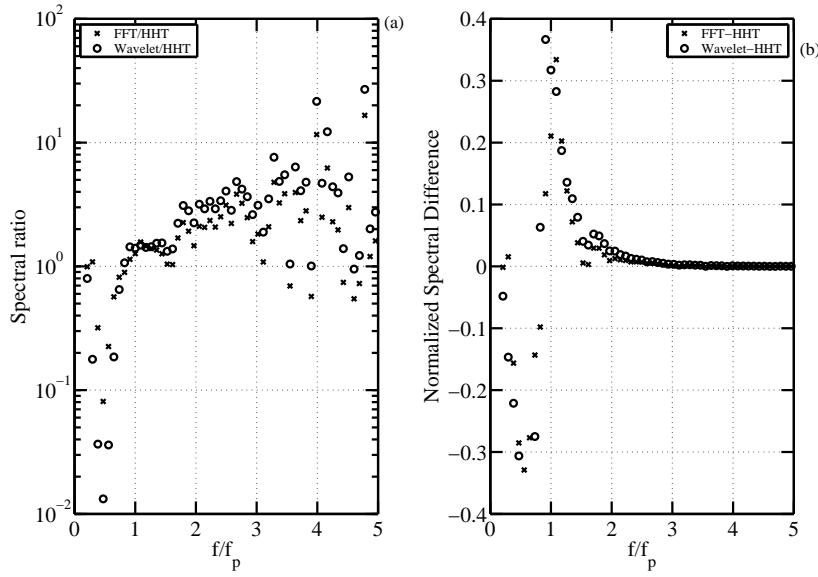


Figure 10.6: (a) The ratios of wavelet and Fourier spectra normalized by the Hilbert spectrum. (b) The difference spectra normalized by the Hilbert peak spectral density. The average wind speed is 5 m s^{-1} . Similar results are found for 2 m s^{-1} wind condition.

modulation, the mean frequency of the ocean-wave spectrum is about 1.2 times lower than that given by Fourier analysis.

10.4. Statistical properties and group structure

As a result of the spectral downshift in the Hilbert spectrum in comparison with the Fourier-based spectra as shown in Fig. 10.5, the characteristics of the spectral bandwidth may differ when different processing methods are used. The investigation of the spectral bandwidth is of some interest because it is closely related to the statistical properties of ocean waves. The dimensionless frequency bandwidth of a wave spectrum is defined (Longuet-Higgins 1952, 1980; Huang et al. 1983) by

$$\nu^2 = \frac{m_0 m_2 - m_1^2}{m_1^2} = \frac{f_2^2 - f_1^2}{f_1^2}, \quad (10.5)$$

where m_i is the i th moment of the wave spectrum, and f_i is the characteristic frequency defined by the i th moment of the wave spectrum. As we commented earlier, wave nonlinearity results in harmonic generation in

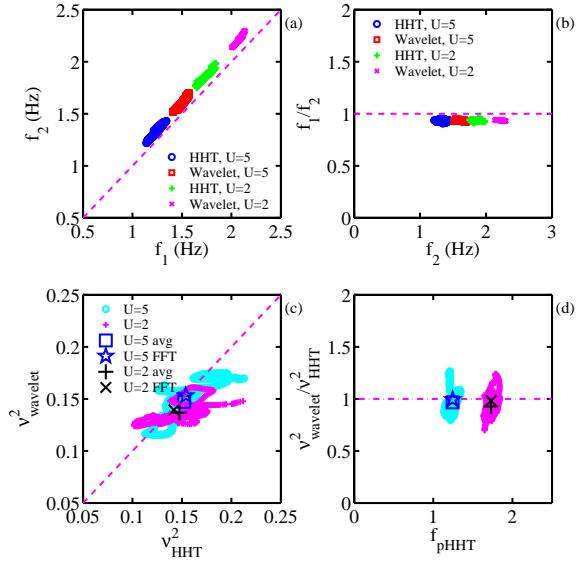


Figure 10.7: (a) Comparison of the characteristic frequencies f_2 and f_1 derived from the HHT and wavelet analyses, (b) ratio f_1/f_2 plotted as a function of f_2 , (c) dimensionless bandwidth ν^2 derived from the HHT and wavelet analyses, (d) ratio $\nu^2_{wavelet}/\nu^2_{HHT}$ plotted as a function of f_p , where f_p is the spectral peak frequency.

Fourier-based processing; therefore, for a wind wave spectrum, $f_2 > f_1$ is always true. With the Hilbert spectrum, the nonlinearity is seen as a frequency modulation near the spectral peak. It is less clear whether $f_2 > f_1$ would hold true. We apply the HHT and wavelet analysis to two wind wave data sets with nominal wind speeds of 2 and 5 m s⁻¹. The results are displayed in Figs. 10.7a and 10.7b. If we use either processing method, $f_2 > f_1$ for both cases, and the ratios of f_1/f_2 from both processing methods are quite compatible. The calculated ν^2 from the two methods and their ratio are shown in Figs. 10.7c and 10.7d, and also show good agreement. In Figs. 10.7c and 10.7d, the bandwidth computed from the ensemble-averaged spectra from the FFT, wavelet and HHT are also displayed. The results from the HHT and FFT are almost identical. This finding suggests that the study of the statistical properties of ocean waves using wave spectral functions is probably not affected by using the Hilbert spectral technique.

As displayed in Fig. 10.4, ocean waves in nature almost always exhibit a group structure, strongly suggesting the presence of energy in the low-

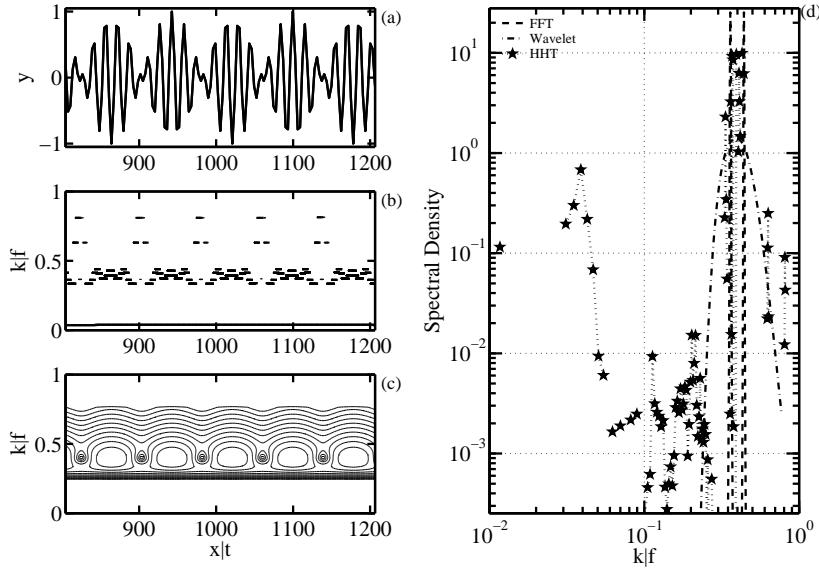


Figure 10.8: (a) Time/space series of an amplitude modulation signal showing the group structure. (b) The temporal/spatial Hilbert spectrum, (c) the temporal/spatial wavelet spectrum, and (d) the average Hilbert, wavelet, and Fourier spectra. The carrier period is $5\Delta t$, and the modulation period is $50\Delta t$, where Δt is the sampling time interval.

frequency components. The Hilbert transform has been widely used in the investigation of the envelope and group structure of surface waves (Melville 1983; Bitner-Gregersen and Gran 1983; Hwang et al. 1989; Veltchva 2002; Veltchva et al. 2003).

The low-frequency energy can be produced, for example, by the interaction of two wave components with a small frequency difference (frequency beating). When FFT-based techniques are used, such low-frequency energy cannot be easily detected. As an illustrative example, let us examine the following simple case of amplitude modulation (AM) of carrier waves with an angular frequency of ω_0 ,

$$y = \cos(\delta\omega t) \cos(\omega_0 t) = \cos\left[\left(\omega_0 + \frac{\delta\omega}{2}\right)t\right] + \cos\left[\left(\omega_0 - \frac{\delta\omega}{2}\right)t\right]. \quad (10.6)$$

Because of the mathematical equivalence of the two expressions in the right side of (10.6), Fourier decomposition interprets the signal as two sinusoidal oscillations of equal amplitude with frequencies at $\omega_0 + \delta\omega/2$ and $\omega_0 - \delta\omega/2$. The HHT, with its mode decomposition nature, will identify the

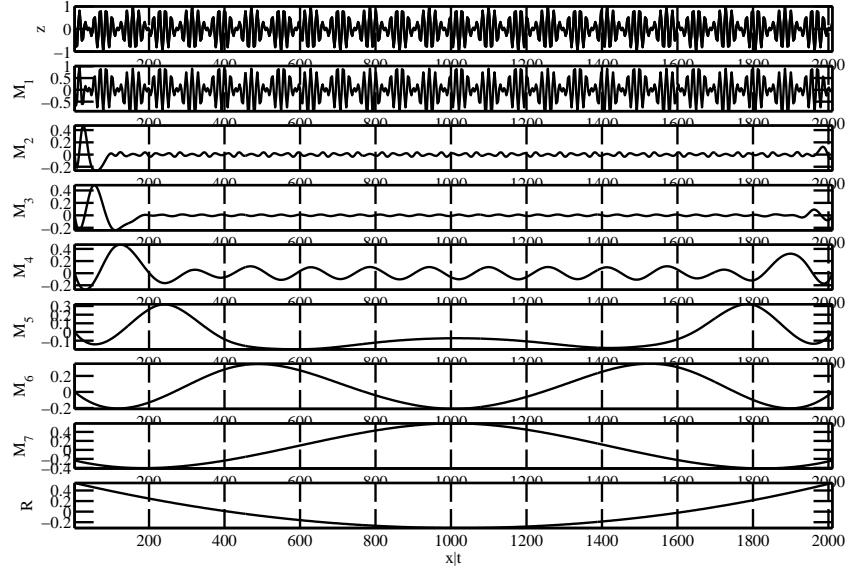


Figure 10.9: EMD processing of the signal shown in Fig. 10.8(a). The top panel is the original signal, M_1 to M_7 , are the modes and R is the remainder.

low-frequency component. Figure 10.8a shows the AM time series with a normalized (by the Nyquist value) carried wave frequency of 0.4 and a modulation frequency one tenth of the carrier frequency. The spectra produced by three difference analysis techniques are shown in Fig. 10.8d. The FFT processing produces two spectral components at 0.36 and 0.44. The small frequency difference is usually difficult for the wavelet processing to distinguish because of the relatively short dynamic windows used. In comparison, the HHT is able to identify both the carrier and the modulation frequency components at the correct frequencies of 0.4 and 0.04. The Hilbert spectrum, however, also contains energy at subharmonic and superharmonic frequencies, although the energy levels are relatively small: about –20 dB or smaller in comparison with the peak energy level. These sub- and superharmonic components occur at the nodal points of the time series (Fig. 10.8b) and suggest that they are the results of a low signal-to-noise ratio.

The ability of the HHT to identify the modulation (group) frequency, as mentioned earlier, is attributed to the nature of mode decomposition. Figure 10.9 plots the time series of the original signal and the decomposed empirical modes. The modulation frequency component shows up in mode 4

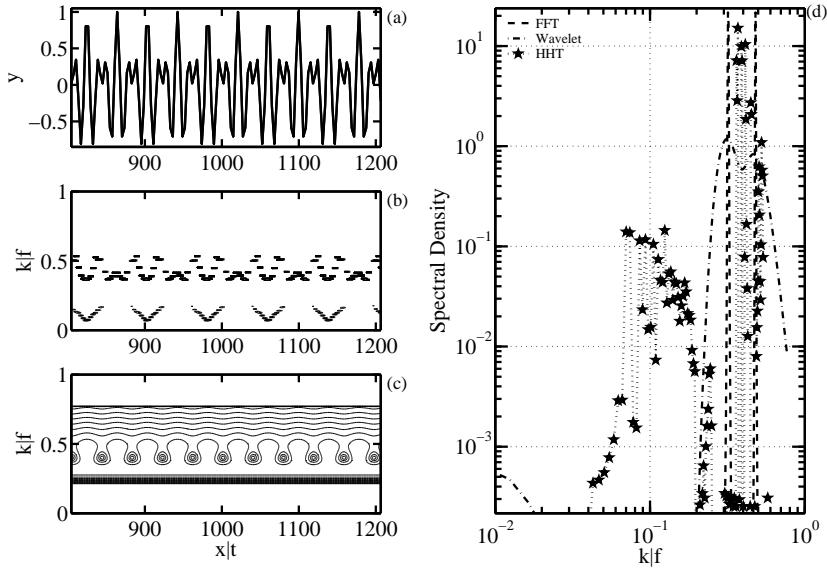


Figure 10.10: Same as Fig. 10.8, but the carrier period is $5\Delta t$, and the modulation period is $25\Delta t$, where Δt is the sampling time interval.

for the present example. From our experience, artificial subharmonics usually occur in the presence of groupiness in the signal. The problem of artificial subharmonic components is especially serious when the group structure is weak. For example, Fig. 10.10 plots a case in which the modulation frequency is 5 times lower than the carrier frequency. The spectral densities at the frequency components between the carrier and modulation frequencies are comparable to that at the modulation frequency (Fig. 10.10d). Based on our experience, it is not unusual that the modulation frequency component disappears whereas the Hilbert spectrum invariably shows low-frequency energy while the Fourier and wavelet spectra show energy only near the carrier frequency.

10.5. Summary

Analyzing nonlinear and nonstationary signals remains a very challenging task. Presently, most methods developed to deal with nonstationarity are based on the concept of Fourier decomposition; therefore, all the shortcomings associated with Fourier transformation are inherent in those methods also. The recent introduction of empirical mode decomposition by Huang et

al. (1998, 1999) represents a fundamentally different approach for decomposing nonlinear and nonstationary signals. The associated spectral analysis (HHT) provides superior spatial (temporal) and wavenumber (frequency) resolution for handling nonstationarity and nonlinearity (section 10.2). The Hilbert spectrum also results in a considerably different interpretation of nonlinearity (frequency modulation). Applying the technique to the problems of wind-generated ocean waves, we found that the spectral function derived from the HHT is markedly different from those obtained by using the Fourier-based techniques. The difference in the resulting spectral functions is attributed to the interpretation of nonlinearity. The Fourier techniques decompose a nonlinear signal into sinusoidal harmonics; therefore, some of the spectral energy at the base frequency is distributed to the higher-frequency components. The HHT interprets nonlinearity in terms of frequency modulation, and the spectral energy remains in the neighborhood of the base frequency. This difference results in a considerably higher spectral energy at lower frequencies and sharper dropoff at higher frequencies in the Hilbert spectrum in comparison with the Fourier-based spectra. The mean frequency computed from the Hilbert spectrum is 13 to 21% lower than those derived from Fourier-based spectra. On the other hand, for the basic statistical measures such as the spectral bandwidth or spectral moments, the results computed from the Hilbert and Fourier spectra are similar, suggesting that the study of the statistical properties of ocean waves is probably not affected by using the Hilbert spectral technique. Finally, wave group structures in the time series of surface displacement represent low-frequency oscillations. When processing the results from Fourier analysis, however, one may interpret the wave group as the interaction of two wave components with a slight difference in their frequencies, and the spectral energy does not appear at the observed frequency of wave groups. The Hilbert analysis places the spectral energy of the wave group at the correct frequency band if it is distinct enough from the carrier frequency.

Acknowledgements

This work was supported by the Office of Naval Research (Naval Research Laboratory Program Elements N61153 and N62435, PAH, DWW, and JMK) and the NASA RTOP program on micro-scale ocean dynamics (NEH). This chapter is NRL contribution BC/7330-03-0002.

References

- Bitner-Gregersen, E. M., and S. Gran, 1983: Local properties of sea waves derived from a wave record. *Appl. Ocean Res.*, **5**, 210–214.
- Chapman, R. D., and F. M. Monaldo, 1991: The APL wave gauge system. Rep. S1R-91U-041, Applied Physics Laboratory, The Johns Hopkins University, Baltimore, MD, 25 pp.
- Guillaume, D. W., 2002: A comparison of peak frequency-time plots produced with Hilbert and wavelet transforms. *Rev. Sci. Instrum.*, **73**, 98–101.
- Huang, N. E., S. R. Long, C. C. Tung, Y. Yuan, and L. F. Bliven, 1983: A non-Gaussian statistical model for surface elevation of nonlinear random wave fields. *J. Geophys. Res.*, **85**, 7597–7606.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Hwang, P. A., and D. W. Wang, 2004: Field measurements of duration limited growth of wind-generated ocean surface waves at young stage of development. *J. Phys. Oceanogr.*, **34**, 2316–2326.
- Hwang, P. A., D. Xu, and J. Wu, 1989: Breaking of wind-generated waves: Measurements and characteristics. *J. Fluid Mech.*, **202**, 177–200.
- Liu, P. C., 2000: Is the wind wave frequency spectrum outdated? *Ocean Eng.*, **27**, 577–588.
- Longuet-Higgins, M. S., 1952: On the statistical distribution of the heights of sea waves. *J. Mar. Res.*, **11**, 245–266.
- Longuet-Higgins, M. S., 1980: On the distribution of the heights of sea waves: Some effects of nonlinearity and finite band width. *J. Geophys. Res.*, **85**, 1519–1523.
- Massel, S. R., 2001: Wavelet analysis for processing of ocean surface wave records. *Ocean Eng.*, **28**, 957–987.
- Melville, W. K., 1983: Wave modulation and breakdown. *J. Fluid Mech.*, **128**, 489–506.
- Shen, Z., and L. Mei, 1993: Equilibrium spectra of water waves forced by intermittent wind turbulence. *J. Phys. Oceanogr.*, **23**, 2019–2026.
- Shen, Z., W. Wang, and L. Mei, 1994: Finestructure of wind waves analyzed with wavelet transform. *J. Phys. Oceanogr.*, **24**, 1085–1094.
- Veltcheva, A. D., 2002: Wave and group transformation by a Hilbert spectrum. *Coast. Eng. J.*, **44**, 283–300.

Veltcheva, A. D., P. Cavaco, and C. Guedes Soares, 2003: Comparison of methods for calculation of the wave envelope. *Ocean Eng.*, **30**, 937–948.

Paul A. Hwang

*Oceanographic Division, Naval Research Laboratory, Stennis Space Center,
MS 39529, USA*

paul.hwang@nrlssc.navy.mil

Norden E. Huang

*Goddard Institute for Data Analysis, Code 614.2, NASA/Goddard Space
Flight Center, Greenbelt, MD 20771, USA*

norden.e.huang@nasa.gov

David W. Wang

*Oceanographic Division, Naval Research Laboratory, Stennis Space Center,
MS 39529, USA*

david.wang@nrlssc.navy.mil

James M. Kaihatu

*Oceanographic Division, Naval Research Laboratory, Stennis Space Center,
MS 39529, USA*

james.kaihatu@nrlssc.navy.mil

CHAPTER 11

EMD AND INSTANTANEOUS PHASE DETECTION OF STRUCTURAL DAMAGE

Liming W. Salvino, Darryll J. Pines, Michael Todd and Jonathan M. Nichols

In this chapter, a new structural health-monitoring and damage-detection method is presented. A general time-frequency data analysis technique (empirical mode decomposition and the Hilbert-Huang spectrum) in conjunction with a wave-mechanics-based concept is developed to provide a diagnostic tool for detecting and interpreting adverse changes in a structure. Sets of simple basis function components, known as intrinsic mode functions (IMF), are extracted adaptively from the measured structural response time series data. These IMFs are amplitude- and phase-modulated signals and are used to define the instantaneous phases of structural waves. The state of a structure is evaluated, and damage is identified based on these instantaneous phase features. Furthermore, fundamental relationships are developed connecting the instantaneous phases to a local physics-based structural representation in order to infer damage in terms of physical parameters, such as structural mass, stiffness, and damping. Damage-detection applications are investigated by using numerical simulations and a variety of laboratory experiments with simple structures. Several different types of excitation mechanisms are used for dynamic input to the structures. The time series output of the structural response is then analyzed by using the new method. The instantaneous phase relationships are extracted and examined for changes which may have occurred due to damage. These results are compared to those from other newly developed detection methods, such as an algorithm based on the geometric properties of a chaotic attractor. The studies presented here show that our method, without linear-system or stationary-process assumptions, can identify and locate structural damage and permit the further development of a reliable real-time structural health-monitoring and damage-detection system.

11.1. Introduction to structural health monitoring

The nondestructive evaluation and monitoring of the health of a large and complex structure and its components are of great interest to the aerospace,

defense, civil, mechanical, and marine engineering communities. One of the key goals of a comprehensive structural health monitoring (SHM) system is the ability to identify a structural anomaly at the earliest possible stage and to evaluate the behavior of the structure as damage accumulates. Almost all damage-detection methods currently in use are either visual or localized experimental methods, such as acoustic or ultrasonic techniques. These local detection methods are not only time-consuming to perform, but also require the general location of the damage to be known in advance and the structure to be taken out of service for inspection. The need for an effective and robust diagnostic method that can be applied to complex structures has led to the development of global detection methods which examine changes in the measured vibration characteristics of the structure. The basic premise of global damage-identification methods is that damage will alter the stiffness, mass, or damping properties of a system, and that this change, in turn, will alter the measured dynamic response of the structure.

The traditional approach of a global method is to examine changes in modal properties or changes in quantities derived from modal properties. The central idea of modal methods is that a change in a structure, such as damage, should imply a change in the modal parameters of the structure, such as frequencies, damping, mode shapes, and mode shape curvatures. These techniques are usually based on classical linear theory, and most studies assume that the structure can be modeled as a linear system before and after damage. In general, these modal techniques have not been widely accepted for practical monitoring applications due to their insensitivity to localized damage, linearity assumptions, boundary conditions, sensitivity to sensor locations, and other environmental effects. To overcome these fundamental limitations of the modal method, an extensive amount of research in recent years has been associated with the development of other diagnostic and prognostic methods and algorithms. A review of these many techniques and their applications can be found in Doebling et al. (1996, 1998) and Chang (2001, 2003).

The use of an empirical or data-driven approach, intended to assess the state of a structure from measured structural responses, is rapidly developing and increasing in popularity. Rather than fitting the data to a specific analytical model of a structure whose true complexity is extremely difficult to capture mathematically, one uses measured data, collected from a structural response at some reference state, as an undamaged baseline or empirical model of the structure. When damage occurs and progresses, the dynamic behavior of the structure will differ from that of the base-

line empirical model. Certain features can then be chosen to measure this discrepancy, detect damage, and evaluate the structural state. For example, the damage-detection method introduced here uses the instantaneous phase of measured structural waveforms as indicators of damage. Another example of a recently developed method uses a prediction error, a concept taken from the nonlinear dynamics community, as the damage indicator (see subsection 11.4.5). Direct comparisons between some of these methods will be given and further discussed in section 11.4.

Another popular approach is a wavelet-based damage-detection method. Wavelet analysis permits vibration signals from structures to be decomposed into fundamental basis functions that are used to characterize the vibration response. By tracking changes in these fundamental basis functions, structural defects can be detected by inspecting the time-frequency properties of the vibration signal. However, a disadvantage of current wavelet-based diagnostic algorithms is the use of currently available wavelet dictionaries, which are not necessarily appropriate for analyzing the behavior of a particular structural system. In addition, wavelet methods, designed to accommodate the transient nature of the data, may not work well if the underlying physical process cannot be approximated by linear processes. Finally, another time domain method is auto-regressive (AR) modeling, which also assumes that a measured response coming from a structure can be accurately represented by a linear model. The AR model and an example of damage detection using the AR model are presented in subsection 11.4.4.

The effects of damage on a structure can be classified as linear or nonlinear. Nonlinear damage, defined as the case when the initially linear-elastic structure behaves in a nonlinear manner after the damage has been introduced, is ubiquitous in nature. A robust damage-detection method must be applicable to, or at least sensitive to, both these general types of damage. Because of the inherent difficulty of analyzing nonlinear processes, the vast majority of past and even most present works address only the problem of linear damage detection.

In this chapter, we discuss a fundamentally new data-driven approach where damage features are identified and evaluated based on the instantaneous phases of structural waves. This approach is implemented by means of the empirical mode decomposition (EMD) and Hilbert-Huang spectrum technique (Huang et al. 1998). Instantaneous phases are defined by utilizing the unique characteristics of intrinsic mode functions (IMFs), which are amplitude- and phase-modulated basis functions extracted directly from the

measured signal. In addition, this IMF function representation is coupled with a local physics-based model of structures in order to infer damage in terms of physical parameters, such as structural mass, stiffness, and damping. Based on the intrinsic nature of the EMD method, this approach does not assume a linear system or a stationary process for the measured data.

The success of SHM and damage-detection methodologies involves many important issues such as excitation and measurement considerations, including the selection of types and locations of sensors and excitations. The examples used in this chapter to demonstrate EMD and instantaneous phase damage detection include several different types of excitation mechanisms such as a broadband noise or random excitations, deterministic chaos signals, as well as realistic signals recorded during an earthquake event. The results of these studies show that our method and algorithm perform well without any knowledge of excitation sources. The detailed description of the phase detection method and the results for several examples will be given in sections 11.3 and 11.4. We begin the discussion in the next section by introducing the concept of “instantaneous phase” and the method used to extract it from a measured time series in order to describe the dynamics of a structural system.

11.2. Instantaneous phase and EMD

The fundamental physical quantities of a structural system, such as acceleration, strain, and pressure, can be obtained through measurements of the time waveforms, which are denoted by time series data or the signal of $x(t)$. In principle, such signals can have any functional form and most likely will have extraordinary richness and complexity. One of the logical attempts to express such signals is to generalize the simplicity of the sinusoid functions with the time-dependent variables

$$x(t) = a(t) \cos[\theta(t)], \quad (11.1)$$

where the amplitude $a(t)$ and phase $\theta(t)$ are arbitrary functions of time. To emphasize that they generally change in time, the terms “amplitude modulation” and “phase modulation” are often used. Difficulties arise immediately. Nature gives us only the left side of (11.1). How do we break up a signal in terms of time-dependent amplitude and phase? Moreover, even if the data were generated numerically by way of (11.1) with chosen $a(t)$ and $\theta(t)$, the choices are somewhat arbitrary because there are an infinite number of ways of choosing different pairs of amplitudes and phases that

generate the same data $x(t)$. How do we unambiguously define an amplitude and phase? Why is a complex signal used, and how do we define a complex signal corresponding to a real signal? The general mathematical background needed to address these questions as well as the development of an analytic signal and its equations can be found in Cohen (1995).

The focus in this section is to utilize the mathematical groundwork given in Cohen (1995) and further impose constraints such that the phase function $\theta(t)$ can be used as a generalized angle associated with the time evolution of the dynamic system in space to describe structural waves. In the context of wave propagation, the time-dependent phase angle should be one of the important quantities for a variety of engineering applications. Unfortunately, the phase information is often ignored mainly because of a fundamental problem associated with extracting phases from measured data by way of the analytic signal without the important constraint. This topic will be the subject of discussion in the next few subsections.

11.2.1. Instantaneous phase

An instantaneous phase $\theta(t)$ can be defined by the analytic signal $z(t)$,

$$z(t) = x(t) + i \mathcal{H}[x(t)] = a(t)e^{i\theta(t)}, \quad (11.2)$$

where the given time series $x(t)$ is the real part of (11.2), and the imaginary part is the Hilbert transform of $x(t)$,

$$\mathcal{H}[x(t)] = \frac{1}{\pi} \text{PV} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau. \quad (11.3)$$

In (11.3), PV denotes the Cauchy principal value of the integral. An analytic signal represents a rotation in the complex plane with the radius of rotation $a(t)$ and the instantaneous phase angle $\theta(t)$, where

$$a(t) = \sqrt{[x(t)]^2 + \{\mathcal{H}[x(t)]\}^2} \quad \text{and} \quad \theta(t) = \arctan \left\{ \frac{\mathcal{H}[x(t)]}{x(t)} \right\}. \quad (11.4)$$

To uniquely describe the time evolution of a dynamic system flow such as a measured structural response, the phase functions need to be restricted to a definite evolving direction (e.g., either clockwise or counterclockwise) and a unique center of rotation at any time t . This requirement guarantees that the instantaneous angular velocity $\omega(t)$ of the rotation

$$\omega(t) = \frac{d\theta(t)}{dt} = 2\pi f(t), \quad (11.5)$$

or instantaneous frequency $f(t)$ remains positive for all time t . Obviously, only a positive $f(t)$ can be physically meaningful and can be used to construct the time-frequency spectrum.

A fundamental difficulty in defining an instantaneous phase $\theta(t)$ by using (11.2) and (11.3) based on a measured time series data $x(t)$ is that $\theta(t)$ may exhibit different evolving directions in the complex plane and/or multiple centers of rotations, even for simple continuous dynamic system flows. For example, the numerical solutions of the Lorenz system of equations,

$$\begin{aligned}\dot{y}_1 &= 16(y_2 - y_1) \\ \dot{y}_2 &= 40y_1 - y_2 + y_1y_3 \\ \dot{y}_3 &= -4y_3 + y_1y_2,\end{aligned}\tag{11.6}$$

given in Fig. 11.1 illustrate this difficulty. Figure 11.1a is the output time series $x(t) \equiv y_1(t)$, and Fig. 11.1 shows its trajectory in the complex plane of $z(t)$, which is the analytic signal constructed by using $x(t)$.

No phase function can be properly defined for the rotation characteristics displayed in Fig. 11.1b. A phase function obtained directly from the analytic signal does not uniquely represent the time evolution of the dynamics. Further explanations and examples of such phase characteristics will be given in subsection 11.2.3. This problem remains unsolved without the introduction of a special type of function, the intrinsic mode function (IMF). An IMF is a key concept associated with the method of EMD and Hilbert-Huang transform (HHT).

11.2.2. EMD and HHT

The empirical mode decomposition and Hilbert-Huang transform is a general time-frequency method, which consists of two stages in analyzing time series data. The first stage involves the use of EMD, which decomposes any given time series data into a set of simple oscillatory functions, defined as intrinsic mode functions (IMFs). An IMF must satisfy the conditions that a number of extrema must be equal to the number of zero crossings (or differ by one at most), and the mean value of the envelope functions, defined by local extrema, must be zero. The IMF components can be obtained by a repeated application of an iterative procedure called “sifting.” For example, the first IMF $c_1(t) = \text{sifting}[x(t)]$ can be obtained by sifting the $x(t)$, where $x(t)$ is the measured time series data. The second IMF is

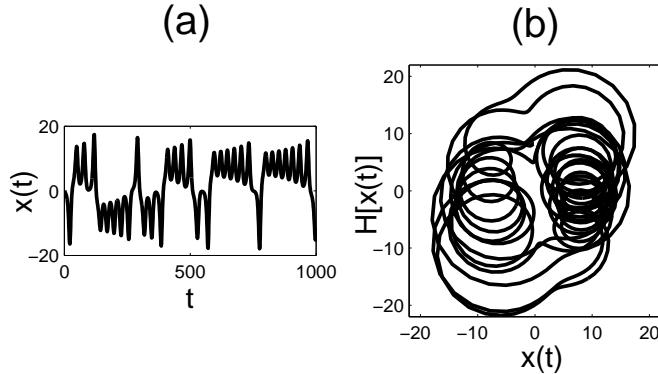


Figure 11.1: (a) Lorenz chaotic time series and (b) Lorenz trajectory in the complex plane.

then equal to $c_2(t) = \text{sifting}[x(t) - c_1(t)]$, and so on. As a result,

$$x(t) = \sum_{k=0}^n c_k(t); \quad (11.7)$$

that is, the measured time series data can be decomposed into n empirical modes, each satisfying the conditions of an IMF, plus a residual term $c_0(t)$. In general, n is a small number (often less than 10). This method of decomposition is very efficient. The decomposition base is directly derived from the data and is established from a simple assumption that any data consist of different intrinsic modes of oscillation. As a result, each IMF is an amplitude- and frequency-modulated signal; i.e.,

$$c_k(t) = a_k(t) \cos[\theta_k(t)]. \quad (11.8)$$

By these criteria, $c_k(t)$ is a mono-component signal (Cohen 1995), which has a monotonically increasing phase and a positive instantaneous frequency. A complete description of the sifting procedure can be found in Huang et al. (1998), and examples of $c_k(t)$ obtained from the structural response of an undamaged structure in comparison with that of a damaged structure are given in later subsections.

The second stage of the method defines the time-dependent amplitudes (or energies) and frequencies of the empirical modes by using a Hilbert transform. For a given $c_k(t)$, an analytic signal $z_k(t)$ can then be defined

by using the k th IMF component $c_k(t)$ and its Hilbert transform $d_k(t) = \mathcal{H}[c_k(t)]$. An instantaneous amplitude and frequency can then be defined accordingly by using (11.4) and (11.5). For simplicity, the mode index k is omitted in all equations and text below.

The instantaneous (time-dependent) frequency defined by this approach is based on the local oscillation over a characteristic time scale of the measured data, which is represented by n empirical modes, and the time derivative of its phase relationship is obtained from the Hilbert transform. This approach is fundamentally different compared to the conventional Fourier transform and Fourier representation of the frequency domain information.

Because both the amplitude and frequency of each IMF are a function of time, a three-dimensional space or ordered triplet $[t, \omega(t), a(t)]$ can be defined. This space is generalized by means of a function of two variables $H(\omega, t)$ to $[t, \omega, H(\omega, t)]$, where $a(t) = H[\omega(t), t]$. This final representation is referred to as the “Hilbert (amplitude) spectrum” $H(\omega, t)$ in Huang et al. (1998). $H(\omega, t)$ or $H(f, t)$ is also known as the “Hilbert-Huang transformation” or “HHT”. Similarly, the Hilbert energy spectrum can be defined by computing the amplitude squared to represent the energy density if desired. In addition, a Hilbert marginal spectrum $H(\omega)$ can be defined by the integration of $H(\omega, t)$ over the entire data length T . The marginal spectrum $H(\omega)$ or $H(f)$ is a frequency spectrum and measures the total amplitude contribution from each frequency value of the data.

11.2.3. Extracting an instantaneous phase from measured data

As stated in previous subsections, each $\theta(t)$ computed by $c(t)$ exhibits a definite evolving direction, either clockwise or counterclockwise, and a unique center of rotation at any time t . The left column of Fig. 11.2 shows the first four IMFs plotted as a function of time t , computed from the Lorenz output data given in Fig. 11.1a. The right column of Fig. 11.2 shows their corresponding trajectory in the complex plane of $z(t)$, or equivalently, in the polar coordinate $[a(t), \theta(t)]$. Collectively, a total phase function $\theta(t)$ can be defined as

$$\theta(t) = \sum_{k=0}^n \arctan \left\{ \frac{\mathcal{H}[c_k(t)]}{c_k(t)} \right\}. \quad (11.9)$$

While $\theta_k(t)$ [see (11.8)] represents the number of rotations for the k th IMF function, (11.9) gives the total number of rotations of the measured signal

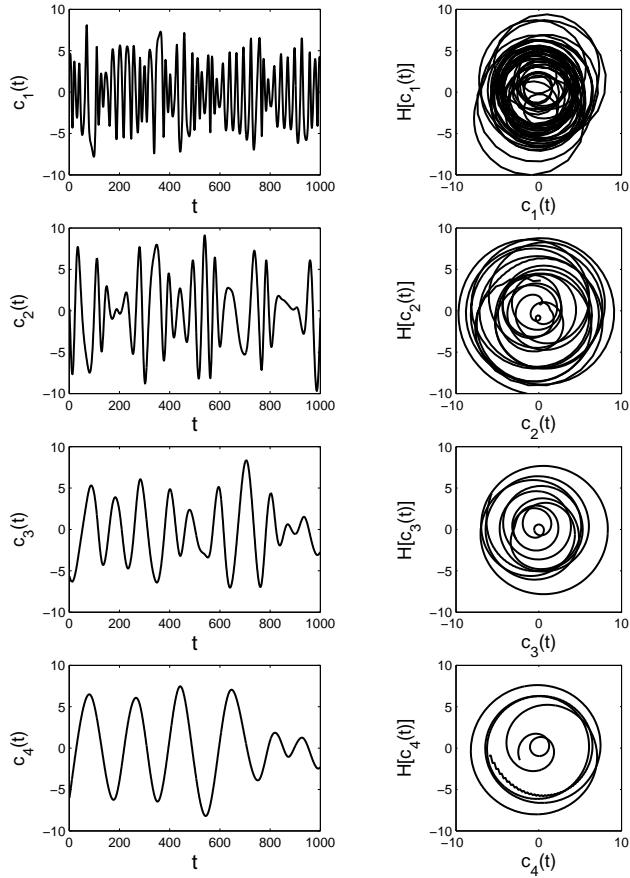


Figure 11.2: The first four IMFs (left column) and their corresponding trajectory in the complex plane (right column).

$x(t)$ in the complex plane. When $\theta(t)$ is unwrapped with radian phases in time, it assigns the total number of rotations in the complex plane for a unique time value t (multiplied by 2π). To further illustrate the properties of instantaneous phase functions, Fig. 11.3 is used to display $\theta(t)$ by using unwrapped radian phases with changing absolute jumps greater than π to their 2π complements. In Fig. 11.3a, $\theta(t)$, defined directly by the Lorenz output data, is compared with Fig. 11.3b, where $\theta(t)$ is defined by IMF and

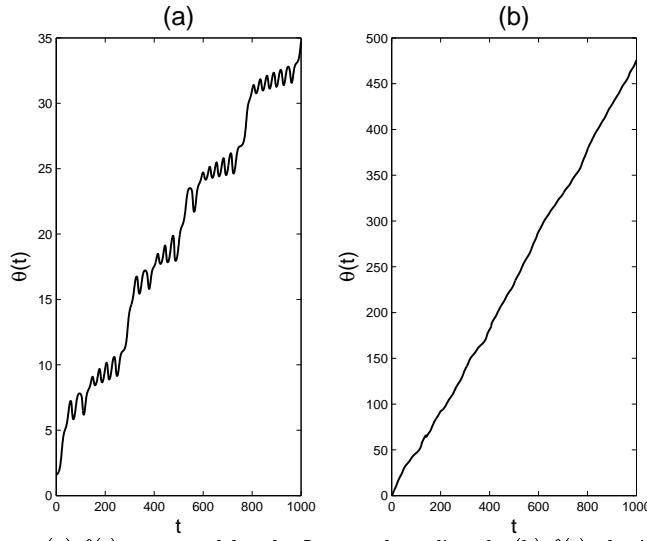


Figure 11.3: (a) $\theta(t)$ computed by the Lorenz data directly (b) $\theta(t)$ obtained by IMF and (11.9).

(11.9). Note the significant difference in vertical axes. In Fig. 11.3a, $\theta(t)$ misrepresents many important local oscillatory modes; i.e., it associates many different intrinsic modes of oscillation with the same time variable. Only in Fig. 11.3b, does $\theta(t)$ represent a well-defined phase function with monotonically increasing values for all time t . As a result, this phase function properly represents all local oscillations in the data.

11.3. Damage detection application

An instantaneous phase, defined in (11.9), can be a natural physical variable to describe the dynamics of a system. When this description is applied to a structural system, the phase is interpreted as a generalized angle associated with the time evolution of a measured structural response. In particular, $\theta(t)$ is used to represent the phase of traveling structural waves of any dynamically measurable quantity, such as the acceleration, strain, or displacement. The value of this phase function at a location p on a structure can be written as $\theta_p(t)$. If point zero on a structure is chosen as a reference point, the phase function relative to the reference point can be defined by

$$\varphi_s(t) = |\theta_p(t) - \theta_0(t)|. \quad (11.10)$$

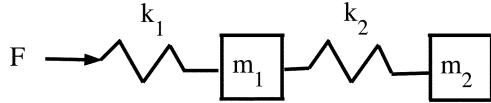


Figure 11.4: Three-degrees of freedom (DOF), one-dimensional spring-mass system.

The variable \$\varphi_s(t)\$ describes the relative phase relationship of a traveling structural wave for a given state of a structure \$s\$. The basic idea of detecting damage by using \$\varphi_s(t)\$ is that damage in a structure will alter the speed at which energy traverses through the structure. As a result, the \$\varphi_s(t)\$ values will be altered compared to those of an undamaged state \$\varphi_u(t)\$, with the damaged state \$\varphi_d(t)\$ between the same locations \$p\$ and 0 on the structure. The characteristics of the damage features identified through \$\varphi_s(t)\$ can then be combined with the concept of phase dereverberation in order to infer damage in terms of physical parameters such as structural stiffness, mass and damping (Purekar and Pines 2000; Ma and Pines 2003). Our previous work has suggested that tracking the changes in the wave speed of response measurements by using (11.10) through each individual structural element between successive degrees of freedom for an impulse excited one-dimensional structure is an effective method for identifying as well as locating damage (Pines and Salvino 2002; Salvino and Pines 2002). A brief description of this approach is given in the next few subsections.

11.3.1. One-dimensional structures

The concept of phase dereverberation can be interpreted as obtaining the response of a structure to the incident energy imparted to the system. Consider the three-degrees of freedom (DOF) spring-mass system shown in Fig. 11.4.

The equation of motion for this system is given generically by

$$M\ddot{x} + Kx = F, \quad (11.11)$$

where

$$M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & m_1 & 0 \\ 0 & 0 & m_2 \end{bmatrix}, \quad K = \begin{bmatrix} k_1 & -k_1 & 0 \\ -k_1 & k_1 + k_2 & -k_2 \\ 0 & -k_2 & k_2 \end{bmatrix}, \quad \text{and} \quad F = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \quad (11.12)$$

By using the concept of virtual control, the dereverberated transfer functions can be represented by

$$\begin{bmatrix} u_1/F_1 \\ u_2/F_1 \\ u_3/F_1 \end{bmatrix} = \begin{bmatrix} k_1 & -k_1 & 0 \\ -k_1 & k_1 + k_2 - m_1\omega^2 + G_1 & -k_2 \\ 0 & -k_2 & k_2 - m_2\omega^2 + G_2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \quad (11.13)$$

The virtual controllers G_1 and G_2 are given by

$$G_1 = k_1 (1 - e^{-\mu_1}) - k_2 (1 - e^{-\mu_2}), \quad (11.14)$$

and

$$G_2 = k_2 (1 - e^{-\mu_2}), \quad (11.15)$$

where μ_1 and μ_2 are the propagation coefficients. An impulse F is applied from the left side, and the baseline system response is simulated by considering parameters $m_1 = m_2 = 1$, $k_1 = 1$, and $k_2 = 2.5$. The damaged cases are generated by reducing the spring constants by 25%, 50% and 75% for each k_1 and k_2 separately. The displacement response of the dereverberated time series at DOF 2 is displayed in Fig. 11.5 comparing baseline and three simulated damage cases in the first structural element with varying spring stiffness k_1 .

For the example given in Fig. 11.5, the time series of the DOF 2 reveals a phase lag associated with the incident impulse as it traverses the structure from left to right. As the stiffness of element 1 decreases, the speed at which energy traverses this element decreases. This finding suggests that one approach for determining the presence of damage in a structure is to track the changes in wave speed through each individual structural element between successive degrees of freedom. Thus, one physical approach to damage identification in a system is to use the phase relationship of successive degrees of freedom to track the variation in structural properties.

The simulated time series for the displacement response, as shown in Fig. 11.5, are first decomposed into a set of IMFs. These IMFs are then used in (11.9) to calculate the time-dependent phases of the three DOF spring-mass system. The value of this phase function at DOF 2 is written as $\theta_2(t)$. If DOF 1 is chosen as a reference point, the phase function relative to the reference point can be defined by

$$\varphi_s(t) = |\theta_2(t) - \theta_1(t)|. \quad (11.16)$$

Figure 11.6 illustrates the changes of phase relationship due to the presence of damage in the system. The left side of Fig. 11.6 shows the relative phases

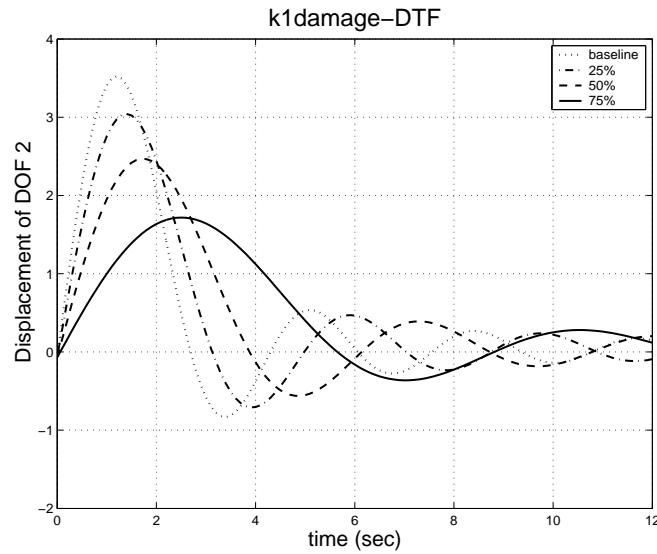


Figure 11.5: Displacement responses for the three DOF system of undamaged and three k_1 -damage cases.

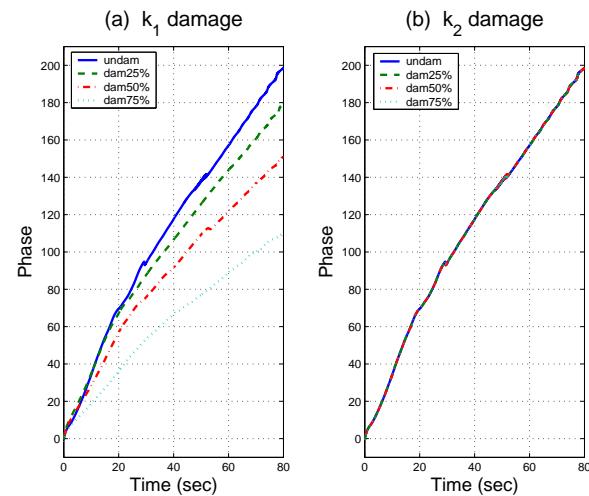


Figure 11.6: Relative phases of dereverberated response between DOF 2 and 1. (a) damage occurred in k_1 , (b) damage occurred in k_2 .

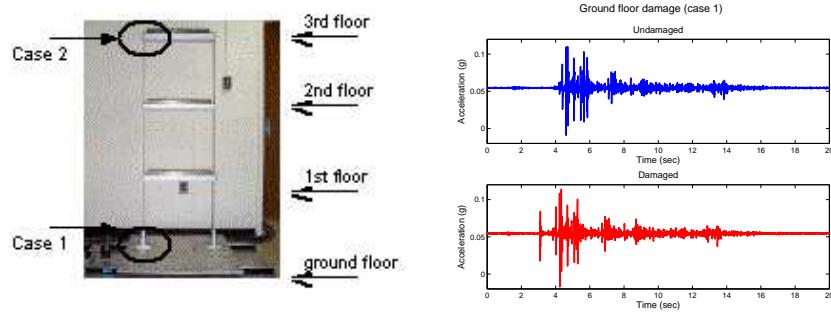


Figure 11.7: (a, left) Scaled building model and (b, right) measured acceleration time series data (in the unit of g) for undamaged and damaged ground floor response.

$\varphi_s(t)$ for simulated damage in k_1 . The state of system in the baseline stage ($s = 1$) is compared with three simulated damage cases ($s = 0.75, 0.5, 0.25$) in this figure. The phase lags between DOF 2 and DOF 1 due to increasingly softened springs are clearly evident. The right side of Fig. 11.6 shows the relative phases for simulated damage in k_2 . The relative phases between DOF 2 and DOF 1 remain constant for all four cases because the damage occurring in k_2 (between DOF 3 and DOF 2) does not affect the incident energy between DOF 2 and DOF 1 for the dereverberated response.

The use of the EMD method to obtain IMFs is essential to define and compute instantaneous phases, even for the simplest system response such as that shown in this subsection. Without employing the EMD method, the necessary property of proper rotation defined in the previous subsection cannot be guaranteed.

11.3.2. Experimental validations

To develop a better physical understanding of the method, a laboratory test of a scaled building model, with and without damage, is used to validate our numerical results. The building model, shown in Fig. 11.7a consists of four floors, with the ground floor driven by a hydraulic shake system. On each floor, PCB accelerometers are mounted to measure the vibratory response of the building before and after damage. To simulate seismic loading, a time series from the El Centro earthquake was used as input to the hydraulic actuators. The identified undamaged natural frequencies were determined to be 2.5, 7, and 12 Hz, respectively, for the first three modes of the structure. Two damage cases were simulated by physically changing the properties

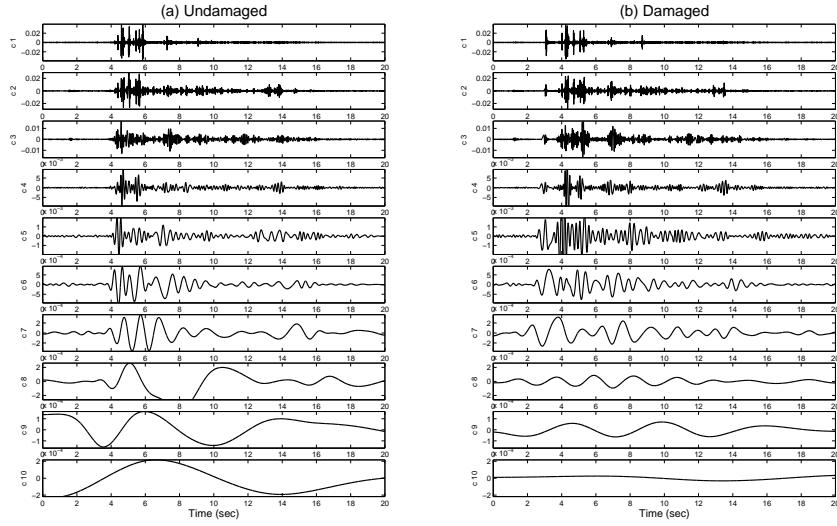


Figure 11.8: IMF components for (a) undamaged and (b) damaged response data (in units of g).

of the structure. In this study, stiffness damage was simulated by removing two bolts in the bottom flange near the ground level (Case 1), and by removing two bolts from the top floor connection (Case 2). An example of the measured response time series data, undamaged and damaged, for Case 1 of the ground floor is shown in Fig. 11.7b. These measured acceleration time series are first decomposed into sets of IMFs, and examples of IMFs are given in Fig. 11.8. Each IMF component can be viewed as an adaptive basis function derived directly from measured data. These basis functions are both amplitude- and frequency-modulated signals, and they reveal unique time-frequency features through HHT.

Figure 11.9 displays the comparison of the HHT spectrum of undamaged (top) and damaged (bottom) ground-floor response as joint functions of time (horizontal axes) and frequency (vertical axes). The spectra amplitudes are displayed as scaled images, and their normalized values are indicated by the color bars on the right side of each image. The time-frequency characteristics of the top and bottom plots, due to the richness of the structural response at the ground level, are similar, with the exception of the appearance of a noticeable distribution of broadband signal energy at about

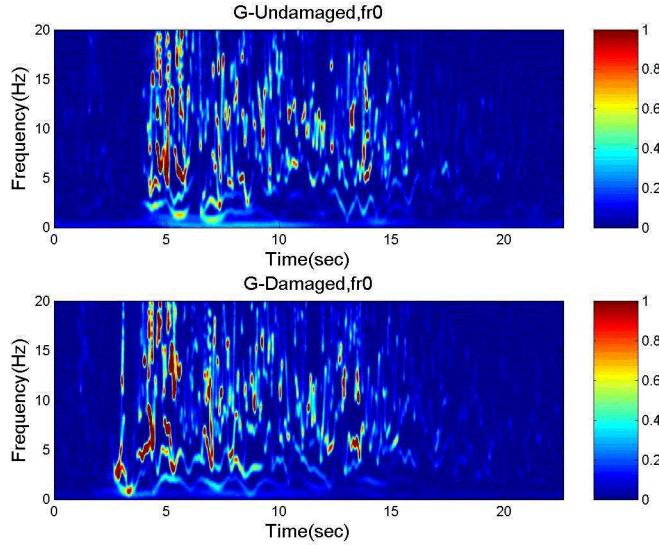


Figure 11.9: HHT spectrum of ground floor acceleration response for undamaged (top) and damaged (bottom) cases.

3 s in the damage case. This feature is due mainly to the loosening of the bolts of the ground floor brace used to support the upper stories. As the seismic wave enters the ground floor, the brace bangs against the loose bolts and vertical legs of the structures, causing an impact to occur prior to the full seismic wave reaching the ground floor accelerometer.

As the seismic wave continues to traverse up the building, noticeable changes start to appear in the Hilbert response spectrum displayed in Fig. 11.10. Notice the significant loss in intensity between the undamaged and damaged responses. This loss is particularly clear if one tracks the third modal frequency band of energy between the damaged and undamaged spectra.

As the intensity in this frequency band diminishes, a new phenomenon becomes apparent in the time frequency response. At approximately 12 s, the third mode band starts to spread as the intensity decreases, suggesting the presence of a structural nonlinearity in the system caused by the loosening of the two bolts. Additional analysis by extracting the damping loss factors (Salvino 2000; Pines and Salvino 2002) confirms that this drop in intensity is consistent with an increase in damping in the system as a result of members sliding back and forth against one another.

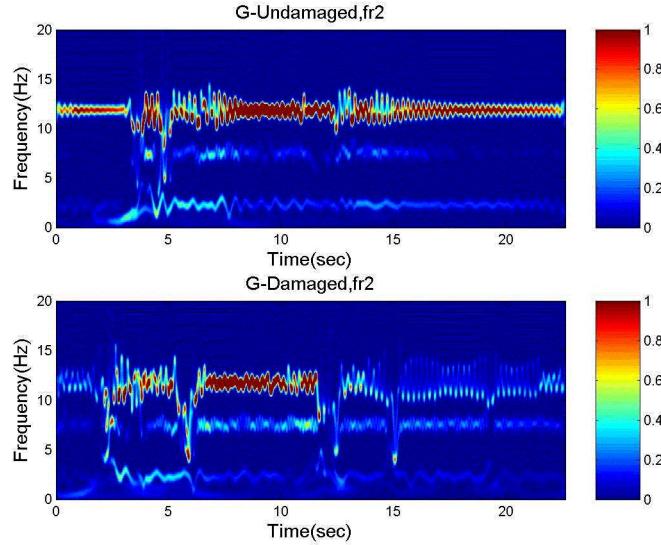


Figure 11.10: HHT spectrum of second floor acceleration response for undamaged (top) and damaged (bottom) measurements.

Since the nonlinear/inelastic behavior of a damaged structure should be expected to affect the frequency composition of the structural response, and the novel technique of EMD and HHT spectrum is known to offer superb temporal and frequency resolution compared to that of more conventional wavelet and windowed FFT analysis methods, it is not surprising that damage in structures can be identified by using this method. Others have also studied EMD and HHT applications in system identification and damage detection for civil structural applications (Yang et al. 2004a, 2004b).

This result is particularly important when the examination of the signals recorded during a damaging event requires consideration of nonlinear behavior. These time-frequency changes must be linked to physical parameters; in other words, changes identified from response characteristics must be linked to changes in the physical properties due to damage. This problem is extremely difficult because the mathematical model never captures the true complexity of the real system. No well-established solution is available at present.

In the next subsection, we demonstrate a promising step to link the time-frequency features identified by using the empirical mode analysis to the

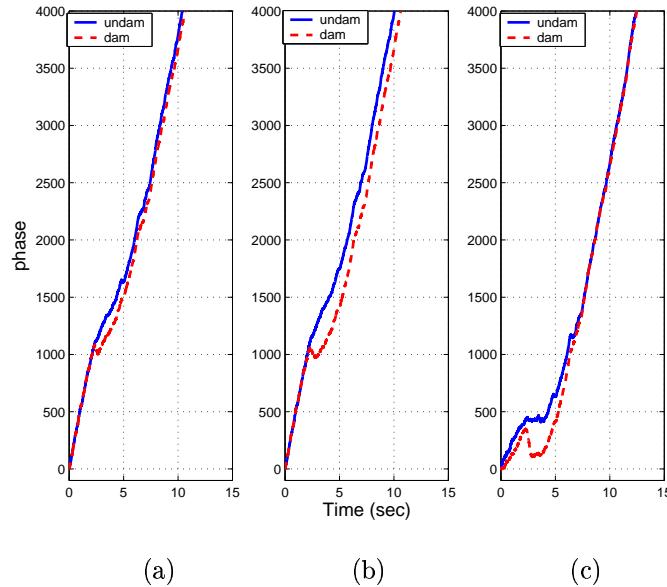


Figure 11.11: Relative phases of measured response for ground floor damage of a scaled building.

concept of damage identification by using phase dereverberation, a concept derived from the wave mechanics for the above building model.

11.3.3. Instantaneous phase detection

The IMF functions such as those given in Fig. 11.8 are then used in (11.9) and (11.10) to directly obtain the phase lags as functions of time.

Figure 11.11 shows the relative phases defined in (11.10) for the undamaged baseline and the ground-floor damage case. Figure 11.11a is obtained from data on the first floor, Fig. 11.11b is from the second floor, and Fig. 11.11c is from the third floor. All sets of data use the reference point of the ground floor measurement data.

The phase lag due to damage shows up in the measured signal for each floor. These phase lags can be directly attributed to the stiffness loss of the structure based on the numerical study given in subsection 11.3.1. In addition, the locations of the measurements clearly indicate that the damage must have occurred between the ground and first floors since the magnitude

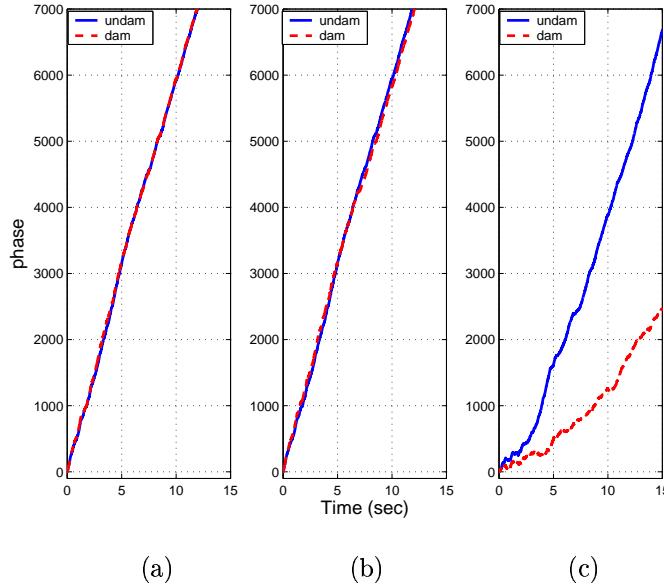


Figure 11.12: Relative phases of measured response for top-floor damage of a scaled building.

of the phase lags are proportional as the seismic wave continues traveling from the ground floor to the top floor.

Figure 11.12 shows the top floor damage case where the relative phases were computed in the same way as those for Fig. 11.11. The phase lag due to damage shows up in the measured signal of the top floor only, indicating the structural damage is located between the top two accelerometers or between the second and third floors.

To further demonstrate the concept of combining an instantaneous phase of a measured structural response with the values of the physical parameters of a structure, a linear approximation of a one-dimensional asymmetric structural element was studied analytically. The results can be found in Pines and Salvino (2004). The phase delay, which is expressed by physical quantities such as mass and stiffness in the closed-form solutions of linear approximation, is identical to the phase functions given in (11.3)–(11.5) below the cutoff frequency.

The basic concept of a new structural-damage-detection method is presented in this subsection. This method tracks the changes in wave speed

through each individual structural element between successive degrees of freedom. The examples given, both the simple numerical simulation and the experimental results, have shown that the method can be an effective technique for identifying and locating damage. Quantifying the amount of structural damage present in the structure requires correlating these relative phases with the known values of stiffness between each degree of freedom. The procedure is similar to that of previous work (Ma and Pines 2003).

11.4. Frame structure with multiple damage

The EMD and HHT technique in conjunction with a wave-mechanics based structural health monitoring method has been presented in the previous subsection for a one-dimensional structure. Here, in this subsection, this approach is extended through experimental demonstrations to a more generalized structure with damage in multiple locations. Furthermore, the instantaneous phase detection method is compared with a more conventional time-domain approach using an auto-regressive modeling technique as well as another recently proposed damage-detection method using geometric properties of deterministic dynamics. These comparative studies show the wide applicability and adaptive nature of the instantaneous phase detections approach. In addition, direct comparisons of these three methodologies provide the reader with examples of the rapidly evolving SHM approach, which is based on data-driven or empirical modeling techniques.

11.4.1. Frame experiment

A scaled three-story frame structure was used to perform the experimental demonstration. A photograph of the test structure is shown on the left side of Fig. 11.13. The floors were 1.3-cm-thick (0.5-in-thick) aluminum plates with two bolt connections to brackets on the aluminum unistrut columns. The base was a 3.8-cm-thick (1.5-in-thick) aluminum plate. Support brackets for the columns were bolted to this plate. Further details of this structure may be found in Fasel et al. (2002). Damage was created by removing the bolts for various combinations of the connections between floor 1 and the three unistruts (right side of Fig. 11.13). The baseline condition (undamaged) and five damage scenarios are summarized in Table 11.1.

An electromechanical shaker was connected to the structure at the mid-height of the base plate so that translational motion could be imparted. The structure was instrumented with six piezoelectric accelerometers on floor 1 only. Accelerometers were mounted on blocks glued to three of the unistrut

Table 11.1: Damage case summary description.

Damage Case	Damage Location
0 (baseline)	All bolts at 12.4 N·m (110 in-lbf) torque
1	Joint D bolts removed
2	Joint A bolts removed
3	Joint B bolts removed
4	Joints A and D bolts removed
5	Joints A, B, and D bolts removed

columns (joints A, B and D) at the floor level, in both in-plane directions. Additionally, a force transducer was mounted between the stinger and the base plate and used to measure the input to the base of the structure. A commercial-data acquisition system controlled from a laptop PC was used to digitize the accelerometer and force transducer analog signals.

Two methods of interrogation were used to excite the structure in order to perform the comparative studies mentioned previously. The excitation signals include both a deterministic chaotic input source and a random white noise input source. The chaotic excitation signal was generated by the well-known Lorenz oscillator explicitly given by the vector $\mathbf{y}(t) = [y_1, y_2, y_3]$ in (11.6).

A simple input/output MATLAB SIMULINKTM model along with XPC TargetTM and a commercially available signal conditioner was used to send the pre-digitized chaotic Lorenz waveform to the shaker. The waveform was sent at a 5000 Hz shaker update rate, and the analog sensors were sampled at 1500 Hz with $N_t = 16\,384$ total sample points. Five sets of baseline undamaged datasets were recorded before damage was introduced to the structure. Similarly, the test procedure was repeated for each of the damage cases, resulting in $N_r = 5$ runs at each of $N_d = 6$ total damage cases (including the baseline) to give 30 total datasets. The sample shaker input (channel 1), force transducer response (channel 2), and acceleration response signals (channel 3 and channel 4) are shown in Fig. 11.14, along with their corresponding phase portraits generated by delay coordinate constructions (Kantz and Schreiber 1999).

The driving signal and the force felt by the structure closely track the input Lorenz waveform, Figs. 11.14a and 11.14b, with some differences in the latter primarily due to structural feedback. The response acceleration time series, however, are less clearly “structured” and illustrate the filtering

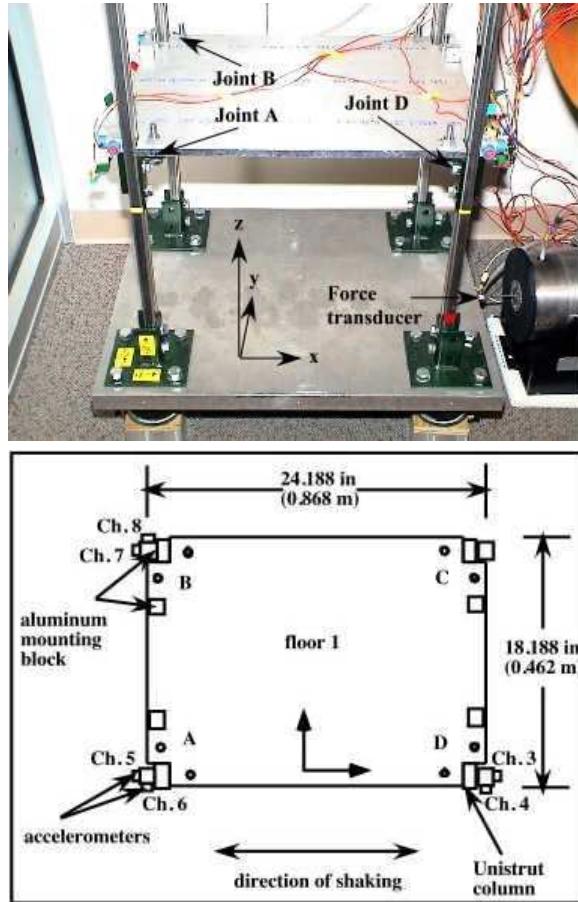


Figure 11.13: Frame experimental set-up showing instrumented floor at joints A, B, and D.

effect of the structure on the input signal. Both co-directional (channel 3 in Fig. 11.14c is in the same direction as the excitation) and contra-directional (channel 4 in Fig. 11.14d is perpendicular to the direction of excitation) sensor data are shown in order to highlight the differences in response data from these two directions. Although the damage identification analyses were performed for many different locations on the frame, or, in other words, all channels of measured data were studied, we show the results only for channel 3 data in subsections 11.4.2–11.4.5, providing direct comparisons of three entirely different data-driven SHM methods.

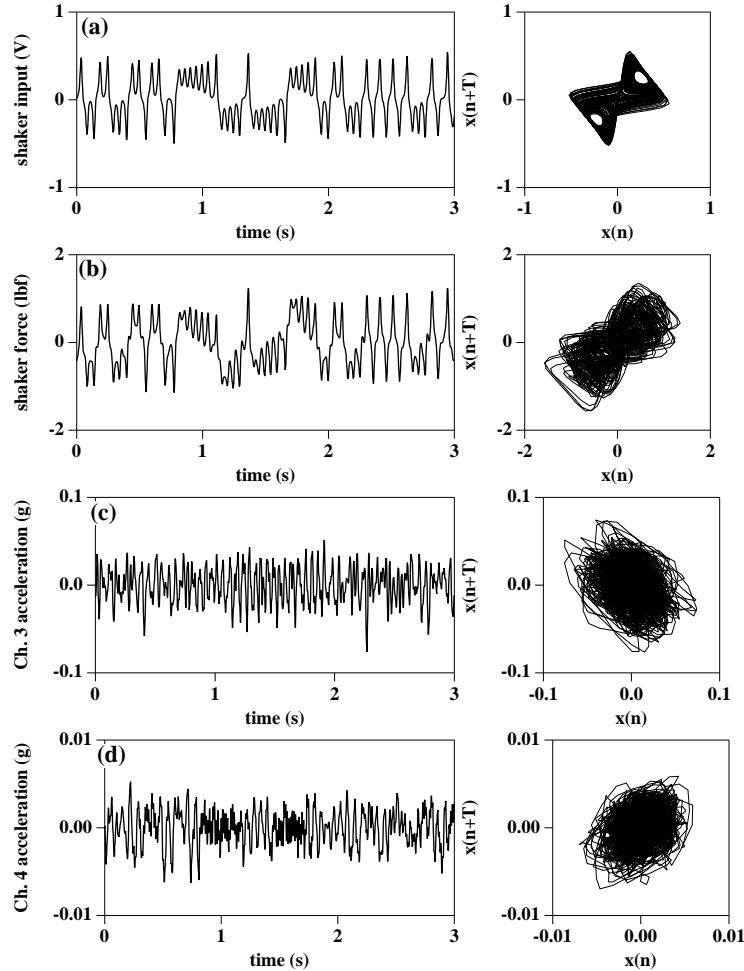


Figure 11.14: Typical 3-s time windows (left column) and their delay coordinate constructions (right column) for (a) raw voltage input to shaker, (b) force transducer output (channel 2), (c) channel 3 acceleration, and (d) channel 4 acceleration.

Additionally, broadband random signals were also applied to excite the frame. The test procedure was repeated for the baseline and for each of the damage cases. The data were sampled under the same conditions as those of the chaotic input except $N_t = 8\,192$ total sample points were recorded for each of the 30 datasets. Examples of the power spectra density (PSD) obtained from the baseline structural response measurement (channel 3

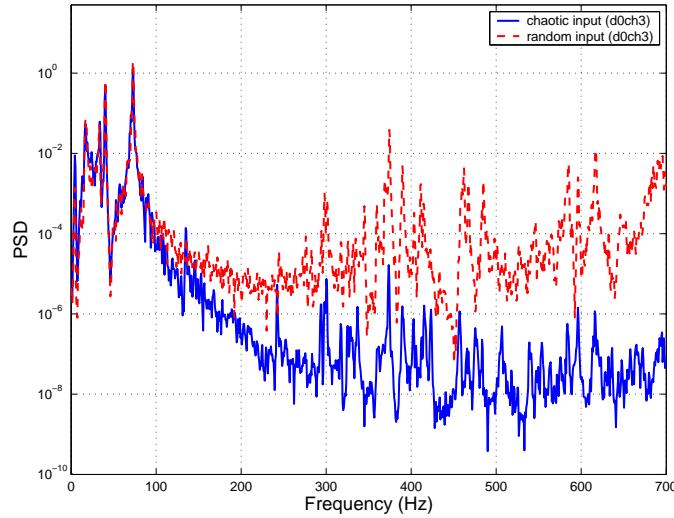


Figure 11.15: Power spectral density of channel 3 data for baseline frame excited by chaotic input and excited by random input.

time series) are given in Fig. 11.15.

The characteristic comparisons displayed on the PSD plot for chaotic and random forcing indicate that distinct structural response differences occur in the energy distributions. Even though the fundamental structural modes being excited are somewhat similar for the lower frequencies, the floor level of the higher frequency region is much lower for the chaotic driving condition. This finding may indicate a high signal-to-noise ratio or more clean structural-response data. We must point out that although the Lorenz oscillator's input appears fairly complex in the time domain, its dynamics in phase space is low-dimensional. The power spectra density cannot reveal its unique chaotic dynamic nature.

11.4.2. Detecting damage by using the HHT spectrum

The time series from the frame experiment described in subsection 11.4.1 were analyzed initially by using the EMD and HHT method. The first step in this procedure is to decompose the data into a set of IMF functions; in other words, the measured time series is represented by a sum of n amplitude- and frequency-modulated functions, each satisfying the condi-

tion of monotonically increasing phase and positive frequency. This step, as stated in a previous subsection, is the fundamental aspect of the method. These IMF functions are then used to compute the Hilbert transform and to construct the HHT, $H(f, t)$. Figure 11.16 displays the comparison of the HHT for the channel 3 data of the baseline (Fig. 11.16a) and damage cases 1–5 (Figs. 11.16b–f, respectively) given in Table 11.1.

Figure 11.16 provides a good initial visual comparison of the structural response as joint functions of time (horizontal axes) and frequency (vertical axes). As in Figs. 11.9 and 11.10, the spectra amplitudes are displayed as scaled images, and their normalized values are indicated by the color bars on the right side of each image. Many features in $H(f, t)$ can be interpreted to explain the frame behavior for each damage case. Generally, many more temporal features are noticeable when the structure becomes damaged. The damaged frame exhibits different degrees of frequency lowering, widening, and intensity loss. This finding suggests the presence of a structural nonlinearity in the system caused by stiffness loss at the joints. For example, the baseline frame (Fig. 11.16a) is quite different compared to each of the damaged cases at approximately 40 Hz. Note that the channel 3 measurements displayed in Fig. 11.16a–f are the accelerometer outputs closest to the driving location and to joint D. The frequency band between 140–160 Hz that is clearly present in Figs. 11.16b and 11.16e but not in the others appears to be due to the damage caused by removing bolts at joint D (damage cases 1 and 4 in Table 11.1).

The Hilbert spectra analysis can be beneficial for understanding the overall frame behavior resulting from each of the damage cases. However, the relative phase relationship introduced in the previous subsections can provide further understanding of the time-frequency changes in terms of physical parameters. Natural connections can be made from the changes identified through the response characteristics to changes in the physical properties due to damage.

11.4.3. Detecting damage by using instantaneous phase features

In this subsection, the damage identification for the frame structure is based on the comparison of the relative phase values of frame vibratory response before and after damage. As explained previously, when damage in a structure is present, this damage will alter the speed at which energy traverses through the structure. For the baseline undamaged frame $s = 0$, variable

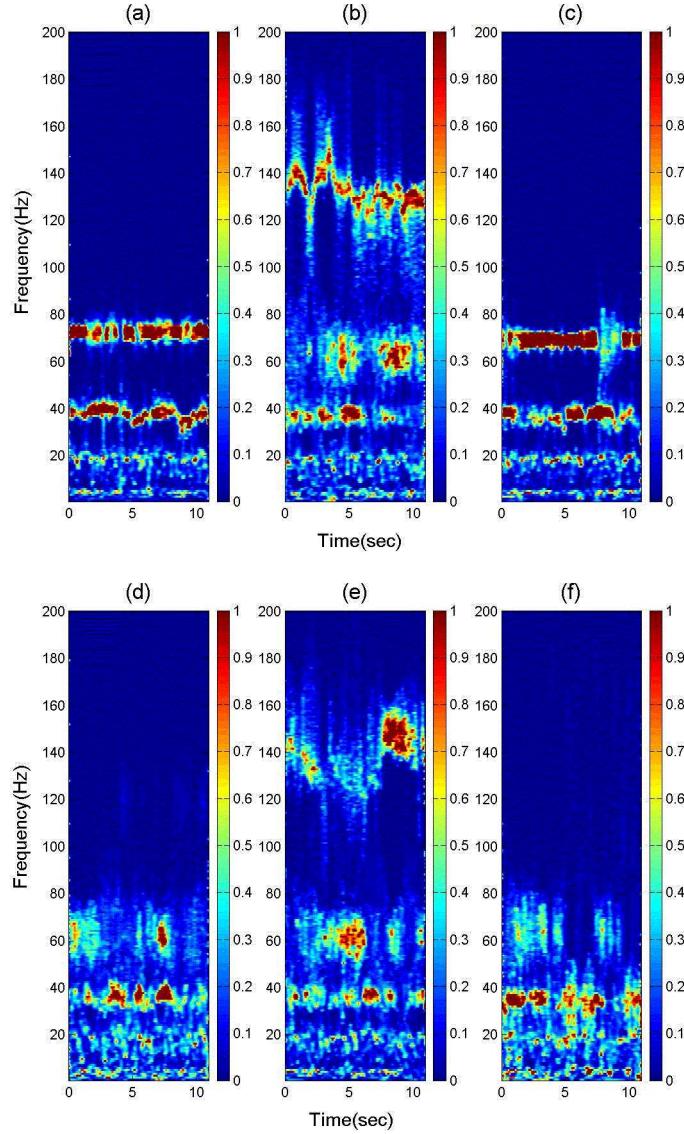


Figure 11.16: Hilbert spectra for baseline and different cases of frame damage.

$\varphi_s(t)$ given in (11.10) is calculated by using the measured acceleration signal. The result is then compared with the $\varphi_s(t)$ values obtained by using the acceleration data at the same locations for a variety of different dam-

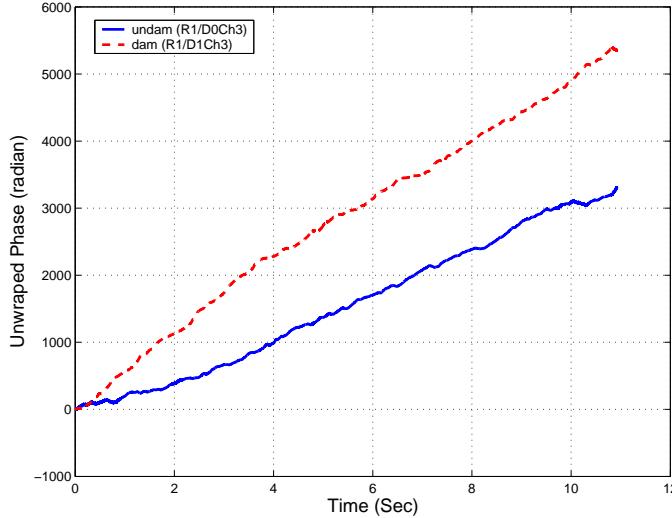


Figure 11.17: Variable $\varphi_0(t)$ for baseline frame is compared with $\varphi_1(t)$, frame with damage case 1 (joint D damage).

age cases $s = 1, 2, 3, 4, 5$. Figure 11.17 displays examples of comparing $\varphi_0(t)$ (baseline) and $\varphi_1(t)$ (damage case 1 - Joint D bolt removed).

The vertical axis is given by unwrapping radian phases with changing absolute jumps greater than π to their 2π complement, just as we did in the figures of the instantaneous phase. The two curves in Fig. 11.17 provide a comparison of $\varphi_1(t)$ and $\varphi_0(t)$, where $s = 0$ indicates the baseline frame, and $s = 1$ indicates the damage case 1 (joint D damage in this example). Channel 3 data for the baseline frame and for the frame with damage case 1 are used to calculate $\theta_p(t)$. Channel 2 data are used to calculate $\theta_0(t)$.

Figure 11.17 clearly illustrates the changes of the phase relationships relative to the driving point due to the presence of damage in the system. The reference point used here is chosen to be the excitation force measurement point (channel 2). This choice of the reference phase function $\theta_0(t)$ gives larger values compared with $\theta_p(t)$, for all t , for both undamaged and damaged cases. Because of this choice for $\theta_0(t)$, the phase difference $\theta_p(t) - \theta_0(t)$ will be negative for all t , again for both the damaged and undamaged cases. However, the relative phase $\varphi_s(t)$ defined in (11.10) is the absolute value of the phase difference. The curve for the damaged case lies above the curve

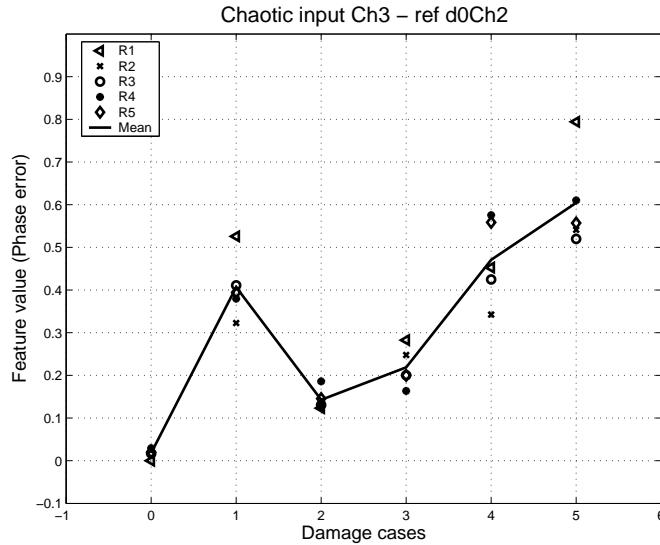


Figure 11.18: Damage index (feature value) calculated from channel 3 data for all five runs of each of the 6 total damage cases including the baseline case.

for the undamaged case in Fig. 11.17 because of the choice for $\theta_0(t)$ and the use of the absolute value in the definition (11.10).

The choice of a reference point (channel 2) is based purely on considerations of the signal-to-noise ratio and the measurement quality. When the reference points on a structure are varied, such as by using channel 5 and channel 7 data to calculate $\theta_0(t)$, damage can be located based on the geometry of the structure. More systematic research will be conducted on this topic in the future.

To measure the average phase change between the baseline and the damaged structure, a mean phase error is introduced as a feature value or damage index:

$$E = \sqrt{\frac{\sum_{t=1}^{N_t} [\varphi_1(t) - \varphi_0(t)]^2}{\sum_{t=1}^{N_t} \varphi_0^2(t)}}. \quad (11.17)$$

The feature values can be computed for all $N_r = 5$ runs of each of the $N_d = 6$ total damage cases (including the baseline). The results for the channel 3 data are shown in Fig. 11.18.

The E value provides an average phase difference measured between

two locations on a structure for two different structural conditions such as the baseline and damaged frame. In other words, the E value indicates the average deviation for the phases of a structural wave from its baseline state. A greater E value shows larger phase differences and may indicate more severe damage or a larger alteration from the baseline state. The phase error E value for the baseline frame (damage case=0) was calculated by using the channel 3 data from all 5 runs to obtain $\theta_p(t)$ and channel 2 data from the first run (R1) to obtain $\theta_0(t)$. As a result, $\theta_p(t)$ calculated by using the first run (R1) gives an identically zero E value as expected. However, the E values calculated from runs 2 to 5 (R2-R5) provide a good range of statistical error. As can be seen, all the E values for the baseline case are at least an order of magnitude smaller than the E values for any of the damaged cases.

The feature E value displayed in Fig. 11.18 discriminates among all damage cases except between cases 1 and 4. Channel 3, as indicated in Fig. 11.13, is located at joint D and provides overwhelmingly strong signatures for case 1 damage (bolts removed at joint D). Better discrimination can be expected between damage cases 1 and 4 when data from other channels are used. Similar analyses were performed for other measurement channels, and the results were consistent with those for channel 3.

As mentioned previously, broadband random signals are also applied to excite the frame. The frame response magnitudes are in the same range as those for the chaotic excitations. The output response time series from random forcing are also used to calculate the feature E value. Direct comparisons of random and chaotic excitation E values for the channel 3 data can be found in Salvino et al. (2003). Although the averaged E follows somewhat the same trend compared to that of the chaotic excitation cases, large fluctuations are observed between different runs. The use of broadband forcing present a much greater challenge as far as signal processing is concerned. Improved E values can be obtained when the data are filtered. The improvement and the final results critically depend on the characteristics of the filter used. Obviously, this approach is not desirable either for gaining a basic understanding of the physics involved or for practical application and implementation. A more conventional modeling technique that lends itself to stochastic excitations is presented below in subsection 11.4.4.

11.4.4. Auto-regressive modeling and prediction error

A pure time-domain approach for empirical modeling of time series data is autoregressive (AR) modeling. Here, the future system output is approximated by regressions of the past values of the dynamics. This approach is also data-driven in that all parameters are estimated from the response signal alone. The only assumption is that the output is a linear combination of the past values, such that the underlying equation of the AR model is given by

$$x(n) = a_0 + \sum_{j=1}^{\beta} a_j x(n-j) + \eta_n, \quad n = 1, 2, \dots, N. \quad (11.18)$$

The $x(n)$ are observed measurements, and the a_j values denote the AR regression coefficients, and the η_n are taken as Gaussian white noise. Therefore, (11.18) is most appropriately used for the modeling of linear stochastic processes. Coefficients for the AR model are computed by solving the over-determined least-squares problem

$$\begin{bmatrix} x(\beta+1) \\ x(\beta+2) \\ \vdots \\ x(\beta+N) \end{bmatrix} = \begin{bmatrix} 1 & x(\beta) & x(\beta-1) & \cdots & x(1) \\ 1 & x(\beta+1) & x(\beta) & \cdots & x(2) \\ \vdots & \vdots & & \ddots & \\ 1 & x(N-1) & \cdots & \cdots & x(N-\beta) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_\beta \end{bmatrix}. \quad (11.19)$$

The order of the model is denoted by β and defines the number of past observations used in determining the present value. Numerous methods for choosing an appropriate model order appear in the literature. The model order should be at least as large as the number of frequencies one hopes to capture, for roots of the polynomial defined by the coefficients give the poles of the system's rational transfer function (Owen et al. 2001). While a certain minimal number of past values are necessary to accurately predict any future values, too many parameters can lead to overfitting. Overfitting occurs when the peculiarities of the actual signal, such as noise and intrinsic fluctuations, are interpreted as global features of the system (Kantz and Schreiber 1999).

The autocorrelation function of the data is often a good indicator for choosing an appropriate order. Fast decay rates of the autocorrelation function imply that the time series has a “short memory”; hence, a lower order model is appropriate. Probably the most common approach is to form a cost function for the selection of β that involves a “goodness of fit” term and a penalty term for a high model order. One such example is the final

prediction error (FPE) criteria given by Akaike (1969) and Owen et al. (2001). For health-monitoring applications, the importance of model order is somewhat diminished. The goal in SHM is not necessarily to make accurate forecasts, but to construct a model that breaks down in the presence of damage.

In this subsection, the idea is to use the time series taken from the “pristine” structure to reconstruct a reference, or baseline, system AR model. This AR model is then used to forecast or predict the system’s response by using data taken at later times, when the structure may possibly have been damaged. Such an approach falls within the context of supervised learning where data from known particular damage cases must be obtained for later classification purposes. The working hypothesis is that damage accumulation will alter the structure’s dynamic response, causing these baseline models to lose their ability to make predictions. The error formed between prediction and actual measurement, called the *prediction error*, is proposed as a good candidate feature for quantifying both the presence and magnitude of structural degradation.

A group of N_r time series measurements $x_i(n), i = 1, \dots, N_r$, from a given sensor location are taken from the structural vibration response (subject to random Gaussian excitation). Recording N_r such independent responses allows for the inclusion of ambient variation in the set of baseline data. Such variation is known to occur in practice and must be accounted for if damage-induced changes are to be distinguished from those due to environmental factors. The AR coefficients are determined for each of these reference datasets. The coefficients are then used to create forecasts on a subsequently measured time series $y_i(n)$. For a given time series, the future value \hat{y} of a randomly chosen fiducial point $y(f)$ is predicted by using the model generated with the baseline data $x(n)$ as

$$\hat{y}(f+1) = a_0 + \sum_{j=1}^{\beta} a_j y(f-j+1), \quad (11.20)$$

shown here for prediction horizon $s = 1$. Subsequent predictions for $s > 1$ may be made by incorporating the predicted values into (11.20) in order to make continued forecasts. The residual errors between the actual and predicted values are the damage-sensitive feature, computed at the fiducial point f by

$$\gamma_f = |\hat{y}(f+s) - y(f+s)| / \sigma_x^2, \quad (11.21)$$

where the metric has been normalized by the variance of the reference

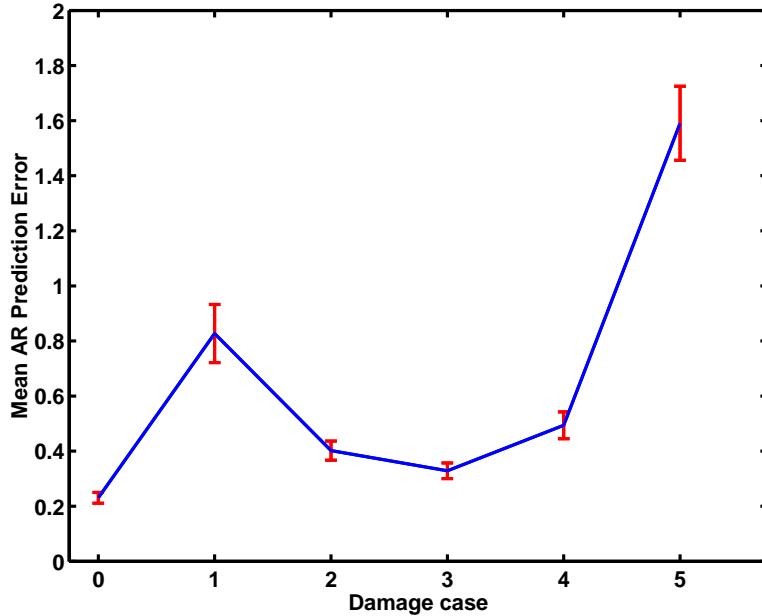


Figure 11.19: AR mean prediction error as a function of damage cases.

signal, σ_x^2 . This metric tests the ability of a model derived from one set of data [in this case, the a_j computed from the $x(n)$] to make predictions on the other time-series $y(n)$. This process is applied to each of the N_r^2 possible pairs of time-series and for some number of fiducial points to generate a distribution of feature values .

This procedure was applied to the channel 3 sensor data of the test structure described previously. The excitation signal was taken to be broadband, Gaussian noise in accordance with (11.18). Each time series (for all damage levels) was first normalized by subtracting the mean and dividing by the standard deviation. The structure's response for the undamaged scenario was then used to build an AR model by solving (11.19) with model order $\beta = 5$ (based on the autocorrelation structure of the baseline time series). Subsequent data from the damaged structure were then tested against this model by means of (11.20) and the prediction error γ recorded for $f = 1, 2, \dots, 5000$ randomly chosen fiducial points.

The probability density functions (PDFs) associated with the prediction errors for each damage case were first computed, and, as expected, the distributions appeared Gaussian. The mean values from these distributions

along with 95% data quantiles for AR model prediction error plotted as a function of the damage cases are shown in Fig. 11.19. Not surprisingly, damage scenarios 1 and 5 show up clearly as they both involved the removal of the bolts at joint D (the location of channel 3). Other damage scenarios show less deviation from the baseline. The PDFs of the prediction errors also show that larger damage results in a larger variance associated with the prediction. As the baseline model breaks down, not only do predictions become less accurate, but the errors have a much wider spread. This result is consistent with those of previous observations using prediction error metrics in SHM.

This simple implementation of an AR approach is effective in diagnosing damage-induced changes to the dynamics. More complicated variants of the AR modeling can be used for SHM purposes including addition of exogenous inputs (ARX model, see Sohn et al. 2001), addition of a moving average term to better account for noise (ARMA, ARMAX models, see Fassois 2001), and the extension to multivariate signals (see Bodeux and Golinval 2003 for an ARMAV approach to SHM). Building empirical models and then searching for model breakdown as a damage indicator can be an effective method for monitoring complicated structures for which no analytical model exists.

11.4.5. Chaotic-attractor-based prediction error

Many of the time series methods used to do structural damage assessment rely upon analyzing transient dynamic or stochastic time series, e.g., on using modal analysis (the modal method) or the AR modeling discussed in the previous subsection. Nichols et al. (2003a,b) and Todd et al. (2001) have proposed assessing the steady-state dynamics imposed by a deterministic waveform. Specifically, these authors used a chaotic waveform to excite a structure and then developed methods for comparing the resulting attractors obtained as the structure became damaged over time.

A variety of metrics exist for detecting differences in dynamical attractors. Much work has focused on using the correlation dimension (see Logan and Mathew 1996; Craig et al. 2000; Wang et al. 2001), although this method's numerous important shortcomings and pitfalls were discussed in Nichols (2003c). In a supervised learning context similar to that described in the last subsection, the idea is to use data taken from the “pristine” structure to reconstruct reference, or baseline, attractors. These data are to be compared to the data collected from the structure at subsequent times,

when damage may or may not have occurred to the structure. By using a simple attractor geometry prediction scheme, points on the “damaged” attractors are forecast by using the baseline data as a model. Similarly to how the errors is used in the AR approach, the error formed between prediction and measurement is used here to detect and quantify damage-induced changes in the dynamics. The models make no assumptions about the underlying system, in contrast to the models used in AR modeling techniques, which implicitly assume a response coming from a structure for which a linear model is an accurate descriptor.

As with the AR approach, the method begins by recording a set of N_r baseline time series when the structure is in its baseline condition. This baseline set of experiments is represented by $X = \{x_1(n), x_2(n), \dots, x_{N_r}(n)\}$, where each $x_u(n)$, $u = 1, 2, \dots, N_r$, is a vector of discretely sampled values of structural response time series consisting of $n = 1, 2, \dots, N$ points. The discrete time index n refers to the value recorded at sample time $t_s = n\Delta t$, where $1/\Delta t$ is the sampling frequency. New “test” data are collected at some later time from the same location on the structure, denoted by $Y = \{y_1(n), y_2(n), \dots, y_{N_r}(n)\}$, in the same fashion that the set X was created. The next step is to use the baseline data X to empirically generate attractor-based models of the baseline structure’s dynamics and observe the degree to which they predict the dynamics of subsequent “test” datasets.

The algorithm used here is adopted from Schreiber (1997), where it was used to detect non-stationarity in time-series data. A simple attractor-based prediction scheme is used on the undamaged data to forecast the values of the damaged data some number of time steps s into the future. First, each of the baseline time series in X and Y are used to reconstruct attractors in phase space by using the embedding theorems (Takens 1981; Sauer et al. 1991). The method proceeds by comparing any of the baseline attractors to any of the “test” attractors in the following way: given a randomly selected trajectory with time index f on a reconstructed “test” attractor $\mathbf{y}(f)$, the algorithm selects the set of points on the baseline attractor $\mathbf{x}(n)$ that are within some radius ε of that trajectory

$$U_\varepsilon^{\mathbf{x}(p)} \mathbf{y}(f) = \mathbf{x}(p) : \|\mathbf{x}(p) - \mathbf{y}(f)\| < \varepsilon, \quad (11.22)$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. No temporal relationship need exist between the indices p and f , as the set is constructed purely by geometry; in other words, the $\mathbf{x}(n)$ are selected via their Euclidean distance to the fiducial point with the time index of that point passively

carried along for bookkeeping. This selection method contrasts with that of an autoregressive approach, which is based on temporal relationships. In fact, this current method may be thought of as regressing in phase space geometry, similarly to how an autoregressive approach regresses in time. The idea here is to describe the evolution of the neighborhood $U_\epsilon^{\mathbf{x}(p)}[\mathbf{y}(f)]$ and to use this description as a predictor for how subsequent data should evolve. The predicted value for $\mathbf{y}(f)$ at s time steps into the future, denoted $\hat{\mathbf{y}}(f + s)$, becomes

$$\hat{\mathbf{y}}(f + s) = \frac{1}{|U_\epsilon^{\mathbf{x}(p)}[\mathbf{y}(f)]|} \sum_{\mathbf{x}(p) \in U_\epsilon^{\mathbf{x}(p)}[\mathbf{y}(f)]} \mathbf{x}(p + s), \quad (11.23)$$

where the quantity $|U_\epsilon^{\mathbf{x}(p)}[\mathbf{y}(f)]|$ denotes the number of points in the neighborhood. The predicted value is the average of the predicted values for the neighborhood. In this sense, the baseline attractors are used as “look-up” tables which contain the various patterns present in the data. The working hypothesis is that these tables will lose their ability to serve as an accurate database as the dynamics are altered by damage. The prediction horizon s will depend on the rate at which the data are acquired and the specific application. For health-monitoring purposes, $s = 1$ will suffice if the data is reasonably sampled. While more complicated prediction schemes exist, this model is among the simplest ones that can be used to quantify the evolution of the dynamics. Because simple attractor discrimination is the objective, the absolute quality of the predictions is of diminished importance, and the simplest, most computationally efficient scheme is considered optimal. Variations of this algorithm have been used for prediction and data cleansing (see Kantz and Schreiber 1999; Nichols and Nichols 2001).

Once the predictions has been made, the s -step prediction error for trajectory f is quantified by

$$\gamma = \frac{1}{N} \sum_{j=1}^N \|\hat{\mathbf{y}}(f + s) - \mathbf{y}(f + s)\| / \sigma_x^2, \quad (11.24)$$

where σ_x^2 is again the baseline variance. Because the forecast is made in multi-dimensional phase space, the Euclidean norm is used to resolve the error to a single component. Equation (11.24) is simply the multi-dimensional analogue to (11.21). This process is repeated for some randomly selected subset of the total number of points on the attractor, resulting in a vector of prediction errors for each pair of attractors under consideration. In contrast to the one-dimensional predictions made with AR modeling (i.e.,

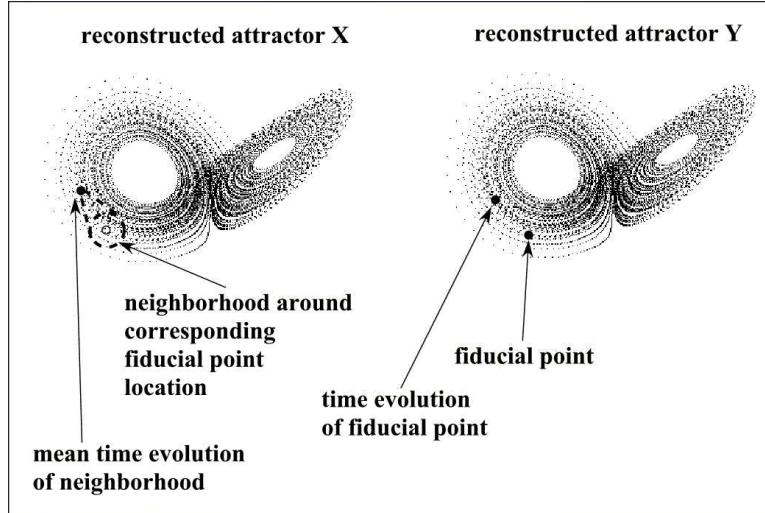


Figure 11.20: A schematic of attractor-based prediction error

a single time series at a time), this approach is multi-dimensional, incorporating information from all the degrees of freedom by using the embedding approach.

In computing this feature, the auto-prediction error must also be computed by replacing $\mathbf{y}(f)$ with $\mathbf{x}(f)$ in (11.23). The resulting γ gives some idea of the prediction error one would expect to find if the dynamics were not changing. In order to make these “auto-comparisons,” duplicate pairings must be eliminated. For example, comparing \mathbf{x}_1 and \mathbf{x}_2 , for example, followed by comparing \mathbf{x}_2 and \mathbf{x}_1 could potentially bias the data, so such a pairing is disregarded in the second comparison. Although comparisons between data from different damage scenarios do not suffer this effect (each baseline attractor \mathbf{x} could be used to predict each test attractor \mathbf{y}) the constraint is maintained across all cross-predictions for consistency. A schematic of this algorithm is shown in Fig. 11.20.

This procedure was applied to the test structure described previously by using the x -coordinate [$y_1(t)$ in (11.6)] of a chaotic Lorenz oscillator as the excitation waveform. The structural response attractors were embedded in four dimensions with a delay of five time steps. For each baseline/test attractor pair, 3 276 predictions were made, resulting in 32 760 total values of prediction error across the multiple runs that should include ambient variation in the environmental conditions for statistical robustness. Each

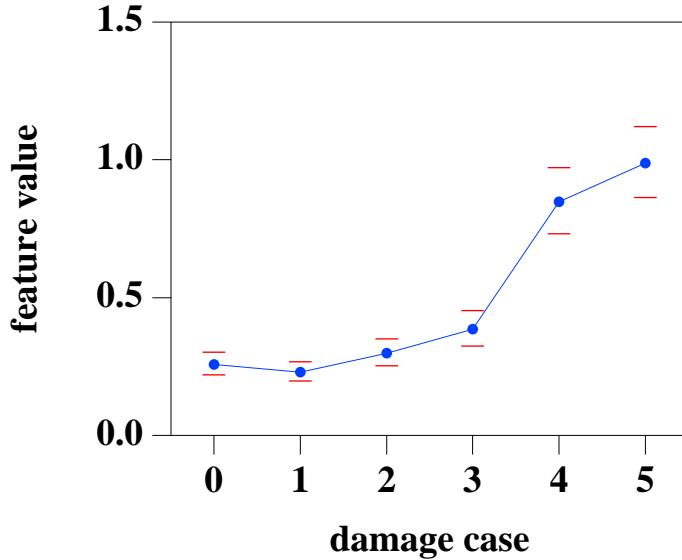


Figure 11.21: Chaotic attractor-based prediction error as a function damage cases.

of the prediction error feature sets was resampled by using the mean value from 1000-member random subsets to generate 5000 “new” feature values. The resulting estimated probability density functions for the channel 3 data, similar to those for the AR model described in the previous subsection, are also fairly Gaussian after the resampling procedure. The mean values from these distributions along with 95% data quantile limits are shown in Fig. 11.21. In general, there is a reasonable separation of the PDFs, especially at damage cases 3–5, so that the classification of data into these damage cases is possible. This possibility is further reflected in the quantile limits, which do not overlap across the baseline condition and damage cases 3, 4 and 5. However, cases 4 and 5 cannot be distinguished from each other, and other data channels or a multivariate analysis combining data from all the channels was used to increase the level of sensitivity.

11.5. Summary and conclusions

We have presented a unique methodology for assessing structural damage. This method is based on instantaneous phase functions extracted from the measured time series by using EMD and IMFs. The characteristics of the damage features are then combined with phase dereverberation, a concept

derived from wave mechanics. An instantaneous phase function of the measured time series is interpreted as a wave response traversing through the structure under investigation. As a structure becomes damaged, the nature of the wave response changes and results in phase lags between designated locations on a structure when it is compared with its baseline state. Numerical examples and experimental validation for a simple one-dimensional system with, and without, damage are used to illustrate the effectiveness of the method.

Experimental studies of a simple frame structure with multiple damage cases are also presented. A simple damage-feature value (damage index) is constructed based on the average phase lags to detect frame damage for a variety of damage states. The damage-feature value is a simple and effective index that distinguishes all the damage cases presented in section 11.4. Figures 11.18, 11.19, and 11.21, as described in subsections 11.4.3–11.4.5, are calculated by using entirely different time series analysis methods or different types of data-driven modeling techniques. The quantities used by these methods (feature values or damage indexes) to identify damage are also very different. As a result, the numerical values in the vertical axes are very dissimilar in these three figures. However, as in any data-driven approach, structural damage is identified based on comparisons of the feature-value change from its baseline state. In this sense, the results obtained in the three figures are quite similar for most of the damage cases studied. In particular, the EMD and phase-detection method's results shown in Fig. 11.18 are similar to the AR modeling results in Fig. 11.19 except for those for damage case 4. Figure 11.18 is also similar to Fig. 11.21 (chaotic prediction error) except for damage case 1.

The EMD and instantaneous phase detection method does not require predefined decomposition basis functions and does not assume linear system behavior or the type of input (transient or steady state). This method, using adaptive time-varying signals (IMFs) as basis functions, can fully accommodate the physics of a system's behavior. In particular, this method is suitable for detecting structural damage by using signals measured during extreme loading events, such as earthquakes and shocks, where the non-linear response associated with structural damage may be important to consider.

References

- Akaike, H., 1969: Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, **21**, 243–249.
- Bodeux, J. B., and J. C. Golinval, 2003: Modal identification and damage detection using the data-driven stochastic subspace and ARMAV methods. *Mech. Syst. Signal Process.*, **17**, 83–89.
- Chang, F.-K., 2001: *Structural Health Monitoring 2000*. CRC Press, 1062 pp.
- Chang, F.-K., 2003: *Structural Health Monitoring 2003: From Diagnostics & Prognostics to Structural Health Management*. DEStech Publications, Inc, 1552 pp.
- Cohen, L., 1995: *Time-Frequency Analysis*. Prentice Hall, 299 pp.
- Craig, C., R. D. Neilson, and J. Penman, 2000: The use of correlation dimension in condition monitoring of systems with clearance. *J. Sound Vib.*, **231**, 1–17.
- Doebling, S. W., C. R. Farrar, M. B. Prime, and D. W. Shevitz, 1996: *Damage Identification and Health Monitoring of Structural and Mechanical Systems from Changes in Their Vibration Characteristics: A Literature Review*. Tech. Report. LA-13070-MS, Los Alamos, NM, 118 pp.
- Doebling, S. W., C. R. Farrar, M. B. Prime, and D. W. Shevitz, 1998: A review of damage identification methods that examine changes in dynamic properties. *Shock Vib. Digest*, **30**, 91–105.
- Fassois, S. D., 2001: MIMO LMS-ARMAX identification of vibrating structures—Part I: The method. *Mech. Syst. Signal Process.*, **15**, 723–735.
- Fasel, T. R., S. Gregg, T. Johnson, C. R. Farrar, and H. Sohn, 2002: Experimental modal analysis and damage detection in a simulated three story building. *Proc. IMAC XX: Conf. on Struct. Dynam.*, Los Angeles, CA, Soc. Exper. Mech., 590–595.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Kantz, H., and T. Schreiber, 1999: *Nonlinear Time Series Analysis*. Cambridge University Press, 304 pp.
- Logan, D., and J. Mathew, 1996: Using the correlation dimension for vibration fault diagnosis of rolling element bearings. *Mech. Syst. Signal Process.*, **10**, 241–264.

- Ma, J., and D. J. Pines, 2003: Damage detection in a building structure model under seismic excitation using dereverberated wave mechanics. *Eng. Struct.*, **25**, 385–396.
- Nichols, J. M., and J. D. Nichols, 2001: Attractor reconstruction for nonlinear systems: A methodological note. *Math. Biosci.*, **171**, 21–32.
- Nichols, J. M., M. D. Todd, and J. R. Wait, 2003a: Using state space predictive modeling with chaotic interrogation in detecting joint preload loss in a frame structure experiment. *Smart Mater. Struct.*, **12**, 580–601.
- Nichols, J. M., M. D. Todd, M. Seaver, and L. N. Virgin, 2003b: Use of chaotic excitation and attractor property analysis in structural health monitoring. *Phys. Rev. E*, **67**, Art. No. 016209.
- Nichols, J. M., L. N. Virgin, M. D. Todd, and J. D. Nichols, 2003c: On the use of attractor dimension as a feature in structural health monitoring. *Mech. Syst. Signal Process.*, **17**, 1305–1320.
- Owen, J. S., B. J. Eccles, B. S. Choo, and M. A. Woodings, 2001: The application of auto-regressive time series modelling for the time-frequency analysis of civil engineering structures. *Eng. Struct.*, **23**, 521–536.
- Pines, D. J., and L. W. Salvino, 2002: Health monitoring of one-dimensional structures using empirical mode decomposition and the Hilbert-Huang transform. *Proc. 9th Annu. SPIE Smart Struct. Mater. Conf.*, San Diego, CA, SPIE, 127–143.
- Pines, D. J., and L. W. Salvino, 2004: Structural health monitoring using empirical mode decomposition and the Hilbert phase. *J. Sound Vibr.*, in press.
- Purekar, A. S., and D. J. Pines, 2000: Detecting damage in non-uniform beams using the dereverberated transfer function response. *Smart Mater. Struct.*, **9**, 429–444.
- Salvino, L. W., 2000: Empirical mode analysis of structural response and damping. *Proc. 18th Int. Modal Analysis Conf.*, San Antonio, TX, Soc. Exper. Mech., 503–509.
- Salvino, L. W., and D. J. Pines, 2002: Structural damage detection using empirical mode decomposition and HHT. *6th World Multi-Conf. on Systemics, Cybernetics and Informatics*, Orlando, FL, Int. Inst. Informatics Systemics (IILS), 293–298.
- Salvino, L. W., D. J. Pines, M. D. Todd, and J. M. Nichols, 2003: Signal processing and damage detection in a frame structure excited by chaotic input force. *Proc. 10th Annu. SPIE Smart Mater. and Struct. Conf.*, San Diego, CA, SPIE, Vol. 5049, 639–650.
- Sauer, T., J. A. Yorke, and M. Casdagli, 1991: Embedology. *J. Stat. Phys.*,

- 65, 579–616.
- Schreiber, T., 1997: Detecting and analyzing nonstationarity in a time series using nonlinear cross predictions. *Phys. Rev. Lett.*, **78**, 843–846.
- Sohn, H., C. R. Farrar, N. F. Hunter, and K. Worden, 2001: Structural health monitoring using statistical pattern recognition techniques. *J. Dynam. Syst.*, **123**, 706–711.
- Takens, F., 1981: Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence*, D. A. Rand and L. S. Young, Eds., Lecture Notes in Mathematics, Vol. 898, Springer-Verlag, 366–381.
- Todd, M. D., J. M. Nichols, L. M. Pecora, and L. N. Virgin, 2001: Vibration-based damage assessment utilizing state space geometry changes: Local attractor variance ratio. *Smart Mater. Struct.*, **10**, 1000–1008.
- Wang, W. J., J. Chen, X. K. Wu., and Z. T. Wu, 2001: The application of some non-linear methods in rotating machinery fault diagnosis. *Mech. Syst. Signal Process.*, **15**, 697–705
- Yang, J. N., Y. Lei, S. Lin, and N. Huang, 2004a: Hilbert-Huang based approach for structural damage detection. *J. Eng. Mech. Div. (Amer. Soc. Civ. Eng.)*, **130**, 85–95.
- Yang, J. N., Y. Lei, S. L. Lin, and N. Huang, 2004b: Identification of natural frequencies and damping ratios of in-situ tall buildings using ambient vibration data. *J. Eng. Mech. Div. (Amer. Soc. Civ. Eng.)*, **130**, 570–577.

Liming W. Salvino

Code 652, Carderock Division, Naval Surface Warfare Center, West Bethesda, MD 20817-5700, USA
salvinolw@nswccd.navy.mil

Darryll J. Pines

Smart Structures Laboratory, Alfred Gessow Rotorcraft Center, University of Maryland, College Park, MD 20742, USA
djppter@eng.umd.edu

Michael Todd

Department of Structural Engineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0085, La Jolla, CA 92093-0085, USA
mdt@ucsd.edu

Jonathan M. Nichols

Code 5673, U.S. Naval Research Laboratory, Washington, DC 20375, USA

pele@ccs.nrl.navy.mil

CHAPTER 12

HHT-BASED BRIDGE STRUCTURAL HEALTH-MONITORING METHOD

Norden E. Huang, Kang Huang and Wei-Ling Chiang

A new Hilbert-Huang transform (HHT)-based method for nondestructive instrument structure health monitoring is developed. When applied to bridges, this new method depends on a transient test load and simple data collection. The essence of the method is the newly developed HHT for nonstationary and nonlinear time series analysis, which consists of the empirical mode decomposition and Hilbert spectral analysis. The final decision on the health of the bridge structure is based on two criteria. The first criterion detects the nonlinear characteristic of the intra-wave frequency modulations of the bridge response, which usually appears during comparisons of light with heavy loads. The second criterion detects the frequency downshift as an indication of structural yield. This new method enjoys many advantages: no *a priori* data required, simple data collection, minimum traffic disruption, and precise and nuance quantitative answers. The result of a case study is presented, which establishes the feasibility of this new approach for structural health monitoring.

12.1. Introduction

With the inevitable aging of civil infrastructures, structural health monitoring has become an urgent problem worldwide (Chang 1997, 1999, 2001). To safeguard the safety performance of a bridge, regular inspections are essential (see, for example, Mori and Ellingwood 1994a,b). At the present time, the inspection method is primarily visual: a technician has to go through a bridge to examine each member and certify its safety. This method is subjective and flawed, lacking rigorous and objective standards. For example, for a bridge deteriorating from fatigue or aging, the damage is not clear-cut at any time. Therefore, any call is judgmental. Furthermore, using this method for complicated bridge structures is not feasible: some members might be located at positions too awkward to access; the members might require too much time to inspect; and the damage might be too subtle to

be detected visually. Because of these limitations, the visual-inspection results are known to be not totally reliable, yet we are forced to rely on them today.

According to Doebling et al. (1996, 1998), an ideal inspection method would have to satisfy the following conditions:

- (1) be robust, objective, and reliable,
- (2) be able to identify the existence of damage,
- (3) be able to locate the damage,
- (4) be able to determine the degree of damage, and
- (5) be able to provide data to estimate the remaining period of service.

The visual inspection method certainly fails the first requirement, and thus creates uncertainties in the remaining four conditions. These requirements lead us to conclude that non-destructive inspection methods that employ precise scientific sensors coupled with rigorous data analysis would be clearly preferred if available. Developing alternatives for the visual inspection method has been the central theme of the research in the Bridge Management Program, Turner-Fairbank Highway Research Center of the Federal Highway Administration (Chase and Washer 1997). A large program of research and development in new technologies for the nondestructive evaluation of highway bridges has been initiated. The objectives are to locate, quantify, and assess the degree of damage in bridges. Although various technologies, such as infrared thermography, ground-penetrating radar, acoustic emission monitoring, eddy current detection and others have been developed and are feasible, none of them is really practical. The difficulties of these systems are many. The first difficulty is due mainly to their limited field of view. One would have to locate the damage first before using the sophisticated imaging devices to examine it in detail. For a complicated structure, locating the damage is difficult enough. Secondly, the more daunting task involves evaluating the damage in terms of structural safety. Relating what can be measured in terms of structural safety is an almost impossible task without dynamic tests. As a result, even with the advances made by these sophisticated and esoteric techniques, their complexity and expense prevent their routine use. The data used routinely in bridge management today are still based almost entirely upon the unreliable visual-inspection methods, followed by expensive static load testing. The only viable alternative lies in the structural damage identification and health monitoring from changes in the bridge's vibration characteristics (Salawu 1997; Farrar et al. 1999), but past efforts to use vibration as a structural health-monitoring

tool have met with great difficulties, mainly because of the lack of a better data-analysis method. The New Nondestructive Instrument Bridge Safety Inspection System (Huang 1998a) is based on such a new data-analysis method: the Hilbert-Huang transform (HHT), which consists of empirical mode decomposition (EMD) and the Hilbert spectral analysis (HSA) developed by Huang et al. (1996, 1998d, 1999, 2003) and Wu and Huang (2004). This new data-analysis method immediately found applications in a wide variety of geophysical engineering and biomedical problems (Huang 1998b,c; 1999). The details of the specific application to bridge-safety inspection have been described in a patent filed by Huang (1998a).

For bridge-safety inspection, Huang (1998a) proposed to use a transient load and to examine the nonlinear characteristics in the bridge vibration data to identify the damage. Huang et al. (1998d) clearly pointed out that, for a faithful representation of the nonlinear and nonstationary data, a different approach other than Fourier or Fourier-type wavelet analysis is needed. Huang's (1998a) approach is based on Huang et al.'s (1998d) newly developed HHT, which was designed specially for nonlinear and nonstationary data analysis. This new method requires two steps for analyzing data: the first step is to use the EMD method, with which any data can be decomposed, according to the characteristic scales, into a number of intrinsic mode functions (IMFs). In this way, the data are expanded in a basis derived from the data. The second step, the HSA, is to apply the Hilbert transform to the IMF components and construct the time-frequency-energy distribution, designated as the "Hilbert spectrum." In this form, the time location of events will be preserved, for the frequency and energy defined by the Hilbert transform has intrinsic physical meaning at any point. Huang (1998a) used precisely these characteristics of this new data-analysis method. Before presenting the new method, it will be instructive to review the present state-of-the-art methods used for bridge inspection.

12.2. A review of the present state-of-the-art methods

The approach of using dynamic response and vibration characteristics for nondestructive structural damage identification is the theoretical foundation of instrumental safety-inspection methods. This approach has also been the mainstream of research for more than thirty years (Natke et al. 1993; Chang 1997). Doebling et al. (1996, 1998) and Salawu (1997) have reviewed the available literature on this approach, and Farrar and Doebling (1997,

1998) and Felber (1997) have reviewed the practical problems associated with it. In principle, each structure should have its proper frequency of vibration under dynamic loading. The value of this proper frequency can be computed based on the elasticity properties of the structure (see, for example, Clough and Penzien 1993 or Chopra 1995).

Sound as this argument is, the dynamic instrument inspection has never worked successfully. The reasons are many: first, the precision sensors needed to measure the detailed dynamic response of the structure under loading are lacking. Secondly, historical data about individual bridges do not exist. Thirdly, proper data-processing methods to process the structural response have not been developed. Finally, the sensitivity of the structure in global response due to the local damage is unknown, and this problem is further gravitated by the large built-in safety factor. For example, damage of up to 50% of the cross-section locally can result in only a few percentages of vibration-frequency change. Such a small frequency shift, when processed by using the conventional methods, would be totally lost in the inevitable noise in all real situations. In the final analysis, many of the difficulties can be alleviated if the data-processing method can be made more versatile to handle vibration signals from highly transient loads and nonlinear responding waveforms. This problem will be addressed presently.

According to the traditional approach, the proper frequency can be computed only through Fourier analyses, from which the time domain data are reduced to purely frequency domain results. In this approach, the data have to be assumed as stationary and linear. Under such a restriction, if one has a perfect record of the undamaged structure as a reference, one will be able to detect the change of the proper frequency and its harmonics, but such a result still will not reveal the location of the damage. In what follows, a review of the state-of-the-art data-analysis methods, loading conditions, and use of the transient load will be presented. As the data-processing method is the driving force in determining the testing conditions, including the sensor types and their deployment schemes, and loading strategy, these problems will be examined first.

12.2.1. Data-processing methods

As discussed by Huang et al. (1998d), using the Fourier analysis method for nonlinear and nonstationary data involves fundamental problems, yet for lack of alternatives, it is still used extensively, although seldom in its bare form as in Basseville et al. (1993) and Hanagud and Luo (1997). Fourier

analysis, however, appears in almost all other data-analysis methods such as the Wigner-Ville distribution, modal, wavelet, and even Hilbert analyses, as will be discussed later. In fact, the frequency determination from any data is almost always based on Fourier analysis, yet it is physically meaningful only for linear and stationary data. All real structures are seldom linear, especially when the structure is damaged to the extent where the response could be plastic. The lack of a nonlinear and nonstationary signal-processing method has made the random-vibration approach inconclusive as a method for non-destructive tests of bridges and other structures. Detailed summaries of the lessons learned from the Fourier approach are given by Salawu and Williams (1995a,b), Farrar and Doebling (1997) and Felber (1997). The applications of these Fourier-based methods are summarized as follows:

Modal analysis: The modal-analysis method was proposed as an adaptive approach for analyzing stationary random data (see, for example, Pandit 1991). When applied to analyzing deformations of a structure, the modes involved have been reduced drastically. Stationarity (or homogeneity) is assumed. This assumption should not be a serious problem for the lower modes. The most serious limitation is that the modal analysis depends on the global deformation of the whole structure. Consequently, deformation due to the local damages can be detected only in higher mode variations, but to determine the mode shape of the higher modes, a large number of sensors must be used to collect the necessary data. Even with the detailed data, the sensitivity is still low for local damages, due to the ubiquitous noise problem. As a result, this method is not very sensitive to the existence of damage (Farrar and Doebling 1997).

Recent developments, however, have alleviated some of the difficulties mentioned above. For example, Kim and Bartkowicz (1997) proposed a method with limited instruments; Stubbs and Kim (1996) suggested a method to infer the reference state from measured data; Vakakis (1997) proposed nonlinear normal mode expansion; and Fahy (1994) and Doebling et al. (1997) suggested an energy-based method to improve damage location. All these improvements notwithstanding, the real test of bridges and large structures by using modal analysis still presents great problems (Alampalli et al. 1997; Juneja et al. 1997). Even with good reference data, the noise from the real system and measurements can still cloud the picture and render the detection and location of the damage difficult. Another serious drawback of modal analysis is its requirement for prior analytical or test data of the undamaged state as a reference. Usually, such data are

not available. The main problem when using modal analysis to locate damage is the requirement for the higher mode of deformation, which requires very detailed deformation measurements from many sensors. Therefore, this method is quite expensive and complicated to implement. Furthermore, noise removal and simplification of the structural deformation to a finite number of modes both present problems (Kim and Stubbs 1996). To overcome them, modal analysis is usually conducted jointly with Wigner-Ville distribution methods, or wavelet analysis as well.

Wavelet analysis: The wavelet analysis method is an adaptive window Fourier analysis (see, for example, Chui 1992); therefore, it can accommodate nonstationary but not nonlinear data (Dalpiaz and Rivola 1997). It is well known that the discrete wavelet is not useful for extracting special features from data. On the other hand, the continuous wavelet suffers redundancy and can hardly give quantitative results. Furthermore, wavelet analysis is still a Fourier-type transform; therefore, its use makes sense only when the data are from linear systems. Any nonlinear distortion of the waveform will require harmonics to be involved. Harmonics are mathematical entities with no physical meaning. Another problem of the wavelet approach is the uncertainty principle: the fundamental flaw of continuous wavelet analysis is the conflicting requirement of localization (with a narrow window) and frequency resolution (with a wide window) as discussed by Huang et al. (1998d). Even with these flaws, wavelet analysis has been used by Surace and Ruotolo (1994), Staszewski et al. (1997), Basu and Gupta (1997), Al-Khalidy et al. (1997), and Hou et al. (2000) to detect damage in various structures. Due to the poor frequency resolution, this detection method suffers a serious signal-to-noise ratio problem (Al-Khalidy et al. 1997). One possible usage of the wavelet analysis is to detect the singularity on the signal due to a sudden change of signal properties. Such changes, however, are rare in the bridge-damage problem.

Wigner-Ville distribution: The Wigner-Ville distribution was thoroughly discussed by Cohen (1995). Brancaleoni et al. (1993) and Feldman and Braun (1995) have tried to use it in damage detection with some limited success. As Wigner-Ville distribution is also Fourier-based, it suffers all the shortcomings of the Fourier analysis. Furthermore, its result is not strictly local; therefore, its ability to identify the location of the damage is also limited.

Neural network: Application of the neural network technique found application in damage detection as early as 1991, by Wu et al. (1991). Many investigators have extended the application, such as Tsou and Shen

(1994), Manning (1994), Pandey and Barai (1995), and Barai and Pandey (1995, 1997). Most of these applications train the program to construct the reference modes or to reduce the noise in the data (Barai and Pandey 1995); therefore, this application is still mode-based. Any drawbacks of modal analysis cannot be fully eliminated and can be only partially ameliorated. For a true solution, a method that can produce localized analysis as well as accommodate nonlinear variation in the data must be found. The Hilbert transform certainly fits these requirements, but it also has limitations. These will be discussed in the following section.

Hilbert transform: The application of the Hilbert transform to non-stationary data was proposed long ago (see, for example, Bendat and Perisic 2000). Its application in damage identification has been tried by Feldman (1991, 1994a,b), Feldman and Braum (1995), Braum and Feldman (1997), and Feldman (1997). In all these studies, the signal was limited to a “mono-component” condition, i.e., without the superposition of any smaller, riding waves, and the signals had to also be symmetrical with respect to the zero-mean. Thus, the applications were limited to cases of simple free vibrations. Although Prime and Shevitz (1996) and Feldman (1997) used Hilbert transforms to identify some of the nonlinear characteristics through the frequency modulation in a nonlinear structure, the limitations imposed on the data properties render the method of little practical use in both identifying and locating the damage. Among all the Hilbert transform applications, the most relevant one was due to Brancaloni et al. (1993) who has employed a transient load over a damaged bridge. Confronted by the limitations of a straightforward application of the Hilbert transform to arbitrary data (as discussed by Huang et al. 1998d), Brancaloni et al. resorted to filtering of the data to separate the data into different modes. As the filtering process is Fourier-based, it alters the nonlinear properties of the data drastically. Thus, the filtering method renders the results questionable. The real value of the Hilbert transform had to wait until Huang (1996) and Huang et al. (1996, 1998d) introduced the empirical mode decomposition (EMD) method as a pre-processing step, to be discussed presently. Before discussing the EMD method, another limitation of the present bridge-inspection method, the loading conditions, will be discussed.

12.2.2. Loading conditions

Because the previous data-processing methods are limited to mainly linear and stationary processes, the loading conditions will have to be designed to

produce datasets tailored to fit the available analysis methods. As reviewed by Salawu and Williams (1995a,b), two loading conditions are frequently used: free and forced vibrations.

Free vibrations: In the free-vibration test, the structure is not under any live load other than the one that triggers the vibrations at the beginning, and is free of load thereafter. That load can be an impulse or residual vibrations from a transient loading. The free vibrations of bridges are usually assumed to be linear but with time-varying amplitude. The global mean frequency can be determined to a high degree of accuracy with Fourier analysis, yet without a reference state from a healthy bridge, the frequency from the free vibration is not very informative. The traditional vibration analysis might work under special conditions when a structure becomes highly nonlinear due to damage. Then the Fourier analysis will show many harmonics as indicators of the nonlinear deformation of the vibration waveforms. Even then, for lack of the phase information of the harmonics, the various harmonics cannot be uniquely used to reconstitute the data; therefore, it will be impossible to determine the damage scenario or the degree of damage. Furthermore, the Fourier-based analysis also makes vibration analysis unable to determine the locations of the damage, for vibration analysis works only in frequency space.

Forced vibrations: In forced vibrations, the structure is under some loading throughout the period of vibration, which can be due to artificial or ambient forces. Ambient forces include those from the traffic, the wind, and earthquakes. All these loads are assumed to be linear and stationary, an assumption that is hardly ever true: since the wind force is ever fluctuating and is proportional to the squared velocity, this force is certainly neither linear nor stationary. The ground motions from earthquakes are never stationary; for a strong earthquake, these motions cannot only be nonstationary, but also highly nonlinear. One of the loading conditions is a special artificially-induced vibration from a point source of a vibrator. The data from such a condition, though relatively easy to interpret, are hard to generate effectively, for the application points are usually different from the unknown damage location (Felber 1997). Therefore, the force will not produce diagnostic data as effective and sensitive as the date produced by the forces that are applied just at the damaged spot. Whenever the ambient loads are uniformly applied to all the structure, the load will certainly visit all the damage locations, but the loading conditions are not controllable, and the nonstationary properties also make the data analysis difficult. Furthermore, the signal-to-noise ratio becomes a critical issue: if a light load

is considered as in microseisms, the response is not sensitive to the local damage. Only under a large load will the deformation be larger, and the responses nonlinear, but then, the random nature of the loading condition will make the signal (from the damage) and noise (from the ambient load) ratio too low to reveal the damage clearly.

The most effective loading condition should be the one with a transient load. This is equivalent to a point source applied to every load-bearing part of the structure. The data obtained from such a load, however, are certainly nonstationary. If the loading is up to the designed standard, the deformation of the damaged bridge should be linear. If the structure is damaged, the load-bearing capacity will decrease. Under such conditions, the structure will behave nonlinearly. Thus, a moving design load might produce nonlinear and nonstationary data for a damaged structure, a problem that the present available methods are unable to handle.

12.2.3. *The transient load*

Given all the shortcomings of the available methods, Huang (1998a) proposed a totally different approach: use a transient load and record the dynamic response, and then analyze the data by using the newly invented Hilbert spectral analysis (Huang 1996; Huang et al. 1998d). This method is designed for both nonstationary and nonlinear data.

Huang's (1998a) approach is based on the following two observations:

- (1) When a bridge suffers structural deficiency, the stress-strain relationship will go beyond the linear limit. Then, we can visualize that the vibration waveform should be deformed. As discussed by Huang et al. (1998d), this nonlinear deformation will show up as an intra-wave frequency modulation.
- (2) The response of the structure will be the strongest if the load is applied directly at the damaged location. This result is a logical consequence of the influence line, which represents the influence of the loading at any given point of the structure from the load applied at any other point of the structure.

The justifications of these claims are briefly summarized as follows: any structure under a design load should respond linearly and elastically. Under loading, the structure should reveal its proper frequency as well as free vibration. When the structure is damaged, its strength will decrease. Such a structure, even under the design load, will have an abnormally large

deformation and behave nonlinearly. The nonlinearity could be the consequence of the non-elastic response of the material, or could be due to the non-uniform cross-section of the load-bearing members that suffered damage. Because of the unique capability of the Hilbert-Huang transform for analysis of transient and nonlinear signals, the precise location of the damage can be determined from the time domain variations of the vibration characteristics without any *a priori* knowledge of the damage location.

With these observations in mind, the ideal method to detect damage at an unknown location will be a transient load, which will visit every point of the bridge. When this load is directly over the damage point, the response will be strongest. This loading condition can be easily implemented with a moving vehicle. Unfortunately, until now, such a loading condition has been actively avoided because no proper method has been available for analyzing the data obtained from this condition. With the invention of the HHT technique, the transient data are no longer a problem. Before proceeding further, the problems associated with nondestructive structural health-monitoring methods should be considered.

Limited by the data-analysis methods, the structural vibration data were always assumed to be linear and stationary. These assumptions are not true, especially when the structure is damaged. Even if one assumes that one has a complete knowledge of the structure, in fact, any knowledge will be limited and only empirical. Even if the complete design plan is available, the anomalies introduced in construction would make such a design plan an approximation at best. Because of the lack of complete knowledge of the structure, any model is also an approximation. When measurements of the structure are made, further limitations are imposed by the number of sensors available, accessibility to the structure everywhere, and noise in the data from other sources. Finally, the unbridgeable gap between the damage response and the damage scenario must be mentioned, as well as the low sensitivity of damage to loadings, and the fussiness of damage thresholds. All these difficulties point to a need to change the damage-detection paradigm. No perfect solution exists for the myriad problems, but an attractive alternative is now available: the HHT-based nondestructive method, which, though having not solved all the problems, has certainly ameliorated most of the difficulties listed here. As the HHT method is a central part of this approach, a brief summary is given below.

12.3. The Hilbert-Huang transform

As discussed by Huang et al. (1996, 1998d), the HHT method is necessary to deal with data from both nonstationary and nonlinear processes. Contrary to almost all the previous methods, this new method is intuitive, direct, and adaptive, with the basis of the decomposition based on and derived from the data. HHT consists of two parts: the empirical mode decomposition (EMD) and the Hilbert spectral analysis (HSA). Details of this procedure can be found in Huang et al. (1998d, 1999).

With the EMD's initial processing, any data can be decomposed into a finite set of intrinsic mode functions (IMFs). These IMF components are usually physically meaningful, for the characteristic scales are defined by the physical data. Additionally, we can also identify a new use of the IMF components, that of filtering. Traditionally, filtering is carried out in frequency space only, but applying any frequency filtering when the data are either nonlinear or nonstationary is very difficult, for both nonlinear or nonstationary data cause these methods to generate harmonics of all ranges in order to match the nonlinear shape and nonstationary occurrence of the data. Therefore, any filtering will eliminate some of the harmonics, and this result, in turn, will cause deformation of the data filtered. Using IMF, however, a time domain filtering can be devised. For example, a low pass filtered result of a signal having n -IMF components can be simply expressed as

$$x_{lk}(t) = \sum_{j=k}^n c_j + r_n; \quad (12.1)$$

a high pass result can be expressed as

$$x_{hk}(t) = \sum_{j=1}^k c_j; \quad (12.2)$$

and a band pass result can be expressed as

$$x_{bk}(t) = \sum_{j=b}^k c_j. \quad (12.3)$$

The advantage of this time domain filtering is that the results preserve the full nonlinearity and nonstationarity in the physical space. This Hilbert-Huang transform method has been used recently by an increasing number of investigators for structural health monitoring (Liu 1999; Yang et al. 2002, 2003a, 2003b, 2003c, 2004). The HHT approach will also be demonstrated

here for health monitoring of a bridge structure, but before providing the details, the damage-detection criteria must first be clarified.

12.4. Damage-detection criteria

Limited by the previous data-analysis methods, the past practice of damage identification was based primarily on modal analysis, as seen in the work of Doebling et al. (1996, 1998d), for example. The problems with modal analysis are as listed above. Here, the ability of the HHT to determine the instantaneous frequency precisely will be utilized. Furthermore, the ability of the HHT to distinguish between the linear vs. the nonlinear vibrations will also be used. This ability is crucial for the success of the HHT approach.

As the HHT can clearly define nonlinearly deformed waveforms, this definition will be used as the first indication of the existence of damage. The following approaches will be shown:

- (1) comparing the fundamental frequencies of the new and the current structure,
- (2) comparing the fundamental frequencies of the healthy and damaged structure, and
- (3) comparing the fundamental frequencies of light and heavy loadings.

All structures are designed to perform in the elastic limit. Therefore, the structure should behave linearly under the design load when it is newly finished and healthy. If such data are available, the data could serve as a valuable reference. In most cases, such data are not available, and using an old and healthy structure as a reference will also be impossible, as will be discussed presently. A new and practical approach introduced here is to utilize the differences between the linear and nonlinear responses. Doing so is critical, for the HHT can be used to identify nonlinear distorted waveforms. To take advantage of these characteristics of structure, both light and heavy loading on the structure will be used to generate different responses. If the structure is healthy, the response should be linear and have clean symmetrical sinusoidal waveforms irrespective of the loading conditions, as long as the load is within the design limit. If the structure is damaged, the response of the structure will then be different: under a light load, the structure might still behave linearly; once the stress increases, however, the damaged structure might behave nonlinearly. Then the vibration waveforms will be distorted, and the distortion will give intra-wave frequency modulation. The consequence is a broadening of the marginal Hilbert spectrum.

The nonlinear behavior of the structure is a clear warning sign of structural deficiency. Appropriate actions should be taken: limiting loads or speed, for example, might be necessary.

If the damage is more severe, the stiffness of the structure might be permanently changed. In that case, the fundamental frequency will change. To take the full advantage of the linear and nonlinear response under different loading conditions, a frequency ratio test must be done. A frequency ratio measurement can be defined as follows: let ω denote the frequency of a bridge, and ω_0 denote the frequency when the bridge is new; then the frequency ratio S is defined as

$$S = \frac{\omega}{\omega_0}. \quad (12.4)$$

According to Nishimura (1990),

$0.85 \leq S \leq 1.00$	<i>safe</i>
$0.70 \leq S \leq 0.85$	<i>caution</i>
$0.00 \leq S \leq 0.70$	<i>detailed inspection</i>

(12.5)

The rationale is simple: according to the beam theory, the fundamental frequency of the beam is proportional to the squared-root of the stiffness of the beam. Therefore, when the frequency ratio drops to 0.70, the stiffness should have reduced to 49%, a condition that certainly warrants a detailed inspection. Under such conditions, the original safety factor is reduced to unity.

As reasonable as these criteria are, they are unfortunately practically useless, for the frequency records for the mint state are available for few, if any, bridges. For bridges, a saving grace is that there are usually repeated spans, and all the spans are unlikely to suffer the same damage simultaneously. To utilize this fact, Li and Yen (1997) proposed a modification of the frequency ratio by using ω_0 as the frequency for a healthy, old but undamaged span. They also modified Nishimura's criteria to read

$0.90 \leq S \leq 1.00$	<i>safe</i>
$0.80 \leq S \leq 0.90$	<i>caution</i>
$0.00 \leq S \leq 0.80$	<i>detailed inspection</i>

(12.6)

This modification is reasonable, for all structures will soften slightly with age. Given the lack of data for a bridge in its mint state, using the slightly lower ω_0 and, therefore, a reduced ratio criterion is a logical alternative. Further research should be conducted to give this modification quantitative

validation. In summary, the variation of the instantaneous frequency will be used here to detect damage, for the variation of the instantaneous frequency can arise from the following two causes.

First, the variation is due to the nonlinear behavior of the bridge; damage will cause a change of stiffness of the structure. This damage may or may not be over the linear elastic limit, but the sudden change of stiffness will cause the vibrational waveform to deform or the frequency to decrease, and the results will introduce either intra-wave frequency modulations or frequency downshift. These are critical indicators of local structure damage. As discussed by Huang et al. (1998d) these intra-wave frequency modulations are a clear indication of nonlinear oscillations.

Second, the transient characteristic of the load can help us to locate the damaged spot. After the initial variation of the frequency due to the transient response when the test load first arrives on the bridge, the rest will be a forced vibration. As the test load passes along the bridge, this load will visit all the points, including the damaged spots, where the response should be the strongest as the test load moves over the damaged spots. Therefore, the first time the frequency goes nonlinear, or when it is below the free oscillation value, damage is indicated. The damage location is thus found from the time record and is identified when the testing load passes over the damaged location. This result is another crucial discriminator for damage detection.

12.5. Case study of damage detection

To demonstrate the use of the HHT analysis, Huang (1998a) used the deflection of a bridge under a transient load through computer simulation. The case studied was a simply supported bridge with a span of 8.5 m subjected to various damages under the standard design load. Clearly, the damaged bridge under the design load revealed the nonlinear characteristics of structural yielding at precisely the starting point when the load was over the damaged spot. From this time on, the period of oscillation of the damaged bridge was much longer than that of the undamaged bridge. This result was true also for the free oscillation after the truck had passed across the bridge. Fourier analysis of these data produced the spectra, which also showed the damaged bridge with a frequency down shift (indicating the increasing of the oscillation period). Since Fourier analysis totally lacks time information, there was no way one could determine the location of the damage from the Fourier analysis.

By considering a real bridge case study, the ability of this new approach with real measurements over a bridge with live transient loads can be demonstrated. This example also serves to show that data from a simple accelerometer can also be used for detecting damage. This demonstration will further show that HHT results can provide critical information for damage identification.

A bridge in southern Taiwan was selected for a field test. This bridge is a two-lane pre-stressed concrete girder bridge with 12 spans each 30 m in length. The girders are simply supported between piers, but the bridge deck is continuous with 15 cm reinforced concrete over three spans. There is a construction joint at every three spans.

During this test, a tri-axial force balance Kinematics EpiSensor accelerometer was used. For the bridge deck, the sensor was installed at the mid span; for the piers, the sensor was installed on the pier cap. This sensor has an extended bandwidth of DC to 200 Hz with a user-selected full-scale recording range covering ± 0.25 to ± 4 g. The output was field selected and set at the ± 2.5 V range.

The test loads employed were the regular traffic flow with vehicles of various sizes and speeds. The procedure of the test was to wait for a single vehicle to pass by and record the vibration generated by it. By recording the various sizes of vehicles, a comparison of the bridge vibration characteristics under different loading conditions could be made. Figures 12.1a–c give the recorded vertical component of the acceleration as recorded by the accelerometers located at the middle of span numbers 2, 3 and 4. The recorded acceleration signals vary over a large range. In fact, the signal from a small light car is not even as big as the noise of the large heavy truck. Light or heavy, all transient vehicles triggered the bridge system's vibration signals. This result could have been realized only as a consequence of the transient loading. According to the linear vibration theory, if the loading is stationary, the system should respond with the frequency of the forcing function.

An example of the Hilbert spectrum for the heavy loads over span number 4 is given in Fig. 12.2. This figure reveals that the strongest bridge responses under the transient loads are concentrated around the 4 Hz range. Although concentrated around 4 Hz for this load, the energy distribution is diffused under the heavier loads, indicating the onset of intra-wave frequency modulations. The diffusing of energy is more obvious in the marginal spectra to be discussed next.

According to the beam theory, the response frequency should be a func-

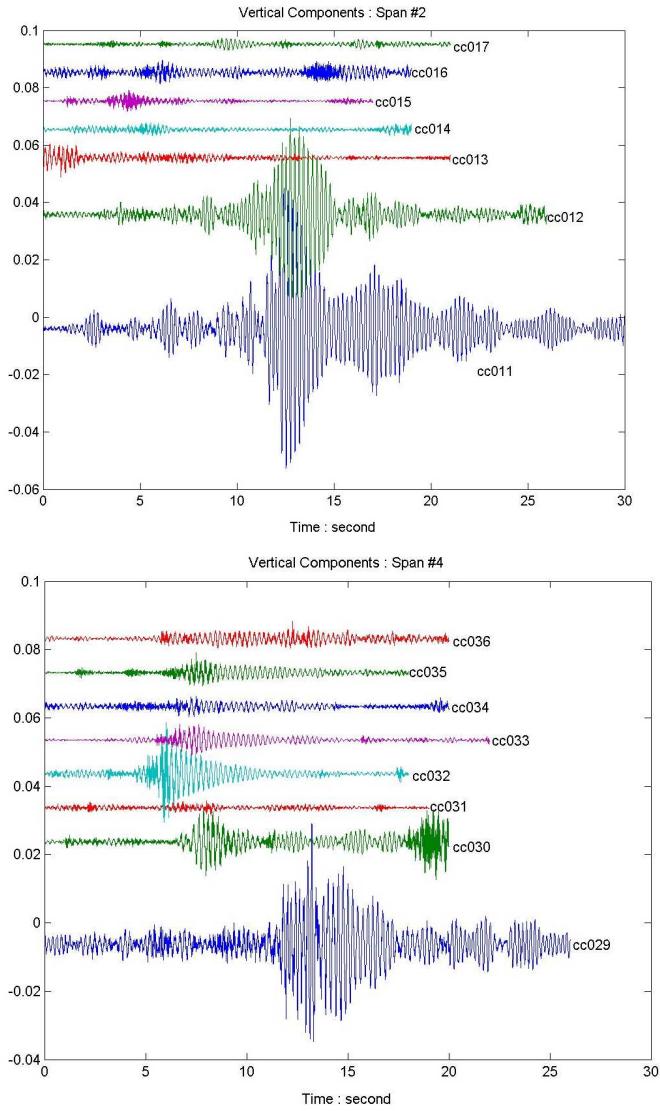


Figure 12.1: The raw vibration data collected with an accelerometer at the mid-span.
(a, top) From Span 2. (b, bottom)

tion of the elastic modulus and the moment of inertia of the cross-section of the beam. When the reaction of the beam is linear, the elastic modulus, or the ratio of stress-strain, should be constant. If the beam suffers any

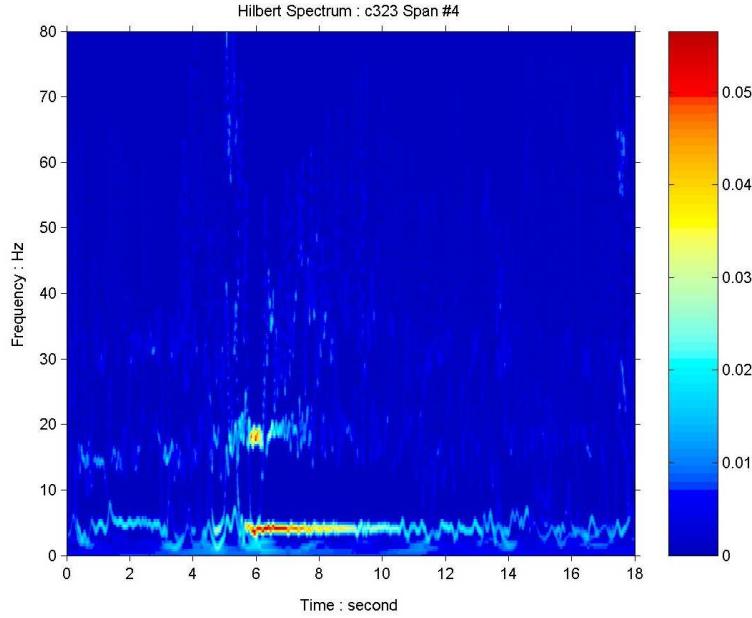


Figure 12.2: An example of the Hilbert spectra for the acceleration data for span 2.

damage, it will become softened; i.e., the beam will deform more than the undamaged span under the same load. Thus, the reaction becomes nonlinear. Based on these considerations, the determination of whether the reaction of a structural beam is or is not linear should be possible by examining the reaction of the beam under varied loading conditions. When the load is light, the structure will usually react linearly. If the structure is sound, the reaction of the beam should be linear even under a heavy load as long as the loading is within the design limit. On the other hand, if the structure is damaged, the beam will become softened permanently or yield exceedingly under heavy loads. Then the reaction will vary according to the loading condition: the reaction might be linear under a light load, but become nonlinear under a heavy load.

Following these observations, the Hilbert spectra of the beam under different loads for the two representative spans can be examined. The results of the marginal spectra for spans 2 and 4 are shown in Figs. 12.3a–b. Figure 12.3a for span 2 is clearly linear under all the loads. For span 4, the peak of the energy is still centered around 4 Hz. Only under the heaviest load, does the spectral form widen slightly, indicating a nonlinear reaction by the onset

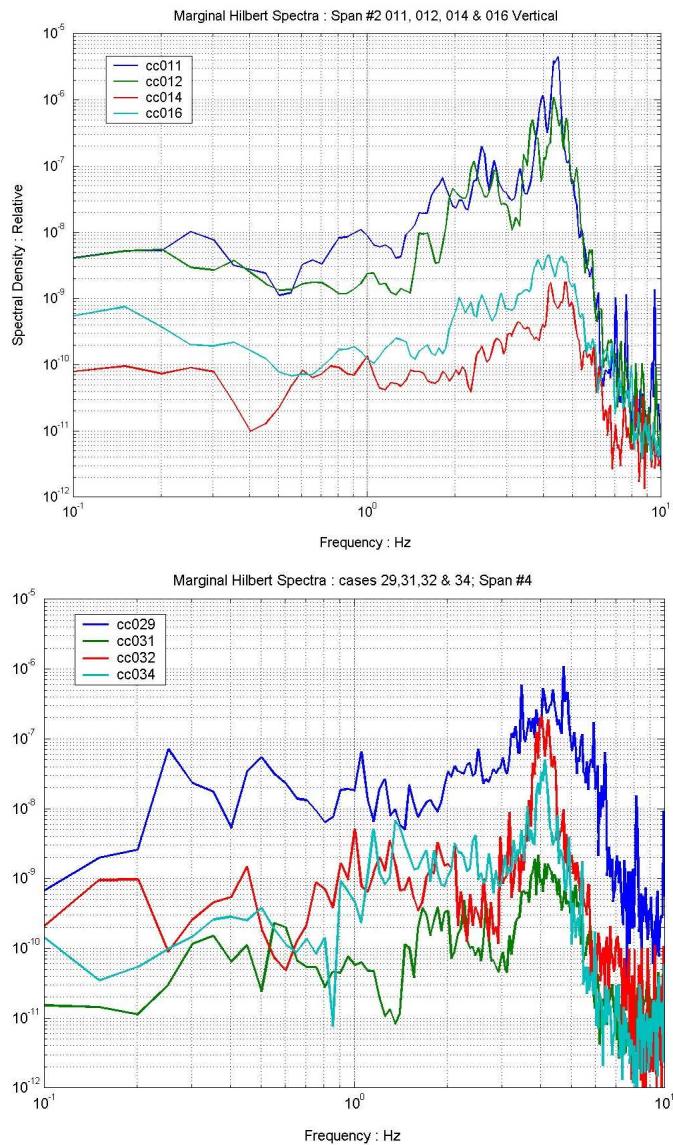


Figure 12.3: The marginal Hilbert spectra for the acceleration data. (a, top) Span 2. (b, bottom) Span 4.

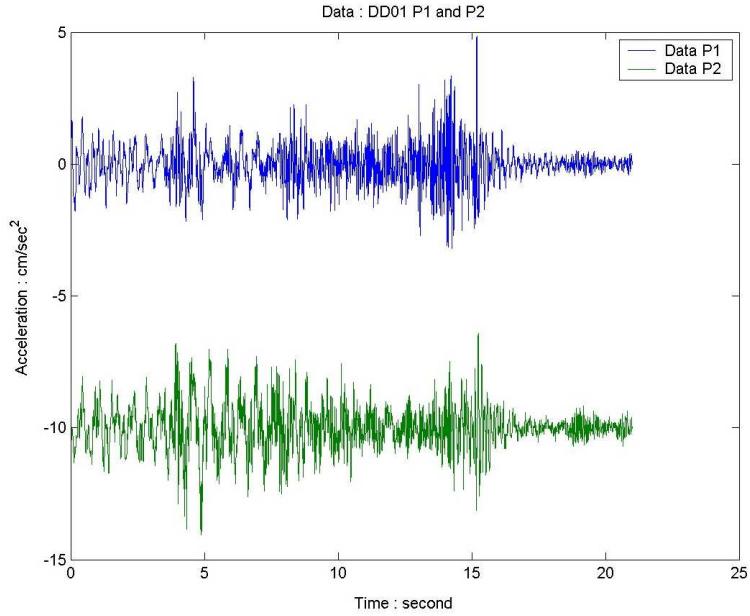


Figure 12.4: The raw vibration data collected with an accelerometer at the pier caps #1 (top) and #2 (bottom) under a heavy load. The “P2” data has been offset by –10.

of intra-wave frequency modulations: this reaction indicates the deformation of the vibration waveform under the heavy load. As the central value stays at the same location, the structure should still be sound, although it is under extreme stress, and care should be paid to its operational condition. Perhaps a limitation on load and/or speed should be considered.

Next, the test for the piers will be demonstrated. The data obtained when a heavy truck passed over the bridge are given for both pier 1 and 2 in Fig. 12.4. These data do not indicate any drastic difference between the different piers. Figures 12.5a–b show the Hilbert spectra of the two piers under an identical heavy load. Notice the time-frequency variation with the energy concentrated around 1.5 Hz. At around 5 s after the onset of data taking, a very high energy density in both the spectra appears. These densities represent the time when the load is over the pier. On careful examination of the frequency difference, one can detect the slight decrease of frequency for pier number 2 relative to pier number 1. The difference shows up even more clearly when we examine the marginal Hilbert spectra together with the Fourier spectra in Fig. 12.6. The Fourier spectra reveals no

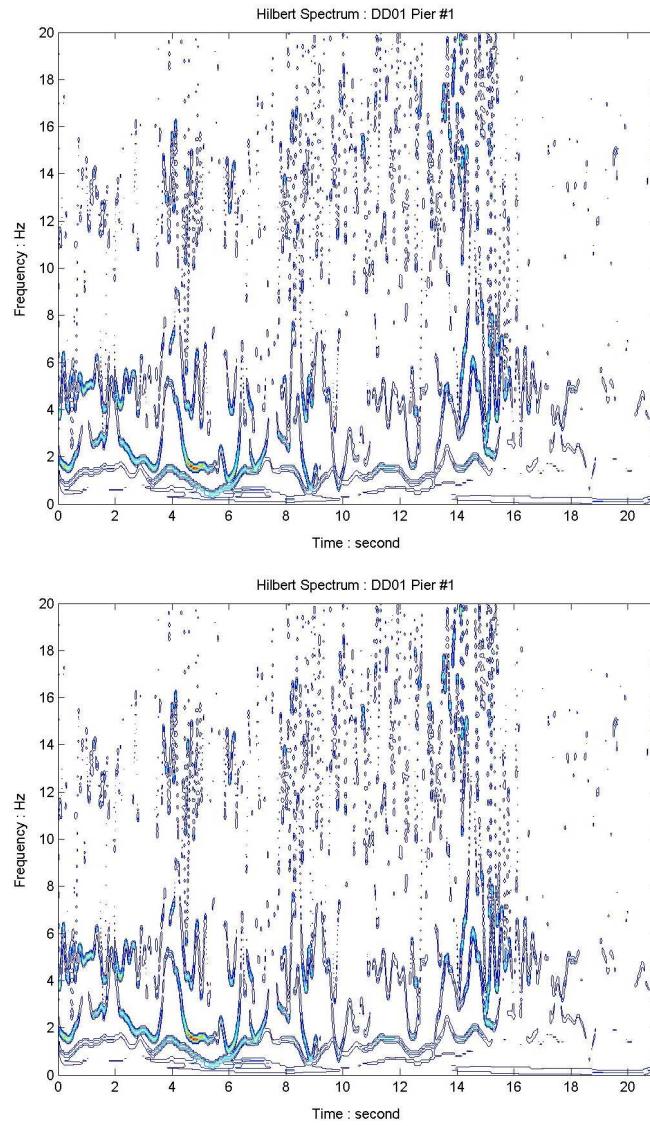


Figure 12.5: The Hilbert spectra for the acceleration data. (a, top) Pier 1. (b, bottom) Pier 2.

substantial difference between the two piers. The spectra might have small amplitude changes, but unfortunately, amplitude is not a good discrimi-

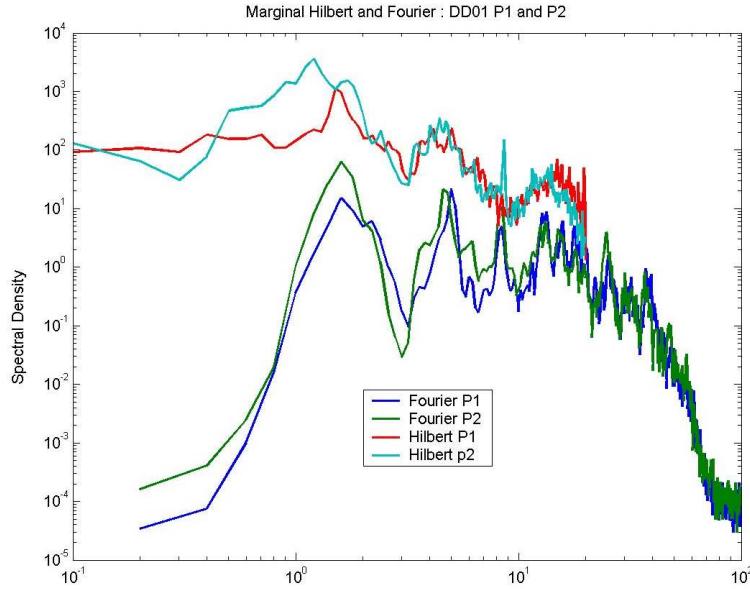


Figure 12.6: The marginal Hilbert and the Fourier spectra from the acceleration data collected at the pier caps. Note the frequency downshift revealed from the Hilbert spectra, but not in the corresponding Fourier spectra.

nator for problems. The Hilbert marginal spectra, however, shows drastic changes: the peak frequency for pier 2 is at 1.2 Hz, while that for pier 1 is at 1.6 Hz, a clear downshift of the vibration frequency and a clear indication of softening of the structure. As the frequency is proportional to the square root of the stiffness of the structure, the ratio of 1.6 Hz to 1.2 Hz indicate that the stiffness has changed in proportionality from 16 to 9, a reduction of almost 50%. Assuming that pier 1 is healthy but not in mint condition, the frequency ratio is only at 0.75, a value already under the empirical criterion set by Li and Yen (1997) and Li et al. (2003). Visual inspection of the pier, as shown in Fig. 12.7, indeed found the piling of pier 2 exposed through erosion, a common problem of bridges in Taiwan, where rivers are short and riverbeds are steep, so that the rains coming with typhoons always produce torrential flash floods. The reduction of the stiffness for pier 2 is a serious condition, warranting immediate remedial actions. Bridge scouring is also a serious problem in many locations, which contributes to many a bridge failure. The collapse of the Kaoping Bridge in 1990 is an example.



Figure 12.7: A photograph of pier #2. Note the pilings exposed by scouring.

12.6. Conclusions

The HHT method for structural health monitoring introduced here is both new and practical. The case study serves to illustrate the idea of data collection for safety inspection: use a constant speed transient load that will pass through the bridge to produce forced and free vibrations. This case study is of one of the simplest bridges in the real world. To isolate the contribution of the different structural members, the test should consist of sensors located on different parts of the bridge. With additional sensors on the piers, girders, and beams, the vibrations can be better separated than they were in the case discussed here. Future research should be concentrate along this direction. The analysis here has shown that the vibrations can be separated into

different modes, and that the proper modes with structural and dynamic significance can be extracted. From this analysis, the weak structural members have been identified through either intra-wave frequency modulations or outright frequency downshifts. Therefore, the feasibility of the method has been established beyond any reasonable doubt. The present method is based on the most logical test load condition, the transient load, and the most effective data analysis method, the HHT. This method requires no special force-generating machines. The only requirement is that traffic be restricted for as long as needed for the test load, traveling at the normal speed, to pass across the bridge spans under observation. This requirement will cause only minimal traffic disruption, and the test could even be done at night or whenever normal traffic is at a minimum. The test load can be a fully loaded truck, or a roller, which is even better, for the load will be even more concentrated. The data-analysis method is the key to success, and the HHT approach used here is the most unique method and at the forefront of the research in data analysis. This approach utilizes not only the nonlinear characteristics of the response to determine the damage, but also the transient properties of the load to determine the damage location. Additionally, the free vibration frequency can be used to determine the extent of the damage. Considering the low number of the sensors required, and the efficient way of utilizing the data, the HHT presents a new, viable alternative for bridge-damage identification.

The similarity between the marginal and Fourier spectra further demonstrates the power of the Hilbert spectral analysis: it can produce whatever the Fourier analysis can and more, but Fourier-based analysis will never produce the time-frequency-energy analysis as the Hilbert spectral analysis does. Furthermore, any locally generated large amplitude vibration with a short period will be smeared by Fourier spectral analysis, as the case of pier 2 indicated. Such smearing of energy to a wide frequency range effectively obscures the frequency change of the structure. This factor is another indication that the Hilbert-Huang transform should be the method of choice.

Acknowledgements

The authors would like to express their deep appreciation to the graduate students and technicians from the Bridge Research Center, National Central University, for collecting the vibration data. WLC is supported in part by a program from the Ministry of Transportation, ROC. NEH would like to acknowledge the support from the Special Program Office at the Turner-

Fairbank Research Center of the Federal Highway Administration, USA.

References

- Alampalli, S., G. Fu, and E. W. Dillon, 1997: Signal versus noise in damage detection by experimental modal-analysis. *J. Struct. Eng. (Amer. Soc. Civ. Eng.)*, **123**, 237–245.
- Al-Khalidy, A., M. Moori, Z. Hou, S. Yamamoto, A. Masuda, and A. Sone, 1997: Health monitoring systems of linear structures using wavelet analysis. *Structural Health Monitoring: Current Status and Perspectives*, F.-K. Chang, Ed., Technomic, 164–175.
- Barai, S. V., and P. C. Pandey, 1995: Vibration signature analysis using artificial neural networks. *J. Comput. Civ. Eng.*, **9**, 259–265.
- Barai, S. V., and P. C. Pandey, 1997: Time-delay neural networks in damage detection of railway bridges. *Adv. Eng. Softw.*, **28**, 1–10.
- Bacry, E., A. Arnéodo, U. Frisch, Y. Gagne and E. Hopfinger, 1991: Wavelet analysis of fully developed turbulence data and measurement of scaling exponents. *Proc. Turbulence89: Organized Structures and Turbulence in Fluid Mechanics*, M. Lesieur and O. Métais, Eds., Kluwer, 203–215.
- Basdevat, M., A. Benveniste, B. Gach-Devauchelle, M. Goursat, D. Bonnecase, P. Dorey, M. Prevosto, and M. Olagnon, 1993: In-situ damage monitoring in vibration mechanics: Diagnostic and detective maintenance. *Mech. Syst. Signal Process.*, **7**, 401–423.
- Basu, B., and V. K. Gupta, 1997: Nonstationary seismic response of MDOF systems by wavelet transform. *Earth. Eng. Struct. Dynam.*, **26**, 1243–1258.
- Bendat, J. S., and A. G. Piersol, 2000: *Random Data Analysis and Measurement Procedures*. John Wiley & Sons, 602 pp.
- Brancaleoni, F., D. Spina, and C. Valente, 1993: Damage assessment from the dynamic response of deteriorating structures. *Safety Evaluation Based on Identification Approaches*, H. G. Natke, G. R. Tomlinson, and J. T. P. Yao, Eds., Vieweg, 276–291.
- Braun, S., and M. Feldman, 1997: Time-frequency characteristics of nonlinear-systems. *Mech. Syst. Signal Proc.*, **11**, 611–620.
- Carmona, R., W. L. Hwang, and B. Torresani, 1998: *Practical Time-Frequency Analysis: Gabor and Wavelet Transform with an Implementation in S*. Academic Press, 490 pp.
- Chang, F.-K., 1997: *Proceedings of the First International Workshop on Structural Health Monitoring*. Technomic, 801 pp.

- Chang, F.-K., 1999: *Proceedings of the Second International Workshop on Structural Health Monitoring*. Technomics, 1062 pp.
- Chang, F.-K., 2003: *Proceedings of the Third International Workshop on Structural Health Monitoring*. DEStech, 1552 pp.
- Chase, S. B., and G. Washer, 1997: Nondestructive evaluation for bridge management in the next century. *Public Roads*, **61**, 16–25.
- Chiu, C. K., 1992: *An Introduction to Wavelets*. Academic Press, 266 pp.
- Chopra, A. K., 1995: *Dynamics of Structure: Theory and Applications to Earthquake Engineering*. Prentice Hall, 729 pp.
- Clough, R. W., and J. Penzien, 1993: *Dynamics of Structures*. McGraw-Hill, 738 pp.
- Cohen, L., 1995: *Time Frequency Analysis*. Prentice Hall, 299 pp.
- Doebling, S. W., C. R. Farrar, M. B. Prime, and D. W. Shevitz, 1996: *Damage Identification and Health Monitoring of Structural and Mechanical Systems from Changes in their Vibration Characteristics: A Literature Review*. Report LA-13070-MS, Los Alamos National Laboratory, NM, 118 pp.
- Doebling, S. W., F. M. Hemez, L. D. Peterson, and C. Farhat, 1997: Improved damage location accuracy using strain energy-based mode selection criteria. *AIAA J.*, **35**, 693–699.
- Fahy, F. J., 1994: Statistical energy analysis—a critical overview. *Phil. Trans. R. Soc. London, Ser. A*, **346**, 431–447.
- Farra, C. R., W. E. Baker, T. M. Bell, K. M. Cone, T. W. Darling, T. A. Duffey, A. Eklund, and A. Migliori, 1994: *Dynamic Characterization and Damage Detection in the I-40 Bridge over the Rio Grande*. Los Alamos Report LA-12767-MS UC-906, 153 pp.
- Farrar, C. R., and S. W. Doebling, 1997: Lessons learned from applications of vibration-based damage identification methods to large bridge structure. *Structural Health Monitoring: Current Status and Perspectives*, F.-K. Chang, Ed., Technomic, 351–370.
- Farra, C. R., S. W. Doebling, and D. A. Nix, 1999: Vibration-based structural damage identification. *Phil. Trans. R. Soc. London, Ser. A*, **359**, 131–149.
- Felber, A., 1997: Practical aspects of testing large bridges for structural assessment. *Structural Health Monitoring: Current Status and Perspectives*, F.-K. Chang, Ed., Technomic, 577–588.
- Feldman, M., 1991: Method for determination of vibratory system modal parameters using Hilbert transform. Application for Patent No. 098985, Israel, 28 July 1991.

- Feldman, M., 1994a: Nonlinear system vibration analysis using Hilbert transform: I. Free-vibration analysis method, "FREEVIB." *Mech. Syst. Signal Process.*, **8**, 119–127.
- Feldman, M., 1994b: Nonlinear system vibration analysis using Hilbert transform: II. Forced vibration analysis method, "FORCEVIB." *Mech. Syst. Signal Process.*, **8**, 309–318.
- Feldman, M., 1997: Nonlinear free-vibration identification via the Hilbert transform. *J. Sound Vib.*, **208**, 475–489.
- Feldman, M., and S. Braun, 1995: Identification of non-linear system parameters via the instantaneous frequency: Application of the Hilbert transform and Wigner-Ville techniques. *Proc. 13th Int. Modal Analysis Conf.*, Nashville, TN, 637–642.
- Hahn, S., 1995: *Hilbert Transforms in Signal Processing*. Artech House, 299 pp.
- Hanagud, S., and H. Luo, 1997: Damage detection and health monitoring based on structural dynamics. *Structural Health Monitoring: Current Status and Perspectives*, F.-K. Chung, Ed., Technomic, 715–726.
- Hou, Z., M. Noori, and R. St. Amand, 2000: Wavelet-based approach for structural damage detection. *J. Eng. Mech. Div. (Amer. Soc. Civ. Eng.)*, **126**, 677–683.
- Huang, N. E., 1996: Computer Implemented Empirical Mode Decomposition Method, Apparatus, and Article of Manufacture. (US Provisional Application Serial Number 60/023,411, August 14, 1996 and Serial No. 60/023,822 filed on August 12, 1996, Patent allowed March 1999).
- Huang, N. E., Z. Shen, and S. R. Long, 1996: The mechanism for frequency downshift in nonlinear wave evolution. *Adv. Appl. Mech.*, **32**, 59–117.
- Huang, K. 1998a: A New Instrumental Method for Bridge Safety Inspection Based on a Transient Test Load. US Patent Application Serial No. 09-210693, filed December 14, 1998; Patent No. US 6,192,758 B1 issued February 27, 2001.
- Huang, N. E. 1998b: Computer Implemented Empirical Mode Decomposition Method, Apparatus, and Article of Manufacture Utilizing Curvature Extrema, U.S. App. No. 09/082,523, filed May 21, 1998.
- Huang, N. E., 1998c: Computer Implemented Empirical Mode Decomposition Method, Apparatus, and Article of Manufacture for Two-Dimensional Signals, U.S. App. No. 09/150,671, filed September 10, 1998.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998d: The empirical mode decompo-

- sition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., 1999: Implemented Empirical Mode Decomposition Method, Apparatus, and Article of Manufacture for Analyzing Biological Signals and Performing Curve Fitting, U. S. App. No. 09/282,424., filed March, 31, 1999.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of water waves—The Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Huang, N. E., M. C. Wu, S. R. Long, S. S. P. Shen, W. Qu, P. Gloersen, and K. L. Fan, 2003: A confidence limit for empirical mode decomposition and Hilbert spectral analysis. *Proc. R. Soc. London, Ser. A*, **459**, 2317–2345.
- Juneja, V., R. T. Haftka, and H. H. Cudney, 1997: Damage detection and damage detectability—analysis and experiments. *J. Aerosp. Eng.*, **10**, 135–142.
- Kim, H. M., and T. J. Bartkowicz, 1997: A 2-step structural damage detection approach with limited instrumentation. *J. Vib. Acous.*, **119**, 258–264.
- Kim, J. T., and N. Stubbs, 1995: Model-uncertainty impact and damage-detection accuracy in plate girder. *J. Struct. Eng. (Amer. Soc. Civ. Eng.)*, **121**, 1409–1417.
- Li, Y. F., and J. C. Yen, 1997: Vibration resistance evaluation of Chuang-Sha Bridge after erosion induced by typhoons. *Struct. Eng.*, **12**, 89–105.
- Li, Y. F., W. C. Tseng, K. Huang, and T. Y. Ho, 2003: A study of ambient vibration test of bridge column by using Hilbert-Huang transform. *J. Chinese Inst. Civ. Hydr. Eng.*, **15**, 21–33.
- Manning, R. A., 1994: Structural damage detection using active members and neural networks. *AIAA J.*, **32**, 1331–1333.
- Mori, Y., and B. R. Ellingwood, 1994a: Maintaining reliability of concrete structures. 1. Role of inspection repair. *J. Struct. Eng. (Amer. Soc. Civ. Eng.)*, **120**, 824–845.
- Mori, Y., and B. R. Ellingwood, 1994b: Maintaining reliability of concrete structures. 2. Optimum inspection repair. *J. Struct. Eng. (Amer. Soc. Civ. Eng.)*, **120**, 846–862.
- Natke, H. G., G. R. Tomlinson, and J. T. P. Yao, 1993: *Safety Evaluation Based on Identification Approaches*. Vieweg, 324 pp.
- Nishimura, A., 1990: Examination of bridge substructure for integrity. *Jap. Railway Eng.*, **114**, 13–17.
- Pandit, S. M., 1991: *Modal and Spectrum Analysis: Data Dependent System*

- in State Space. Wiley-Interscience, 415 pp.
- Pandey, P. C., and S. V. Barai, 1995: Multilayer perception in damage detection of bridge structures. *Comput. Struct.*, **54**, 597–608.
- Prime, M. B., and D. W. Shevitz, 1996: Linear and nonlinear methods for detecting cracks in beams. *Proc. 14th Int. Modal Analysis Conf.*, Dearborn, MI, 1437–1443.
- Salawu, O. S., 1997: Detection of structural damage through changes in frequency—A review. *Eng. Struct.*, **19**, 718–723.
- Salawu, O. S., and C. Williams, 1995: Bridge assessment using forced-vibration testing. *J. Struct. Eng. (Amer. Soc. Civ. Eng.)*, **121**, 161–173.
- Salawu, O. S., and C. Williams, 1995: Review of full-scale dynamic testing of bridge structures. *Eng. Struct.*, **17**, 113–121.
- Staszewski, W. J., S. G. Pierce, W. R. Phip, and G. R. Tomlinson, 1997: Wavelet signal processing for enhanced Lamb wave defect detection in composite plates using optical fiber. *Opt. Eng.*, **36**, 1877–1888.
- Stubbs, N., and J. T. Kim, 1996: Damage localization in structures without base-line modal parameters. *AIAA J.*, **34**, 1644–1649.
- Surace, C., and R. Ruotolo, 1994: Crack detection of a beam using the wavelet transform. *Proc. 12th Int. Modal Analysis Conf.*, Honolulu, HI, 1141–1147.
- Titchmarsh, E. C., 1948. *Introduction to the Theory of Fourier Integrals*. Oxford University Press, 394 pp.
- Tsou, P. Y., and M. H. H. Shen, 1994: Structural damage detection and identification using neural networks. *AIAA J.*, **32**, 176–183.
- Vakakis, A. F., 1997: Nonlinear normal-modes (NNMS) and their applications in vibration theory—An overview. *Mech. Syst. Signal Proc.*, **11**, 3–22.
- Wu, X., J. Ghaboussi, and J. H. Garrett, 1992: Use of neural networks in detection of structural damage. *Comput. Struct.*, **42**, 649–659.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.
- Yang, J. N., S. Lin, and S. Pan, 2002: Damage detection of a health monitoring building using Hilbert-Huang spectral analysis. *Advances in Building Technology*, **2**, 1017–1024.
- Yang, J. N., Y. Lei, and N. Huang, 2003a: Damage identification of civil engineering structure using Hilbert-Huang transform. *Structural Health Monitoring: The Demands and Challenges*, F.-K. Chang, Ed., DEStech,

544–553.

- Yang, J. N., Y. Lei, S. Pan, and N. Huang, 2003b: Identification of linear structures based on Hilbert-Huang transform. Part I: Normal modes. *Earthq. Eng. Struct. Dynam.*, **32**, 1443–1467.
- Yang, J. N., Y. Lei, S. Pan, and N. Huang, 2003c: Identification of linear structures based on Hilbert-Huang transform. Part II: Complex modes. *Earthq. Eng. Struct. Dynam.*, **32**, 1533–1554.
- Yang, J. N., Y. Lei, S. Lin, and N. Huang, 2004: Hilbert-Huang based approach for structural damage detection. *J. Eng. Mech. Div. (Amer. Soc. Civ. Eng.)*, **130**, 85–95.

Norden E. Huang

*Code 614.2, NASA/Goddard Space Flight Center, Greenbelt, MD 20771,
USA*

norden.e.huang@nasa.gov

Kang Huang

*DART, ACT 21, 1201 Main Street, Suite 800, Dallas, TX 75202, USA
KangHuang@sbcglobal.net*

Wei-Ling Chiang

*Bridge Research Center, Department of Civil Engineering, National Central University, Chuang-Li, Taiwan, Republic of China
chiang@cc.ncu.edu.tw*

CHAPTER 13

APPLICATIONS OF HHT IN IMAGE ANALYSIS

Steven R. Long

The recently developed empirical mode decomposition/Hilbert-Huang transform (EMD/HHT) for the analysis of nonlinear and non-stationary data has been extended to include the analysis of image data. Because image data can be expressed in terms of an array of rows and columns, this robust concept is applied to these arrays row by row. Each slice of the data image, either row or column-wise, represents local variations of the image being analyzed. Just as much of the data from natural phenomena are either nonlinear or non-stationary, or both, so are the data that form images of natural processes. Thus, the EMD/HHT approach (or HHT in abbreviated form) is especially well-suited for image data, giving frequencies, inverse distances or wavenumbers as a function of time or distance, along with the amplitudes or energy values associated with these, as well as a sharp identification of embedded structures. The varied products of this new analysis approach are joint and marginal distributions to be viewed as isosurfaces, contour plots, and surfaces that contain information on frequency, inverse wavelength, amplitude, energy and location in time, space, or both. Additionally, component images representing the intrinsic scales and structures embedded in the data will be described, as well as a technique for obtaining frequency variations.

13.1. Introduction

Images have long held a fascination for us, for our eyes constantly input a stream of data into our minds from the world around us, images we process to obtain distance, size, color, orientation, along with beauty or even a sense of danger. We have always analyzed and characterized drawings, inscriptions, and paintings and assigned them a level of value and appreciation accordingly. Our minds are capable of image processing of the highest order.

In recent times, it has become technically possible to obtain images that are more than just a picture, images that are actually an array of

numbers of high precision that represent point-wise measurements over an area, not just the gray scale value in a scanned photograph, but a detailed measure of electromagnetic wavelength and intensity representing color, heat, or x-ray intensity, to name just a few, all with underlying physical significance. Images are also routinely acquired from sources other than electromagnetic waves, such as magnetic resonance images. These arrays of numbers are handled easily by computer and can thus be displayed, printed, and viewed as an image, while representing a reality that our own eyes could never see directly. In this sense, modern imaging technology and techniques have expanded our vision, allowing us to “see” new things never observed before. Such is also the case when applying a totally different method to image analysis, a method such as the EMD/HHT technique. New types of results can be expected, allowing our minds to see the reality around us in entirely new ways. In fact, this process has already begun in the many and various applications of EMD/HHT reported to date on “one-dimensional” data such as time records of earthquakes, ocean waves, rogue water waves, sound analysis, and length-of-day measurements. The approach has been successful with nonlinear and non-steady data because its basis functions are time varying and adaptive.

If this data in general can be written as $X(d)$, then the first step in this new approach is to “sift” or decompose the data into n empirical components such that

$$X(d) = \sum_{j=1}^n c_j + r_n, \quad (13.1)$$

where c_j is the j th component, r_n is the residue, and d denotes the axis over which the data varies, such as time or spatial distance. The sifting stops when either the last component c_n or the residue r_n reaches a value lower than a predetermined level, which is a small value of no consequence, or when r_n becomes a monotonic function from which no more components can be extracted. Even if the data have a zero mean, the final residue can be different from zero. If the data contain a trend (such as a slow drift in the instrument calibration or the tide changing during ocean-wave measurements), then when the sifting is completed, the residue r_n will be that trend.

Just as in earlier work on the EMD/HHT analysis, this sifting will be employed here on data that make up each row or column of an image array. Thus, the initial processing used here is identical with that used in earlier reports. The application to images, however, requires this new sifting

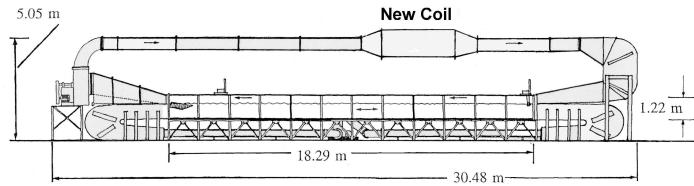


Figure 13.1: The wind, wave, and current interaction research tank at NASIRF.

processing and subsequent analysis to be repeated perhaps hundreds, even thousands of times, depending on image resolution and size. For example, a square image array of 512 pixels on a side such as will be presented here would require 512 repetitions of the sifting process, and each repetition would produce a complete component set and residue as outlined in (13.1). Once the component sets are obtained, they become the input to further processing steps that can produce unique products from the initial image. After a brief overview, these various products will be discussed, in turn, and examples used to illustrate the new possibilities opened up by the application of this robust technique to images.

13.2. Overview

Image processing in general has always made full use of available computer capabilities. Because of the shear size of the data to be analyzed, the needs of image processing have often “raised the bar” on hardware requirements, such as the need for more memory, increased computational speed, increased internal bus rates, and greater and better storage. Just as the images to be processed come from widely varied sources such as satellite imagery, aerial photography, microscope imagery, industrial imaging for quality control, and CT scan data for medical applications, so do the methods of image processing vary. It is almost always driven by a need to reveal or detect some feature within the image, to clarify the feature or resolution, or to make possible the measurement of a feature of the image, to name reasons why image processing is used. Therefore, image processing encompasses a wide range of mathematical techniques with a proven track record of effectiveness and established mathematical foundations.

The approach taken here is to refer the reader to well-established texts such as Castleman (1996) or Russ (2002) as a starting point in a vast literature already published on the subject, and to outline the new and important

methods that can now be added to the available tools for producing new and unique image products. The descriptive and mathematical groundwork for this new approach has already been established in the series of articles denoted here as “foundation articles,” by Long et al. (1995), Huang et al. (1998, 1999, 2000, 2001, 2003a, 2003b, 2004), and Wu and Huang (2004). Others have also made valuable contributions to this concept of image data decomposition, as can be seen in the work of Nunes et al. (2003).

13.3. The analysis of digital slope images

13.3.1. *The NASA laboratory*

The laboratory used for acquiring the images discussed in this section is the NASA Air-Sea Interaction Research Facility (NASIRF) located at the NASA Goddard Space Flight Center/Wallops Flight Facility, at Wallops Island, VA. The test section is 18.3 m long and 0.9 m wide, filled to a depth of 0.76 m of water, leaving 0.45 m for air flow if needed. The facility can produce wind- and/or paddle-generated waves over a water current in either direction, and its capabilities, instruments, and software have been described in detail by Long (1992, 1993, 2004) and Long and Klinke (2002). The basic facility is shown in Fig. 13.1, with an additional new feature, indicated as “New Coils,” recently installed to control the temperature and humidity of a cooled air flow over heated water.

13.3.2. *The digital camera and set-up*

The device used to acquire the laboratory images presented here was a Silicon Mountain Design M-60 camera, capable of acquiring images of up to 1024×1024 pixels at a resolution of 4096 intensity levels and at a rate of up to 60 images s^{-1} . For the examples shown here, the resolution was set at 512×512 pixels, at a rate of 60 images s^{-1} . The camera was mounted to look vertically down at the water surface, so that its 512×512 pixel image area covered a physical square on the water surface of 26.54 cm \times 26.54 cm. Each pixel thus covered a square of about 0.518 mm each throughout the image area. To obtain an image array of surface slope values from each 512×512 image, the configuration illustrated in Fig. 13.2 was used. The light source was mounted at the bottom of the tank and initially produced a uniform field of light. A thin film that varied in transparency between clear and black along the downwind direction was placed on top of the light source, resulting in a light intensity output that varied linearly with fetch.

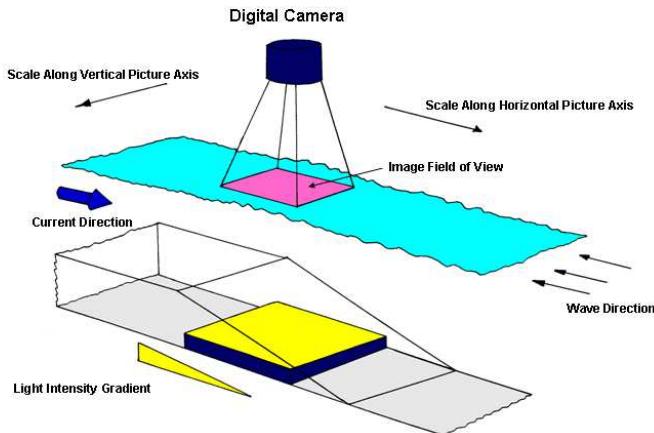


Figure 13.2: The digital slope imaging system at NASIRF, illustrating the false bottom used to create shear zones in the water current flow. A subsurface light source emits light with an intensity gradient along the direction of wave motion. The current flows against the waves over the false bottom in a depth of 38.1 cm, which increases down the slope of the full tank depth of 76.2 cm. Thus, the waves encounter increasing current strength until they are blocked. The pixel indices of the image area run as shown on the horizontal and vertical scales.

Because of this variation and the laws of refraction, each image acquired by the camera had the downwind surface slope at each pixel location on the surface recorded in the intensity level of that pixel in the image array. By using calibration lookup tables obtained through simple geometry and the application of Schnell's law of refraction, this intensity level was then converted to the down channel surface slope for each pixel, an array of 512×512 numerical values.

13.3.3. Acquiring experimental images

With this imaging system in place, steps were taken to acquire interesting images of wave blocking due to an opposing water current. This system is illustrated in Fig. 13.2, which shows the presence of a false bottom. A transparent window was placed over the light source and allowed the light

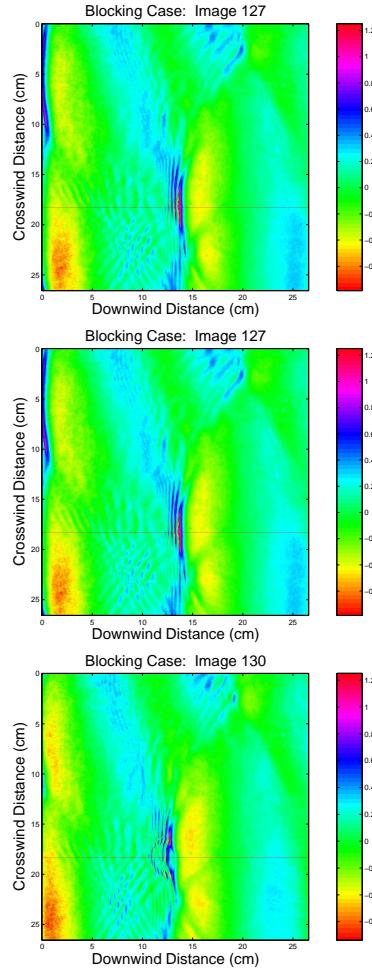


Figure 13.3: Wave blocking case. (a) Top. Current flows from left to right, and paddle waves move from right to left. The direction of the scale for downwind distance is an artifact of the matrix storage and indexing. The colorbar at right gives the slope intensity values. (b) Middle. Development continues at two time intervals after (a), or $\frac{1}{30}$ s later. The dark line along the direction of wave motion (right to left) shows the slice of data to be further processed. The current flows from left to right, and the colorbar at right shows the slope intensity. The downwind direction scale (horizontal) above is tied to the index of stored matrix elements and runs opposite to the wave direction. (c) Bottom. $\frac{1}{60}$ s later than (b). Note the structure extending down the face of the wave.

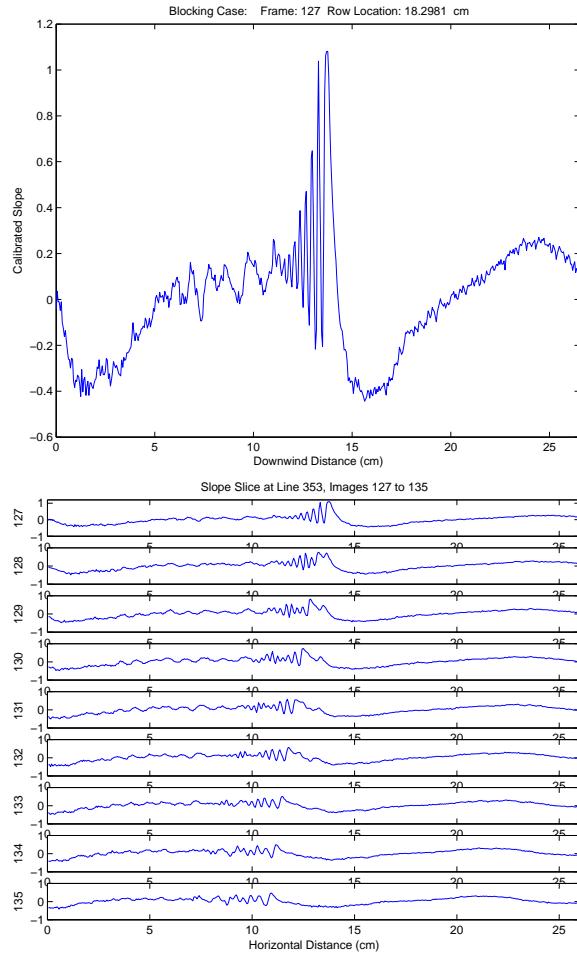


Figure 13.4: (a) Top. Horizontal slice at line 353 in Fig. 13.3a. (b) Bottom. Slices of slope at horizontal line 353 in images 127 to 135.

source to function, while maintaining the slope of the false bottom for the water flow. When water flowed against the wave direction, an area of current shear was set up over the light box. As discussed in Long et al. (1993), waves can become trapped in such a shear zone and shift to higher frequencies in the absence of wind. To illustrate what happens, Figs. 13.3a-c present a series of images of this phenomena. Although the camera acquired images at 60 s^{-1} , only a selection from the sequence is shown here. To help our eyes see the images, the 4096 levels of gray have been converted to a color

variation. A horizontal line down channel is also included to mark an area of interest, where a very rapid development occurs during the time covered by these images. Even though $\frac{1}{60}$ s variations may seem to be rapid for a water surface, one gets the impression from these images that something is developing faster than the camera acquisition rate. The images obtained do, however, capture significant stages in this rapid development.

Using the horizontal line through the phenomena in image 127, Fig. 13.4a illustrates the detail contained in the actual array of data values. Fig. 13.4b shows a comparison of consecutive image slices at $\frac{1}{60}$ s steps, through the entire 9-image sequence of images, from 127 to 135. These are horizontal slices through the region of an interesting phenomena and show the complex changes that can occur in surface slope over a time period covering only $\frac{8}{60}$ s.

13.3.4. Using EMD/HHT analysis on images

From Fig. 13.4b, the slice from image 129 (image Fig. 13.3b) was processed following the EMD sifting procedures, and the results are shown in Fig. 13.5, which shows that the complex surface may be represented by comparatively few components. The sifting was done via the extrema approach discussed in the foundation articles and produced a total of eight components, the first seven of which are shown in Fig. 13.5.

13.3.5. The digital camera and set-up

Using the components illustrated in Figure 13.5, the next step with Hilbert-Huang transform (HHT) analysis produces a result that can be visualized as in Fig. 13.6a showing the contour plot of the result.

In Fig. 13.6a, a standing wave pattern can be seen between about 4 cm and 10 cm on the horizontal scale of the image slice. This pattern persists through many images preceding and following this moment in the sequence, and within the group of standing waves, different peaks take a turn at being the largest. Because they persist in time, however, they can be thought of as “standing waves” trapped in the blocking conditions of the current shear flow when the images were acquired. The shorter capillary wave group that appears on the leading edge of the crest and suddenly bursts out radially is represented in the higher values on the cm^{-1} scale. To produce a true wavenumber, one only has to convert by using

$$k = 2\pi/\lambda, \quad (13.2)$$

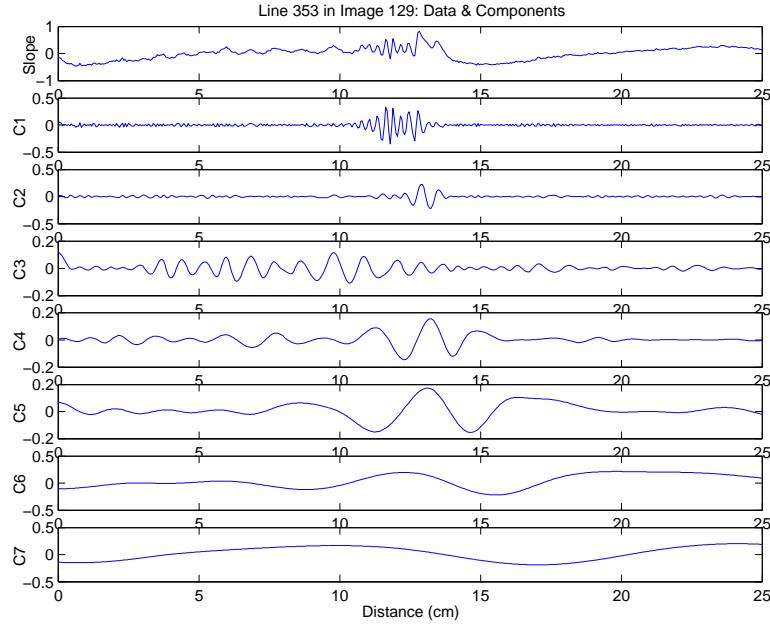


Figure 13.5: The horizontal slice of data at line 353 of image 129 in Fig. 13.3b, followed by the first seven components of EMD extrema sifting. Note the removal of intrinsic scales starting with the shortest (C_1) and increasingly longer scales through C_7 .

where k is the wave number (in cm^{-1}), and λ is the wavelength (in cm). The largest slopes occur where certain wavelengths are dominating in the data and its values are entered by the HHT process in the resulting array by location (horizontal location on the slice in cm), by wavelength inversed on the vertical (cm^{-1}) scale, and by amplitude or intensity of slope on the color scales.

To visualize how rapidly the capillary wave burst affects the result, a comparison is presented in Figs. 13.6b. It represents a combination of the result from the first and last images of the image sequence illustrated partially in Figs. 13.3a-c and 13.4a-b. By adding the result of image 127 to that of image 135, the change caused by the capillary wave burst, seen as a developing semi-circular feature on the leading edge of the wave, over a period of $\frac{8}{60}$ (0.1333) s can be seen. The standing wave pattern appears stable in the 5-to-11 cm horizontal area of the combined image in the vertical scale around 1 (cm^{-1}), while the incoming wave carrying the capillary wave's semi-circular burst is noted as moving left at a scale of 0.5 (cm^{-1})

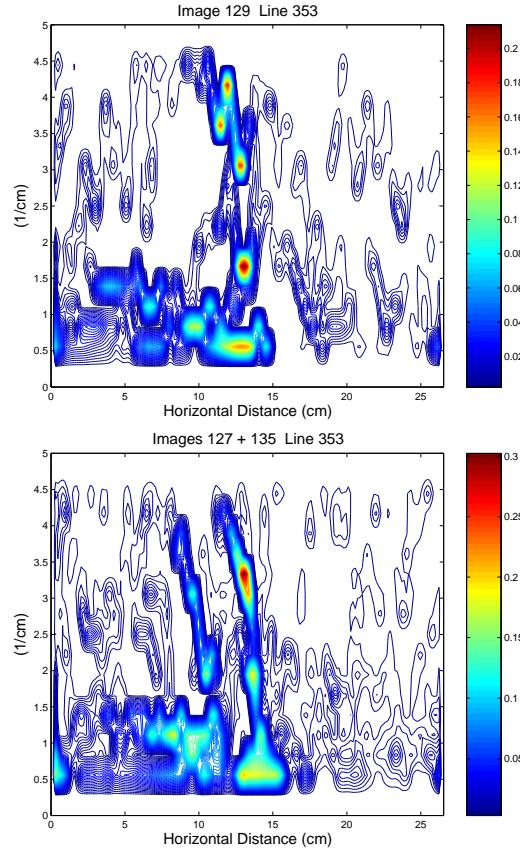


Figure 13.6: (a) Top. Contour plot of the results of EMD/HHT analysis on the slope slice taken horizontally from image 129, Fig. 13.3b. Note the standing wave pattern between 4 and 10 cm approximately, and the incoming wave at about 13 cm that is producing the burst of shorter waves appearing at 1.6, 3, 3.5, and 4.2 on the cm^{-1} inverse wavelength scale. Waves move from right to left into increasing current strength. The horizontal distance scale above is tied to the matrix elements, so that the waves progress through decreasing distance values. (b) Bottom. The analysis of the result of the sum of images 127 and 135. The standing wave pattern can be seen to be stable in the 5 to 11 cm horizontal area of the combined image, while the incoming wave carrying the capillary burst is noted as moving left at a scale of 0.5 (cm^{-1}) from 15 to about 13 cm. Above this, the rapidly moving burst is seen in the vertical scales between 2 and 4.5 (cm^{-1}).

from 15 to about 13 cm. Component wavelengths of the rapidly moving burst is seen in the vertical scales between 2 and 4.5 (cm^{-1}). By using the difference in horizontal distance covered by the burst in 0.1333 s, it can be shown that the burst is moving approximately 60 cm s^{-1} over the surface of an opposing, rapid flow. When the opposing current speed is also taken into account, this speed appears to be beyond the usual speeds of capillary waves of this wavelength.

13.3.5.1. Volume computations and isosurface visualization

Many interesting phenomena happen in the flow of time; thus, it is of interest to explain how changes occur in the images as time passes.

By starting with a single horizontal line from the image, a contour plot as was shown in Figs. 13.6a can be computed from the EMD/HHT analysis. By using a longer set of images, such as from 110 to 150 (41 images covering $\frac{40}{60}$ or $\frac{2}{3}$ s), and a slice through the capillary wave's burst location, a set of 41 numerical arrays can be obtained from the EMD/HHT analysis. Each array can be visualized by means of a contour plot as shown before. The entire set of 41 arrays can now be combined in sequence to form an array volume, or an array of dimension 3. Within the volume, each element of the array contains the amplitude or intensity of the surface slope from the image sequence. One axis (call it x) of the volume represents the horizontal distance down the slice as before, in cm. Another axis (call it y) represents the resulting inverse length scale (cm^{-1}) that signifies the inverse wavelength of the slope values associated with waves in the data. The additional axis (call it z), produced by laminating the 41 arrays together, represents time, because each image was acquired in steps of $\frac{1}{60}$ s. Thus the position of the element in the volume gives location x (cm) along the horizontal slice, inverse wavelength (cm^{-1}) along the y axis, and time (s) along the z axis.

To visualize the data stored in a three-dimensional array, isosurface techniques are needed. This process could be compared to peeling an onion, except that the different layers, or spatial contour values, are not bound in spherical shells. After a value of slope amplitude or intensity is specified, the isosurface visualization will make transparent all array elements outside of the level of the value chosen, while shading in the chosen value so that the elements inside that level (or behind it) can not be seen.

Some examples of isosurface representations are seen beginning in Fig. 13.7a, where the horizontal line at 353 through the middle of the capillary wave's semi-circular burst splitting off the front face of the incoming carrier

wave and then rejoining the standing waves around the 8 to 10 cm area. These occurrences are the most energetic ones and are seen in the absence of lower levels. To illustrate the effect of the peeling level, a slow variation in peeling value is presented as Figs. 13.7a–d. At the peeling level of 0.03 shown in Fig. 13.7b, the incoming wave packet is seen to start around the 14 cm area and to progress up and left, denoting movement down the horizontal pixel line with time. As the burst to shorter wavelengths (larger cm^{-1} values) occurs, those slope intensity levels are taken by the capillary wave burst, which extends rapidly back along the cm^{-1} axis, and then return to the path taken by the carrier wave packet around the 0.5 s vertical level. With a peeling level of 0.025 as shown in Fig. 13.7c, the standing wave area around 8 cm and 1.9 on the cm^{-1} scale now appears as a column along time, initially shifted toward a shorter wavelength (compressed) until it is enveloped by the incoming carrier wave and capillary wave burst extending out along the y axis (cm^{-1}) as the shorter wavelengths suddenly appear. Fig. 13.7d illustrates a peeling level of 0.02. Here the standing wave column finally emerges near the end of the sequence (0.66 s), as a combination of standing wave and incoming wave energy. Further details can also be seen of the effect the incoming wave packet has on the standing wave before being merged with it and the capillary wave burst it is carrying. Also, other standing waves around 4 cm and 6 cm make an appearance at various times at this intensity level. With further decreases in the peeling level, a point is reached where the outer surface grows so large and complex that it starts to hide the underlying features. Nevertheless, the sequence shown in Figs. 13.7a-d demonstrates what is now possible by applying EMD/HHT to images.

The example shown here was developed by using a single line from each image in the time sequence of images. Other approaches may be used as well, such as averaging over a subset of adjacent lines within each image, if doing so is justifiable based on the content of the image and the range of lines chosen over which to average. The process described here could also be repeated for each of the 512 lines across the wave tank of the present images, from top to bottom in each image, through the 41 images in the time sequence shown, or for even more lines by including more images. Doing so would produce an array with a dimension beyond 3, which would be no problem for the computers, but difficult for us to visualize. Each slice along the across tank axis from the higher dimensional array would be a volume result similar to the ones shown here, where isosurface techniques could reveal interesting processes and features.

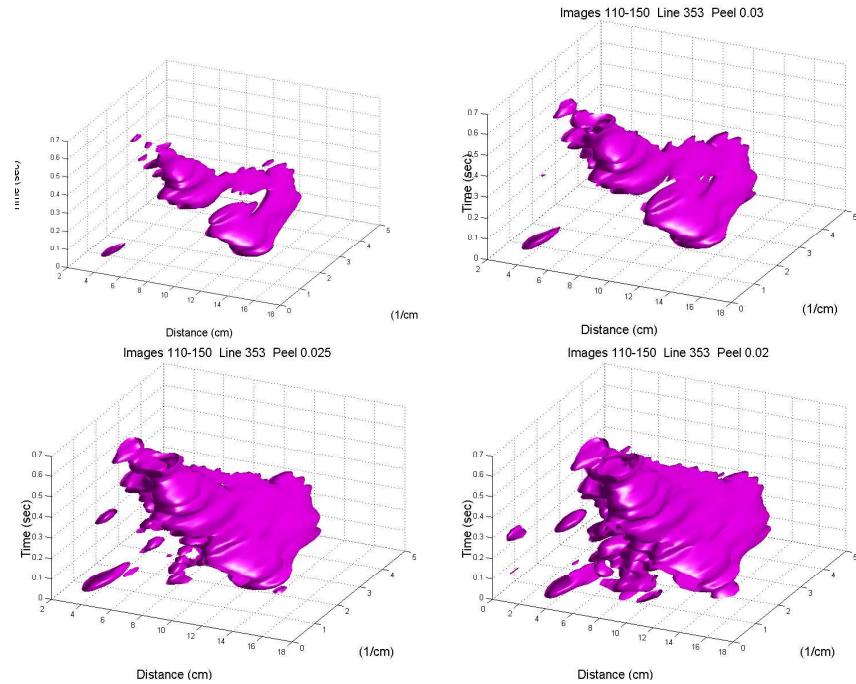


Figure 13.7: (a) Upper left. The result along the 353 through the center of the capillary wave burst from images 110 to 150. A fairly restrictive peel level (0.035) reveals details of the burst splitting off the front face of the incoming carrier wave and then rejoining the standing waves around the 8 to 10 cm area. (b) Upper right. Re-computing (a) for a peeling value of 0.03. The incoming carrier wave is seen around 14 cm. With the passage of time (up the z axis), the capillary wave burst extends the width to higher values along the y (cm^{-1}) axis. (c) Lower left. Peeling level now at 0.025, for (a) re-computed. Note the appearance of the stalk at coordinates of about 8 cm distance, 1.9 on the cm^{-1} scale. These coordinates correspond to a standing wave that was compressed by the approach of the incoming wave, bending to larger cm^{-1} values with the passage of time, and then being engulfed by the more energetic incoming wave and its capillary burst. (d) Lower right. Peeling level of 0.02, major details still evident. The combination of the incoming and standing wave column finally emerges near the end of the sequence (0.66 s), now moved to a slightly longer wavelength (smaller cm^{-1} value).

Another approach for the analysis of images is to re-assemble the image components in a different way. Again starting with a single horizontal line, this time at line 256 through the center of the image, each center line from the sequence of images is laminated to its predecessor to build up an array

that is 512 along one edge, in units of cm, and the number of images along the other axis, in units of time (s). Once complete, this two-dimensional array can be split into 512 slices along the time axis. Each of these time slices, representing the variation in slope with time at a single pixel location, can then be processed by using EMD/HHT techniques. For example, consider Figs. 13.8a-b. Fig. 13.8a represents the change in slope values at pixel 250 on line 353, or a location of 12.96 cm on the distance axis over a time period of just over seven seconds. Fig. 13.8b presents similar data, but at a location on line 353 at pixel location 300, or at a location of 15.55 cm. Both figures show the passage of wave groups of longer wavelength through the chosen pixel as a function of time. By using this data, the EMD/HHT techniques reveal variations in the frequency of the waves passing through this location. Consider Fig. 13.9a, which shows the change of frequency with time by using line 353 and pixel 250. Note the capillary wave bursts to higher frequency occurring very rapidly around the 1 s area, and also the persistent 2 Hz wave with fluctuating frequency. The sequence of images examined in earlier figures, images 110 through 150, occurs in time between 1.83 and 2.5 s, and the shorter 9 image sequence between 127 and 135 occurs on this time scale between 2.12 and 2.25 s.

By using the slope versus time data of Figs. 13.8a and 13.8b, EMD/HHT can be used to reveal the frequency variations, as shown in Figs. 13.9a and 13.9b, corresponding to the data of Figs. 13.8a and 13.8b, respectively.

Changes in curvature may also be examined by selecting two pixels along the same line. The spatial distance along this line between the pixels selected can be adjusted to vary the resolution and sensitivity to wavelength. The curvature along the direction of wave motion is then $\Delta\text{slope} / \Delta\text{separation}$, where the change of slope is the difference between the two measured values at the two selected pixels, and the change in separation is the length difference along the chosen horizontal line between the two selected pixels. By repeating this procedure for each image in a sequence, the time variation of curvature at the selected location can be obtained and studied. This capability is another avenue opened up by the application of the EMD/HHT methods to images.

13.3.5.2. Use of EMD/HHT in image decomposition

Another application of EMD/HHT is a process similar to the filtering of images. A more accurate description of this process would refer to it as a separation of the complete image into component images, from the shortest

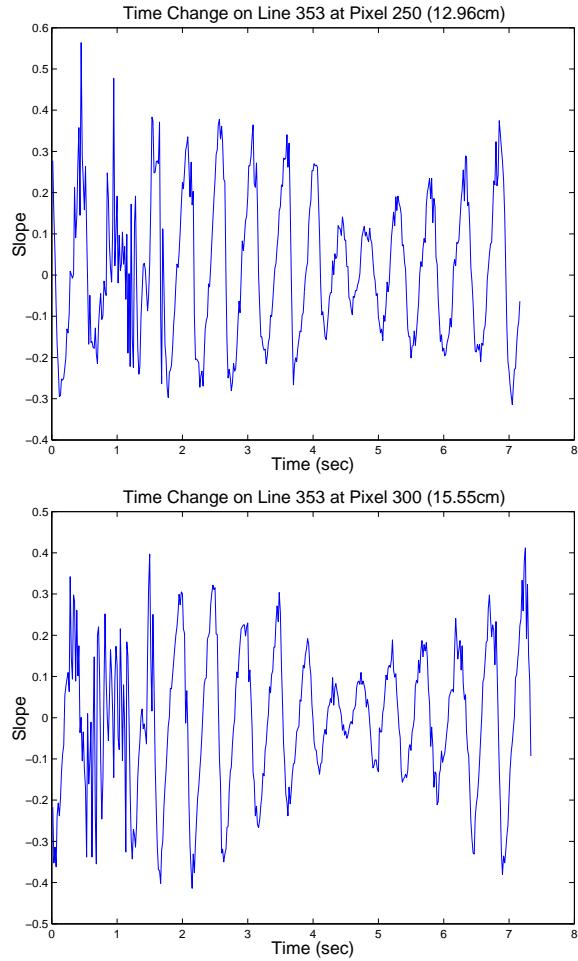


Figure 13.8: (a) Top. Change in slope over a time period of just over seven seconds at a location on line 353 through the burst at pixel 250 for a location at 12.96 cm. (b) Bottom. Same as (a), except through the burst at pixel 300, or 15.55 cm.

scales to the longer scales. Again, as in the earlier analysis, the summation of these components produces the original data, in this case an image, with minimal computational differences from the original.

Starting with Fig. 13.3c, image 130 is again processed line by line, so that each horizontal slice produces a set of components. The difference in this approach is that now the first component is taken from each com-

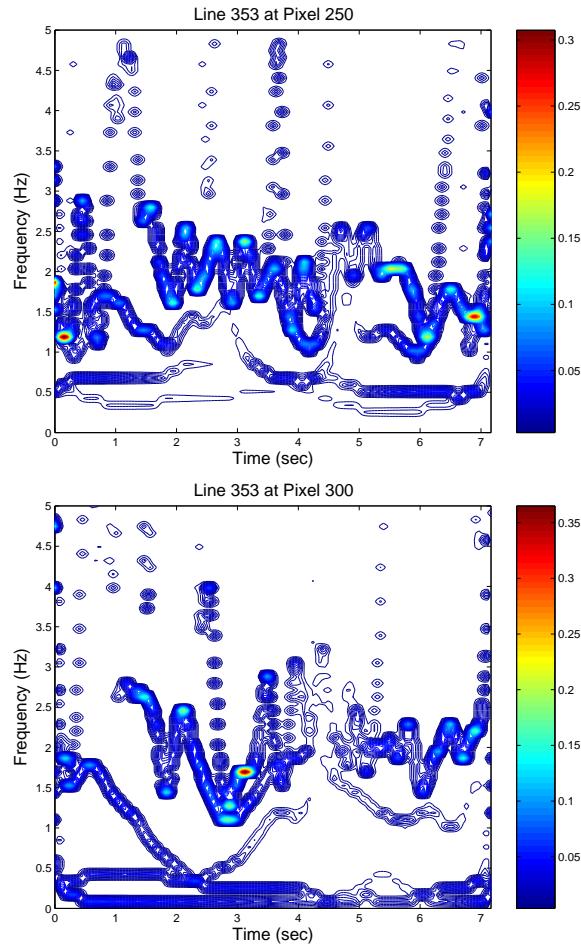


Figure 13.9: (a) Top. Change of frequency with time by using line 353 and pixel 250 from Fig. 13.8a. Note the bursts to higher frequency occurring very rapidly around the one-second area and also the persistent 2 Hz wave with fluctuating frequency. (b) Bottom. Same as (a), except now using pixel 300 from Fig. 13.8b.

ponent set, or 512 different first components from the 512 complete sets. These 512 first components are laminated back together in the correct order, producing a 512×512 array, which can then be viewed as an image, the first component image. Fig. 13.10a shows the first component image of Fig. 13.3c. Note that only the shortest scales are seen, and in this case, they are the capillary wave burst seen in the earlier images. The decomposition

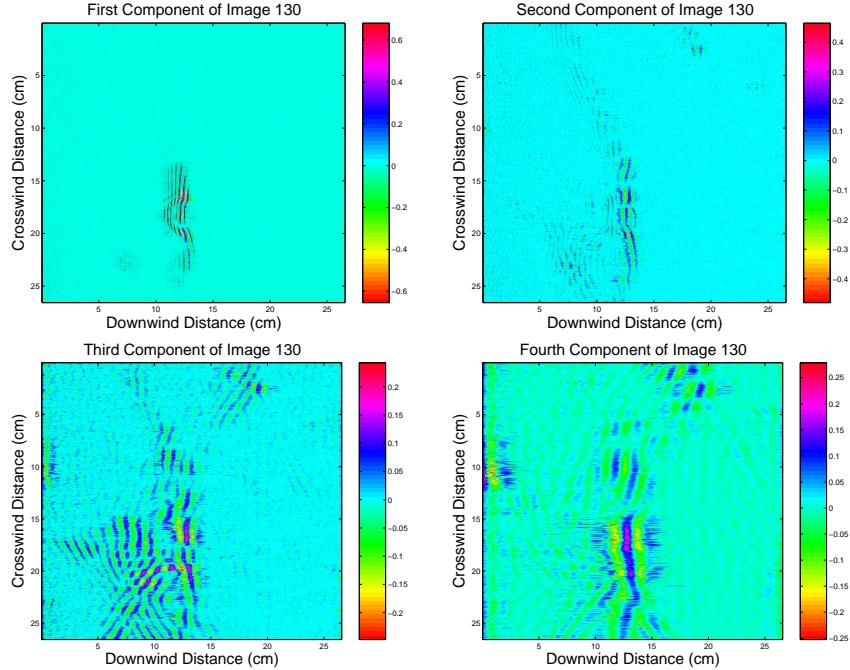


Figure 13.10: (a) Upper left. First component image from image 130 shown earlier, and assembled from the first components obtained from each of the 512 horizontal lines and slope data making up the image. Note that the shortest scale is present, and may be studied in the absence of all other scales. (b) Upper right. The next longer scale, the second component image from image 130, assembled as before, but now using the second component only. (c) Lower left. The third component image developed as before. Note the standing wave patterns now evident, and that the incoming carrier wave has now crossed into the area of the standing waves. (d) Lower right. The fourth component image, revealing the carrier wave scale and some of the persistent standing wave pattern. Note the apparent attachment to a bifurcation, seen in the top right quadrant. Lamination effects are becoming evident here (some horizontal mismatch), illustrating the need for some matching computations between the slices at this and higher component levels.

(EMD) continues with Fig. 13.10b, illustrating that the slightly longer scale now resides with the incoming wave packet that carried and produced the capillary burst. Fig. 13.10c, the third component image, reveals the scales associated with the standing wave field, and it can be seen to project into the incoming wave, or put another way, the incoming wave has crossed into the domain of the standing waves. This result raises the possibility of the

standing wave pattern being the trigger for the burst that arose out of a momentary but extreme steepness (slope > 1) on the incoming wave packet's front face near the crest. Fig. 13.10d shows the fourth component image containing the longer scale associated with the incoming wave, and some smaller contributions from the standing waves, along with an attachment to a bifurcation, seen in the top right quadrant. At this level, the lamination effects are becoming evident (some horizontal mismatch), illustrating the need for some matching computations between the slices at this and higher component levels. Up to now, the raw result of processing has just been assembled back together to form the image array.

13.4. Summary

As has been illustrated here, the application of the EMD/HHT techniques to image processing opens up new and exciting frontiers in image analysis. It is hoped that this brief review of some of these new possibilities will raise still others in the minds of readers, as well as point to new and interesting applications.

Images of water waves were used here because they happen to be the author's research field of interest. However, this present study in no way limits the wide application of the steps and results discussed here to interesting images of other processes. The new views into the complex interactions occurring routinely at the interface between the atmosphere and earth's oceans were made possible entirely by the power and versatility of the EMD/HHT breakthrough technology. If data from irregular heart beats, brain wave patterns during epileptic seizures, images from CT, MRI, or x-ray images of patients with a medical problem were analyzed by researchers in other fields, it is certainly possible that new and useful results and techniques would result. That work has indeed already begun, and not only in the medical fields, but in science and engineering applications as well. Such is the case with useful tools. They can simplify existing tasks and help to accomplish new ones that we thought were not even possible.

Acknowledgments

The author wishes to express his continuing gratitude and thanks to his colleague Dr. Norden E. Huang, Senior Fellow at NASA Goddard Space Flight Center, Director of the Goddard Institute for Data Analysis, and inventor of the EMD/HHT techniques for his help and discussions. The author also wishes to acknowledge and sincerely thank Dr. Dean G. Duffy

of NASA/GSFC for his discussions on improvements to the manuscript and its presentation. Support from NASA Headquarters is also gratefully acknowledged, specifically Dr. Eric Lindstrom and Dr. William Emery, for their encouragement and support of the work.

References

- Castleman, K., 1996: *Digital Image Processing*. Prentice Hall, 667 pp.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-steady time series analysis. *Proc. R. Soc. London, Ser. A*, **454**, 903–995.
- Huang, N. E., Z. Shen, and S. R. Long, 1999: A new view of water waves – The Hilbert spectrum. *Annu. Rev. Fluid Mech.*, **31**, 417–457.
- Huang, N. E., H. H. Shih, Z. Shen, S. R. Long, and K. L. Fan, 2000: The ages of large amplitude coastal seiches on the Caribbean coast of Puerto Rico. *J. Phys. Oceanogr.*, **30**, 2001–2012.
- Huang, N. E., C. C. Chern, K. Huang, L. W. Salvino, S. R. Long, and K. L. Fan, 2001: A new spectral representation of earthquake data: Hilbert spectral analysis of Station TCU129, Chi-Chi, Taiwan, 21 September 1999. *Bull. Seism. Soc. Am.*, **91**, 1310–1338.
- Huang, N. E., M. C. Wu, S. R. Long, S. S. P. Shen, W. Qu, P. Gloersen, and K. L. Fan, 2003a: A confidence limit for empirical mode decomposition and Hilbert spectral analysis. *Proc. R. Soc. London, Ser. A*, **459**, 2317–2345.
- Huang, N. E., M.-L. C. Wu, W. Qu, S. R. Long, S. S. P. Shen, and J. E. Zhang, 2003b: Applications of Hilbert-Huang transform to non-stationary financial time series analysis. *Appl. Stoch. Model. Bus.*, **19**, 245–268.
- Huang, N. E., Z. Wu, S. R. Long, K. C. Arnold, K. Blank, and T. W. Liu 2004: On instantaneous frequency, *Proc. R. Soc. London, Ser. A*, under revision.
- Long, S. R., R. J. Lai, N. E. Huang, and G. R. Spedding, 1993: Blocking and trapping of waves in an inhomogeneous flow. *Dynam. Atmos. Oceans*, **20**, 79–106.
- Long, S. R., N. E. Huang, C. C. Tung, M.-L. C. Wu, R.-Q. Lin, E. Mollo-Christensen, and Y. Yuan, 1995: The Hilbert techniques: An alternate approach for non-steady time series analysis. *IEEE GRSS Newsletter*, **3**, 6–11.

- Long, S. R., and J. Klinke, 2002: A closer look at short waves generated by wave interactions with adverse currents. *Gas Transfer at Water Surfaces, Geophysical Monograph 127*, AGU, 121–128.
- Nunes, J. C., Y. Bouaoune, E. Delchelle, N. Oumar, and Ph. Bunel, 2003: Image analysis by bidimensional empirical mode decomposition. *Image Vision Comput.*, **21**, 1019–1026.
- Russ, J. C., 2002: *The Image Processing Handbook*. CRC Press, 732 pp.
- Wu, Z., and N. E. Huang, 2004: A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. London, Ser. A*, **460**, 1597–1611.

Steven R. Long

*NASA/GSFC/WFF, Ocean Sciences Branch, Code 614.2, Wallops Island,
VA 23337, USA*
Steven.R.Long@nasa.gov

INDEX

- acceleration data, 321, 322, 324, 325
- adaptive basis, 2
- adaptive data analysis, 22
- advanced very high resolution
 - radiometer, 193
 - drift, 194, 196, 197, 203
- air-sea interaction, 245
- aliasing, 89, 108, 113, 115, 120, 121
- amplitude, 335
- amplitude envelope, 107, 111, 117
- amplitude modulation, 91, 104, 266
- amplitude step signal, 112
- amplitude-modulated signal, 89, 106, 121
- analysis of nonlinear and
 - non-stationary data, 335
- analytic signal, 36, 93, 97, 151
- annual cycle, 174, 176, 178–180, 184, 218–221, 226, 231, 235
- Appalachian Mountains, 160
- array volume, 345
- atmosphere, 173, 177, 186, 187, 189
- auto-regressive, 78, 265, 292–297, 300
- automotive gearbox, 54
- autoregressive error, 182
- average marginal spectrum, 156
- AVHRR
 - kee advanced very high resolution radiometer
- B-spline, 33
- B-spline algorithm of EMD, 37
- B-spline EMD, 35
- band limited, 151
- bandwidth, 114, 118
- baroclinic instability, 157, 165
- barographic data, 164
- basis functions, 199
 - a posteriori*, 199
 - empirical, 199
- basis pursuit, 34
- Bedrosian theorem, 17, 33, 35, 46, 93, 101, 111
- Bessel functions, 109
- binary hypothesis testing, 126
- bispectrum, 253
- blocking conditions, 342
- bridge deck, 319
- bridge safety inspection, 307
- BS-EMD
 - see B-spline EMD
- calibration
 - on-board, 194
- capillary wave, 343
- cardinal B-splines, 41
- carrier frequency, 104
- carrier wave, 251, 256
- central limit theorem, 133
- chaos, 266, 283, 286, 291, 295, 300
- characteristics of noise, 126
- Chesapeake Bay Bridge and Tunnel, 156
- chi-square distribution, 133
- climate, 173, 174, 177, 179, 187, 189
- climate data, 218, 220
- cloud screening, 196
- coastal storms, 160
- cold front, 165, 166
- component images, 335, 348

- confidence levels, 138
- confidence limit, 19
- contour plot, 345
- convolution, 41, 102
- cubic B-spline approximation, 38
- cubic spline, 7, 25, 26, 37, 75, 77, 151, 177
- curvature, 348
- cyclones, 160
- daily maximum temperature, 227, 233, 237
- damage features, 265, 273, 299
- damage identification, 327
- data
 - nonlinear, 199
 - nonstationary, 199
- data-driven, 264, 265, 282, 284, 292, 300
- degrees of freedom, 133
- denoising, 67, 81, 82, 84, 86
- Des Moines, IA, 161
- detrending, 67, 81, 83, 84, 86
- differentiating filter, 103
- digital slope images, 338
- diurnal tide, 157
- down-sampling, 122
- Duffing equation, 3, 151
- El Chichón, 164
- El Niño, 12, 20, 221, 229, 241
- El Niño/Southern Oscillation, 157, 174–176, 184–186, 189
- Election Day Flood, 158
- EMD
 - filtering, 203
 - see empirical mode decomposition
- empirical components, 336
- empirical correction, 137
- empirical mode decomposition, 2, 33, 34, 67–69, 73–75, 78, 80, 81, 83–85, 97, 127, 173–175, 177, 179–182, 184, 189, 193, 219, 221–223, 227, 229, 231, 240, 247, 265, 266, 268, 276, 279, 282, 286, 299, 307, 315, 335
- empirical model, 264, 265, 295
- empirical orthogonal function, 150
- end effect, 24
- end point criteria, 177
- end-point analysis, 95
- energy density, 128
- ENSO
 - see El Niño/Southern Oscillation
- envelope, 6, 18, 35
- Euler polynomials, 40
- Euler spline, 33, 40
- extrema, 5, 6, 9, 11, 24–26, 151, 199
- extrema-counting mean period, 129
- failure diagnosis, 56
- fatigue test, 54
- fault diagnosis, 35
- feature values, 290, 294, 300
- fGn
 - see fractional Gaussian noise
- filter bank, 33, 67, 72, 73, 75, 78, 85
- filter transfer functions, 100
- filtering, 89, 91, 103, 109
- filters
 - time-space, 202
- finite impulse response (FIR) filter, 91
- flat spectrum, 71
- forced vibration, 312
- Fourier analysis, 245
- Fourier spectra, 126, 217, 222, 238, 239
- Fourier transform, 33, 90
- fractional Brown motion, 68
- fractional Gaussian noise, 20, 67–70, 72, 73, 75, 78, 79, 81–84
- frame structure, 282
- free vibrations, 312
- frequency domain, 114, 118
- frequency downshift, 305, 318, 325, 327
- frequency modulation, 91, 104, 246
- frequency response, 100
- frequency step discontinuity, 117
- frequency-modulated component, 108
- frequency-modulated signal, 89, 109, 121

- Gabor transform, 90
 Gaussian white noise, 127–129
 GDCN
 see Global Daily Climatology Network
 georeferenced satellite data, 195
 Global Daily Climatology Network,
 225
 global detection method, 264
 global spectrum, 71
 globally averaged surface temperature
 anomaly, 140
 greenhouse gases, 174, 176, 184, 189
 group structure, 245, 246, 251, 254,
 255, 258, 259
 harmonic generation, 246
 heart-rate variability, 84
 Hermite interpolation, 98
 HHT
 see Hilbert-Huang transform
 hidden peaks, 99
 high pressure, 160
 high-pass filter, 100
 Hilbert amplitude spectrum, 13
 Hilbert spectral analysis, 2, 13, 222,
 307, 315
 Hilbert spectrum, 13, 35, 154, 245
 Hilbert transform, 4, 33, 93, 150, 178,
 222, 267, 269, 270, 287, 311
 Hilbert transform filter, 102
 Hilbert-Huang transform, 2, 34, 89,
 219, 222, 268, 270, 277, 279, 282,
 286, 307, 335
 HT
 see Hilbert transform
 Hurricane Gloria, 152
 Hurricane Opal, 165
 Hurst exponent, 68, 70, 75–79, 81, 84,
 86
 image analysis, 335
 image data, 335
 image decomposition, 348
 IMF
 see intrinsic mode function
 impulse response, 75, 77
 incremental HHT algorithm, 90, 97,
 98
 index of energy conservation, 53
 information extraction, 126
 instantaneous amplitude, 4, 154
 instantaneous angular velocity, 267
 instantaneous frequency, 4, 34, 92,
 107–109, 112, 117, 120, 154, 199,
 246
 instantaneous phase, 265–268, 271,
 272, 276, 281, 282, 289, 299
 interannual variation, 196
 intermittent components, 122
 intra-wave frequency modulation, 313
 intrinsic mode decomposition, 75, 76
 intrinsic mode function, 6, 35, 68–74,
 76–84, 86, 97, 127, 129, 151, 177,
 180, 196, 220, 222, 229, 265, 268,
 307, 315
 intrinsic scales, 335
 isosurface techniques, 345
 iteration convergence, 98
 iteration scheme, 97
 joint time-frequency distribution, 90
 landcover, 194
 length-of-day, 11, 20
 linear damage, 265
 loading
 heavy, 313, 316, 319, 321, 323
 light, 312, 316, 321
 loading condition, 308, 311–314, 316,
 317, 319, 321
 local detection method, 264
 local mean, 6, 24, 37
 local wavenumber, 246
 Lorenz equation, 151, 268, 270, 271,
 283, 286, 298
 low-pass filter, 100
 Makassar Straits, 157
 marginal spectrum, 15, 156
 matching pursuit, 34
 mean energy density, 136

- mean period, 128, 129
- measures for orthogonality, 52
- mechanical systems, 54
- meteorological dataset, 149
- modal analysis, 309
- modal methods, 264, 295
- monocomponent, 92
- monocomponentness, 92
- Monte-Carlo, 179–181
- Morlet wavelet, 161, 164, 167
- NAC
 - see nonlinear-and-non-stationary annual cycle
- NASA Air-Sea Interaction Research Facility, 338
- National Climatic Data Center, 225
- National Oceanic and Atmospheric Administration, 165, 193
- NDVI
 - see normalized difference vegetation index
- neural network, 310
- New England, 160
- NOAA
 - see National Oceanic and Atmospheric Administration
- noise, 125
- non-destructive test, 309
- non-uniform sampling, 122
- nonlinear damage, 265
- nonlinear deformation, 313
- nonlinear phase, 111
- nonlinear system identification, 22
- nonlinear-and-non-stationary annual cycle, 219, 220, 225, 227, 229, 233, 236, 237, 240
- nonlinearity, 245
- nonstationary, 126, 259, 307
- normal distributions, 133
- normalized difference vegetation index, 193, 194, 196
 - maximum composite, 194
- North Atlantic, 160
- North Atlantic Oscillation, 140
- null hypothesis, 126
- Nuttal theorem, 18
- Nyquist frequency, 101
- ocean engineering, 245
- ocean waves, 245, 246, 254, 255, 259
- one-dimensional structures, 273
- orbital drift, 194
- Pacific Decadal Oscillation, 188
- PDO
 - see Pacific Decadal Oscillation
- peeling level, 346
- period doubling, 180
- phase, 36
- phase dereverberation, 273
- phase error, 290, 291
- phase function, 4, 16, 18, 27, 267, 268, 270, 272, 274, 289
- phase lag, 274, 276, 280, 300
- phase modulation, 266
- photosynthesis, 194
- piling, 325
- power spectra, 70
- power spectral density, 69, 70, 76, 285, 286
- power-law spectrum, 71
- prediction error, 265, 292–295, 297–300
- probability density functions (PDFs), 294
- probability density functions of IMFs, 132
- projection
 - Albers equal-area, 196
- proper frequency, 308
- PSD
 - see power spectral density
- QBO
 - see Quasi-Biennial Oscillation
- quadratic B-spline approximation, 38
- quadrature model, 93
- Quasi-Biennial Oscillation, 174–176, 184–187, 189

- radiance, 194
 - near infrared (NIR), 194
 - red (VIS), 194
- random excitations, 266, 283, 285, 286, 291, 293
- real-time HHT algorithm, 89, 90, 97
- recursive formula, 37
- red noise, 179–182
- reflectance, 194
 - surface bidirectional properties, 194
 - near infrared, 194
 - red, 194
 - soil background, 194
- regression techniques, 155
- residue, 336
- resolution
 - spatial, 194
 - temporal, 196
- rolling element bearing, 56
- sampling function, 117
- SAT
 - see surface air temperature
- satellite transitions, 197
- satellites, new generation, 194
 - SeaWiFS, SPOT, MODIS, 213
- scaled building model, 276
- sea-level height, 156
- semidiurnal tide, 157
- sensor degradation, 196
- separating amplitude and phase, 103
- Shapiro filter, 152
- shear zone, 341
- short-time Fourier transform, 34, 149
- sifting, 37, 89, 91, 96, 108, 151, 199, 247, 336
- sifting process, 5, 7, 9, 10, 26
- signal processing, 67
- signal separation process, 96
- simple oscillating mode, 6
- solar cycle, 174, 182–184, 186, 187, 189
- solar radiation, 161
- solar zenith angle, 193
- Southern Oscillation Index, 140
- space of splines, 37
- spatial coherence, 195
- spectral analysis, 67, 69, 70, 75, 80
- spectral bandwidth, 194, 254
- spectrum-weighted mean period, 129, 131, 136
- spline, 199
- spline interpolation, 98
- spread of the energy densities, 135, 137
- standing waves, 342
- statistical significance, 12, 20, 26, 127
- step discontinuity, 112
- STFT
 - see short-time Fourier transform
- stiffness, 317, 318, 325
- Stokes's equation, 109
- stopping criterion, 9, 39, 152, 178, 223
- stratospheric volcanic aerosols, 163, 196
- strong northerly winds, 160
- structural health monitoring, 264, 266, 282, 293, 295, 305
- structures, 335
- subharmonic frequency, 257
- sun-target-view geometry, 194
- sunspot cycle, 176, 184
- superharmonic frequency, 257
- surface air temperature, 217, 218, 225
- surface slope, 342
- system solution
 - closure, 195
 - observational approach, 195
 - parametric approach, 195
- SZA
 - see solar zenith angle
- TAC
 - see thirty-year-mean annual cycle
- Teager energy operator, 23
- Teager's energy operator, 94
- temporal-frequency analysis, 149
- TFR
 - see time-frequency representation
- thirty-year-mean annual cycle, 218, 221, 227, 229, 236, 237, 240
- tight-frame analysis, 105

- time domain filtering, 315
time variation of curvature, 348
time-frequency analysis, 33, 89, 199
time-frequency representation, 34, 149
time-warp analysis, 99, 111
transient amplitude changes, 121
transient detection, 53
transient load, 313
trend, 173–176, 178, 179, 184, 189, 200, 201
trispectrum, 253
trough, 166

uncertainty, 13, 16
uncertainty principle, 310
uncertainty problem, 90, 92
under-sampling, 120
unit step transient, 105
up-sampling, 122

vegetation dynamics, 194
vegetation indices, 194
Virginia, 158

warm front, 166
warped filter, 101
warped sampling, 101
warped signal, 100
wave blocking, 339
wave energy, 346
wave extension, 178
wave group, 245, 251, 259
wavelength, 335
wavelet, 34, 68, 71, 74, 78, 246
wavelet analysis, 150, 310
wavelet packet, 34
wavelet transform, 90, 222, 229
wavenumber, 335
West Virginia, 158
Wigner-Ville distribution, 34, 310
wind-generated waves, 245, 246, 251, 252, 259
WT
 see wavelet transform

zero-crossing, 5, 6, 9, 24, 71, 151, 199