Structural Safety 52 (2015) 150–160

Contents lists available at ScienceDirect

Structural Safety

journal homepage: www.elsevier.com/locate/strusafe

An artificial neural network approach for stochastic process power spectrum estimation subject to missing data

Liam Comerford [a,*], Ioannis A. Kougioumtzoglou [b], Michael Beer [a]

[a] Institute for Risk and Uncertainty, University of Liverpool, Liverpool L69 3GH, UK

[b] Department of Civil Engineering & Engineering Mechanics, The Fu Foundation School of Engineering & Applied Science, Columbia University, 610 S.W. Mudd Building, 500 West 120th Street, New York, NY 10027, USA

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form 10 September 2014
Accepted 2 October 2014
Available online 18 November 2014

Keywords:
Neural network
Stochastic process
Missing data
Evolutionary power spectrum

ABSTRACT

An artificial neural network (ANN) based approach is developed for estimating the power spectrum of stochastic processes subject to missing/limited data. In this regard, an appropriately defined ANN is utilized to capture the stochastic pattern in the available data in an average sense. Next, the extrapolation capabilities of the ANN are exploited for generating realizations of the underlying stochastic process. Finally, power spectrum estimates are derived based on established frequency (e.g. Fourier analysis), or versatile joint time–frequency analysis techniques (e.g. wavelets) for the cases of stationary and non-stationary stochastic processes, respectively. One of the significant advantages of the approach relates to the fact that no a priori knowledge about the data is assumed, while the approach is applicable for treating non-stationary processes not only with separable but non-separable in time and frequency evolutionary power spectra as well. Comparisons of several target power spectra with Monte Carlo simulation based power spectrum estimates demonstrate the versatility and reliability of the approach for up to 50% missing data.

1. Introduction

Probabilistic engineering simulations often require models for the engineering system excitation/response processes. In this regard, evolutionary power spectrum (EPS) estimation can be a central part of this modelling process [1,2]. Further, for stochastic process model based Monte Carlo simulations to be reliable, modelling/estimation techniques often require a significant amount of data and/or some prior knowledge of the underlying physics of the process; i.e. the more data on which a model is built, the more statistically accurate the simulation is likely to be. Nevertheless, in several engineering applications large amounts of data can be difficult to acquire for several reasons, such as cost (e.g. expensive sensor maintenance in harsh conditions/ remote areas), frequency and unpredictability of the effect (e.g. earthquakes), and sensor failures (e.g. damage from the effect itself, using cheap sensors, etc). Further, available data can often be highly limited and irregularly sampled. When working with limited data, standard Fourier techniques for spectral estimation, e.g. [3], can demonstrate poor performance, and without any prior knowledge of the underlying statistics of the process, alternative (less general) analysis techniques can be problematic in certain cases. For instance, autoregressive methods may require significant amounts of data upon which to build models, e.g. [4,5]. Furthermore, most spectral estimation approaches require uniformly sampled data, whereas the majority of available methods for dealing with non-uniform sampling perform satisfactorily mostly in cases of idealized scenarios where missing data are infrequent [6–10], or when a very limited number of significant harmonic components is assumed [11,12]. Additional challenges arise when dealing with non-stationary data. In this regard, to estimate the power spectrum of a non-stationary process, the Gabor transform [13], wavelets [14–18], chirplets [19] and the Wigner–Ville distribution [20,21] present a means of analyzing the non-stationary spectral content of a signal. Nevertheless, many of the approaches for addressing missing data in the stationary case cannot be applied, at least in a straightforward manner, for non-stationary cases, or assume that the process is locally stationary [22].

In general, if a measured realization of a stochastic process is available for analysis with no further information than that contained within the sample, it is impossible to predict with certainty what lies beyond the known time interval. Similarly, if data points are missing within the same sample, it is impossible to predict them with certainty. Further, if the focus is not on predicting

* Corresponding author.
E-mail addresses: l.comerford@liv.ac.uk (L. Comerford), iak2115@columbia.edu (I.A. Kougioumtzoglou), m.beer@liv.ac.uk (M. Beer).

http://dx.doi.org/10.1016/j.strusafe.2014.10.001
0167-4730/© 2014 The Authors. Published by Elsevier Ltd.

exactly what happens within intervals of missing data, but instead, on determining a complete time history to be used in a Monte Carlo simulation analysis for instance, then predicting the missing data in an average sense can be satisfactory.

For instance, in a Monte Carlo analysis it is often required that millions, or even billions of samples are considered to estimate rare events probabilities reliably [23]. Obviously, it is not always possible to collect that many real time histories; thus, they must be simulated in an appropriate manner. In this regard, an appropriate spectral analysis can be conducted on the few available measured time histories, and then new time histories with the same statistical characteristics can be further generated for a larger scale analysis. If the available measured time histories contain gaps, the problem is not that the time histories no longer contain important underlying process information (and, thus, they should be discarded), but that traditional spectral analysis techniques are not equipped to identify this information.

In this regard, an artificial neural network (ANN) approach, e.g. [24] is developed herein to facilitate estimating the power spectrum (stationary or non-stationary) of stochastic processes subject to missing data. First, an appropriately defined ANN is employed to capture the stochastic pattern in the available data in an average sense. Next, the ANN, having stored process trends within its connection weights, is exploited for generating new data to fill sampling gaps fitting with the underlying stochastic process. Note that the ANN is utilized only to reconstruct data in the time domain and not to produce spectra directly; once the record is reconstructed, a range of spectral estimation methods become immediately applicable (Fourier, Wigner–Ville, maximum entropy etc). In this paper, Power spectrum estimates are derived by utilizing standard Fourier analysis (stationary case), or recently developed wavelet based EPS estimation approaches (non-stationary case). Several numerical examples are included to demonstrate the reliability of the approach.

2. Technical methodology

2.1. Stochastic process representation and spectral estimation

In this section, a concise review on stationary and non-stationary stochastic process representation is included for completeness. Further, a recently developed wavelet based evolutionary power spectrum estimation approach is delineated as well. The latter is instrumental in assessing the reliability of the ANN approach for generating complete realizations, and eventually, for estimating the underlying stochastic process EPS.

For any real-valued stationary process, $X(t)$, there exists a corresponding complex orthogonal process $Z(\omega)$ such that $X(t)$ can be written in the form of Eq. (1), e.g. [3,25,26].

$$X(t) = \int_0^\infty e^{i\omega} dZ(\omega) \tag{1}$$

where

$Z(\omega)$ has the properties

$$E\left(\left|dZ^2(\omega)\right|\right) = 4S_X(\omega)d\omega \tag{2}$$

and

$$E(dZ(\omega)) = 0. \tag{3}$$

In Eq. (2), $S_X(\omega)$ is the two-sided power spectrum of the process $X(t)$. Further, a versatile formula for generating realizations compatible with the stationary stochastic process model of Eq. (1) is given by [27]

$$X(t) = \sum_{j=0}^{N-1} \sqrt{4S_X(\omega_j)\Delta\omega \sin(\omega_j t + \Phi_j)} \tag{4}$$

where Φ_j are uniformly distributed random phase angles in the range $0 \leqslant \Phi_j < 2\pi$. Furthermore, regarding estimation of the power spectrum of the process of Eq. (1) based on available realizations, this is given by the ensemble average of the square of the absolute Fourier transform amplitudes of the realizations. In this context, standard established Fast Fourier Transform algorithms can be utilized, e.g. [28].

Next, for the case of non-stationary stochastic processes, a similar to Eq. (1), rigorous process representation of non-stationary stochastic processes needs to be employed. In this regard, Nason et al. [29] developed a framework for representing non-stationary stochastic processes by utilizing a time/frequency-localized wavelet basis as opposed to the Fourier decomposition of Eq. (1). The representation reads

$$X(t) = \sum_j \sum_k w_{j,k}\psi_{j,k}(t)\xi_{j,k}, \tag{5}$$

where $\xi_{j,k}$ is a stochastic orthonormal increment sequence; $\psi_{j,k}(t)$ is the chosen family of wavelets and j and k represent the different scales and translation levels, respectively. Further, the local contribution to the variance of the process of Eq. (5) is given by the term $\left|w_{j,k}\right|^2$. The wavelet-based model of Eq. (5) relies on the theory of locally stationary processes (see also [30]). The aforementioned wavelet based representation can be viewed as a natural extension in the wavelet domain of earlier work related to the representation of non-stationary stochastic processes, e.g. [1,30].

Focusing next on generalized harmonic wavelets [17], these have a box-shaped frequency spectrum, whereas a wavelet of (m,n) scale and (k) position in time attains a representation in the frequency domain of the form

$$\Psi^G_{(m,n),k}(\omega) = \begin{cases} \frac{1}{(n-m)\Delta\omega} e^{\left(-i\omega\frac{kT_0}{n-m}\right)}, & m\Delta\omega \leqslant \omega \leqslant n\Delta\omega \\ 0, & otherwise \end{cases}, \tag{6}$$

where m, n and k are considered to be positive integers and $\Delta\omega = \frac{2\pi}{T_0}$, where T_0 is the total time duration of the signal under consideration. A collection of harmonic wavelets of the form of Eq. (6) spans adjacent non-overlapping intervals at different scales along the frequency axis. The inverse Fourier transform of Eq. (6) gives the time-domain representation of the wavelet which is equal to

$$\Psi^G_{(m-n),k}(t) = \frac{e^{\left(in\Delta\omega\left(t-\frac{kT_0}{n-m}\right)\right)} - e^{\left(im\Delta\omega\left(t-\frac{kT_0}{n-m}\right)\right)}}{i(n-m)\Delta\omega\left(t-\frac{kT_0}{n-m}\right)} \tag{7}$$

Furthermore, the continuous generalized harmonic wavelet transform (GHWT) is defined as

$$W^G_{(m,n),k} = \frac{n-m}{kT_0} \int_{-\infty}^{\infty} f(t)\overline{\psi^G_{(m,n),k}(t)}dt, \tag{8}$$

and projects the function $f(t)$ on this wavelet basis. Next, utilizing the generalized harmonic wavelets, Eq. (5) becomes (see [31])

$$X(t) = \sum_{(m,n)} \sum_k \left(X_{(m,n),k}(t)\right), \tag{9}$$

where

$$X_{(m,n),k}(t) = \sqrt{S_{(m,n),k}(n-m)\Delta\omega}\psi_{(m,n),k}(t)\xi_{(m,n),k} \tag{10}$$

Eq. (10) represents a localized process at scale (m,n) and translation (k) defined in the intervals $[m\Delta\omega, n\Delta\omega]$ and $\left[\frac{kT_0}{n-m}, \frac{(k+1)T_0}{n-m}\right]$, whereas $S_{(m,n),k}$ represents the EPS $S_X(\omega,t)$ at scale (m,n) and translation (k). In [31] it has been shown that under certain assumptions, Eq. (10) can be written in the form

$$X_{(m,n),k}(t) = \int_{m\Delta\omega}^{n\Delta\omega} e^{i\omega\left(t-\frac{kT_0}{n-m}\right)} dZ_{(m,n),k}(\omega), \tag{11}$$