Maria Sklodowska-Curie Actions (MSCA)
Innovative Training Networks (ITN)
H2020-MSCA-ITN-2018
Grant number 813137

Project number 813137

URBASIS-EU
New challenges for Urban Engineering Seismology

DELIVERABLE

Work Package: WP4
Number: D4.5 – Use of Sparse information from Earthquakes in urban area

Authors:	**Chen, Yu**	(UoL)
Co-authors:	Edwards, Ben	(UoL)
	Patelli, Edoardo	(UoL)

Reviewer	Beer, Michael	(UoL)
Approval	Management Board	
Status	Final Version	
Dissemination level	Public	
Delivery deadline	31.04.2021	
Submission date	07.06.2021	
Intranet path	https://urbasis-eu.osug.fr/Scientific-Reports-157	

URBASIS-EU

Use of sparse information from earthquakes in urban area

Yu Chen, Institute for Risk and Uncertainty, University of Liverpool

April 2021

Contents

1 Introduction

While we are marching into a Big Data world with an exponential increase in seismic recordings (Kong *et al.*, 2019), nevertheless, a large amount of data does not mean a large amount of information. Data might be irrelevant, corrupted, noisy or incomplete. It seems inevitable and even more necessary to deal with bad data for a risk-informed design of structures and infrastructure in earthquake prone regions (Beer *et al.*, 2013; Ching & Phoon, 2019). Inaccuracies in the data might be propagated through the model and affect further analyses (for example spectral analyses) and finally compromise the credibility of the analysis. Reliable methodologies and relevant procedures/tools that can identify and even handle those data problems are of great interests to both earthquake engineers and seismologists. Specifically, attentions are paid to reflect the degree of indeterminacy from the sparse information, and to account for the uncertainties with the model and the imperfect data. For example, evolutionary power spectrum models serve as potent characterization for ground motion . Their estimation requires both an in-depth understanding of the underlying physics of the problem and a relatively significant amount of data. However, both components are only available with a remaining degree of indeterminacy.

To understand the challenges related to modeling and characterization of ground motions given scarce and limited data, this deliverable reviews recent development of methodologies that improve spectral estimation from ground motion recordings (Section 2), simulate ground motion based on stochastic processes models and Generative modeling techniques (Section 3), and also present spectral analyses that calculate uncertainties over spectral estimation with incomplete data (Section 4).

2 Modeling ground motion as random processes

Owing to the influence of a series of uncontrollable factors like the mechanism of the seismic source, propagation paths and geotechnical media distribution at the engineering site, the ground motion is typically considered as stochastic processes (Liu, 1968, 1970). Actual earthquake records show that the time history of the ground motion accelerations usually includes three stages of vibrations: the initial, the strong and the attenuating stages. Therefore, the ground motion is a typical nonstationary process (Liu, 1970), whose statistical characteristics are changing with time. But stationary process model is also used to establish the ground motion models, it is usually believed that this only reflects its strong motion stage (Li & Chen, 2009).

2.1 Spectral representation of random processes

For a deterministic continuous signal $x(t)$, the Fourier transform is used to describe its spectral content. Similarly, in order to describe the random processes in the frequency domain, power spectral density (PSD) is adopted to describe how power is distributed over frequency. Summation or integration of the spectral contents thus yields the total power, identical to what would be obtained by integrating $x^2(t)$ over the time domain, as suggested by Parseval's theorem.

2.1.1 PSD of a stationary process

Given a random process $\{X(t) : t \in T\}$, a truncated version could be defined as (Miller & Childers, 2012):

$$X_{t_0}(t) = \begin{cases} X(t), & |t| \leq t_0 \\ 0, & |t| > t_0 \end{cases} \tag{1}$$

Based on Parseval's theorem, the time-averaged power can be given by :

$$P_{X_{t0}} = \frac{1}{2t_0} \int_{-\infty}^{\infty} X_{t0}(t)^2 dt = \frac{1}{2t_0} \int_{-\infty}^{\infty} |X_{to}(f)|^2 \mathrm{d}f \tag{2}$$

Since $P_{X_{t0}}$ is random variable, to get the ensemble averaged power, an expectation is taken:

$$\mathbb{E}[P_{X_{t0}}] = \frac{1}{2t_0} \int_{-\infty}^{\infty} \mathbb{E}[|X_{t_0}(f)|^2] \mathrm{d}f \tag{3}$$

The power in the (untruncated) random process $X(t)$ is then found by passing to the limit as $t_0 \to \infty$,

$$\overline{P_X} = \lim_{t_0 \to \infty} \frac{1}{2t_0} \int_{-\infty}^{\infty} \mathbb{E}[|X_{t_0}(f)|^2] \mathrm{d}f = \int_{-\infty}^{\infty} \lim_{t_0 \to \infty} \frac{\mathbb{E}[|X_{t_0}(f)|^2]}{2t_0} \mathrm{d}f \tag{4}$$

Define the PSD as the integrand of the power (i.e. $\overline{P_X}$) in the last equation:

$$S_{XX}(f) = \lim_{t_0 \to \infty} \frac{\mathbb{E}[|X_{t_0}(f)|^2]}{2t_0} \tag{5}$$

Therefore $S_{XX}(f)$ has the units of power per unit frequency and hence the name *power spectral density* of the random process.

2.1.2 EPSD of a non-stationary process

Clearly, the time-dependent frequency content of a signal cannot be precisely captured by ordinary Fourier analysis as the associated transform provides only the average spectral composition of a signal. Correspondently, The *evolutionary power spectral density* (EPSD) reflects the time-varying frequency-domain energy distribution of the nonstationary stochastic process. A zero-mean non stationary process can be represented by (Priestley, 1967; Stoica *et al.*, 2005):

$$f(t) = \int_{-\infty}^{\infty} A(\omega, t) e^{i\omega t} \mathrm{d}Z(\omega) \tag{6}$$

where $A(\omega, t)$ represents a deterministic modulating function and $Z(\omega)$ is a spectral process with orthogonal increments. Then the EPSD of the nonstationary process is defined as (Wang *et al.*, 2018):

$$S_{ff}(\omega, t) = |A(\omega, t)|^2 S_{\bar{f}\bar{f}}(\omega) \tag{7}$$

where $S_{\bar{f}\bar{f}}(\omega)$ is the power spectral density of the associated stationary process:

$$\bar{f}(t) = \int_{-\infty}^{\infty} e^{i\omega t} \mathrm{d}Z(\omega) \tag{8}$$

and $Z(\omega)$ satisfies:

$$\mathbb{E}[|\mathrm{d}Z(\omega)|^2] = S_{\bar{f}\bar{f}}(\omega)\mathrm{d}\omega \tag{9}$$

2.2 Spectral estimation from earthquake recordings

2.2.1 Non-parametric PSD estimators

To apply advanced spectral analysis techniques for systems and structures, power spectra provide a potent load characterisation. The periodogram estimate of estimating the power spectra based on a ground motion record $x_L(n)$ is to find the discrete time Fourier transform and appropriately scale the magnitude squared of the results.

$$P_{xx}(f) = \frac{1}{LF_s} \left| \sum_{n=0}^{L-1} x_L(n) \cdot e^{-2jfn\pi/F_s} \right|^2 \tag{10}$$

where F_s is the sampling frequency and L is the length of the recording.

To mitigate the effect of spectral leakage, windowing functions have been introduced to modify traditional periodogram method, at the cost of reducing resolution though. In comparison, Welch's method, which computes the modified periodogram of the segments of the recording and take the average, results in a smooth PSD estimate. The averaging tends to decrease the variance of the estimate relative to a single periodogram estimate of the entire data record. Still, the combined use of data segments and windowing function lead to reduced resolution. Varying the parameters in Welch method represents the tradeoff between variance reduction and resolution.

2.2.2 Evolutionary spectra estimation via Wavelet Transform

Classical Fourier analysis can provide only the average spectral composition of a signal, hence it cannot capture the time-dependent frequency change of a non-stationary signal. In comparison, several other techniques have been employed to estimate the EPSD from, for example, a ground motion time-history record, such as Wigner-Ville method (WVM) or STFT. However, it's argued that both WVM and STFT have certain limitations (Spanos & Failla, 2004). For example, a WVM time-dependent spectrum cannot reflect the actual local behavior of the process at time t, while STFT has the resolution dilemma. As such, wavelet-based spectral estimation methods have been developed (Iyama & Kuwamura, 1999). The harmonic wavelets (HW) and the generalized harmonic wavelets (GHW), which were both proposed by Newland, have been applied later in the EPSD estimation of the nonstationary stochastic process (Spanos *et al.*, 2005; Beer *et al.*, 2019).

Consider a function $f(t)$ with finite energy (i.e. $\int_{-\infty}^{\infty} |X(t)|^2 \mathrm{d}t < \infty$), the continuous wavelet transform is defined as:

$$W(u, s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(t) \cdot \Psi^*(\frac{t-u}{s}) \mathrm{d}t \tag{11}$$

in which $\Psi(t)$ is the mother wavelet function, u is the scale parameter and s is the time parameter. The asterisk denotes the complex conjugate. The function $f(t)$ can be reconstructed from $W(u, s)$ via the double integral representation:

$$f(t) = \frac{1}{2\pi C_\Psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{s^2} W(u, s) \Psi(\frac{t-u}{s}) \mathrm{d}u \mathrm{d}s \tag{12}$$

with the assumption:

$$C_\Psi = \int_{-\infty}^{\infty} \frac{|\hat{\Psi}(w)|^2}{|w|} \mathrm{d}w < \infty \tag{13}$$

where $\hat{\Psi}(w)$ represents the Fourier transform of $\Psi(t)$, defined as:

$$\hat{\Psi}(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Psi(t) e^{-iwt} \mathrm{d}t \tag{14}$$

It may be noted that wavelet transform decomposes signal $f(t)$ over dilated and translated wavelets. $W(u,s)$ represents the contribution of the function $f(t)$ in the neighborhood of $t = u$ and in the frequency band corresponding to scale s. It can also be shown that:

$$\int_{-\infty}^{\infty} |f(t)^2| \mathrm{d}t = \int_{-\infty}^{\infty} \left[\frac{1}{2\pi C_\Psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{s^2} |W(u,s)|^2 \mathrm{d}u \mathrm{d}s \right] \times |\hat{\Psi}_{u,s}(w)|^2 \mathrm{d}w \tag{15}$$

where $\hat{\Psi}_{u,s}(w)$ represents the Fourier transform of $\Psi[(t-u)/s]$ and can be given as $\hat{\Psi}_{u,s}(w) = \sqrt{s}\hat{\Psi}(sw)e^{iwu}$. Then by using Parseval's identity, one can write:

$$|F(w)|^2 = \frac{1}{2\pi C_\Psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{s^2} |W(u,s)|^2 |\hat{\Psi}_{u,s}(w)|^2 \mathrm{d}u \mathrm{d}s \tag{16}$$

where $F(w)$ represents the Fourier transform of $f(t)$. As the wavelet coefficients $W(u,s)$ provides the localized information of signal at $t = u$, the EPSD $S_{f_0 f_0}(t,w)$ can be expressed as:

$$|F(t,w)|^2 = \frac{1}{2\pi C_\Psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{s^2} |W(t,s)|^2 |\hat{\Psi}_{t,s}(w)|^2 \mathrm{d}s \tag{17}$$

It may be noted that the above equation obeys total energy equilibrium. Therefore, any wavelet basis that satisfies $\int_{-\infty}^{\infty} |\hat{\Psi}_{u,s}(w)|^2 \mathrm{d}w = 1$ can be used.

3 Stochastic modeling and simulation of ground motions

In order for the characterization, simulation, and response evaluation of ground motion processes, stochastic ground motion model is often used (Douglas & Aochi, 2008). A prevailing application is for these models to synthesize artificial spectrum-compatible ground motion processes (to provide stochastic citations) for structural nonlinear response analysis and seismic reliability evaluation. Current practices suffer from scarcity of recorded ground motions for specified earthquake scenarios and require careful scaling of recorded motion spectrum matching, which is sometimes criticized as unrealisitc (Wang *et al.*, 2018; Vlachos *et al.*, 2018; Rofooei *et al.*, 2001; Rezaeian & Der Kiureghian, 2010).

3.1 Stationary models - Kanai Tajimi model

Based on Kanai's investigation on the pattern of the spectra from many past earthquake records, Tajimi proposed the following relation for the spectral density function of the

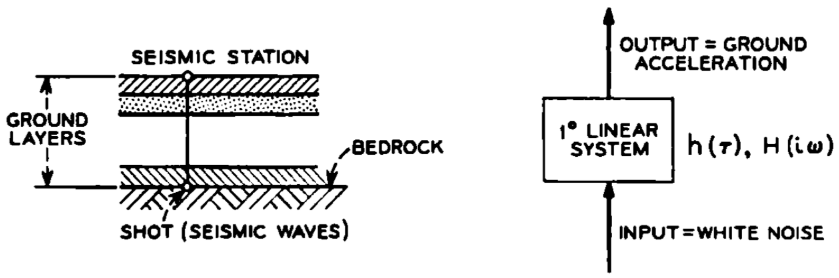


Figure 1: Ground layers represented by a linear system (Liu, 1968)

strong ground motion with a distinct dominant frequency (Tajima, 1960; Kramer, 1996; Liu, 1968).

$$G(\omega) = \frac{1 + 4\xi_g^2(\omega/\omega_g)^2}{1 - (\omega/\omega_g)^2 + 4\xi_g^2(\omega/\omega_g)^2} G_0 \tag{18}$$

where G_0 is the constant power spectral density of the input white noise process; ω_g and ξ_g can be interpreted as the characteristic frequency and damping ratio of the ground (Rofooei *et al.*, 2001). The Kanai-Tajimi power spectral density function may be interpreted as a filtered white noise process (see Fig 1), with stationary white noise excitation at the bedrock and ground layer as a single-degree-of-freedom system (Liu, 1968).

3.2 Non-stationary models for ground motion

Obviously, the most significant limitation of the Kanai-Tajimi model is its stationarity. Several attempts have therefore been made to extended it into non-stationary models (Rofooei *et al.*, 2001; Lin & Yong, 1987; Vlachos *et al.*, 2018). In more details, recorded earthquake ground motions usually exhibit nonstationarity both in frequency contents and also their intensity (Der Kiureghian & Crempien, 1989). Temporal nonstationarity refers to the variation in the intensity of the ground motion in time while spectral non-stationarity means the variation in the frequency contents. Fig. 2 shows a model that accounts for both the temporal and spectral nonstaionarities of the ground motion (Rezaeian & Der Kiureghian, 2008). It can be seen, a simulated time-history is passed through a high-pass filter to assure zero residual velocity and displacement, as well as to produce reliable response spectral values at long periods. In more details, before high-pass filtering, the ground acceleration process $x(t)$ is obtained by time-modulating a normalized filtered white-noise process with the filter having time-varying parameters. In a continuous form, the model is defined as:

$$x(t) = q(t, \boldsymbol{\alpha}) \left\{ \frac{1}{\sigma_h(t)} \int_{-\infty}^{t} h[t-\tau, \boldsymbol{\lambda}(\tau)] w(\tau) \mathrm{d}\tau \right\} \tag{19}$$

where $q(t, \boldsymbol{\alpha})$ is a deterministic, nonnegative, time-modulating function with parameters $\boldsymbol{\alpha}$ controlling shape and intensity, which has completely defined the temporal characteristics of the process. $w(\tau)$ is a white-noise process, and the integral inside the curved brackets is a filtered white noise process with $h[t-\tau, \boldsymbol{\lambda}(\tau)]$ denoting the impulse-response function of the filter with time varying parameters $\boldsymbol{\lambda}(\tau)$, which define the spectral characteristics of the process.

Over the years, many more stochastic models have been proposed from different perspectives. In summary, those representative models can be categorized into four main classes (Rezaeian & Der Kiureghian, 2008; Douglas & Aochi, 2008): (a) Processes obtained by passing a white noise through a filter, with subsequent modulation in time to achieve temporal nonstationarity. But these processes have time-invariant frequency content. (b) Processes obtained by passing a train of Poisson pulses through a filter. Through modulation in time, these processes can possess both temporal and spectral nonstationarities. However, matching with recorded ground motions is difficult. (c) Auto-regressive moving average models. By allowing the model parameters to vary with time, these models can have both temporal and spectral nonstationarity. However, it's difficult to relate the model parameters to any physical aspects of ground motion. (d) Developing a time-varying spectral representation, which requires extensive processing of the target recorded ground motion.

3.3 Simulation of ground motions from stochastic process models

Due to the scarcity of recorded ground motions for specified earthquake scenarios (magnitude, distance, type of faulting, site conditions, etc), stochastic models have been widely used to simulate ground motions to provide input excitations for structural analysis (Rofooei *et al.*, 2001; Liu, 1970; Shinozuka & Deodatis, 1991; Liang *et al.*, 2007; Wang *et al.*, 2018). Examples are the prediction of ground motion at a certain site where no past records are available, and statistical analyses of structural responses based upon very limited actual ground-motion records. (Liu, 1968) As suggested by the random vibration theory, the randomness involved in the excitations should be taken into account. In this sense, by applying the spectral representation method (see below), the EPSD models

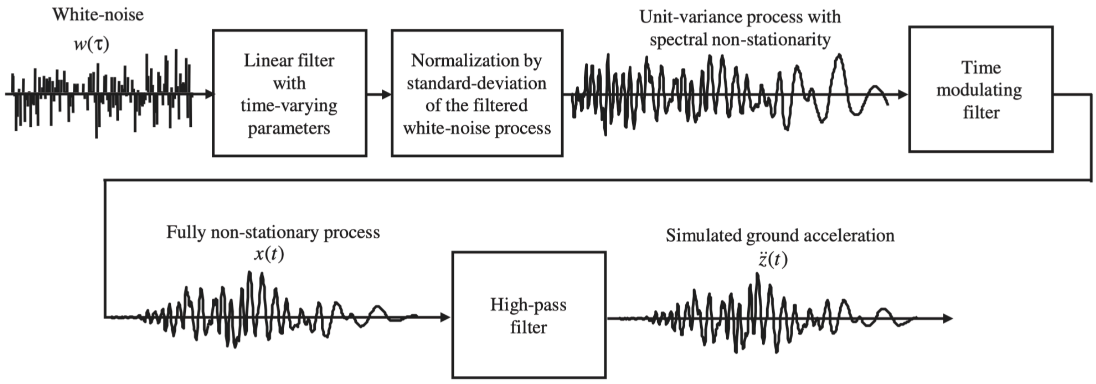


Figure 2: A stochastic ground motion model with separable temporal and spectral non-stationarities (Rezaeian & Der Kiureghian, 2008)

are able to simulate nonstationary and spectrum-compatible ground motion processes for structural nonlinear response analysis and seismic reliability evaluation.

3.3.1 Spectral representation method

Consider a 1D-1V stationary stochastic process $f_0(t)$ with mean value equal to zero, autocorrelation function $R_{f_0 f_0}(\tau)$ and a two-sided power spectral density function $S_{f_0 f_0}(\omega)$. The stochastic process can be simulated by the following series as $N \to \infty$ (Shinozuka & Deodatis, 1991):

$$f(t) = \sqrt{2} \sum_{n=0}^{N-1} A_n \cos(\omega_n t + \Phi_n) \tag{20}$$

where $A_n = \sqrt{2S_{f_0 f_0}(\omega)\Delta\omega}$, $w_n = n\Delta\omega$ in which $\Delta\omega$ is calculated by a cut-off frequency value ω_u via $\Delta\omega = \omega_u/N$, such that beyond ω_u the power spectral density function may be assumed to be zero.

A sample function $f^{(i)}(t)$ of the simulated stochastic process can be obtained by replacing the sequence of random phase angles with their respective realization:

$$f^{(i)}(t) = \sqrt{2} \sum_{n=0}^{N-1} A_n \cos(\omega_n t + \phi_n^{(i)}) \tag{21}$$

Similarly, when it comes to simulating nonstationary stochastic processes, consider a stochastic process $f_0(t)$ with mean value equal to zero, autocorrelation function $R_{f_0 f_0}(t, t+\tau)$ and a two-sided PSDF $S_{f_0 f_0}(t, \omega)$. The stochastic process can be simulated by the following series as $N \to \infty$ (Liang *et al.*, 2007):

$$f(t) = \sqrt{2} \sum_{n=0}^{N-1} \sqrt{2S_{f_0 f_0}(t, \omega_n)\Delta\omega} \cos(\omega_n t + \Phi_n) \tag{22}$$

where Φ_n are uniformly distributed random phase angles in the range of $0 \leq \Phi_n \leq 2\pi$ and N related to the discretization of the frequency domain.

3.3.2 Simulation of ground motions for certain earthquake scenarios

Many stochastic ground motion models, either stationary or nonstationary, work by fitting a parameterized stochastic model to a (target) recorded ground motion, thereby generating similar synthetic time histories. But in recent years, improved models have been proposed to model and simulate ground motions for specified earthquakes and site characteristics (Rezaeian & Der Kiureghian, 2010; Vlachos *et al.*, 2018). In order to relate the stochastic model parameters to earthquake and site characteristics of recorded motions, regression models (i.e. empirical predictive equations) have been constructed for each of the model parameters in terms of the earthquake characterization variables (e.g. F, M, R_{rep}, V_{s30}) through regression analysis of the fitted parameter values.

Rezaeian & Der Kiureghian (2008) proposed an approach to assign probability distribution to stochastic model parameters based on empirical data obtained from fitting the model to a subset of NGA strong motion database. Let θ_i denote the ith model parameter and $F_{\theta_i}(\theta_i)$ denote the marginal cumulative distribution fitted to the data. The marginal transformation is given by:

$$v_i = \Phi^{-1}[F_{\theta_i}(\theta_i)] \tag{23}$$

where $\Phi^{-1}[.]$ is the inverse of the standard normal CDF; v_i is the transformed standard Gaussian random variable, which are used as a surrogate to relate to the variables defining earthquake and site characteristics, as shown below:

$$v_i = \mu_i(\text{Earthquake, Site}, \boldsymbol{\beta_i}) + \epsilon_i \tag{24}$$

where μ_i represents a selected predictive formula for the conditional mean of v_i given the earthquake and site characteristics, and $\boldsymbol{\beta_i}$ represents the vector of regression coefficients.

3.4 Simulation of ground motion with GANs

The booming of Deep Leaning has unlocked a whole new perspective of simulating ground motions. Considering the recent explosion in seismological data collection and the many exciting developments in modern machine learning, the prospects that an artificial intelligence system could provide on-demand, accurate and realistic ground motion time histories for engineering purposes is tempting (Kong *et al.*, 2019; Florez *et al.*, 2020).

Generative models are a class of statistical models that attempt to capture the underlying probability distribution of a dataset. In particular, they are trained to produce data that looks as if it was sampled from the original training set. Much progress and successful, some even revolutionary applications, have been seen in other domains such as generation of high-resolution realistic-looking images of human faces, audio and video sequences; text generation; artistic style transfer, etc (LeCun *et al.*, 2015; Goodfellow *et al.*, 2016).

In short, GANs (Generative Adversarial Networks) enable the generation of fairly realistic synthetic images by forcing the generated images to be statistically almost indistinguishable from real ones. Conceptually, it works as two trained networks competing with each other — the first one is a *generator* that transforms a random input into a synthetic image or sequence, while the other one *discriminator* trying to differenciate between the output of the generator and real images from a training dataset. The generator network is trained to be able to fool the discriminator network, and thus it evolves toward generating increasingly realistic images as training goes on. It should be noted that, unlike classical dense neural networks that only involve gradient descent during backpropagation, GANs is a dynamic system where the optimization process is seeking not a minimum, but an equilibrium between two forces. Therefore, theoriotically speaking, GANs is much harder to train.

Some pioneering efforts have been seen towards applying GANs to generate synthetic seismograms (Li *et al.*, 2018; Wang *et al.*, 2019; Mosser *et al.*, 2020). As an effort to generate time histories with certain earthquake scenarios, some physical parameters are first identified to characterize earthquake scenarios, common choices are magnitude M, event-station distance R and shear wave speed V_{s30}.

Under the formulation of Wasserstein Generative Adversarial Networks, the discrimi-

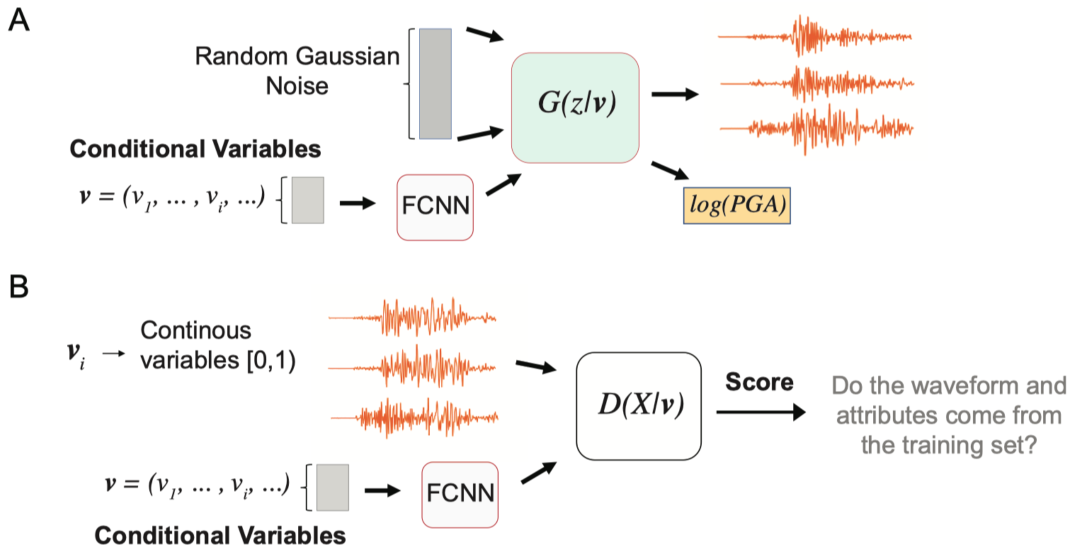


Figure 3: Diagram of the conditional GAN model (Florez *et al.*, 2020)

nator is trained to solve an optimization problem:

$$\max_D \mathbb{E}_{x \sim \mathbb{P}_r}[D(x)] - \mathbb{E}_{z \sim p}[D(G(z))] \tag{25}$$

where $D(x) \in \mathbb{R}$ and $z \sim p$ implies that z is sampled from a Gaussian distribution p. By adding a regularization term to the discriminator objective function:

$$L_D = \mathbb{E}_{z \sim p}[D(G(z))] - \mathbb{E}_{x \sim \mathbb{P}_r}[D(x)] + \lambda \mathbb{E}_{x' \sim \mathbb{P}_{x'}}[(||\nabla_{x'} D(x')||_2 - 1)^2] \tag{26}$$

where λ is a constant, and x' is uniformly sampled along straight lines connecting points in $\mathbb{P}_\mathrm{r}$ (the data distribution) and $\mathbb{P}_\mathrm{g}$ (the distribution defined by the generator model). The generator is adversarially trained by minimizing the following objective function:

$$L_G = -\mathbb{E}_{z \sim p}[D(G(z))] \tag{27}$$

Building on the above framework, Florez *et al.* (2020) proposed to combine the GAN model with a couple of earthquake-related parameters. As shown in Fig.3, a random white noise z and these extra parameters (i.e. as conditional variables $\mathbf{v}$) are combined into a mapping to output accelerogram via the generator — $G : \{\mathbf{v}, z\} \to w$. The mapping then implicitly defines a conditional probability distribution $\mathbb{P}_\mathrm{g}(\mathbf{w}|\mathbf{v})$.

4 Spectral analysis with incomplete data

A realistic characterisation of earthquake loads that reflects all their uncertainties is the key requirement, and at the same time, the greatest challenge, for a risk-informed design of structures and infrastructure in earthquake prone regions. To apply advanced spectral analysis techniques for systems and structures, evolutionary power spectra provide a potent load characterization. Their estimation requires both an in-depth understanding of the underlying physics of the problem and a relatively significant amount of evenly sampled data. However, both components are only available with a remaining degree of indeterminacy. Our models are approximate and do not capture all phenomena completely. In reality, our data are often sparse, have gaps in the records, and offer only limited insight into dependency structures.

4.1 Fitting spectrum to incomplete recording with Maximum Likelihood

Spectral analysis of earthquake recordings is one of the most fundamental analysis in many seismological applications. But in reality missing data in measurements is frequently an unavoidable situation. Therefore, a challenge remains as how to reliably fit a parametric spectrum to earthquake recordings with incomplete data.

In order to fit a parametric spectral model to an observed seismogram, the likelihood of seeing measurements $\mathbf{Y}$ given a spectrum $\mathbf{u}$, the conditional probability distribution is given as (Maranò *et al.*, 2017):

$$p(\mathbf{Y}|\mathbf{u}) = \beta \mathcal{N}(\mathbf{F}^{-1}\mathbf{y}; \mathbf{u}, \mathbf{F}^{-1}\mathbf{V}_Z\mathbf{F}^{-\mathrm{T}}) \tag{28}$$

in which β is constant, $\mathbf{V}_Z$ is the noise variance matrix and $\mathbf{F}$ represents an operation akin to an inverse Fourier transform. The maximum likelihood estimate of the parameter vector can be given as:

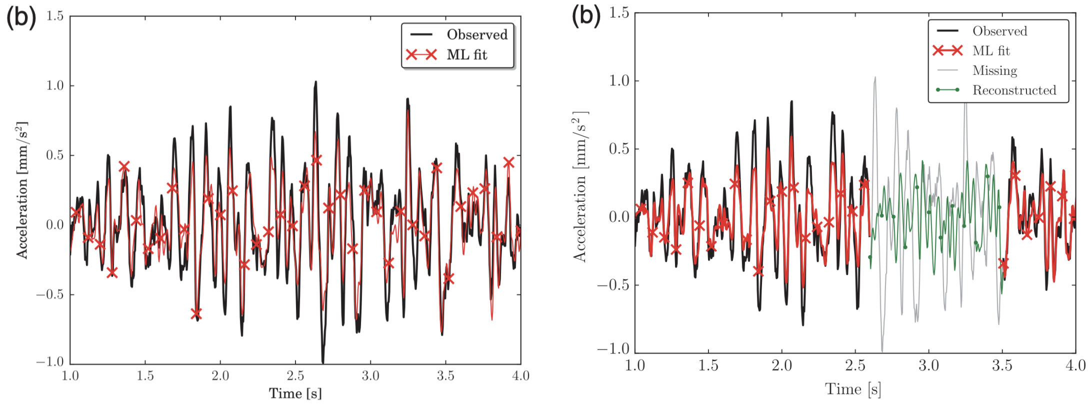
$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} (\mathbf{y} - \mathbf{F}\mathbf{u}(\boldsymbol{\eta}))^{\mathrm{T}} \mathbf{V}_Z^{-1} (\mathbf{y} - \mathbf{F}\mathbf{u}(\boldsymbol{\eta})) \tag{29}$$

Moreover, in the presence of data gaps, treat the spectrum as a random variable $\mathbf{U}$ with a Gaussian prior:

$$\mathbf{U} \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I}) \tag{30}$$

in which α is a regularization parameter to be estimated from data. The likelihood of observations as a function of α is:

$$p(\mathbf{Y} = \mathbf{y}|\alpha) = \int p(\mathbf{Y} = \mathbf{y}, \mathbf{u}|\alpha) \mathrm{d}\mathbf{u} \tag{31}$$



(a) Observed seismogram and its MLE fit

(b) Reconstructed incomplete seismogram

Figure 4: Spectra fitting using MLE with incomplete data (Maranò *et al.*, 2017)

The Maximum likelihood estimation of $\boldsymbol{\eta}$ in the incomplete case is analogous in the complete case. The conditional PDF of $\mathbf{Y}$ given $\boldsymbol{\eta}$ is:

$$p(Y|\eta) = \mathcal{N}(\mathbf{Y}; \mathbf{PFu}(\boldsymbol{\eta}), \mathbf{V}_Z) \tag{32}$$

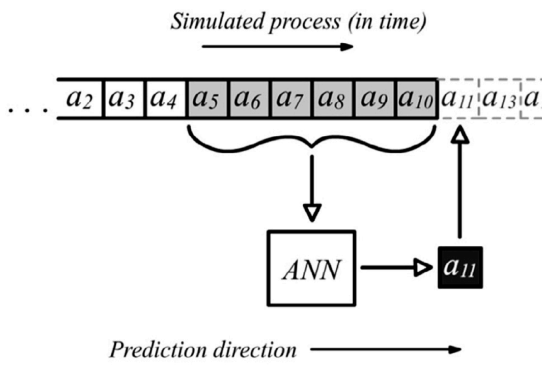
A maximum likelihood estimate of $\boldsymbol{\eta}$ is found by maximizing the likelihood, i.e. $\arg\max p(Y|\eta)$. Fig. 4 displays an example of the spectral fitting method applied to a real seismogram with incomplete data, see for (Maranò *et al.*, 2017) details.

4.2 Estimation of PSD with incomplete data using Dense neural networks

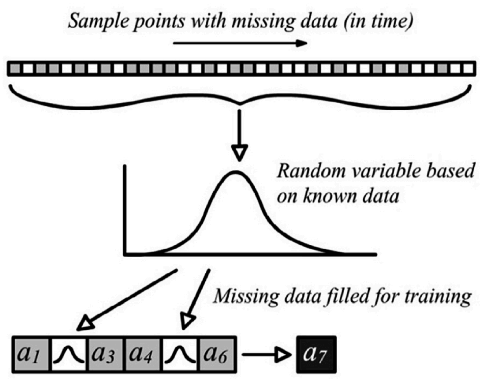
In the time domain, neural networks can be used in an autoregressive manner, trying to predict the missing time points. A statistical mapping can be established on lagged values of the variable of interest. In a simple one-step-ahead setting, based on a continuous recording, windows of data that has the length of $p + 1$ numbers of values, with the first p as features and the rest one value as target, can be extracted. However, a real challenge is we might need full data sequence to train a model in the first place. Comerford *et al.* (2015) proposed to fill in the missing data points first by drawing random values from a distribution based on the known data.

The cumulative distribution function (CDF) of known data is estimated by sorting the data in order of size, joining the CDF with a polynomial spline. Random values are drawn from:

$$a_i = F^{-1}(U_{(0,1)}) \tag{33}$$



(a) Diagram for an autoregressive model



(b) Filling missing data on a stationary process

Figure 5: Illustrations of the dense neural network model (Comerford *et al.*, 2015)

where F^{-1} is the inverse CDF of the known data and $U_{(0,1)}$ represents random values drawn from a uniform distribution.

With a gap of missing data in the middle of a data sequence, the known data before and after the gap can be adopted to train models to predict inward. Phan (2020) accordingly suggested a univariate imputation approach. The main idea of such a method can be seen in the diagram (see Fig. 6).

4.3 Uncertainty quantification with missing data imputation

Besides these efforts mentioned above, there are other deterministic imputation methods that have been proposed over years, such as Compressive Sensing (Comerford *et al.*, 2016), least squares method (Lomb, 1976; Scargle, 1982), ARIMA (Broersen & Bos, 2006), Singular spectrum analysis (Kondrashov & Ghil, 2006), CLEAN algorithm (Baisch & Bokelmann, 1999), etc. Among them, many methods first reconstruct the missing data in the time domain and then use the classical spectral analysis methodologies (such as Fourier or wavelet transform) to obtain PSD estimates. Since no method can perfectly reconstruct the missing data, the inaccuracies from the imputation will be propagated to further spectral estimations. Therefore, these estimates could be misleading as they provide no information about the degree of uncertainty related to the original incomplete data. This leads to the need of considering the uncertainty in spectral estimates when confronting missing data in the signals.

Based on one type of interpretation of a stochastic process that considers it as a collection of random variables indexed by time (in the context of discrete-time stochastic processes). Naturally, one would think about modeling each missing data point as random

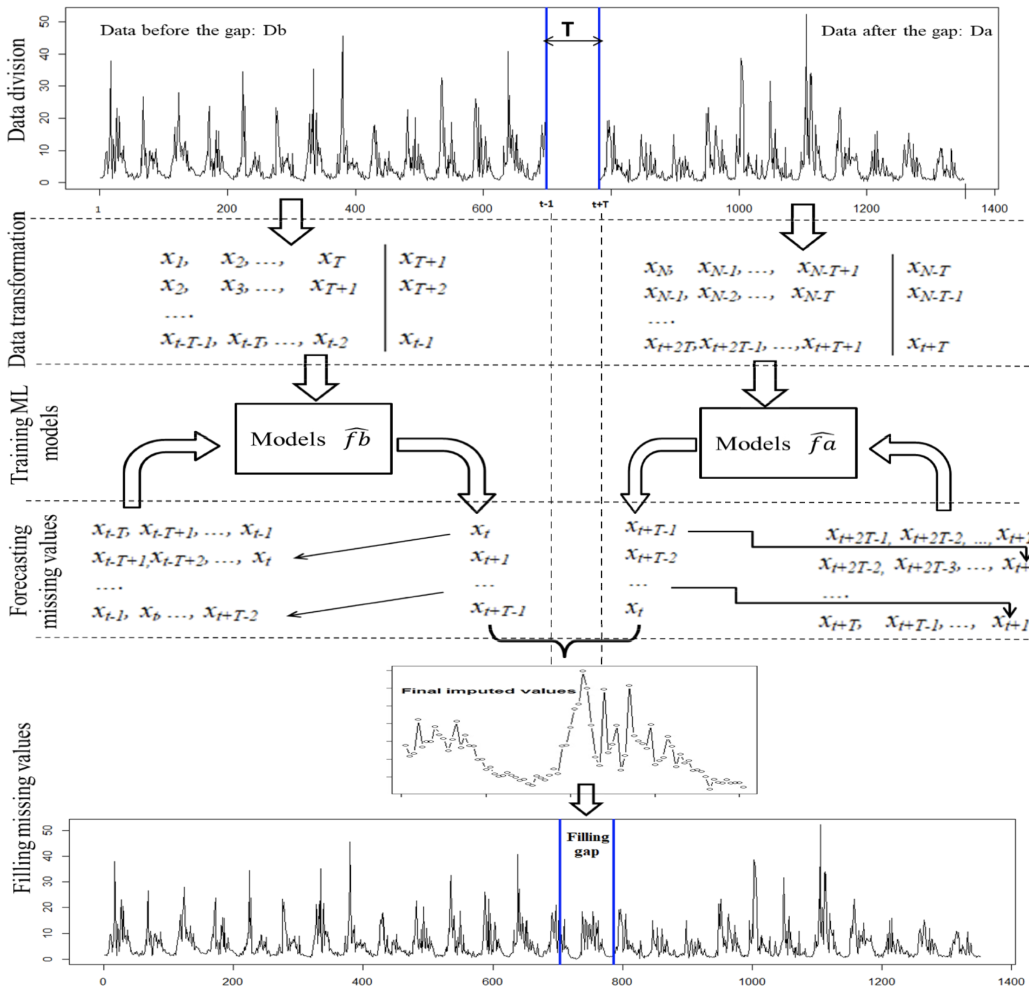


Figure 6: Scheme of univariate missing data imputation approach (Phan, 2020)

variable and thus filling in the missing value with a random sample from its underlying probability distribution. The question, however, being how to determine the PDF of the random variable X_i for the stochastic process $\{X(t)\}_{t \in T}$ so that one can draw samples from it? Naively, Comerford *et al.* (2015) proposed to fill in missing data points using samples from standard Gaussian distribution, provided the original signal being normalized first. In this manner, many realizations of imputation can be obtained by Monte Carlo simulations and thus an ensemble of PSD estimates regarding each frequency component can be further obtained. Such a probabilistic power spectrum then provides a tool to express the uncertainty on the PSD estimates under missing data. Importantly, a closed-form expression has been derived for the PDF of the power spectrum value corresponding to a specific frequency value, see (Comerford *et al.*, 2015) for details.

Building on the assumption of modeling missing data as random variables, Zhang

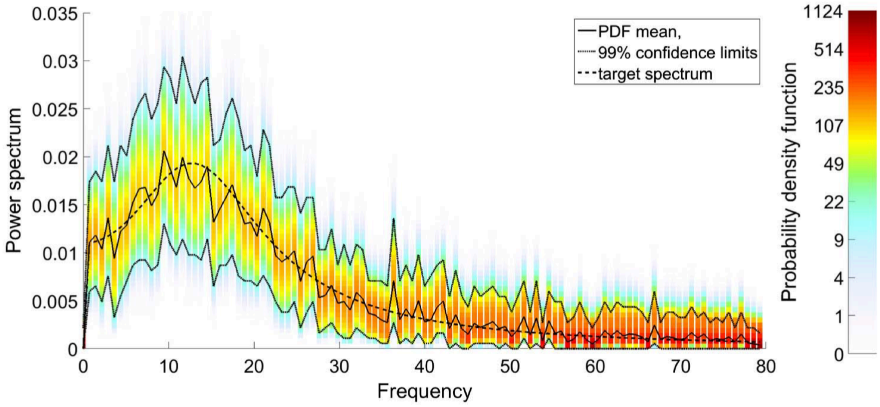


Figure 7: PSD with 10% missing data replaced by correlated Gaussian random variables (Zhang *et al.*, 2017)

et al. (2017) model the missing data using a multivariate Gaussian distribution in which correlation between missing points (i.e. random variables) are taken into account, and Kriging model is used to estimate the mean and covariance matrix of the missing data. Fig. 7 displays the results of filling in missing data to a stationary realization of time history generated from a Kanai-Tajimi PSD model. Notably, the PDF of the power spectral density estimates corresponding to each frequency component can also be seen in this figure.

5 Final remarks and conclusions

This deliverable presents a concise review of the recent literature regarding simulating ground motions via stochastic models and spectral estimation methods given incomplete ground motion recordings.

Due to the scarcity of recorded ground motions for specified earthquake scenarios, there's a trend, building on many stochastic process models (stationary and non-stationary), to link a stochastic model with certain earthquake scenarios. For example they usually relate model parameters with selected earthquake characterization variables, such as M, R, V_{s30}. It can also be noticed the latest attempt of using trending Generative modeling techniques (such as GANs) in Deep Learning in this direction. As a method known for excellent performance in many other domains, such as human face generation, music generation, etc., Generative modeling techniques, which are also inherently probabilistic, might provide another perspective to classical stochastic process methods.

For the challenge of spectral representation with incomplete data, it's suggested that an important part of this challenge is quantifying the uncertainties over the imputation as no reconstruction is perfect and such uncertainties will be further propagated to further spectral analyses. An attempt of using Dense neural networks in an autoregressive manner, which is in essence doing time series forecasting, is summarized in this report. But the prospects of using a Recurrent neural network, which is another Deep Learning architecture renowned for dealing with sequence data such as text generation, hasn't been examined yet. Also, a Bayesian version of these neural network models seem to be even more suitable considering the goal of uncertainty quantification. Implicit to the use of these Bayesian neural networks that indeed work in time domain is the challenge of how to train these models given incomplete data since, in normal settings, we may need complete data in the first place to train a neural network model.

Alternative to methods that reconstruct the incomplete recording in the time domain, the booming developments in the field of Generative modeling has provided new perspectives for doing missing data imputation based on two-dimensional data, such as spectrograms (a two-dimensional time-frequency representation). Similarly, the challenge exists as how to train a reliable model given the scarcity of recorded ground motions for specified earthquake scenarios. Simulations of ground motion by using a stochastic model or a GAN model might be handy in this regard, but under this scheme, a comprehensive characterization of the uncertainty propagation will be indispensable.

References

Baisch, Stefan, & Bokelmann, Götz H. R. 1999. Spectral analysis with incomplete time series: an example from seismology. *Computers & Geosciences*, **25**(7), 739–750.

Beer, Michael, Zhang, Yi, Quek, Ser Tong, & Phoon, Kok Kwang. 2013. Reliability analysis with scarce information: Comparing alternative approaches in a geotechnical engineering context. *Structural Safety*, **41**(41), 1–10.

Beer, Michael, Gholamy, Afshin, & Kreinovich, Vladik. 2019. A Theoretical Explanation for the Efficiency of Generalized Harmonic Wavelets in Engineering and Seismic Spectral Analysis.

Broersen, P.M.T., & Bos, R. 2006. Time-series analysis if data are randomly missing. *IEEE Transactions on Instrumentation and Measurement*, **55**(1), 79–84.

Ching, Jianye, & Phoon, Kok-Kwang. 2019. Modeling Multivariate,Uncertain,Sparse, and Incomplete Site Investigation Data with Spatial Variation (MUSIC-X). *In: Proceedings of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR 2019)*.

Comerford, Liam, Kougioumtzoglou, Ioannis A, & Beer, Michael. 2015. An artificial neural network approach for stochastic process power spectrum estimation subject to missing data. *Structural Safety*, **52**, 150–160.

Comerford, Liam, Kougioumtzoglou, Ioannis A., & Beer, Michael. 2015. On quantifying the uncertainty of stochastic process power spectrum estimates subject to missing data. *International Journal of Sustainable Materials and Structural Systems*, **2**, 185.

Comerford, Liam, Kougioumtzoglou, Ioannis A., & Beer, Michael. 2016. Compressive sensing based stochastic process power spectrum estimation subject to missing data. *Probabilistic Engineering Mechanics*, **44**, 66–76.

Der Kiureghian, Armen, & Crempien, Jorge. 1989. An evolutionary model for earthquake ground motion. *Structural safety*, **6**(2-4), 235–246.

Douglas, John, & Aochi, Hideo. 2008. A Survey of Techniques for Predicting Earthquake Ground Motions for Engineering Purposes. *Surveys in Geophysics*, **29**(3), 187–220.

Douglas, John, & Aochi, Hideo. 2008. A survey of techniques for predicting earthquake ground motions for engineering purposes. *Surveys in geophysics*, **29**(3), 187.

Florez, Manuel A., Caporale, Michaelangelo, Buabthong, Pakpoom, Ross, Zachary E., Asimaki, Domniki, & Meier, Men-Andrin. 2020. Data-driven Accelerogram Synthesis using Deep Generative Models. *arXiv preprint arXiv:2011.09038*.

Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Iyama, Jun, & Kuwamura, Hitoshi. 1999. Application of wavelets to analysis and simulation of earthquake motions. *Earthquake Engineering & Structural Dynamics*, **28**(3), 255–272.

Kondrashov, Dmitri, & Ghil, Michael. 2006. Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, **13**(2), 151–159.

Kong, Qingkai, Trugman, Daniel T., Ross, Zachary E., Bianco, Michael J., Meade, Brendan J., & Gerstoft, Peter. 2019. Machine Learning in Seismology: Turning Data into Insights. *Seismological Research Letters*, **90**(1), 3–14.

Kramer, Steven Lawrence. 1996. *Geotechnical Earthquake Engineering.*

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey. 2015. Deep learning. *Nature*, **521**(7553), 436–444.

Li, Jie, & Chen, Jianbing. 2009. *Stochastic Dynamics of Structures.*

Li, Zefeng, Meier, Men-Andrin, Hauksson, Egill, Zhan, Zhongwen, & Andrews, Jennifer. 2018. Machine Learning Seismic Wave Discrimination: Application to Earthquake Early Warning. *Geophysical Research Letters*, **45**(10), 4773–4779.

Liang, Jianwen, Chaudhuri, Samit Ray, & Shinozuka, Masanobu. 2007. Simulation of Nonstationary Stochastic Processes by Spectral Representation. *Journal of Engineering Mechanics-asce*, **133**(6), 616–627.

Lin, Y. K., & Yong, Yan. 1987. Evolutionary Kanai-Tajimi Earthquake Models. *Journal of Engineering Mechanics-asce*, **113**(8), 1119–1137.

Liu, S. C. 1968. Statistical analysis and stochastic simulation of ground-motion data. *The Bell System Technical Journal*, **47**(10), 2273–2298.

Liu, SC. 1968. Statistical Analysis and Stochastic Simulation of Ground-Motion Data. *Bell System Technical Journal*, **47**(10), 2273–2298.

Liu, Shin-Chi. 1970. Evolutionary power spectral density of strong-motion earthquakes. *Bulletin of the Seismological Society of America*, **60**(3), 891–900.

Lomb, N. R. 1976. Least-Squares Frequency Analysis of Unequally Spaced Data. *Astrophysics and Space Science*, **39**(2), 447–462.

Maranò, Stefano, Edwards, Benjamin, Ferrari, Graziano, & Fäh, Donat. 2017. Fitting Earthquake Spectra: Colored Noise and Incomplete Data. *Bulletin of the Seismological Society of America*, **107**(1), 276–291.

Miller, Scott L., & Childers, Donald G. 2012. *Probability and Random Processes : With Applications to Signal Processing and Communications.*

Mosser, Lukas, Dubrule, Olivier, & Blunt, Martin J. 2020. Stochastic Seismic Waveform Inversion Using Generative Adversarial Networks as a Geological Prior. *Mathematical Geosciences*, **52**(1), 53–79.

Phan, Thi-Thu-Hong. 2020. Machine Learning for Univariate Time Series Imputation. *In: 2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR).*

Priestley, MB. 1967. Power spectral analysis of non-stationary random processes. *Journal of Sound and Vibration*, **6**(1), 86–97.

Rezaeian, Sanaz, & Der Kiureghian, Armen. 2008. A stochastic ground motion model with separable temporal and spectral nonstationarities. *Earthquake Engineering & Structural Dynamics*, **37**(13), 1565–1584.

Rezaeian, Sanaz, & Der Kiureghian, Armen. 2010. Simulation of synthetic ground motions for specified earthquake and site characteristics. *Earthquake Engineering & Structural Dynamics*, **39**(10), 1155–1180.

Rofooei, FR, Mobarake, Aghababaii, & Ahmadi, G. 2001. Generation of artificial earthquake records with a nonstationary Kanai–Tajimi model. *Engineering Structures*, **23**(7), 827–837.

Scargle, J. D. 1982. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, **263**, 835–853.

Shinozuka, Masanobu, & Deodatis, George. 1991. Simulation of stochastic processes by spectral representation.

Spanos, Pol D, & Failla, Giuseppe. 2004. Evolutionary spectra estimation using wavelets. *Journal of Engineering Mechanics*, **130**(8), 952–960.

Spanos, Pol D., Tezcan, Jale, & Tratskas, Petros. 2005. Stochastic processes evolutionary spectrum estimation via harmonic wavelets. *Computer Methods in Applied Mechanics and Engineering*, **194**(12), 1367–1383.

Stoica, Petre, Moses, Randolph L, *et al.* 2005. Spectral analysis of signals.

Tajima, H. 1960. A Statistical Method of Determining the Maximum Response of a Building Structure During an Earthquake. *Proc. 2nd WCEE*, 781–797.

Vlachos, Christos, Papakonstantinou, Konstantinos G, & Deodatis, George. 2018. Predictive model for site specific simulation of ground motions based on earthquake scenarios. *Earthquake Engineering & Structural Dynamics*, **47**(1), 195–218.

Wang, Ding, Fan, Zenglei, Hao, Shengwang, & Zhao, Dahai. 2018. An evolutionary power spectrum model of fully nonstationary seismic ground motion. *Soil Dynamics and Earthquake Engineering*, **105**, 1–10.

Wang, Ding, Fan, Zenglei, Hao, Shengwang, & Zhao, Dahai. 2018. An evolutionary power spectrum model of fully nonstationary seismic ground motion. *Soil Dynamics and Earthquake Engineering*, **105**, 1–10.

Wang, Tiantong, Trugman, Daniel, & Lin, Youzuo. 2019. SeismoGen: Seismic Waveform Synthesis Using Generative Adversarial Networks. *arXiv: Learning.*

Zhang, Yuanjin, Comerford, Liam, Kougioumtzoglou, Ioannis A., Patelli, Edoardo, & Beer, Michael. 2017. Uncertainty Quantification of Power Spectrum and Spectral Moments Estimates Subject to Missing Data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, **3**(4), 4017020.