Capturing Epistemic Uncertainties in the Power Spectral Density for Limited Data Sets

Marco Behrendt[a], Matthias G.R. Faes[b], Marcos A. Valdebenito[c], and Michael Beer[a,d,e]

[a] Institute for Risk and Reliability, Leibniz Universität Hannover, Germany, behrendt@irz.uni-hannover.de

[b] Chair for Reliable Engineering, Technische Universität Dortmund, Germany

[c] Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Viña del Mar, Chile

[d] Institute for Risk and Uncertainty, University of Liverpool, United Kingdom

[e] International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, Shanghai, China

Abstract: In stochastic dynamics, it is indispensable to model environmental processes in order to design structures safely or to determine the reliability of existing structures. Wind loads or earthquakes are examples of these environmental processes and may be described by stochastic processes. Such a process can be characterised by means of the power spectral density (PSD) function in the frequency domain. Based on the PSD function, governing frequencies and their amplitudes can be determined. For the reliable generation of such a load model described by a PSD function, uncertainties that occur in time signals must be taken into account. In this paper, an approach is presented to derive an imprecise PSD model from a limited amount of data. The spectral densities at each frequency are described by reliable bounds instead of relying on discrete values. The advantages of the imprecise PSD model are illustrated and validated with numerical examples in the field of stochastic dynamics.

1. INTRODUCTION

The consideration of uncertainties in data sets is of paramount importance in engineering and especially in the field of stochastic dynamics. Buildings and structures are subject to random vibrations induced, for example, by environmental processes such as earthquakes or wind loads. In order to obtain reliable simulation results for existing structures or for the design of future buildings, it is essential to consider uncertainties. A commonly used model in stochastic dynamics is the power spectral density (PSD) function, which can be used to represent stochastic processes in the frequency domain and thus identify dominant frequencies. The PSD can also be used to generate time signals representing the characteristics of this underlying PSD [1-3].

For reliability analysis of buildings and structures, real data or artificially generated data can be used. While artificial data can only represent an environmental process realistically to a certain degree, real data are often subject to uncertainties. These can result from measurement errors, damaged sensors or simply from limited data. In order to obtain reliable simulation results, these uncertainties must be taken into account into the representation of the physical process. If these uncertainties are not taken into account or are incorrectly quantified, this can lead to fatally incorrect interpretations of the results, possibly classifying a building as safe under a particular load when in fact it poses a high risk of damage or collapse. Therefore, every attempt must be made to account for uncertainties when generating load models and simulations [4].

Some approaches for estimating PSD functions have already been introduced that can account for uncertainty in the data, as well as quantify it. For example, in [5,6] the problem of missing data is addressed. These missing data are reconstructed and assumed to be normally distributed. The probability distributions of the missing are then propagated through the discrete Fourier transform to quantify the uncertainties in the frequency domain. In [7], a large set of accelerograms is used to determine interval parameters for a semi-empirical PSD function. Thus, different representations of the

PSD functions result, depending on the bounds used for the derived interval parameters. A relaxed PSD function, based on a large data set of similar signals transformed into the frequency domain, is derived in [8]. Since it is possible to extract robust statistical information from a large amount of data, the relaxed PSD provides a probabilistic representation of the data in the frequency domain. Although these are different approaches, they all have in common that the PSD functions are not treated as purely deterministic functions, as it is usually the case.

In this work, the problem of limited data is considered. Commonly used estimators of the PSD function, such as the periodogram, could lead to a highly unrepresentative model under scarce data, so that the simulation results may not reflect the actual response behaviour of the system under investigation. If not sufficient data are available, the actual underlying PSD function cannot be estimated with certainty from the data sets. Since reliable statistical information cannot be derived from a small amount of data, this paper proposes an interval approach to define reliable bounds without considering the distribution within these bounds. The estimation of the proposed imprecise PSD is carried out entirely in the frequency domain, using a radial basis function (RBF) network [9]. An interval parameter will be added to the basis functions to obtain an approximation of an upper and lower bound, which needs to be optimised considering the actual minimum and maximum of the available data. After the optimisation, reliable bounds of the data set are derived that reflect the physics of the data and also take into account dependencies between frequencies.

This paper is structured as follows: A brief overview of PSD estimation, stochastic processes and RBF networks is given in Section 2. The estimation of the imprecise PSD is elaborated in Section 3. In Section 4, the derived model is employed on numerical examples to demonstrate its applicability and strengths. The conclusions are given in Section 5.

2. PRELIMINARIES

This section introduces some basic theoretical concepts that are relevant for the derivation and understanding of the imprecise PSD model introduced later in this work.

2.1. PSD Estimation and Stochastic Processes

A stochastic process is affected by random occurrences; therefore, it cannot be described in a purely deterministic way. The stochastic process at any time is determined by random variables, see e.g., [10].

If no data are available or do not meet the requirements for the simulation, artificially generated stochastic processes can be used for the simulations as an approximation to real stochastic processes. Such a process can be generated using the Spectral Representation Method (SRM) [11]. SRM requires an analytical or empirical function of a PSD S_X to construct a stochastic process X_t with their underlying characteristics. SRM reads as follows

$$X_t = \sum_{n=0}^{N_\omega - 1} \sqrt{4 S_X(\omega_n)\Delta\omega}\,(\omega_n t + \varphi_n) \tag{1}$$

where

$$\omega_n = n\Delta\omega, \quad n = 0,1,2,\ldots,N_\omega - 1 \tag{2}$$

with N_ω as the total number of frequency points, $\omega_n$as the frequency vector, $\Delta\omega$ as frequency step size, φ_n as uniformly distributed random phase angles in the range $[0, 2\pi]$ and t as time vector. This provides a suitable method for generating compatible time signals derived from and carrying the characteristics of the underlying PSD function S_X.

The estimation of the PSD function of a stationary stochastic process can be obtained by the periodogram [12,13], which is formed by the squared absolute value of the discrete Fourier transform of the signal x_t. The periodogram reads as follows

$$\hat{S}_X(\omega_k) = \frac{\Delta t^2}{T}\left|\sum_{t=0}^{T-1} x_t e^{-\frac{i2\pi kt}{T}}\right|^2 \tag{3}$$

where Δt is the time step size, T is the total length of the record, t describes the data point index in the data record and k is the integer frequency for $\omega_k = \frac{2\pi k}{T}$.

2.2. RBF Networks

A radial basis function (RBF) network is a class of artificial neural networks [9]. It typically consists of three layers, namely the input layer, the hidden layer and the output layer. It is used to interpolate or approximate functions from the n-dimensional input space to the scalar output space but can be extended to a multi-output network. Thus, in this work the RBF network is a mapping of y: $\mathbb{R}^{N_\omega} \to \mathbb{R}$.

The input layer of an RBF network passes the input data to the hidden layer. The hidden layer consists of a number of m neurons whose activation functions are RBFs, which are characterised by the fact that they are symmetrical around their assigned centre c_i. In this work, the RBF

$$\phi_i(x) = e^{-\left(\|x-c_i\|\cdot b_{\phi_i}\right)^2} \tag{4}$$

is used, where $\|x - c_i\| \cdot b_{\phi_i}$ describes the Euclidean distance from the input x to the designated centre c_i multiplied with a bias b_{ϕ_i}, which indicates the spread of the neuron.
The function values of the RBFs based on the input data are propagated to the output layer, where a weighted linear combination of all neurons takes place. The weights w_i of all neurons can be determined with a linear least squares method. In addition, to manipulate the sensitivity of a neuron, a bias b_0 can be employed. Thus, the RBF network results in

$$y(x) = \sum_{i=1}^{m} w_i\, \phi_i\left(\|x - c_i\| \cdot b_{\phi_i}\right) + b_0 \qquad x \in \mathbb{R}^{N_\omega}. \tag{5}$$

For an exact interpolation of a function, the number of RBFs m must be equal to the number of data points N_ω. In general, however, exact function interpolation is not necessary. Often the input data are noisy, so it is advisable to approximate a smoother function and thus average out the noise. In addition, for an exact interpolation the number of neurons can be prohibitively high, which leads to a significantly higher computational effort. In the case of an approximation, the number of RBFs m is usually less than the number of data points N_ω.

For more information on RBF networks, such as training and validation of the network, the reader is referred to [14-16] and the references therein.

3. THE IMPRECISE PSD FUNCTION

For the derivation of the imprecise PSD function, the Kanai-Tajimi PSD function of the form

$$S_{KT}(\omega) = S_0 \frac{1 + 4\xi^2 \frac{\omega^2}{\omega_p^2}}{\left(1 - \frac{\omega^2}{\omega_p^2}\right)^2 + 4\xi^2 \frac{\omega^2}{\omega_p^2}} \tag{6}$$

is utilised in this section and throughout this work. In this equation, S_0 is a constant, ω_p describes the peak frequency and ξ indicates the sharpness of the peak [17,18]. In this work $S_0 = 0.25$, $\omega_p = 3\pi$ rad/s and $\xi = 0.5$ are utilised to describe the PSD function. Furthermore, the upper cut-off frequency is defined to be $\omega_u = 50$ rad/s.

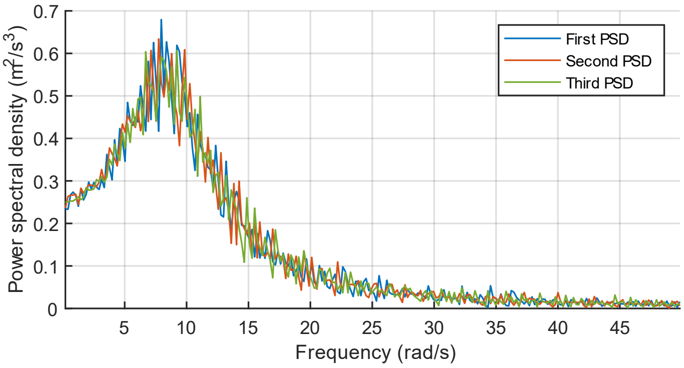
3.1. Estimation of an Imprecise PSD Function

For robust simulation results considering uncertainties introduced by the limited number of data and the PSD estimation processes in general, it is proposed to derive an imprecise PSD function, i.e., a PSD function with reliable upper and lower bound. The estimation process is carried out entirely in the frequency domain. After transformation of the excitation/signals given in the time domain, an ensemble of PSD functions results, see Fig. 1. Based on such an ensemble, the imprecise PSD function can be derived performing the following steps:

1. Identification of the basis power spectrum of the ensemble
2. Fitting an RBF network to the basis power spectrum
3. Adding/subtracting an interval parameter to/from each basis function to obtain a first approximation of the bounds
4. Optimisation of the interval parameter with minimum of constrained non-linear multi-variable function to obtain reliable bounds depending on the ensemble minimum and maximum

These steps will be discussed in the subsequent sections in details.

Figure 1: Ensemble consisting of 3 PSD functions utilised to estimate the imprecise PSD function.



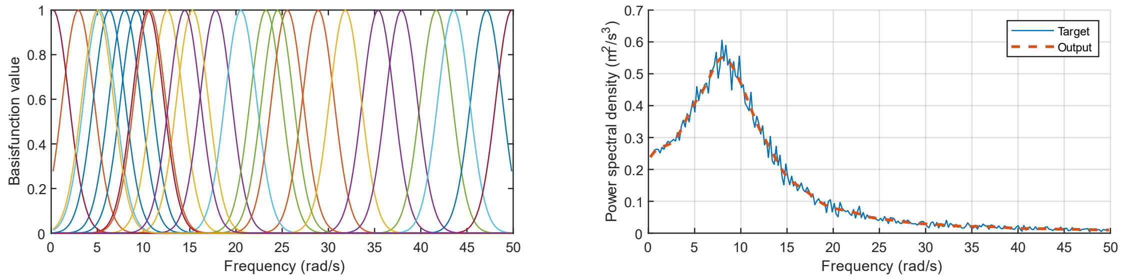
3.1.1. Basis Power Spectrum

The basis power spectrum can be identified using different approaches. As the imprecise PSD function delivers an upper and lower bound at each frequency regardless of any distribution of the data within those bounds, the midpoint spectrum, i.e., the midpoint between minimum S_{min} and maximum S_{max} of the ensemble at each frequency component, is suggested in this work.

$$S_{basis} = \frac{1}{2}(S_{max} + S_{min}) \tag{7}$$

As the choice of the basis power spectrum also depends on the shape of the data, other basis power spectra such as mean power spectrum could be possible as well.

Figure 2: Basis functions used to approximate the midpoint spectrum.



3.1.2. Fitting an RBF Network

For an exact interpolation of the basis spectrum, it is required to use as many neurons (i.e., basis functions) as frequency points in the ensemble. As such a representation will often yield in a highly spiky power spectrum and the subsequent optimisation of the bounds will yield in the minimum and maximum value of the ensemble at each frequency, it is advisable to choose a lower number of neurons. This results in a smoother approximated midpoint spectrum. The number of neurons as well as the spread of the neurons are crucial here, as an unfavourable choice of these parameters can lead to unreliable results. For example, the bounds of the imprecise PSD may be too wide or too narrow and therefore not correspond to the data set. This will falsify the simulation results. In addition, the RBF network thus operates as a smoother for its realisations.

A reasonable first approximation for the number of the basis functions m is about 5% - 15% of the total number of frequency points N_ω and for the bias the range $b_{\phi_i} = [0.25, 0.45]$. However, it should be noted that the choice of m and b_{ϕ_i} depend on the appearance of the PSD function to be fitted and may deviate from the proposed values. There are several approaches in the literature to find a set of optimal parameters, such as pruning methods, see e.g., [16,19] and references therein. In the given example the number of frequency points is $N_\omega = 238$, while the number of neurons is $m = 25$ and the bias is $b_{\phi_i} = 0.4$. In Fig. 2 the basis functions of the RBF network for deriving the approximated basis power spectrum S_{basis} are depicted, as well as the resulting basis spectrum and target spectrum itself.

3.1.3. Optimisation of the Bounds

The optimisation of the bounds requires the definition of a vector $\delta \in \mathbb{R}^{N_\omega}$ as optimisation parameter, which controls the sensitivity of the respective frequency components and thus the distance between the basis power spectrum S_{basis} and the upper and lower bound. At each discrete frequency ω_n the corresponding component δ_n is added to the basis functions to obtain the upper bound

$$\overline{S_{opt}}(\omega_n; \delta_n) = \sum_{i=1}^{m} w_i \left(\phi_i(\omega_n) + \delta_n\right) + b_0 \tag{8}$$

and subtracted from the basis functions to derive the lower bound

$$\underline{S_{opt}}(\omega_n; \delta_n) = \sum_{i=1}^{m} w_i \left(\phi_i(\omega_n) - \delta_n\right) + b_0 \tag{9}$$

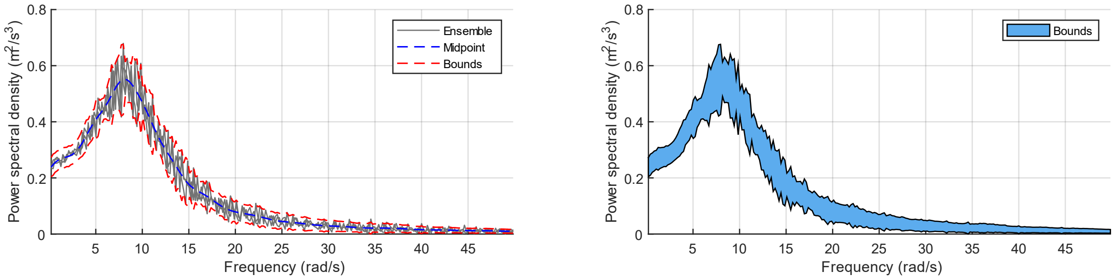
with ω_n and n as defined in Eq. 2. The weights w_i and the bias b_0 result from fitting the RBF network to the basis power spectrum S_{basis}.

To ensure that representative and reliable bounds are derived for the data set, the norm of the difference between the upper and lower bound will be the objective function for the optimisation. This optimisation is subject to the condition, such that the resulting upper bound shall be larger than the maximum of the ensemble and the resulting lower bound shall be smaller than the minimum of the ensemble to ensure that all data points are included in the bounds. For physical reasons the lower bound must not be smaller than 0 by all means as negative values are not possible in terms of power spectral densities. The optimisation problem results as follows

$$\begin{aligned} \min \quad & \left|\overline{S_{opt}}(\omega;\delta) - \underline{S_{opt}}(\omega;\delta)\right| \\ s.t. \quad & \overline{S_{opt}}(\omega;\delta) \geq S_{max}(\omega) \\ & \underline{S_{opt}}(\omega;\delta) \leq S_{min}(\omega) \\ & \underline{S_{opt}}(\omega;\delta) \geq 0 \end{aligned} \tag{10}$$

When the parameter δ is optimised, reliable bounds can be provided. The imprecise PSD function for the ensemble given in Fig. 1 is shown in Fig. 3.

Figure 3: Bounds of the estimated imprecise PSD function.



3.2. Sampling from the Imprecise PSD Function

The derived model of the imprecise PSD function is directly applicable for Monte Carlo (MC) simulations. Any arbitrary PSD function sampled within the bounds is a possible PSD function can be used to generate a stochastic process with SRM (Eq. 1). Each interval is considered as a uniform distribution from which samples are drawn. Although this is a very rudimentary way of propagating intervals and is not an appropriate representation of an interval, the uniform distribution is only used as a tool to evenly sample the solution space in the uncertain input space and to illustrate how the solution is approximated.

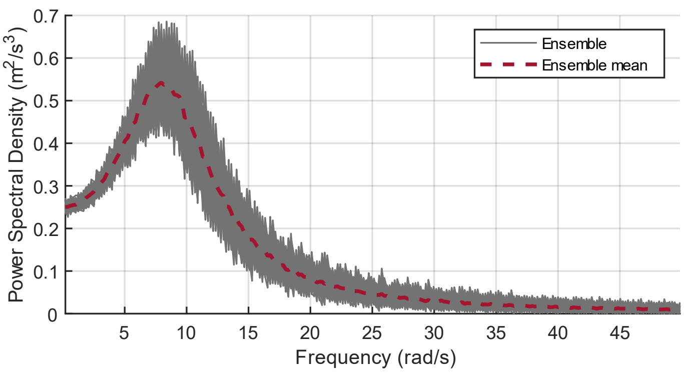
The total energy of the sampled PSD function S_{sample}, i.e., the integral of the PSD function, needs to match the energy of the midpoint spectrum S_{basis} with a certain tolerance. This guarantees that no unrealistic PSD functions are sampled. For instance, a sampled PSD function consisting only of the upper value at each frequency (i.e., the upper bound) does not realistically reflect the stochastic process. This is particularly important for imprecise PSD functions with wide bounds. The sampled PSD function can be utilised to generate an adequate stochastic process by using Eq. 1. By taking into account the total energy, PSD functions that realistically represent the stochastic process can be approximated while still covering the entire range of the imprecise PSD function, which, on the other hand, provides a full consideration of the epistemic uncertainty.

$$\left|\sum_{n=0}^{N_\omega - 1} S_{basis}(\omega_n) - \sum_{n=0}^{N_\omega - 1} S_{sample}(\omega_n)\right| \leq \Delta_{tol} \tag{11}$$

4. NUMERICAL EXAMPLES

To validate the results of the imprecise PSD function, a comparison is made using an ensemble of 100 artificially generated time signals transformed into the frequency domain, from which the mean value is calculated, see Fig. 4. The reason for this approach is that the mean of a large data set approximates the true underlying power spectrum. Furthermore, this procedure ensures that the stochastic fluctuations introduced by the random variables in Eq. 1 have a negligible influence. The sparsity of the data is emulated by selecting only 3 randomly chosen PSD functions from this ensemble to estimate the imprecise PSD function.

Figure 4: Ensemble consisting of 100 PSD functions utilised to estimate the mean.



For the numerical examples a Single-Degree-of-Freedom (SDOF) system is utilised. The equation of motion of such a system reads as follows:

$$m\ddot{x}(t) + c\dot{x}(t) + kx(t) = F(t) \tag{12}$$

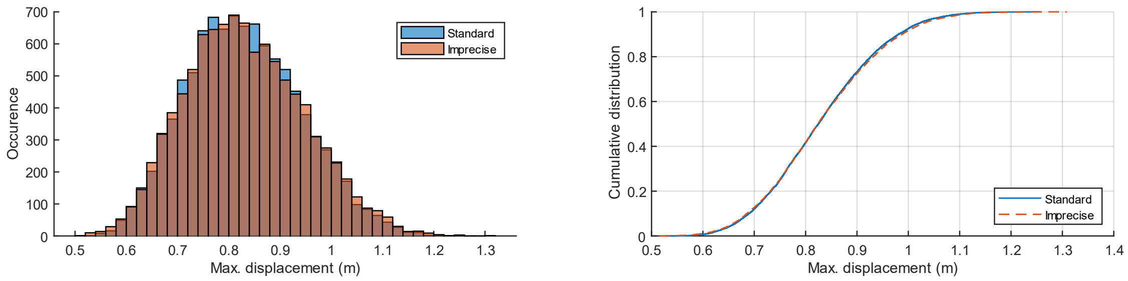
In this equation, m, c and k denote the mass, the damping coefficient and the spring constant, respectively, and $\ddot{x}$, $\dot{x}$ and x describe the acceleration, the velocity and the displacement of the system, respectively. $F(t)$ denotes the external excitation force.

For comparison, the numerical examples for both the ensemble mean and the imprecise PSD function are performed with a total of 10,000 MC samples in the time domain to obtain reliable results. The maximum system displacement is adopted as the quantity of interest. The histogram and the cumulative distribution function (CDF) are drawn from the absolute maximum system displacement for each individual MC sample to obtain an approximation of the probability of failure.

4.1. Example 1

In the first examples, the following system parameters are used: $m = 50$ kg, $c = 5$ Ns/m and $k = 4441.3$ N/m. The natural frequency is thus $\omega = \sqrt{k/m} \approx 3\pi$ rad/s and therefore approximately the peak frequency ω_p of the excitation. The results of the MC simulation are shown in Fig. 5. On the one hand, the histogram of the maximum system displacements after the excitation of the system with the ensemble mean as well as the imprecise PSD function, and on the other hand the corresponding CDF are shown. It can be seen that there is very little difference between ensemble mean and the imprecise PSD function. In the histogram, the range of maximum system displacements is almost identical, ranging from about 0.5 m to 1.2 m for both models. The frequency of occurrence of the displacements is also identical and has its highest probability at a system displacement of about 0.8 m. The resulting CDF of the imprecise PSD function shows only minimal differences to the CDF of the ensemble mean. As a result, the probabilities of failure at a given system displacement are identified to be similar for both excitation models. These results show that valid results can be obtained with the imprecise PSD function based on sparse data and that they are not falsified.

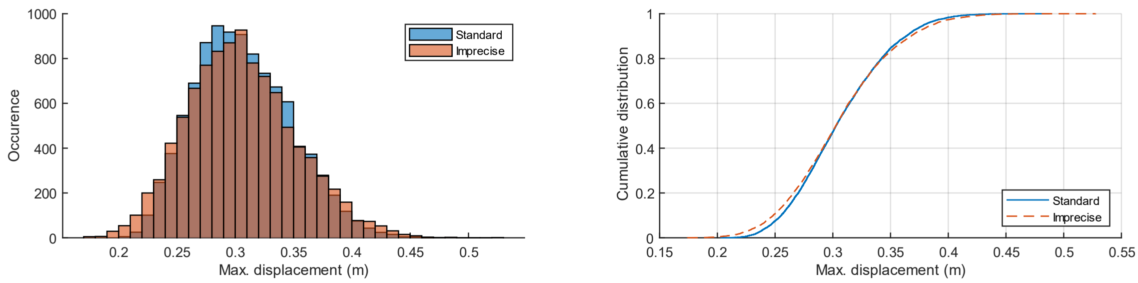
Figure 5: Histogram and CDF of the MC simulation for the standard approach and the imprecise PSD for example 1.



4.2. Example 2

In the second example, the following system parameters are used: $m = 50$ kg, $c = 3$ Ns/m and $k = 12337$ N/m. The natural frequency is thus $\omega = \sqrt{k/m} \approx 5\pi$ rad/s. The second example has been changed by a modified natural frequency compared to the first example. However, the comparison between the ensemble mean and the imprecise PSD function yields similar results, see histogram and CDF in Fig. 6. In the histogram, the MC samples of the imprecise PSD cover a slightly higher range. Both smaller and larger system displacements can be obtained. This minimally stretches the histogram, resulting in a lower occurrence of the mean system displacement of about 0.3 m. In both models, however, 0.3 m is the highest occurrence of the displacements. Therefore, the CDF also provides only minimally different results for the two models. The failure probabilities change only slightly as a result. Particularly in the case of smaller system displacements, small differences can be seen; in the case of larger system displacements, the failure probabilities are almost identical. It can thus be stated that a limited data set can provide robust and reliable results utilising the imprecise PSD function.

Figure 6: Histogram and CDF of the MC simulation for the standard approach and the imprecise PSD for example 2.



5. CONCLUSION

Accounting for uncertainties in data sets to obtain reliable simulation results is of paramount importance in engineering. Especially when only limited data are available, uncertainties can have a large impact on the results and can easily distort them. This might can lead to disastrous consequences, e.g., when an actually catastrophic result is shifted into an acceptable range due to incorrect consideration of uncertainties or limited data. In such a case, it is important to correctly interpret the data and generate appropriate load models that account for those uncertainties. From a large amount of data, it is often possible to derive a robust model that provides reliable simulation results. However, in order to estimate robust models from limited data, an imprecise model of a PSD function is proposed in this paper, which provides a reliable upper and lower bound of the data set. Moreover, by using an RBF network, the physics of the underlying stochastic process is reflected and dependencies between frequencies are taken into account. In this work the samples are drawn from a uniform distribution, future work will involve the employment of appropriate interval distribution schemes as crude MC sampling is not a viable propagation strategy and is only used here to illustrate an approximated solution. The resulting imprecise PSD function was validated by numerical examples and its strengths were demonstrated.

References

[1] A. Powell, S. Crandall, *Random Vibration*, The Technology Press of the Massachusetts Institute of Technology, Cambridge, 1958.

[2] Y.-K. Lin, G.-Q. Cai, *Probabilistic Structural Dynamics: Advanced Theory and Applications*, McGraw-Hill New York, 1995.

[3] L. D. Lutes, S. Sarkani, *Random Vibrations: Analysis of Structural and Mechanical Systems*, Butterworth-Heinemann, 2004.

[4] M. Beer, S. Ferson, V. Kreinovich, *Imprecise probabilities in engineering analyses*, Mechanical Systems and Signal Processing 37 (2013) 4 – 29, http://dx.doi.org/10.1016/j.ymssp.2013.01.024.

[5] L. Comerford, I. A. Kougioumtzoglou, M. Beer, *On quantifying the uncertainty of stochastic process power spectrum estimates subject to missing data*, International Journal of Sustainable Materials and Structural Systems 2 (2015) 185–206, http://dx.doi.org/10.1504/IJSMSS.2015.078358.

[6] Y. Zhang, L. Comerford, I. A. Kougioumtzoglou, E. Patelli, M. Beer, *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 (2017) 04017020, http://dx.doi.org/10.1061/AJRUA6.0000925.

[7] G. Muscolino, F. Genovese, A. Sofi, *Reliability bounds for structural systems subjected to a set of recorded accelerograms leading to imprecise seismic power spectrum*, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 8 (2022) 04022009, http://dx.doi.org/10.1061/AJRUA6.0001215.

[8] M. Behrendt, M. Bittner, L. Comerford, M. Beer, J. Chen, *Relaxed power spectrum estimation from multiple data records utilising subjective probabilities*, Mechanical Systems and Signal Processing 165 (2022) 108346, http://dx.doi.org/10.1016/j.ymssp.2021.108346.

[9] D. S. Broomhead, D. Lowe, *Radial basis functions, multi-variable functional interpolation and adaptive networks*, Technical Report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[10] M. Priestley, *Spectral Analysis and Time Series*, *Probability and mathematical statistics: A series of monographs and textbooks*, Academic Press, 1982.

[11] M. Shinozuka, G. Deodatis, *Simulation of stochastic processes by spectral representation*, Applied Mechanics Reviews 44 (1991) 191–204, http://dx.doi.org/10.1115/1.3119501.

[12] J. Li, J. Chen, *Stochastic Dynamics of Structures*, John Wiley & Sons, 2009.

[13] D. E. Newland, *An Introduction to Random Vibrations, Spectral & Wavelet Analysis*, Courier Corporation, 2012.

[14] M. D. Buhmann, *Radial Basis Functions: Theory and Implementations*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2003, http://dx.doi.org/10.1017/CBO9780511543241.

[15] J. Ghosh, A. Nag, *An Overview of Radial Basis Function Networks*, Physica-Verlag HD, Heidelberg, 2001, pp. 1–36, http://dx.doi.org/10.1007/978-3-7908-1826-0_1.

[16] S. Chen, C. F. N. Cowan and P. M. Grant, *Orthogonal least squares learning algorithm for radial basis function networks*, IEEE Transactions on Neural Networks, vol. 2, no. 2, pp. 302-309, March 1991, http://dx.doi.org/10.1109/72.80341.

[17] K. Kanai, *Semi-empirical formula for the seismic characteristics of the ground*, Bulletin of the Earthquake Research Institute 35 (1957) 309–325.

[18] H. Tajimi, *A statistical method of determining the maximum response of a building structure during an earthquake*, in: Proceedings of the 2nd world conference of earthquake engineering, Volume 11, 1960, pp. 781–797.

[19] P. L. Narasimha, W. H. Delashmit, M. T. Manry, J. Li, F. Maldonado, *An integrated growing-pruning method for feedforward network training*, Neurocomputing 71 (2008) 2831–2847, https://dx.doi.org/10.1016/j.neucom.2007.08.026.