Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)
Edited by Maria Chiara Leva, Edoardo Patelli, Luca Podofillini, and Simon Wilson
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-5183-4_S14-05-243-cd

Assessing the severity of missing data problems with the interval discrete Fourier transform algorithm

Marco Behrendt
Institute for Risk and Reliability, Leibniz Universität Hannover, Germany.
E-mail: behrendt@irz.uni-hannover.de

Marco de Angelis
Institute for Risk and Uncertainty, University of Liverpool, United Kingdom.
E-mail: marco.de-angelis@liverpool.ac.uk

Liam Comerford
Institute for Risk and Uncertainty, University of Liverpool, United Kingdom.
E-mail: l.comerford@liverpool.ac.uk

Michael Beer
Institute for Risk and Reliability, Leibniz Universität Hannover, Germany.
Institute for Risk and Uncertainty, University of Liverpool, United Kingdom.
International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, Shanghai, China.
E-mail: beer@irz.uni-hannover.de

The interval discrete Fourier transform (DFT) algorithm can propagate in polynomial time signals carrying interval uncertainty. By computing the exact theoretical bounds on signal with missing data, the algorithm can be used to assess the worst-case scenario in terms of maximum or minimum power, and to provide insights into the amplitude spectrum bands of the transformed signal. The uncertainty width of the spectrum bands can also be interpreted as an indicator of the quality of the reconstructed signal. This strategy must however, assume upper and lower values for the missing data present in the signal. While this may seem arbitrary, there are a number of existing techniques that can be used to obtain reliable bounds in the time domain, for example Kriging regressor or interval predictor models. Alternative heuristic strategies based on variable (as opposed to fixed) bounds can also be explored, thanks to the flexibility and efficiency of the interval DFT algorithm. This is illustrated by means of numerical examples and sensitivity analyses.

Keywords: Missing data, Exact bounds, Interval discrete Fourier transform, Power spectral density estimation, Interval uncertainty, Uncertainty quantification.

1. Introduction

The consideration and quantification of uncertainties in real data are of paramount importance for the design and simulation of buildings and structures. Even small measurement errors can lead to a wrong consideration of the input data and result in a disastrous interpretation of the simulation results, e.g. if an actually catastrophic result is shifted into an acceptable range by not taking uncertainties into account. Uncertainties should therefore be considered in any case and included in the simulation, also in order to determine possible safety margins. In order to safely design or to assess the reliability and robustness of buildings and structures that are subject to environmental processes such as wind, earthquakes or waves and thus exhibit dynamic behaviour, simulations are indispensable. The discrete Fourier transform (DFT) is an important tool in this field to determine the frequency components and their amplitude of the environmental processes. Consideration of uncertainties in the data, such as missing data, should be combined with the DFT to obtain reliable results.

Some approaches for estimating power spectral density (PSD) functions from signals with missing data have already been proposed. In particular, approaches treating missing data as normal distributed random variables and propagating them through the DFT (Comerford et al., 2015; Zhang et al., 2017). Another approach was presented by Liu and Kreinovich (2010), where the fast Fourier transform (FFT) and convolution were studied for signals with interval and fuzzy uncertainty. An algorithm to propagate interval signals through the DFT to obtain exact bounds on the Fourier amplitude was proposed by the authors of this work in De Angelis et al. (2021), while the algorithm is described in details and applied to an example involving dynamic structural analysis in Behrendt et al. (2022). The algorithm enables the quantification of uncertainties in time signals and to project them into the frequency domain. No assumptions are made about the dependence and distribution of the error over the time steps. Thus, a bounded Fourier amplitude and a PSD function can be computed, which can be used to analyse system responses in the frequency domain, taking into account these uncertainties. The objective of this work is to investigate the capabilities of the interval DFT in missing data problems. It also aims to determine the severity of the missing data and the possible impact on the interval DFT algorithm. The quantity used to measure uncertainty in this work is the area between the upper and lower bounds. Contrarily, a Fourier amplitude without uncertainty will have such an area equal to zero.

This work is organised as follows: Some theoretical background that is relevant for this work is provided in Section 2. The problem of missing data is elaborated in Section 4. In Section 3 the interval DFT algorithm is described briefly. The capabilities of said algorithm in combination with missing data problems are explored in Section 5, while the final conclusions are given in Section 6.

2. Preliminaries

This section introduces some fundamental theoretical concepts that will be required in this work.

2.1. *Power spectrum estimation*

Given a signal x_n, represented as a zero mean stochastic process. To examine the signal for its frequency components, it can be transformed into the frequency domain using the periodogram. The periodogram is the squared absolute value of the Fourier transform and reads as follows

$$\hat{S}_X(\omega_k) = \frac{\Delta t^2}{T} \left| \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{2\pi i k n}{N}} \right|^2, \quad (1)$$

where Δt is the time step size, T is the total length of the record, n describes the data point index in the record, N is the total number of data points in the signal and k is the frequency number of $\omega_k = \frac{2\pi k}{T}$.

2.2. *Generation of artificial time signals*

To generate an artificial time signal for simulation purposes, the Spectral Representation Method (SRM) can be utilised (Shinozuka and Deodatis, 1991). The SRM generates a time signal X_t based on an underlying PSD function S_X and carries their characteristics. The SRM is

$$X_t = \sum_{m=0}^{M-1} \sqrt{4 S_X(\omega_m) \Delta\omega} \cos(\omega_m t + \varphi_m), \quad (2)$$

with $\omega_m = m\Delta\omega$, $m = 0, 1, 2, \ldots, M-1$, where M is the total number of frequency points, t as time vector and φ_m as uniformly distributed random phase angles in the range $[0, 2\pi]$.

As the underlying PSD function, a spectrum derived within the Joint North Sea Wave Observation Project (JONSWAP) (Hasselmann et al., 1973) will be used throughout this work. The JONSWAP spectrum is an extension of the Pierson-Moskowitz PSD function (Pierson Jr. and Moskowitz, 1964) and is utilised to describe the dynamic behaviour of sea waves in the frequency domain.

3. The interval DFT algorithm

To transform signals into the frequency domain, the DFT can be employed. The DFT converts a signal $x = x_0, x_1, ..., x_{N-1}$ to a Fourier sequence $z = z_0, z_1, ..., z_{N-1}$ for $k = 0, ..., N-1$. Since many signals are subject to missing data, these

must be taken into account during the transformation in order to obtain reliable results. One possibility is to reconstruct the data before the transformation. However, since the DFT is very sensitive to changes in the signal, as shown in Section 4, it is more reasonable to fill the missing data gaps with intervals and propagate them through the DFT. However, since the DFT is not able to transform such uncertainties, an algorithm was proposed that is capable to propagate interval uncertainties through the DFT and thus calculate exact bounds on the Fourier amplitude. This *interval DFT algorithm* is briefly described here, for a detailed explanation and examples the reader is referred to Behrendt et al. (2022) and De Angelis et al. (2021).

Based on the *interval extension* of the DFT, obtained by replacing the real signal with their interval values for each frequency number k

$$\overline{\underline{z}}_k = \sum_{n=0}^{N-1} \overline{\underline{x}}_n e^{-ia_k} = \sum_{n=0}^{N-1} \overline{\underline{x}}_n \cdot [\cos a_k - i \cdot \sin a_k], \quad (3)$$

with $a_k = 2\pi kn/N$, the algorithm computes two vertices for each iteration n, resulting from the interval values of the n-th data point of the signal. In each iteration step, the vertices are added to the previous vertices. The convex hull is calculated from these. The points of the convex hull are passed on to the next iteration step, while the remaining vertices have no influence on the calculation and are discarded. Once all points of the signal have been iterated, the minimum and maximum distance of the convex hull to the origin of the coordinate system is determined, which defines the interval bounds of the absolute value of the transform

$$\overline{\underline{A}}_k = |\overline{\underline{z}}_k| = \sqrt{\left[\sum_{n=0}^{N-1} \overline{\underline{x}}_n \cos a_k\right]^2 + \left[\sum_{n=0}^{N-1} \overline{\underline{x}}_n \sin a_k\right]^2}. \quad (4)$$

The absolute values of the points in the convex hull are calculated for this purpose. If the origin of the coordinate system is within the convex hull, the lower bound is 0, otherwise it is defined by the minimum absolute value. The upper bound is always determined by the maximum absolute value. Thus, an upper and lower bound of the Fourier amplitude can be computed for each frequency number k.

4. Missing data

A common problem when using real data records is that of missing data. The causes of missing data range from simple measurement errors to total sensor failure. It is possible that the sensor is damaged by the event it is supposed to record, e.g. an earthquake, and makes incorrect recordings or stops recording completely. In addition, the sensors may be temporarily unavailable due to maintenance. If the period of unavailability is sufficiently short, intervals can be used to bridge this gap. These causes introduce uncertainty into the data series. Although there are various reconstruction methods, e.g. least squares method, compressive sensing or autoregressive methods, the method for reconstructing the signal in case of missing data will not be considered further. Here, the focus is on the performance of the proposed algorithm rather than the reconstruction method. The reconstructed data are represented by intervals, accounting for uncertainties induced through the reconstruction. Thus, the reconstructed signal is passed to the interval DFT algorithm as an interval signal. Fig. 1 shows the signal under investigation in this work with two examples each with missing data. In this work, the same signal is used throughout to ensure maximum comparability of the different cases. If a signal in time domain is certainly known, it can be transformed to the frequency domain via the DFT without loss of information. In fact, the DFT is sensitive to small changes in the signal. To demonstrate the sensitivity, the signal in Fig. 1, consisting of 64 data points, is investigated. The target spectrum, i.e. the Fourier amplitude of the signal without missing data computed with Eq. 1, is depicted with the Fourier amplitudes of the same signal with 5%, 10% and 25% missing data, which are reconstructed by linear interpolation between the two adjacent non-missing data points, see Fig. 2.

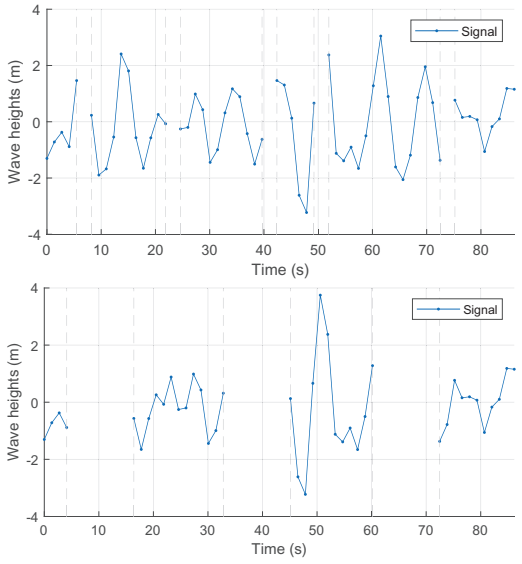

Fig. 1. The signal under investigation with two examples each with missing data.

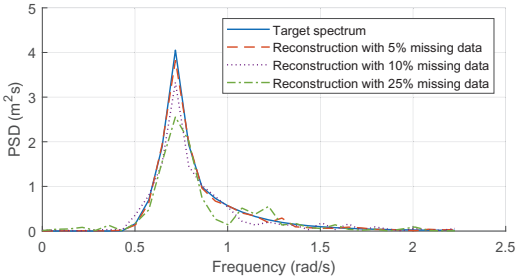

Fig. 2. Influence of the linear interpolation on the amplitude of the DFT.

The position of the missing data is randomly chosen. The interpolated values are treated as discrete values instead of intervals first. Although linear interpolation is not considered as a reconstruction method in this work, it can be used to illustrate the sensitivity. It can be clearly seen that the transformations have the same shape and peak frequency, but are in part very different from the target spectrum and are not as smooth. Since reconstructed data accordingly do not allow a reliable transformation into the frequency domain and do not take uncertainties into account, it is reasonable to derive bounds in which the actual spectrum may be located. The algorithm presented in 3 is applicable for this purpose.

Two reconstruction methods are utilised in this paper:

(1) A method based on artificial inflation of the "true" data point using the sample standard deviation s of the entire signal. An interval of height $[-s, s]$ replaces the missing data.
(2) A method that replaces the missing data by an interval determined by the minimum and maximum value of the entire signal.

The sample standard deviation of the signal is defined as:

$$s = \sqrt{\frac{\sum_{n=0}^{N-1}(x_n - \tilde{x})^2}{N-1}},$$

where $\tilde{x}$ is the sample mean of the signal.

5. Case studies

In this section, the influence of missing data on the bounds of the estimated PSD is investigated. Specifically, interval width, the number of missing data, the gap length, and the distribution of missing data within the signal are examined. The signal under investigation is generated by SRM (Eq. 2) with the underlying PSD function given in (Hasselmann et al., 1973). The positions of the missing data in the signal are artificially generated in random order. It is assumed that the missing data is uniformly distributed within the signal. A study is also conducted to investigate the influence of the position of the missing data, comparing the uniformly distributed missing data with binomially distributed missing data. In order to obtain the best possible comparison, the same signal is used in all studies of this work.

5.1. *Sensitivity to interval uncertainty*

Let ξ be the height of the interval gap, a.k.a. interval uncertainty. To investigate the sensitivity of the interval uncertainty ξ in time domain to the interval uncertainty in the frequency domain, the signal is randomly equipped with missing data gaps of length $l_g \in \{1, 3, 5, 7, 9, 11\}$, where the gap length is given as the number of missing time points.

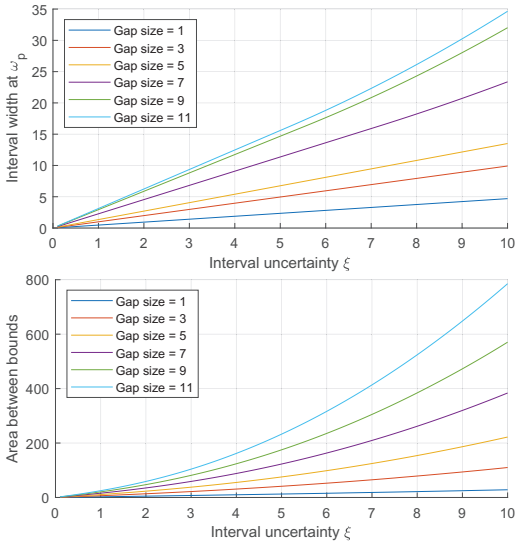

Fig. 3. Interval width of the bounded PSD function at the peak frequency ω_p (top) and the area between upper and lower bound (bottom) for increasing interval uncertainty ξ and different lengths of the gap $l_g \in \{1, 3, 5, 7, 9, 11\}$.

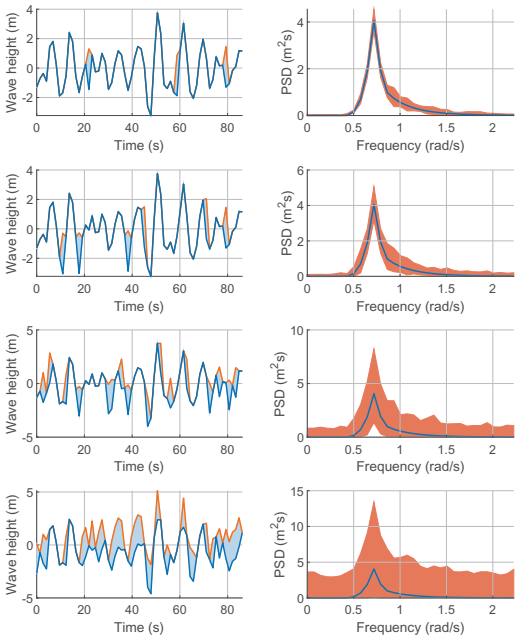

Fig. 4. Signal with 5%, 10%, 25% and 50% missing data reconstructed using method (1) and corresponding bounded PSD functions.

The interval uncertainty ξ of these gaps is successively increased from 0.1 to 10. To determine the sensitivity, the interval width of the Fourier amplitude at the peak frequency ω_p, as well as the area between the upper and lower bound are determined. The results are depicted in Fig. 3. For smaller gaps with low interval uncertainty, a linear trend in the interval width at ω_p appears at the beginning, which later changes to a non-linear trend. This occurs as soon as the lower bound of the Fourier amplitude reaches 0 and only the upper bound contributes to the interval width. The increase is nevertheless moderate and not extremely strong. The area between the bounds, on the other hand, has a non-linear trend even with low interval uncertainty and small gaps. This non-linearity becomes stronger the larger the gap becomes. This is due to the fact that the entire frequency range is considered instead of only the peak frequency ω_p. At many frequency points, the lower bound has already reached 0, while it still higher at the peak frequency. For larger gaps, the lower bound is mostly zero, which explains why in Fig. 3 the start of the non-linear behaviour is appreciated for lower interval uncertainty.

5.2. *Number of missing data*

In the following example, the interval uncertainty has been kept constant and corresponds to the sample standard deviation s of the signal. The number of missing data points, on the other hand, has been gradually increased to investigate the influence of the number of missing data on the bounds of the PSD. In Fig. 4, the reconstructed signals and the the bounds of the estimated PSDs are shown for 5%, 10%, 25% and 50% missing data in the signal. The results show that a small amount of missing data (e.g. 5% or 10%) can be captured well with the interval DFT algorithm. The bounds enclose the estimated PSD function of the discrete signal relatively tightly and are therefore very useful for quantifying the uncertainties. Also, the bounds of the PSD for a higher amount of missing data in the signal (up to 50% in this example) can still be considered, despite the relatively wide bounds, e.g. for a worst-case consideration where only the upper bound is used.

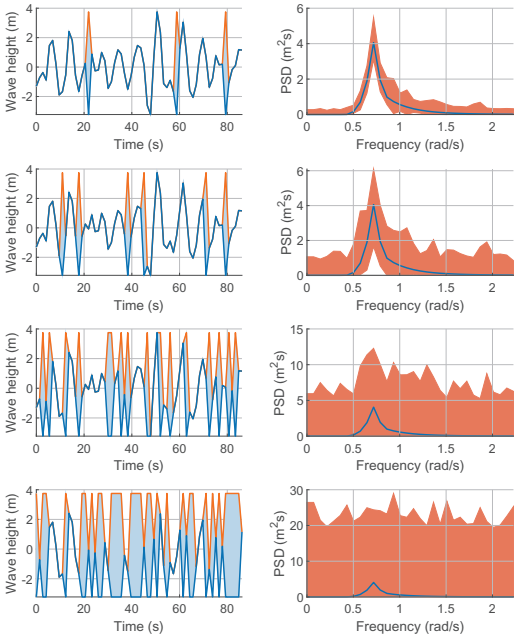

Fig. 5. Signal with 5%, 10%, 25% and 50% missing data reconstructed using method (2) and corresponding bounded PSD functions.

In the following, the same example is shown, but the data was reconstructed using method (2), see Fig. 5 for the reconstructed signals and the bounds of the PSDs in frequency domain.

The results also show here that small amounts of missing data can be mapped well in the frequency domain even with reconstruction method (2). With higher numbers of gaps, however, the determination of the bounds in the frequency domain reaches its limitation, as the computed bounds are very high and can no longer be used for practical purposes. For example, the bounds from the previous example with 50% missing data have a lower interval uncertainty than the signal with 25% missing data in this example. This yields in particular that if there is little missing data, reconstruction can be carried out conservatively with wide intervals. Conversely, if the number of missing data is large, a method with a more accurate reconstruction is required. As a measure for uncertainty, the area between upper and lower bound is utilised. Fig. 6 shows this for an increasing number of missing data recovered with the two

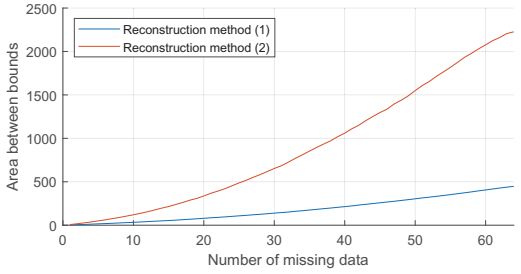

Fig. 6. Area between upper and lower bound investigated for the number of missing data.

reconstruction methods. Due to possible random fluctuations, as the position of missing data is randomly chosen, this simulation was carried out 100 times in order to average out these fluctuations. As expected, there is a significantly higher area between the bounds when using reconstruction method (2) compared to reconstruction method (1).

5.3. *Gap size of missing data*

Recall that gap size is given as the number of missing time points, and it is also referred to as gap length. To determine the influence of the gap length, different scenarios were evaluated. The gap lengths $l_g \in \{1, 20, 40, 60\}$ were artificially inserted into the signal. The gaps were first reconstructed with method (1). The signals and the corresponding transformations are shown in Fig. 7. It can be seen that small gaps filled with the intervals provide a good transformation and the bounds are relatively tight around the exact spectrum. The interval DFT algorithm can also handle larger gaps well, although the bounds of the transformation are comparatively large. Nevertheless, these can be used, for example, to design for a worst-case when only the upper bound with the largest energy content is used for planning and simulation. For these investigations, all examples with reconstruction method (2) are omitted, since it has been shown that large gaps already lead to extremely large bounds with reconstruction method (1) and are practically no longer useful. Since the length of the gap naturally corresponds to the number of missing data, no significant differences between Fig. 8 and Fig. 6 in the previous section can be

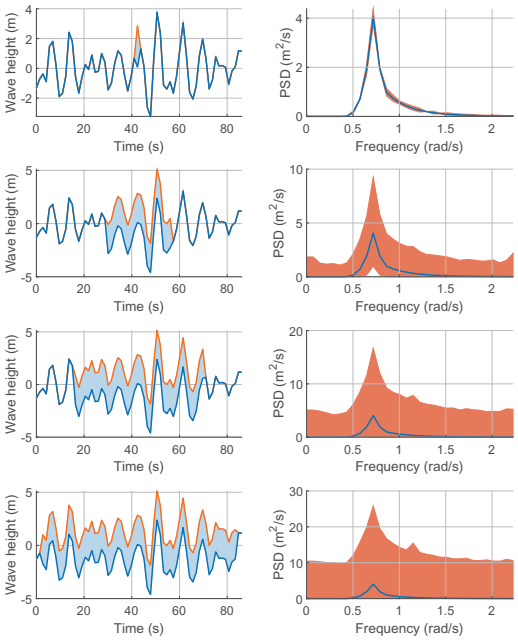

Fig. 7. Signals with a gap of length $l_g \in \{1, 20, 40, 60\}$ of missing data reconstructed by method (1) and corresponding bounded PSD functions.

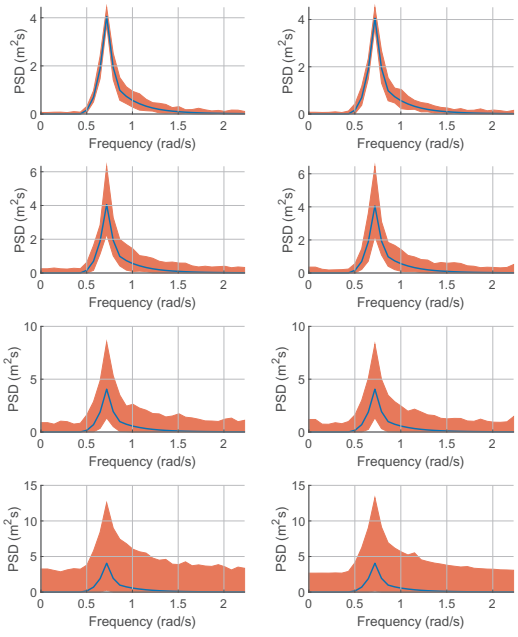

Fig. 9. Influence of the distribution of missing data within the signal for 4, 8, 16 and 32 missing data reconstructed with method (1).

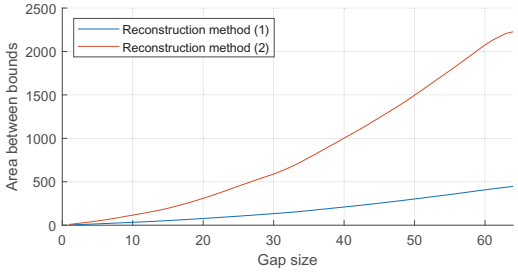

Fig. 8. Area between upper and lower bound investigated for the length of the gap.

detected. This indicates that the position of the missing data has a minor role in determining the uncertainty, but the number has a major role. Nevertheless, for completeness the influence of the distribution of the missing data will be investigated in the next section.

5.4. *Distribution of missing data*

This example is used to investigate the influence of the distribution of the missing data within the signal. For the sake of brevity, only the results for reconstruction method (1) are shown. In addition, it has been shown in the previous sections reconstruction method (2) cannot be used for real phenomena if the number of missing data is sufficiently high, as the bounds are extremely large. For the investigations a uniform distribution and a binomial distribution were utilised to randomly generate the missing data and to investigate their influence on the transformation to the frequency domain. The interval PSDs of the reconstructed signal with 4, 8, 16 and 32 missing data are depicted in Fig. 9. It can be seen that the influence of the position of the missing data is of minor relevance. Although the transformed signals shown are only specific cases, they are nevertheless representative of the general case. This statement can be supported by the fact that this simulation has been carried out several times, but the results are always identical. The transforms look almost identical in each case, regardless of the distribution of the missing data. In addition, the interval widths at the respective frequencies, such as the peak frequency, are very similar and the area

between the bounds are also almost identical.

6. Conclusions

In this work, the interval DFT algorithm has been investigated for its ability to transform signals with missing data reconstructed by intervals. Different scenarios have been considered, such as the influence of the interval width, the number of missing data, the length of the gap of missing data and the distribution of the missing data in the signal. It was shown, that the largest influence was exerted by the interval uncertainty in the signal and the number of missing data, while the distribution of the data and their position is of minor importance. In addition, no indications could be found of an influence whether the data are missing at individual points or appear as a large gap. It was found that too large intervals often lead to extremely wide bounds, which are usually no longer usable for practical purposes. If the number of missing data is sufficiently small, however, a good transformation can be computed even with a conservative estimation of the intervals, in which the bounds are close to the actual spectrum. With a larger number of missing data or larger gaps, it is also possible to plan for the worst-case by considering only the upper bound, provided that the interval width is reasonably chosen. It has also been shown that the potential energy content of the PSD can change significantly depending on the choice of intervals. In summary, the interval DFT algorithm provides excellent results for uncertain data. However, it should be noted that the results are highly dependent on the reconstruction of the data. Thus, it is highly recommended that in the case of missing data, the interval DFT algorithm should be employed with an advanced reconstruction method in order to obtain reliable results.

Replicability

The software for computing the interval DFT can be accessed in a single instance via GitHub at: `https://github.com/interval-fourier-transform/application-to-missing-data`. The code, examples and numerical results presented in this paper are therefore fully replicable.

Acknowledgement

This research is funded by the Engineering & Physical Sciences Research Council (EPSRC) with grant no. EP/R006768/1. The EPSRC is greatly acknowledged for their funding and support.

References

Behrendt, M., M. de Angelis, L. Comerford, Y. Zhang, and M. Beer (2022). Projecting interval uncertainty through the discrete Fourier transform: An application to time signals with poor precision. *Mechanical Systems and Signal Processing 172*, 108920.

Comerford, L., I. A. Kougioumtzoglou, and M. Beer (2015). On quantifying the uncertainty of stochastic process power spectrum estimates subject to missing data. *International Journal of Sustainable Materials and Structural Systems 2*(1-2), 185–206.

De Angelis, M., M. Behrendt, L. Comerford, Y. Zhang, and M. Beer (2021). Forward interval propagation through the discrete Fourier transform. In *The 9th international workshop on Reliable Engineering Computing*, pp. 39–52.

Hasselmann, K. F., T. P. Barnett, E. Bouws, H. Carlson, D. E. Cartwright, K. Eake, J. Euring, A. Gicnapp, D. Hasselmann, P. Kruseman, et al. (1973). Measurements of wind-wave growth and swell decay during the Joint North Sea Wave Project (JONSWAP). *Ergaenzungsheft zur Deutschen Hydrographischen Zeitschrift, Reihe A*.

Liu, G. and V. Kreinovich (2010). Fast convolution and Fast Fourier Transform under interval and fuzzy uncertainty. *Journal of Computer and System Sciences 76*(1), 63–76. Special Issue on Intelligent Data Analysis.

Pierson Jr., W. J. and L. Moskowitz (1964). A proposed spectral form for fully developed wind seas based on the similarity theory of S. A. Kitaigorodskii. *Journal of Geophysical Research (1896-1977) 69*(24), 5181–5190.

Shinozuka, M. and G. Deodatis (1991). Simulation of stochastic processes by spectral representation. *Applied Mechanics Reviews 44*(4), 191–204.

Zhang, Y., L. Comerford, I. A. Kougioumtzoglou, E. Patelli, and M. Beer (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), 04017020.