Nonlinearity in data with gaps: Application to ecological and meteorological datasets

SANDIP V. GEORGE and G. AMBIKA

Indian Institute of Science Education and Research, Pune 411 008, India
*Corresponding author. E-mail: g.ambika@iiserpune.ac.in

Abstract. Datagaps are ubiquitous in real-world observational data. Quantifying nonlinearity in data having gaps can be challenging. Reported research points out that interpolation can affect nonlinear quantifiers adversely, artificially introducing signatures of nonlinearity where none exist. In this paper we attempt to quantify the effect that datagaps have on the multifractal spectrum \((f(\alpha))\) in the absence of interpolation. We identify tolerable gap ranges, where the measures defining the \(f(\alpha)\) curve do not show considerable deviation from the evenly sampled case. We apply this to the multifractal spectra of multiple datasets with missing data from the SMEAR database. The datasets we consider include ecological datasets from SMEAR I, namely CO\(_2\) exchange variation, photosynthetically active radiation levels and soil moisture variation time series, and meteorological datasets from SMEAR II, namely dew point variation and air temperature variation. We could establish multifractality due to deterministic nonlinearity in two of these datasets, where the gaps were within tolerable limits.

Keywords. Datagaps; multifractal spectrum; SMEAR; photosynthesis; meteorology.

PACS Nos 05.45.–a; 05.45.Tp; 87.23.–n; 92.60.–e

1. Introduction

The presence of gaps in observational data is one of the major problems that affects real-world datasets. Such gaps or the consequent uneven sampling in time series can be a major impediment in detecting nonlinearity in them. The most common way to deal with datagaps has been to interpolate through them to make an evenly sampled dataset. Past studies have shown that interpolation can have adverse effects on nonlinear time series quantifiers [1]. Some recent studies on the effect that datagaps have on nonlinear quantifiers when the analyzed time series is not interpolated through, showed that there are certain values of gap size and frequency which affect correlation dimension calculations significantly [2]. We propose to extend these studies to the multifractal spectrum. Such an analysis helps us identify how susceptible the multifractal spectrum is to gaps in data.

We note that astrophysical data, ecological data, meteorological data, etc., are often plagued by gaps [3, 4]. They may occur due to multiple reasons like instrument failure, cloud cover, adverse weather conditions. Most nonlinear quantifiers require large evenly sampled datasets in order to make predictions [5, 6]. One can exploit the presence of multiple, smaller, evenly sampled segments in order to calculate quantifiers like bicoherence [7]. However no analysis has been done so far on the effect of datagaps on the multifractal spectra. Such an analysis is important in the context of quantifying the multifractal nature of large numbers of datasets that come with missing data.

The multifractal spectrum, \(f(\alpha)\), is a detailed characterization of the fractal structure of the system in a reconstructed phase space. Unlike an average measure like the correlation dimension, \(D_2\), the \(f(\alpha)\) spectrum takes into consideration the local contributions of different regions of the attractor. The idea of \(f(\alpha)\) spectrum can be applied beyond attractors in phase space, to describe many natural objects [8]. A complete characterization of the multifractal curve was shown to be achieved with a set of four parameters. In this study we aim to check how much variation these parameters show when the time series used for phase space reconstruction through embedding, contains missing data [9].

To determine the susceptibility or resilience of the \(f(\alpha)\) spectrum to datagaps, we first start with an evenly

CrossMark
sampled time series of a standard system. We subject the evenly sampled time series to multifractal analysis using the algorithm described in Ref. [9]. We then remove data from the evenly sampled time series, progressively. The size of the data removed and frequency of the data removal are drawn from two Gaussian distributions. The time series with datagaps are also subjected to multifractal analysis. The difference in the multifractal quantifiers, like the width of the spectrum, in the two analyses is treated as an index of susceptibility of that quantifier to datagaps. We can hence identify tolerable ranges of parameters that give reliable conclusions about the multifractal nature of time series.

As an illustration, we proceed to analyze real-world datasets with missing data. The datasets are obtained from the Station for Measuring Forest Ecosystem-Atmosphere Relation (SMEAR) in Finland [10]. We consider ecological and meteorological datasets obtained from SMEAR I and SMEAR II stations respectively. The ecological datasets are related to photosynthesis, which acts as a proxy for photosynthesis. We also obtain the ecological datasets from SMEAR II station in the Värriö forest. We look into the multifractal spectrum of CO$_2$ exchange in pine shoots, which acts as a proxy for photosynthesis. We also obtain the $f(a)$ curves for the time series of levels of photosynthetically active radiation and soil moisture, both of which are known to influence photosynthesis rates [11, 12]. For the meteorological datasets we consider the time series of air temperature and dew point, obtained from the SMEAR II station in the Hyytiala forest. Both the time series give a continuous $f(a)$ curve, indicative of multifractality in its phase space structure.

2. Analysis of synthetic datasets

In this section we consider the effect of datagaps on the structure of the $f(a)$ curves for standard nonlinear systems like Rössler and Lorenz systems. We start with large evenly sampled datasets of the $x$ variable of these systems and introduce gaps into the data. Since the gaps typically arise from multiple independent sources, we take that both the position and size of the gaps follow Gaussian distributions,

$$G_s(s; m_s, \omega_s) = \frac{1}{\omega_s \sqrt{2\pi}} e^{-(s-m_s)^2 / 2\omega_s^2};$$

$$G_p(p; m_p, \omega_p) = \frac{1}{\omega_p \sqrt{2\pi}} e^{-(p-m_p)^2 / 2\omega_p^2}.\tag{2}$$

$G_s$ with mean $m_s$ and standard deviation $\omega_s$ determines the size of the gap and $G_p$ with mean $m_p$ and standard deviation $\omega_p$ determines the position of the gap.

Starting from any point in the time series, $G_p$ determines the position of the next gap, i.e., it determines the extent of gapless data. Hence it is an inverse measure of frequency of gaps. $G_s$ determines how big the gap is at a particular position.

2.1 Effect of datagaps on $f(a)$ spectrum

A detailed characterization of the non uniformities of the attractor is captured by the generalized dimensions, $D_q$ [8, 9]. In an analogous formulation, if one covers the attractor with boxes of size $R$, the probability $p_i$ for points to fall inside the $i^{th}$ box is found to scale as

$$p_i(R) = R^{a_i(R)}.$$ \tag{3}

Then the number of such boxes that have $a$ between $a$ and $a + \Delta a$ is [13]

$$n(a, R) \propto R^{-\gamma}.$$ \tag{4}

$D_q$ and $f(a)$ are related through Legendre transformations. This provides a method to compute the $f(a)$ spectrum numerically from the $D_q$ values. We can characterize the $f(a)$ curve using the following function fit

$$f(a) = A(a - a_1)^{\gamma_1}(a_2 - a)^{\gamma_2},\tag{5}$$

and $a_1$, $a_2$, $\gamma_1$ and $\gamma_2$ serve as unique parameters that can characterize the $f(a)$ completely [9]. The $f(a)$ spectrum finds widespread use across a variety of different fields [14–17].

For the datasets of the Rössler system, with gaps introduced, the delay time $\tau$ is first calculated as the point where the auto-correlation falls to $\frac{1}{e}$. The embedding dimension $M$ is chosen as the integer value above which the correlation dimension $D_2$ saturates. For the ranges of $m_s$ and $m_p$ under consideration, $M$ remains close to 3, and hence $M$ for our analysis is chosen to be 3. In our study both $m_s$ and $m_p$ are varied in units of $\tau$. We consider the variation of $\Delta a = a_2 - a_1$, $\gamma_1$ and $\gamma_2$ with $m_s$ and $m_p$. As a sample case, we show the calculated $f(a)$ spectrum for the evenly sampled time series and two cases with datagaps in figure 1, for the Rössler system. The full variation of $\Delta a$ with $m_s$ and $m_p$ for the Rössler system is shown in figure 2, where we plot the relative variation, $\delta a = a_2 - a_1$. Here $\Delta a$ is the value of $\Delta a$ for the evenly sampled case. For a fixed value of $m_p$, as we increase $m_s$, the value of $\Delta a$ deviates heavily from the evenly sampled value and reaches a peak at around 1 $\tau$. On increasing $m_s$ further, it falls again and saturates at a value close to the evenly sampled value. We find that in the critical region identified, the $f(a)$ spectrum widens considerably for the Rössler system, corresponding to the peak value of $\Delta a$. Since the shape of the curve
changes considerably when gaps are present, we find that \( \gamma \) values fluctuate heavily. The variation of \( \delta\gamma^{rel} = \frac{\gamma_{skip} - \gamma_{even}}{\gamma_{even}} \) with \( m_s \) and \( m_p \) is shown in figure 2b and c. Here again, \( \gamma_{even} \) is the value of \( \gamma \) for the evenly sampled case. The large fluctuations in the figures make it clear that \( \gamma \) values derived from a time series contaminated with gaps cannot be trusted, in general.

The analysis was conducted for the Lorenz system as well, where the variation is much smaller in comparison. We, however, conclude that while the width of the \( f(\alpha) \) spectrum may remain broadly unchanged, except at small \( m_p \) and for \( m_s \approx 1\tau \), the overall shape of the curve does change, as signified by the change in the \( \gamma \) values. We point out that widening of the \( f(\alpha) \) curve can also happen due to noise contamination [9].

2.2 Surrogate analysis

A wide \( f(\alpha) \) spectrum is indicative of multifractality in the system. However this is not conclusive evidence of deterministic nonlinearity. Linear stochastic processes were shown to have saturating \( D_2 \) and wide multifractal curves in the literature [9, 18]. A surrogate analysis is required to confirm the existence of multifractality due to deterministic nonlinearity in the system. For this we initially produce 5 Iterative Amplitude Adjusted Fourier Transform (IAAFT) surrogates of the evenly sampled data implemented using the TISEAN package [19, 20]. The \( f(\alpha) \) curves obtained from the Lorenz system, Rössler system, red noise, white noise and their surrogates are shown in figure 3. The \( \alpha_1 \) and \( \alpha_2 \) values in all the cases studied are tabulated in table 1. It is clear that for the deterministic systems, the data and surrogates have clearly differing \( f(\alpha) \) curves, while for noise, they are not distinguishable.

We repeat the analysis using data with gaps. The same profile of gaps is used to remove data from the surrogates. We find that in the presence of gaps, \( \alpha \) values from the original data and surrogates become increasingly difficult to differentiate. In figure 4 we show function fits for two cases of the Rössler system from a region where the gaps are tolerable and where gaps cause high
Figure 3. \( f(\alpha) \) vs \( \alpha \) curves for (a) Lorenz system (b) Rössler system and (c) red and white noises and their surrogates. (a) and (b) show distinct difference between the original data and surrogates, while they are identical for noises.

Table 1. \( \alpha_1, \alpha_2 \) values for Lorenz, Rössler, red noise, white noise and their surrogates. \( \alpha_1^d \) and \( \alpha_2^d \) represent the values for data while \( \alpha_1^s \) and \( \alpha_2^s \) represent the corresponding values for surrogates.

<table>
<thead>
<tr>
<th>Data</th>
<th>( \alpha_1^d )</th>
<th>( \alpha_2^d )</th>
<th>( \alpha_1^s )</th>
<th>( \alpha_2^s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lorenz</td>
<td>1.54 ± 0.04</td>
<td>2.84 ± 0.04</td>
<td>3.58 ± 0.04</td>
<td></td>
</tr>
<tr>
<td>Rössler</td>
<td>1.60 ± 0.03</td>
<td>2.99 ± 0.19</td>
<td>3.44 ± 0.19</td>
<td></td>
</tr>
<tr>
<td>Red noise</td>
<td>2.43 ± 0.04</td>
<td>3.68 ± 0.07</td>
<td>3.81 ± 0.07</td>
<td></td>
</tr>
<tr>
<td>White noise</td>
<td>2.93 ± 0.03</td>
<td>3.09 ± 0.03</td>
<td>3.13 ± 0.03</td>
<td></td>
</tr>
</tbody>
</table>

\( \alpha_2 \) for data and the corresponding values averaged over five surrogate time series. Hence we conclude that in the regions with gap distribution we identify to be non tolerable, multifractality alone cannot confirm deterministic nonlinearity.

Figure 4. Function fits for \( f(\alpha) \) for the Rössler system and surrogates for (a) \( m_s = 8 \tau, m_p = 7 \tau \) and (b) \( m_s = 1 \tau, m_p = 7 \tau \). In (a) the data and surrogates can be differentiated, but they merge with each other in (b). (Some curves do not touch the x-axis as the function (eq. 5) varies too fast for small \( \gamma \) values.)

variation. We see that data merges with surrogates in the region where gaps are not tolerable. This is also evident from table 2 which lists the values for \( \alpha_1 \) and \( \alpha_2 \).
Table 2. $\alpha_1$, $\alpha_2$ values for the Rössler system and its surrogates for two different values of $m_s$ and $m_p$. We notice that in both cases, $\alpha_1^s$ and $\alpha_2^s$ remain different from each other, whereas $\alpha_1^p$ and $\alpha_2^p$ remain identical.

<table>
<thead>
<tr>
<th>$m_s$</th>
<th>$m_p$</th>
<th>$\alpha_1^s$</th>
<th>$\alpha_2^s$</th>
<th>$\alpha_1^p$</th>
<th>$\alpha_2^p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8$\tau$</td>
<td>7$\tau$</td>
<td>1.73</td>
<td>2.02 ± 0.03</td>
<td>2.98</td>
<td>3.06 ± 0.18</td>
</tr>
<tr>
<td>1$\tau$</td>
<td>7$\tau$</td>
<td>1.81</td>
<td>1.95 ± 0.07</td>
<td>4.04</td>
<td>3.82 ± 0.43</td>
</tr>
</tbody>
</table>

3. Analysis of datasets with gaps

The analysis conducted in the sections above are with the aim of studying variations in $f(\alpha)$ in the presence of datagaps and identifying whether the regions with gaps are tolerable and the validity of the conclusions drawn from the spectrum will hold. We now apply the analysis to real-world datasets. We use datasets with missing data, from the Station for Measuring Forest Ecosystem-Atmosphere Relation (SMEAR), Finland. We look into the time series of air temperature and dew point data measured by the SMEAR II station in the Hyrylä forest, and datasets of variation of CO$_2$ exchange, photosynthetically active radiation and soil moisture from the SMEAR I station situated in the Värriö forest.

In all the datasets, we calculate the correlation time $\tau$ and estimate $m_s$ and $m_p$ in terms of $\tau$. In order to find the dimension used to embed the attractor in, we calculate saturated $D_2$ for the dataset and pick the smallest integer greater than $D_2$ as the embedding dimension. The calculation of $D_2$ filters out all non saturating datasets (i.e. noise processes). Hence the aim of calculating the $f(\alpha)$ spectrum is not just to detect nonlinearity, but to characterize the multifractal properties of the attractor.

3.1 Datasets from SMEAR I

The SMEAR I station in Värriö, Finland measures a number of meteorological and ecological parameters. The station is situated about 200 km north of the Arctic circle. The datasets produced by the instruments are however not evenly sampled and contain datagaps [4]. Some of the primary reasons for gaps in data in this context include thunderstorms and breaks in electricity [21].

In this work we consider the time series of CO$_2$ exchange in pine trees, which acts as a proxy for photosynthesis rate [22] and two factors that affect the photosynthesis rate, namely variation of photosynthetically active radiation on top of the measurement chamber and variation of soil moisture. The rate of photosynthesis has already been suspected of chaotic variation by multiple authors [23, 24]. However, to the best of our knowledge no study has been conducted to determine the multifractal characteristics of its embedded phase space structure. Photosynthesis is well known to be dependent on ambient temperature, concentration of atmospheric CO$_2$, amount of photosynthetically active radiation (PAR), soil moisture (SM), etc. [11, 12]. In this context the study of the nonlinear properties of these variables becomes important as well. We also study the multifractal properties of the embedded attractors of the time series of variation of photosynthetically active radiation and soil moisture.

All three time series are averaged over half an hour for a period of 12 years starting from 2005 to 2017 and the $m_1$, $m_2$, $\tau$ and $D_2$ values are shown in table 3.

We find that barring the time series for soil moisture, most of the $m_1$ and $m_2$ are outside the identified critical region. Hence in the $f(\alpha)$ computed, $\alpha_1$ and $\alpha_2$ are likely to be largely unaffected. In the case of soil moisture it is possible that the width of the curve in the absence of datagaps may be much smaller than the one calculated. The plots of the $f(\alpha)$ spectra are shown in figure 5. We then proceed to do surrogate analysis to check for signs of deterministic nonlinearity. Since the time series has missing data we cannot directly apply the method of IAAFT mentioned above. Instead we take evenly sampled segments of the data to construct surrogates. We first take segments of data that are continuously sampled. Surrogates are constructed for each

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
Datagap & $\tau$ & $m_1$ & $m_2$ & $D_2$ & \\
\hline
CO$_2$ & 0.07 & 0.18 & 3.18 & 1.15 & \\
PAR & 0.07 & 0.18 & 3.18 & 1.15 & \\
SM & 0.07 & 0.18 & 3.18 & 1.15 & \\
\hline
\end{tabular}
\end{table}

Figure 5. $f(\alpha)$ vs $\alpha$ for time series of CO$_2$ exchange in a pine shoot, photosynthetically active radiation (PAR) and soil moisture (SM) from the SMEAR I station. All three show multifractality.
segment, and joined together to act as a surrogate for the whole data. We caution that this may not always be possible and is dependent upon the availability of gap-less chunks in the dataset. As before, we find the \( \alpha_i \) and \( \alpha_s \) values for data and surrogates, presented in table 4. We find that for CO\(_2\) exchange and photosynthetically active radiation, values of \( \alpha_i \) and \( \alpha_s \) are distinctly different from their surrogates, suggesting that multifractality due to underlying deterministic nonlinearity can be concluded. This adds weight to the claims in the literature that photosynthesis may be varying chaotically [23, 24]. No such conclusion can be drawn for the variation of soil moisture, as the data and surrogates lie very close to each other. This could be a result of the values of \( m_i \) and \( m_p \) lying very close to the identified critical range or due to a lack of deterministic nonlinearity or noise contamination.

### 3.2 Datasets from SMEAR II

The SMEAR II station is in the Hyytiälä forest in Finland and measures atmospheric aerosols, eco-physiology, soil and water data, solar and terrestrial radiation and meteorological data. In this work we primarily consider two meteorological datasets, the dew-point and air temperature. These were chosen partially due to the high instance of datagaps in them [4]. Climate time series have been often subject to studies of multifractality in the past [14, 16]. We consider the data sampled every half hour over the period 2008–2017. As before we first determine \( m_i \), \( m_p \), \( \tau \) and \( D_2 \) (shown in table 5). We notice that \( m_i \) for the air temperature is in the critical region identified and hence the \( f(\alpha) \) curve may be significantly affected. The \( f(\alpha) \) curves for both are plotted in figure 6.

With surrogates constructed as before, we perform a surrogate analysis for datasets from SMEAR II. The results are presented in table 6. For the dew point variation data, both \( \alpha_i \) and \( \alpha_s \) lie close to the value for the surrogate datasets. Since the \( m_i \) and \( m_p \) values are away from the identified critical region, it appears that dew point variation has no underlying deterministic nonlinearity and is probably a linear stochastic process. The air temperature time series has \( \alpha_i \) and \( \alpha_s \) values close to the surrogates, but not within error bars. This could be the result of the values of \( m_i \) and \( m_p \) lying in the identified critical region.

### 4. Conclusion

We investigate the effect of gaps in data on the \( f(\alpha) \) spectrum of the reconstructed attractor. We use data from two standard systems, the Rössler and Lorenz systems, to check the effect of datagaps on the computed multifractal spectrum. From the analysis on standard chaotic data we could arrive at tolerable limits on gap parameters, within which conclusions from
nonlinear time series analysis can be meaningful. Since these limits are linked to the embedding technique itself, we assume that the same can be valid for real-world datasets analyzed using the same procedure of embedding. In this work, we illustrate this by analyzing 5 ecological and meteorological datasets with gaps of varying ranges. We support our conclusions by carrying out detailed surrogate analysis on all of them.

We specifically study the variations in the four quantifiers of $f(\alpha)$, viz. $\alpha_1$, $\alpha_2$, $\gamma_1$ and $\gamma_2$, in all the datasets. To the best of our knowledge, our work for the first time establishes multifractality due to deterministic nonlinearity in photosynthesis variation and factors affecting it.

The absence of large evenly sampled datasets is a major impediment for calculating measures applying nonlinear time series analysis techniques and especially deriving quantifiers from $f(\alpha)$. The pitfalls of interpolation make it an undesirable technique in these cases. Hence, it becomes important to consider how much deviation from the evenly sampled case actually occurs if we ignore gaps. We find that certain broad conclusions about the data can still be retained even in the presence of missing data. We suggest the following procedure while dealing with time series data contaminated with gaps. The means of the gap distributions, $m_s$ and $m_p$, should initially be quantified. If they are not within tolerable limits, binning can be used as a method to change the values of $m_s$ and $m_p$ and bring them into the tolerable range identified [2]. Once $m_s$ and $m_p$ are within tolerable ranges, reliable conclusions can be drawn about $\alpha_1$ and $\alpha_2$. For any conclusions about underlying deterministic nonlinearity, one needs to conduct an analysis for comparison with phase randomized surrogates. If the surrogates give very different values for quantifiers, as compared to those from the data, one can conclude the existence of an underlying nonlinearity. If surrogates and data merge within tolerable range, the underlying process can be assumed to be purely stochastic or deterministic with high noise contamination. In cases where, even with considerable binning, the $m_s$ and $m_p$ values cannot be brought within tolerable ranges, no conclusions can be drawn based on multifractal analysis.

Acknowledgements

We acknowledge with thanks the data from the SMEAR database that are used in this research. SVG thanks Ron Sunny, IISER, Pune for help with the SMEAR datasets.

References

[12] D Shimsh, Plant physiology 38(6), 713 (1963)
[22] G W Schneider and N F Childers, Plant physiol. 16(3), 565 (1941)