Analysis of scale invariance in nonuniformly sampled processes

Stéphane G. ROUX[1] Janka LENGYEL[1,3] Ptashanna THIRAUX[1] Stéphane JAFFARD[2] Olivier BONIN[3] Patrice ABRY[1]

[1]ENS de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France

[2]Univ Paris Est Creteil, Univ Gustave Eiffel, CNRS, LAMA, F-94010 Créteil, France

[3] Univ Gustave Eiffel, Ecole des Ponts, LVMT, F-77454 Marne-la-Vallée, France

Résumé – Nous proposons une méthode d'estimation d'exposants d'invariance d'échelle pour des processus portés par des supports lacunaires ou non-homogènes. Son originalité est de permettre d'extraire des informations relatives à un large éventail d'échelles spatiales et temporelles sans qu'il soit nécessaire d'extrapoler le processus en dehors de son support. À partir de données synthétiques (mouvement brownien fractionnaire restreint à un support obtenu comme distributions de Poisson), nous comparons l'approche proposée à l'analyse d'invariance d'échelle classique, appliqués aux mêmes données après extrapolation des donnés manquantes sur une grille régulière et homogène.

Abstract – We propose a strategy for estimating the parameters of scale invariance for processes defined on lacunary or nonhomogeneous supports. The originality of this strategy is that it allows us to extract information over a wide range of spatial and temporal scales without having to extrapolate the process outside its support. Using synthetic data (generated from fractional Brownian motion restricted to supports from Poisson distributions), we evaluate the proposed method by comparing it to the classical scale invariance approach applied to the same data after filling in missing samples (through various interpolation methods) to create a regular homogeneous grid.

1 Motivation

Context. Multifractal analysis [1, 2, 3] - or the general study of scale invariance - has already found numerous applications in a variety of fields, including finance, ecology, geography, and geology [4, 5]. Multifractal analysis is mainly concerned with measuring scaling exponents, which can then be further used for, e.g., feature detection and classification tasks. An intrinsic limitation is that most classical mathematical formulations of scale invariance analysis rely on the assumption that data are defined either everywhere on a homogeneous grid or on an exactly self-similar fractal set. However, a significant number of applications actually entail data that are only defined on nonhomogeneous sets; this is notably the case in geography, where collected data come with precise, georeferenced locations and are defined only on restricted and irregular subsets (cf., e.g., [5]). This work aims to discuss issues related to practical scale invariance analysis for such data.

Related works. Most studies dealing with scale invariance and multifractal analysis on irregular grids base the discussion on missing samples. This problem is particularly relevant in neuroscience [6] or in the context of geospatial data [7, 8]. In most cases, missing samples are handled by *interpolation*, followed by the use of *classical* scale invariance and multifractal analysis. Beyond classical interpolation schemes (cf., e.g., [6, 7]), energy-consuming neural network-based interpolation strategies have also been studied (cf., e.g., [8]). However, in the case of geospatial and environmental datasets, non-existing values are due to the very nature of the phenomena and cannot be interpreted as a simple missing value problem. Thus, the present work provides preliminary contributions to alternative scale invariance and multifractal analysis specific to data sampled irregularly in nature.

Goals, contributions and outlines. The contributions and goals of the present work are twofold. First, in the context of nonhomogeneous 1D and 2D processes, we develop a method that can extract information across a wide range of spatiotemporal scales (notably for *fine* length scales) without requiring a prior interpolation of the data outside of their natural support (cf. Section 2.1). Second, the proposed method, which uses only the original data points, is compared empirically against classical multifractal analysis, in which missing entries are interpolated, using synthetic 1D and 2D synthetic data (defined in Section 3). Comparative results (cf. Section 4) can be used to quantify the advantages of the proposed approach in terms of the level of lacunarity of the data. Python codes, devised by the authors, used for this analysis are open-source and available at `https://pypi.org/project/lompy/`.

2 Multifractal analysis for processes on nonhomogeneous supports

2.1 Classical Multifracal analysis

Multifractal analysis. Multifractal analysis aims to characterize the pointwise regularity of a process, the Hölder exponent being the most widely used in practice. This information is then encapsulated through the *multifractal spectrum* $D(h)$, which yields the fractal dimensions of the set of points where the regularity exponent takes the value h. In practice, the estimation of the multifractal spectrum requires a set of procedures originally inspired by thermodynamic formalism (see, e.g., [3]). These are based on estimating the different moments of multiscale coefficients and observing their evolution across scales.

Multiscale (wavelet analysis). The wavelet coefficients $T(a, \underline{x})$ of a process V are defined as $T(a, \underline{x}) =$

$\int_{\underline{x}\in\mathbb{R}^d} V(\underline{y})\psi_a(\underline{y}-\underline{x})d\underline{y}$. The mother wavelet ψ needs to have at least one vanishing moment [2], $\int_{\underline{x}\in\mathbb{R}^d} \psi(\underline{x})d\underline{x} = 0$ [9].

Scale invariance. A process V is said to possess scale invariance or scaling properties if, for some statistical orders q, the time/space averages of $|T_V(a,\underline{x})|^q$ display power law behaviors with respect to scales a

$$S_q(a) = \mathbb{E}\{|T_V(a,\underline{x})|^q\} \sim F_q|a|^{\zeta_q}. \tag{1}$$

Multifractal formalisms. The scaling exponents ζ_q are related by a Legendre transform to the multifractal spectrum $D(h) \leq \min_q(qH - \zeta_q)$. A key parameter is ζ_2, which provides a summary of the properties of scaling invariance in terms of both the global correlation (or spectral density) and the Hurst exponent, $H = \zeta_2/2$.

Rationale behind the *Haar* and *poor* wavelets. Let us now focus on the examples of the so-called *Haar* and *poor* wavelets [9], defined as, with $\Pi_{x_0,a}(x)$ a rectangular function centered at X_0 and of size a:

$$\psi^1(x) = \Pi_{1/2,1}(x) - \Pi_{-1/2,1}(x), \tag{2}$$

$$\psi^2(x) = \delta(x) - \Pi_{0,1}(x). \tag{3}$$

The corresponding (L^1-norm) wavelet coefficients read:

$$T_V^1(a,x) = \frac{1}{a}\int_{x-a}^{x} V(x')dx' - \frac{1}{a}\int_{x}^{x+a} V(x')dx' \tag{4}$$

$$T_V^2(a,x) = V(x) - \frac{1}{a}\int_{x-a/2}^{x+a/2} (x')dx' \tag{5}$$

$T_V^1(a,x)$ can be interpreted as a derivative of the approximated signal. $T_V^2(a,x)$ can be read as the averages of all increments of size $a' < a$ originated at x.

In a discrete time/space frameworks, these wavelet coefficients are practically computed using $B_{x,a}^V$, the ball centered in x and of radius a, as:

$$T_V^1(a,x) = \frac{1}{a}\sum_{x_i\in B_{x-a/2,a}^V} V_i - \frac{1}{a}\sum_{x_i\in B_{x+a/2,a}^V} V_i, \tag{6}$$

$$T_V^2(a,x) = V(x) - \frac{1}{a}\sum_{x_i\in B_{x-a/2,x+a/2}^V} V_i. \tag{7}$$

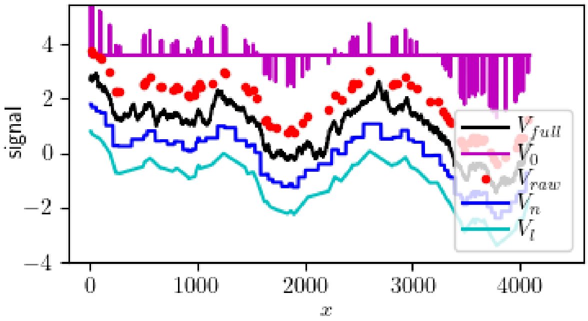

Figure 1 – **1D signal**. Example of a “full” signal known everywhere (V_{full}, in black), its restriction to a “nonhomogeneous” set of points (V_{raw}, red dots), and the different signals where the unknown values are replaced by the sample mean of the known signal (V_0, magenta), the nearest interpolation (V_n, blue) and the linear interpolation (V_l, cyan).

2.2 Processes with nonhomogeneous supports

One-dimensional case. Processes defined on nonhomogeneous supports can be modeled as a sum of Dirac distributions, $V(x) = \sum_{i\in\mathcal{V}} V_i\delta(x-x_i)$, where $S \subset \mathbb{R}$ denotes the support of s, i.e., the set of points where s is defined. The wavelet coefficients thus become $T_V(a,x) = \sum_{i\in\mathcal{V}} V_i\psi_a(x-x_i)$.

From the intuitions recalled above for the *Haar* and *poor* wavelets, the original proposition of this work is to define two tentative types of wavelet coefficients specifically suited to nonhomogeneous support, strongly relying on a new and original quantity. If $N_x^V(a) = \#\left\{B_x^V(a)\right\}$ is the number of samples in the support within a ball of radius a centered on x, then the multiresolution quantity is defined as:

$$T_V^1(a,x) = \frac{1}{N_{x+a/2}^V(a)}\sum_{x_i\in B_{x-a/2,a}^V} V_i - \frac{1}{N_{x-a/2}^V(a)}\sum_{x_i\in B_{x+a/2,a}^V} V_i, \tag{8}$$

$$T_V^2(a,x) = V(x) - \frac{1}{N_x^V(a)}\sum_{x_i\in B_{x-a/2,x+a/2}^V} V_i. \tag{9}$$

Obviously, $N_x^V(a)$ is directly proportional to the analysis scale a if the process is defined everywhere or is on a homogeneous support, and thus recovers straightforwardly, up to the multiplicative constant, the classical wavelet coefficients. The corresponding wavelets (equations 8 and **??**) are implicitly associated with the wavelet coefficients in the equations 6 and 7 and have a vanishing first-moment independent of the scale a and the position $x \in \mathrm{Supp}_s$.

Two-dimensional case. For 2D process analysis, instead of the Haar wavelet, it is classical to consider an extension of the 1D so-called *top-hat* wavelet, as proposed in[5]:

$$\psi^1(x) = \Pi_0(x,1) - \frac{1}{\sqrt{2}}\Pi_{0,\sqrt{2}}(x) \tag{10}$$

and 2D-wavelet coefficients read for this *top-hat* extension and for the *poor* wavelets:

$$T_V^1(a,x) = \frac{1}{N_x^V(a)}\sum_{x_i\in B_{x-a/2,a}^V} S_i - \frac{1}{N_x^V(\sqrt{2}a)}\sum_{x_i\in B_{x+a/2,a}^V} V_i, \tag{11}$$

$$T_V^2(a,x) = V(x) - \frac{1}{N_x^V(a)}\sum_{x_i\in B_{x-a/2,x+a/2}^V} V_i. \tag{12}$$

2.3 Scaling exponents

Multifractal analysis then consists of estimating moments $S_q^i(a) = \frac{1}{\#\{\mathrm{Supp}_s\}}\sum_{\underline{x}\in\mathrm{Supp}_s} |T_V^i(a,\underline{x})|^q$ of the obtained coefficients. Linear regressions in a log-log representation of $S_q^i(a)$ vs. $\log a$ permit to estimate scaling exponents $\hat{\zeta}^i{}_q$ (eq. 1). The scaling function of order $q = 2$ thus yields the estimated Hurst exponent: $\hat{H}^i = \hat{\zeta}^i{}_2/2$.

3 Data on nonhomogeneous supports

Processes are defined in analogous manners for the 1D and 2D cases. In what follows, *support* refers to the temporal/spatial distribution of available samples, while *mark* denotes

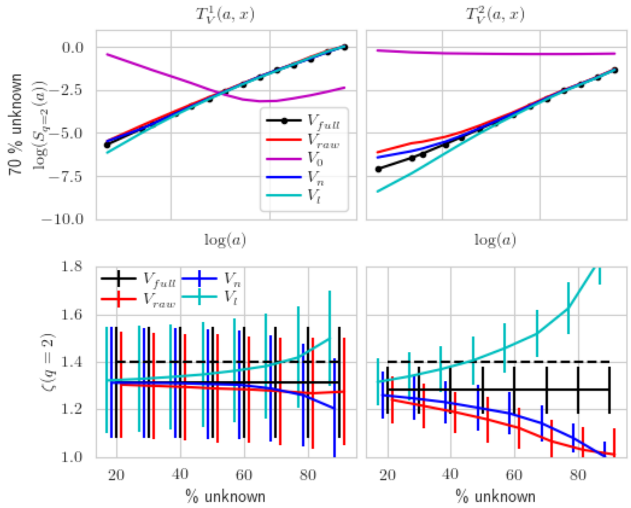

Figure 2 – **1D analysis**. Scaling of the moment of order two of $T_V^1(a,x)$ (left) and $T_V^2(a,x)$ (right) obtained on the full signal (V_{full}, in black), and the nonhomogeneous signal (V_{raw}, red), as well as on the interpolated signal by the sample mean (V_0, magenta), nearest (V_n, blue), and linear (V_l, cyan) approximations. The nearest approximation performs much better than the other interpolation methods but still shows a deviation from the original scaling. In contrast, our proposed method (V_{raw}, red) provides a suitable approximation.

the amplitude value of the process.

Nonhomogeneous support. There are many ways to define data on nonhomogeneous supports, as deviation from homogeneous data can be captured in at least two ways: The complexity of the form of the support itself and the degree to which the support is lacunar. In the present work, we are mainly concerned with the latter issue. We study a nonhomogeneous support of a simple form defined according to the Poisson distribution, where the parameter λ controls the degree of lacunarity. Scale invariance analysis for processes defined on nonhomogeneous supports (V_{raw}) is compared against scale invariance analysis for processes defined on homogeneous supports, i.e., regular-grid sampling. Such signals are referred to as V_{full} for full-interval time series in 1D or full-raster images in 2D.

Mark. The synthetic process used here is fractional Brownian motion (FBM), characterized by a unique regularity exponent H. This is a Gaussian-centered signal with variance one and stationary increments. This *full* signal V_{full}, of size N, containing only known values, is used as the reference for comparing the different analysis methods discussed here.

Interpolation. In the case of processes defined on a nonhomogeneous support, V_{raw}, the original scale invariance analysis proposed is compared against classical multifractal analysis applied to the same data, yet interpolated on a regular grid, thus on a homogeneous support. Three types of widely used interpolation schemes are compared:

- V_0: Each missing sample is replaced with V_{raw} sample mean.
- V_n: Each missing sample is replaced with the value of its nearest neighbor.
- V_l: Each missing sample is replaced with the sample of the field obtained by linear interpolation of available samples.

These interpolation schemes are implemented using the SciPy package. Examples of the signals are shown in Figures 1 (1D) and 3 (2D). Considering point processes in general, there are essentially two possible computational paths; the convolution or k-nearest neighbors methods. The first chooses a resolution and rasterizes the data into an image (which contains a predefined percentage of unknown values). Using the *convolution* method, for each scale a, one then computes the $T_V^1(a,x)$ and $T_V^2(a,x)$ coefficients according to (8) and (9), and for the 2D case, $T_V^1(a,x)$ and $T_V^2(a,x)$ with equations (11) and (12). Rasterizing the image is computationally challenging for highly nonhomogeneous samples, such as in geographic and environmental studies. In addition, rasterization can result in the loss of fine-scale information. Therefore, a second computational method based on the *k-nearest neighbors* algorithm, also implemented in the Python package `LomPy`, can be used, which allows the analysis of nonhomogeneous data (V_{raw}) without the need for rasterization.

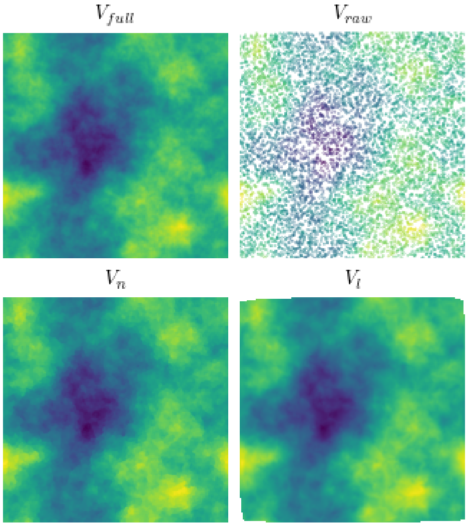

Figure 3 – **2D signal**. The homogeneous 2D fractional Brownian field (V_{full}) and the nonhomogeneous (V_{raw}) support as well as the nearest (V_n) and linear approximation (V_l) procedures performed to fill the missing values of the nonhomogeneous support (V_{raw}).

4 Performance assessment

One-dimensional case. This section performs the multifractal analysis by following the procedure described above. The results for the one-dimensional case are shown in Fig. 2 using $T_V^1(a,x)$ (left column) and $T_V^2(a,x)$ (right column). The logarithm of the second-order moments of the full signal (V_{full} in black) shows a linear behavior with respect to the logarithm of the scale over the whole range of the observed scale range. Scaling degraded for the three interpolated signals compared to the full signal at small scales. The sample mean (V_0 in magenta) drastically changes the slope toward -1. Interpolation with linear approximation also deteriorates scaling (V_l in cyan) where the slope goes to 2. The nearest approximation (V_n in blue) and the method introduced here (without interpolation, V_{raw}) give highly satisfactory results when $T_V^1(a,x)$ is used but deteriorate significantly at small scales when $T_V^2(a,x)$ is applied. This is confirmed by the bottom line of Fig. 2, where we show the slope obtained by linear regression using all scales.

We average the results for 20 different realizations and for six different densities of the support. As the sparseness of the data increases, biases in the estimation occur for all interpolation methods. The sample mean approximation yields such poor results that they fall beyond the reported value range. The results in the one-dimensional context suggest that interpolation of the signal is unnecessary since working with the original set of non-homogeneously distributed points alone (together with the coefficient $T_V^1(a,x)$ introduced here) yields similarly accurate results.

Two-dimensional case. In this section, we compare the performance of the two proposed 2D wavelets (eq. 11 and 12) on an FBM $\upsilon(\underline{x})$ process originally generated on a spatially homogeneous support (V_{full}). The latter full image, the resulting nonhomogeneous point process (V_{raw}), and the corresponding interpolated images (V_n and V_l) are shown in Fig. 3. Note that we do not consider the sample mean interpolation (V_0) in the two-dimensional case due to the previously obtained unsatisfactory results (see the performance assessment in 1D). The obtained scaling of the $T_V^1(a,x)$ (eq. 11) and $T_V^2(a,x)$ coefficients (eq. 12) are displayed in the first line of Figure 4 where the dashed black line indicates the expected behavior. In line with the 1D analysis in Fig. 2, the average slope (and standard deviation) obtained by linear regression from twenty independent 2D realizations are shown in the second line in Fig. 4, where we applied an increasing percentage of missing values. The border effect at large scales is significant with $T_V^1(a,x)$ and degrades the overall scaling. In contrast, results with $T_V^2(a,x)$ demonstrate linear behavior over a broader range of scales. Like the 1D case, interpolation with linear approximation deteriorates scaling, especially for small scales. The scaling obtained by the method without interpolation using $T_V^2(a,x)$ (in red) follows the scaling obtained for the full image (in black) very well. The same applies to the nearest approximation method using $T_V^1(a,x)$.

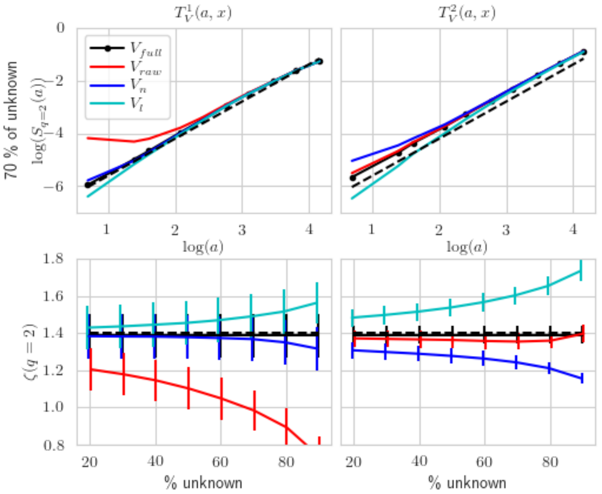

Figure 4 – **2D analysis**. Long-range dependence analysis using $T_V^1(a,x)$ (left column) and $T_V^2(a,x)$ (right column). The first line shows the scaling obtained on the images (V_{full} in black, nearest approximation V_n in blue, linear approximation V_l in cyan) and the no-interpolation method (V_{raw} in red). The dashed black line shows the theoretical behavior. The second line shows the mean and standard deviation of the slope estimates obtained from twenty independent realizations.

5 Conclusions

This article compared scale invariance analysis for processes on nonhomogeneous supports, with and without interpolating missing samples on a regular grid. For 1D signal, it was shown that interpolation is unnecessary since the analysis of the original set of (nonhomogeneously distributed) samples alone yield similarly accurate results. For 2D images, the outcome depends on the lacunarity of the support. For very lacunar support (when the density of missing values is large), best results are obtained without interpolation and using $T_V^2(a,x)$ (eq. 12). For raster-like supports (thus with few missing samples), nearest neighbor interpolation and classical analysis with $T_V^1(a,x)$ wavelet coefficients (eq. 11) is the most favorable procedure.

References

[1] E. Bacry, J. Muzy, and A. Arneodo, “Singularity spectrum of fractal signals from wavelet analysis: Exact results” J. Stat. Phys., vol. 70, pp. 635–674, 1993.

[2] S. Jaffard, “Wavelet techniques in multifractal analysis,” in Fractal Geometry and Applications: A Jubilee of Beno^ıt Mandelbrot, M. Lapidus and M. van Frankenhuijsen, Eds., Proc. Symposia in Pure Mathematics. 2004, vol. 72(2), pp. 91–152, AMS.

[3] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, “Wavelets for the analysis, estimation and synthesis of scaling data” in Self Similar Network Traffic Analysis and Performance Evaluation, K. Park and W. Willinger, Eds., Wiley, pp. 39–88, 2000.

[4] T., Hirabayashi, K., Ito and T., Yoshii, 1993. Multifractal analysis of earthquakes. Fractals and Chaos in the Earth Sciences, pp.591-610.

[5] J., Lengyel, S. G. Roux, P. Abry, F. Sémécurbe, and S. Jaffard. "Local multifractality in urban systems—the case study of housing prices in the greater Paris region." Journal of Physics: Complexity 3(4) (2022): 045005.

[6] Gao, X., Wang, X. (2018). Exploring the effects of missing data on the estimation of fractal and multifractal parameters based on bootstrap method. Nonlinear Processes in Geophysics Discussions, 1-28.

[7] V. A., Dergachev, A. N., Gorban, A. A., Rossiev, L. M., Karimova, E. B., Kuandykov, N. G., Makarenko, P., Steier (2001). The filling of gaps in geophysical time series by artificial neural networks. Radiocarbon, 43(2A), 365-371.

[8] A. N., Pavlov, O. N., Pavlova, A. S., Abdurashitov, O. A., Sindeeva, O. V., Semyachkina-Glushkovskaya, J., Kurths (2018). Characterizing scaling properties of complex signals with missed data segments using the multifractal analysis. Chaos: An Interdisciplinary Journal of Nonlinear Science, 28(1), 013124.

[9] S. Mallat, A wavelet tour of signal processing, 1999, Elsevier.