Mini-Project (ML for Time Series) - MVA 2023/2024

Marius Alonso marius.alonso@etu.minesparis.psl.eu
Paul Bonin paul.bonin@etu.minesparis.psl.eu

January 8, 2024

1 Introduction and contributions

The solar energy received at ground level, also called Surface Solar Irradiance (SSI), is a key weather parameters in climate simulations. Understanding intrinsic timescales of geophysical signals such as SSI is critical, notably when it comes to predicting them [1]. SSI being nonlinear and nonstationar, [2] uses the Hilbert Huang Transform to properly extract the different timescale components from the signal. In-depth analysis of the results allows the reader to characterize the underlying physical and stochastic processes, as well as the intrinsic non-linear entanglements between them.

This work reproduces the experiments from [2] as faithfully as possible, apart from the ones of Part 3.3 that were based on the 'ML for Time Series' class material. Experiments were performed on a sample of the original dataset [4], and also on a new dataset of discharge data for the Ance river in France.[1]

Please note that:

- Both students contributed equally to this work.
- The source code of the paper was not reused in this study. The Empirical Mode Decomposition developed in Part 2.1 was made using a convenient Python library [3].

2 Method

The Hilbert-Huang Transform is made of two steps : the empirical mode decomposition which outputs intrinsic mode functions (IMFs), and the Hilbert spectral analysis, performed on each of the IMFs.

2.1 Empirical Mode Decomposition (EMD)

An empirical mode decomposition of the time series is performed, as a means of retrieving the local periodicities of the signal. The decomposition progressively sifts the signal into intrinsic mode functions (IMFs), starting from higher frequencies and ending with lower ones.

The sifting is done by progressively constructing upper and lower envelope of the signal, and subtracting the mean of the two envelopes to the signal. The process is repeated until the signal obtained is eligible to be an IMF. Then, the sifting is restarted, with the original signal subtracted from all the already computed IMF components.

To improve the quality of the decomposition and reduce mode mixing[2], an ensemble sifting [6] with 1024 copies of the signal is performed. A random noise of variance 1 is added to each of the copies.

[1] Data was downloaded from: `https://portal.grdc.bafg.de/`, the website of The Global Runoff Data Centre. GRDC number for the Ance river: 6123760. More information about the GRDC: `https://www.bafg.de/GRDC/EN`.

[2] Mode mixing is the phenomenon where some of the IMFs share the same characteristic frequencies and overlap each-other.

The resulting IMFs of a sifting process for the river discharge dataset are displayed in appendix on Figure 10.

2.2 Hilbert Spectral Analysis (HSA)

To retrieve the characteristic timescales of each of the IMFs, we use the Hilbert Spectral Analysis. Each IMF and its Hilbert transform are used to construct a complex analytic signal, described by an amplitude-modulation-frequency-modulation (AM-FM) model. This transformation outputs both an instantaneous amplitude and an instantaneous frequency series, in the temporal space. This decomposition enables the identification of how much power occurs at which timescale. The results of this operation can be displayed in an Hilbert Spectrum (see appendix, Figure 13).

The Hilbert transform of each real-valued IMF $c_k(t)$ can be written as

$$\sigma_k(t) = \mathcal{H}\left(c_k(t)\right) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{c_k(\tau)}{t-\tau} \mathrm{d}\tau,$$

where subscript k designates the k th IMF and P indicates the Cauchy principal value. From each IMF and its Hilbert transformed version, a unique complex-valued analytic signal can be obtained:

$$z_k(t) = c_k(t) + i \cdot \sigma_k(t) = a_k(t) \cdot e^{i \cdot \theta_k(t)},$$

in which $a_k(t) = \sqrt{c_k^2(t) + \sigma_k^2(t)}$ is the instantaneous amplitude and $\theta_k(t) = \tan^{-1}\left(\frac{\sigma_k(t)}{c_k(t)}\right)$ is the instantaneous phase. The instantaneous frequency $\omega_k(t) = \frac{1}{2\pi}\frac{\mathrm{d}\theta_k(t)}{\mathrm{d}t}$ is the first time derivative of the instantaneous phase.

The instantaneous frequencies can be summed over temporal space, to retrieve the Hilbert marginal spectrum of the signal (Figure 1). The summation may be weighted by the instantaneous amplitude (Figure 2).

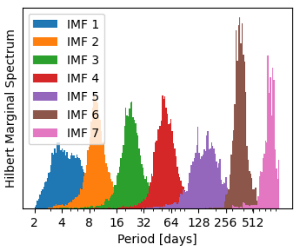

Figure 1: Hilbert Marginal Spectrum of the IMFs (river data)

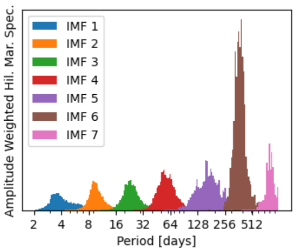

Figure 2: Amplitude-weighted Hilbert Marginal Spectrum of the IMFs

2.3 Distinguishing periodical physical processes from stochastic background processes

In order to be able to know if a given IMF is the result of stochastic noise or of a real, physical process, time series are resampled. When the sampling step is increased (i.e., the sampling frequency is reduced), noise components cannot preserve their original locations in the spectral domain and are shifted towards lower frequencies. Hence, significance testing of IMFs is done by verifying if the IMF remains unchanged in the time-frequency representation of the signal during resampling of the signal.

The retrieval of the amplitude weighted marginal spectrum $S_k(\omega)$ of the kth IMF allows the computation of the spectrum-weighted mean frequency (SWMF) $\bar{\omega}_k = \int S_k(\omega)\omega \mathrm{d}\omega / \int S_k(\omega)\mathrm{d}\omega$ of each IMF. This value is then used to characterize the potential drift.

3 Data

3.1 Source of datasets

The dataset (see Figure 4) of the reference paper is a set of daily mean irradiance, that are resampled versions of minute time series of the BSRN database, for various weather stations. The variable of interest is 'GHI' (Global Horizontal Irradiance).

Our test data is a daily discharge dataset for a river (see Figure 3). The variable of interest is the river discharge.

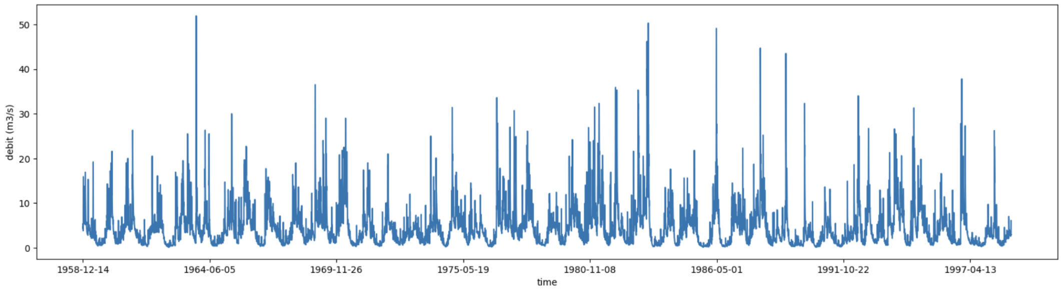

Figure 3: Daily discharge of the Ance river

3.2 Data processing

Irradiance data is cropped to span of years with almost complete availability[3], which gives time spans of approximately 10 - 15 consecutive years. Then, the daily mean is computed, ignoring missing values if less than 20% of the values of the day is missing. For days with more missing data, the daily SSI average is linearly interpolated from neighboring days.

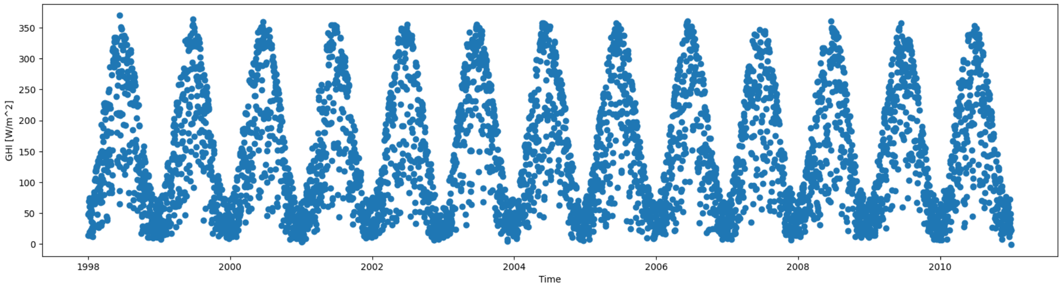

Figure 4: Daily mean irradiance for the PAY station of the BSRN dataset

For the discharge data, no missing data interpolation is needed.

3.3 Trend + Seasonality + Residual decomposition based on a yearly periodic model

The power spectrum has been computed for the two datasets. In both cases, we retrieve a clear yearly periodicity (see appendix, Figures 10 and 11).

Then, we isolate trend (if there is one), yearly periodicity T_{year} and residual signal. To do so, we perform least squares regression, with a dictionary of $p+1$ trend atoms $(1, n, \dots, n^p)$, and $2q$ periodic atoms $(\sin(nT_{\text{day}}/T_{\text{year}}), \cos(nT_{\text{day}}/T_{\text{year}}), \dots, \sin(qnT_{\text{day}}/T_{\text{year}}), \cos(qnT_{\text{day}}/T_{\text{year}}))$. The hyper-parameters p and q are determined using Bayesian Information Criterion. See result in Figure 5.

[3] Periods of at most 10 days missing.

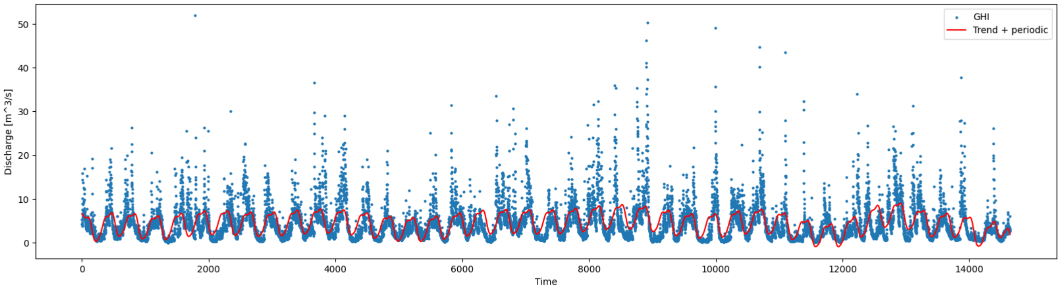

Figure 5: Discharge signal and extracted trend + yearly periodic component

4 Results

4.1 Results of the HHT

We perform HHT on the river discharge data, and obtain the results in Figures 12 and 13 (Appendix).

4.2 Spectrum-Weighted Mean Frequency (SWMF)

Figures 6 and 7 illustrate the spectral shifts of IMFs of the SSI and river dataset, respectively.

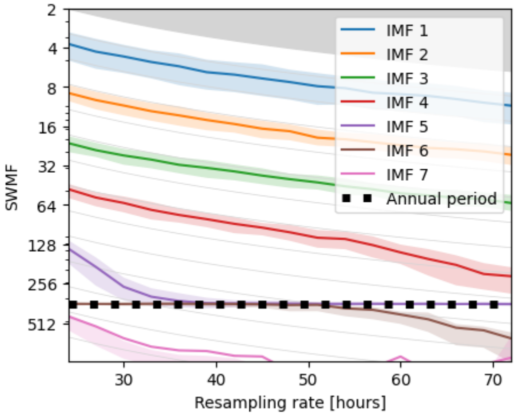

Figure 6: Observed SWMF for the 7 IMFs of BSRN-PAY.

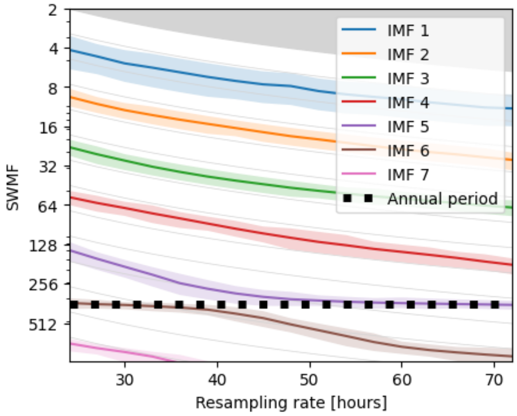

Figure 7: Observed SWMF for the 7 IMFs of the river dataset.

First, look at the IMFs that carry the annual component of the signal. Initially IMF 6 for both signals, it becomes IMF 5 after a 3-day resampling. The transition takes place in two distinct ways for the two signals: for BSRN-PAY, both IMF 5 and IMF 6 carry the temporal component for resampling rates ranging from 40 to 50 hours, whereas for river flow this superposition does not occur, and there is instead a short delay before IMF 5 can "catch up" with IMF 6. Indeed, the annual component has a greater weight in the decomposition of the BSRN-PAY series than in that of the river flow series.

Observe also the 4 first IMFs for both series. They behave very closely, and correspond in fact to the first IMFs obtained when performing the HHT of a pure noise. While their mean frequency shifts progressively in absolute terms, it stays constants relative to the resampling frequency (see the grey lines on the graph, indicating constant frequency relative to the sampling rate). These relative mean frequencies are close to the decreasing sequence 1/6, 1/12, 1/24, 1/48, and we find back the dyadic filtering property of the HHT on noisy signals [5].

Hence, for the original sampling rate of 24h, we can clearly attribute IMF 6 to the yearly periodic component of the signal due to the local absence of drifting of the SWMF. On the other hand, IMF 1, IMF 2, IMF 3 and IMF 4, behaving like results of dyadic filter, model noise-like components of the signal. It is harder to characterize IMF 5, in a transition from modelling noise to carrying the annual component.

4.3 Amplitude modulation of sub-yearly components

In part 3.3, we noted that there seems to be some correlation between the value of the periodic signal and the amplitude of the residual signal. To properly assess that, we perform rank correlation between IMF and AM signals. The statistic used here is Kendall's τ rank correlation. See results in Figures 8 and 9 (the color denotes the correlation, while the `p-value` for the hypothesis of independent normal distribution for the two variables is printed on the {column, line} intersection).

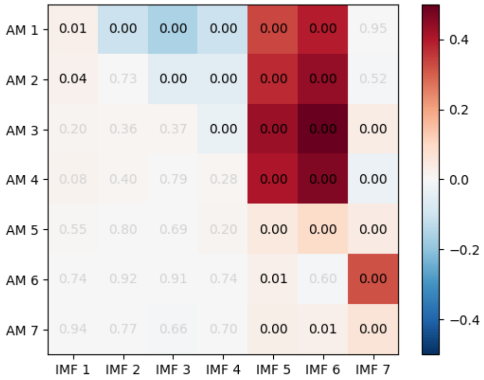

Figure 8: Amplitude modulation of the IMFs for the SSI dataset

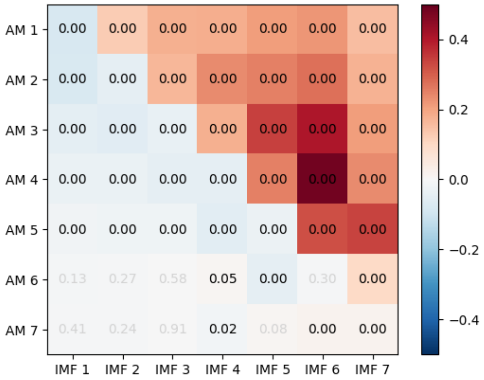

Figure 9: Amplitude modulation of the IMFs for the river dataset

In Figure 8 (SSI time series), we observe a clear positive rank correlation between the value of IMF 6 (the yearly physical signal), and the amplitudes of the sub-yearly components AM 1, AM 2, AM 3 and AM 4. Indeed, the amplitude of the daily stochastic variations in irradiance are higher during high solar irradiance months (May, Jun and Jul) than during the low solar irradiance months (Nov, Dec and Jan). Among the sub-yearly components, note also the negative rank correlation between x (IMF) and y (AM), where $1 \leq y < x \leq 4$. These amplitude modulations highlight the non linear physical intrications in the various timescales of the SSI signal.

In Figure 9 (river time series), we observe a positive correlation between an IMF of a certain order and the AM of the IMF of lower order (higher frequency). This is linked to the large peaks in the data which require several IMFs to be properly modelled. One could also perform the same methodology on the logarithm of the signal, to alleviate the weight of the decomposition of the peaks, and put more emphasis on the interpretation of the lower magnitude effects in the data.

These findings characterize in a way the two signals, their decomposition into intrinsic periodic components and noisy stochastic components, but also the (non-)linear entanglements between these components.

References

[1] Marc Bengulescu, Philippe Blanc, Alexandre Boilley, and Lucien Wald. Do modelled or satellite-based estimates of surface solar irradiance accurately describe its temporal variability? *Advances in Science and Research*, 14:35–48, 2017.

[2] Marc Bengulescu, Philippe Blanc, and Lucien Wald. On the intrinsic timescales of temporal variability in measurements of the surface solar radiation. *Nonlinear Processes in Geophysics*, 25(1):19–37, 2018.

[3] Andrew J. Quinn, Vitor Lopes-dos Santos, David Dupret, Anna C. Nobre, and Mark W. Woolrich. Emd: Empirical mode decomposition and hilbert-huang spectral analyses in python. *Journal of Open Source Software*, 6(59):2977, 2021.

[4] Laurent Vuilleumier and Alain Heimo. Basic and other measurements of radiation at station Payerne (1992-10 et seq), 2022.

[5] Zhaohua Wu and Norden E Huang. A study of the characteristics of white noise using the empirical mode decomposition method. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460(2046):1597–1611, 2004.

[6] Zhaohua Wu and Norden E Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in adaptive data analysis*, 1(01):1–41, 2009.

A Appendix

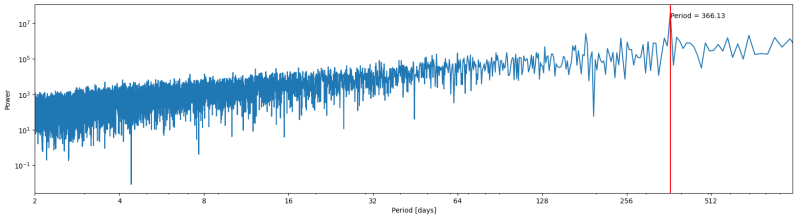

Figure 10: Power Spectrum of the river discharge signal

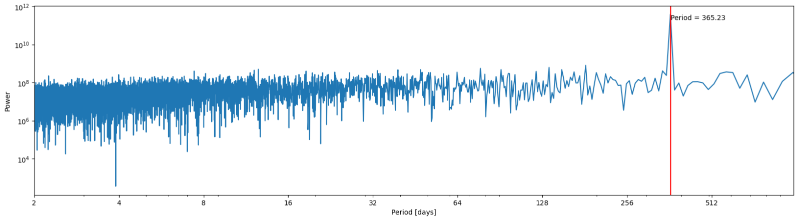

Figure 11: Power Spectrum of the SSI signal

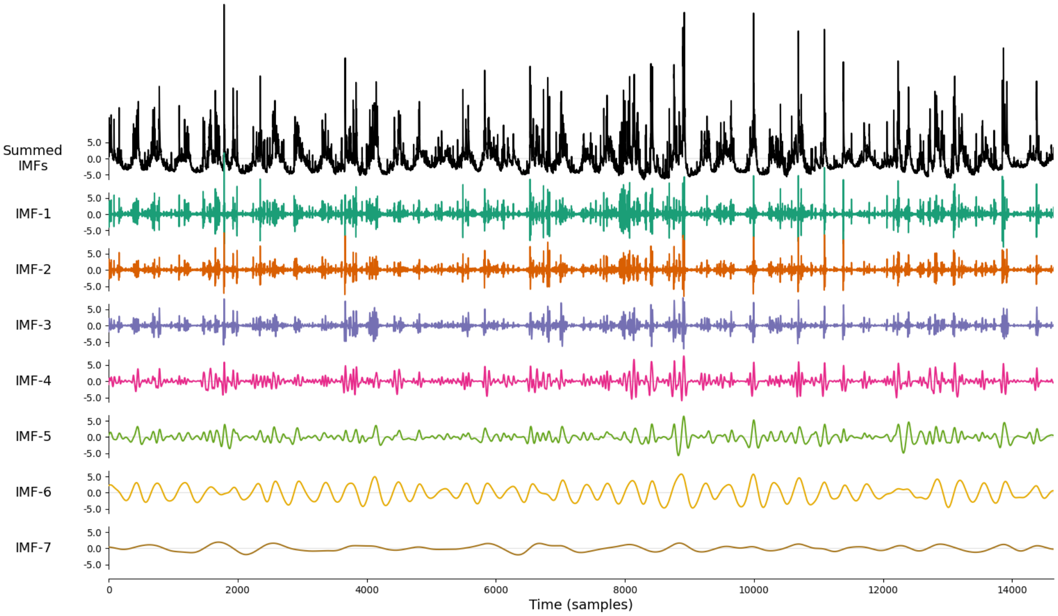

Figure 12: Decomposition of river discharge data into 8 intrinsic mode functions.

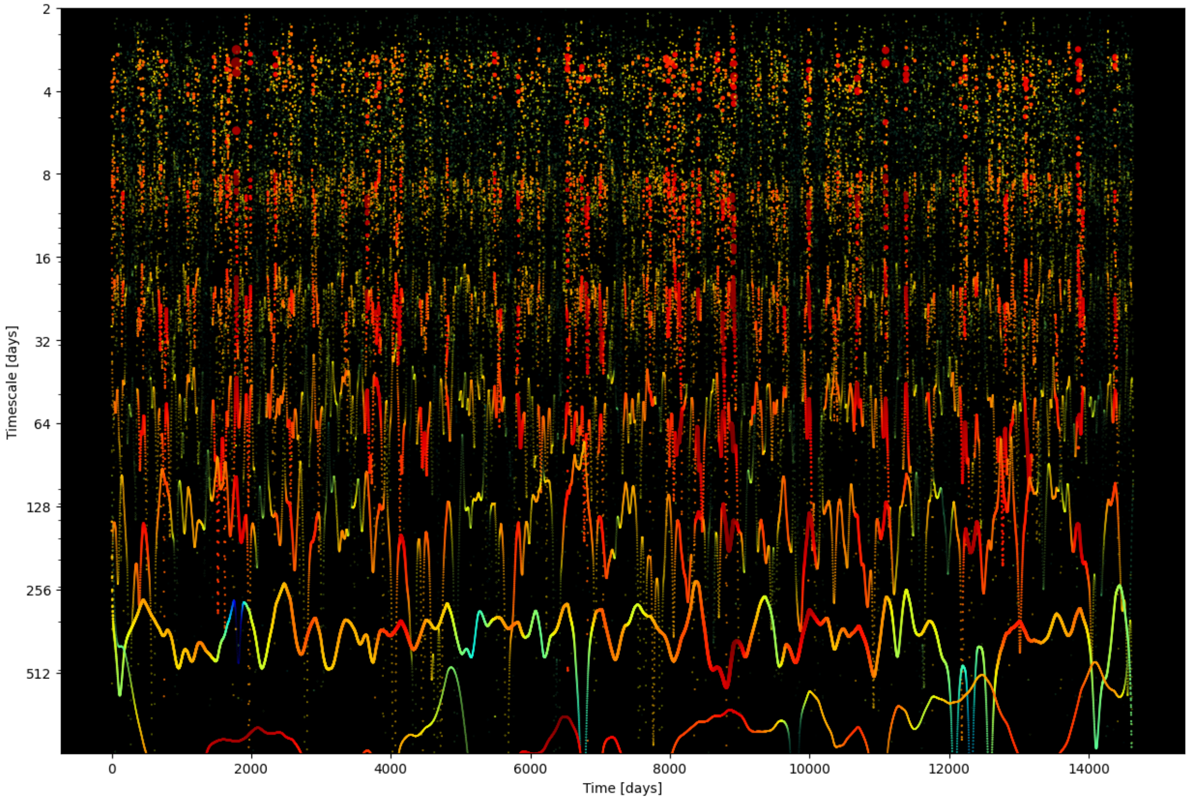

Figure 13: Hilbert spectrum of SSI data. Color and size depend on amplitude.