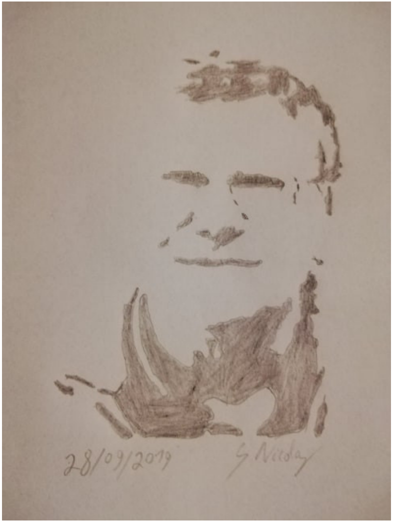
A time-filling talk about fractals

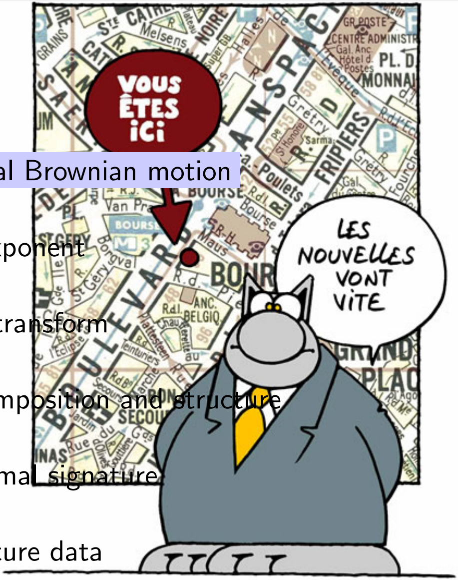
S. Nicolay

GdR 2019



To Alain...

We are here



- Fractional Brownian motion
- Hölder exponent
- Wavelet transform
- DNA: composition and structure
- Nucleosomal signature
- Temperature data

La Brownian motion: The Drunkard's Walk

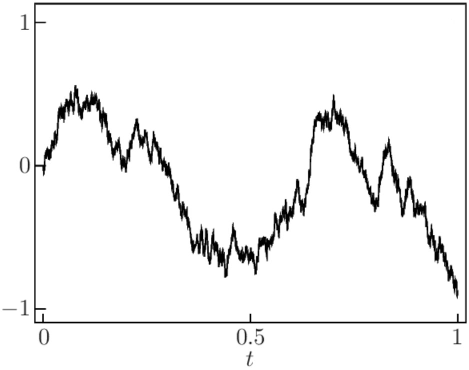
Video

Wiener process

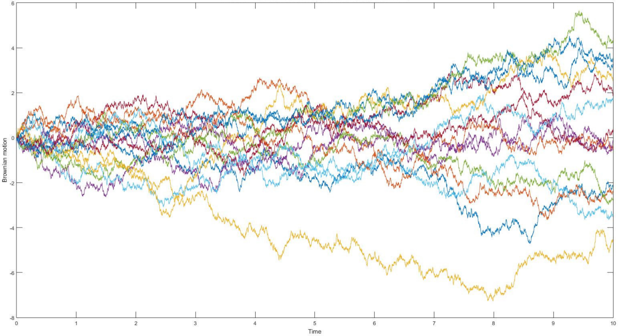
A Brownian motion is stochastic process B s.t.

- $B(0) = 0$,
- B is a.s. continu,
- B has independant increments,
- $B(t) - B(s) \sim N(0, t - s)$.

A realization of a Brownian motion



A realization of a Brownian motion



The fractional Brownian motion

Unlike classical Brownian motion, the increments of a fraction Brownian motion need not be independent. It is a continuous-time Gaussian process $B_H(t)$, that starts at zero, has expectation zero for all t, and has the following covariance function:

$$E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}),$$

where H is a real number in $(0, 1)$, called the Hurst index.

The fractional Brownian motion

Unlike classical Brownian motion, the increments of a fraction Brownian motion need not be independent. It is a continuous-time Gaussian process $B_H(t)$, that starts at zero, has expectation zero for all t, and has the following covariance function:

$$E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}),$$

where H is a real number in $(0,1)$, called the Hurst index.

The value of H determines what kind of process the fractional Brownian motion is:

- if $H = 1/2$ then the process is in fact a Brownian motion,
- if $H < 1/2$ then the increments of the process are negatively correlated,
- if $H > 1/2$ then the increments of the process are positively correlated.

The fractional Brownian motion

Unlike classical Brownian motion, the increments of a fraction Brownian motion need not be independent. It is a continuous-time Gaussian process $B_H(t)$, that starts at zero, has expectation zero for all t, and has the following covariance function:

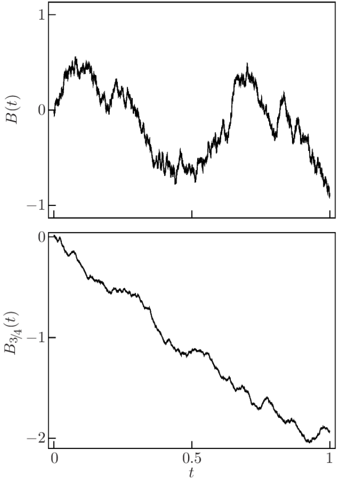
$$E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}),$$

where H is a real number in $(0,1)$, called the Hurst index.

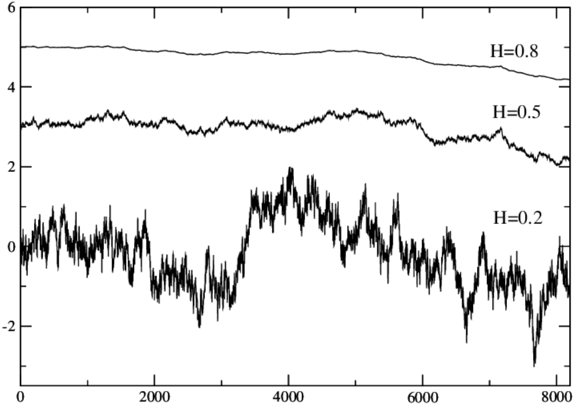
For $H > 1/2$ the process exhibits long-range dependence,

$$\sum_{j=1}^{\infty} E[B_H(1)(B_H(j+1) - B_H(j))] = \infty.$$

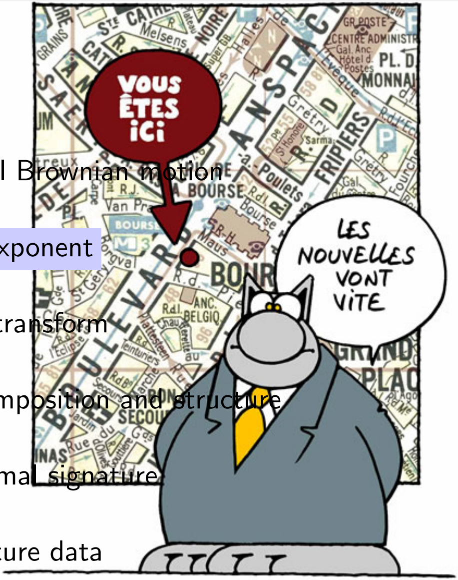
Représentations...



Représentations...



We are here



- Fractional Brownian motion
- Hölder exponent
- Wavelet transform
- DNA: composition and structure
- Nucleosomal signature
- Temperature data

Definition

Let $f : \mathbb{R} \to \mathbb{R}$ be a locally bounded function, $x \in \mathbb{R}$ and $\alpha \geq 0$; $f \in \Lambda^\alpha(x)$ if there exist $R, C > 0$ and a polynomial P_x of degree less than α such that

$$|h| < R \Rightarrow |f(x+h) - P_x(h)| \leq C|h|^\alpha. \qquad (*)$$

A function f belongs to C^α if there exists $C > 0$ such that $(*)$ holds for all x with $R = \infty$.

Definition

Let $f : \mathbb{R} \to \mathbb{R}$ be a locally bounded function, $x \in \mathbb{R}$ and $\alpha \geq 0$; $f \in \Lambda^\alpha(x)$ if there exist $R, C > 0$ and a polynomial P_x of degree less than α such that

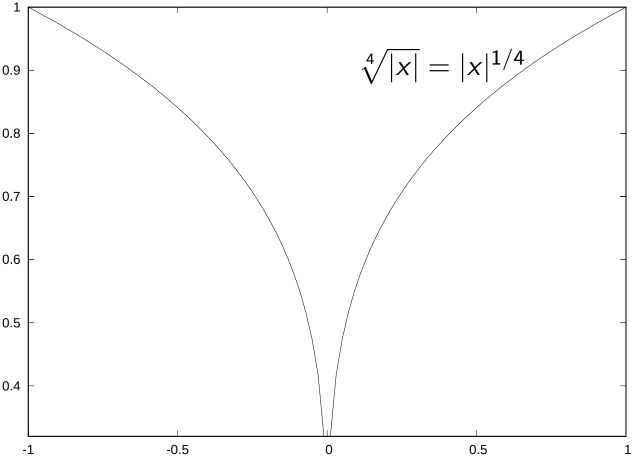
$$|h| < R \Rightarrow |f(x+h) - P_x(h)| \leq C|h|^\alpha. \qquad (*)$$

A function f belongs to C^α if there exists $C > 0$ such that $(*)$ holds for all x with $R = \infty$.

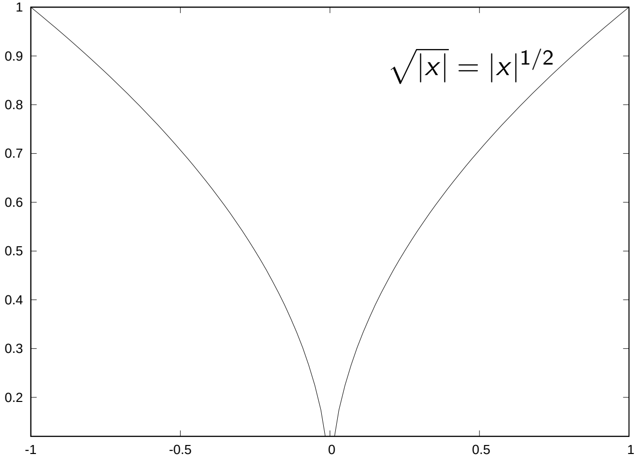
Definition

The Hölder exponent of f at x is $h(x) = \sup\{\alpha : f \in \Lambda^\alpha(x)\}$.

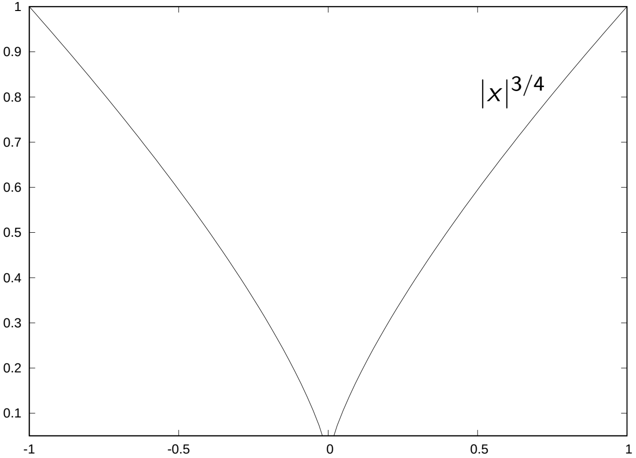
Example: $h(0) = 1/4$



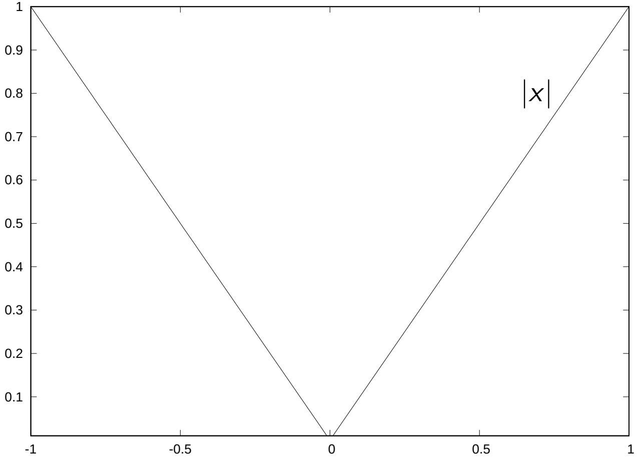
Example: $h(0) = 1/2$



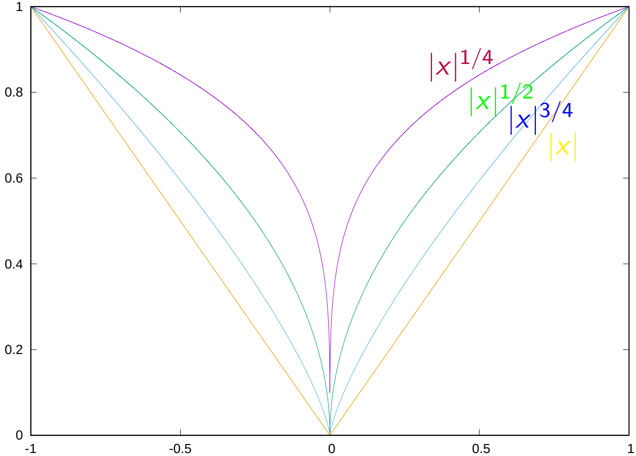
Example: $h(0) = 3/4$



Example: $h(0) = 1$



Hölder exponents from $1/4$ to 1 at 0 :



Théorème

The Hölder exponent of a realization of a fractional Brownian motion B_H with index H is almost surely equal to $h(x) = H$ for every x.

Théorème

The Hölder exponent of a realization of a fractional Brownian motion B_H with index H is almost surely equal to $h(x) = H$ for every x.

Thus for the fractional Brownien motion there is a direct link between correlation et regularity.

For very irregular functions, estimating $h(x)$ for every x is pointless.

For very irregular functions, estimating $h(x)$ for every x is pointless.

One rather tries to give global informations about the Hölder exponents by computing the size of the sets of points sharing the same Hölder exponent:

$$\mathcal{D}(h) = \dim_{\mathcal{H}}(\{x : h(x) = h\}),$$

where $\dim_{\mathcal{H}}$ stands for the Hausdroff dimension.

$\mathcal{D}$ is called the multifractal spectrum of f.

For very irregular functions, estimating $h(x)$ for every x is pointless.

One rather tries to give global informations about the Hölder exponents by computing the size of the sets of points sharing the same Hölder exponent:

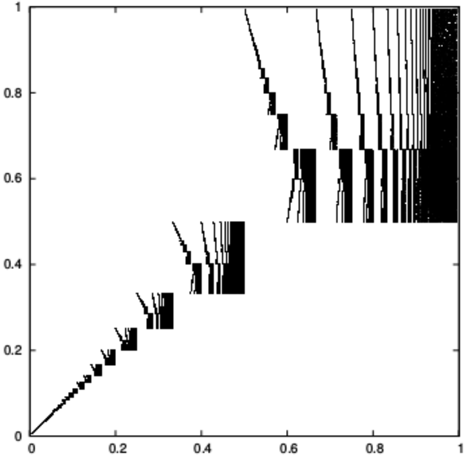
$$\mathcal{D}(h) = \dim_{\mathcal{H}}(\{x : h(x) = h\}),$$

where $\dim_{\mathcal{H}}$ stands for the Hausdroff dimension.

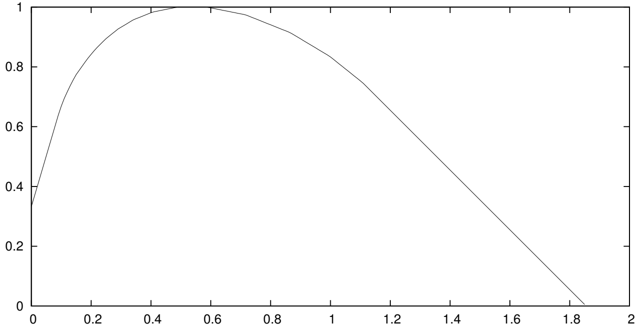
$\mathcal{D}$ is called the multifractal spectrum of f.

If $\mathrm{supp}(f)$ is reduced to a unit set, f is said to be monofractal.

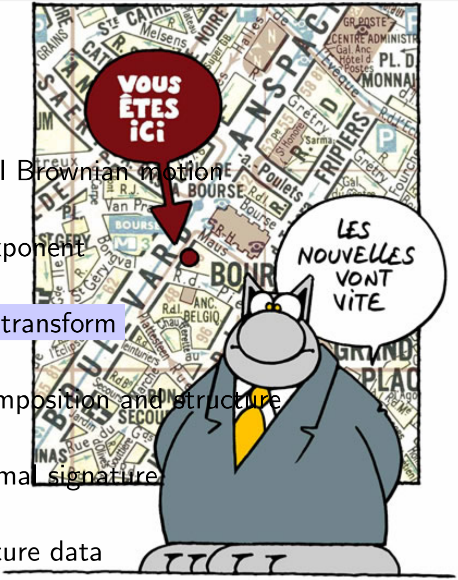
An example of multifractal function:



An example of multifractal function:

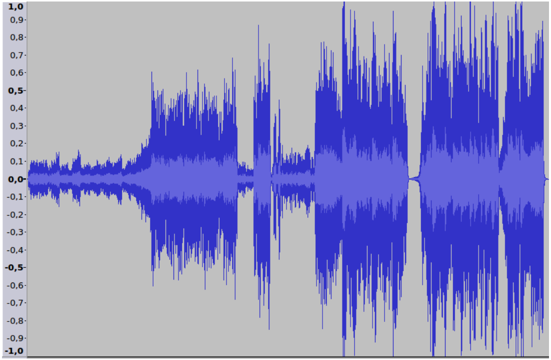


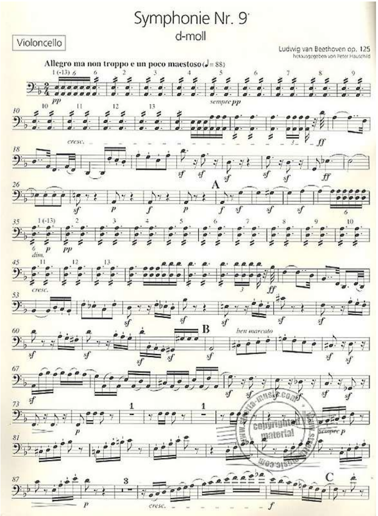
We are here



- Fractional Brownian motion
- Hölder exponent
- Wavelet transform
- DNA: composition and structure
- Nucleosomal signature
- Temperature data

Representation of a musical signal:



Representation of a musical signal:

Under some general conditions, there exist a function ϕ and $2^d - 1$ functions $\psi^{(i)}$ called wavelets s.t.

$$\{\phi(\cdot - k) : k \in \mathbb{Z}^d\} \bigcup \{\psi^{(i)}(2^j \cdot - k) : k \in \mathbb{Z}^d, j \in \mathbb{N}_0\}$$

forms an orthogonal basis of $L^2(\mathbb{R}^d)$.

Under some general conditions, there exist a function ϕ and $2^d - 1$ functions $\psi^{(i)}$ called wavelets s.t.

$$\{\phi(\cdot - k) : k \in \mathbb{Z}^d\} \bigcup \{\psi^{(i)}(2^j \cdot - k) : k \in \mathbb{Z}^d, j \in \mathbb{N}_0\}$$

forms an orthogonal basis of $L^2(\mathbb{R}^d)$.
Any function $f \in L^2(\mathbb{R}^d)$ can be decomposed as follows,

$$f(x) = \sum_{k \in \mathbb{Z}^d} C_k \phi(x - k) + \sum_{j \geq 0, k \in \mathbb{Z}^d, 1 \leq i < 2^d} c_{j,k}^{(i)} \psi^{(i)}(2^j x - k),$$

with

$$C_k = \int f(x)\phi(x - k)\, dx, \quad c_{j,k}^{(i)} = 2^{dj} \int f(x)\psi^{(i)}(2^j x - k)\, dx.$$

We assume

- $\phi, \psi^{(i)} \in \Lambda^n(\mathbb{R}^d)$ with $n > M$,
- $D^\beta \phi$, $D^\beta \psi^{(i)}$ ($|\beta| \leq n$) have fast decay,
- $\mathrm{supp}(\psi^{(i)}) \subset 2^{-j_0} B$ for some j_0.

We assume

- $\phi, \psi^{(i)} \in \Lambda^n(\mathbb{R}^d)$ with $n > M$,
- $D^\beta \phi$, $D^\beta \psi^{(i)}$ $(|\beta| \leq n)$ have fast decay,
- $\mathrm{supp}(\psi^{(i)}) \subset 2^{-j_0} B$ for some j_0.

We set

- $\lambda = \lambda(i, j, k) = \dfrac{k}{2^j} + \dfrac{i}{2^{j+1}} + [0, \dfrac{1}{2^{j+1}})^d$
- $c_\lambda = c_{j,k}^{(i)}$
- $\psi_\lambda = \psi^{(i)}(2^j \cdot -k)$.

The wavelet leaders are defined by

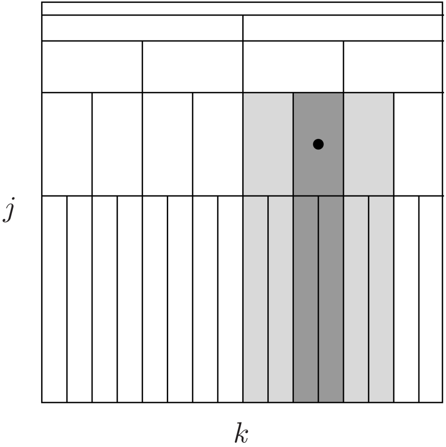
$$d_\lambda = \sup_{\lambda' \subset \lambda} |c_{\lambda'}|$$

The wavelet leaders are defined by

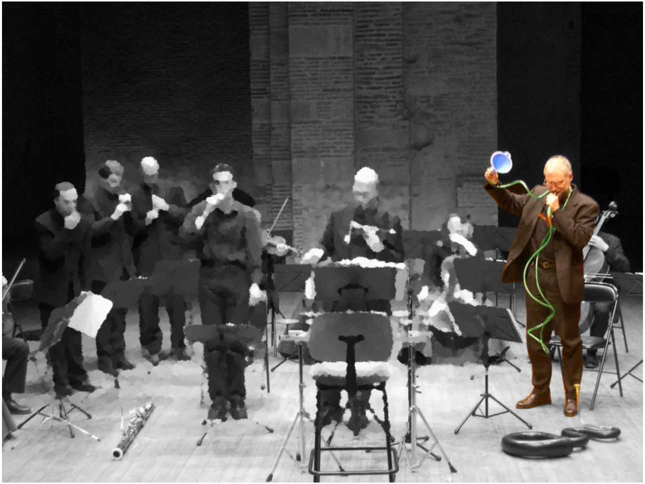
$$d_\lambda = \sup_{\lambda' \subset \lambda} |c_{\lambda'}|$$

If 3λ denotes the 3^d dyadic cubes adjacent to λ and $\lambda_j(x_0)$ the dyadic cube of length 2^{-j} containing x_0, one sets

$$d_j(x_0) = \sup_{\lambda \subset 3\lambda_j(x_0)} d_\lambda$$


j
k

One selects the coefficients related to x_0 up to scale j and then take the sup.

One selects the coefficients related to x_0 up to scale j and then take the sup.

Theorem

If $f \in \Lambda^{\alpha}(x_0)$, then there exists $C > 0$ s.t.

$$d_j(x_0) \leq C2^{-j\alpha},$$

for j sufficiently large.

Theorem

If $f \in \Lambda^\alpha(x_0)$, then there exists $C > 0$ s.t.

$$d_j(x_0) \leq C 2^{-j\alpha},$$

for j sufficiently large.

Conversely, if $f \in \Lambda^\epsilon(\mathbb{R}^d)$ for some $\epsilon > 0$ and f satisfies the previous relation, then $f \in \Lambda^\alpha(x_0)$, up to a logarithmic correction.

A formalism to obtain the multifractal spectrum: One sets

$$S(q,j) = 2^{-dj} \sum_{\lambda \in \Lambda_j} d_\lambda^q,$$

where Λ_j represents the cubes at scale j;

A formalism to obtain the multifractal spectrum: One sets

$$S(q,j) = 2^{-dj} \sum_{\lambda \in \Lambda_j} d_\lambda^q,$$

where Λ_j represents the cubes at scale j;

$$\omega(q) = \liminf_{j \to \infty} \frac{\log S(q,j)}{\log 2^{-j}},$$

A formalism to obtain the multifractal spectrum: One sets

$$S(q,j) = 2^{-dj} \sum_{\lambda \in \Lambda_j} d_\lambda^q,$$

where Λ_j represents the cubes at scale j;

$$\omega(q) = \liminf_{j \to \infty} \frac{\log S(q,j)}{\log 2^{-j}},$$

to estimate the spectrum as follow:

$$D(h) = \inf_q \{qh - \omega(q)\} + d.$$

Theorem

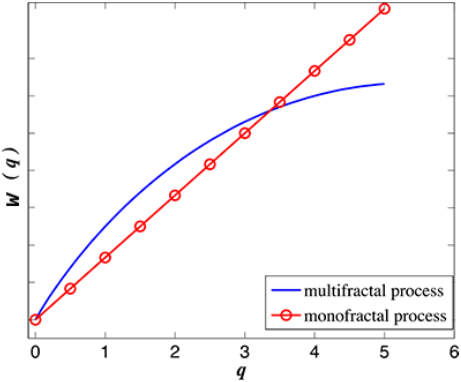
One has

$$\mathcal{D}(h) \leq \inf_{q \in \mathbb{R}} \{qh - \omega(q)\} + d.$$

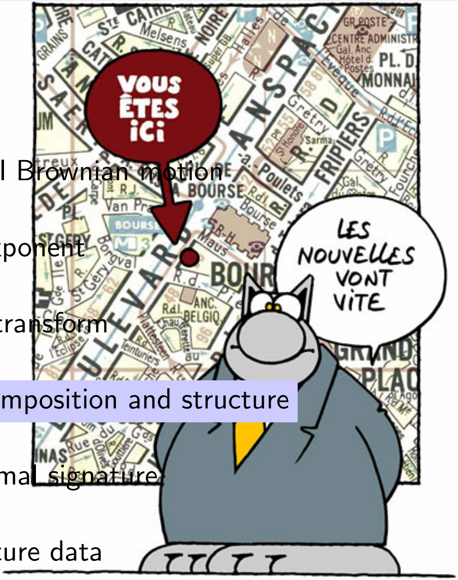
Theorem

One has

$$\mathcal{D}(h) \leq \inf_{q\in\mathbb{R}} \{qh - \omega(q)\} + d.$$



We are here



- Fractional Brownian motion
- Hölder exponent
- Wavelet transform
- DNA: composition and structure
- Nucleosomal signature
- Temperature data

To compose a macromolecule (polymer), we need repeated subunits (monomers)...

To compose a macromolecule (polymer), we need repeated subunits (monomers)...

(4 for the DNA)

These units are brought together to form a macromolecule...

These units are brought together to form a macromolecule...

These units are brought together to form a macromolecule...

In the chain, 🤝 représents a “chemical bond” (for DNA: a phosphodiester bond)

In 1953 Watson and Crick came up with their historic model of the shape of DNA: a double helix structure.

In 1953 Watson and Crick came up with their historic model of the shape of DNA: a double helix structure.

The DNA molecule is composed of two chains (polynucleotidic strands) that coil around each other to form a double helix; the two strands of DNA run in opposite directions to each other.

In 1953 Watson and Crick came up with their historic model of the shape of DNA: a double helix structure.

The DNA molecule is composed of two chains (polynucleotidic strands) that coil around each other to form a double helix; the two strands of DNA run in opposite directions to each other.

A unit from a chain is specifically paired with another unit from the other chain (with hydrogen bonds).

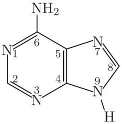
In 1953 Watson and Crick came up with their historic model of the shape of DNA: a double helix structure.

The DNA molecule is composed of two chains (polynucleotidic strands) that coil around each other to form a double helix; the two strands of DNA run in opposite directions to each other.

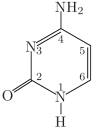
A unit from a chain is specifically paired with another unit from the other chain (with hydrogen bonds).

As a consequence, the two strands are complementary to each other: the sequence on a chain entirely determine the sequece on the other chain.

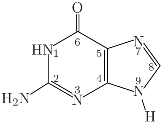
For the DNA: the units are nucleotides.


↔


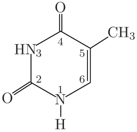
(adenine)


↔


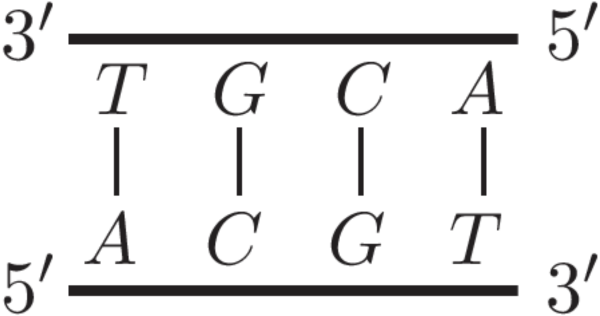
(cytosine)

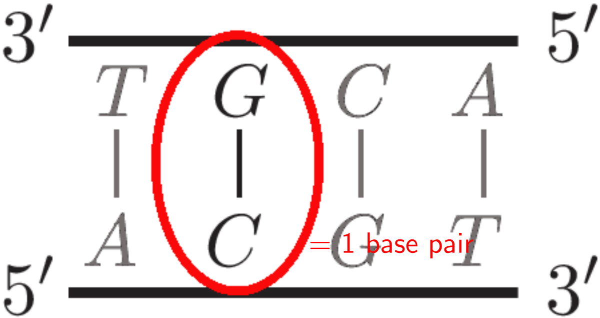

↔


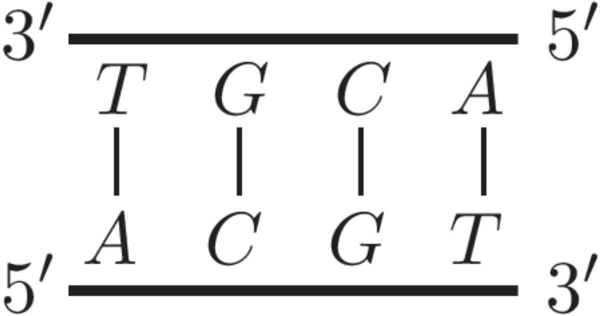
(guanine)


↔


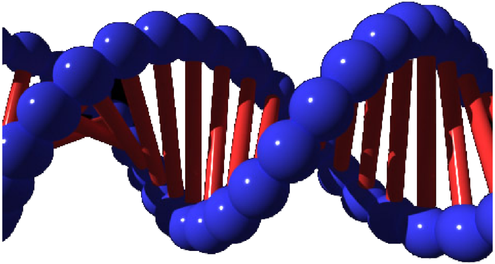
(thymine)


$3'$
$5'$
T
G
C
A
A
C
G
T
$5'$
$3'$


3′
5′
T
G
C
A
A
C
G
T
= 1 base pair
5′
3′



← 4 base pairs (bp) →



Transcription is the first step of DNA based gene expression, in which a particular segment of DNA is copied into RNA (by the enzyme RNA polymerase).
A gene is a sequence of nucleotides in DNA (or RNA) that codes for a molecule that has a function.

The RNA can be directly functional or be the intermediate template for a protein that performs a function.
More precisely, a succession of nucleotides determine a sequence of amino acids that make up proteins.

Transcription is the first step of DNA based gene expression, in which a particular segment of DNA is copied into RNA (by the enzyme RNA polymerase).
A gene is a sequence of nucleotides in DNA (or RNA) that codes for a molecule that has a function.

The RNA can be directly functional or be the intermediate template for a protein that performs a function.
More precisely, a succession of nucleotides determine a sequence of amino acids that make up proteins.
A triplet of nucleotides (codon) generally corresponds to a single amino acid; this mechanism is highly similar among all organisms (from bacteria to humans).

Transcription is the first step of DNA based gene expression, in which a particular segment of DNA is copied into RNA (by the enzyme RNA polymerase).
A gene is a sequence of nucleotides in DNA (or RNA) that codes for a molecule that has a function.

The RNA can be directly functional or be the intermediate template for a protein that performs a function.
More precisely, a succession of nucleotides determine a sequence of amino acids that make up proteins.
The genetic material of an organism (consisting of DNA) is called genome.
For the human : $\sim 3\,000\,000\,000$bp nucleotides
corresponding to $\sim 30\,000$ genes.

Transcription is the first step of DNA based gene expression, in which a particular segment of DNA is copied into RNA (by the enzyme RNA polymerase).
A gene is a sequence of nucleotides in DNA (or RNA) that codes for a molecule that has a function.

Although unicellular eukaryotes (such as yeast) only have coding sequences, metazoans and especially vertebrate genomes have a large fraction of non-coding DNA.
For instance, in the human genome only 1.1% of the genome is spanned by sequences encoding for the final RNA (exons), whereas 24% is removed from the final RNA (introns),
with 75% of the genome being intergenic DNA.

Transcription is the first step of DNA based gene expression, in which a particular segment of DNA is copied into RNA (by the enzyme RNA polymerase).
A gene is a sequence of nucleotides in DNA (or RNA) that codes for a molecule that has a function.

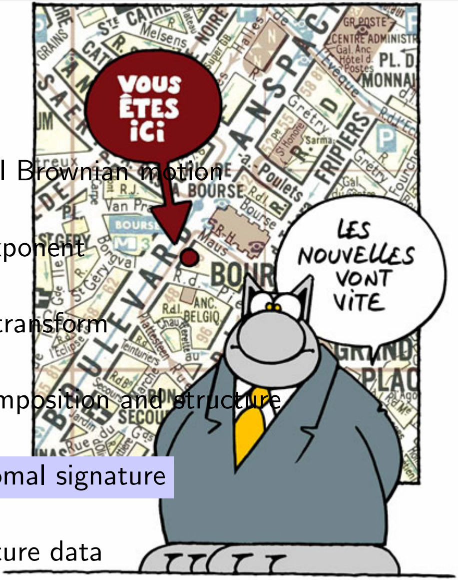
A lot of nucleotides has no active role in the transcription!
$=$ trash bin?

Transcription is the first step of DNA based gene expression, in which a particular segment of DNA is copied into RNA (by the enzyme RNA polymerase).
A gene is a sequence of nucleotides in DNA (or RNA) that codes for a molecule that has a function.

A lot of nucleotides has no active role in the transcription!
= trash bin?

Our goal: to study these regions.

We are here



- Fractional Brownian motion
- Hölder exponent
- Wavelet transform
- DNA: composition and structure
- Nucleosomal signature
- Temperature data

In signal analysis, one consider a DNA strand as a word made of four letters (A, C, G and T).

ACGAACGTTACGGAATGCCAGGGTACCC...

In signal analysis, one consider a DNA strand as a word made of four letters (A, C, G and T).

ACGAACGTTACGGAATGCCAGGGTACCC...

There is too much information:
One has to transform this word into a function.

The PNuc coding is made by watching the bending profiles of the nucleosomal DNA.

This coding tries to explain the natural curvature of the double helix (in a nucleosomal context).

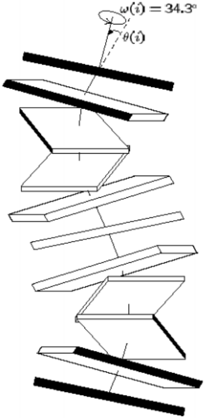
The PNuc coding is made by watching the bending profiles of the nucleosomal DNA.

This coding tries to explain the natural curvature of the double helix (in a nucleosomal context).

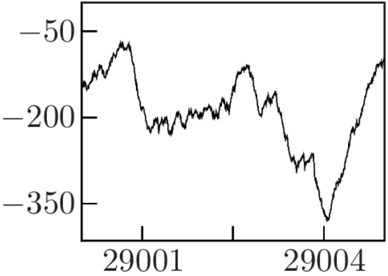
Tri-base	PNuc	Tri-base	PNuc	Tri-base	PNuc	Tri-base	PNuc
AAA	0	CAA	3.3	GAA	3	TAA	2
AAC	3.7	CAC	6.5	GAC	5.4	TAC	3.7
AAG	5.2	CAG	4.2	GAG	5.4	TAG	2.2
AAT	0.7	CAT	6.7	GAT	5.3	TAT	2.8
ACA	5.2	CCA	5.4	GCA	6	TCA	5.4
ACC	5.4	CCC	6	GCC	10	TCC	3.8
ACG	5.4	CCG	4.7	GCG	7.5	TCG	8.3
ACT	5.8	CCT	5.4	GCT	7.5	TCT	3.3
AGA	3.3	CGA	8.3	GGA	3.8	TGA	5.4
AGC	7.5	CGC	7.5	GGC	10	TGC	6
AGG	5.4	CGG	4.7	GGG	6	TGG	5.4
AGT	5.8	CGT	5.4	GGT	5.4	TGT	5.2
ATA	2.8	CTA	2.2	GTA	3.7	TTA	2
ATC	5.3	CTC	5.4	GTC	5.4	TTC	3
ATG	6.7	CTG	4.2	GTG	6.5	TTG	3.3
ATT	0.7	CTT	5.2	GTT	3.7	TTT	0

The PNuc coding is made by watching the bending profiles of the nucleosomal DNA.

This coding tries to explain the natural curvature of the double helix (in a nucleosomal context).



The PNuc signal for chromosome 21 (in kpb):



Using the wavelet transform, one can estimate the Hölder exponent of these walks.

Using the wavelet transform, one can estimate the Hölder exponent of these walks.
They are monofractal and the Hölder exponent H depends on the considered organism.

Using the wavelet transform, one can estimate the Hölder exponent of these walks.
They are monofractal and the Hölder exponent H depends on the considered organism.

For eukaryotes, H is close to 0.6.

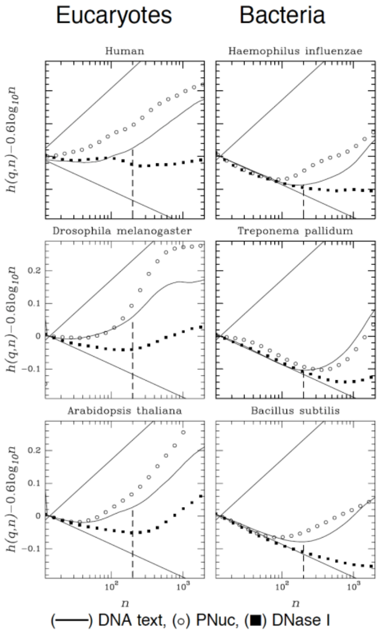
Using the wavelet transform, one can estimate the Hölder exponent of these walks.
They are monofractal and the Hölder exponent H depends on the considered organism.

For eukaryotes, H is close to 0.6.
Following the "Brownian model", these organisms display long range correlations.

Using the wavelet transform, one can estimate the Hölder exponent of these walks.
They are monofractal and the Hölder exponent H depends on the considered organism.

For eukaryotes, H is close to 0.6.
Following the "Brownian model", these organisms display long range correlations.

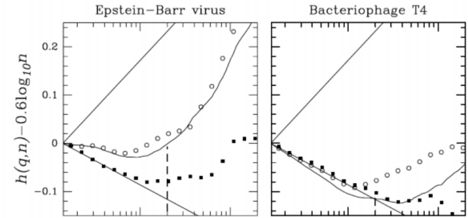
For bacteria,the value is closer to $1/2$.
In this case, with the same model, no correlation is observed.


Eucaryotes
Bacteria
Human
Haemophilus influenzae
Drosophila melanogaster
Treponema pallidum
Arabidopsis thaliana
Bacillus subtilis
$h(q,n)-0.6\log_{10}n$
0.2
0.1
0
−0.1
10^2
10^3
n
(——) DNA text, (○) PNuc, (■) DNase I

In these tests, some bacteria were associated to a value H significantly larger than $1/2$.

In these tests, some bacteria were associated to a value H significantly larger than $1/2$.
Indeed, these are Pox viruses.

In these tests, some bacteria were associated to a value H significantly larger than $1/2$.
Indeed, these are Pox viruses.



In these tests, some bacteria were associated to a value H significantly larger than $1/2$.

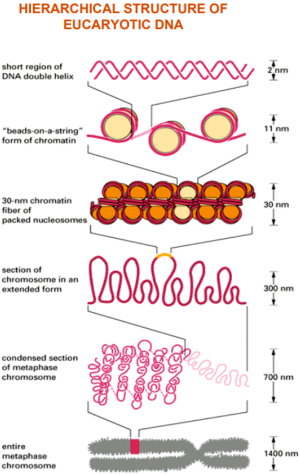
Indeed, these are Pox viruses.

As bacteria, these organisms are prokaryotes,

In these tests, some bacteria were associated to a value H significantly larger than $1/2$.
Indeed, these are Pox viruses.
As bacteria, these organisms are prokaryotes,
but display a packing mechanism more similar to the packing mechanism of the eukaryotes,
involving preoteins comparable to histones.

In these tests, some bacteria were associated to a value H significantly larger than $1/2$.
Indeed, these are Pox viruses.
As bacteria, these organisms are prokaryotes,
but display a packing mechanism more similar to the packing mechanism of the eukaryotes,
involving preoteins comparable to histones.

A Hölder exponent larger than $1/2$ is thus related to nucleosome like structures!


HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA
short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in an extended form
300 nm
condensed section of metaphase chromosome
700 nm
entire metaphase chromosome
1400 nm

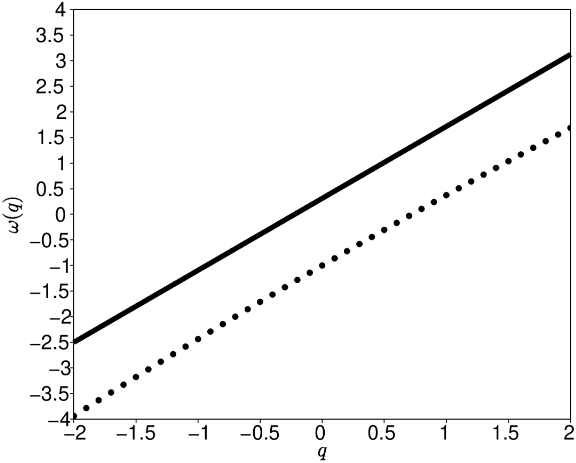
We are here

- Fractional Brownian motion
- Hölder exponent
- Wavelet transform
- DNA: composition and structure
- Nucleosomal signature
- Temperature data

We whish here to analyze multifractal properties of the air temperature data from 1951 and 2003 (the World Meteorological Organization was established in 1950).

We whish here to analyze multifractal properties of the air temperature data from 1951 and 2003 (the World Meteorological Organization was established in 1950).

For Roma:

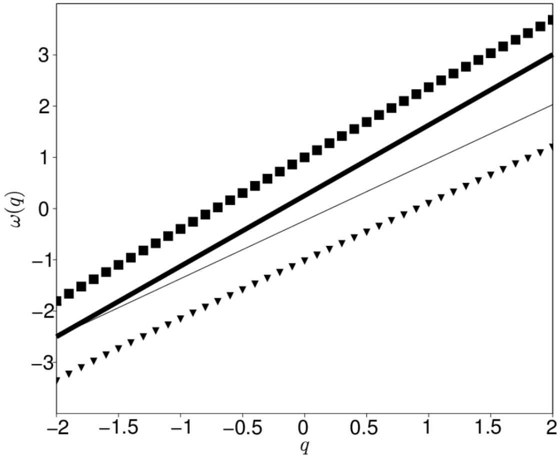


We whish here to analyze multifractal properties of the air temperature data from 1951 and 2003 (the World Meteorological Organization was established in 1950).

These signals are clearly monofractal.

We whish here to analyze multifractal properties of the air temperature data from 1951 and 2003 (the World Meteorological Organization was established in 1950).

For Roma and Armagh:



We whish here to analyze multifractal properties of the air temperature data from 1951 and 2003 (the World Meteorological Organization was established in 1950).

The Hölder exponents range from 1.093 to 1.43 (1.239 ± 0.087).

Are these observations random?

Are these observations random?

Let's to correlate the Hölder exponents with pressure anomalies (data: NCEP-NCAR).

Are these observations random?

Let's to correlate the Hölder exponents with pressure anomalies (data: NCEP-NCAR).

To do so, the map of Europe is gridded into roughly 200 kilometers square pixels.
We compare the map of the inverses of standard deviations of surface pressure anomalies with the map made of the measured Hölder exponents.
We used the Frobenius distance

$$d = \sqrt{\sum_{i,j} (x_{i,j} - x'_{i,j})^2}$$

where $x_{i,j}$ is a pixel of the first map, $x'_{i,j}$ is the corresponding pixel of the second map and where the sum is taken over all pixels.

We get $d_1 = 2.68$ ...

We get $d_1 = 2.68$ ... is it big?

We get $d_1 = 2.68$ ... is it big?

In order to check if these maps are akin, we use a standard Monte-Carlo method: the “Hölder map” is randomly shuffled $10,000$ times.
For each realization, the distance with the original anomalies map is computed in order to get a distribution of these random distances. In this way, one can look where d_1 lies in the distribution of the distances, and one can associate a p-value to this particular distance d_1.

We get $d_1 = 2.68$ ... is it big?

Based on the $10,000$ observations, the probability $1 - p$ to have a randomly shuffled map with a distance smaller than d_1 is lower than 10^{-4},
which shows that the null hypothesis can be rejected with a high confidence level.
In other words, the higher the standard deviation of pressure anomalies, the lower the Hölder exponents.

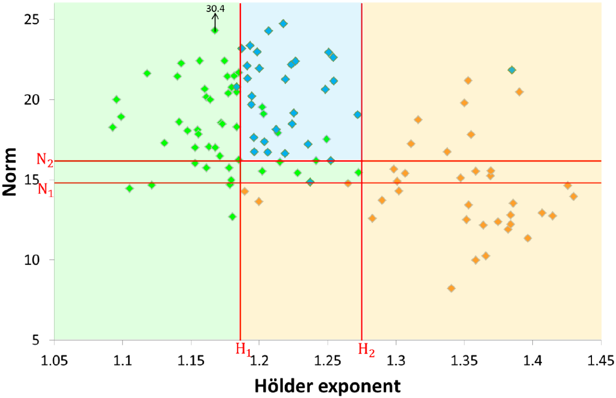
In order to link the Hölder exponents of the stations with the climate they are associated to,
we determine the climate type of each station according to a simplified version of the Köppen-Geiger climate classification.
We focus only on temperature variability (mainly based on minimum and maximum monthly mean temperatures).

In order to link the Hölder exponents of the stations with the climate they are associated to,
we determine the climate type of each station according to a simplified version of the Köppen-Geiger climate classification. We focus only on temperature variability (mainly based on minimum and maximum monthly mean temperatures).

These induce four different types of climate: Mediterranean (Ca-type), Oceanic (Cb-type), Continental (D-type).

In order to check if the Köppen-Geiger climate classification can be recovered from the Hölder analysis,
we plot the points cloud representing the Hölder exponents (abscissa) and norms (ordinate) of the stations,
where each point is colored according to its climate type.

In order to check if the Köppen-Geiger climate classification can be recovered from the Hölder analysis,
we plot the points cloud representing the Hölder exponents (abscissa) and norms (ordinate) of the stations,
where each point is colored according to its climate type.



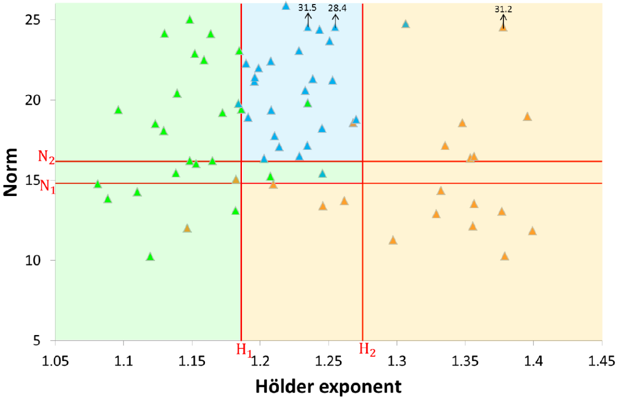
In order to check if the Köppen-Geiger climate classification can be recovered from the Hölder analysis,
we plot the points cloud representing the Hölder exponents (abscissa) and norms (ordinate) of the stations,
where each point is colored according to its climate type.

Since points with the same color are concentrated in the same areas, one may try to cut the plane into rectangles to isolate the three climates.
A simple possibility is to use only two vertical cuts (let us say at H_1 and H_2) and two horizontal cuts (let us say at N_1 and N_2), which gives 9 rectangles. Each of them is then associated to the climate type that is the most abundant therein,
which gives rise to a new kind of climate classification.
In this case, 93.9% of the stations are correctly classified, i.e. their Köppen climate type is recovered.

In order to validate the results, we performed a blind test on 69 other weather stations in the same area, with the same Hölder-based classification.
Of course, the values of H_1, H_2, N_1 and N_2 are left unchanged.

In order to validate the results, we performed a blind test on 69 other weather stations in the same area, with the same Hölder-based classification.

Of course, the values of H_1, H_2, N_1 and N_2 are left unchanged.



In order to validate the results, we performed a blind test on 69 other weather stations in the same area, with the same Hölder-based classification.
Of course, the values of H_1, H_2, N_1 and N_2 are left unchanged.

One can see that most of the points (88.4%) are correctly classified, which validates the "Hölder-based classification".

In order to validate the results, we performed a blind test on 69 other weather stations in the same area, with the same Hölder-based classification.

Of course, the values of H_1, H_2, N_1 and N_2 are left unchanged.