Assignment 3 (ML for TS) - MVA 2021/2022

Gabriel ATHENES gabriel.athenes@polytechnique.edu
Benoît ROUSSEL btroussel@gmail.com

March 18, 2022

1 Introduction

Objective. The goal is to present (i) a model selection heuristics to find the number of change-points in a signal and (ii) wavelets for graph signals.

Warning and advice.

- Use code from the tutorials as well as from other sources. Do not code yourself well-known procedures (e.g. cross validation or k-means), use an existing implementation.
- The associated notebook contains some hints and several helper functions.
- Be concise. Answers are not expected to be longer than a few sentences (omitting calculations).

Instructions.

- Fill in your names and emails at the top of the document.
- Hand in your report (one per pair of students) by Friday 18th March 11:59 PM.
- Rename your report and notebook as follows:
  `FirstnameLastname1_FirstnameLastname1.pdf` and
  `FirstnameLastname2_FirstnameLastname2.ipynb`.
  For instance, `LaurentOudre_CharlesTruong.pdf`.
- Upload your report (PDF file) and notebook (IPYNB file) using this link: dropbox.com/request/5DKPDBVAJ25hon0ZZsnn.

2 Model selection for change-point detection

Notations. In the following, $\|x\|$ is the Euclidean norm of x if x is a vector and the Frobenius norm if x is a matrix. A set of change-points is denoted by a bold $\boldsymbol{\tau} = \{t_1, t_2, \dots\}$ and $|\boldsymbol{\tau}|$ (the cardinal of $\boldsymbol{\tau}$) is the number of change-points. By convention $t_0 = 0$ and $t_{|\boldsymbol{\tau}|+1} = T$. For a set of change-points $\boldsymbol{\tau}$, $\Pi_{\boldsymbol{\tau}}$ is the orthogonal projection onto the linear subspace of piecewise constant signals with change-points in $\boldsymbol{\tau}$: for a signal $x = \{x_t\}_{t=0}^{T-1}$,

$$(\Pi_{\boldsymbol{\tau}} x)_t = \bar{x}_{t_k..t_{k+1}} \quad \text{if } t_k \leq t < t_{k+1} \tag{1}$$

where $\bar{x}_{t_k..t_{k+1}}$ is the empirical mean of the subsignal $x_{t_k..t_{k+1}} = \{x_t\}_{t_k}^{t_{k+1}-1}$.

Model selection. Assume we observe a $\mathbb{R}^d$-valued signal $y = \{y_t\}_{t=0}^{T-1}$ with T samples that follows the model

$$y_t = f_t + \varepsilon_t \tag{2}$$

where f is a deterministic signal which we want to estimate with piecewise constant signals and ε_t is i.i.d. with mean 0 and covariance $\sigma^2 I_d$.

The ideal choice of $\boldsymbol{\tau}$ minimizes the distance from the true (noiseless) signal f:

$$\boldsymbol{\tau}^\star = \arg\min_{\boldsymbol{\tau}} \frac{1}{T} \|f - \Pi_{\boldsymbol{\tau}} y\|^2 . \tag{3}$$

The estimator $\boldsymbol{\tau}^\star$ is an *oracle* estimator because it relies on the unknown signal f. Several model selection procedures rely on the "unbiased risk estimation heuristics": if $\hat{\boldsymbol{\tau}}$ minimizes a criterion $\operatorname{crit}(\boldsymbol{\tau})$ such that

$$\mathbb{E}\left[\operatorname{crit}(\boldsymbol{\tau})\right] \approx \mathbb{E}\left[\frac{1}{T} \|f - \Pi_{\boldsymbol{\tau}} y\|^2\right] \tag{4}$$

then

$$\frac{1}{T} \|f - \Pi_{\hat{\boldsymbol{\tau}}} y\|^2 \approx \min_{\boldsymbol{\tau}} \frac{1}{T} \|f - \Pi_{\boldsymbol{\tau}} y\|^2 \tag{5}$$

under some conditions. In other words, the estimator $\hat{\boldsymbol{\tau}}$ approximately minimizes the oracle quadratic risk.

Here, we will consider penalized criteria:

$$\operatorname{crit}(\boldsymbol{\tau}) = \frac{1}{T} \|y - \Pi_{\boldsymbol{\tau}} y\|^2 + \operatorname{pen}(\boldsymbol{\tau}) \tag{6}$$

where pen is a penalty function. In addition, let

$$\hat{\boldsymbol{\tau}}_{\text{pen}} := \arg\min_{\boldsymbol{\tau}} \left[\frac{1}{T} \|y - \Pi_{\boldsymbol{\tau}} y\|^2 + \operatorname{pen}(\boldsymbol{\tau})\right] . \tag{7}$$

Question 1 *Ideal penalty*

- Calculate $\mathbb{E}[\|\varepsilon\|^2/T]$, $\mathbb{E}[\|\mu^\star - \Pi_{\boldsymbol{\tau}} y\|^2/T]$ and $\mathbb{E}[\|y - \Pi_{\boldsymbol{\tau}} y\|^2/T]$.
- What would be an ideal penalty $\mathrm{pen}_{\mathrm{id}}$ such that Equation (4) is verified?

Answer 1

If X is a gaussian vector $\sim \mathcal{N}(0, \sigma^2 I)$ and P_F is an orthogonal projection on a vector space F of dimension p, then $\frac{\|P_F(X)\|^2}{\sigma^2} \sim \chi^2(p)$ and $\mathbb{E}(\|P_F(X)\|^2) = \sigma^2 p$.

From this theorem, taking $P_F = I$ we have that $\frac{\mathbb{E}(\|\epsilon\|^2)}{T} = \frac{\mathbb{E}(\|\sum_{i=1}^d \epsilon\|^2)}{T} = \frac{\sum_{i=1}^d \mathbb{E}(\|\epsilon_i\|^2)}{T} = d\sigma^2$.

Noting that Π_t is an orthogonal projection and that therefore on a space of dimension $|\tau| + 1$, we have using Pythagore's theorem that

$$\frac{\mathbb{E}(\|f - \Pi_t(y)\|^2)}{T} = \frac{\mathbb{E}(\|f - \Pi_t(f) - \Pi_t(\epsilon)\|^2)}{T} = \frac{\mathbb{E}(\|f - \Pi_t(f)\|^2) + E(\|\Pi_t(\epsilon)\|^2)}{T}$$

and so using Cochran's theorem for the second term of the component of the right term

$$\frac{\mathbb{E}(\|f - \Pi_t(y)\|^2)}{T} = \frac{\mathbb{E}(\|f - \Pi_t(f)\|^2)}{T} + d\sigma^2 \frac{|\tau| + 1}{T}$$

and similarly, using the fact that $\mathbb{E}[((I - \Pi_t) \cdot f)^T (I - \Pi_t) \cdot \epsilon] = ((I - \Pi_t) \cdot f)^T \mathbb{E}[\epsilon] = 0$, we have

$$\frac{\mathbb{E}(\|y - \Pi_t(y)\|^2)}{T} = \frac{\mathbb{E}(\|f - \Pi_t(f)\|^2)}{T} + \frac{\mathbb{E}(\|(I - \Pi_t)(\epsilon)\|^2)}{T} = \frac{\mathbb{E}(\|f - \Pi_t(f)\|^2)}{T} + d\sigma^2 (1 - \frac{|\tau| + 1}{T})$$

.

Using equation (4) we have that the ideal penalty is $pen_{id}(\tau) = 2d\sigma^2(\frac{|\tau|+1}{T}) - d\sigma^2$ or we have that the ideal penalty is $pen_{id}(\tau) = 2d\sigma^2 \frac{|\tau|}{T}$ as the constant does not depend on τ.

Question 2 *Mallows' C_p*

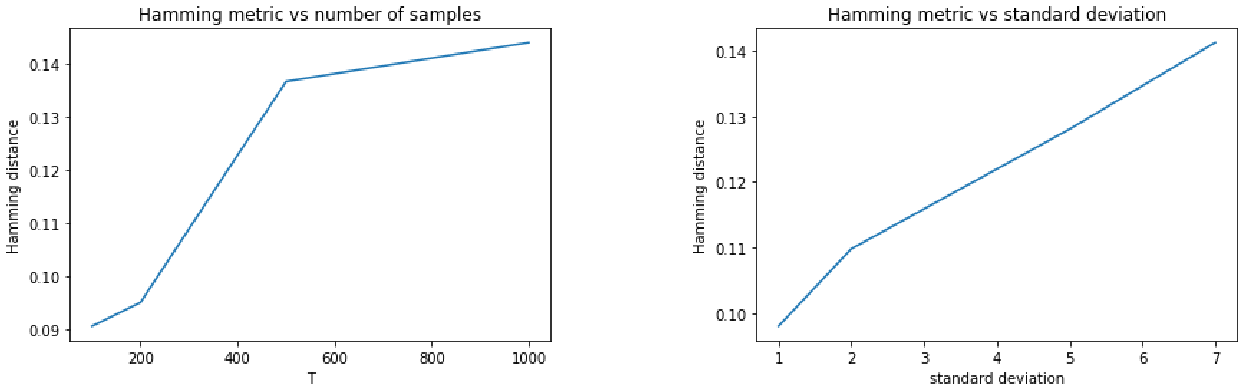
The ideal penalty depends on the unknown value of σ. Pluging an estimator $\hat{\sigma}$ into pen_{id} yields the well-known Mallows' C_p. Use the empirical variance on the first 10% of the signal as an estimator of σ^2.

Simulate two noisy piecewise constant signals with the function `ruptures.pw_constant` (set the dimension to $d = 2$) for each combination of parameters: `n_bkps`$\in \{2, 4, 6, 8, 10\}$, $T \in \{100, 200, 500, 1000\}$ and $\sigma \in \{1, 2, 5, 7\}$.

Using Mallows' C_p,

- for $\sigma = 2$ and $T \in \{100, 200, 500, 1000\}$, compute the Hamming metric between the true segmentation and the estimated segmentation and report the average on Figure 1-a;
- for $T = 500$ and $\sigma \in \{1, 2, 5, 7\}$, compute the Hamming metric between the true segmentation and the estimated segmentation and report the average on Figure 1-b.

Answer 2



(a) Hamming metric vs the number T of samples

(b) Hamming metric vs the standard deviation σ

Figure 1: Performance of Mallows' C_p

Question 3 *Slope heuristics*

The ideal penalty is of shape $\text{pen}(\boldsymbol{\tau}) = Cd|\boldsymbol{\tau}|/T$ where $C > 0$. The slope heuristics is a procedure to infer the best C without knowing σ.

Slope heuristics algorithm.

- Estimate the slope of $\hat{s}$ of $\min_{\boldsymbol{\tau},|\boldsymbol{\tau}|=K} \|\Pi_{\boldsymbol{\tau}} - y\|^2$ as a function of K for K "large enough". Define $\hat{C}_{\text{slope}} := -T\hat{s}$.
- Estimate $\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \|y - \Pi_{\boldsymbol{\tau}} y\|^2 / T + \hat{C}_{\text{slope}} d|\boldsymbol{\tau}|/T$.

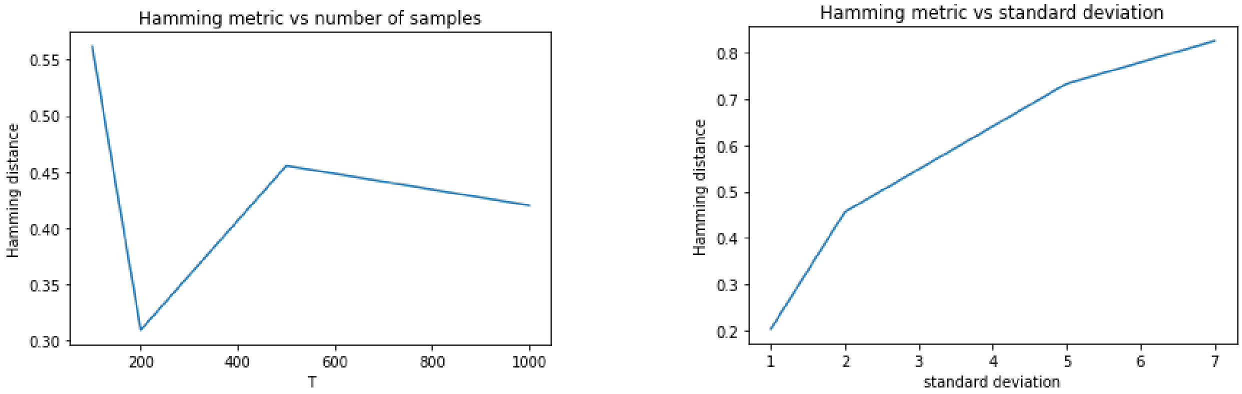
In simulations, "large enough" means for K between 15 and $0.4T$.

Simulate two noisy piecewise constant signals with the function `ruptures.pw_constant` (set the dimension to $d = 2$) for each combination of parameters: `n_bkps`$\in \{2, 4, 6, 8, 10\}$, $T \in \{100, 200, 500, 1000\}$ and $\sigma \in \{1, 2, 5, 7\}$.

Using the slope heuristics,

- for $\sigma = 2$, $T \in \{100, 200, 500, 1000\}$, compute the average Hamming metric between the true segmentations and the estimated segmentations and report the average on Figure 2-a;
- for $T = 500$ and $\sigma \in \{1, 2, 5, 7\}$, compute the average Hamming metric between the true segmentations and the estimated segmentations and report the average on Figure 2-b.

Answer 3



(a) Hamming metric vs the number T of samples

(b) Hamming metric vs the standard deviation σ

Figure 2: Performance of the slope heuristics

3 Wavelet transform for graph signals

Let G be a graph defined a set of n nodes V and a set of edges E. A specific node is denoted by v and a specific edge, by e. The eigenvalues and eigenvectors of the graph Laplacian L are $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $u_1, u_2, \ldots, u_n$ respectively.

For a signal $f \in \mathbb{R}^n$, the Graph Wavelet Transform (GWT) of f is $W_f : \{1, \ldots, M\} \times V \longrightarrow \mathbb{R}$:

$$W_f(m, v) := \sum_{l=1}^{n} \hat{g}_m(\lambda_l) \hat{f}_l u_l(v) \tag{8}$$

where $\hat{f} = [\hat{f}_1, \ldots, \hat{f}_n]$ is the Fourier transform of f and $\hat{g}_m$ are M kernel functions. The number M of scales is a user-defined parameter and is set to $M := 9$ in the following. Several designs are available for the $\hat{g}_m$; here, we use the Spectrum Adapted Graph Wavelets (SAGW). Formally, each kernel $\hat{g}_m$ is such that

$$\hat{g}_m(\lambda) := \hat{g}^U(\lambda - am) \quad (0 \leq \lambda \leq \lambda_n) \tag{9}$$

where $a := \lambda_n / (M + 1 - R)$,

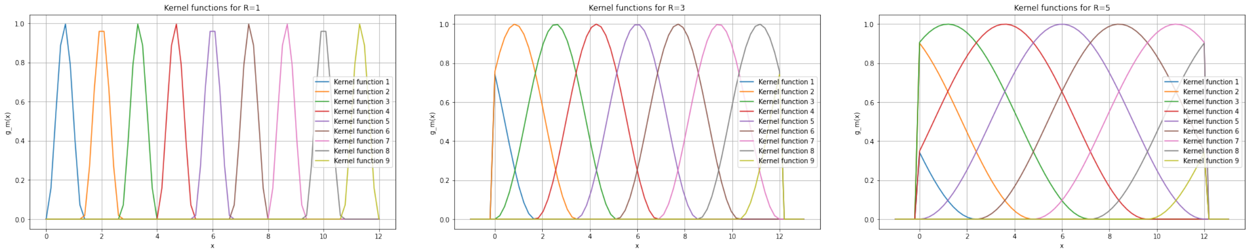
$$\hat{g}^U(\lambda) := \frac{1}{2} \left[1 + \cos\left(2\pi \left(\frac{\lambda}{aR} + \frac{1}{2} \right) \right) \right] \mathbb{1}(-Ra \leq \lambda < 0) \tag{10}$$

and $R > 0$ is defined by the user.

Question 4

Plot the kernel functions $\hat{g}_m$ for $R = 1$, $R = 3$ and $R = 5$ (take $\lambda_n = 12$) on Figure 3. What is the influence of R?

Answer 4



(a) $R = 1$ (b) $R = 3$ (c) $R = 5$

Figure 3: The SAGW kernels functions

We will study the Molene data set (the one we used in the last tutorial). The signal is the temperature.

Question 5

Construct the graph using the distance matrix and exponential smoothing (use the median heuristics for the bandwidth parameter).

- Remove all stations with missing values in the temperature.
- Choose the minimum threshold so that the network is connected and the average degree is at least 3.
- What is the time where the signal is the least smooth?
- What is the time where the signal is the smoothest?

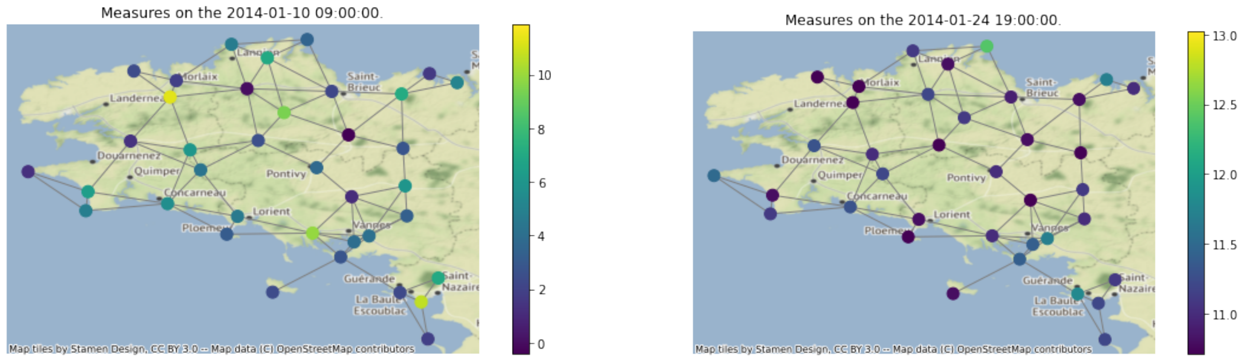
Answer 5

The stations with missing values are ['Arzal, Batz, Begmeil, Brest-guipavas, Brignogan, Camaret, Landivisiau, Lannaero, Lanveoc, Ouessant-stiff, Plouay-sa, Ploudalmezeau, Plougonvelin, Quimper, Riec sur belon, Sizun, St nazaire-montoir, Vannes-meucon'].

The threshold is equal to 0.83.

The signal is the least smooth at 9 am the 10/01/2014.

The signal is the smoothest at 7pm the 24/10/2014.



(a) Least smooth signal

(b) Smoothest signal

Figure 4: Stations with their temperature with respect to two time stamps corresponding to smoothest and least smooth signal. The bar color indicates temperatures.

Question 6

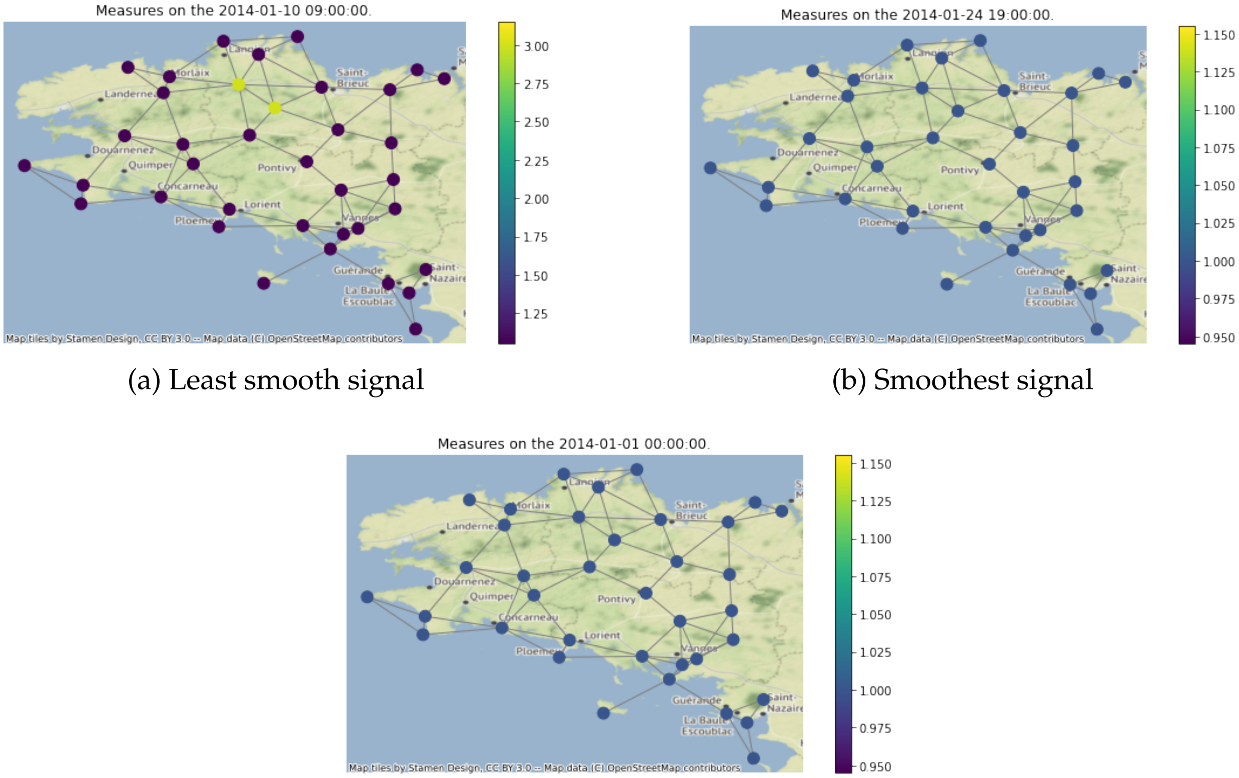
(For the remainder, set $R = 3$ for all wavelet transforms.)

For each node v, the vector $[W_f(1, v), W_f(2, v), \dots, W_f(M, v)]$ can be used as a vector of features. We can for instance classify nodes into low/medium/high frequency:

- a node is considered low frequency if the scales $m \in \{1, 2, 3\}$ contain most of the energy,
- a node is considered medium frequency if the scales $m \in \{4, 5, 6\}$ contain most of the energy,
- a node is considered high frequency if the scales $m \in \{6, 7, 9\}$ contain most of the energy.

For both signals from the previous question (smoothest and least smooth) as well as the first available timestamp, apply this procedure and display on the map the result (one colour per class).

Answer 6



Measures on the 2014-01-10 09:00:00.
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
Measures on the 2014-01-24 19:00:00.
1.150
1.125
1.100
1.075
1.050
1.025
1.000
0.975
0.950
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
Measures on the 2014-01-01 00:00:00.
1.150
1.125
1.100
1.075
1.050
1.025
1.000
0.975
0.950
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

(a) Least smooth signal

(b) Smoothest signal

(c) First available timestamp

Figure 5: Classification of nodes into low/medium/high frequency. Yellow (=3) is high frequency and dark blue (=1) is low frequency.

Question 7

Display the average temperature and for each timestamp, adapt the marker colour to the majority class present in the graph (see notebook for more details).

Answer 7

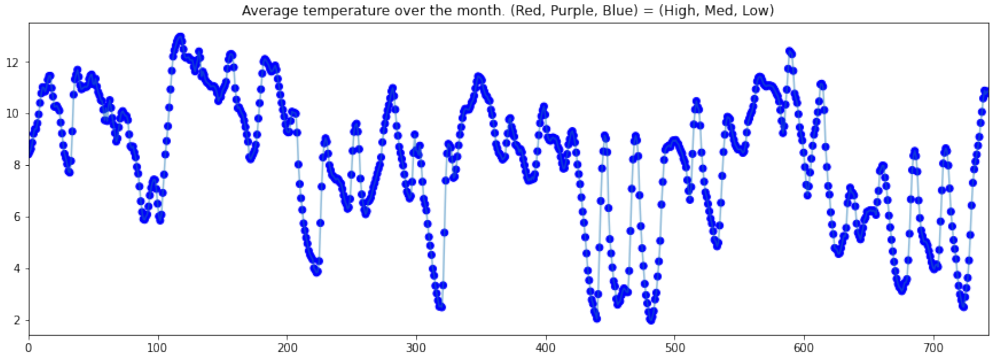


Figure 6: Average temperature. Markers' colours depend on the majority class.

Interpretation :

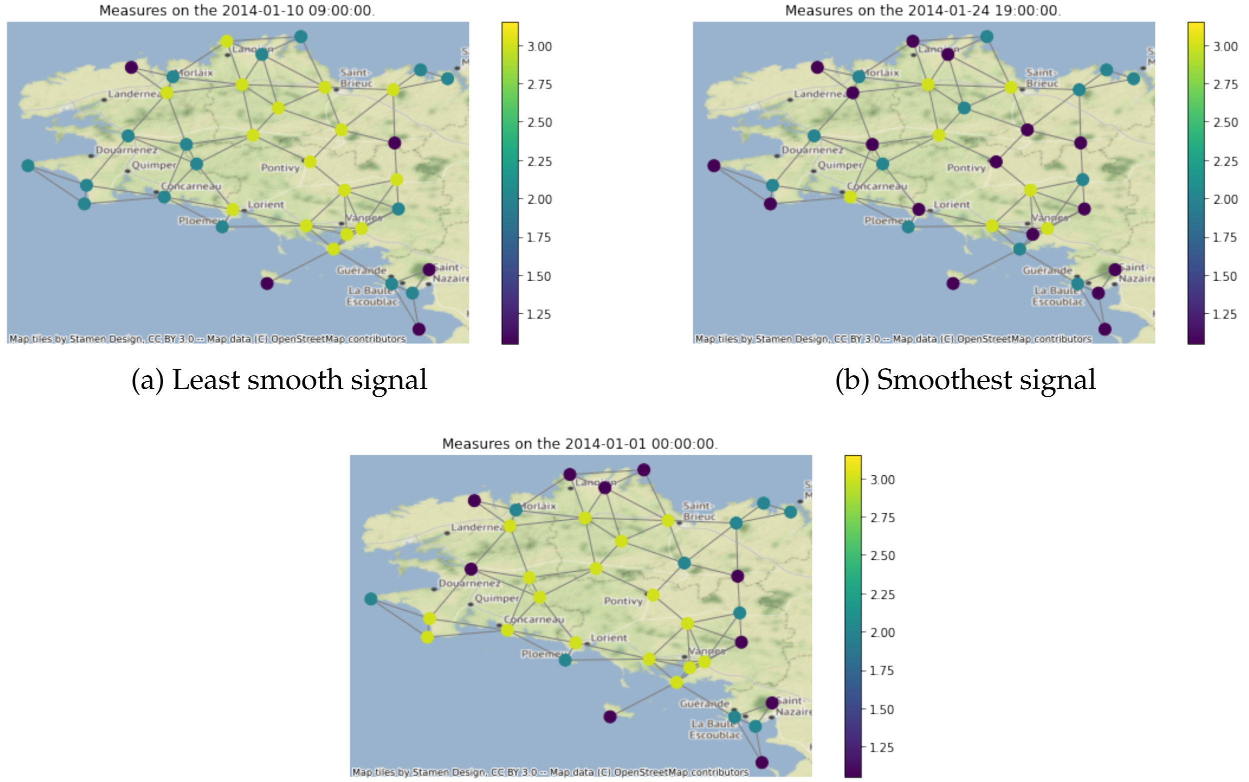
Let's try to explain why there are only low frequencies in this case. As we choose a threshold to connect our graph, the first eigenvector is the eigenvector of 0 is constant (only one connected component), and therefore does not provide any information on the nodes. Now let's look back at the formula defined by

$$W_f(m,v) := \sum_{l=1}^{n} \hat{g}_m(\lambda_l)\hat{f}_l u_l(v) \tag{11}$$

Looking at the plot for $R = 3$ question 4 we see that $\hat{g}_m$ has its weight centered on values close to $\frac{\lambda_n m}{M}$. Now for a signal f, $W_f(m,v) = \hat{g}_m(0)\hat{f}_1 C + \sum_{l=2}^{n} \hat{g}_m(\lambda_l)\hat{f}_l u_l(v)$ with $C = u_l(v)$ constant and $\hat{f}_1$ the average of f. Now looking at the plot 3 for $R = 3$ we see that $\hat{g}_m(0)$ is large for $m = 1$ and $m = 2$. As the nodes have to "share" the weight of u_l in the sum $\sum_{l=2}^{n} \hat{g}_m(\lambda_l)\hat{f}_l u_l(v)$ it is normal that no values of m will dominate others in this sum. Therefore the constant $\hat{g}_m(0)\hat{f}_1 C$ adding to the sum give an advantage to $m = 1$ and $m = 2$, and we have only low frequencies. What would be interesting now would be to remove the first eigenvector of 0 and define

$$W_f^*(m,v) = \sum_{l=2}^{n} \hat{g}_m(\lambda_l)\hat{f}_l u_l(v)$$

Let's do questions 6 and 7 with this approach :


Measures on the 2014-01-10 09:00:00.
Measures on the 2014-01-24 19:00:00.
Measures on the 2014-01-01 00:00:00.
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
Morlaix
Landerneau
Lannion
Saint-Brieuc
Douarnenez
Quimper
Concarneau
Pontivy
Lorient
Ploemeur
Vannes
Guérande
Saint-Nazaire
La Baule-Escoublac
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

(a) Least smooth signal

(b) Smoothest signal

(c) First available timestamp

Figure 7: Classification of nodes into low/medium/high frequency without first eigenvector. Yellow (=3) is high frequency and dark blue (=1) is low frequency.

As we can see the frequencies are much better balanced. Let's now look at question 7 and plot the frequencies against the temperatures and the graph-smoothness of each time stamp.

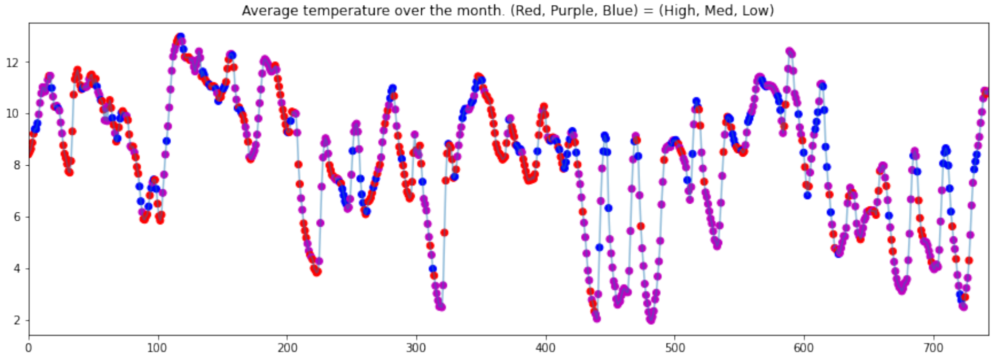


Figure 8: Average temperature. Markers' colours depend on the majority class computed with W^*.

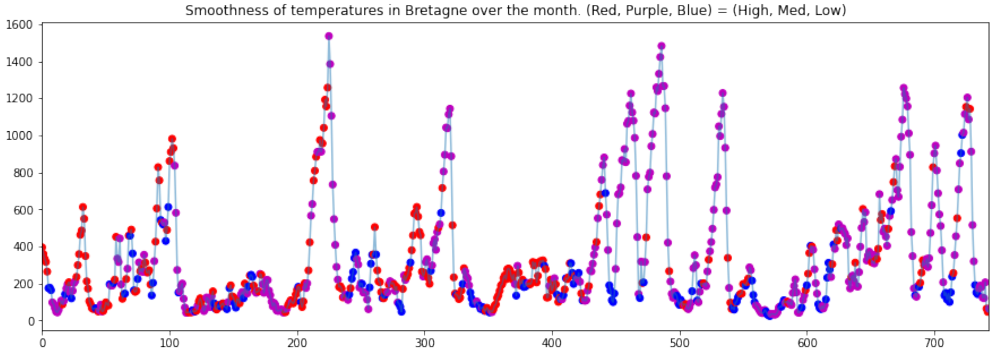


Figure 9: Average temperature. Markers' colours depend on the majority class computed with W^*.

Given these two figures we notice two things. First, figure 8 shows that high frequencies tend to happen when temperatures decrease. From the course we know that local phenomena are depicted by higher frequencies, meaning that a decrease of temperature is correlated with local phenomena. If we assume that gulf stream brings hot air and that high temperatures are correlated with Atlantic wind, an absence of wind could both explain a larger amount of local phenomena as the weather in each geographic region is subject to its characteristics in terms of movement of air, and a decrease of temperature. Second, Figure 9 shows that when the graph smoothness is high, the frequencies are either medium or high. This means that when temperatures are very disparate, local phenomena are more likely to be depicted by our graph.

Question 8

The previous graph G only uses spatial information. To take into account the temporal dynamic, we construct a larger graph H as follows: a node is now *a station at a particular time* and is connected to neighbouring stations (with respect to G) and to itself at the previous timestamp and the following timestamp. Notice that the new spatio-temporal graph H is the Cartesian product of the spatial graph G and the temporal graph G' (which is simply a line graph, without loop).

- Express the Laplacian of H using the Laplacian of G and G' (use Kronecker products).
- Express the eigenvalues and eigenvectors of the Laplacian of H using the eigenvalues and eigenvectors of the Laplacian of G and G'.
- Compute the wavelet transform of the temperature signal.
- Classify nodes into low/medium/high frequency and display the same figure as in the previous question.

Answer 8

We first need to choose a basis for the laplacian of H $L(H)$. We chose to enumerate the vertices from left to right of $L(H)$ this way : $x_1^1, \ldots, x_1^m, \ldots, x_n^1, \ldots, x_n^m$ with n the number of stations and m the number of time stamps. (We realized too late that this numbering was not the best suited for implementation.)

Spatial connections : The matrix is made of $n^2 blocks$ of size $m \times m$. Each block ij is equal to $L(G)_{ij} I_m$
Temporal connections : The matrix is made of $n^2 blocks$ of size $m \times m$. Each block ij is equal to $0_{m\times m}$ if $i \neq j$ and $L(G')$ if $i = j$.
Finally

$$L(H) = L(G) \bigotimes I_m + I_n \bigotimes L(G')$$

Then let e_i^G be the i-th eigenvector of G and λ_i^G its i-th eigenvalue and use the same notations for G'.

$$L(H) \cdot e_i^G \bigotimes e_j^{G'} = (L(G) \bigotimes I_m)(e_i^G \bigotimes e_j^{G'}) = (L(G) \cdot e_i^G) \bigotimes (I_m \cdot e_j^{G'}) + (I_n \cdot e_i^G) \bigotimes (L(G') \cdot e_j^{G'})$$

therefore

$$L(H) \cdot e_i^G \bigotimes e_j^{G'} = (\lambda_i^G + \lambda_j^{G'})(e_i^G \bigotimes e_j^{G'})$$

therefore eigenvalues and eigenvectors of $L(H)$ are $\{\lambda_i^G + \lambda_j^{G'}, e_i^G \otimes e_j^{G'}\}$ for $i \in [\![1, n]\!]$ and $j \in [\![1, m]\!]$.

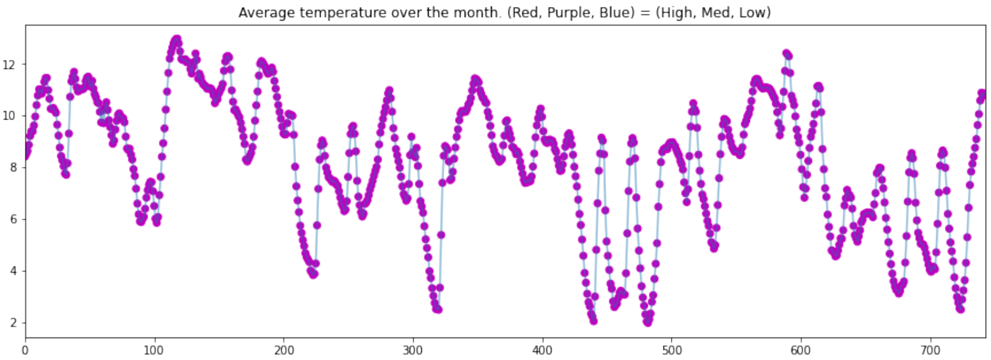


Figure 10: Average temperature. Markers' colours depend on the majority class.

In this case, the first term of the sum is less important as the sum has thousands of terms (compared to 37 in question 7) which might explain why there are no low frequencies.