École polytechnique

MAP565 - Modélisation Statistique

Mémoire de projet

Analyse de données climatiques sur différentes villes de France

Élèves
LACOMBE Armand
BROUX Lucas

30 mars 2018

Table des matières

1 Introduction

Le but de ce projet est d'appliquer les méthodes étudiées en cours de modélisation statistique (MAP565) à l'étude de l'évolution du climat sur plusieurs villes de France.

Cette analyse est motivée par plusieurs facteurs. D'une part, le réchauffement climatique est un phénomène avéré de ces dernières décennies ; il convient donc de le surveiller et l'analyser avec des méthodes de modélisation plus ou moins développées. D'autre part, certaines stations météorologiques relèvent des données de manière journalières depuis des dizaines d'années et distribuent leurs bases de données sur internet. Enfin, les méthodes vues en cours nous semblent particulièrement pertinentes dans ce cadre.

Nous abordons cette étude sous un angle triple. Dans un premier temps, nous emploierons une approche de type "série temporelle" afin de modéliser l'évolution de la température à Bordeaux, modélisation que nous testerons en estimant par prédiction les valeurs obtenues en 2017. Dans une deuxième partie, nous nous intéresserons à la statistique des extrêmes de cette même série temporelle, afin de proposer une analyse des risques de canicules. Enfin, nous utiliserons des méthodes de modélisation de dépendances afin de déterminer si les risques de canicules à Paris et à Bordeaux sont liés.

Nous utilisons les données fournies par le site internet `https://www.ecad.eu/`, que nous avons traitées et converties en format .csv afin de pouvoir les étudier. Nous implémentons nos scripts en Python, et utilisons pour cela les *packages* suivants :

-
-

Toutes les données, ainsi que les scripts utilisés, sont disponibles sur le repository github du projet : `https://github.com/lucas-broux/Projet-Map565`.

2 Analyse de la série temporelle

Étudions dans un premier temps les données selon une approche "série temporelle".

2.1 Méthodologie

2.2 Visualisation des données

2.3 Identification de trend

2.4 Modélisation de la série

2.5 Prédictions

3 Statistique des extrêmes

Nous souhaitons désormais estimer les quantiles extrêmes de la distribution de températures, notre objectif étant de déterminer le "risque de canicule" entre le 15 juillet et le 15 aout à Bordeaux.

3.1 Méthodologie

Nous conservons la même base de données que dans la partie précédente, mais nous considérons dans cette partie les mesures de températures maximales journalières prises entre le 15 juillet et le 15 août de chaque année, afin d'effacer le caractère saisonnier mis en évidence dans la première partie.

Nous supposons ainsi que ces données correspondent à n observations i.i.d. $X_1, ..., X_n$ d'une loi $\mathbb{P}$ inconnue. L'objectif est d'estimer, pour $\alpha \in [0, 1]$ proche de 1, le quantile d'ordre α de cette loi. Cela nous donnera la valeur de température qui ne sera pas dépassée - avec un niveau de confiance α. Notre mesure de "risque de canicule" sera alors la valeur de α pour laquelle cette température maximale est 35 °C.

Pour cela, comme mis en évidence dans le cours, nous ne considérons pas de méthodes de type paramétriques ou de quantiles empiriques, mais préférons une approche par domaine d'attraction. Nous supposons ainsi que $X_1, ..., X_n$ sont dans le domaine d'attraction d'une certaine loi max-stable, ce qui d'après le cours implique l'existence de $\xi \in \mathbb{R}$ caractérisant cette loi max-stable sous la forme

$$H_\xi = \begin{cases} e^{-(1+\xi x)^{-\frac{1}{\xi}}} & \text{si } \xi \neq 0 \\ e^{-e^{-x}} & \text{si } \xi = 0 \end{cases}$$

La première chose est de déterminer une estimation du paramètre ξ, ce que nous faisons dans la section suivante. Nous tâcherons ensuite de proposer une estimation du quantile désiré.

En revanche, il est difficile de chercher à vérifier les résultats, sachant que nous ne connaissons pas la loi exacte de X_1. Il est donc difficile de proposer un test statistique de vérification, et les résultats que nous obtenons restent spéculatifs.

3.2 Estimation de ξ

Au lieu d'appliquer aveuglément des calculs d'estimateurs à nos données, étudions-les. Nous pouvons représenter graphiquement les valeurs de températures :

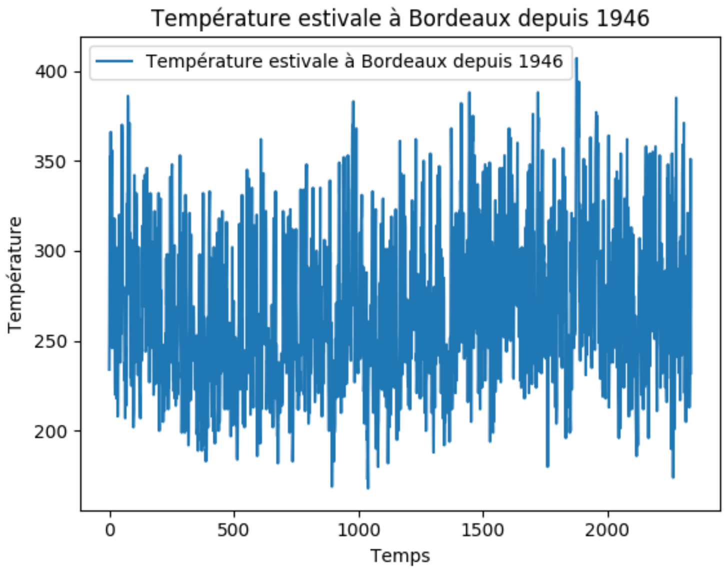

FIGURE 1 – Données

Nous constatons que les données présentent une grande variance, et qu'il y a peu d'événements extrêmes. Pour confirmer cette impression, nous traçons le diagramme quantile-quantile des données, comparant la distribution de celles-ci avec celle de la loi normale :

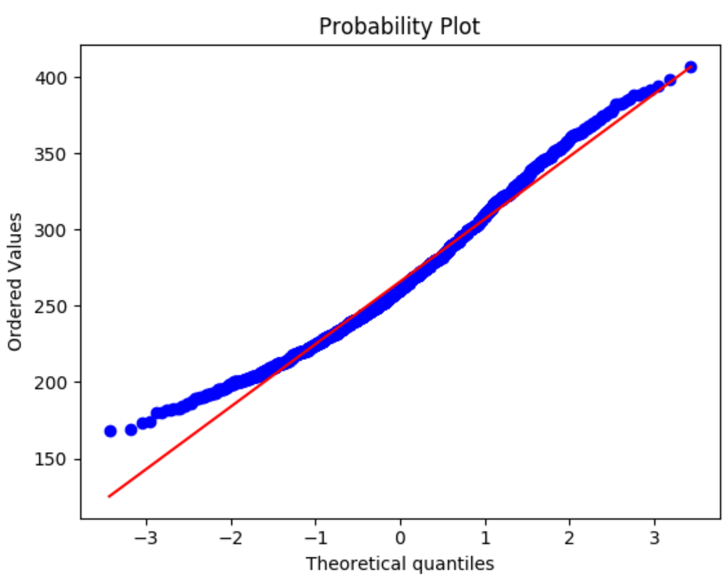

FIGURE 2 – Diagramme quantile-quantile

Nous remarquons une asymétrie des queues de distribution : la queue de distribution correspondant aux valeurs extrêmes négatives est épaisse, mais celle correspondant aux valeurs extrêmes positives est relativement fine. Ainsi, les événements extrêmes les plus récurrents sont ceux de faibles températures et non de forte températures. Or notre problématique est celle des canicules, ce qui signifie que la loi que nous considérons n'est pas une loi à queue de distribution forte. Nous nous attendons donc heuristiquement à une valeur de ξ négative, puisque nous savons que les lois H_ξ pour $\xi > 0$ correspondent sont *heavy-tailed*.

NB : Nous constatons un phénomène symétrique lorsque nous étudions les températures minimales en hiver : dans ce cas ce sont les occurrences de températures chaudes qui sont plus récurrentes que les températures froides.

Nous devons donc adapter les méthodes du cours puisque celles-ci correspondaient à $\xi > 0$. Notamment, nous ne pouvons pas utiliser l'estimateur de Hill.

Comme mentionné dans le cours, nous pouvons en revanche dans cette situation utiliser l'estimateur de Pickands :

$$\hat{\xi}^P_{n,k(n)} = \frac{1}{\log(2)} \log\left(\frac{X_{(n-k(n)+1,n)} - X_{(n-2k(n)+1,n)}}{X_{(n-2k(n)+1,n)} - X_{(n-4k(n)+1,n)}}\right) \qquad \text{(Estimateur de Pickands)}$$

où $X_{(i,n)}$ correspond à la i-ème plus grande valeur parmi tous les X_j.

Des résultats théoriques montrent que l'estimateur converge en probabilités vers la vraie valeur sous les conditions

$$\begin{cases} k(n) & \xrightarrow[n\to+\infty]{} +\infty \\ \frac{k(n)}{n} & \xrightarrow[n\to+\infty]{} 0 \end{cases}$$

Il s'agit donc de trouver un compromis entre la valeur de k et celle de n. En pratique, nous traçons le graphe de $\hat{\xi}_{n,k(n)}$ en fonction de k (la valeur de n est fixée et correspond au nombre d'observations), et nous cherchons une "zone de stabilité" correspondant à la valeur estimée de ξ. Nous obtenons le graphe suivant :

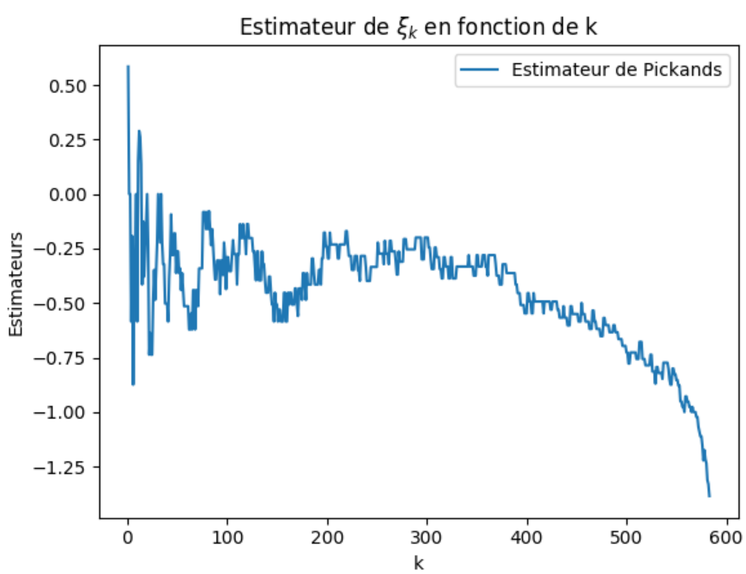

FIGURE 3 – Diagramme quantile-quantile

Nous pouvons constater que comme prédit par l'analyse heuristique précédente des données, cette méthode propose un estimateur négatif. Nous observons un domaine de stabilité pour $k \in [200; 400]$ et nous pouvons choisir pour valeur de ξ l'estimation associée :

$$\hat{\xi} = -0.3$$

Notons qu'en vertu des résultats du cours, le fait que ξ soit négatif implique l'existence d'une constante x_F telle que pour $x \geq x_F$, $\mathbb{P}(X \geq x) = 1$. L'interprétation physique du phénomène semble confirmer ce phénomène, mais l'existence d'un tel x_F rend difficile et moins pertinentes les analyses de risque.

Nous ne pouvons donc pas exploiter le fait que - selon un théorème vu en cours - il existe une fonction L à variations lentes telle que la fonction de répartition voulue s'écrive

$$\bar{F}\left(x_F - \frac{1}{x}\right) = x^{\frac{1}{\xi}} L(x)$$

puisque nous ne savons pas estimer x_F. Nous allons devoir employer une autre approche.

3.3 Méthode *peak over threshold*

Rappelons l'algorithme de la méthode *peak over threshold* présentée durant le cours.

Dans un premier temps, nous estimons $u > 0$ tel que la fonction empirique e_n définie par

$$e_n(u) := \frac{1}{N_u} \sum_{i, X_i > x} (X_i - x)$$

(ù $N_u := card\{i \in [1; n], X_i > u\}$), soit à peu près linéaire.

Nous notons ensuite $Y_i := X_i - u$ les excès, et nous calculons (numériquement) $\hat{\xi} \in \mathbb{R}$ et $\hat{\beta} > 0$ maximisant le maximum de vraisemblance

$$L = -n \log(\beta) - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{n} \log\left(1 + \frac{\xi Y_i}{\beta}\right)$$

Alors pour tout $y > 0$, nous pouvons estimer la fonction de répartition recherchée par

$$\hat{\bar{F}}(u + y) = \frac{N_u}{n} \left(1 + \frac{\hat{\xi} y}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}$$

4 Modélisation des dépendances

5 Conclusion