École polytechnique

MAP565 - Modélisation Statistique

Mémoire de projet

Analyse de données climatiques sur différentes villes de France

Élèves
LACOMBE Armand
BROUX Lucas

29 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 ne considérons que les données de températures 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 30 °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

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

Nous utilisons et comparons deux estimateurs de ξ : l'estimateur de Hill et l'estimateur de Pickands, puis utilisons la valeur obtenue pour calculer les quantiles voulus en vertu des résultats du cours que nous rappellerons.

Nous implémentons en outre la méthode dite "*peak over threshold*" et comparons les résultats obtenus avec les précédentes méthodes.

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 Estimateurs de Hill et Pickands

Nous estimons le paramètre ξ selon les formules suivantes :

$$\hat{\xi}^H_{n,k(n)} = \frac{\sum_{i=n-k(n)+1}^{n} \left(\log\left(X_{(i,n)}\right) - \log\left(X_{(n-k(n)+1,n)}\right)\right)}{k\left(n\right)} \qquad \text{(Estimateur de Hill)}$$

et

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

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

Dans les deux cas, 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.

3.3 Méthode *peak over threshold*

4 Modélisation des dépendances

5 Conclusion