École polytechnique

MAP565 - Modélisation Statistique

Mémoire de projet

Analyse de données climatiques

Élèves
LACOMBE Armand
BROUX Lucas

30 mars 2018

Table des matières

1 Introduction **2**

2 Analyse de la série temporelle **3**
2.1 Décomposition de la série 3
2.2 Interrogation sur la période de la série 4
2.3 Validité du modèle additif 4
2.4 Modélisation 5
2.5 Ordre de différentiation 5
2.6 Saisonnalité 6
2.7 Étude des ordres p et q 6

3 Statistique des extrêmes **7**
3.1 Méthodologie 7
3.2 Estimation de ξ 7
3.3 Méthode *peak over threshold* 10

4 Modélisation des dépendances **13**
4.1 Étude de la *cross-correlation* 13
4.2 Estimation d'une copule 14

5 Conclusion **16**

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

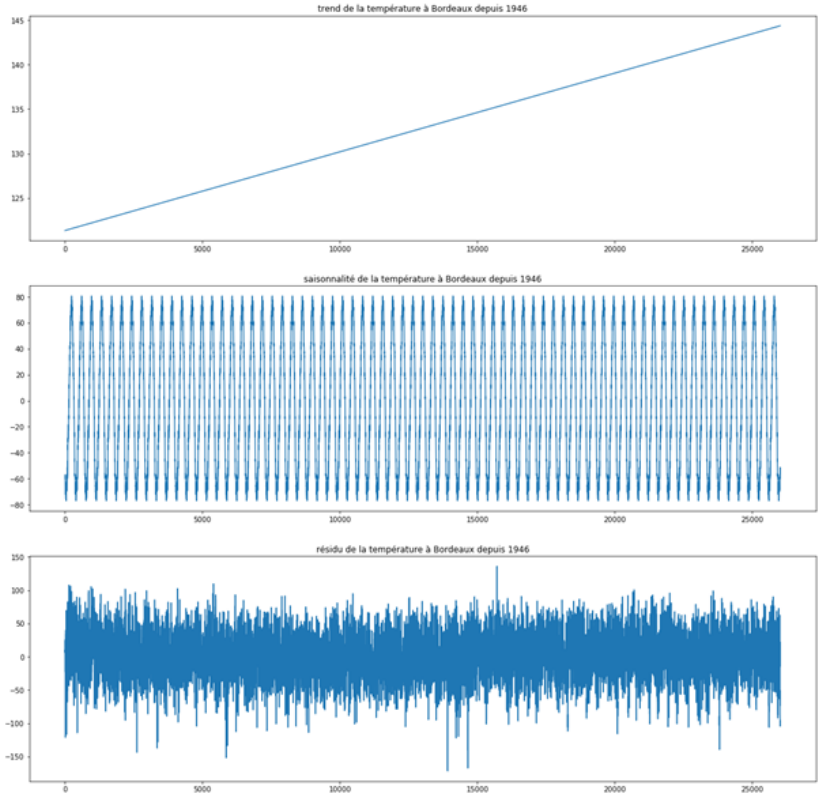
2 Analyse de la série temporelle

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

2.1 Décomposition de la série

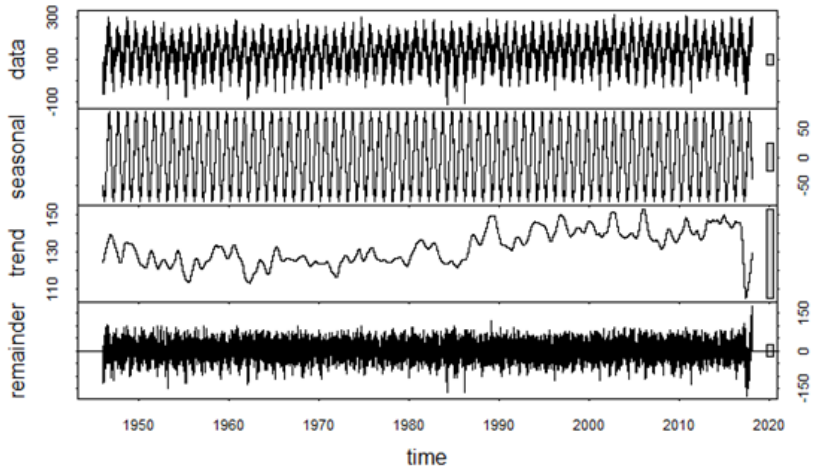
On commence par décomposer la série selon un trend et une saisonnalité. Pour cela, plusieurs méthodes s'offrent à nous. Nous choisirons dans un premier temps la méthode de Buys-Ballot généralisée, qui consiste à écrire chacune des composantes déterministes comme une combinaison linéaire de fonctions connues du temps.

On obtient alors la décomposition suivante :



La variance des résidus est élevée en comparaison de la différence du trend entre le début et la fin de la série temporelle ; les données sont fortement bruitées.

On aurait également pu envisager une décomposition de la forme Seasonal decomposition of Time series by Loess, ce qui est aisé avec R. On choisit une fenêtre de la largeur d'une période et l'on obtient la décomposition suivante :



On obtient pour cette décomposition un écart type de 33,7, à comparer avec le 33,5 de la méthode de Buys-Ballot.

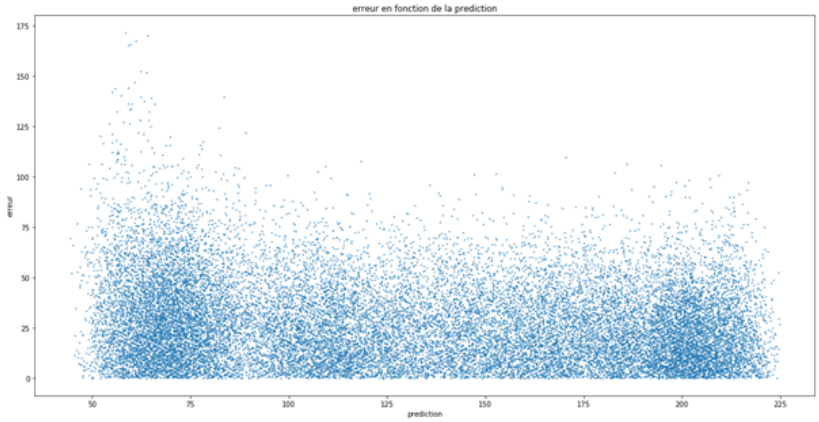
2.2 Interrogation sur la période de la série

On s'interroge alors sur un point. On sait qu'une année dure 365,24 jours et non pas 365 comme on en fait souvent l'approximation. On a ici pris en compte le décalage dans la formule des $\hat{S}$, mais qu'en serait-il si on l'omettait ?

En refaisant une simulation qui fait cette approximation, on obtient une variance de 33,8 sur les résidus, contre 33,5 dans le cas plus rigoureux. Par la suite on se permettra de désigner la période par 365 ou 365,24 compte tenu de la proximité des deux valeurs.

2.3 Validité du modèle additif

On a jusque alors envisagé un modèle de type additif pour la série temporelle : la valeur temporelle s'écrirait comme la somme d'une tendance, d'une saisonnalité et d'une erreur que l'on suppose modélisable par un bruit gaussien faible. On va chercher à vérifier cette hypothèse d'additivité en traçant la courbe de l'erreur obtenue (en valeur absolue) en fonction de la valeur prédite par Buys-Ballot généralisé.



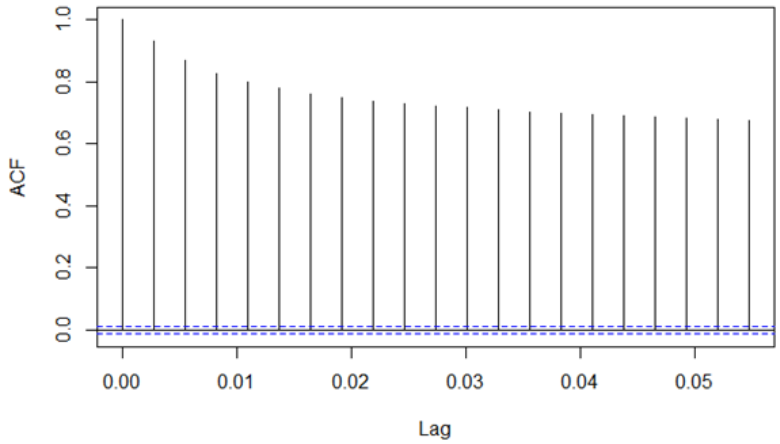
On ne distingue pas de relation nette entre la variance du bruit et la grandeur de la prédiction ; aussi le modèle additif sera-t-il considéré comme pertinent par la suite.

2.4 Modélisation

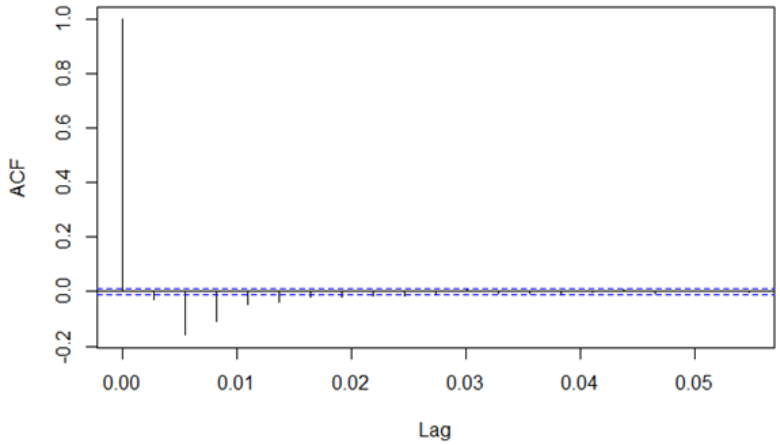
La présence d'une tendance et d'une saisonnalité nous invite à envisager la modélisation de la série temporelle par un processus de type SARIMA.

2.5 Ordre de différentiation

Par acquis de conscience, examinons la fonction d'autocorrélation afin de déterminer l'ordre de différentiation requis.



Comme présumé, le processus est non-stationnaire. On va alors tracer l'autocorrélogramme de la série temporelle différentiée.



Cet autocorrélogramme nous invite à considérer qu'un ordre de différentiation 1 est suffisant pour le modèle SARIMA, ce qui concorde avec les décompositions que l'on a obtenu précédemment (les restes ne semblaient pas présenter de tendance quadratique.)

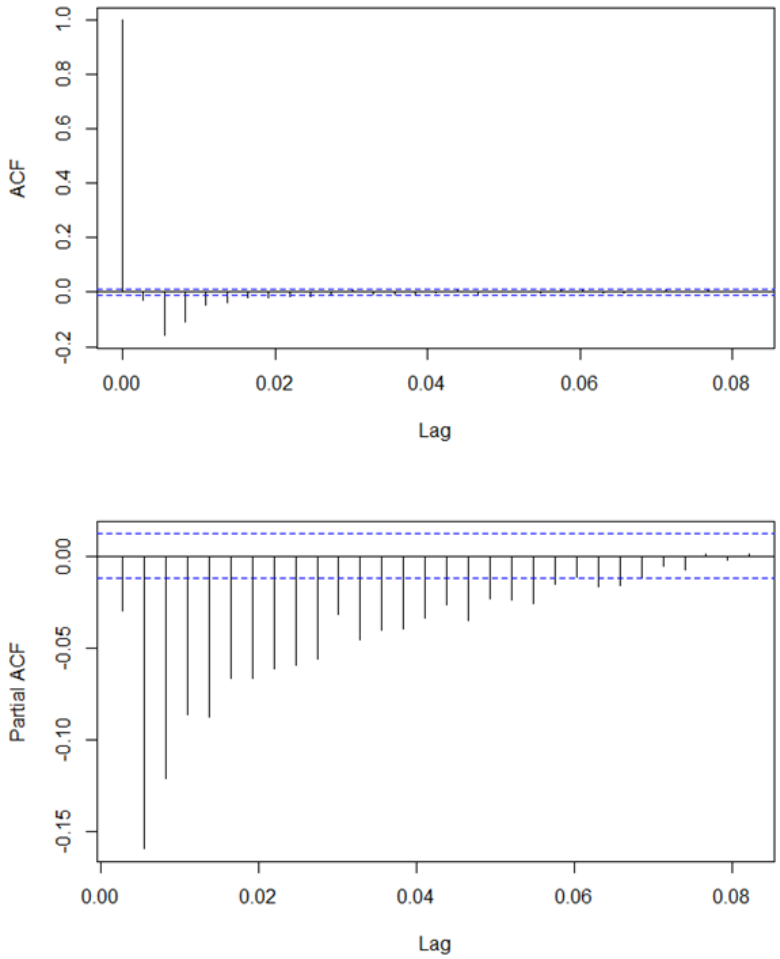
2.6 Saisonnalité

Les études précédentes et le bon sens climatique nous invitent à considérer une saisonnalité d'ordre 365.

2.7 Étude des ordres p et q

Un processus AR d'ordre p se caractérise par sa fonction d'autocorrélation partielle qui s'annule à partir de l'ordre $p+1$, et un processus MA d'ordre q se caractérise par sa fonction d'autocorrélation qui s'annule à partir de l'ordre $q+1$.

Traçons donc les autocorrélogramme et autocorrélogramme partiel de la série différentiée.



On peut dès lors envisager de choisir $q = 6$. En revanche le choix de p est problématique puisque la décroissance est très lente et les calculs sont trop lourds si l'on choisit $p = 21$ par exemple.

On peut ainsi considérer un modèle SARIMA365(26,1,6)(0,1,0). Néanmoins le temps de calcul est trop grand pour de telles valeurs, sans compter un risque important d'overfitting. On pressent que le tracé de autocorrélogramme partiel est pathologique, et un traitement plus important des données aurait été nécessaire afin d'obtenir une modélisation satisfaisante.

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 faisons l'approximation - peut-être grossière - que ces valeurs sont indépendantes. Rappelons que les températures sont fournies sous l'unité 10 * °C.

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 :

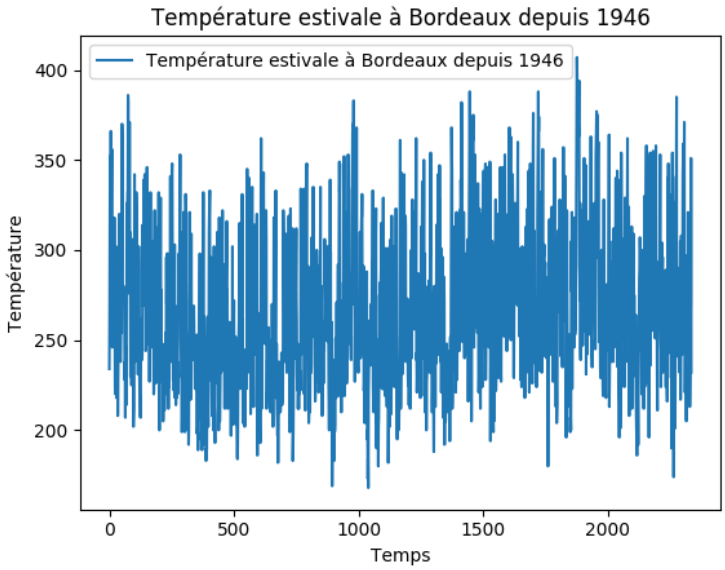


FIGURE 1 – Données

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 :

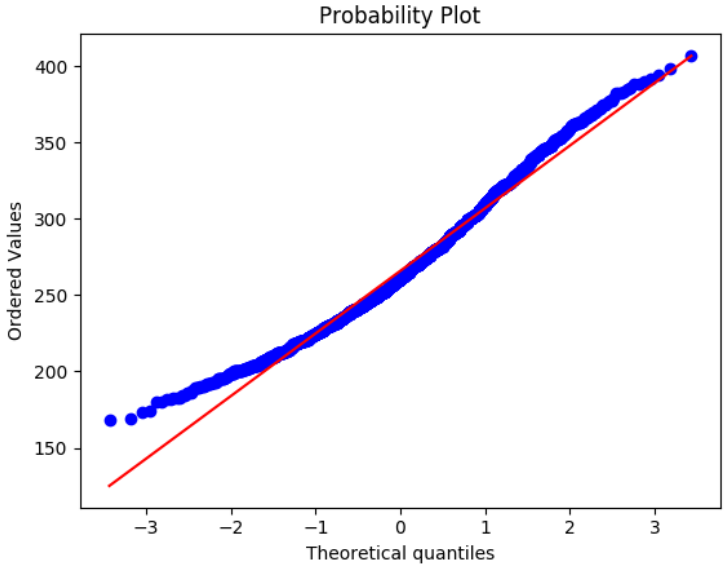


FIGURE 2 – Diagramme quantile-quantile

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 :

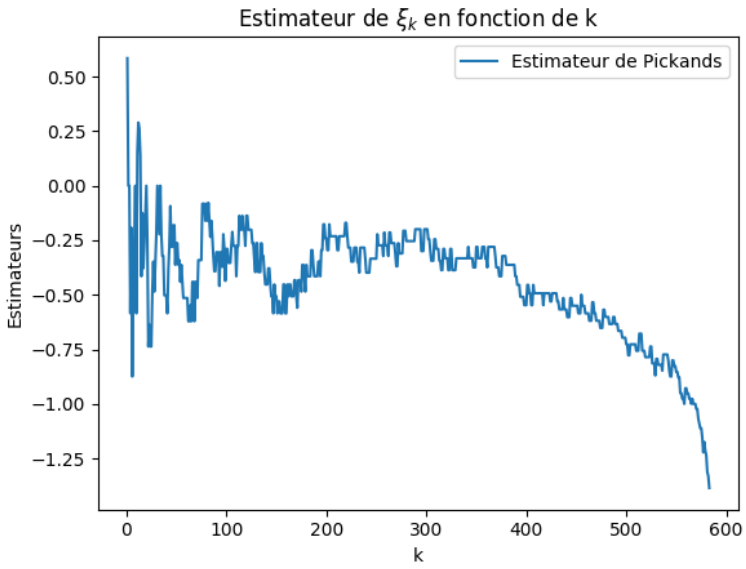


FIGURE 3 – Diagramme quantile-quantile

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 de la canicule pourrait 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

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

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

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

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

Nous traçons d'abord le graphe de $e_n(x)$ en fonction de la température x et obtenons :

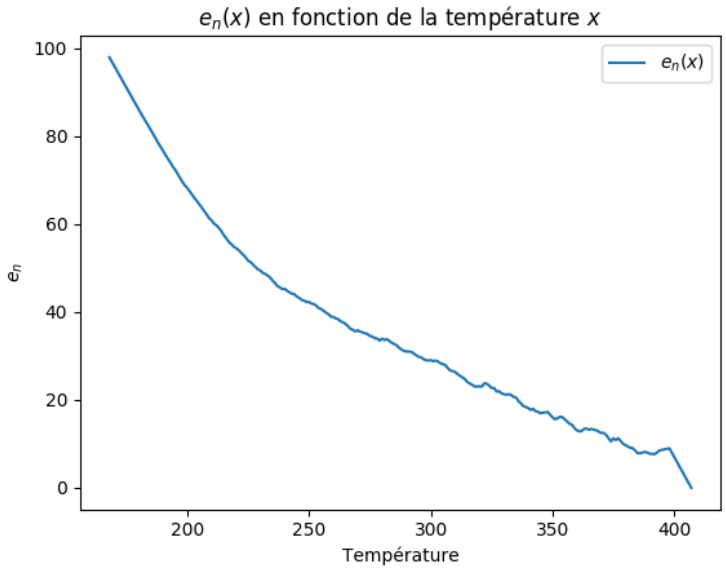


FIGURE 4 – Graphe de e_n

Nous observons que la croissance semble être quadratique dans un premier temps, puis linéaire à partir de $u = 225$. Nous conservons cette valeur de u et réalisons numériquement l'optimisation de la vraisemblance. Nous trouvons

$$\begin{cases} \hat{\xi} & = -0.764 \\ \hat{\beta} & = 150 \end{cases}$$

Cela semble bien confirmer le phénomène établi dans la section précédente d'une fine queue de distribution.

Estimant ξ et β par régression linéaire sur la fonction $e_n(u) \simeq \frac{\beta+\xi u}{1-\xi}$, nous obtenons des résultats ayant le même ordre de grandeur :

$$\begin{cases} \hat{\xi} & = -0.376 \\ \hat{\beta} & = 152 \end{cases}$$

Nous pouvons dès lors tracer la fonction de répartition estimée selon les deux méthodes, que nous comparons avec la fonction de répartition empirique :

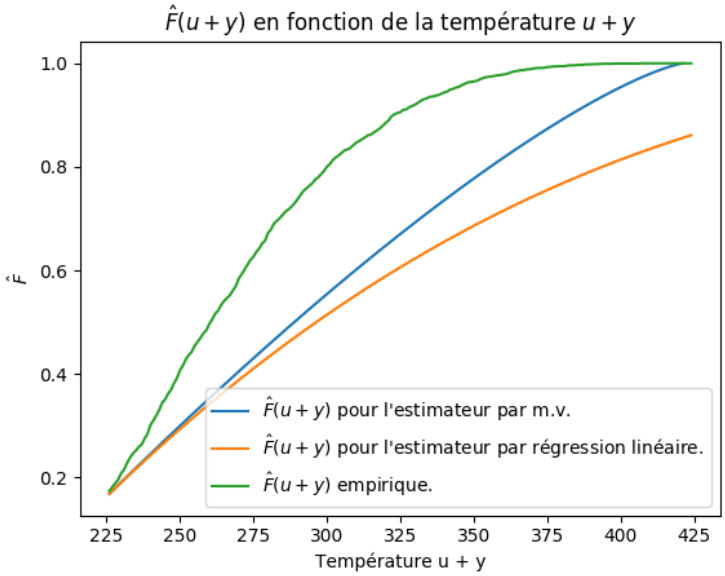


FIGURE 5 – Fonctions de répartition

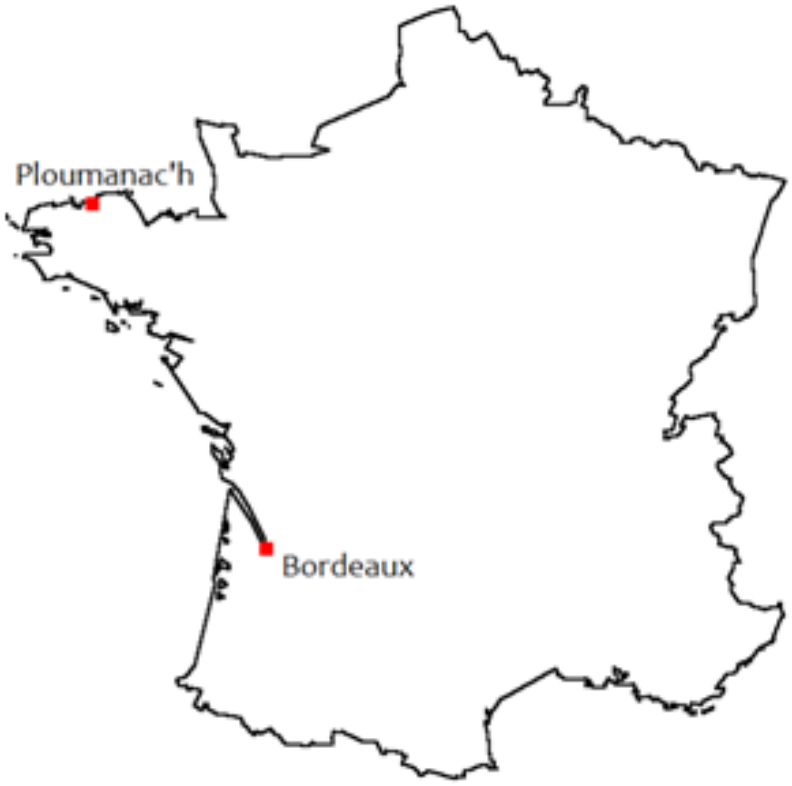
Malheureusement, les résultats ne sont pas très pertinents, car les fonctions estimées n'approximent pas très bien la fonction de répartition empirique, même si ils restent relativement consistants avec le fait que la queue de distribution observée est fine. Réalisant l'application numérique, nous trouvons

$$\mathbb{P}\left(X \geq 35\,^{\circ}\mathrm{C}\right) = \begin{cases} 0.78 & \text{pour l'estimateur du maximum de vraisemblance} \\ 0.69 & \text{pour l'estimateur par régression linéaire} \\ 0.96 & \text{pour l'estimateur empirique} \end{cases}$$

Ainsi les modèles proposés, même si ils respectent le fait que la queue de distribution soit fine, ne permettent pas d'obtenir des résultats précis pour des événements rares. On peut en revanche imaginer qu'ils sont plus pertinents pour les événements très rares car on constate que l'estimateur du maximum de vraisemblance semble devenir proche de l'estimateur empirique pour des températures supérieures à 41 °C.

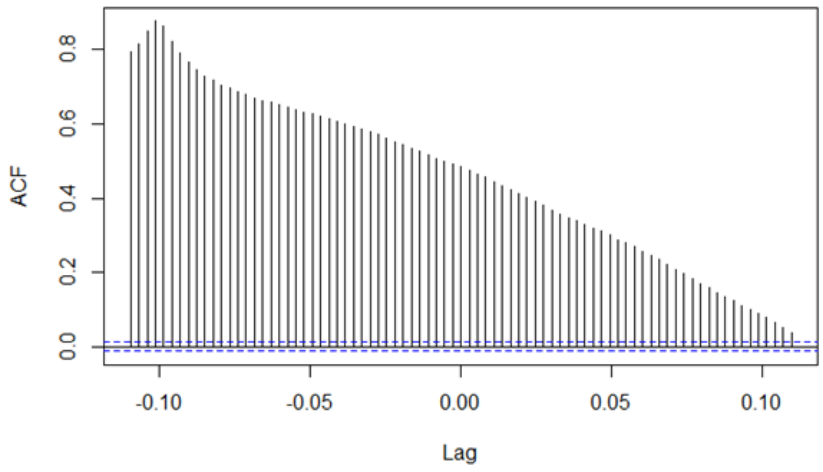
4 Modélisation des dépendances

Dans une dernière partie, on va s'attacher à étudier les corrélations entre les séries temporelles de température à Bordeaux et à Ploumanac'h.



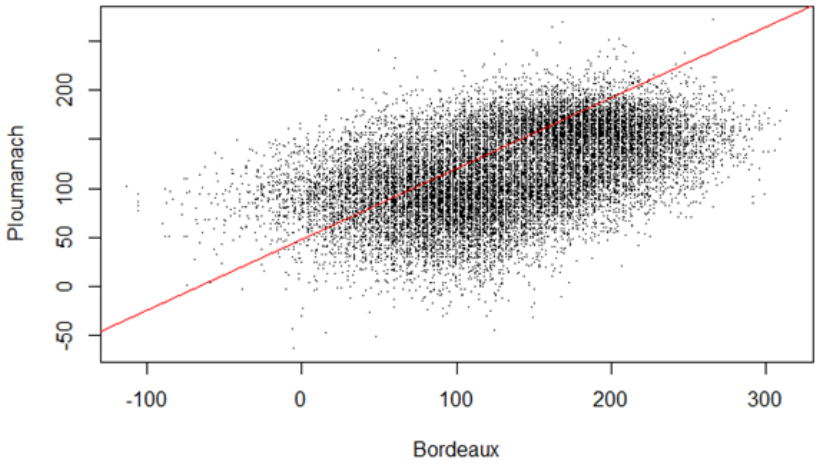
4.1 Étude de la *cross-correlation*

Le tracé de la *cross-correlation* entre les deux séries temporelles révèle une corrélation forte (coefficient de 0.88) avec un maximum atteint pour un décalage de trois jours.



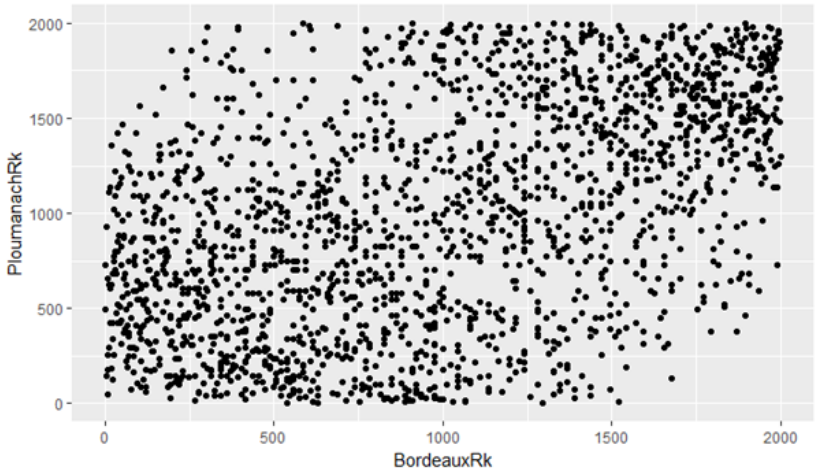
4.2 Estimation d'une copule

On trace tout d'abord le graphe des températures obtenues à Ploumanac'h en fonction des températures relevées à Bordeaux.



La corrélation entre les deux séries est sensible et fait écho au sens physique. La corrélation de Spearman est de 0.51, celle de Kandall de 0,34, confirmant que les deux séries sont corrélées. On note à ce sujet des p-values très basses qui sont gages d'un résultat robuste : elles sont inférieures à 2e-16 dans les deux cas.

On peut tracer le graphe des rangs des températures à Ploumanac'h en fonction des rangs des températures à Bordeaux pour se figurer l'allure de la relation des deux séries temporelles. On ne considère ici que 2000 points par souci de lisibilité.

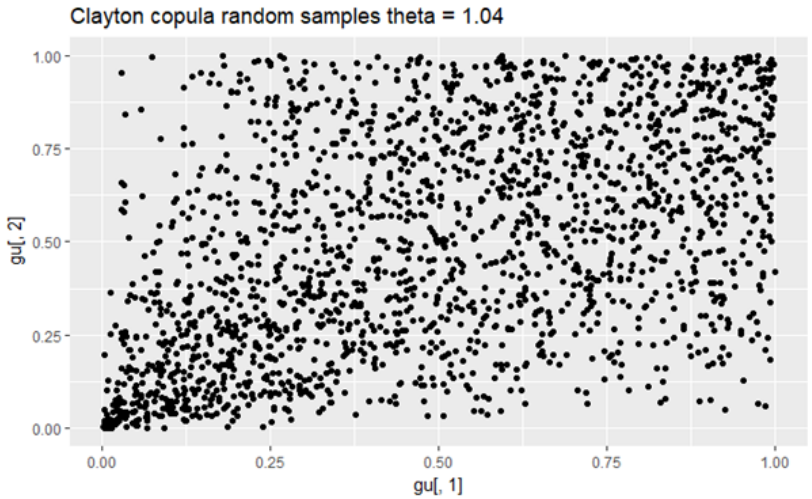


Nous allons à présent choisir une copule pour modéliser la corrélation entre les deux grandeurs. On a vu que les événements extrêmes ne sont pas fortement corrélés, aussi on privilégie la copule de Clayton qui découple plus les canicules dans un premier temps.

On sait que pour elle en dimension 2, le tau de Kandall vaut theta/(theta+2). On en déduit la valeur du paramètre : ici ce sera 1,04.

On en déduit une copule qui modélise la corrélation des deux séries temporelles :

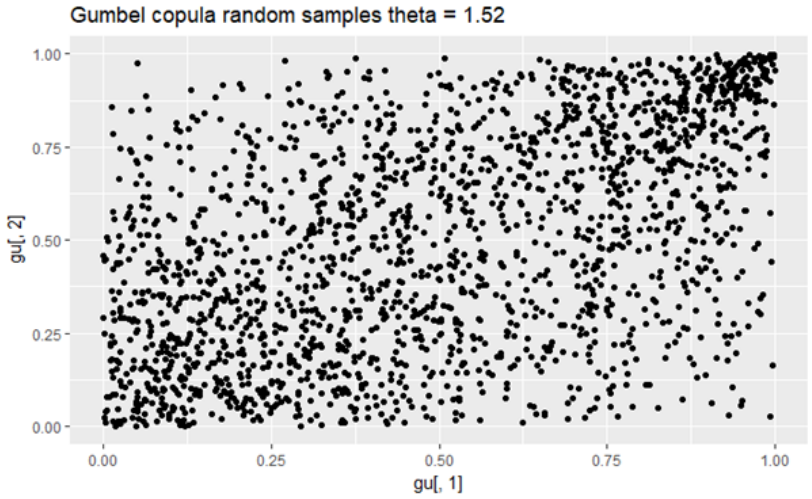
$$\bar{F}C_{\theta}^{C_l}(U_1, U_2) = \Big(U_1^{(} - \theta) + U_2^{(} - \theta) - 1\Big)^{(} \frac{1}{\theta})$$



Dans un second temps, on aborde la copule de Gumbel. On sait que pour elle en dimension 2, le tau de Kandall vaut 1 – 1/theta. On en déduit la valeur de theta dans notre cas : on obtient une valeur de 1,52.

On en déduit une copule qui modélise la corrélation des deux séries temporelles :

$$\bar{F}C_{\theta}^{G_u}(U_1, U_2) = e^{-(-\log(U_1)^{\theta} - \log(U_2)^{\theta})^{(}\frac{1}{\theta})}$$



Ces tracés révèlent les limites des modélisations choisies, pour lesquelles les corrélations des événements extrêmes sont visiblement trop importantes.

5 Conclusion

Ainsi, nous avons tâché de répondre à plusieurs problématiques climatiques en appliquant différentes méthodes de modélisation statistique vues en cours.

Dans un premier temps, nous avons essayé de proposer une modélisation de la série temporelle des températures moyennes journalières observées à Bordeaux depuis 1940.

Nous nous sommes ensuite intéressés à l'analyse des comportements extrêmes des températures. Pour cela, nous avons considéré les températures maximales journalières observées à Bordeaux depuis 1940 en cherchant à étudier le comportement des températures hautes i.e. des risques de canicule. Contrairement à notre intuition originale, nous avons constaté que la distribution des températures hautes présentait une queue relativement fine. Les estimateurs proposés ont permis de confirmer numériquement ce phénomène notamment en estimant un ξ négatif tel que la distribution appartienne à une loi max-stable H_ξ. En revanche, nous avons constaté que les méthodes présentées en cours sont moins pertinentes pour ce genre de distribution car aucune des méthodes appliquées n'a permis d'établir un estimateur conforme aux données. Plusieurs facteurs ont peut-être influencé ces phénomènes :

- D'une part, nous avons supposé les données indépendantes, or les relevés d'une journée influencent potentiellement les suivants.
- D'autre part, certains phénomènes de stabilité numérique dans le calcul d'optimisation par maximum de vraisemblance peuvent fausser les estimations.

Enfin, nous avons cherché à étudier les dépendances entre les relevés de températures de différentes villes relativement éloignées.

Ce projet nous a permis de mettre en oeuvre les méthodes travaillées durant le cours, et a été l'occasion de nous mesurer à l'étude de données réelles. Notamment, nous nous sommes rendus compte à quel point il est important de ne pas appliquer aveuglément les formules théoriques sans étudier le comportement empirique et heuristique des données. Nous avons ainsi pu constater la nécessité de penser de manière critique les méthodes employées et les estimateurs calculés : le processus de modélisation statistique a été itératif, et il est illusoire de vouloir chercher à l'automatiser ou de se contenter des premiers résultats.