MAP565 - Rapport de projet

Éloise Berthier, Clément Mantoux

Mars 2018

Introduction

Nous avons choisi d'étudier en premier lieu un jeu de données issu d'un partenariat entre la mairie de Paris et l'entreprise Cisco : durant plusieurs mois, entre 2016 et 2017, des capteurs placés à divers endroits de la place de la Nation ont récolté des données sur la fréquentation de la place et sur l'environnement alentour. Ces données ont fourni un support intéressant pour l'application des notions vues en cours sur les copules et la théorie des valeurs extrêmes.

Toutefois ces données, bien que chronologiques, sont trop lacunaires et s'étalent sur une période de temps trop courte pour fournir un matériau intéressant à l'application des séries temporelles. Nous nous sommes donc tournés vers l'étude d'autres données : la température mensuelle moyenne aux États-Unis, depuis plus d'un siècle.

1 Données de la place de la Nation

1.1 Pré-traitement des données

On dispose des données suivantes, mesurées par intervalles de dix minutes :
— Flux piétons
— Flux véhicules
— Nombre de connexions au wifi public
— Valeur moyenne du bruit sur 10 minutes
— Pression ambiante
— Température ambiante

Les données sont recueillies par différents capteurs répartis sur la place, et chaque capteur effectue une mesure toutes les dix minutes. Il se trouve cependant que certains capteurs, en raison de leur position ou d'un dysfonctionnement, produisent des données dont la distribution ne ressemble pas à celle des autres capteurs. Ces mesures ne représentant pas les phénomènes que nous souhaitons observer, nous avons donc tout d'abord retiré les mesures provenant de capteurs défaillants. Sur la figure 1, on voit par exemple que le capteur 004 comporte une part significative de points aberrants (températures de -150 $^\circ$).

Par ailleurs, les capteurs ne sont pas tous en fonctionnement aux mêmes moments, et il manque des données sur certains capteurs. Aussi, pour régulariser les distributions de valeurs des capteurs, nous avons agrégé les mesures de différents capteurs correspondant aux mêmes intervalles de temps, puis regroupé ces résultats par heures. Une fois ce nettoyage terminé, on obtient par exemple la figure 2 pour la température agrégée par heures.

1.2 Copules

1.2.1 Distribution bivariée empirique

On se propose d'étudier la structure de dépendance entre les différentes mesures recueillies par les capteurs sur la place de la Nation. Ainsi, on pourra déterminer s'il existe une dépendance entre la météo, les

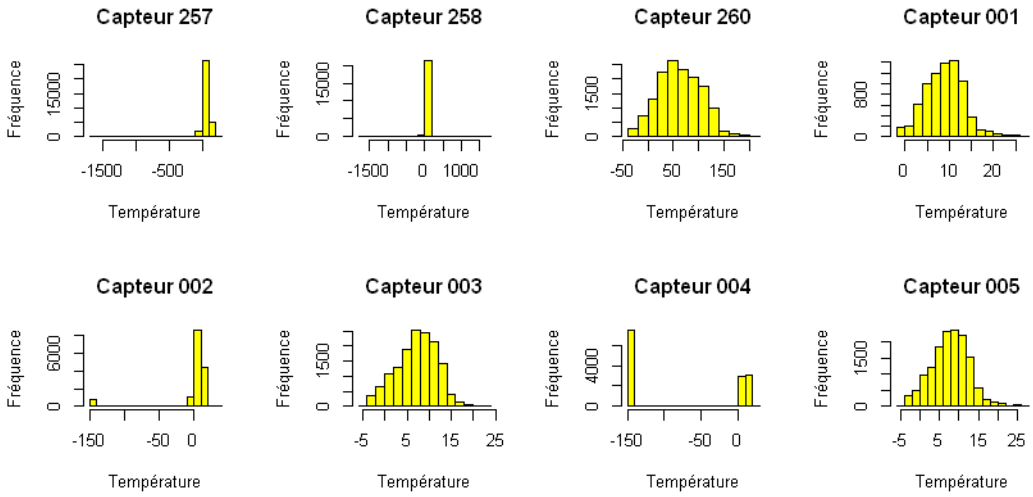


FIGURE 1 – La température selon différents capteurs

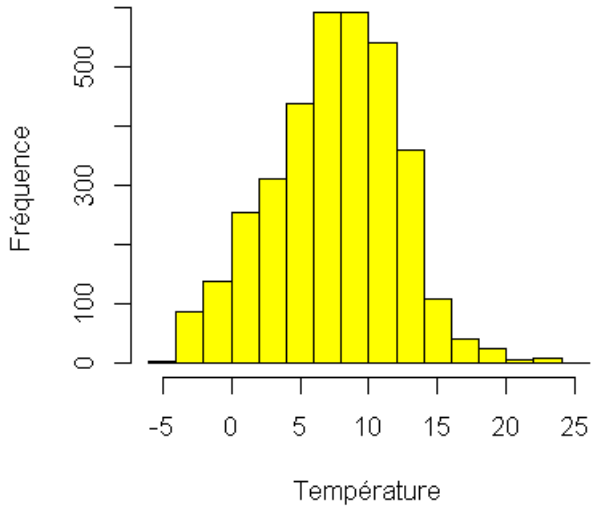


FIGURE 2 – La température filtrée et agrégée par capteurs et par heures

différents flux et le bruit dans un même lieu.

Afin de visualiser les copules, on calcule les rangs de chaque observation pour obtenir la fonction de répartition empirique $\hat{F}_n$ de chaque type de données, approximation de la fonction de répartition F. On a en effet, pour n observations $X_1, ..., X_n$ de l'une des séries de données :

$$\hat{F}_n(x) = \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}$$

Et notamment, en notant $X_{1:n}, ..., X_{n:n}$ la statistique d'ordre associée, on obtient

$$\hat{F}_n(X_{k:n}) = \frac{k}{n}$$

De plus, pour les variables $(X_i)_{1 \leq i \leq n}$, on a $F(X_i) \sim \text{Unif}[0,1]$. La copule de deux observations (bruit et pression par exemple) est alors définie comme la loi du couple $(F_{\text{bruit}}(X_{\text{bruit}}), F_{\text{pression}}(X_{\text{pression}}))$.

Afin d'étudier les copules bivariées, on trace les points $(F_{\text{bruit}}(X^i_{\text{s1}}), F_{\text{pression}}(X^j_{\text{s2}}))$ où s1 et s2 parcourent les types de données. On obtient le résultat de la figure 3. On remarque tout d'abord que la pression et la température (correspondant aux deux dernières lignes de la figure) sont indépendantes des autres données. Pour s'en assurer, on effectue un test du chi-deux d'indépendance entre la température, et les données de fréquentation. On obtient pour les flux piétons une p-valeur de $0.24 > 0.05$, et des valeurs similaires pour les flux véhicules, le wifi et le bruit : on ne peut donc pas rejeter l'hypothèse d'indépendance. On obtient le même résultat pour la pression.

1.2.2 Mesures de dépendance en dimension 2

On s'intéresse maintenant plus spécifiquement aux données directement liées à la fréquentation de la place : les flux piétons, les flux des véhicules, les connexions au wifi et le bruit moyen. Comme le montre la figure 3, ces variables sont fortement corrélées positivement, ce qui est cohérent. Ce lien se retrouve dans les coefficients de corrélation linéaire et le τ de Kendall :

	Piéton	Véhicule	Wifi	Bruit
Piéton	-	0.69	0.67	0.68
Véhicule	0.69	-	0.63	0.66
Wifi	0.67	0.63	-	0.63
Bruit	0.68	0.66	0.63	-

FIGURE 4 – Valeurs du τ de Kendall

Observons à présent les dépendances aux extrêmes. On distingue deux types de comportements : certains croisements, comme par exemple (flux piétons, bruit), sont corrélés positivement, mais sont faiblement dépendants aux extrêmes. D'autres, comme (flux véhicules, bruit) sont fortement corrélés aux extrêmes inférieurs, et plus indépendants pour les extrêmes supérieurs. Dans le premier cas, on observe des grandeurs qui ont comme seul lien l'activité dans l'ensemble sur la place. Dans le second cas, un lien direct de cause à effet fait que, en l'absence de l'une des grandeurs, l'autre est très faible : ainsi, il est rare d'avoir beaucoup de bruit sans aucune voiture.

Ces deux types de comportement nous amènent à nous intéresser aux copules de Clayton et de Frank. Ces copules sont définies comme suit :

$$C^{\text{Cl}}_\theta(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$$
$$C^{\text{Fr}}_\alpha(u,v) = -\frac{1}{\alpha} \ln\left(1 + \frac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^\alpha - 1}\right)$$

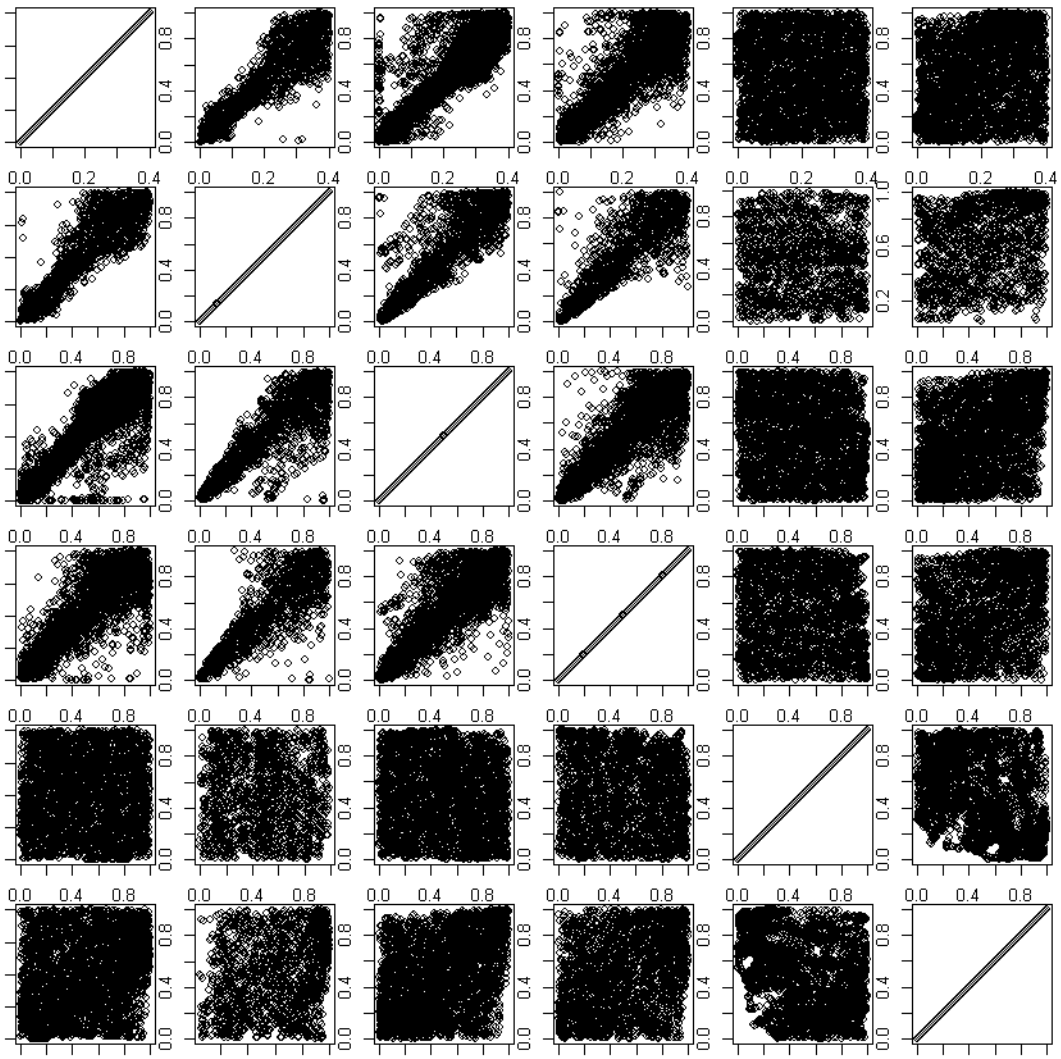
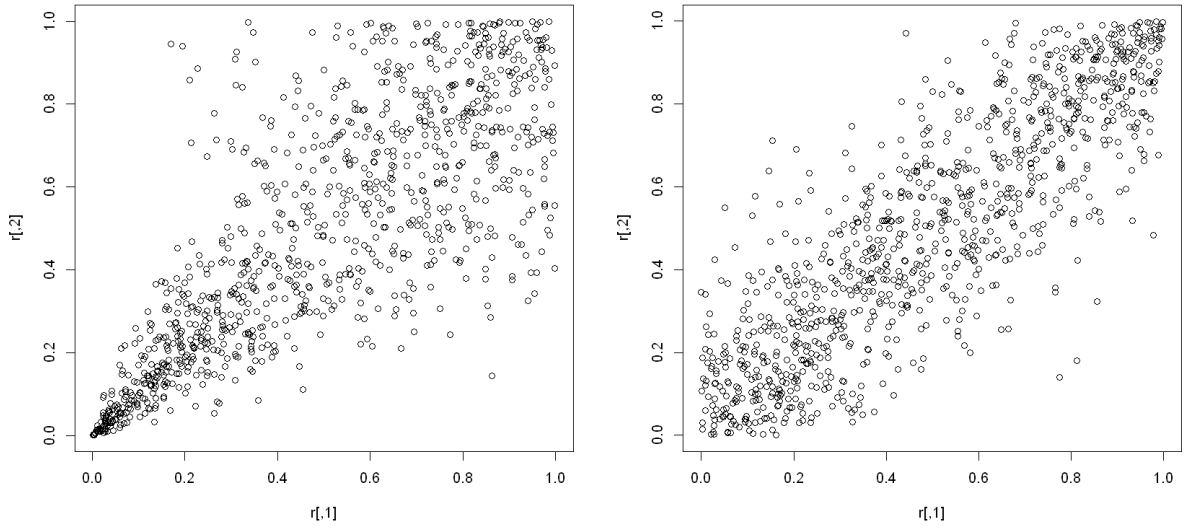


FIGURE 3 – Distributions croisées. Dans l'ordre : flux piétons, flux véhicules, wifi, bruit, pression et température



(a) Copule de Clayton ($\theta = 3$) (b) Copule de Frank ($\alpha = 10$)

FIGURE 5 – Échantillons de taille 1000 pour les copules de Clayton et de Frank

Pour un échantillon de 1000 points tirés selon ces copules, on obtient les distributions de la figure 5. Notamment, pour une copule de Clayton, on a vu en cours les coefficients de dépendance extrêmes :

$$\begin{cases} \lambda_l = 2^{-1/\theta} \\ \lambda_u = 0 \end{cases}$$

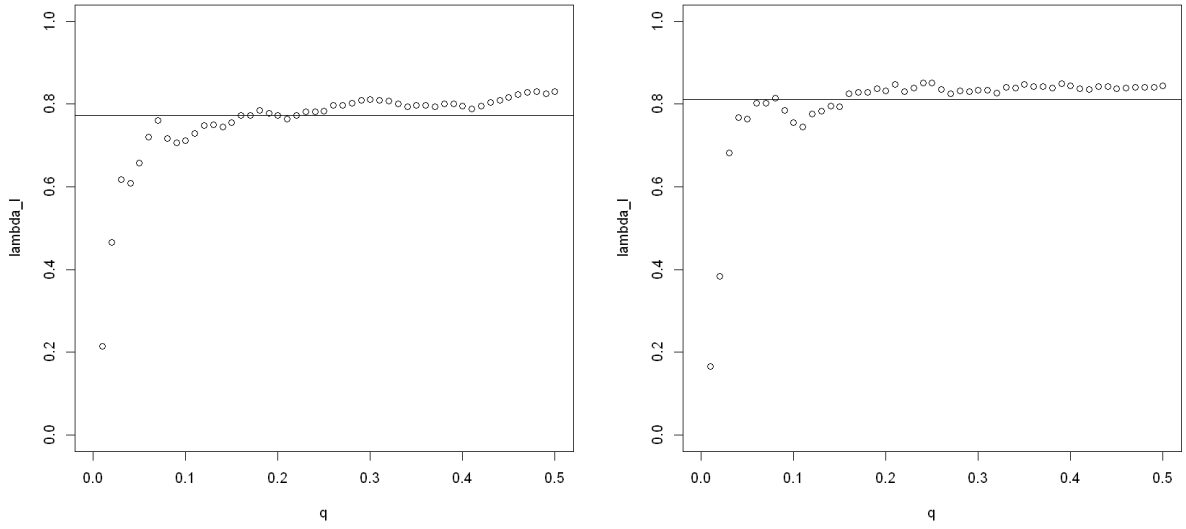
Pour la copule de Frank, on peut utiliser la relation $\lambda_l = \lim_{u \to 0^+} \dfrac{C(u,u)}{u}$:

$$\begin{aligned} \lambda_l &= \lim_{u \to 0^+} -\frac{1}{\alpha u} \ln\left(1 + \frac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^{\alpha} - 1}\right) \\ &= \lim_{u \to 0^+} -\frac{1}{\alpha u} \ln\left(1 + \frac{(\alpha u + O(u^2))(\alpha u + O(u^2))}{e^{\alpha} - 1}\right) \\ &= \lim_{u \to 0^+} -\frac{1}{\alpha u} \ln\left(1 + \frac{u^2 + O(u^3)}{e^{\alpha} - 1}\right) \\ &= \lim_{u \to 0^+} -\frac{1}{\alpha u} \frac{u^2 + O(u^3)}{e^{\alpha} - 1} \\ &\xrightarrow{u \to 0^+} 0 \end{aligned}$$

On obtient bien un coefficient de dépendance inférieur nul, ce qui modélise la différence que l'on souhaite faire avec la copule de Clayton. Les copules de Clayton et de Frank sont donc naturellement adaptées pour la modélisation des données de fréquentation de la place de la Nation. Il reste maintenant à déterminer leurs paramètres.

1.2.3 Sélection de modèle

Il s'agit donc maintenant de déterminer, pour chaque paire de données, si la dépendance se modélise par une copule de Clayton ou une copule de Frank, et avec quels paramètres. Pour ce faire, on utilise le package R VineCopula, qui dispose d'une fonction BiCopSelect qui, à partir d'un ensemble de types de



(a) Copule (Flux véhicules, Wifi)

(b) Copule (Flux véhicules, Bruit)

FIGURE 7 – Évolution de $\hat{\lambda}_l$ en fonction de q. Les lignes horizontales représentent les valeurs théoriques des copules de Clayton obtenues par la formule $\lambda_l = 2^{-1/\theta}$.

copules, détermine par maximum de vraisemblance la famille de copule la mieux adaptée, et le paramètre correspondant. L'exécution de cette fonction sur chaque paire de données produit le résultat suivant :

	Piéton	Véhicule	Wifi	Bruit
Piéton	-	Fr(10.99)	Fr(10.45)	Fr(10.79)
Véhicule	Fr(10.99)	-	Cl(2.79)	Cl(3.3)
Wifi	Fr(10.45)	Cl(2.79)	-	Fr(8.81)
Bruit	Fr(10.79)	Cl(3.3)	Fr(8.81)	-

FIGURE 6 – Copules sélectionnées par BiCopSelect

1.2.4 Étude de la dépendance extrême

Afin de valider notre modèle, on peut à présent se demander si l'on retrouve dans la pratique le résultat théorique pour la copule de Clayton : $\lambda_l = 2^{-1/\theta}$. Ce coefficient est défini, pour X^1 et X^2 deux variables aléatoires, par la quantité :

$$\lambda_l = \lim_{q \to 0^+} \mathbb{P}(X^2 \leq F_2^t(q) \mid X^1 \leq F_1^t(q))$$

Où F_1^t et F_2^t sont les inverses généralisées des fonctions de répartition de X^1 et X^2. On peut calculer empiriquement cette probabilité par l'estimateur

$$\hat{\lambda}_l(q) = \frac{|\{i \mid \hat{F}_n^1(X_i^1) \leq q\} \cap \{i \mid \hat{F}_n^2(X_i^2) \leq q\}|}{|\{i \mid \hat{F}_n^1(X_i^1) \leq q\}|}$$

Autrement dit, on calcule de nombre de couple où les deux conditions sont vérifiées divisé par le nombre de couples où seule la première condition est vérifiée. On espère ainsi observer, quand q se rapproche de 0, une convergence vers le coefficient théorique paramétré par θ. On obtient ainsi la figure 7.

On constate que $\hat{\lambda}_l$ semble effectivement converger vers sa valeur théorique. Mais, quand q est trop proche de zéro, le nombre de points impliqué dans l'estimation devient trop réduit, et la qualité de l'estimation se

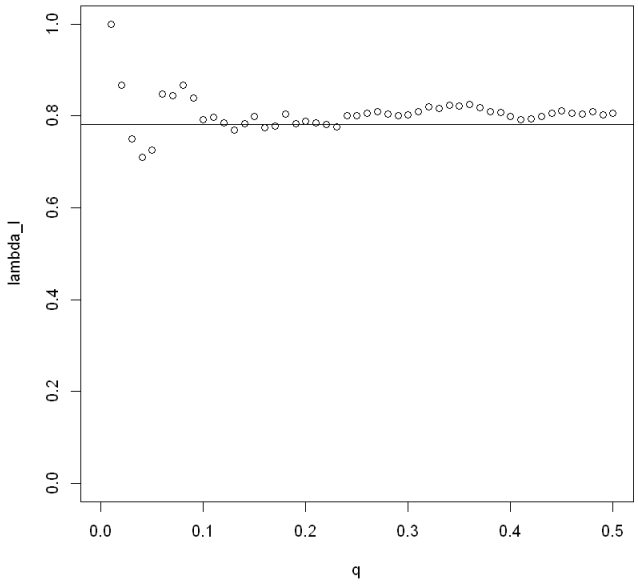


FIGURE 8 – Convergence du coefficient de dépendance extrême empirique de Clayton

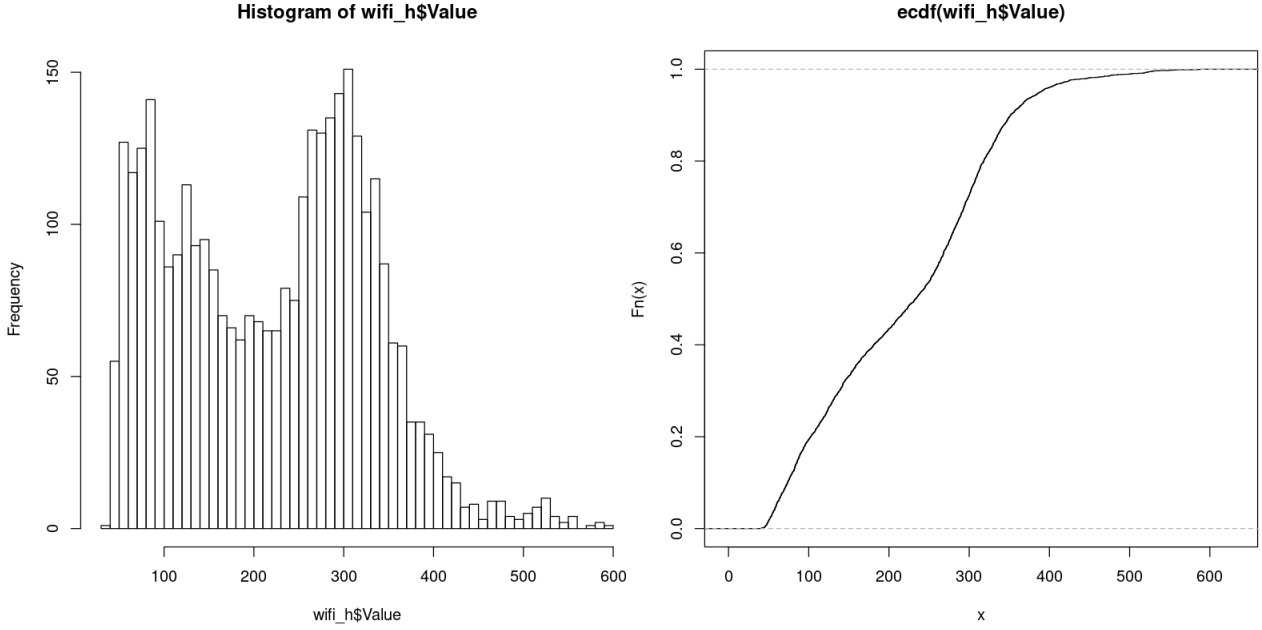
dégrade. A titre de comparaison, nous avons également tracé le même graphique pour un échantillon tiré simulé à partir d'une "vraie" copule de Clayton de paramètre 2.8. On peut observer, sur la figure 8, que la convergence se dégrade de manière similaire à l'approche de zéro, à partir de $q \simeq 0.1$.

Ce résultat confirme que le modèle de la copule de Clayton est adapté à la description de ces phénomènes. (Nous avons effectué la même expérience pour le coefficient de dépendance extrême supérieur, et obtenu chaque fois la convergence vers zéro de l'estimateur empirique).

1.3 Théorie des valeurs extrêmes

On cherche ici à observer des données susceptibles de représenter des phénomènes d'événements extrêmes. Pour cela, nous choisissons d'utiliser des données ayant une distribution à "queue lourde". C'est le cas notamment (visuellement) des relevés horaires de logs au service de wifi de la place de la Nation.

Notre problématique est la suivante : nous souhaitons savoir si le nombre de logs horaires sur le service wifi peut excéder un certain seuil, et estimer le risque d'un tel événement.



(a) Histogramme des logs wifi agrégés par heures

(b) Fonction de répartition empirique

FIGURE 9 – Exploration visuelle des données

Sur la fonction de répartition empirique, on voit que pour x au delà d'un certain seuil (pour $x \geq 350$), celle-ci peut s'approximer par une fonction polynomiale. On peut donc conjecturer que la fonction de répartition F appartient au domaine d'attraction de Fréchet, c'est-à-dire que $\overline{F} = x^{-\alpha}L(x)$, pour $x > 0$, avec $\alpha = \dfrac{1}{\xi} > 0$ et L une fonction à variations lentes.

On va essayer de confirmer ou d'infirmer cette hypothèse en utilisant différentes techniques d'exploration et d'estimation.

1.3.1 Estimateur de Hill

Soient $X_1, ..., X_n$ des variables aléatoires iid de fonction de répartition F que l'on suppose appartenir au domaine d'attraction de la loi de Fréchet de paramètre $\alpha > 0$. L'estimateur de Hill permet d'estimer α à l'aide des observations. Il s'écrit sous la forme :

$$\hat{\alpha}_{k,n}^{(H)} = \left(\frac{1}{k} \sum_{j=1}^{k} \ln X_{j:n} - \ln X_{k:n} \right)^{-1}$$

Cet estimateur dépend du paramètre $k(n)$ qui doit vérifier $k(n) \underset{n\to+\infty}{\longrightarrow} +\infty$ (pour utiliser une statistique d'ordre suffisant) et $\dfrac{k(n)}{n} \underset{n\to+\infty}{\longrightarrow} 0$ (pour considérer seulement la queue de la distribution). Dans ce cas, l'estimateur est consistant.

Nous utilisons le package R evir pour calculer cet estimateur. Nous avons $n = 3441$ observations. Nous choisissons une valeur de k qui vérifie les deux conditions précédentes et qui se trouve dans une zone de stabilité. C'est le cas pour $k = 450$, c'est-à-dire en considérant les valeurs au delà du seuil $u = 340$. On obtient alors $\hat{\alpha} = 7.3$.

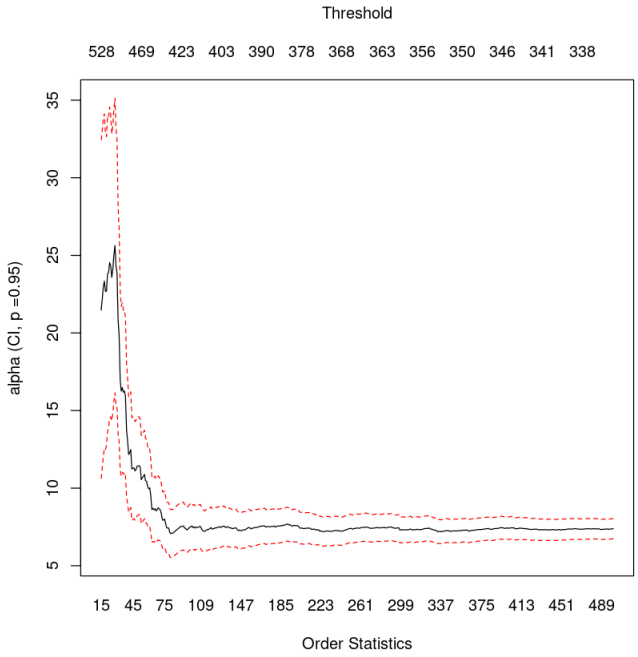
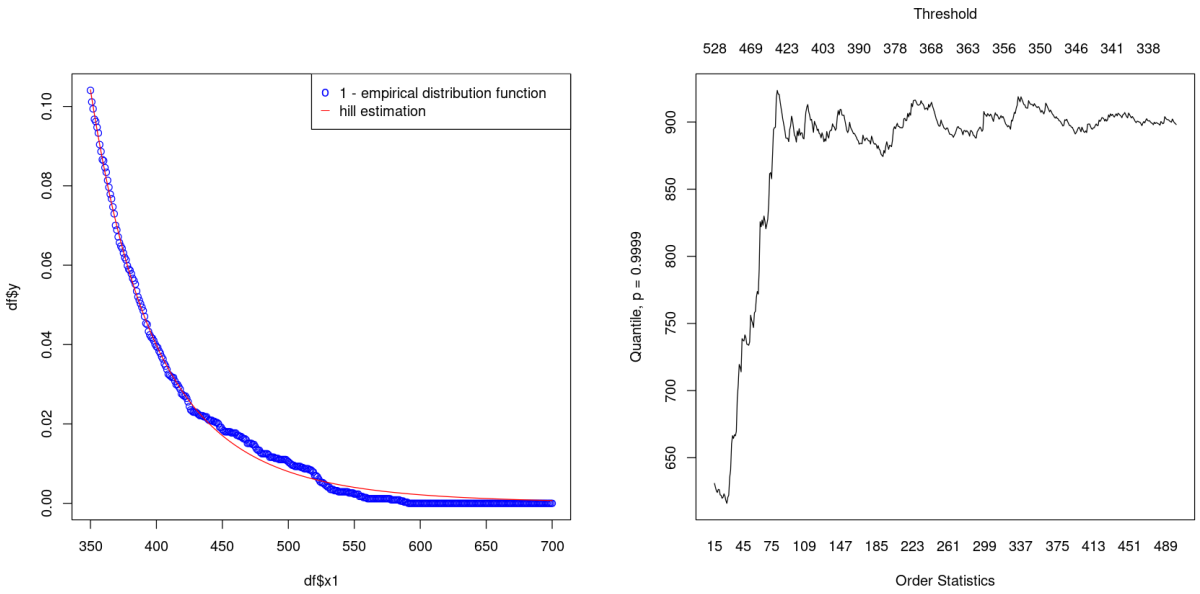


FIGURE 10 – Estimateur de Hill de α en fonction de l'ordre k

Cet estimation nous permet de déduire directement deux autres estimateurs. Le premier est l'estimateur de la fonction de survi au delà du seuil u. Il s'écrit sous la forme

$$\overline{F}(\hat{x}) = \frac{k}{n}\left(\frac{x}{X_{k:n}}\right)^{-\hat{\alpha}_{k,n}^{(H)}}$$

pour $x \geq X_{k:n}$.



(a) Estimateur de Hill de 1 - la fonction de répartition

(b) Estimateur de Hill du quantile 0.9999

FIGURE 11 – Estimateurs dérivés de $\hat{\alpha}_{k,n}^{(H)}$

Cela nous permet de vérifier graphiquement que notre estimation est pertinente. En effet, nous obtenons une queue de distribution de la forme $\hat{\overline{F}}(x) = \hat{C}x^{-\hat{\alpha}_{k,n}^{(H)}}$ pour $x \geq 350$, avec $\hat{\alpha}_{k,n}^{(H)}$ obtenu précédemment et $\hat{C}$ une (très grande) constante obtenue à partir des données. Cette estimation parait pertinente graphiquement.

Enfin, l'estimateur de Hill nous donne un estimateur du p-quantile de F :

$$\hat{x_p} = \left(\frac{n}{k}(1-p)\right)^{-1/\hat{\alpha}_{k,n}^{(H)}} X_{k:n}$$

Par exemple, nous pouvons obtenir une estimation quantile à 99.99%, alors que nous ne pouvons pas l'observer avec nos n réalisations (nous obtenons un quantile empirique égal à 589, le maximum de l'échantillon). En prenant $k = 450$, nous obtenons un quantile estimé à environ 900.

Nous pouvons déjà répondre en partie à notre problématique : si l'hypothèse d'appartenance au domaine d'attraction de Fréchet est valide, le nombre de logs wifi n'excédera pas 900/heure, avec une probabilité supérieure à $1 - 10^{-4}$.

1.3.2 Estimateur de Pickands

Contrairement à l'estimateur de Hill qui se limite au domaine d'attraction de Fréchet, l'estimateur de Pickands permet d'estimer ξ quelque soit son signe, c'est-à-dire pour les lois des valeurs extrêmes généralisées. Une telle loi de paramètre ξ a pour fonction de répartition :

$$H_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right) & \text{si } \xi \neq 0 \\ \exp(-\exp(-x)) & \text{si } \xi = 0 \end{cases}$$

pour $1 + \xi x > 0$.

$\xi = \alpha^{-1} > 0$ correspond à la distribution de Fréchet, $\xi = 0$ correspond à la distribution de Gumbel et $\xi = -\alpha^{-1} < 0$ à la distribution de Weibull. Ici, on ne fait aucune hypothèse sur le signe de ξ et nous allons vérifier que ξ est bien positif.

L'estimateur de Pickands s'écrit :

$$\hat{\xi}_{k,n}^{(P)} = \frac{1}{\ln 2} \ln\left(\frac{X_{k:n} - X_{2k:n}}{X_{2k:n} - X_{4k:n}}\right)$$

Cela vient du fait que si F appartient au domaine d'attraction de H_ξ, alors $U(t) = F^{\leftarrow}(1 - t^{-1})$ vérifie

$$\frac{U(2t) - U(t)}{U(t) - U(t/2)} \underset{t\to+\infty}{\longrightarrow} 2^\xi$$

Avec le package R smoothtails, on obtient l'estimateur de Pickands (pour comparaison, en rouge, le paramètre $\hat{\xi}_{450,n}^{(H)} = 1/\hat{\alpha}_{450,n}^{(H)}$ issu de l'estimateur de Hill). On voit que les valeurs sont relativement proches pour l'ordre $k = 450$ choisi précédemment. Cependant cet estimateur est moins stable que l'estimateur de Hill, donc il est difficile d'obtenir une valeur précise. On peut voir néanmoins que l'on obtient bien une valeur de ξ positive.

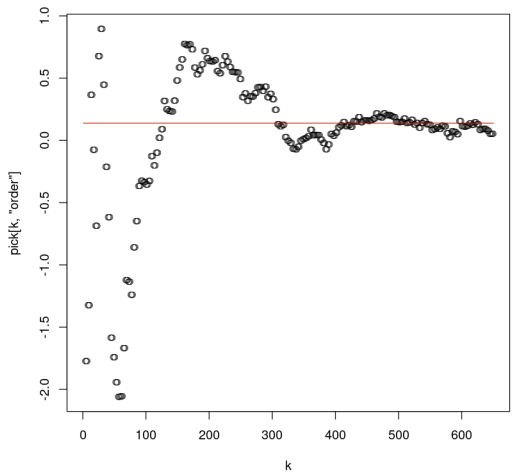
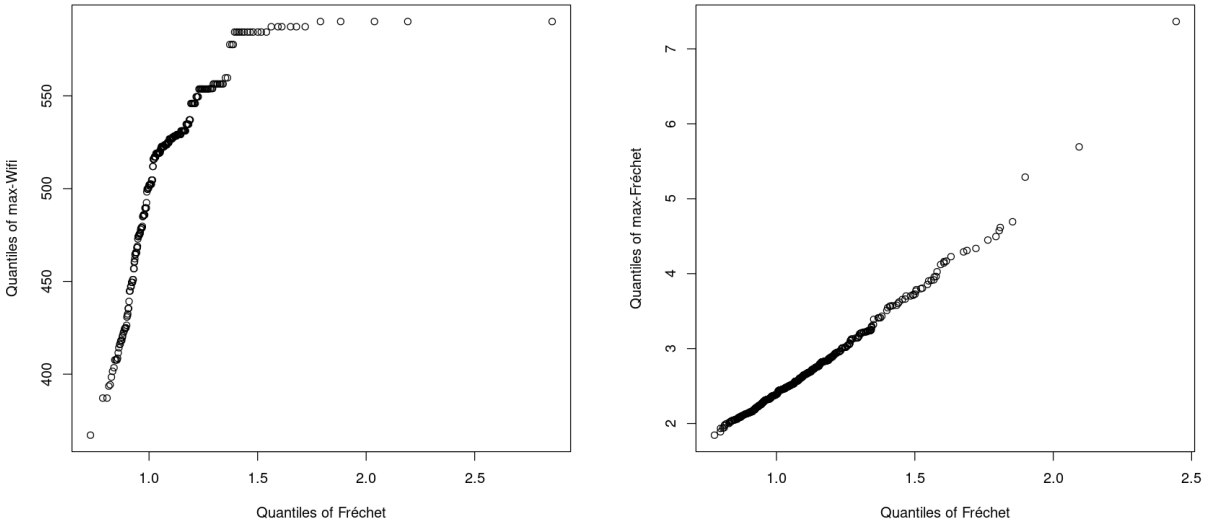


FIGURE 12 – Estimateur de Pickands de ξ en fonction de l'ordre k

1.3.3 Domaine d'attraction de Fréchet

Nous pouvons maintenant supposer que la fonction de répartition des logs wifi appartient au domaine d'attraction de la distribution de Fréchet de paramètre $\alpha = 7.2$. Nous cherchons à illustrer ce phénomène en observant le comportement du maximum M_n de n échantillons $X_1, ..., X_n$. Nous devons obtenir un comportement max-stable, c'est-à-dire que $c_n^{-1}(M_n - d_n) \xrightarrow[\mathcal{L}]{} \Phi_\alpha$.

Pour cela, nous piochons au hasard m échantillons parmi les $n = 3441$ observations et nous en calculons le maximum M_m. Nous répétons l'expérience 500 fois et nous comparons la distribution des maxima avec celle d'une loi de Fréchet. Pour visualiser ces distributions, nous traçons un QQ-plot, qui a l'avantage d'être invariant (à l'échelle près) par transformation affine. Ainsi, il n'est pas nécessaire de calculer les constantes c_n et d_n pour comparer la forme des quantiles.



(a) maxima d'observations / loi de Fréchet

(b) maxima loi de Fréchet / loi de Fréchet

FIGURE 13 – QQ-plot pour 500 répétitions de l'expérience

Nous devrions obtenir une droite. Ce n'est pas réellement le cas car nous ne pouvons pas faire tendre m vers $+\infty$. En effet, $m \leq n$ et nous ne pouvons pas choisir m trop proche de n, sous peine de récupérer toujours la même observation maximale (ie le maximum des observations).

Pour comparer, nous utilisons des simulations de lois de Fréchet dont nous calculons le maximum. Cette fois, nous pouvons choisir n suffisamment grand et nous obtenons un QQ-plot proche d'une droite.

En conclusion, le nombre limité d'observations dont nous disposons ne nous permet pas de mettre en évidence la propriété de max-stabilité d'une loi appartenant au domaine d'attraction de Fréchet.

2 Analyse d'une série temporelle

2.1 Approche par décomposition tendance - saisonnalité - bruit

Étalée de 1895 à 2018, la série temporelle que nous étudions donne la température moyenne dans l'ensemble des États-Unis chaque mois.

On observe de prime abord sur la figure 14 une forte saisonnalité annuelle (donc d'ordre 12), et un tendance a priori très faible. La série temporelle peut donc s'exprimer sous la forme $X_t = m_t + S_t + Y_t$, avec m_t une tendance déterministe, S_t une saisonnalité annuelle et Y_t des résidus.

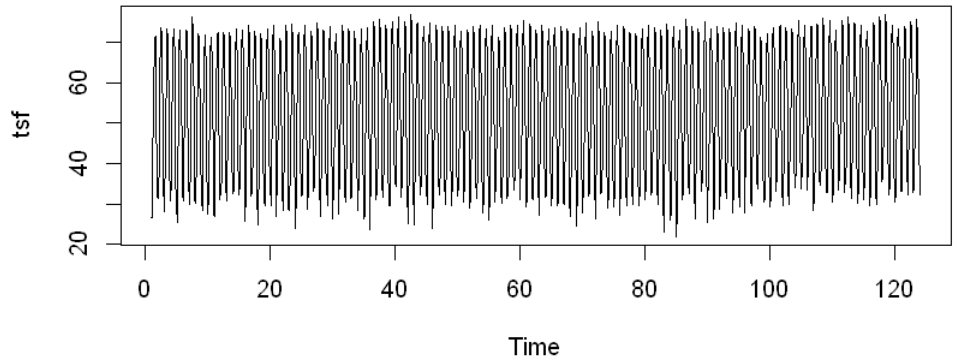


FIGURE 14 – La température moyenne mensuelle aux États-Unis entre 1895 et 2018

Afin de séparer le tendance de la saisonnalité, on applique une moyenne mobile $\{[13]; \frac{1}{24}[1,2,2,2,2,2,2,2]\}$ qui annule la saisonnalité d'ordre 12. Nous obtenons une tendance approximativement linéaire, que nous modélisons par une tendance linéaire obtenue par régression. Nous obtenons ainsi une tendance linéaire $\hat{m}_t = 0.001221 \times t + 51.274747$ (voir figure 15).

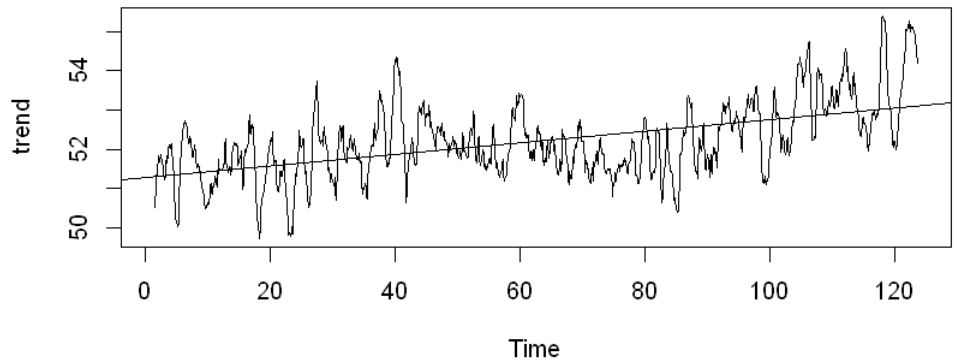


FIGURE 15 – Tendance obtenue par moyenne mobile et son approximation linéaire obtenue par régression

Ensuite, nous soustrayons $(\hat{m}_t)$ au signal, puis nous cherchons à isoler la saisonnalité annuelle. Pour cela, nous appliquons une moyenne mobile symétrique $\{[25]; \frac{1}{3}[1, 0, ..., 0, 1]\}$. Nous obtenons alors une série temporelle $(\hat{Y}_t)$ dont nous traçons l'autocorrélogramme.

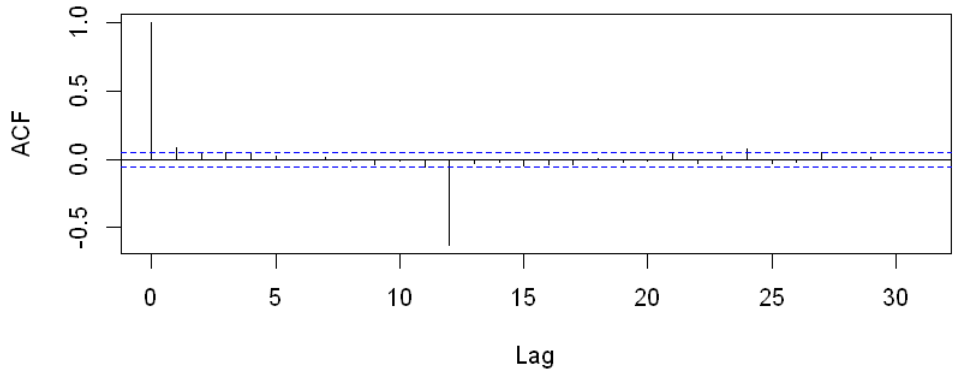


FIGURE 16 – Autocorrélogramme de $(\hat{Y}_t)$

Nous modélisons $(\hat{Y}_t)$ comme une série temporelle linéaire. L'autocorrélation pour un lag de 12 nous suggère d'utiliser un modèle SARMA_{12}. Le modèle le plus satisfaisant est le $\text{SARMA}_{12}(1,1)(0,2)$. Les résidus (ε_t) ont un autocorrélogramme qui suggère qu'il s'agit d'un bruit blanc.

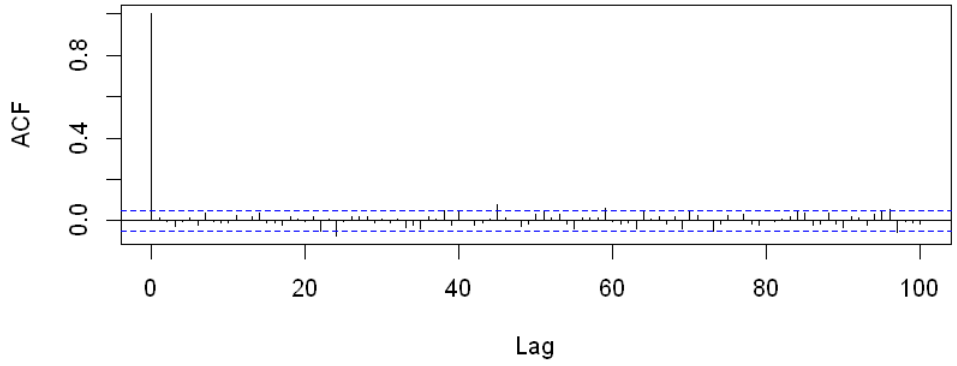


FIGURE 17 – Autocorrélogramme de (ε_t)

Nous effectuons finalement un test de Ljung-Box pour tester l'hypothèse H_0 de non corrélation des observations ε_t. Le test ne nous permet pas de rejeter cette hypothèse au seuil de 5% (voir figure 18).

En conclusion, nous obtenons une modélisation de la série temporelle par la somme de :

— une tendance linéaire déterministe,
— une saisonnalité déterministe,
— un modèle $\text{SARMA}_{12}(1,1)(0,2)$.

2.2 Modèle SARIMA

Nous souhaitons à présente comparer le modèle précédent avec un modèle plus simple, c'est-à-dire qui ne nécessite pas d'effectuer une décomposition tendance - saisonnalité - bruit. Pour cela, nous utilisons un modèle de type SARIMA_{12} pour modéliser directement (X_t).

Le modèle qui annule le mieux les autocorrélations est le modèle $\text{SARIMA}_{12}(1,0,1)(0,1,1)$. Nous obtenons des résidus qui semblent également non corrélés. Cependant, et c'était le cas également pour le modèle précédent, on note que l'hypothèse H_0 de non corrélation est rejetée pour des valeurs de lags élevées (> 150),

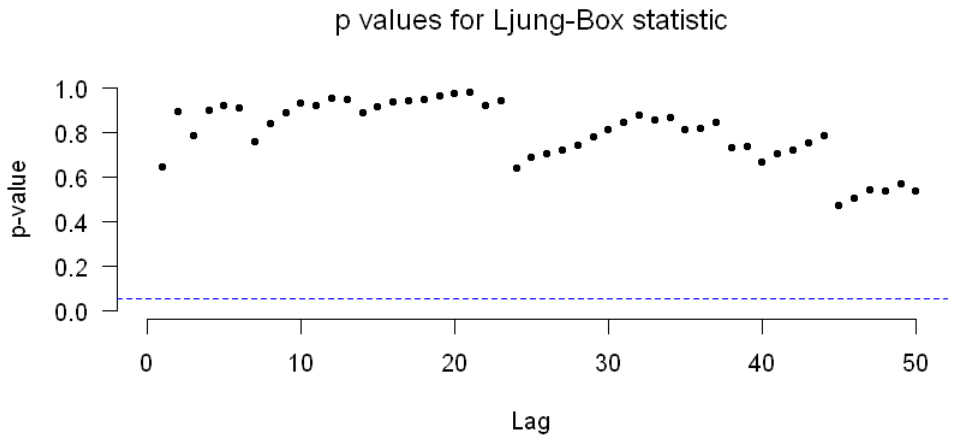


FIGURE 18 – p-valeurs du test de Ljung-Box sur les résidus (ε_t) du $\text{SARMA}_{12}(1,1)(0,2)$

comme on le voit sur la figure 19. Cela pourrait s'expliquer notamment par une annulation imparfaite des effets de la saisonnalité, ou par le fait qu'on teste l'hypothèse d'indépendance linéaire d'un grand nombre de variables.

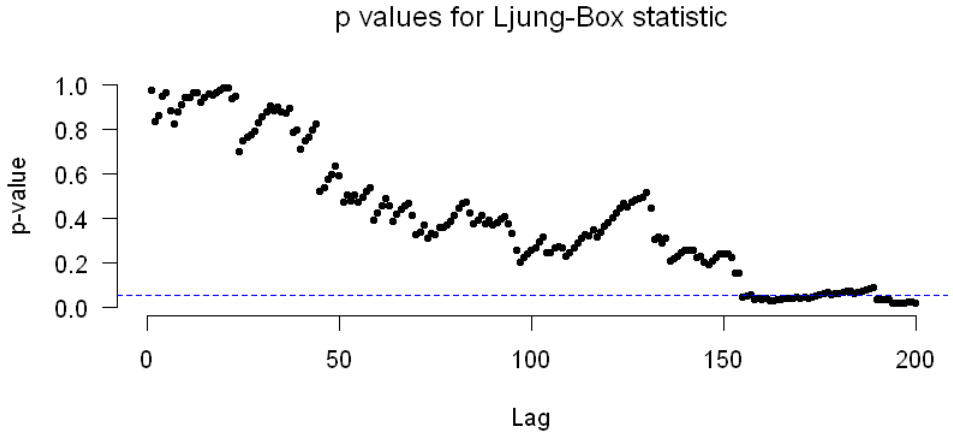


FIGURE 19 – p-valeurs du test de Ljung-Box sur les résidus du $\text{SARIMA}_{12}(1,0,1)(0,1,1)$ pour des valeurs de lag élevées

2.3 Prédiction

Nous cherchons finalement à comparer les capacités de prédiction des deux modèles présentés. Nous ajustons les modèles sur les mois 1 à 1200 (100 ans), et nous les évaluons sur les mois 1201 à 1478 (23 ans). Nous prenons pour critère d'évaluation l'erreur quadratique moyenne $\mathcal{E} = \frac{1}{278} \sum_{t=1201}^{1478} (X_t - \hat{X}_t)^2$.

Pour le modèle tendance + saisonnalité + SARMA, nous prolongeons la saisonnalité et étendons la tendance linéaire, puis nous utilisons la fonction R forecast. Pour le modèle SARIMA, nous utilisons uniquement forecast.

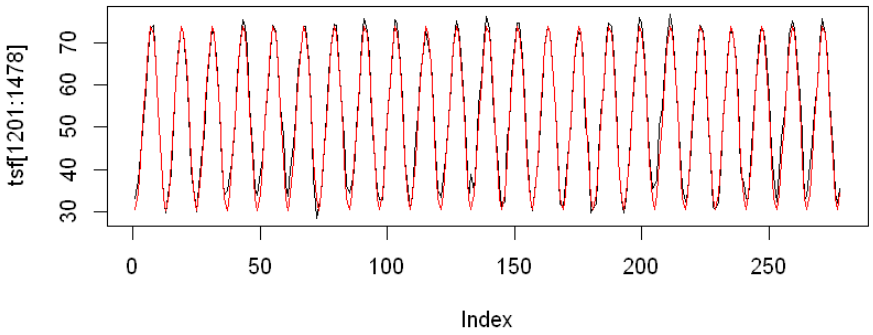


FIGURE 20 – Prédictions du $\text{SARIMA}_{12}(1,0,1)(0,1,1)$ en rouge, vraies valeurs en noir

Modèle	Erreur quadratique moyenne
tendance + saisonnalité + $\text{SARMA}_{12}(1,1)(0,2)$	4.1036
$\text{SARIMA}_{12}(1,0,1)(0,1,1)$	5.0394

FIGURE 21 – Erreur quadratique moyenne de prédiction des modèles

Le modèle ayant la meilleure performance prédictive est le modèle tendance + saisonnalité + $\text{SARMA}_{12}(1,1)(0,2)$. Cependant, il faut noter que les prédictions par de ces modèles de séries temporelles ne sont significatives que pour des temps proches. En effet (voir figure 22), au delà d'une trentaine de valeurs, la meilleure prédiction de la perturbation Y_t dans le modèle SARMA est 0. L'essentiel de la prédiction réside donc dans la tendance et la saisonnalité. Une manière alternative de procéder serait d'effectuer une prédiction glissante, c'est-à-dire de prédire mois par mois et de réinjecter les vraies valeurs. Néanmoins, cette méthode n'est pas envisageable si l'on veut prédire des évolutions du climat à plus long terme.

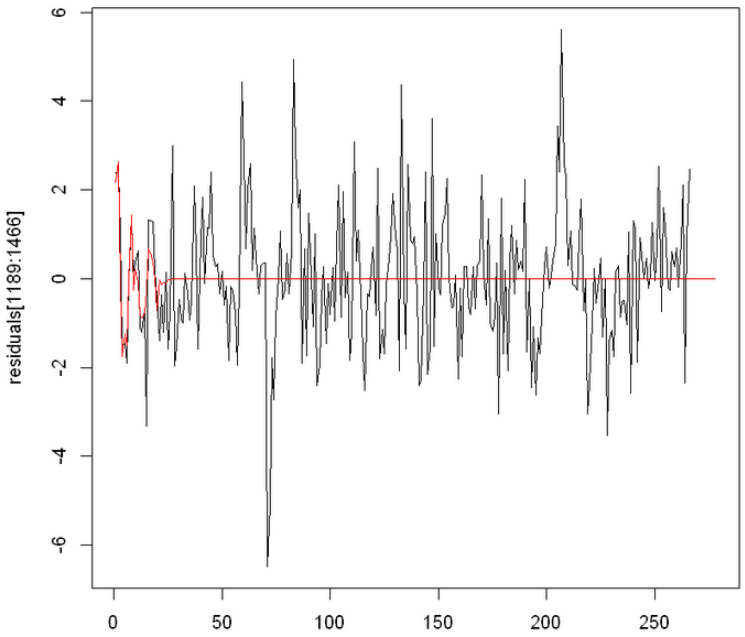


FIGURE 22 – Prédictions du SARMA en rouge, vraies valeurs en noir après soustraction de la tendance et de la saisonnalité

Références

Jeux de données :

ParisData, Mairie de Paris, Mars 2018, `https://opendata.paris.fr/explore/?q=nation`

NOAA National Centers for Environmental information, Climate at a Glance : National Time Series, published March 2018, retrieved on March 29, 2018 from `http://www.ncdc.noaa.gov/cag/`

Documents :

Embrechts, P., Klüppelberg, C., & Mikosch, T. (2013). *Modelling extremal events : for insurance and finance* (Vol. 33). Springer Science & Business Media.