

Modélisation des séries de températures avec différentiation géographique et impact climatique

Table des matières

1 Descriptif des données	2
2 Traitement des données	5
2.1 Estimation des signaux déterministes	5
2.1.1 Estimation de la tendance	5
2.1.2 Estimation de la saisonnalité	6
2.1.3 Estimation de la tendance de la variance	6
2.1.4 Estimation de la saisonnalité de la variance	7
2.2 A propos de la Partitioned Cross Validation (PCV)	8
2.3 A propos de l'implémentation	8
3 Etude des résidus	9
3.1 Etude descriptive	10
3.2 Analyse en composantes principales	13
3.3 Sur l'hypothèse de stationnarité de nos résidus	15
4 Modélisation des résidus et des composantes principales.	16
4.1 Généralités et estimation des supports	16
4.2 Modélisation des composantes principales par des lois Kappa	18
4.3 Modélisation des composantes principales par des lois Beta	20
4.4 Sur l'indépendance des composantes principales	21
5 Modélisation des trajectoires	24
5.1 Modèle de diffusion basé sur la loi marginale et l'autocorrélogramme [Bibby-Sorensen]	24
5.1.1 Généralités et aspects théoriques	24
5.1.2 Calibration de la fonction d'autocorrélation	26
5.1.3 Implémentation (temporaire)	26
5.2 Modèle de diffusion avec mean-reversion dépendant d'un paramètre stochastique.	28
5.2.1 Idée générale	28
5.2.2 Exemple	29
5.2.3 Estimation non paramétrique de la persistance	33
5.2.4 A DEPLACER	38
A Liste des stations	40

Chapitre 1

Descriptif des données

Les données utilisées sont issues de l'European Climate Assessment & Dataset. Elles sont composées des relevés journaliers en degrés celsius des températures maximales pour 44 stations réparties à travers la France.

L'étude des valeurs manquantes sur l'intervalle 1950-2017, plus détaillée dans le fichier MissingValues.html, indique que toutes les stations ne sont pas utilisables en fonction du choix de l'intervalle de temps.



FIGURE 1.1 – En bleu les stations valides, en rouge les stations invalides

La solution choisie, en raison du maillage conséquent du territoire français (à l'exception de la région est), est de ne pas tenir compte des stations invalides sur l'intervalle 1950-2017. On ne considère ainsi finalement que 30 stations, dont la liste est fournie en annexe.

Dans la suite, on note Y_t la température relevée au temps t . On ne précise pour l'instant pas la station dont cette mesure est issue, ni s'il s'agit de la température minimale, maximale ou moyenne.

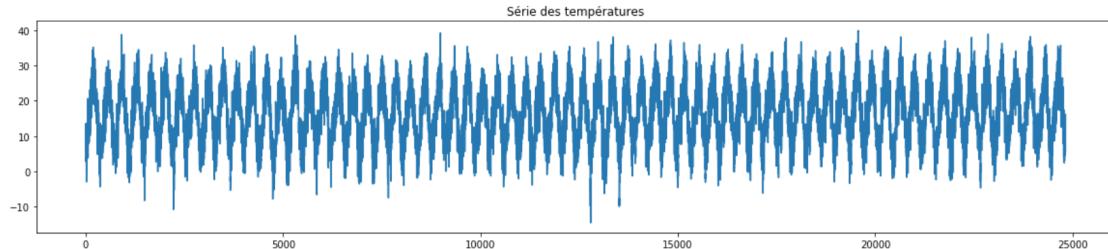


FIGURE 1.2 – Série des températures maximales (station 32)

On fixe aussi les conventions suivantes :

- On considère que chaque année contient 365 jours, on supprime ainsi les 29 février des années bissextiles.
- On associe chaque date à un identifiant numérique unique, incrementé de 1. La date initiale (01/01/1950) est donc la date 1, et la date finale (31/12/2017) la date 24820.

Chapitre 2

Traitement des données

2.1 Estimation des signaux déterministes

2.1.1 Estimation de la tendance

La méthode LOESS (régression non paramétrique pondérée) est utilisée pour l'estimation de la tendance.

Considérons $h > 0$ une fenêtre et t_1, \dots, t_n nos dates ordonnées. Considérons de plus K un kernel. On souhaite obtenir une tendance de la forme :

$$\hat{m}(t_k) = \beta_0^*(t_k) + \sum_{j=1}^d \beta_j^*(t_k) t_k^j$$

où les $\beta_j^*(t_k)$ sont obtenus en minimisant, pour tout k , la fonction de régression pondérée suivante :

$$\sum_{i=1}^n K\left(\frac{t_k - t_i}{h}\right) (Y_{t_i} - \beta_0 - \beta_1 t_i - \dots - \beta_d t_i^d)^2$$

A noter que cette procédure peut être effectuée plusieurs fois pour augmenter le fitting aux données. Il s'agit alors d'effectuer une régression LOESS sur les résidus successifs $Z_i - \hat{m}_{t_i}$.

Dans la bibliographie, il est fixé $d = 1$ et le kernel utilisé est le tricube :

$$K(x) \propto (1 - |x|^3)^3 \mathbb{1}_{[-1,1]}(x)$$

La recherche du paramètre h optimal se fait par minimisation du score de Partitioned Cross Validation (PCV) associé à l'erreur quadratique moyenne, pour G=10 groupes (voir en fin de section).

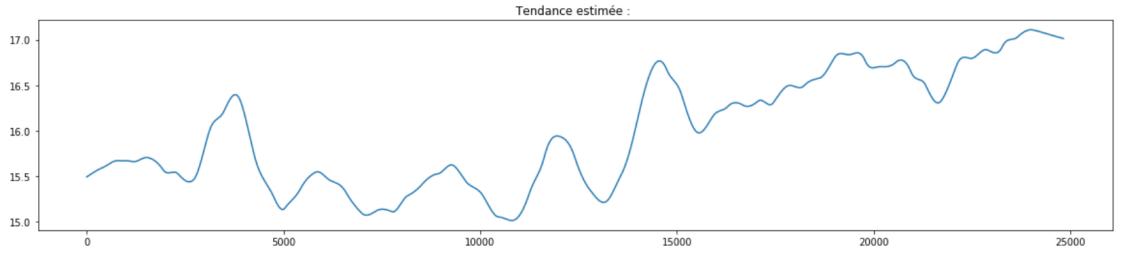


FIGURE 2.1 – Tendance des températures maximales (station 32)

2.1.2 Estimation de la saisonnalité

L'idée est de fitter par régression ordinaire, pour chaque station, la température désaisonnalisée $Z_t = Y_t - \hat{m}(t)$ par rapport à une fonction trigonométrique de la forme :

$$\sum_{i=0}^p \left(\theta_{i,1} \cos\left(\frac{2\pi i t}{365}\right) + \theta_{i,2} \sin\left(\frac{2\pi i t}{365}\right) \right)$$

Ceci pour $p = 0, \dots, 6$. La fonction finale est choisie par minimisation du critère AIC. On note $\hat{s}(\cdot)$ la fonction de saisonnalité ainsi obtenue.

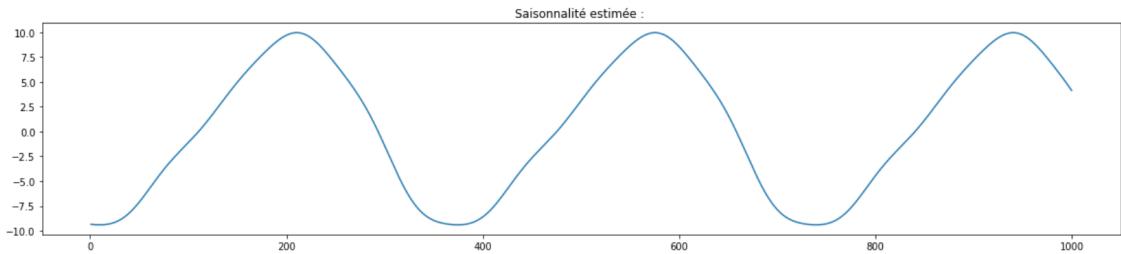


FIGURE 2.2 – Saisonnalité des températures maximales (station 32)

2.1.3 Estimation de la tendance de la variance

On pose $v_t^2 = (Y_t - \hat{m}(t) - \hat{s}(t))^2$. On cherche à estimer une tendance multiplicative par régression LOESS, toujours avec $d = 1$ et le kernel tricube.

On cherche ainsi une fonction \widehat{m}_v de la forme :

$$\widehat{m}_v(t_k) = \beta_0^*(t_k) + \sum_{j=1}^d \beta_j^*(t_k) t_k^j$$

Où les $\beta_j^*(t_k)$ sont obtenus par minimisation de :

$$\sum_{i=1}^n K \left(\frac{t_k - t_i}{h} \right) (v_{t_i}^2 - \beta_0 - \beta_1 t_i - \dots - \beta_d t_i^d)^2$$

A noter que, comme dans le script tendais.R, la fenêtre h utilisée est la même que celle utilisée pour l'estimation de la tendance des Y_t .

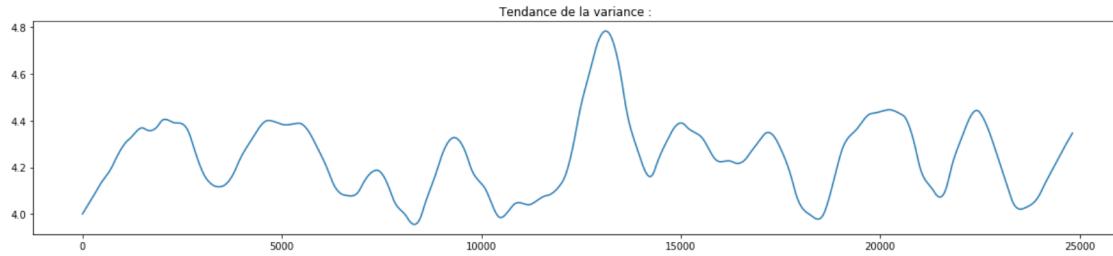


FIGURE 2.3 – Tendance de la variance des températures maximales (station 32)

2.1.4 Estimation de la saisonnalité de la variance

On estime la saisonnalité de la variance par régression linéaire ordinaire de notre variance désaisonnable $\frac{v_t^2}{\widehat{m}_v(t)}$ contre les fonctions trigonométriques de la forme :

$$\sum_{i=0}^p \left(\theta_{i,1} \cos \left(\frac{2\pi i t}{365} \right) + \theta_{i,2} \sin \left(\frac{2\pi i t}{365} \right) \right)$$

Où $p = 0,..6$. On sélectionne le meilleur modèle par minimisation de l'AIC. On note la fonction obtenue $\widehat{s}_v^2(\cdot)$

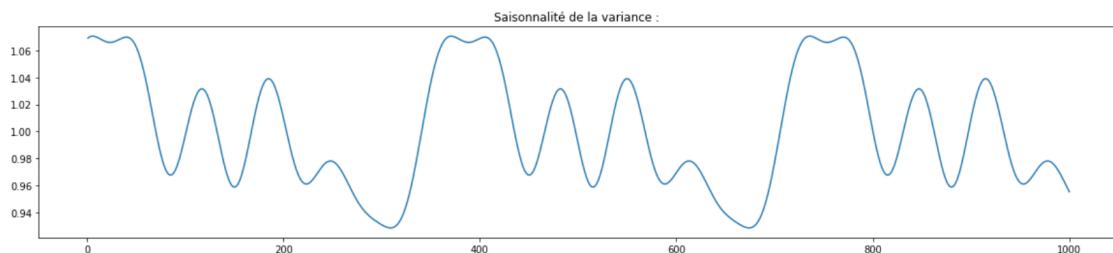


FIGURE 2.4 – Saisonnalité de la variance des températures maximales (station 32)

2.2 A propos de la Partitioned Cross Validation (PCV)

La sélection de la fenêtre h optimale pour la régression LOESS s'effectue par minimisation du score de PCV associée à l'erreur quadratique moyenne.

Procédure, pour chaque h appartenant à un ensemble de valeurs potentielles :

- On fixe $G=10$ un nombre de groupes.

- Pour chaque $g = 1, \dots, G$, on divise notre ensemble D de dates d'entraînement en un groupe d'entraînement et un groupe de test.

Plus précisement on pose $test_g = [g + xG]_{\{x|g+xG \in D\}}$ et $train_g = D \setminus test_g$

- Pour chaque $T \in test_g$, on "prédit" la tendance au temps T grâce aux dates contenues dans $train_g$. Mathématiquement, on cherche $\hat{m}(T) = \beta_0^*(T) + \sum_{j=1}^d \beta_j^*(T)T^j$ où les $\beta_j^*(t_k)$ minimisent :

$$\sum_{t \in train_g} K\left(\frac{T-t}{h}\right) (Y_t - \beta_0 - \beta_1 t - \dots - \beta_d t^d)^2$$

- On calcule l'erreur quadratique moyenne associée à nos prédictions :

$$MSE(g) = \frac{1}{card(test_g)} \sum_{T \in test_g} (Y_T - \hat{m}(T))^2$$

On sélectionne la valeur de h qui minimise $PCV = \frac{1}{G} \sum_{g=1}^G MSE(g)$.

2.3 A propos de l'implémentation

Les estimations sont réalisées, en l'absence -pour l'instant- d'implémentation python satisfaisante pour la régression LOESS, grâce au script `tendsais.R` fourni par Sylvie Parey. Aussi, le paramètre h optimal, représentant cette fois-ci la proportion optimale de données à utiliser, est de 0.08. Ce paramètre s'applique à toutes les stations, aussi bien pour l'estimation de la tendance des températures que pour celle de la tendance de la variance.

Chapitre 3

Etude des résidus

On considère dans cette partie nos résidus renormalisés, c'est à dire, en notant $Y_i(t)$ la température maximale relevée au temps t pour la station i , et en utilisant la même convention pour nos estimations des tendances et saisonnalités, on étudie les X_i définis par :

$$X_i(t) = \frac{Y_i(t) - \hat{m}_i(t) - \hat{s}_i(t)}{\sqrt{\hat{s}_{v,i}^2(t) \hat{m}_{v,i}^2(t)}}$$

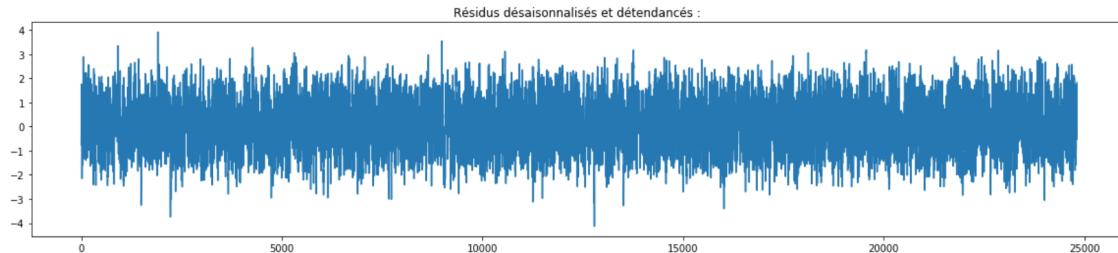


FIGURE 3.1 – Résidus normalisés des températures maximales (station 32)

D'un point de vue distribution, les résidus semblent suivre une loi à support compact. Cela fait sens d'un point de vue climatologique : les températures sont en effet à valeurs dans un intervalle borné, et ainsi la partie purement aléatoire de ces températures ne peut prendre ses valeurs que dans un tel intervalle.

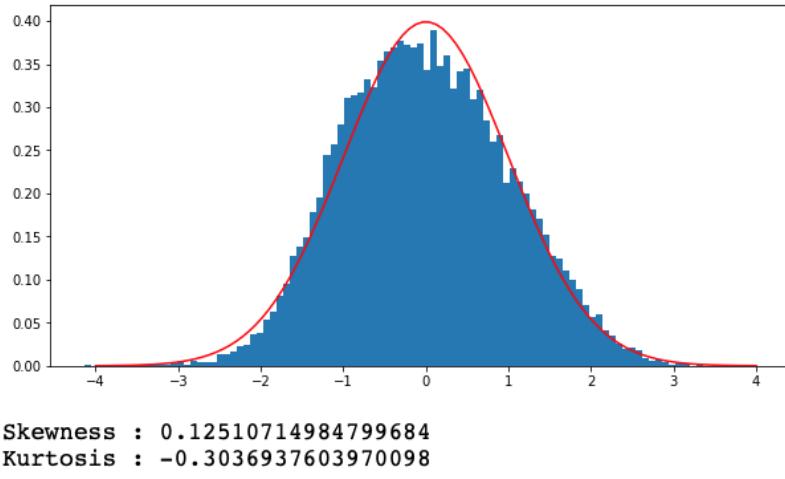


FIGURE 3.2 – Histogramme, skewness et kurtosis des résidus normalisés (station 32).

3.1 Etude descriptive

L'étude des autocorrélogrammes de nos résidus normalisés montre une corrélation temporelle positive entre les valeurs. On constate une décroissance exponentielle de la corrélation, qui devient négligeable à partir du rang 15. Une étude des décalages plus prononcés montre aussi une légère corrélation négative, négligeable, pour un décalage d'environ 6 mois.

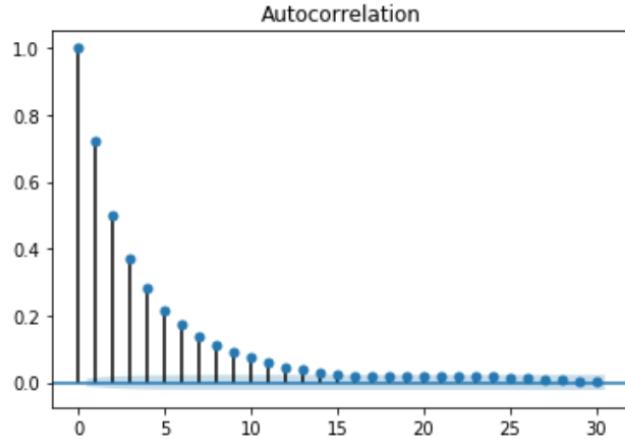


FIGURE 3.3 – Autocorrélogramme des résidus normalisés (station 32).

L'étude des autocorrélations partielles, cependant, montre que la vaste majorité des corrélations mentionnées au-dessus peuvent être expliquées par les valeurs intermédiaires. Ainsi, les autocorrélations partielles indiquent une corrélation non-négligeable uniquement dans le cas d'un décalage de taille 1.

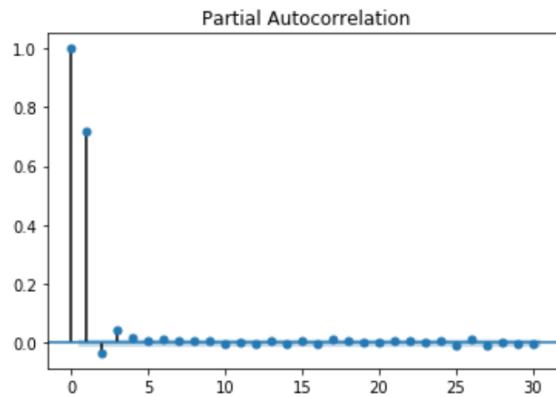


FIGURE 3.4 – Autocorrélogramme partiel des résidus normalisés (station 32).

D'un point de vue spatial, l'analyse purement graphique des résidus montre que la corrélation entre deux stations n'est pas constante par changement de stations. Ce phénomène est constatable par utilisation de scatterplot sur la station 32, en faisant varier la seconde station du plot.

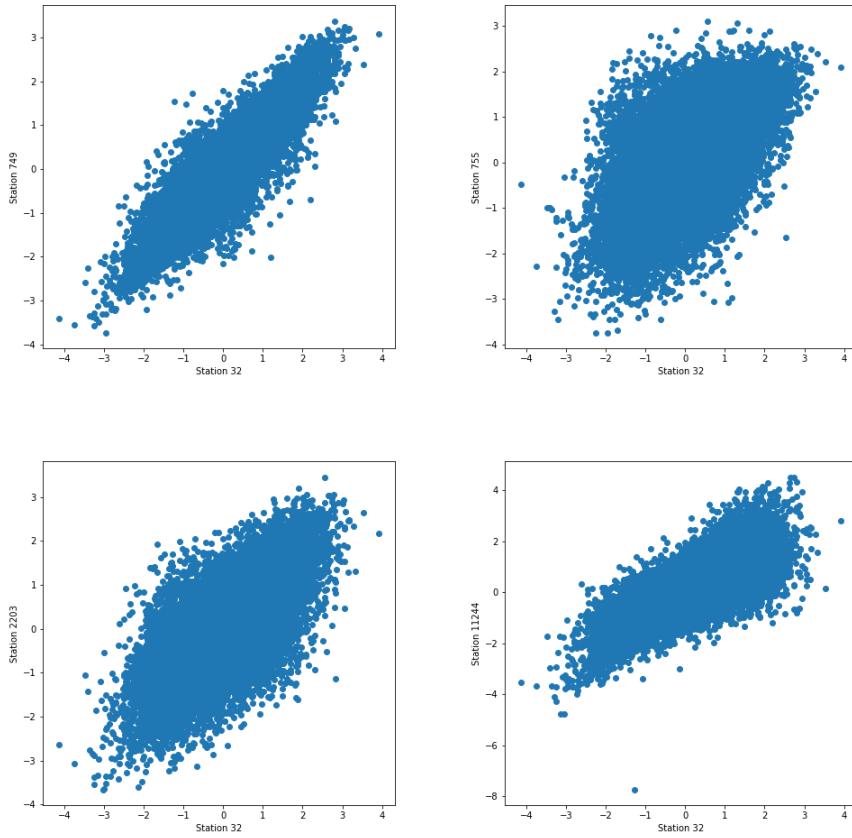


FIGURE 3.5 – Scatterplots de la station 32 contre différentes stations.

Une étude plus quantitative montre que les corrélations prennent valeur dans l'intervalle [0.20,0.90].

Une hypothèse tout à fait naturelle est que la corrélation de la température entre deux stations dépend de la distance qui les sépare. On confirme cette hypothèse en traçant, pour chaque station, la corrélation entre cette station et les autres par ordre de distance.

On obtient pour chacune des stations une tendance de corrélation décroissante. A noter que cette tendance est relativement irrégulière.

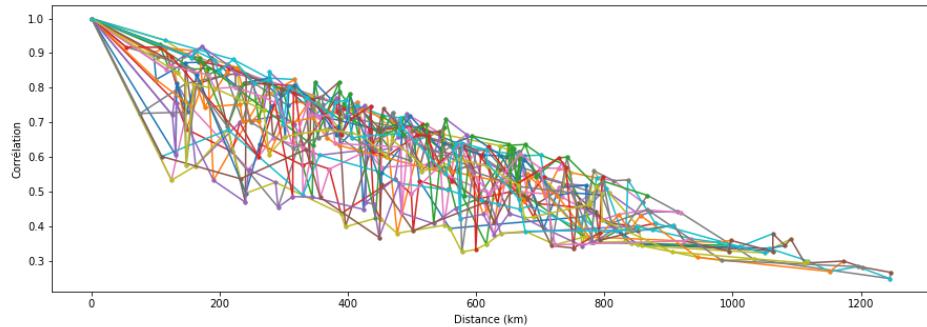


FIGURE 3.6 – Corrélation des stations en fonction de la distance entre elles.

3.2 Analyse en composantes principales

L'analyse en composantes principales de nos résidus spatialisés, réalisée à l'aide de la bibliothèque scikit, indique que la vaste majorité de la variance peut être expliquée par un nombre réduit de "directions".

Ainsi, plus précisément, plus de 90% de la variance a lieu dans un sous-espace de taille 9, bien inférieur à notre espace initial de taille 30. A noter que la première composante explique à elle seule 62% de la variance.

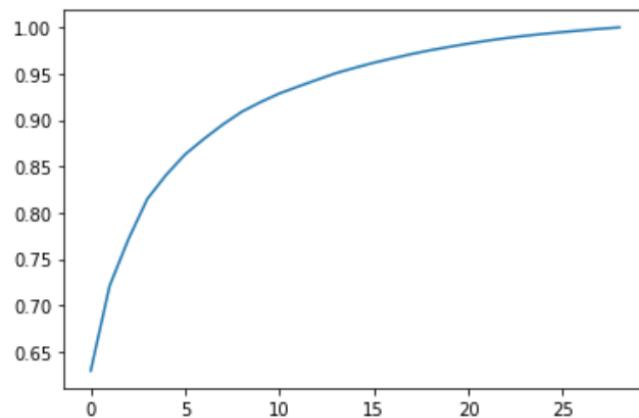


FIGURE 3.7 – Variance expliquée cumulée en fonction du nombre de directions.

On obtient ainsi les valeurs propres suivantes suite à l'ACP :

$[18.25, 2.66, 1.47, 1.26, 0.75, 0.64, 0.48, 0.45, 0.4, 0.3, 0.27, 0.22, 0.21, 0.21, 0.17, 0.16, 0.14, 0.14, 0.12, 0.11, 0.1, 0.09, 0.08, 0.07, 0.06, 0.06, 0.05, 0.05, 0.04]$

Les critères d'inertie de Kaiser-Guttman ($\lambda > 1$) et de Karlis-Saporta ($\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$) [Sap06] nous indiquent que seules les 4 premières composantes sont significatives.

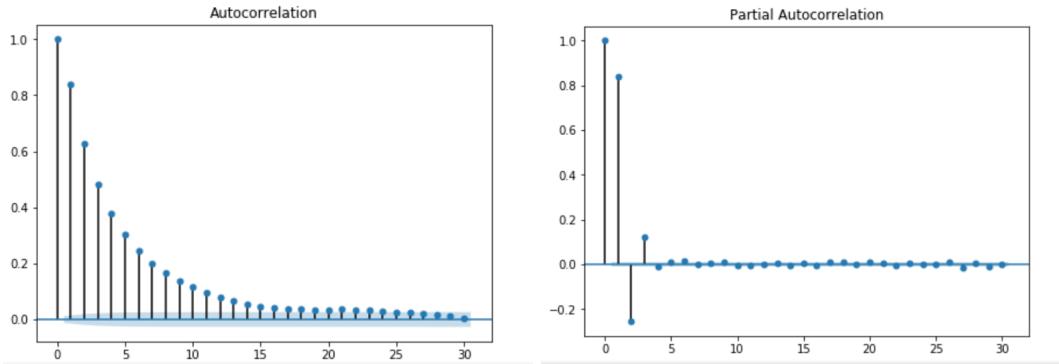


FIGURE 3.8 – Autocorrélation et autocorrélation partielle de la première composante principale.

En traçant l'erreur de reconstruction en fonction du nombre de composantes, on obtient une très nette décroissance lorsque le nombre de composantes augmente. Cependant les graphiques suivants indiquent qu'il est délicat de réduire la dimension "stochastique" sans fausser drastiquement nos résultats de par l'erreur de reconstruction.

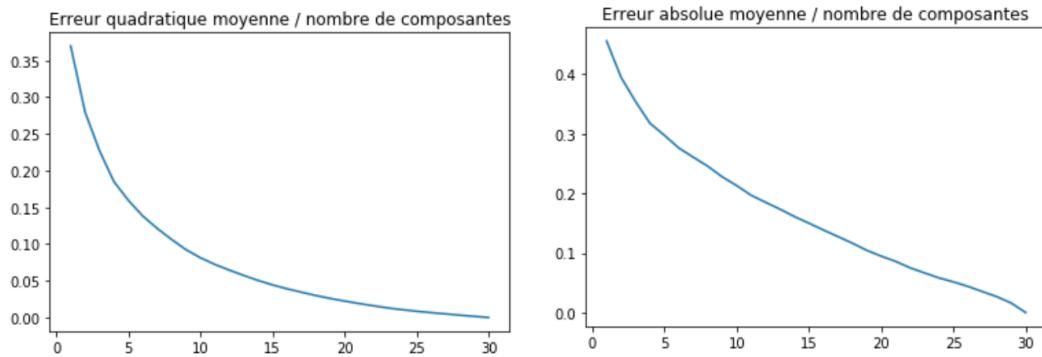


FIGURE 3.9 – Erreurs de reconstruction empiriques.

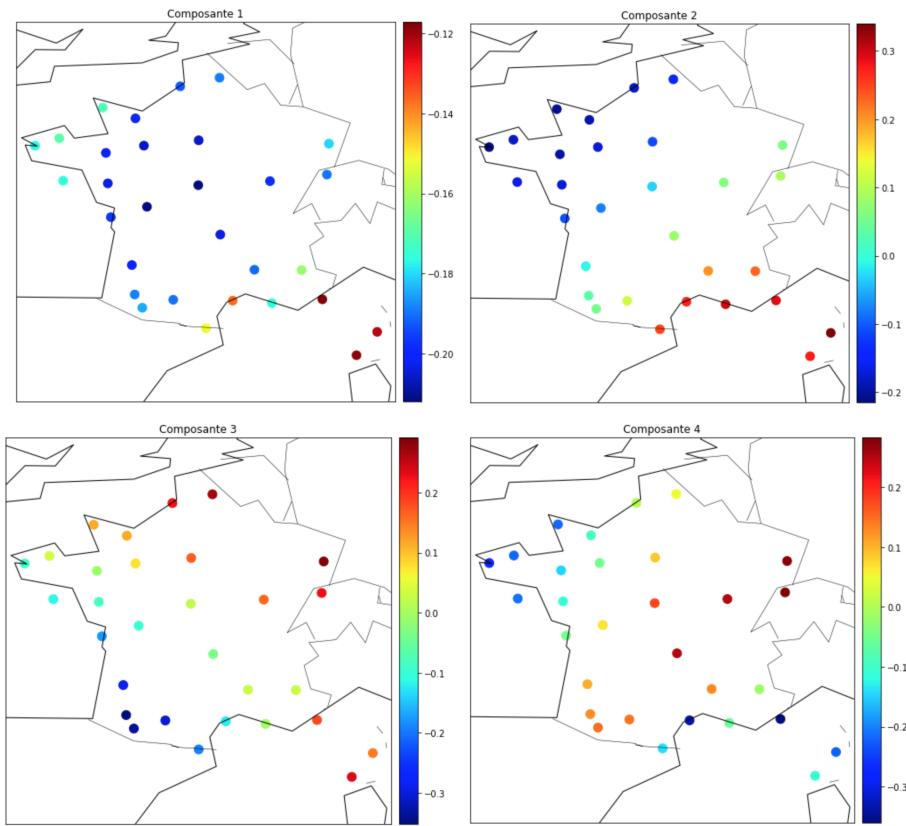


FIGURE 3.10 – Représentation géographique des composantes sélectionnées.

Les graphiques précédents nous permettent d'interpréter géographiquement nos facteurs principaux significatifs de la manière suivante :

- Le premier facteur principal représente le niveau global sur la majorité du territoire français, à l'exception particulière de la côte sud, de la corse, et des côtes bretonnes et normandes.
- Le second facteur principal représente le mouvement de variation nord-ouest/sud-est.
- Le troisième facteur principal représente le mouvement de variation sud-est/est.
- Le quatrième facteur principal représente le mouvement de variation de la diagonale [sud-ouest et centre]/[nord-ouest et sud-est]

3.3 Sur l'hypothèse de stationnarité de nos résidus

Chapitre 4

Modélisation des résidus et des composantes principales.

4.1 Généralités et estimation des supports

Nous pouvons supposer, de manière logique, que les températures sont bornées. Ainsi il est souhaitable de modéliser nos résidus de températures et nos composantes principales par des lois à support compact. La question se pose ainsi de l'estimation du support de ces lois.

Un estimateur naturel est d'utiliser le maximum et le minimum historique observé sur nos séries de températures. Cette estimateur, bien que convergent, pose un problème majeur : il empêche les événements climatiques plus extrêmes que ceux observés historiquement.

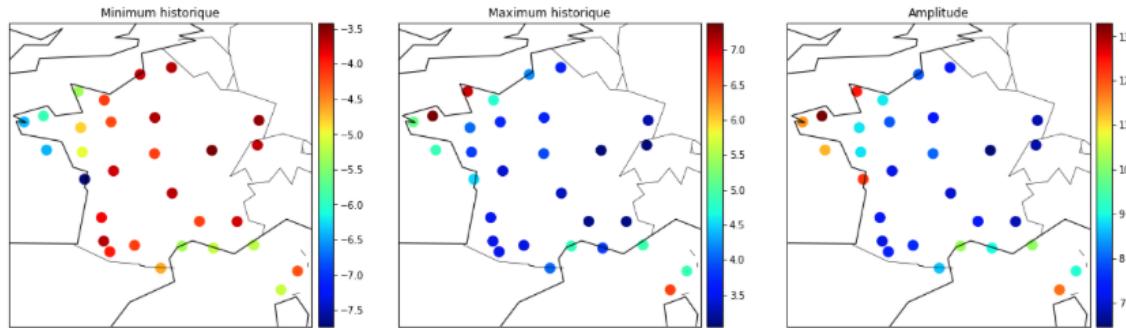


FIGURE 4.1 – Support historique des résidus.

Afin d'obtenir un support plus cohérent, il est nécessaire de se tourner vers la théorie des valeurs extrêmes, qui fournit des estimateurs censés converger plus rapidement sous de bonnes conditions [LdH06].

Définition 4.1. Soit $\gamma \in \mathbb{R}$. On dit qu'une loi de probabilité, identifiée par sa fonction de répartition F , appartient à $D(G_\gamma)$ le domaine d'attraction de G_γ s'il existe deux suites de réels $a_n > 0$ et b_n telles que :

$$\forall x \text{ tel que } 1 + \gamma x > 0, \quad F^n(a_n x + b_n) \rightarrow G_\gamma(x)$$

On suppose désormais que les lois de nos composantes normalisées appartiennent au domaine d'attraction de $G_\gamma(x) = \exp\left(-(1 + \gamma x)^{\frac{-1}{\gamma}}\right)$, avec $\gamma < 0$

Sous cette hypothèse, [IFA16] nous indique que l'estimateur :

$$x^* = X_{n:n} + X_{n-k:n} + \frac{1}{\log 2} \sum_{i=0}^{k-1} \log \left(1 + \frac{1}{k+1}\right) X_{n-k-i:n}$$

où $X_{i:n}$ désigne la statistique d'ordre i , converge presque-sûrement vers la borne supérieur du support pour toute suite $k(n)$ telle que

$$k(n) \rightarrow +\infty \quad \text{et} \quad \frac{k(n)}{n} \rightarrow 0$$

Cet estimateur possède une propriété intéressante comparativement aux estimateurs présentés dans [LdH06] (principalement dans les chapitres 2 et 3) basés sur l'estimation des différents paramètres a_n , b_n et γ : on a toujours $x^* \geq X_{n:n}$

On utilise le même estimateur, mais sur $-(X_n)$ pour l'estimation de la borne inférieure du support, en supposant à nouveau que cet échantillon suit une loi appartenant à $D(G_\gamma)$.

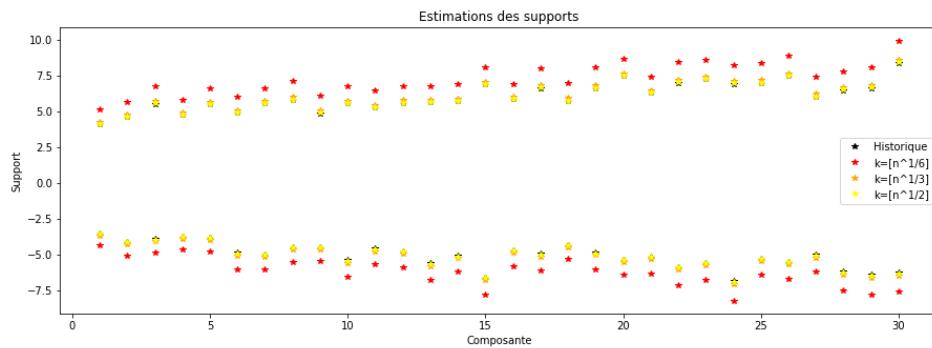


FIGURE 4.2 – Supports estimés des composantes normalisées.

4.2 Modélisation des composantes principales par des lois Kappa

Afin de reproduire le comportement particulier de chaque résidu et de chaque composante principale, nous choisissons de les modéliser par des lois à support compact. Il est relativement naturel de modéliser chaque composante, et d'autant plus chaque résidu, par une loi spécifique plutôt que par une loi générique. En effet il apparaît que les résidus ont un comportement statistique bien distinct en fonction de la localisation géographique étudiée.

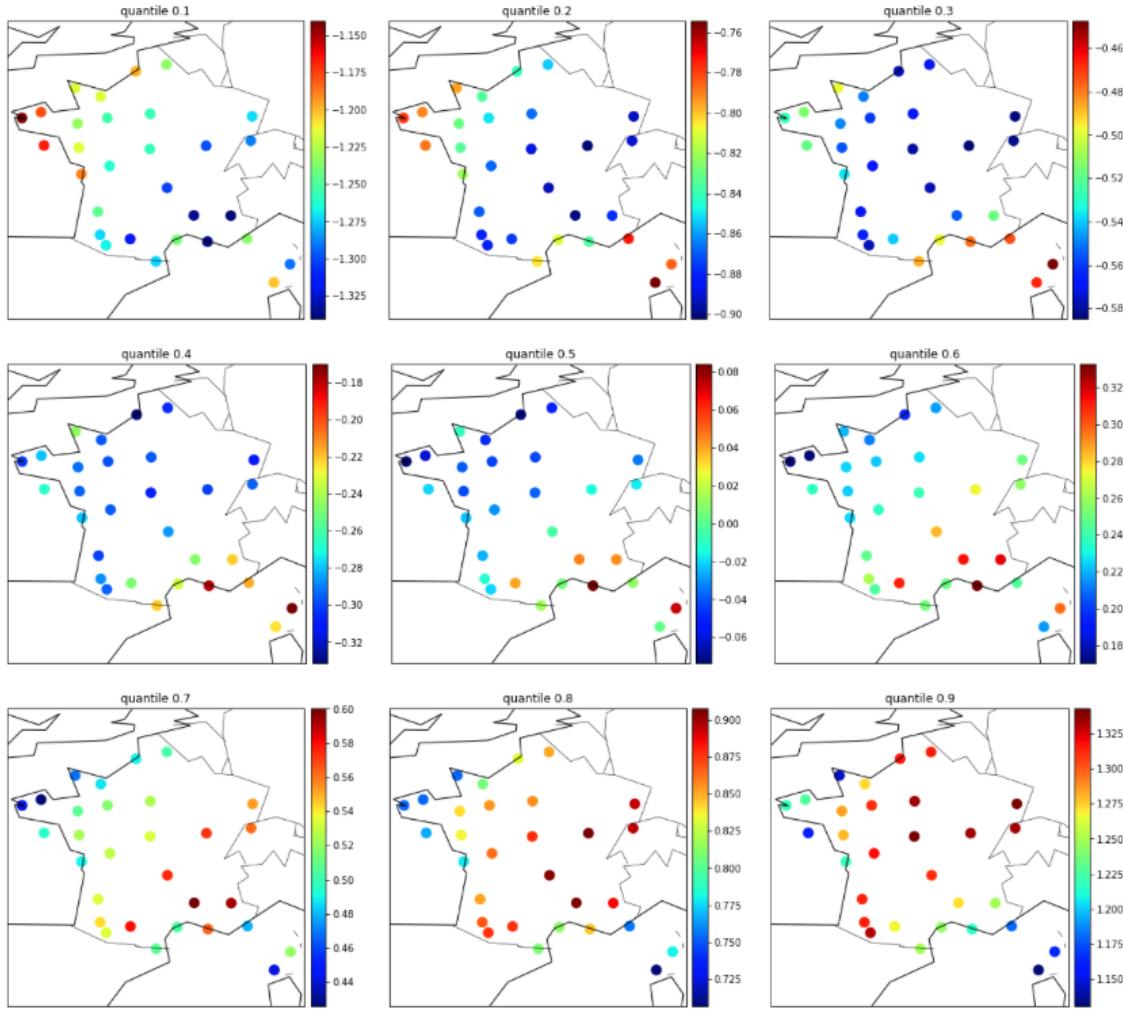


FIGURE 4.3 – Quantiles empiriques des résidus.

La loi Kappa à 4 paramètres $\kappa(\xi, \alpha, k, h)$ [Hos94], est définie par la fonction de répartition suivante, où $\xi, k, h \in \mathbb{R}$ et $\alpha > 0$ (en considérant les cas limites $k = 0$ et $h = 0$) :

$$F(x) = \{1 - h[1 - k(x - \xi)/\alpha]^{1/k}\}^{1/h} \quad \forall x \in I$$

Avec :

- $I =] -\infty, +\infty[\quad \text{si } k = 0 \text{ et } h \leq 0$
- $I = [\xi + \alpha \log(h), +\infty[\quad \text{si } k = 0 \text{ et } h > 0$
- $I = [\xi + \alpha/k, +\infty[\quad \text{si } k < 0 \text{ et } h \leq 0$
- $I = [\xi + \alpha(1 - h^{-k})/k, +\infty] \quad \text{si } k < 0 \text{ et } h > 0$
- $I =] -\infty, \xi + \alpha/k] \quad \text{si } k > 0 \text{ et } h \leq 0$
- $I = [\xi + \alpha(1 - h^{-k})/k, xi + \alpha/k] \quad \text{si } k > 0 \text{ et } h > 0$

On peut ainsi en déduire la densité de la loi Kappa :

$$f(x) = \alpha^{-1}[1 - k(x - \xi)/\alpha]^{\frac{1}{k}-1} F^{1-h}(x) \mathbb{1}_I(x)$$

Cette loi a l'avantage de pouvoir reproduire un grand nombre de densités. Les lois GEV et Gumbel, par exemple, sont ainsi des cas particuliers de la loi Kappa. On constate empiriquement que la loi Kappa reproduit assez fidèlement le comportement de "milieu" de distribution de nos composantes principales.

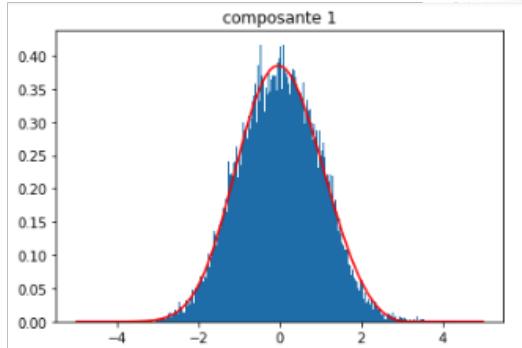


FIGURE 4.4 – Densité de la loi Kappa estimée pour la première composante principale.

Le principal problème de modélisation lié à la loi Kappa est ainsi son support. Pour des raisons de cohérence, nos résidus de températures ne peuvent excéder une certaine valeur. C'est pourquoi nous introduisons la loi Kappa tronquée sur $[a, b]$, définie par sa fonction de répartition :

$$\forall x \in [a, b] \quad F_{a,b}(x) = \frac{F(x) - F(a)}{F(b) - F(a)}$$

Cette loi admet évidemment une densité, qui à la différence de sa version non tronquée, a deux points de discontinuité (en a et en b). L'intérêt de cette loi tronquée, d'un point de vue modélisation, est qu'elle respecte le support estimé tout en ne s'éloignant que très peu de la loi non tronquée au niveau du "milieu" de la distribution.

COMMENTAIRE (A REVOIR - QUE SE PASSE-T-IL AVEC DES COMPOSANTES RESCALEES?) : Dans la pratique, un problème se pose avec la loi Kappa : il apparaît finalement parfois (et en particulier sur la première composante principale) que le minimum ou le maximum théorique entre en conflit avec le support obtenu par estimation. Le notebook LoiResidusComposantes est plus précis à ce sujet. Un autre souci se pose aussi, comme précisé dans mes rapports, au niveau de l'estimation des paramètres pour certaines composantes principales, c'est pourquoi je ne développe pas l'aspect estimation pour l'instant.

4.3 Modélisation des composantes principales par des lois Beta

La loi Kappa tronquée, explorée dans la section précédente, donne des résultats très satisfaisants au niveau du corps de la distribution, mais souffre d'un manque de fléxibilité et de cohérence au niveau du support.

On s'intéresse désormais à la modélisation de nos composantes principales par des lois Béta. Rappelons que la loi Béta est caractérisée par deux paramètres de forme $\alpha > 0$ et $\beta > 0$. Il s'agit d'une loi continue par rapport à la mesure de Lebesgue sur \mathbb{R} , de densité :

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbb{1}_{[0,1]}(x)$$

Où Γ est la fonction gamma usuelle définie par $\Gamma(x) = \int_{\mathbb{R}^+} z^{x-1} \exp(-z) dz$

Ainsi la loi bêta est à support $[0, 1]$. Nous avons cependant vu précédemment que nos composantes principales ont des supports variés, et tous différents de celui de la loi bêta. C'est pourquoi nous procédons de la façon suivante, en notant (C_i) les variables issues d'une composante principale arbitraire (normalisée), et s_1, s_2 les estimations respectivement des bornes inférieures et supérieures du support de la loi de notre composante principale (réduite) :

On pose $(\tilde{C}_i) = \frac{(C_i) - s_1}{s_2 - s_1}$. On a bien que les (\tilde{C}_i) suivent une loi à valeurs dans $[0, 1]$. On estime les paramètres optimaux d'une loi bêta considérant les observations (Y_i) par maximisation de la vraisemblance (Non respect de l'hypothèse d'indépendance. A améliorer ?)

Par cette procédure, on peut ainsi obtenir une loi bêta généralisée adaptée aux (C_i) initiaux. En notant f la densité de la loi bêta adaptée aux (Y_i) , on peut déduire par changement de variable que la densité de la loi généralisée associée au (C_i) s'écrit :

$$g(x) = \frac{1}{s_2 - s_1} f\left(\frac{x - s_1}{s_2 - s_1}\right)$$

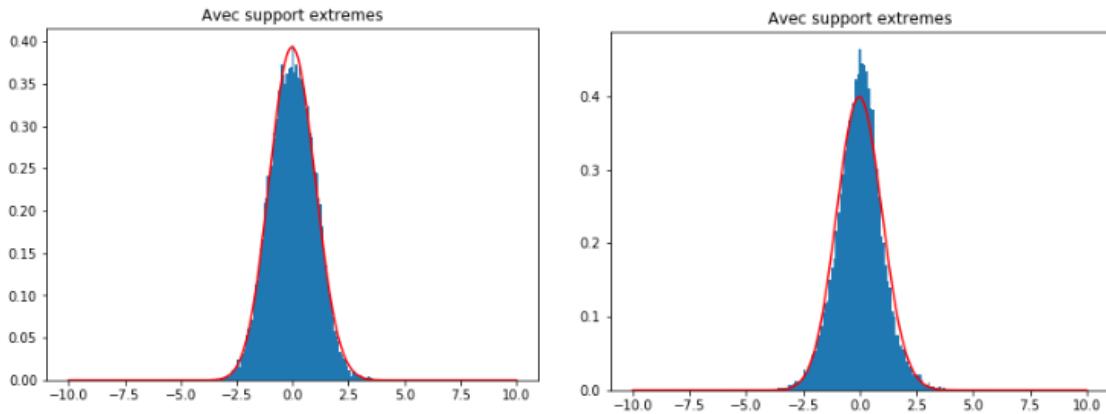


FIGURE 4.5 – Densités des betas généralisées estimées pour la 1^{ere} composante principale (gauche) et la 3^{ème} composante principale (droite). Les supports ont été estimé avec le paramètre $k(n) = n^{1/2}$.
GRAPHE A REFAIRE

Les graphiques précédents permettent de réaliser, graphiquement, que la concordance de la loi beta généralisée au milieu de distribution est très hétérogène en fonction des stations. Concernant les queues de distribution, la renormalisation de nos données à l'aide des supports estimés par les valeurs extrêmes (au lieu des support historique) permet en effet de générer, avec une probabilité particulièrement faible, des valeurs plus extrêmes que les valeurs historiques. Il serait intéressant d'observer la différence entre les quantiles empiriques de nos données et les quantiles théoriques des lois bêta généralisées obtenues (à faire).

4.4 Sur l'indépendance des composantes principales

Par propriété de la PCA, nous savons que nos composantes principales sont décorrélées : il n'existe pas de fonction linéaire reliant les valeurs de nos composantes principales. Cependant, cela ne nous indique pas l'inexistence de structures de dépendance plus complexes.

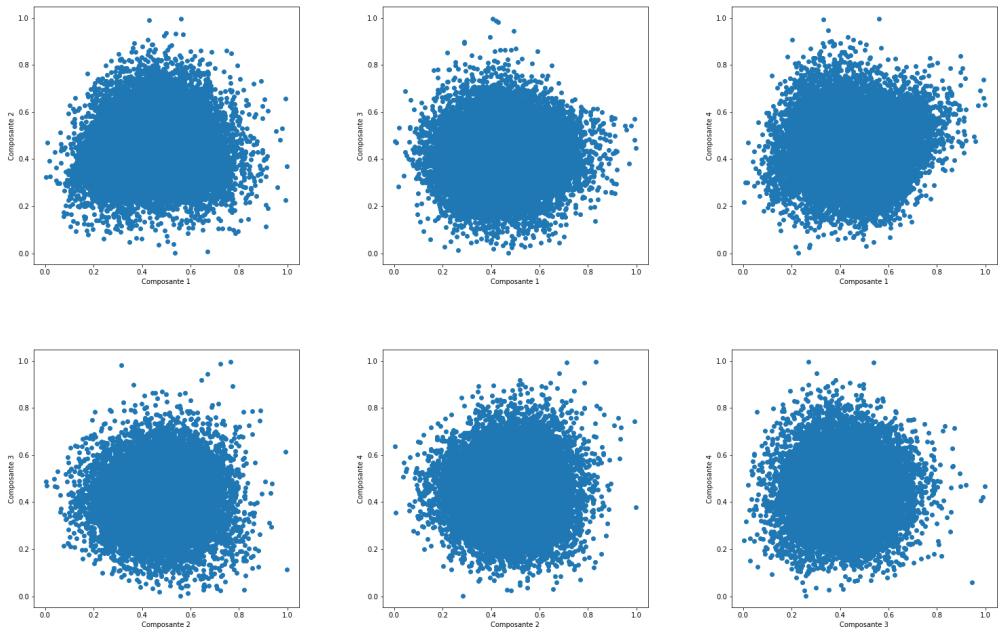


FIGURE 4.6 – Scatterplots des composantes principales (normalisées et rescalées) "significatives".

Notre but étant de reproduire les vagues de chaleur, il est intéressant d'observer la structure de dépendance lors de la période d'été (20 juin - 22 septembre) :

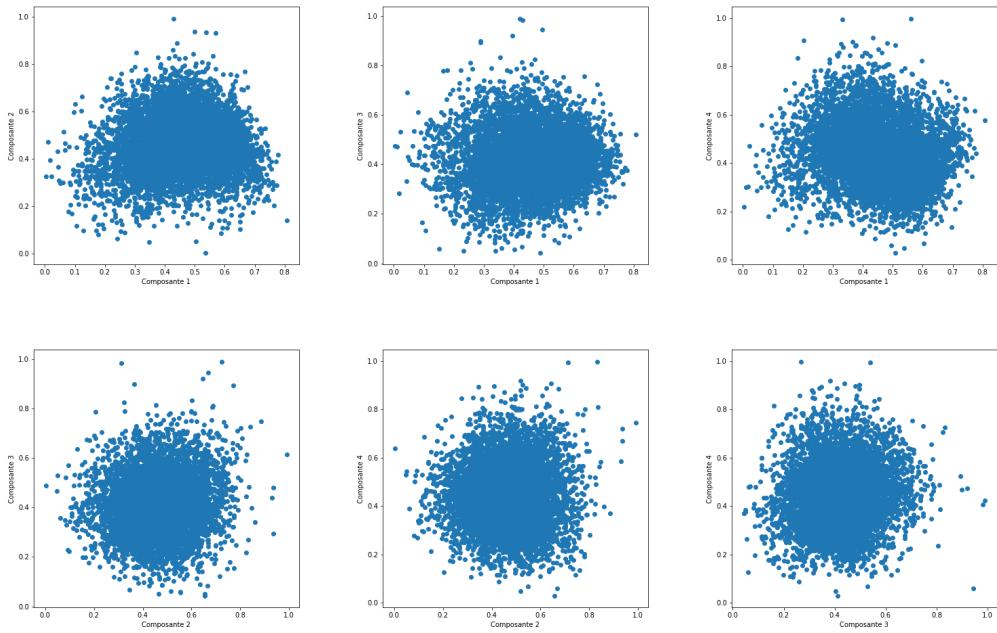


FIGURE 4.7 – Scatterplots des composantes principales (normalisées et rescalées) "significatives" (été uniquement).

L'étude des scatterplots précédents permet de constater, graphiquement, qu'il ne semble pas y avoir de dépendance significatives entre nos composantes principales. On peut ainsi considérer, à des fins de modélisation, que nos composantes principales sont indépendantes entre elles.

Chapitre 5

Modélisation des trajectoires

Dans les chapitres précédents, nous avons étudié le comportement empirique de nos résidus dans leur ensemble, qui présentent une structure de processus de par la présence de corrélations aussi bien temporelles que géographiques. Nous avons aussi étudié la modélisation "marginale" de nos résidus, par le biais de la modélisation des composantes propres. Dans ce nouveau chapitre, nous abordons la modélisation des trajectoires, à l'aide des résultats obtenus précédemment.

5.1 Modèle de diffusion basé sur la loi marginale et l'autocorrélogramme [Bibby-Sorensen]

5.1.1 Généralités et aspects théoriques

Notre première approche est de considérer les trajectoires de nos composantes principales comme la discrétisation d'un processus de diffusion. A cette fin, [BMB05] nous donne une méthode de construction d'un processus de diffusion ayant, premièrement, la bonne loi marginale, deuxièmement un autocorrélogramme cohérent avec nos observations.

Plus précisément, le théorème 2.1 de [BMB05] nous indique que si une densité de probabilité f par rapport à la mesure de Lebesgue λ sur \mathbb{R} respecte l'hypothèse **(H)** suivante :

(H) : f est bornée, continue sur \mathbb{R} , strictement positive sur son support $]s_1, s_2[$ ($-\infty \leq s_1 < s_2 \leq +\infty$), et telle que :

$$\int_{\mathbb{R}} x^2 f(x) d\lambda(x) < +\infty$$

Alors :

Théorème 5.1. Soit l'équation différentielle **(E)** définie par :

$$(E) \quad dX_t = -\theta(X_t - \mu)dt + \sqrt{v(X_t)}dB_t$$

où f est une densité de probabilité d'espérance μ satisfaisant (H) et telle que $\theta > 0$, (B_t) est un mouvement Brownien unidimensionnel standard, et v est définie pour tout $x \in]s_1, s_2[$ par :

$$v(x) = \frac{2\theta \int_{s_1}^x (\mu - z)f(z)d\lambda(z)}{f(x)}$$

Alors (E) a les propriétés suivantes :

1. (E) admet une unique solution (faible) markovienne.
2. Le processus $(X_t)_t$ solution de (E) est ergodique. La loi stationnaire de $(X_t)_t$ admet f comme densité par rapport à la mesure de Lebesgue.
3. Si $X_0 \sim f$, alors $(X_t)_t$ est stationnaire et admet une fonction d'autocorrélation donnée par :

$$\gamma(h) = \exp(-\theta|h|) \quad \forall h \in \mathbb{R}$$

Remarque 5.1. La condition (H) sur la fonction v peut se réécrire :

$$v(x) = 2\theta \frac{\mu F(x) - \int_{s_1}^x z f(z) d\lambda(z)}{f(x)} \quad \forall x \in]s_1, s_2[$$

Où F est la fonction de répartition associée à la densité f .

Ainsi, en utilisant les résultats du chapitre précédent concernant la modélisation des composantes principales, il est possible de construire pour chaque composante principale un processus de diffusion ayant de bonne propriétés de loi marginale et de structure de corrélation. L'idée principale est ainsi de construire de tels processus pour simuler les trajectoires de nos composantes principales, et d'ensuite appliquer la transformation inverse de notre ACP aux trajectoires simulées, afin d'obtenir une simulation des trajectoires des résidus.

A noter que nous supposons ici, en utilisant des mouvements Browniens indépendants pour chaque composante que les composantes principales sont indépendantes entre elles. C'est une hypothèse forte mais qui est admissible, comme observé en section 4.4

D'un point de vue pratique, se pose la question du nombre de composantes principales à utiliser. Nous sommes en effet confrontés ici à deux erreurs qui évoluent de manières opposées en fonction du nombre de composantes principales :

- L'erreur de reconstruction, qui comme vu précédemment, décroît avec le nombre de composantes principales.
- Une erreur d'approximation, liée à l'approximation du support et des lois de nos composantes principales ainsi qu'à l'approximation des diffusions utilisées par notre modèle, qui croît avec le nombre de composantes principales.

Il est ainsi nécessaire, dans l'idéal, de simuler les trajectoires des 30 composantes principales, et d'utiliser par la suite une métrique permettant de quantifier l'erreur commise en fonction du nombre de composantes sélectionnées. Cette métrique, toujours dans l'idéal, devrait tenir compte des facteurs suivants :

- Adéquation à la loi marginale des résidus.
- Correspondance de la structure de corrélation temporelle (fonction d'autocorrélation).
- Correspondance de la structure de corrélation géographique.

5.1.2 Calibration de la fonction d'autocorrélation

Pour chaque composante principale, nous souhaitons obtenir le paramètre θ^* tel que $t \rightarrow \exp(-\theta^*|t|)$ reproduise au mieux l'autocorrélogramme de notre composante principale. On procède pour cela par régression linéaire.

Posons γ la fonction d'autocorrélation de notre composante principale considérée, nous pouvons nous ramener à un problème linéaire en considérant le logarithme de nos autocorrélations. Ainsi, en ne considérant que les autocorrélations de lag inférieur ou égale à 12, il suffit de prendre :

$$\theta^* = \arg \min_{\theta \in \mathbb{R}} \sum_{t=1}^{12} (\log(\gamma(t)) + \theta t)^2$$

5.1.3 Implémentation (temporaire)

Considérons comme précédemment une composante principale réduite dont on dénomme (\tilde{C}_i) les valeurs renormalisées telle qu'expliquée précédemment dans le cadre de la modélisation par loi béta. On souhaite simuler une trajectoire de cette composante principale, composée de 1500 variables.

Pour cela, on associe à (\tilde{C}_i) la meilleure loi béta classique tel que décrit précédemment. Notons f la densité associée à cette loi. On génère $S(0) \sim f$ et l'on obtient la suite de la trajectoire par l'utilisation d'un schéma d'Euler de pas $p = 1/25$

$$S\left(\frac{t}{p}\right) = S\left(\frac{t-1}{p}\right) - \theta^* \left(S\left(\frac{t-1}{p}\right) - \mu \right) p + \sqrt{pv\left(S\left(\frac{t-1}{p}\right)\right)} Z_t \quad \forall 1 \leq t \leq \frac{1500}{p}$$

Où $\mu = \mathbb{E}(S(0))$ et θ^* estimé par régression linéaire, et (Z_t) est une suite de variables aléatoires indépendantes suivant une loi normale centrée réduite.

On ne conserve que $S(t)$ pour $t = 0, \dots, 1500$, et on considère la trajectoire dénormalisée définie par :

$$S_d(t) = (s_2 - s_1)S(t) + s_1$$

Où s_1 et s_2 sont les bornes estimées (par méthode des valeurs extrêmes) du support de notre composante principale.

On répète la procédure précédente pour toutes les composantes principales. Une fois fait, pour obtenir la trajectoire des résidus dans leur ensemble, on applique la transformation inverse de notre ACP aux trajectoires dénormalisées (et multipliées par $\sqrt{\lambda_i}$ où λ_i est la valeur propre associée à la composante i).

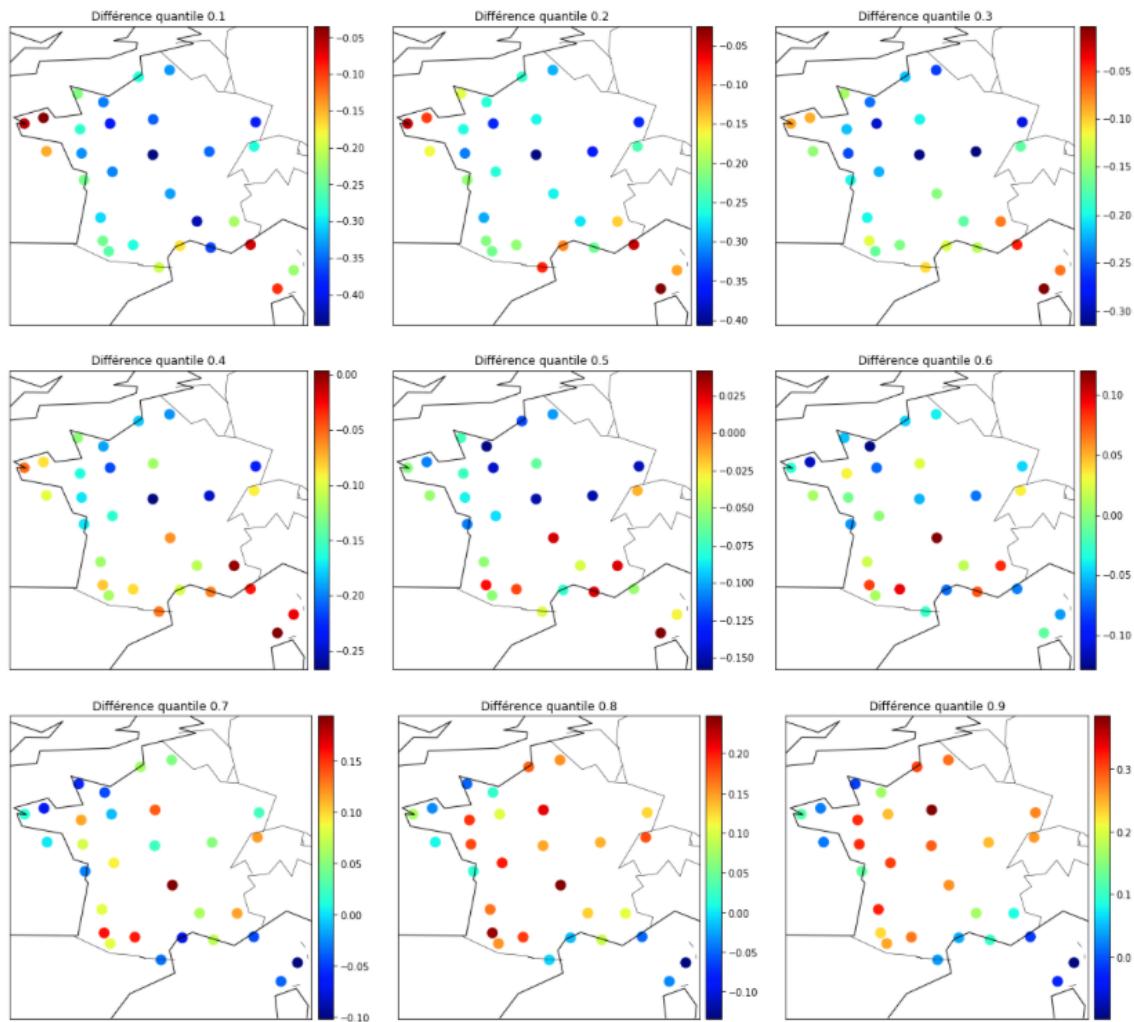


FIGURE 5.1 – Différence entre quantiles empiriques des résidus et quantiles empiriques de la trajectoire simulée (Euler, taille : 1500, pas : 1/25)

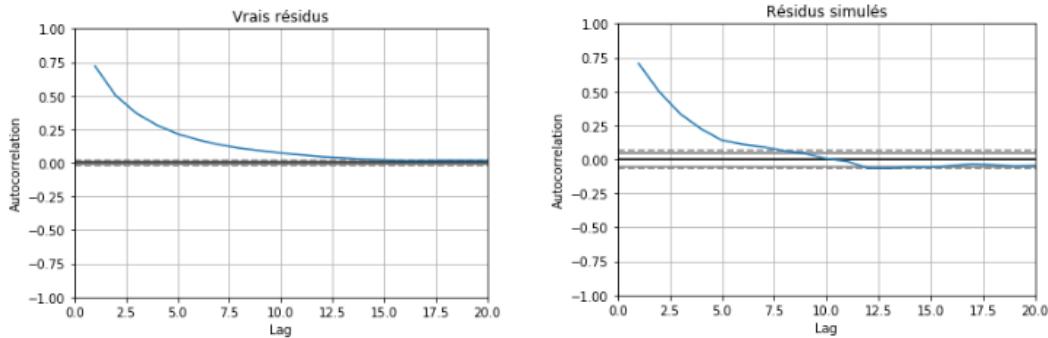


FIGURE 5.2 – Autocorrélation des résidus réels et des résidus simulés (Euler, taille : 1500, pas : 1/25) pour la station 32

5.2 Modèle de diffusion avec mean-reversion dépendant d'un paramètre stochastique.

5.2.1 Idée générale

Lors de l'implémentation du modèle précédent, il a été constaté que le modèle, bien qu'ayant la bonne loi stationnaire, reproduisait de manière assez défectiveuse les trajectoires des températures maximales. Il s'avère en effet que de par son drift induisant une propriété de mean-reversion, il était difficile d'obtenir deux jours d'affilé des résidus de température considérés comme extrêmes.

Pour pallier à ce défaut, une option est de considérer un processus ayant une force de rappel dépendant d'un paramètre lui-même aléatoire. Plus précisément, en considérant (X_t) une composante principale arbitraire, renormalisée et rescalée, on souhaite la modéliser grâce au système d'EDS (\mathbf{E}') suivant :

$$\begin{aligned} dX_t &= -R_t(X_t - \mu) + \sqrt{v(X_t, R_t)} dB_t \\ dR_t &= b(R_t)dt + \sigma(R_t)dW_t \end{aligned}$$

Où (B_t) et (W_t) sont deux mouvements browniens standards indépendants, $b()$ et $\sigma()$ deux fonctions bien choisies de telle sorte que (R_t) soit à valeurs dans $]0, +\infty[$, et $v(., .)$ est définie pour tout $(x, r) \in]s_1, s_2[\times]0, +\infty[$ par :

$$v(x, r) = \frac{2r \int_{s_1}^x (\mu - z)f(z)d\lambda(z)}{f(x)}$$

Où f est la densité de la loi stationnaire souhaitée pour notre composante principale, à support $]s_1, s_2[$, dans la même pensée que pour le modèle de Bibby-Sorensen.

Remarque 5.2. On admet pour l'instant qu'un tel système d'EDS admet non seulement une solution, mais aussi que le processus (X_t) admet la bonne loi stationnaire. Cela est cohérent empiriquement, voir l'exemple.

Il est ainsi imaginable, que sous de bonnes conditions, si (R_t) est stationnaire, (X_t) le soit aussi, avec comme loi invariante f . Il y aura cependant perte de propriétés sur l'autocorrélogramme, qui ne sera probablement plus de la forme $\exp(-\theta t)$

5.2.2 Exemple

On considère dans la suite (X_t) une composante principale arbitraire renormalisée et rescalée, f la loi bêta cible, et θ le paramètre optimal au sens de la régression linéaire pour fitter l'autocorrélogramme. On choisit de définir le processus (R_t) comme un processus de type "Bibby-Sorensen", de la manière suivante :

$$dR_t = -\beta(R_t - \theta)dt + \sqrt{v_R(R_t)}dW_t$$

Avec, pour tout $r \in]0, +\infty[$

$$\begin{aligned} v_R(r) &= \frac{2\beta \int_{s_1}^r (\mu - z)g(z)d\lambda(z)}{g(r)} \\ &= 2\theta\beta r \end{aligned}$$

Où g est la densité de probabilité d'une loi exponentielle de paramètre $\frac{1}{\theta}$. Il s'agit ainsi, toujours selon [BMB05] d'un processus admettant g comme loi invariante.

L'intérêt de cette expression pour le processus (R_t) est que, pour des valeurs suffisamment élevées de β (on considère dans la suite $\beta = 1$), l'autocorrélogramme se comportera approximativement comme la fonction $\exp(-\theta t)$

On génère une trajectoire de la composante principale à l'aide d'un schéma d'Euler de pas $p = 1/100$. C'est-à-dire : on génère $S(0) \sim f$, $R(0) \sim g$ et on définit $\forall 1 \leq t \leq \frac{24820}{p}$:

$$\begin{aligned} S\left(\frac{t}{p}\right) &= S\left(\frac{t-1}{p}\right) - R\left(\frac{t-1}{p}\right)\left(S\left(\frac{t-1}{p}\right) - \mu\right)p + \sqrt{pv\left(S\left(\frac{t-1}{p}\right), R\left(\frac{t-1}{p}\right)\right)}Z_t \\ R\left(\frac{t}{p}\right) &= R\left(\frac{t-1}{p}\right) - \beta\left(S\left(\frac{t-1}{p}\right) - \theta^*\right)p + \sqrt{pv_R\left(R\left(\frac{t-1}{p}\right)\right)}\tilde{Z}_t \end{aligned}$$

Où (Z_t) et (\tilde{Z}_t) sont deux suites indépendantes de variables aléatoires normales centrées réduites. Comme précédemment, on ne conserve que $S(t)$ pour $t = 1, \dots, 24820$ et on considère finalement le processus :

$$S_d(t) = (s_2 - s_1)S(t) + s_1$$

Où s_1 et s_2 sont les bornes estimées (par méthode des valeurs extrêmes) du support de notre composante principale.

On répète la procédure précédente pour toutes les composantes principales. Une fois fait, pour obtenir la trajectoire des résidus dans leur ensemble, on applique la transformation inverse de notre ACP aux trajectoires dénormalisées (et multipliées par $\sqrt{\lambda_i}$ où λ_i est la valeur propre associée à la composante i).

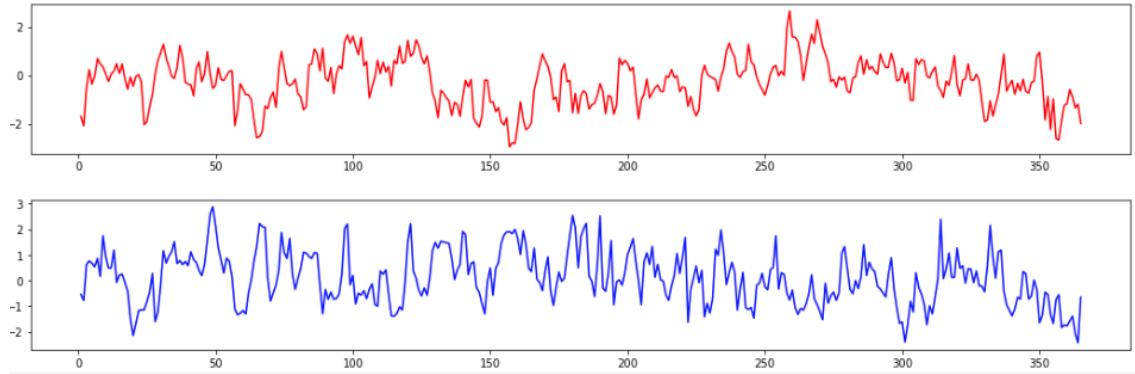


FIGURE 5.3 – Exemple de trajectoire obtenue sur l'année 1960 pour la station 32. Trajectoire historique en bleu, simulée en rouge.

On semble obtenir, graphiquement, de bonnes propriétés de persistance. On peut cependant s'interroger sur l'effet de changement de régime dans les résidus de températures : les résidus historiques semblent avoir des changements de régime plus brusques et plus fréquents que nos résidus simulés.

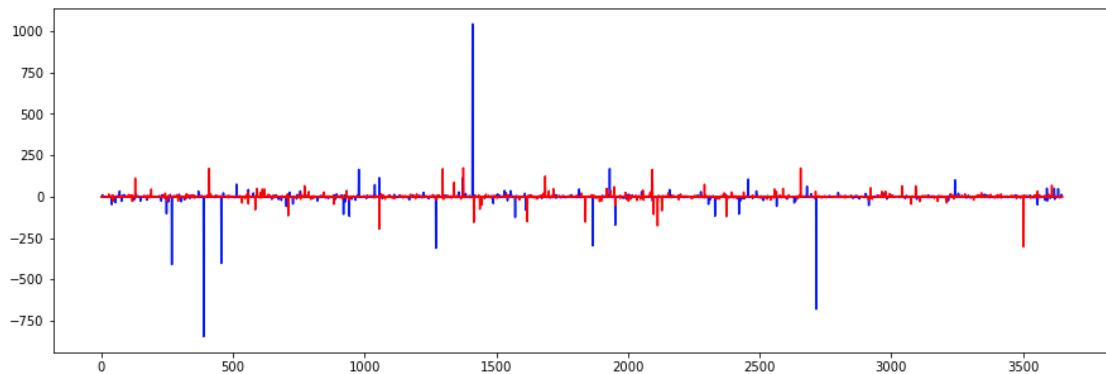


FIGURE 5.4 – Taux d'accroissement de la **station 33** sur la période 1960-1970. En bleu pour la trajectoire historique, en rouge pour la trajectoire simulée.

Cette analyse graphique peut-être "confirmée" par l'étude de la variation quadratique de nos processus historiques et simulés.

Station numéro 32
 Variation quadratique historique : 0.5592527183856387
 Variation quadratique simulée : 0.39913411816456723

Station numéro 33
 Variation quadratique historique : 0.6609728703667058
 Variation quadratique simulée : 0.39913411816456723

Station numéro 34
 Variation quadratique historique : 0.6408074577147225
 Variation quadratique simulée : 0.39913411816456723

FIGURE 5.5 – Variation quadratique historique et simulée pour une trajectoire de 24820 dates.

Il est cependant à noter que le taux d'accroissement "classique", c'est-à-dire en dehors des points de changement de régime, semble correspondre graphiquement à celui de la trajectoire historique.

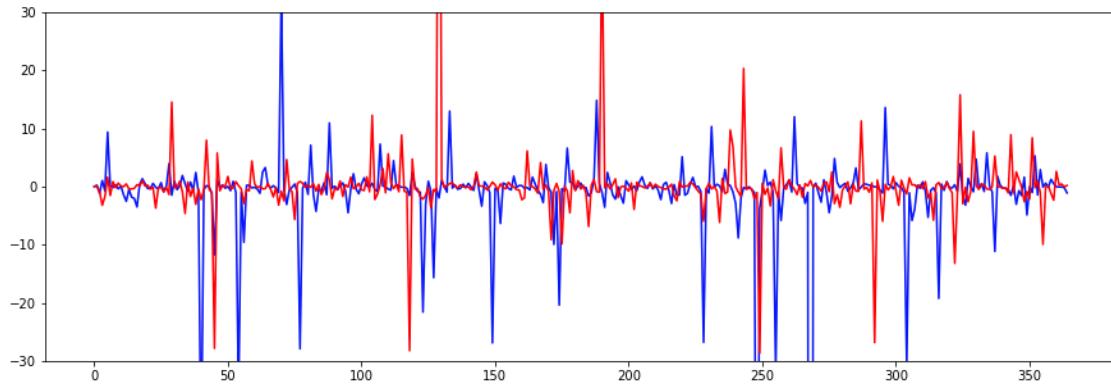


FIGURE 5.6 – Taux d'accroissement de la **station 33** sur l'année 1960. En bleu pour la trajectoire historique, en rouge pour la trajectoire simulée. **Ordonnées tronquées à [-30,30]**

On peut par ailleurs remarquer que plus le paramètre β est grand, plus le taux d'accroissement de nos composantes principales est volatile. Cependant, la taille de ce paramètre influe aussi sur la qualité de la simulation, et tend à influer sur les extrêmes obtenus.

Station 32 :
min empirique/simulé : -4.138779409368775 / -3.7200374464069537
max empirique/simulé : 3.9131710192157776 / 3.4264651425005566

Station 33 :
min empirique/simulé : -4.145378424884074 / -4.013527177729356
max empirique/simulé : 3.432461776138982 / 3.950965192157517

Station 34 :
min empirique/simulé : -3.8672875162250255 / -4.1220056822409035
max empirique/simulé : 3.4895549561895605 / 3.43356892642321

FIGURE 5.7 – Extrêmes historiques et simulés pour quelques stations (Euler, taille : 24820, pas : 1/100, $\beta = 1$)

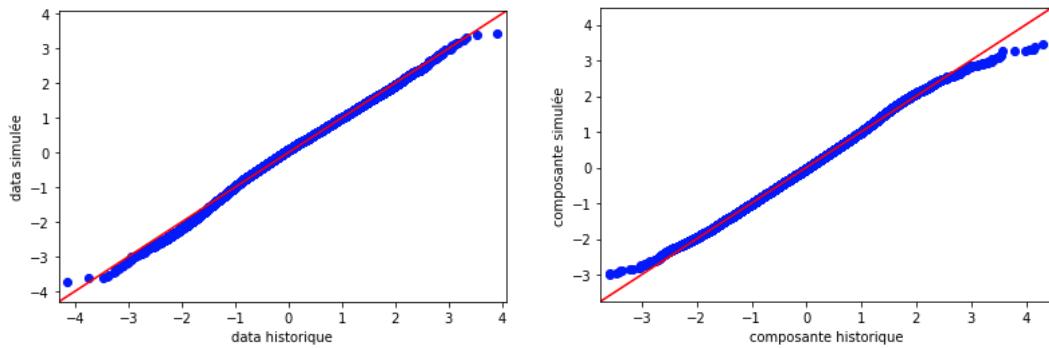


FIGURE 5.8 – QQ-plot des quantiles simulés contre les quantiles empiriques. Station 32 (gauche), première composante principale (droite)

A noter qu'on obtient un comportement assez atypique sur certaines stations. Cela est dû au fait qu'elles sont plus ou moins dépendantes des résultats de certaines composantes principales mal modélisées (la loi bêta ne prend pas en compte l'effet de skew de certaines composantes principales).

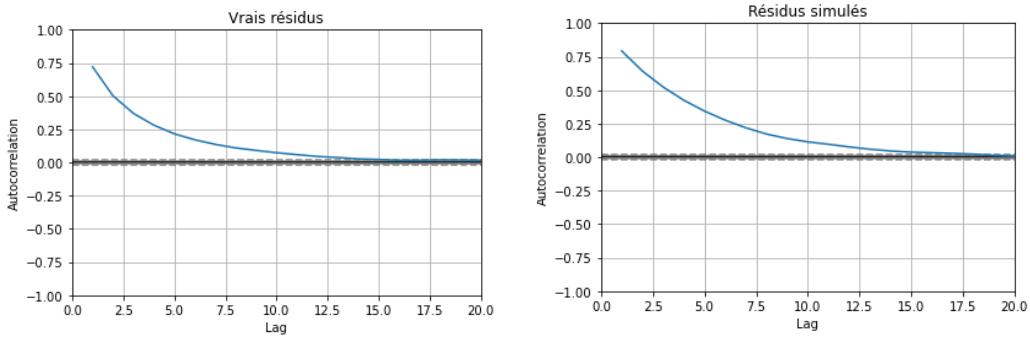


FIGURE 5.9 – Autocorrélation des résidus réels et des résidus simulés (Euler, taille : 24820, pas : 1/100, $\beta = 1$) pour la station 32

5.2.3 Estimation non paramétrique de la persistance

Supposons que le modèle exposé précédemment soit "vrai" au sens où, si l'on considère \tilde{C} une composante principale renormalisée et rescalée, alors elle suit l'équation différentielle stochastique définie précédemment :

$$d\tilde{C}_t = -R_t(\tilde{C}_t - \mu) + \sqrt{v(\tilde{C}_t, R_t)} dB_t$$

où (R_t) est un processus de diffusion..

Alors par le lemme d'Ito, la composante non-rescalée $C = (s_2 - s_1)\tilde{C} + s_1$ suit l'équation différentielle :

$$\begin{aligned} dC_t &= -(s_2 - s_1)R_t(\tilde{C}_t - \mu) + (s_2 - s_1)\sqrt{v(\tilde{C}_t, R_t)} dB_t \\ &= -R_t(C_t - s_1 - (s_2 - s_1)\mu) + (s_2 - s_1)\sqrt{v(\tilde{C}_t, R_t)} dB_t \end{aligned}$$

En remarquant que $(s_2 - s_1)\mu + s_1 \approx 0$ on peut approximer le comportement de cette diffusion par un schéma d'Euler de pas 1 :

$$C_i = (1 - R_i)C_{i-1} + (s_2 - s_1)v(R_i, C_{i-1})Z_i$$

où (Z_i) est une suite de variables indépendantes de loi normale centrée réduite.

On observe que ce schéma d'Euler définit un AR(1) à coefficients variant dans le temps. On peut alors extrapoler en supposant que (R_t) est déterministe et estimer sa trajectoire en chaque point i avec l'estimateur :

$$\hat{R}_i = - \left(\frac{\sum_{j=1}^n K\left(\frac{j-i}{b_n}\right) C_j C_{j-1}}{\sum_{j=1}^n K\left(\frac{j-i}{b_n}\right) C_{j-1}^2} - 1 \right)$$

où K est le noyau d'Epanechnikov défini par :

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{[-1,1]}(u)$$

et $b_n = n^{-8/11}$ (permet d'estimer sur un intervalle de 31 jours).

Remarque : Nous faisons ici certains écarts avec la théorie, en particulier sur le fait qu'on est forcé de supposer $|1 - R_i| < c$ avec $c < 1$.

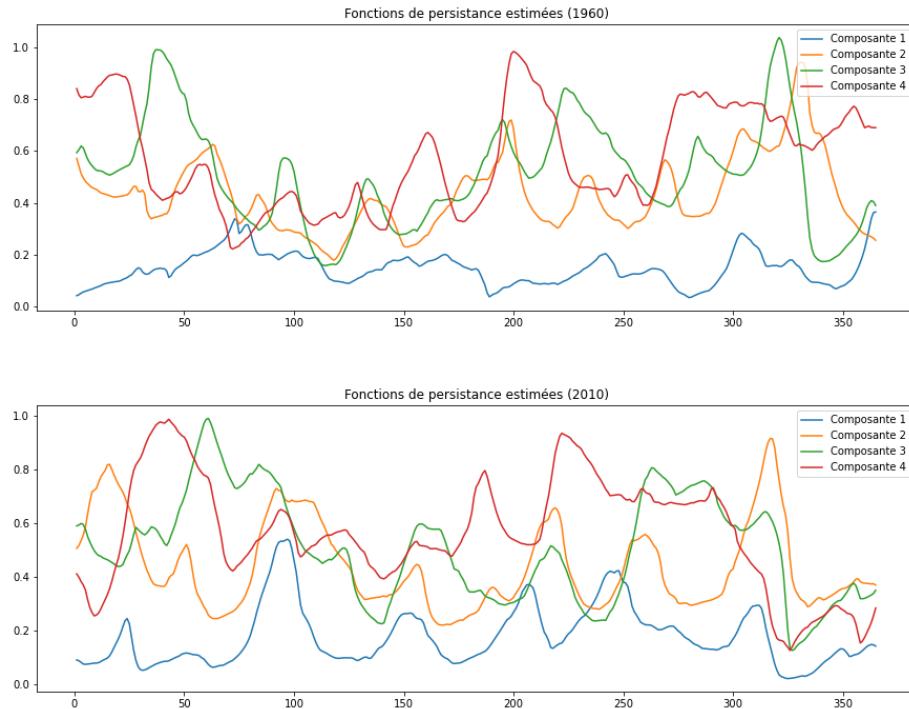


FIGURE 5.10 – Trajectoires obtenues sur l'année 1950 (haut) et l'année 2000 (bas) pour quelques processus de persistance.

Un phénomène intéressant est que tous les processus de persistance semblent avoir la même structure d'autocorrélation.

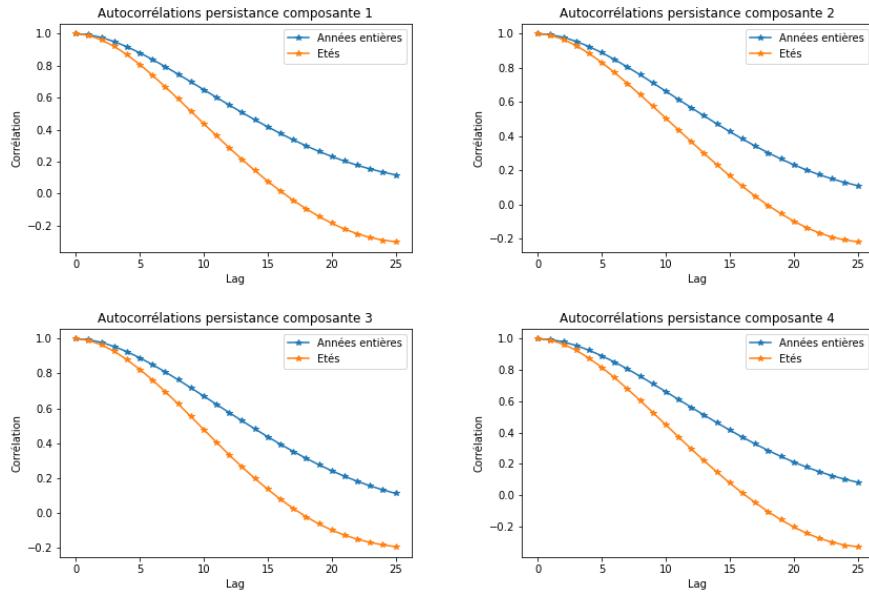


FIGURE 5.11 – Autocorrélations empiriques des processus de persistance.

Il ne semble pas y avoir de corrélation, ni de structure de dépendance entre les processus de persistance :

1.000000	0.052528	0.171141	0.198653
0.052528	1.000000	0.120231	0.138944
0.171141	0.120231	1.000000	0.198507
0.198653	0.138944	0.198507	1.000000

FIGURE 5.12 – Matrice de corrélation des quatre premiers processus de persistance.

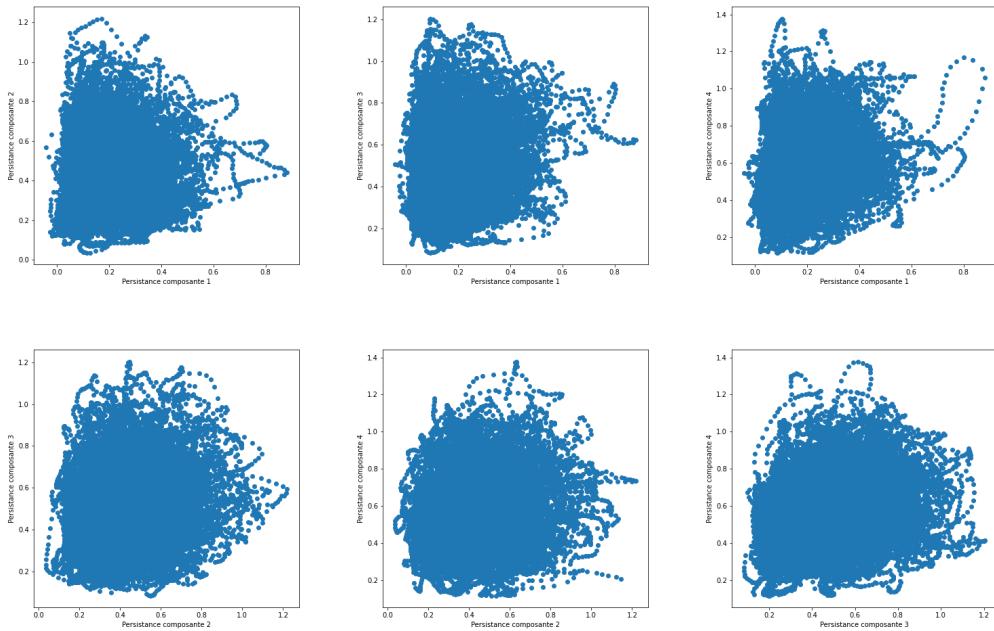


FIGURE 5.13 – Scatterplots des processus de persistance associés aux composantes "significatives".

Notre objectif étant de modéliser les vagues de chaleur, on peut s'intéresser de plus près à la corrélation et la dépendance entre ces processus uniquement sur la période 20 juin-22 septembre :

1.000000	0.044117	0.191347	0.172917
0.044117	1.000000	0.084693	0.107475
0.191347	0.084693	1.000000	0.183156
0.172917	0.107475	0.183156	1.000000

FIGURE 5.14 – Matrice de corrélation des quatre premiers processus de persistance (été uniquement).

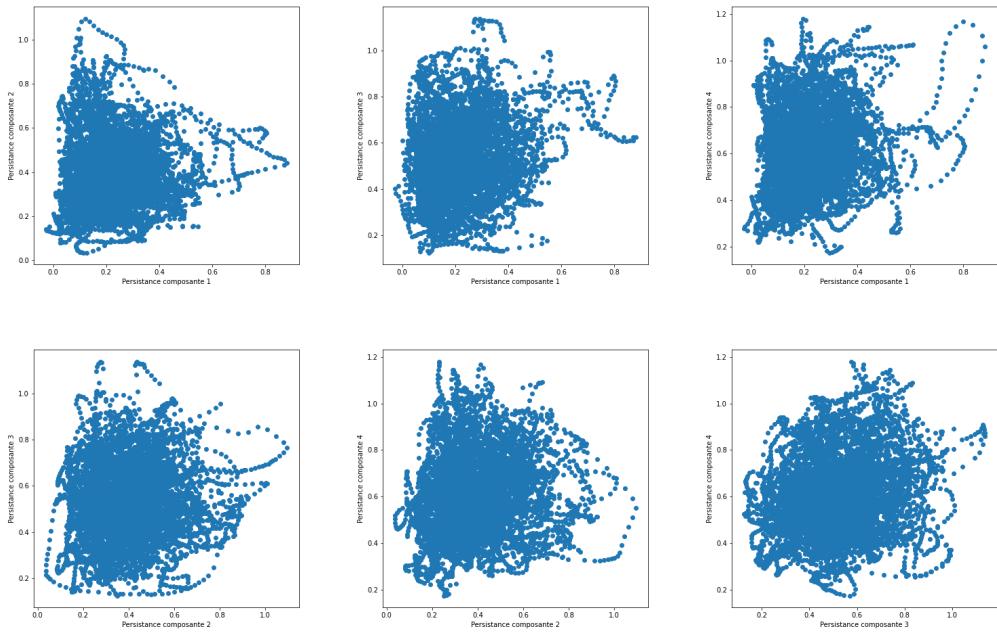


FIGURE 5.15 – Scatterplots des processus de persistance associés aux composantes "significatives" (étés uniquement).

5.2.4 A DEPLACER

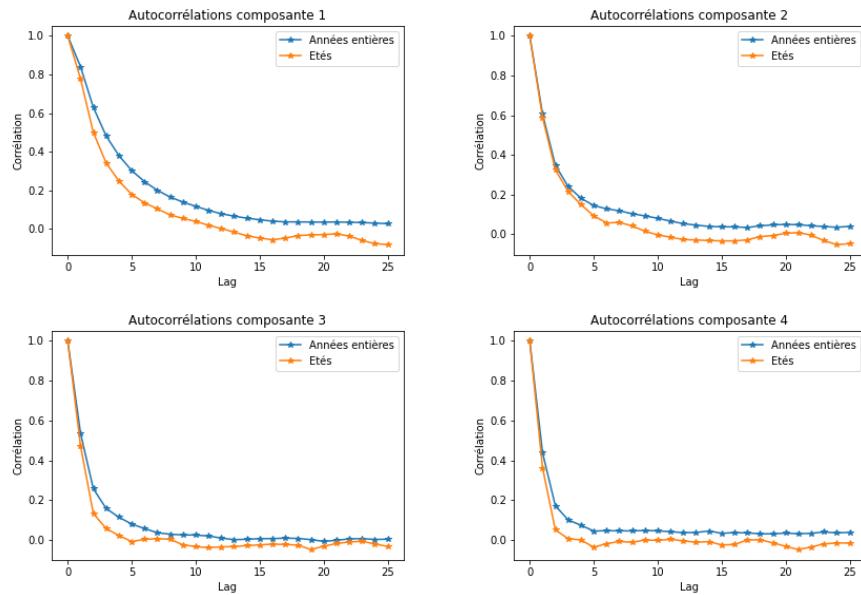


FIGURE 5.16 – Autocorrélations des composantes principales "significatives".

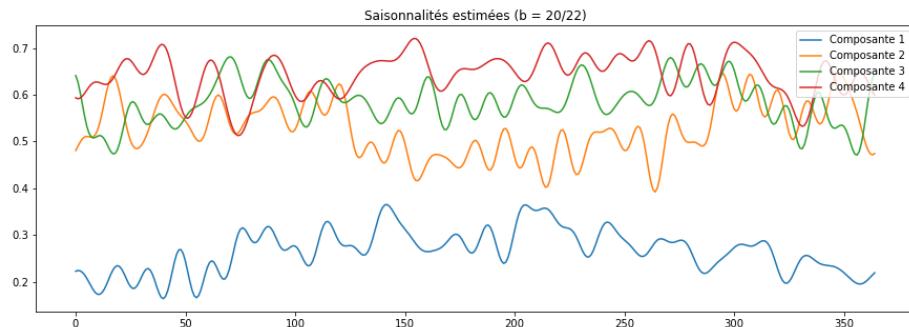


FIGURE 5.17 – Saisonnalités estimées de la persistance des composantes principales "significatives".

Bibliographie

- [BMB05] Michael Sorensen Bo Martin Bibby, Ib Michael Skovgaard. Diffusion-type models with given marginal distribution and autocorrelation function. *Bernoulli* 11(2), 2005.
- [Hos94] J. R. M. Hosking. The four-parameter kappa distribution. *IBM Journal of Research and Development*, may 1994.
- [IFA16] Pedro Rosario Isabel Fraga Alves, Claudia Neves. A general estimator for the right endpoint. *arXiv :1412.3972v3*, jun 2016.
- [LdH06] Ana Ferreira Laurens de Haan. *Extreme Value Theory : An Introduction*. Springer, 2006.
- [Sap06] Gilbert Saporta. *Probabilités, analyse des données et Statistique*. Technip, 2006.

Annexe A

Liste des stations

Stations retenues :

- 32 - BOURGES, FRANCE
- 33 - TOULOUSE-BLAGNAC, FRANCE
- 34 - BORDEAUX-MERIGNAC, FRANCE
- 36 - PERPIGNAN, FRANCE
- 39 - MARIGNANE, FRANCE
- 322 - RENNES-ST JACQUES, FRANCE
- 323 - STRASBOURG-ENTZHEIM, FRANCE
- 434 - BREST-GUIPAVAS, FRANCE
- 736 - ABBEVILLE, FRANCE
- 737 - LILLE-LESQUIN, FRANCE
- 738 - CAEN-CARPIQUET, FRANCE
- 740 - ALENCON, FRANCE
- 742 - NANTES-BOUGUENAIS, FRANCE
- 745 - DIJON-LONGVIC, FRANCE
- 749 - POITIERS - BIARD, FRANCE
- 750 - CLERMONT-FERRAND, FRANCE
- 755 - EMBRUN, FRANCE
- 756 - TARBES - OSSUN, FRANCE
- 757 - NICE, FRANCE
- 758 - BASTIA, FRANCE
- 786 - MONTELIMAR, FRANCE
- 793 - PTE DE LA HAGUE, FRANCE
- 2192 - BASEL-MULHOUSE, FRANCE
- 2203 - MONT-DE-MARSAN, FRANCE
- 2207 - MONTPELLIER-AEROPORT, FRANCE
- 2209 - AJACCIO, FRANCE
- 11244 - PTE DE CHASSIRON, FRANCE
- 11245 - PLOUMANAC'H, FRANCE
- 11247 - BELLE ILE - LE TALUT, FRANCE
- 11249 - ORLY, FRANCE