Statistiques Appliquées
Rapport de Projet

Risque en température - ENGIE

BERNARD Foucauld
LEWINSKI Grégory
STIZI Nina

Encadrant : DORAY Franck

15 mai 2023

Table des matières

Introduction

Né en 2008 de la fusion de Gaz de France (GDF) et de Suez, le Groupe GDF-Suez devient **ENGIE** en 2015, afin de faire face aux défis climatiques et environnementaux (décarbonation, développement des énergies renouvelables, réducation des consommations, digitalisation).
Le Groupe est présent sur toute la chaîne de valeur de l'énergie : de la production d'un mix peu carboné à la fourniture de solutions de production d'énergie et de services pour tous ses clients.

Dans le **secteur du gaz**, **ENGIE se positionne comme leader européen**. En effet, le groupe se classe :

- 1er réseau européen de transport de gaz (naturel),
- 1er réseau européen de distribution de gaz (naturel),
- 1ère capacité européenne de stockage de gaz (naturel).

Comme fournisseur de gaz européen, et a fortiori français, **ENGIE** doit remplir les obligations du service public (OSP) et de la Commission de Régulation de l'Énergie (CRE), en particulier **l'obligation de servir du gaz en pointe et en aléa**, c'est-à-dire être capable de :

- **aléa** : garantir l'approvisionnement en gaz même en cas d'hivers très rudes.
- **pointe** : garantir l'approvisionnement pendant 3 jours consécutifs extrêmement froids.

Ces problématiques, entre autres, obligent **ENGIE** à réaliser des **simulations de températures afin de déterminer les risques en température, et y répondre (anticiper l'approvisionnement, le stockage, etc.)**.

L'objet de ce projet de statistiques appliquées est donc de calculer des quantiles de températures dans le cadre de ces deux problématiques.

Pour ce faire, nous commencerons par calculer des quantiles pour des lois connues par Monte-Carlo, et illustrerons deux théorèmes de convergence. Puis, nous construirons un simulateur de températures, ce qui nous permettra *in fine* de calculer tout quantile pour répondre aux Obligations de Service Public.
Nous utiliserons les notations standards en mathématiques telles que rappelées en annexes.

I Rappels sur les convergences

Avant de réaliser le simulateur de températures, nous vérifions les principes de convergence de deux théorèmes fondamentaux de statisiques : la loi forte des grands nombres et le théorème de la limite centrale.

Afin de s'assurer de notre compréhension de ces théorèmes cruciaux, nous constatons empiriquement les convergences qu'ils garantissent. Nous ferons également un rappel sur la construction d'un intervalle de confiance et la convergence des quantiles empiriques vers les quantiles théoriques par méthode de Monte-Carlo. Cette méthode de calcul de quantiles, par Monte-Carlo, est utilisée à la fin du projet pour obtenir les risques en température.
Pour des raisons de concision, seuls des exemples avec des lois normales seront réalisés, sans prendre en compte le fait que dans le cas normal, les calculs sont explicites.

I.1 Loi forte des grands nombres

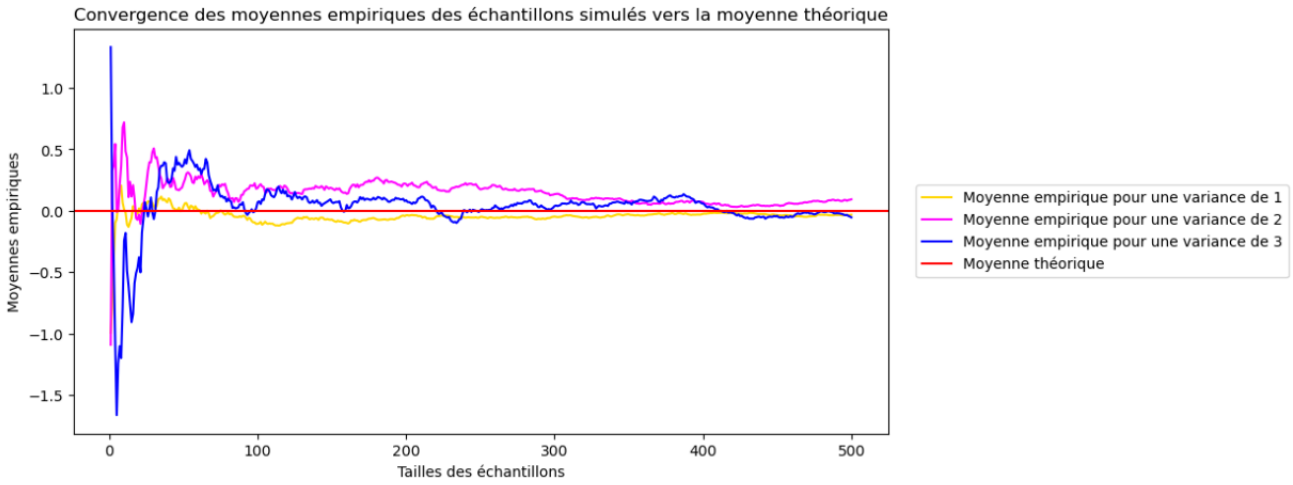
I.1.1 Énoncé de la loi forte des grands nombres

Soient $X_1, X_2, ..., X_n$ des variables aléatoires indépendantes, identiquement distribuées, et intégrables (i.e telles que $\mathbf{E}(|X_1|) < \infty$).
Alors :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s} \mathbf{E}(X_1).$$

I.1.2 Constatations empiriques : loi normale

A l'aide de Python, nous simulons un échantillon de variables aléatoires indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(\mu, \sigma^2)$ et observons l'évolution de la moyenne empirique de cet échantillon à chaque fois que sa taille augmente, c'est-à-dire à chaque fois que l'échantillon de taille n se voit ajouter une nouvelle variable gausienne pour devenir un échantillon de taille $n + 1$.
Par la loi forte des grands nombres, nous nous attendons à ce que la moyenne empirique de l'échantillon se rapproche de sa moyenne théorique μ à mesure que sa taille n croît. Par le théorème centrale limite, nous nous attendons à ce que la vitesse de convergence de l'échantillon soit de l'ordre de $\frac{\sqrt{n}}{\sigma^2}$. Le graphique ci-dessous montre que les résultats empiriques vont en ce sens et vérifient les attentes théoriques.



Nous constatons que la vitesse de convergence d'un échantillon de réalisations indépendantes et identiquement distribuées selon une loi normale d'espérance μ dépend de la variance de cette loi : plus la variance est grande, plus la vitesse de converge des moyennes empiriques des échantillons vers la moyenne théorique est petite. En d'autres termes, **plus la variance est petite, moins la taille de l'échantillon doit être grande pour s'approcher de l'espérance théorique**.

I.2 Théorème centrale limite et intervalles de confiance

I.2.1 Énoncé du théorème de la limite centrale

Soient $X_1, X_2, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées, et de carrés intégrables (i.e telles que $\mathbf{E}(X_1^2)<\infty$).
Alors on a :

$$\sqrt{n}\Big(\overline{X_n} - \mathbf{E}(X_1)\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\big(0,\, \mathbf{V}(X_1)\big).$$

I.2.2 Construction des intervalles de confiance

On souhaite construire un intervalle de confiance asymptotique à 95% de confiance pour $\overline{X_n}$.
On sait que pour $\alpha = 0.05$, $\quad q_{1-\frac{\alpha}{2}} = 1.96$
D'après ce qui précède (théorème centrale limite, loi forte des grands nombres, et lemme de Slutsky),

$$P\left(\left|\frac{\sqrt{n}\left(\overline{X_n}-\mu\right)}{\sigma}\right| \leq 1.96\right) \underset{n\to+\infty}{\longrightarrow} 95\%.$$

Or,

$$\left|\frac{\sqrt{n}\left(\overline{X_n}-\mu\right)}{\sigma}\right| \leq 1.96 \iff -1.96\cdot\sigma \leq \sqrt{n}\left(\overline{X_n}-\mu\right) \leq 1.96\cdot\sigma \iff \mu - \frac{1.96\cdot\sigma}{\sqrt{n}} \leq \overline{X_n} \leq \mu + \frac{1.96\cdot\sigma}{\sqrt{n}} \tag{1}$$

Donc, asymptotiquement, $I = [\overline{X_n} - \frac{1.96\cdot\sigma}{\sqrt{n}}, \overline{X_n} + \frac{1.96\cdot\sigma}{\sqrt{n}}]$ contient la moyenne empirique avec une probabilité de 0.95.

Ainsi, pour avoir une distance entre $\overline{X}_n$ et μ inférieure à 10^{-3}, il faut et il suffit d'avoir :

$$2\times\frac{1.96\cdot\sigma}{\sqrt{n}} \leq 2\cdot 10^{-3} \iff \frac{1.96\cdot\sigma}{\sqrt{n}} \leq 10^{-3} \iff \sqrt{n} \geq \frac{1.96\cdot\sigma}{10^{-3}} \iff n \geq \frac{(1.96\cdot\sigma)^2}{10^{-6}}$$

I.2.3 Conclusions sur l'analyse du théorème centrale limite

L'analyse menée précédemment nous permet de souligner que la taille nécessaire de l'échantillon pour que la moyenne empirique approche l'espérance théorique avec une précision de 10^{-3} ne dépend pas de la valeur de l'espérance théorique μ, mais seulement de la valeur de la variance σ^2 de la loi qui a servi à générer l'échantillon, et également de la taille de ce dernier (i.e. n). Des les faits, **le nombre de réalisations nécessaires à cette approche ne dépend même pas du type de loi qu'on utilise, mais bien uniquement de la variance choisie et de la taille de l'échantillon**. Le résultat est donc le même que ce soit un loi normale, une loi Gamma [1],...

On remarque aussi que **plus la variance est grande, l'échantillon doit être grand pour approcher l'espérance théorique**, comme nous l'avions déjà observé dans la partie I.1.2.
Ainsi, si $\sigma^2 = 1$ par exemple, il faut que $n \geq \frac{1.96^2}{10^{-6}} \iff n \geq 3841600$ (plus de 3 millions de réalisations).
Si $\sigma^2 = 100$ par exemple, il faut que $n \geq \frac{1.96^2\cdot 100}{10^{-6}} \iff n \geq 38416000$ (plus de 38 millions de réalisations).

I.3 Quantiles empiriques et convergence

I.3.1 Quantiles empiriques : définition et propriétés

Les quantiles empiriques sont des statistiques permettant d'estimer les quantiles d'une distribution inconnue, et *in fine* estimer la distribution elle-même.
Pour un niveau α, le α-quantile empirique est la valeur $x_{(\alpha)}$ telle que :

1. Durant ce stage, le même travail a été réalisé avec des lois Gamma.

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}[X_i \leq x_{(\alpha)}]$$

où $X_1, X_2, ..., X_n$ sont les données observées, et n est le nombre de données.
La fonction ci-dessus est la probabilité cumulée de la distribution estimée par les données observées, et le quantile empirique est la valeur pour laquelle la probabilité cumulée est égale à α.
En d'autres termes, **le quantile empirique est un estimateur non-paramétrique des quantiles** d'une distribution à partir de données observées.

La convergence du quantile empirique vers le quantile théorique est un résultat important en statistiques. Cela signifie que lorsque nous avons suffisamment de données observées, l'estimation du quantile empirique devient de plus en plus proche de la valeur théorique.
Nous pouvons démontrer cela en utilisant le théorème centrale limite, qui précise que pour une série de données provenant d'une population avec une moyenne finie et une variance finie, la moyenne des échantillons aléatoires tend vers la moyenne de la population. De manière similaire, nous pouvons montrer que le quantile empirique tend vers le quantile théorique pour une série de données assez grande.

Formellement, nous pouvons écrire cela comme suit :

$$\sqrt{n}(\hat{F}_n(x(\alpha)) - \alpha) \xrightarrow{d} \mathcal{N}(0, F'(x_{(\alpha)})^2)$$

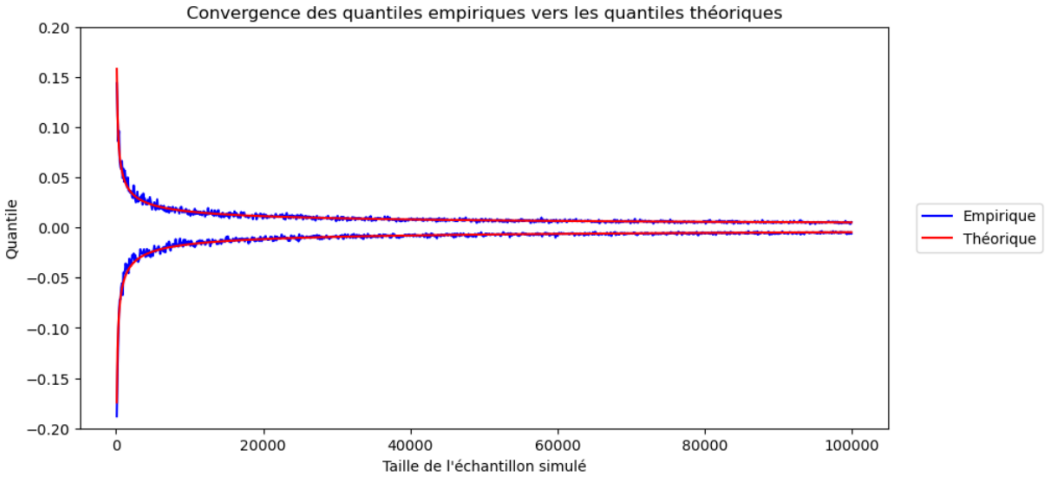
où :
- $\hat{F}_n(x(\alpha))$ est le quantile empirique pour un niveau de confiance α,
- $F(x_{(\alpha)})$ est la fonction de distribution théorique correspondante, $F'(x_{(\alpha)})$ est la dérivée de la fonction de distribution théorique,
- et $\mathcal{N}(0, F'(x_{(\alpha)})^2)$ décrit la distribution normale centrée autour de 0 avec une variance égale à $F'(x_{(\alpha)})^2$.

Cela signifie que **la différence entre le quantile empirique et le quantile théorique suit une distribution normale, et que la variance de cette différence tend vers 0 lorsque le nombre de données augmente**. Par conséquent, pour suffisamment de données, **l'estimation du quantile empirique devient de plus en plus proche de la valeur théorique**.

I.3.2 Convergence des quantiles empiriques

Nous pouvons dessiner plusieurs trajectoires de convergence des quantiles empiriques vers les quantiles théoriques. C'est la méthode de Monte-Carlo.

A l'aide de Python, de la même manière qu'en partie I.1.2, nous simulons des échantillons gaussiens et étudions la convergence des quantiles empiriques 5% et 95% vers leur pendant théorique. Nous obtenons les résultats suivants. Les résultats ci-dessous nous permettent de confirmer la convergence des quantiles empiriques vers les quantiles théoriques.



La courbe rouge représente la vitesse de convergence en $\frac{1}{\sqrt{n}}$, obtenue après calibration.

II Simulateur de températures

II.1 Motivation

Les obligations de service public (OSP) fixées par la Commission de Régulation de l'Énergie (CRE) précise qu'ENGIE doit être capable de fournir du gaz en tout temps sauf en cas de froid extrêmement intense [2]. Ceci implique une limite basse de températures qu'ENGIE doit savoir gérer, autrement dit, une limite haute de gaz devant être fourni. Ainsi, ENGIE, comme tout énergéticien doit être capable de passer des hivers froids, et pouvoir le montrer aux autorités compétentes.
Comme nous venons de voir, **la méthode de Monte-Carlo utilisée pour obtenir les quantiles nécessite un nombre conséquent de *samples* pour avoir un résultat proche de la valeur théorique**. En ce sens, il serait dommageable de calculer les quantiles (c'est-à-dire les risques en température) avec un historique de températures (en général de 20-30 ans). Si nous disposons, en effet, d'**un simulateur de températures**, nous bénéficierons d'un **nombre quasi illimité de scénari de températures, et ainsi pourrons-nous avoir suffisament de *samples* pour calculer un quantile par la méthode de Monte-Carlo avec une bonne précision**.

Nous allons donc créer un simulateur statistique, et le calibrer, et obtenir ainsi énormément de scénari de températures plausibles.

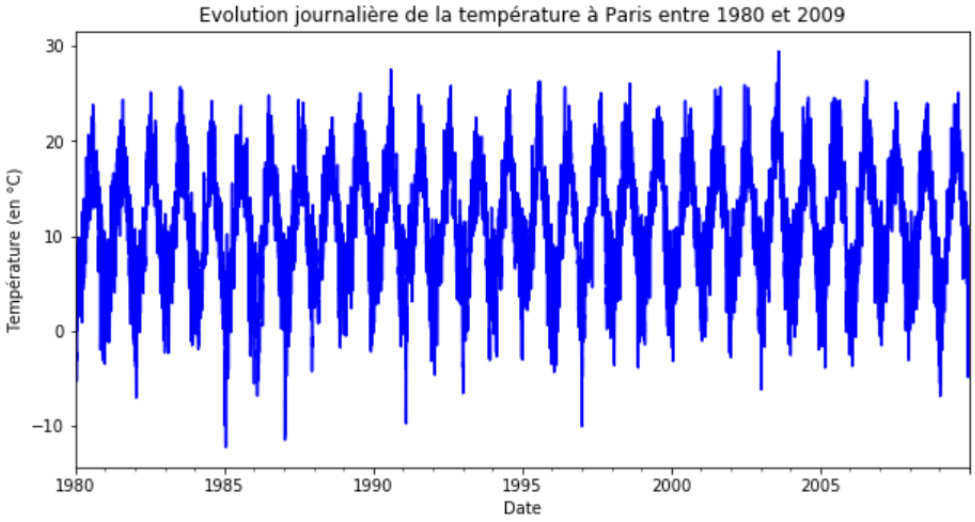
II.2 Données météorologiques et simulations

L'organisation mondiale de la météorologie impose d'avoir un historique de 20 ans minimum de températures pour calibrer un simulateur.
ENGIE nous a fourni 30 ans d'historique (1980-2009) de températures journalières à Paris. Notons que les températures fournies ont été vérifiées par des météorologues afin d'éviter toute erreur de relève. Aussi, la base de données suit des conventions précises (données rangées dans l'ordre chronologique, aucune donnée manquante...) qui nous épargne le nettoyage du jeu de données à notre disposition. **En somme, nous pouvons faire confiance aux données qui nous ont été fournies par ENGIE**.
Nous négligerons le réchauffement climatique [3]

II.3 Première analyse de l'historique de températures

A l'aide de Python, nous visualisons la série temporelle des températures, que l'on notera dès à présent $(X_t)_t$, et obtenons les premières éléments d'analyse statistique.

2. On consultera les textes de loi en annexes.
3. Il existe en effet des algorithmes pour réchauffer l'historique et ainsi prendre en compte cet effet.



Par ce premier niveau d'analyse, nous pouvons dire que :

- $(X_t)_t$ semble périodique de période égale à une année.
- $(X_t)_t$ semble suivre une "sinusoïde" de période un an.
- si la distribution des températures pouvait à première vue sembler symétrique (voire gaussienne), la réalité est autre puisque les moyennes et les médianes des températures ne coïncident pas. Cette représentation est disponible en annexes.
- si l'augmentation des températures est une réalité empirique, elle n'est pas linéaire ni triavallement prévisible. Comme évoqué précédemment, nous ne prendrons pas en compte cet effet.

Des graphiques complémentaires, disponibles en annexes, permettent d'apporter une justification et un fondement aux conclusions ci-avant établies.

II.4 Normales de saison

II.4.1 Écriture décomposée de $(X_t)_t$ en termes déterministe et stochastique

Si nous notons X_t la temperature à la date t, nous pouvons la réécrire $X_{a,j}$, ou j désigne le jour entre $\{1, \dots, 365\}$ et a l'année.
On suppose en effet dans toute la suite du projet que toutes les années ne comportent que 365 jours en se passant des 29 février. Cette approximation est possible puisqu'en 30 ans, se passer des 29 février représente une perte de 8 données sur presque 11 000, soit une perte largement négligeable.

Notons $S_j = \mathbf{E}_a[X_{a,j}]$, la moyenne journalière de températures, que l'on appelle **normale de saison** ou moyenne saisonnière.

Nous définissons alors $A_{a,j}$, l'anomalie de températures comme l'écart entre $X_{a,j}$, la température relevée au jour j de l'année a, et S_j, la normale saisonnière. Nous pouvons écrire :

$$X_{a,j} = S_j + A_{a,j},$$

où, par construction, le processus A est centré.

II.4.2 Approximation des normales de saison par séries de Fourier

Nous souhaitons rendre déterministe le terme de normale saisonnière. Ainsi, nous déterminons une fonction $f : j \mapsto f(j)$ qui approxime les températures moyennes journalières S_j pour tout jour $j \in \{1, \dots, 365\}$. Une telle approximation donne une **fonction déterministe et plus lisse** qui nous permettra, ensuite, de **simuler les températures en ayant uniquement l'anomalie de températures comme terme stochastique**.

Par la trajectoire périodique des températures, nous avons l'intuition que **la normale saisonnière peut être approximée par les séries de Fourier**.
Dans le cas général, si $f : \mathbf{R} \to \mathbf{R}$ est périodique de périodique $T > 0$, il existe des scalaires a_n et b_n tels que :

$$\forall x \in \mathbf{R},\ f(x) = \sum_n a_n \cos\left(\frac{2\pi n}{T}x\right) + b_n \sin\left(\frac{2\pi}{T}x\right),$$

sous de bonnes conditions de régularités sur f.
En particulier, la série converge. Ceci implique que des fonctions périodiques régulières peuvent être approximées par des polynômes trigonomtriques (en tronquant la série convergente ci-dessus). Autrement dit, on aura, pour un N fixé :

$$\forall x,\ f(x) = \sum_{n \leqslant N} a_n \cos\left(\frac{2\pi n}{T}x\right) + b_n \sin\left(\frac{2\pi n}{T}x\right) + R_N(x),$$

où R_N est "petit".

Si nous appliquons ces éléments théoriques au cas précis des températures, alors, en suivant l'exemple de décomposition des fonctions périodiques (applicable ici puisque les grandes variations de températures sont données par la saison), nous **cherchons la décomposition de la normale de saison comme une somme de polynômes trigonométriques**.

Par exemple en prenant $N = 2$:

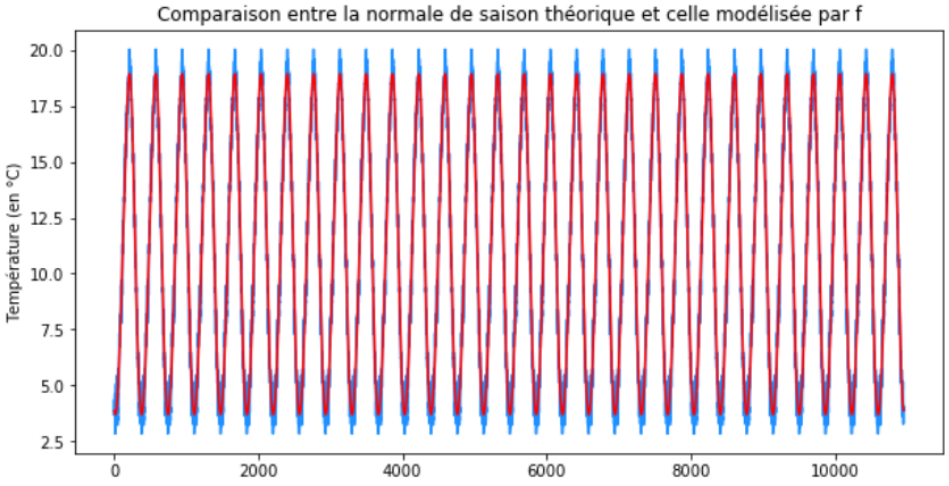
$$\begin{aligned} X_t = a_0 + & \\ & a_1 \cos\left(\frac{2\pi}{365}t\right) + b_1 \sin\left(\frac{2\pi}{365}t\right) + \\ & a_2 \cos\left(\frac{2\times 2\pi}{365}t\right) + b_2 \sin\left(\frac{2\times 2\pi}{365}t\right) + \\ & a_3 \cos\left(\frac{3\times 2\pi}{365}t\right) + b_3 \sin\left(\frac{3\times 2\pi}{365}t\right) + \\ & A_{a,j}, \end{aligned}$$

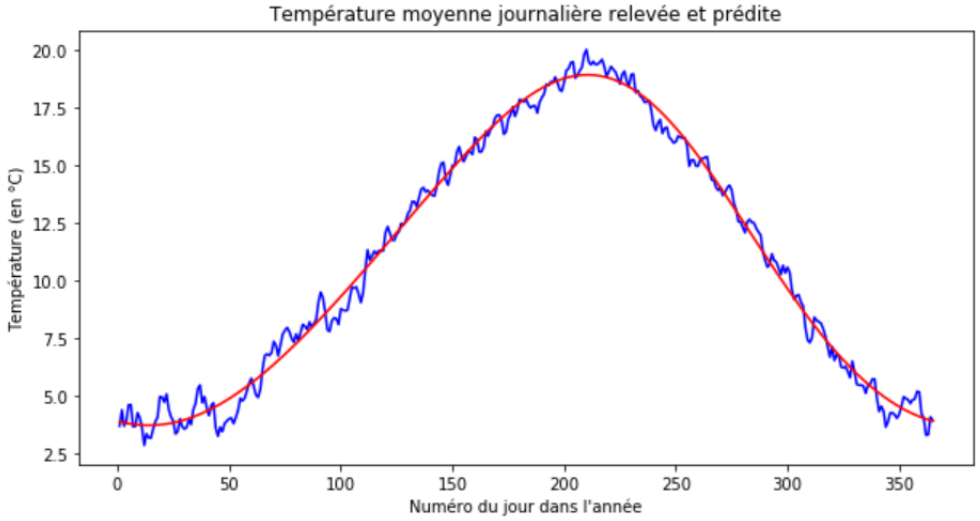
où A joue le rôle de R ci-dessus.

A l'aide de Python, nous observons que, passée le troisième harmonique de Fourier, l'approximation de la normale saisonnière par f est inchangée. Aussi, **nous nous contenterons de trois harmoniques de Fourier** dans la décomposition proposée ci-avant de la normale de saison.
A l'aide du *package* "sklearn", **nous déterminons par régression linéaire les coefficients** a_1, a_2, a_3 **et** b_1, b_2, b_3. Cela nous permet d'obtenir la fonction f qui approxime les normales saisonnières.
Les détails graphiques et numériques de notre démarche sont apportés en annexes.

II.4.3 Résultats et conclusions sur les normales de saison

En comparant f, en rouge, à la courbe représentative des moyennes journalières de températures empiriques, en bleu, nous constatons que les résultats sont très proches. f est donc une **excellente approximation des normales saisonnières**.





En résumé :
Comme anticipé, nous obtenons une fonction qui associe chaque jour $j \in \{1, ..., 365\}$ à sa normale de saison qui soit :

- déterministe,
- plus lisse que la fonction S obtenu par moyenne empirique sur l'historique de températures.

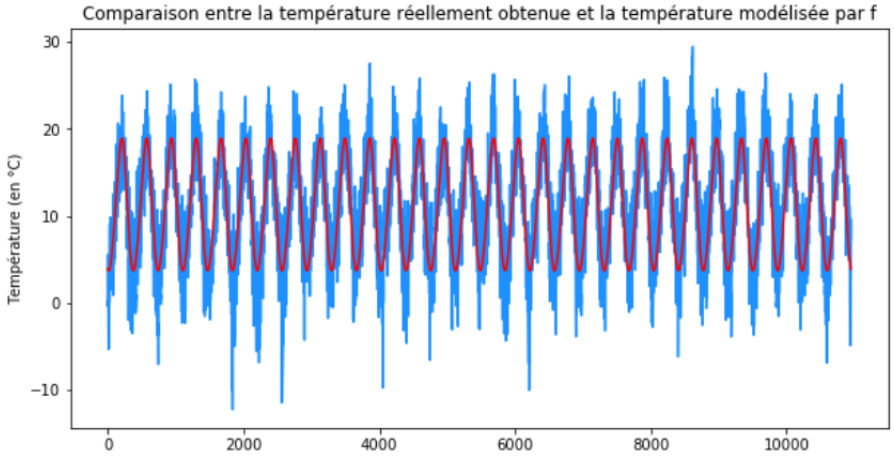
A partir de maintenant, nous appelerons

- normale de saison, l'image du jour $j \in \{1, ..., \}$ par la fonction f,
- anomalie de températures $A_{a,j}$ au jour j de l'année a, la différence entre la température prélevée $X_{a,j}$ à cette date et la normale de saison $f(j)$.

Nous pouvons donc réécrire le modèle :

$$\forall j \in \{1, ..., 365\}, \forall a \in \mathbf{Z}, \quad X_{a,j} = f(j) + A_{a,j}$$

L'analyse de l'anomalie de températures $(A_t)_t$ est indispensable à la bonne modélisation de la température $(X_t)_t$. En effet, même si f est une excellente approximation de la normale saisonnière, f **approxime très mal la température**, comme le montre le graphique ci-après.



Comprendre **l'anomalie de températures $(A_t)_t$ nous permettra de mieux modéliser la température $(X_t)_t$ et donc, de mieux les simuler**.

II.5 Anomalies de températures

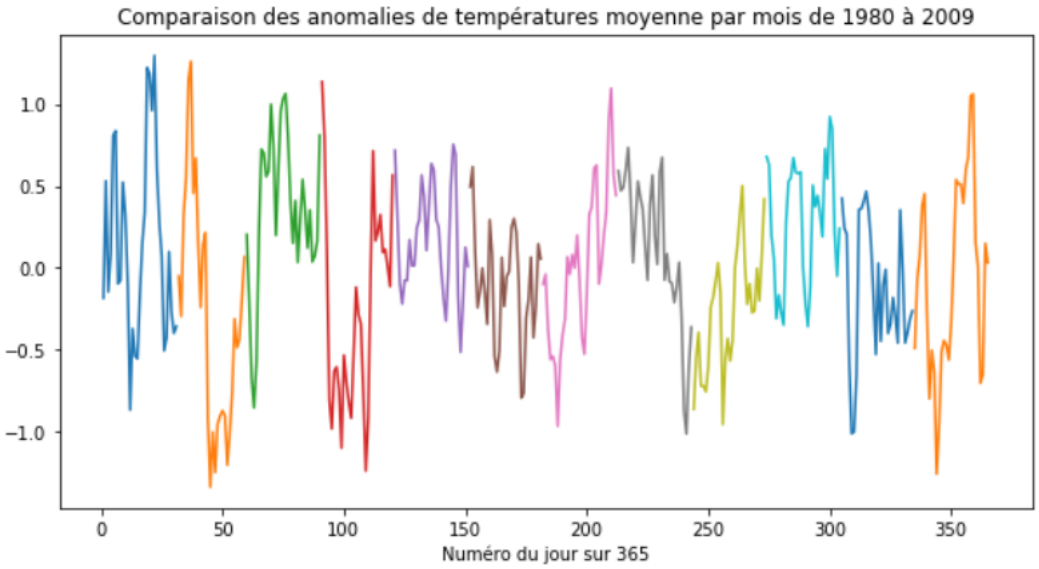
II.5.1 Description du processus d'anomalie de températures

Une première analyse graphique et de statistique descriptive nous permet de dire que :

- **l'anomalie de températures semble périodique, de période égale à 1 an (365 jours)**, en ce que les mêmes mouvements se répètent chaque année. Les graphiques en annexes et particulièrement le zoom visuel l'anomalie de températures sur 5 ans nous permettent de le voir encore plus clairement. Ainsi, l'anomalie de températures apparaît **365-périodique et ne semble pas être déterminée en fonction de l'année**.
- **empiriquement, l'anomalie de températures est de moyenne très proche de 0**, ce qui est en accord avec la construction centrée de l'anomalie $(A_t)_t$ que nous avons conduite au préalable.
- **l'anomalie de températures est très volatile** puisque son écart moyen à la moyenne est de -3°C ; +3°C.

Plus en détail, le graphique ci-dessous, qui identifie l'anomalie de températures moyenne mois par mois (chaque mois étant représenté par une couleur différente), nous montre que :

- **l'anomalie de températures est fonction du temps et n'est pas stationnaire sur l'année**.
- **l'anomalie de températures semble stationnaire par mois**



II.5.2 Stationnarité mois par mois de l'anomalie de températures

La stationnarité des anomalies de températures est indispensable à la construction d'un modèle auto-régressif pour leur prédiciton par la suite. La vérifier est donc indispensable. **Pour vérifier la stationnarité mensuelle des anomalies de températures, nous réalisons un test de Dickey-Fuller augmenté sur chaque processus d'anomalie de températures découpé de manière mensuel**. Autrement dit, nous découpons le processus $(A_t)_t$ en douze processus $(A_{m,j})_j$ pour m variant dans $\{1, ..., 12\}$, et réalisons le test de Dickey-Fuller augmenté sur chacun de ces processus. Le test de Dickey-Fuller augmenté pour tester que les racines unitaires de nos données sont bien nulle. Plus précisement, nous réalisons le test :

$\mathbf{H_0}$: Le processus $(A_{m,j})_j$ comporte au moins une racine unitaire.
$\mathbf{H_1}$: Le processus $(A_{m,j})_j$ ne comporte pas racine unitaire. Il est stationnaire.

Nous pouvons écrire mathématiquement l'hypothèse nulle comme s'en suit :

$$\mathbf{H_0} : 1 - \sum_{i=1}^{p} \phi_i.z^i = (1-z).(1 - \sum_{i=1}^{p-1}(-\phi_{i+1} - ... - \phi_p).z^i), \forall z \leq 1$$

où les ϕ_i sont les coefficients autorégressifs du processus $(A_{m,j})_j$.

Nous réalisons ce test avec Python, en utilisant la fonction "adfuller". Pour tout mois m, les résultats du test indique que les p-values sont toutes largement inférieures à 0.0001. **Nous pouvons donc considérer avec un niveau de confiance de plus de 99.99% que les processus $(A_{m,j})_j$ sont bien stationnaires, pour tout mois $m \in \{1, ..., 12\}$.** En annexes, nous proposons de vérifier la stationnarité des processus $(A_{m,j})_j$ pour tout mois m par la méthode des autocorrélations partielles. Les graphiques montrent également la stationnarité de ces processus et nous permettent d'identifier que l'ordre autoregressif de ces processus est égal à 1. Aussi, ce résultat nous indique que **pour tout mois m, le processus $(A_{m,j})_j$ peut être simulé selon un modèle autorégressif d'ordre 1 AR(1)**.

II.5.3 Modèle AR(1) pour les processus d'anomalie de températures mois par mois

Motivation

Un modèle autorégressif (AR) est un modèle de série temporelle qui décrit comment les valeurs passées d'une variable particulière influencent sa valeur actuelle. En d'autres termes, **un modèle AR tente de prédire la valeur suivante d'une série en incorporant les valeurs passées les plus récentes et en les utilisant comme données d'entrée**. Les modèles autorégressifs reposent sur l'idée que **les événements passés peuvent nous aider à prédire les événements futurs**.
La modélisation autorégressive consiste à former un modèle de régression sur la valeur de la variable de réponse elle-même. Le terme autorégressif est composé des mots "Auto" et "Régressif" qui représentent la régression linéaire sur elle-même (auto).
Dans le contexte des simulations de températures, **la modélisation autorégressive consiste à créer un modèle dans lequel l'anomalie de températures prédite**

pour une date t, A_t, dépend des anomalies de température précédentes A_s tel que $s < t$ avec un décalage temporel constant prédéterminé, ici égal à une année (365 jours) : il fait plus ou moins la même température au 1er janvier de chaque année, 2 janvier de chaque année,..., 31 décembre de chaque année, et les températures sont cohérentes (nous n'avons pas 35°C à Paris un 2 janvier).
L'utilisation d'un modèle auto-régressif est justifié par la corrélation certaine entre la température à la date t X_t et la température à la date $t-1$ X_{t-1}, et donc entre l'anomalie de températures à la date t et à la date $t+1$: les températures ne font pas un bond de 30°C entre aujourd'hui et demain.
Enfin, les températures dans un même mois ou d'une même saison sont plutôt homogènes et les processus d'anomalies de températures par mois sont stationnaires. **Il convient donc de construire un modèle auto-régressif pour chaque mois**.

Construction du modèle AR(1) pour chaque processus $(A_{m,j})_j$:
A partir de l'historique de températures fourni (et d'anomalies de températures construites préalablement), nous sommes alors en mesure de construire un modèle AR(1) pour chaque date t :

$$A_t = \alpha_m + \beta_m A_{t-1} + \epsilon_t$$

où :

- $t = (j, m, a)$ est la date t décomposée au format jour j, mois m, année a.
- ϵ_t est le résidu de la régression et est stochastique.
- α_m et β_m sont les coefficients de régression pour chaque mois $m \in \{1, ..., 12\}$ et sont déterministes.

Démarche suivie pour construire les modèles AR(1) :
À chaque date $t \geq 1$, nous effectuons la régression de A_t sur A_{t-1} et obtenons alors les coefficients de régression α_m et β_m recherchés pour chaque pour $m \in \{1, ..., 12\}$ ainsi que les résidus ϵ_t à chaque date $t \geq 1$.

Nous créons ensuite une matrice qui va stocker toutes les anomalies de températures prédites par régression linéaire de A_t sur A_{t-1} contenant les coefficients α_m et β_m correspondant à chaque date $t = (a, m, j)$ et les résidus ϵ_t à chauque date t.
Dans cette matrice :

- chaque colonne de la matrice représente un scénario (par exemple, N = 2 simule deux scénari de températures, c'est-à-dire deux années ou deux scénari pour une même année),
- chaque ligne de la matrice représente, elle, le jour (par exemple, i = 2 pour le 2 janvier),
- chaque coefficient est égale à l'anomalie de température prédite sous le mode

$$A_t = \beta_m \times A_{t-1} + \epsilon_m$$

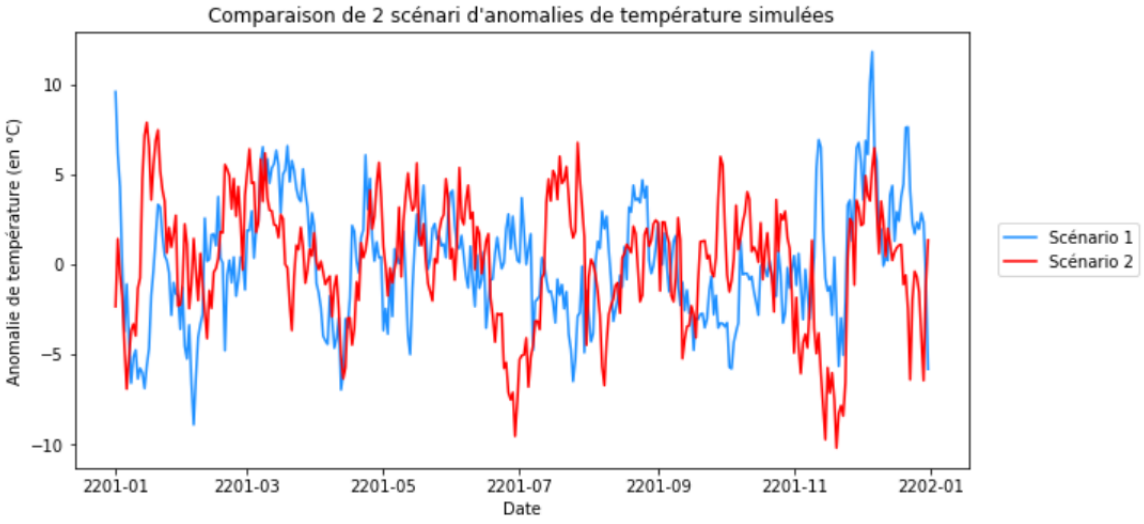
.

Concernant les coefficients de cette matrice :

- le coefficient (i, j) de la matrice représente l'anomalie au jour i ($i \in \{1, ..., 365\}$) du scénario j ($j \in \{1, ..., N\}$).
- le coefficient $(i-1, j)$ de la matrice représente l'anomalie de températures de la veille.
- les premiers coefficients de chaque colonne $2 \leq j$, c'est-à-dire chaque premier janvier, sera simulé grâce au coefficient $(365, j-1)$.
- le premier coefficient de la première colonne, c'est-à-dire le premier janvier de la simulation, ne peut lui être simulé par une valeur antérieur. En conséquence, on attribue la valeur 0 à l'anomalie de la veille ($A_0 = 0$) ce qui peut être justifié par le fait que l'anomalie est centrée par construction, et attribuer sa valeur espérée en moyenne lorsque la donnée est inconnue paraît assez cohérant.

Ajoutons que **les coefficients β_m sont soigneusement attribués à chaque date t.** Par exemple, les 31 jours de janvier sont tous simulés avec les valeurs β_1.
Enfin, **les résidus, ϵ_t sont tirés aléatoirement dans la liste de $365 \times N$ valeurs de résidus** créée à partir de chacune des régressions linéaires de A_t sur A_{t-1}, en prenant en compte le mois et ajoutés dans la prédiction. Autrement dit, pour construire A_t, on utilise le terme précédent A_{t-1} déjà calculé, puis on le multiplie par le bon β_m (si t appartient au mois m), et on lui ajoute un résidu tiré aléatoirement dans le paquet du mois m.

Le graphique ci-dessous est le résultat que nous obtenons lorsque nous simulons l'anomalie de températures pour N = 2 scénari pour l'année 2201 par exemple. **En somme, nous avons donc bel et bien simulé une anomalie de températures stochastique donnant différents scénari possibles à une même date donnée**.



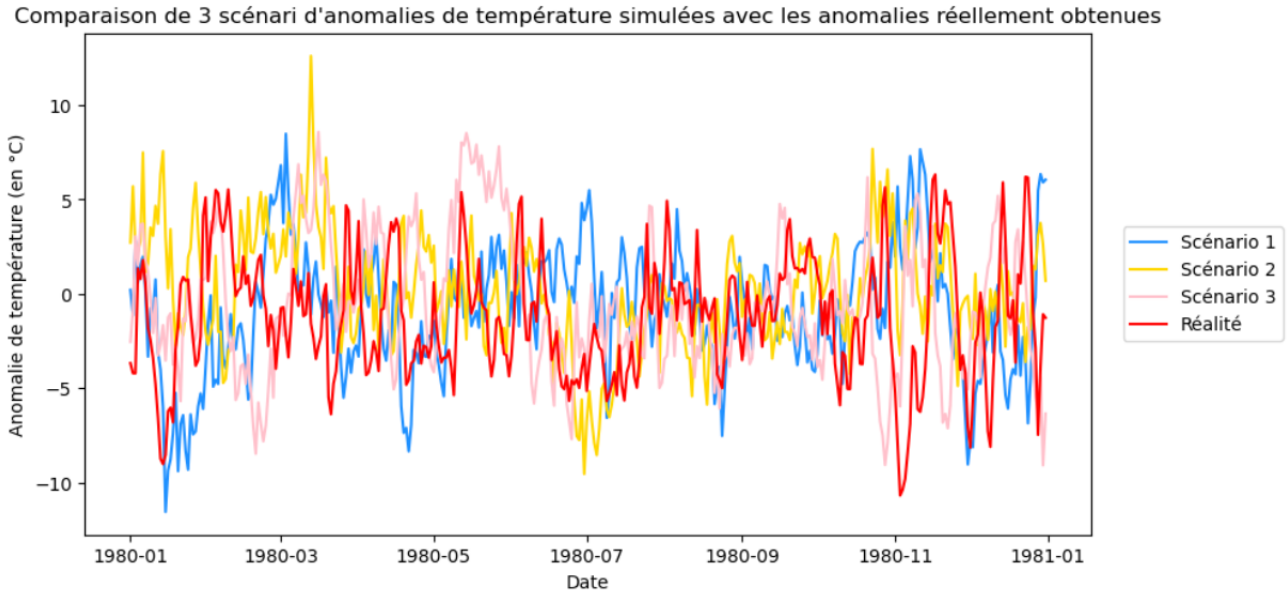
II.6 Simulateur de températures

Pour construire notre simulateur de températures à l'aide de Python, nous raisonnons en deux étapes.

1. Nous créons une fonction "SimuAnomalie" qui prend en entrée un nombre N de scenari de températures souhaités, une date de début et une date de fin de la simulation (quand veut-on que commence et s'arrête la prédicition ?) et permet de sortir N scenari d'anomalies de températures simulées telles que ci-avant.

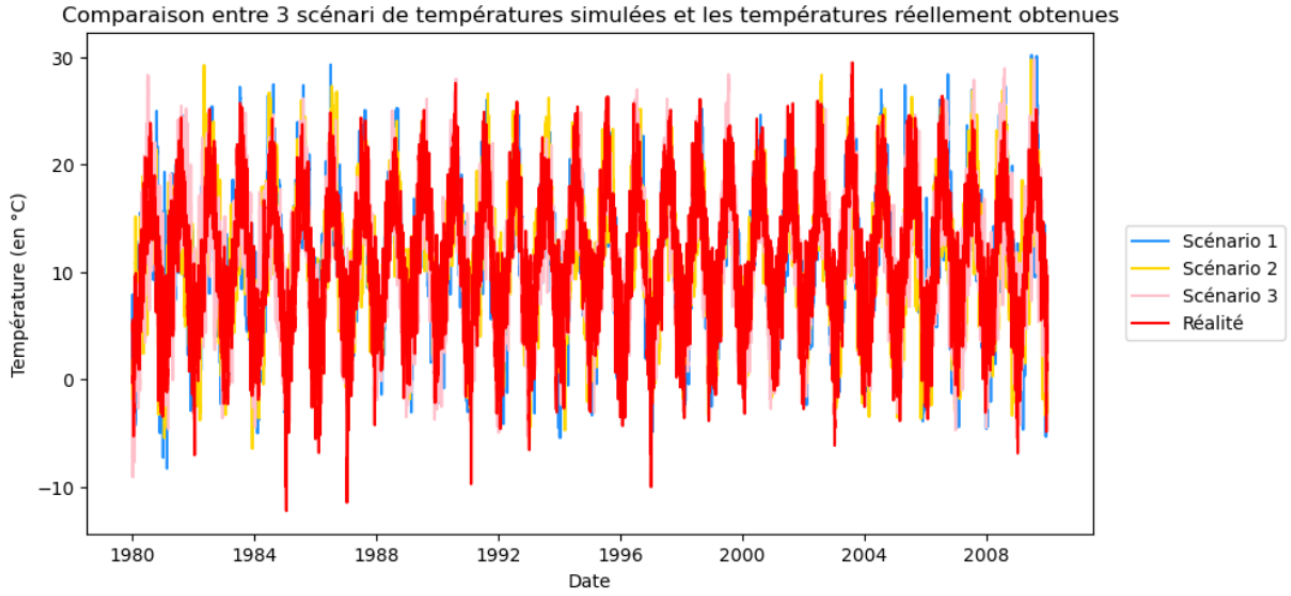
2. Nous définissons ensuite la fonction "simulateur" qui prend en entrée un nombre N de scenari de températures souhaités, une date de début et une date de fin de la simulation, et permet de sortir N scenari de températures prédites pour chaque date t comprise entre les bornes temporelles. Cette fonction est simplement la somme du terme déterministe de la normale de saison obtenue par la fonction f avec le terme stochastique de l'anomalie de température obtenue par la fonction "SimuAnomalie".

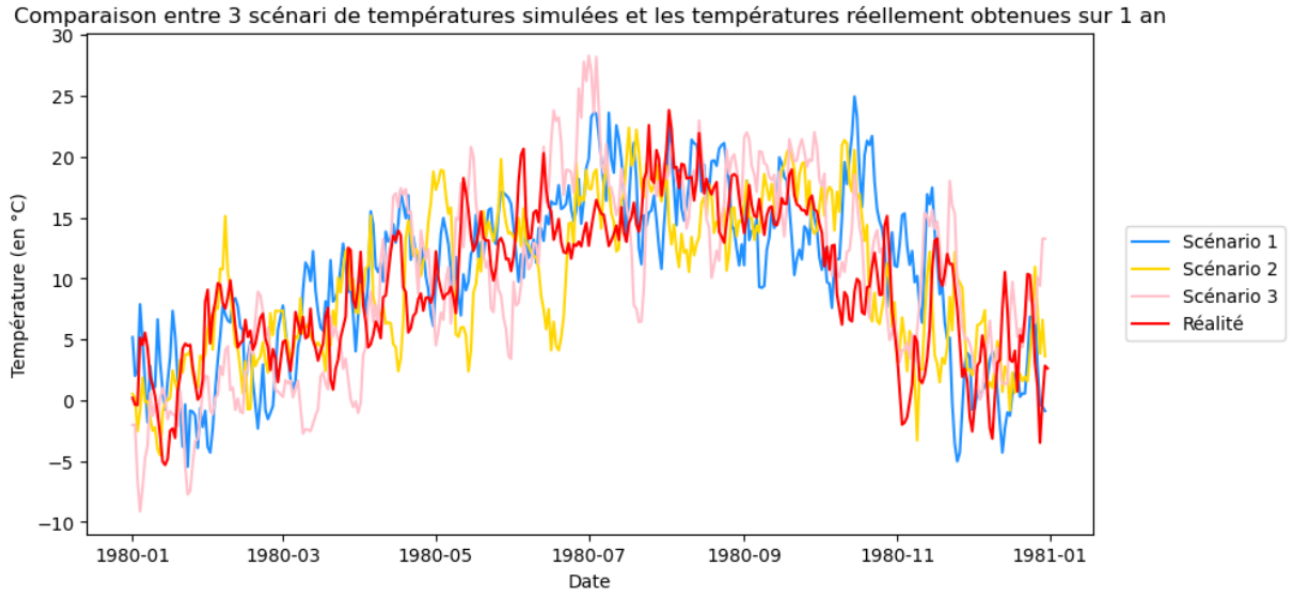
Après avoir créé la fonction "SimuAnomalie" et comparé les anomalies de températures générées par le simulateur de températures pour trois scénari différents sur l'année 1980 aux anomalies de températures empiriquement relevées pour cette même année (1980), nous constatons que **les anomalies de températures simulées miment très bien la réalité, au point où on ne sait plus quelle était l'anomalie prélevée dans la réalité et l'anomalie de températures simulée par notre simulateur**.
Le graphique ci-après, où les anomalies de températures empiriques pour l'année 1980 apparaissent en rouge et les anomalies de températures simulées en bleu, jaune et rose (une couleur par scénario), en est la preuve.



Après avoir créé la fonction "simulateur", nous pouvons **comparer les températures simulées aux températures relevées dans la réalité d'abord sur toute la série**

(du 1er janvier 1980 au 31 décembre 2009) puis en faisant un zoom sur une année : 1980.
Comme pour le simulateur de l'anomalie de températures, nous constatons que **les températures simulées miment très bien la réalité, au point où on ne sait plus distinguer la température prélevée dans la réalité de la température simulée par notre simulateur**.
Les représentations graphiques ci-après permettent de faire ce constat.





D'autres tests auraient pu être réalisés pour valider ce simulateur : quantiles, qqplot, variance, trajectoires... Nous considérerons désormais que le simualteur ainsi créé est correct pour être utilisé pour le calcul des OSP.

III Risque 2% en température

Comme tout énergéticien, ENGIE se doit de respecter les OSP fixées par la CRE (connaître les températures 2%, c'est-à-dire les températures extrêment froides telles qu'elles n'arrivent que tous les 50 ans, calculer des scénari d'hiver très froids, plutôt froids, etc.) pour vérifier que les stockages et approvisionnements sont capables de répondre à la demande en gaz..

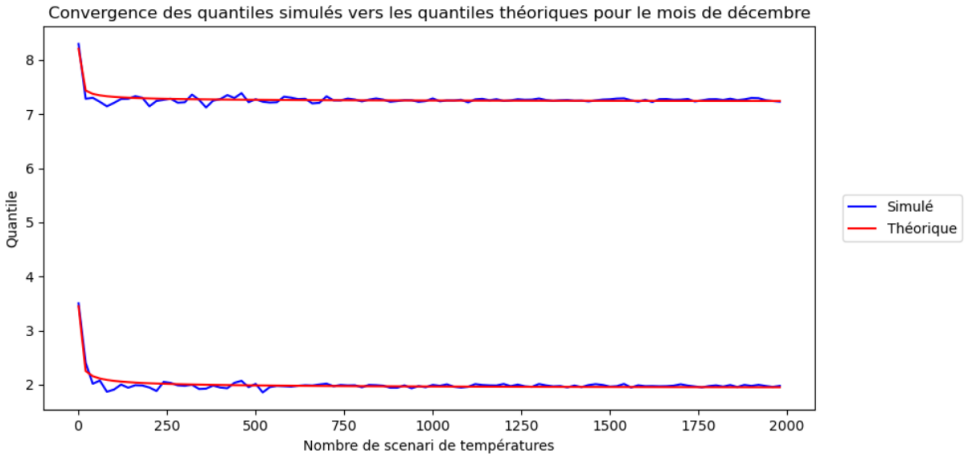
III.1 Application du simulateur

À partir du simulateur de températures construit dans la partie précédente, nous pouvons appliquer la méthode de Monte-Carlo de convergence des quantiles rappelée dans la première partie.
Nous avons construit 2000 scénari de températures sur trois ans. Cela nous donne un total de quasi 150 000 jours simués par mois. Grâce à la partie I.3, nous avons que ce sera suffisant pour obtenir les quantiles souhaités (20% et 2% ici) avec une précision acceptable.

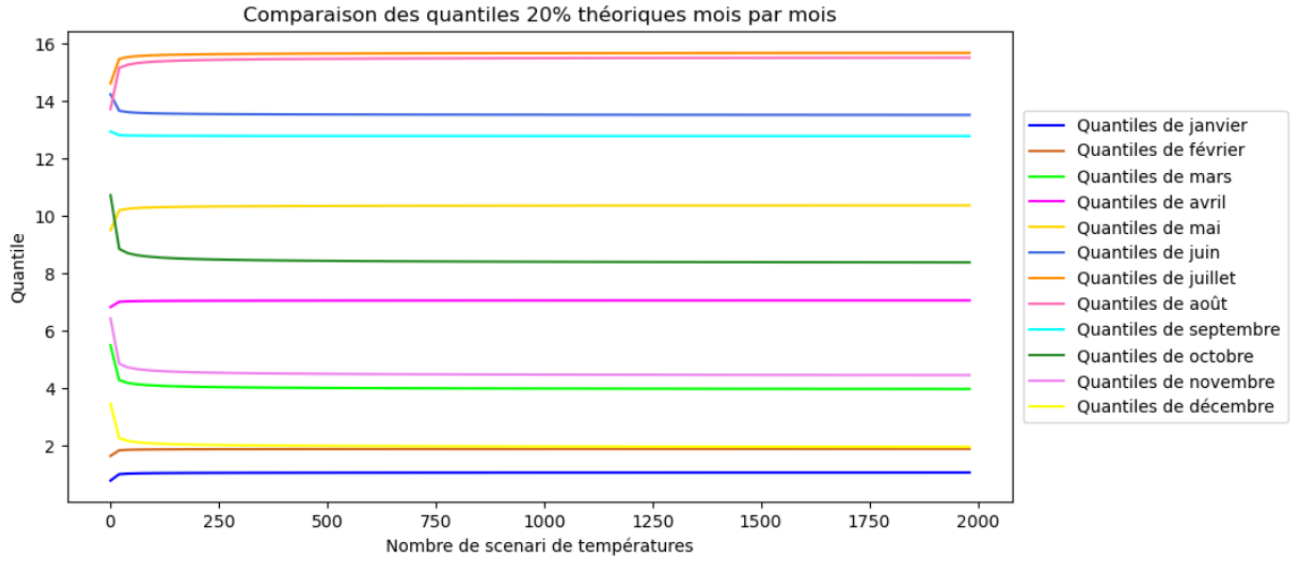
III.2 Quantile 20% mois par mois

Par exemple, nous pouvons calculer le quantile à 20% froid et 20% chaud des températures journalières, mois par mois, en montrant la convergence par méthode de Monte-Carlo, comme en partie I.3.

Les résultats ci-dessous affichent la convergence des quantiles 20% froid et 20% chaud pour le mois de décembre vers respectivement des quantiles théoriques, que nous pouvons identifier pour déterminer les risques en températures recherchés.

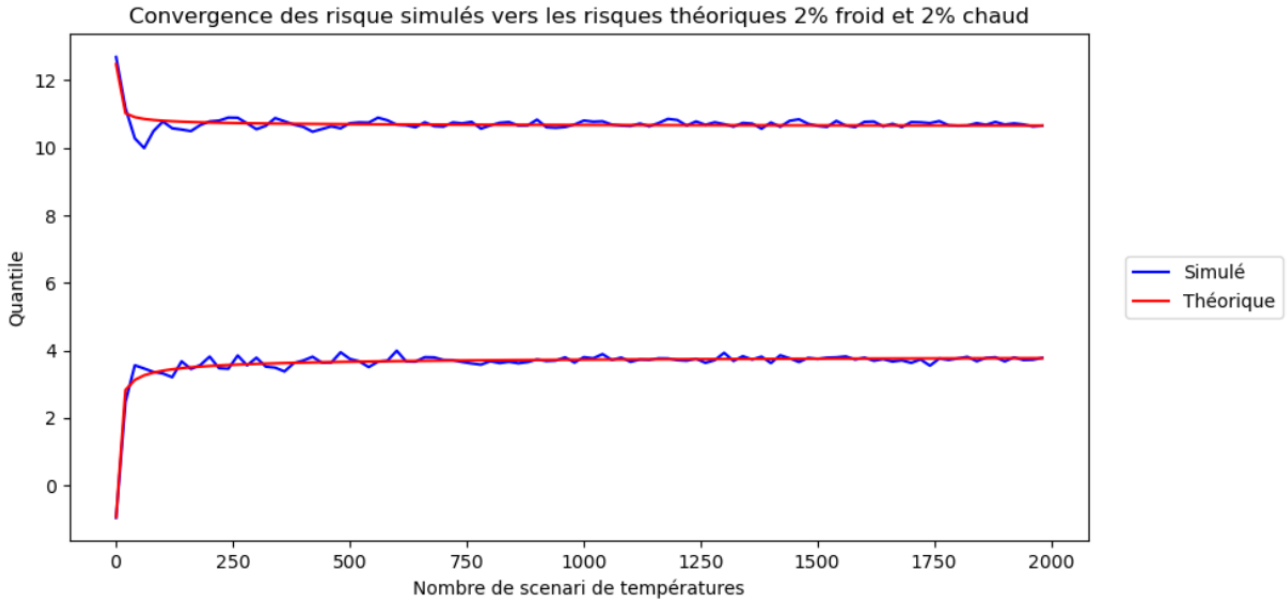


Ainsi, en décembre, il y a une chance sur cinq d'avoir une température en-deçà de 2°C.
Nous pouvons aussi comparer tous les quantiles 20% froid mois par mois, **les mois qui sont alors le plus à risques sont novembre, décembre, janvier, février** :



III.3 Quantile 2%, risque 2% sur l'hiver

L'esprit de la réglementation sur le risque de pointe est de connaître le risque 2% sur 3 jours glissants durant l'hiver. C'est la limite inférieure en-deçà de laquelle il est licite de ne pas fournir de gaz aux particuliers. Cette température est assez froide ce qui oblige les énergéticiens à avoir de gros stocks de gaz.



En hiver, il y a donc une chance sur cinquante que la température soit en-deçà de $+3.771°C$ pendant trois jours consécutifs.

Conclusion

L'objet de ce projet de statistiques appliquées était de calculer des quantiles de températures pour répondre aux questions suivantes :

À quelle température doit-on s'attendre avec un risque de 2% ?
Comment calculer les différents risques par mois ?

Pour ce faire, nous avons créé un simulateur de températures permettant de simuler N scenari de températures pour une même date à partir d'un historique de températures relevées quotidiennement du 1er janvier 1980 au 31 décembre 2009. Ce simulateur a été créé en remarquant que la série temporelle des températures $(X_t)_t$ pouvait être modélisée par :

$$X_t = f(j) + A_t, \forall t$$

où :

- t est la date d'entrée pouvant être décomposée comme (j, m, a) pour le format jour, mois, année.
- X_t est la température à la date t, ce que l'on cherche à prédire.
- $f(j)$ est la normale de saison donnée au jour j du mois m de cette date S_t, et est donnée par la fonction f définie préalablement par régression linéaire et approximation par les séries de Fourier.
- A_t est l'anomalie de température à la date t, c'est-à-dire l'écart à la normale de saison. Elle est modélisée par un modèle AR(1) tel que donné précédemment.

La partie I a montré que la méthode de Monte-Carlo donne une approximation des quantiles.
Nous avons utilisé cette méthode pour calculer différents risques, avec une bonne approximation.

Avec un risque 2%, nous pouvons nous attendre à ce que les températures soient en dessous de +3.771°C pendant trois jours consécutifs entre novembre et avril.

Annexes

Notations

Les notations mathématiques suivantes seront utilisées dans ce travail :

R est l'ensemble des nombre réels.
$\mathbf{E}(X)$ désigne l'espérance d'une variable aléatoire X, lorsqu'elle existe.
$\mathbf{V}(X)$ désigne la variance d'une variable aléatoire X, lorsqu'elle existe.
$\overline{X_n}$ désigne la moyenne arithmétique $\frac{1}{n}\sum_{1i\leqslant n} X_i$ de variables aléatoires $X_1, \ldots, X_n$ pour un $n \geqslant 1$ donné.
$\mathcal{N}(\mu, \sigma^2)$ est les lois normales de moyenne $mu > 0$ et de variance σ^2.
$[X < \alpha]$ désigne pour une variable aléatoire réelle X et $\alpha \in \mathbf{R}$, l'indicatrice de l'ensemble $\{\omega,\, X(\omega) < \alpha\}$.

Textes réglementaires

L'Obligation de Service Public que doit respecter **ENGIE** se fonde principalement sur l'articles R121-4 du Code de l'Énergie :

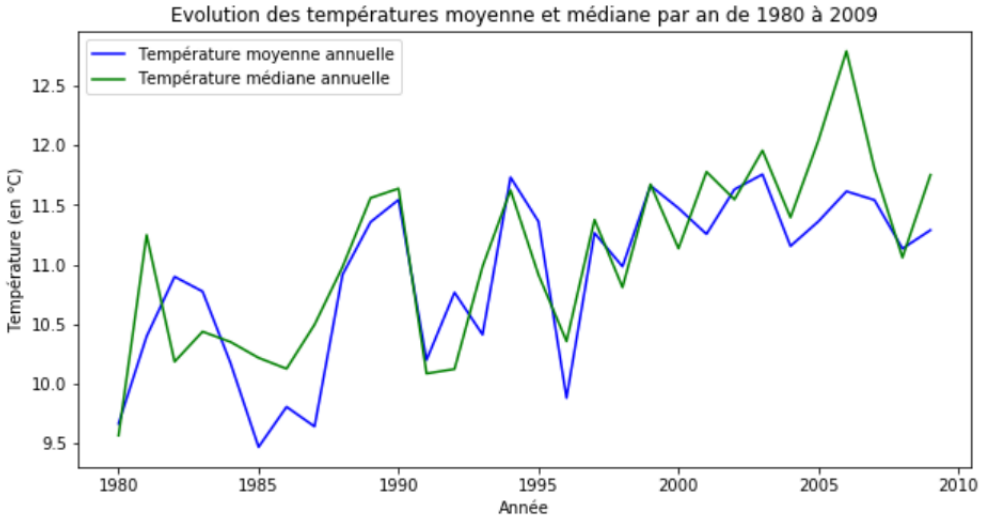
(...) le fournisseur doit être en mesure d'assurer la continuité de fourniture même dans les situations suivantes :
1° Disparition pendant six mois au maximum de la principale source d'approvisionnement dans des conditions météorologiques moyennes ;
2° Hiver froid tel qu'il s'en produit statistiquement un tous les cinquante ans ;
3° Température extrêmement basse pendant une période de trois jours au maximum telle qu'il s'en produit statistiquement une tous les cinquante ans.

Le point 3 ci-dessus est le *risque en pointe*, le point 2 quant à lui est le *risque en aléa*. Ce dernier nécessite de construire différents scénarios d'hiver à 2%. En particulier, **ENGIE** produit plusieurs hivers-types, et ceci oblige à calculer des risques de températures, ou quantiles, sur plusieurs périodes, et selon plusieurs pourcentages.

Compléments à l'analyse des données de températures

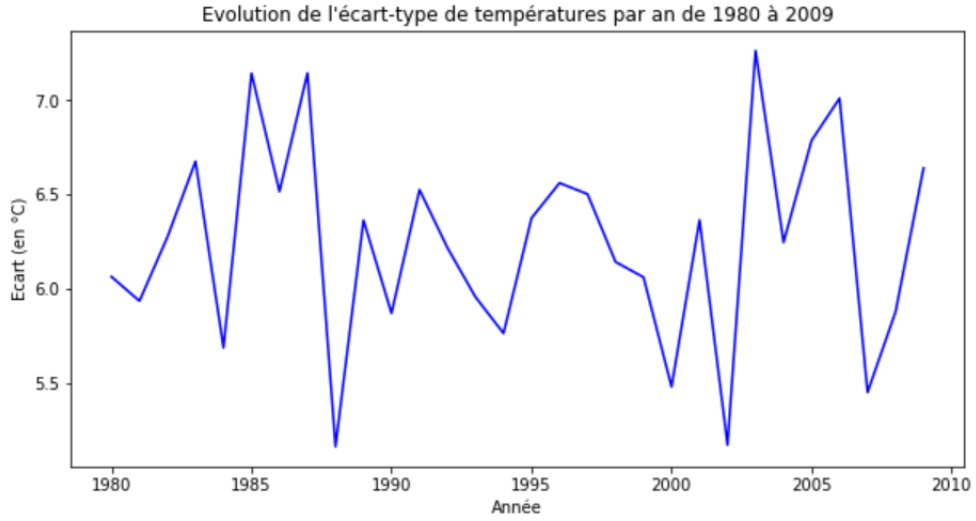
Comparaison médiane / moyenne

Par le graphique ci-après, nous comparons la médiane des températures annulles à la moyenne des températures annuelles entre 1980 et 2009. Nous remarquons que la médiane et la moyenne ne coïncident pas. Ainsi, si la distribution des températures pouvait à première vue sembler symétrique (voire gaussienne), la réalité est autre.



Ecart-types significatifs

Par le graphique ci-après, nous montrons que les écarts de températures à la moyenne des températures sont importants. Ces écart-types significatifs permettent d'appuyer l'hypothèse de saisonnalité et également d'introduire le concept d'anomalie de températures (écart entre la normale de saison et la température relevée).



Compléments à l'obtention de la fonction lissée des normales de saison

Nous commençons par décomposer la série temporelle des températures entre le 1er janvier 1980 et le 31 décembre 2009 en termes trigonométriques sinus et cosinus. Nous obtenons les résultats suivants :

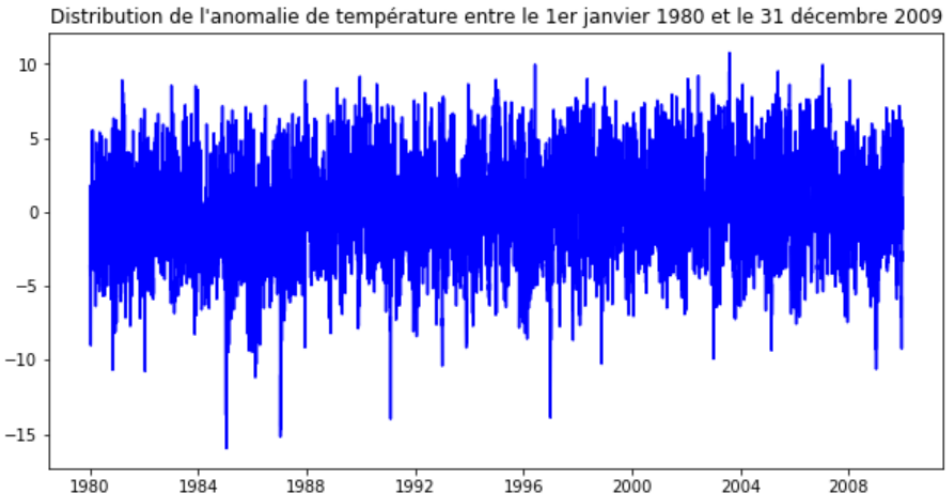
	COS1	SIN1	COS2	SIN2	COS3	SIN3
0	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000
1	0.999852	0.017213	0.999407	0.034422	0.998667	0.051620
2	0.999407	0.034422	0.997630	0.068802	0.994671	0.103102
3	0.998667	0.051620	0.994671	0.103102	0.988023	0.154309
4	0.997630	0.068802	0.990532	0.137279	0.978740	0.205104

Puis, à l'aide du module "Linear Regression" du package "sklearn", nous réalisons la régression linéaire des moyennes journalières de températures sur la décomposition trigonométrique proposée ci-avant et les températures de l'historique de températures disponible, et obtenons ainsi la fonction f.

Compléments à la description de l'anomalie de températures

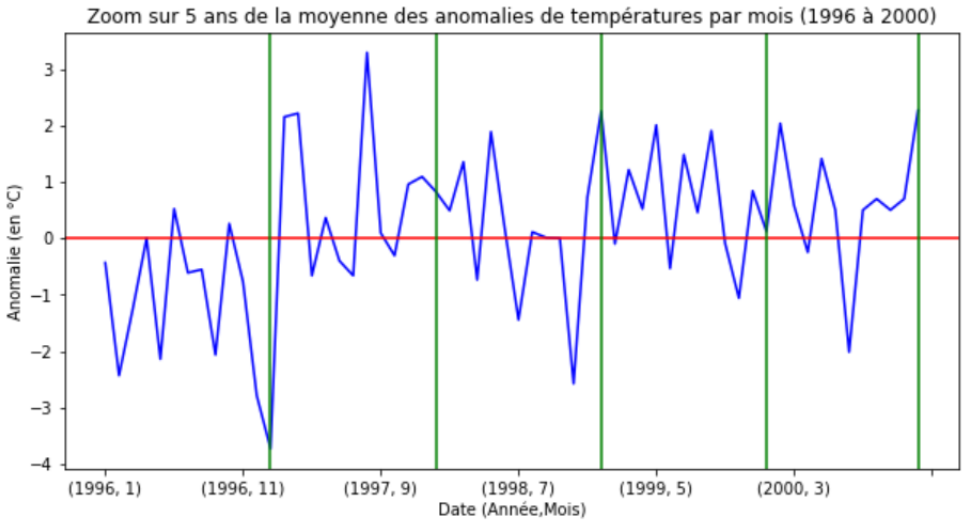
Réprésentation graphique de l'anomalie de températures

Le graphique ci-dessous est représente du comportement de l'anomalie de températures $(A_t)_t$ à toutes les dates t disponibles dans l'historique de températures fourni par ENGIE.



Zoom sur 5 ans du comportement de l'anomalie de températures

Le graphique ci-dessous est un zoom sur cinq ans du comportement de l'anomalie de températures. Les lignes vertes délimitent les années. Nous pouvons alors observer le caractère assimilable à une périodicité de 365 jours (1 an) de l'anomalie de températures $(A_t)_t$.



Stationnarité de l'anomalie de températures par mois par le test de Dickey-Fuller Augmenté

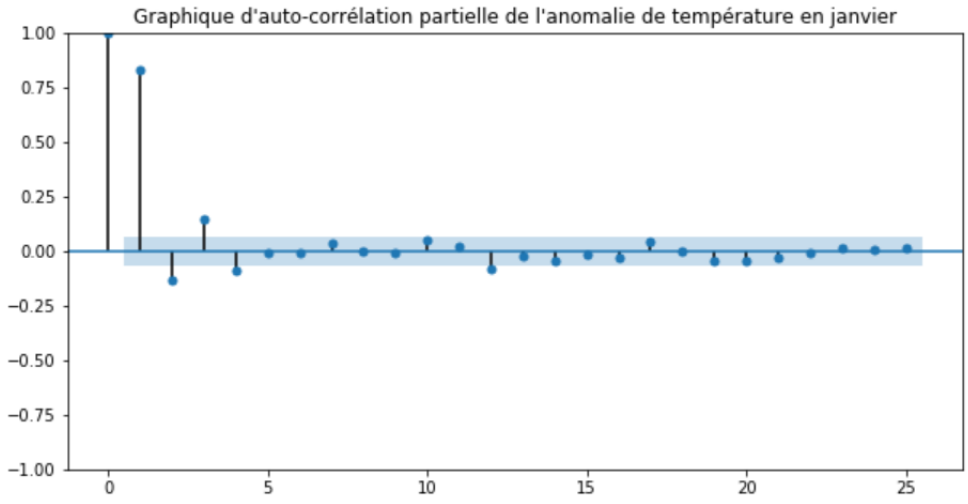
Voilà les résultats du test de Dickey-Fuller Augmenté conduit avec Python sur les processus $(\mathbf{A_{m,j}})_\mathbf{j}$ pour chaque mois $m \in \{1, ..., 12\}$.

```
P-value de l'anomalie du mois de janvier :  2.426283201052311e-14
P-value de l'anomalie du mois de février :  1.877415332485163e-13
P-value de l'anomalie du mois de mars :  1.0276455334687085e-18
P-value de l'anomalie du mois de avril :  2.01285405388709e-16
P-value de l'anomalie du mois de mai :  8.761525083127981e-20
P-value de l'anomalie du mois de juin :  4.2927839866401665e-19
P-value de l'anomalie du mois de juillet :  2.1274603090529366e-18
P-value de l'anomalie du mois de août :  8.017872649307874e-19
P-value de l'anomalie du mois de septembre :  2.759162299679457e-11
P-value de l'anomalie du mois de octobre :  4.803504215131005e-16
P-value de l'anomalie du mois de novembre :  3.601129818545732e-18
P-value de l'anomalie du mois de décembre :  1.1077746446993655e-17
```

Toutes les p-values sont bien largement inférieures à 0.0001, ce qui nous permet d'affirmer avec un niveau de confiance de 99.99% que les processus $(\mathbf{A_{m,j}})_\mathbf{j}$ pour chaque mois $m \in \{1, ..., 12\}$ sont stationnaires.

Stationnarité de l'anomalie de températures par mois par la méthode des autocorrélations partielles

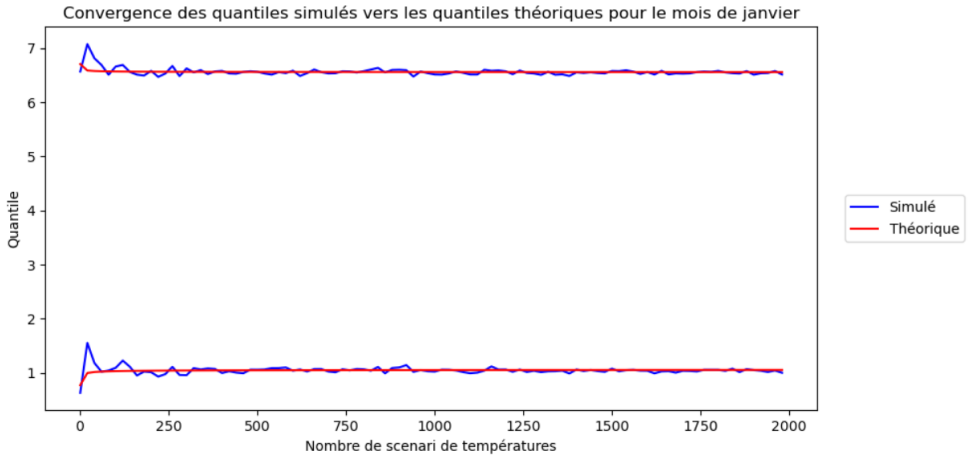
La corrélation entre la série à l'instant t et la série à l'instant t-k est appelée l'autocorrélation de retard k.
L'autocorrélation partielle entre les termes d'une série stationnaire permet de mesurer la corrélation entre de deux observations en éliminant l'effet de toutes les autres observations intermédiaires. **Tracer, pour chaque mois $m \in \{1, ..., 12\}$, le graphique de l'autocorrélation de la série temporelle stationnaire $(\mathbf{A_{m,j}})_\mathbf{j}$ en fonction du retard $\mathbf{k \in \{1, ..., 25\}}$ grâce à Python** nous permet d'**identifier le nombre de retards à partir duquel l'autocorrélation partielle devient non significative**. Ce nombre correspond au degré du polynôme autorégressif, **c'est-à-dire l'ordre p** du modèle AR(p) dont le modèle $(A_{m,j})_j$ est solution.

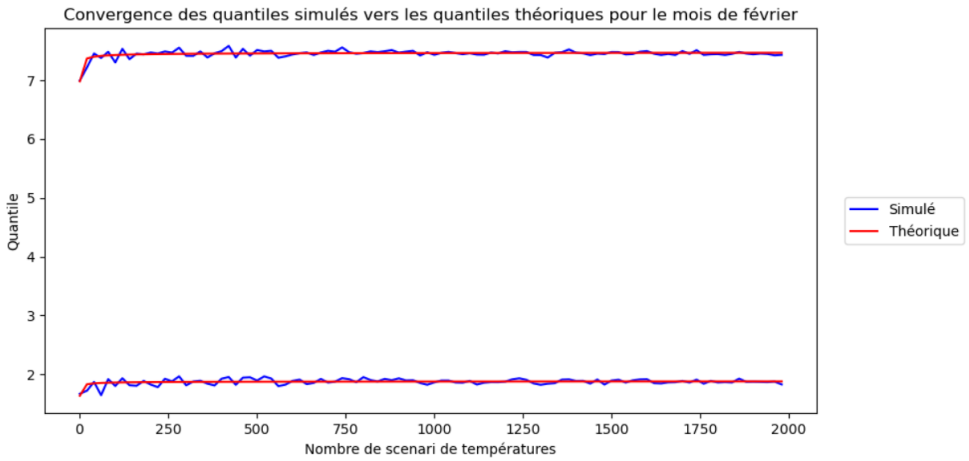


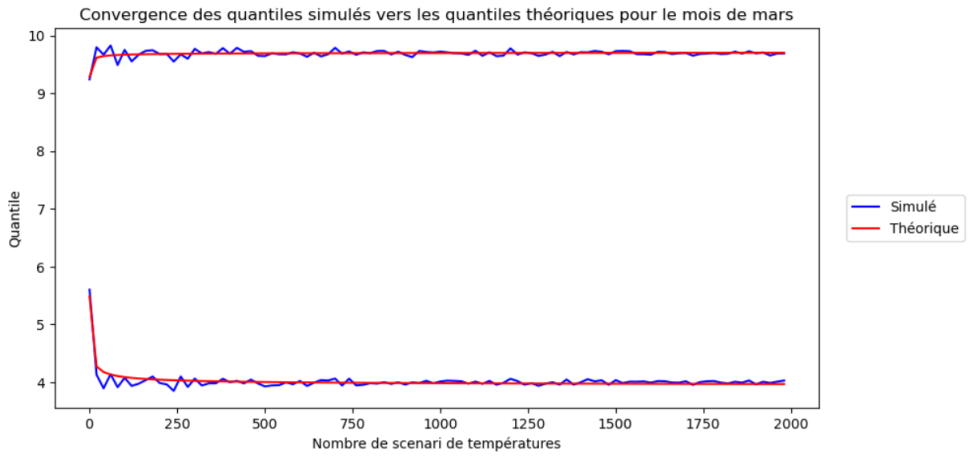
Le graphique ci-dessous trace les autocorrélations partielles pour le processus $(A_{1,j})_j$, c'est-à-dire l'anomalie de températures pour le mois de janvier. Les mêmes graphiques peuvent être observés pour les mois de février, mars, ..., décembre et suivent la forme prise par le mois de janvier.
Le graphique des autocorrélations partielles montre que l'autocorrélation devient non significative à partir du retard k = 1. En conséquence, nous pouvons **choisir l'ordre p = 1 pour le modèle AR duquel la série temporelle $(\mathbf{A_{m,j}})_\mathbf{j}$ est solution**, et ceci pour tout mois $m \in \{1, ..., 12\}$.

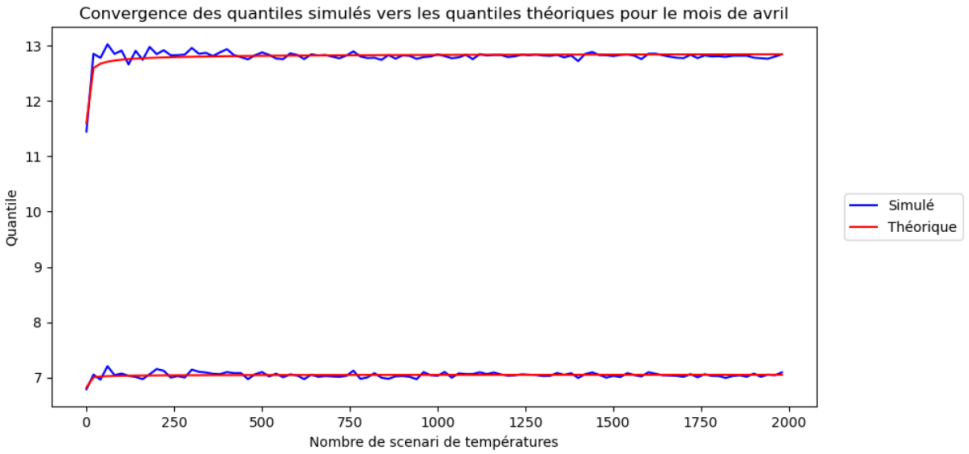
Résultats des convergences des quantiles 20% froids

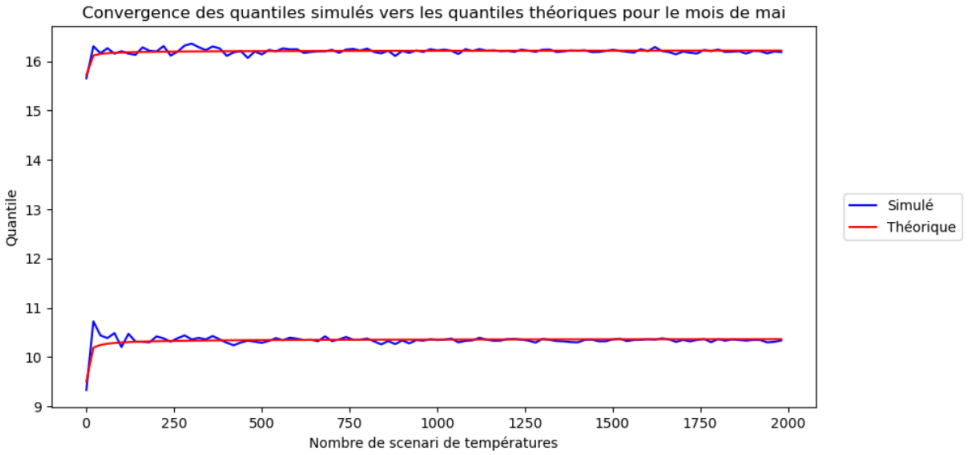
Les résultats, obtenus par méthode de Monte-Carlo comme en partie I.3, ci-dessous montrent la convergence des quantiles 20% froids pour tous les mois, de janvier à décembre.

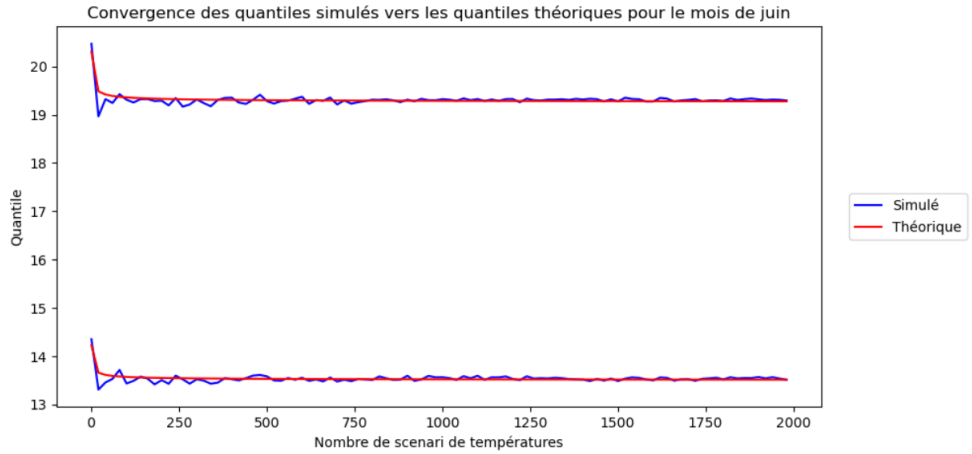


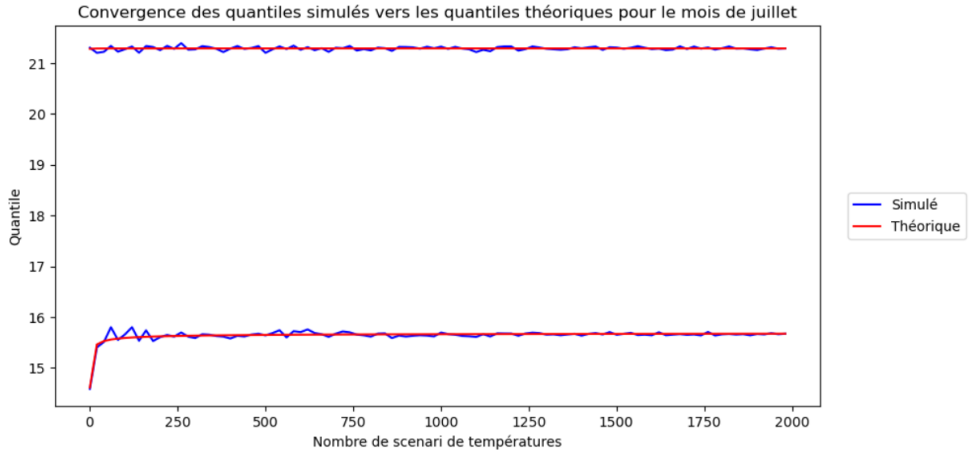


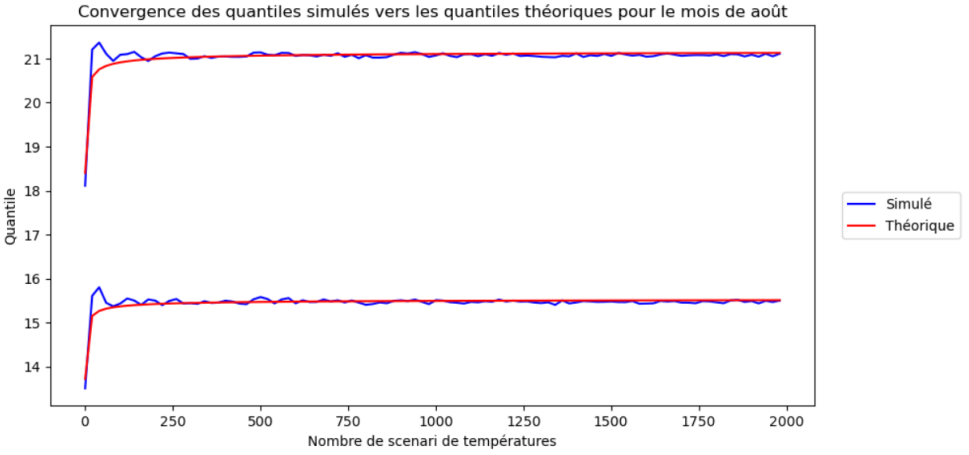


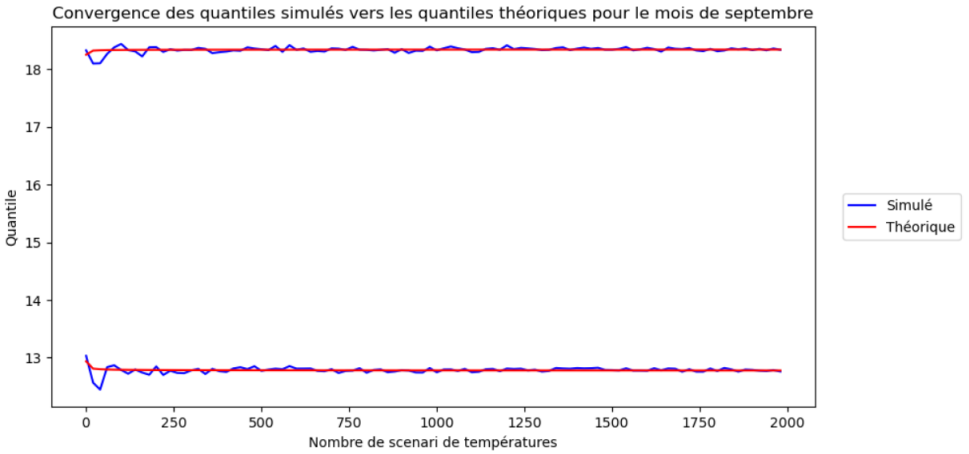

Convergence des quantiles simulés vers les quantiles théoriques pour le mois de avril
Quantile
Nombre de scenari de températures
Simulé
Théorique

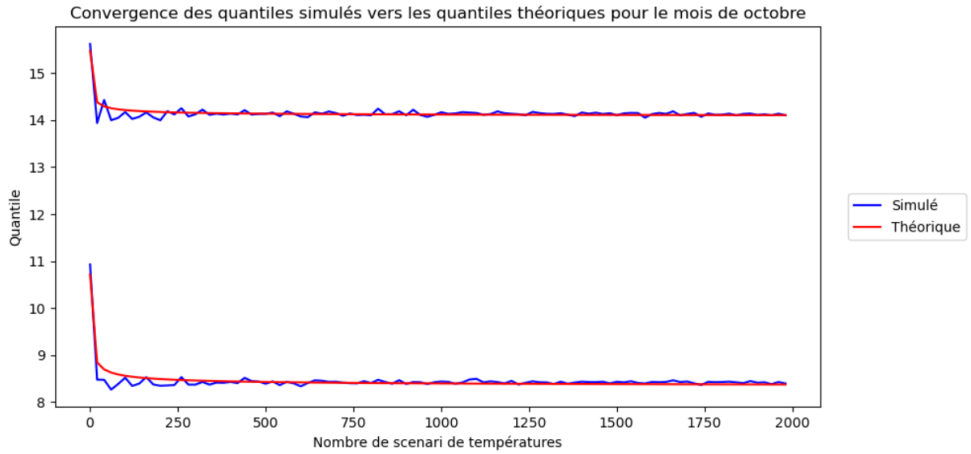

Convergence des quantiles simulés vers les quantiles théoriques pour le mois de mai
Quantile
Nombre de scenari de températures
Simulé
Théorique

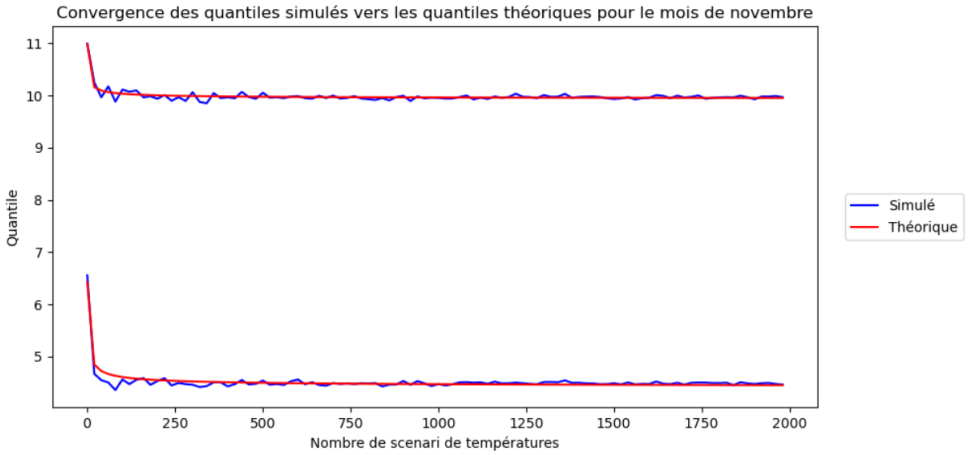

Convergence des quantiles simulés vers les quantiles théoriques pour le mois de juin
Quantile
Nombre de scenari de températures
Simulé
Théorique

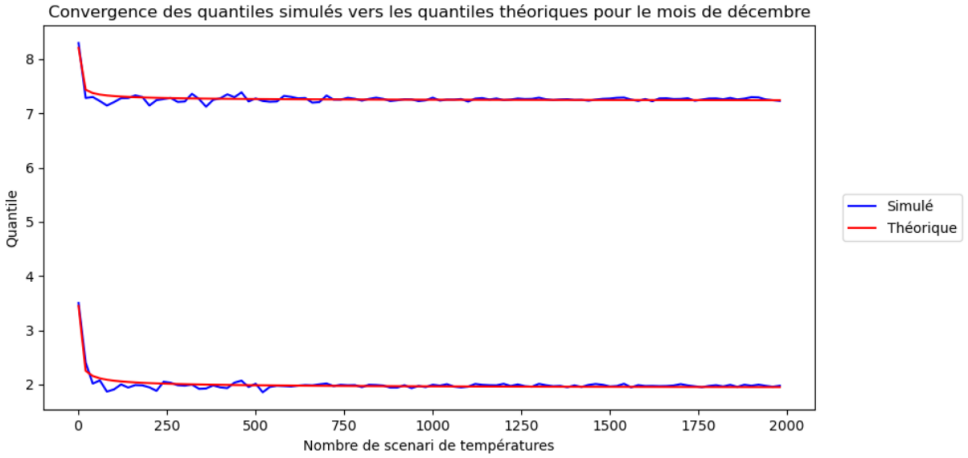

Convergence des quantiles simulés vers les quantiles théoriques pour le mois de juillet
Quantile
Nombre de scenari de températures
Simulé
Théorique


Convergence des quantiles simulés vers les quantiles théoriques pour le mois de août
Quantile
Nombre de scenari de températures
Simulé
Théorique


Convergence des quantiles simulés vers les quantiles théoriques pour le mois de septembre
Quantile
Nombre de scenari de températures
Simulé
Théorique


Convergence des quantiles simulés vers les quantiles théoriques pour le mois de octobre
Quantile
Nombre de scenari de températures
Simulé
Théorique


Convergence des quantiles simulés vers les quantiles théoriques pour le mois de novembre
Quantile
Nombre de scenari de températures
Simulé
Théorique


Convergence des quantiles simulés vers les quantiles théoriques pour le mois de décembre
Quantile
Nombre de scenari de températures
Simulé
Théorique

Résultat de la convergence du risque 2% froid en hiver

Le tableau ci-dessous est la représentation numérique des quantiles théoriques obtenus par méthode de Monte-Carlo.
En hiver, pendant trois jours consécutifs, le risque 2% froid converge vers +3.771°C et le risque 2% chaud vers +10.661°C.

	QQ 98% - hiver	QQ 2% - hiver
0	12.480130	-0.937797
1	11.025718	2.827902
2	10.910292	3.126757
3	10.857947	3.262288
4	10.826458	3.343816
...	...	...
95	10.662418	3.768541
96	10.662195	3.769118
97	10.661976	3.769685
98	10.661760	3.770244
99	10.661548	3.770795