Mémoire de DEA TIS

Estimation des températures journalières à l'aide de techniques markoviennes

Julien Chiquet

Université de Technologie de Compiègne - Gaz de France

5 septembre 2003

Sommaire

Avant-Propos

Résumé

Ce mémoire présente les résultats de l'utilisation d'outils markoviens dans le cadre de l'estimation et de la simulation de données températures journalières, sujet motivé par Gaz de France.

Le mémoire s'organise en 2 chapitres :

Le premier chapitre définit les enjeux motivant la modélisation climatique et l'utilisation des outils markoviens. Une section s'attachera à présenter un premier modèle illustrant les principes des outils utilisés.

Le second chapitre présente le modèle markovien non-homogène finalement retenu et les résultats obtenus.

On trouvera en annexe une description structurelle de l'implémentation numérique de ce modèle.

Mots-Clés

chaîne de Markov, chaîne de Markov non-homogène, chaîne de Markov discrète, chaîne de Markov d'ordre 1, Estimation, modélisation climatique, température

Laboratoires d'accueil

– Direction de la recherche de Gaz de France *Centre de recherche de Saint-Denis*
Pôle E2S (Pôle Economie, Statistiques et Sociologie)

– LMAC *Université de Technologie de Compiègne*
Laboratoire de Mathématiques Appliquées de Compiègne

Responsable UTC et suiveur DEA **Nikolaos Limnios**
Responsable Gaz de France **Karine Vernier**

Chapitre I

Des enjeux aux premiers modèles markoviens

I.1 Problématique

I.1.1 Enjeux

Modéliser les phénomènes climatiques

La modélisation mathématique des phénomènes météorologiques peut sembler être un sujet d'étude ambitieux de par la complexité des phénomènes étudiés. Ainsi, nous garderons à l'esprit que la modélisation et la simulation de phénomènes naturels ne tendent pas à les reproduire exactement mais plutôt à proposer des systèmes qui soient capables de fournir un certain nombre de sorties identifiées approximant au mieux certaines des caractéristiques du phénomène réel. Gageons que ce champ d'investigation est porteur d'enjeux multiples et très actuels :

- **des enjeux fondamentaux - comprendre -** La modélisation des phénomènes naturels et climatiques suscite l'intérêt et la curiosité de la communauté scientifique particulièrement ces 50 dernières années puisqu'elle s'inscrit dans l'émergence de nouvelles théories scientifiques prometteuses telle que la très médiatique théorie du Chaos : l'attracteur de Lorentz, premier objet "chaotique" est né de données météorologiques. Cet intérêt est d'autant plus vif depuis que le grand public a pris conscience des phénomènes de réchauffement climatique et de détérioration de l'équilibre écologique : la prévision climatique à court et moyen terme, tout comme l'identification des causes et des conséquences d'un éventuel réchauffement climatique, constitue un enjeux déterminant dans l'optique de l'adaptation des activités humaines, préventivement et curativement.

- **des enjeux industriels et économiques - anticiper -** Les enjeux sont ici évidents : pour les industriels et les commerciaux, il est essentiel de pouvoir jauger l'impact de l'occurence, de la durée ou de l'intensité de phénomènes climatiques sur leurs activités. Production, ventes, prix ou fréquentation sont autant d'enjeux industriels et commerciaux qui sont très corrélés aux phénomènes climatiques. Pouvoir prévoir les risques des phénomènes climatiques (comme la probabilité de passage dans des températures extrêmes, du nombre de jours de pluie, etc... avec un certain degré de confiance), c'est pouvoir adapter son activité et anticiper les secousses économiques encourues.

- **des enjeux spécifiques à Gaz de France -** Gaz de France a de multiples raisons (logistiques et économiques) de disposer de statistiques sur les données météorologiques, en particulier au niveau de la gestion et du dimensionnement des stockages et des canalisations mais également au niveau de ses ventes (usage chauffage dominant) [11].

  Cette dépendance aux aléas climatiques est d'autant plus actuelle que l'ouverture du capital de Gaz de France est envisagée dans un futur proche. La fluctuation annuelle des revenus en fonction du climat, paramètre non-contrôlable, n'est pas pour rassurer les futurs actionnaires du groupe. L'entreprise se doit donc aujourd'hui d'identifier très précisement la part de cette fluctuation aléatoire dans ses revenus et de se "couvrir" contre ce risque.

Par ailleurs, l'ouverture du marché à la concurrence a imposé de nouvelles contraintes de gestion (et en imposera encore d'autres). En terme d'organisation, Gaz de France a dû séparer ses activités de transport et de commercialisation de gaz naturel. L'accès au réseau de transport est dorénavant ouvert à tout tiers souhaitant commercialiser du gaz en France. Bien évidemment, cet accès est réglementé par une juridiction stricte, d'autant qu'il existe de nombreuses contraintes auxquelles doivent se soumettre tant les transporteurs que les revendeurs. Côté transporteur, les prévisions à court terme (1 à 5 jours) des quantités à acheminer permettent de gérer les demandes des commercialisateurs et en particulier de les valider ou de les refuser en cas de saturation du réseau. Un refus en raison de saturation doit bien évidemment être argumenté par un modèle fiable reliant la consommation de gaz naturel aux prévisions de climat. On voit ici tout l'intérêt de prévisions à 5 jours fiables. Le transporteur est également responsable du dimensionnement des canalisations de transport et des stockages souterrains de gaz naturel. Il doit dans ce domaine disposer des distributions de probabilités des températures, paramètre climatique considéré comme "dimensionnant". Le stage réalisé contribue à l'évolution des connaissances dans ce domaine. Concernant le commercialisateur, il a en charge les achats et ventes de gaz naturel. L'achat de gaz est basé aujourd'hui à 90% sur des contrats de longs termes avec les producteurs en mer du nord, Afrique ou Russie. Ces contrats de longs termes engagent Gaz de France sur des quantités à enlever pour un horizon de 3 à 5 ans. Ces quantités sont liées à une clause dite *"take or pay"* qui impose à l'acheteur d'enlever les quantités contractuelles ou de les payer. Le revendeur doit s'engager à acheter des volumes de gaz alors même que leur vente dépend largement du climat. On voit bien ici l'intérêt de connaître le climat et ses probables évolutions pour négocier des flexibilités d'enlèvement liées au climat et choisir au plus juste les volumes soumis à la clause de *"take or pay"*.

Bien évidemment, ces quelques exemples ne dressent pas une liste exhaustive des activités de Gaz de France soumises aux aléas climatiques. Ils permettent néanmoins d'identifier les forts enjeux associés à une bonne connaissance du climat et de ses évolutions probables à court et moyen terme.

Les réponses à ces enjeux demandent des modèles climatiques de plus en plus complexes puisque les exigences nouvelles de l'étude des réchauffements climatiques, des simulations et des prévisions à court et moyen terme impliquent l'utilisation d'outils mathématiques nouvellement utilisables avec l'explosion des capacités de calcul et justifient pleinement l'actualité de l'étude de la modélisation des phénomènes naturels.

REMARQUE I.1.1

Les données que nous avons utilisées pour mettre au point les modèles markoviens sont des données de températures journalières de la station météorologique Paris-Montsouris *sur un historique de 52 années de températures, de 1950 à 2001.*

I.1.2 Le choix d'un type de modélisation

Nous distinguerons deux grandes familles de modèles pour les phénomènes météorologiques [11] :

– Les modèles *déterministes*

  Il s'agit d'établir un système d'équations (pour la plupart issues de la mécanique des fluides), pour lequel les paramètres à l'instant initial sont déterminés par les observations météorologiques. Ce type de modèle est dédié à la *prévision du climat*. Il se doit d'être réinitialisé fréquement avec les observations disponibles, et les calculs doivent pouvoir se faire d'une manière proche du temps réel.

– Les modèles *statistiques et probabilistes*

  Il s'agit ici de créer un système dont les comportements sont du même type que le système réel. Pour autant, ils ne doivent pas coïncider exactement dans le temps, mais en convergence. On s'attache ici à faire un modèle numérique dont les caractéristiques globales vont tendre en moyenne vers celles du système réel. Ce type de modèle est plutôt dédié à la *simulation*. On utilise dans ce cas des outils issus des mathématiques probabilistes, permettant d'évaluer les *risques* d'atteindre telle température, qu'il pleuve, etc... après avoir appris le comportement réel sur un *ensemble d'apprentissage*. C'est le type de modélisation que nous avons choisi, de par les exigences du stage et son actualité en recherche [5].

Parmi le vaste champ des mathématiques probabilistes, nous avons retenu un outil qui se répand de manière grandissante dans la modélisation faisant intervenir les phénomènes temporels (modélisation financière, climatique, biologique, fiabilité, et même musicale !) : les outils issus du Markovien. Ils permettent d'énoncer des relations et des probabilités quant au voisinage (spatial ou temporel) de phénomènes. Très grossièrement, on pourra dire qu'ils permettent de définir l'état d'un système à un instant t uniquement à partir d'un certain nombre de ses états précédents : instant $(t-1)$ pour du markovien *du premier ordre*, instants $(t-1)$ et $(t-2)$ pour du markovien *du second ordre*, etc ...

I.1.3 Eléments essentiels sur les chaînes de Markov

Définition et propriétés d'une chaîne de Markov

Nous commencerons par donner la définition mathématique d'une chaîne de Markov d'ordre 1 [8] :

DÉFINITION I.1.2

La suite $X = (X_n, n \in \mathbb{N})$ de variables aléatoires sur un espace probabilisé $(\Omega, \mathbb{F}, P)$ à valeurs dans un ensemble E, fini ou dénombrable, est une chaîne de Markov si, pour tout entier positif n, pour tout $i, j, i_0, \ldots, i_n \in E$, on a :

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i) = p_n(i, j)$$

REMARQUE I.1.3

Dans le cas où $p_n(i, j) = p(i, j)$, la chaîne de Markov ne dépend pas du temps n et est dite homogène. La fonction $p(i, j)$ sera notée $\mathbb{P}_{ij}$

REMARQUE I.1.4

Dans ce cas, la discrétisation du temps est faite sur l'ensemble des entiers naturels $\mathbb{N}$. On ne parle plus de chaînes de Markov mais de Processus de Markov dans le cas continu, c'est-à-dire quand le temps est modélisé sur $\mathbb{R}^+$.

REMARQUE I.1.5

Plus intuitivement, la propriété essentielle des chaînes de Markov peut se traduire par le fait que toute l'information du passé est contenue dans le présent. Ainsi, la définition précédente peut s'écrire :

$$P(\textit{Futur} | \textit{Passé}, \textit{Présent}) = P(\textit{Futur} | \textit{Présent})$$

On parle de ***Propriété de Markov****.*

Les processus à sauts

Les chaînes de Markov permettent de modéliser très facilement n'importe quel processus dit *"à sauts"* (cf fig I.1) : ceux-ci permettent de représenter l'évolution d'un système dans le temps. On y suppose que le système peut évoluer dans le temps en un certain nombre d'états définis. La modélisation consiste à expliciter comment le système transite d'un état à un autre et comment il se comporte lorsqu'il reste dans le même état : on parle de *loi de changements d'états* d'une part et de *loi du temps de séjour* d'autre part.

Pour l'exemple des températures, les états seront les différentes valeurs possibles en degré. L'élément le plus important à développer sera la loi de changement d'état. En effet, de part la nature des données (des températures moyennes journalières) on considéra que la loi de temps

de séjour sera constante : lorsque le système entrera dans un état, il y restera une journée, soit un pas de temps, ce qui n'empêche pas le système de rester dans le même état (la même température) au pas de temps (au jour) suivant.

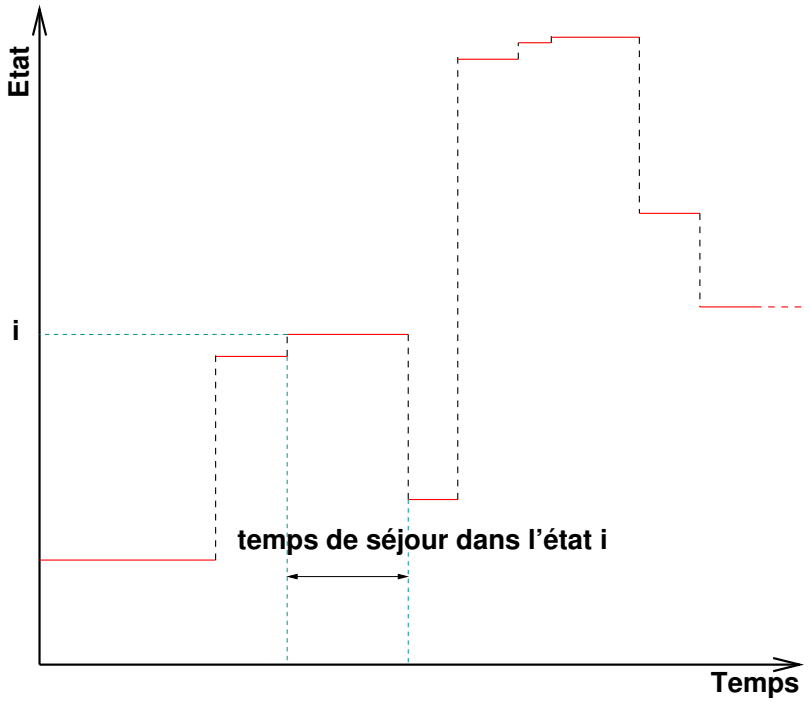


FIG. I.1 – Schéma d'un processus à sauts classique

Profondeur ou Ordre d'une chaîne de Markov

Lorsque l'on fait dépendre l'instant futur uniquement de l'instant précédent (*voisinage d'ordre 1*), la profondeur de la chaîne est 1. Il est légitime de penser que la température à un instant t est dépendante au sens probabiliste des températures des instants précédents.

On pourrait donc imaginer un modèle dont la probabilité d'état à l'instant t est dépendante non pas uniquement de l'instant précédent, mais des r instants précédents : ainsi, on parle d'une chaîne de Markov d'ordre r.

DÉFINITION I.1.6

Soit $X = (X_n, n \in \mathbb{N})$ *une chaîne de Markov. La chaîne est d'ordre* r *si on a :*

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)$$

$$= P(X_{n+1} = j | X_n = i, \ldots, X_{n-r+1} = i_{n-r+1})$$

Nous nous baserons sur ces éléments dans le cadre du stage réalisé, en vue de la modélisation markovienne des températures journalières.

I.2 Modèle markovien homogène d'ordre 1

On applique dans cette section les principes de base de la modélisation markovienne évoqués précédement. Ce modèle est l'occasion de bien assimiler ces principes à l'aide des hypothèses markoviennes les plus simples. Ainsi, nous avons choisi de commencer à modéliser les températures journalières à l'aide d'une chaîne de Markov homogène d'ordre 1.

I.2.1 Notations

REMARQUE I.2.1

Ce premier modèle est plutôt destiné à fixer les idées, nous donnant des premiers indices sur le comportement des chaînes de Markov appliquées à des données de température. Il nous permettra de mieux nous orienter par la suite.

On propose un premier modèle (modélisation des moyennes journalières des températures) pour lequel les températures sont discrétisées au pas de $\frac{1}{2}$ degré.

Soit une variable aléatoire X, et $(X_n)_{n\in\mathbb{N}}$ un processus décrivant la température à un instant n. L'unité (ou *pas*) de temps est **le jour**

On introduit les notations suivantes :

- E l'espace d'état (l'ensemble des états dans lesquels va pouvoir rentrer le système)
- $\mathbb{P}$ la matrice de transition du système (elle définit la loi de changement d'état du système)
- $\mathbb{P}_{ij}$ la fonction de transition
- μ la loi de X_0 (loi initiale)

$$\rightarrow (X_n) \text{ est une } \textit{chaîne de Markov Homogène}.$$

Après avoir identifié la température maximale (30.5) et minimale (-12.5), on en déduit l'espace d'états suivant :

$$E = \{-12.5, -12.0, \ldots, 30.0, 30.5\}, \text{ soit 87 états.}$$

On fait l'hypothèse markovienne à l'ordre 1, c'est-à-dire,

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$$

avec

$$P(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \neq 0$$

et on note

$$\mathbb{P}_{ij} = \boldsymbol{P(X_{n+1} = j | X_n = i)} \tag{I.1}$$

où $\mathbb{P}_{ij}$ est le $i^{\text{ème}}$ élément de la $j^{\text{ème}}$ colonne de la matrice de transition de la chaîne de Markov $(X_n)_{n\in\mathbb{N}}$.

I.2.2 Estimation de la fonction de transition $\mathbb{P}_{ij}$

La matrice de transition est, avec la loi initiale μ de X_0, l'élément essentiel d'une chaîne de Markov. En effet toute chaîne peut être définie intrinsèquement par le couple $(\mu, \mathbb{P}_{ij})$. Il est donc absolument nécessaire d'estimer pertinemment la fonction de transition d'une chaîne de Markov. C'est ce à quoi nous nous attacherons dans ce grain.

Notations

Soit $X^t = (X_0, \ldots, X_t)$ la réalisation de la chaîne X_n pour $n = 0, 1, \ldots\ t$ (jusqu'à l'instant t).

Notons

$$N_{ij}^t = \sum_{k=1}^{t} \mathbf{1}_{\{X_{k-1}=i, X_k=j\}}$$

$$N_i^t = \sum_{k=1}^{t} \mathbf{1}_{\{X_k=i\}}$$

avec

- $\boldsymbol{N_{ij}^t}$*le nombre de transitions de* i *à* j *sur l'intervalle de temps* $[1, t]$.
- $\boldsymbol{N_i^t}$*le temps de passage dans l'état* i *sur l'intervalle de temps* $[1, t]$.

On se propose de donner une estimation de la fonction de transition $\mathbb{P}_{ij}$ en utilisant la méthode du maximum de vraisemblance.

Estimateur du Maximum de Vraisemblance

La fonction du maximum de vraisemblance à l'instant $t \in \mathbb{N}$ est donnée par :

$$\boldsymbol{L_t(\mathbb{P}) = p(X_0)p(X_0, X_1) \ldots\ldots p(X_{t-1}, X_t),}$$

où $p(X_{k-1}, X_k) = \mathbb{P}_{X_{k-1}, X_k}$ et $p(X_0)$, la loi initiale de (X_n) (noté (μ_{X_0})) .

De là il vient facilement

$$L_t(\mathbb{P}) = p(X_0) \prod_{k=1}^{t} p(X_{k-1}, X_k),$$

$$L_t(\mathbb{P}) = p(X_0) \prod_{(i,j)\in E} \mathbb{P}_{ij}^{N_{ij}^t}.$$

En passant par la log-vraisemblance, on a

$$L_t(\mathbb{P})_{log} = log(p(X_0)) + \sum_{(i,j)\in E} N_{ij}^t.log(\mathbb{P}_{ij})$$

$$\text{avec } \left\{ \begin{array}{l} \sum_{i\in E} \mathbb{P}_{ij} = 1 \\ N_i^t = \sum_{i\in E} N_{ij}^t \end{array} \right. \tag{I.2}$$

Puis, par maximisation de la log-vraisemblance, on établit le théorème suivant :

THÉORÈME I.2.2

Si E est un ensemble régulier, alors l'estimateur du maximum de vraisemblance pour $\mathbb{P}_{ij}$ est donné par :

$$\widehat{\mathbb{P}}_{\boldsymbol{ij}} = \frac{\boldsymbol{N_{ij}}}{\boldsymbol{N_i}} \tag{I.3}$$

[9], [4]

On estime ainsi que la probabilité de passage de l'état i vers l'état j (par exemple le passage de la température 12 degré Celsius à la température 8 degré Celsius) est égale au nombre de passages de i vers j de l'ensemble d'apprentissage divisé par le nombre de passages dans l'état i (Cf. Bartlett et Billingsley).

On utilisera ce théorème dans l'algorithme d'estimation de la matrice de transition.

I.2.3 Premières Simulations

Documents :
Annexe B.1

Précision sur les tracés des simulations :

Ce grain [1] propose les premiers tracés des simulations obtenues à l'aide de la modélisation décrite précédement. La programmation s'est faite en environnement `MatLab` avant d'être réécrite en langage `R`.

On s'aperçoit sur les graphes suivants que l'allure générale des courbes n'est pas représentative des courbes réelles de températures. Cependant, l'étude d'un modèle stochastique se fait en convergence et on ne peut juger d'un modèle sur une seule ou quelques simulations. Le prochain grain s'attachera à fournir ce type de résultats. Les figures de simulations proposés ici tendent simplement à vérifier qu'il n'y a pas de défaillance "grave" numérique et ne peuvent fournir des résultats considérés comme significatifs.

Résultats :

- La figure I.2 représente le tracé d'une année de simulation.
- La figure I.3 représente les données réelles suivies d'un an de simulation.
- A titre indicatif, nous montrons ici les résultats obtenus pour une modélisation du même type, mais avec une chaîne de Markov d'ordre 2 : la figure I.4 représente un an de simulations en profondeur 2. Notons que de petits paliers se forment, mais la courbe est loin d'être une courbe en cloche, comme le seraient des relevés classiques de températures. D'autres solutions sont envisageables, comme le passage à une chaîne non homogène par exemple (cf. prochain chapitre).

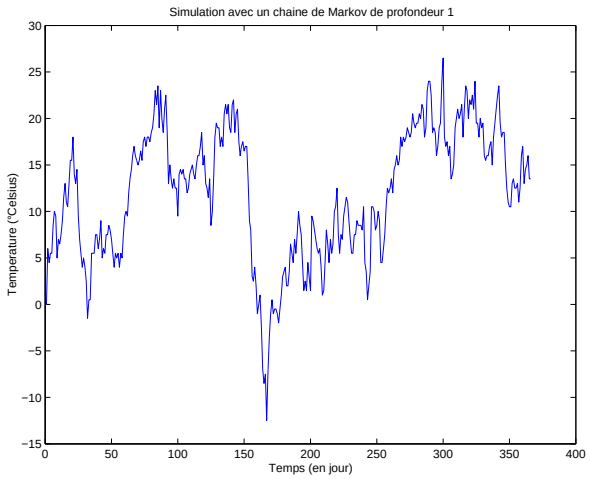


FIG. I.2 – Tracé d'une année de simulations (ordre 1)

[1] un grain est l'unité de sens de plus bas niveau en typographie

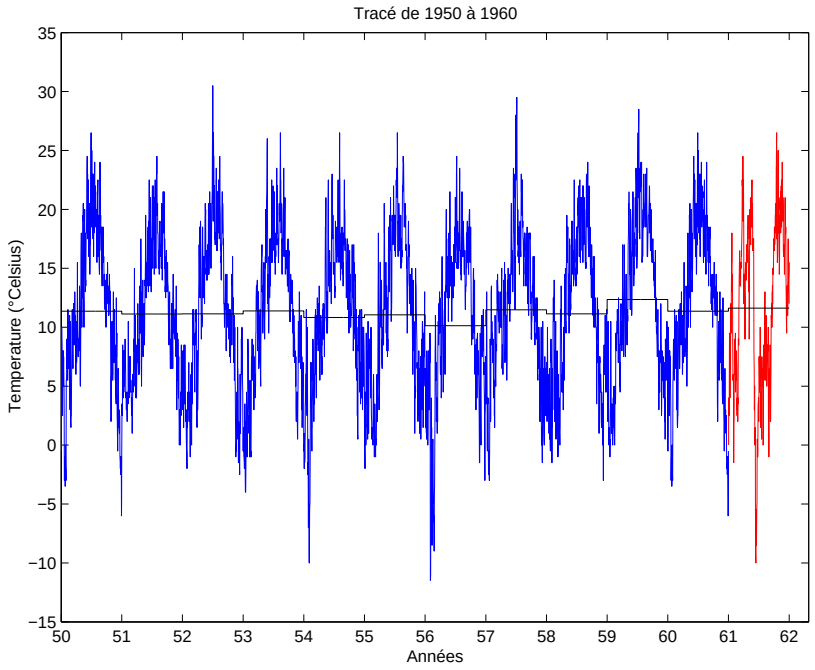


FIG. I.3 – Tracé réel suivi d'un an de simulations

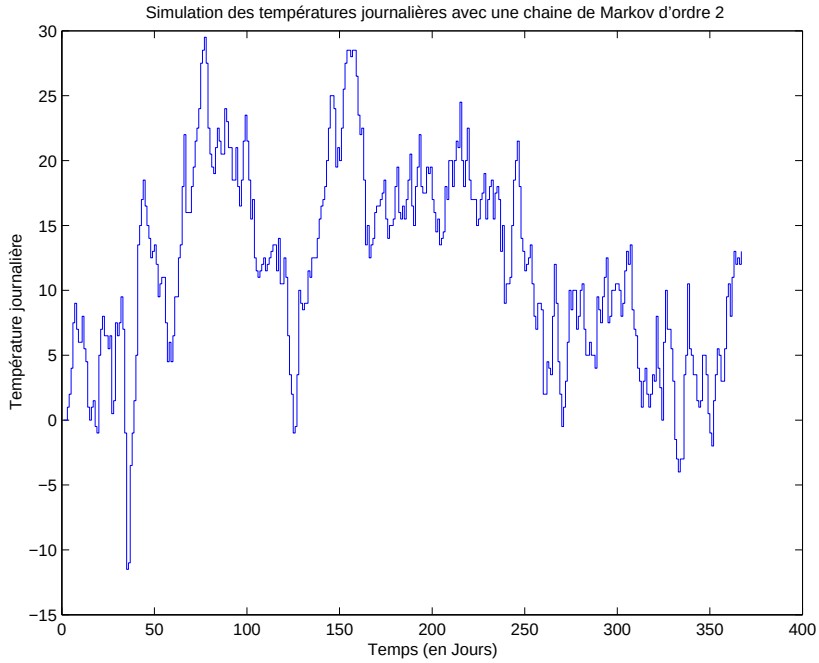


FIG. I.4 – Tracé d'une année de simulations (ordre 2)

I.2.4 Estimation de l'espérance de la température moyenne

PRINCIPE I.2.3

On se propose de comparer la moyenne calculée analytiquement sur notre modèle avec la moyenne réelle. Pour ce faire, on estime la matrice de transition sur x années de données (par exemple : $[1950; 1990]$). On prend pour loi initiale de la simulation l'instant correspondant au dernier jour des données. On calcule ensuite la moyenne analytique pour ce type de modèle d'ordre 1 sur les y années qui suivent (par exemple : $[1991; 2000]$) que nous allons comparer avec la moyenne issue des données réelles.

Analytiquement, la formule de l'espérance à l'instant t est fonction d'une puissance de la matrice de transition, de la loi intiale de la chaîne et des états $i \in E$. Si l'on note $\mathbb{E}_{i_0}[X_t]$ l'espérance de X à l'instant t sachant qu'on a pris pour instant initial l'état i_0, on aura :

$$\mathbb{E}_{i_0}[X_t] = \sum_{i \in E} P_{i_0}(X_t = i).i$$

et donc

$$\mathbb{E}_{i_0}[X_t] = \sum_{i \in E} \mathbb{P}^t(i_0, i).i$$

D'où l'estimateur

$$\widehat{\mathbb{E}}_{i_0}[X_t] = \sum_{i \in E} \widehat{\mathbb{P}}^t(i_0, i).i \tag{I.4}$$

La moyenne analytique du modèle d'ordre 1 de 1991 à 2000 sera obtenue en calculant à chaque instant t compris entre 1991 et 2000 l'espérance avec la formule (I.4). Une fonction numérique permet le calcul analytique pour tous les instants au-delà du dernier instant de l'apprentissage. On prend pour loi intiale $(\text{en } i_0)$ la valeur du dernier instant des données d'apprentissage. On trace ainsi les points obtenus pour chaque instant et on compare avec la moyenne réelle, issue des données.

Résultats

- La figure I.5 illustre le principe décrit ci-dessus
- La figure I.6 fournit les résultats obtenus pour une matrice $\mathbb{P}_{ij}$ estimée entre 1950 et 1990.
- La figure I.7 fournit les résultats obtenus pour une matrice $\mathbb{P}_{ij}$ estimée entre 1950 et 1960.

REMARQUE I.2.4

Pour le calcul de l'espérance, les résultats obtenus avec une estimation sur la période $[1950 - 1960]$ (fig I.7) sont meilleurs car cette période climatique est plus stable que la totalité de la période $[1950 - 1990]$ (fig I.6) : les 10 dernières années s'inscrivent dans un réchauffement climatique notable, à la vue des données fournies. Du fait de l'instabilité des données, les résultats paraîssent moins bons pour une estimation réalisée paradoxalement sur une plus longue période. [7]

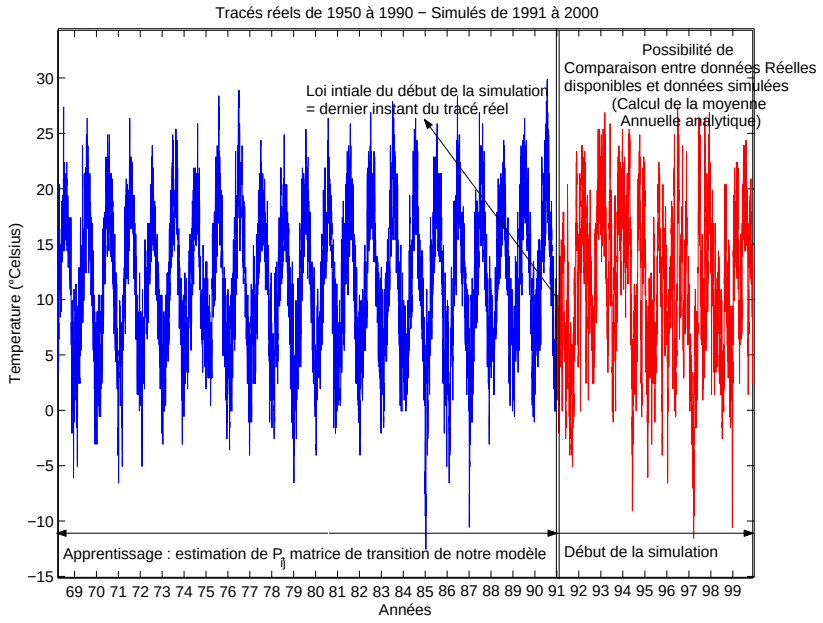


FIG. I.5 – Schéma de Principe

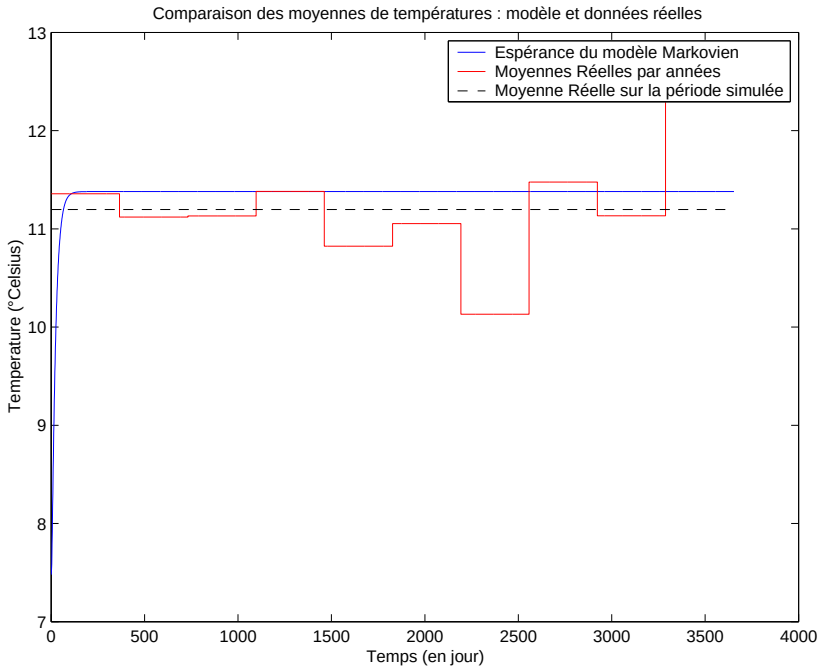


FIG. I.6 – Comparaison après apprentissage sur 40 ans (-> 1990)

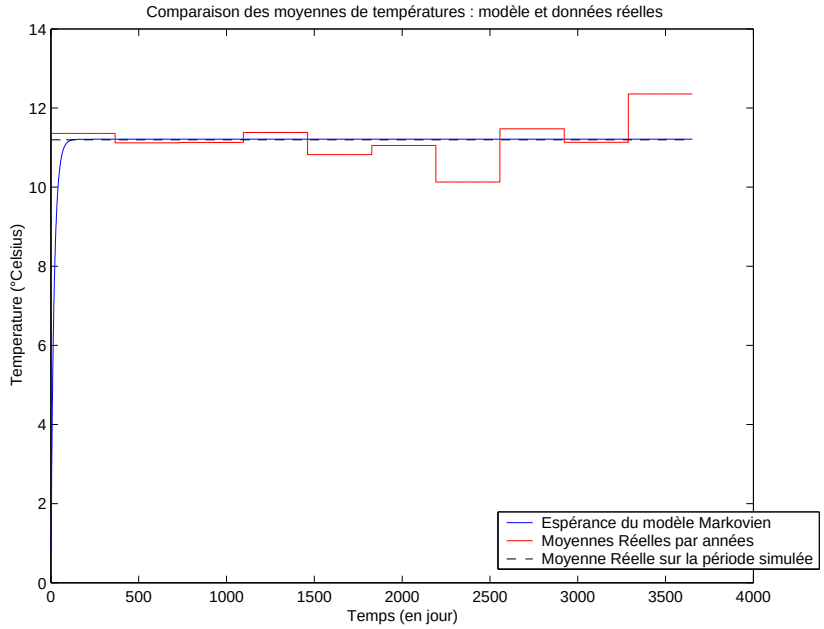


FIG. I.7 – Comparaison après apprentissage sur 10 ans (-> 1960)

REMARQUE I.2.5

En ce qui concerne les résultats I.6 et I.7, on tend en moyenne journalière vers la moyenne théorique annuelle : la distribution de la chaîne n'épouse pas du tout la forme d'une courbe de températures réelle (en "cloche" dans les pays au climat tempéré). Si l'on souhaite obtenir des résultats qui s'apparentent plus à la réalité, il nous faut dans ce cas utiliser des hypothèses markoviennes plus forte sur l'homogénéité ou sur la profondeur.

Plusieurs questions se posent alors :

- Comment s'affranchir du problème d'instabilité des données due au réchauffement climatique observé sur l'ensemble d'apprentissage ? (fig I.6, remarque I.2.4)
- Faut-il augmenter la profondeur de la chaîne ?
- Faut-il passer au cas non-homogène, c'est à dire changer de matrice de transition au cours du temps ? (cf. remarque I.2.5)

Le passage à un modèle markovien non-homogène permettrait de traduire une hypothèse physique intéressante en matière de phénomène climatique : les probablités de transition vers un état (dans ce cas une température) dépendraient de la période de l'année et on ferait ainsi apparaître la notion de saisonalité dans notre modèle. Augmenter la profondeur de la chaîne pourrait traduire les phénomènes de paliers observables sur les données réelles, mais ferait perdre les calculs analytiques, d'autant que l'information sur les paliers de températures peut être considérée comme partiellement traduite par le passage à une chaîne non-homogène. C'est sur ces réflexions que nous avons opté pour le modèle qui suit, présenté dans le prochain chapitre : un modèle se basant sur une chaîne de Markov non-homogène de profondeur 1.

Chapitre II

Estimation des températures par chaîne de Markov non-homogène

II.1 Modélisation

Physiquement, il est logique de penser que les probabilités de transition changent selon les périodes de l'année (les saisons). Il nous semble légitime de penser que les distributions de probabilités des températures ne sont pas les mêmes en hiver et en été. Le fait de changer de loi se traduit par le passage au cas du Markov non-homogène. Dans notre étude les lois sont estimées de manière non-paramétrique sur un échantillon défini sur une période donnée (une "saison"), par estimation du maximum de vraisemblance. Ainsi nous aurons autant de matrices de transitions que de périodes de découpage (de "saisons") estimées chacune sur la période lui correspondant.

Après avoir testé différents modèles, nous avons choisi de changer de matrice de transition toutes les semaines. C'est pour cette périodicité que l'on obtenait les résultats les plus corrects sans pour autant faire du sur-apprentissage sur les données.

→ Modélisation journalière des températures avec une chaîne de Markov non-homogène d'ordre 1

II.1.1 Notations

PRINCIPE II.1.1

Par rapport au premier modèle, l'idée essentielle est de changer de matrice de transition au cours du temps. Dans ce cas, la fonction de transition n'est plus indépendante du temps : on parle alors de chaîne de Markov non-homogène.

Soit la chaîne de Markov non-homogène $(X_n)_{n\in\mathbb{N}}$ décrivant les températures aux instants n :

- On choisit de changer de matrice de transition chaque semaine de l'année. On construit donc 52 matrices, notées $\mathbb{P}_{S1}, \ldots, \mathbb{P}_{Sk}, \ldots, \mathbb{P}_{S52}$.
- La fonction de transition est notée $\mathbb{P}_n(ij)$, où n est discrétisé par unité de temps journalière.

  ainsi,

$$\begin{cases} \mathbb{P}_n(ij) = \mathbb{P}_{S1}(ij) & \text{pour } n = 1 \text{ à } 7 \\ \mathbb{P}_n(ij) = \mathbb{P}_{S2}(ij) & \text{pour } n = 8 \text{ à } 14 \\ \vdots & \\ \text{jusqu'à } \mathbb{P}_{S52}(ij) & \text{pour } n = 358 \text{ à } 365 \end{cases}$$

- On se donne de nouveau une loi initiale pour la chaîne : μ_{X_0}
- On discrétise cette fois encore l'espace d'état sur des $\frac{1}{2}$ degrés :

$$E = \{-12.5, -12, \ldots, 29.5, 30, 30.5\}$$

→ (X_n) est une *chaîne de Markov non-homogène d'ordre 1*

II.1.2 Estimation de la fonction de transition $\mathbb{P}_{Sn}(ij)$

Méthode d'estimation / apprentissage :

PRINCIPE II.1.2

Dans le cadre du premier modèle markovien homogène, on ne considérait qu'une seule matrice de transition constituant la fonction de transition du modèle. On considère ici 52 matrices de transition. Pour le premier modèle, l'estimation de l'unique matrice $\mathbb{P}$ se faisait sur tout l'ensemble d'apprentissage. L'estimation des matrices $\mathbb{P}_{Sn}$ est faite sur les semaines n de chaque année de l'ensemble d'apprentissage.

Exemple :

Soit l'ensemble d'apprentissage les années 1950 à 2000. Pour l'estimation de $\mathbb{P}_{S1}$, on isolera toutes les semaines 1 des années 1950 à 2000 (cf. schéma II.1). On crée ensuite une matrice de transition correspondant à la semaine 1, en comptant la fréquence des passages des états i vers j pour estimer les N_i et N_{ij} de cette semaine. Nous les noterons donc N_{iSk} et $N_{ij\,Sk}$

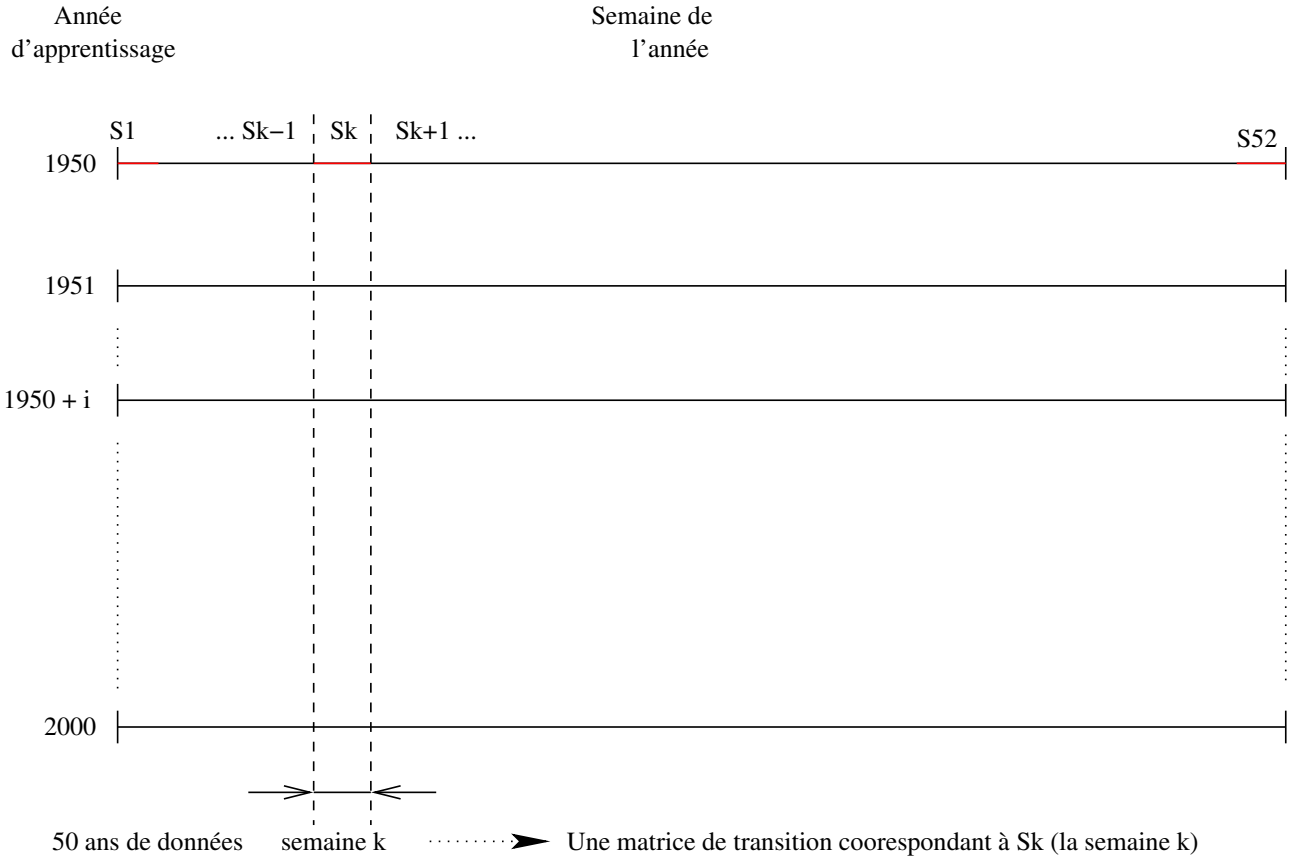


FIG. II.1 – Schéma explicatif de la méthode d'estimation des matrices de transition

Modification de l'estimateur :

Dans le modèle précédent, si un état n'existait pas dans l'ensemble d'apprentissage (i. e. si $N_i = 0$), on considérait qu'il était "absorbant" dans la matrice de transition. Dès lors, on avait :

$$\forall i \text{ tels que } N_i = 0 \begin{cases} \widehat{\mathbb{P}}(ij) = 1 & \text{si } i = j \\ \widehat{\mathbb{P}}(ij) = 0 & \text{si } i \neq j \end{cases}$$

Ce choix n'était pas forcément logique si l'on réfléchit au problème qui nous est posé : ça n'est pas parce qu'une température n'est pas observée sur l'ensemble d'apprentissage qu'elle est physiquement impossible, or c'est comme cela que nous la traitions jusqu'à présent. Ceci est d'autant plus vrai que le fait de passer au markovien non-homogène et d'estimer chaque matrice de transition semaine par semaine réduit chacun des ensembles d'apprentissage. Nous avons donc pensé qu'il était judicieux d'estimer les probabilités de transition de ces états non-visités autrement qu'en les considérant comme des états récurrents.

Notre nouvelle méthode d'estimation des matrices de transition est la suivante :

– Pour les états "visités" dans l'ensemble d'apprentissage (i. e. $N_i \neq 0$), pour une matrice $\mathbb{P}_{Sk}$, on conserve l'estimateur de Billingsley, qui reste l'estimateur du maximum de vraisemblance, i.e. :
$$\widehat{\mathbb{P}_{Sk}}(ij) = \frac{N_{ij\,Sk}}{N_{i\,Sk}}$$

– Pour les états que l'on n'a pas "rencontrés" dans l'ensemble d'apprentissage (i. e. $N_i = 0$), on estime la ligne i de la matrice $\mathbb{P}_{Sk}$ par la loi stationnaire de cette matrice. Nous justifierons cet estimateur dans le grain suivant.

Nouvel estimateur de $\widehat{\mathbb{P}_{Sk}}$:

Pour résumer, on aura :

$$\widehat{\mathbb{P}_{Sk}}(ij) = \begin{cases} \frac{N_{ij\,Sk}}{N_{i\,Sk}} & \text{pour } i \text{ tel que } N_i \neq 0 \\ \Pi_{Sk}(j) & \text{pour } i \text{ tel que } N_i = 0 \end{cases}$$

où on a noté Π_{Sk} la loi stationnaire de la matrice $\mathbb{P}_{Sk}$ estimée au sens du maximum de vraisemblance (cette loi existe bien malgré l'existence d'états absorbants puisqu'ils ne peuvent être atteints depuis les autres états).

$$\text{avec la contrainte } \sum_j \mathbb{P}_{Sk}(i,j) = 1, \forall i \in E, \forall k = 1 \ldots 52$$

On garantit ainsi la *connexité* de la chaîne, lors d'un changement de semaine, c'est-à-dire lors du passage d'une matrice de transition à une autre.

II.1.3 Estimation des distributions empiriques des températures

Idée

PRINCIPE II.1.3

L'idée est d'approximer la distribution de probabilités des températures de chaque semaine à partir des données réelles par la loi stationnaire de la semaine correspondante, afin d'estimer les probabilités de transition pour les états non visités sur l'ensemble d'apprentissage (états pour lesquels on a a priori aucune information). Il est aisé, à partir des données journalières, de tracer une distribution de probabilités d'une semaine Sk sur l'espace d'état E.

Commençons par rappeler la définition de la loi stationnaire d'une chaîne de Markov :

Loi Stationnaire : définition et propriétés

DÉFINITION II.1.4

Une loi de probabilité Π sur un espace d'état E est dite stationnaire pour la chaîne de Markov X si, $\forall j \in E$ on a

$$\sum_{j \in E} \Pi(i)p(i,j) = \Pi(j)$$

On peut traduire cette relation matriciellement :

$$\Pi\mathbb{P} = \Pi$$

On parle de loi stationnaire car pour une chaîne de Markov X_n ayant Π pour la loi initiale et vérifiant $\Pi\mathbb{P} = \Pi$ pour sa matrice de transition, on a :

$$\mathbf{Loi}(X_n) = \Pi, \forall n \geq 0$$

Ceci est vrai pour une chaîne de Markov homogène. Notre modèle est basé sur une chaîne non-homogène sur une annéee, mais qui peut être considérée comme homogène par morceaux sur chaque semaine de l'année ; d'autant que l'espace d'état reste le même, malgré le changement hebdomadaire de matrice de transition.

La propriété essentielle suivante de la loi stationnaire permet de comprendre pourquoi on a songé à l'utiliser pour estimer les probabilités de transitions pour les états non-visitées sur l'ensemble d'apprentissage :

THÉORÈME II.1.5

Si X_n est une chaîne de Markov ergodique, alors

$$lim_{n \to \infty} \mathbb{P}_{ij}^n = \Pi_j, \forall i,j \in E$$

[12]

Cette propriété permet de bien comprendre que la probabilité stationnaire Π_i représente donc *la proportion de temps que la chaîne de Markov passe à l'état i à la longue.* D'où l'idée qu'il est judicieux de l'utiliser pour estimer les probabilités de passage pour les états non estimés.

Ainsi, on aura , $\forall k = 1, \ldots, 52$, une loi stationnaire par semaine et par matrice de transition :

$$\Pi_{Sk}\mathbb{P}_{Sk} = \Pi_{Sk}$$

Application à l'estimation des distributions empiriques hebdomadaires

La figure II.2 fournit un exemple comparant les distributions empiriques issues des données réelles avec les distributions issues des lois stationnaires correspondantes.

On réalise la même identification pour chaque semaine de l'année : on dispose ainsi d'une loi stationnaire pour chaque matrice Sk de chaque semaine de l'année. Ainsi,

$$\forall i \text{ tel que } N_i = 0, \forall j \in E \text{ et } \forall k \in \{1, \ldots, 52\}$$
$$\widehat{\mathbb{P}_{Sk}}(i,j) = \widehat{\Pi}_{Sk}$$

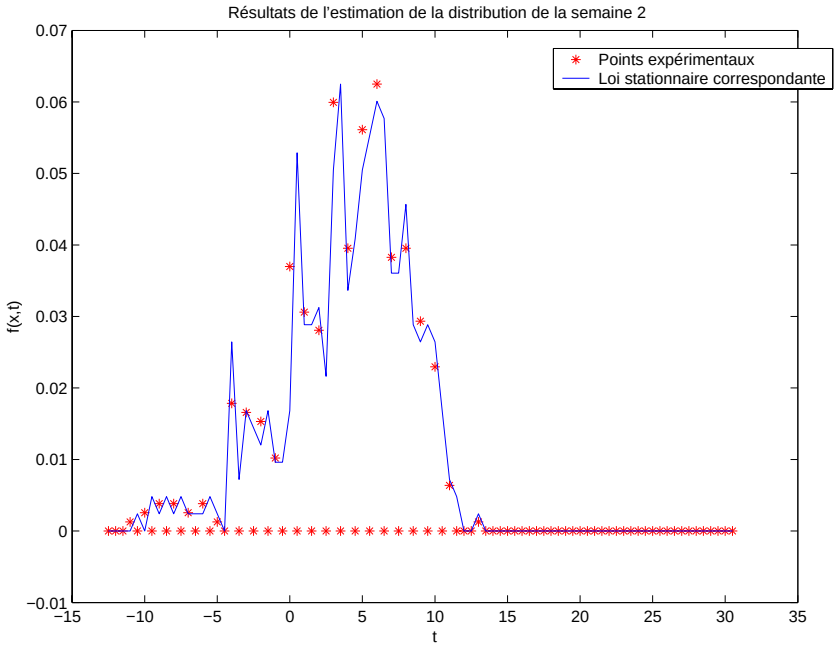


FIG. II.2 – Estimation des distributions de probabilité par loi stationnaire - semaine 2

Quel intérêt de modifier l'estimateur pour les états non-définis ?

Le fait d'estimer les probabilités de transition pour les états non visités prend tout son intéret du fait de la non-homogénéité de la chaîne : en effet, on ne risque pas, lors du passage d'une matrice de transition Sk à une matrice $Sk+1$ de rencontrer un passage d'un état i à j qui ne serait pas défini. Ainsi, on obtient des résultats pour les calculs analytiques (qui utilisent les puissance $n^{\text{ième}}$ des matrices de transition) beaucoup plus satisfaisants, approchant beaucoup mieux les courbes réelles de températures.

II.1.4 La Méthode de Monte-Carlo

Documents :
Annexe A.3

Nous énonçons dans ce grain le principe de la méthode de Monte-Carlo que nous avons utilisée de manière récurrente pour vérifier nos calculs analytiques et estimer des probabilités dont le calcul analytique s'avérait trop complexe ou impossible.

Le principe de la méthode

PRINCIPE II.1.6

La méthode de Monte-Carlo est une méthode très puissante de simulation Stochastique qui va permettre d'approcher le problème d'une autre manière que par résolution analytique.

Imaginons que l'on veuille calculer une probabilité du style $P_{i_0}(X_n = j)$, c'est à dire la probabilité que la température au temps $t = n$ soit j alors que la température initiale, à l'instant $t = 0$ était i_0.

La méthode de Monte-Carlo consiste à répéter un grand nombre de fois l'expérience aléatoire par simulation numérique, et compter $+1$ à chaque fois que l'événement dont on veut estimer la probabilité se produit, et de diviser le résultat obtenu par le nombre d'expériences réalisées.

La loi forte des grands nombres pour les chaînes de Markov

Intuitivement, on sent bien que ce nombre va tendre vers l'espérance de $\mathbb{E}_i(\mathbf{1}_{\{X_n=j\}})$. Cela découle en fait directement de la loi forte des grands nombres pour les chaînes de Markov, que nous rappelons :

THÉORÈME II.1.7

Soit $(X_n)_{n\in\mathbb{N}^*}$ une chaîne de Markov ergodique de loi stationnaire Π et Φ une fonction réelle sur E, alors

$$\frac{1}{n+1}\sum_{k=0}^{n}\Phi(X_k) \longrightarrow \sum_{k\in E}\Phi(k)\Pi(k), n \to \infty, \textit{p.s.}$$

Posons la variable Y telle que, à l'expérience k, Y_k vaut 1 si $X_n = j$ et 0 sinon, c'est-à-dire $Y_k = \Phi(j) = \mathbf{1}_{\{X_n^k=j\}}$. Les Y_k suivent donc une loi de *Bernouilli* de paramètre p, avec p la probabilité d'atteindre l'état j à l'instant n en partant de l'état i_0 initialement. On peut appliquer le théorème de la loi forte des grands nombres pour les chaînes de Markov et l'on peut approcher cette probabilité par Monte-Carlo.

$$\text{On a : } \frac{S_n}{n} = \frac{Y_1 + \ \ldots\ + Y_n}{n} \longrightarrow \mathbb{E}(Y_k), \text{ si } n \longrightarrow \infty$$

La méthode Monte-Carlo consiste à calculer $\frac{S_n}{n}$ pour n suffisamment grand, ce qui estime $\mathbb{E}(Y_k)$, c'est à dire $P_{i_0}(X_n = j)$.

Monte-Carlo et simulation numérique

On peut ainsi estimer toute probabilité numériquement, sans avoir forcément recours à un calcul analytique. L'explosion de la puissance de calcul des ordinateurs personnels et non-plus uniquement celle de super-calculateurs dédiés à cet emploi en font une méthode très puissante pour l'estimation des probabilités dont la résolution analytique est impossible. Mais dans le cas où le calcul analytique est possible, son résultat est donc beaucoup plus intéressant, puisqu'il est la convergence des simulations par Monte-Carlo pour un nombre en théorie infini d'itérations. Dans le cadre du stage, la méthode de Monte-Carlo a donc été utilisée pour vérifier la justesse des calculs analytiques, garantissant par ailleurs que l'estimation des matrices de transition était faite correctement.

Soulignons enfin que l'on peut déterminer le nombre d'expériences à réaliser avec Monte-Carlo en fonction d'une précision donnée sur le résultat en utilisant le théorème de *Bienaymée-Tchebychev*, dont le résultat est extrêment utilisé dans le cadre des simulations numériques (Voir le document associé : annexe A.3).

II.2 Résultats analytiques et numériques

Cette section vise à introduire quelques résultats qui vont nous permettre de juger la qualité de notre modèle.

II.2.1 Application au calcul de l'Espérance et de la Variance d'une année de simulation

Documents :
Annexe A.4

Tout comme au paragraphe I.2.4, on souhaite calculer l'espérance analytiquement à partir des puissances des matrices de transition. On espère cette fois-ci que l'espérance ne tendera pas simplement vers la moyenne annuelle, mais aura une forme plus proche d'une courbe de températures (courbe en cloche), du fait de la non-homogénéité de la chaîne.

Expressions de $\mathbb{E}[X]$ et de $Var[X]$

Rappelons la formule analytique de l'espérance d'une chaîne de Markov X à l'instant t sachant qu'on est parti à l'instant initial de l'état i_0.

$$\widehat{\mathbb{E}}_{i_0}[X_t] = \sum_{i \in E} \widehat{\mathbb{P}^t}(i_0, j).j \tag{II.1}$$

N'oublions pas que la chaîne est non-homogène, ainsi

$$\begin{cases} \text{pour } t = 1 \dots 7, & \widehat{\mathbb{P}^t} = \widehat{\mathbb{P}}_{S1} \\ \text{pour } t = 8 \dots 14, & \widehat{\mathbb{P}^t} = \widehat{\mathbb{P}}_{S2} \\ \text{etc..., jusqu'à } t \text{ final} & \end{cases}$$

$$\text{avec } \widehat{\mathbb{E}}_{i_0}[X_t] = \sum_{j \in E} \widehat{\mathbb{P}^1 \mathbb{P}^2} \dots \widehat{\mathbb{P}^t}(i_0, j).j$$

Pour le calcul de la variance, il suffit de se rappeler de la formule suivante pour le moment d'ordre 2 centré :

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{II.2}$$

$$\text{Soit } Var_{i_0}[X_t] = \mathbb{E}_{i_0}[X_t^2] - \mathbb{E}_{i_0}[X_t]^2 \tag{II.3}$$

Sur le même principe que pour l'espérance (II.1), on calcule l'écart-type analytique du modèle en utilisant la formule (II.3).

Convergence des estimateurs

Les quantités $\mathbb{E}_{i_0}[X_t]$ et $Var_{i_0}[X_t]$ sont des valeurs exactes obtenues analytiquement sur notre modèle, mais restent des *estimateurs* de l'espérance et de la variance *réelles* des températures journalières. Il est donc indispensable d'en étudier la convergence :

Pour ce faire, nous allons réécrire différemment espérance et variance de notre modèle (II.1) et (II.3), estimateurs des espérances et variances réelles.

Nous les noterons $\widehat{\mu_{i_0}^t}$ et $\widehat{\sigma_{i_0}^t}^2$.

Ecrivons par ailleurs l'instant t comme le multiple d'un numéro de semaine sommé au numéro du jour de la semaine courante, soit :

$$t = 7(k-1) + l, \text{ avec } k = 1, \ldots, 52 \text{ et } l = 1, \ldots, 7$$

– Pour l'estimateur de l'espérance $\widehat{\mu_{i_0}^t}$ on aura pour tout $t > 0$

$$\widehat{\boldsymbol{\mu_{i_0}^t}} = \begin{cases} \sum_{j \in E} \prod_{i=1}^{k-1} (\widehat{\mathbb{P}}_{Si})^7 . (\widehat{\mathbb{P}}_{Sk})^l (i_0, j).j & \text{pour } \boldsymbol{k > 1} \\ \sum_{j \in E} (\widehat{\mathbb{P}}_{Sk})^l (i_0, j).j & \text{pour } \boldsymbol{k = 1} \end{cases} \tag{II.4}$$

– De la même manière, on a pour l'estimateur de la variance $\widehat{\sigma_{i_0}^t}^2$, $\forall k > 0$:

$$\widehat{\boldsymbol{\sigma_{i_0}^t}}^{\boldsymbol{2}} = \begin{cases} \sum_{j \in E} (j - \widehat{\boldsymbol{\mu_{i_0}^t}})^2 \prod_{i=1}^{k-1} (\widehat{\mathbb{P}}_{Si})^7 . (\widehat{\mathbb{P}}_{Sk})^l (i_0, j).j & \text{pour } \boldsymbol{k > 1} \\ \sum_{j \in E} (j - \widehat{\boldsymbol{\mu_{i_0}^t}})^2 (\widehat{\mathbb{P}}_{Sk})^l (i_0, j).j & \text{pour } \boldsymbol{k = 1} \end{cases} \tag{II.5}$$

Rappelons que notre chaîne de Markov non-homogène est *homogène par morceaux* sur chaque semaine de l'année S_k. Ainsi, pour un nombre de trajectoires M de l'ensemble d'apprentissage tendant vers l'infini, du fait que les matrices $\mathbb{P}_{S_k}$ sont estimées au sens du maximum de vraisemblance, on peut écrire successivement :

$$\forall i, j \in E, \forall k = 1 \ldots 52 \text{ et pour } M \to \infty,$$

$$\widehat{\mathbb{P}}_{Sk}(i,j) \longrightarrow \mathbb{P}_{Sk}(i,j), \text{p.s.}$$

$$\widehat{\mathbb{P}}_{Sk}^7 \longrightarrow \mathbb{P}_{Sk}^7, \text{p.s.}$$

$$\widehat{\mathbb{P}}_{Sk}^l(i,j) \longrightarrow \mathbb{P}_{Sk}^l(i,j), \text{p.s.}$$

REMARQUE II.2.1

Voir l'annexe A.4 pour un rappel de la convergence presque sûre

Puisque l'on a exprimé les estimateurs $\widehat{\mu_{i_0}^t}$ et $\widehat{\sigma_{i_0}^t}^2$ dans les équations (II.4) et (II.5) uniquement à l'aide de puissances des matrices de transitions estimées dont on est assuré de la convergence presque sûre, on peut écrire

$$\left.\begin{array}{ll} \widehat{\mu_{i_0}^t} \longrightarrow \mu_{i_0}^t & \text{p.s} \\ \widehat{\sigma_{i_0}^t}^2 \longrightarrow \sigma_{i_0}^t & \text{p.s} \end{array}\right\} \text{pour } M \to \infty$$

où $\mu_{i_0}^t$ et ${\sigma_{i_0}^t}^2$ sont les espérance et variance réelles.

Résultats

Nous proposons les résultats pour l'espérance et l'écart-type calculés analytiquement et ceux calculés avec la méthode de Monte-Carlo pour 200 simulations de 50 ans chacune (soit 10000 ans de simulations, ce qui est très peu comparativement à la borne que l'on obtiendrait en utilisant le théorème de Bienaymé-Tchebichev).

On superpose aux résultats la moyenne et l'écart-type journaliers réels obtenus sur l'ensemble d'apprentissage.

- Les figures II.3 et II.4 fournissent les résultats pour l'espérance.
  On constate que Monte-Carlo converge vers la solution analytique assez rapidement, puisque les deux figures sont très proches, pour seulement 200 fois 50 années de simulations, soit 10000 ans. L'erreur donnée est calculée au sens des moindres carrés. Elle n'a que pour intérêt de comparer les deux figures, puisqu'elle n'est pas normée, et une erreur seule serait complètement arbitraire. Les résultats sont assez satisfaisants : on obtient une courbe assez lisse qui suit bien l'orientation globale de la courbe issue des données réelles. Le découpage en semaines semble être approprié : nous avons testé un découpage plus précis (sur 2 jours, par exemple) et dans ce cas le modèle colle "trop" aux données qui sont volatiles. Au contraire, il est préférable d'obtenir une courbe de l'espérance la plus lisse possible. Un découpage plus large (sur un mois, par exemple) donne des sauts trop importants entre chacune des périodes, dûs au passage d'une matrice de transition à une autre. La figure II.4 représente les différences journalières en valeur absolue entre l'espérance issue des données réelles et celle issue des calculs : la convergence du modèle est bonne, puisque la différence moyenne entre les données réelles et simulées et de 0.4 degré.
- Les figures II.5 et II.6 fournissent les résultats pour l'écart-type.
  Les résultats pour l'écart-type sont corrects d'autant que les résultats réels sont extrêment volatiles. Cependant, on pourrait être en droit d'attendre une courbe analytique encore plus lisse. Mais dès lors, on perdrait la bonne simulation des queues de distributions, c'est-à-dire les valeurs extrêment chaudes et surtout froides qui nous intéressent tout particulièrement dans le cardre de notre étude pour Gaz de France. Pour la différence entre les résultats par calculs analytiques et réels, on perçoit des pics dûs justement à la très grande volatilité de la variance réelle. Globalement, on conserve une erreur en valeur absolue correcte. Notons que le premier trimestre de l'année est le plus difficile à simuler, car le moins stable.

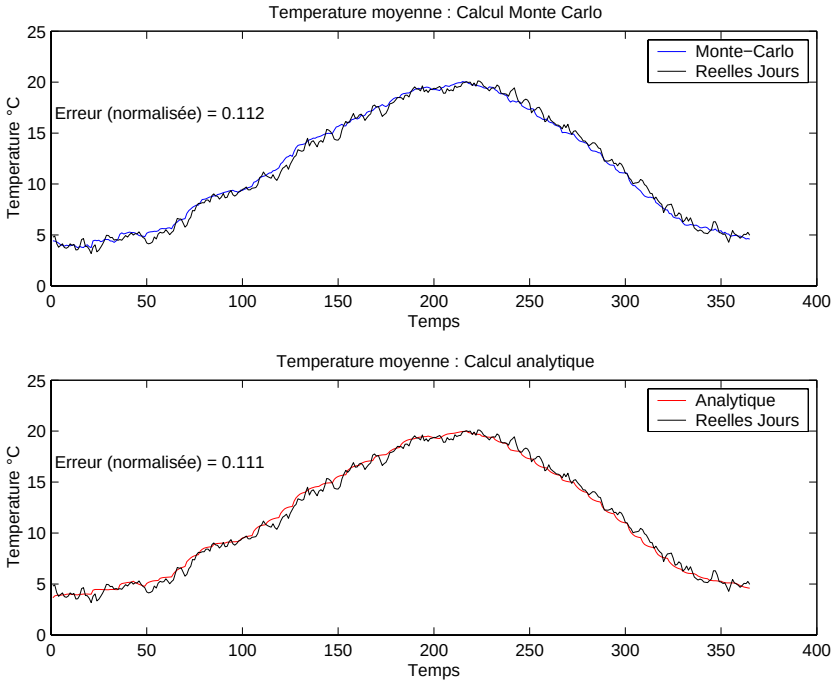


FIG. II.3 – Espérance d'une année de température

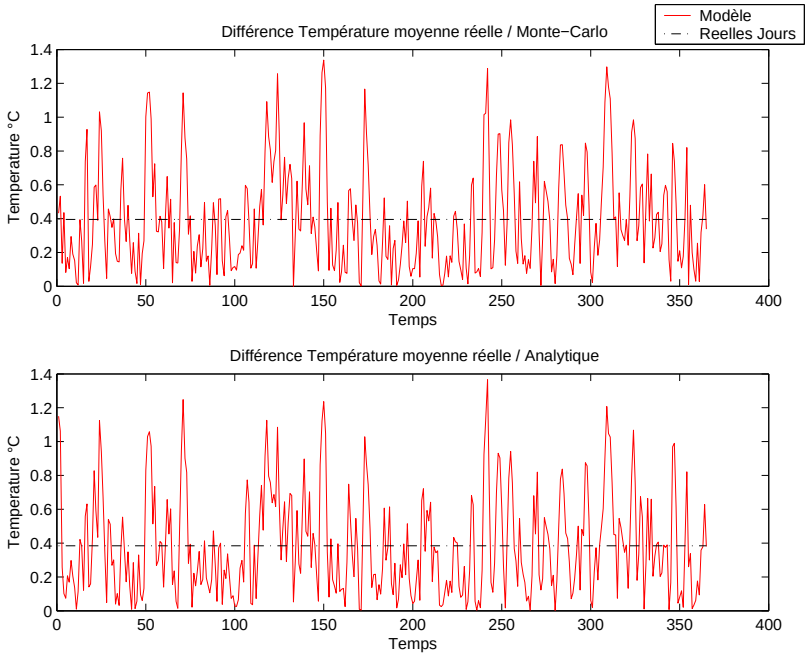


FIG. II.4 – Erreur journalière pour l'espérance d'une année de température

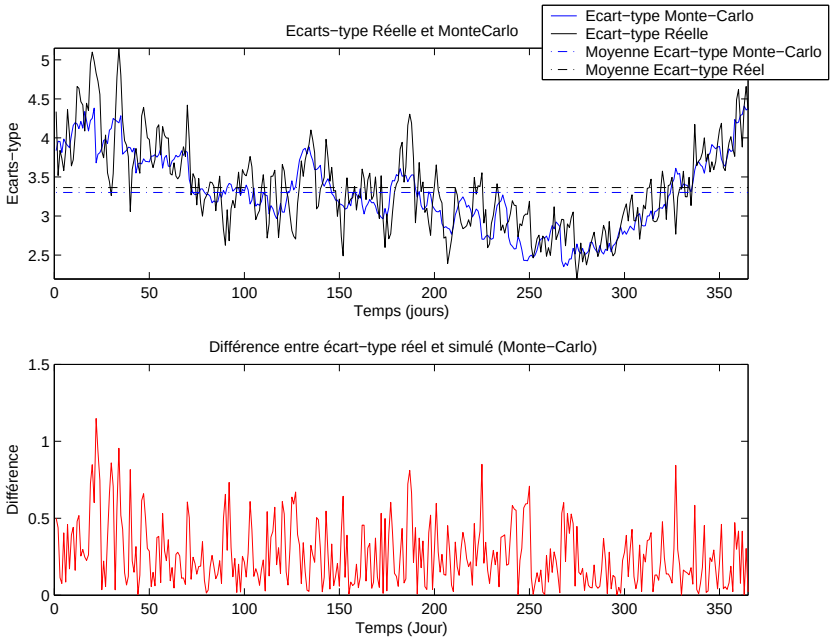


FIG. II.5 – Ecart-type et erreur journalière d'une année de température (Monte-Carlo)

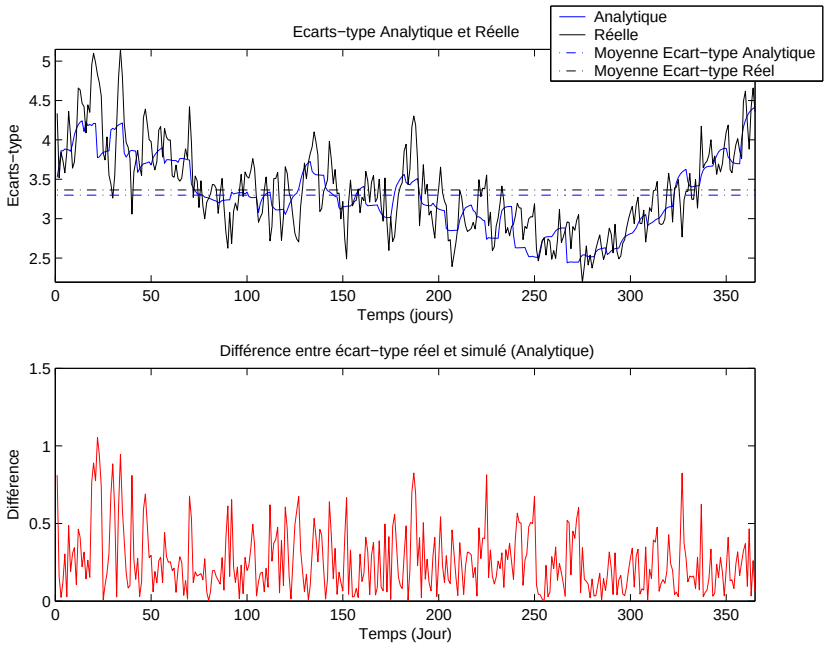


FIG. II.6 – Ecart-type et erreur journalière d'une année de température (Analytique)

II.2.2 Application au calcul des degrés jours cumulés mensuels

Des élements statistiques intéressants dans le cadre de la modélisation des phénomènes climatiques sont les degrés jours (DJ) cumulés mensuels (CDD et HDD) dont nous rappelons les définitions [1] :

Définitions : Degrés Jours Cumulés, CDD et HDD

DÉFINITION II.2.2

Soit T_i la température au jour i.

On définit HDD_i("heating degree-days") et CDD_i("cooling degree-days") de la manière suivante :

$$HDD_i = max\{17 - T_i, 0\}$$

$$CDD_i = max\{T_i - 17, 0\}$$

17 *est la température seuil de chauffage donnée par Gaz de France.*

Nous proposons de donner les HDD et CDD *"cumulés mensuels"*, c'est à dire $\sum_j HDD_j$, respectivement $\sum_j CDD_j$ où les j décrivent tous les jours pour un mois donné. Les DJ cumulés mensuels représentent donc la somme de tous les degrés d'un même mois où l'on estime que la population chauffera ou climatisera (selon que l'on s'intéresse aux HDD ou aux CDD).

Calcul analytique des Degrés Jours Cumulés Mensuels

Pour un mois k donné, la formule des degrés jours cumulés mensuels s'écrit :

$$DJ_{mois_k} = \begin{cases} \sum_{j=1}^{nbJmois_k} HDD_j \\ \sum_{j=1}^{nbJmois_k} CDD_j \end{cases}$$

soit

$$DJ_{mois_k} = \begin{cases} \sum_{j=1}^{nbJmois_k} max\{17 - T_j, 0\} \\ \sum_{j=1}^{nbJmois_k} max\{T_j - 17, 0\} \end{cases}$$

Pour le calcul analytique des DJ, on se souviendra simplement de la définition de l'espérance de toute variable aléatoire discrète X à valeur dans E, et de sa propriété principale, à savoir :

$$\mathbb{E}[X] = \sum_{x \in E} P(X = x).x$$

et

$$\mathbb{E}[\varphi(X)] = \sum_{x \in E} P(X = x).\varphi(x)$$

En posant successivement

$$\varphi(x) = HDD(x) = max\{17 - T_x, 0\}$$

et

$$\varphi(x) = CDD(x) = max\{T_x - 17, 0\}$$

Puis en adaptant le calcul de l'espérance au cas des chaînes de Markov, on trouve pour formule analytique des DJ cumulés mensuels :

$$\mathbb{E}[DJ_{mois_k}(HDD)] = \sum_{t=1}^{nbJmois_k} \sum_{j \in E} \widehat{\mathbb{P}}^t(i_0, j).max\{17 - T_j, 0\}, \text{ pour les } HDDs$$

$$\mathbb{E}[DJ_{mois_k}(CDD)] = \sum_{t=1}^{nbJmois_k} \sum_{j \in E} \widehat{\mathbb{P}}^t(i_0, j).max\{T_j - 17, 0\}, \text{ pour les } CDDs$$

Résultats

- La figure II.7 fournit les résultats des HDD et CDD cumulés mensuels calculés avec les T_i simulés par la méthode de Monte-Carlo et les T_i observés. Les résultats ont été obtenus en réalisant 200 simulations de 50 années chacune, soit 10000 ans de simulations. Sur ces tracés, on a calculé pour chaque année les HDD et CDD dont on a fait une moyenne par la suite. La figure II.8 donne les résultats du calcul analytique sur le modèle pour les HDD et CDD, soit la convergence de Monte-Carlo.
- Le tableau II.9 donne les valeurs numériques des HDD calculées sur les données réelles (l'historique des températures de 1950 à 2001), sur les simulations obtenues avec le modèle markovien (Monte-Carlo), les résultats analytiques du modèles et les valeurs actuellement utilisées par Gaz de France. [6]

Commentaires des résultats

REMARQUE II.2.3

Pour le graphe II.7, les résultats des calculs des DJ cumulés mensuels sont corrects, même si l'on pourrait souhaiter avoir une précision plus grande encore. On voit bien ici tout l'intérêt des résultats analytiques II.8, qui confirment les résultats de Monte-Carlo : ils sont non seulement plus précis mais ils permettent en plus de savoir vers quoi converge exactement le modèle.

REMARQUE II.2.4

En ce qui concerne le tableau comparatif des valeurs des Degrés journaliers cumulés HDD, les résultats sont très satisfaisants, dans le sens où l'on retrouve des valeurs proches des HDD cumulés issus des données réelles et de ceux utilisés par Gaz de France.

Notons que les données de GDF ont été obtenus avec le logiciel Prosper, mis au point par GDF en utilisant un modèle à persistance [2] pour les données températures, dans le but d'estimer des données telles les DJ Cumulés ou les premiers moments (espérance et variance centrée réduite). Les résultats du modèle markovien semblent bien compléter les résultats du modèle à persistance et seront donc utilisés lors de la réécriture prochaine du logiciel Prosper.

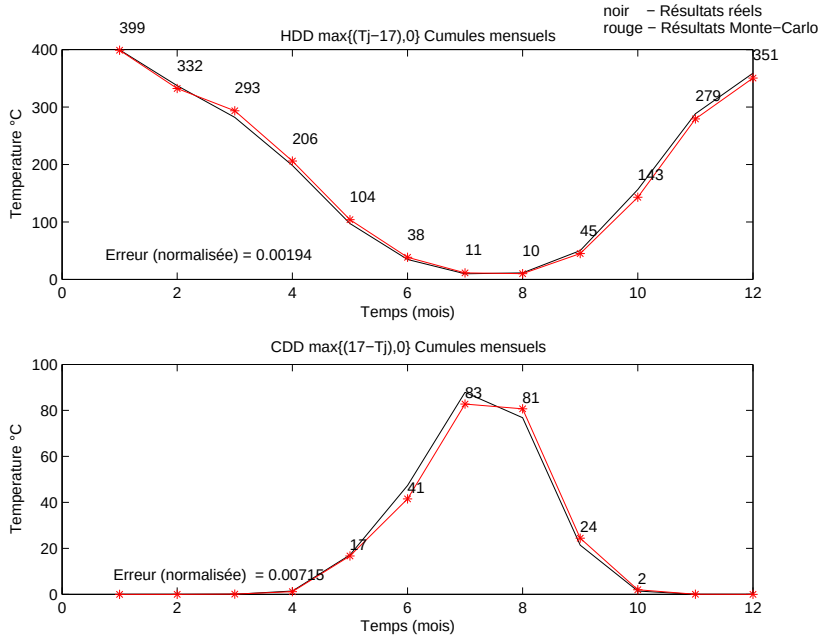


FIG. II.7 – DJ cumulés mensuels : HDD et CDD (Approximation Monte-Carlo)

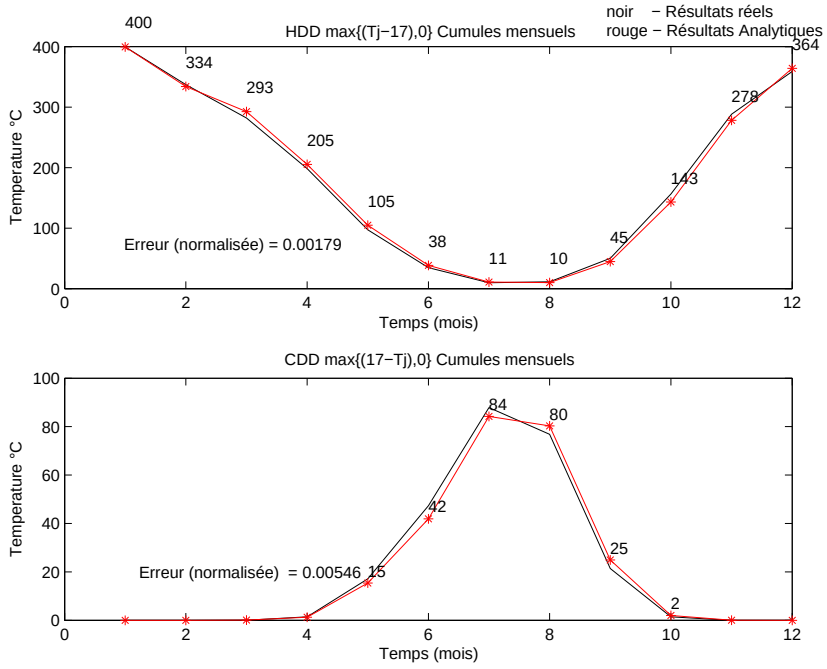


FIG. II.8 – DJ cumulés mensuels : HDD et CDD (Calcul Analytique)

Résultats	Jan	Fév	Mar	Avr	Mai	Jun	Jul	Aou	Sep	Oct	Nov	Déc	Année
Réels	400	337	282	198	97	35	10	11	50	156	288	359	2223
Simulés	399	333	293	206	104	39	11	10	45	143	279	351	2213
Analyt.	400	334	293	205	105	38	11	10	45	143	278	364	2225
GDF	399	333	288	196	110	43	19	16	55	160	282	376	2277

FIG. II.9 – Tableau de résultats comparatifs pour les HDDs

II.2.3 Application à l'estimation des Quantiles et des Courbes en U

Dans ce paragraphe, nous allons appliquer le modèle précédent au calcul des courbes en U, dont nous rappelons la définition [2] :

Définitions : Quantiles et Courbes en U

DÉFINITION II.2.5

Soit un jour $j \in \{1 \dots 365\}$ et une probabilité $\alpha \in [0,1]$. On définit le quantile d'ordre α de la température au jour j comme la quantité $F_T(\alpha, j)$ telle que :

$$P(T_j < F_T(\alpha, j)) = \alpha$$

La courbe en U d'ordre α est la représentation graphique de la fonction

$$j \longrightarrow F_T(\alpha, j)$$

Il est possible d'estimer les courbes en U à partir d'échantillons de relevés de températures obtenus par simulation, en utilisant l'estimateur des quantiles ainsi que de les calculer analytiquement avec les matrices de transition. Nous les avons calculées pour $\alpha = 0.02, 0.1, 0.5, 0.9, 0.98$.

- Le tableau II.10 donne les quantiles calculés sur les minimaux annuels (on parle de *températures à risque*) pour différents modèles. [3]
- La figure II.11 fournit les résultats des courbes en U (analytiques, Monte-Carlo et réelles).
- La figure II.12 fournit les différences journalières entre les résultats réels et analytiques d'une part, et entre les résultats réels et Monte-Carlo d'autre part.

Modèle Utilisé | Valeur de α	**0.02**	**0.1**	**0.5**	**0.9**	**0.98**
Réels	-11.838	-9.845	-3.985	-0.595	0.603
GDF	-9.485	-6.920	-3.637	-1.326	-0.247
Autorégressifs	-10.129	-7.454	-3.977	-1.494	-0.326
Autorégressifs + résidus	-10.528	-7.761	-4.175	-1.623	-0.424
Markov	-11.500	-10.000	-4.000	-1.500	0.000

FIG. II.10 – Tableau de résultats comparatifs pour les quantiles sur les températures minimums annuelles ("à risque")

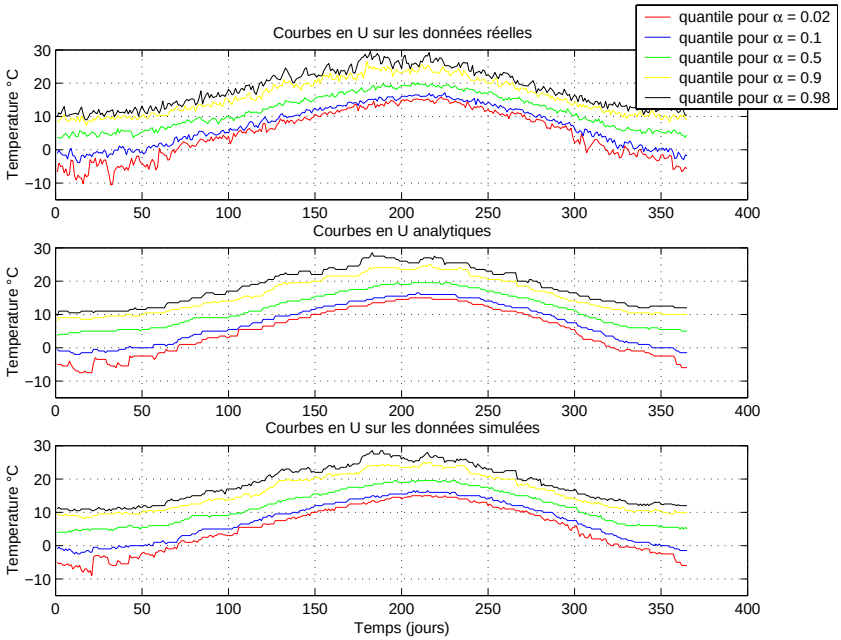


FIG. II.11 – Courbes en U de l'historique des températures obtenues par calculs analytiques et Monte-Carlo

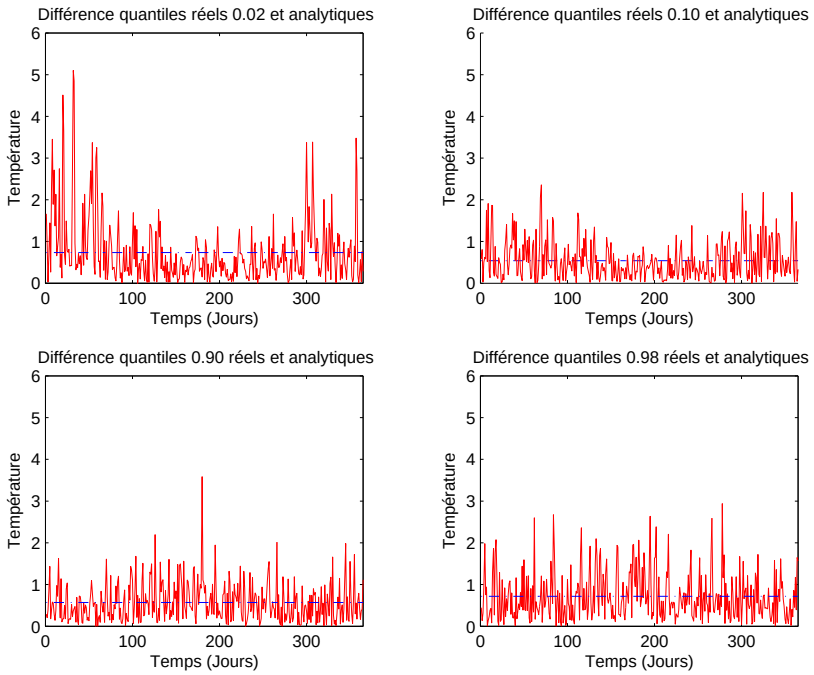


FIG. II.12 – Différences journalières courbes en U de l'historique et celles issues du modèle

Commentaires des résultats

REMARQUE II.2.6

Les résultats issus des courbes en U et du calcul des températures à risque doivent être interprétés en même temps : les courbes en U simulées ont une allure correcte, mais on a l'impression que le modèle "lisse" les températures, et que les températures les plus froides et les plus chaudes (températures extrêmes) ne sont pas correctment simulées. En fait, il faut faire plusieurs remarques sur les courbes en U avant de juger hâtivement le modèle :

- *Premièrement, les courbes en U issues de l'historique des températures (de 1950 à 2001) sont assez accidentées, ce qui laisse penser qu'une estimation des quantiles sur un historique plus grand aboutirait à des courbes beaucoup plus lisses. Ainsi, il suffit que par pure coïncidence, il ait fait 3 ou 4 fois -12 degrés le même jour de l'année sur un historique de 50 ans pour que les quantiles soient perturbés (comme on peut penser que cela soit le cas aux alentours du 35ème jour de l'année sur la figure II.11). Il est donc plutôt rassurant et logique que nos résultats simulés donnent des courbes assez lisses, étant donné que pour l'estimation Monte-Carlo, cela n'est pas sur 50 ans que les quantiles ont été calculés (comme pour l'historique), mais sur 10 000 ans !*
- *Si l'on regarde les différence au jour le jour entre courbes en U réelles et simulées (fig II.12), on s'aperçoit clairement que c'est pour le quantile le plus faible ($\alpha = 0.02$) que les différences sont les plus fortes. Notons encore que les écarts importants se situent au début de l'année, période à laquelle les températures sont les plus froides et surtout, les plus instables. Ainsi, il est normal, du fait de l'instabilité de l'historique sur cette période, que les simulations markoviennes soient moins précises.*
- *Alors que les deux remarques précédentes pourraient laisser à penser que le modèle markovien ne simule pas mieux les températures extrêmes que les autres types de modélisation, les résultats du tableau II.10 montrent au contraire qu'il simule très bien les températures à risque. Si l'on ne retrouve pas ces valeurs sur les courbes en U, cela vient du fait que le modèle visite bien les états extrêmes, mais pas exactement au jour de l'année observé sur les données réelles. En effet, les pics de températures froides sur les données réelles sont extrêmement localisées sur les courbes en U, ce qui engendre une forte volatilité des valeurs. Le fait de changer de matrice de transition chaque semaine représente déjà une trop grande période. Si l'on veut reproduire très exactement les courbes en U réelles avec notre modèle, il faut passer à un niveau de découpage des périodes encore plus faible : non plus en de semaines mais en terme de jours. Nous avons testé un modèle pourvu d'un tel découpage. Celui-ci colle effectivement bien aux courbes en U réelles, mais fait du sur-apprentissage. Or il n'est pas dans notre intérêt de produire un modèle qui simule très exactement les données observées, car il reproduirait également l'extrême volatilité de la variance, des quantiles, et proposerait des phénomènes beaucoup trop localisés : rappelons que nous avons pour objectif d'élaborer un système qui puisse fournir des sorties du même ordre que le système réel.*

II.2.4 Calcul Analytique de $P_{i_0}(X_t \leq T_{Seuil})$

Idée

On se propose de calculer la probabilité que la température à un jour t donné de l'année soit inférieure à une température seuil. Si on se donne un état initial (une *température* dans le cadre de notre modèle) i_0, cette probabilité s'exprime bien par :

$$P_{i_0}(X_t \leq T_{Seuil})$$

Du fait de la complémentarité des événements, on peut dès lors calculer facilement la quantité

$$P_{i_0}(X_t > T_{Seuil}) = 1 - P_{i_0}(X_t \leq T_{Seuil})$$

On peut ainsi évaluer analytiquement *dans le cadre de notre modèle* les risques d'être en-dessous ou au-dessus d'une température à un jour donné de l'année.

Fonction de répartition

Pour évaluer cette probabilité, nous allons devoir évaluer la fonction de répartition des températures pour chaque jour de l'année. En effet, si on note $F_{X_t}(T_{Seuil})$ la fonction de répartition au jour t de l'année, on peut écrire par définition :

$$F_{X_t}(T_{Seuil}) = P_{i_0}(X_t \leq T_{Seuil}),$$

si on suppose être parti initialement de l'état i_0. La figure II.13 illustre ce principe pour le jour $t = 32$, soit le 1^{er} février.

Ainsi, pour tout $t = 1 \ldots 365$, pour tout $T_{Seuil} \in E$, et pour tout $T_k \in E$ défini tel que $T_k \leq T_{Seuil}\ \forall\ k$, on a :

$$F_{X_t}(T_{Seuil}) = P_{i_0}(X_t <= T_{Seuil}) = P_{i_0}(X_t = T_1) + P_{i_0}(X_t = T_2) + \cdots + P_{i_0}(X_t = T_{Seuil})$$

Pour les chaînes de Markov, on peut finalement écrire

$$F_{X_t}(T_{Seuil}) = \sum_{T_k \leq T_{Seuil}} \mathbb{P}^t(i_0, T_k)$$

Rappelons que la chaîne $(X_n)_{n\in\mathbb{N}}$ est non-homogène ; il faudra donc changer de matrice de transition $\mathbb{P}$ toutes les semaines, donc à chaque puissance de 7, tout comme lors du calcul de l'espérance ou de la variance dans le grain II.2.1.

Le tableau II.14 propose quelques applications numériques de ce résultat.

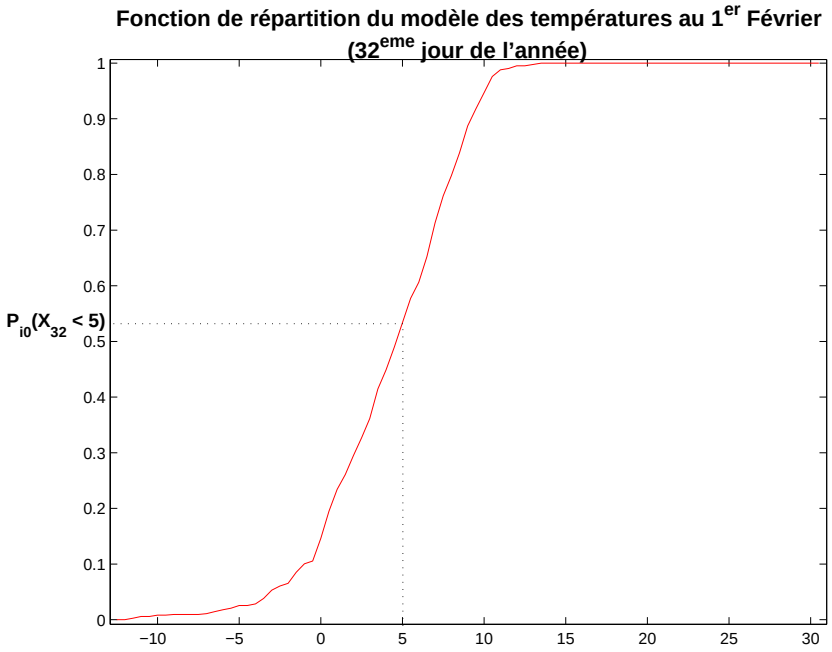


FIG. II.13 – Principe du calcul de $P_{i_0}(X_t \leq T_{Seuil})$ via les fonctions de répartition journalière.

Probabilité	Jour de l'année correspondant
$P_{i_0}(X_{10} \leq -5) = 0.02549$	10 janvier
$P_{i_0}(X_{32} \leq 5) = 0.53367$	1^{er} février
$P_{i_0}(X_{182} > 15) = 0.82582$	1^{er} juillet
$P_{i_0}(X_{182} > 25) = 0.05834$	1^{er} juillet
$P_{i_0}(X_{213} \leq 15) = 0.03243$	1^{er} Août
$P_{i_0}(X_{335} \leq 0) = 0.04919$	1^{er} décembre

FIG. II.14 – Quelques résultats numériques du calcul analytique de $P_{i0}(X_t \leq T_{Seuil})$

II.2.5 Test de l'homogénéité de la chaîne

Idée

On souhaite tester l'homogénéité de la chaîne de Markov au niveau hebdomadaire. La chaîne est par définition non-homogène au niveau annuel puisque l'on change de matrice de transition au cours du temps. En revanche, on la considère homogène par morceau sur chaque semaine de l'année. On se propose de le vérifier ici par un *test d'homogénéité*.

PRINCIPE II.2.7

On observe m processus sur un intervalle $[1;n]$, tel que $m \leq n$ et $m, n \in \mathbb{N}$. On veut tester l'homogénéité hebdomadaire, donc on prendra $n = 7$ puisque les semaines comportent 7 jours. On réalisera le test pour chacune des 52 semaines de l'année.
Posons

$$N_{ij}^m = \text{nombre de transitions de } i \rightsquigarrow j \text{ sur } [1;m]$$

$$N_i^m = \text{nombre de processus dans l'état } i \text{ sur } [1;m]$$

Le but du test est de savoir si

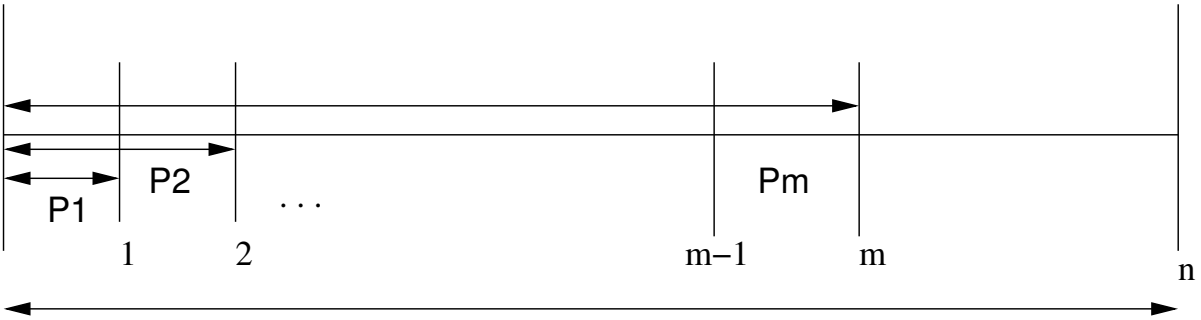
$$\widehat{\mathbb{P}}_{ij}^m = \widehat{\mathbb{P}}_{ij}, \text{ pour } m = 2, \ldots, n$$

pour une semaine considérée, où

$$\mathbb{P}_{ij}^m \text{ est la matrice de transition estimée sur } [1;m]$$

$$\mathbb{P}_{ij} \text{ est la matrice de transition estimée sur } [1;n]$$

Pour une semaine donnée, $\mathbb{P}_{ij}$ est donc la matrice de transition issue de l'estimateur du maximum de vraisemblance.



Une Semaine numérotée de chacune des années de l'ensemble d'apprentissage

FIG. II.15 – Principe du test d'homogénéité hebdomadaire

Le schéma II.15 présente comment s'opère le test : pour différentes valeurs de m (comprises entre 1 et 7), on va estimer une matrice de transition sur les m premiers jours d'une

semaine. Il s'agit de comparer ensuite le niveau d'égalité entre une matrice estimée sur les m premiers jours et celle estimée sur la semaine entière : c'est ce en quoi consiste le test d'homogénéité.

le Test

Le test est le suivant :

$$\begin{cases} H_0 : & \mathbb{P}_{ij}^m = \mathbb{P}_{ij} \\ H_1 : & \mathbb{P}_{ij}^m \neq \mathbb{P}_{ij} \end{cases} \text{, pour } m = 2, \ldots, n$$

Sous l'hypothèse H_0, on a :

$$-2log\Lambda \sim \chi^2_{(n-1)[m(m-1)]}$$

où on a posé Λ le rapport des vraisemblances

$$\Lambda = \frac{L(\mathbb{P}_{ij}^m)}{L(\mathbb{P}_{ij})}$$

et donc :

$$-2log\Lambda = 2\sum_{l=1}^{m}\sum_{i,j\in E} N_{ij}^l log\frac{N_{ij}^l}{N_i^{l-1}\mathbb{P}_{ij}}$$

On obtient une région critique de la forme :

$$-2log\Lambda \geq C$$

Pour accepter l'hypothèse d'homogénéité H_0 on doit vérifier :

$$-2log\Lambda \geq \chi^2_{(n-1)[m(m-1)],1-\alpha}$$

où α définit le niveau de confiance du test.

Résultats

Même en prenant un α très petit (jusqu'à 0.005), donc à un niveau de confiance de 99.5 %, le test d'homogénéité est vérifié, c'est-à-dire que l'hypothèse H_0 est acceptée. On peut donc bien considérer la chaîne homogène au niveau hebdomadaire.

II.2.6 Espérance du temps d'entrée

On se bornera à fournir des résultats analytiques applicables aux calculs numériques pour les chaînes de Markov non-homogènes.

Notations

Soit la chaîne $X = (X_n)_{n\in\mathbb{N}}$ avec un espace d'état E. Pour une chaîne de Markov non-homogène, on a la propriété suivante, déjà évoquée :

THÉORÈME II.2.8

$\forall\, j \in E,\ n \geq 0,$

$$P(X_{n+1} = j/X_k, k \leq n) = P(X_{n+1}/X_n)$$

On note, pour $m < n$

$$p_{m,n}(i,j) = P(X_n = j/X_m = i)$$

$$p_n = p_{n,n+1}$$

Un premier résultat : Chapman-Kolmogorov

Nous proposons une propriété essentielle des fonctions de transition des chaînes de Markov, bien connue dans le cas homogène et dont nous donnons l'équivalent dans le cas non-homogène : il s'agit de l'équation de **Chapman-Kolmogorov**.

THÉORÈME II.2.9

$\forall i,j \in E, \forall r$ *tel que* $m < r < n$

$$p_{m,n}(i,j) = \sum_k p_{m,r}(i,k)p_{r,n}(k,j)$$

Démonstration : page 44, Annexe A.1

Calcul de l'espérance du temps d'entrée dans un état

Ce résultat est particulièrement intéressant : il va nous permettre de déterminer analytiquement au bout de combien de temps le système entre, en moyenne, dans un état donné. Donnons pour commencer la définition d'un **temps d'entrée** :

DÉFINITION II.2.10

Soit $i \in E$. La variable aléatoire $T_i = inf\{n \in \mathbb{N}^*, X_n = i\}$ est appelé temps d'entrée dans i. Si $P(X_0 = i) = 1$, alors T_i est appelé Temps de récurrence.

Pour calculer l'espérance du temps d'entrée, nous partirons du théorême suivant :

THÉORÈME II.2.11

[10]

Soit T le temps d'entrée dans un sous-ensemble A de E. On note $B = E\backslash A$. On a :

$$\mathbb{E}_i^n[T] = 1 + \sum_{j\in B} p_n(i,j)\mathbb{E}_j^{n+1}[T]$$

Démonstration : page 44, Annexe A.2

Si on note maintenant $p^B_{0,k}$ la restriction sur B de la fonction de transition $p_{0,k}$, on a le résultat matriciel :

$$\mathbb{E}_i[T] = (I + \sum_{k=0}^{\infty} p^B_{0,k}).e_i$$

e_i est un vecteur de taille égale au nombre d'états dans la chaîne, nul partout sauf en i, tel que $e_i(i) = 1$.

De plus, il vient facilement que

$$p^B_{0,k} = \prod_{r=0}^{k} p^B_r$$

De là vient le résultat :

THÉORÈME II.2.12
[8]

$$\mathbb{E}_i[T] = (I + \sum_{k=0}^{\infty} \prod_{r=0}^{k} p^B_r).e_i$$

Démonstration : page 46, Annexe A.14

Calcul de l'espérance du temps de passage

N^n_j est la durée totale passée dans l'état j sur l'intervalle de temps $[1; n]$. On a

$$N^n_j = \sum_{k=1}^{n} \mathbf{1}_{\{X_k=j\}}$$

On se propose de calculer $\mathbb{E}^t_i[N^n_j]$, c'est à dire le temps moyen passé dans l'état j sur l'intervalle de temps $[1; n]$.

On a facilement

$$\mathbb{E}_i[N^n_j] = \sum_{k=1}^{n} P_i(X_k = j)$$

De là il vient

$$\mathbb{E}_i[N^n_j] = (\sum_{k=1}^{n} \prod_{r=0}^{k} p^B_r)(i, j)$$

Conclusion

L'utilisation des outils markoviens pour la simulation des températures s'avère satisfaisante : nous avons mis en évidence l'intérêt d'avoir recours à l'utilisation des chaînes de Markov non-homogènes au dépend de la profondeur, tout au moins dans ce cadre de développement. Ce choix permet d'implémenter un simulateur assez léger, demandant un temps de calcul et un espace mémoire raisonables. L'utilisation de `Matlab` et de `R` permet de surcroît un programme compact, facilement exportable et modifiable. Les résulats obtenus avec le modèle non-homogène d'ordre 1 donnent de bons résultats sur les statistiques que l'industriel Gaz de France nous avait demandé d'évaluer pour juger le modèle :

- Les trajectoires de simulation sont semblables à des relevés réels de températures
- La convergence du modèle (l'espérance analytique) est en accord avec les résultats réels et permet en plus d'obtenir une trajectoire plus lisse que la trajectoire moyenne réelle, accidentée du fait de la grande instabilité des données.
- La variance (sous forme d'écart-type ici) est correcte, mais on pourrait souhaiter la lisser davantage : nous pensons en effet que la volatilité des données températures n'est pas à reproduire, car due à un manque de données.
- Les résultats sur les Degrés Jours Cumulés sont bons puisqu'ils complètent très bien les résultats obtenus par Gaz de France avec d'autres modèles.
- On peut être satisfait de la manière dont le modèle simule les températures extrêmes par rapport aux modèles AR ou à persistance. Les résultats obtenus sont meilleurs (tableau II.12) pour la simulation des températures extrêment froides que ceux obtenus avec les modèles utilisés jusqu'ici. Par ailleurs, nous pensons qu'il n'est pas possible de reproduire l'occurence de ces valeurs extrêmes (comme en dénotent les courbes en U). Ce n'est d'ailleurs pas l'objectif de la modélisation probabiliste, et l'on touche ici à la limite prévision de températures / simulation de données températures.

Une autre preuve de l'intérêt du recours au markovien est la satisfaction de l'industriel, puisque Gaz de France va utiliser le programme élaboré dans le cadre de la réécriture de son logiciel *Prosper* destiné à l'évaluation des risques sur les températures extrêmes.

Pour conclure, nous faisons ici quelques remarques sur les évolutions que nous avons testées mais pas détaillées dans ce mémoire, qui permettent d'évaluer les limites de l'utilisation des outils markoviens pour ce type de données et dans ce cadre de modélisation :

- **Augmenter la profondeur de la chaîne -** Le lecteur pourrait s'interroger sur le fait que les modèles markoviens que nous avons utilisés ne vont pas au-delà de l'ordre 1, alors que les phénomènes de températures sont physiquement dépendants de plus d'un jour sur l'autre (les autres modèles utilisés par Gaz de France comme le modèle à persistance reposent justement sur ce type de relation). Dans le cas de la modélisation markovienne des températures, il s'avère que le passage au non-homogène garantit cette relation de

profondeur, par le fait que la mutiplicités des matrices de transition assure de ne pas passer d'une température à une autre sans que cela soit abérant. Par ailleurs l'ordre 1 permet des calculs numériques en terme de puissance matricielle qui restent raisonables ; augmenter l'ordre de la chaîne rendait même certains calculs impossibles ou tout du moins problèmatiques : pour un ordre 3, il faut calculer des puissances $n^{\text{ième}}$ de matrices de dimension 4 ! Dans ce cas les statistiques d'un modèle d'ordre supérieur à 1 s'obtenaient uniquement par simulation Monte-Carlo, ce qui n'est pas satisfaisant.

- **Apprendre sur des données Bootstrappées -** Nous avons également envisagé d'estimer les matrices de transition sur des données Bootstrappées : il s'agit de générer des données avec les matrices de transition issues de l'ensemble d'apprentissage initial, puis de réapprendre et de réestimer ces matrices sur l'ensemble données réelles/données simulées ; a priori il semble intéressant de disposer du plus de données possible pour affiner les estimateurs. Malheureusement dans ce cas le modèle issu des données bootstrappées avait tendance à tasser les lois de probabilités : on simulait mal les queues de distributions, c'est à dire les valeurs extrêmes, qui sont justement l'enjeu principal de la modélisation des températures.
- **Rééquilibrer les données avant l'apprentissage -** Comme nous avons pu nous en rendre compte avec le tout premier modèle markovien homogène (grain I.2.4, les données températures ne sont pas stables et toutes les études de Gaz de France dénotent d'un réchauffement climatique et d'une rupture depuis 1987, ce qui fausse quelque peu l'efficacité de l'estimation des matrices de transition. Une étude Gaz de France a donc proposé un redressement des données températures [7]. Nous les avons utilisées dans l'estimation des matrices de transition : comme on pouvait s'y attendre, l'effet est le même qu'avec les données bootstrappées, ce qui n'est pas acceptable dans le cadre de cette étude.

Annexe A
Documents associés

Annexe A.1 Démonstration du théorème II.2.9

> **Retour Grain** :
> Théorème II.2.9

On a par définition $\forall m < n$

$$p_{m,n}(i,j) = P(X_n = j / X_m = i)$$

Puis, comme $P(\cup_n A_n) = \sum_n P(A_n)$, il vient

$$= \sum_k P(X_n = j, X_r = k / X_m = i)$$

En appliquant la propriété $P(A \cap B/C) = P(A/B,C).P(B/C)$, on trouve

$$p_{m,n}(i,j) = \sum_k P(X_n = j / X_r = k, X_m = i).P(X_r = k / X_m = i)$$

$$= \sum_k P(X_n = j / X_r = k).P(X_r = k / X_m = i)$$

et finalement,

$$p_{m,n}(i,j) = \sum_k p_{m,r}(i,k) p_{r,n}(k,j)$$

■

Annexe A.2 Démonstration du théorème II.2.11

> **Retour Grain** :
> Théorème II.2.11

On part de la relation suivante (évidente)

$$1_{\{X_n \in E\}} = 1_{\{X_n \in A\}} + 1_{\{X_n \in B\}}$$

d'où

$$T = T.1_{\{X_1 \in E\}} = T.(1_{\{X_n \in A\}} + 1_{\{X_n \in B\}})$$

et

$$T = \sum_{j \in A} T.1_{\{X_n = j\}} + \sum_{j \in B} T.1_{\{X_n = j\}}$$

Soit pour l'espérance

$$\mathbb{E}_i^n[T] = \sum_{j \in A} \mathbb{E}_i^n[T.1_{\{X_n = j\}}] + \sum_{j \in B} \mathbb{E}_i^n[T.1_{\{X_n = j\}}]$$

– pour $j \in A$

$$\mathbb{E}_i^n[T.1_{\{X_n = j\}}] = \mathbb{E}_i^n[T / X_n = j].P_i(X_n = j) = P_i(X_n = j) = p_n(i,j)$$

– pour $j \in B$

$$\mathbb{E}_i^n[T.1_{\{X_n=j\}}] = \mathbb{E}_i^n[T/X_n = j].P_i(X_1 = j)$$

On opère ensuite une translation de l'état intial i vers j :

$$\mathbb{E}_j^n[T+1].\mathbb{P}(i,j) = 1 + \mathbb{E}_j^{n+1}.p_n(i,j)$$

Ainsi, on retrouve

$$\mathbb{E}_i^n[T] = \sum_{j\in A} p_n(i,j) + \sum_{j\in B}(1 + \mathbb{E}_j^{n+1}[T])p_n(i,j)$$

puis

$$\mathbb{E}_i^n[T] = \sum_{j\in A\cup B} \mathbb{P}(i,j) + \sum_{j\in B} \mathbb{E}_j^{n+1}[T]p_n(i,j)$$

et

$$\boldsymbol{\mathbb{E}_i^n[T] = 1 + \sum_{j\in B} \mathbb{E}_j^{n+1}[T]p_n(i,j)}$$

■

Annexe A.3 Bienaymé-Čebyčev

Retour Grain :
Monte-Carlo

Ce document propose une utilisation de l'inégalité de Bienaymé-Cebycev dans le cadre de simulations par Monte-Carlo : cette proposition permet de déterminer le nombre de simulations à effectuer pour obtenir un résultats par Monte-Carlo avec une marge d'erreur que l'on impose. Commençons donc par rappeller l'inégalité.

THÉORÈME A.13
Soit X une variable aléatoire réelle absolument continue admettant un moment d'ordre 2. Alors pour tout $\varepsilon > 0$, on a :

$$P(|X - \mathbb{E}[X]| > \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$$

Admettons que l'on veuille estimer pour notre modèle la probabilité que la température le 31 janvier soit inférieur à 5 degrés, soit formellement $P(X_31 < 5)$. Comme nous l'avons vu dans la section rappelant le principe de la méthode de Monte-Carlo, on utilise une variable aléatoire $S_N = \sum_1^N Y_k$ où les Y_k suivent une loi de Bernouilli. L'estimation par Monte-Carlo de $P(X_31 < 5)$ est donnée par $\frac{S_N}{N}$.

Si on applique l'inégalité A.13 à cette variable aléatoire, on obtient :

$$P(|\frac{S_N}{N} - \mathbb{E}[X]| > \varepsilon) \leq \frac{Var(\frac{S_N}{N})}{\varepsilon^2}$$

Pour la variance, vu que les variables sont *i.i.d.*, on a

$$Var(\frac{S_N}{N}) = \frac{1}{N^2} \times Var(S_N) = \frac{1}{N^2} \times N \times Var(Y_k) = \frac{Var(Y_1)}{N}$$

d'où

$$P(|\frac{S_N}{N} - \mathbb{E}[X]| > \varepsilon) \leq \frac{Var(Y_1)}{N\varepsilon^2}$$

Si l'on veut un nombre d'itérations N tel que :

$$P(|\frac{S_N}{N} - P(X_3 1 < 5)| > \varepsilon) \leq \alpha$$

on doit prendre N tel que :

$$\frac{Var(Y_1)}{N\varepsilon^2} \leq \alpha \Leftrightarrow N \geq \frac{Var(Y_1)}{\alpha\varepsilon^2}$$

Il suffit enfin de se rappeler que les Y_k suivent une loi de Bernouilli, de variance égale à $p(1-p)$. Or on vérifie aisément que la fonction $p(1-p)$ est majorée par $\frac{1}{4}$ quand $p = \frac{1}{2}$

Ainsi, en prenant $\varepsilon = 10^{-1}$ et une précision $\alpha = 10^{-}2$ (1%), on trouve un N raisonnable en majorant la variance par $\frac{1}{4}$. L'application numérique donne :

$$N = \frac{Var(Y_1)}{\alpha\varepsilon^2} = \frac{1}{4} \times \frac{1}{10^{-1}(10^{-2})^2}$$

Soit $\boldsymbol{N = 25000}$ années de simulations, d'où l'intérêt de trouver un résultat analytique !

Annexe A.4 Convergence stochastique presque sûre

> **Retour Grain** :
> Espérance et variance analytique

On rappelle la définition de la convergence presque sûre :

DÉFINITION A.14
On dit que la suite de variable aléatoire $(X_n)_{n\geq 1}$ converge presque sûrement vers X et on note $X_n \longrightarrow X$, p.s. quand $n \to \infty$ si :

$$P\Big(\omega \in \Omega : lim_{n\to\infty} X_n(\omega) = X(\omega)\Big) = 1$$

Annexe B
Implémentation logicielle du modèle

Annexe B.1 Présentation du programme

Environnement de programmation

L'implémentation informatique des différents modèles markoviens testés, et en particulier le modèle à chaîne de Markov discrète non-homogène d'ordre 1 développé dans le chapitre 2, a été réalisé dans un premier temps avec le logiciel `Matlab`. Celui-ci est tout particulièrement adapté à la manipulation de vecteurs et de matrices, favorisant la concision et l'aspect conceptuel et mathématique de la modélisation et facilitant les sorties graphiques.

Pour les besoins de Gaz de France, et à la vue des résultats obtenus avec le modèle markovien, le modèle finalement choisi a également été implémenté en langage `R`, logiciel libre permettant un développement dans le même esprit que `Matlab` et possédant une bibliothèque de calculs statistiques avancée. Le programme `R` a été inséré au logiciel de simulation *Prosper* utilisé par Gaz de France.

Organisation sommaire

Bien que des modèles à base de chaînes de Markov d'ordres supérieurs aient été développés, nous présentons dans le grain suivant uniquement la structure du programme `R` implémentant le modèle décrit au chapitre 2 puisqu'il est celui obtenant les meilleurs résultats tout en concervant la possibilité de faire des calculs analytiques et utilisant une place de stockage mémoire raisonable.

Le programme réalisé s'organise en 3 temps, auxquels correspondent 4 scripts, chacun faisant appel à ses propres fonctions :

- une phase de préparation des données, les formatant de manière à être plus facilement utilisables pour la manipulation vectorielle, et de création des objets du modèle (en particulier l'estimation des matrices de transition hebdomadaires à partir de l'ensemble d'apprentissage). Ces fonctionnalités sont réalisées par le script `Initialisation.R`.
- une phase de simulation, utilisant un moteur de simulation pour les chaînes de Markov d'ordre 1 et les variables formatées dans le script `Initialisation.R` : le script `simpleSimu.R` réalise simplement une simulation type sur un nombre d'années définies ; le script `monteCarlo.R` lance un grand nombre de simulations en vue de l'estimation des paramètres de sortie du modèle.
- une phase utilisant les variables du script d'initialisation et les résultats des simulations Monte-Carlo pour réaliser des calculs analytiques et les estimations de statistiques permettant de juger de la qualité du modèle. Cette fonctionnalité est remplie par le script `calculStat.R`

Le grain suivant propose les organigrammes ainsi qu'une brève description des fonctions utilisées par ces différents scripts.

Annexe B.2 Organigrammes du programme

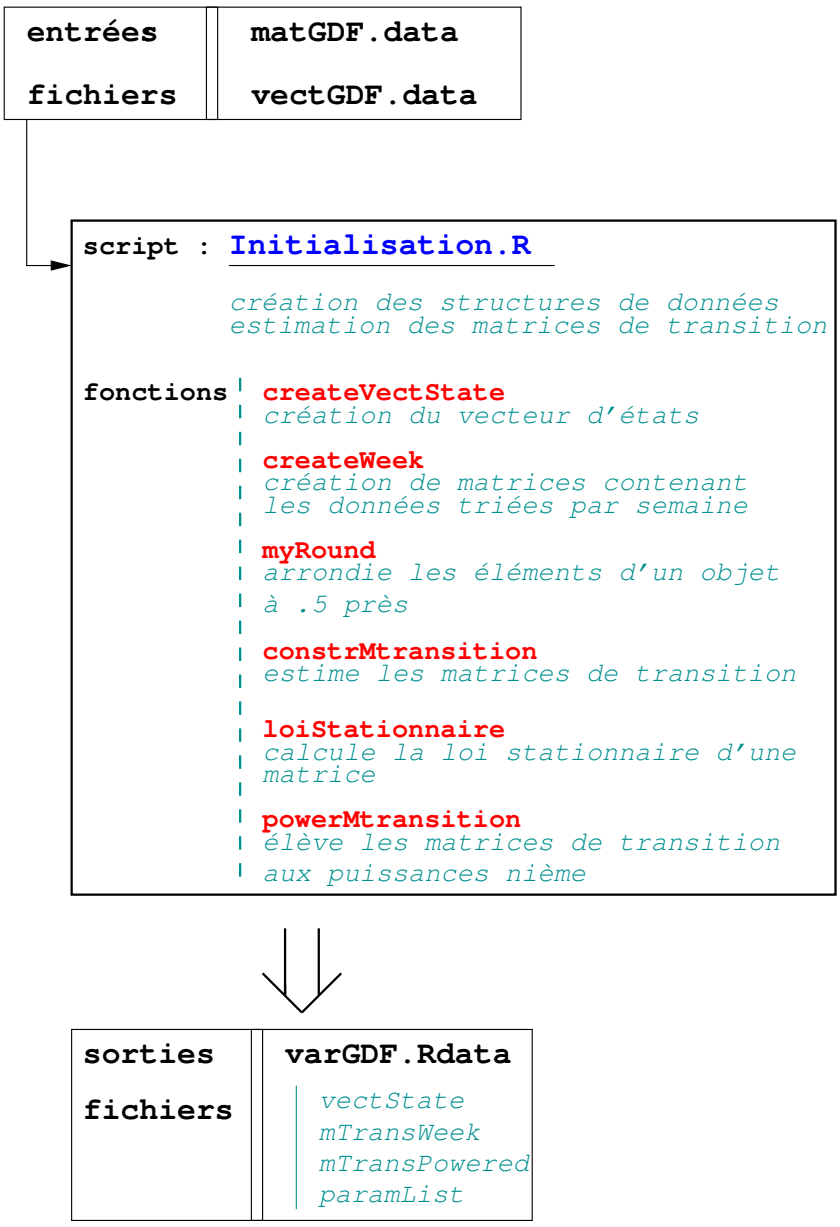


FIG. B.1 – Organigramme du script d'initialisation

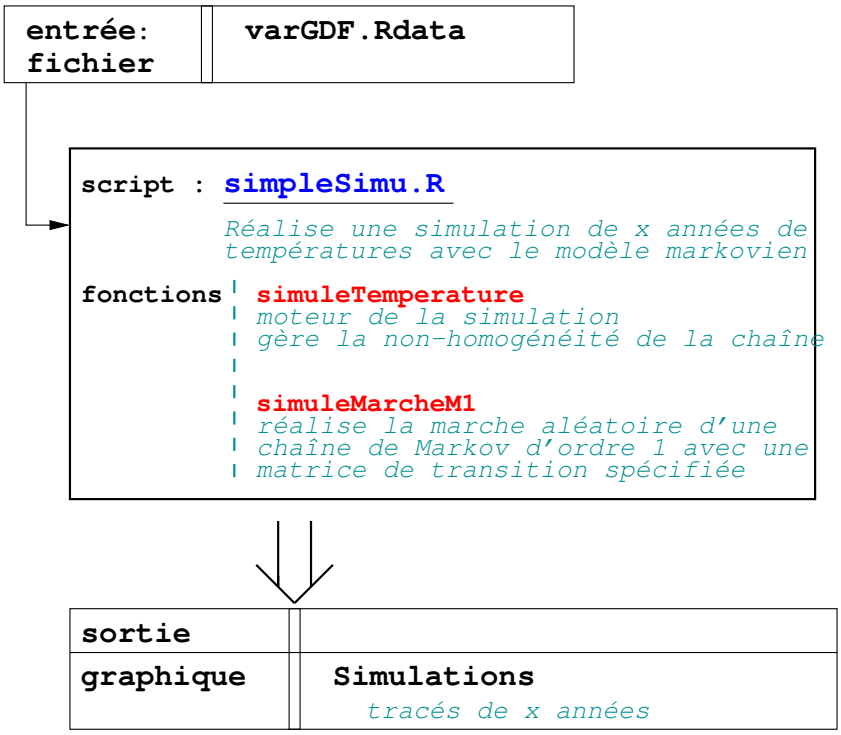


FIG. B.2 – Organigramme du script de simple simulation

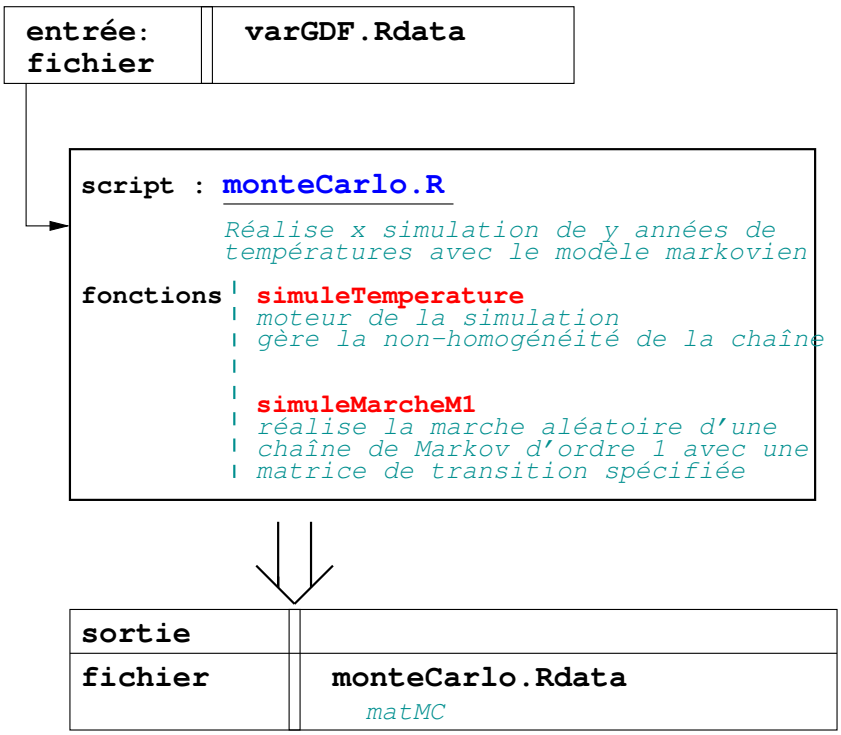


FIG. B.3 – Organigramme du script de simulation Monte-Carlo

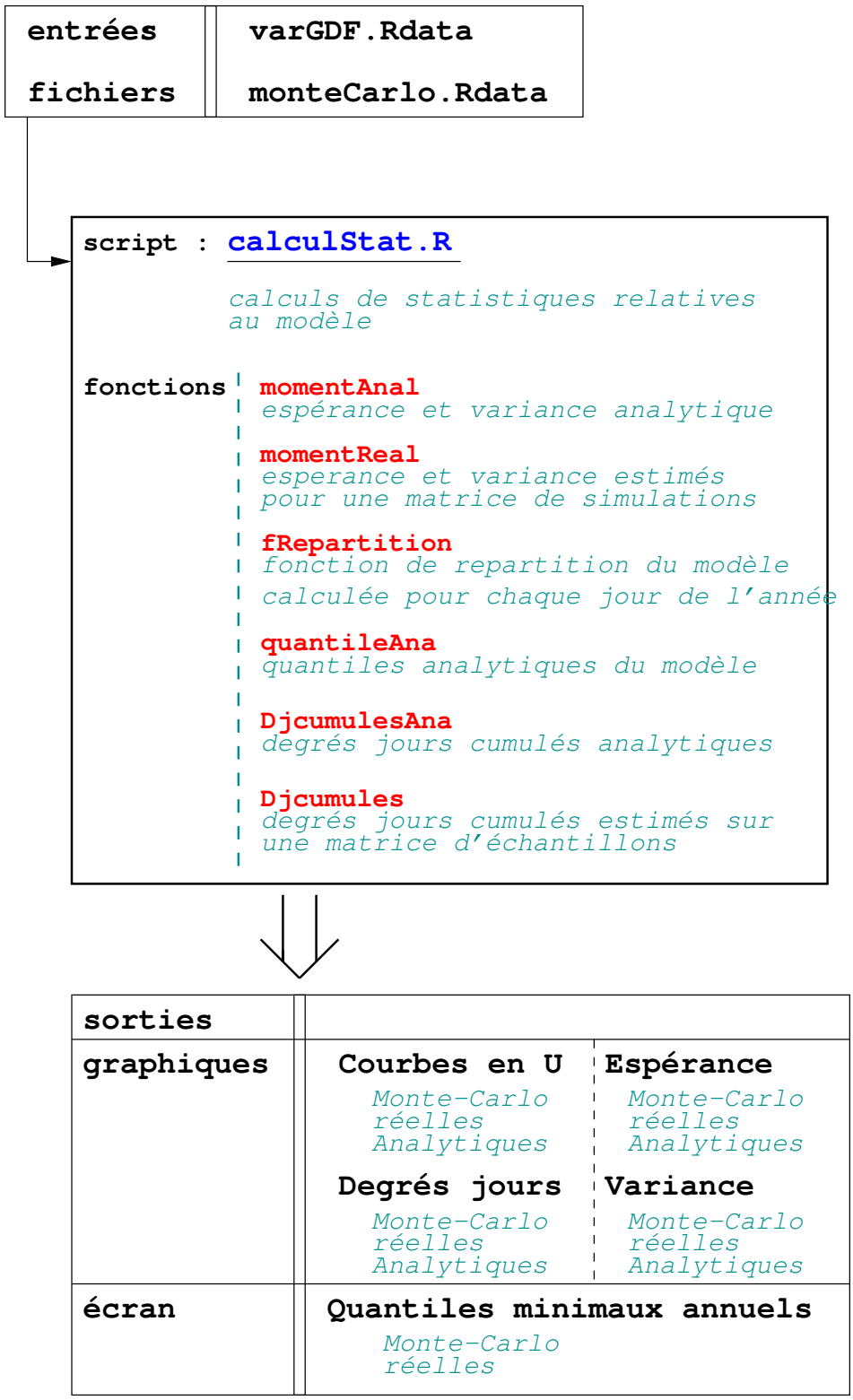


FIG. B.4 – Organigramme du script de calcul des statistiques

Index des concepts

Bibliographie

[1] P. Alaton. On modelling and pricing weather derivatives. pages 1–5, Stockholm, suède, 2000.

[2] G. Benmenzer. Prosper et le processus à persistance. pages 1–17, GDF Saint Denis, France, 2003.

[3] G. Benmenzer. Synthèse des résultats pour prosper du lundi 10 mars 2003. pages 1–4, GDF Saint Denis, France, 2003.

[4] P. Billingsley. Statistical methods in markov chains. In *Stanford meetings of the Institute of Mathematical Statistics*, Chicago, USA, 1960.

[5] I. Chèze and S. Jourdain. Calcul des quantiles de durées de retour de la température par la méthode gev. In *Calcul des températures à risque*, pages 1–50, Météo France DP/SERV/BEC Toulouse, France, 2003.

[6] Direction de la production et du transport de Gaz de France Service Etude. Etude statistique des températures journalières relevées à la station de paris-montsouris. In *Document M43*, pages 6–11, GDF Saint-Denis, France, 1999.

[7] A. Lenormand. Méthodes de sélection et d'estimation d'un changement climatique dans une chronique de température. In *Projet PREMICC*, pages 1–30, GDF Saint Denis, France, 2003.

[8] N. Limnios. Processus stochastiques et fiabilité. In *Cours de DEA - Module TI08*, pages 1–12, Laboratoire de Mathématiques Appliquées de Compiègne - UTC, France, 2003.

[9] N. Limnios and A. Sadek. Asymptotic properties for maximum likehood estimators for reliability and failure rates of markov chains. In *Statistics - Theory and Methods 31*, pages 1837–1861, Laboratoire de Mathématiques Appliquées de Compiègne - UTC, France, 2002.

[10] A. Platis, M. Le Du, and N. Limnios. hitting time in a finite non-homogeneous markov chains with applications. In *Applied Stochastic Models and Data Analysis 14*, pages 241–253, Laboratoire de Mathématiques Appliquées de Compiègne - UTC, France, 1998.

[11] J. Portes. Modélisation du climat et réchauffement climatique. pages 4–10, GDF Saint-Denis, France, 2002.

[12] Girardin V. and N. Limnios. Chapitre 6 : chaînes de markov. In *Probabilités*, pages 249–300, 2002.