République Algérienne Démocratique et Populaire

الجمهورية الديمقراطية الشعبية الجزائرية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

وزارة التعليم العالي والبحث العلمي

Université Saad Dahlab, Blida 1
USDB.

Faculté des sciences.
Département mathématiques

Mémoire de Fin d'Etudes

En vue de l'obtention du diplôme de MASTER

En : mathématiques

Option : Modélisation stochastique et statistique

Par : Aouicha Asma

Laib Rahma

THEME

Tests des ruptures des séries chronologiques.

Application sur des données météorologiques.

Température à Soumaa W. Blida

Soutenu le 10 /07/2019, devant le jury composé de :

Mr.	O.TAMI	MAA	**USDB**	**Président**
Mr.	A. RASSOUL	MCA	**ENSH**	**Promoteur**
Mr.	R.FRIHI	MAA	**USDB**	**Examinateur**

DÉDICACE

Je dédie ce travail, comme preuve de respect, de gratitude, et de reconnaissance :

A mes chers parents, Je vous remercie pour tout le soutien et l'amour que vous me portez depuis mon enfance et j'espère que votre bénédiction m'accompagne toujours. Que ce modeste travail soit l'exacement de vos voeux tant formulés, le fruit de vos innombrables sacrifices, bien que je ne m'en acquitterai jamais.

À mon mari CHRAF EDDIN et ma petite NOUR EL HOUDA.

À mon cher frère ZAKARIA, et ma chère sœur MARIEM EL BATOUL, KHADIDJA, TKWA et KAOULA.

Merci pour m'avoir toujours supporté dans mes décisions. Merci pour tout votre amour et votre confiance.

LAIB RAHMA

DÉDICACE

A MA CHERE MAMAN

Honorable aimable tu représente pour moi le symbole de la bonté par excellence, la source de la tendresse et l'exemple du dévouement qui n'a pas cessé de m'accompagner

Je te dédie ce travail en témoignage de mon profond amour

A MON CHER PAPA

Autant de phrases et d'expressions ne sauraient exprimer ma gratitude et ma reconnaissance, tu as su graver en moi le sens de la responsabilité de l'optimisme et de la confiance en moi

Je te dois ce que je suis aujourd'hui et ce que je serai demain.

AOUICHA ASMA

REMERCIEMENT

Tout d'abord,nous remercions **ALLAH** tout puissant tout nous avoir donné la volontée et le courage d'accomplir ce modeste travail.

En second lieu, nous tenons à remercier notre encadreur **Dr. : A. RASSOUL**, son précieux conseils et son aide durant toute la période du travail.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Enfin, nous tiendrons à remercier tous ceux qui, de prés ou de loin, nous ont soutenus et aidé à réaliser et qui ont contribués, de prés ou de loin, au bon déroulement et à réalisation de ce travail.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉS

ملخص

يركز هذا العمل على اختبارات الكسر للسلاسل الزمنية. نقدم اولا السلاسل الزمنية ونماذجها (العمليات العشوائية,الثابتة و الغير ثابتة) ووظائف الارتباط التلقائي و الربط التلقائي الجزئي.

ثم نقدم اختبارات الكسر الحدودي والغير حدودي.

تنهي هذه المذكره بامثلة من محاكاه نماذج سلسلة زمنية مع برنامج R وتطبيقها على بيانات معطاه على مستوى منطقة الصومعة.

0.1 Résumé

Le présent travail porte sur les tests de rupture sur les séries chronologiques et ses applications sur les séries de temperature ; nous présentons dans un premier temps les séries tomporelles et ses modèles (processus aléatoire, stationnaire et non stationnaire) et les deux fonctions d'autocorélation et autocorélation partielle,en suite nous présentons les tests de rupture paramétriques et non paramétriques. Nous terminons ce mémoire par des exemples de simulation des modèles d'une série chronologique avec le logiciel R, et une application sur des données de température aux niveau de la région de SOUMAA tel que nous testons sa tendance et sa saisonalité à l'aide de logiciel R.

0.2 Abstract

The present work deals with time series repetitive tests and its applications on temperature series ; we first present the tomporal series and its models (random, stationary and nonstationary processes) and the two autocorrelation and partial autocorrelation functions, then we present the parametric and non-parametric rupture tests. We end this paper with examples of simulation of the models of a time series with the software R, and an application on temperate data at the levels of the region of SOUMAA as we test its tendency and its seasonality with the help of R. software

INTRODUCTION GÉNÉRALE

Ce mémoire présente l'une des applications très important tel que nous appliqués les teste de rupture sur les séries chronologique, ou séries temporelles : est une suite des valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. De telles suites de variables aléatoires peuvent être exprimées mathématiquement afin d'en analyser le comportement, généralement pour comprendre son évolution passée et pour en prévoir le comportement futur. Une telle transposition mathématique utilise le plus souvent des concepts de probabilités et de statistique. L'analyse des séries chronologiques est un outil couramment utilisé de nos jours pour la prédiction de données futures. Ce domaine possède beaucoup d'applications en finance, en médecine, en économétrie et en météorologie et dans bien d'autres domaines. Sous travaille diviser par trois chapitres :

— Le premier chapitre présente les séries chronologiques. Cette dernière est une suite formée d'observations au cours du temps, est appliquée de nos jours dans des domaines par exemple en astronomie (1906) et en météorologie (1968)...ect. L'objet des séries temporelles est l'étude des variables au cours du temps, comprendre le passé (analyser et expliquer les valeurs observées), bâtir des prévisions pour les valeurs non encore observées et étudier le lien avec d'autres séries chronologiques. Nous touchons aussi les modèles d'une série chronologique est partager par des processus aléatoire stationnaire(Processus autorégressif AR, Processus à moyenne mobile MA et les processus mixtes ARMA), processus aléatoire non stationnaire(processus ARIMA et SARIMA) et processus stockastique non linéaire(ARCH et GARCH) et on définit l'autocorrélation et l'autocorrélation partielle comme des fonction qui mesure le lien entre les valeurs du processus à deux dates distantes.

— Nous avons consacré le deuxième chapitre pour exposer deux types de tests de rupture paramétriques et non paramétriques. Ceux-ci seront comparés aux tests présentés à l'aide de simulations de plusieurs types de rupture qui tiendront compte de plusieurs facteurs. Et nous présenterons quelques méthodes de détection de rupture souvent utilisées dans les séries chronologiques. Ces méthodes permettent principalement d'effectuer l'identification et la modélisation des changements des

paramètres statistiques et stochastiques liés à la série chronologique. Ces paramètres statistiques sont éventuellement caractérisés dans notre cas par le changement des moyennes et des variances d'une série temporelle.

— Au dernier chapitre, on effectué une étude pratique et une simulation des modèles d'un série chronologique ,tel que on traite la tendance et la saisonnalité de la température de l'atmosphère au niveau de la région de SOUMAA depuis 2002 jusqu'à 2015. Un traitement à l'aide du logiciel R est exécuter sur une série de plus de 4000 valeurs de températures relevées au pare avant, afin d'extraire l'irrégularité et la rupture au cours de ces années.

CHAPITRE 1

LES SÉRIES CHRONOLOGIQUES

1.1 Introduction

Une série chronologique est une suite formée d'observations au cours du temps. L'analyse des séries chronologiques est un outil couramment utilisé de nos jours pour la prédiction des données futures. Ce domaine possède beaucoup d'applications en finance, en médecine, en économétrie, en météorologie et dans bien d'autres domaines. Par exemple en finance : on s'intéresse à modéliser le taux de change d'une devise.

L'idée est de prendre un échantillon de données et de construire le meilleur modèle qui ajuste ces données. Ce modèle nous permet de tirer certaines conclusions sur la série. Par exemple, on peut établir une formule pour la prédiction de données, détecter certains pics ou modéliser la tendance (orientation) de la série. Un autre aspect important de la série est la composante saisonnière, c'est-à-dire la présence de cycles. Un autre concept intéressant serait le phénomène de causalité, c'est-à-dire l'influence d'une série sur une autre. Cependant il n'est pas toujours évident de choisir le bon modèle, car normalement plusieurs peuvent être de bons candidats.

La plupart de ceux-ci auront tendance à minimiser la variance des résidus. Il faut aussi considérer le nombre de paramètres à estimer. Normalement on utilise le principe de parcimonie qui nous propose de choisir le modèle demandant le moins de paramètre avec une variance faible des résidus. Il y a donc plusieurs critères à considérer dans la création d'un modèle.

1.2 Présentation d'une série chronologique

1.2.1 Définitions

Définition 1.1 *On appelle série chronologique, ou bien encore chronique ou série temporelle, une suite finie de données quantitatives indexée par le temps. L'indice temps peut être selon les*

cas, la seconde, la minute, l'heure, le jour, le mois, l'année..

Définition 1.2 *La théorie des séries chronologiques (ou temporelles) est appliquée de nos jours dans des domaines aussi variés que l'économétrie, la médecine ou la démographie, pour n'en citer qu'une petite partie.*

On s'intéresse à l'évolution au cours du temps d'un phénoméne, dans le but de décrire, expliquer puis prévoir ce phénomène dans le futur.

On dispose ainsi d'observations à des dates différentes,c'est à dire d'une suite de valeurs numériques indicées par le temps.

Exemple 1.1 *On peut songer par exemple à l'évolution du nombre de voyageurs utilisant le train, à l'accroissement relatif mensuel de l'indice des prix ou encore à l'occurence d'un phénomène naturel (comme le nombre de taches solaires). Cette suite d'observations d'une famille de variables aléatoires réelles notées* $(x_t)_{t\in\Theta}$ *est appelée série chronologique (ou temporelle). nous la noterons*

$$(x_t)_{t\in\Theta} \quad ou \qquad (x_t, t\in\Theta) \tag{1.1}$$

où l'ensemble est appelé ***espace des temps*** *qui peut être :*

— ***Discret*** *(nombre de voyageurs SNCF quotidien, température maximale...). Dans ce cas,* $\Theta \subset Z$*. Les dates d'observations sont le plus souvent équidistantes : par exemple relevés mensuels, trimestriels... Ces dates équidistantes sont alors indexées par des entiers :* $t = 1, 2, ..., T$ *et* T *est le nombre d'observations. On dispose donc des observations des variables* $X_1, X_2, ..., X_T$ *issues de la famille* $(X_t)_{t\in\Theta}$ *où* $\Theta \subset \mathbb{Z}$ *(le plus souvent =* $\mathbb{Z}$*). Ainsi si* h *est l'intervalle de temps séparant deux observations et* t_0 *l'instant de la première observation, on a le schéma suivant :*

$$\begin{array}{ccc} t_0 & t_0+h... & t_0+(T-1)h \\ \downarrow & \downarrow & \downarrow \\ X_{t_0} & X_{t_0+h}... & X_{t_0+(T-1)h} \\ \downarrow & \downarrow & \downarrow \\ X_1 & X_2... & X_T \end{array}$$

— **Continu** (signal radio, résultat d'un électrochardiogramme...). L'indice de temps est à valeurs dans un intervalle de $\mathbb{R}$ et on dispose (au moins potentiellement) d'une infinité d'observations issues d'un processus $(X_t)_{t\in\Theta}$ où est un intervalle de $\mathbb{R}$. Un tel processus est dit à temps continu. Les méthodes présentées dans ce cadre sont différentes de celles pour les séries chronologiques à temps discret et présentées dans la suite.

Nous considérerons uniquement des processus stochastiques $(X_t)_{t\in\Theta}$ à temps discret et unidimensionnels : chaque observationest X_t un réel. On peut également s'intéresser à des séries chronologiques multidimensionelles, c'est à dire telles que X_t soit un vecteur de$\mathbb{R}^d$.

1.2.2 Représentation graphique

On représente graphiquement la série chronologique $\{X_t\}_{t\in T}$

— en dessinant le nuage formée par les points $\left(t_j, X_j\right)_{1\leq j\leq n}$
— en reliant les points entre eux par des segments de droite, pour indiquer la chronologique.

1.2.3 Quelque exemple

Exemple 1.2 *On considére la série du nombre annuel de tâches solaires entre 1700 et 2005.*

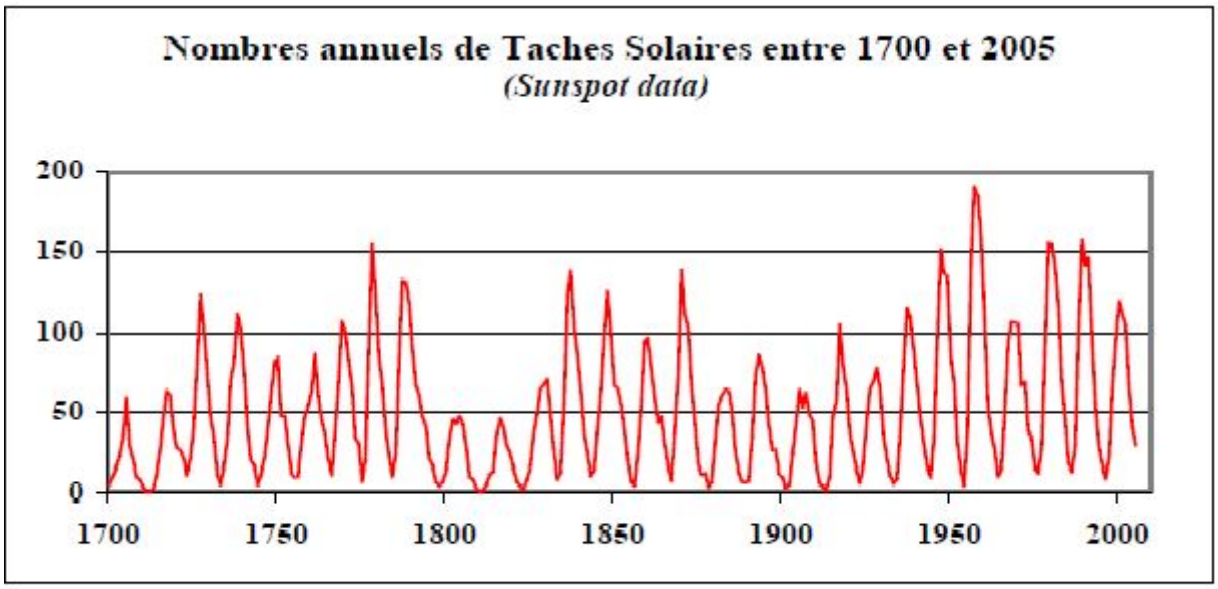

FIGURE 1.1 – Réprésenter graphique de nombre annuel des taches solaires de 1700 à2005.

Exemple 1.3 *:On s'intéresse a l' evolution de la population des USA. On dispose de données recueillies tous les 10 ans entre 1790 et 1990.*

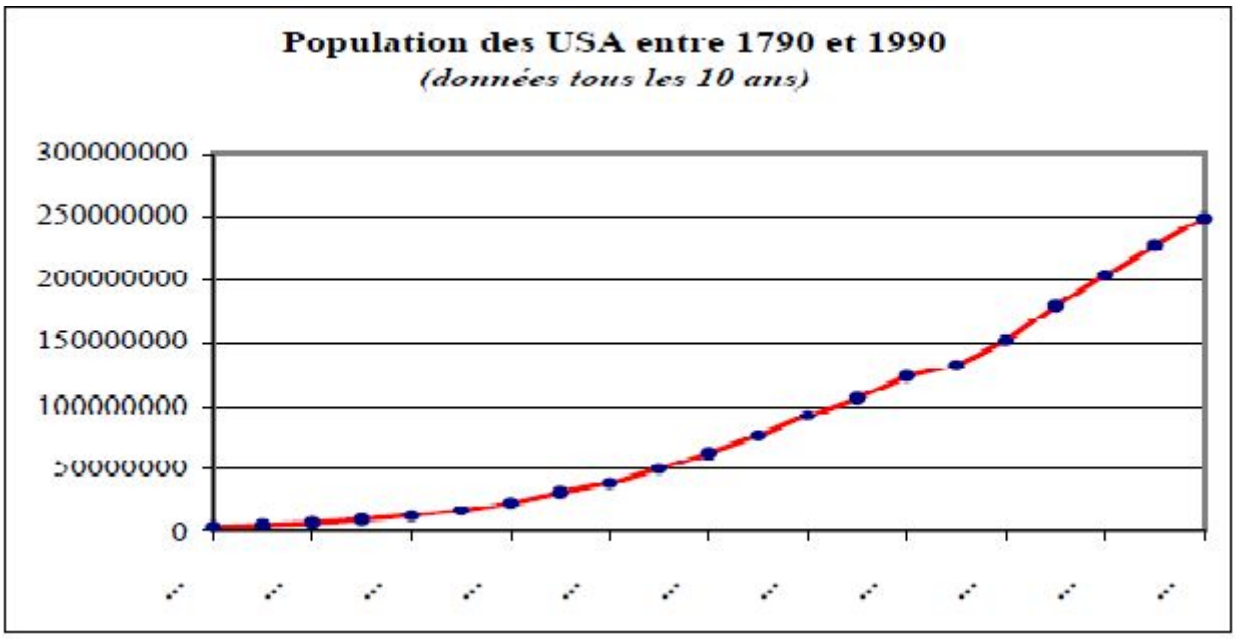

FIGURE 1.2 – Réprésenter graphique de la population des USA entre 1790 et 1990.

Remarque 1.1 *Les dates d'observations sont généralement ordonnées de manière régulière dans le temps : on manipule des série.*

- ***Journalières****(cours d'une action en bourse)*
- ***Mensuelles****(consommation mensuelle d'électricité)*
- ***Trimestrielles****(nombre trimestriel de chômeurs)*
- ***Annuelles****(chiffre annuel des bénéfices des exportations).*

1.3 Notation de base en série chronologique

1.3.1 Décomposition d'une série temporelle

Le but de la décomposition d'une série chronologique est de distinguer dans l'évolution de la série, une **tendance** « générale », des **variations saisonnières** qui se répètent chaque année, et des **variations accidentelles** imprévisibles.

L'intérêt de ceci est d'une part de mieux comprendre, de mieux décrire l'évolution de la série, et d'autre part de prévoir son évolution (à partir de la tendance et des variations saisonnières).

1.3.1.1 La tendance

La tendance est une fonction déterministe à variation que l'on espère lente, qui capte les variations de niveau et que l'on espère assez lisse (variations à long terme de la série)

1.3.1.2 Les variations saisonnières

Les variations saisonnières sont des fluctuations périodiques à l'intérieur d'une année, et qui se reproduisent de façon plus ou moins permanente d'une année sur l'autre.

Ces variations sont dues au rythme des saisons : matières premières, congés

Deux principes :

— Principe de répétition à l'identique : Les variations saisonnières sont périodiques de période p (nombre de mois)

$$s_{t+p} = s_t \tag{1.2}$$

— Principe de conservation des aires : Par année, l'influence des variations saisonnières est nulle.Cela sera traduit à l'aide de la moyenne des S_t. On en reparlera lorsqu'on aura défini les modèles de composition.

1.3.1.3 Les variations accidentelles ou résiduelles

Les variations accidentelles sont des fluctuations irrégulières et imprévisibles.Elles sont supposées en général de faible amplitude.Elles proviennent de circonstances non prévisibles : catastrophes naturelles, crise boursière, grèves.

1.3.2 Modèles de composition

1.3.2.1 Modèle additif

Dans un modèle additif, on suppose que les 3 composantes : tendance, variations saisonnières et variations accidentelles sont indépendantes les unes des autres. On considère que la série Y_t s'écrit comme la somme de ces 3 composantes :

$$y_t = c_t + s_t + \varepsilon_t \tag{1.3}$$

Graphiquement, l'amplitude des variations est constante autour de la tendance

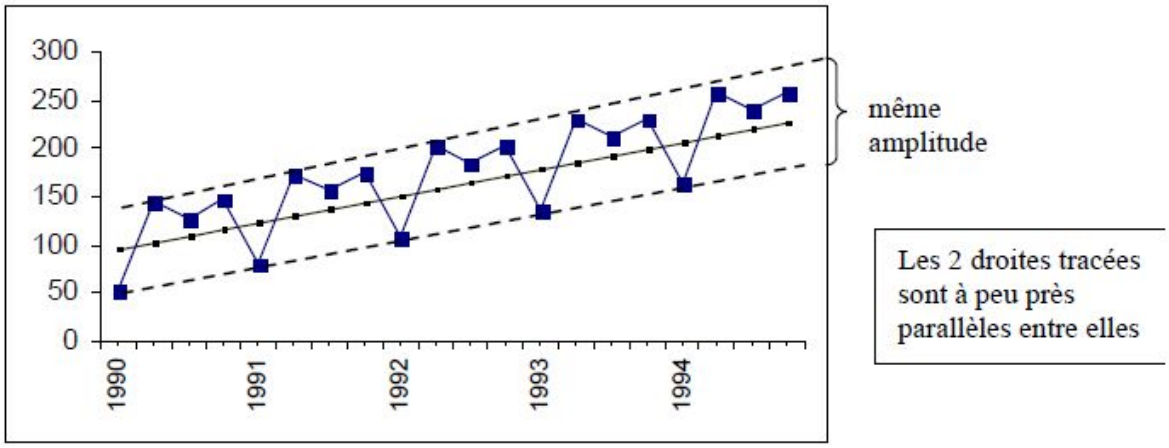

FIGURE 1.3 – Réprésentation graphique d'une modèle additif.

1.3.2.2 Modèle multiplicatif

1.3.2.2.1 Le premier forme de modèle multiplicatif : On suppose que les variations saisonnières dépendent de la tendance.Et on considère que Yt s'écrit de la manière suivante :

$$y_t = c_t \times s_t + \varepsilon_t \tag{1.4}$$

Graphiquement, l'amplitude des variations (saisonnières) varie.

1.3.2.2.2 La deuxième forme de modèle multiplicatif : On suppose que les variations saisonnières et les variations accidentelles dépendent de la tendance.

Et on considère que Y_t s'écrit de la manière suivante :

$$y_t = c_t \times s_t \times \varepsilon_t \tag{1.5}$$

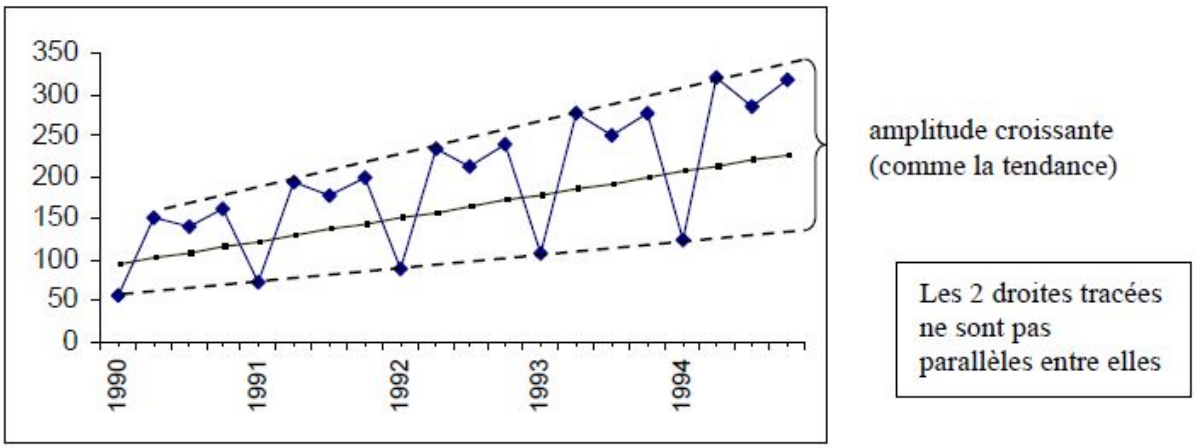

FIGURE 1.4 – Réprésentation graphique d'une modèle multiplicatif.

— Dans le cas d'une série (Yt) à valeurs positives, ce 2^{eme} modèle multiplicatif se ramène à un modèle additif en considérant la série $(ln(Y_t))$

$$ln(Y_t) = ln(C_t) + ln(S_t) + ln(\varepsilon_t) \tag{1.6}$$

— La seule différence entre les deux modèles multiplicatifs est dans l'estimation des ε_t, qui n'a pas une grande importance.

1.3.3 Autocorrélation de série temporelle (chronologique)

1.3.3.1 Autocorrélation

Définition 1.3 *Pour un processus (X_t) en calculant son autocorrélation de retard d noté ρ_d :*

$$\rho_d = corr(X_t, X_{t-d}) = \frac{cov(X_t, X_{t-d})}{\sqrt[2]{v(X_t), v(X_{t-d})}} \tag{1.7}$$

qui mesure le lien entre les valeurs du processus à deux dates distantes de d. Pour un **processus stationnaire** , d prend une forme plus simple :

$$\rho_d = \frac{cov(X_t, X_{t-d})}{v(X_t)} = \frac{\gamma_d}{\gamma_0} \tag{1.8}$$

1.3.3.2 Autocorrélation partielle

Définition 1.4 *De même, on définit l'autocorrélation partielle de retard d comme la corrélation entre $(X_t - X_t^*)$ et $(X_{t-d} - X_{t-d}^*)$*

où X_t^* désigne la régression de X_t sur les $(d-1)$ valeurs $\{X_{t-1}, X_{t-2}, ..., X_{t-d+1}\}$

$$\tau_d = corr(X_t - X_t^*, X_{t-d} - X_{t-d}^*) = \frac{cov(X_t - X_t^*, X_{t-d} - X_{t-d}^*)}{\sqrt[2]{v(X_t - X_t^*)v(X_{t-d} - X_{t-d}^*)}} \tag{1.9}$$

Avec

$$X_t^* = \Sigma_{k=1}^{d-1} \alpha_k X_{t-k} \text{ et } \; X_{t-d}^* = \Sigma_{k=1}^{d-1} \beta_k X_{t-k} \tag{1.10}$$

Où les α_k et β_k est coefficients des régressions. Cette quantité rend compte de l'intensité de la liaison entre X_t et X_{t-h} en supprimant les liaisons induites par des variables intermédiaires $\{X_{t-1}, X_{t-2}, ...X_{t-h+1}\}$. On peut ainsi remarquer que pour tout processus, $\rho_1 = \tau_1$, puisque qu'il n'y a aucune variable intermédiaire entre X_t et X_{t-1}. Comme en régression multiple, l'estimation des d nous permet de mesurer le retard qu'il faut remonter pour trouver une information originale sur X_t.

Exemple 1.4 *soit j est un vecteur à valeur réel. Dans ce exemple en à définie la fonction autocorrélation et la fonction autocorrélation partielle.*

```
> j=c(4.6, 6.1, 7.5, 7.6, 9.2, 10.3, 9.3, 8.9, 12.6)
> j
[1]  4.6  6.1  7.5  7.6  9.2 10.3  9.3  8.9 12.6
> acf(j,plot = FALSE)

Autocorrelations of series 'j', by lag

     0      1      2      3      4      5      6      7      8
 1.000  0.378  0.190  0.160 -0.121 -0.296 -0.187 -0.260 -0.362
> acf(diff(j), plot = FALSE)

Autocorrelations of series 'diff(j)', by lag

     0      1      2      3      4      5      6      7
 1.000 -0.123 -0.474  0.106  0.144 -0.269  0.026  0.091
> acf(diff(j, differences = 2), plot = FALSE)

Autocorrelations of series 'diff(j, differences = 2)', by lag

     0      1      2      3      4      5      6
 1.000  0.006 -0.455  0.059  0.200 -0.248 -0.063
>
```

FIGURE 1.5 – L'execution de la fonction ACF et PACF de vecteur j en R.

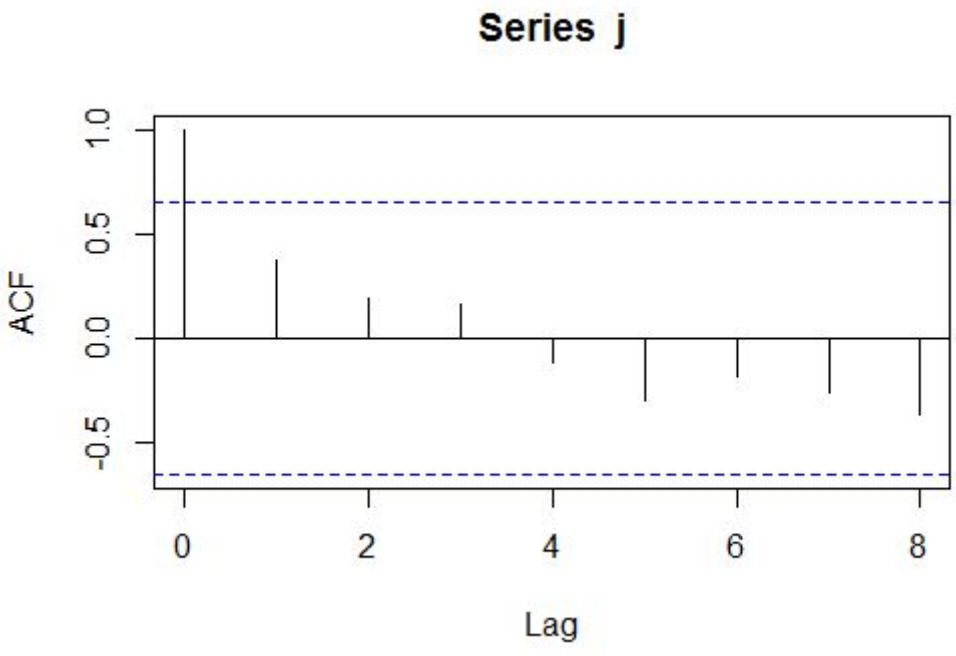

FIGURE 1.6 – Réprésentation graphique de fonction d'autocorrélation.

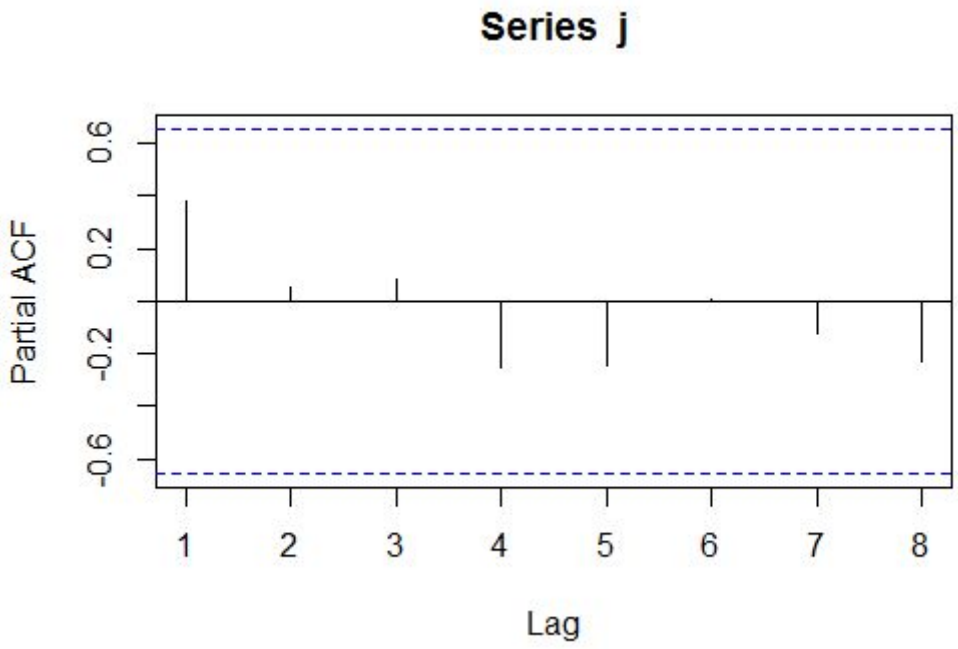

FIGURE 1.7 – Répresentation graphique de fonction d'autocorrélation partielle.

Nous allons nous intéresser aux fonctions d'autocorrélations ACF et PACF , dans le but d'en tirer des résultats pouvant nous guider à choisir le bon modèle.

1.3.4 Opérateurs définis sur une série chronologique

1.3.4.1 Opérateur de retard

Définition 1.5 *L'opérateur de retard B se définit de la manière suivante*

$$B(X_t) = X_{t-1}. \tag{1.11}$$

Remarque 1.2 $B_n(X_t) = X_{t-n}$ *pour tout* $n \in \mathbb{N}$

1.3.4.2 Opérateur de différence d'ordre d

Définition 1.6 *On définit l'opérateur* Δ_d *de différence d'ordre d comme l'opérateur linéaire tel que :*

$$\Delta_d(X_t) = X_t - X_{t-d} = \left(1 - B^d\right) X_t \tag{1.12}$$

On peut aussi prendre l'opérateur d'ordre un et l'appliquer plusieurs fois :

$$\Delta^2(X_t) = \Delta(\Delta(X_t)) = \Delta(X_t - X_{t-1}) = (1-B)(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2} \tag{1.13}$$

Ces opérateurs peuvent être utilisés afin de transformer un processus de ***moyenne non nulle*** *en un processus de* ***moyenne nulle****. On peut aussi s'en servir pour enlever la composante saisonnière de la série. Dans la prochaine section,on les utilise pour mieux représenter les modèles.*

1.4 Objective

Mise en place de techniques mathématiques pour l' etude des séries chronologiques, dans le but de :

— **Comprendre le passé** : analyser et expliquer les valeurs observées.
— **Prédire le futur** : bâtir des prévisions pour les valeurs non encore observées.
— **Etudier le lien avec d'autres séries chronologiques**.

1.5 Modèles d'une série chronologique

1.5.1 Processus aléatoire stationnaires

Définition 1.7 *Un processus aléatoire $(X_t)_{t\geq 0}$ est stationnaire s'il est d'espérance constante ($E(X_t)$ ne dépend pas de t) et si les* ***covariance sont stables par translation dans le temps*** *($Cov(X_t; X_{t+h})$ ne dépend pas de t, quel que soit $h \geq 0$). Attention, ce n'est pas la définition usuelle des ouvrages de probabilités. Pour un processus $(X_t)_{t\geq 0}$ stationnaire.*

1.5.1.1 Processus autorégressif AR(p)

Définition 1.8 *La suite $(X_t)_{t>0}$ est un processus autorégressif d'ordre p ($p > 0$) s'il peut s'écrire sous la forme suivante :*

$$X_t = \sum_{K=1}^{p} \phi_k X_{t-K} + \varepsilon_t ou : \varepsilon_t \to \mathcal{N}(0, \sigma_t^2) \tag{1.14}$$

où $\phi_k (k = 1, 2, ..., p)$ constituent les paramètres du modèle, dans ce cas, on note $\{X_t\} \sim AR(p)$. De la même façon, on peut réécrire un processus AR(p) avec un polynôme $\phi(B)$ qui multipliera X_t cette fois-ci :

$$\phi(B)X_t = \varepsilon_t \; avec : \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p. \tag{1.15}$$

Un processus AR(1) prend la forme suivante :

$$X_t = \phi X_{t-1} + \varepsilon_t : où \; \varepsilon_t \to \mathcal{N}(0, \sigma_t^2) \tag{1.16}$$

1.5.1.1.1 Autocorré On montre que les autocorrélations sont solutions des équations Yule et Walker :

$$\rho(h) = \sum_{K=1}^{p} \phi_k \rho(h - K) = 0 \tag{1.17}$$

1.5.1.1.2 Autocorrélation partielle Dans un tel processus, X_t et X_{t-p+1} sont indépendants conditionnellement aux valeurs intermédiaires $\{X_{t-1}, X_{t-2}, ..., X_{t-p}\}$ et donc : $h > p \Rightarrow \tau_h = 0$

La valeur a la date t dépend des p dates précédentes et pas des autres.

1.5.1.2 Processus à moyenne mobile MA(q)

Définition 1.9 *On dit que la suite $(X_t)_{t>0}$ est un processus à moyenne mobile d'ordre q ($q > 0$) si celui-ci peut s'écrire sous la forme suivante :*

$$X_t = \sum_{k=1}^{q} \theta_k \varepsilon_{t-k} + \varepsilon_t : où\ \varepsilon_t \to \mathcal{N}(0, \sigma_t^2) \tag{1.18}$$

où $\theta_k, (k = 1, 2, ..., q)$ constituent les paramètres du modèle ; dans ce cas, on note $\{X_t\} \sim MA(q)$.

Exemple 1.5 *Un processus MA(1) prend la forme suivante :*

$$X_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t : où\ \varepsilon_t \to \mathcal{N}(0, \sigma_t^2). \tag{1.19}$$

1.5.1.2.1 Autocorrélation Pour un tel processus, on peut montrer que l'autocorrélation est $\rho(h)$ nulle pour $h > q$:

$$\rho(h) = \begin{cases} \frac{\theta_h + \sum_{K=1}^{K=q} \theta_K \theta_{h-K}}{1 + \sum_{K=1}^{K=h} \theta_K^2}, & si : h \leq K \\ 0, & \text{sinon} \end{cases} \tag{1.20}$$

Cette propriété est évidemment trés précieuse pour l'identification du modéle et la détermination de l'ordre q d'un processus MA(q).

1.5.1.3 Processus mixtes ARMA(p,q)

Définition 1.10 *Un processus auto-régressif en moyenne mobile d'ordres p, q (tels que p et q différent de 0) est un processus qui peut s'écrire :*

$$X_t = \sum_{k=1}^{p} a_q X_{t-k} + \sum_{j=0}^{q} b_j \epsilon_{t-j}; \forall t \geq 0. \tag{1.21}$$

où les (ϵ_j) sont des bruits blancs centrés de variance σ. Par convention : $X_{-k} = \epsilon_{-k}$ pour tout k dans $\mathbb{N}$.

On dira aussi que (X_t) est un processus $ARMA(p, q)$.

1.5.1.3.1 Prévision Si le modèle est un processus ARMA, la prédiction pour X_{n+h}, sachant $X_1, ..., X_n$ est :

$$\hat{X}_{n,h} = c_1 X_1 + ... + C_n X_n \tag{1.22}$$

où les coefficients sont choisis de manière à minimiser l'erreur quadratique :

$$E\left[(X_{n+h} - c_1 X_1 - ... - c_n X_n)^2\right] \tag{1.23}$$

TABLE 1.1 – Tableau des propriétés

Modèle	$MA(q)$	$AR(p)$	$ARMA(p;q)$
auto-corrélation	$\rho(h) = 0$ si $h > q$	$\rho(h) \underset{h\longrightarrow+\infty}{\longrightarrow} 0$	$\rho(h) \underset{h\longrightarrow+\infty}{\longrightarrow} 0$
auto-corrélation partielle	$r(h) \underset{h\longrightarrow+\infty}{\longrightarrow} 0$	$r(h) = 0$ si $h > q$	$r(h) \underset{h\longrightarrow+\infty}{\longrightarrow} 0$

Ces propriétés servent à identifier la nature des séries temporelles.

Remarque 1.3 *Sous R, on utilisera les fonctions ACF, PACF qui tracent, respectivement, les $\rho(h)$ et les $r(h)$ (les auto-corrélations empiriques et les auto-corrélations partielles empiriques).*

1.5.2 Processus aléatoire non stationnaires

1.5.2.1 Les processus ARIMA et SARIMA

Nous disposons d'une série temporelle $(X_1,...,X_n)$ qui a une saisonnalité de période 12. On étudie

$$y_t = X_t - X_{t-12} \tag{1.24}$$

(pour supprimer la saisonnalité). L'ajustement d'un modèle ARMA et les prévisions sont réalisées sur la série (y_t). On écrit ensuite les (X_t) en fonction des (y_t) :

$$\begin{aligned} X_t &= y_t + X_{t-12} \\ &= y_t + y_{t-12} + X_{t-24} \\ &= ..... \\ &= y_t + y_{t-12} + .. + y_{r+12} + X_r \end{aligned} \tag{1.25}$$

Avec $r = t$ modulo 12 (le reste de la division euclidienne de t par 12). Ce sont des généralisation des processus ARMA aux cas non stationnaires, avec tendance polynômiale (ARIMA) ou avec une saisonnalité (SARIMA). Ce sont les processus directement utilisés par R.

Définition 1.11 *Le processus $(X_t)_{t\geq 0}$ est un processus **ARIMA(p ; d ; q)** si le processus $y_t = \Delta_1^d X_t$ tq X_t est une processus ARMA(p ; q).*

Les processus ARIMA(p ; d ; q) sont donc bien adaptés à l'étude des séries temporelles présentant une tendance polynômiale de degré d-1.

Définition 1.12 *Le processus $(X_t)_{t\geq 0}$ est un processus **SARIMA(p ; d ; q ; T)** si le processus $y_t = \Delta_t \Delta_1 X_t$ est un processus ARMA(p ; q).*

Les processus SARIMA(p ; d ; q ; T) sont donc bien adaptés à l'étude des séries temporelles qui présentent une saisonnalité de période T et qui ont une tendance polynômiale de degré $d-1$.

1.5.3 Processus stockastique non linéaire

1.5.3.1 Processus ARCH

Commençant tout abord à présenter le modèle ARCH(1) introduit par Engle (1982) [**?**]. Le processus x_t est un processus ARCH (1) si $x_t = \sqrt{h_t}\varepsilon_t$ tel que ε_t est un bruit blanc gaussien, $\varepsilon_t \to \mathcal{N}(0,1)$ et $h_t = w_0 + w_1 x_{t-1}^2$ représente la variance conditionnelle du processus x_t.

Les moments conditionnels se représentent comme :

$$E[x_t/h_t] = E[\sqrt{h_t}\varepsilon_t/h_t] = \sqrt{h_t}E[\varepsilon_t/h_t] = 0. \tag{1.26}$$

La variance conditionnelle s'exprime de la manière suivante :

$$V[x_t/h_t] = V[\sqrt{h_t}\varepsilon_t/h_t] = \sqrt{h_t}V[\varepsilon_t/h_t] = \sqrt{h_t}E[\varepsilon_t^2/h_t] = 0. \tag{1.27}$$

On détermine **l'espérance non conditionnelle** du processus x_t comme suit :

$$E[x_t] = E[E[x_{tt}]] = E[0] = 0. \tag{1.28}$$

La variance du processus à l'instant t est donnée par :

$$V[x_t] = \frac{w_0}{1 - w_0} \tag{1.29}$$

De même **la covariance de ce processus** est définie par :

$$cov[x_t, x_{t+h}/x_{t-m}] = 0 \text{ ou } \forall h > 0, \forall m > 0 \tag{1.30}$$

— Cas générale : Un processus **ARCH (p)** est un processus x_t défini par :

$$x_t = \sqrt{h_t}\varepsilon_t \text{ ou } h_t = w_0 + \sum_{i=1}^{p} w_i x_{t-i}^2. \tag{1.31}$$

L'espérance du processus est donnée par : $E[x_t] = 0$. De ce qui est la variance, il suffit de définir une formule de récurrence de façon à exprimer explicitement cette variance en passant par le calcul de la limite. A ce stade on ne ferait pas les calculs, on se contente de fournir le résultat et de supposer que si $\sum_{i=1}^{p} w_i < 0$, alors **la variance du processus** existe et elle s'exprime par la formule suivante :

$$V[X_t] = \frac{w_0}{1 - \sum_{i=1}^{p} w_i} \tag{1.32}$$

— **Les Conditions de la stationnarité** Un processus ARCH(p) est dit stationnaire si $\sum_{i=1}^{p} w_i < 1$

1.5.3.2 Processus GARCH

Un processus GARCH (1,1) s'écrit de la forme :

$$X_t = \sqrt{h_t}\varepsilon_t \text{ ou } h_t = w_0 + w_1 X_{t-1}^2 + \beta_1 h_{t-1} \tag{1.33}$$

Avec $\varepsilon_t \to \mathcal{N}(0,1)$ comme c'est précédemment défini et h_t représente la variance conditionnelle du processus. Ainsi, on fournit les moments conditionnels : $E[X_t/F_t] = 0$, où F_t représente la filtration engendrée par les valeurs passées de X_t, X_t^2 et de h_t . De ce fait.

La variance conditionnelle du modèle sera alors :

$$V[X_t/F_t] = V[\sqrt{h_t}\varepsilon_t/F_t] = h_t V[\varepsilon_t/h_t] = 0. \tag{1.34}$$

Pour ce qui est des **moments non conditionnels**, l'espérance ainsi que **la variance du processus** s'obtiennent de la façon suivante :

$$E[X_t] = E[E[x_t/h_t]] = 0 \tag{1.35}$$

De même

$$V[X_t] = \frac{w_0}{1-(w_1+\beta_1)}. \tag{1.36}$$

— Cas générale : Un modèle **GARCH (p, q)** s'écrit de la façon suivante :

$$X_t = \sqrt{h_t}\varepsilon_t \text{ ou } h_t = w_0 + \sum_{i=1}^{p} w_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j} \tag{1.37}$$

L'espérance s'obtient simplement à partir de la loi des espérances itérées :

$$E[X_t] = E[E[X_t/h_t]] = 0 \tag{1.38}$$

Si le processus $GARCH(p,q)$ est stationnaire au second ordre, on aura alors : $(\sum_{i=1}^{p} w_i + \sum_{j=1}^{q} \beta_j) < 1$. Cette condition s'avère nécessaire pour définir la variance non conditionnelle par :

$$V[X_t] = \frac{w_0}{1-(\sum_{i=1}^{p} w_i + \sum_{j=1}^{q} \beta_j)} \tag{1.39}$$

CHAPITRE 2

TESTS DE RUPTURE

2.1 Introduction

Une rupture est définie comme un changement, dans une série temporelle, dans le comportement des données à partir d'un certain point. La majeure partie du temps, la moyenne des données avant et après la rupture ne seront pas les mêmes. Cela s'appelle une rupture abrupte. Certaines ruptures se passent en deux points où il y a une transition entre ceux-ci. Dans d'autres cas, une tendance apparait après la rupture, d'où le nom de déclenchement de tendance. Dans ce mémoire, nous nous intéressons particulièrement aux ruptures abruptes, aux déclenchements de tendance et aux tendances.

Plusieurs tests de ruptures existent déjà dans la littérature, tous aussi presenter uns des autres, dans ce chapitre deux types de tests de rupture paramétriques et non paramétriques. Ceux-ci seront comparés aux test présentés à l'aide de simulations de plusieurs types de rupture qui tiendront compte de plusieurs facteurs.

Et nous présenterons quelques méthodes de détection de rupture souvent utilisées dans les séries chronologiques. Ces méthodes permettent principalement d'effectuer l'identification et la modélisation des changements des paramètres statistiques et stochastiques liés à la série chronologique. Ces paramètres statistiques sont éventuellement caractérisés dans notre cas par le changement des moyennes et des variances d'une série temporelle.

2.2 Tests de rupture

2.2.1 Les conditions d'application des méthodes

Les procédures statistiques qui sont étudiées regroupent des tests statistiques classiques, une méthode bayésienne et une technique de segmentation de séries chronologiques. Leurs conditions d'application sont ici précisées.

2.2.1.1 Les tests classiques

— **Le test de corrélation sur le rang** ne suppose aucune propriété particulière de la série chronologique étudiée.
— **Le test de Pettitt** s'applique à des séries non autocorrélées et requiert implicitement que la variance de la série ne soit pas affectée par la rupture si une rupture en moyenne est prioritairement recherchée.

2.2.1.2 Procédure bayésienne

La procédure bayésienne de Lee et Heghinian impose normalité, non-autocorrélation et constance de la variance.

2.2.1.3 Segmentation

La segmentation qui fait intervenir le test de Scheffé sous-entend implicitement la normalité de la série chronologique.

2.3 Tests de rupture non paramétriques

2.3.1 Test de Mann-Whitney

Le test Mann-Whitney est un test non paramétrique pouvant être utilisé à la place d'un test t-test non apparié. Il est utilisé pour tester l'hypothèse nulle selon laquelle deux échantillons proviennent du un même population (c'est-à-dire avoir la même médiane) ou, alternativement, si des observations dans une échantillon ont tendance à être plus grande que les observations dans l'autre. Bien que ce soit un paramétrique test, il suppose que les deux distributions ont une forme similaire.

2.3.1.1 Réalisation du test U de Mann-Whitney

Supposons que nous ayons un échantillon n_x d'observations $\{X_1, X_2, ..., X_n\}$ dans un groupe (c'est-à-dire d'un groupe population) et un échantillon de n_y observations $\{Y_1, Y_2, ..., Y_n\}$ dans un autre groupe (c'est-à-dire de une autre population).

Le test de Mann-Whitney est basé sur une comparaison de chaque observation X_i du premier échantillon avec chaque observation Y_j dans l'autre échantillon. Le nombre total de paires les comparaisons qui peuvent être faites est $n_x\ n_y$.

Si les échantillons ont la même médiane, chaque X_i à une chance égale (c'est-à-dire une probabilité$\frac{1}{2}$)d'être plus grand ou plus petit que chaque Y_j).

Donc, sous l'hypothèse nulle :

$$H_0 = P\left(X_i > Y_j\right) = \frac{1}{2} \tag{2.1}$$

et sous l'hypothèse alternative :

$$H_1 = P\left(X_i < Y_j\right) \neq \frac{1}{2} \tag{2.2}$$

Nous comptons le nombre de fois où un X_i de l'échantillon 1 est supérieur à un Y_j de l'échantillon 2.ce nombre est noté U_X. De même, le nombre de fois qu'un X_i de l'échantillon 1 est plus petit qu'un Y_j de l'échantillon 2 est noté U_Y. Sous l'hypothèse nulle, nous voudrions s'attendre à ce que U_X et U_Y soient approximativement égaux.

2.3.1.2 Procédure pour effectuer le test

1. Disposez toutes les observations par ordre de grandeur.
2. Sous chaque observation, écrivez X ou Y (ou un autre symbole pertinent) pour indiquer de quel échantillon ils proviennent.
3. Sous chaque X, écrivez le nombre de Y_S qui se trouvent à sa gauche (c.-à-d. Plus petit que cela) ; cela indique $X_i > Y_j$. Sous chaque Y, notez le nombre de X qui sont à sa gauche (c'est-à-dire plus petit que celle-ci) ; cela indique $X_i > Y_j$
4. Additionnez le nombre total de $X_i > Y_j$ - dénoté par U_X. Ajouter le nombre total des temps $Y_j > X_i$- désignent par U_Y. Vérifiez que $U_X + U_Y = n_x\, n_y$.
5. Calculer $U = min(U_X, U_Y)$.
6. Utilisez les tableaux statistiques du test U de Mann-Whitney pour trouver la probabilité d'observer une valeur de U ou inférieure. Si le test est unilatéral, il s'agit de votre p-valeur ; si le test est un test bilatéral, doublez cette probabilité pour obtenir la valeur p.

Remarque 2.1 *Si le nombre d'observations est tel que n_x, n_y est assez grand (> 20), un normale approximation peut être utilisé avec*

$$\mu_U = \frac{n_x n_y}{2}, \qquad \sigma_U = \sqrt{\frac{n_x n_y\,(N+1)}{12}} \tag{2.3}$$

avec : $N = n_x + n_y$.

2.3.1.3 Traiter avec des cravates

Il est possible que deux ou plusieurs observations soient identiques. Si ce est le cas, nous pouvons toujours calculer U en allouant la moitié du lien à la valeur X et la moitié de la lier à la valeur Y. Cependant, si tel est le cas, alors l'approximation normale doit être utilisé avec un ajustement à l'écart type. Cela devient :

$$sigma_U = \sqrt{\frac{n_x n_y}{N\,(N-1)}\left[\frac{N^3 - N}{12} - \sum_{j=1}^{g}\frac{t_j^3 - t_j}{12}\right]} \tag{2.4}$$

avec :

$N = n_x + n_y$.

g : le nombre de groupes de liens.

t_j : le nombre de rangs liés dans le groupe j.

2.3.2 Test de Pettitt

Ce dernier a considéré les séries $X_1, X_2, ..., X_N$ comme étant une séquence de variables aléatoires indépendantes. La séquence $X_1, X_2, ..., X_N$ est supposée contenir un point de rupture à l'instant τ et ceci en ayant la condition que les séries $X_{1,...,\tau}$ aillent une distribution commune $F_1(x)$, et aussi que les séries $X_{\tau+1,...,N}$ aillent une distribution commune $F_2(x)$ tel que :

$$F_1(x) \neq F_2(x) \tag{2.5}$$

2.3.2.1 L'hypothèses de ce test

L'hypothèse nulle qui correspond au cas de non-rupture est désignée par H_0 telle que $\tau = N$, contrairement à cette hypothèse, c'est à dire celle qui correspond à l'alternative de rupture est désignée par H_1 telle que $1 \leq \tau < N$ Le test de Pettitt est considérée comme un test statistique non paramétrique. Pour lui, on estime qu'aucune condition particulière n'est obligatoire pour les formes fonctionnelles de $F_1(x)$ et $F_2(x)$.

Pettitt a prouvé efficacement comment une formulation appropriée du test de MannWhitney peut être utilisée pour effectuer le test des hypothèses de non-rupture H_0 et de rupture H_1.

La statistique $U_{1,N}$ est considérée pour les valeurs de t bornée entre 1 et N. Pour le test de l'hypothèse H_0 contre celle H_1 , Pettitt a proposé d'utiliser la variable définie par :

$$K_N = \max_t \mid U_{1,N} \mid \tag{2.6}$$

En utilisant la théorie des rangs, Pettitt a donné la probabilité de dépassement approximative d'une valeur k par :

$$P(K_N > K) = 2\exp(\frac{-6K^2}{(N^3 + N^2)}) \tag{2.7}$$

Pour un risque α de première espèce donnée, H_0 est rejetée si cette probabilité est inférieur α. Dans ce cas, la série présente une rupture du temps $t = \tau$ définissant.

Nous pouvons expliquer ce traitement par l'organigramme suivant :

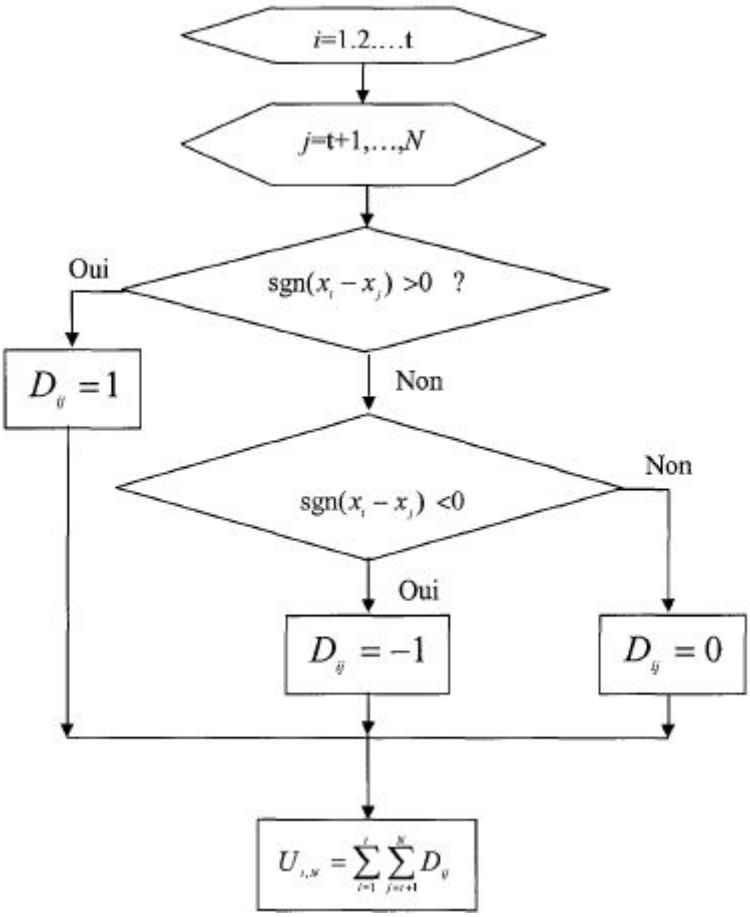

FIGURE 2.1 – Statistique de pettitt.

2.3.2.2 Test de Mann-Kendall

Le test de tendance Mann-Kendall (Mann, 1945 ; Kendall, 1975) est sur la base de la corrélation entre les rangs d'une série chronologique et leur ordre de temps. Pour une série chronologique $X = \{X_1, X_2, ..., X_n\}$, l'hypothèse nulle H_0 est l'hypothèse de stationnarité de la série (absence de tendance). L'hypothèse alternative H_1 correspond à la non stationnarité de la série. La statistique de test est donnée par :

$$S = \sum_{i<j} a_{ij} \tag{2.8}$$

Avec :

$$a_{ij} = sign(X_j - X_i) = signe(R_j - R_i) = \begin{cases} 1 \; si \;\; X_i < X_j \\ 0 \; si \; X_i = X_j \\ -1 \; si \; X_i > X_j \end{cases} \tag{2.9}$$

Et R_i et R_j sont les rangs des observations X_i et X_j du série chronologique, respectivement. La statistique de test ne dépend que du rang des observations, plutôt que leurs valeurs réelles, ce qui entraîne une distribution libre statistique de test. C'est vrai parce que si les données devaient être transformés en n'importe quelle distribution, les rangs

des observations resterait le même. Les tests sans distribution ont l'avantage que leur puissance et leur signification ne sont pas affectées par la distribution réelle des données.

En supposant que les données soient indépendante et identiquement distribuée, la moyenne et la variance de la statistique S:

$$E(S) = 0 \tag{2.10}$$

$$V_0(S) = n(n-1)(2n+5)/18 \tag{2.11}$$

Où n est le nombre d'observations. L'existence de rangs liés (observations égales) dans les résultats de données dans un réduction de la variance de S pour devenir :

$$V_0^* = (n-1)(2n+5)/18 - \sum_{j=1}^{m} t_j\Big(t_j-1\Big)\Big(2t_j+5\Big)/18 \tag{2.12}$$

Où m est le nombre de groupes de rangs liés, chacun avec t_j observations liées. Kendall (1975) montre également que la distribution de S tend à la normalité car le nombre d'observations devient grand.L'importance des tendances peut être testée en comparant les variable normalisée U.avec la norme normale varier au niveau de signification souhaité α où la soustraction ou addition d'unité dans l'équation$U$ est une correction de continuité :

$$U = \begin{array}{cc} (S-1)/\sqrt{V_0(S)} & S>0 \\ 0 & S=0 \\ (S+1)/\sqrt{V_0(S)} & S<0 \end{array} \tag{2.13}$$

Si X est distribué normalement avec la moyenne μ et la variance σ^2, alors$(X_j - X_i)$ sera aussi normalement distribué avec zéro moyen et variance $2\sigma^2$.Nous avons donc :

$$\begin{aligned} E(S) &= E\left(\sum_{i<j} a_{ij}\right) = E\left(\sum_{i<j} signe(X_j - X_i)\right) && (2.14)\\ &= \sum_{i<j} E\Big(signe(X_j - X_i)\Big) && (2.15)\\ &= P\Big(X_j - X_i > 0\Big) - P\Big(X_j - X_i < 0\Big) = 0 && (2.16) \end{aligned}$$

Cette résultat est découle du fait que $(X_j - X_i)$ est normalement distribué avec une moyenne nulle et la variance $2\sigma^2$, qui est symétrique autour de l'origine.La variance de S est donnée par :

$$V(S) = E\Big(S^2\Big) = E\left(\sum_{i<j} a_{ij}\right)^2 \tag{2.17}$$

Pour une série chronologique à n observations, la somme en équation présédente im-

plique $n^2(n-1)^2/4$ termes (Kendall, 1955). On peut montrer que si les éléments de X sont indépendants et ordonné au hasard,danc cette somme est réduit aux trois termes suivants (Kendall, 1955[3]) :

$$Var(S) = \binom{n}{2} E\left({a_{ij}}^2\right) + 6\binom{n}{3} E\left(a_{ij}a_{ik}\right) + 6\binom{n}{4} E\left(a_{ij}a_{kl}\right)$$

1. La première attente dans cette l'équation comptes pour termes avec des suffixes identiques i et j.
2. La seconde attente dans cette l'équation comptes pour les termes avec un commun suffixe.
3. La troisième attente dans cette l'équation comptes pour termes avec suffixes distincts.

Pour $i \neq j$;${a_{ij}}^2 = 1$ et donc $E\left({a_{ij}}^2\right)$ est toujours égal à l'unité.Aussi, sous le null hypothèse X_i, X_j, X_k, X_l, sont indépendants pour différents suffixes i, j, k, l de sorte que $Y = \left(X_j - X_i\right)$ et $Z = (X_l - X_k)$ sont également indépendants et $E\left(a_{ij}a_{kl}\right) = E\left(a_{ij}\right)E(a_{kl}) = 0$.

2.3.2.3 Formule approximative pour calculer V (S)

Bayley et Hammersley (1946) [1] ont étudié la variance de la moyenne d'un échantillon de taille n lorsque les données sont autocorrélées. En cas de non corrélation la variance de la moyenne est donnée par :

$$Var\left(\overline{X}\right) = \frac{\sigma}{n} \tag{2.18}$$

Où σ est la variance de X. Cependant, lorsque les données sont autocorrélées. La variance correcte de la moyenne dans le cas de données autocorrélées est montré par Bayley et Hammersley (1946) à donner par :

$$V\left(\overline{X}\right) = \frac{\sigma}{n_b^*} = Var\left(\overline{X}\right)\frac{n}{n_b^*} \tag{2.19}$$

Où :

$$\frac{1}{n_b^*} = \frac{1}{n} + \frac{2}{n}\sum_{j=1}^{n-1}(n-j)\rho(j) \tag{2.20}$$

Où : $\rho(j)$ est l'autocorrélation fonction des données et n est le réel nombre d'observations. La valeur de n_b^* est considérée comme un 'efficace'nombre d'observations, pour tenir compte de l'autocorrélation dans les données.

Par analogie avec l'affaire considérée par Bayley et Hammersley, nous proposons une formule empirique pour calculer la variance de S dans le cas de données autocorrélées similaires à celles de l'équation suivante :

$$V^*(S) = Var(S)\frac{n}{n_S^*} = \frac{n(n-1)(2n+5)}{18}\frac{n}{n_S^*} \tag{2.21}$$

Où :

$$\frac{n}{n_S^*} = 1 + \frac{2}{n(n-1)(n-2)}\sum_{i=1}^{n-1}(n-i)(n-i-1)(n-i-2)\rho_S(i) \tag{2.22}$$

Où : n est le nombre réel d'observations et $\rho_S(i)$ est la fonction d'autocorrélation des rangs de les observations

2.3.3 Test de Mann-Kendall modifié

Une version modifiée du test Mann-Kendall qui est robuste en présence d'autocorrélation est proposé, sur la base de la variance modifiée de S donnée par équation 2.17 et son approximation dans l'équation 2.21. l'autocorrélation entre les rangs des observations, $\rho_S(i)$ est d'abord évalué. Les valeurs de $\rho_S(i)$, cependant, doit être calculé après soustraction d'une valeur non paramétrique appropriée estimateur de tendance (Sen, 1968 [2]; Zetterqvist, 1991 [4]). Des exemples d'estimations de l'autocorrélation ont un écart d'ordre $\frac{1}{n}$ (Kendall, 1955 [3]).

En raison de nature des calculs dans les équations (2.17) et (2.21), qui impliquent un grand nombre de termes, il a été constaté que des valeurs non significatives de $\rho_S(i)$ auront un effet négatif effet sur la précision de la variance estimée de S.

Ceci est réalisé en exigeant un niveau de signification prédéfini pour les autocorrélations être inclus dans les calculs, qui peuvent être pris égale à celle du test.

Il y a deux propriétés importantes d'un statistique test qui sont étudiés pour évaluer ses performances :

1. La première propriété est le niveau de signification empirique du test, défini comme la probabilité de rejeter l'hypothèse nulle H_0 d'aucune tendance alors qu'en fait H_0 est vrai. Cette probabilité est le pourcentage d'échantillons rejetés par l'essai sous H_0, et doivent être égal au niveau de signification nominal α du test,à condition que la distribution normale avec zéro moyen et la variance égale à $V(S)$ est la distribution correcte de S.
2. La deuxième propriété est la puissance du test, définie comme la probabilité de rejeter H_0 lorsque le L'hypothèse alternative est vraie. Cette probabilité est la pourcentage d'échantillons rejetés par le test lorsqu'un tendance d'une certaine pente existe dans les données.La puissance d'un bon test devrait augmenter rapidement car la pente de la tendance part de zéro.

Un nombre de numérique des simulations ont été utilisées pour évaluer la performance du test modifié par rapport au test de Mann original Test de tendance Kendall.

Échantillons de taille 2000 de données non corrélées, modèles $AR(1)$ et $ARMA(1,1)$ont été générés. Le modèle $AR(1)$ est caractérisé par ϕ comme suit :

$$X_t = \phi X_{t-1} + e_t \tag{2.23}$$

Dans ce cas :

$$\rho(i) = \phi^{|i|} \tag{2.24}$$

Le modèle $ARMA(1,1)$ est du forme de :

$$X_t = \phi X_{t-1} + e_t + \theta e_{t-1} \tag{2.25}$$

Le modèle $ARMA(1,1)$ est caractérisé par ϕ et $\rho(1)$, où $\rho(1)$ est donné par :

$$\rho(1) = \frac{(1-\phi\theta)(\phi-\theta)}{1-2\phi\theta+\theta} \tag{2.26}$$

2.3.4 Test de corrélation sur le rang

Dans la méthode de Mann-Kendall, le test de corrélation sur le rang est basé sur le calcul du nombre P de paires $\left(X_i, X_j\right)$ pour lequel $X_j > X_i \quad (j > i; i = 1, ..., N-1)$. Le nombre de pair P est obtenu en comparant la valeur du premier terme de la série X_1 avec les autres termes suivants jusqu'au dernier X_N et le nombre des termes dont la valeur dépasse X_1 est compté, et ainsi de suite pour les autres termes jusqu'à X_{N-1}.

2.3.4.1 L'hypothèses de ce test

Donc à partir du nombre de paire$P$et en posant l'hypothèse nulle H_0 de stationnarité de la série on définit la variable w de Mann-Kendall comme suit :

$$w = \frac{4p}{N(N-1)} - 1 \tag{2.27}$$

qui suit une distribution normale de moyenne nulle et de variance égale à :

$$\sigma_t = \frac{2(2N+5)}{9N(N-1)} \tag{2.28}$$

Pour un risque α de première espèce donnée, l'acceptation de H_0 est défini par l'appartenance de w à l'intervalle $\left[-U_{1-\frac{\alpha}{2}}\sigma_t \ , U_{1-\frac{\alpha}{2}}\sigma_t \ \right]$ où $U_{1-\frac{\alpha}{2}}$ est la valeur de la variable normale réduite de probabilité de non dépassement $1-\frac{\alpha}{2}$.

l'hypothèse alternative H_1 de ce test est celle dans le cas où la série chronologique est non stationnaire.

2.3.5 Test de somme cumulative

La technique de la somme cumulative est appliquée au problème de la détection de tout changement soudain dans un système dynamique stochastique. La procédure de test

proposée est non paramétrique et aucune hypothèse concernant l'ordre ou la nature du modèle est faite.

2.3.5.1 Introduction

L'un des problèmes importants des systèmes dynamiques stochastiques est de détecter si un changement soudain s'est produit dans le système.Des illustrations Bohlin a indiqué où de tels changements soudains doivent être détectés (1976), Ishii (1980). L'approche suggérée par Bohlin (1976) [5] consiste à modéliser la relation entrée-sortie du système linéaire à l'aide d'un modèle à moyenne mobile autorégressif à coefficients temporels. En supposant que les coefficients dépendant du temps puissent être décrits par un modèle de marche aléatoire, un algorithme récursif similaire au filtre de Kalman,[12] obtenu pour l'état du système, est conçu pour les paramètres.Ces paramètres ont été utilisés pour obtenir une fonction de «densité spectrale changeante», et ce spectre sera similaire au spectre lorsque les coefficients sont déterministes et ne dépendent pas du temps.

Soit $y(t)$ le signal observé, à l'instant t, $y(t)$ vérifie l'équation des différences.

$$y(t) + a_1(t)\,y(t-1) + .. + a_n(t)\,y(t-n) \tag{2.29}$$
$$= \lambda\left(e(t) + c_1 e(t-1) + ... + c_m e(t-m)\right) + D'(t) \tag{2.30}$$

Où les coefficients $a_i(t)$ sont supposés être un processus aléatoire,$D'(t)$ est un bruit observé.

En définissant certaines variables auxiliaires, Ishii et autre (1980) ont montré que 2.29 peut être écrit sous la forme d'espace d'état :

$$\begin{aligned} \theta(t+1) &= \theta(t) + V(t) \\ x(t) &= X(t)\,\theta(t) + e(t) \end{aligned} \tag{2.31}$$

Où $\theta(t)$ est un vecteur de paramètre, $X(t)$ est un vecteur d'observation, etc. Pour détecter les changements brusques dans les séries chronologiques, ils ont obtenu les fonctions de vraisemblance des innovations (et des processus d'erreur) sur de courts intervalles de temps.Le changement des fonctions de vraisemblance peut être considéré en raison d'un changement dans la structure de la série chronologique.

Bien entendu, le logarithme de la probabilité est fonction des variances de l'innovation, et donc du test proposé par Ishii et al(1980) pour détecter les changements équivaut à tester un changement à divers intervalles des variances de l'innovation.

2.3.5.2 Processus stochastiques évolutifs

Considérons la classe des processus gaussiens non stationnaires à paramètres discrets $\{X_t\}$.

avec $E(X_t) = 0$ $E(X_t) = \sigma_{x,t} < \infty$ tous t et ayant la représentation :

$$X_t = \int_{-w}^{w} \exp(itw) A_t(w) dz(w) \qquad (t = 0, \pm 1, ...) \tag{2.32}$$

où $z(w)$ est un processus orthonormal avec

$$E(dz(w)) = 0 \qquad E|dz(w)| = f(w)dw \tag{2.33}$$

pour chaque fixe w.

Nous supposons que la séquence $\{A_t(w)\}$ à une généralisation Transformée de Fourier dont le module à un maximum absolu à l'origine.

La fonction de densité spectrale évolutive au temps t est alors définie comme étant :

$$f_{x,t}(w) = |A_t(w)| f(w) \tag{2.34}$$

$$\sigma_{x,t} = \int_{-w}^{w} f_t(w) dw. \tag{2.35}$$

Maintenant, il est évident que $\sigma_{x,t}$. peut être interprété comme une mesure de la «puissance totale» du processus $\{X_t\}$ à l'instant t . Et la contribution de la fréquence w, $w + dw$ à l'énergie totale étant t, $f_t(w)dw$.

L'estimation de $f_t(w)$ sur la base d'une réalisation unique a été considéré par Priestley (1965). La procédure d'estimation est basée sur la "double fenêtre technique" et il est comme suit :

Choisissez une "séquence de poids" $\{g_u\}$ satisfaisante :

$$2\pi \sum_{-\infty}^{\infty} |g_u| = \int_{-w}^{w} |T(w)| dw = 1 \tag{2.36}$$

où :

$$T(w) = \sum_{-\infty}^{\infty} g_u \exp(-iuw) \tag{2.37}$$

Pour toute fréquence w, écrivez :

$$U_t(w) = \sum_{-\infty}^{\infty} g_u X_{t-u} \exp(-iw(t-u)) \tag{2.38}$$

Choisissez une autre séquence de poids $\{W_{T'}(t)\}$, en fonction du paramètre T', satisfaisant :

$W_{T'}(t) \geq 0$pour tout T', t.

$$W_{T'}(t) \geq 0 \qquad \forall T', t \qquad \lim_{|t| \to \infty} W_{T'}(t) \to 0 \qquad \forall T' \tag{2.39}$$

$$\sum_{-\infty}^{\infty} W_{T'}(t) = 1 \quad \forall T'; \quad \sum_{-\infty}^{\infty} W_{T'}(t) < \infty \quad \forall T' \tag{2.40}$$

L'estimation de $f_t(w)$est obtenue comme :

$$\overline{f_t}(w) = \sum_{u=-\infty}^{\infty} W_{T'}(u)\,|U_{t-u}(w)|^2 \tag{2.41}$$

Priestley (1965)[9] a montré qu'environ :

$$\begin{aligned} E\left(\overline{f_t}(w)\right) &\rightarrow f_{x,t}w \\ var\left(\overline{f_t}(w)\right) &\rightarrow \sigma f_{x,t}^2(w) \end{aligned} \tag{2.42}$$

Où

$$\left\{ \begin{array}{c} \sigma = \frac{C}{T'}\int_{-w}^{w}|T(w)|^4\,dw \\ C = \lim_{T'\rightarrow\infty}\left\{T'\int_{-w}^{w}|W_{T'}(\lambda)|\,d\lambda\right\} \end{array} \right\}. \tag{2.43}$$

Et

$$W_{T'}(\lambda) = \sum \exp(-i\lambda t)\,W_{T'}(\lambda) \tag{2.44}$$

L'expression 2.41 pour $var\left(\overline{f_t}(w)\right)$ doit être doublée lorsque $w = 0$ ou π. Priestley (1965) [9] a également montré que :

$$cov\left(\overline{f_{t_1}}(w_1), \overline{f_{t_2}}(w_2)\right) = 0 \tag{2.45}$$

En d'autres termes, si les points de fréquence et les points de temps sont choisis suffisamment écartés, les estimations spectrales évolutives sont approximativement non corrélées :

Nous obtenons maintenant la relation entre la variance de l'innovation et la fonction de densité spectrale évolutive. Sous certaines conditions de $f_{x,t}(w)$ Abdrabbo et Priestley (1967) [6] ont montré qu'un processus oscillatoire $\{X_t\}$ (supposé être gaussien ici) a un moyenne mobile unilatérale (avec coefficients dépendants) de la forme :

$$X_t = \sum_{u=0}^{\infty} b_t(u)\,e_{t-n} \tag{2.46}$$

Où Ie e_t est une séquence de indépendante, à distribution identique $N(0,\sigma_e)$ variables. On peut montrer que la fonction de densité spectrale évolutive de X, est donc donné par :

$$\begin{aligned} f_{x,t}(w) &= \left|\sum_{u=0}^{\infty} b_t(u)\exp(-iuw)\right|^2 f_0(w) \\ &= \frac{\sigma_e}{2\pi}\left|\sum_{u=0}^{\infty} b_t(u)\exp(iuw)\right|^2 \end{aligned} \tag{2.47}$$

Considérons maintenant la prédiction de X_t, étant donné l'ensemble $\{X_s, s \leq t-1\}$. le prédicteur qui minimise l'erreur quadratique moyenne est l'espérance conditionnelle de X, étant donné $\{X_s, s \leq t-1\}$. D'où le prédicteur de X_t quand X_t satisfait 2.46est donné par :

$$E\left(X_t/X_s\ , s \leq t-1\right) = \sum_{u=1}^{\infty} b_t\left(u\right) e_{t-u} \tag{2.48}$$

et l'erreur dans le prédicteur est :

$$X_t - E\left(X_t/X_s\ , s \leq t-1\right) = b_t\left(0\right) e_t \tag{2.49}$$

2.3.6 Test de Lombard (1987)

Supposons que l'on a des variables aléatoires $y_1, ..., y_T$ indépendantes qui suivent chacune une distribution continue $F(y, \theta_1), ..., F(y, \theta_T)$. On essait d'étudier s'il y a un changement tel que les paramètres θ_i soient de la forme $\theta_1 = ... = \theta_\tau \neq \theta_{\tau+1} = ... = \theta_T$. Mais on a besoin d'un modèle plus général si les paramètres θ changent graduellement dans un intervalle.

Alors, Lombard (1987) [10] a introduit ce modèle :

$$\theta_i = \begin{array}{c} \rho \text{ si} i = 1, ..., \tau_1 \\ \rho + \frac{i-\tau_1}{\tau_2-\tau_1}\delta \quad \text{si } i = \tau_1 + 1 .... \tau_2 \\ \rho + \delta \text{si} i = \tau_2 + 1, ..., T \end{array} \tag{2.50}$$

pour $\tau_1 < \tau_2$, ρ, δ inconnus .lombard(1987) utilise l'hypothèse suivante :

$$\begin{array}{l} H_0 : \ \delta = 0 \\ H_1 : \ \delta \neq 0 \end{array} \tag{2.51}$$

Pour commencer, voici un peu de notation :

— r_i est le rang de la variable aléatoire y_i,

— ϕ représente une fonction de score avec :

$$0 < \int_0^1 \phi^2\left(\omega\right) d\omega \ < \infty \tag{2.52}$$

$$\bar{\phi} = \frac{1}{T} \sum_{i=1}^{T} \phi\left(\frac{i}{T+1}\right) \tag{2.53}$$

$$A^2 = \frac{1}{T-1} \sum_{i=1}^{T} \left[\phi\left(\frac{i}{T+1}\right) - \bar{\phi}\right]^2 \tag{2.54}$$

Alors, le **score de classement** de y_i, noté $s(r_i)$ est :

$$s(r_i) = \frac{1}{T}\left[\phi\left(\frac{r_i}{T+1}\right) - \bar{\phi}\right] \text{ pour } i = 1,...,T. \tag{2.55}$$

Donc **Lombard (1987)** a introduit le test statistique suivant :

$$U_{\tau_1,\tau_2} = \sum_{j=\tau_1+1}^{\tau_2}\sum_{i=1}^{j-1} s(r_i) = (\tau_2 - \tau_1)\sum_{i=1}^{\tau_1} s(r_i) + \sum_{i=\tau_1+1}^{\tau_2-1}(\tau_2 - i)s(r_i) \tag{2.56}$$

Qui est le test de rang localement le plus puissant pour tester l'hypothèse nulle versus une hypothèse alternative unilatérale si τ_1 et τ_2 sont connus.

La statistique dépend indirectement des y_i en utilisant le rang de y_i, $i = 1,...,T$. S'ils sont inconnus, Lombard propose le test suivant :

$$q_{1.T} = \sum_{t_1=1}^{T-1}\sum_{t_2=t_1+1}^{T} U_{t_1,t_2}^2 \tag{2.57}$$

Si on observe une **grande valeur**, on **rejette l'hypothèse nulle**. Lorsque T tend vers. l'infini ($T \to \infty$), il est à **noter que** :

$$\frac{q_{1.T}}{T^5} \to q_1 = \sum_{n=1}^{\infty}\frac{1}{(n\pi)^4}Z_n^2 \text{ sous } H_0 \tag{2.58}$$

Où $Z = (Z_1, Z_2, ...)^t$ est un vecteur de variables aléatoires indépendantes et identiquement distribuées de loi normale standard. Les points critiques pour différentes valeurs de α et pour toute valeur de T sont présentés dans le tableau suivant :

TABLE 2.1 – Valeurs de test de Lombardl de loi normale pour différentes valeurs de α

α	0.1	0.075	0.05	0.025	0.01
q_1	0,0287	0,0334	0,0403	0,0525	0,0690
$\frac{q_{1.T}}{T^5}$	0.0289	0.0334	0.0402	0.0515	0.0662

Si on est intéressé à tester l'hypothèse nulle versus une hypothèse alternative unilatérale ($\delta > 0$ ou $\delta < 0$).

Lombard propose trois différents scores :

- **Le score de Wilcoxon :** $\Phi_1(u) = 2u - 1$, pour tester des changements dans la position (une tendance centrale),
- **Le score de Mood :** $\Phi_2(u) = (2u-1)^2$, pour tester les changements d'échelle
- **Le score logarithmique :** $\Phi_3(u) = log(1-u)$ (les auteurs utilisent $-log(1-u)$),pour tester des changements d'échelle dans une distribution ayant pour domaine $[0, \infty[$.

Pour le score de Wilcoxon, la fonction de score $s(r_i)$ est équivalente à :

$$s(r_i) = \sqrt{\frac{12}{T(T+1)}}\left(r_i - \left[\frac{T+1}{2}\right]\right) \tag{2.59}$$

Lorsqu'on veut tester s'il y a l'apparition d'une tendance, c'est-à-dire que $\tau_2 = T$ et que $\tau_1 = T$, la statistique du test de Lombard est modifiée tel que $t_2 = T$:

$$q^*_{1.T} = \sum_{t=1}^{T-1} U^2_{t,T} \tag{2.60}$$

Et, lorsque T tend vers l'infini ($T \to \infty$), sous l'hypothèse nulle :

$$\frac{q^*_{1.T}}{T^4} \to q^*_1 = \sum_{n=1}^{\infty} \gamma_n Z^2_n \tag{2.61}$$

Ou $\gamma_1 > \gamma_2 > 0$ sont les solutions positives réelles de $tan(\gamma^{-1/4}) + tanh(\gamma^{-1/4}) = 0$. Les points critiques pour différentes valeurs de T et pour toute valeur de T sont présentés dans le tableau suivant :

TABLE 2.2 – Valeurs de test de Lombardl

α	0.1	0.075	0.05	0.025	0.01
q^*_1	0.0879	0.1027	0.1242	0.162	0.2135
$\frac{q^*_{1.T}}{T^4}$	0.0882	0.1026	0.1241	0.158	0.2035

Lorsqu'on veut tester s'il y a un changement abrupte, c'est-à-dire que $\tau_1 = \tau$ et $\tau_2 = \tau + 1$, toujours en utilisant la fonction de score de Wilcoxon, la statistique de test de Lombard devient :

$$q^0_{1.T} = \sum_{t=1}^{T-1} U^2_{t,t+1} = \sum_{t=1}^{T-1} \left[\sum_{i=1}^{t} S(r_i) \right]^2 \tag{2.62}$$

Et sous l'hypothèse nulle, la distribution limite de $\frac{q^0_{1.T}}{T^2}$ converge vers la distribution limite du test d'adéquation de Cramer von Mises (q^0_1) Une fois de plus, pour toute valeur de T, les points critiques sont présentés dans la tableau suivant :

TABLE 2.3 – Valeurs de test de Lombard2

α	0,1	0.075	0.05	0.025	0.01
$q^0_{1.T}$	0,3473	0,3939	0,4614	0,5608	0,7435
$\frac{q^0_{1.T}}{T^2}$	0,3431	0,3870	0,4521	0,5596	0,7022

2.4 Tests de rupture paramétriques

2.4.1 Test de Jarušková (1997)

Le test de Jarušková (1997) [11] est un test paramétrique de type «maximum». Tout comme le test de Lombard (1987), supposons que l'on observe des variables aléatoires

$y_1, ..., y_T$ .L'hypothèse nulle stipule que

$$H_0 : \mu_1 = \mu_2 \tag{2.63}$$

L'hypothèse alternative stipule qu'il existe un point dans le temps $\tau \in \{1, .., T-1\}$ qui fait en sorte que le modèle est :

$$Y_i \rightarrow \begin{array}{l} N(\mu_1; \delta^2) \text{ si } i = 1, ..., \tau \\ N(\mu_2; \delta^2) \text{ si } i = \tau, ..., T \end{array} \tag{2.64}$$

Où $\mu_1 \neq \mu_2$. En supposant δ^2 inconnu, **le test de Jarušková** , noté $J(T)$ est le suivant :

$$J(T) = \max_{1 \leq \tau < T} |J_T| = \max_{1 \leq \tau < T} \frac{1}{s_\tau} \sqrt{\frac{(T-\tau)\tau}{T}} |\bar{y}_\tau - \bar{y}^*_\tau| \tag{2.65}$$

où :

$$\bar{y}_\tau = \frac{\sum_{j=1}^{\tau} y_i}{\tau} \tag{2.66}$$

$$\bar{y}^*_\tau = \frac{\sum_{j=\tau+1}^{T} y_i}{T-\tau}$$

$$s_\tau = \sqrt{\frac{\sum_{j=1}^{\tau} (y_i - \bar{y}_\tau)^2 + \sum_{j=\tau+1}^{T} (y_i - \bar{y}^*_\tau)^2}{T-2}} \tag{2.67}$$

L'hypothèse nulle est rejetée lorsque la statistique $J(T)$ est **plus grande qu'une (supérieur à 1)** certaine valeur critique (voir le tableau suivant). Si **l'hypothèse nulle est vraie**, alors $\bar{y}_\tau$ et $\bar{\mathbf{y}}^*_\tau$ devraient être d'environ la même valeur. Jaruškov, (1997) á obtient celles-ci par simulations. Une statistique tronquée peut aussi être utilisée :

$$J_1(T) = \max_{t_0 T \leq \tau < (1-t_0)T} |J_T| \tag{2.68}$$

Où $t_0 \in (0, 0.5)$ (Jarušková utilise $t_0 = 0.05$). Jarušková (1997) obtient aussi les valeurs critiques par simulations et sont aussi présentées dans le tableau suivant.

pour $\alpha = 0.05$

TABLE 2.4 – Valeurs de test de Jaruskova pour $\alpha = 0.05$

T	50	100	200	300	500
$J(T)$	3.15	3.16	3.19	3.21	3.24
$J_1(T)$	3.08	3.06	3.07	3.08	3.09

pour $\alpha = 0.01$

TABLE 2.5 – Valeurs de test de Jaruskova pour $\alpha = 0.01$

T	50	100	200	300	500
$J(T)$	3.76	3.71	3.72	3.73	3.73
$J_1(T)$	3.69	3.62	3.61	3.62	3.62

2.4.2 Tests de Reeves et al (2007) TPR

Les tests de Reeves et al. (2007)[7] utilisent les **sommes au carré des résidus** et comparent leur statistique de test à une **loi de Fisher** dont les degrés de liberté changent dépendamment du test utilisé. Tout comme les autres tests, supposons que l'on a des variables aléatoires $y_1, ..., y_T$ et que celles-ci, sous l'hypothèse nulle, soient indépendantes et identiquement distribuées tandis que, sous l'hypothèse alternative, **il existe un point tel que la moyenne des variable** $y_1, ..., y_\tau$ **soit différente des variables** $y_{\tau+1}, ..., y_T$

Reeves et al. (2007) proposent les tests (méthodes) **LR** (Modified Lund and Reeves TPR method) et **XLW** (Modified Wang's TPR method) où la méthode TPR (two-phase regression) avec un point de rupture au point τ est simplement :

$$Y_i = \begin{cases} \mu_1 + \beta_1 t_i + \varepsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \beta_2 t_i + \varepsilon_i & \text{si } i = \tau + 1, ..., T \end{cases} \tag{2.69}$$

Ainsi que $t_1 \leq ... \leq t_T$, $\varepsilon_i \rightsquigarrow N(0; \delta^2)$ pour $i = 1, ..., T$ et β_1, β_2, μ_1, μ_2, τ sont inconnus. Il y a cependant une contrainte de continuité de la régression au point τ se traduisant par $\mu_2 = \mu_1 + (\beta_1 - \beta_2)t_\tau$. Les deux tests suivants proposent
quelques modifications au test TPR, tel que l'égalité de β_1 et β_2.

2.4.3 Test Lund et Reeves (2007) (LR)

Lund et Reeves (2007) ont modifié le modèle TPR en cessant d'imposer la contrainte de continuité. Le modèle LR suppose que les t_i sont des entiers. Le modèle LR est donc :

$$Y_i = \left\{ \begin{array}{ll} \mu_1 + \beta_1 t_i + \varepsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \beta_2 t_i + \varepsilon_i & \text{si } i = \tau + 1, ..., T \end{array} \right\} \tag{2.70}$$

où on peut alors détecter une rupture dans la moyenne ($\mu_1 \neq \mu_2$) et dans la tendance ($\beta_1 \neq \beta_2$). Les erreurs suivent une loi normale de moyenne nulle et de variance inconnue $\{\varepsilon_i \rightsquigarrow N(0, \delta^2)\}$. Ce qui revient donc à confronter les hypothèses suivantes :

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \text{ et } \beta_1 = \beta_2 \\ H_1 &: \mu_1 \neq \mu_2 \text{ et/ou } \beta_1 \neq \beta_2 \end{aligned} \tag{2.71}$$

— **Si le point de rupture** τ est connu et fixé, le test devient simplement :

$$F_\tau = \frac{(SSE_0 - SSE_A)/2}{SSE_A/(T-4)} \rightsquigarrow F_{2,T-4} \tag{2.72}$$

Où SSE_0 et SSE_A sont respectivement **les sommes des erreurs au carré** sous l'hypothèse nulle et alternative. Sous H_0, la statistique de test devrait suivre une **loi de Fisher** à $(2, T-4)$ degrés de liberté.

On **rejette** l'hypothèse nulle pour de **grandes valeurs de F**.

— **Si le point de rupture τ est inconnu**, la statistique devient :

$$F_{max} = max_{1 \leq \tau \leq T} F_\tau \tag{2.73}$$

Et on rejette H_0 pour une grande valeur de $F_{\max}$. Comme cette statistique ne suit pas de **loi connue**, Lund et Reeves (2007) obtiennent les valeurs critiques par simulations et sont présentées dans le tableau suivant avec $\alpha = 0.05$:

TABLE 2.6 – Valeurs de Test Lund et Reeves pour $\alpha = 0.05$

T	25	50	75	100	200	500
$LR(F_{\max})$	11.67	11.07	11.06	11.09	11.21	11.54

2.4.4 Test Wang (2007) XLW

Wang (2007)[8] a apporté une petite modification au test LR car il trouvait que celui-ci expliquait mal les phénomènes climatiques. Son modèle devient donc :

$$Y_i = \begin{array}{ll} \mu_1 + \beta_1 t_i + \varepsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \beta_2 t_i + \varepsilon_i & \text{si } i = \tau + 1, ..., T \end{array} \tag{2.74}$$

où les termes sont définis précédemment. Les hypothèses deviennent donc :

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned} \tag{2.75}$$

— **Si le point de rupture τ est connu et fixé**, le test devient simplement :

$$F_\tau = \frac{(SSE_0 - SSE_A)}{SSE_A/(T-3)} \rightsquigarrow F_{1,T-3} \tag{2.76}$$

où les termes sont aussi définis précédemment.

— **Si le point de rupture τ est inconnu** la statistique $F_{\max}$ est une fois de plus utilisée. Wang (2007) obtient aussi les valeurs critiques par simulations et sont présentées dans le tableau suivant avec $\alpha = 0.05$:

TABLE 2.7 – Valeurs de Test Wang pour $\alpha = 0.05$

T	25	50	75	100	200	500
$XLW(F_{\max})$	7.37	6.92	6.88	6.91	7.01	7.24

2.5 Méthode bayésienne de Lee et Heghinian

La méthode bayésienne de Lee et Heghinian ne s'exprime pas comme un test statistique classique. Il s'agit plutôt d'une approche paramétrique qui demande une distribution normale des variables étudiées. L'interprétation de cette méthode vise à confirmer ou infirmer l'hypothèse d'un changement de moyenne dans la série chronologique. Le modèle de base de la procédure est le suivant :

$$X_j = \mu + \varepsilon_j, j = 1\tau\mu + \delta + \varepsilon_j, j = \tau + 1, ..., N \tag{2.77}$$

où les ε_j sont des variables indépendants et normalement distribués, de moyenne nulle et de variance σ, τ et δ représentent respectivement la position dans le temps et l'amplitude d'un changement éventuel de moyenne.

Lee et Heghinian ont posé les distributions à priori des paramètres $\tau, \delta, \mu, \sigma$ comme suit :

Notation 1

1. $P(\tau) = \frac{1}{N-1}; \tau = 1, 2, ..., N-1$.
2. *$P(\delta)$ est une probabilité qui suit la loi normale dont la variance est σ_δ*
3. *$P(\mu)$ est une probabilité qui suit la loi normale dont la variance σ_μ*
4. *$p(\sigma) = \sigma^{-1}$: est la probabilité de l'écart type inversement proportionnelle à celui-ci, ces paramètres sont statistiquement indépendants.*

Le vecteur $X = [X_1, X_2, ..., X_N]^t$ a la forme d'une probabilité normale avec une moyenne δc_τ avec :

c_τ : est un vecteur identité lequel les premières τ composantes sont nulles et les derniers $N - \tau$ composantes sont des identités, et une matrice de covariance M de la forme suivante :

$$M = \sigma \left[I_n + \left(\frac{\sigma_\mu}{\sigma} \right) 1_n 1_n^t \right] \tag{2.78}$$

avec : 1_n est un vecteur à n dimension dont les éléments sont des unités et I_n est une matrice identité d'ordre n.

Et on a la matrice inverse qui est M^{-1} :

$$M^{-1} = \sigma^{-2} \left[I_n - (1+r)^{-1} \frac{1_n 1_n^T}{n} \right] \tag{2.79}$$

avec :

$$r = \frac{(\sigma/n)}{\sigma_\mu} \tag{2.80}$$

et le déterminant qui est :

$$|M| = n\sigma^{2(n-1)}\sigma_{\mu}(1+r) \tag{2.81}$$

Si p la **densité de probabilité** de τ est uniforme, elle est alors définie par :

$$P\left(\frac{\tau}{X}\right) = \left[\frac{n}{\tau(n-\tau)}\right]^{\frac{1}{2}} [R(\tau)]^{-\left(\frac{n-2}{2}\right)} \tag{2.82}$$

et $0 \leq \tau \leq n-1$ avec :

$$R(\tau) = \frac{\left[\sum_{i=1}^{\tau}(x_i - \overline{x_\tau}) + \sum_{i=\tau+1}^{n}(x_i - \overline{x_{n-\tau}})\right]}{\sum_{i=1}^{N}(x_i - \overline{x_n})} \tag{2.83}$$

$$\begin{aligned} \overline{x_n} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ \overline{x_\tau} &= \frac{\sum_{i=1}^{\tau} x_i}{\tau} \\ \overline{x_{n-\tau}} &= \frac{\sum_{i=\tau+1}^{N} x_i}{(N-\tau)} \end{aligned} \tag{2.84}$$

Donc la méthode établit la distribution de probabilité a posteriori de la position dans le temps d'un changement. Le mode de la distribution à posteriori de τ sert à estimer la date de la rupture, lorsque la distribution est unimodale. Plus la dispersion de la distribution est faible, plus l'estimation est précise. La méthode bayésienne de Lee et Heghinian impose normalité, constance de la variance et non-autocorrélation.

2.6 Procédure de segmentation de Pierre Hubert

La méthode de Pierre Hubert est basée sur une procédure originale de segmentation des séries hydrométéorologiques qui représentent le plus souvent la température, les précipitations, et le débit des grands lacs. Cette méthode permet de mettre en évidence le problème de la stationnarité et de la modélisation des séries hydrométéorologiques. Pierre Hubert a testé sa méthode de segmentation en modélisant de courtes séries empiriques.

Il a utilisé une procédure de segmentation lui permettant de partitionner une série hydrométéorologique donnée en m segments de façon à ce que la moyenne de chaque segment soit nettement différente de celle du segment adjacent.La segmentation de Pierre Hubert est acceptée lorsque l'écart quadratique entre celle-ci et de la série est mini-

male. La méthode de ce dernier augmente la capacité de rechercher les changements des moyennes des segments des séries hydrométéorologiques.

Supposons que l'on ait une série chronologique de n valeurs numériques $X_i = 1, 2, ..., n$. Un segment donnée extrait de la série initiale X_i est désigné par $X_i = i_1, i_2$ avec $1 \leq i_1 < i_2 \leq n$.

La partition de la série initiale en m segments représente une segmentation d'ordre m. Pierre Hubert a défini $i_{k=1,2,...,m}$ comme étant le rang de la série initiale de l'extrémité du kième segment avec :

$$i_0 = 0 < i_1 < ... < i_k < ... < i_{m-1} < i_m < n \tag{2.85}$$

On désigne par n_k la longueur du kième segment défini comme suit :

$$n_k = i_k - i_{k-1} \tag{2.86}$$

Sa moyenne locale est exprimée par :

$$\overline{x_N} = \frac{\sum_{i=i_{k-1}+1}^{i=i_k} x_i}{n_k} \tag{2.87}$$

On définit les deux quantités suivantes :

$$d_k = \sum_{i=i_{k-1}+1}^{i=i_k} (x_i - \overline{x_k})^2 \tag{2.88}$$

$$D_m = \sum_{k=1}^{K=m} d_k \tag{2.89}$$

Avec D_m représentant l'écart quadratique entre la série et la segmentation considérée. Le test de Scheffé s'applique pour vérifier si les moyennes de deux segments successifs sont nettement différentes. Il a défini ψ_s comme étant le contraste du segment initial tel que :

$$\psi_s = \sum_{k=1}^{K=m} c_k \mu_k \tag{2.90}$$

Pierre Hubert estime qu'une segmentation d'ordre m n'est acceptable que si tous les contrastes définis par :

$$\psi_{s,k} = \overline{x_k} - \overline{x}_{k-1}, k = 1, 2, ..., m-1. \tag{2.91}$$

Sont différents de zéro, au niveau de signification du test de Scheffé.

CHAPITRE 3

SIMULATIONS ET APPLICATIONS

3.1 Introduction

Après l'étude théorique des séries chronologiques et les tests de rupture dans les chapitres 1 et 2 respectivement.

Dans ce chapitre, on effectué une étude pratique et une simulation des modèles d'un série chronologique, tel que on traite la tendance et la saisonnalité de la température de l'atmosphère au niveau de la région de SOUMAA depuis 2002 jusqu'à 2015. Un traitement à l'aide du logiciel R est exécuter sur une série de 5000 valeurs de températures relevées au pare avant, afin d'extraire l'irrégularité et la rupture au cours de ces années.

3.2 Série temporelle avec R

3.2.1 Présentation du logicioel R

Définition 3.1 *R est un logiciel et un langage de programmation gratuits et à source ouverte développés en 1995 à l'Université d'Auckland en tant qu'environnement informatique statistique et graphique (Ikaha et Gentleman, 1996). Depuis lors, R est devenu l'un des environnements logiciels dominants pour l'analyse de données et est utilisé par diverses disciplines scientifiques, notamment les sciences du sol, l'écologie et la géoinformatique (vue Environmetrics CRAN; vue Spatial CRAN). R est particulièrement populaire pour ses fonctionnalités graphiques, mais il est également prisé pour ses fonctionnalités SIG, qui facilitent relativement la génération de modèles à base de raster. Plus récemment, R a également acquis plusieurs packages spécialement conçus pour l'analyse des données de sol.*

— un environnement logiciel : statistiques, graphique,la programmation,calculatrice,SIG,etc...
— un langage pour explorer, résumer et modéliser des données
fonctions = verbes
objets = noms

3.2.1.1 Origines

Le logiciel R est un logiciel de statistique créé par **Ross Ihaka** & **Robert Gentleman**. Il est à la fois un langage informatique et un environnement de travail : les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple, les résultats sont affichés sous forme de texte et les graphiques sont visualisés directement dans une fenêtre qui leur est propre. C'est un clone du logiciel **S-plus** qui est fondé sur le langage de programmation orienté objet S, développé par **AT&T Bell** Laboratories en 1988. Ce logiciel sert à manipuler des données, à tracer des graphiques et à faire des analyses statistiques sur ces données.

3.2.2 Pourquoi utiliser R ?

Tout d'abord R est un logiciel gratuit et à code source ouvert (opensource). Il fonctionne sous **UNIX** (et Linux), **Windows** et **Macintosh**. C'est donc un logiciel multi-plates-formes. Il est développé dans la mouvance des logiciels libres par une communauté sans cesse plus vaste de bénévoles motivés.

Tout le monde peut d'ailleurs contribuer à son amélioration en y intégrant de nouvelles fonctionnalités ou méthodes d'analyse non encore implémentées. Cela en fait donc un logiciel en rapide et constante évolution.

Le logiciel R est particuliérement performant pour la manipulation de données, le calcul et l'affichage de graphiques. Il posséde, entre autres choses :

- un systéme de documentation intégré trés bien conçu (en anglais).
- des procédures efficaces de traitement des données et des capacités de stockage de ces données.
- une suite d'opérateurs pour des calculs sur des tableaux et en particulier sur des matrices.
- une vaste et cohérente collection de procédures statistiques pour l'analyse de données.
- des capacités graphiques évoluées.
- un langage de programmation simple et efficace intégrant les conditions, les boucles, la récursivité, et des possibilités d'entrée-sortie.

Remarque 3.1 *R est plus difficile d'accés que certains autres logiciels du marché. Il faut passer du temps à en apprendre la syntaxe et les commandes.*

3.2.3 R et les statistiques

R est un logiciel dans lequel de nombreuses techniques statistiques modernes et classiques ont été implémentées. Les méthodes les plus courantes permettant de réaliser une analyse statistique telles que :

- statistique descriptive.

— tests d'hypothéses.
— analyse de la variance
— méthodes de régression linéaire (simple et multiple).
— etc...

Sont enchâssées directement dans le coeur du systéme. Notez egalement que la plupart des méthodes avancées de statistique sont aussi disponibles au travers de modules externes appelés **packages**. Ceux-ci sont faciles à installer directement à partir d'un menu du logiciel. Ils sont tous regroupés sur le site internet du Comprehensive R Archive Network (CRAN) (http ://cran.r-project.org) sur lequel vous pouvez les consulter. Ce site fournit aussi, pour certains

grands domaines d' etude, une liste commentée des packages associés à ces thémes (appelée Task View), ce qui facilite ainsi la recherche d'une méthode statistique particuliére. Par ailleurs, une documentation d etaill ee en anglais de

chaque package est disponible sur le CRAN. Il est par ailleurs utile de noter que les méthodes statistiques les plus récentes y sont réguliérement ajoutées par la communauté statistique elle-même.

3.2.4 R et les graphiques

Une des grandes forces de R réside dans ses capacités, bien supérieures à celles des autres logiciels courants du marché,à combiner un langage de programmation avec la possibilité de réaliser des graphiques de qualité. Les graphiques usuels s'obtiennent aisément au moyen de fonctions prédéfinies. Ces derniéres possédent de trés nombreux paramétres permettant par exemple d'ajouter des titres, des légendes, des couleurs, etc. Mais il est également possible d'effectuer des graphiques plus sophistiqués permettant de représenter des données complexes telles que des courbes de surface ou de niveau, des volumes affichés avec un effet 3D, des courbes de densité, et bien d'autres choses encore.

Il vous est également possible d'y ajouter des formules mathématiques. Vous pouvez aussi agencer ou superposer plusieurs graphiques sur une même fenêtre,et utiliser de nombreuses palettes de couleur.

Vous pouvez obtenir une démonstration des possibilités graphiques de R en tapant les commande suivantes :

— *demo(graphics)*
— *demo(persp)*
— demo(plotmath)

3.3 Les structures de séries temporelles dans R

3.3.1 La fonction ts()

La fonction ts() fait partie de stats, chargé au lancement de R. Elle permet de créer des séries temporelles à temps régulier.**ts** fournit la syntaxe et de très utiles exemples. Illustrons la fonction en fabriquant différentes séries temporelles.

Exemple 3.1 *(**Série temporelle mensuelle**) Fabriquons une série mensuelle supposée commencer en février 2005 à partir d'un vecteur x obtenu par simulation de 30 valeurs indépendantes d'une loi normale de moyenne 2, de variance 1,et dont on ne conserve que 3 décimales. L'unité de temps est l'année et il y a 12 observations par unité de temps. Février est la deuxième observation dans une unité de temps, d'où la syntaxe*

3.4 Tendance et ruptures

3.4.1 Objective

Ce modèle permet d'appliquer des tests statistiques de détection de tendance et de rupture sur les chroniques.

En fonction des caractéristiques des chroniques (nombre d'analyses, normalité de la distribution, autocorrélation....) l'outil détermine automatiquement les tests les plus robustes à appliquer. L'illustration 29 présente l'arbre décisionnel appliqué dans l'outil HYPE.

Dans ce module, les résultats d'analyses reportés inférieurs à une limite (de quantification ou de détection) sont substitués par la valeur correspondante de la limite (de quantification ou de détection) sont substitués par la valeur correspondante de la limite de quantification ou de direction. Il convient alors d'interpréter avec grand prudence les chroniques présentant des taux de quantification faible.

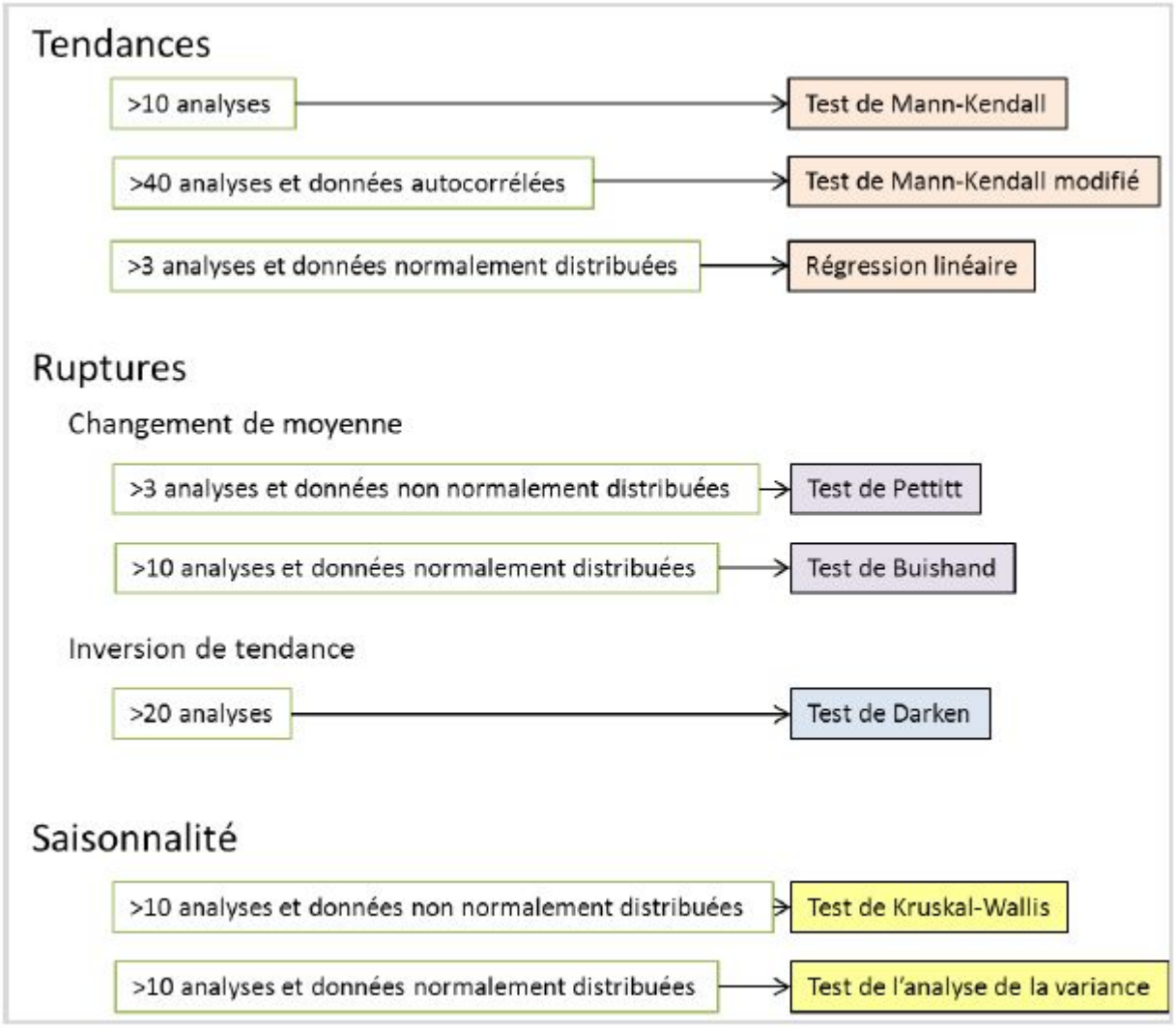

FIGURE 3.1 – Récapitulatif des crétéres seléction automatique des tests.

3.4.2 Recherche de tendance

Deux tests de tendance peuvent être appliqués. Dans le cas où les données sont normalement distribuées, une régression linéaire Est appliquée si la chronique comporte au moins 3 données et un test de Mann-Kendall est également appliqué si la chronique comporte au moins 10 données. Dans le cas ou ne sont pas normalement distribuées, seul le test de Mann-Kendall est applique si la chronique comporte au moins 10 données.

De plus, si les données présentent une autocorrélation significative et si la chronique dépose d'au moins 40 données un test de Mann-Kendall modifier est appliqué.la p-value de ce test est différente de celle de test non modifie ; elle tient compte de l'autocorrélation. En ce qui concerne le pont de sen, ce test est équivalent de test de Mann-Kendall non modifier elle n'est don recalculer.

3.4.3 Recherche de ruptures

Deux types de ruptures sont recherchés dans les chroniques : la présence d'un changement significatif de moyenne est recherchée à l'aide d'un test d'homogénéité :

Test de Buishand si les données sont normalement distribuées.

Test de Pettitt si les données ne sont pas normalement distribuées.

La présence d'une inversion de tendance est également recherchée. Pour cela une méthode tirée des travaux de Darken est appliquée si la chronique dispose d'au moins 20 données.

3.5 SIMULATION

3.5.1 Package"changepoint"en 4 October 2016

3.5.1.1 Description

Implémente diverses méthodes de point de changement classiques et spécialisées pour la recherche d'un ou de plusieurs points de changement dans les données. De nombreuses méthodes populaires non paramétriques et fréquentistes sont incluses.

Les fonctions cpt.mean (), cpt.var (), cpt.meanvar () devraient être votre premier point d'appel.

3.5.1.2 Information

Type Package
Title Methods for Changepoint Detection
Version 2.2.2
Date 2016-10-04
Maintainer Rebecca Killic : r.killick@lancs.ac.uk
BugReports : changepoints@lancs.ac.uk
URL https ://github.com/rkillick/changepoint//
Dépôt CRAN.

3.5.1.3 Test de changment de variance sur R

```
set.seed(1)
x=c(rnorm(100,0,1),rnorm(100,0,10))
ansvar=cpt.var(x)
plot(ansvar)
print(ansvar) # identifies 1 changepoint at 100
```

l'execution : Dans ce cas en changer la variance telle que les variables aléatoires de premier échantillion suit la loi normale (0,1) et l'autre suit loi normale(0,10).

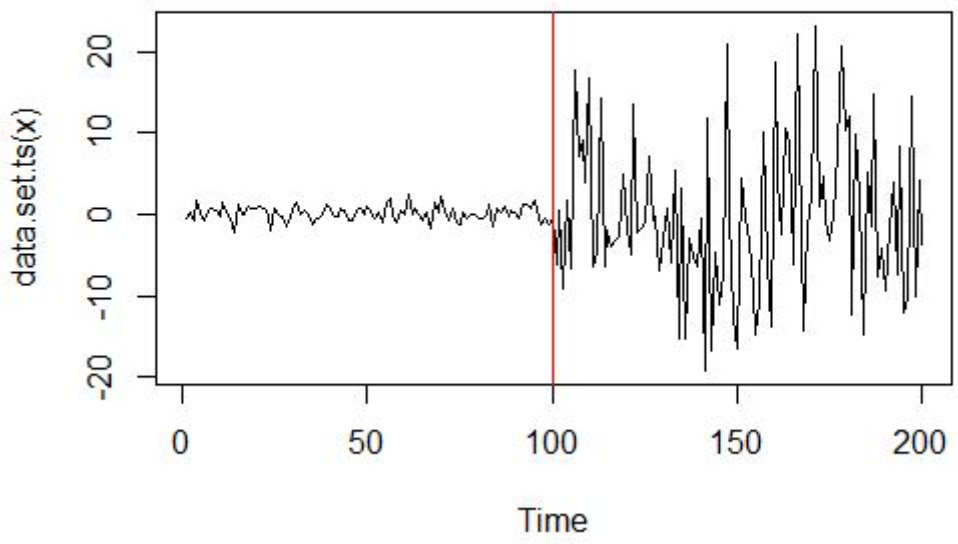

FIGURE 3.2 – L'execution de Test de changment de variance sur R.

3.5.1.4 Test de changment de moyenne sur R

y=c(rnorm(100,0,1),rnorm(100,5,1))
ansmean=cpt.mean(y)
plot(ansmean,cpt.col='blue')
print(ansmean)

l'execution : Dans ce cas en changer la moyenne telle que les variables aléatoires de premier échantillion suit la loi normale (0,1) et l'autre suit loi normale(5,1).

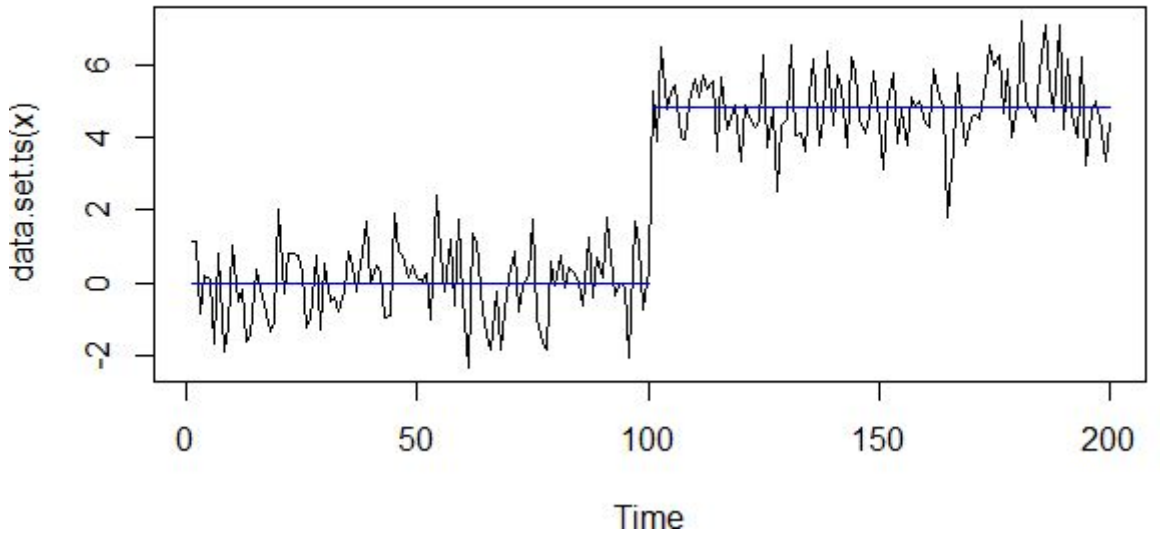

FIGURE 3.3 – L'execution de Test de changment de moyenne sur R.

Remarque 3.2 *On remarque aux milieux de la série il y a une rupture.*

3.5.1.5 Test de changement de moyenne et variance sur R

z=c(rnorm(100,0,1),rnorm(100,2,10))
ansmeanvar=cpt.meanvar(z)
plot(ansmeanvar,cpt.width=3)
print(ansmeanvar)

l'exécution : Dans ce cas en changer la moyenne et la variance telle que les variables aléatoires de premier échantillion suit la loi normale (0,1) et l'autre suit loi normale(2, 10).

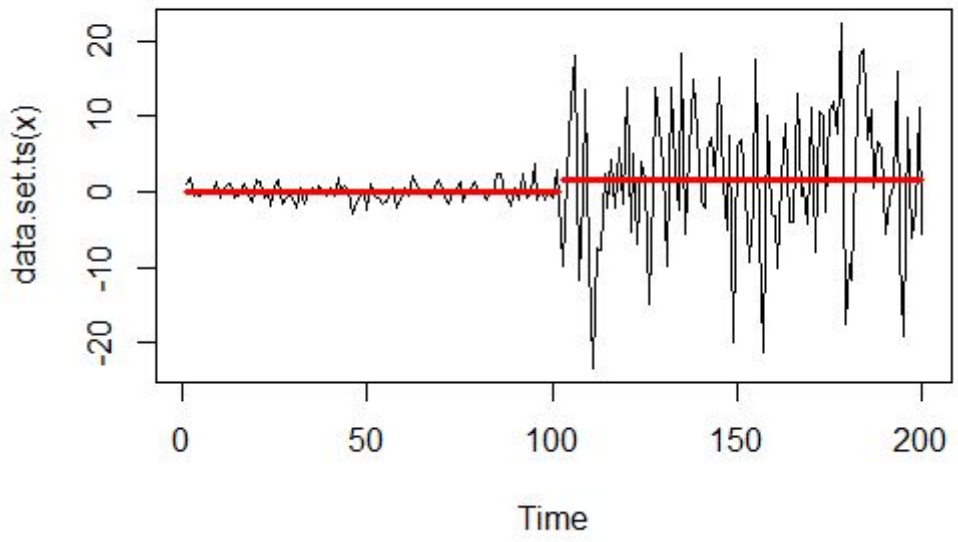

FIGURE 3.4 – Le graphe du changement de la moyenne et la variance.

Remarque 3.3 *On remarque aux milieux de la série il y a une rupture.*

3.5.2 Simulation des modèles

3.5.2.1 Modèle stationnaire ARMA

Dans la suite, on simulant certains processus ARMA, puis on estimant les différents paramètres l'identifiant (paramètres qui

sont les coefficients des 2 polynômes P et Q) et son ordre. $X = arma.sim(100, model = list(ar = -.3, ma = .7))$ la forme suivent pour simuler un processus ARMA.

3.5.2.2 Modèle stationnaire ARIMA

ord=c(1,0,1) l'ordre supposé de ce modéle.

coef=arima(X,ord)

var.coef, nous à donn matrice suivante :

Table 3.1 – Estimation d'un modèle ARMA sous logiciel R

	$ar1$	$ma1$	$intercept$
$ar1$	0.0596010884	−0.0453351484	0.0008804861
$ma1$	−0.0453351484	0.0410339004	−0.0007626542
$intercept$	0.0008804861	−0.0007626542	0.0144308540

acf(X) : donne la fonction auto-correlation de modéle(ACF)

pacf(X) : donne la fonction auto-correlation partielle de modéle (PACF)

tsdiag(coef)

MA=acf(X) : Donne ACF.

AR=pacf(X) : Donne PACF.

paramAR=ar(X)

paramAR. Call : ar(x=X)

Coefficients : 0.3966, -0.2243

Order selected 2 σ^2 estimated as 0.8339

mk.test(X) : application du test de Mann-Kendall sur le modèle

data : X

Table 3.2 – Application du test de Mann-Kendall sur le modèle ARMA

Mann-Kendall trend test	z	n	p-value
valeur	-0.84876	100	0.396

alternative hypothesis : true S is not equal to 0

sample estimates :

S varS tau

-2.860000e+02 1.127500e+05 -5.777778e-02

Après utilisation de la commend fit=auto.arima(T2), on trouvons les coefficients de modèle ARIMA (0,1,1), et leur AIC=1199.26.

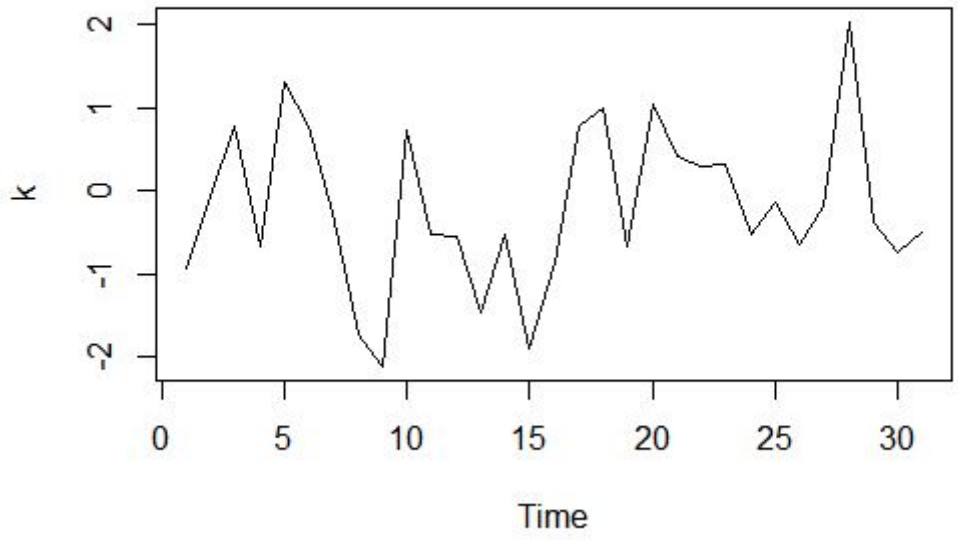

FIGURE 3.5 – Représentation graphique de modèle ARMA .

3.5.2.3 Modèle stochastique non linéaire ARCH et GARCH

v=rnorm(500) : un échantillon normale de taille 500 plot.ts(v) : donneé le plot de densité v.

acf(v).la fonction ACF de v. pacf(v).la fonction PACF de v.

arch.x=garch(v,order=c(0,1))

summary(arch.x).Call

garch(x = v, order = c(1, 1)

GARCH(1,1)

Residuals :

TABLE 3.3 – Résumé d'applique le modèle garch

Min	1Q	Median	3Q	Max
-2.97439	-0.64191	-0.06434	0.66465	3.76764

Coefficient(s) :

TABLE 3.4 – Coefficient de modèle garch(1 ;1)

	Estimate	Std	Error t	value Pr(> |t|)
a0	9.585e-01	6.264e+01	0.015	0.988
a1	8.176e-14	4.665e-02	0.000	1.000
b1	6.283e-02	6.125e+01	0.001	0.999

Diagnostic Tests : Jarque Bera Test

data : Residuals

X-squared = 0.34287, df = 2, p-value = 0.8425

Box-Ljung test

data : Squared.Residuals

X-squared = 0.37639, df = 1, p-value = 0.5395

3.5.3 Test AIC du modèle

On a estimé la température max. Appliquer le modèle arch et le modèle garch.

— Après utilisation de la commanda fit, nous trouvons les coefficients de modèle Arch (0,1), et leur AIC=1443.971.

— Après utilisation de la commande fit, on trouvons les coefficients de modèle Garch (1,1), et leur AIC=1445.844.

TABLE 3.5 – AIC de modèles

modèle	arma	arch	garch
AIC	1199.26	1443.971	1445.844

On remarque que $AIC(arma) < AIC(arch) < AIC(garch)$. Donc le meilleur modéle est ARMA.

3.6 APPLICATION

3.6.1 Introduction

Dans le contexte actuel de changements climatiques et environnementaux, la tendance des paramètres météorologiques, notamment celle de la température, est souvent discutée.

Alors que les évènements de forte chaleur sont bien moins étudiés en domaine tropical qu'en domaine tempéré, cette région connaît régulièrement ces aléas, à l'instar des fortes chaleurs enregistrées à SOUMAA wilaya de Blida Algérie. En 01 septembre 2002 jusqu'à 31 décembre 2015.

3.6.2 Données et méthodes statistiques

3.6.2.1 Résumé des tableaux

Résume du tabluax : Notre tableaux de donnéé représente les degrés de température dans la région de Soumaa de Blida en 01/09/2002 au 31/12/2015.

TABLE 3.6 – Paramètres discriptives des données

summary	min	Q1	Median	Mean	Q3	Max
T1(moy)	0.00	13.20	18.10	18.60	24.10	74.10
T2(max)	0.00	18.00	24.10	24.46	30.70	77.20
T3(min)	-29.60	8.80	13.10	13.49	18.30	68.10

3.6.2.2 Températures de l'air

Les données utilisées sont des observations quotidiennes de température minimale (min),maximale (max) et moyenne (moy) de l'air à Soumaa (2002-2015). Ces données météorologiques ont été fournies par la Direction de la Météorologie Algérien. Cette figure donnée la boîte à moustaches des valeurs qui présenter leur minimum, maximum et moyenne.

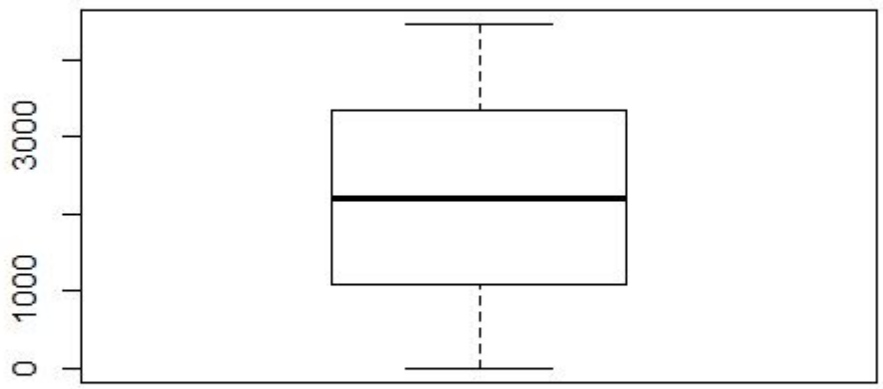

FIGURE 3.6 – Boîte à moustaches des valeurs données.

3.6.2.3 Graphes des fonctions ACF et PACF

La fonction générique pour tracer des objets R de séries temporelles. Ou bien utilisé la commende "plot.ts(donnees)" en a donnée même résultat(même plot).

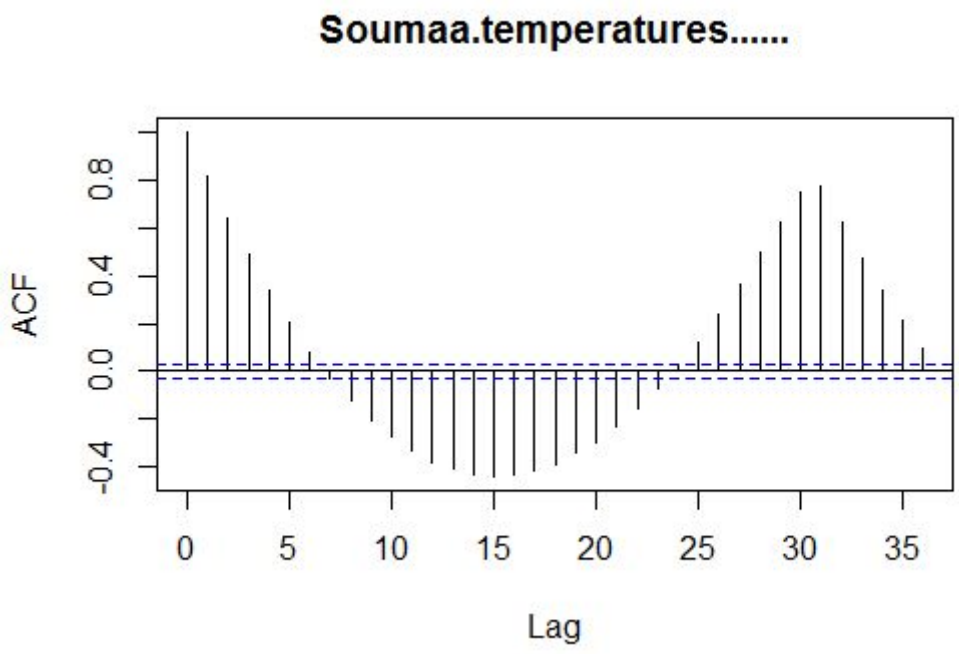

FIGURE 3.7 – ACF des données.

La fonction d'autocorrélation (La fonction acf calcul (et par défaut, les graphiques) des estimations de la fonction d'autovariance ou d'autocorrélation). On remarque que la série n'est pas stationnaire d'où on passe à l'autocorrélation partiel.

La figure suivante représente la fonction d'autocorrélation partiel.

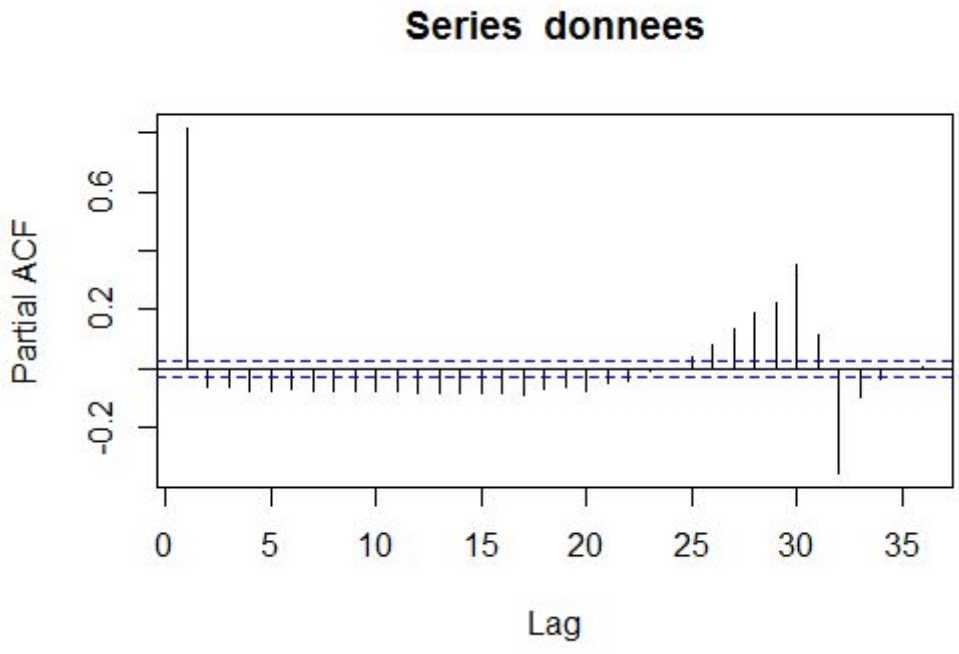

FIGURE 3.8 – PACF des données.

3.6.3 Application de test Mann-Kendall avec le package trend sur température de SOUMAA

3.6.3.1 Description de tendance

L'analyse des données environnementales nécessite souvent la détection des tendances et des points de changement.

Ce paquet comprend des tests pour la détection de tendance (Test de tendance de Cox-Stuart, Test de tendance de Mann-Kendall, (corrélé) test de Hirsch-Slack, test partiel de tendance de Mann-Kendall, multivarié (multisite) Test de tendance de Mann-Kendall, pente de Sen (saisonnière), test de tendance de corrélation partielle de Pearson et Spearman), détection de point de changement (procédures de test de Lanzante, Test de Pettitt, test de Buishand Range, Test de Buishand U, test d'homogénéité normale normale), détection du caractère non aléatoire (test de fréquence de phase de Wallis-Moore, Bartels classe le test de von Neumann (test de Wald-Wolfowitz)

3.6.3.2 Sur les températures maximales

mk.test(data)

data : $XZ\$max$ lire notre vecteur max.

z	n	p-value
2.8372	4468	0.004552

alternative hypothesis : true S is not equal to 0.

$p-value 0.2525 \neq 0$ donc on accepte l'hypothèse $H1$ d'où existe une tendance dans notre série de max. et nous avons la tendance n'est pas significatif car $p-value > 0.05$.

sample estimates :

S	varS	tau
$2.824900e^5$	9.913686e+09	2.835846e-02

3.6.3.3 Sur les températures moyenne

data : XZ$moy

z	n	p-value
1.1442	n = 4468	p-value = 0.2525

alternative hypothesis : true S is not equal to 0. $p-value = 0.2525 \neq 0$ donc on accepte l'hypothèse $H1$ d'où existe une tendance dans notre série de max et nous avons la tendance n'est pas significatif car $p-value > 0.05$.

sample estimates :

S	varS	tau
$1.139320e^5$	$9.913856e^9$	$1.141789e^{-2}$

3.6.4 Application test de Pettitt avec le package trend sur température de SOUMAA

3.6.4.1 Sur les températures moyenne

pettitt.test(XZ$moy)

Pettitt's test for single change-point detection

data : XZ$moy

TABLE 3.7 – Résultats du test de Pettitt

U*	p-value
365760	0.0002475

alternative hypothesis : two.sided

sample estimates :

probable change point at time K 2423

$$p_K = 2\exp\left[-\frac{6K^2}{(n^3+n^2)}\right] = 0,1999$$

on à $\alpha > p_K$ d'où H_0 est rejetée, la série présente alors une rupture, la tendance est significatif car $p-value < 0.05$.

3.6.4.2 Sur les températures max

Pettitt's test for single change-point detection

data : XZ$max

U*	p-value
469970	7.076e-07

alternative hypothesis : two.sided

sample estimates :

probable change point at time K 2423

$$p_K = 2\exp\left[-\frac{6K^2}{(n^3+n^2)}\right] = 0,1999$$

On a $\alpha > p_K$ d'où H_0 est rejetée. la série présente alors une rupture, la tendance est significatif car $p-value < 0.05$.

3.6.5 Représentation graphique des températures moy :

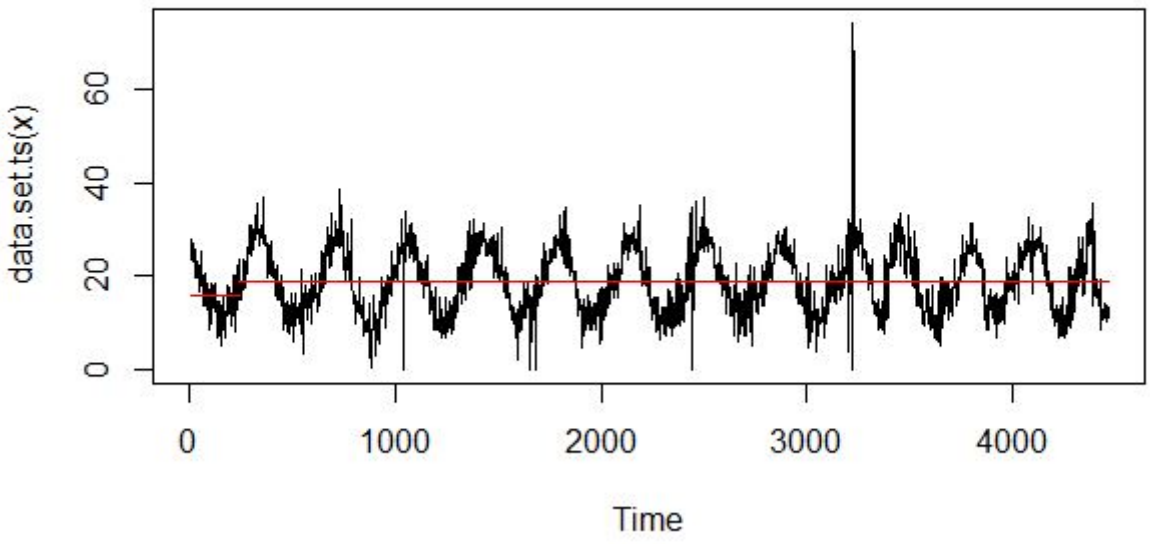

FIGURE 3.9 – Représentation graphique des températures moyenne.

3.6.6 Conclusion :

Aprés l'étude de la série avec les tests de rupture, nous avons découvert une rupture à l'observation 365 de date 30/08/2003, cette rupture est expliquée par une augmentation de la valeur de la température moyenne annuelle.

CONCLUSION GÉNÉRALE

L'importance des tests de rupture par est illustrés par les résultats que nous avons aboutis à travers des données réelles qui sont des séries statistiques qui mesurent de la température dans la région de SOUMAA depuis 2002 jusqu'à 2015. L'analyse des séries chronologiques est un outil couramment utilisé de nos jours pour la prédiction de données futures.

Premièrement, une recherche bibliographique a été établie sur les notions de bases sur les séries chronologiques, tel que nous avons étudié les fonctions de base (acf, pacf, autcorrelation, saisoniéries temporelles et ses modèles comme ARMA,ARIMA,AR,ARCH...), ainsi en deuxième partie nous pésentons les différents tests des ruptures des séries chronologiques, notamment les tests de rupture paramétrique et non paramétrique.

Dans la troisième partie, nous avons manipuler les séries chronologiques sous le logiciel statistique R et réaliser des simulation des séries sous différents modèles existent, par la suite nous avons démontré l'existence d'une rupture dans une série chronologique, où découvert une rupture à l'observation 365 de date 30/08/2003, cette rupture est expliquée par une augmentation de la valeur de la température moyenne annuelle.

BIBLIOGRAPHIE

[1] Bayley, G. V., Hammersley, J. M. (1946). The" effective" number of independent observations in an autocorrelated time series. Supplement to the Journal of the Royal Statistical Society, 8(2), 184-197.

[2] Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American statistical association, 63(324), 1379-1389.

[3] Kendall, M.G., (1955). Rank Correlation Methods. New York : Hafner Publishing Co.

[4] Zetterqvist, L. (1991). Statistical estimation and interpretation of trends in water quality time series. Water Resources Research, 27(7), 1637-1648.

[5] Bohlin, T. (1976). Four cases of identification of changing systems. In Mathematics in Science and Engineering (Vol. 126, pp. 441-518). Elsevier.

[6] ABDRABBO, N. A. et PRIESTLEY, M. B. On the prediction of non-stationary processes. Journal of the Royal Statistical Society : Series B (Methodological), 1967, vol. 29, no 3, p. 570-585.

[7] Reeves, G. K., Pirie, K., Beral, V., Green, J., Spencer, E., & Bull, D. (2007). Cancer incidence and mortality in relation to body mass index in the Million Women Study : cohort study. Bmj, 335(7630), 1134.

[8] Reeves, J., Chen, J., Wang, X. L., Lund, R., & Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. Journal of applied meteorology and climatology, 46(6), 900-915.

[9] Priestley, M. B., & Rao, T. S. (1969). A test for non-stationarity of time-series. Journal of the Royal Statistical Society : Series B (Methodological), 31(1), 140-149.

[10] Lombard, F. (1987). Rank tests for changepoint problems. Biometrika, 74(3), 615-624.

[11] Jarušková, D. (1997). Some problems with application of change-point detection methods to environmental data. Environmetrics : The official journal of the International Environmetrics Society, 8(5), 469-483.

[12] Grewal, M. S. (2011). Kalman filtering (pp. 705-708). Springer Berlin Heidelberg.

[13] Khaled, H., Hamed, A., (1998) A modified Mann-Kendall trend test for autocorrelated data. Journal of Hydrology

[14] Khale H. Hamed, 2007 Trend detection in hydrologic data : The Mann-Kendall trend test under the scaling hypothesis

[15] Yves Aragon, 2012 Séries temporelles avec R. Méthodes et cas.

[16] URL https ://github.com/rkillick/changepoint/

[17] Package 'sac'.February 20, 2015

[18] Thorsten Pohlert, 2018 Non-Parametric Trend Tests and Change-Point Detection .July 30, 2018