République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Mouloud Mammeri, Tizi-Ouzou

Faculté des Sciences
Département de Mathématiques

Mémoire de fin d'étude

En vue de l'obtention du Diplôme de Master Professionnel en Mathématiques Appliquées à la Gestion

Thème

Etude et modélisation d'une évolution d'une pathologie.

Membres du jury :

Réalisé par :

Président :	Pr : AIDEN	KESRAOUI Mahrez
Examinateur :	Pr : OUKACHA	NZISABIRA Cléophas
Promoteurs :	Dr : TALEB et Dr ISSIAKHEM	

Promotion : 2016/2017

Remerciements

Le plus important dans la vie d'un homme est d'être reconnaissant et son plus grand malheur est de n'être utile à personne. Il n'est pas de notre habitude d'être ingrats. Il nous est donc agréable d'adresser nos premiers remerciements au Dieu Tout-Puissant, Lui qui nous a donné la volonté et courage pour la réalisation de ce modeste travail. Après, nos remerciements vont à l'endroit de nos encadreurs, Dr TALEB Youcef, Dr Issiakhem et Pr Toudeft, eux qui n'ont ménagé aucun effort pour nous encadrer, nous guider, nous aider dans ce travail, ils ont été toujours à nos côtés. Nos remerciements vont à l'endroit de l'équipe du service d'épidémiologie au CHU de Tizi Ouzou. Nous leur disons merci surtout pour la confiance et l'attention qu'ils nous ont accordées. Que l'examinateur trouve ici l'expression de nos sincères remerciements pour l'honneur qu'il nous fait en acceptant de juger notre travail. Nos remerciements vont aussi à l'endroit des enseignants qui nous ont enseignés. Nous remercions également tous nos amis et connaissances et tous ceux et celles qui, de loin ou de près nous ont aidés et/ou soutenus au cours de ce travail.

Dédicace

Je dédie ce modeste travail à ....

Mes très chers parents

A la mémoire de ma défunte maman, à la plus belle créature que Dieu a créée sur terre, reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments et mon éternelle gratitude.

A mon père qui peut être fier et trouver ici le résultat de longues années de sacrifices.Merci pour les valeurs nobles, l'éducation et le soutient permanent venu de toi.

Mes très chers frères et sœurs(*Younes,Lila,Taoues,Ghani,Ghenima,Rachid,Kahina,Slimane et Smail*)

Qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité. Dieu vous garde pour moi.

Ma chère Chanez

Toi qui n'as cessée de croire en moi et de m'encourager et qui as toujours été à mes cotés.

A mon binôme NZISABIRA Cleophas

A nos encadreurs Dr TALEB Youcef et Dr ISSIAKHEM

Au Professeur Toudeft, la responsable du service d'épimédiologie au CHU de Tizi Ouzou

A mes amis et camarades de classe(*Salim, Mounir, Hocine, Samia, Massi, Nawel, Chahinez, Dihia, Azedine, Zoher, Bachir, Youcef, Amar, Isguem et Smail...*)

Je vous dédie ce travail en témoignage de mon profond amour. Puisse Dieu, vous préserver et vous accorder santé, longue vie et bonheur.

Mahrez KESRAOUI

Je dédie ce mémoire à ....

A mes très chers parents

Affables, honorables, aimables : vous représentez pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager

Aucune dédicace ne saurait être assez éloquente pour exprimer ce que vous méritez pour tous les sacrifices que vous n'avez cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte.

A mes très chers frères et sœurs

Vous avez toujours été présents dans les bons moments
comme dans les durs, on a partagé de merveilleux moments.
Votre affection et votre soutien ont fait ma force et,
m'ont encouragé toujours à aller vers l'avant
Dieu vous garde pour moi

A mon binôme KESRAOUI Mahrez

A nos encadreurs Dr TALEB Youcef et Dr ISSIAKHEM

Au Professeur Toudeft, la responsable du service d'épimédiologie au CHU de Tizi Ouzou

A mes amis et camarades de classe

Je vous dédie ce travail en témoignage de mon profond amour. Puisse Dieu, vous préserver et vous accorder santé, longue vie et bonheur.

Cléophas NZISABIRA

Table des matières

Résumé

On a toujours voulu prévoir les valeurs futures afin de prendre des décisions sur la stratégie qui sera suivie à l'avenir .
Notre mémoire portera sur la modélisation de l'évolution du cancer colo-rectal sur la période allant de 2003 à 2011, pour cela on utilisera des modèles des séries chronologiques et tout particulièrement ceux associés à la classe des processus ARMA qui constituent un aspect important de l'application de la statistique, et cela en vue de prévoir le nombre de malades à prendre en charge en matière de suivi et de traitement.
Pour aboutir à cela on a utilisé 2 méthodes à savoir le lissage exponentiel et la méthode de Box Jenkins.
Nous terminerons notre travail par une analyse des résultats obtenus par les 2 méthodes précédemment citées, afin de déterminer la méthode la plus adéquate pour prédire l'évolution de cette pathologie.

Introduction générale

Connaître le futur, ou du moins avoir une idée du futur est l'un des soucis de l'Homme depuis toujours. De nos jours aussi, les raisons socioéconomiques poussent à anticiper l'avenir. Une question importante est de savoir sur quoi nous appuyer pour prédire l'avenir. Il est donc primordial d'arriver à prévoir le mieux possible le futur en s'appuyant sur le passé. D'une façon mathématique, on peut formuler le problème de la prévision en supposant avoir N observations $(x_1, x_2, ..., x_N)$ issues d'un processus (un ensemble des variables aléatoires) quantifiant une certaine activité, dans notre cas l'évolution du cancer et on souhaite connaître la valeur à une date future. Si on note le processus de l'évolution du cancer colo-rectal évoluant au cours du temps par F, on peut alors écrire : $x_t = \eta(F, t)$ où $\eta(F, t)$ est la mesure du processus F au temps t (x_t une variable qui quantifie le nombre de cas du cancer colo-rectal). S'il était possible de connaître F pour tout instant t, on ne pourrait plus avoir besoin de prédire x_t. Car en s'appuyant sur cette hypothèse, quelque soit le temps considéré, on pourrait connaître, et toutes les valeurs passées et toutes les valeurs futures. Tout le travail se résume dans la modélisation qui est une démarche statistique consistant à la représentation simplifiée d'observations. L'attention sera focalisée sur les propriétés évolutives d'une variable aléatoire, tant pour sa prévision que dans sa relation avec son passé.
Suivre périodiquement l'évolution du nombre de cas du cancer colo-rectal au CHU de Tizi Ouzou, définir quelques causes et la fréquence de cette pathologie sont certains objectifs du service Epidémiologie du CHU. L'objectif de ce mémoire est de modéliser l'évolution du cancer colo-rectal en vu de prévoir le nombre de malades à prendre en charge en matière de suivi et de traitement. Nous travaillerons avec les données allant de 2003 à 2011. Nous organisons ce travail en deux parties : Dans la première partie, on fera un rappel sur les processus stochastiques et les séries chronologiques(temporelles) ; dans la deuxième partie, on fera la présentation du CHU de Tizi Ouzou, en particulier le service d'Epidémiologie et on passera à l'application des théories présentées à la première parie sur les données recueillies au service concerné.

Chapitre 1

Processus stochastiques et modèle de séries chronologiques

1.1 Variables aléatoires et processus stochastiques :

On considère un espace de probabilité (Ω, A, P) où Ω est un espace des évènements, A une tribu adaptée à Ω et P une mesure de probabilité sur A. Alors on a les définitions suivantes :

Définition 1.1[8]
Une variable aléatoire réelle X est une application définie par :

$X : \Omega \to \mathbb{R}$ telle que pour tout réel $c, Ac = \{\omega \in \Omega \backslash X(\omega) \leq c\} \in A$. En d'autres termes, Ac est un événement tel que : $P(Ac) = F(c)$, F étant la fonction de répartition de X définie par : $F : \mathbb{R} \to [0, 1]$
$c \longmapsto F(c) = P(X(\omega) \leq c)$

Remarque 1.1. Dans toute la suite on désigne par T un ensemble d'indexation dénombrable contenu dans $\mathbb{N}$ ou dans $\mathbb{Z}$.

Définition 1.2 On appelle processus stochastique une fonction $X : T \times \Omega \longrightarrow \mathbb{R}$
telle que pour réel t donné, $X_t(.)$ soit une variable aléatoire. Un processus aléatoire est une suite de variables aléatoires $\{X_t(\omega), t \in T, \omega \in \Omega\}$ telle que pour tout $t \in T, X_t(\omega)$ soit une variable aléatoire sur Ω et que pour tout $\omega \in \Omega, X_t(\omega)$ est une réalisation du processus aléatoire sur l'ensemble d'indexation T

1.1.1 Classification des processus stochastiques

Soit $(X_t)_{t\in T}$ un processus stochastique tel que :

$$X : T \times \Omega \to E$$

$$(t,\omega) \longmapsto X_t(\omega)$$

T :ensemble de temps(espace d'indexation discret ou continu)
Ω :ensemble d'évènements
E : espace d'états pouvant être continu ou discret (fini ou infini)
En se basant sur la nature des ensembles T et E, on distingue les processus stochastiques suivants :

1. Processus à temps discret et à espace d'état discret
2. Processus à temps continu et à espace d'état discret
3. Processus à temps discret et à espace d'état continu
4. Processus à temps continu et à espace d'état continu

La trajectoire du processus $(X_t)_{t\in T}$, est une suite des réalisations des variables aléatoires X_t,$t \in T$. Pour t fixé, X_t représente une variable aléatoire à valeurs dans E. Pour $\omega \in \Omega$ fixe,$X_t(\omega)$ est une fonction du temps définie sur T.

1.1.2 Processus stationnaires

La notion stationnarité joue un rôle capital dans la théorie des processus aléatoires surtout en analyse des séries chronologiques. Généralement on considère deux types de stationnarités, la stationnarité stricte et la stationnarité faible.

1. **Processus stationnaire au sens strict (stationnarité forte)**[5]
   Soit un processus aléatoire réel $X_t, t \in T$.
   Le processus aléatoire X_t, est dit strictement (ou fortement) stationnaire si pour tous les n-uplets du temps $(t_1 < t_2 < ... < t_n)$ tel que $t_i \in T$ et pour tout temps $h \in T$ avec $t_i + h \in T, \forall i, i = 1, ..., n$, la suite $(X_{t_{1+h}}, X_{t_{2+h}}, ..., X_{t_{n+h}})$ a la même loi de probabilité que la suite $(X_{t_1}, X_{t_2}, ..., X_{t_n})$.
   La loi de probabilité qui correspond à la suite $(X_{t_1}, X_{t_2}, ..., X_{t_n})$ est caractérisée par sa fonction de répartition, d'où la définition équivalente de la stationnarité forte :
   $\forall (x_1, ..., x_n), \forall (t_1, ..., t_n)$, et $\forall h \in T$.
   $P(X_{t_1} < x_1, ..., X_{t_n} < x_n) = P(X_{t_{1+h}} < x_1, ..., X_{t_{n+h}} < x_n)$

Cette définition implique que tous les moments d'un tel processus sont invariants dans le temps. Cependant, la stationnarité au sens strict exige la connaissance de la loi conjointe du processus, ce qui est très difficile en pratique. Pour cette raison, on présente une stationnarité moins forte qui est la stationnarité faible.

2. **Processus stationnaire au sens faible :**[5]
   Le processus aléatoire X_t,$t \in T$ est dit faiblement stationnaire si seuls les moments d'ordre 1 et 2 sont stationnaires. Par exemple, $E(X_t^3)$ dépend du temps t alors le processus est faiblement stationnaire.
   Les processus stationnaires d'ordre 2 sont des processus générateurs des chroniques sans tendance en moyenne et sans tendance en variance mais cela ne signifie pas que les séries temporelles ont une représentation graphique stable.
   La définition peut s'écrire comme suit :
   - $\forall t \in T$, $E(X_t^2) < \infty$ (le processus est de second ordre)
   - $\forall t \in T$, $E(X_t) = m$ (m est une constante indépendante du temps)
   - $\forall t \in T$, $\forall h \in \mathbb{Z}$, $cov(X_t, X_{t+h}) = E[(X_t - m)(X_{t+h} - m)] = \gamma(h)$ (indépendant de t)

- Il est évident que la stationnarité forte implique la stationnarité faible
- La propriété d'invariance des moments dans le temps pour un processus stationnaire se conserve quand on prend la combinaison linéaire de plusieurs processus aléatoires stationnaires. C'est d'ailleurs un moyen commode pour définir de nouveaux processus stationnaires.
- Dans la suite, on désigne par processus stationnaire le processus faiblement stationnaire.

1.1.3 Processus Bruit Blanc (White Noise Process)

Définition 1.2. *"Bruit blanc faible"*

Le processus $\{X_t,t \in T\}$ est un bruit blanc **faible** si et seulement si :

$$\begin{cases} E(X_t) = m \\ V(X_t) = \sigma^2 & \forall t \in \mathbb{Z} \\ Cov(X_t, X_{t+h}) = 0 & \forall h \in \mathbb{Z}^*, t \in \mathbb{Z} \end{cases}$$

$X \perp Y \Rightarrow cov(X,Y) = 0$ mais l'inverse n'est pas forcément vraie. $X \perp Y$ veut dire que les variables X et Y sont orthogonales

Notation : On note $X_t \sim BB(0, \sigma^2)$

Définition 1.3. *"Bruit blanc fort"*

Le processus $\{X_t, t \in T\}$ est un bruit blanc **fort** si et seulement si :

$$\begin{cases} E(X_t) = m \\ V(X_t) = \sigma^2 \\ Cov(X_t, X_{t+h}) \sim iid \quad c'est - \text{à} - dire \quad les \quad accroissements(X_{t+h} - X_t) \quad sont \quad indépendants \end{cases}$$

Notation : On note $X_t \sim BB(m, \sigma^2)$. Le bruit blanc $(X_t)_{t \in T}$ est dit gaussien si sa distribution de probabilité suit une loi normale

1.1.4 Fonction d'autocovariance

Définition 1.4. Soit $\{X_t, t \in T\}$ un processus aléatoire, la fonction **d'autocovariance** est la fonction qui mesure la covariance pour un couple de valeurs associées à des dates différentes dont l'intervalle est de longueur h notée $\gamma(h)$

$$\gamma(h) = cov(X_t, X_{t+h}) = E[(X_t - E(X_t))(X_{t+h} - E(X_{t+h}))] \qquad \forall t, h \in \mathbb{Z}$$

Propostion 1.1.[7] " Propriétés de la fonction d'autocovariance"
Soit $(X_t)_{t \in T}$ un processus stationnaire et $\gamma(h)$ sa fonction d'autocovariance, on a :

1. *pour* $h = 0, \gamma(0) = V(X_t) \geq 0, \quad \forall t \in T$
2. $\gamma(h) = \gamma(-h), \quad \forall t \in T$
3. $| \gamma(h) | < \gamma(0), \quad \forall t \in T$

Preuve

1. $\gamma(0) = Cov(X_t, X_t) = Var(X_t) \geq 0$(par définition de la variance
2. $\gamma(-h) = Cov(X_{t-h}, X_t) = Cov(X_t, X_{t+h}) = \gamma(h)$
3. En utilisant l'inégalité de **Cauchy-Schwarz** on peut écrire

$$| Cov(X_{t+h}, X_t) | \leq (Var(X_{t+h}))^{\frac{1}{2}} (Var(X_t))^{\frac{1}{2}}$$

or $(X_t)_{t \in T}$ est stationnaire $\Rightarrow (Var(X_{t+h})) = (Var(X_t)), \quad \forall t, h \in T$
D'où $| Cov(X_{t+h}, X_t) | \leq (Var(X_t))^{\frac{1}{2}} (Var(X_t))^{\frac{1}{2}} \Rightarrow | \gamma(h) | \leq V(X_t) = \gamma(0)$

Estimation de la fonction d'autocovariance :
La fonction d'autocovariance n'est pas connue donc il faux l'estimer sur la base d'un vecteur aléatoire $(X_1, X_2, ..., X_T)$ par :

$$\widehat{\gamma}(h) = \frac{1}{T-h} \sum_{t=1}^{T-|h|} (X_t - \overline{X}_t)(X_{t-h} - \overline{X}_{t-h})$$

avec :

$$\overline{X}_t = \frac{1}{T}\sum_{t=1}^{T} X_t \qquad et \qquad \overline{X}_{t-h} = \frac{1}{T-h}\sum_{t=1}^{T-|h|} X_{t-h}$$

1.1.5 Fonction d'autocorrélation et d'autocorrélation partielle

Fonction d'autocorrélation

Définition 1.5. La fonction d'autocorrélation $\rho(h)$ d'un processus stationnaire du second ordre $\{X_t, t \in T\}$ de moyenne $E(X_t) = m$ est de variance $V(X_t) = \gamma(0)$ et donnée par :

$$\rho(h) = \frac{cov(X_t, X_{t+h})}{\sqrt{V(X_t)V(X_{t+h})}} = \frac{\gamma(h)}{\gamma(0)}, \qquad \forall h \in \mathbb{Z}$$

Propriétés

$$-1 \leq \rho(h) \leq 1$$

D'après la propriété 3 de la fonction d'autocovariance on a

$$-\gamma(0) \leq \gamma(h) \leq \gamma(0)$$

si $\gamma(h) = -\gamma(0)$ on a $\rho(h) = -\gamma(0)/\gamma(0) = -1$ et si $\gamma(h) = \gamma(h)$ on a $\rho(h) = \gamma(0)/\gamma(0) = 1$
D'où

$$-1 \leq \rho(h) \leq 1$$

Estimation de la fonction d'autocorrélation :
L'estimateur de la fonction d'autocorrélation est notée par : $\widehat{\rho}(h)$. Il est obtenu pour un échantillon de T réalisations du processus $\{X_t, t \in T\}$ en remplaçant dans l'expression de $\rho(h), \gamma(h)$ et $\gamma(0)$ par leurs estimateurs $\widehat{\gamma}(h)$ et $\widehat{\gamma}(0)$ on obtient :

$$\forall h \in \mathbb{Z}, \qquad \widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}$$

La fonction d'autocorrélation partielle

Définition 1.6. L'autocorrélation partielle désigne la corrélation entre X_t et X_{t-h} obtenue lorsque l'influence des variables X_{t-h-i}(avec $0 < i < h$) a été retirée. On note $\rho(h)$ et ϕ_{hh} les

fonctions respectivement d'autocorrélation et d'autocorrélation partielle de X_t au retard h. Soit ρ_h la matrice symétrique formée des $(h-1)$ premières autocorrélations de $\{X_t, t \in T\}$.

$$\rho_h = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{h-1} \\ \rho_1 & 1 & & & \rho_{h-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{h-3} \\ \vdots & \vdots & & \ddots & \\ \rho_{h-1} & \rho_{h-2} & \cdots & & 1 \end{pmatrix} \qquad \rho_h^* = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_1 \\ \rho_1 & 1 & & & \rho_2 \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_3 \\ \vdots & \vdots & & \ddots & \\ \rho_{h-1} & \rho_{h-2} & \cdots & & \rho_n \end{pmatrix}$$

$$o\grave{u} \qquad \phi_{hh} = \frac{\mid \rho_h^* \mid}{\mid \rho_h \mid}$$

$\mid \rho_h^* \mid =$ le déterminant de la matrice (ρ_h) dans laquelle on remplace la dernière colonne par le vecteur $[\rho_1, ..., \rho_h]$

Estimation de la fonction d'autocorrélation partielle :

On obtient l'estimateur de la fonction d'autocorrélation partielle $\widehat{\phi}_{hh}$ d'un processus $\{X_t, t \in T\}$, en utilisant les estimateurs des autocorrélations $\widehat{\rho}_h$ de la manière suivante :

$$\widehat{\phi}_{hh} = \frac{\mid \rho_h^* \mid}{\mid \rho_h \mid}$$

$$\rho_h = \begin{pmatrix} 1 & \widehat{\rho}_1 & \widehat{\rho}_2 & \cdots & \widehat{\rho}_{h-1} \\ \widehat{\rho}_1 & 1 & & & \widehat{\rho}_{h-2} \\ \widehat{\rho}_2 & \widehat{\rho}_2 & 1 & \cdots & \widehat{\rho}_{h-3} \\ \vdots & \vdots & & \ddots & \\ \widehat{\rho}_{h-1} & \widehat{\rho}_{h-2} & \cdots & & 1 \end{pmatrix} \qquad \rho_h^* = \begin{pmatrix} 1 & \widehat{\rho}_1 & \widehat{\rho}_2 & \cdots & \widehat{\rho}_1 \\ \widehat{\rho}_1 & 1 & & & \widehat{\rho}_2 \\ \widehat{\rho}_2 & \widehat{\rho}_1 & 1 & \cdots & \widehat{\rho}_3 \\ \vdots & \vdots & & \ddots & \\ \widehat{\rho}_{h-1} & \widehat{\rho}_{h-2} & \cdots & & \widehat{\rho}_n \end{pmatrix}$$

1.2 Séries chronologiques

Les définitions et théorèmes de cette section sont tirés du polycopié [3] de Agnès LAGNOUX.

La théorie des séries temporelles(chronologiques) abordée dans ce travail 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 phénomène, dans le but de décrire, expliquer puis prévoir ce phénomène dans le futur.

Définition 1.7. Une série chronologique(temporelle) est un processus stochastique. Elle est une suite d'observations d'une grandeur aléatoire liée à un phénomène. Ces observations

sont habituellement faites à des dates différentes, c'est-à-dire une suite de valeurs numériques indicées par le temps. Dans la suite de notre travail, on notera $(X_t, t \in T)$ où T est appelé espace de temps qui peut être :

-discret : dans ce cas $T \subseteq \mathbb{Z}$. Les dates d'observations sont le plus souvent équidistantes par les relevés mensuels, trimestriels, ... Ces dates équidistantes sont indexées par les entiers $t = 1, 2, ..., n$ où n est le nombre d'observations. On aura alors les variables $X_1, X_2, ..., X_n$ issues de la famille $(X_t, t \in T)$ où $T \subseteq \mathbb{Z}$ (le plus souvent $T = \mathbb{Z}$).

-continu (signal radio, résultat d'un électrocardiogramme...). 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 $(X_t, t \in T)$ où T 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.

Dans la suite de ce travail on considèrera uniquement des processus stochastiques $(X_t, t \in T)$ à temps discret.

1.2.1 Description d'une série chronologique

Il est important de signaler qu'une série chronologique(X_t) est une résultante de différentes composantes fondamentales suivantes :

- **tendance(trend)** Z_t qui représente l'évolution à long terme. Elle traduit le comportement moyen de la série.
- **la composante saisonnière** (ou saisonnalité) (S_t) qui correspond à un phénomène qui se répète à intervalles de temps réguliers (périodiques). En général, c'est un phénomène saisonnier d'où le terme de variations saisonnières.
- **la composante résiduelle** (ou bruit ou résidu) ε_t qui correspond à des fluctuations irrégulières, en général de faible intensité mais de nature aléatoire. On parle aussi d'aléas.
- **Des phénomènes accidentels** (grèves, conditions météorologiques exceptionnelles, crash financier) peuvent notamment intervenir. Dans notre cas, on les inclut dans ε_t

1.2.2 Objectifs d'étude de séries temporelles

L'étude d'une série chronologique permet d'analyser, de décrire et d'expliquer un phénomène au cours du temps et d'en tirer des conséquences pour des prises de décision (marketing...). Cette étude permet aussi de faire un contrôle, par exemple pour la gestion des stocks, le contrôle d'un processus chimique... Plus généralement, nous pouvons déjà poser

quelques problèmes lorsqu'on étudie une série chronologique. Mais l'un des objectifs principaux de l'étude d'une série chronologique est la prévision qui consiste à prévoir les valeurs futures $X_{n+h}(h = 1, 2, 3, ...)$ de la série chronologique à partir de ses valeurs observées jusqu'au temps $n : X_1, X_2, ..., X_n$. La prédiction de la série chronologique au temps $n + h$ est notée $\widehat{X_n}(h)$ et, en général, est différente de la valeur réelle X_{n+h} que prend la série au temps $n + h$.

L'intervalle de prévision, défini par les valeurs $\widehat{X}_n^1(h)$ et $\widehat{X}_n^2(h)$, est susceptible de contenir la valeur inconnue X_{n+h}.

1.2.3 Description schématique de l'étude complète d'une série chronologique

Comme on vient de le voir, l'un des objectifs principaux de l'étude d'une série chronologique est la prévision des valeurs futures de cette série. Pour cela, on a besoin de connaître ou tout au moins de modéliser le mécanisme de production de la série chronologique.

Notons que les variables X_t ne sont le plus souvent ni indépendantes (on peut s'attendre en effet à ce que des observations relativement proches dans le temps soient liées) ni identiquement distribuées (dans la plupart des cas, le phénomène évolue, se modifie au cours du temps ce qui entraîne que les variables le décrivant ne sont pas équidistribuées). Cela nécessite des méthodes statistiques de traitement et de modélisation spécifiques puisqu'en particulier dans un cadre standard (celui de la description d'un échantillon) les méthodes statistiques classiques sont basées sur des hypothèses d'indépendance. Schématiquement, les principales étapes de traitement d'une série chronologique sont les suivantes :

1. correction des données
2. observation de la série
3. modélisation (avec un nombre fini de paramètres)
4. analyse de la série à partir de ses composantes
5. diagnostic du modèle - ajustement au modèle
6. prédiction (= prévision)

On explique ces étapes :

1. **correction des données :**Avant de se lancer dans l'étude d'une série temporelle, il est souvent nécessaire de traiter, modifier les données brutes.

2. **observation de la série :** Une règle générale en Statistique Descriptive consiste à commencer par regarder les données avant d'effectuer le moindre calcul. Ainsi, une fois la

série corrigée et pré-traitée, on trace son graphique c'est-à-dire la courbe de coordonnées (t, X_t).

3. **modélisation (avec un nombre fini de paramètres) :**Un modèle est une image simplifiée de la réalité qui vise à traduire les mécanismes de fonctionnement du phénomène étudie et permet de mieux les comprendre. Un modèle peut être meilleur qu'un autre pour décrire la réalité et bien sûr, plusieurs questions se posent alors : comment mesurer cette qualité ?

comment diagnostiquer un modèle ? Nous présentons dans cette section une petite liste qui sert à résumer et classifier les différents modèles envisagés dans ce travail. On distingue principalement deux types de modèles :

– **-les modèles déterministes**. Ces modèles relèvent de la Statistique Descriptive. Ils ne font intervenir que de manière sous-jacente le calcul des probabilités et consistent à supposer que l'observation de la série à la date t est une fonction du temps t et d'une variable centrée faisant office d'erreur au modèle, représentant la différence entre la réalité et le modèle proposé : $X_t = f(Z_t, S_t, \varepsilon_t)$
On suppose de plus qu'elles sont décorrélées. Les deux modèles de ce type les plus utilisés sont les suivants
1. le modèle additif. C'est le " modèle classique de décomposition " dans le traitement des modèles d'ajustement. La variable X_t s'écrit comme la somme de trois termes :$X_t = Z_t + S_t + \varepsilon_t$ où Z_t représente la tendance (déterministe), S_t la saisonnalité (déterministe aussi) et les composantes (" erreurs au modèle ") aléatoires iid.

2. le modèle multiplicatif. La variable s'écrit au terme d'erreur près comme le produit de la tendance et d'une composante de saisonnalité $X_t = Z_t(1 + S_t)(1 + \varepsilon_t)$. Cet ajustement intervient dans les modèles (G)ARCH.

– **-les modèles stochastiques**.

Ils sont du même type que les modèles déterministes à ceci près que les variables de bruit ε_t ne sont pas iid mais possèdent une structure de corrélation non nulle : ε_t est une fonction des valeurs passées et d'un terme d'erreur η_t, $\varepsilon_t = g(\varepsilon_{t1}, \varepsilon_{t2}, ..., \eta_t)$.

Les modèles de ce type les plus utilisés sont les modèles SARIMA (et les sous-modéles ARIMA, ARMA,...). Comme on l'a vu, la série chronologique est l'observation d'un processus stochastique : la modélisation porte ici sur la forme du processus (ε_t). Le cas particulier où la relation fonctionnelle g est linéaire est très important et très utilisé. Il mène aux modèles auto-régressifs linéaires, par exemple un modèle d'ordre 2 avec des coefficients auto-régressifs a_1, a_2 est donné par $\varepsilon_t = a_1 X_{t-1} + a_2 X_{t-2} + \eta_t$, où (η_t) est un bruit blanc c'est-à-dire une

variable aléatoire de moyenne nulle non corrélée.

Les deux types de modèles ci-dessus conduisent à des techniques de prévision bien particulières. Schématiquement, on s'intéresse tout d'abord à la tendance et à la saisonnalité éventuelle(s) que l'on isole tout d'abord. Après on essaie de les modéliser, les estimer. Enfin on les élimine de la série : ces deux opérations s'appellent la détendancialisation et la désaisonnalisation de la série. Une fois ces composantes éliminées on obtient la série aléatoire ε_t :
- pour les modèles déterministes, cette série sera considérée comme décorrélée et il n'y a plus rien à faire.
- pour les modèles stochastiques, on obtient une série stationnaire (ce qui signifie que les observations successives de la série sont identiquement distribuées mais pas nécessairement indépendantes) qu'il s'agit de modéliser.

1.2.4 Modélisation déterministe

Modèle additif

On considère dans cette partie une série $X = (X_t)_{t\in T}$ admettant une décomposition additive $X_t = Z_t + S_t + \varepsilon_t, t = 1, ..., n$, où Z_t est la composante tendancielle, S_t la composante saisonnière et ε_t représente l'erreur ou l'écart au modèle(fluctuations aléatoires). Comme nous l'avons dit ,

- **la tendance** Z_t exprime un mouvement à moyen terme de la série. Elle est le plus souvent modélisée par une fonction polynomiale du temps.
- **la composante saisonnière** exprime un phénomène qui se reproduit de manière analogue sur chaque intervalle de temps successif. L'étendue de cet intervalle qui est constante est appelée période et on la note P dans la suite. La plupart du temps, on suppose que la composante saisonnière est constante sur chaque période P, c'est-à-dire $S_{t+P} = S_t, \forall t$. Cela revient à dire que l'effet net du saisonnier sur une période est nul ; ce qui est trivial puisqu'il est repris dans la tendance générale de la série chronologique. Il s'agit là du modèle le plus simple dans lequel le saisonnier est caractérisé par P coefficients $c_1, ..., c_P$. Lorsque $P = 4$, la série est trimestrielle ; lorsque $P = 12$, la série est mensuelle...On suppose par ailleurs que l'effet du saisonnier est en moyenne nul sur une période, ce qui signifie que $\sum_{i=1}^{P} c_i = 0$.
- **les erreurs** sont des variables aléatoires centrées. On considère le plus souvent un bruit blanc, c'est-à-dire une suite de v.a.r. telles que $E(\varepsilon_t) = 0$ et $E(\varepsilon_t \varepsilon_{t'}) = \sigma^2 \delta_{tt'}$. Les v.a.r. sont alors non corrélées et lorsque le bruit blanc est gaussien c'est-à-dire que $\varepsilon_t \longrightarrow N(0, \sigma^2)$, on a de plus l'indépendance des ε_t.

Modèle multiplicatif

On considère dans cette section une série $X = (X_t)_{t\in T}$ admettant une décomposition multiplicative $X_t = Z_t(1 + S_t)(1 + \varepsilon_t), t = 1...n$,

Là encore, la composante saisonnière vérifie $\sum_{i=1}^{P} c_i = 0$

L'amplitude de la série n'est plus constante au cours du temps : elle varie au cours du temps proportionnellement à la tendance Z_t au bruit près. Dans ce modèle, on considère que les amplitudes des fluctuations dépendent du niveau.

Modèle mixte

Il s'agit de modèles où addition et multiplication sont utilisées. On peut supposer par exemple que la composante saisonnière agit de façon multiplicative alors que les fluctuations irrégulières sont additives : $X_t = Z_t S_t + \varepsilon_t, t = 1, ..., n$

Choix du modèle

Avant toute modélisation et étude approfondie du modèle, on tente d'abord de déterminer si on est en présence d'une série dans laquelle pour une observation X donnée
– la variation saisonnière S_t s'ajoute simplement à la tendance Z_t ; c'est le modèle additif.
– la variation saisonnière S_t est proportionnelle à la tendance Z_t ; c'est le modèle multiplicatif. Afin de faire cette distinction, on peut se baser sur une méthode graphique ou utiliser une méthode analytique.

- **Méthode du profil**

Pour faire la détermination entre modèle additif et modèle multiplicatif graphiquement, on peut par exemple superposer les saisons représentées par des courbes de profil sur un même graphique. Si ces courbes sont parallèles, le modèle est additif, autrement le modèle est multiplicatif.

- **Méthode de la bande**

On fait un graphique représentant la série chronologique, puis on trace une droite passant respectivement par les minima et par les maxima de chaque saison. Si ces deux droites sont parallèles, nous sommes en présence d'un modèle additif. Dans le cas contraire, c'est un modèle multiplicatif.

- **Méthode analytique**

On calcule les moyennes et les écarts-types pour chacune des périodes considérées puis la

droite des moindres carrés $\sigma = a\overline{x} + b$. Si a est nul, c'est le modèle additif, sinon c'est le modèle multiplicatif.

Important : Il faut bien tester avec les trois méthodes pour décider du modèle !

1.2.5 Analyse de la saisonnalité

Dans le traitement d'une série temporelle, il est indispensable d'étudier la saisonnalité. Si cette composante existe il est convenable de l'isoler afin de pouvoir étudier les autres composantes. Car une désaisonnalisation systématique sans tester l'existence de cette composante peut conduire à la création d'un bruit qui serait nuisible à l'analyse de la série, ce qui dégraderait la qualité de la prévision.

la détection de la saisonnalité

1. **La représentation graphique et le tableau de Buys-Ballot**
   On peut détecter la saisonnalité par l'analyse du graphique de la chronique. Au cas où le graphique ne serait pas révélateur de cette composante ou en cas de doute on peut passer par le tableau de Buys-Ballot qui permet d'analyser finement l'historique.
2. **Analyse de la variance et test de Fisher**
   L'examen visuel du graphique ou du tableau Buys-Ballot ne permet pas toujours de déterminer l'existence d'une saisonnalité. Mais l'analyse de variance et le test de Fisher permettent de pallier ces inconvénients. Ce test fait la comparaison des moyennes des échantillons issus d'une même population dont on suppose qu'elle suit une distribution normale.

(Pour plus de détails voir [5] p.11)

1.2.6 Analyse de la tendance

Dans cette section on se place dans le cadre d'un modèle composé uniquement d'une tendance et de fluctuations irrégulières et on donne différentes méthodes permettant d'estimer la tendance.

Rappels sur la régression linéaire

Soient deux variables quantitatives X et Y de taille n. Le problème de régression consiste à chercher une relation pouvant éventuellement exister entre x et y, par exemple de la forme $y = f(x)$. Lorsque la relation recherchée est affine, c'est-à-dire de la forme $y = ax+b$, on parle de régression linéaire. Mais même si une telle relation est effectivement présente, les données

mesurées ne vérifient pas en général cette relation exactement. Pour tenir compte dans le modèle mathématique des erreurs observées,on considère les données $y_1, y_2, ..., y_n$ comme étant de réalisations de la variable aléatoire Y et parfois aussi les données $x_1, x_2, ..., x_n$ comme étant de réalisations d'une variable aléatoire X. On dit que la variable Y est la variable dépendante ou variable expliquée et que la variable X est la variable explicative.

Droite des moindres carrés

Les données $(x_i, y_i), i = 1, ..., n$ peuvent être représentées par un nuage de n points dans le plan (x, y), le diagramme de dispersion. Le centre de gravité de ce nuage peut se calculer facilement : il s'agit du point de coordonnées

$$(\overline{x}, \overline{y}) = (\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i)$$

. Rechercher une relation affine entre les variables X et Y revient à rechercher une droite qui s'ajuste le mieux possible à ce nuage de points. Parmi toutes les droites possibles, on retient celle qui jouit d'une propriété remarquable :c'est celle qui rend minimale la somme des carrés des écarts des valeurs observées y_i à la droite $\widehat{y}_i = ax_i + b$. Si ε_i représente cet écart, appelé aussi résidu, le principe des moindres carrés ordinaires (MCO) consiste à choisir les valeurs de a et de b qui minimisent

$$E = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

.

On montre en minimisant la fonction de deux variables

$$E = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

que le couple solution $(\widehat{a}, \widehat{b})$ est donné par :

$\widehat{a} = \frac{Cov(X,Y)}{V(X)}$ et $\widehat{b} = \overline{Y} - a\overline{X}$

avec

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} y_i x_i - (\frac{1}{n}\sum_{i=1}^{n} y_i)(\frac{1}{n}\sum_{i=1}^{n} x_i)$$

et

$$V(X) = (\frac{1}{n}\sum_{i=1}^{n} x_i^2) - (\frac{1}{n}\sum_{i=1}^{n} x_i)$$

La droite d'équation $y = \widehat{a}x + \widehat{b}$ est appelée droite de régression de Y en X et est notée : $\nabla_{Y/X}$.

1.2.7 Différents types d'ajustement tendanciel

Il existe plusieurs types d'ajustement de la tendance, on peut citer entre autres :

1. Ajustement tendanciel linéaire par la droite de moindres carrés
2. Ajustements tendanciels non linéaires tels que l'ajustement exponentiel, l'ajustement hyperbolique, ajustement par une fonction logarithmique, ajustement par une fonction puissance.
3. Estimation non paramétrique : Dans certaines situations, il n'est pas toujours facile de trouver le degré du polynôme d'ajustement pour Z_t ou de changement de variable adéquat. Par ailleurs, on ne sait pas non plus déterminer l'allure de cette fonction. Dans cette situation, on a recours à la théorie non paramétrique de l'estimation de la tendance qui ne suppose rien sur celle-ci à priori et on approxime la tendance par la moyenne mobile d'ordre $p, X_t^* = M_p(t)$

Il est important de signaler qu'ici deux cas peuvent se présenter :
- Si $p = 2m + 1$ on aura $X_t^* = M_p(t) = (\frac{1}{p})(\sum_{i=-m}^{m} \theta_i X_{t+i})$
- Si $p = 2m$ on aura $X_t^* = M_p(t) = (\frac{1}{p})(0.5X_{t-m} + \sum_{i=-m+1}^{m-1} \theta_i X_{t+i} + 0.5X_{t+m})$

1.2.8 Les moyennes mobiles

Dans la précédente section, nous avons vu quelques types d'ajustement tendanciel dans un modèle composé simplement d'une tendance et de variations irrégulières. Dans la partie présente, nous nous intéressons à un ensemble d'outils, les moyennes mobiles ou filtres linéaires, transformations de séries chronologiques. Le but sera de lisser une série temporelle, en gardant la tendance et en supprimant la saisonnalité pour ensuite procéder à l'estimation de ces deux composantes. Nous présentons dans la suite les moyennes mobiles et leurs propriétés.

Définition des moyennes mobiles

Soit une série chronologique X_t, on appelle moyenne mobile, une transformation de X_t s'écrivant comme combinaison linéaire finie des valeurs de la série correspondant à des dates entourant t. La série transformée s'écrit comme suit : $M_{p+q+1}(X_t) = \sum_{i=-q}^{p} \theta_i X_{t+i} = \theta_{-q}X_{t-q} + ... + \theta_{-1}X_{t-1} + \theta_0 X_t + \theta_1 X_{t+1} + ... + \theta_p X_{t+p}$ où $\theta_{-q}, ..., \theta_p$ sont des réels et $q, p \in \mathbb{N}$ et on appelle ordre de la moyenne le nombre $(q + p + 1)$
Étant donné qu'une moyenne mobile en t est une combinaison linéaire finie des valeurs de

la série correspondant à des dates entourant t, elle réalise donc un lissage de la série, une moyennisation.

Notation : On peut réécrire la moyenne mobile en termes d'opérateurs. On définit pour cela l'opérateur B, appelé opérateur retard, qui à tout processus $(X_t)_{t\in T}$ associe le processus $(Y_t)_{t\in T}$ défini par
$\forall t \in T, \quad Y_t = BX_{t-1}$
On peut composer B avec lui-même et on a B^2 tel que $\forall \quad t \in T, B^2X_t = X_{t-2}$

Par itération et par récurrence on peut définir $B^kX_t = X_{t-k}, k \in \mathbb{N}$. Par convention, B^0 est l'opérateur identité I. L'opérateur B est linéaire et inversible. Son inverse $B^{-1} = F$ est défini par $\forall t \in T, FX_t = X_{t+1}$

L'opérateur F est appelé opérateur avance. On peut réécrire la moyenne mobile en termes d'opérateurs B et F : $M_{p+q+1} = \sum_{i=-q}^{p} \theta_i B^{-i} = \sum_{i=-q}^{p} \theta_i F^i$.
En factorisant par B^q et en faisant $j = i+q$, on obtient la forme canonique suivante $M_{p+q+1} = B^q \sum_{j=0}^{p+q} \theta_{j-q} F^j = B^q \sum_{i=q}^{p} \theta_i F^{q+i} = B^q P(F)$ où $P(F)$ est appelé polynôme caractéristique de la moyenne mobile.

Effet d'une moyenne mobile sur une tendance

L'application d'une moyenne mobile arithmétique (paire ou impaire) ne modifie pas une tendance constante. L'application d'une moyenne mobile arithmétique (paire ou impaire) conserve une tendance linéaire.

Effet d'une moyenne mobile sur une composante saisonnière

Si la série X_t possède une composante saisonnière de période P alors l'application d'une moyenne mobile d'ordre P supprime cette saisonnalité. La série $M_P(X_t), t \in T$ ne possède plus de composante saisonnière de période P.

Effet d'une moyenne mobile sur les fluctuations irrégulières

Jusqu'à présent on s'est intéressé qu'à l'effet d'une moyenne mobile sur la partie déterministe de la série (tendance et saisonnalité). On va étudier maintenant l'effet d'une moyenne mobile sur le résidu lorsque celui-ci est un bruit blanc. Par construction, une moyenne mobile consiste à faire des moyennes partielles de proche en proche. On obtient donc un lissage de la série. L'effet de la composante irrégulière est d'autant plus atténué que l'ordre de la moyenne mobile est grand.

1.2.9 Décomposition d'une série chronologique

On dispose maintenant des outils de base permettant la décomposition d'une série chronologique. Il paraît clair qu'afin de pouvoir estimer la tendance, c'est-à-dire le mouvement du phénomène observé sur un grand intervalle de temps, il faut disposer d'une série statistique sur une longue période. Disposant de ces données, comme on l'a dit précédemment, le premier travail consiste à effectuer une représentation graphique adéquate permettant d' avoir une vue globale du phénomène en question. Afin d'éliminer ou d'amortir les mouvements cycliques, saisonniers et accidentels, on utilise donc la technique des moyennes mobiles et on procède ainsi en quelque sorte au lissage de la courbe. D'après ce qu'on a vu, la méthode des moyennes mobiles arithmétiques peut être utilisée pour tout type de modèle. Cependant, d'après ses propriétés, elle est particulièrement adaptée pour le modèle déterministe additif lorsque

- la tendance est sensiblement linéaire,
- la composante saisonnière est périodique,
- le bruit est de variance faible.

La méthode des moyennes mobiles peut être préconisée comme technique de lissage de la série quelle que soit la forme de la tendance. On peut cependant utiliser d'autres techniques plus adaptées pour lisser la série. La décomposition et l'étude de la série statistique (X_t) en vue de la prédiction se font selon les étapes suivantes :

1. application d'une moyenne mobile d'ordre judicieusement choisi.
2. estimation de la saisonnalité.
3. estimation de la tendance.
4. itération éventuelle de la procédure.
5. . prévision des valeurs futures.
6. analyse des résidus.

• La série lissée par moyenne mobile

Tout d'abord, on applique une moyenne mobile arithmétique d'ordre $2p+1$ dans le cas d'une saisonnalité d'ordre impair $2p+1$ ou une moyenne mobile arithmétique modifiée d'ordre $2p+1$ dans le cas d'une saisonnalité de période paire $2p$. Dans chaque cas, on a vu que la saisonnalité est ainsi annulée et que la variance du bruit est diminuée. Si le modèle est bon, la série transformée ne contient plus aucun mouvement saisonnier. Notons que la série ainsi obtenue est de longueur $n-2p$. Et on trouvera une série $M_k(X_t) = X_t^*$

•Estimation de la saisonnalité

On calcule la série S_t^* diminuée de sa tendance en retranchant de la série de départ la série

transformée X_t^* . On estime ensuite les P coefficients du saisonnier par la moyenne des valeurs de S_t^* correspondant à chaque temps de la période. On peut affiner la méthode en retranchant ensuite de chaque coefficient estimé la moyenne des coefficients afin que la condition de nullité de la moyenne des coefficients sur la période soit respectée. Plus précisément, supposons que les observations soient périodiques de période $P = 2p$ paire et réalisées sur N périodes. A l'issue du premier filtrage, on dispose de la série

$$S_t^* = X_t - X_t^*, t = p+1, ..., PN-p.$$

Cette série est appelée série corrigée de la tendance

L'estimation du coefficient saisonnier c_j^* est donc

$$\frac{1}{N-1}\sum_{i=2}^{N-1} S^*_{j+P(i-1)}, \quad 1 \leq j \leq p$$

$$\frac{1}{N-1}\sum_{i=1}^{N-1} S^*_{j+P(i-1)}, \quad 1 \leq j \leq N)$$

Cette opération consiste appliquer à la série $(S_t^*)_{t=p+1,...,PN-p}$ à laquelle on ajoute P observations nulles pour $t = 1, ..., p$ et $t = PN-p+1, ..., PN$, une moyenne mobile d'ordre $P(N-1)+1$ et de coefficients

$$\frac{1}{N-1}(\overbrace{1,\underbrace{0,...,0}_{p-1},1,0,...,0,1,\underbrace{0,...,0}_{p-1},1}^{})$$

Les estimateurs des P coefficients saisonniers
$\widehat{c_j} = c_j^* - \frac{1}{P}\sum_{j'=1}^{P} c_{j'}^*$ de façon à vérifier la condition d'une somme nulle sur une période.

•Estimation de la tendance

Une fois les coefficients saisonniers estimés à l'étape précédente, on retranche l'estimation du saisonnier à la série X_t. La série ainsi obtenue est appelée série corrigée des valeurs saisonnières :
$X_{CVS,t} = X_t - \widehat{S}_t$ avec $\widehat{S}_t = \widehat{c_j}$

On procède alors à l'estimation du terme représentant la tendance par une méthode de régression comme précédemment : on modélise le plus souvent la tendance par un polynôme $Q(t)$

Puis on ajuste au sens des moindres carrés un polynôme $\widehat{Q}_t$ à la série corrigée des valeurs saisonnières $X_{CVS,t}$, avec les paramètres suivants : $\widehat{a} = \frac{Cov(t, X_{CVS,t})}{V(t)}$ et $\widehat{b} = \overline{X_{CVS,t}} - \widehat{a}\overline{t}$

• **Itération de la procédure**

On procède parfois à une itération de la procédure : on estime à nouveau la tendance à partir de la série $X_{CVS,t}$ en utilisant une moyenne mobile d'ordre différent de celui utilisé à l'étape 1. Soit $\widehat{Z}_t$ cette estimation. On retranche à la série X_t cette tendance et on revient éventuellement à l'étape 2 pour une seconde estimation du saisonnier.

• **Prévision des valeurs futures** Pour prévoir les valeurs futures de la série, on utilise l'estimation de la tendance et celle de la composante saisonnière. Si on souhaite prévoir une valeur de la série à l'instant $n+h$ où $h \geq 1$, c'est-à-dire à l'horizon h, on utilise les estimations de la tendance et de la saisonnalité et on pose $\widehat{X_t}(h) = \widehat{Q}_{n+h} + \widehat{c}_j$, $n + h \equiv j[P]$.

• **Analyse des résidus**

Une fois qu'on a estimé les composantes du modèle, on pourra contrôler la pertinence du modèle par une analyse des résidus ε_t. Ceux-ci sont définis par $\widehat{\varepsilon}_t = X_t - S_t^* - \widehat{Q}_t = X_{CVS,t} - \widehat{Q}_t$. Si le modèle est bon, il ne doit rester dans les résidus aucune trace du saisonnier. Pour le vérifier, on trace le corrélogramme des résidus c'est-à-dire le graphe d'un estimateur de la fonction d'autocorrélation.

Le corrélogramme n'est tracé en théorie que dans le cas où la série est stationnaire, ce qui implique en particulier qu'il n'y ait dans cette série ni tendance ni saisonnalité. En pratique, on s'en sert (dans le cas de l'analyse des résidus) pour vérifier justement l'absence de saisonnalité dans les résidus. Si c'est le cas et si le modèle est bon, le corrélogramme ne doit présenter que des valeurs faibles, indiquant une faible corrélation entre les erreurs. Si au contraire, le corrélogramme présente des pics régulièrement espacés, cela indique que le saisonnier n'a pas été complètement éliminé et c'est donc le signe que le modèle proposé a échoué. On peut alors réitérer la procédure ci-dessus ou proposer un autre modèle. Dans le cas où le corrélogramme des résidus n'indique pas la présence d'un mouvement saisonnier, on trace le graphe des résidus $(t, \widehat{\varepsilon}_t)$ qui sert à repérer d'éventuelles observations exceptionnelles, un mouvement tendanciel,...

1.2.10 Lissages exponentiels

Les définitions et le théorèmes présentés dans cette section sont tirés du polycopié [4] de Sylvain Rubenthaler.

Lissage exponentiel simple

On dispose de $x_1, ..., x_n$ et on veut estimer $x_{n+h} (h \in \mathbb{N}^*)$. Pour $\alpha \in]0,1[$, on définit la prévision par lissage exponentiel simple

$$\widehat{x}_n(h) = \alpha \sum_{j=0}^{n-1} (1-\alpha)^j x_{n-j}$$

(En fait, on fait une moyenne pondérée des observations passées).

Remarque 1 :

Plus α est petit plus on accorde d'importance aux observations anciennes. On remarque que $\widehat{x}_n(h)$ ne dépend pas de h.

On peut calculer par récurrence à l'aide de la formule suivante dite de mise à jour :

$$\widehat{x}_n(h) = \alpha x_n + (1-\alpha)\widehat{x}_{n-1}(h)$$

Ces formules permettent de calculer rapidement les prédictions .

Lemme 1 : La prédiction $\widehat{x}_n(h)$ est (asymptotiquement) la solution de l'équation

$$\widehat{x}_n(h) = argmin_x \sum_{j=0}^{n-1} (1-\alpha)^j (x_{n-j} - x)^2$$

. C'est un problème de moindres carrés.

Démonstration. Soit

$$f : x \longrightarrow \sum_{j=0}^{n-1} (1-\alpha)^j (x_{n-j} - x)^2$$

on commence par chercher les points critiques. Nous avons :

$$f^{'}(x) = -2 \sum_{j=0}^{n-1} (1-\alpha)^j (x_{n-j} - x)$$

Donc $f^{'}$ s'annule en

$$x_0 = \frac{\sum_{j=0}^{n-1}(1-\alpha)^j x_{n-j}}{\sum_{j=0}^{n-1}(1-\alpha)^j} = \frac{\alpha \sum_{j=0}^{n-1}(1-\alpha)^j x_{n-j}}{1-\alpha^n} \xrightarrow[n \longrightarrow +\infty]{} \sum_{j=0}^{n-1} (1-\alpha)^j x_{n-j}$$

Une étude rapide du tableau de variation de f nous indique que x_0 est le minimum absolu. C'est à cause de l'équivalent que l'on dit que $\widehat{x}_n(h)$ est la solution « asymptotique » du problème des moindres carrés.

Remarque 2 :

Pour choisir α, on peut se servir de la remarque précédente ou calculer, pour tout α, les estimateurs calculés avec le paramètre $\alpha : x_t(h, \alpha)$. On regarde ensuite l'erreur quadratique

$$E_2 = \sum_{t=1}^{n-h} (x_{t+h} - x_t(h, \alpha))^2$$

Si cette erreur est petite, c'est que le paramètre α fournit des prédictions perfomantes, au vu des données $x_1, ..., x_n$.

Lissage exponentiel double

On cherche à ajuster à l'instant t une droite d'équation :$y_t = a_1 + a_2(t - n)$. La prévision par le lissage exponentiel double est la suivante

$$\widehat{x}_n(h) = \widehat{a_1} + \widehat{a_2}h$$

où $(\widehat{a_1}, \widehat{a_2})$ est solution de

$$(\widehat{a_1}, \widehat{a_2}) = argmin_{(a_1,a_2)\in\mathbb{R}} \sum_{j=0}^{n-1} (1-\alpha)^j (x_{n-j} - (a_1 + a_2 j))^2$$

Lemme 2 :
Les solutions du problème de minimisation ci-dessus sont asymptotiquement
$\begin{cases} \widehat{a_1} = -L_2(n) + 2L_1(n) \\ \widehat{a_2} = \dfrac{\alpha}{1-\alpha}(-L_2(n) + L_1(n)), \end{cases}$
où
$\begin{cases} L_1(n) = \alpha \sum_{j=0}^{n-1} (1-\alpha)^j x_{n-j} \\ L_2(n) = \alpha \sum_{j=0}^{n-1} (1-\alpha)^j L_1(n-j) \end{cases}$

Il faut remarquer qu'on a fait deux lissages exponentiels, en effet L_2 est un lissage exponentiel de L_1. D'où l'appellation de lissage exponentiel double.

Démonstration du lemme
On note

$$C(a_1, a_2) = \sum_{j=0}^{n-1} (1-\alpha)^j (x_{n-j} - (a_1 + a_2 j))^2$$

On va commencer à chercher les points critiques de C. On a

$$\frac{\partial C}{\partial a_1}(a_1, a_2) = -2 \sum_{j=0}^{n-1} (1-\alpha)^j (x_{n-j} - (a_1 + a_2 j))$$

$$\frac{\partial C}{\partial a_2}(a_1, a_2) = -2\sum_{j=0}^{n-1}(1-\alpha)^j(x_{n-j} - (a_1 + a_2 j))$$

On cherche donc (a_1, a_2) solution du système

$$\mathbf{1}\begin{cases} -\sum_{j=0}^{n-1}(1-\alpha)^j x_{n-j} - \sum_{j=0}^{n-1}(1-\alpha)^j a_1 + \sum_{j=0}^{n-1}(1-\alpha)^j a_2 j = 0 \\ -\sum_{j=0}^{n-1}(1-\alpha)^j x_{n-j} - \sum_{j=0}^{n-1}(1-\alpha)^j a_1 j + \sum_{j=0}^{n-1}(1-\alpha)^j a_2 j^2 = 0 \end{cases}$$

Il convient de rappeler les formules

$$\sum_{j=0}^{+\infty}(1-\alpha)^j = \frac{1}{\alpha}$$

$$\sum_{j=0}^{+\infty} j(1-\alpha)^j = \frac{1-\alpha}{\alpha^2}$$

$$\sum_{j=0}^{+\infty} j^2(1-\alpha)^j = \frac{(1-\alpha)(2-\alpha)}{\alpha^3}$$

On peut les retrouver en manipulant des séries entières. On obtient donc, en remplaçant certaines sommes partielles de séries par leurs limites,

$$\mathbf{2}\begin{cases} -\alpha\sum_{j=0}^{n-1}(1-\alpha)^j x_{n-j} - a_1 + \frac{1-\alpha}{\alpha}a_2 = 0 \\ -\alpha^2\sum_{j=0}^{n-1}(1-\alpha)^j j x_{n-j} - (1-\alpha)a_1 + \frac{(1-\alpha)(2-\alpha)}{\alpha}a_2 = 0 \end{cases}$$

Les systèmes 1 et 2 ne sont pas équivalents. Plus n est grand, plus ils se ressemblent. On va continuer le calcul à partir du système 2. C'est pour cela qu'on parle de solution asymptotique dans l'énonce du lemme. On pose pour tout n

$$\begin{cases} L_1(n) = \alpha\sum_{j=0}^{n-1}(1-\alpha)^j x_{n-j} \\ L_2(n) = \alpha\sum_{j=0}^{n-1}(1-\alpha)^j L_1(n-j) \end{cases}$$

On remarque que

$$L_2(n) = \alpha\sum_{j=0}^{n-1}(1-\alpha)^j L_1(n-j)$$

$$= \sum_{j=0}^{n-1}(1-\alpha)^j \sum_{i=0}^{n-j-1}(1-\alpha)^i x_{n-j-i}$$

$$= \alpha^2\sum_{k=0}^{n-1} x_{n-k}(1-\alpha)^k(k+1)$$

où

$$k = i + j)$$

Le système devient
$$\begin{cases} -L_1(n) - a_1 + \frac{1-\alpha}{\alpha} a_2 = 0 \\ -L_2(n) + \alpha L_1(n) - (1-\alpha)a_1 + \frac{(1-\alpha)(2-\alpha)}{\alpha} a_2 = 0 \end{cases}$$
D'où

$$-L_1(n)(2-\alpha) - a_1(2-\alpha) + L_2(n) - \alpha L_1(n) + (1-\alpha)a_1 = 0$$

$$L_2(n) - 2L_1(n) = a_1$$

et

$$\begin{aligned} a_2 &= \frac{\alpha}{1-\alpha}(L_1(n) + a_1) \\ &= \frac{\alpha}{1-\alpha}(L_2(n) - L_1(n)) \end{aligned}$$

On montre que la fonction C est convexe. Pour (a_1, a_2) et (b_1, b_2) dans $\mathbb{R}^2$ et $\lambda \in [0,1]$, on a

$$C(\lambda(a_1,a_2) + (1-\lambda)(b_1,b_2)) = \sum_{j=0}^{n-1} (1-\alpha)^j x_{n-j} - (\lambda a_1 + (1-\lambda)b_1 - (\lambda b_1 + (1-\lambda)b_2 j))^2$$

$$\begin{aligned} &= \sum_{j=0}^{n-1} (1-\alpha)^j (\lambda(x_{n-j} - (a_1 - ja_2)) + (1-\lambda)x_{n-j} - (b_1 - jb_2))^2 \\ &\leq \sum_{j=0}^{n-1} (1-\alpha)^j [\lambda(x_{n-j} - (a_1 - ja_2))^2 + (1-\lambda)(x_{n-j} - (b_1 - jb_2))^2 \\ &= \lambda C(a_1,a_2) + (1-\lambda)C(b_1;b_2) \end{aligned}$$

D'où la fonction C est convexe. L'unique point critique est le minimum absolu.

Les formules mise à jour sont

3
$$\begin{cases} \widehat{a}_1(n) = \widehat{a}_1(n-1) + \widehat{a}_2(n-1) + (2\alpha - \alpha^2))(x_n - \widehat{x}_{n-1}(1)) \\ \widehat{a}_2(n) = \widehat{a}_2(n-1) + \alpha^2(x_n - \widehat{x}_{n-1}(1)) \end{cases}$$
avec l'initialisation
$$\begin{cases} \widehat{a}_1(0) = x_1 \\ \widehat{a_2}(0) = x_2 - x_1 \end{cases}.$$

Ces formules aident à comprendre comment calculer les coefficients, ce qui limite la complexité du calcul.

Méthode de Holt-Winters

1.Méthode non saisonnière

On cherche au voisinage de l'instant n, à ajuster une droite d'équation

$$y_t = a_1 + a_2(t-n)$$

. La prévision en $n+h$ sera

$$\widehat{x}(h) = \widehat{a}_1(n) + \widehat{a}_2(n)h$$

On choisit deux constantes de lissage α et β dans $]0,1[$. Les $\widehat{a}_1, \widehat{a}_2$ sont calculés récursivement par les équations suivantes :

4
$$\begin{cases} \widehat{a}_1(n) = \alpha x_n + (1-\alpha)(\widehat{a}_1(n-1) + \widehat{a}_2(n-1)) \\ \widehat{a}_2(n) = \beta(\widehat{a}_1(n) - (\widehat{a}_1(n-1)) + (1-\beta)\widehat{a}_2(n-1) \end{cases}$$

Remarque 3 :

Cette méthode est souple que la précédente. Le paramètre α joue un rôle dans l'estimée de l'ordonnée en n et β joue un rôle dans l'estimée de la pente. Plus α et β sont petits, plus on tient compte du passé lointain.

Méthode saisonnière additive

Au voisinage de n, on cherche à ajuster une courbe d'équation

$$y_t = a_1 + a_2(t-n) + s_t$$

(avec s périodique, de période T). On choisit α, β, γ dans $]0,1[$ et aussi T. Les formules de récursion sont

$$\begin{cases} \widehat{a}_1(n) = \alpha(x_n - \widehat{s}_{n-T}) + (1-\alpha)(\widehat{a}_1(n-1) + \widehat{a}_2(n-1) \\ \widehat{a}_2(n) = \beta(\widehat{a}_1(n) - (\widehat{a}_1(n-1)) + (1-\beta)\widehat{a}_2(n-1) \\ \widehat{s}_n = \gamma(x_n - \widehat{a}_1(n)) + (1-\gamma)\widehat{s}_{n-T} \end{cases}$$

La prévision prend la forme suivante :

$$\begin{cases} \widehat{x}_n(h) = \widehat{a}_1(n) + h\widehat{a}_2(n) + \widehat{s}_{n+h-t}, 1 \leq h \leq T \\ \widehat{x}_n(h) = \widehat{a}_1(n) + h\widehat{a}_2(n) + \widehat{s}_{n+h-2T}, T+1 \leq h \leq 2T... \end{cases}$$

Les valeurs initiales suivantes permettent de commencer le calcul à $T+2$

$$\begin{cases} \widehat{a}_1(T+1) = x_{T+1} \\ \widehat{a}_2(T+1) = \frac{x_{T+1} - x_1}{T} \\ \widehat{s}_j = x_j - (x_1 + (T-1)\widehat{a}_2(T+1)), 1 \leq j \leq T \end{cases}$$

Méthode saisonnière multiplicative

Au voisinage de n,on cherche à ajuster une courbe d'équation

$$y_t = (a_1 + a_2(t-n)) * s_t$$

(avec s périodique, de période T). On choisit α, β, γ dans $]0,1[$. Les formules de récursion sont

$$\begin{cases} \widehat{a}_1(n) = (1-\alpha)\dfrac{x_n}{\widehat{s}_{n-T}} + \alpha(\widehat{a}_1(n-1) + (\widehat{a}_2(n-1)) \\ \widehat{a}_2(n) = (1-\beta)(\widehat{a}_1(n) - (\widehat{a}_1(n-1)) + \beta\widehat{a}_2(n-1) \\ \widehat{s}_n = (1-\gamma)\frac{x_n}{\widehat{a}_1(n)} + \gamma\widehat{s}_{n-T} \end{cases}$$

La prévision prend la forme

$$\begin{cases} \widehat{x}_n(h) = (\widehat{a}_1(n) + h\widehat{a}_2(n)) \times \widehat{s}_{n+h-t}, 1 \leq h \leq T \\ \widehat{x}_n(h) = (\widehat{a}_1(n) + h\widehat{a}_2(n)) \times \widehat{s}_{n+h-2T}, T+1 \leq h \leq 2T \\ ... \end{cases}$$

1.3 Processus autoregressif à moyenne mobile(ARMA)"

Les définitions de cette section tirées de l'ouvrage [2] Yves ARAGON et l'ouvrage [5] de Bourbonnais et Terraza.

1.3.1 Processus AR(p) et MA(q)

Processus AR(p)

Soit $(X_t)_{t\in T}$ un processus stochastique. (X_t) est dit $AR(p)$ s'il obéit à

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \varepsilon_t = c + \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t$$

Où ε_t est un bruit blanc de moyenne nulle et de variance σ_ε^2, $\phi_p \neq 0$ et un processus ε_t est stationnaire.

Processus MA(q)

Soit $(X_t)_{t\in T}$ un processus. (X_t) est dit $MA(q)$ s'il obéit à

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q} = \mu + \sum_{j=0}^{q} \theta_j\varepsilon_{t-j}, \theta_q \neq 0$$

Où ε_t est un bruit blanc de moyenne nulle et de variance σ_ε^2

En introduisant l'opérateur moyenne mobile

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q,$$

on peut noter d'une façon équivalente :

$$X_t = \mu + \Theta(B)\varepsilon_t$$

Un $MA(q)$ est toujours stationnaire quelles que soient les valeurs de θ ; il est de moyenne μ

Processus ARMA(p,q)

Les processus ARMA sont des mélanges des processus AR et MA. Ils sont nécessairement, en pratique, finis.
X_t obéit à un modèle $ARMA(p,q)$ s'il est stationnaire et vérifie :

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} (*)$$

ε_t est un bruit blanc avec c constante arbitraire, $\phi_p \neq 0, \theta_q \neq 0$ et les polynômes

$$1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p \quad et \quad 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$$

n'ont pas de racines communes.
En utilisant l'opérateur retard, ce processus ARMA peut s'écrire comme suit :

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) X_t = c + (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)\varepsilon_t$$

X_t obéissant à l'égalité $(*)$ est stationnaire si les racines du polynôme dit d'autorégression $1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p = 0$ sont en module strictement supérieures à 1

1.4 Opérateurs sur les séries temporelles

Les définitions et les propriétés de cette section sont tirées de l'ouvrage [2] de Yves Aragon.

La manipulation pratique ou théorique des séries temporelles se trouve considérablement simplifiée par l'usage de l'opérateur retard (*Lag operator*). On donne dans la suite ses propriétés élémentaires.

1.4.1 L'opérateur retard

On note indifféremment B(Backward) ou L(lag), l'opérateur retard qui fait passer de X_t à X_{t-1} :

$$BX_t = X_{t-1}.$$

On a :

$$B^2X_t = B(BX_t) = BX_{t-1} = X_{t-2}.$$

Par l'établissement d'une relation de récurrence, on a

$$B^n = B(B(...(BX_t)) = X_{t-n}$$

1.4.2 L'opérateur différence

La différence première est

$$\Delta X_t = X_t - X_{t-1} = X_t - BX_t = (1-B)X_t,$$

c'est la série des accroissements, alors que la différence seconde Δ^2 donne la série des "accroissements des accroissements". On a :

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

Cet opérateur permet d'éliminer la composante tendancielle sans la calculer.

Elimination de la tendance

L'opérateur différence Δ élimine les tendances. Par exemple pour un processus de la forme :

$$X_t = a + bt + \varepsilon_t$$

On a $\Delta X_t = b + \varepsilon_t - \varepsilon_{t-1}$

1.4.3 L'opérateur différence saisonnière

Etant donné une série mensuelle, il peut être important d'en examiner les accroissements d'une année sur une autre (janvier sur janvier,...). L'opérateur différence saisonnière $\Delta_{12} = 1 - B^{12}$ est utile dans ce cas

$$\Delta_{12} X_t = (1 - B^{12})X_t = X_t - X_{t-12}$$

Elimination de la saisonnalité

L'opérateur de différence saisonnière noté Δ_s élimine une saisonnalité de période s. L'opérateur Δ_s associé au processus $X_t, t \in T$ est telle que :

$$\forall t \in T, \quad \Delta_s X_t = X_t - X_{t-s} = (1 - B^s) X_t$$

L'opérateur retard simplifie l'écriture des équations relatives aux séries. Il permet d'établir une équation de récurrence comme un polynôme de l'opérateur appliqué à une série.

Et par conséquent on peut obtenir l'opérateur de la d-ième différence noté Δ^d tel que :

$$\forall t \in T, \quad \Delta^d X_t = (1 - B)^d X_t$$

1.5 Séries non stationnaires

Les définitions utilisées dans cette section sont tirées de l'ouvrage [5]

Les chroniques économiques sont rarement des réalisations de processus aléatoires stationnaires. La non stationnarité des processus peut concerner aussi bien le moment du premier ordre (espérance mathématique) que celui du second ordre (variance et covariance du processus). Cette non stationnarité peut être repérée graphiquement (tendance, cycle long, saisonnalité,...) ou encore au moyen de la fonction d'auto-corrélation (fonction d'auto-corrélation lentement décroissante). Mais la plupart des résultats et des méthodes utilisées dans l'analyse des séries temporelles reposent sur la notion de stationnarité du second ordre, ce qui nous mène à appliquer à la chronique non stationnaire, certaines transformations (différence ordinaire, différence saisonnière,...) pour la rendre stationnaire. Parmi les processus aléatoires non stationnaires ; nous pouvons distinguer deux grandes classes, à savoir les processus TS(Stationnary Trend) et les processus DS(Differency Stationnary).

Définition et description des processus TS et DS

Définition du TS

Un processus TS est un processus s'écrivant comme suit : $X_t = f_t + \varepsilon_t$ où f_t est une fonction polynomiale qui dépend du temps, linéaire ou non linéaire, et ε_t est un processus stationnaire de type ARMA. Le processus TS le plus simple est représenté par un polynôme de degré 1. Ce processus s'écrit :$X_t = a_0 + a_1 t + \varepsilon_t$

Si ε_t est un bruit blanc, les caractéristiques de ce processus sont alors les suivantes :

$$\begin{cases} E(X_t) = a_0 + a_1 t + E(\varepsilon_t) = a_1 t + a_0 \\ Var(X_t) = 0 + Var(\varepsilon_t) = \sigma_s^2 \\ Cov(X_t, X_{t'}) = 0; \forall t \neq t' \end{cases}$$

Ce processus TS est non stationnaire car $E(X_t)$ dépend du temps. Etant donné que cette espérance est égale à $a_0 + a_1 t$, il s'agit à l'instant t d'un chiffre certain. Dans ce cas, on peut estimer de façon efficace les paramètres a_0 et a_1 de la tendance, en utilisant la méthode des moindres carrés ordinaires(MCO) sur les couples (X_t, t). Ces estimateurs peuvent être employés par suite pour réaliser une prévision de la série temporelle. Si on connait a_0 et a_1, le processus X_t peut être stationnarisé en retranchant la valeur estimée $\widehat{a}_0 + \widehat{a}_1 t$ de la valeur X_t en t.

Définition du DS

Les processus DS sont des processus non stationnaires que l'on peut rendre stationnaires par l'usage d'un filtre aux différences : $(1-B)^d X_t = \beta + \varepsilon_t$ où ε_t est un processus stationnaire de type ARMA ou encore bruit blanc, β une constante réelle et d l'ordre du filtre aux différences. Ces processus sont souvent représentés en utilisant le filtre aux différences premières $(d = 1)$. Le processus est dit alors processus du premier ordre. Il s'écrit : $(1-B)X_t = \beta + \varepsilon_t$;
$X_t = X_{t-1} + \beta + \varepsilon_t$
Où ε_t est un processus stationnaire de type bruit blanc(gaussien ou non). L'introduction de la constante β dans le processus DS fait qu'on définisse deux processus différents :
si $\beta = 0$; le processus DS est dit sans dérive et $\beta \neq 0$ le processus DS est dit avec dérive.

1.5.1 Processus autorégressif intégré à moyenne mobile ARIMA

Les modèles $ARIMA$ sont des modèles non stationnaires et ont une structure proche des modèles ARMA, ils sont intégrés et modélisables par des processus $ARMA$.

Définition

Un processus intégré est un processus qui peut être rendu stationnaire par différenciation. Si un processus doit être différencié d fois pour atteindre la stationnarité, il est dit intégré d'ordre d ou $I(d)$; par conséquent les processus stationnaires sont $I(0)$

1.5.2 Identification d'un modèle ARIMA(p,d,q)

Cette section est basée sur le mémoire [1] de Abdou Niandou Daouda.

La première étape dans l'analyse d'une série chronologique (c'est-à-dire la donnée de n réalisations $x_1, x_2, ..., x_n$ d'un processus stochastique) est de pouvoir trouver un modèle approprié qui représente la série. Ceci suppose, pour une modélisation ARMA, de chercher

les ordres p et q convenables. Néanmoins, cette étape dénommée identification du modèle doit être précédée d'une étude préalable de la stationnarité de la série observée. En cas de non stationnarité, il sera préconisé un modèle $ARIMA(p,d,q)$. Ce qui implique la détermination du degré d qui représente le nombre de fois qu'il est nécessaire de différencier la série pour aboutir à une série stationnaire.

Détermination du degré de différentiation

Dans la démarche de modélisation, on a besoin de vérifier au préalable si les données observées proviennent d'un processus stationnaire. Pour ce faire, on peut observer le comportement de la fonction d'autocorrélation empirique, et/ou procéder par des tests de stationnarité (tests de racine unitaire).

Approche par autocorrélogramme

On considère n observations $(x_1, x_2, ..., x_n)$ issues d'un processus $(X_t)_{t\in T}$. Une série stationnaire doit rester dans un intervalle borné et ne pas s'écarter durablement de la moyenne de la série. Donc si X_t est stationnaire, alors $\gamma(h) \to 0$ et $\rho(h) \to 0$ quand $h \to \infty$ avec une décroissance exponentielle. Telle est la morphologie d'une série stationnaire. Par conséquent, si le graphique d'une série ne satisfait pas à ce type de morphologie, on peut soupçonner d'une non stationnarité. Un autre argument utilisé dans cette approche se base sur la fonction d'autocorrélation (voir [9]). Si $(X_t)_{t\in T}$ n'est pas stationnaire, alors sa fonction d'autaucorrélation va dépendre du temps et on peut l'écrire comme suit :

$$\rho(h,t) = \frac{Cov(X_t, X_{t+h})}{(V(X_t)V(X_{t+h}))^{\frac{1}{2}}}$$

Si on s'intéresse au comportement de cette fonction au voisinage de l'infini on voit que : $\rho(h,t) \to 1$ quand $t \to \infty$. Ainsi, si la fonction d'autocorrélation estimée $\widehat{\rho}(h)$reste relativement proche de 1 pour un assez grand nombre de valeurs de h, on peut penser à une non stationnarité du processus. En pratique, on préfère le critère de proximité des valeurs de $\widehat{\rho}(h)$ entre elles par rapport au critère de proximité de 1 des valeurs de $\widehat{\rho}(h)$). Donc on peut différencier la série pour la rendre stationnaire dès qu'on observe que les premiers coefficients de $\widehat{\rho}(h)$ sont suffisamment proches les uns des autres même si le coefficient $\widehat{\rho}(1)$ est assez différent de 1. Dès qu'on opte pour la différenciation, on applique le critère ci-dessus à la série différenciée pour savoir s'il convient de la différencier une seconde fois et ainsi de suite. En pratique, les valeurs de d sont le plus souvent égales 0 ou 1 et quelque fois 2. La série différenciée d fois peut alors être modélisée par un ARMA et le modèle résultant sera un $ARIMA(p,d,q)$.

1.5.3 Tests de la non stationnarité

Ce sont des tests qui permettent de détecter la présence ou l'absence d'une racine unitaire qui caractérise les processus non stationnaires de type DS.

Test de Dickey-Fuller

Dickey et Fuller ont construit leur test à partir des modèles de base suivants :

- Modèle(1) : $X_t = \rho X_{t-1} + \varepsilon_t$ AR(1)
- Modèle(2) : $X_t = c + \rho X_{t-1} + \varepsilon_t$ AR(1) avec constante.
- Modèle(3) : $X_t = c + bt + \rho X_{t-1} + \varepsilon_t$ AR(1) avec constante et tendance.

.

Où ε_t est un bruit blanc de variance σ^2 et b, c des constantes réelles. Le principe du test consiste à tester l'hypothèse nulle de racine unitaire contre l'hypothèse alternative d'absence de racine unitaire. $\begin{cases} H_0 : \rho = 1 \\ H_1 : \mid \rho \mid < 1 \end{cases}$

Les modèles de base du test étant théoriques, l'application du test requiert l'estimation en pratique de modèles :

- Modèle$(1)^{'}$: $\Delta X_t = \phi X_{t-1} + \varepsilon_t$
- Modèle$(2)^{'}$: $\Delta X_t = c + \phi X_{t-1} + \varepsilon_t$
- Modèle$(3)^{'}$: $\Delta X_t = c + bt + \phi X_{t-1} + \varepsilon_t$

. Où on pose $\phi = \rho - 1$ dans tous les trois modèles. On calcule la t-statistique $t_{\widehat{\phi}}$ qui est donnée par :

$$t_\phi = \frac{\phi - 1}{\widehat{\sigma}^2}$$

$t_{\widehat{\phi}}$ sera comparée à la valeur critique tabulée notée t_{tab} et on applique la règle suivante :

Si $t_{\widehat{\phi}} < t_{tab}$ on rejette H_0

Si $t_{\widehat{\phi}} \geq t_{tab}$ on accepte H_0

En pratique, on n'effectue pas ce test sur les trois modèles mais on procède par une stratégie séquentielle en trois étapes suivantes :

Etape 1

On estime le modèle $(3)^{'}$ et on teste la significativité de la tendance déterministe(test de Student sur le paramètre b).

• Si cette tendance estimée n'est pas significativement différente de zero(donc la t-statistique de la tendance est inférieure aux valeurs critiques de la tendance tabulée par **Dickey-Fuller**) alors on passe à l'étape 2.

• Si la tendance est différente de zero, on teste l'hypothèse nulle unitaire :

Si on accepte H_0, X_t est non stationnaire de type DS
Si on rejette H_0, X_t est non stationnaire de type TS.

Etape 2

On aura à appliquer cette étape que si à l'étape 1 on a rejeté l'idée d'une tendance significative. On estime le modèle $(2)'$ et on teste la sigificativité de la constante c.

• Si H_0 est acceptée X_t est non stationnaire de type DS
• Si H_0 est rejetée X_t est stationnaire.

Etape 3

Si l'étape 2 detecte une constante nulle, alors on estime le modèle(1) et on effectue le test de racine unitaire tel que :

• Si H_0 est acceptée X_t est non stationnaire de type DS
• Si H_0 est rejetée X_t est alors stationnaire.

Test de Dickey-Fuller augmenté

Ce test se déroule exactement comme dans le cas de la version simple, à la seule différence que les modèles de base à la construction de ce test sont les suivants :

- Modèle(1) : $\Delta X_t = \phi X_{t-1} + \sum_{j=1}^{p} \phi X_{t-j} + \varepsilon_t$
- Modèle(2) : $c + \Delta X_t = \phi X_{t-1} + \sum_{j=1}^{p} \phi X_{t-j} + \varepsilon_t$
- Modèle(3) : $c + bt + \Delta X_t = \phi X_{t-1} + \sum_{j=1}^{p} \phi X_{t-j} + \varepsilon_t$

.

Test de Dickey et Pantula(1987)

Ce test est tiré de l'ouvrage [5] Bourbonnais et Terraza.

En passant au test de Dickey-Fuller, l'existence d'une racine unitaire conduit donc à la différenciation de la série en question. Une fois la série différenciée, on peut se demander si la nouvelle série obtenue après différenciation est stationnaire ou non. On applique donc à nouveau le test de Dickey-Fuller et ainsi de suite. Selon Dickey et Pantula, cette procédure dite séquentielle ascendante peut donner des résultats faux car les distributions statistiques diffèrent suivant qu'il existe une ou deux racines unitaires. Ainsi, ils ont proposé, en se réferant aux tables de Dickey-Fuller, une nouvelle procédure dite séquentielle descendante. Cette procédure permet de tester en même temps l'existence de plusieurs racines unitaires. Supposons que l'on veuille tester l'existence de deux racines unitaires, alors le test se basera sur le modèle suivant :

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-1} + \phi_1 \phi_2 X_{t-2} + \varepsilon_t$$

Et en pratique on estime le modèle suivant : $\Delta^2 X_t = \phi_2 \Delta X_{t-1} + \phi_1 X_{t-1} + \varepsilon_t$ Où l'on pose $\theta_1 = -(\phi_1 - 1)(\phi_2 - 1)$ et $\theta_2 = (\phi_1 \phi_2 - 1)$. Le test se fait en deux étapes suivantes :

Etape 1

On teste l'hypothèse nulle de deux racines unitaires contre l'alternative d'une seule racine unitaire

$$\begin{cases} H_0 : \theta_1 = \theta_2 = 0 \\ H_1 : \theta_1 = 0 \end{cases}$$

Sous l'hypothèse H_1, on estime $\Delta^2 X_t = \theta_2 \Delta X_{t-1} + \varepsilon_t$ et on fait l'usage de la règle suivante :

• Si $t_{\theta_2} \leq t_{tab}$ on rejette H_0 et on passe à l'étape 2

• Si $t_{\theta_2} > t_{tab}$ on accepte H_0, le processus contient donc deux racines unitaires.

Etape 2

Si à l'étape 1 on rejette l'hypothèse H_0, alors on peut appliquer l'étape 2. On teste donc l'hypothèse nulle de la présence d'une racine unitaire contre l'hypothèse alternative d'aucune racine unitaire dans $\Delta X_t = (\phi_1 - 1) X_{t-1} + \varepsilon_t$. Ce test sur $(\phi_1 - 1)$ est l'équivalent de celui sur θ_1 dans le modèle $\Delta^2 X_t$.

Dickey et Fuller ont tabulé les valeurs des t-statistiques $t_{\widehat{\phi}}$ selon les trois modèles. Il existe beaucoup d'autres tests de stationnarité entre autres le test de Phillips et Perron, le test KPSS(1992),...)

1.5.4 La méthodologie de Box et Jenkins

Cette section est tirée de l'ouvrage [5] de Bourbonnais et Terraza et du mémoire [11] de Oukacha et Lounis.

Box et Jenkins(1970) ont proposé une méthodologie de modélisation d'une série chronologique univariée basée sur les modèles linéaires ARMA, ARIMA. Cette méthodologie possède trois étapes : identification, estimation et validation. Il s'agit tout d'abord d'étudier la courbe représentative pour repérer la saisonnalité et la tendance éventuelle. Dans la phase d'identification, on doit identifier les ordres de différenciation et de saisonnalité puis les paramètres p et q en faisant appel à des outils tels que les fonctions d'auto-corrélations. A la fin de cette méthode on a plusieurs modèles parmi lesquels il faudra choisir lors des phases suivantes. Puis la phase d'estimations des différents modèles retenus pour les valeurs de p, q, d. Plus précisément on estime les coefficients. Dans certains cas on utilise la méthode de maximum de vraisemblance. Enfin, on soumet les différents ajustements à un certain nombre de tests et on applique un certain nombre de critères(Critères d'Akaike (AIC) et le critère bayesien (BIC) pour choisir le modèle final retenu. L'intérêt de cette approche est qu'une modélisation

ARMA conduit à des prévisions optimales si la variance de l'erreur de prévision est minimale.

Stationnarisation

Pour stationnariser un processus TS, la bonne méthode est celle des moindres carrés ordinaires ; pour un processus DS il faut utiliser le filtre aux différences. L'opérateur de différentiation $\Delta = 1 - B$ tel que l'opérateur retard B élimine les tendances, d est estimé en effectuant des tests de stationnarité sur la série brute puis sur les séries résiduelles. Cette opération rend la série stationnaire et donne une estimation du nombre d. La série résiduelle supposée stationnaire sera modélisée par un $ARMA(p, q)$.

Identification du modèle

En premier lieu, on examine le graphe représentatif de la série temporelle, ceci peut donner une idée préliminaire sur le comportement de la série (stationnarité, tendance, saisonnalité ,...). Si la série présente une tendance et/ou une saisonnalité , des transformations adéquates doivent être appliquées afin de stationnariser la série. L'idée générale de l'identification dans la méthodologie Box-Jenkins, consiste à comparer la structure des corrélations estimées que présente la série à travers le corrélogramme (diagramme représentatif des autocorrélations estimées) avec la structure de corrélation théorique exhibée par des modèles bien connus. Ainsi l'étude du corrélogramme est très utile pour la détermination des ordres p et q, puisque les fonctions d'autocorrélation simples et partielle peuvent indiquer la présence d'un modèle moyenne mobile ou auto-régressif respectivement. Plus précisément si la fonction d'autocorrélation simple décroit rapidement vers 0 et la fonction d'autocorrélation partielle présente un cut-off après p retard, on peut conclure que la série provient d'un processus AR d'ordre $p(AR(p))$. Si la fonction d'autocorrélation simple présente un cut-off après q retards et que la fonction d'autocorrélations partielles décroit rapidement vers 0, alors on peut conclure que la série est générée à partir d'un modèle moyenne mobile d'ordre $q(MA(q))$. On note que si les fonctions d'autocorrélations simples et partielles présentent une forme exponentielle ou sinusoïdale, on constate qu'on est en présence d'un processus auto-régressif à moyenne mobile $ARMA(p, q)$. Cette étape n'est pas aisée et demande beaucoup d'expertise, il existe cependant des méthodes d'identifications automatiques, basées sur le critère d'information.

Critère d'information

Il existe des critères d'informations qui sont utilisés, comme guide, dans le choix du modèle, ce qui nous permet d'éviter la sélection arbitraire des paramètres p et q du modèle. Parmi ces critères, il existe les critères d'information qui mesurent l'écart entre la vraie loi inconnue et celle du modèle proposé ; les estimations de la qualité d'information qui ont été proposées sont :

1. Critère d'Akaike(1969) appelé aussi AIC, il est défini comme suit :

$$AIC(p,q) = log\sigma^2 + \frac{2(p+q)}{N}$$

2. Critère baysien(1977) appelé aussi BIC, il est défini comme suit :

$$BIC(p,q) = log\sigma^2 + 2(p+q)\frac{log(N)}{N}$$

Remarque

Il y a d'autres critères qui sont orientés surtout vers la mesure de la performance prévisionnelle des modèles. Ces critères sont appelés critères de pouvoir prédictif. On peut citer :

1. Le coefficient de détermination R^2

$$R^2 = 1 - \frac{\sigma^2}{V}$$

   Avec V la variance de la série initiale

2. La statistique de Fisher F

$$F = \frac{\frac{V-\sigma^2}{(p+q)}}{\frac{\sigma^2}{N-p-q}}$$

Ces critères doivent être maximisés et le modèle qui a la meilleure performance prévisionnelle est celui qui rend maximal l'un ou les critères considérés.

Estimation des paramètres et validation du modèle

Les définitions,les tests et les remarques dans cette section sont basées de l'ouvrage [5] de Bourbonnais et Terraza.

Estimation des paramètres

Après avoir terminé l'identification, il convient d'estimer les paramètres qui sont les coefficients des polynômes AR et MA et la variance des résidus ε_t. La méthode d'estimation la plus utilisée est celle du maximum de vraisemblance ou la méthode des moindres carrés .Le principe consiste à construire une fonction dite de fonction de vraisemblance et par la suite à maximiser son logarithme par rapport aux paramètres θ_i, θ_j,(avec $i = 1,...,p; j = 1,...,q$),

permettant ainsi de trouver la valeur numérique la plus vraisemblable pour ces paramètres. L'étape d'estimation finie, l'étape suivante va nous permettre de valider le modèle estimé.

Vérification et validation

Au début de cette étape on dispose de plusieurs processus ARMA dont on a estimé les paramètres. Il faut maintenant valider ces modèles afin de les départager. Pour cela, on applique des tests sur les paramètres et sur les résidus. Si plusieurs modèles sont validés, l'étape de validation doit se poursuivre par une comparaison de qualité de ces derniers.

Tests concernant les paramètres

Après avoir estimé les paramètres d'un modèle, on peut se poser la question de savoir si ces paramètres sont significativement différents de zéro. Ces tests sont aussi appelés tests sur le modèle, car si le test de significativité des paramètres détecte des paramètres non significatifs, cela entraine automatiquement un changement dans l'ordre du modèle. Soit un modèle $ARMA(p;q)$.

1. On peut par exemple tester $p^{'} = p-1$ et $q^{'} = q-1$. Ici il est question de savoir si, on peut diminuer d'une unité l'ordre de la partie AR. Pour cela, on utilise le test de Student qui va tester la signigicativité du coefficient ϕ_p. Soit $\widehat{\phi}_p$ l'estimateur de ϕ_p et $\widehat{V}(\widehat{\phi}_q)$ sa variance estimée. En supposant que les estimateurs sont normalement distribués au risque de 0.05 on compare la valeur de la statistique t_c donnée par,

$$t_c = \frac{\mid \widehat{\phi} \mid}{(V(\widehat{\phi}_p))^{\frac{1}{2}}}$$

   à la valeur critique 1.96. Si t_c est supérieure à 1.96, on rejette l'hypothèse $\phi_p = 0$. Dans le cas contraire on accepte l'hypothèse de nullité de ϕ_p.

2. On peut également tester $p = p+1$ et $q^{'} = q$. Dans ce cas, il s'agit de connaître la possibilité d'augmenter l'ordre de la partie autoregressive. Donc il faut tester la significativité du coefficient ϕ_{p+1}. Comme précédemment, on utilise la statistique de Student. On compare alors le rapport

$$t_c = \frac{\mid \phi_{p+1} \mid}{(\widehat{V}(\widehat{\phi}_{p+1}))^{\frac{1}{2}}}$$

   à 1.96 (au risque de 0.05)

   Tests concernant le bruit blanc

   L'une des hypothèses qui doivent être vérifiées de manière rigoureuse pour la validité d'un modèle ajustée est celle de bruit blanc. En effet, si les résidus ne forment pas un bruit blanc on peut penser à une mauvaise stationnarisation des données ou même à un mauvais choix du modèle.

(a) **Test Portmanteau (Box et Pierce(1970))**

C'est un test qui permet de déduire si vraiment les résidus forment un bruit blanc. Il se base sur le fait que la fonction d'autocorrélation empirique d'un bruit blanc ne doit pas révéler d'autocorrélation non significativement différente de zéro. La statistique du test est donnée par :

$$Q = n \sum_{h=1}^{K} \widehat{\rho}_{\varepsilon}^{2}(h)$$

(Somme des carrés des autocorrelations empiriques du bruit blanc ε_t) K est un entier(choisi en général entre 15 et 20). On teste alors :
•$H_0 : \rho_\varepsilon(1) = \rho_\varepsilon(2) = ... = \rho_\varepsilon(K) = 0$
•$H_1 : \exists$ au moins un j tel que $\rho_\varepsilon(j) \neq 0$

On accepte H_0 si $Q \leq \chi^2_{1-\alpha}(K-p-q)$ et on rejette H_0 si $Q > \chi^2_{1-\alpha}(K-p-q)$.

(b) **Test de Durbin-Watson**
Ce test permet de détecter la présence d'une corrélation à l'ordre 1, sous la forme

$$\widehat{\varepsilon}_t = \rho \widehat{\varepsilon}_{t-1} - \mu_t$$

où

$$\mu_t \sim N(0, \sigma_\mu^2)$$

Et $\widehat{\varepsilon}_t = y_t - \widehat{y}_t$ est le résidu de l'estimation du modèle, $\widehat{y}_t$ la prévision de y_t faite à l'instant $t-1$.

Donc il s'agit de tester

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

La statistique du test est donnée par : $DW = \frac{1}{\sum_{t=1}^{n} \widehat{\varepsilon}_t^2} \sum_{t=2}^{n} (\widehat{\varepsilon}_t - \widehat{\varepsilon}_{t-1})^2$
DW est comprise entre 0 et 4 et vaut 2 lorsque $\rho = 0$. Durbin et Watson ont établi des valeurs critiques de DW au seuil de 0.05 en fonction de la taille de la série de variables explicatives.

(c) **Test de Ljung et Box(amélioration du test du Portemanteau)**
Ce test est d'une part comme une amélioration du test de Box et Pierce. D'autre part, si on ne connaît rien sur la structure du processus comme souvent c'est le cas, le test de Ljung-Box est plus général que le test de Durbin-Watson. Il est basé

sur la statistique :

$$Q^{'} = n(n+2)\sum_{h=1}^{K}\frac{\widehat{\rho}_{\varepsilon}^{2}(h)}{n-h}$$

$Q^{'}$ suit une loi de χ^2 à $(K-p-q)$degrés de liberté et le test se déroule comme celui de Box-Pierce.

On peut résumer la méthodologie de Box et Jeninks par le schéma ci-dessous.

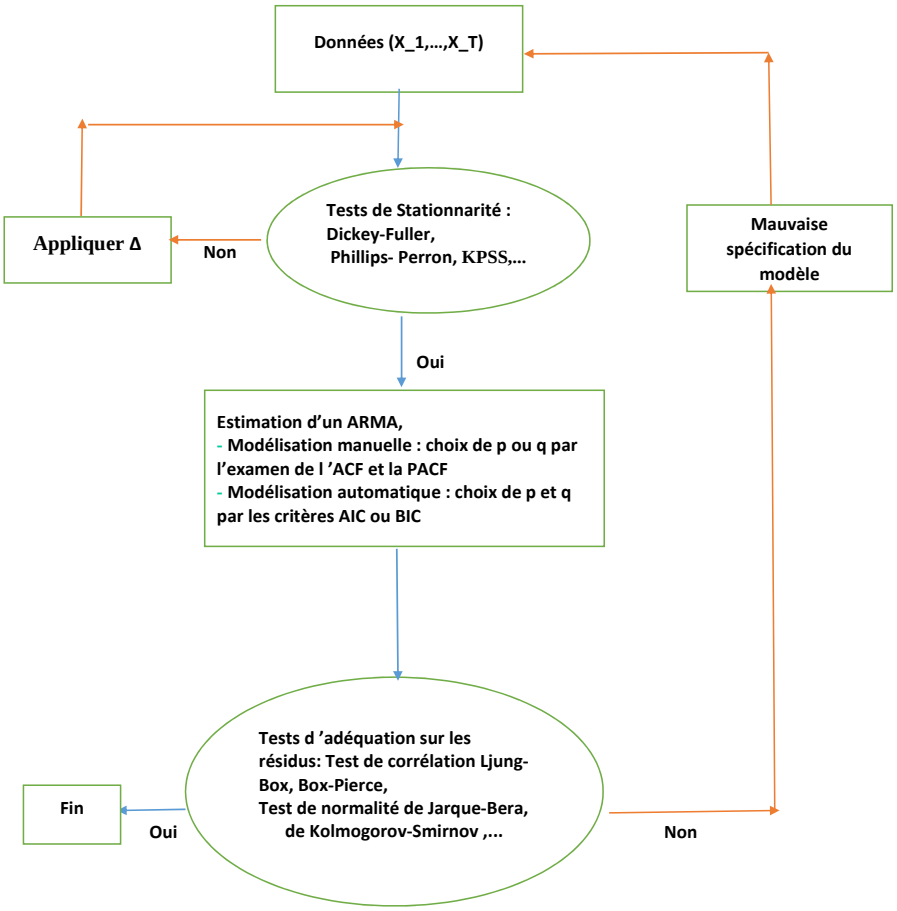

FIGURE 1.1 – Schéma général de la modélisation d'une série temporelle par un modèle ARIMA.

Prévision

Soit $(X_t)_{t\in T}$, un processus au second ordre réel et centré. On a que la fonction d'autocovariance est notée $\gamma(i,j) = E(X_iX_j)$. Soit $X_1, X_2, ..., X_n$ un échantillon de $(X_t)_{t\in T}$, on note $H_n = X_1, X_2, ..., X_n$ le sous espace fermé de $L^2(\Omega, A, P)$ engendré par $(X_t)_{1\leq j\leq n}$. On pose que $\widehat{X} = 0$ et $\widehat{X}_j = H_{n-1}(X_j)$. On suppose que la matrice $K(i,j)$ est définie positive, on montre $H(n) = [(X_1 - \widehat{X_1}), (X_2 - \widehat{X_2}), ..., (X_n - \widehat{X_n})]$

Et on déduit que $\widehat{X}_{n+1} = \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \widehat{X}_{n+1-j})$ pour $n \geq 1$.

Pour des prévisions d'ordre $h \geq 1$

$$\widehat{X}_{n+h} = \sum_{j=1}^{n+h-1} \theta_{n+h-1,j}(X_{n+1-j} - \widehat{X}_{n+1-j})$$

Si on écrit (X_t) sous forme d'une moyenne mobile infinie :

$$X_t = \varepsilon_t + \sum_{j=1}^{+\infty} (b_j \varepsilon_{t-j})$$

On a l'intervalle de confiance au seuil de $\alpha = 0.05$ d'où

$$X_t \in [\widehat{X}_{t+h} - 1.96\sigma_\varepsilon(\sum_{j=0}^{h-1} b_j^2), \widehat{X}_{t+h} + 1.96\sigma_\varepsilon(\sum_{j=0}^{h-1} b_j^2)]$$

Chapitre 2

Partie Application

2.1 Objectifs du sujet

2.1.1 Objectif principal

Proposer un modèle mathématique d'évolution de la pathologie cancéreuse colorectal au niveau de la wilaya de Tizi Ouzou

2.2 Matériels et méthodes

2.2.1 Type d'étude

Étude d'une série chronologique (temporelle) appliquée sur la pathologie cancéreuse colorectale de la wilaya de TIZI OUZOU.
Le choix a été porté sur cette pathologie du fait de sa chronicité, sa lourdeur de prise en charge, sa mortalité et sa fréquence sur l'échelle nationale et mondiale.

2.2.2 Population d'étude

Le choix a été porté sur une population de patients détectés à la wilaya de TIZI OUZOU atteints du cancer colo-rectal, hospitalisés et/ou suivis en consultations dans les structures de la santé publiques ou privées du territoire national durant la période 2003 à 2011.

2.3 Source de données

Les cas ont été récupérés sur la base des données du registre des tumeurs de la wilaya domicilié au SEMEP du CHU de TIZI OUZOU.
Le registre des cancers de la wilaya de Tizi Ouzou, constitue une source de données. Ces

dernières sont recueillies sur tous les nouveaux cas de cancer survenant dans une population géographiquement définie, il est dirigé par le Pr Toudeft, depuis 2003.

Le registre est l'un des outils indispensables de la santé environnementale et de l'épidémiologie du cancer, c'est une structure qui réalise un recueil systématique, continu et exhaustif de données nominatives concernant les individus atteints de cancer dans une population géographiquement définie, à des fins de recherche et de santé publique. Il permet, via des études statistiques et épidémiologiques, de :

1. Estimer l'incidence et son évolution dans l'espace et le temps pour mieux estimer les besoins en matière de prévention, diagnostic et soins,
2. Mesurer mieux la gravité d'une situation régionale ou locale
3. Comprendre mieux les causes de certains cancers
4. Comprendre mieux et prévenir les facteurs de risques, ou conditions d'apparition de certains cancers
5. Détecter de manière plus précoce l'émergence de nouveaux cancers ;
6. Évaluer l'efficacité d'actions préventives ou curatives.

2.4 Description du lieu de déroulement

L'hôpital NEDIR Mohamed a été inaugurée ; précisément le 28 juillet 1955. A cette époque, ce dernier comportait un nombre restreint de disciplines médicales. En 1974, hôpital régional de Tizi Ouzou devient un secteur sanitaire grâce aux différentes unités de santé qui lui étaient reliées. En 1982, le secteur sanitaire de Tizi Ouzou se voit transformer en Secteur Sanitaire Universitaire (SSU) et ceci par l'ouverture de la formation biomédicale pluridisciplinaire. Le CHU est une institution publique à caractère administratif rattaché au ministre de la santé, crée par le décret n°86/25 du 11 Février 1986, complété et modifier par Le décret n°86/294 du 16 décembre 1986. Le siège du CHU de Tizi Ouzou est fixé à l'hôpital NEDIR Mohamed. Il est composé de deux unités, l'une sise au centre ville dénommé NEDIR Mohamed, l'autre à 4 Km du chef-lieu de la wilaya, hôpital Sidi Balloua, sise Redjaouna (Sanatorium). L'unité NEDIR Mohamed est chargée en relation avec l'établissement d'enseignement et de formation supérieur en sciences médicales concerné, des missions de diagnostic, d'exploration, de soin, de prévention, de formation, d'études et de recherche.
(source :site officiel de CHU Tizi Ouzou)

1. **En matière de santé** elle est chargée de :
   (a) Assurer les activités de diagnostic, de soins d'hospitalisation et des urgences médico-chirurgicales, de prévention ainsi que de toute activité concourant à la protection et à la promotion de la santé de la population.

(b) Appliquer les programmes nationaux, et locaux de santé.

(c) Contribuer à la promotion et à la protection de l'environnement dans les domaines relevant de la prévention, de l'hygiène, de la salubrité et de la lutte contre les nuisances et les fléaux sociaux.

2. **En matière de recherche :**

(a) Effectuer, dans le cadre de réglementation en vigueur, travaux d'études et de recherche dans les domaines des sciences de santé.

(b) Organiser des séminaires, colloques, journées d'études et autre manifestation techniques et scientifiques en vue de promouvoir les activités de soins, de formation et de recherche en science de santé.

2.4.1 Le Service d'Epidémiologie et de Médecine Préventive du CHU de TIZI OUZOU

Le service épidémiologie est domicilié au niveau de l'unité Nedir Mohamed du CHU de Tizi Ouzou.

Définition du mot "épidémiologie"

L'origine grecque du mot épidémiologie est simple : EPI - veut dire « sur » ; DEMOS - veut dire « peuple - population » ; LOGOS - veut dire « Étude ou connaissance » ; Par conséquent : l'épidémiologie est l'étude de ce qui arrive aux individus.

Avant,l'épidémiologie ne s'intéressait qu'aux maladies infectieuses et épidémiques, avec l'apparition d'études sur les maladies non transmissibles l'épidémiologie est considérée comme une discipline à part entière de la médecine. La méthodologie épidémiologique s'est même élargie à d'autres domaines même en dehors de la médecine.

Les tâches d'épidémiologie

Les tâches en épidémiologie :

1. Surveillance épidémiologique : pour reconnaître l'existence d'un problème de santé dans la communauté. Ceci nécessite un système de recueil de données spécifique et sensible qui procure des informations rapides et sûres permettant de donner l'alerte sur un problème réel ou potentiel.
2. Enquête épidémiologique : La surveillance épidémiologique ayant fourni des faits prouvant l'existence d'un problème de santé, l'enquête va permettre de rechercher les circonstances de survenue du problème.
3. Analyse épidémiologique : Suite à l'enquête et à la collecte des données, il s'agit : D'analyser les données ; De tirer les conclusions ; De faire des recommandations pour la prévention et la lutte contre les maladies.

4. Recherche scientifique.
5. Évaluation : L'évaluation des techniques de prévention et de lutte, des modalités thérapeutiques et des interventions utilisées pour la décroissance de la mortalité et de la morbidité sont sous la responsabilité de l'épidémiologique.
6. Information-Communication :les conséquences et des résultats des investigations épidémiologiques doivent être communiquées par les épidémiologistes.

2.4.2 Durée de l'étude

six mois de travail du mois d'avril à septembre 2017

2.5 Moyens

2.5.1 Moyens humains

Il s'agit de deux étudiants du Département de Mathématiques de l'Université Mouloud Mammeri de Tizi Ouzou, en master2, spécialité mathématiques appliquées à la gestion.

2.5.2 Moyens matériels

- Microordinateurs.
- Logiciels informatiques : SPSS20, Excel 2013, Langage R 3.2.3, LaTex

1. **Langage R** :est un système interactif et convivial de calcul et de visualisation graphique destiné aux scientifiques. Il est communément appelé langage et logiciel, il permet de réaliser des analyses statistiques. En particulier, il comporte des moyens qui rendent possibles la manipulation des données, les calculs et les représentations graphiques. R a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes. En effet R possède :
   - Un système efficace de manipulation et de stockage des données,
   - Différents opérateurs pour le calcul sur tableaux, en particulier les matrices,
   - Un grand nombre d'outils pour l'analyse des données et les méthodes statistiques,
   - Des moyens graphiques pour visualiser les analyses,
   - Un langage de programmation 'a la fois puissant et simple d'utilisation qui intègre des fonctions d'analyse de calcul matriciel, . . . etc.
2. **Logiciel SPSS20 :** (Statistical Package for the Social Sciences) : est un logiciel utilisé pour l'analyse statistique. C'est aussi le nom de la société qui le revend(SPSS Inc). c'est un logiciel spécialisé de traitement statistique comprenant les modules suivants : système de base, modèle de régression(regression models), modèles avancés,

tableaux(tables), tests exacts, catégories, tendances(trend), autres modules spécialisés. Il peut faire la saisie des données et la gestion des bases de données, le traitement, l'analyse des données et le traitement graphique des résultats.

3. **LaTeX :** est un langage de mise en page permettant de produire des documents de format pdf d'une grande qualité typographique et rigoureusement homogènes dans leur présentation.

2.6 Déroulement

2.6.1 Phase préparatoire

Une demande de stages a été déposée au niveau du service Épidémiologie au CHU de Tizi Ouzou et un avis favorable fut accordé par le responsable du dit service afin de pouvoir commencer le travail concernant l'étude de l'évolution de la pathologie cancéreuse.

2.6.2 Phase de réalisation

Après la récupération des données du registre sur logiciel SPSS, il y a eu vérification des données, codage et organisation en fonction du sexe, années et mois. Pour chaque année, les cas indéterminés(cas dont la date de diagnostic n'est pas déterminée) ont été répartis selon le sexe en vérifiant les prénoms des patients. Après, ces cas ont été répartis d'une façon aléatoire c'est-à-dire ils ont été ajoutés à des mois où il y avait peu de cas. Pour les données globales, le nombre de cas(féminin et masculin) a été donné par combinaison du nombre de cas féminin et le nombre de cas masculin. On les a converties par Excel en format csv pour pouvoir les importer par le logiciel R sous forme de data.frame, afin de les utiliser pour proposer le modèle mathématique que suit l'évolution du cancer colo-rectal.

2.7 Analyse de données

Variables : Les variables utilisées sont les variables qualitatives telles que le sexe des patients et les variables quantitatives telles que les dates de diagnostic, l'année de diagnostic, le nombre de cas.

2.7.1 Tests utilisés

Pour l'étude de la série de l'évolution du cancer colo-rectal, nous avons utilisé les tests suivants :

1. Test de Student

2. Test de Fisher
3. Test de Portemanteau, qui se présente comme un test du khi-deux.
4. Test de normalité des résidus
5. Test de Dickey-Fuller augmenté pour tester la stationnarité.

2.8 Plan d'analyse

Caractéristiques de la population

Pathologie= Cancer colo-rectal

L'analyse a été faite en fonction :

1. sexe
2. âge moyen
3. le nombre de cas par année
4. le nombre de cas par année et par sexe

Dans notre travail seul le caractère "Nombre de cas mensuels" a été pris en charge pour étudier la série afin de faire des prévisions.

2.9 Étude de la série chronologique et résultats

2.9.1 Description de la population :

Dans notre série on a enregistré 1068 cas dont 50.5 pourcent de sexe féminin, avec un âge moyen de 59.35 et un écart-type de 16 ans.

Année	Effectifs(N)	Pourcentage(%)
2003	106	9,9
2004	126	11,8
2005	41	3,8
2006	104	9,7
2007	99	9,3
2008	176	16,5
2009	181	16,9
2010	133	12,5
2011	102	9,6
Total	1068	100,0

FIGURE 2.1 – Taux des cas de cancer colo-rectal en fonction

Le pourcentage de cancer colo-rectal varie de 3.8 pourcent en 2005 à 16.9 en 2009.

Dans cette partie nous allons appliquer les outils mathématiques sur les données de l'évolution du cancer dans la région de Tizi Ouzou entre l'année 2003 et 2011. Notre but dans cette étape d'étude de la série qui sera notée par "cancer", est d'appliquer différentes méthodes de prévision sur la série.

Présentation de la chronique

La chronique de l'évolution du cancer colo-rectal au de la wilaya de Tizi-Ouzou est représentée par la figure ci dessous :

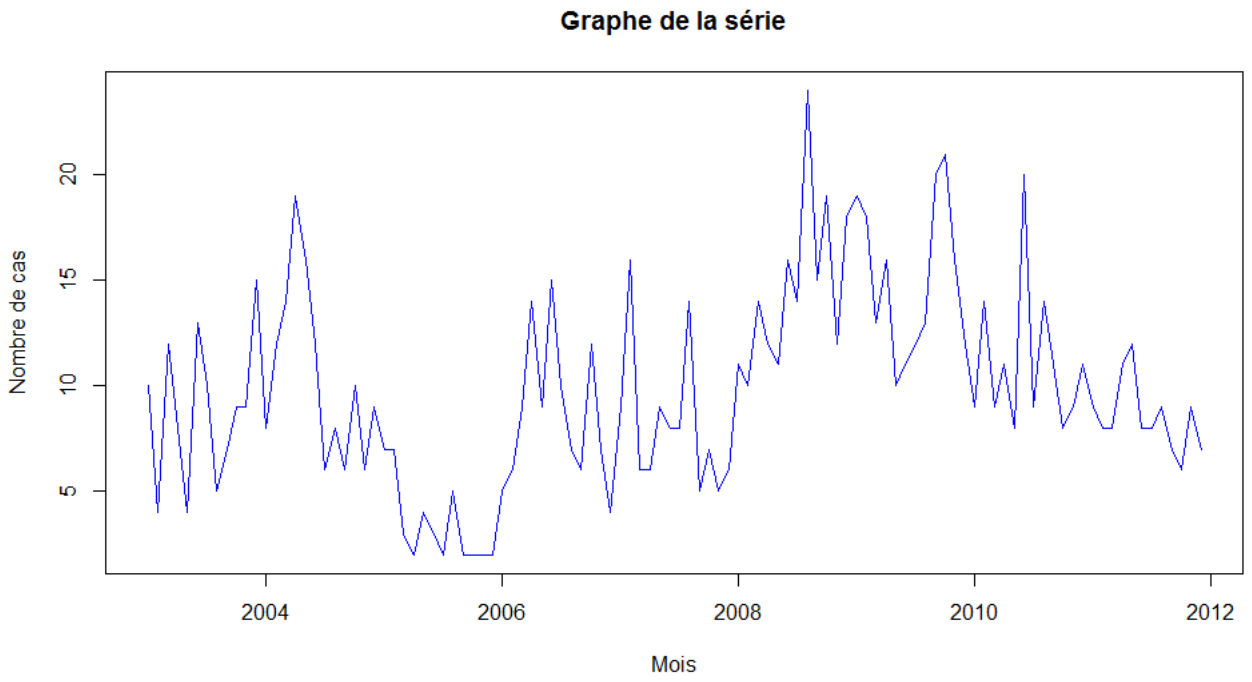

FIGURE 2.2 – Evolution du nombre des cas du cancer colo-rectal de la wilaya de Tizi Ouzou de 2003 à 2011

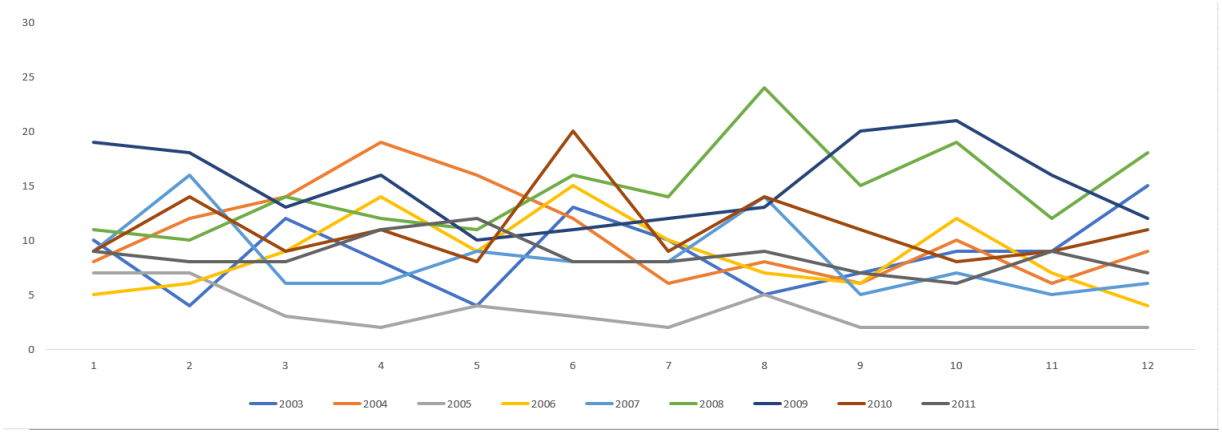

FIGURE 2.3 – Evolution du cancer colo-rectal par année

2.9.2 Etude de la saisonnalité et la tendance

En examinant le chronogramme de la série il n'est pas évident de déceler l'existence de la saisonnalité et l'existence de la tendance. Par ailleurs, on va faire l'analyse de la variance qui est un test de comparaison des moyennes des échantillons issus d'une même population. Et on va appliquer le test de Fisher. Ce test sera fait avec le langage R suivant le programme ci-dessous.

Calcul des moyennes générale, par année et par le mois

On calcule les ces moyennes pour pouvoir faire le test de l'analyse de variance.

```
> ##Calcul de la moyenne générale##
> MoyGen=mean(P)
> MoyGen
[1] 9.888889
> ##Calcul de moyenne période##
> moyennemois=c(apply(P,2,mean))
> moyennemois
 [1]  9.666667 10.555556  9.777778 11.000000  9.222222 11.777778  8.777778
 [8] 11.000000  8.777778 10.444444  8.333333  9.333333
>
> moyenneannée=c(apply(P,1,mean))
> moyenneannée
[1]  8.833333 10.500000  3.416667  8.666667  8.250000 14.666667 15.083333
[8] 11.083333  8.500000
```

Calcul des variances par année, par mois et la variance résiduelle

```
> Sp=nrow(P)*sum((moyennemois-MoyGen)^2)##la somme des carrés periode(mois)##
> Sp
[1] 112.4444
> Vp=Sp/11     ##la variance période##
> Vp
[1] 10.22222

> Sa=ncol(P)*sum((moyenneannée-MoyGen)^2)##la somme des carrées année##
> Sa
[1] 1208.667
> Va=Sa/8    ##la variance année##
> Va
[1] 151.0833

> n_annee <- nrow(P)
> n_mois <- ncol(P)
> #Calcul de la variance résiduelle#
> Sr=0
> for(i in 1:n_annee)
+ {
+ for(j in 1:n_mois)
+ {
+ Sr=Sr+(P[i,j]-moyenneannée[i]-moyennemois[j]+MoyGen)^2
+ }
+ }
> Sr
[1] 1027.556
>
> Vr=Sr/88
> Vr
[1] 11.67677
```

Tests d'existence de la saisonnalité et de tendance

```
> ##Test de l'existence de la saisonnalité##
> Fc1=Vp/Vr
```

```
> alpha=0.05
> n11=ncol(P)-1
> n2=(nrow(P)-1)*(ncol(P)-1)
>
> ficherlu1=qf(1-alpha,n11,n2)
> if (Fc1>ficherlu1){
+   print("La série est saisonnière")
+ }else
+ {
+ print("La série n'est pas saisonnière")
+ }
[1] "La série n'est pas saisonnière"


> ##Test de l'existence de la tendance##
>
> alpha=0.05
> n12=nrow(P)-1
> n2=(nrow(P)-1)*(ncol(P)-1)
>
> Fc2=Va/Vr
> ficherlu2=qf(1-alpha,n12,n2)
>
> if (Fc2>ficherlu2){
+   print("La série a une tendance")
+ }else {
+ print("La série n'a pas de tendance")
+ }
[1] "La série a une tendance"
```

2.9.3 Détermination du modèle(additif ou multiplicatif)

En analysant les graphes annuels de la série(figure 1.2), on voit qu'ils ne sont pas parallèles. Malgré ça on ne peut donc pas conclure directement à un schéma additif. Il faut passer aux autres tests, nous proposons le test de Buys-Ballot(méthode analytique). Ce dernier sera fait avec le langage R. Il consiste à faire une régression linéaire des écarts annuels sur

les moyennes annuelles et estimer par la méthode des moindres carrés les paramètres de l'équation suivante : $\sigma_i = a\overline{x}_i + b$ où σ_i et $\overline{x}_i$ sont respectivement l'écart-type et la moyenne de l'année i. Si le coefficient a n'est pas significativement différent de zero on va conclure à un schéma additif, dans le cas contraire c'est le modèle multiplicatif. Le test de la significativité de a fait intervenir le test de Student avec les hypothèses $H_0 : a$ n'est pas significativement différent de zero et $H_1 : a$ est significativement différent de zero

```
> cancer_matr=matrix(cancer,nrow=12,ncol=9)
> P=t(cancer_matr)
> P
```

```
> DonneeCarré=P^2
> DonneeCarré

> som=c(apply(DonneeCarré,1,sum))
> som
[1] 1070 1518  181 1038  949 2764 2885 1607  898
```

Calcul du vecteur des variances et des écarts annuels

```
> VarAnn=(1/ncol(P))*som-(moyenneannée)^2
> VarAnn
[1] 11.138889 16.250000  3.409722 11.388889 11.020833 15.222222 12.909722
[8] 11.076389  2.583333
> Ecart=c(sqrt(VarAnn))
> Ecart
[1] 3.337497 4.031129 1.846543 3.374743 3.319764 3.901567 3.593010 3.328121
[9] 1.607275
```

Régression des écarts annuels sur les moyennes annuelles

```
> regre1=lm(Ecart~moyenneannée)
> regre1

Call:
lm(formula = Ecart ~ moyenneannée)

Coefficients:
 (Intercept)  moyenneannée
      1.5317        0.1635

> summary(regre1)

Call:
lm(formula = Ecart ~ moyenneannée)

Residuals:
     Min       1Q   Median       3Q      Max
-1.31445 -0.24389 -0.01606  0.42577  0.78234

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.53170    0.68876   2.224   0.0615 .
moyenneannée  0.16353    0.06598   2.479   0.0423 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.6622 on 7 degrees of freedom
Multiple R-squared:  0.4674,    Adjusted R-squared:  0.3913
F-statistic: 6.144 on 1 and 7 DF,  p-value: 0.0423

> coef(regre1)
 (Intercept) moyenneannée
   1.5316964    0.1635324
```

$$\sigma_i = 0.1635\overline{x}_i + 1.5317 + e_i$$

$$n = 9$$

En faisant le test de Student sur la significativité du coefficient $a = 0.1635$ avec la statistique correspondante $t_{\widehat{a}} = 2.479$ que l'on compare avec $t_7^{0.05}$. Si $t_{\widehat{a}} = 2.479 < t_7^{0.05}$, on accepte H_0 sinon on rejette H_0. Comme $t_7^{0.05} = 2.365 < t_{\widehat{a}}$, on accepte H_1. Donc le coefficient a est significativement différent de 0. On conclut à un schéma multiplicatif sans saisonnalité qu'on peut écrire comme suit :

$$X_t = Z_t * \varepsilon_t$$

.

La série ne présentant pas de saisonnalité, déterminons la tendance par une droite de moindres carrés qui sera faite par la régression des observations sur le temps. La droite de tendance est d'équation

$$Z_t = 0.4754637 * t - 944.5847123$$

Le code de l'estimation est le suivant :

```
> temps=time(cancer)
> temps
> reglin=lm(cancer~temps)
> coef(reglin)
 (Intercept)        temps
-944.5847123    0.4754637
```

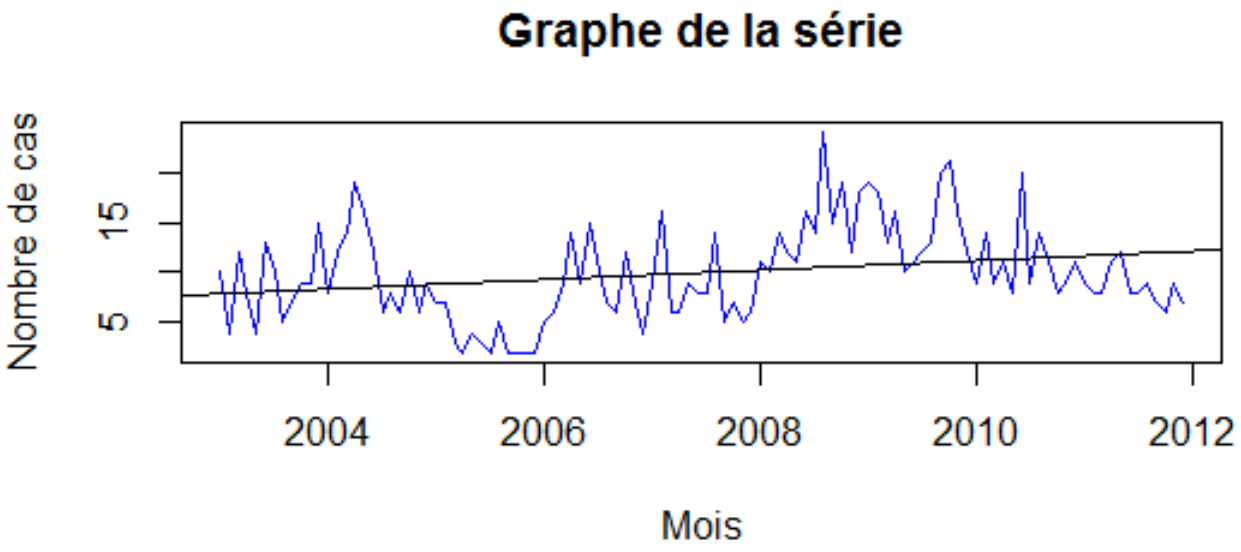

FIGURE 2.4 – Tendance évolutive du nombre de cas du cancer colo-rectal de la wilaya de Tizi Ouzou de 2003 à 2011

Après l'estimation de la tendance linéaire, on détermine les résidusε_t et en étudier les propriétés suivantes pour valider cet ajustement. Les résidus doivent :

1. être d'espérance nulle i.e $E(\varepsilon_t) = 0$;
2. avoir la même variance pour tout t, $var(\varepsilon_t) = \sigma_\varepsilon^2$;
3. être non corrélés entre eux, $corr(\varepsilon_t, \varepsilon_s) = 0, t \neq s$;

4. être normalement distribués.

```
> #moyenne des résidus#
> m=mean(resi.mco)
> m
> #Etude de correlation des résidus#
> lag.plot(resi.mco,lag=20,do.lines=FALSE)
> par(mfrow=c(1,2))
> acf(resi.mco)
> pacf(resi.mco)
> hist(resi.mco)
```

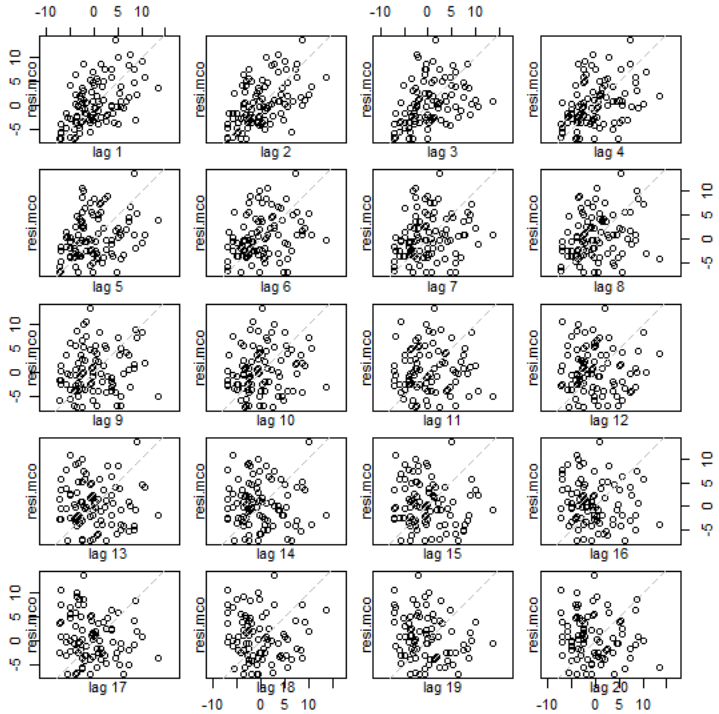

FIGURE 2.5 – Le lag.plot des résidus, Etude de corrélation entre les résidus

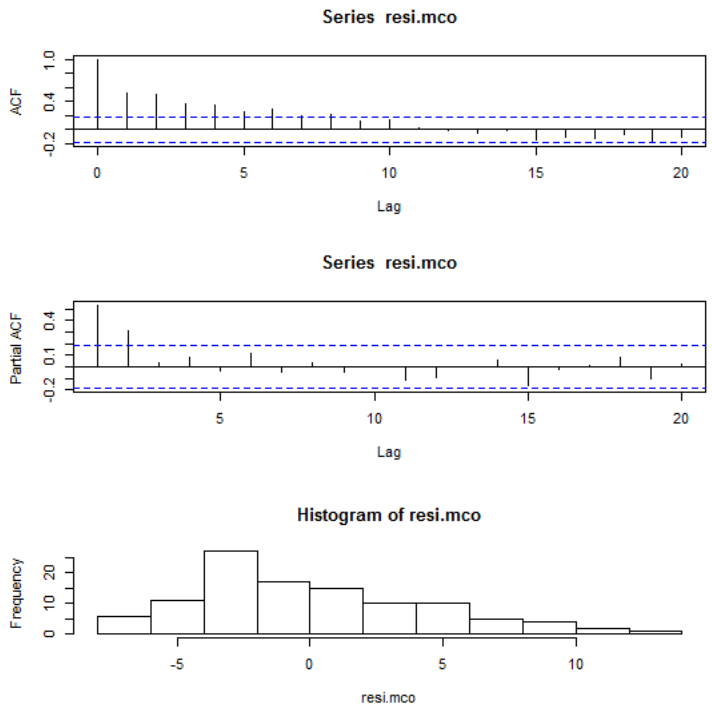

FIGURE 2.6 – Autocorrelogrammes simple et partiel et l'histogramme des résidus

Comme on le remarque sur l'autocorrelogramme simple et le lag plot, les résidus présentent une corrélation. D'où on peut dire que ε_t n'est pas un bruit blanc. Cela peut être aussi montré par le test de portemanteau avec les hypothèses suivantes

$H_0 : \varepsilon_t$ est un bruit blanc
$H_1 : \varepsilon_t$ n'est pas un bruit blanc

```
> Box.test(resi.mco, lag = 20, type = c("Ljung-Box"))

        Box-Ljung test

data:  resi.mco
X-squared = 133.06, df = 20, p-value < 2.2e-16
```

On a que la statistique de test $Q > \chi^2_{20;0.05}$ c'est-à-dire que les ε_t sont corrélés entre eux. Alors on rejette H_0

```
> #Etude de la normalité#
```

```
> qqnorm(resi.mco)
> abline(mean(resi.mco),sd(resi.mco))
```

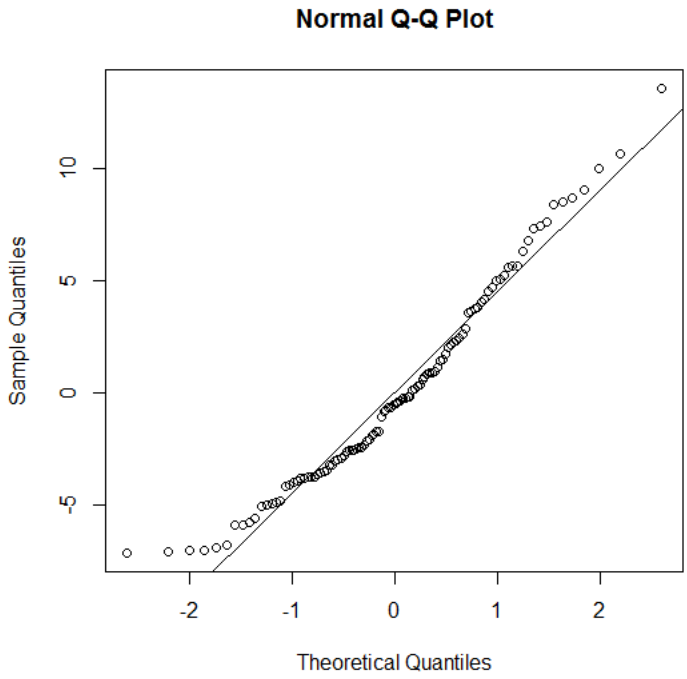

FIGURE 2.7 – Le QQ-plot des résidus

On constate les résidus de la régression ne suivent pas une loi normale. Le coefficient de détermination $R^2 = 0.07016$ est non significatif donc on conclut que la régression n'est pas significative. On peut aussi le montrer par le test en posant comme hypothèse nulle H_0 : la régression est significative contre l'hypothèse alternative H_1 : la régression n'est pas significative. La statistique F de test(F statistique de Fisher) est donnée dans la fonction

```
> summary(reglin)
```

par $F - statistic$. On va la comparer avec la valeur du quantile q de Fisher aux ddl $k-1$ et $n-k$ où k est le nombre de paramètres estimés et n le nombre total d'observations. On rejette H_0 si $F > q$.

```
Call:
lm(formula = cancer ~ temps)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1559 -3.4686 -0.5276  2.6847 13.5762
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -944.5847   337.4917  -2.799  0.00609 **
temps          0.4755     0.1681   2.828  0.00560 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.539 on 106 degrees of freedom
Multiple R-squared:  0.07016,   Adjusted R-squared:  0.06139
F-statistic: 7.998 on 1 and 106 DF,  p-value: 0.005599
```

D'après les résultats ci-dessus on a $7.998 = F > q = 3.930 (q = F_{1,106}^{0.05})$. Donc on rejette l'hypothèse nulle H_0. D'où l'ajustement par une droite de moindres carrés n'est pas adéquat.

Toutes ces propriétés nous conduisent à conclure que les résidus de la régression sont non stationnaires. Et on peut dans ce cas les étudier par les modèles de classe ARMA. Ce qui nous pousse à étudier directement la série brute par les méthodes de Box et Jenkins.

2.10 Étude de la série chronologique par les méthodes de lissage exponentiel et de Box-Jenkins

En vue de la prédiction, on met en réserve la dernière année, 2011. On utilisera alors les données de 2003 à 2010 pour faire les prédictions de 2011 afin de faire les comparaisons avec les données réelles de 2011.

2.10.1 Méthode de lissage exponentiel double

choix des paramètres

On va procéder par une méthode empirique qu'on peut décrire ainsi. On commence par fixer des valeurs de α et β au "hasard" et une fois que nous avons dégagé les valeurs autour des quelles on peut avoir des bonnes prévisions, on a alors choisi trois modèles à comparer. Ces trois modèles correspondant respectivement aux couples $(\alpha;\beta), (0.7;0.3), (0.7;0.2), (0.7,0.1)$ seront comparés par rapport à leurs différentes qualités de prévision des douze dernières valeurs. Cela revient à prendre le modèle qui minimise la somme des carrés des erreurs de prévisions réelles. Ainsi, on retiendra finalement le modèle de lissage exponentiel avec comme constantes $\alpha = 0.7$ et $\beta = 0.2$.

Prédiction avec le lissage exponentiel double La série présente une tendance. Il convient d'utiliser le lissage exponentiel double pour la lisser et faire des prévisions.

```
> pred <- ts(cancer10, start = c(2003, 1), frequency = 12)
> plot(pred)
> hw <- HoltWinters(pred,alpha=0.7,beta=0.2,gamma=FALSE)
> hw

Holt-Winters exponential smoothing with trend and without seasonal component.

Call:
HoltWinters(x = pred, alpha = 0.7, beta = 0.2, gamma = FALSE)

Smoothing parameters:
 alpha: 0.7
 beta : 0.2
 gamma: FALSE

Coefficients:
        [,1]
a 10.1662034
b -0.2015456
call
> forc <- predict(hw, n.ahead = 12, prediction.interval = TRUE, level = 0.95)
> forc
              fit      upr         lwr
Jan 2011 9.964658 19.06793   0.8613856
Feb 2011 9.763112 21.65186  -2.1256359
Mar 2011 9.561566 24.42529  -5.3021564
Apr 2011 9.360021 27.38450  -8.6644580
May 2011 9.158475 30.52306 -12.2061114
Jun 2011 8.956929 33.83403 -15.9201664
Jul 2011 8.755384 37.31065 -19.7998846
Aug 2011 8.553838 40.94665 -23.8389760
Sep 2011 8.352293 44.73624 -28.0316538
Oct 2011 8.150747 48.67411 -32.3726201
Nov 2011 7.949201 52.75543 -36.8570265
```

```
Dec 2011 7.747656 56.97574 -41.4804295
> plot(hw,forc)
```

Le graphe des prévisions : On donne ci-dessous une représentation graghique des prévisions

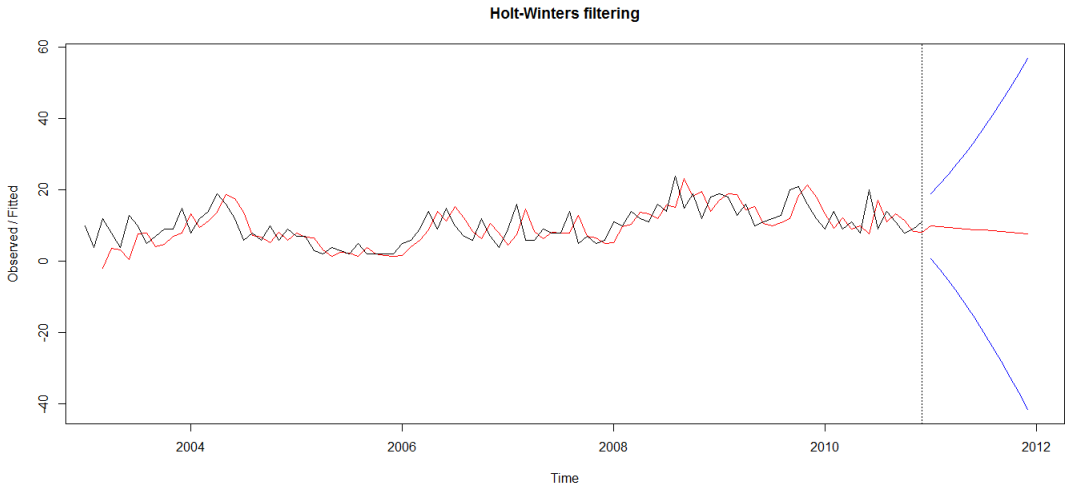

FIGURE 2.8 – Le graphe des prévisions de l'année 2011 avec la série lissée

obtenues avec le lissage (série cancer en noir, des valeurs lissées en rouge , et l'intervalle de prévision en bleu).

2.10.2 Méthode de BOX JENKINS :

Étude de stationnarité de la série :

Graphiques de la série :

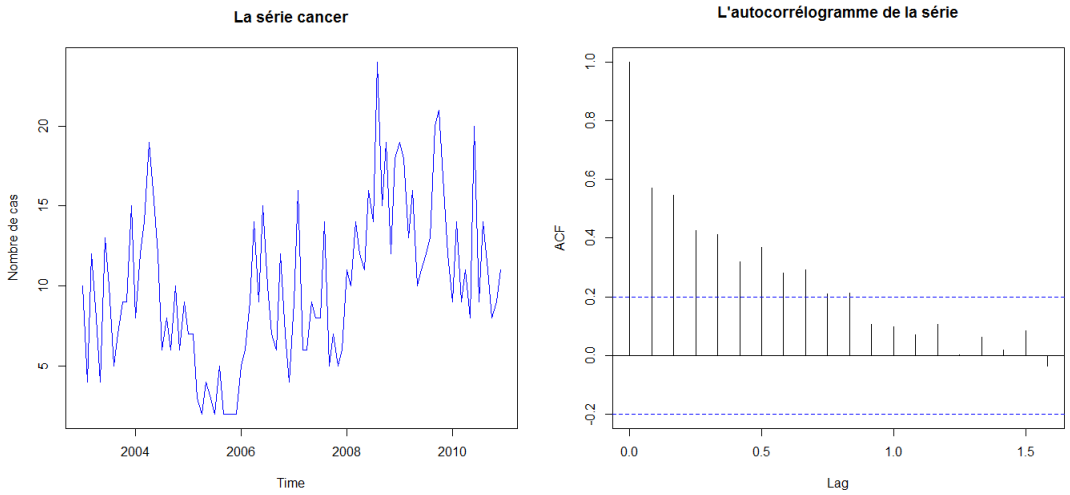

FIGURE 2.9 – Le graphe et l'autocorrélogramme de la série

En regardant le graphe de la série "cancer" ainsi que sont autocorrélogramme simple, ils montrent clairement que la série n'est pas stationnaire car la fonction d'autocorrelation présente lentement, d'où la présence d'une tendance dont on étudiera la nature à l'aide de test de ***Dickey-Fuller.***

Test de Dickey-Fuller augmenté :

En appliquant ce test sur la série "cancer" on obtient les résultats suivants :

```
        Augmented Dickey-Fuller Test

data:  cancer
Dickey-Fuller = -2.6286, Lag order = 4, p-value = 0.3167
alternative hypothesis: stationary
```

Hypothèses :
H_0 :Existence d'une racine unitaire (non stationnaire)
H_1 :La série est stationnaire
La p-value est supérieure à 0.05, donc on retient l'hypothèse H_0 d'une racine unitaire. On note par "dcancer" la série obtenue après une différenciation. Le test de **Dickey-Fuller** sur la nouvelle série nous donne le résultat suivant :

```
        Augmented Dickey-Fuller Test

data:  d.cancer
Dickey-Fuller = -6.173, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(d.cancer, alternative = "stationary") :
  p-value smaller than printed p-value
```

La valeur $t_{\widehat{\phi}}$ estimée (-6.173) est inférieure à la valeur tabulée (-2.62), ainsi que la p-value est inférieure à 0.05, d'où on conclue que la série est stationnaire. On peut donc procéder à une modélisation ARMA de la série "cancer".

Estimation du modèle :

En utilisant d'abord l'autocorrélogramme et l'autocorrélogramme partiel de la série "d.cancer" donnés ci-dessous :

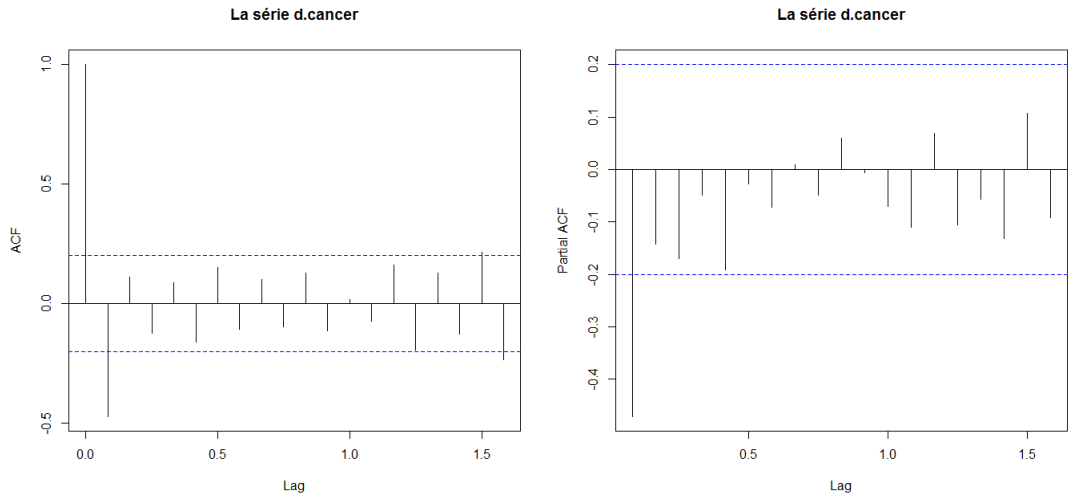

FIGURE 2.10 – l'autocorrélogramme et l'autocorrélogramme partiel de la série "d.cancer"

Après l'analyse de l'acf (autocorrelation function) et de la pacf (partiel autocorrelation function) de la série "d.cancer", on peut proposer le modèle ARIMA(1 ;1 ;1). Or après estimation et évaluation des différents modèles pouvant être proposés, il se trouve que c'est ARIMA(0 ;1 ;1) qui minimise le critère aic. Le résultat de son estimation est donné ci-dessous.

```
Call:
arima(x = cancer, order = c(0, 1, 1))

Coefficients:
          ma1
      -0.6077
s.e.   0.0909

sigma^2 estimated as 15.16:  log likelihood = -264.16,  aic = 532.32
```

Validation du modèle :

test sur le bruit blanc

test de Box-Pierce (portmanteau) : hypothèses :

H_0 :Les résidus sont un bruit blanc

H_1 :Les résidus ne sont pas un bruit blanc

```
        Box-Pierce test

data:  estim_resid
X-squared = 6.5715, df = 12, p-value = 0.8846
```

On accepte alors l'hypothèse H_0 que les résidus sont non corrélés entre eux (car $Q < X^2_{1-\alpha}(k-p-q)$ avec $k = 18$).

tests de normalité :

```
   Jarque Bera Test

data:  estim_resid
X-squared = 4.3871, df = 2, p-value = 0.1115
  Lilliefors (Kolmogorov-Smirnov) normality test

data:  estim_resid
D = 0.073262, p-value = 0.2321
```

Hypothèses :

H_0 :Les résidus suivent une loi normale

H_1 :Les résidus ne suivent une loi normale

Les tests de Jarque-Bera et de Kolmogorov-Smirnov confirment l'hypothèse de normalité des résidus car la p-value est supérieure à 0.05 dans les deux tests. De plus lautocorrélogramme et l'autocrrélogramme partiel des résidus montrent bien l'absence de corrélation entre eux.

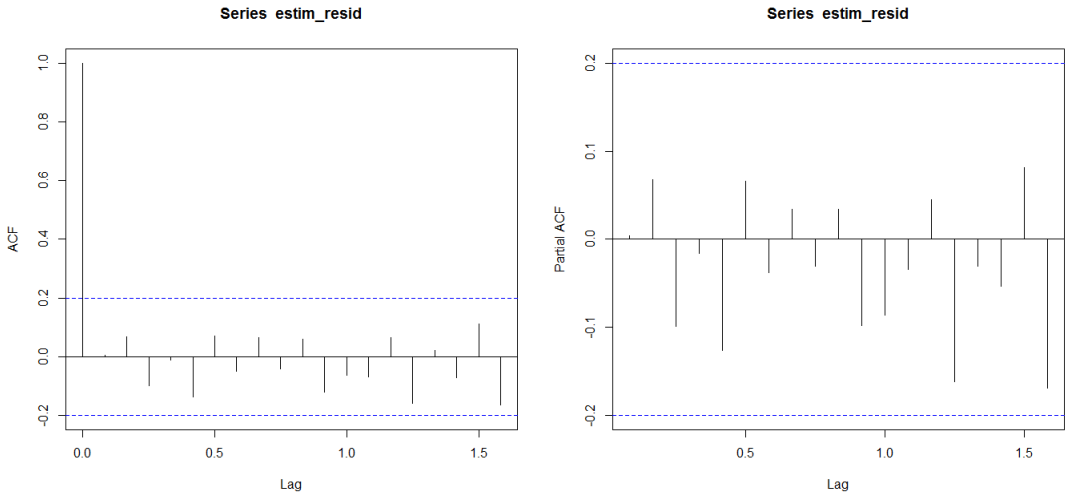

FIGURE 2.11 – l'autocorrélogramme et l'autocorrélogramme partiel des résidus

Donc le bruit blanc suit une loi normale.

prévision avec le modèle ARIMA(0,1,1) les prévision à l'horizon h = 12 (donc pour l'année 2011), le nombre de cas de malades estimé au niveau de la wilaya de Tizi-Ouzou ont donné les résultats suivant :

```
   Point Forecast    Lo 80    Hi 80      Lo 95    Hi 95
Jan 2011       10.33202 5.342491 15.32155  2.7011933 17.96285
Feb 2011       10.33202 4.972243 15.69180  2.1349477 18.52909
Mar 2011       10.33202 4.625969 16.03807  1.6053670 19.05868
Apr 2011       10.33202 4.299539 16.36450  1.1061351 19.55791
May 2011       10.33202 3.989888 16.67415  0.6325649 20.03148
Jun 2011       10.33202 3.694667 16.96938  0.1810640 20.48298
Jul 2011       10.33202 3.412030 17.25201 -0.2511925 20.91523
Aug 2011       10.33202 3.140492 17.52355 -0.6664738 21.33052
Sep 2011       10.33202 2.878840 17.78520 -1.0666354 21.73068
Oct 2011       10.33202 2.626068 18.03797 -1.4532175 22.11726
Nov 2011       10.33202 2.381328 18.28271 -1.8275154 22.49156
Dec 2011       10.33202 2.143899 18.52014 -2.1906307 22.85467
```

Le graphe des prévisions :

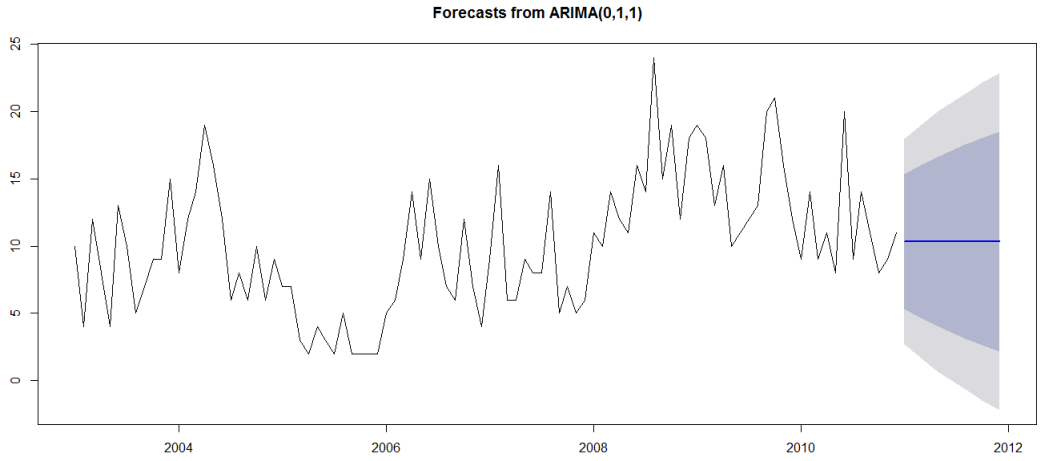

FIGURE 2.12 – Le graphe des prévisions

On donne ci-dessus une représentation graphique des prévisions obtenues avec le modèle ARIMA(0,1,1) avec les intervalles de prédiction à 80 pourcent en bleu et à 95 pourcent en gris.

2.11 Discussion des résultats

2.11.1 Contraintes et biais

1. Problème de l'horizon
2. Présence des cas indéterminés par rapport à la date de diagnostic

2.11.2 Discussion

Dans notre il y a eu problème de prévisions par la droite de tendance par la méthode des moindres carrés car l'ajustement n'était pas adéquat car le coefficient de détermination(R^2) n'est pas significatif. On a opté pour d'autres modélisations telles le lissage exponentiel double qui est une méthode adaptée quand il y a une tendance et pas saisonnalité et grâce à ses formules de mise à jour qui permettent de faire des prédictions.
Ayant mis en réserve l'année 2011 pour prédire les valeurs afin de comparer les données réelles avec celles prédites. Puis on a utilisé la modélisation de Box-Jenkins pour comparer les différents modèles proposes et on a constaté que le lissage exponentiel serait plus adéquat pour prédire des valeurs futures car il donne une bonne precision. Donc le lissage exponentiel paraît meilleure méthode pour prédire le futur mais il convient de limiter l'horizon car la qualité se dégrade rapidement avec l'accroissement de l'horizon.

Le tableau récapitulatif des prévisions :

Données 2011	Lissage	Box-Jenkins
9	9.96	10.332
8	9.76	10.332
8	9.56	10.332
11	9.36	10.332
12	9.15	10.332
8	8.95	10.332
8	8.75	10.332
9	8.55	10.332
7	8.35	10.332
6	8.15	10.332
9	7.94	10.332
7	7.74	10.332
102	99	120

FIGURE 2.13 – Tableau récapitulatif

Prévisions avec le lissage exponentiel double pour l'année 2012 :

Après avoir fait la comparaison entre les deux méthodes de prévision, et après avoir jugé que le lissage exponentiel double nous donne de meilleurs résultats, on a donc prédit pour l'année 2012. A l'aide de logiciel R on obtient les résultats suivants :

Prévisions :

```
> pred <- ts(cancer, start = c(2003, 1), frequency = 12)
> hw <- HoltWinters(pred,alpha=0.7,beta=0.2,gamma=FALSE)
> hw
Holt-Winters exponential smoothing with trend and without seasonal component.

Call:
HoltWinters(x = pred, alpha = 0.7, beta = 0.2, gamma = FALSE)

Smoothing parameters:
 alpha: 0.7
 beta : 0.2
 gamma: FALSE

Coefficients:
        [,1]
a  7.2841967
b -0.2197186

> forecast <- predict(hw, n.ahead = 12, prediction.interval = T, level = 0.95)
> forecast
              fit      upr        lwr
Jan 2012 7.064478 15.73955  -1.610594
Feb 2012 6.844760 18.17428  -4.484765
Mar 2012 6.625041 20.78960  -7.539522
Apr 2012 6.405322 23.58197 -10.771321
May 2012 6.185604 26.54524 -14.174035
Jun 2012 5.965885 29.67281 -17.741042
Jul 2012 5.746167 32.95825 -21.465919
Aug 2012 5.526448 36.39557 -25.342673
Sep 2012 5.306730 39.97925 -29.365788
Oct 2012 5.087011 43.70424 -33.530218
Nov 2012 4.867292 47.56592 -37.831340
Dec 2012 4.647574 51.56007 -42.264920
```

```
> plot(hw,forecast)
```

Graphe des prévision :

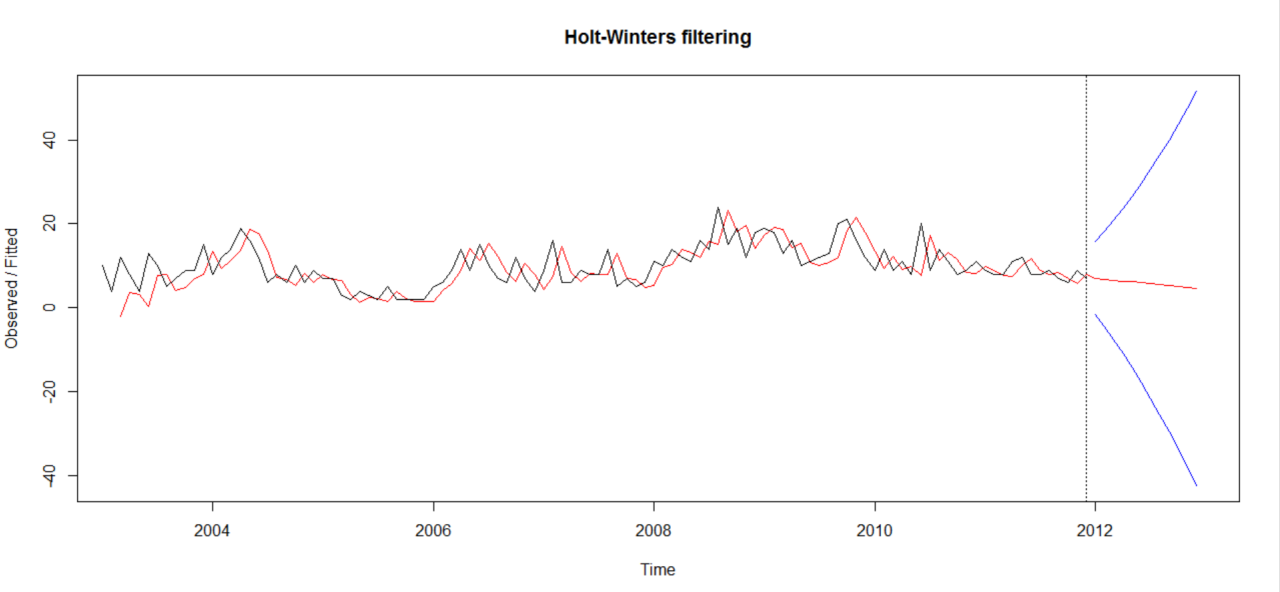

FIGURE 2.14 – Le graphe des prévisions de l'année 2012 avec la série lissée

La série cancer en noir, les valeurs lissées en rouge, et l'intervalle de prévision en bleu.

Conclusion générale

Le cancer colo-rectal est l'une des pathologies qui prennent actuellement de l'ampleur sur le plan tant national qu'international. Dans ce contexte, il est important de prédire l'évolution de cette maladie afin de pouvoir prendre en charge les patients atteints d'autant plus que la prise en charge et le traitement supposent les dépenses énormes. Dans ce cas il est extrêmement important de modéliser adéquatement l'évolution du cancer, dans le but d'éviter les mauvaises prises de décisions.

Dans notre travail, une modélisation sur l'évolution du cancer colo-rectal a été effectuée, nous avons utilisé une modélisation qui traduit une analyse son évolution sous forme de séries temporelles. Le domaine des séries temporelles est en pleine expansion et les notions présentées dans notre travail ne constituent qu'une petite partie des connaissances actuelles sur le sujet. Deux approches de résolution de la problématique ont été étudiées, mises en œuvres et comparées : lissage exponentiel et la méthode de Box-Jenkins, qui nous ont permis de dresser un bilan des réponses apportées à la question posée. Ainsi à la question de savoir la meilleure méthode de prévision, on ne peut y répondre sans préciser la relativité de la réponse à un critère de comparaison et qu'il n'existe pas d'approche qui n'ait pas à la fois des avantages et des inconvénients.

Au terme de notre travail, nous invitons à généraliser la modélisation des pathologies dans le domaine médical à savoir les pathologies cancéreuses et chroniques, approfondir l'étude en prenant en compte plusieurs paramètres(sexe, age ,...)

Bibliographie

[1] ABDOU NIANDOU Daouda. Etude et comparaison des méthodes de prévision des séries chronologiques. Mémoire d'ingénieur en Recherche Opérationnelle. UMMTO, Tizi Ouzou,2009.

[2] Yves Aragon, aragon@Cici.fr. Introduction aux Séries temporelles. Septembre 2004.

[3] Agnès Lagnoux(lagnoux@univ-tlse2.fr). Séries chronologiques : Cours ISMAG MASTER 1 - MI00141X, Université de Toulouse

[4] Sylvain Rubenthaler(Université Nice Sophia Antipolis). Séries chronologiques (avec R), (Cours et exercices) Master1 IM, 2016-2017

[5] Regis Bourbonnais, Michel Terraza. Analyse des Séries temporelles : application à l'économie et à la gestion. Ed Dunod, Paris 2004.

[6] Mohamed Boutahar. Séries temporelles, estimation paramétrique et non paramétrique avec le logiciel R. Département de Mathématiques, Marseille, Décembre 2007.

[7] J. Brokwell, Richard A. Davis. Time Series : Theory and Methods. Second ed Springer, 1991.

[8] Guillaume Chevillon. Pratique des séries temporelles. OFCE et Université d'oxford, décembre 2004.

[9] Ch. Gourieroux et A. Monfort.Séries temporelles et Modèles Dynamiques. Collection "Economie et statistique avancées " seconde éd Economica, 1995.

[10] RAINER VON SACHS ET SEBASTIEN VAN BELLEGEM. STAT 2414, Séries chronologiques Université Catholique de Louvain. 4ème éd, 26 Septembre 2005.

[11] LOUNIS Abbes et OUKACHA Fazia. Modélisation des séries chronologiques. Application aux consommations budgétaires de la wilaya de Tizi Ouzou.
Mémoire de master en Recherche Opérationnelle. UMMTO, Tizi Ouzou,2016