



## INTRODUCTION AUX SÉRIES CHRONOLOGIQUES

Youssef SALMAN

École Nationale Supérieure des Mines de Saint-Etienne

Ce cours est une adaptation de celui rédigé par O. Roustant, à laquelle nous avons ajouté  
une partie sur l'hétéroscédasticité.

Septembre 2025

# Contents

<b>1 Quelques techniques descriptives</b>	<b>4</b>
1.1 Utilisation d'une transformation . . . . .	5
1.1.1 Interprétation . . . . .	5
1.1.2 Démonstration . . . . .	5
1.2 Estimation de la Tendance et de la Saison . . . . .	6
1.2.1 Définitions . . . . .	6
1.2.2 Estimation de la Tendance . . . . .	6
1.2.3 Filtres Moyenne Mobile (MA) . . . . .	6
1.2.4 Preuve . . . . .	7
1.3 Estimation de la saisonnalité . . . . .	9
1.4 Filtrage de la tendance et de la saisonnalité . . . . .	13
1.4.1 Filtrage de la tendance . . . . .	13
1.4.2 Filtrage de la saisonnalité . . . . .	14
1.5 Un deuxième exemple . . . . .	15
<b>2 Prévision par les méthodes de lissage exponentiel</b>	<b>18</b>
2.1 Introduction . . . . .	18
2.2 Lissage exponentiel simple (SES pour Single Exponential Smoothing) . . . . .	18
2.2.1 Description . . . . .	18
2.2.2 Algorithme itératif . . . . .	19
2.2.3 Choix du paramètre . . . . .	19
2.2.4 Cadre d'application . . . . .	19
2.3 Méthode de Holt . . . . .	20
2.3.1 Initialisation . . . . .	21
2.3.2 Choix des paramètres . . . . .	21
2.4 Méthodes de Holt-Winters . . . . .	21
2.4.1 Holt-Winters, version multiplicative . . . . .	21
2.4.2 Holt-Winters, version additive. . . . .	22
2.5 Critique des méthodes de lissage exponentiel . . . . .	23
<b>3 Cadre probabiliste. Quelques modèles probabilistes.</b>	<b>25</b>
3.1 Notions générales. . . . .	25
3.1.1 Stationnarité. . . . .	25
3.1.2 Fonction d'autocovariance. Autocorrélations. . . . .	26
3.1.3 Autocorrélations partielles. . . . .	27

3.2	Modèles SARIMA . . . . .	28
3.2.1	Bruit Blanc . . . . .	28
3.2.2	Marche au hasard . . . . .	28
3.2.3	Modèle autorégressif. . . . .	29
3.2.4	Modèle à moyenne mobile. . . . .	30
3.2.5	Modèle mixte ARMA . . . . .	32
3.2.6	Modèles ARMA intégrés : ARIMA et SARIMA . . . . .	32
<b>4</b>	<b>Méthodologie de Box et Jenkins</b>	<b>34</b>
4.1	Préparation des données : stationnarisation . . . . .	34
4.2	Sélection d'un modèle . . . . .	34
4.2.1	ACF et PACF . . . . .	35
4.2.2	Autres outils de décision . . . . .	38
4.3	Estimation . . . . .	38
4.4	Validation . . . . .	40
4.4.1	Vérifications graphiques . . . . .	40
4.4.2	Tests statistiques . . . . .	42
4.5	Retour sur la série 'airline'. . . . .	44
<b>5</b>	<b>Prévision avec un modèle probabiliste</b>	<b>47</b>
5.1	Qu'est-ce qu'une prévision ? . . . . .	47
5.2	Calcul des prévisions . . . . .	48
5.2.1	Calcul explicite . . . . .	48
5.2.2	Simulation . . . . .	51
5.2.3	Bootstrap . . . . .	53
5.3	Optimalité des méthodes de lissage exponentiel . . . . .	53
5.4	Performances en termes de prévision. Analyse post-sample . . . . .	54
<b>6</b>	<b>Hétéroscédasticité</b>	<b>56</b>
6.1	Introduction et définition générale . . . . .	56
6.2	Détéction de l'hétéroscédasticité - Test de White . . . . .	57
6.3	Test de Breusch-Pagan . . . . .	58
6.4	Modèle ARCH(q) . . . . .	59
6.4.1	Définition et représentation . . . . .	61
6.5	Autre Modèle . . . . .	62

# Introduction

Une série chronologique, ou série temporelle, est une série d'observations ordonnées chronologiquement<sup>1</sup>. Elles se rencontrent naturellement dans une grande variété de domaines. On peut citer : l'économie (taux de chômage, PNB ...), la finance (cours d'action, taux d'intérêt, ...), l'écologie (pollution à l'ozone, au CO, ...), le transport (avec l'exemple célèbre du trafic aérien international), la démographie...

Les objectifs d'étude sont multiples. La prévision est sans doute le but le plus fréquent. Il s'agit de prévoir les valeurs futures d'une variable grâce aux valeurs observées dans le présent et le passé de cette même variable ; la problématique n'est donc pas la même qu'en régression où l'on cherche à prévoir le niveau d'une variable (la réponse) en fonction du niveau d'autres variables (les prédicteurs). Parmi les autres objectifs avoués de l'étude des séries temporelles, figure le problème de l'estimation d'une tendance ; par exemple on peut se demander si une variation observée du chômage est le fait d'une fluctuation saisonnière, ou bien est le reflet d'une tendance. Cela nécessite donc le filtrage des variations saisonnières. En finance, c'est en général tout simplement (!) la dynamique<sup>2</sup> de la série qui est au centre des études ; la modélisation d'un cours d'action est quasiment sans intérêt sur la prévision mais est essentielle pour l'évaluation (le "pricing") des produits financiers complexes construits sur l'action (on parle de "produits dérivés"). Un autre problème consiste à évaluer l'impact d'un événement sur une variable : comment quantifier l'influence de la ceinture de sécurité sur le nombre de tués sur la route? du changement d'horaire sur la consommation d'énergie ? ...

L'étude systématique des séries temporelles remonte à la fin du 2ème conflit mondial et n'a cessé de s'intensifier depuis (la révolution informatique a même donné un sérieux coup d'accélérateur pour tout ce qui relève des applications pratiques<sup>3</sup>!). Le nombre de techniques d'études et de modèles est maintenant colossal. Ce cours est une modeste introduction à quelques techniques de base reconnues par le monde scientifique et employées par les praticiens. Il se limite aux séries temporelles univariées (une seule variable) et discrètes (par opposition au temps continu) ; quelques compléments (notamment l'aspect "fréquence") et extensions sont évoqués dans le tout dernier chapitre.

---

<sup>1</sup>Par la suite, il s'agira d'une suite de variables aléatoires.

<sup>2</sup>C'est-à-dire la façon dont influent les valeurs passées sur le présent.

<sup>3</sup>...mais aussi du développement théorique : la croissance fulgurante de la puissance de calcul a ainsi ouvert la voie à l'estimation non paramétrique.

# Chapter 1

## Quelques techniques descriptives

Nous allons illustrer quelques techniques descriptives utiles en séries chronologiques sur quelques exemples. Commençons par la célèbre série du trafic aérien international, que nous dénommerons désormais 'airline' (source : <http://go.to/forecasting/>).

	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
Janvier	112	115	145	171	196	204	242	284	315	340	360	417
Février	118	126	150	180	196	188	233	277	301	318	342	391
Mars	132	141	178	193	236	235	267	317	356	362	406	419
Avril	129	135	163	181	235	227	269	313	348	348	396	461
Mai	121	125	172	183	229	234	270	318	355	363	420	472
Juin	135	149	178	218	243	264	315	374	422	435	472	535
Juillet	148	170	199	230	264	302	364	413	465	491	548	622
Août	148	170	199	242	272	293	347	405	467	505	559	606
Septembre	136	158	184	209	237	259	312	355	404	404	463	508
Octobre	119	133	162	191	211	229	274	306	347	359	407	461
Novembre	104	114	146	172	180	203	237	271	305	310	362	390
Décembre	118	140	166	194	201	229	278	306	336	337	405	432

Table 1.1: Trafic aérien international de janvier 1949 à décembre 1960 (milliers)

La première étape consiste à tracer les données, ce qui est fait sur la figure ci-après. On peut déjà faire quelques remarques préliminaires :

- Augmentation régulière du trafic ;
- Fluctuation saisonnière : augmentation de novembre à juillet-août, avec un creux vers le mois d'avril, puis diminution jusqu'en novembre.
- Les données sont de plus en plus dispersées.

Cependant certains points mériteraient d'être éclaircis ; par exemple :

- L'augmentation se fait-elle de façon constante, exponentielle, etc.?
- La fluctuation saisonnière est-elle constante au fil du temps ?

- Que se passe-t-il, indépendamment de la tendance à la hausse et des fluctuations saisonnières?

Pour faire vite, disons que les deux premières questions reviennent à étudier la partie déterministe de la série que l'on visualise aisément ; la dernière vise à analyser la structure aléatoire - "le bruit" - qui reste une fois que l'on a extrait la partie déterministe. Dans ce chapitre on étudiera essentiellement la partie déterministe en préparant le terrain pour la partie aléatoire.

## 1.1 Utilisation d'une transformation.

Bien que cela ne soit pas complètement indispensable ici, il est commode de faire subir une transformation aux données dans le but de stabiliser la variance. Après quelques essais, la transformation qui semble la mieux adaptée être la fonction logarithme.

On constate que la transformation a bien l'effet escompté de rendre la variance à peu près constante. Le logarithme a également le mérite d'être relié à une interprétation simple. On constate que la transformation a bien l'effet escompté de rendre la variance à peu près constante. Le logarithme a également le mérite d'être relié à une interprétation simple.

### 1.1.1 Interprétation

Notons  $\log(y_t) = \log(x_t)$ . Alors pour de petites variations de  $x_t$ ,

$$\frac{x_{t+1} - x_t}{x_t} \approx y_{t+1} - y_t.$$

Autrement dit, l'accroissement sur la courbe transformée est approximativement le pourcentage d'accroissement sur la courbe initiale.

### 1.1.2 Démonstration

Faire un développement limité...

La transformation logarithmique fait partie de la famille des transformations de Box-Cox :

$$g(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \log(x) & \text{si } \lambda = 0, \end{cases}$$

qui sont également fréquemment citées. (En dehors du cas  $\lambda = 0$ , il est difficile de donner un sens à ces transformations). On peut citer également la transformation logistique adaptée aux séries qui varient dans un intervalle constant de temps :

$$g(x) = \log\left(\frac{x}{1-x}\right).$$

La valeur transformée varie entre  $-\infty$  et  $+\infty$ .

## 1.2 Estimation de la Tendance et de la Saison

### 1.2.1 Définitions

Qu'appelle-t-on tendance et saisonnalité ? Il est bien difficile de répondre et on se limitera à une définition approximative : la *tendance* correspond à l'évolution au cours du temps indépendamment de fluctuations saisonnières ; la *saisonnalité* aux variations saisonnières "pures". Cependant, tendance et saisonnalité semblent souvent liées et il est parfois difficile de les extraire. Ici cependant, on peut suggérer que la série, une fois transformée, résulte simplement de l'addition de la tendance et de la saisonnalité. Autrement dit, on propose un modèle de décomposition additive pour  $y_t = \log(x_t)$  :

$$y_t = m_t + s_t + u_t,$$

où  $m_t$  représente la tendance,  $s_t$  la saisonnalité et  $u_t$  un terme aléatoire. Il en résulte une décomposition multiplicative pour  $x_t$  :

$$x_t = M_t \times S_t \times U_t,$$

Avec  $\exp(m_t) = M_t$ ,  $\exp(s_t) = S_t$  et  $\exp(u_t) = U_t$ . La transformation logarithme a permis de visualiser la forme d'une décomposition adéquate; ceci fait, il est plus naturel (mais pas obligatoire) de travailler directement sur la série de trafic aérien et de considérer la décomposition multiplicative. Supposons alors que  $M_t$  soit connu. Alors  $S_t$  et  $U_t$  s'interprètent comme des indices;  $S_t$  est l'indice saisonnier et  $U_t$  l'indice aléatoire par lesquels on doit multiplier le niveau actuel de tendance pour obtenir le nombre de passagers. On peut donc exprimer  $S_t$  et  $U_t$  en pourcentage.

### 1.2.2 Estimation de la Tendance

Elle peut se faire par exemple :

- Soit en imposant une forme paramétrée, par exemple ici une fonction affine  $m_t = dt + e$ , ou de type exponentiel  $m_t = \exp(dt + e)$ ;
- Soit en filtrant la saisonnalité. Ceci peut être réalisé au moyen d'un lissage par moyenne mobile.

### 1.2.3 Filtres Moyenne Mobile (MA)

Définition : un filtre moyenne mobile (ou MA pour Moving Average) est une application de la forme :

$$M_t = \sum_{i=-m}^m \alpha_i x_{t+i}.$$

Les filtres MA centrés les plus simples sont de la forme :

$$MA_{(2h+1)} = \frac{1}{2h+1} \sum_{i=-h}^h x_{t+i} \quad \text{et} \quad MA_{2h} = \frac{1}{2h} (x_{t-h} + x_{t+h}).$$

L'appellation  $(2h + 1)MA$  fait référence à la largeur de la fenêtre utilisée pour lisser; il en est de même de l'appellation  $2hMA$  même si la largeur de la fenêtre est  $2h + 1$  car ce filtre s'obtient comme la moyenne des deux filtres "naturels" de taille  $2h$ .

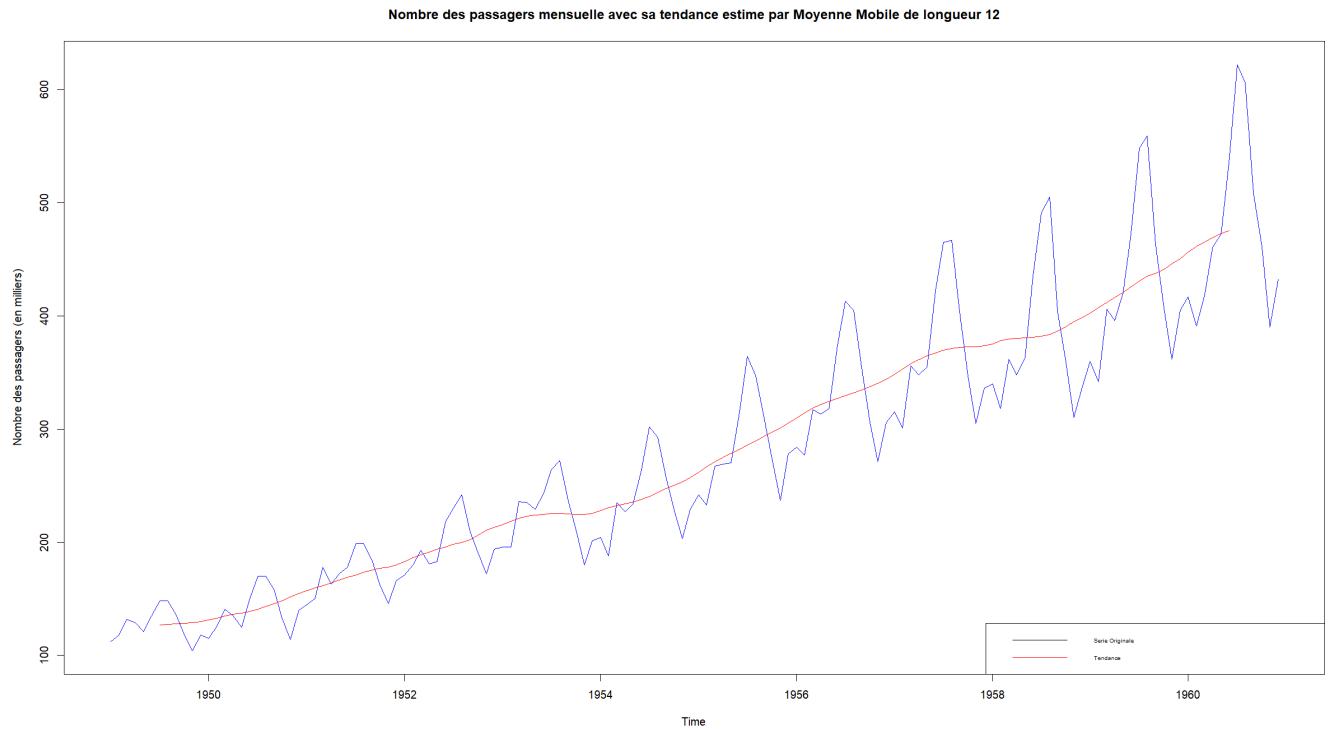
### 1.2.3.1 Propriété des Filtres MA

Les filtres  $(2h + 1)MA$  et  $2hMA$  laissent invariants les polynômes de degré 1.

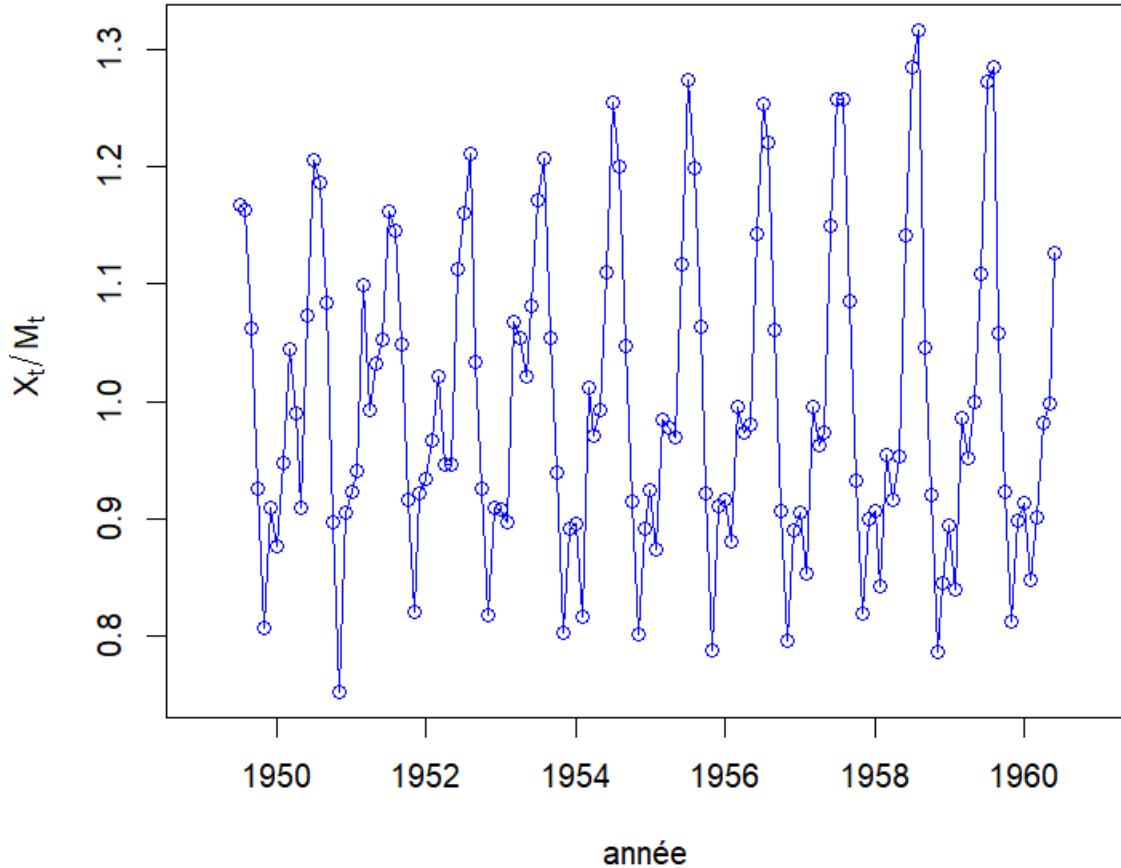
### 1.2.4 Preuve

Preuve laissée en exercice.

La largeur de la fenêtre doit être choisie en fonction de l'objectif souhaité. S'agissant de filtrer la saisonnalité, il est recommandé de choisir la taille de la fenêtre égale à la périodicité. Dans notre cas, on utilisera le filtre  $12MA$ . Le résultat est donné sur la figure ci-dessous, ainsi que la série résiduelle  $x_t/M_t$ .



### Series sans tendance



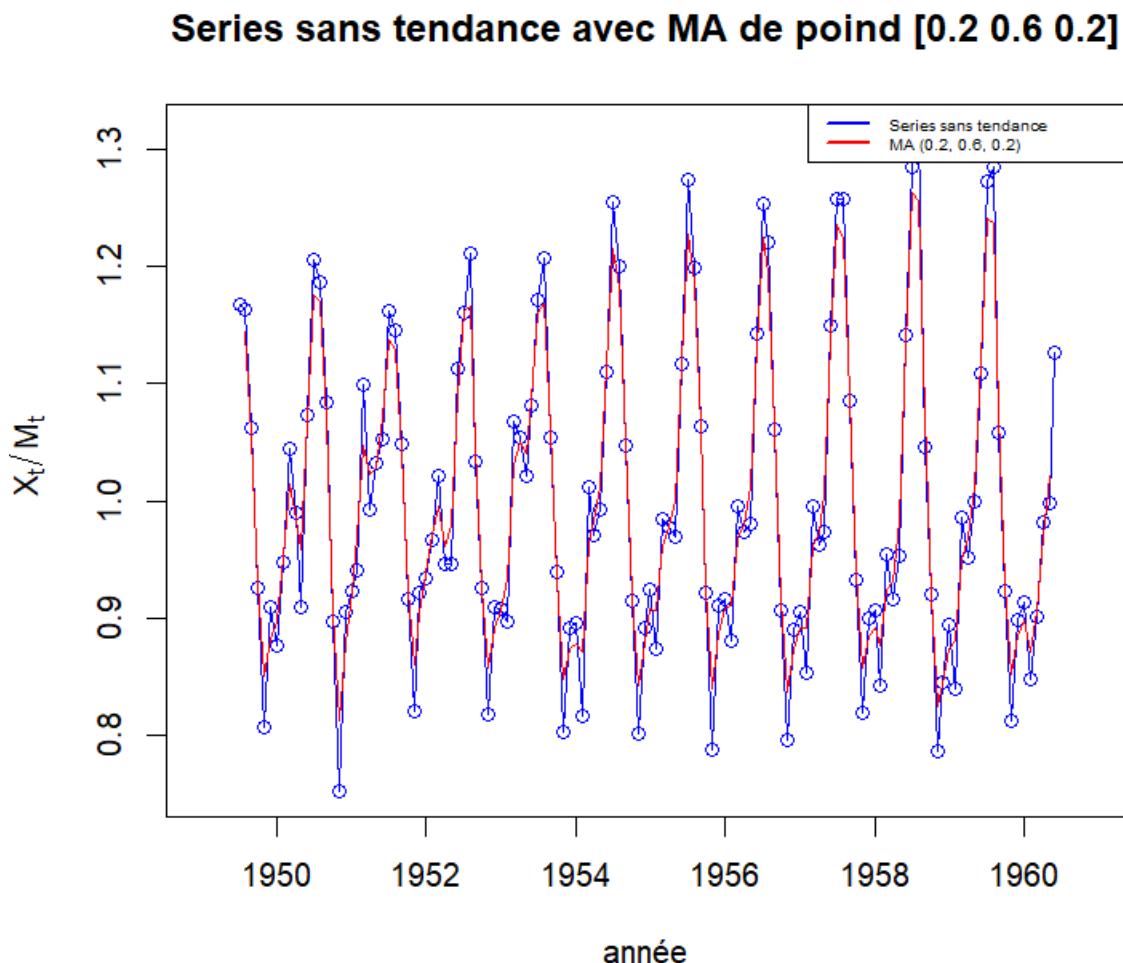
Avant toute chose, remarquons que les lissages utilisés suppriment les valeurs du bord. Ils ne sont donc pas à recommander pour la prévision ! Toutefois, diverses techniques existent pour pallier ce problème (lissage sur la plus grande fenêtre possible, prévision des valeurs futures avant lissage, etc.). Venons-en aux résultats eux-mêmes. On constate que l'augmentation régulière du trafic est entrecoupée – semble-t-il – de deux paliers, un correspondant grosso modo à l'été 1953, l'autre à l'hiver 1958. Il serait intéressant à ce stade d'avoir des informations supplémentaires pouvant (peut-être) expliquer ces paliers. Si le lissage utilisé a bien filtré toute la saisonnalité, alors la série résiduelle  $x_t/M_t$  correspond comme indiqué à une série à laquelle on a enlevé la tendance : il reste la saisonnalité "bruitée" par un terme aléatoire. Malgré ce bruit, certaines caractéristiques semblent apparaître :

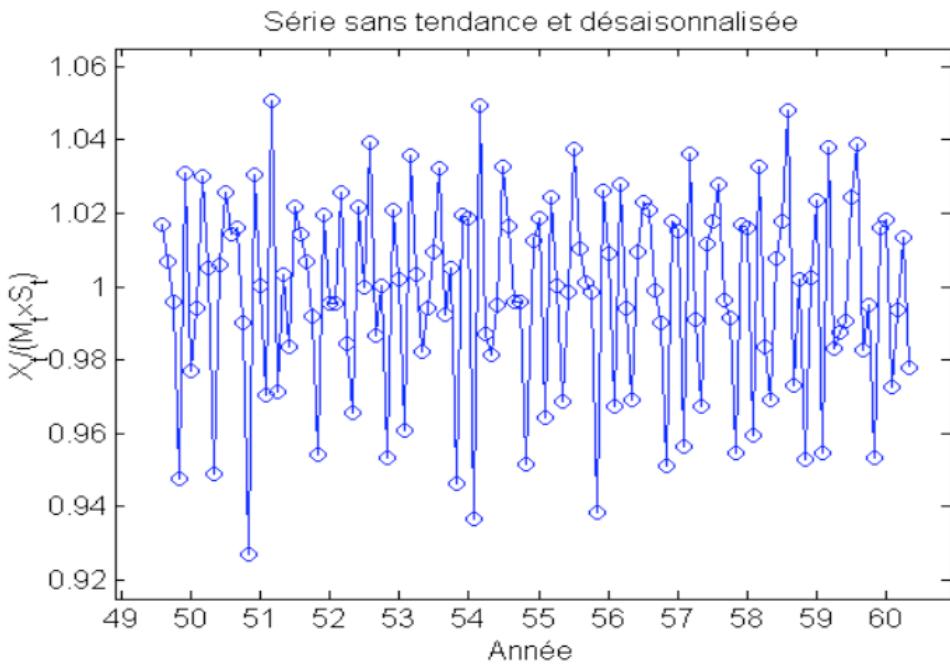
- ★ D'août à novembre, la décroissance du trafic semble régulière ; en revanche, la croissance du trafic de novembre à juillet-août semble être perturbée par deux légers creux en février et en avril ;

\* D'autre part, concernant les séries saisonnières (c'est-à-dire les séries du type  $S_{t+12}$  où  $t$  décrit un mois donné) : celles des mois de juillet-août semblent indiquer une croissance de la fluctuation estivale (bien entendu : indépendamment de la tendance haussière) tandis que celle de mars (2ème "pic") semble indiquer une baisse du trafic pour ce mois-ci au cours du temps.

### 1.3 Estimation de la saisonnalité

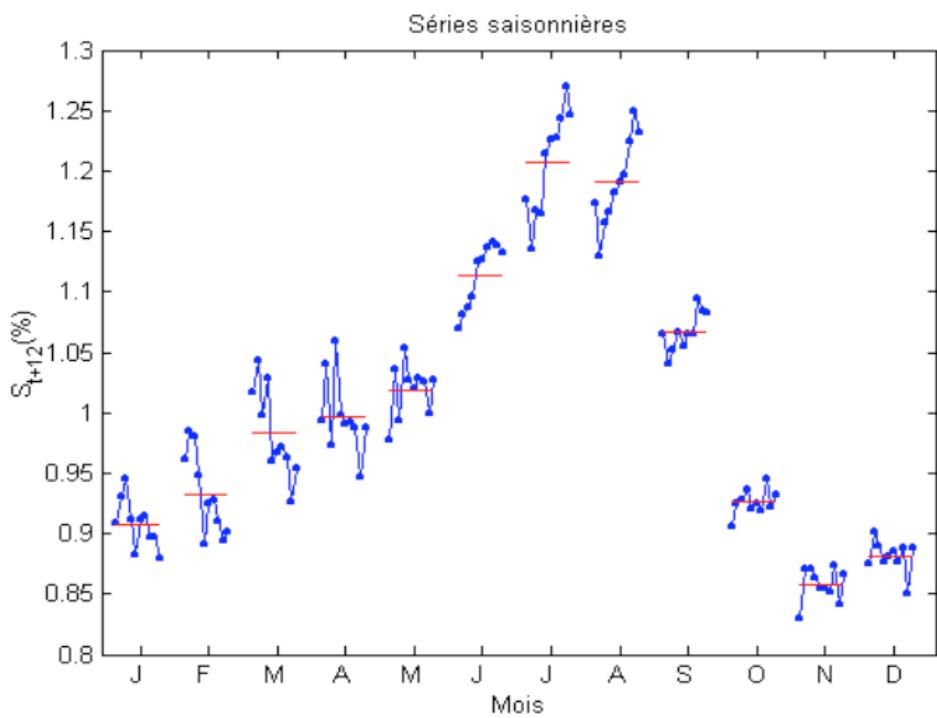
Pour estimer la saisonnalité, on va chercher à filtrer la composante aléatoire encore présente (on pourrait également utiliser d'autres techniques : ajustement d'une courbe périodique, "moyennage" saisonnier, etc.). Afin de lisser suffisamment mais en évitant de gommer les creux en février et avril que l'on juge important, on a choisi un filtre 3 MA avec poids [0.2 0.6 0.2]. Le résultat est montré ci-dessous, ainsi que la série résiduelle qui devrait donc pouvoir s'interpréter comme une série "purement aléatoire".



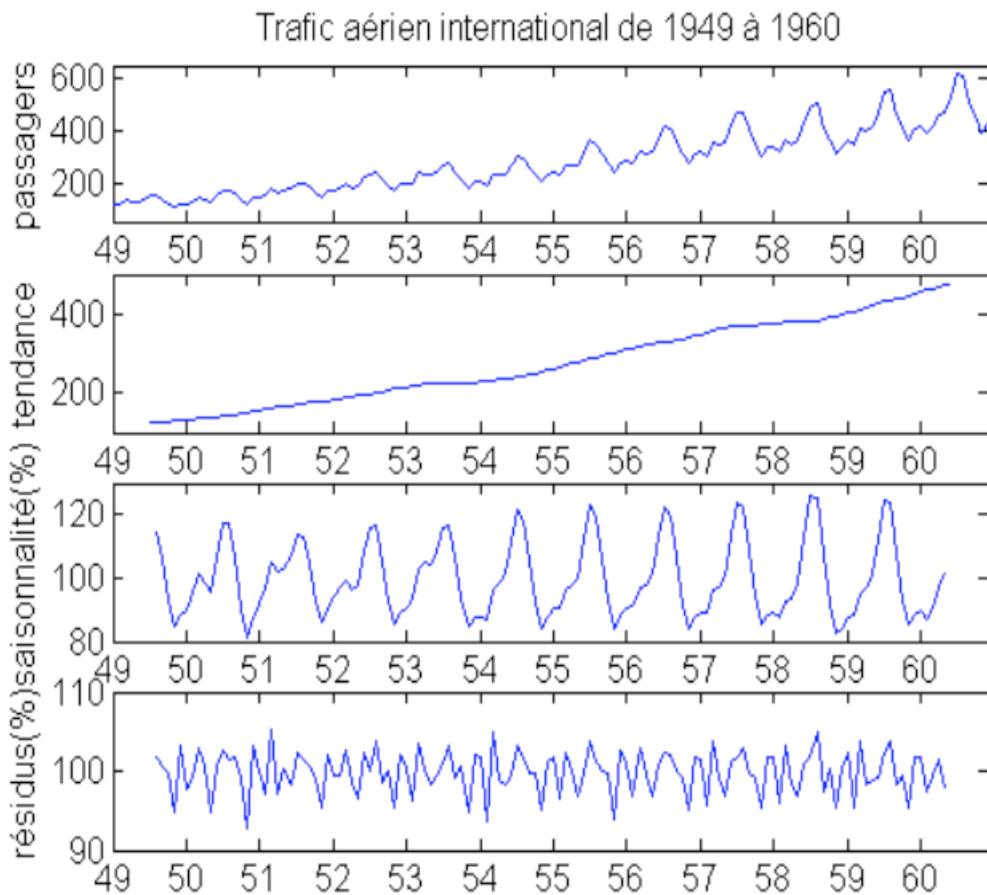


Avant d'aller plus loin à propos de la saisonnalité, on peut constater que le lissage a gommé, malgré tout, une partie des creux de février et avril. Ceci peut être évité mais demande un peu plus de travail. Voir (Makridakis et al.), chapitre "Time Series Decomposition". D'autre part, la série résiduelle semble assez imprévisible, ce qui fait dire que notre étude n'est pas de trop mauvaise qualité : il ne reste pas de terme de tendance ou de saisonnalité flagrants.

On peut maintenant étudier le comportement des séries saisonnières à partir de l'estimation de la saisonnalité par lissage. On les a représentées sur la figure ci-dessous ; le trait horizontal représente la valeur moyenne de chaque série.



On constate comme supposé que les indices saisonniers de juillet-août - mais aussi juin – ont tendance à augmenter ; d'autre part, ceux de février et mars ont tendance à diminuer. Finalement la décomposition obtenue pour la série 'airline' est la suivante :



*Remark 1. :*

\* Notons que, bien que donnant les caractéristiques essentielles, il s'agit d'une décomposition relativement grossière. En effet, on peut vouloir affiner l'estimation de la tendance une fois que l'on a estimé la saisonnalité. Ainsi on peut considérer  $x_t/S_t$  pour obtenir une nouvelle estimation de la tendance  $M'_t$ . Puis on peut vouloir réestimer la saisonnalité, puisque l'on a une estimation que l'on espère plus fine de la tendance, en considérant  $x_t/M'_t$ , etc. Bon nombre de logiciels existent sur ce principe, avec plus ou moins de raffinements (traitements des valeurs aberrantes, utilisation de filtres sophistiqués, etc.).

Le lecteur intéressé pourra consulter (Makridakis et al.), chapitre "Time Series Decomposition".

\* Le problème de la perte de valeurs à chaque lissage peut-être résolu en estimant les valeurs futures ou passées. Voir aussi (Makridakis et al.), même chapitre.

\* Les accros n'auront pas manqué de remarquer que le terme aléatoire  $U_t$  a bon dos et autorise toute sorte de "bidouillage" ; cela vient du fait que l'on ne fait aucune hypothèse sur la nature probabiliste de  $U_t$ . C'est l'un des mérites de la modélisation probabiliste que de donner un cadre rigoureux au traitement des séries chronologiques.

## 1.4 Filtrage de la tendance et de la saisonnalité

Dans certaines circonstances, il sera utile non pas d'estimer la tendance et la saisonnalité mais plutôt de les filtrer afin d'estimer directement le terme aléatoire. La technique précédente permet d'estimer ce dernier mais les lissages utilisés ont le mauvais goût d'introduire des artefacts sous forme de corrélations "parasites" (voir TD).

### 1.4.1 Filtrage de la tendance

Il est important cette fois de partir de la série transformée  $\log(y_t) = x_t$ , car les techniques que l'on va voir correspondent à des opérations additives. Au premier regard, la tendance semble être assez proche d'une droite. Pour l'éliminer, l'idée est de dériver. Pour les séries chronologiques, les données sont souvent espacées de façon régulière et l'intervalle de temps qui les sépare est choisie comme unité. Par conséquent, l'opération de dérivation correspond simplement à regarder la différence  $y_{t+1} - y_t$  entre deux valeurs consécutives.

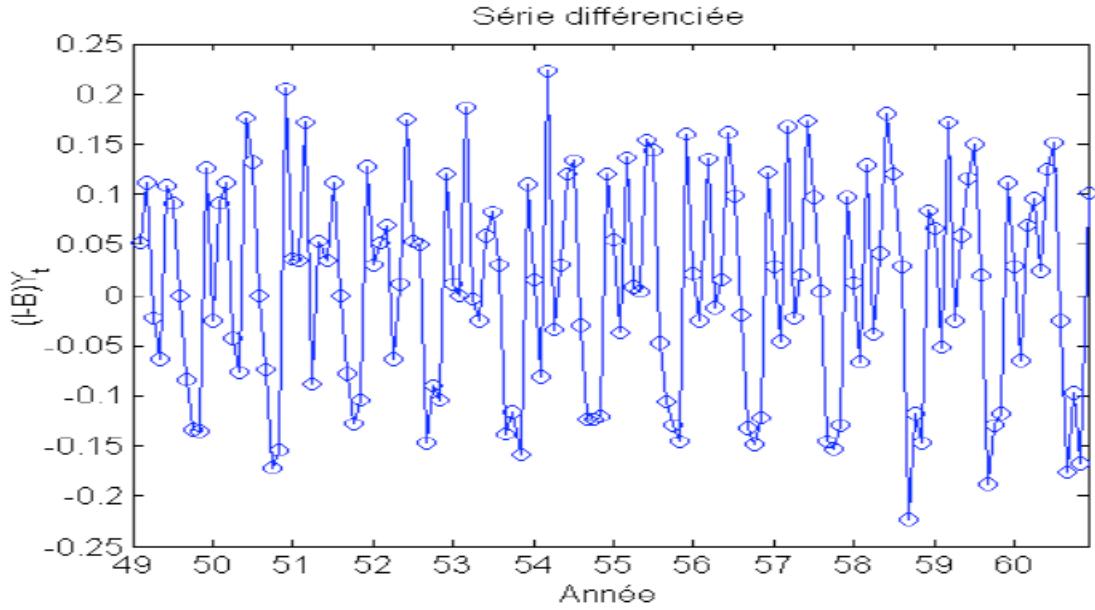
**Definition 1.** L'opérateur de différentiation, noté  $\nabla$ , est  $y_t \xrightarrow{\nabla} y_t - y_{t-1}$ .

Cet opérateur peut s'écrire au moyen de l'opérateur retard :

**Definition 2.** L'opérateur retard, noté  $B$  (pour backward), est  $y_t \xrightarrow{B} y_{t-1}$ .

On a alors  $\nabla = I - B$ , où  $I$  est l'application identique.

Différentiant une première fois, on obtient le résultat ci-dessous. Comme prévu, la tendance a été en bonne partie éliminée ; il reste peut-être un terme constant. D'autre part, la transformation n'a pas éliminé la saisonnalité : on voit de façon évidente des cycles de longueur 12. On pourrait ensuite différencier une deuxième fois mais cela n'est pas nécessaire car le filtrage de la saisonnalité aura la même conséquence.



*Remark 2.* Notons  $y'_t = \nabla y_t$ . Il est très simple de revenir à  $y_t$  à partir de  $y'_t$ . En temps continu, il suffirait d'intégrer et on écrirait

$$y(t) = y(0) + \int_0^t y'(u) du$$

Ici on a la formule analogue  $y_t = y_0 + y'_1 + y'_2 + \cdots + y'_t$ . On revient donc à  $y_t$  par intégration discrète de  $y'_t$  avec la condition initiale  $y_0$ .

#### 1.4.2 Filtrage de la saisonnalité

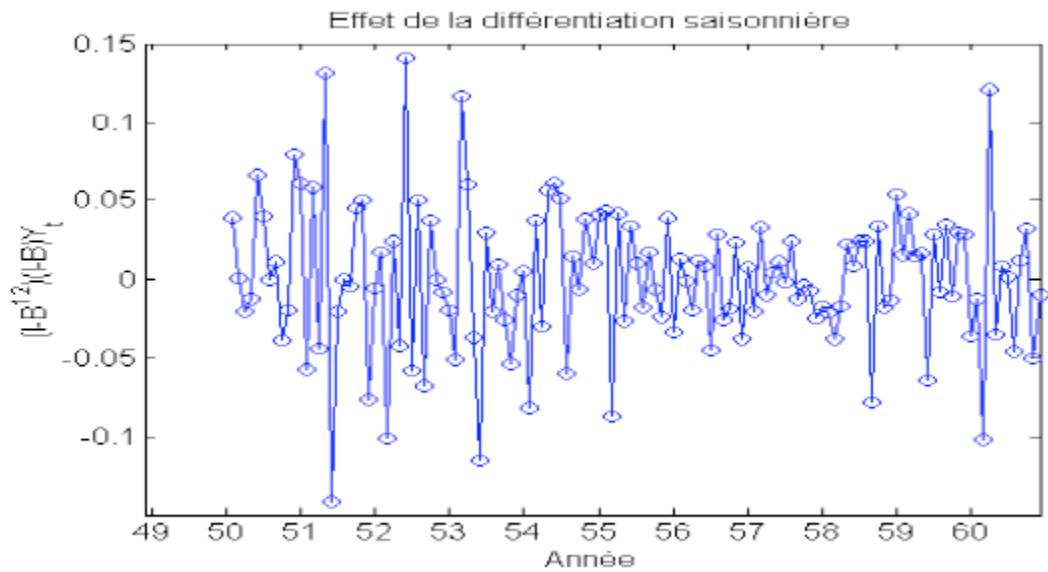
Une façon simple de filtrer la saisonnalité est de considérer les variations d'une année à l'autre, ce qui correspond à effectuer une différentiation saisonnière :

**Definition 3.** L'opérateur de différentiation saisonnière pour une période  $s$ , noté  $\nabla_s$ , est  $x_t \xrightarrow{\nabla_s} x_t - x_{t-s}$ .

On a donc  $\nabla_s = I - B^s$ , avec  $B^s = B \circ B \circ \cdots \circ B$  ( $s$  fois).

*Remark 3* (Ordre des différentiations). Sous cette forme, on remarque que  $\nabla$  et  $\nabla_s$  commutent. L'ordre dans lequel on effectue les différentiations "simple" et saisonnière n'a donc aucune importance.

La différentiation saisonnière  $\nabla_{12}$  conduit à la série  $z_t = \nabla_{12}(\nabla y_t) = (I - B)(I - B^{12})y_t$ , représentée ci-dessous. La série obtenue paraît effectivement "aléatoire" au sens où il n'y a pas, de façon évidente, de terme déterministe. Au chapitre 4, on précisera sa nature aléatoire.



## 1.5 Un deuxième exemple

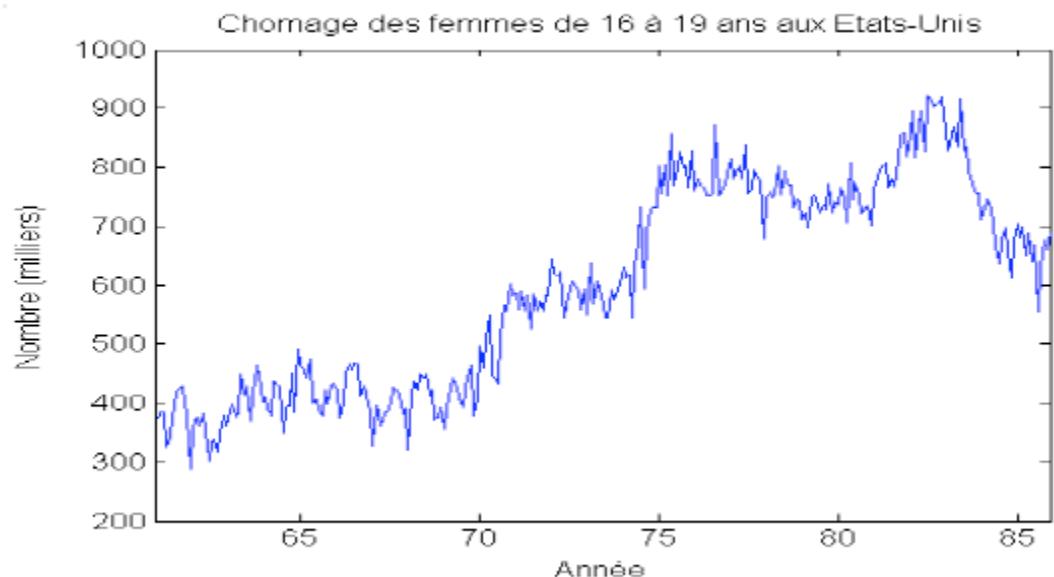
La table ci-dessous contient les données mensuelles de janvier 1961 à décembre 1985 du nombre de jeunes femmes sans emploi des Etats-Unis entre 16 et 19 ans. Nous dénommerons cette série 'unemp' (source : <http://go.to/forecasting/>).

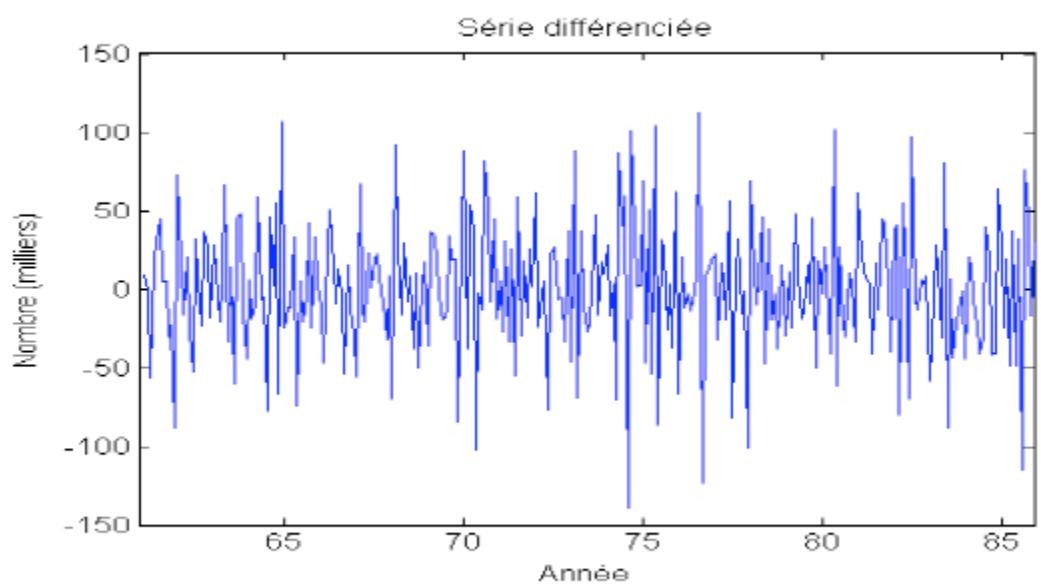
Année	Jan.	Fev.	Mars	Avril	Mai	Juin	Juil.	Août	Sep.	Oct.	Nov.	Dec.
1961	375	384	383	326	344	375	419	424	429	399	376	288
1962	360	376	360	381	354	301	333	339	316	352	378	360
1963	388	398	377	383	449	415	429	369	414	462	447	403
1964	409	390	380	438	431	426	348	394	396	451	384	491
1965	466	454	442	475	401	406	385	380	422	397	430	433
1966	421	374	401	451	465	456	469	466	412	427	414	384
1967	328	395	381	360	383	383	403	425	422	414	382	390
1968	320	412	437	421	450	442	450	412	422	372	375	392
1969	356	392	426	442	426	406	392	426	445	464	379	409
1970	497	459	513	549	447	445	432	514	565	557	601	582
1971	587	560	590	556	582	527	585	556	574	556	582	583
1972	644	620	618	623	546	568	595	605	598	592	558	595
1973	549	637	568	605	594	567	545	545	592	576	593	603
1974	631	614	617	546	632	673	732	593	693	730	731	733
1975	802	755	805	751	855	769	800	825	799	802	765	827
1976	760	781	769	766	752	751	761	873	750	758	772	791
1977	813	781	797	802	782	838	756	764	796	781	780	679
1978	748	759	749	756	802	754	792	772	769	731	746	741
1979	712	723	698	746	754	735	722	737	728	773	723	741

1980	738	765	748	707	808	746	773	751	721	731	735	701
1981	762	783	796	803	806	765	781	768	812	854	858	818
1982	856	897	817	872	895	825	922	915	902	908	911	919
1983	861	827	855	867	836	916	828	835	792	771	757	756
1984	712	733	746	728	707	666	636	676	696	654	613	677
1985	705	680	699	650	687	638	670	555	631	676	659	689

Le graphe de la série est représenté ci-après. Il est bien difficile de déceler une tendance ou un phénomène cyclique. Néanmoins, on ne peut espérer étudier cette série sans effectuer de transformation car il y a un manque de stabilité au niveau des valeurs elle-même. En revanche, le problème disparaît lorsqu'on s'intéresse à la série des variations du nombre de jeunes femmes sans emploi obtenue par différentiation. La différentiation a pour but de se ramener à une série "stationnaire" (la définition rigoureuse est donnée au chapitre 3).

L'utilisation de transformations est la première étape de la méthodologie de Box et Jenkins d'étude des séries temporelles.





## Chapter 2

# Prévision par les méthodes de lissage exponentiel

### 2.1 Introduction

Étant donnée une série d'observations  $x_1, x_2, \dots, x_N$ , on s'intéresse aux prévisions qu'on peut donner à la date  $N$  pour les dates futures. De façon générale, la prévision faite à une date  $t$  pour l'horizon  $h$ , c'est-à-dire pour la date  $t + h$ , sera notée  $\hat{x}(t, h)$ . La méthode de lissage exponentiel simple procède par filtrage de la série de données avec les particularités suivantes:

- le filtre utilisé fait intervenir tout le passé (il est donc décentré à gauche, contrairement à ceux employés au chapitre précédent).
- les poids attribués aux observations décroissent de façon exponentielle en fonction de l'ancienneté de ces observations.

Le lissage exponentiel simple ne s'applique qu'aux séries sans tendance ni saisonnalité. Les extensions de la méthode - méthodes de Holt et de Holt-Winters – permettent de tenir compte de la présence d'une tendance et/ou d'une saisonnalité.

### 2.2 Lissage exponentiel simple (SES pour Single Exponential Smoothing)

#### 2.2.1 Description

La prévision à l'horizon 1 est donnée ici par la moyenne des observations passées, avec des poids décroissant avec l'ancienneté de façon géométrique:

$$\hat{x}(N, 1) = c_0 x_N + c_1 x_{N-1} + \dots$$

avec  $c_{t+1} = (1 - \alpha)c_t$ ,  $0 < \alpha < 1$ . Avec la contrainte que la somme des poids fasse 1, on en déduit la forme des poids comme une fonction exponentielle de l'ancienneté:

$$c_t = \alpha(1 - \alpha)^t, \quad t = 0, 1, \dots$$

La prévision à l'horizon  $h$  est, par définition, la même qu'à l'horizon 1:

$$\hat{x}(N, h) = \hat{x}(N, 1), \quad h = 1, 2, \dots$$

### 2.2.2 Algorithme itératif

Sous la forme précédente, l'évaluation des prévisions comme une moyenne de toutes les observations passées peut-être très coûteuse en temps de calcul. Heureusement on a la relation:

$$\hat{x}(N, 1) = \alpha x_N + (1 - \alpha)\hat{x}(N - 1, 1)$$

ce qui permet de calculer les prévisions à la date  $N$  de proche en proche. Pour initialiser l'algorithme, on adopte généralement le choix  $\hat{x}(1, 1) = x_1$ . La formule ci-dessus donne en outre une autre interprétation de la méthode: la prévision à la date  $N$  "corrige" la prévision antérieure avec l'observation présente. Le paramètre  $\alpha$  régit l'importance du présent dans cette correction; par exemple, pour  $\alpha = 0$  la prévision est la valeur la plus ancienne tandis que pour  $\alpha = 1$ , la prévision est donnée par l'observation présente.

### 2.2.3 Choix du paramètre

Le choix de  $\alpha$  dépend du but recherché. Supposons par exemple que l'objectif soit la prévision à l'horizon 1. Alors il est naturel de minimiser un critère faisant intervenir les erreurs de prévision à l'horizon 1 jusqu'à la date  $N$ ,  $e_t = x_t - \hat{x}(t - 1, 1)$ ,  $t = 2, 3, \dots, N - 1$ . On choisit souvent un critère de moindre carrés:

$$f(\alpha) = \sum_{t=2}^N e_t^2$$

et la valeur de  $\alpha$  correspond au minimum du critère<sup>1</sup>.

### 2.2.4 Cadre d'application

Le lissage exponentiel simple n'est rien d'autre qu'un filtrage local de la série de donnée; rien n'est prévu pour prendre en compte un terme tendanciel ou un phénomène cyclique, et les résultats sont en effet très décevants dans ces cas. On se bornera donc à l'utiliser lorsque aucune tendance ou saisonnalité n'est visible.

*Exemple:*

Reprendons la série 'unemp'. Supposons que l'on ignore les données des 6 derniers mois de l'année 1985 et que l'on souhaite prévoir au 30 juin 1985 les chiffres du chômage relatif à la catégorie sociale considérée pour les 6 derniers mois de l'année. Ici la série ne présente

---

<sup>1</sup>Dans un cadre probabiliste le critère serait choisi, de préférence, de façon à maximiser la vraisemblance.

pas de tendance évidente ni de saisonnalité; on peut donc employer la technique de lissage exponentiel simple. Ci-dessous, on donne les résultats obtenus pour 3 valeurs du paramètre  $\alpha$ .

Mois	Index	Valeurs observées	Prévisions à l'horizon 1		
			$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
Janvier 1961	1	375			
Février	2	384	375.0	375.0	375.0
Mars	3	383	375.9	379.5	383.1
Avril	4	326	376.6	381.3	383.0
Avril 1985	292	650	712.6	690.0	697.3
Mai	293	687	706.3	670.0	654.7
Juin	294	638	704.4	678.5	683.8
JUILLET	295	(670)*	697.8	658.2	642.6
Août	296	(555)*	697.8	658.2	642.6
Septembre	297	(631)*	697.8	658.2	642.6
Octobre	298	(676)*	697.8	658.2	642.6
Novembre	299	(659)*	697.8	658.2	642.6
Décembre	300	(689)*	697.8	658.2	642.6
<b>Analyse des erreurs.</b> (Janvier 1961 à Juin 1985)					
RMSE**			51.06	36.93	39.80

- \* Données inconnues au moment de la prévision.
- \*\*Root Mean Square Error : racine carrée de la somme des carrés des erreurs.

Les prévisions à la date  $t$  à l'horizon 1 sont placées à la date  $t + 1$ . Parmi les trois valeurs  $\alpha = 0.1$ ,  $\alpha = 0.5$  ou  $\alpha = 0.9$ , c'est la valeur  $\alpha = 0.5$  qui minimise la somme des carrés des erreurs. sur la période d'estimation (Janvier 1961 à Juin 1985). En fait la valeur optimale de  $\alpha$  est 0.4931.

## 2.3 Méthode de Holt

Il s'agit d'une adaptation du lissage exponentiel simple aux séries présentant une tendance mais sans saisonnalité évidente. Elle opère au plan local le lissage simultané du "niveau" de la série  $L_t$  et de la pente  $b_t$  de la tendance, au moyen des équations récursives:

$$\begin{aligned} L_t &= \alpha x_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \\ b_t &= \beta(L_t - L_{t-1} + (1 - \beta)b_{t-1}). \end{aligned}$$

$L_t$  s'interprète comme une estimation de la tendance à la date  $t$ , et  $b_t$  comme une estimation de la pente. La prévision à l'horizon  $h$  est définie par:

$$\widehat{x}(t, h) = L_t + hb_t.$$

On retrouve le lissage exponentiel simple pour  $\beta = 0$ , et  $b_1 = 0$ . Dans ce cas on a tout simplement  $L_t = \widehat{x}(t, 1)$ .

### 2.3.1 Initialisation

Le plus simple consiste à prendre  $L_1 = x_1$  et  $b_1 = x_2 - x_1$ , mais d'autres techniques peuvent être envisagées, par exemple une régression linéaire sur les premières valeurs pour donner une estimation locale de la tendance initiale.

### 2.3.2 Choix des paramètres

On peut choisir  $\alpha, \beta$  de façon à minimiser, par exemple, un critère de moindres carrés des erreurs de prévisions  $e_t(\alpha, \beta) = \widehat{x}(t, 1) - x_{t+1}$ .

## 2.4 Méthodes de Holt-Winters

Ce sont les méthodes à privilégier parmi les techniques de lissage exponentiel dans le cas de séries d'observations présentant à la fois un terme de tendance et une saisonnalité. Elles opèrent le lissage simultané de 3 termes correspondant respectivement à des estimations locales du niveau de la série *désaisonnalisée*  $L_t$ , de la pente de la tendance  $b_t$  et de la saisonnalité  $S_t$ . On peut citer au moins deux méthodes dont l'une est adaptée aux séries admettant une décomposition multiplicative et l'autre correspondant aux décompositions additives.

### 2.4.1 Holt-Winters, version multiplicative

En notant  $s$  la périodicité naturelle de la série, les équations sont les suivantes:

$$\begin{aligned} L_t &= \alpha \frac{x_t}{S_{t-s}} + (1 - \alpha)(L_{t-1} - b_{t-1}) \\ b_t &= \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \\ S_t &= \gamma \frac{x_t}{L_t} + (1 - \gamma)S_{t-s}. \end{aligned}$$

La prévision à l'horizon  $h$  est donnée par:

$$\widehat{x}(t, h) = (L_t + hb_t)S_{t-s+h}$$

### 2.4.1.1 Initialisation

L'initialisation de l'algorithme requiert cette fois 3s valeurs:  $L_1, \dots, L_s, b_1, \dots, b_s, S_1, \dots, S_s$ . Il est naturel de choisir pour  $t = 1, \dots, s$ :

$$\begin{aligned}L_t &= \frac{x_1 + \dots + x_s}{s} \\b_t &= \frac{1}{s} \left[ \frac{x_{1+s} - x_1}{s} + \dots + \frac{x_{2s} - x_s}{s} \right] \\S_t &= \frac{x_t}{L_t}\end{aligned}$$

mais d'autres choix restent possibles.

### 2.4.1.2 Choix des paramètres

Le choix de  $\alpha, \beta, \gamma$  peut être fait là encore en minimisant un critère des moindres carrés des erreurs de prévision  $e_t(\alpha, \beta, \gamma) = \hat{x}(t, 1) - x_{t+1}$ .

## 2.4.2 Holt-Winters, version additive.

Le système d'équations est donné par:

$$\begin{aligned}L_t &= \alpha(x_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1}) \\b_t &= \beta(L_t + L_{t-1}) + (1 - \beta)b_{t-1} \\S_t &= \gamma(x_t - L_t) + (1 - \gamma)S_{t-s}\end{aligned}$$

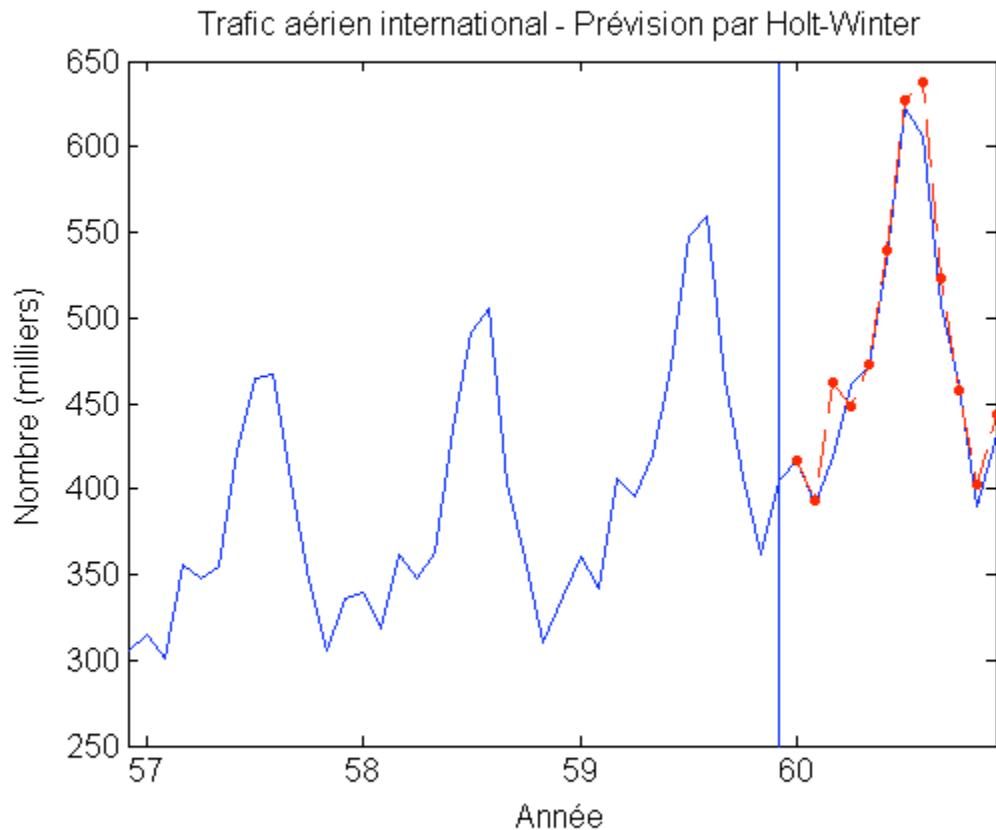
et la prévision à l'horizon  $h$  par:

$$\hat{x}(t, h) = L_t + hb_t + S_{t-s+h}$$

Le choix des valeurs initiales et des paramètres se fait de façon tout à fait analogue au cas multiplicatif.

*Exemple:*

Reprendons la série 'airline'. Plaçons-nous au 31 décembre 1959 et supposons que l'on ait à prévoir le trafic pour l'année suivante. D'après l'étude descriptive réalisée au chapitre 1, c'est la méthode de Holt-Winters multiplicatif qui paraît la plus adaptée parmi les techniques de lissage exponentiel. L'estimation des paramètres par moindres carrés donne:  $\alpha = 0.319$ ,  $\beta = 0.049$ ,  $\gamma = 0.986$ . Les prévisions obtenues sont montrées sur la figure cidessous; on observe qu'elles sont "proches" des valeurs réellement observées - ce qui relève de la chance; mais surtout elles sont la "suite logique" de ce qui précède, ce qui est du à la technique utilisée.



## 2.5 Critique des méthodes de lissage exponentiel

L'avantage des méthodes vues dans ce chapitre pour la prévision, est de fournir une prévision "bon marché" (peu coûteuse en moyens) et parfois très satisfaisante comme dans l'exemple précédent.

Les inconvénients les plus flagrants sont de deux ordres. Tout d'abord, rien ne garantit l'*optimalité* de la méthode sur une série de donnée : les méthodes de lissage exponentiel sont parfois loin d'être les mieux adaptées (encore faut-il s'en apercevoir). D'autre part, elles sont incapables de fournir des *intervalles de prévision*, c'est-à-dire un intervalle contenant la prévision avec une probabilité donnée. Et pour cause, aucun cadre probabiliste n'a été défini pour le moment.

Pour pallier ces insuffisances, on est amené à réaliser des *prévisions au moyen de modèles probabilistes*. Il est à noter que les méthodes de lissage exponentiel correspondent (à l'exception de la version multiplicative de Holt-Winters) à des modèles probabilistes particuliers. On peut donc voir les méthodes probabilistes comme des techniques plus générales permettant de justifier l'emploi des méthodes élémentaires et d'en élargir le champ d'application.

## Chapter 3

# Cadre probabiliste. Quelques modèles probabilistes.

Jusqu'à présent, nous avons regardé les données sans aucune ambition de modélisation. Si l'on prend l'image d'un projectile propulsé par un canon, notre démarche se serait limitée à observer une forme parabolique de la trajectoire. On aimeraient maintenant aller plus loin et proposer un modèle capable de reproduire le "comportement" des données de façon analogue que le modèle Newtonien explique la forme de la chute du boulet de canon. Ici cependant la tâche est bien différente puisque les données ne sont pas déterministes. La démarche consiste à supposer que les données observées  $x_1, \dots, x_n$  forment un extrait d'une trajectoire d'un processus stochastique  $(X_t(\omega))_{t \in \mathbb{Z}}$ , donc il existe  $\omega$  tel que  $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ .

L'ambition est alors de proposer, lorsque cela est possible, un modèle "plausible" pour le processus  $(X_t)$ . En fait on ne sait vraiment faire quelque chose que lorsqu'il est possible de supposer que le processus est *stationnaire*. Dans ce cas on peut associer deux caractéristiques essentielles au processus, la *fonction d'autocorrélation* et la *fonction d'autocorrélation partielle*. La nécessaire comparaison des autocorrélations estimées à partir des données avec celles calculées pour des modèles connus sera parfois même suffisante pour avoir une bonne idée d'un modèle adéquat. Cela suppose naturellement d'avoir une connaissance solide des modèles probabilistes les plus courants. Dans ce cours, on se restreint à des modèles linéaires de type "SARIMA"; leur présentation fait l'objet de la deuxième partie du chapitre.

### 3.1 Notions générales.

#### 3.1.1 Stationnarité.

La connaissance d'un processus équivaut à connaître la loi de tout vecteur  $(X_{t+1}, \dots, X_{t+h})$ ,  $h$  entier. La notion de stationnarité au sens strict, analogue à celle de régime permanent en physique, est donnée par la

**Definition 4.**  $(X_t)$  est *stationnaire* au sens strict si et seulement si la loi de  $(X_{t+1}, \dots, X_{t+h})$  dépend seulement de  $h$ .

Cette notion est parfois trop restrictive et l'on préfère alors la *stationnarité* au second ordre :

**Definition 5.**  $(X_t)$  est *stationnaire* au second ordre ou simplement : *stationnaire*, si et seulement si pour tout entier  $h$ ,  $\mathbb{E}X_{t+h}$  et  $\text{Cov}(X_t, X_{t+h})$  ne dépendent que de  $h$ .

Notons que ces définitions sont équivalentes lorsque  $(X_t)$  est une processus Gaussien (c'est-à-dire lorsque la loi de tout vecteur  $(X_{t+1}, \dots, X_{t+h})$  est Gaussienne).

Pour un processus stationnaire, les moments d'ordre 1 et 2,  $\mathbb{E}X_t$  et  $\text{Var}(X_t)$ , sont donc constants au cours du temps. Toutefois cette propriété n'est pas suffisante. Quitte à retrancher  $\mu = \mathbb{E}X_t$ , on peut toujours se ramener à un processus centré. *Dans la suite tous les processus sont centrés, et la notion de stationnarité doit être comprise au sens faible.* (En revanche on ne suppose pas dans ce chapitre que les processus sont gaussiens).

### 3.1.2 Fonction d'autocovariance. Autocorrélations.

La connaissance d'un processus centré stationnaire se ramène entièrement à l'étude de la fonction d'autocovariance  $\gamma(h) = \text{Cov}(X_t, X_{t+h})$ , ou bien à la connaissance de la variance  $\sigma^2 = \gamma(0) = \text{Var}(X_t)$  et à la *fonction d'autocorrélation*  $\rho(h) = \text{Cor}(X_t, X_{t+h}) = \frac{\gamma(h)}{\gamma(0)}$ .

Notons que ces notions n'ont de sens que pour les processus stationnaires. Les propriétés sont semblables pour  $\gamma$  et  $\rho$ . Dans le cas de la fonction d'autocorrélation, on a

#### 3.1.2.1 Propriétés.

- $\rho(0) = 1$ ,
- $|\rho(h)| \leq 1$ , (d'après l'inégalité de Cauchy-Schwartz)
- $\rho(h) = \rho(-h)$ ,  $\rho$  est une fonction paire.

#### 3.1.2.2 Interprétation géométrique.

$\rho(h)$  est le cosinus de l'angle entre  $X_t$  et  $X_{t+h}$ .

Signalons une difficulté d'ordre théorique : deux processus stationnaires distincts peuvent avoir la même fonction d'autocorrélation. Considérons par exemple :

$$X_t = Z_t + \theta Z_{t-1} \quad \text{et} \quad Y_t = Z_t + \frac{1}{\theta} Z_{t-1}$$

où  $\theta$  est un réel non nul fixé, et les v.a.  $Z_t$  sont i.i.d. Il est facile de voir que les fonctions d'autocorrélations  $\rho_X$  et  $\rho_Y$  sont égales.

D'autre part, sur un plan pratique cette fois, il serait miraculeux de pouvoir identifier un processus à la seule vue de la fonction d'autocorrélation (estimée). Heureusement, nous avons un deuxième outil à notre disposition.

### 3.1.3 Autocorrélations partielles.

Considérons le processus

$$X_t = 0.8X_{t-1} + Z_t, \quad (3.1)$$

où les v.a.  $Z_t$  sont i.i.d. avec  $Z_t$  indépendant de  $X_{t-1}$ .

De par la définition de  $(X_t)_t$ , il y a une forte corrélation entre  $X_t$  et  $X_{t-1}$  ( $\rho(1) = 0.8$ ), qui se répercute entre  $X_t$  et  $X_{t-2}$  ( $\rho(2) = 0.8^2 = 0.64$ ), entre  $X_t$  et  $X_{t-3}$  ( $\rho(3) = 0.8^3 = 0.512$ ), ...

Pourtant la formule (3.1) semble indiquer qu'il n'y a pas de corrélation "directe" entre  $X_t$  et  $X_{t-h}$  pour  $h > 1$ . L'autocorrélation partielle permet en fixant le niveau des variables intermédiaires, de mesurer cette dépendance.

**Definition 6.** Si  $(X_t)$  est un processus stationnaire, l'autocorrélation partielle d'ordre  $h$ , notée  $\pi(h)$ , est définie par

$$\pi(h) = \text{Cor}(X_{t+h} - \mathbb{E}_L(X_{t+h}|X_{t+1}, \dots, X_{t+h-1}); X_t - \mathbb{E}_L(X_t|X_{t+1}, \dots, X_{t+h-1})), \quad \text{pour } h \geq 2 \quad (3.2)$$

où  $\mathbb{E}_L$  désigne l'espérance conditionnelle linéaire.  $\mathbb{E}_L(\cdot|X_{t+1}, \dots, X_{t+h-1})$  est donc l'opérateur de projection orthogonale sur l'espace vectoriel engendré par  $X_{t+1}, \dots, X_{t+h-1}$ .

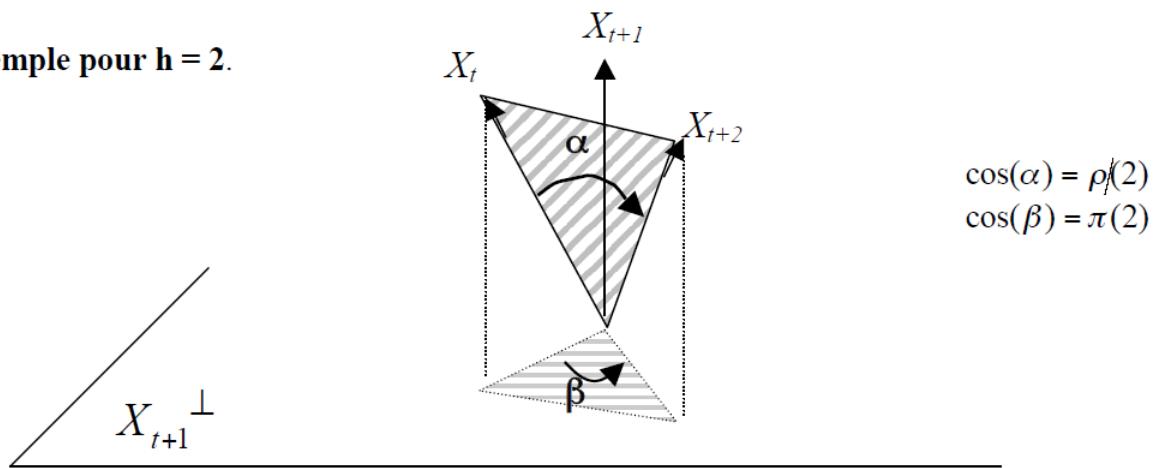
On convient que  $\pi(0) = 1$  et  $\pi(1) = \rho(1)$ .

Le coefficient  $\pi(h)$  jouit des mêmes propriétés qu'un coefficient de corrélation ordinaire et on a :

#### Interprétation géométrique.

$\pi(h)$  est le cosinus de l'angle entre  $\tilde{X}_t$  et  $\tilde{X}_{t+h}$ , les projections sur l'orthogonal de l'espace vectoriel engendré par  $X_{t+1}, \dots, X_{t+h-1}$ .

#### Exemple pour $h = 2$ .



Pour le processus  $(X_t)$  précédent, on a  $\mathbb{E}_L(X_{t+h}|X_{t+1}, \dots, X_{t+h-1}) = 0.8X_{t+h-1}$ , si bien que pour  $h \geq 2$ ,  $\pi(h) = \text{Cor}(Z_{t+h}, X_t - \mathbb{E}_L(X_t|X_{t+1}, \dots, X_{t+h-1})) = 0$ . Il n'y a donc pas, comme on s'y attendait, de corrélation linéaire directe entre  $X_t$  et  $X_{t-h}$  pour  $h \geq 2$ . Le résultat suivant donne un autre moyen de calcul de  $\pi(h)$ .

**Proposition 1.** Considérons la projection de  $X_{t+h}$  sur l'espace vectoriel engendré par  $X_t, X_{t+1}, \dots, X_{t+h-1}$ :

$$\mathbb{E}_L(X_{t+h}|X_t, \dots, X_{t+h-1}) = \phi_{h,1}X_{t+h-1} + \dots + \phi_{h,h-1}X_{t+1} + \phi_{h,h}X_t$$

Alors  $\pi(h) = \phi_{h,h}$ .

Le calcul pratique des coefficients  $\pi(h)$  est basé sur ce résultat et utilise astucieusement le fait que lorsque  $h$  augmente la projection sur l'espace augmenté de la nouvelle variable se déduit de la projection précédente. C'est l'algorithme de Durbin-Levinson (voir Brockwell and Davis (2002) §5.2.).

## 3.2 Modèles SARIMA

### 3.2.1 Bruit Blanc

C'est le modèle le plus simple où l'on suppose qu'il n'y a aucune dépendance temporelle, soit au sens strict, soit seulement au second ordre, ce qui donne les deux notions :

#### Definition 7. Bruit blanc fort et faible

1.  $(X_t)$  est un bruit blanc fortssi les v.a.  $X_t$  sont identiquement distribuées et indépendantes. On note  $(X_t) \sim IID(0, \sigma^2)$  où  $\sigma^2$  est la variance de  $X_t$ .
2.  $(X_t)$  est un bruit blanc (faible)ssi les v.a.  $X_t$  sont identiquement distribuées et non linéairement corrélées :  $\text{Cov}(X_t, X_{t+h}) = 0$ . On note  $(X_t) \sim WN(0, \sigma^2)$

Dans le cas de processus gaussiens, les deux notions coïncident. On parle alors de bruit blanc gaussien.

#### Propriétés du second ordre.

Un bruit blanc est stationnaire et les autocorrélations  $\rho(h)$  et  $\pi(h)$  sont nulles dès que  $h \geq 1$ .

### 3.2.2 Marche au hasard

**Definition 8.**  $(X_t)$  est une marche au hasardssi  $\nabla X_t = X_t - X_{t-1}$  est un bruit blanc.

Il s'agit donc d'un processus dont les accroissements sont un bruit blanc. Lorsque  $(X_t)$  est gaussien, c'est la version discrète du célèbre mouvement brownien.

#### Stationnarité.

Une marche au hasard n'est pas stationnaire car  $\mathbb{V}ar(X_t) = \mathbb{V}ar(X_0) + t\mathbb{V}ar(X_1 - X_0)$ .

### 3.2.3 Modèle autorégressif.

**Definition 9.** Le modèle autorégressif d'ordre  $p$ , noté  $AR(p)$ , est donné par

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$$

avec  $Z_t \sim WN(0, \sigma^2)$ .

Par sa forme le modèle  $AR(p)$  évoque le modèle de régression linéaire. La différence majeure est que les prédicteurs sont eux-mêmes aléatoires et (souvent) corrélés entre eux. Pour étudier ses propriétés, il est commode de commencer par le cas  $p = 1$ .

#### Cas $p = 1$

Le processus  $AR(1)$  est donc défini par

$$X_t = \phi_1 X_{t-1} + Z_t, \quad \text{avec } Z_t \sim WN(0, \sigma^2). \quad (3.3)$$

#### Stationnarité. Causalité.

- Le cas  $\phi = 1$  correspond à la marche au hasard, le processus n'est pas stationnaire.
- Pour  $|\phi| < 1$ , on a de proche en proche la relation

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

qui définit un processus stationnaire.

- Pour  $|\phi| > 1$ , on a de façon analogue une unique solution stationnaire de 3.3 donnée cette fois par

$$X_t = - \sum_{j=1}^{\infty} \frac{1}{\phi^j} Z_{t+j}.$$

Cette dernière solution n'est toutefois pas envisageable dans la pratique car la valeur du processus en un instant fait intervenir tous les instants futurs. On dit que le processus n'est pas *causal*.

**Definition 10.**  $(X_t)$  est dit causal par rapport au bruit blanc  $(Z_t)$  si on peut l'écrire sous la forme

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

On parle alors de la forme  $MA(\infty)$  de  $(X_t)$ . En utilisant l'opérateur retard  $B$ , elle se présente de façon condensée :

$$X_t = \Psi(B) Z_t$$

où  $\Psi$  est la série entière  $\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$

En résumé le processus  $AR(1)$  que nous considérerons sera causal, ce qui implique  $|\psi| < 1$  et l'écriture  $MA(\infty)$ :  $X_t = \Psi(B)Z_t$  avec  $\Psi(z) = 1 + \phi z + \phi^2 z^2 + \dots$ .

### Propriétés du second ordre.

On vérifie aisément :

- $\rho(h) = \phi^h$ , pour tout  $h \geq 0$ ,
- $\pi(h) = 0$ , pour tout  $h > 1$ .

Les autocorrélations décroissent de façon exponentielle tandis que (comme on l'avait déjà remarqué) les autocorrélations partielles sont nulles pour  $h > 1$ .

### Cas général.

Le modèle  $AR(p)$  peut s'écrire sous la forme synthétique

$$\Phi(B)X_t = Z_t,$$

avec  $\Phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ .

On a les résultats suivants :

#### *Théorème 1. Causalité.*

$(X_t)$  est causalessi les racines de  $\Phi$  sont à l'extérieur du cercle unité  $|z| = 1$ .

La série entière  $\Phi(z)$  est alors inversible et  $(X_t)$  admet la forme  $MA(\infty)$  :  $X_t = \Psi(B)Z_t$ , avec  $\Psi(z) = \Phi^{-1}(z)$ .

#### *Autocorrélations.*

Les coefficients d'autocorrélations vérifient la relation, dite de Yule-Walker,

$$\rho(h) = \phi_1\rho(h-1) + \dots + \phi_p\rho(h-p)$$

D'après les résultats classiques sur les suites récurrentes linéaires, on en déduit que  $\rho(h)$  décroît exponentiellement avec  $h$ :

$$\exists \alpha \in ]0, 1[ \text{ et } C > 0 \text{ tels que pour tout } h \geq 0, |\rho(h)| \leq C\alpha^h.$$

#### *Autocorrélations partielles.*

On a  $\pi(p) = \phi_p$  et surtout: pour tout  $h > p$ ,  $\pi(h) = 0$

### 3.2.4 Modèle à moyenne mobile.

**Definition 11.** Le modèle à moyenne mobile d'ordre  $q$ , noté  $MA(q)$ , est défini par

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \dots + \theta_q Z_{t-q}$$

avec  $(Z_t) \sim WN(0, \sigma^2)$ .

Le terme "MA" vient du fait que que  $X_t$  est exprimé comme une moyenne mobile de  $Z_t$  avec les poids  $1, \theta_1, \dots, \theta_q$ .

**cas  $q = 1$ .**

### Stationnarité. Inversibilité.

Le modèle  $MA(1) : X_t = Z_t + \theta Z_{t-1}$  est toujours stationnaire. Cependant il est indispensable de pouvoir exprimer  $Z_t$  en fonction de  $X_t, X_{t-1}, \dots$  ne serait-ce que pour pouvoir calculer  $Z_1, Z_2, \dots, Z_n$  en fonction des données  $X_1, \dots, X_n$ . Or si  $|\theta| < 1$ , on peut écrire

$$Z_t = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \dots$$

Ce dernier cas est peu recommandable et correspond à une situation de *non-inversibilité*.

**Definition 12.**  $(X_t)$  est dit inversible par rapport au bruit blanc  $(Z_t)$  si on peut écrire  $Z_t$  sous la forme

$$Z_t = \sum_{j=0}^{+\infty} \alpha_j X_{t-j}$$

On parle alors de la forme  $AR(\infty)$  de  $(X_t)$ .

On s'intéressera uniquement aux processus inversibles dans la suite pour la raison évoquée plus haut. Pour un MA(1), la condition d'inversibilité est simplement  $|\theta| < 1$ .

### Propriétés du second ordre

Il est aisément vérifiable que  $\rho(1) = \theta$  et  $\rho(h) = 0$  pour  $h > 1$ . Les autocorrélations partielles sont plus difficiles à obtenir ; on montre que pour tout  $h$

$$\pi(h) = \frac{\theta(1 - \theta^2)^{h-1}}{1 - \theta^2}$$

En particulier  $\pi(h)$  décroît exponentiellement avec  $h$ .

*Remark 4.* On n'aura pas manqué d'observer la "dualité" entre un modèle autorégressif et un modèle à moyenne mobile, entre causalité et inversibilité, propriétés de  $\rho$  (resp.  $\pi$ ) pour le modèle AR et celles de  $\pi$  (resp.  $\rho$ ) pour le modèle MA. Cependant on ne peut pas parler de symétrie car dans les deux cas la propriété de bruit blanc est portée par  $Z_t$ .

### Cas général.

Écrivons le modèle MA(q) sous la forme

$$X_t = \Theta(B)Z_t$$

avec  $\Theta(B) = 1 + \theta_1B + \theta_2B^2 + \dots + \theta_qB^q$ . On a les résultats suivants:

*Théorème 2.* Stationnarité, inversibilité.

$X_t$  est toujours stationnaire, et inversible si et seulement si  $\Theta$  a toutes ses racines à l'extérieur du cercle unité.

Autocorrélations. On a  $\rho(h) = 0$  si  $h > q$ . En outre, la réciproque est vraie : si  $X_t$  est un processus stationnaire avec les coefficients d'autocorrélations nuls au-delà de l'ordre  $q$  (et non nul en  $q$ ), alors  $X_t$  suit le modèle MA(q).

Autocorrélations partielles. Il n'y a pas de résultat particulier.

### 3.2.5 Modèle mixte ARMA

Le modèle ARMA(p,q) est formé de “l’assemblage” du modèle AR(p) et du modèle MA(q). Il est défini par

$$\Phi(B)X_t = \Theta(B)Z_t$$

avec  $\Phi(B) = 1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p$  et  $\Theta(B) = 1 + \theta_1B + \theta_2B^2 + \dots + \theta_qB^q$ .

On se limite au cas où les polynômes  $\Phi$  et  $\Theta$  n’ont pas de zéros communs (le cas général peut être traité au prix d’une plus grande complexité; en pratique, il ne se présente pour ainsi dire pas).

On requiert la causalité et l’inversibilité pour  $X_t$ . La condition est la réunion des conditions des modèles AR et MA à savoir que  $\Phi$  et  $\Theta$  ont leurs racines à l’extérieur du cercle unité.

### Propriétés au second ordre

Il n’y a pas de formule explicite simple donnant  $\rho(h)$  ou  $\pi(h)$ . Signalons cependant que la relation de Yule-Walker

$$\rho(h) = \phi_1\rho(h-1) + \phi_2\rho(h-2) + \dots + \phi_p\rho(h-p)$$

est valable pour  $h > q$ . En particulier  $\rho(h)$  décroît de façon exponentielle.

### 3.2.6 Modèles ARMA intégrés : ARIMA et SARIMA

Comme on l’a observé au cours du chapitre 1, les séries temporelles ”brutes” sont rarement stationnaires et l’on peut parfois ”stationnariser” en effectuant des différentiations convenables. Les modèles ARIMA et SARIMA sont conçus en ce sens; les processus de type ARIMA sont adaptés à des séries sans saisonnalité; ceux du type SARIMA peuvent s’utiliser pour les séries les plus générales.

**Definition 13.** 1.  $X_t$  suit le modèle ARIMA(p,d,q) lorsque la série différenciée  $Y_t = (1 - B)^d X_t$  suit le modèle ARMA(p,q). On a donc formellement

$$\Phi(B)(1 - B)^d X_t = \Theta(B)Z_t$$

où  $\Phi$  et  $\Theta$  sont deux polynômes unitaires de degré respectif p et q, et  $Z_t \sim WN(0, \sigma^2)$ .

2.  $X_t$  suit le modèle SARIMA( $p, d, q$ )( $P, D, Q$ )<sub>s</sub> lorsque la série différenciée

$$Y_t = (1 - B)^d(1 - B^s)^D X_t$$

suit un modèle ARMA "saisonnier" du type

$$\Phi(B^s)\Phi(B)Y_t = \Theta(B^s)\Theta(B)Z_t$$

où  $\Phi_s$ ,  $\Phi$ ,  $\Theta_s$ ,  $\Theta$  sont des polynômes unitaires de degré respectif P, p, Q, q et  $Z_t \sim WN(0, \sigma^2)$ . Le modèle s'écrit

$$(I - B^s)^D \times \Phi_s(B^s) \times (I - B)^d \times \Phi(B) \times X_t = \Theta_s(B^s) \times \Theta(B) \times Z_t$$

↑                      ↑                      ↑                      ↑                      ↑                      ↑  
 AR(P)  
saisonnier              AR(p)  
différence              MA(Q)  
saisonnier              MA(q)

Les propriétés du modèle général SARIMA se déduisent de celles du processus ARMA. Il est bien sûr non stationnaire sauf (peut-être) si  $d = D = 0$ .

## Chapter 4

# Méthodologie de Box et Jenkins

Étant donnée une série d'observations, une question essentielle dans l'approche probabiliste est le choix d'un modèle le mieux adapté. Box et Jenkins ont proposé au cours des années 1970 une démarche systématique permettant d'aboutir, si c'est possible, au choix d'un modèle de type (S)ARIMA. Elle repose sur plusieurs étapes qui sont détaillées sur la série 'unemp'. La série du trafic aérien est étudiée ensuite.

### 4.1 Préparation des données : stationnarisation

La première étape consiste à se ramener à une série stationnaire. Elle se décompose en deux phases<sup>1</sup> :

- utilisation d'une transformation;
- différentiation (simple et/ou saisonnière).

Pour la série 'airline', on a utilisé la transformation logarithmique suivie d'une différentiation simple et d'une différentiation saisonnière. Pour la série "unemp", une différentiation simple a suffi. Il est important d'utiliser les transformations à bon escient; dans les exemples précédents, chacune correspondait à la résolution d'une difficulté (présence d'une tendance, d'un terme saisonnier, etc.). La sur-différentiation est à éviter ainsi que la sous-différentiation. Un mauvais usage des différentiations sera repéré, en principe, dans la dernière étape de validation.

### 4.2 Sélection d'un modèle

Il s'agit de choisir pour les données transformées  $y_1, \dots, y_n$  un modèle de type ARMA(p,q) ou de type saisonnier SARIMA(p,0,q)(P,0,Q). L'examen des autocorrelations empiriques "totales" (ACF) ou partielles (PACF) permet souvent à lui seul de sélectionner un bon candidat.

---

<sup>1</sup>Notons que les techniques de décompositions ne font pas partie, stricto sensu, de la démarche de Box et Jenkins. En fait lorsqu'une décomposition a été réalisée, il est souvent possible d'obtenir un résultat analogue par le biais de transformations et de différentiations (comme dans le cas du trafic aérien).

#### 4.2.1 ACF et PACF

On se rappelle qu'un processus MA(q) est caractérisé par le fait que ses autocorrélations théoriques  $\rho(k)$  sont nulles pour  $k > q$ ; par ailleurs les autocorrélations partielles  $\pi(k)$  d'un processus AR(p) sont nulles pour  $k > p$ . On peut donc espérer reconnaître au moins les processus autorégressif ou moyenne mobile purs. Cependant, on doit se contenter d'estimations de  $\rho(k)$  et  $\pi(k)$  (puisque les processus sont inconnus),  $\hat{\rho}(k)$ ,  $\hat{\pi}(k)$ . La sélection s'effectue en examinant les corrélations significatives.

Les estimateurs usuels de  $\rho(k)$  et  $\pi(k)$  sont naturels compte-tenu de leur définition. Pour le coefficient d'autocorrélation d'ordre  $k$ , on utilise :

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{n-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

Le coefficient d'autocorrélation partielle empirique  $\hat{\pi}(k)$  s'obtient par exemple comme l'estimation  $\hat{\phi}_{k,k}$  du coefficient de  $Y_{t-k}$ ,  $\phi_{k,k}$ , dans la régression linéaire de  $Y_t$  sur  $Y_{t-1}, \dots, Y_{t-k}$ :

$$Y_t = \phi_{1,k}Y_{t-1} + \phi_{2,k}Y_{t-2} + \dots + \phi_{k,k}Y_{t-k} + e_t$$

Les autocorrélations empiriques "complètes" ou partielles sont données par les logiciels de statistique.

Même lorsqu'un coefficient d'autocorrélation théorique est nul, le coefficient estimé correspondant est non nul et l'on a recours à des tests statistiques permettant de décider si la valeur obtenue est significative ou non. On ne détaillera pas ici les propriétés statistiques sous-jacentes des estimateurs  $\hat{\rho}(k)$  et  $\hat{\pi}(k)$ <sup>2</sup>.

Disons simplement que celles-ci dépendent du processus ARMA mais dans la pratique, on considère qu'une valeur à l'intérieur des bornes  $\pm 1.96/\sqrt{n}$  n'est pas significative quel que soit le modèle. L'interprétation est la suivante :  $k$  étant fixé, sous l'hypothèse de nullité de  $\rho(k)$  ( $\pi(k)$ ), la probabilité que  $\hat{\rho}(k)$  ( $\hat{\pi}(k)$ ) soit à l'intérieur des bornes est approximativement de 95

Ci-dessous figurent ACF et PACF pour la série "unemp", avec les bandes<sup>3</sup> de confiance  $\pm 1.96/\sqrt{n}$ .

---

<sup>2</sup>Les lecteurs intéressés pourront regarder (Brockwell, Davis), théorèmes 7.2.2. et 8.1.2.

<sup>3</sup>Le terme est trompeur car l'interprétation se fait coefficient par coefficient.

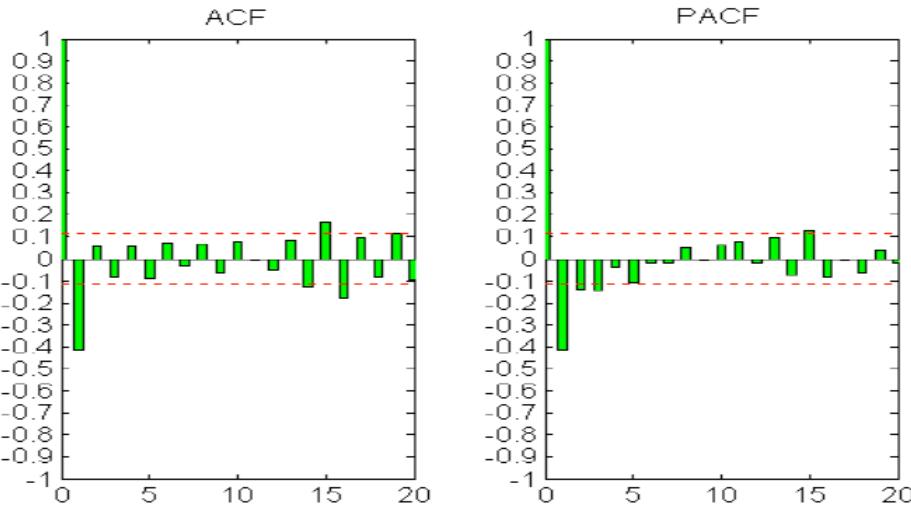


Figure 4.1: ACF et PACF pour la série "unemp" avec les bandes de confiance

Avant d'analyser les résultats, remarquons que l'on a représenté seulement les 20 premiers coefficients estimés. En effet, il faut être conscient que lorsque  $h$  devient grand, l'estimation  $\hat{\rho}(h)$  (par exemple) devient de plus en plus mauvaise car elle utilise de moins en moins de données (dans le cas extrême où  $h = n$ , l'autocorrélation d'ordre  $n$  est estimée seulement avec la première et la dernière donnée !). En règle générale, on préfère se limiter à  $h < n/5$ . On observe que les autocorrélations empiriques ne sont pas significatives à partir du rang 2 (à l'exception de cas isolés autour du rang 15), tandis que les autocorrélations partielles empiriques décroissent très rapidement. Cela évoque donc un processus MA(1). Le fait d'observer des défauts autour du rang 15 n'est pas rédhibitoire: des données issues d'un véritable processus MA(1) peuvent donner des ACF et PACF tout à fait semblables; c'est le nombre relativement peu élevé de données qui explique principalement la qualité médiocre des estimateurs.

### Le piège...

Il est crucial d'avoir une série stationnaire pour pouvoir tirer des enseignements des graphiques précédents. Considérons par exemple le modèle formé d'une composante saisonnière bruitée:

$$X_t = \sin\left(\frac{2\pi t}{30}\right) + 0.5Z_t \quad (*)$$

où  $Z_t$  est un bruit blanc  $N(0, 1)$ . Il est clair que les autocorrélations théoriques sont nulles à partir du rang 1 car le processus (\*) ne diffère d'un bruit blanc qu'au travers d'une partie déterministe. Pourtant voici ce que donne ACF et PACF pour une série de 500 valeurs simulées à partir de (\*) :

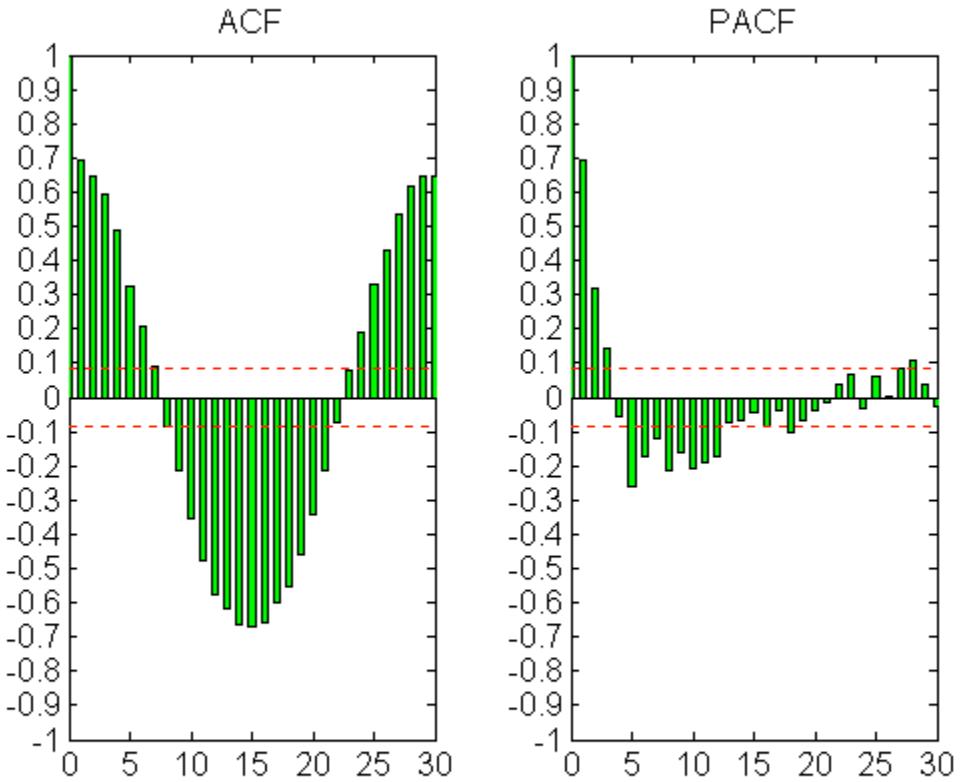


Figure 4.2: ACF et PACF pour une série de 500 valeurs simulées à partir de (\*)

Il ne faut surtout pas conclure qu'il y a de fortes corrélations ! En fait l'estimation des autocorrélations est brouillée par la composante déterministe. Le graphique ci-dessus nous montre la présence d'une saisonnalité mais est inutilisable pour examiner les corrélations au sens probabiliste. **En général le fait que les corrélations ne tendent pas rapidement vers zéro est un signe de non-stationnarité<sup>4</sup>.**

---

<sup>4</sup>Mais la série correspondante peut très bien être associée à un modèle ARMA...

### 4.2.2 Autres outils de décision

Il se peut que l'ACF et la PACF ne soient pas aussi simples que dans l'exemple précédent ou que l'on hésite entre plusieurs modèles. Dans ce cas, on peut avoir recours à l'un des outils suivants.

#### Utilisation d'un critère

Différents critères sont fournis par les logiciels de statistique : AIC (Akaike's Information Criterion), SBC (Schwarz Bayesian Criterion)… Ils sont basés sur la notion de vraisemblance. En principe le modèle pour lequel le critère a la plus faible valeur est le mieux adapté. Cependant le critère AIC favorise les modèles sur-paramétrés; il est donc à utiliser avec précaution. On lui préférera le critère SBC qui a de meilleures propriétés. De façon générale cependant, il vaut mieux éviter d'employer un quelconque critère "en aveugle"; il s'agit plutôt d'un outil complémentaire de sélection.

#### Analyse des résultats d'estimation

Si un ARMA( $p, q$ ) est sélectionné, il est évident qu'un ARMA( $p', q'$ ) avec  $p' \geq p$  et  $q' \geq q$  fera aussi l'affaire. Cependant il est préférable de conserver le modèle avec le moins de paramètres. La raison essentielle est que les paramètres supplémentaires sont souvent très mal estimés (voir un exemple dans la section suivante). À cette fin, la table regroupant les résultats d'estimation (voir ci-après) est d'une grande utilité.

## 4.3 Estimation

Nous considérons ici le cas d'un ARMA(p,q),  $(\Phi(B)Y_t = \Theta(B)Z_t)$  avec  $Z_t \sim N(0, \sigma^2)$ , mais des résultats analogues existent pour les modèles du type SARIMA. Notons  $\beta = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)$  le vecteur des paramètres. L'estimation de  $\beta$  est un problème assez délicat, mais qui est traité par les logiciels de statistique usuels. Parmi les techniques employées, citons la plus séduisante pour l'esprit qui est celle du maximum de vraisemblance. On estime  $\beta$  par la valeur  $\hat{\beta}$  la plus vraisemblable, c'est-à-dire celle qui donne la plus grande densité de probabilité :

$$\hat{\beta} = \arg \max_{\beta} L(y_1, \dots, y_n; \beta)$$

où  $L(y_1, \dots, y_n; \beta)$  est la densité de probabilité du vecteur aléatoire  $(Y_1, \dots, Y_n)$ . Les propriétés asymptotiques<sup>5</sup> sont données dans le théorème suivant :

---

<sup>5</sup>Les propriétés asymptotiques sont identiques pour l'estimateur des moindres carrés.

*Théorème 3.* Soit  $(\Phi(B)Y_t = \Theta(B)Z_t)$  un modèle ARMA(p,q) causal et inversible, tel que  $\Phi$  et  $\Theta$  n'ont pas de zéros communs. Alors si  $(Z_t)$  est un bruit blanc gaussien  $N(0, \sigma^2)$ ,

- (i)  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V(\beta))$  où  $V(\beta)$  est une matrice de taille  $(p+q) \times (p+q)$  connue.
- (ii)  $\frac{\hat{\beta}_i - \beta_i}{\sqrt{v_{ii}(\hat{\beta})}} \sim N(0, 1)$  où  $v_{ii}(\hat{\beta})$  est le coefficient  $(i, i)$  de  $V(\hat{\beta})$ .

Les logiciels fournissent la matrice de variance-covariance de  $\hat{\beta}$ , qui est donc approximativement égale (pour  $n$  "grand") à  $V(\hat{\beta})/n$ . On peut alors, de façon analogue à ce qui est fait en régression linéaire, présenter les résultats dans des tables de la forme suivante :

Paramètre	Estimation	Écart-type	Ratio	p-valeur
$\beta_i$	$\hat{\beta}_{i,\text{obs}}$	$\sqrt{\hat{v}_{ii}/n}$	$\frac{\hat{\beta}_{i,\text{obs}}}{\sqrt{\hat{v}_{ii}/n}}$	$P( \hat{\beta}_i  > \hat{\beta}_{i,\text{obs}})$

On les interprète de la façon suivante:

- La seconde colonne contient une estimation de chaque coefficient;
- L'écart-type représente l'écart-type d'estimation approximatif;
- La colonne ratio est le rapport entre les 2 colonnes précédentes; sous l'hypothèse  $H_0 : \beta_i = 0$ , ce rapport est approximativement de loi  $N(0, 1)$ ;
- La colonne p-valeur contient la probabilité sous  $H_0$  pour une loi  $N(0, 1)$  de dépasser en valeur absolue le ratio calculé précédemment (autrement dit d'obtenir une plus mauvaise statistique).

Pour la série différenciée obtenue à partir de "unemp", on a sélectionné un modèle de type MA(1) :  $Y_t = Z_t + \theta_1 Z_{t-1}$ . Les résultats de l'estimation sont montrés ci-dessous. On constate que le paramètre estimé est très significatif.

Paramètre	Estimation	Écart-type	Ratio	p-valeur
$\theta_1$	-0.5070	0.0519	-9.7676	0.0000

Bien que ACF et PACF ne le suggèrent pas, on pourrait être tenter d'essayer un modèle plus riche de type MA(2) :  $Y_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$ . On obtient alors :

Paramètre	Estimation	Écart-type	Ratio	p-valeur
$\theta_1$	-0.5032	0.0594	-8.4644	0.0000
$\theta_2$	-0.0106	0.0592	-0.1785	0.8584

La valeur estimée pour  $\theta_1$  n'a quasiment pas changé (ce qui est rassurant!) et est toujours très significative; en revanche, l'estimation obtenue pour  $\theta_2$  est entachée d'une erreur plus de 5 fois supérieure et n'est donc pas significative. On n'a donc pas intérêt à rajouter un deuxième paramètre.

Enfin, l'estimation ne serait pas complète sans l'estimation de la variance du bruit. On obtient  $\hat{\sigma}_Z = 37.45$ .

## 4.4 Validation

Tous les résultats d'estimation obtenus précédemment, toutes les prévisions que l'on pourra faire, etc. dépendent de façon cruciale du fait que les résidus forment un bruit blanc gaussien. Il est donc capital d'accorder un soin particulier à la vérification de cette hypothèse.

Les vérifications vont porter non pas sur les résidus théoriques mais sur les résidus estimés ce qui peut modifier les conclusions parfois de façon importante. D'autre part, l'indicateur usuel de dépendance - le coefficient de corrélation - mesure la corrélation linéaire et peut faire manquer des dépendances importantes. C'est pourquoi les tests d'indépendance seront réalisés à la fois sur les résidus et sur les carrés des résidus

Il est clair en effet que si les résidus sont indépendants, leurs carrés le sont également.

### 4.4.1 Vérifications graphiques

Pour un bruit blanc, les coefficients de corrélation sont nuls à partir du rang 1. Il est donc naturel de tracer ACF et PACF pour les résidus et leurs carrés avec les bandes de signification  $\pm 1.9n^{-1/2}$ . Pour le modèle MA(1),  $Y_t = Z_t + \theta_1 Z_{t-1}$  où  $Z_t \sim N(0, \sigma^2)$ , avec  $\hat{\theta}_1 = -0.5070$  et  $\hat{\sigma} = 37.45$ , et les résidus estimés correspondants, on obtient les graphiques ci-dessous.

Figure 4.3: Résidus

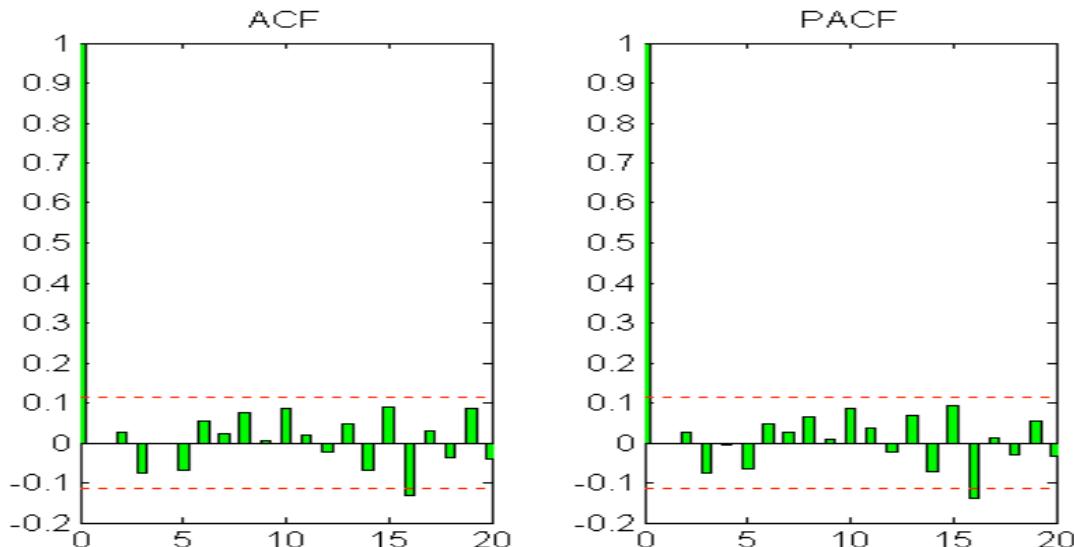
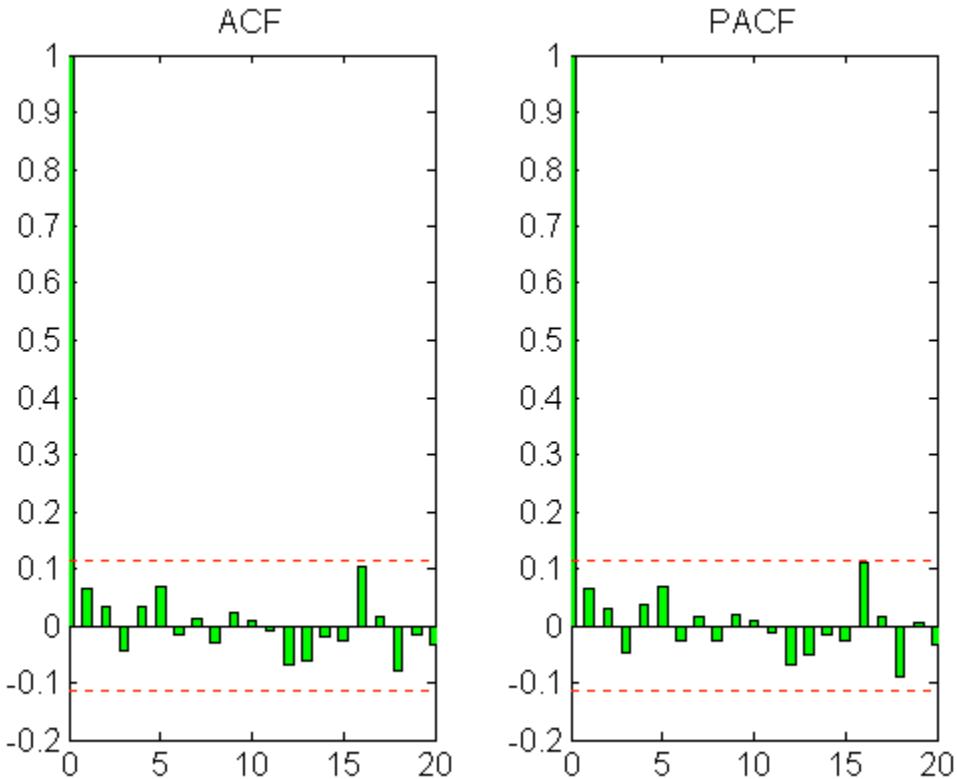


Figure 4.4: Carrés des résidus



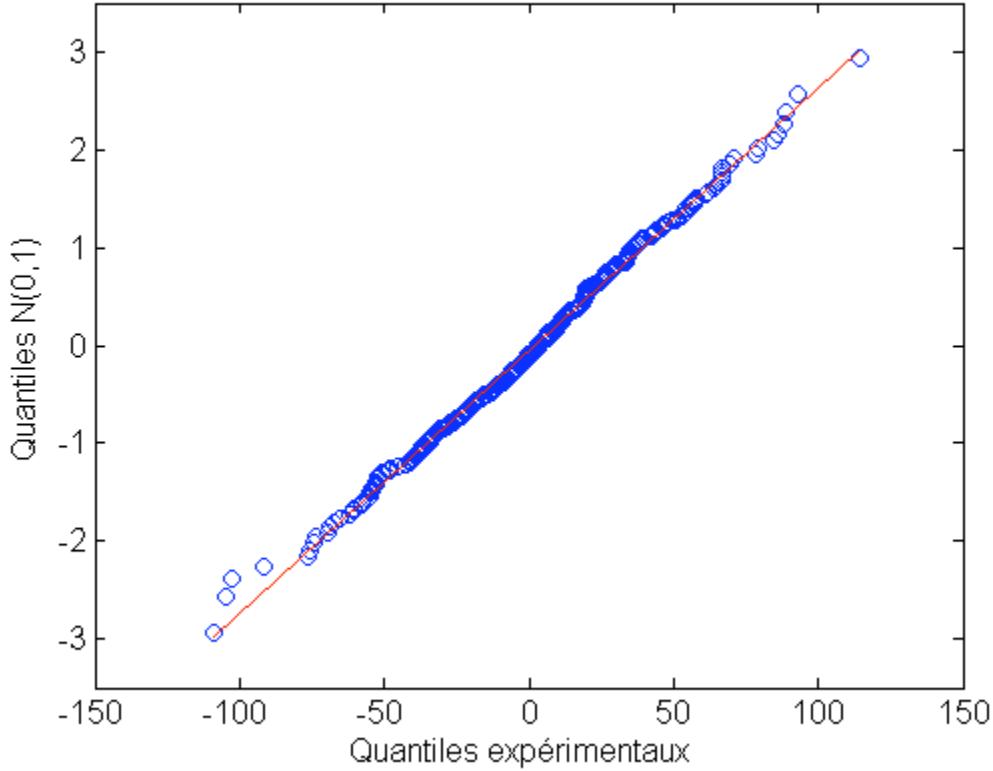
Les résultats sont plutôt encourageants. toutefois, il faut savoir que – travaillant avec des résidus estimés – les bandes de confiance pour les coefficients de corrélation des résidus estimés sont trop larges pour de petits décalages<sup>6</sup> ( $h \leq 10$  en pratique). On doit donc utiliser les conclusions avec prudence: ces graphiques servent surtout à invalider le modèle.

Avant de passer à des tests plus rigoureux, on peut également contrôler visuellement la normalité des résidus en traçant la droite de Henri. On constate ici que les quantiles expérimentaux sont parfaitement alignés<sup>7</sup>.

---

<sup>6</sup>Voir Brockwell and Davis (2002) section 9.4

<sup>7</sup>Un tel alignement est plutôt exceptionnel!



#### 4.4.2 Tests statistiques

L'utilisation des tests va permettre de décider de manière plus objective si le modèle est valide ou non. Notons qu'il est préférable de faire les tests d'indépendance avant le test d'adéquation à la loi normale, car ce dernier concerne les échantillons<sup>8</sup>.

##### Tests de Portmanteau.

Notons  $(\hat{\rho}_Z(h))$  la fonction d'autocorrélation empirique des résidus estimés  $\hat{Z}_t$ .

Le tracé de l'ACF des résidus vise à vérifier que chaque coefficient  $\hat{\rho}_Z(h)$  est, de façon individuelle, à l'intérieur de la bande de confiance. Les tests de Portmanteau sont basés sur une statistique faisant intervenir collectivement les premières autocorrelations. L'hypothèse  $H_0$  est : "  $Z_1, \dots, Z_n$ " sont indépendantes et identiquement distribuées". On se contentera de présenter le test de Ljung-Box, dont la statistique est :

$$Q_Z = n(n+2) \sum_{j=1}^n \frac{\hat{\rho}_Z^2(j)}{n-j}.$$

---

<sup>8</sup>Données issues de v.a. indépendantes.

Intuitivement une valeur élevée de  $Q_Z$  signifie qu'il y a parmi les  $h$  premières autocorrélations des valeurs significatives et va pousser au rejet de  $H_0$ . Pour les carrés des résidus on considère de façon analogue la statistique :

$$Q_{ZZ} = n \sum_{j=1}^n \left( \frac{\hat{\rho}_{ZZ}(j)}{j} \right)^2$$

où  $\hat{\rho}_{ZZ}(h)$  est la fonction d'autocorrélation empirique des carrés des résidus estimés :

$$\hat{\rho}_{ZZ}(h) = \frac{\sum_{t=1}^{n-h} (\hat{Z}_t^2 - \bar{Z}^2)(\hat{Z}_{t+h}^2 - \bar{Z}^2)}{\sum_{t=1}^n (\hat{Z}_t^2 - \bar{Z}^2)^2}$$

avec  $\bar{Z}^2 = \frac{1}{n} \sum_{t=1}^n \hat{Z}_t^2$ . On a alors le résultat suivant, relatif à un modèle ARMA(p,q) sous des hypothèses que l'on supposera vérifiées.

*Théorème 4.* Si les variables aléatoires  $Z_1, \dots, Z_n$  sont i.i.d., alors

- (i)  $Q_Z$  est approximativement de loi  $\chi_{h-(p+q)}^2$
- (ii)  $Q_{ZZ}$  est approximativement de loi  $\chi_h^2$

Le choix de  $h$  est laissé à l'utilisateur. On peut choisir plusieurs valeurs de  $h$  inférieures à  $n$  (ce seuil intervient dans les hypothèses du théorème).

## Test de Kendall

Il est bon d'effectuer un test d'indépendance qui ne fasse pas intervenir le coefficient de corrélation linéaire. Le test de Kendall en est un. L'hypothèse  $H_0$  est inchangée (cependant on applique le test aux résidus estimés). La statistique  $T$  du test est le nombre de "retournements", ou changements de pente, de la série de données  $z_1, \dots, z_n$ . Un retournement désigne une configuration de trois nombres  $a, b, c$  avec  $a < b$  et  $b > c$  ou  $a > b$  et  $b < c$ . On a alors :

*Théorème 5.* Si les variables aléatoires  $Z_1, \dots, Z_n$  sont i.i.d., alors  $T$  est asymptotiquement normale  $(T, \mathcal{N}(\mu_T, \sigma_T^2))$  avec  $\mu_T = \frac{n(n-1)}{4}$  et  $\sigma_T^2 = \frac{n(n-1)(2n+5)}{72}$ .

La règle de décision est cette fois bilatérale. Un trop grand nombre de retournements indique des variations trop fréquentes pour un bruit blanc (corrélation négative); à l'inverse trop peu de changements de pente indique une régularité suspecte (corrélation positive).

## Test de Kolmogorov

On renvoie le lecteur au cours de 1ère année où ce test est présenté en détail.

On peut maintenant présenter le résultat des tests sur la série étudiée.

Test	Largeur	Résidus		Carrés des résidus	
		Statistique	p-valeur	Statistique	p-valeur
Ljung-Box	4	1.9767	0.5772	2.4591	0.6520
	8	6.1874	0.5180	4.2703	0.8319
	12	8.8336	0.6372	5.9859	0.9168
	16	19.0694	0.2106	10.8931	0.8160
Kendall			0.8906		
Kolmogorov		0.0374			

Les p-valeurs pour les tests d'indépendance sont satisfaisantes. La statistique de Kolmogorov  $D_n$  est fournie. Au niveau 5%, on rejette l'hypothèse de normalité si  $\sqrt{n} + 0.85/\sqrt{n} - 0.01)D_n$  est supérieur à 0.895. Le membre de gauche est évalué à 0.64 ce qui est acceptable.

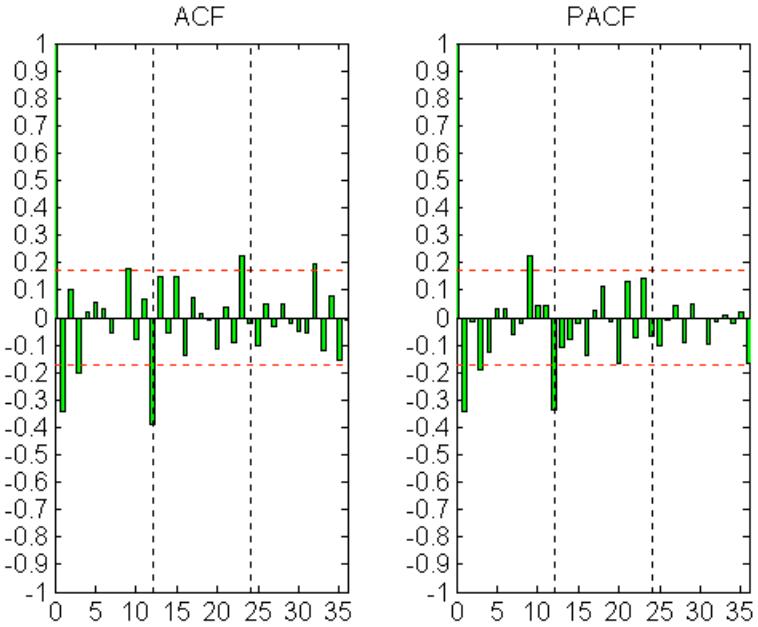
En conclusion, on valide le modèle MA(1) pour la série différenciée. Pour la série initiale du nombre de jeunes femmes sans emploi aux Etats-Unis, on peut donc proposer le modèle ARIMA(0,1,1) :

$$X_t = X_{t-1} + Z_t - 0.51Z_{t-1} \text{ avec } Z_t \sim WN(0, 37, 45^2)$$

## 4.5 Retour sur la série 'airline'.

Que donne la méthodologie de Box et Jenkins sur cette série ? La première étape a été effectuée lors du chapitre 1 où l'on s'est ramené à une série stationnaire ( $Z_t$ ) avec  $Z_t = \nabla\nabla_{12}X_t$ . La sélection d'un modèle probabiliste pour cette ( $Z_t$ ) est plus délicate. ACF et PACF sont représentés ci-dessous jusqu'au rang  $h = 36$ .

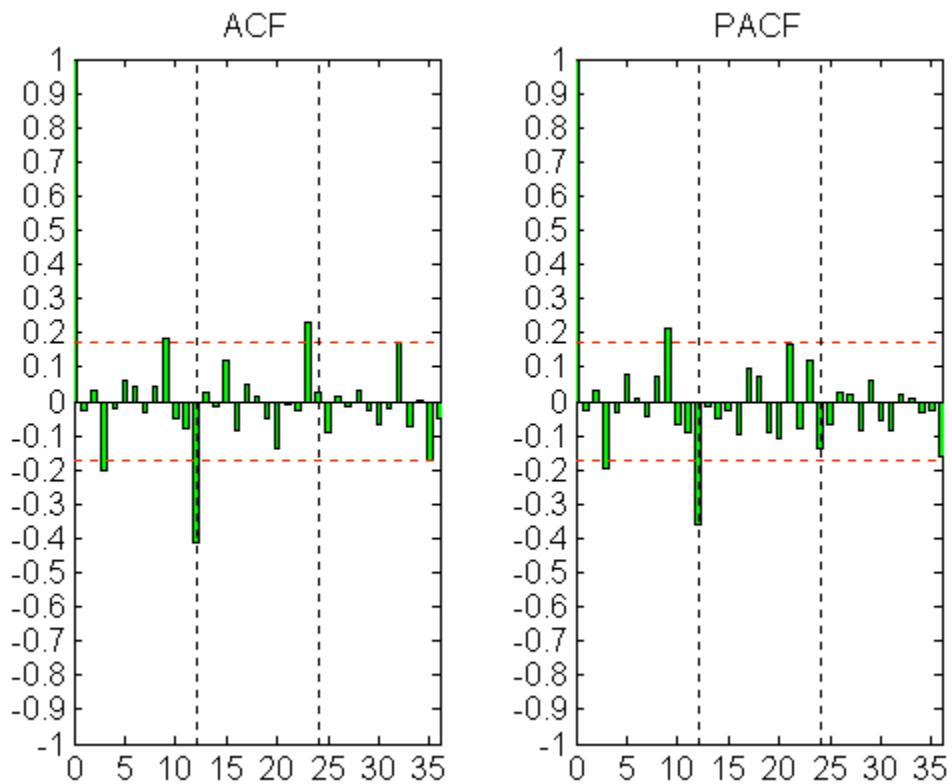
On remarque tout d'abord que parmi les premières autocorrélations, seule la première est



”clairement” significative, ce qui évoque un processus de type MA(1). Cependant, il y a des corrélations importantes au rang 12. Si l’on se souvient qu’il y a une saisonnalité annuelle marquée, cette corrélation a trait aux séries saisonnières. Si l’on regarde donc les autocorrélations de 12 en 12, on remarque que seule la première est significative, ce qui évoque un processus MA(1). Tout ceci invite à proposer un modèle du type SARIMA(0,0,1)(0,0,1)<sub>12</sub> pour  $(Z_t)$  :

$$(I - B)(I - B^{12})\log X_t = (I - 0.40B)(I - 0.53B^{12})U_t \quad \text{avec } U_t \sim WN(0, 0.037^2)$$

Figure 4.5: Corrélogrammes de  $\hat{V}_t$



## Chapter 5

# Prévision avec un modèle probabiliste

Dans le cadre probabiliste fixé au chapitre 3, les données  $x_1, x_2, \dots, x_N$  sont vues comme une réalisation particulière d'un processus aléatoire  $(X_t)_{t \in \mathbb{R}}$ . À l'issue des chapitres 3 et 4, on suppose qu'on connaît la loi du processus, donc la loi de tout vecteur  $(X_t, \dots, X_{t+h})$ . Dans ce contexte, à quoi correspond la notion de prévision ? Comment la calcule-t-on ?

### 5.1 Qu'est-ce qu'une prévision ?

Plaçons-nous à la date  $t$ , et notons  $\mathcal{F}_t$  l'information connue à cette date. Que peut-on prévoir à l'horizon  $h$  ? On peut donner, au moins théoriquement, la loi de probabilité conditionnelle de  $X_{t+h}$  sachant  $\mathcal{F}_t$ . La prévision au sens large est donc cette loi, que l'on appellera "loi de prévision". On retrouve alors la notion de prévision au sens commun comme la valeur moyenne de la loi de prévision, mais l'on a bien plus : on connaît les fluctuations autour de cette valeur.

#### Définition

La prévision en  $t$  à l'horizon  $h$ , notée  $\hat{X}(t, h)$ , est

$$\hat{X}(t, h) = \mathbb{E}(X_{t+h} | \mathcal{F}_t)$$

L'erreur de prévision correspondante, notée  $e(t, h)$ , est définie par

$$e(t, h) = \text{var}(X_{t+h} | \mathcal{F}_t) = \mathbb{E}[(X_{t+h} - \hat{X}(t, h))^2 | \mathcal{F}_t]$$

La prévision et l'erreur de prévision sont des variables aléatoires, fonctions de  $X_t, X_{t-1}, \dots$ . En pratique,  $\mathcal{F}_t$  représente l'information apportée par  $X_t, X_{t-1}, \dots, X_1$  sous la condition  $X_0 = 0, X_{-1} = 0, \dots$ . En revenant au jeu de données, la prévision que l'on donne est :

$$\hat{x}(t, h) = \mathbb{E}(X_{t+h} | X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1, X_0 = 0, X_{-1} = 0, \dots)$$

L'erreur de prévision est généralement insuffisante pour décrire les fluctuations autour de la prévision (cas d'une loi dissymétrique par exemple), sauf dans le cas gaussien. En fait, le prévisionniste recherche un intervalle de prévision.

**Definition 14.** Un intervalle de prévision est un intervalle contenant la prévision avec une probabilité donnée. Ainsi  $I_\alpha(t, h)$  est un intervalle de prévision au niveau  $\alpha$  de  $\hat{X}(t, h)$  ssi :

$$\mathbb{P}(\hat{X}(t, h) \in I_\alpha(t, h) | \mathcal{F}_t) = \alpha$$

Il ne s'agit donc pas d'un intervalle de confiance...  
<sub>1 2</sub>

**Proposition 2.** Si  $X_t$  est un processus gaussien, l'intervalle

$$[\hat{X}(t, h) - 1.96 \cdot e(t, h); \hat{X}(t, h) + 1.96 \cdot e(t, h)]$$

est un intervalle de prévision au niveau 95

## 5.2 Calcul des prévisions

### 5.2.1 Calcul explicite

Dans les cas simples, il est tout à fait possible de calculer analytiquement la prévision et l'erreur de prévision. À l'exception du cas gaussien, la loi de prévision est plus difficile à obtenir. Les résultats suivants sont laissés en exercice.

**Exercice 1.** Prévision pour le modèle AR(p) gaussien

Soit  $X_t$  le processus AR(p) causal et gaussien,

$$\Phi(B)X_t = Z_t$$

avec  $Z_t \sim WN(0, \sigma^2)$  et  $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p$ .

1. Vérifier que :

- a)  $\hat{X}(t, h) = \phi_1 \hat{X}(t, h-1) + \phi_2 \hat{X}(t, h-2) + \dots + \phi_p \hat{X}(t, h-p)$   
avec  $\hat{X}(t, -i) = 0$  pour  $i \geq 0$ .
- b)  $e^2(t, h) = \sigma^2(1 + \psi_1^2 + \psi_2^2 + \dots)$

où  $\psi_1, \psi_2, \dots$  sont les coefficients de la forme MA( $\infty$ ),  $X_t = \Psi(B)Z_t$ .

2. Expliciter la prévision et l'erreur de prévision dans le cas  $p = 1$ . Donner un intervalle de prévision dans le cas gaussien. Que se passe-t-il lorsque  $h$  tend vers l'infini?

**Exercice 2.** Prévision pour le modèle MA(q) gaussien.

Soit  $X_t$  le processus MA(q) inversible et gaussien,

$$X_t = \Theta(B)Z_t$$

avec  $Z_t \sim WN(0, \sigma^2)$  et  $\Theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \dots + \theta_q z^q$ .

---

<sup>1</sup>Il s'agit d'un abus de langage. L'erreur de prévision désigne, à proprement parler, l'écart entre la valeur inconnue  $X_{t+h}$  et la prévision  $\hat{X}(t, h)$ . Il faudrait parler d'erreur quadratique moyenne de prévision.

<sup>2</sup>... pour un processus centré.

Montrer que :

$$a) \quad \hat{X}(t, h) = \sum_{j=0}^{\min(q, h-1)} \theta_j Z_{t+h-j} \quad \text{et vaut donc } 0 \text{ si } h > q.$$

$$b) \quad e(t, h) = \sigma \left( \sum_{j=0}^{\min(q, h-1)} \theta_j^2 \right)^{1/2}$$

### Retour sur la série 'unemp'

Pour cette série, on a proposé le modèle ARIMA(0,1,1) gaussien. On peut facilement déduire de l'exercice précédent la loi de prévision :

**Proposition 3.** Soit  $X_t$  le processus ARIMA gaussien

$$X_t = X_{t-1} + Z_t + \theta Z_{t-1} \quad (*)$$

avec  $Z_t \sim WN(0, \sigma^2)$ . Alors,

$$X_{t+h} | \mathcal{F}_t \sim N\left(X_t + \theta Z_t, \sigma^2(1 + (h-1)(1 + \theta^2))\right)$$

La valeur de  $Z_t$  s'obtient de proche en proche à partir de (\*) et une condition d'initialisation, par exemple  $Z_0 = 0$ . En fait, **les prévisions à un pas<sup>3</sup> sont simplement données par la méthode de lissage exponentiel**. En effet, on a

$$\begin{aligned} \hat{X}(t, 1) &= X_t + \theta Z_t \\ &= X_t + \theta(X_t - (X_{t-1} + \theta Z_{t-1})) \\ &= (1 + \theta)X_t - \theta \hat{X}(t-1, 1) \end{aligned}$$

ce qui correspond à la méthode SES avec  $\alpha = 1 + \theta$  lorsque  $\theta \in ]-1; 0[$ . On vérifie que les prévisions à un horizon quelconque coïncident également.

Autrement dit, les prévisions obtenues par le modèle ARIMA sont les mêmes que celles obtenues par lissage exponentiel. On peut donc dans ce cas donner des intervalles de prévision pour les prévisions obtenues par SES. Soulignons que si le modèle ARIMA(0,1,1) n'avait pas été validé pour la série de données, la notion même d'intervalle de prévision n'aurait pas de sens dans le cadre SES.

Reprendons le problème de prévision consistant à prédire le chômage pour les 6 derniers mois de l'année 1985 au 30 juin 1985. La procédure de Box et Jenkins décrite au chapitre 4 doit être reconduite en enlevant les 6 dernières valeurs de la série. Le modèle ARIMA(0,1,1) est encore valide; on obtient  $\hat{\theta} = -0.5005$  et  $\hat{\sigma} = 37.06$ . On remarque que  $1 + \hat{\theta} = 0.4995$  est différent du paramètre estimé dans SES,  $\hat{\alpha} = 0.4931$ . Cela vient du fait que la procédure d'estimation de  $\theta$  n'est pas un calcul de moindres carrés simples. La différence est cependant

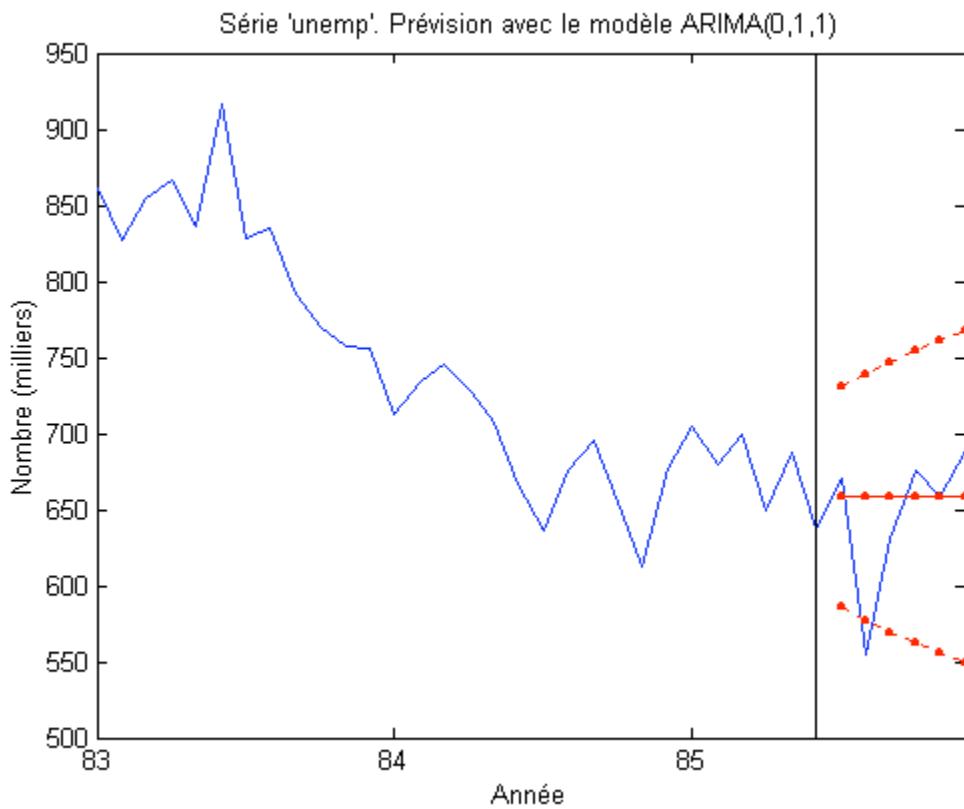
---

<sup>3</sup>... c'est-à-dire à l'horizon 1.

minime (ne pas oublier l'ordre de grandeur de l'erreur d'estimation de  $\theta$ ...). Bien que  $\hat{\theta}$  et  $\hat{\sigma}$  soient estimés, on se contente de faire comme s'il s'agissait des "vrais" paramètres. La loi de prévision est celle indiquée dans la proposition ci-dessus. Finalement, on a :

Mois	Index	Valeur observée	Valeur prédite	Erreur de prévision	Intervalle de prévision	
					Borne inférieure	Borne supérieure
Juillet	295	(670)*	658.3	37.1	585.6	730.9
Août	296	(555)*	658.3	41.4	577.1	739.5
Septembre	297	(631)*	658.3	45.4	569.3	747.2
Octobre	298	(676)*	658.3	49.0	562.2	754.3
Novembre	299	(659)*	658.3	52.4	555.6	760.9
Décembre	300	(689)*	658.3	55.6	549.4	767.2

\*Données inconnues au moment de la prévision.



On observe que les valeurs futures – à l'exception de la donnée du mois d'août - se trouvent dans la fenêtre prévue. Cependant, il faut prendre garde qu'il s'agit des valeurs

réellement observées et non des valeurs futures de la trajectoire du processus. Il faut donc être prudent avec ce genre d'appréciation.

Force est de constater que les prévisions deviennent rapidement très imprécises ! En fait les prévisions pour cette série n'ont de sens qu'à très court terme. D'autre part, en donnant un ordre de grandeur de l'incertitude future, les intervalles de prévision sont sans doute une information plus intéressante que la prévision elle-même (constante...). On ne prévoit pas la baisse du mois d'août 85, mais celle-ci est (relativement) en accord avec la fourchette de prévision à 95%. Il ne faut pas s'étonner de la qualité médiocre des prévisions : peut-on espérer mieux avec aussi peu d'information ? De meilleures prévisions seraient vraisemblablement obtenues en faisant intervenir d'autres variables économiques. On parle alors de séries **multivariées**.

### 5.2.2 Simulation

La loi de prévision peut être approchée au moyen de simulations. D'un point de vue théorique, les simulations permettent d'obtenir des réalisations indépendantes  $X_{t+h}^r, /, r = 1, \dots, R$ , de la loi conditionnelle  $X_{t+h} | \mathcal{F}_t$ . D'après la loi des grands nombres, on dispose d'approximations de la prévision et de l'erreur de prévision, valables pour R "grand" ...

$$\mathbb{E}(X_{t+h} | \mathcal{F}_t) = \bar{X}^R = \frac{1}{R} \sum_{r=1}^R X_{t+h}^r, \quad \text{Var}(X_{t+h} | \mathcal{F}_t) = \frac{1}{R} \sum_{r=1}^R (X_{t+h}^r - \bar{X}^R)^2$$

... mais aussi de la loi de prévision elle-même :

$$P(X_{t+h} \leq x | \mathcal{F}_t) = \frac{1}{R} \sum_{r=1}^R \mathbb{1}_{\{X_{t+h}^r \leq x\}}$$

Pratiquement on utilise les valeurs connues  $x_1, x_2, \dots, x_t$  pour simuler le début d'une *trajectoire* future du processus  $X_{t+1}^r, X_{t+2}^r, \dots, X_{t+h}^r$  conditionnellement aux valeurs passées. On obtient donc d'un seul coup les lois de prévision approchées pour tous les horizons jusqu'à  $h$ .

### Retour sur la série 'airline'

Reprendons le problème consistant à prévoir au 31 décembre 1959 le trafic aérien mensuel de l'année suivante. En suivant la méthodologie de Box et Jenkins pour la série amputée des douze dernières données, on obtient le modèle SARIMA(0,1,1)(0,1,1)<sub>12</sub>:

$$(1 - B)(1 - B^{12}) \log X_t = (1 - \theta_1 B)(1 - \theta_{12} B^{12}) U_t$$

avec  $U_t \sim WN(0, \sigma^2)$ .

Les coefficients estimés sont  $\hat{\theta}_1 = 0.35$ ,  $\hat{\theta}_{12} = 0.53$  et  $\hat{\sigma} = 0.037$ . Pour effectuer les simulations, il est commode de développer. Avec  $\log X_t = Y_t$ , on a

$$Y_{t+k} = Y_t + \theta_1 Y_{t+k-1} - \theta_1 Y_{t+k-13} + \theta_{12} Y_{t+k-12} - \theta_{12} \theta_1 Y_{t+k-13} + U_{t+k} - \theta_1 U_{t+k-1} - \theta_{12} U_{t+k-12} + \theta_{12} \theta_1 U_{t+k-13}$$

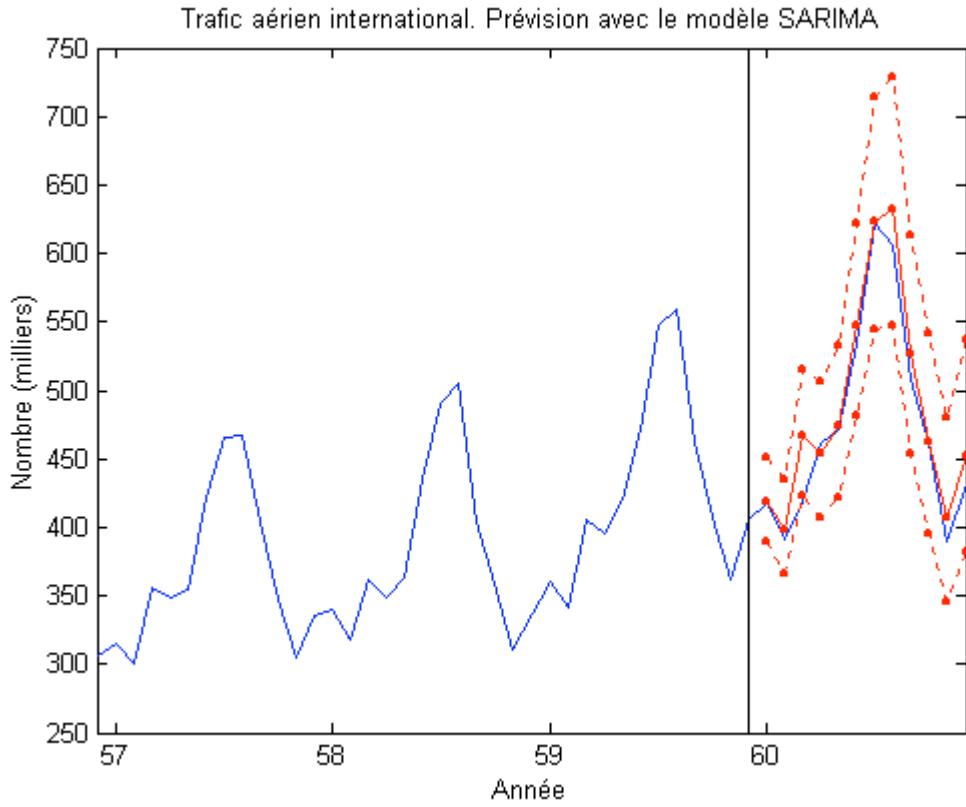
Les valeurs  $x_1, x_2, \dots, x_t$  étant données, on déduit  $y_1, y_2, \dots, y_t$  puis  $z_1, z_2, \dots, z_t$  (avec la condition  $x_i = 0$  pour  $i \leq 0$ ). Conditionnellement à ces valeurs, on obtient en simulant  $U_{t+1}$  selon la loi  $\mathcal{N}(0, \hat{\sigma}^2)$  des réalisations indépendantes de  $Y_{t+1}$ . Puis de proche en proche de  $Y_{t+2}, \dots, Y_{t+h}$ . Le processus étant gaussien, on dispose d'intervalles de prévisions relatifs à  $Y_t$ .

Il faut ensuite revenir à  $X_t$ .

Ci-dessous, on donne les résultats obtenus pour  $R = 10^5$  simulations.

Mois	Index	Valeur observée	Valeur prédictive	Intervalle de prévision	
				Borne inférieure	Borne supérieure
Janvier	133	(417)*	419.6	390.3	451.0
Février	134	(391)*	398.7	365.9	434.4
Mars	135	(419)*	466.6	423.1	514.6
Avril	136	(461)*	454.3	407.5	506.5
Mai	137	(472)*	473.8	421.1	533.1
Juin	138	(535)*	547.4	482.0	621.6
Juillet	139	(622)*	623.2	544.3	713.5
Août	140	(606)*	631.8	547.7	728.9
Septembre	141	(508)*	527.1	453.4	612.7
Octobre	142	(461)*	462.7	395.3	541.6
Novembre	143	(390)*	407.1	345.4	479.7
Décembre	144	(432)*	452.6	381.6	536.7

\*Données inconnues au moment de la prévision.



### 5.2.3 Bootstrap

Les simulations précédentes sont rendues possibles par la connaissance de la loi du bruit blanc  $U_t$ . Dans certaines situations, il peut arriver que l'on n'ait pas de raison de rejeter l'indépendance de  $U_1, \dots, U_t$  alors que l'on observe des défauts évidents de normalité. Dans ce cas, on peut utiliser les résidus estimés  $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_t$ :

- soit en ajustant une loi paramétrique à  $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_t$ <sup>4</sup>, et en effectuant des simulations à partir de cette loi. On parle de bootstrap paramétrique. - soit en rééchantillonnant directement à partir de  $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_t$ <sup>17</sup>; on parle de bootstrap non paramétrique.

Dans les situations où à la fois l'indépendance et la normalité sont douteuses, il est préférable d'envisager un autre modèle.

## 5.3 Optimalité des méthodes de lissage exponentiel

Pour la série 'unemp', modélisée par un processus ARIMA(0,1,1), on a observé que les prévisions sont les mêmes que par lissage exponentiel. En revanche, il est aisément vérifiable que pour un processus de type AR(1), par exemple:

$$X_t = 0.8X_{t-1} + Z_t, \quad Z_t \sim WN(0, 1)$$

---

<sup>4</sup>Voir le cours de 1ère année, paragraphe "simulation d'une loi discrète".

les prévisions probabilistes sont bien plus satisfaisantes. En fait, une méthode de lissage exponentiel est optimale lorsque les erreurs de prévision sont totalement imprévisibles, c'est-à-dire lorsqu'elles forment un bruit blanc. Ici, on ne s'intéresse qu'aux erreurs de prévisions à l'horizon 1.

On peut montrer (voir Granger and Newbold (2014)) Section 5.3 :

- La méthode SES est optimale si et seulement si la série est générée par le modèle ARIMA(0,1,1)

$$(1 - B)X_t = (1 + \theta B)Z_t$$

avec  $\theta = \alpha - 1$ .

- La méthode de Holt est optimale nécessairement pour un modèle de type ARIMA(0,2,2)

$$(1 - B)^2 X_t = (1 + \theta_1 B + \theta_2 B^2) Z_t$$

où  $\theta_1$  et  $\theta_2$  sont des fonctions de  $\alpha$  et  $\beta$ .

- La méthode de Holt-Winters, version additive, est optimale nécessairement pour un modèle de type SARIMA

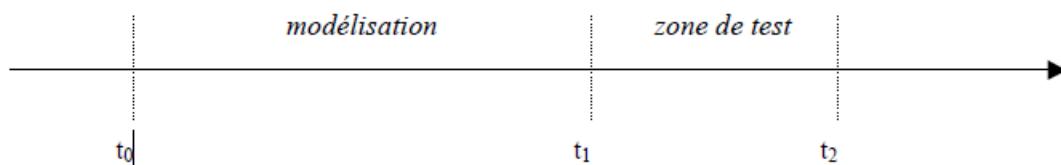
$$(1 - B^s)(1 - B)X_t = (1 + b_1 B + b_2 B^2 + \cdots + b_s B^s) Z_t$$

où  $b_1, b_2, \dots, b_s$  sont des fonctions de  $\alpha, \beta$  et  $\gamma$ . Le modèle est donc contraint (les  $b_k$  sont "liés").

- La méthode de Holt-Winters, version multiplicative, ne correspond à aucun modèle particulier de type SARIMA.

## 5.4 Performances en termes de prévision. Analyse post-sample

La performance des prévisions dépend de l'importance des erreurs de prévision. Cependant jusqu'à présent, seules les erreurs théoriques ont été envisagées mesurant l'écart entre la valeur prédite et la valeur "de modèle". Il serait plus judicieux de comparer les prévisions avec les valeurs réellement observées. Pratiquement, on divise l'intervalle de temps en 2 parties dans l'ordre chronologique :



- La première détermine les données servant à la modélisation ;
- La seconde est utilisée comme zone de test.

On compare alors les prévisions avec les données postérieures à celles utilisées pour la modélisation. On parle de données post-sample et d'évaluation post-sample des erreurs. Pour évaluer la qualité des prévisions à l'horizon  $h$ , on calcule les erreurs réelles

$$er(t, h) = \hat{x}(t, h) - x_t, \quad t_1 < t \leq t_2$$

et on peut mesurer leur importance au moyen d'indicateurs classiques tels que

- l'erreur moyenne absolue (MAE)

$$MAE(h) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2-h} |er(t, h)|$$

- l'erreur quadratique moyenne (MSE)

$$MSE(h) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2-h} er^2(t, h)$$

On peut également vouloir évaluer la qualité des prévisions sans distinction d'horizons. Si  $H$  est l'horizon maximum d'intérêt, un indicateur correspondant à cet objectif est par exemple

$$MSE(t) = \frac{1}{H} \sum_{h=1}^H er(t, h)^2, \quad t_1 < t \leq t_2 - H$$

# Chapter 6

## Hétéroscédasticité

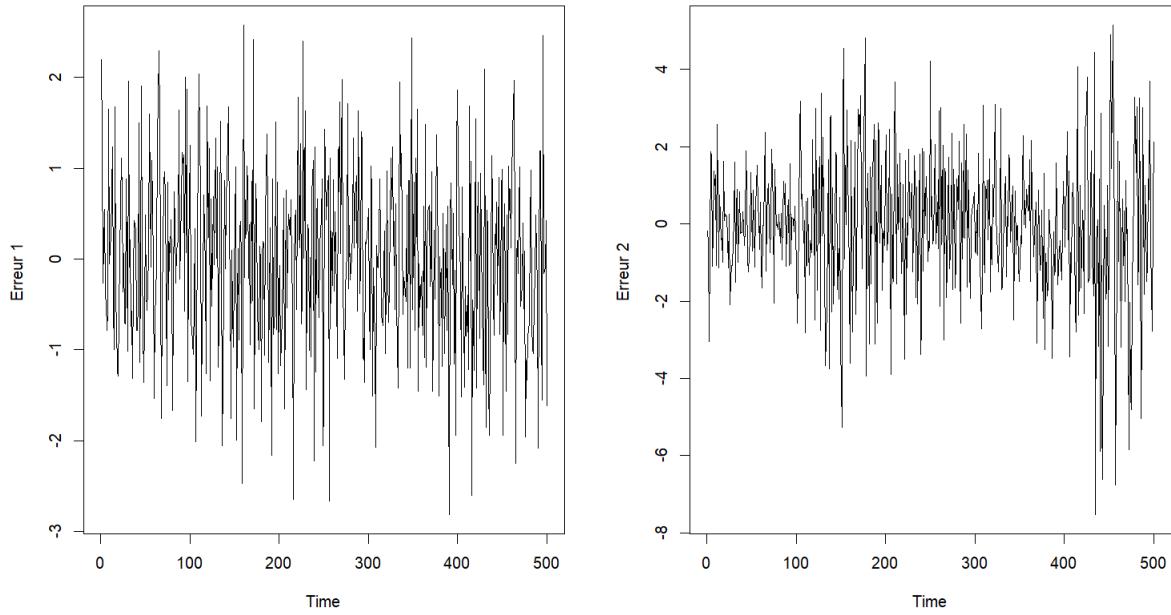
### 6.1 Introduction et définition générale

L'hétéroscédasticité est une caractéristique fondamentale des séries temporelles financières et économiques, où la variance conditionnelle des erreurs n'est pas constante au fil du temps. Contrairement à l'homoscédasticité, qui suppose une variance conditionnelle constante, l'hétéroscédasticité permet de mieux modéliser les phénomènes où les erreurs fluctuent en intensité en fonction de différentes périodes ou événements.

Dans ce chapitre, nous allons explorer la nature de l'hétéroscédasticité et son importance dans l'analyse des séries temporelles. Nous commencerons par définir ce concept et discuter des implications qu'il peut avoir sur les estimations et les prévisions économiques. Ensuite, nous aborderons plusieurs tests statistiques permettant de détecter la présence d'hétéroscédasticité, tels que le test de White et le test de Breusch-Pagan.

Enfin, nous introduirons le modèle GARCH (Generalized Autoregressive Conditional Heteroskedasticity), l'un des modèles les plus utilisés pour capturer la dynamique de la variance conditionnelle dans les séries temporelles. Ce modèle, développé par Robert Engle en 1982 et généralisé par Tim Bollerslev en 1986, offre un cadre robuste pour modéliser et prévoir la volatilité des séries financières.

Pour illustrer cette idée, nous prendrons deux exemples portant sur deux types d'erreurs. Il est facile de constater que les fluctuations, qui nous donnent une idée de la variance conditionnelle d'une série, sont constantes dans la figure à gauche. En revanche, les erreurs dans la figure à droite présentent des fluctuations qui varient au cours du temps. Ce phénomène est appelé hétéroscédasticité.



**Definition 15.** En statistique, une séquence de variables aléatoires est dite conditionnellement hétérosclélastique si ces variables ont des variances différentes. Au contraire, une séquence de variables est appelée homosclélastique si la variance est constante pour chaque variable.

Par exemple, la séquence  $Y_t$ ,  $t = 1, \dots, n$ , est dite hétérosclélastique si la variance conditionnelle de  $Y_t$  sachant  $Y_t$  change avec le temps. **Ainsi, la variance conditionnelle n'est plus constante et doit être modélisée.**

Historiquement, l'analyse de la régression linéaire en présence d'hétérosclélasticité a été étudiée par Engle (1982). L'hétérosclélasticité n'affecte pas seulement les paramètres estimés par la méthode des moindres carrés, mais aussi l'écart-type des estimateurs.

Pour illustrer cela, considérons l'exemple suivant : soit

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$$

où  $\text{Var}(\varepsilon_t | Y_{t-1}, \dots) = \sigma_t^2 = \sigma_\varepsilon^2 Z_t^2$ , et  $Z_t$  est un BB.  $\text{Var}(\varepsilon_t | Y_{t-1}, \dots)$  n'est plus constante, mais dépend du carré de  $Y_{t-1}, \dots$  (si  $Z_t \nearrow \Rightarrow \sigma_t \nearrow$ ).

## 6.2 Détection de l'hétérosclélasticité - Test de White

Ce test respecte les 3 étapes suivant:

1. On estime un modèle de régression multiple:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \varepsilon_t$$

2. On estime la régression auxiliaire. On regresse le carré des résidus  $\hat{\varepsilon}_t^2$  sur les variables  $X_{it}$ ,  $i = 1, \dots, k$ , sur leurs carrés  $X_{it}^2$  et sur leur produit 2 à 2.

$$\begin{aligned}\hat{\varepsilon}_t^2 &= a_0 + a_1 X_{1t} + a_2 X_{2t} + \dots + a_k X_{kt} \\ &\quad + b_1 X_{1t}^2 + b_2 X_{2t}^2 + \dots + b_k X_{kt}^2 \\ &\quad + c_{1,2} X_{1t} X_{2t} + \dots + c_{1,k} X_{1t} X_{kt} \\ &\quad + c_{2,3} X_{2t} X_{3t} + \dots + c_{2,k} X_{2t} X_{kt} \\ &\quad + \dots \\ &\quad + c_{k-1,k} X_{k-1,t} X_{kt} \\ &\quad + u_t, \quad \text{où } u_t \sim BB(0, \sigma_u^2).\end{aligned}$$

On inspecte  $R^2$  le coef. de détermination

$$R^2 = 1 - \frac{\sum u_t^2}{\sum y_t^2} = 1 - \frac{SCR}{SCT}$$

3. On teste l'hypothèse nulle de l'hétéroscléasticité

$$\begin{aligned}H_0 : a_1 &= a_2 = \dots = a_k = b_1 = b_2 = \dots = b_k \\ &= c_{1,2} = \dots = c_{k-1,k} = 0\end{aligned}$$

contre

$$H_1 : \text{il existe au moins un paramètre non nulle.}$$

On calcule la stat de White

$L = TR^2 \sim \chi^2(p)$

où  $T$  est la taille de la série chronologique des erreurs et  $p$  est le nombre des variables.  
Si  $L_{obs} < \chi^2_{theo} \implies$  on rejette  $H_0$  et **on conclut la présence d'hétéroscléasticité**.

*Dans ce cours, nous allons concentrer sur un modèle de séries chronologiques pour le cas d'hétéroscléasticité qui est le modèle **ARCH**.*

### 6.3 Test de Breusch-Pagan

Le test de Breusch-Pagan est un test statistique qui permet de détecter la présence d'hétéroscléasticité dans un modèle de régression linéaire. Ce test repose sur l'idée que, sous l'hypothèse nulle d'homoscléasticité, les résidus au carré devraient être indépendants des variables explicatives. Le test se déroule en deux étapes :

1. On estime le modèle de régression initial et on récupère les résidus  $\hat{\varepsilon}_i$ .

- On effectue ensuite une régression auxiliaire où les résidus au carré sont régressés sur les variables explicatives du modèle initial. Le test de Breusch-Pagan se base sur la statistique de Lagrange Multiplier (LM), qui est asymptotiquement distribuée selon une loi du chi-deux sous l'hypothèse nulle d'homoscédasticité.

La statistique du test de Breusch-Pagan, notée LM, est utilisée pour tester l'hypothèse nulle d'homoscédasticité contre l'alternative où la variance des erreurs est une fonction linéaire des variables explicatives.

- Estimation des résidus :** Considérons le modèle de régression linéaire

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

où  $y_i$  est la variable dépendante,  $x_{i1}, x_{i2}, \dots, x_{ik}$  sont les variables explicatives,  $\beta_0, \beta_1, \dots, \beta_k$  sont les coefficients à estimer, et  $\varepsilon_i$  est le terme d'erreur.

Après avoir estimé les coefficients du modèle par moindres carrés ordinaires (MCO), on obtient les résidus estimés  $\hat{\varepsilon}_i = y_i - \hat{y}_i$ .

- Régression auxiliaire :** On effectue ensuite une régression auxiliaire où les carrés des résidus  $\hat{\varepsilon}_i^2$  sont régressés sur les variables explicatives du modèle original :

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_k x_{ik} + u_i.$$

Ici,  $\alpha_0, \alpha_1, \dots, \alpha_k$  sont les nouveaux coefficients à estimer, et  $u_i$  est le terme d'erreur de cette régression auxiliaire.

- Calcul de la statistique LM :** La statistique du test de Breusch-Pagan est calculée comme suit :

$$LM = \frac{n \times R^2}{2},$$

où  $n$  est le nombre d'observations et  $R^2$  est le coefficient de détermination de la régression auxiliaire. Cette statistique suit asymptotiquement une distribution du chi-deux avec  $k$  degrés de liberté sous l'hypothèse nulle d'homoscédasticité.

- Interprétation :** Si la statistique LM est grande, ou si la p-valeur associée est inférieure à un seuil de signification (généralement 0,05), on rejette l'hypothèse nulle d'homoscédasticité, suggérant la présence d'hétérosclélasticité.

## 6.4 Modèle ARCH(q)

Le modèle autorégressif conditionnellement hétérosclélastique (ARCH) est utilisé dans l'analyse des séries présentant des fluctuations irrégulières, notamment dans les domaines financiers, les sciences sociales, la santé publique, et l'épidémiologie, entre autres.

Le modèle ARCH s'applique lorsque la variance courante du terme d'erreur (innovation) *n'est plus constante*, mais dépend des carrés des innovations des périodes antérieures. Par conséquent, il s'agit d'une forme particulière d'hétérosclélasticité (voir Engle (1982)).

$$\sigma_\varepsilon^2 = f(\varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \dots, \varepsilon_{t-q}^2)$$

Les modèles ARCH sont couramment utilisés pour modéliser des séries financières chronologiques présentant une volatilité marquée, c'est-à-dire des périodes de grandes oscillations suivies par des périodes relativement calmes.

*Prenons un exemple introductif :*

Dans l'analyse traditionnelle de la prévision, la construction des valeurs prévues est fondée sur la moyenne conditionnelle de la série utilisée. Ainsi, la prévision de  $Y_t$  à la date  $t+1$  compte tenu du passé est donnée par  $\mathbb{E}(Y_{t+1} | y_t, y_{t-1}, \dots)$ .

*Exemple numérique:*

Soit  $Y_t \sim AR(1)$  de moyenne 0. En d'autres termes,

$$y_t = \phi y_{t-1} + \varepsilon_t \quad \text{où } \varepsilon_t \sim BB(0, \sigma_\varepsilon^2).$$

Il est facile de voir que la moyenne non conditionnelle de  $y_t$  est

$$\mathbb{E}(y_t) = 0,$$

et la moyenne conditionnelle de  $y_t$  est :

$$\begin{aligned} \mathbb{E}(y_t | y_{t-1}) &= \mathbb{E}(\phi y_{t-1} | y_{t-1}) + \mathbb{E}(\varepsilon_t | y_{t-1}) \quad \text{car } \varepsilon_t \text{ est indépendant de } y_{t-1}, \\ &= \phi y_{t-1}. \end{aligned}$$

L'idée de Engle (1982) est de tenir compte des autres moments conditionnels de ce processus :

- La variance non conditionnelle de  $AR(1)$  est :

$$\mathbb{V}ar(y_t) = \frac{\sigma_\varepsilon^2}{1 - \phi^2}, \quad \phi \in [0, 1].$$

- La variance conditionnelle est égale à :

$$\begin{aligned} \mathbb{V}ar(y_t | y_{t-1}) &= \mathbb{V}ar(\phi y_{t-1} | y_{t-1}) + \mathbb{V}ar(\varepsilon_t | y_{t-1}) \\ &= \mathbb{V}ar(\varepsilon_t) \\ &= \sigma_\varepsilon^2. \end{aligned}$$

*Il est facile de voir que  $\mathbb{V}ar(y_t) > \mathbb{V}ar(y_t | y_{t-1})$ .*

C'est pourquoi Engle (1982) a proposé de modéliser simultanément la moyenne (représentée par le modèle  $AR(1)$ ) et la variance d'une série chronologique.

Le principe consiste à supposer que la variance dépend des informations antérieures. Il propose une spécification ARCH(q) où le carré des perturbations suit un processus autorégressif AR(q).

La famille des modèles ARCH peut se décomposer en deux sous ensemble

Modèle ARCH linéaire	Modèle ARCH non-linéaire
ARCH(q)	EGARCH(p,q)
GARCH(p,q)	TARCH(q)
IGARCH(p,q)	TGARCH (p,q)
caractérisé par des spécification quadratique de la variance conditionnelle des perturbations	caractérisé par des spécification asymétriques des perturbations

#### 6.4.1 Définition et représentation

Un processus  $\varepsilon_t$  est dit ARCH(q) s'il satisfait les deux conditions suivantes:

1.  $\mathbb{E}(\varepsilon_t / \varepsilon_u, u < t) = 0, \quad \forall t \in \mathbb{Z}$
2.  $\sigma_t^2 = \text{Var}(\varepsilon_t / \varepsilon_u, u < t) = \alpha_0 + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2$

Le modèle ARCH sera écrit de la façon suivante:

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t \quad \text{où } z_t \sim BB(0, \sigma_\varepsilon^2) \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^n \alpha_i \varepsilon_{t-i}^2. \end{aligned}$$

Autre écriture:

$$\varepsilon_t = z_t \sqrt{\alpha_0 + \sum_{i=1}^n \alpha_i \varepsilon_{t-i}^2}$$

*Exemple:*

Le modèle ARCH(1) s'écrit sous la forme:

$$\varepsilon_t = \sigma_t z_t$$

avec

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \quad \text{et} \quad z_t \sim BB(0, \sigma^2),$$

où  $\sigma^2$  est une constante.

On peut écrire le modèle aussi de la façon suivante

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + u_t,$$

où  $u_t$  est le processus d'innovation pour  $\varepsilon_t^2$  vérifiant  $\mathbb{E}(u_t / \varepsilon_{t-1}) = 0$ .

## 6.5 Autre Modèle

### 1. GARCH (Generalized Autoregressive Conditional Heteroskedasticity)

- Développé par Bollerslev (1987).
- Généralise le modèle ARCH en incluant également des termes autorégressifs.
- Formule générale pour un modèle GARCH(p, q) :

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$$

- Propriétés :
  - Capable de modéliser une dépendance à long terme dans la variance.
  - La variance conditionnelle dépend des erreurs passées et des variances conditionnelles passées.
  - Souvent utilisé dans la modélisation des séries financières pour capturer des dynamiques complexes de la volatilité.

### 2. EGARCH (Exponential GARCH)

- Proposé par Nelson (1992).
- Utilise une fonction logarithmique pour assurer la positivité de la variance conditionnelle.
- Formule générale pour un modèle EGARCH(p, q) :

$$\log(\sigma_t^2) = \omega + \sum_{i=1}^p \beta_i \log(\sigma_{t-i}^2) + \sum_{j=1}^q \alpha_j \left( \frac{\epsilon_{t-j}}{\sigma_{t-j}} - E \left[ \frac{\epsilon_{t-j}}{\sigma_{t-j}} \right] \right)$$

- Propriétés :
  - Capable de modéliser l'asymétrie dans la volatilité.
  - La variance conditionnelle reste positive grâce à la transformation logarithmique.
  - Utilisé pour des séries temporelles où les chocs positifs et négatifs ont des impacts différents sur la volatilité.

### 3. TGARCH (Threshold GARCH)

- Proposé par Francq and Zakoian (2019).
- Modélise la volatilité en tenant compte des effets seuils asymétriques.
- Formule générale pour un modèle TGARCH(p, q) :

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + \sum_{k=1}^r \gamma_k \epsilon_{t-k}^2 I(\epsilon_{t-k} < 0)$$

- Propriétés :

- Modèle bien adapté pour capturer les asymétries dans les séries temporelles financières.
- La variance conditionnelle dépend des erreurs passées, des variances conditionnelles passées, et des chocs négatifs.

*Le domaine des séries temporelles est très vaste et trouve de nombreuses applications dans divers domaines. Pour plus de détails, voir Brockwell and Davis (2002). Autre références: Autre références:*

- *Campbell et al. (1997)*
- *Chatfield and Xing (2019)*
- *Gouriéroux and Monfort (1995)*

# Bibliography

- Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, 542–547.
- Brockwell, P. J. and R. A. Davis (2002). *Introduction to time series and forecasting*. Springer.
- Campbell, J., A. Lo, and A. MacKinlay (1997). The econometrics of financial markets princeton, nj: Princeton univ.
- Chatfield, C. and H. Xing (2019). *The analysis of time series: an introduction with R*. Chapman and hall/CRC.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, 987–1007.
- Francq, C. and J.-M. Zakoian (2019). *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons.
- Gouriéroux, C. and A. Monfort (1995). *Séries temporelles et modèles dynamiques*. FeniXX.
- Granger, C. W. J. and P. Newbold (2014). *Forecasting economic time series*. Academic press.
- Nelson, D. B. (1992). Filtering and forecasting with misspecified arch models i: Getting the right variance with the wrong model. *Journal of econometrics* 52(1-2), 61–90.