# [Analysez et modélisez des séries temporelles](https://openclassrooms.com/fr/courses/4525371-analysez-et-modelisez-des-series-temporelles)

Table des matières

[Analysez et modélisez des séries temporelles 1](#_Toc57987661)

[**Découvrez l'univers des données temporelles** 3](#_Toc57987662)

[**Qu'est-ce qu'une série temporelle ?** 3](#_Toc57987663)

[**Où trouve-t-on des séries temporelles ?** 3](#_Toc57987664)

[**À l'origine des séries temporelles** 3](#_Toc57987665)

[**Familiarisez-vous avec certaines séries temporelles** 5](#_Toc57987666)

[**1️⃣ Population des USA** 5](#_Toc57987667)

[**2️⃣ Passagers aériens** 5](#_Toc57987668)

[**3️⃣ Production de bière** 7](#_Toc57987669)

[**4️⃣ Lynx capturés au Canada** 8](#_Toc57987670)

[**Revenons à notre premier exemple** 9](#_Toc57987671)

[TP : représentez des séries temporelles 11](#_Toc57987672)

[Comprenez les variations saisonnières 13](#_Toc57987673)

[TP : Désaisonnalisez à l'aide de la régression linéaire 15](#_Toc57987674)

[Désaisonnalisez à l'aide des moyennes mobiles 18](#_Toc57987675)

[Une définition 18](#_Toc57987676)

[Un peu de terminologie 18](#_Toc57987677)

[Deux moyennes mobiles classiques 19](#_Toc57987678)

[Utiliser des moyennes mobiles pour corriger des variations saisonnières 19](#_Toc57987679)

[Evaluer les propriétés d'une moyenne mobile 20](#_Toc57987680)

[Découvrez des algorithmes de traitement des moyennes mobiles 22](#_Toc57987681)

[Commençons par un algorithme simple 22](#_Toc57987682)

[TP : Désaisonnaliser une série temporelle à l’aide de la fonction decompose 26](#_Toc57987683)

[Pour aller plus loin : l’algorithme X11 26](#_Toc57987684)

[Partie 2 29](#_Toc57987685)

[Compétences évaluées 29](#_Toc57987686)

[ Question 1 29](#_Toc57987687)

[ Question 2 29](#_Toc57987688)

[ Question 3 29](#_Toc57987689)

[ Question 4 30](#_Toc57987690)

[ Question 5 30](#_Toc57987691)

[Appréhendez le lissage exponentiel simple 32](#_Toc57987692)

[Méthode 32](#_Toc57987693)

[Formule de mise à jour 32](#_Toc57987694)

[**Appréhendez le lissage exponentiel double et la méthode de Holt-Winters** 34](#_Toc57987695)

[**1. lissage exponentiel double** 34](#_Toc57987696)

[TP : Prévoyez une série à l’aide des méthodes de lissage exponentiel 36](#_Toc57987697)

[Partie 3 39](#_Toc57987698)

[Compétences évaluées 39](#_Toc57987699)

[ Question 1 39](#_Toc57987700)

[ Question 2 39](#_Toc57987701)

[Découvrez les processus stationnaires 40](#_Toc57987702)

[Série temporelle et processus stochastique 40](#_Toc57987703)

[La notion de bruit blanc 40](#_Toc57987704)

[Une notion fondamentale : la stationnarité 40](#_Toc57987705)

[Modéliser par des processus stationnaires 41](#_Toc57987706)

[Mesurer la dépendance temporelle linéaire 42](#_Toc57987707)

[Estimer les moments d'un processus stationnaire 43](#_Toc57987708)

[Tester la blancheur d'un résidu 44](#_Toc57987709)

[Vers les processus ARMA 45](#_Toc57987710)

[Les processus AR, MA et ARMA 46](#_Toc57987711)

[Les processus AR 46](#_Toc57987712)

[Les processus MA 46](#_Toc57987713)

[Les processus ARMA 47](#_Toc57987714)

[Les processus non stationnaires : ARIMA et SARIMA 49](#_Toc57987715)

[Les processus ARIMA 49](#_Toc57987716)

[Les processus SARIMA 49](#_Toc57987717)

[Entraînez des modèles SARIMA 51](#_Toc57987718)

[La synthèse 51](#_Toc57987719)

[Stationnarisation 51](#_Toc57987720)

[Identification a priori de modèles potentiels 53](#_Toc57987721)

[Estimation des modèles potentiels 54](#_Toc57987722)

[Vérification des modèles potentiels 54](#_Toc57987723)

[Choix définitif d'un modèle 55](#_Toc57987724)

[Prévision à l'aide du modèle choisi 55](#_Toc57987725)

[Analyse a posteriori 55](#_Toc57987726)

[TP : Prévoyez une série temporelle à l’aide des méthodes SARIMA 57](#_Toc57987727)

[Stationnarisation de la série 57](#_Toc57987728)

[**Familiarisez-vous avec d'autres modèles connus** 71](#_Toc57987729)

[**Les modèles ARMAX** 71](#_Toc57987730)

[**Les modèles ARCH et GARCH** 71](#_Toc57987731)

[**Les modèles à mémoire longue** 71](#_Toc57987732)

[**Les modèles multivariés** 71](#_Toc57987733)

[**D'autres alternatives** 71](#_Toc57987734)

[Partie 4 73](#_Toc57987735)

[Compétences évaluées 73](#_Toc57987736)

[ Question 1 73](#_Toc57987737)

[ Question 2 73](#_Toc57987738)

[ Question 3 73](#_Toc57987739)

[ Question 4 74](#_Toc57987740)

[ Question 5 74](#_Toc57987741)

Vous baignez, tous les jours parmi les **séries temporelles**. Ce sont tout simplement des relevés, **des mesures d’un même phénomène au cours du temps**. Vous les avez trouvées dans vos manuels d’histoire et de géographie, vous les observez au quotidien dans les médias, et les rencontrerez sans nul doute dans votre environnement professionnel.

Ensemble, nous apprendrons à mieux connaître ces séries dans le passé, vous serez ainsi capable de les **désaisonnaliser** ! Par exemple, les chiffres du chômage publiés tous les mois sont corrigés de leur effet mensuel pour en faciliter les comparaisons.

Mais nous irons au-delà, nous chercherons à **prévoir le futur**. Une bonne analyse des comportements passés sera la clé pour inférer le futur. Et pour cela, aucune boule de cristal ne sera requise, nous nous munirons plutôt de modèles comme les **ARMA**… !

**Objectifs pédagogiques :**

* Découvrir différentes séries temporelles
* Comprendre la problématique posée par le lien temporel
* Corriger une série temporelle des variations saisonnières
* Mettre en œuvre des techniques simples pour prévoir une série temporelle

**Prérequis :**

* Probabilité.
* [**Statistique inférentielle.**](https://openclassrooms.com/courses/initiez-vous-a-la-statistique-inferentielle)
* Régression linéaire.

**Découvrez l'univers des données temporelles**

**Qu'est-ce qu'une série temporelle ?**

Dans ce cours, nous allons nous intéresser **aux séries temporelles.** Ce sont des ensembles d’observations qui se distinguent par le rôle important que joue l’ordre dans lequel elles ont été recueillies.

Au-delà du statisticien qui doit mener des travaux de modélisation sur de telles données, tout un chacun est immergé dans un monde de données temporelles, et se confronte naturellement à la notion de prévision, d’inférence.

La règle de la **persistence** (par exemple quand la météo du lendemain est identique à celle du jour) est sans nul doute la prévision la plus basique, très souvent perfectible !

**Où trouve-t-on des séries temporelles ?**

Voici quelques exemples de séries temporelles :

* **Dans le domaine économique** : le taux de croissance, le taux de chômage, les cours d’actions, etc.
* **Dans le domaine démographique** : le taux de natalité, les flux migratoires, etc.
* Et **beaucoup d’autres** : un électrocardiogramme, les températures, le traffic routier, l’énergie consommée, le niveau d’un cours d’eau, etc.

Toutes ces séries temporelles ne subiront pas les mêmes traitements, certaines seront simplement **décrites** (ex : analyse visuelle de courbes), d’autres **analysées a posteriori** via des modèles déterministes (ex : correction des variations saisonnières), d’autres enfin **prévues** (à l’aide de modèles issus de la physique, la statistique inférentielle, du machine learning, etc.).

**À l'origine des séries temporelles**

Les outils à disposition du statisticien trouvent leurs sources dans trois phases historiques :

1️⃣ La représentation **graphique** : les diagrammes apparaissent en astronomie, le plus ancien connu remontant au 10ème siècle.

2️⃣ Les **techniques temporelles déterministes** (en opposition aux techniques probabilistes) : elles apparaissent aux 18ème et 19ème siècles, on distingue deux voies très importantes :

* L’analyse fréquentielle d’une série temporelle : un signal est approché par une  
  somme de sinusoı̈dales (analyse de Fourier).
* La décomposition d’une série temporelle en composantes tendancielle, cyclique,  
  saisonnière et accidentelle.
* La correction des variations saisonnières d’une série temporelle, qu’on étudiera,  
  repose sur cette décomposition.

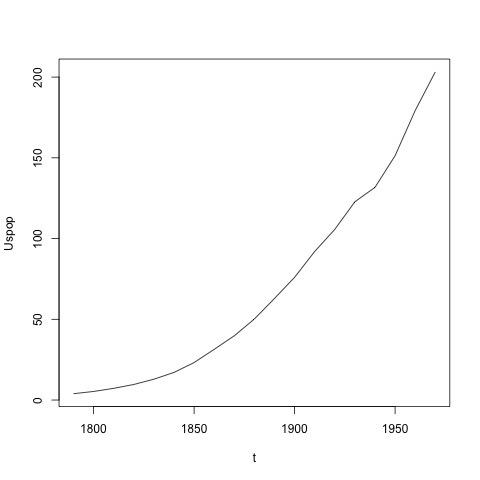
3️⃣ Les méthodes **stochastiques** (ou probabilistes) : elles émergent au 20ème siècle, et ont connu depuis un véritable essor. On étudiera une de ces méthodes : le modèle **ARMA**.

**Familiarisez-vous avec certaines séries temporelles**

Voyons ensemble quelques exemples typiques de séries temporelles.

**1️⃣ Population des USA**

Le graphique suivant représente la population des Etats Unis, uspop, en millions, au pas de temps décennal de 1790 à 1990.



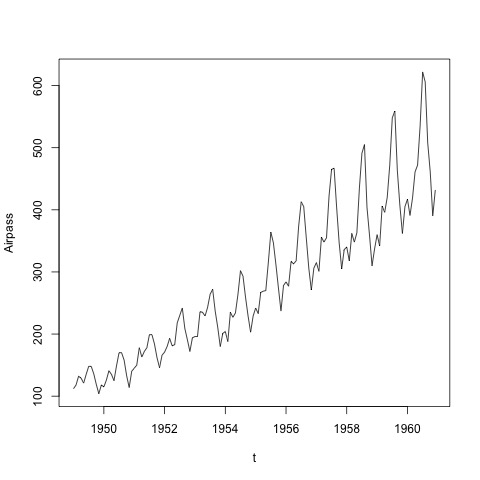
Population des USA entre 1790 à 1990

On constate que son niveau s’accroît avec le temps, de manière quasi-exponentielle : on parle volontiers de tendance (exponentielle) pour décrire ce phénomène.

Par ailleurs, on constate qu’il y a un léger infléchissement dans cette évolution, au moment de la seconde guerre mondiale. On pourrait naturellement considérer ici qu’il s’agit d’un *outlier* (sans pour autant être forcément contraint de le corriger ou de le retirer).

**2️⃣ Passagers aériens**

Le graphique suivant représente le nombre mensuel de passagers aériens, en milliers, de janvier 1949 à décembre 1960. Cette série, fréquemment rencontrées dans les ouvrages traitant des séries temporelles, sera notre fil rouge tout au long de ce cours.

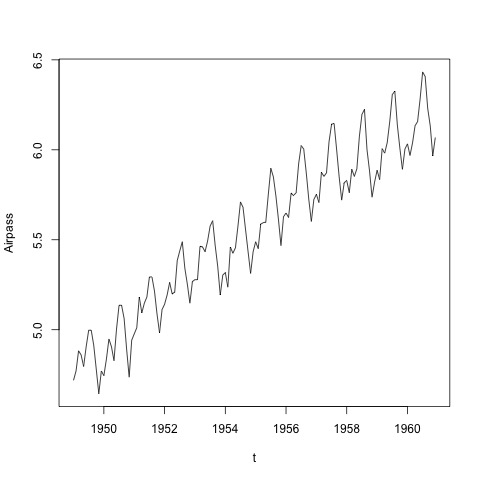


Nombre de passagers aériens entre 1949 et 1960

On constate que son niveau moyen s’accroı̂t avec le temps, de manière quasi-exponentielle ou polynomiale : on traduit encore cette évolution sous le terme de **tendance**.

On observe par ailleurs qu’il y a des variations régulières autour du niveau moyen, de période 12 (correspondant à une période annuelle puisque la série est mensuelle) : on nomme cela l’**effet saisonnier**.

Si on représente cette fois le logarithme de cette série, on obtient :

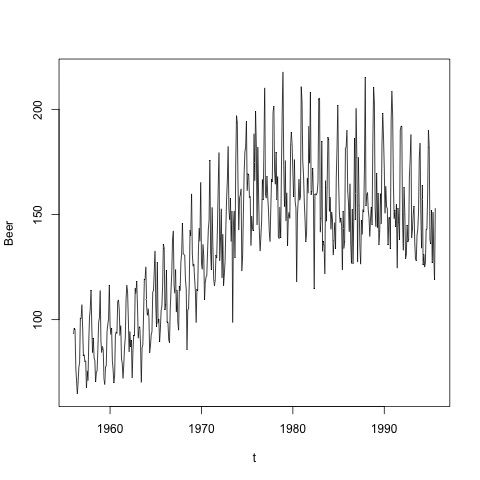


Logarithme du nombre de passagers aériens

On constate que la tendance est **croissante**, linéaire ici**.**

**3️⃣ Production de bière**

Le graphique suivant représente la production mensuelle de bière en Australie, en mégalitres, de janvier 1956 à août 1995.



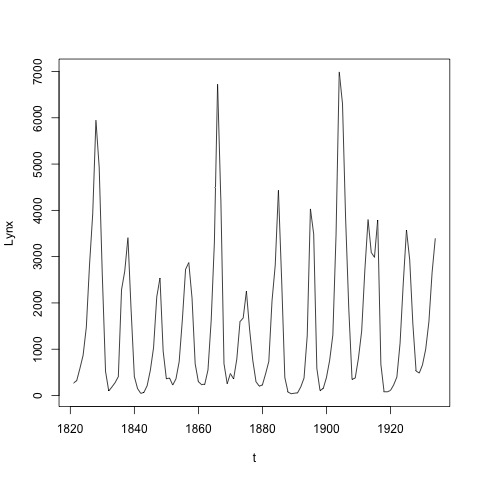
Production mensuelle de bière en Australie de 1956 à 1995

On peut parler ici d’une **tendance** : fortement stable jusqu’à 1964, fortement croissante ensuite jusqu’à 1978, puis stable, etc.

On pourrait également vérifier qu’il y a là-encore une saisonnalité annuelle. Une observation rapide de la série pourrait nous conduire à ne pas forcément considérer la série complète pour prévoir à court terme : il y a sans doute eu un changement de régime.

**4️⃣ Lynx capturés au Canada**

Le graphique suivant représente le nombre annuel de lynx capturés au Canada de 1821 à 1934.



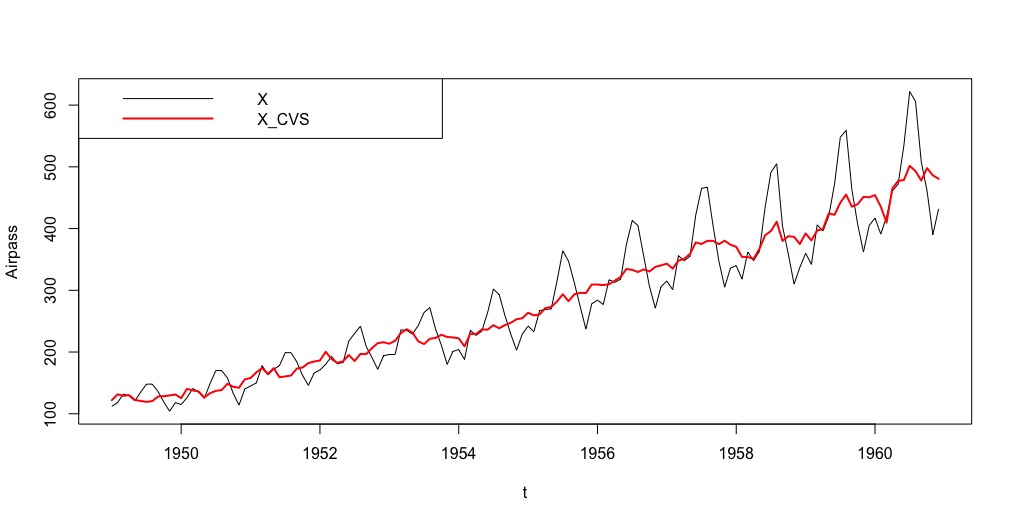
Lynx capturés au Canada de 1821 à 1934

Cette série présente une tendance stable mais une **double saisonnalité** (saisonnalité et cycle).

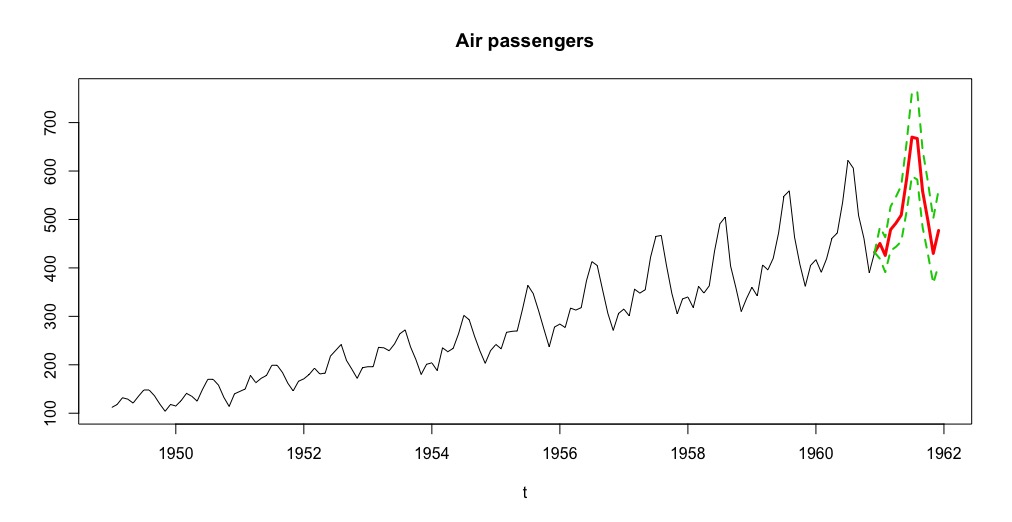
**Revenons à notre premier exemple**

Sur la série temporelle relative aux passagers aériens, notre fil rouge, nous obtiendrons les résultats suivants :

* Une correction en variations saisonnières (à l’aide de la régression linéaire ou de moyennes mobiles).

Correction en variations saisonnières

* Une prévision (à l’aide de lissage exponentiel ou de modèle ARMA) :

Prévision de nombre de passagers aériens

*Voilà donc quelques exemples typiques de séries temporelles. Ces exemples ne sont qu’un panel des cas d’étude auxquels pourront être confrontés les data analysts et data scientists. Observons maintenant de plus près notre exemple fil-rouge.*

## TP : représentez des séries temporelles

Avant de commencer, vous devrez télécharger les jeux de données à [**cette adresse**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/4525371-analysez-et-modelisez-des-series-temporelles/Analysez+et+modelisez+des+series+temporelles.zip). Vous y trouverez aussi le code Python, si vous utilisez ce langage.

Pour représenter les séries vues précédemment, on peut utiliser la commande `plot`. Vous pouvez lancer ces quelques lignes et vérifier le résultat obtenu.

# Série sunspot : nombre annuel de tâches solaires de 1790 à 1970

plot(sunspot.year,xlab="t",ylab="Sunspots")

# Bruit blanc gaussien de loi N(0,3^2)

# Pour les simulations effectuées dans ce document, on fixe arbitrairement la racine (seed) à 1789.

set.seed(1789)

plot(ts(rnorm(100,sd=3),start=1,end=100),xlab="t",ylab="Bruit blanc gaussien de variance 9")

abline(h=0)

# Série uspop : population des Etats-Unis, en millions, de 1790 à 1990 (Pas de temps décennal)

plot(uspop,xlab="t",ylab="Uspop")

# Série airpass : nombre mensuel de passagers aériens, en milliers, de janvier 1949 à décembre 1960

# Série brute :

plot(AirPassengers,xlab="t",ylab="Airpass")

# Logarithme de la série airpass

plot(log(AirPassengers),xlab="t",ylab="Airpass")

# Série beer : production mensuelle de bière en Australie, en mégalitres, de janvier 1956 à aout 1995

beer=read.csv("../Data/beer.csv",header=F,dec=".",sep=",")

beer=ts(beer[,2],start=1956,freq=12)

plot(beer,xlab="t",ylab="Beer")

# Série lynx : nombre annuel de lynx capturés au Canada, de 1821 à 1934

plot(lynx,xlab="t",ylab="Lynx")

Sauf mention contraire, on travaillera dans la suite du cours sur la série temporelle airpass. On la stockera sous la variable x , et son logarithme sous la variable  y  :

x=AirPassengers

y=log(x)

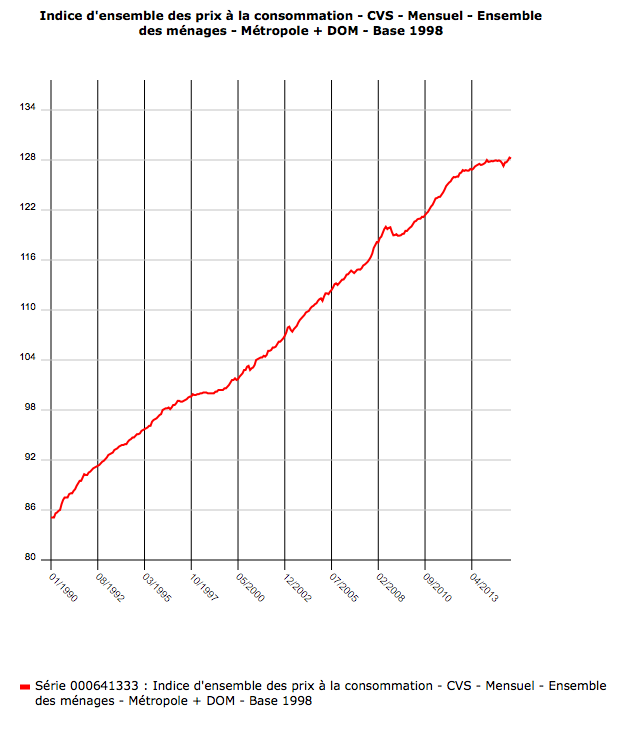
Conservez-bien ces quelques lignes de commande, elles nous seront utiles tout au long de ce cours.

## Comprenez les variations saisonnières

De nombreuses séries économiques présentent des comportements périodiques, rendant difficile la comparaison de deux instants successifs. On traite ainsi fréquemment de séries **trimestrielles** (ex : taux de croissance) ou **mensuelles** (ex : taux de chômage). On fait alors appel à une **désaisonnalisation** permettant d'obtenir des séries dites **corrigées des variations saisonnières (CVS)**.

Ces techniques sont notamment utilisées par les instituts statistiques qui publient des séries temporelles CVS afin d'en faciliter les comparaisons.

La figure suivante illustre le type de courbe désaisonnalisée qu'on peut trouver dans les publications :



Une courbe désaisonnalisée

Les méthodes couramment employées pour désaisonnaliser une série temporelle sont la **régression linéaire** et les filtres **moyennes mobiles** qu'on retrouve dans des algorithmes tels que Tramo-Seats et la famille des Xnum (dont X11 que nous verrons ensuite) ou Xnum-ARIMA.

## TP : Désaisonnalisez à l'aide de la régression linéaire

On souhaite désaisonnaliser la série temporelle airpass à l'aide de la régression linéaire.

On créé à cet effet les bases tendancielle et saisonnière :

t=1:144

for (i in 1:12)

{

su=rep(0,times=12)

su[i]=1

s=rep(su,times=12)

assign(paste("s",i,sep=""),s)

}

On effectue la régression linéaire (le modèle est transformé, comme vu en cours, afin de pallier le problème de colinéarité) sur la série **Yt** :

reg=lm(y~t+s1+s2+s3+s4+s5+s6+s7+s8+s9+s10s11+s12-1)

summary(reg)

On obtient l'estimation des différents paramètres :

##

## Call:

## lm(formula = y ~ t + s1 + s2 + s3 + s4 + s5 + s6 + s7 + s8 +

## s9 + s10 + s11 + s12 - 1)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.156370 -0.041016 0.003677 0.044069 0.132324

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## t 0.0100688 0.0001193 84.4 <2e-16 \*\*\*

## s1 4.7267804 0.0188935 250.2 <2e-16 \*\*\*

## s2 4.7047255 0.0189443 248.3 <2e-16 \*\*\*

## s3 4.8349527 0.0189957 254.5 <2e-16 \*\*\*

## s4 4.8036838 0.0190477 252.2 <2e-16 \*\*\*

## s5 4.8013112 0.0191003 251.4 <2e-16 \*\*\*

## s6 4.9234574 0.0191535 257.1 <2e-16 \*\*\*

## s7 5.0273997 0.0192073 261.7 <2e-16 \*\*\*

## s8 5.0181049 0.0192617 260.5 <2e-16 \*\*\*

## s9 4.8734703 0.0193167 252.3 <2e-16 \*\*\*

## s10 4.7353120 0.0193722 244.4 <2e-16 \*\*\*

## s11 4.5915943 0.0194283 236.3 <2e-16 \*\*\*

## s12 4.7054593 0.0194850 241.5 <2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.0593 on 131 degrees of freedom

## Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999

## F-statistic: 9.734e+04 on 13 and 131 DF, p-value: < 2.2e-16

Les différents coefficients sont contenus dans `reg$coefficients` :

reg$coefficients

## t s1 s2 s3 s4 s5 s6

## 0.0100688 4.7267804 4.7047255 4.8349527 4.8036838 4.8013112 4.9234574

## s7 s8 s9 s10 s11 s12

## 5.0273997 5.0181049 4.8734703 4.7353120 4.5915943 4.7054593

On revient aux coefficients initiaux :

a=mean(reg$coefficients[2:13])

b=reg$coefficients[1]

c=reg$coefficients[2:13]-mean(reg$coefficients[2:13])

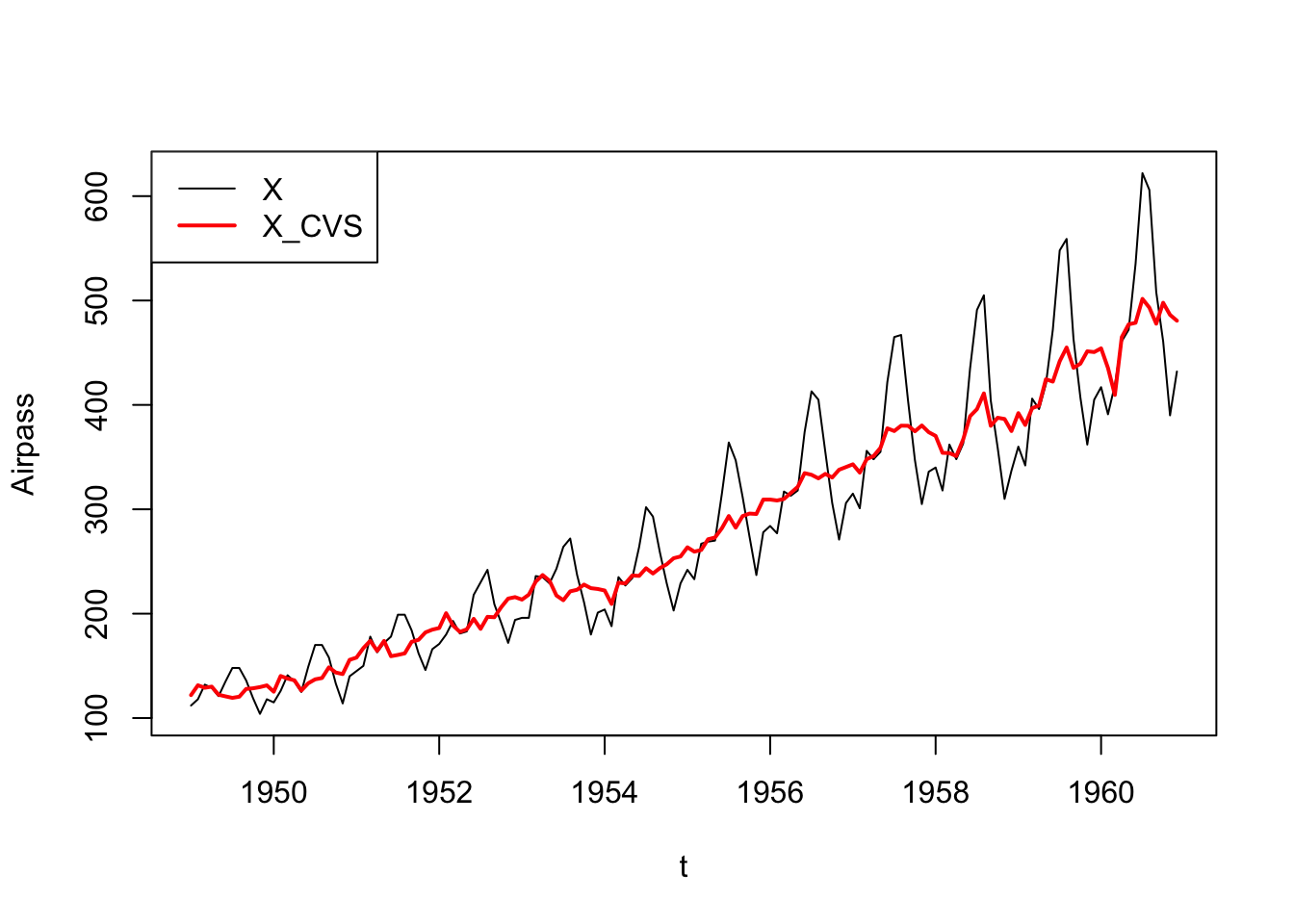
et on obtient la série corrigée des variations saisonnières (en n'oubliant pas de passer à l'exponentiel pour revenir à **Xt** ) :

y\_cvs=y-(c[1]\*s1+c[2]\*s2+c[3]\*s3+c[4]\*s4+c[5]\*s5+c[6]\*s6+c[7]\*s7+c[8]\*s8+c[9]\*s9+c[10]\*s10+c[11]\*s11+c[12]\*s12)

x\_cvs=exp(y\_cvs)

ts.plot(x,x\_cvs,xlab="t",ylab="Airpass",col=c(1,2),lwd=c(1,2))

legend("topleft",legend=c("X","X\_CVS"),col=c(1,2),lwd=c(1,2))



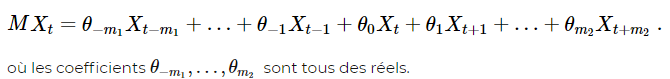
On constate que la série CVS présente des baisses en 1958 et 1960, pas forcément perceptibles à l'oeil nu sur la série brute. On pourra vérifier que la méthode des moyennes mobiles mettra en évidence les mêmes ruptures.

## Désaisonnalisez à l'aide des moyennes mobiles

Si la régression linéaire est une solution "acceptable", on trouve beaucoup plus couramment des méthodes basées sur ce qu'on nomme les moyennes mobiles.

### ****Une définition****

On utilise des filtres afin d'effectuer ces décompositions. De manière peu rigoureuse, un filtre est un "transformateur" d'un signal (temporel). On n'utilise ici que des filtres linéaires, et plus particulièrement les moyennes mobiles.  
  
Une **moyenne mobile M** est une combinaison linéaire d'instants passés et futurs de la série temporelle :



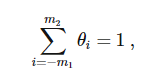
Notons d'ores et déjà que le produit de deux moyennes mobiles est commutatif, associatif et distributif par rapport à l'addition. Cette propriété nous indique notamment que si l'on doit appliquer plusieurs moyennes mobiles sur une série temporelle, l'ordre d'application est indifférent.

### ****Un peu de terminologie****

ll existe de nombreuses moyennes mobiles et il nous faudra les caractériser, comprendre leurs propriétés.

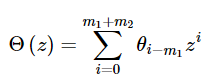
Dans la littérature, on trouve (et manipule) des moyennes mobiles qui peuvent être :

* normalisées (il s'agit alors d'une moyenne au sens où on l'entend habituellement) si :



* centrées (on prend en compte autant d'instants passés que futurs) si **m1=m2=m**.
* symétriques (on donne des poids similaires aux instants passés et futurs de même ordre) si  

On désigne par :



le **polynôme caractéristique** de la moyenne mobile (comme son nom l'indique, il permet de caractériser ses propriétés).

### ****Deux moyennes mobiles classiques****

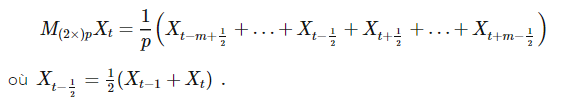
Voici deux des moyennes mobiles les plus employées (on reviendra sur les motivations plus tard).

* Pour **p=2m+1** , on définit la **moyenne mobile simple** (arithmétique) :



C'est une moyenne mobile symétrique et normalisée.

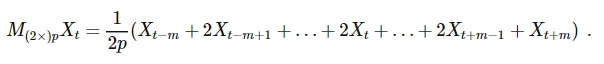
* Pour **p=2m** , on définit la **moyenne mobile centrée** (souvent on la nomme également, abusivement, moyenne mobile simple) :



On peut réécrire :

M(2×)pXt=12p(Xt−m+2Xt−m+1+…+2Xt+…+2Xt+m−1+Xt+m) .M(2×)pXt=12p(Xt−m+2Xt−m+1+…+2Xt+…+2Xt+m−1+Xt+m) .

C'est une moyenne mobile centrée et normalisée.



### ****Utiliser des moyennes mobiles pour corriger des variations saisonnières****

L'idée consiste à appliquer une ou plusieurs moyennes mobiles afin de mettre en évidence, estimer, les différentes composantes d'une série temporelle.  
Supposons que :

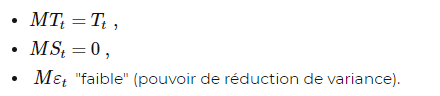


Si on applique une moyenne mobile  sur la série, nous obtenons (grâce à la propriété d'associativité) :



La visualisation de la série nous permet d'émettre des hypothèses sur la tendance (linéaire, quadratique, etc.) et la saisonnalité (période, forme, etc.).

L'enjeu est de trouver une moyenne mobile qui laisse la tendance invariante, qui absorbe la saisonnalité et qui réduit le résidu :



Dans ces conditions   constitue une estimation de la tendance. La saisonnalité peut ensuite être estimée en travaillant sur la différence entre la série et la tendance ainsi estimée.

### ****Evaluer les propriétés d'une moyenne mobile****

On constate donc qu'il nous faut connaitre les propriétés des moyennes mobiles afin de savoir quelles séries elles peuvent laisser invariantes ou absorber.

#### Séries temporelles absorbées et invariantes

* Une série temporelle est **absorbée** (ou annulée) par si :



* Une série temporelle est **invariante** par si :



Notons qu'une série temporelle invariante par  est absorbée par , où est l'opérateur identité.

L'étude des propriétés des moyennes mobiles repose sur le polynôme caractéristique, on est alors amenés à résoudre une équation de récurrence linéaire.

Il existe de nombreux résultats sur les propriétés, citons notamment :

* absorbe une composante saisonnière de période **s**  s'annulant sur une période si et seulement si est divisible .
* Les constantes sont invariantes par  si et seulement si est normalisée.

Revenons à nos deux moyennes mobiles phares :

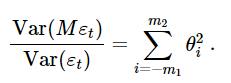
* absorbe les saisonnalités de période et laisse invariantes les tendances linéaires.
* absorbe les saisonnalités de période   et laisse invariantes les tendances linéaires.

Cela pourra donc s'avérer fort utile...

#### Pouvoir de réduction de variance

Une dernière propriété est souvent attendue, celle de réduction de la variance. L'idée est que l'application d'une moyenne mobile sur une série temporelle puisse réduire sa variance, le lisser.

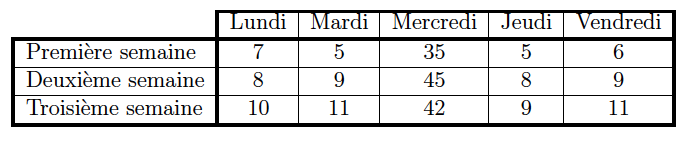
Mathématiquement, ce pouvoir se quantifie en étudiant le rapport de variances entre un bruit blanc fort , c'est-à-dire une suite de variables aléatoires i.i.d, en entrée, et sa résultante en sortie de la moyenne mobile  :



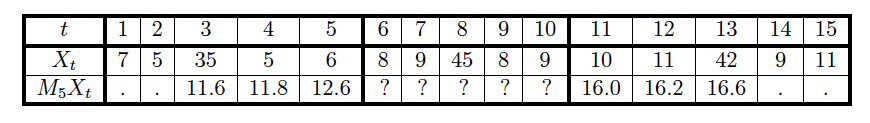
## Découvrez des algorithmes de traitement des moyennes mobiles

### Commençons par un algorithme simple

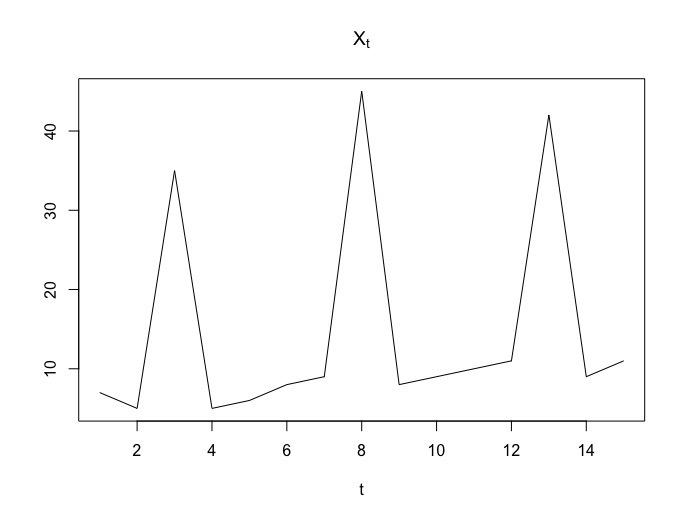
On a comptabilisé pendant trois semaines consécutives le nombre journalier de visiteurs d’un musée dont les jours de fermeture sont le samedi et le dimanche.



Ces données sont encore représentées dans le tableau ci-dessous (avec en sus la série lissée par moyenne mobile).

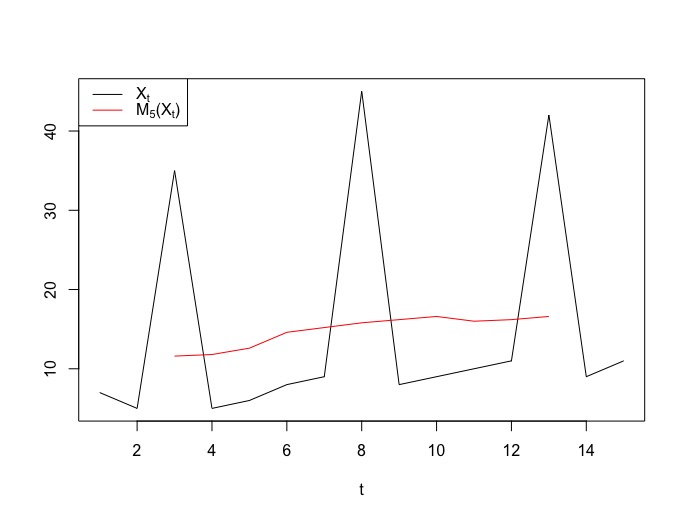


Le graphe semble mettre en évidence une tendance linéaire ainsi qu’une saisonnalité de période 5 :

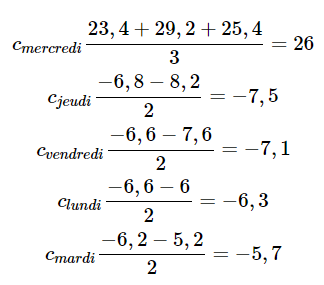
Moyenne mobile

On choisit un modèle additif car les différences (maximum-minimum) semblent similaires sur chaque période (semaine). Les termes tendanciels pour **t**∈**{3,…,13}** figurent dans le tableau ci-après.

On utilise la moyenne mobile  afin d’estimer la tendance (supposée linéaire) de la série temporelle. Pour rappel cette moyenne mobile conserve les tendances linéaires et absorbe les saisonnalités de période 5.



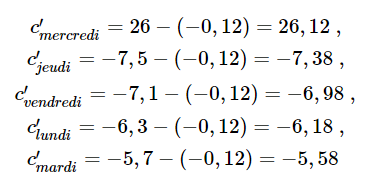
On calcule les coefficients saisonniers comme suit :



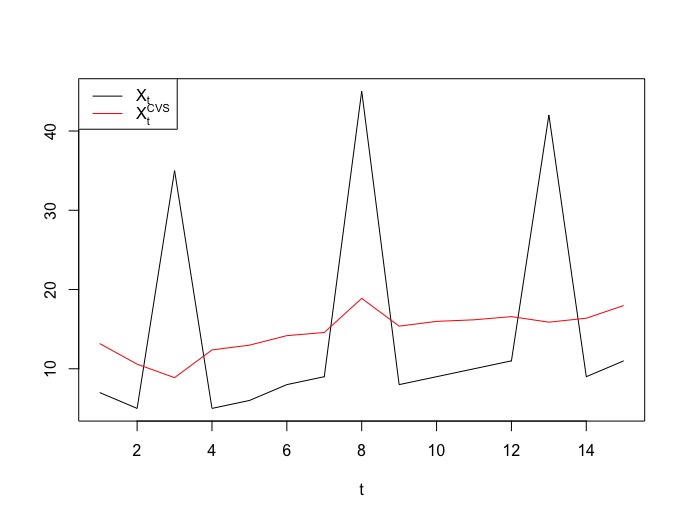
On constate que :



Les coefficients saisonniers centrés sont :



La série corrigée des variations saisonnières, est ainsi représentée :

Série corrigée des variations saisonnières

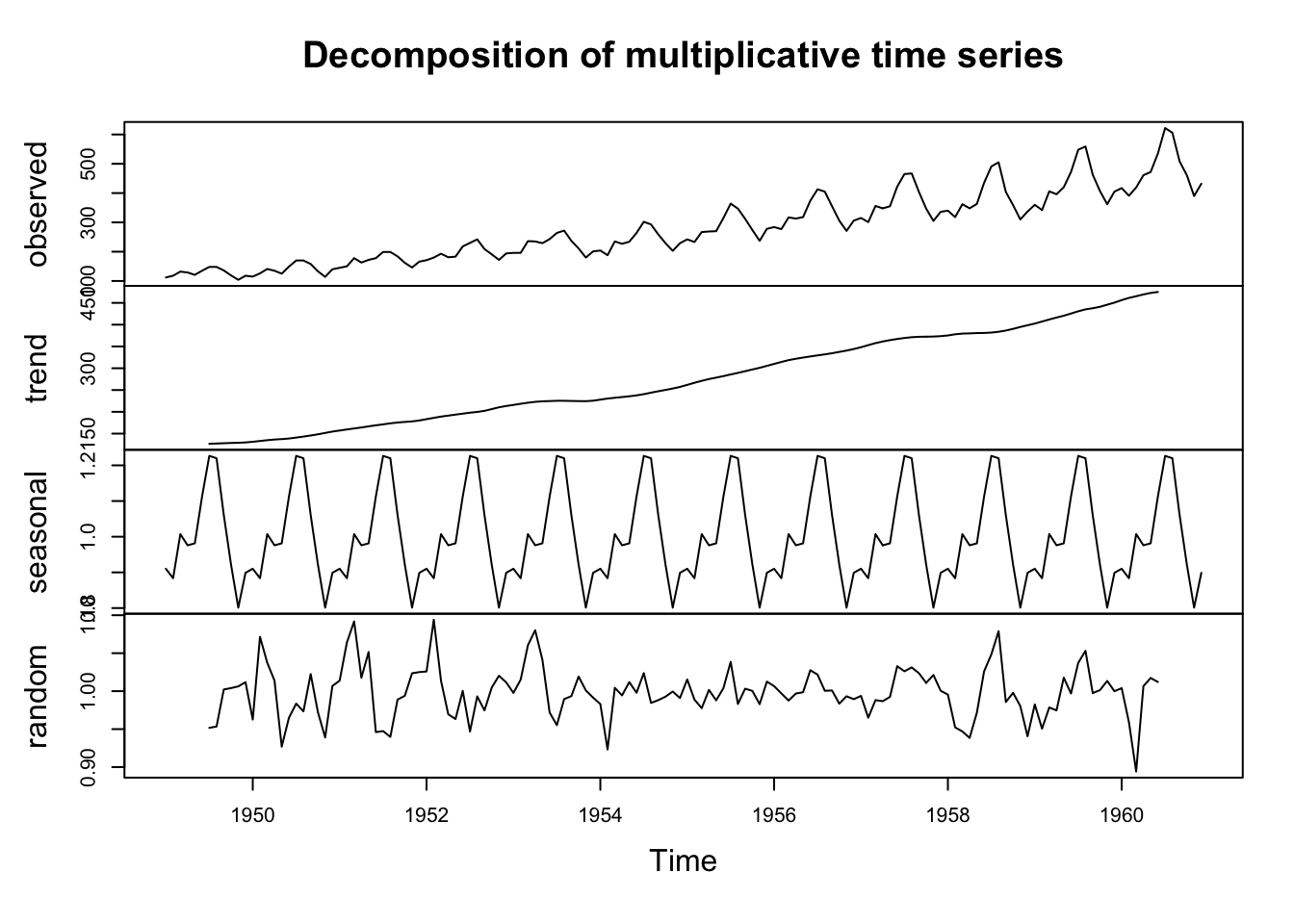
Retrouvez cette série sous forme de tableau à [**ce lien**](https://drive.google.com/open?id=1JMWXeipo2vV6vVtjBEoV92bC-sT20g2Q).

### TP : Désaisonnaliser une série temporelle à l’aide de la fonction decompose

On utilise la fonction `decompose` :

decomp.x=decompose(x,type="multiplicative")

plot(decomp.x)



### Pour aller plus loin : l’algorithme X11

Il existe aujourd’hui de nombreux algorithmes élaborés pour corriger les séries tempo-  
relles des variations saisonnières. Nous n’entrerons pas dans le détail de ces méthodes mais  
présentons ici un de ces représentants les plus simples, l’algorithme X11 développé par le  
Census Bureau. Il s’agit d’un algorithme qui procède à une estimation de la série corrigée  
des variations saisonnières en 2 phases de 4 étapes chacune.

Considérons une série temporelle répondant au modèle additif classique, avec une tendance  
linéaire et une saisonnalité de période 12 (il existe plusieurs versions de cet algorithme :  
pour les modèles additifs et multiplicatifs, pour des séries trimestrielles ou mensuelles, avec  
une saisonnalité annuelle) :

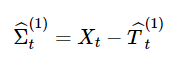


**1. Primo-estimation de la tendance**



(sur le même principe que la méthode simplifiée vue auparavant).

**2. Primo-estimation de la somme composante saisonnière-perturbation**  
En notant :



(on retranche simplement la primo-estimation de la tendance à la série de départ).

**3. Primo-estimation de la composante saisonnière**



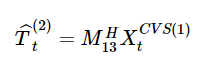
(les coefficients sont ainsi normalisés afin que leur somme soit nulle sur une période  
de 12 instants consécutifs).

**4. Primo-estimation de la série corrigée des variations saisonnières**

****

(on retranche la saisonnalité estimée à la série de départ).

**5. Estimation définitive de la tendance**



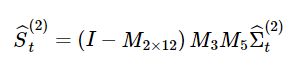
(on utilise une moyenne mobile, nommée de Henderson d’ordre FORMULE\_MATH afin d’extraire une seconde tendance à partir de la première série corrigée des variations saisonnières).

**6. Estimation définitive de la somme composante saisonnière-perturbation (2)**



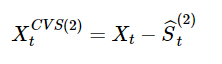
(on retrouve ici la même procédure que précédemment).

**7. Estimation définitive de la composante saisonnière**

****

(on retrouve ici un lissage et une normalisation de la composante saisonnière).

**8. Estimation définitive de la série corrigée des variations saisonnières (2)**



(on retranche la seconde saisonnalité estimée à la série de départ, cette étape clôt  
l’algorithme).

# Partie 2

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Corriger une série temporelle des variations saisonnières
* Comprendre la problématique posée par le lien temporel

### Question 1

**Une moyenne mobile est :**

* + 

un filtre non linéaire,

* + 

un filtre linéaire, pas forcément centré ou symétrique

* + 

un filtre linéaire centré, pas forcément symétrique

* + 

un filtre linéaire normalisé

*Les moyennes mobiles utilisées en pratique sont généralement centrées voire symétriques, mais elles ne le sont pas forcément toutes.*

### Question 2

**La décomposition d'une série temporelle est essentiellement utilisée pour :**

* + 

mettre en évidence la tendance d'une série

* + 

retrancher à une série sa tendance

* + 

retrancher à une série sa saisonnalité

* + 

"stationnariser" une série

### Question 3

**Idéalement, la première moyenne mobile utilisée dans la correction des variations saisonnières d'une série temporelle doit :**

*Attention, plusieurs réponses sont possibles.*

* + 

conserver la tendance

* + 

absorber la tendance

* + 

conserver la saisonnalité

* + 

réduire la variance du bruit

*Elle doit en sus absorber la saisonnalité.*

### Question 4

**La moyenne mobile**M(2×)4M(2×)4**:**

* + 

absorbe les tendances linéaires et conserve les saisonnalités de période 4

* + 

absorbe les tendances linéaires et conserve les saisonnalités de période 8

* + 

conserve les tendances linéaires et absorbe les saisonnalités de période 4

* + 

conserve les tendances linéaires et absorbe les saisonnalités de période 8

### Question 5

**L'algorithme X11 :**

* + 

contient 2 boucles pour évaluer la série CVS

* + 

est l'algorithme le plus sophistiqué connu à l'heure actuelle

* + 

intègre de manière conjointe régression linéaire et moyenne mobile

## Appréhendez le lissage exponentiel simple

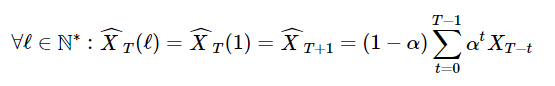
Les méthodes de lissage exponentiel, basées sur les travaux de Holt et Winters, ne font plus l'objet de recherches. Si elles peuvent apparaître "faibles" théoriquement, elles sont faciles à mettre en œuvre, d'où leur utilisation populaire dans le passé. Leur objectif est de prévoir à la date **T** une série temporelle à un horizon **ℓ** à partir de T observations **(X1,…,XT)**.

### Méthode

Le **lissage exponentiel simple** (LES) consiste à supposer que  est approximable autour de **T** par une constante .

Soit  la prévision de à l'instant **T**.

La prévision par la méthode du LES est la suivante :



où **α]0,1[** désigne le coefficient de lissage.

On peut remarquer que :

* Les réalisations ont d'autant plus de poids qu'elles sont récentes (en effet le poids  devant décroît avec **t**).

La prévision ne dépend pas de l'horizon **ℓ**, ce qui peut laisser pantois en l'état.

### Formule de mise à jour

Il n'est pas obligatoire de recalculer entièrement la prévision à chaque instant.

On a en effet :

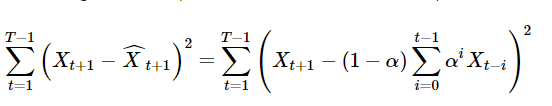


D'où :



On peut ainsi calculer la lissée  à n'importe quel instant **t**en fonction de celle à l'instant **t-1**.

Choix du coefficient de lissage  
On utilise un algorithme d'optimisation afin de minimiser la quantité suivante :



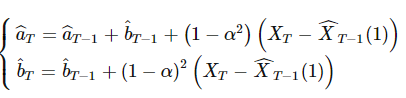
Cette quantité est la somme des écarts quadratiques entre les réalisations et les prévisions de la série temporelle.

**Appréhendez le lissage exponentiel double et la méthode de Holt-Winters**

**1. lissage exponentiel double**

Le **lissage** **exponentiel** **double** (LED) consiste lui à supposer que  est approximable au voisinage de **T** par une droite : .

Les formules de mise à jour sont :



Le paramètre de la méthode du LED est **α]0,1[** .

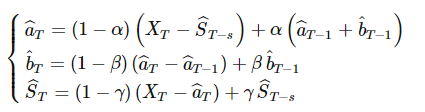
La prévision par la méthode du LED est la suivante :



**2. méthode de Holt-Winters**

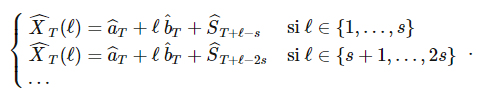
La **méthode** **de** **Holt-Winters, parfois encore utilisée en pratique**, consiste à supposer que est approximable au voisinage de T par .

En désignant par **s** la période du cycle saisonnier de la série temporelle, les formules de mise à jour sont :



Les paramètres de la méthode de Holt-Winters sont α, β et γ  (tous dans ]0,1[ ).

La prévision par la méthode de Holt-Winters est la suivante :



Quelle que soit la méthode employée, on constate qu'une prévision n'est qu'une fonction du passé de la série (linéaire dans les cas traités dans ce cours). Selon la méthode retenue, cette fonction diffère, mais l'idée est toujours la même !

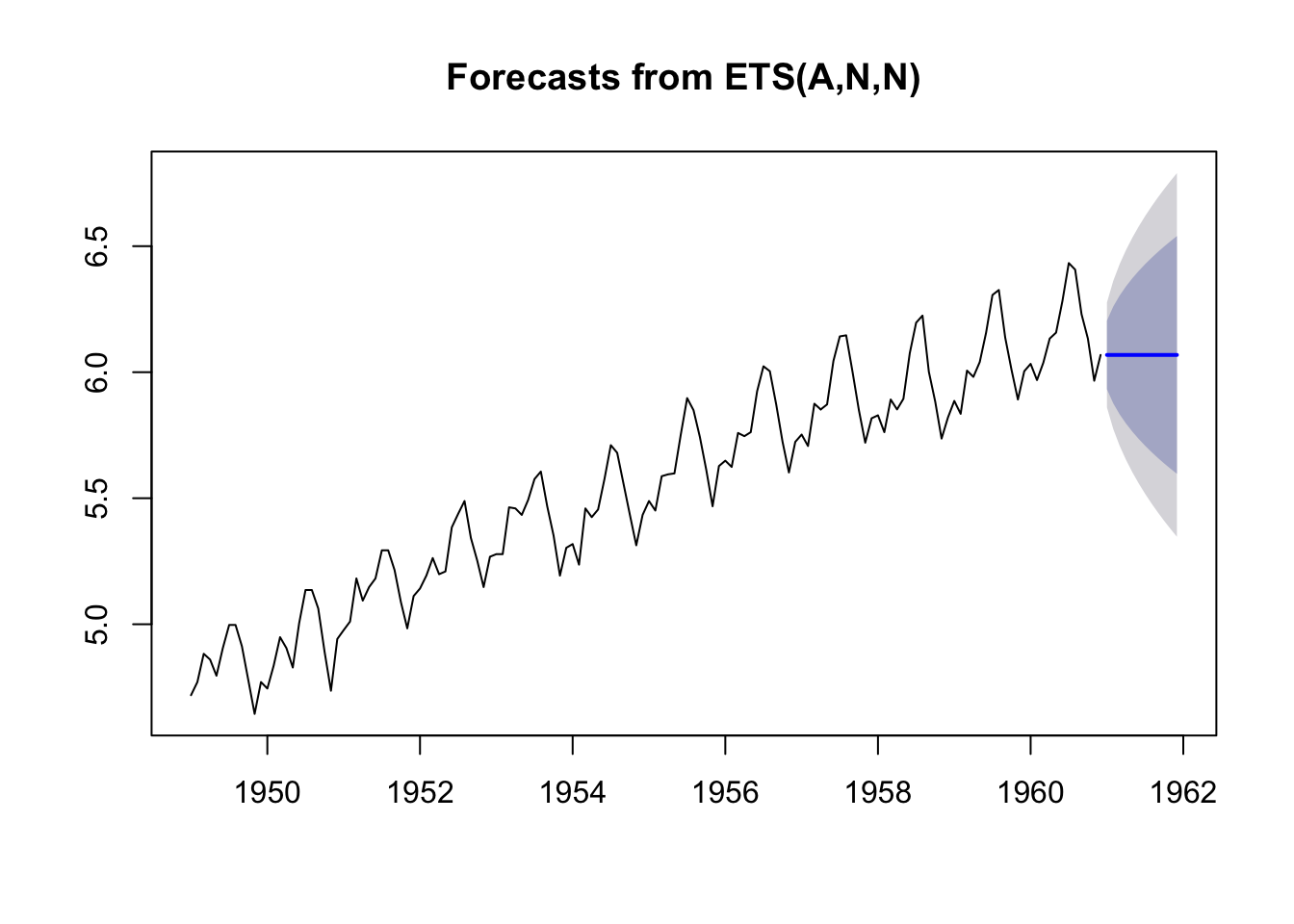
## TP : Prévoyez une série à l’aide des méthodes de lissage exponentiel

Si on souhaite prévoir la série airpass à l'aide du lissage exponentiel simple, on peut utiliser les commandes suivantes :

les=ets(y,model="ANN")

les.pred=predict(les,12)

plot(les.pred)

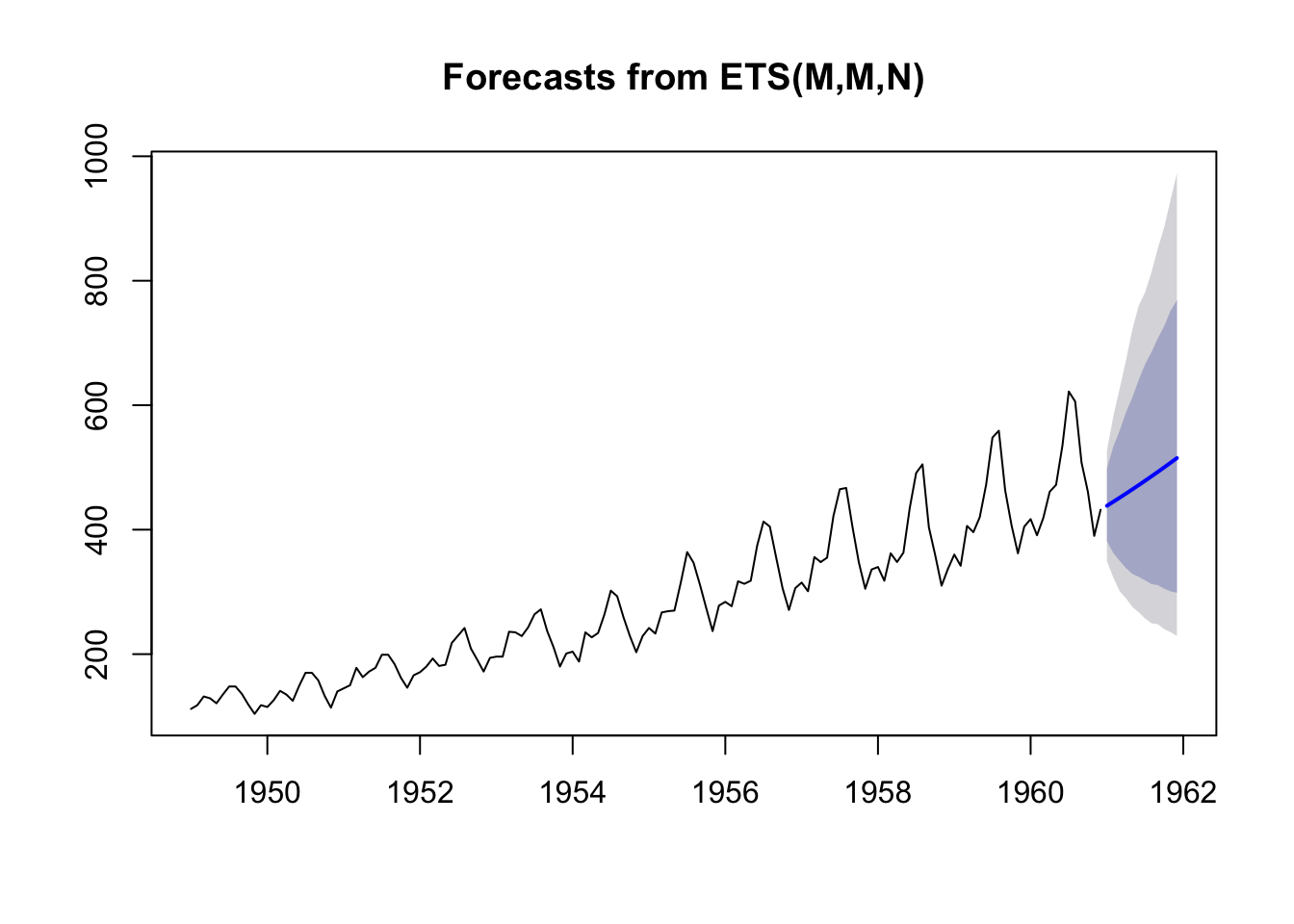


Pour le lissage exponentiel double :

led=ets(x,model="MMN")

led.pred=predict(led,12)

plot(led.pred)

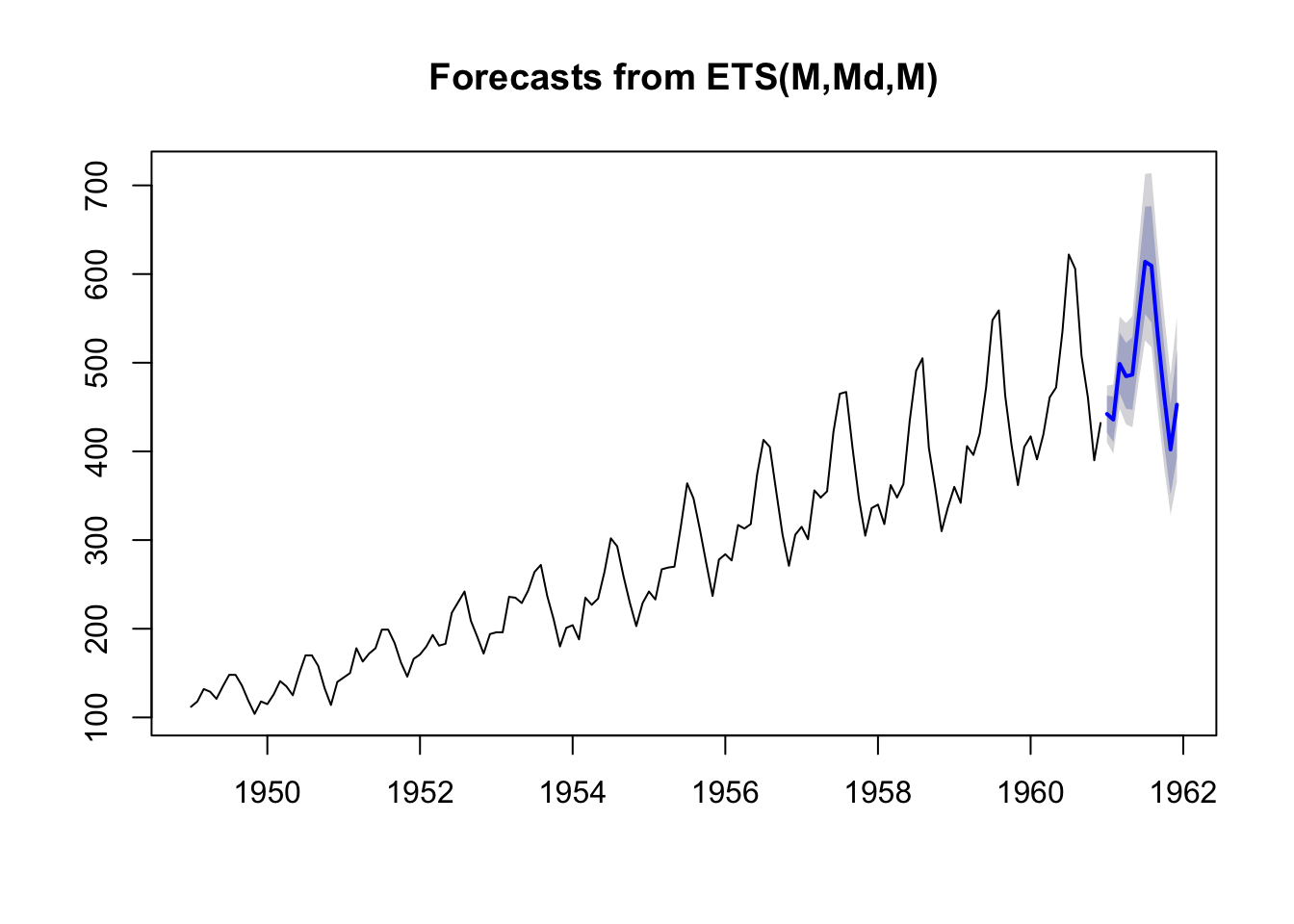


Et enfin pour la méthode de Holt-Winters :

hw=ets(x,model="MMM")

hw.pred=predict(hw,12)

plot(hw.pred)



On constate que seule la dernière prévision est raisonnable, elle intègre la saisonnalité.

Bas du formulaire

# Partie 3

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Mettre en œuvre des techniques simples pour prévoir une série temporelle
* Comprendre la problématique posée par le lien temporel

### Question 1

**La prévision obtenue par lissage exponentiel simple...**

* + 

tient compte de la tendance

* + 

ne tient pas compte de la saisonalité

* + 

varie en fonction de l'horizon de prévision

*La prévision est constante et ne prend pas en compte d'une éventuelle tendance ou saisonnalité.*

### Question 2

**La méthode de Holt-Winters...**

*Attention, plusieurs réponses sont possibles.*

* + 

est plus efficace que les lissages exponentiels simple ou double

* + 

tient compte d'une éventuelle saisonnalité mais pas de tendance

* + 

tient compte d'une éventuelle tendance et saisonnalité

## Découvrez les processus stationnaires

### Série temporelle et processus stochastique

On considère ici un ensemble d'observations dans  enregistrées à un temps spécifique , on parle alors de série temporelle univariée à temps discret.

On considère en Statistique que l'observation **x** est la réalisation d'une variable aléatoire **X**.

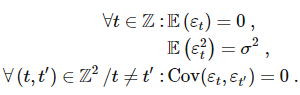
De manière analogue, une série temporelle sera considérée comme la réalisation d'un processus stochastique  .

Ce qu'il faut retenir à ce stade, c'est qu'on parle d'un processus stochastique si, pour tout fixé, est une variable aléatoire réelle.

### La notion de bruit blanc

A l'issue d'une modélisation, il nous faudrait idéalement obtenir un signal résiduel qui ne contient plus d'information temporelle. Dans le cadre des modèles ARMA, on souhaite que le résidu soit un bruit blanc (faible), c'est-à-dire sans dépendance temporelle linéaire.

 est un bruit blanc faible s'il est constitué de v.a.r telles que :

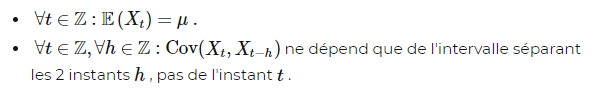


Dans la suite, on désignera par un bruit blanc (faible).

### Une notion fondamentale : la stationnarité

Dans de très nombreux cas, on ne peut pas renouveler la suite de mesures dans des conditions identiques. Pour que le modèle déduit à partir d'une suite d'observations ait un sens, il faut que toute portion de la trajectoire observée fournisse des informations sur la loi du processus et que des portions différentes, mais de même longueur, fournissent les mêmes indications. D'où la notion de stationnarité.

Un processus  est (faiblement) stationnaire si son espérance et ses autocovariances sont invariantes par translation dans le temps :



Un processus stationnaire n'est pas obligatoirement maîtrisable ou "sympathique". On sait par exemple que la somme de deux processus stationnaires n'est pas forcément stationnaire.

### Modéliser par des processus stationnaires

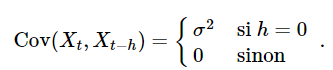
Une série présentant une tendance et/ou une saisonnalité (elle sont nombreuses dans le quotidien du data analyst !) ne pourra pas être modélisée par un processus stationnaire ; une technique communément employée est de travailler non pas sur la série mais sur des différences de la série comme l'illustre l'exemple suivant.  
Considérons le processus  vérifiant :



avec **a≠0**.  
  
On a :



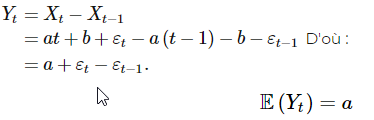
et :



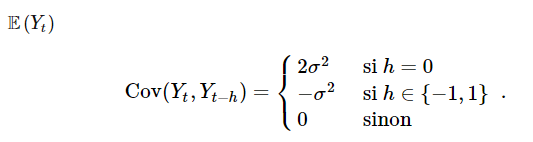
Contrairement à  dépend de **t**.  
Le processus n'est donc pas stationnaire.  
  
Considérons maintenant le processus "différence première" :



On a :



et :



 ne dépendent pas de **t** donc  est bien un processus stationnaire.

On a donc transformé un processus non stationnaire  par différenciation pour obtenir un processus stationnaire .

### Mesurer la dépendance temporelle linéaire

#### La fonction d'autocovariance

Les modèles ARMA sont linéaires, afin de définir leur structure nous utiliserons les autocorrélogrammes simples et partiels basés sur la fonction d'autocovariance.  
On appelle fonction d'autocovariance d'un processus stationnaire **X** la fonction **γ** suivante :



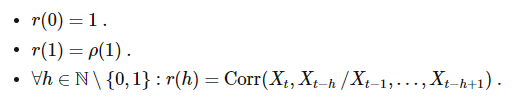
**γ** est une fonction symétrique :



On pourra donc calculer la fonction d'autocovariance pour .

#### L'autocorrélogramme simple

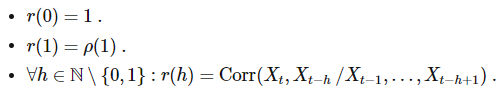
On appelle autocorrélogramme simple d'un processus  stationnaire la fonction **ρ** suivante :



Il s'agit d'une simple normalisation de la fonction d'autocovariance.

#### L'autocorrélogramme partiel

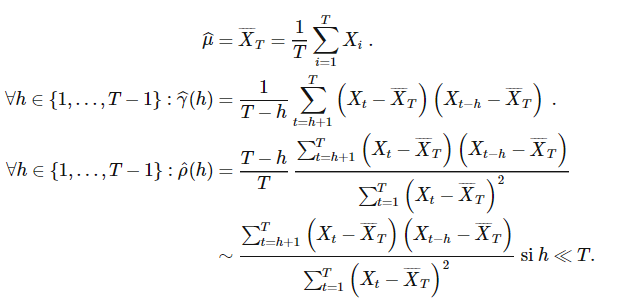
On constate que la corrélation existante entre 2 instants successifs d'un processus stationnaire se "diffuse". Il s'avère intéressant de connaître la corrélation entre 2 instants éloignés d'un processus conditionnellement aux instants intermédiaires, il s'agit là de **l'autocorrélogramme partiel** noté **r**  :



Afin de les déterminer, on calcule au préalable les autocorrélations simples et on les déduit à l'aide de l'algorithme de Durbin-Levinson (non abordé dans ce cours).

### Estimer les moments d'un processus stationnaire

Soit  un processus stationnaire.  
A partir de **(X1,…,XT)**, on peut considérer les estimateurs suivants :



Remarquons que :

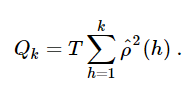
* Estimer  pour des valeurs élevées de **h** peut devenir délicat au vu du nombre d'observations en jeu. En pratique on n'excède pas un quart de la taille de la série temporelle.
* On peut mener les calculs même lorsque le processus n'est pas stationnaire ! Les logiciels le font par défaut, le data analyst devra évaluer la pertinence du calcul.
* Les estimations des autocorrélations partielles se déduisent des estimations des autocorrélations simples grâce à l'algorithme de Durbin-Levinson.

### Tester la blancheur d'un résidu

A l'issue d'une modélisation ARMA, il faudra tester la blancheur du résidu à partir de l'autocorrélogramme simple.  
Soit  un processus stationnaire.

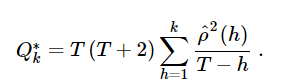
Considérons le test suivant :  


Si on dispose de **(X1,…,XT)**, on considère la statistique de Portmanteau (calculée sur les **k** premières estimations des autocorrélations) :

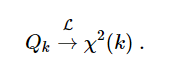


Une trop grande valeur de indique que les autocorrélations sont trop importantes pour être celles d'un bruit blanc (en effet ces autocorrélations sont théoriquement nulles pour un bruit blanc).

Il existe d'autres versions de cette statistique, par exemple la statistique de Ljung–Box :



Sous  (ainsi que ) suit asymptotiquement une loi du Khi-deux à **k** degrés de liberté :



On rejette donc l'hypothèse **H0** au niveau de test **α** si :



où désigne le quantile d'ordre **1−α**  d'une loi du Khi-deux à **k** degrés de liberté.

La p-valeur vaut :



### Vers les processus ARMA

Il existe un résultat théorique, nommé décomposition de Wold, qui nous indique que tout processus stationnaire peut être modélisé par un processus ARMA, ce résultat est remarquable car il rend universel le modèle linéaire sur des processus stationnaires.

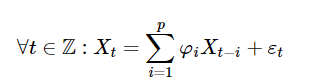
## Les processus AR, MA et ARMA

### Les processus AR

Sauf mention contraire, on considère dans la suite des processus centrés. Si le processus n'est pas centré, on peut se ramener à ce cas de figure en centrant préalablement le processus (on travaillerait alors sur .

#### Définition

On dit qu'un processus  (stationnaire) est un processus processus AR (AutoRegressive) d'ordre **p**, noté **AR(p)**, si :

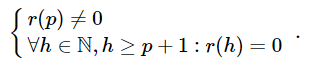


où 

La modélisation de  se résume à une relation linéaire le liant aux **p**  derniers instants.

#### Caractérisation

Si est un processus **AR(p)** alors ses autocorrélations partielles s'annulent à partir du rang **p+1** :

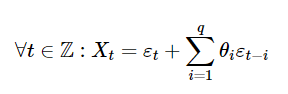


Réciproquement, il s'agit d'une condition nécessaire et suffisante pour qu'un processus  soit un **AR(p)**.  
Les autocorrélations simples, quant à elles, décroissent rapidement vers 0 (de manière exponentielle ou sinusoïdale amortie).

### Les processus MA

#### Définition

On dit qu'un processus est un processus MA (Moving Average) d'ordre **q**, noté **MA(q)**, si :



où .

On considère ici que le processus est la résultante d'une combinaison linéaire de perturbations décorrélées (un bruit blanc et son passé).

#### Caractérisation

Si ****est un processus **MA(q)** alors ses autocorrélations simples s'annulent à partir du rang **q+1** :



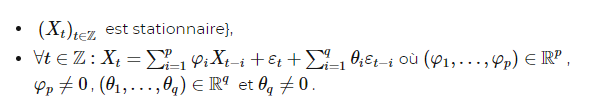
Il s'agit d'une condition nécessaire et suffisante pour qu'un processus ****soit un **MA(q)**.

### Les processus ARMA

#### Définition

Soit un bruit blanc faible de variance .

On dit qu'un processus ****est un processus ARMA (AutoRegressive Moving Average) d'ordre **(p,q)**, noté **ARMA(p,q)**, si :



Il s'agit ici d'une "synthèse" des processus AR et MA.

#### Caractérisation

Les autocorrélations simples décroissent vers 0. Il n'existe pas malheureusement de caractérisation simple, on verra comment procéder en pratique plus tard.

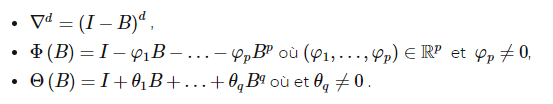
## Les processus non stationnaires : ARIMA et SARIMA

### Les processus ARIMA

On dit qu'un processus est un processus ARIMA (AutoRegressive Integrated Moving Average) d'ordre **(p,d,q)**, noté **ARIMA(p,d,q)**, si :



où :



Notons que :

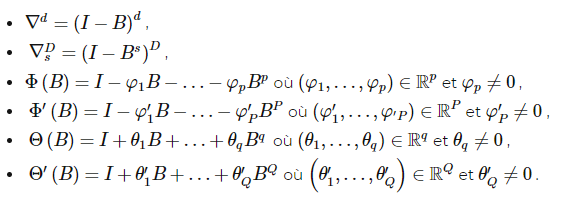
* Les modèles ARIMA permettent de modéliser des séries temporelles qui présentent une tendance polynomiale.
* est équivalent asymptotiquement à un processus **ARMA(p,q)**.
* Le processus  n'est pas stationnaire.

### Les processus SARIMA

On dit qu'un processus est un processus SARIMA (Seasonnal AutoRegressive Integrated Moving Average) d'ordre , noté   si :



où :



Notons que :

* Les modèles SARIMA permettent de modéliser des séries qui présentent une saisonnalité.
* Estimer un modèle SARIMA se ramène en pratique à l'estimation d'un modèle ARMA sur la série différenciée.

## Entraînez des modèles SARIMA

### La synthèse

La démarche adoptée est la suivante :

1. stationnarisation (éventuellement),
2. identification a priori de modèles potentiels,
3. estimation des modèles potentiels,
4. vérification des modèles potentiels,
5. choix définitif d'un modèle,
6. prévision à l'aide du modèle choisi,
7. analyse a posteriori de la prévision.

### Stationnarisation

La plupart des séries temporelles présentent une tendance et/ou une saisonnalité, et ne sont donc pas modélisables par un processus stationnaire.

Afin de se ramener à un processus ARMA, il faut stationnariser la série (en toute rigueur, on devrait parler de stationnariser un processus et non une série temporelle) et différentes méthodes sont envisageables :

#### **Décomposition saisonnière**

Cette méthode permet d'éliminer tendance et saisonnalité, sources évidentes de non-stationnarité. Il se peut néanmoins que la série résultant de la décomposition ne soit toujours pas stationnaire.

#### **Différenciation**

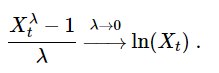
C'est la méthode employée par les modèles ARIMA et SARIMA. On ne modélise pas la série brute mais la série différenciée, en "tendance" à l'aide de , ou la série différenciée en "saisonnalité" à l'aide de . De manière générale,  permet de stationnariser des séries possédant une tendance polynomiale de degré **d**, et   des séries possédant une composante saisonnière de période ss .

Notons que si les processus stationnaires peuvent être approchés par des modèles ARMA, il n'est pas certain que la différenciation permette de stationnariser tous les processus.

#### **Méthode de Box-Cox**

Elle permet une stationnarisation en "variance" (ou encore de stationnariser des séries présentant une tendance exponentielle). On utilise la transformation avec  (il existe des méthodes alternatives si n'est pas une série positive).

Remarquons que :



#### **Méthode empirique**

Dans le cas des modèles SARIMA, on utilise très souvent une méthode empirique basée sur l'autocorrélogramme simple.

On effectue une différenciation en "tendance" si :

* Les autocorrélations sont proches de 1 pour un grand nombre de retards. Attention, il s'agit ici d'un abus de langage car si la sortie nommée "autocorrélations simples" nous permet de douter de la stationnarité du processus sous-jacent, nous ne devrions pas parler alors d'autocorrélations simples.
* Les premières autocorrélations sont proches les unes des autres (même si elles ne sont pas forcément proches de 1).

On parle souvent de décroissance lente des autocorrélations simples.

On effectue une différenciation en "saisonnalité" si des comportements similaires sont observés de manière périodique. Par exemple, si , ……  sont proches de 1, on utilise une différenciation en "saisonnalité" avec **s=12**.

On peut compléter cette méthode empirique par des tests de racines unité qui permettent de détecter des problèmes de stationnarité pour des séries qui ne présentent pas forcément de tendance et/ou de saisonnalité (non-stationnarité dite "stochastique".

* Notons que :  
  Quelle(s) que soi(en)t la(les) méthode(s) retenue(s), on procède de manière itérative : on effectue une première différenciation, si celle-ci d'avère insuffisante, on en effectue une seconde, etc.
* En pratique on considère rarement **d>2** et **D>2**.

### Identification a priori de modèles potentiels

Une fois la stationnarisation effectuée, on peut se consacrer aux choix potentiels des polynômes AR et MA.

Il existe différentes méthodes pour identifier un modèle **ARMA(p,q)** :

#### **Méthode de Box et Jenkins**

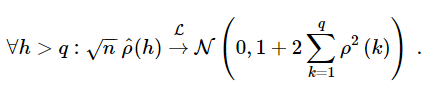
Il s'agit d'une méthode heuristique pour majorer **p** et **q**.

Pour un processus **AR(p)**, on peut montrer que :

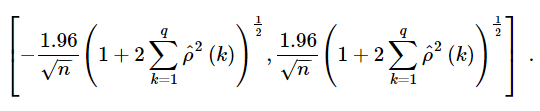


On peut définir un intervalle de confiance à 95% et on recherche à partir de quelle valeur, 95% des sont dans l'intervalle .

Pour un processus **MA(q)**, on peut montrer que :



On peut définir un intervalle de confiance à 95% et on recherche à partir de quelle valeur, 95% des sont dans l'intervalle :



#### **Critère d'information**

Les algorithmes de recherche automatique recherchent les ordres du modèle qui minimisent un critère d'information, avec plus ou moins de succès ……

#### **Heuristique**

En pratique (surtout pour les modèles SARIMA), on essaye d'identifier les autocorrélations simples et partielles "significatives" pour caler ensuite des polynômes AR et MA qui reflètent ces liens temporels.

Afin d'obtenir des modèles potentiels, l'idéal est de regarder l'autocorrélogramme partiel afin d'émettre une hypothèse sur la partie autorégressive (simple et saisonnière), la tester puis regarder l'autocorrélogramme simple (et partiel) du résidu afin d'identifier complètement un modèle. Cette démarche itérative permet en général d'obtenir plusieurs modèles potentiels.

### Estimation des modèles potentiels

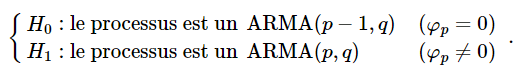
On estime les modèles potentiels à l'aide des méthodes classiques : maximum de vraisemblance (basée sur l'hypothèse gaussienne du résidu) ou moindres carrés. Il n'existe pas de solution explicite et il nous faut utiliser des algorithmes d'optimisation avec une estimation préliminaire (issue des équations dites de Yule-Walker ou de l'algorithme des innovations, non abordé dans ce cours).

### Vérification des modèles potentiels

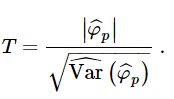
Afin de vérifier la validité des modèles estimés, on doit vérifier au minimum la significativité des paramètres et la blancheur du résidu.

#### **La significativité des paramètres**

Par exemple, pour le coefficient AR d'ordre **p**, on effectue le test suivant :



On utilise pour cela la statistique de test suivante :



Le test de Student permet de rejeter **H0** au niveau 5\% si **|t|** est strictement supérieur à 1.96.

Il existe des résultats similaires pour les coefficients MA.

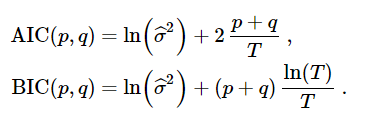
#### **La blancheur et la normalité du résidu**

On vérifie que le résidu est bien un bruit blanc, à l'aide du test de Ljung–Box par exemple, et qu'il est distribué suivant une loi normale.

Si le test de normalité n'est pas vérifié (à l'aide du test de Shapiro-Wilk par exemple), il est utile d'étudier le caractère ARCH ou GARCH du résidu ……

### Choix définitif d'un modèle

Ce choix s'opère entre les modèles potentiels retenus via :

Des critères d'information basés sur l'information de Kullback, par exemple, les critères d'Akaike (

Des critères basés sur le pouvoir prédictif.

Une fois ce choix effectué, le modèle retenu est utilisé à des fins de prévision.

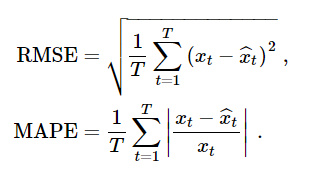
### Prévision à l'aide du modèle choisi

La fonction de prévision s'obtient assez facilement à partir des écritures dites autorégressive ou moyenne mobile.

### Analyse a posteriori

L'analyse a posteriori permet de quantifier les écarts entre les prévisions et les réalisations, en tronquant la série d'un certain nombre de points (notons que le modèle doit être correctement estimé sur la série tronquée.

On utilise des critères d'erreur comme l'erreur quadratique moyenne (Root Mean Square Error : RMSE ou l'erreur relative absolue moyenne (Mean Average Percentage Error : MAPE) :

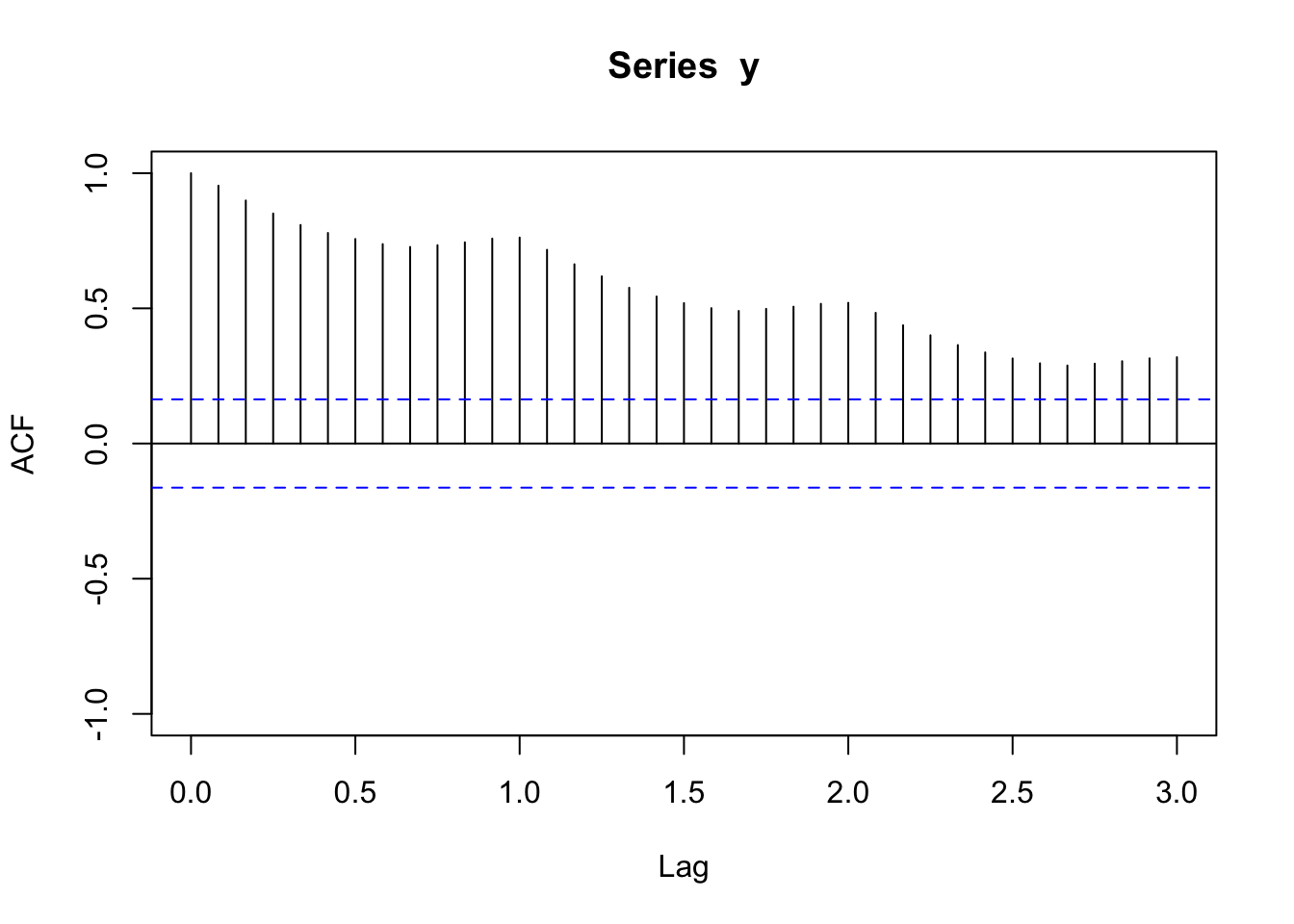


## TP : Prévoyez une série temporelle à l’aide des méthodes SARIMA

### Stationnarisation de la série

On désigne par **Xt** la série airpass, et on considère **Yt=ln(Xt)**. On travaille en effet sur le logarithme de la série afin de pallier l’accroissement de la saisonnalité. On passe ainsi d’un modèle multiplicatif à un modèle additif.

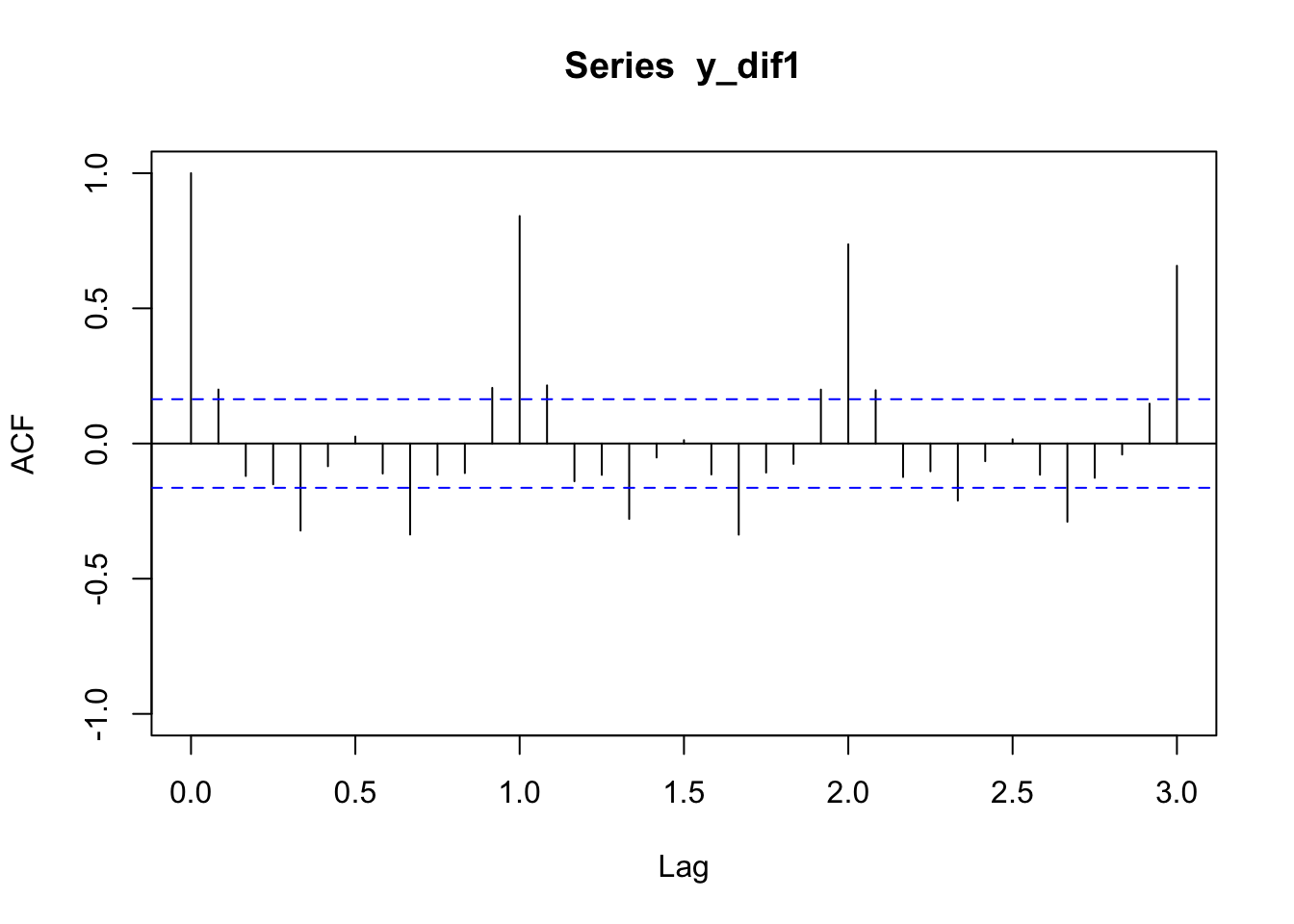
plot(acf(y,lag.max=36,plot=FALSE),ylim=c(-1,1))



La sortie ACF présente une décroissance lente vers 0, ce qui traduit un problème de non-stationnarité. On effectue donc une différenciation **(I−B)**.

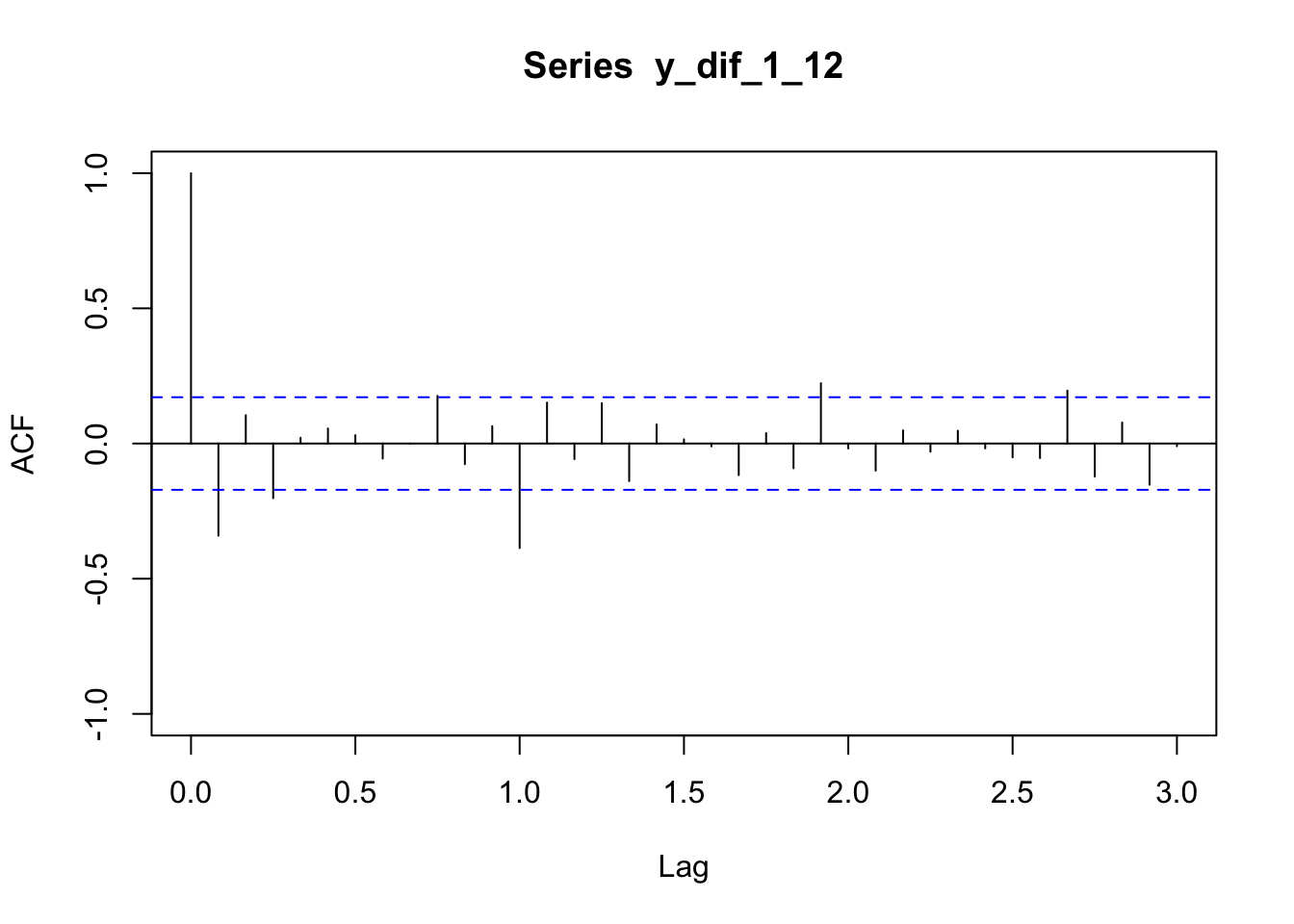
y\_dif1=diff(y,lag=1,differences=1)

plot(acf(y\_dif1,lag.max=36,plot=FALSE),ylim=c(-1,1))

 La sortie ACF de la série ainsi différenciée présente encore une décroissance lente vers 0 pour les multiples de 12. On effectue cette fois la différenciation **(I−B12)**.

y\_dif\_1\_12=diff(y\_dif1,lag=12,differences=1)

plot(acf(y\_dif\_1\_12,lag.max=36,plot=FALSE),ylim=c(-1,1))



La sortiE ACF de la série doublement différenciée semble pouvoir être interprétée comme un autocorrélogramme simple empirique.

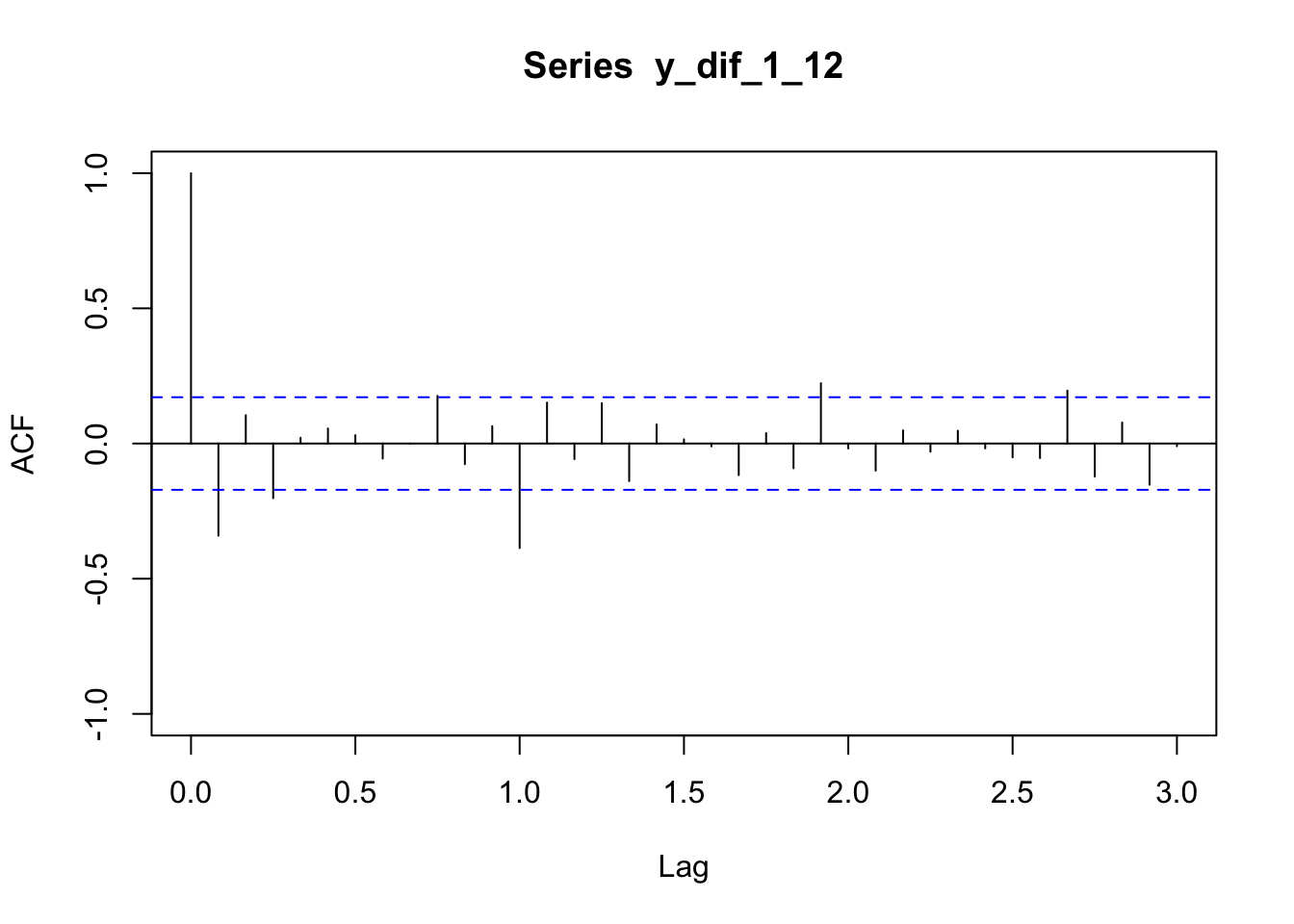
On identifiera donc un modèle ARMA sur la série  (I−B)(I−B12)ln(Xt) .(I−B)(I−B12)ln⁡(Xt) .

#### **Identification, estimation et validation de modèles**

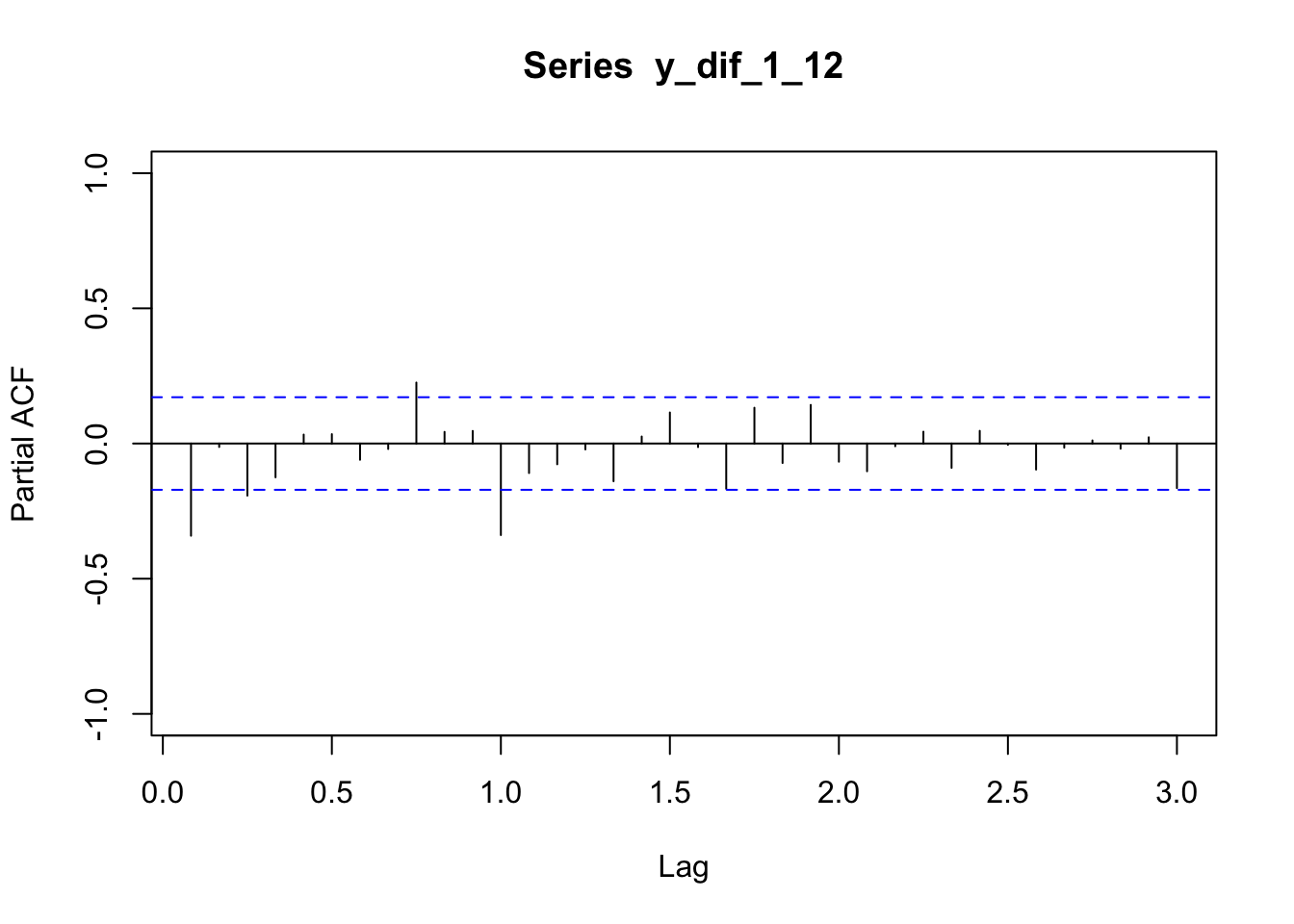
On s’appuie sur les autocorrélogrammes simple et partiels estimés :

y\_dif\_1\_12=diff(y\_dif1,lag=12,differences=1)

plot(acf(y\_dif\_1\_12,lag.max=36,plot=FALSE),ylim=c(-1,1))



plot(pacf(y\_dif\_1\_12,lag.max=36,plot=FALSE),ylim=c(-1,1))



 Tous les tests sont effectués au niveau de test 5%.

##### **Modèle 1**

On estime en premier lieu un modèle  au vu des autocorrélogrammes empiriques simples et partiels. Ce modèle s’écrit :



model1=Arima(y,order=c(1,1,1),list(order=c(1,1,1),period=12),include.mean=FALSE,method="CSS-ML")

summary(model1)

## Series: y

## ARIMA(1,1,1)(1,1,1)[12]

##

## Coefficients:

## ar1 ma1 sar1 sma1

## 0.1666 -0.5615 -0.099 -0.4973

## s.e. 0.2459 0.2115 0.154 0.1360

##

## sigma^2 estimated as 0.00138: log likelihood=245.16

## AIC=-480.31 AICc=-479.83 BIC=-465.93

##

## Training set error measures:

## ME RMSE MAE MPE MAPE

## Training set 0.0006239395 0.03489259 0.02595463 0.01199887 0.4696646

## MASE ACF1

## Training set 0.2144266 -0.01250397

t\_stat(model1)

## ar1 ma1 sar1 sma1

## t.stat 0.677738 -2.654214 -0.642984 -3.657670

## p.val 0.497938 0.007949 0.520235 0.000255

Box.test.2(model1$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)

## Retard p-value

## [1,] 6 0.64051

## [2,] 12 0.81959

## [3,] 18 0.82768

## [4,] 24 0.59646

## [5,] 30 0.75443

## [6,] 36 0.65902

Ce modèle ayant des paramètres non significatifs, on en teste un second.

##### **Modèle 2**

Soit le modèle :



model2=Arima(y,order=c(1,1,1),list(order=c(0,1,1),period=12),include.mean=FALSE,method="CSS-ML")

summary(model2)

## Series: y

## ARIMA(1,1,1)(0,1,1)[12]

##

## Coefficients:

## ar1 ma1 sma1

## 0.1960 -0.5784 -0.5643

## s.e. 0.2475 0.2132 0.0747

##

## sigma^2 estimated as 0.001375: log likelihood=244.95

## AIC=-481.9 AICc=-481.58 BIC=-470.4

##

## Training set error measures:

## ME RMSE MAE MPE MAPE

## Training set 0.0006214569 0.03495868 0.02587023 0.01205357 0.4682205

## MASE ACF1

## Training set 0.2137293 -0.01530519

t\_stat(model2)

## ar1 ma1 sma1

## t.stat 0.792074 -2.712668 -7.554412

## p.val 0.428318 0.006674 0.000000

Box.test.2(model2$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)

## Retard p-value

## [1,] 6 0.65243

## [2,] 12 0.80854

## [3,] 18 0.82136

## [4,] 24 0.57705

## [5,] 30 0.74768

## [6,] 36 0.67642

Ce modèle ayant des paramètres non significatifs, on en teste un troisième.

##### **Modèle 3**

Soit le modèle :



model3=Arima(y,order=c(0,1,1),list(order=c(0,1,1),period=12),include.mean=FALSE,method="CSS-ML")

summary(model3)

## Series: y

## ARIMA(0,1,1)(0,1,1)[12]

##

## Coefficients:

## ma1 sma1

## -0.4018 -0.5569

## s.e. 0.0896 0.0731

##

## sigma^2 estimated as 0.001371: log likelihood=244.7

## AIC=-483.4 AICc=-483.21 BIC=-474.77

##

## Training set error measures:

## ME RMSE MAE MPE MAPE

## Training set 0.0005730622 0.03504883 0.02626034 0.01098898 0.4752815

## MASE ACF1

## Training set 0.2169522 0.01443892

t\_stat(model3)

## ma1 sma1

## t.stat -4.482494 -7.618978

## p.val 0.000007 0.000000

Box.test.2(model3$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)

## Retard p-value

## [1,] 6 0.51519

## [2,] 12 0.72613

## [3,] 18 0.77822

## [4,] 24 0.50077

## [5,] 30 0.68838

## [6,] 36 0.65352

Les tests de significativité des paramètres et de blancheur du résidu sont validés au niveau 5%.

shapiro.test(model3$residuals)

##

## Shapiro-Wilk normality test

##

## data: model3$residuals

## W = 0.98637, p-value = 0.1674

Le test de normalité est également validé pour ce modèle.

##### **Prévision à l’aide du modèle retenu (3) de l’année 1961**

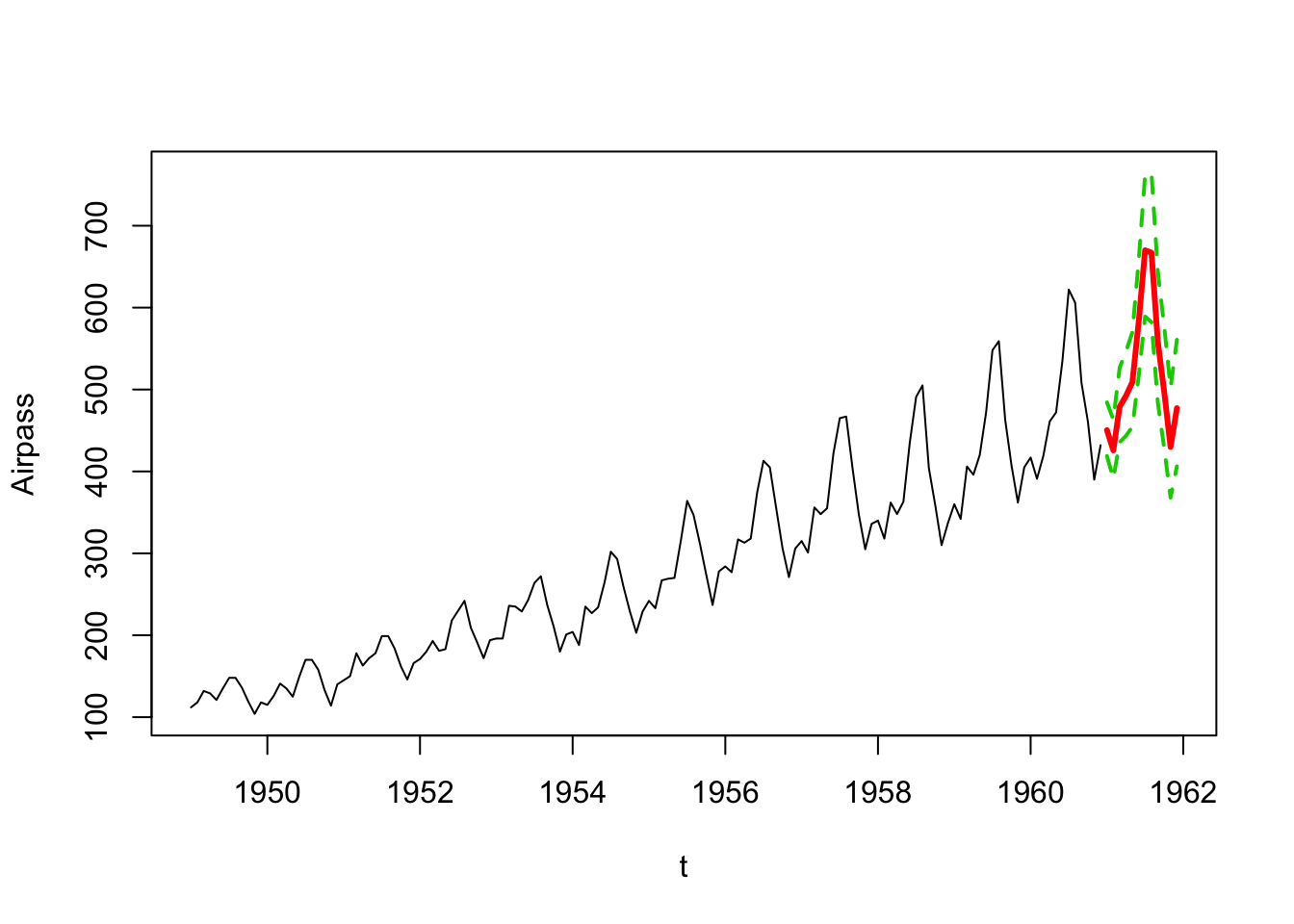
pred\_model3=forecast(model3,h=12,level=95)

pred=exp(pred\_model3$mean)

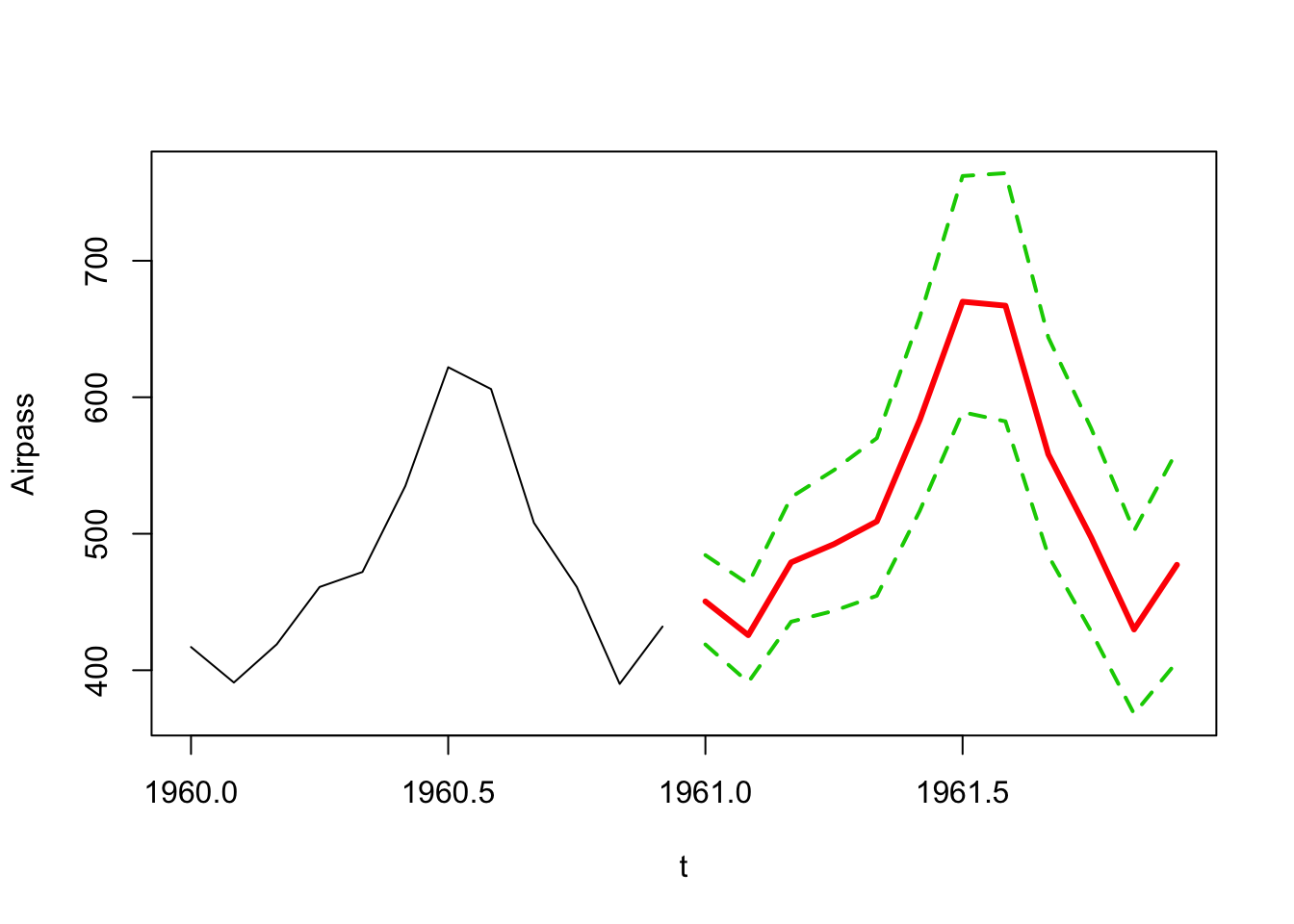
pred\_l=ts(exp(pred\_model3$lower),start=c(1961,1),frequency=12)

pred\_u=ts(exp(pred\_model3$upper),start=c(1961,1),frequency=12)

ts.plot(x,pred,pred\_l,pred\_u,xlab="t",ylab="Airpass",col=c(1,2,3,3),lty=c(1,1,2,2),lwd=c(1,3,2,2))



ts.plot(window(x,start=c(1960,1)),pred,pred\_l,pred\_u,xlab="t",ylab="Airpass",col=c(1,2,3,3),lty=c(1,1,2,2),lwd=c(1,3,2,2))



#### Analyse a posteriori

On tronque la série de l’année 1960, qu’on cherche ensuite à prévoir à partir de l’historique 1949-1959.

x\_tronc=window(x,end=c(1959,12))

y\_tronc=log(x\_tronc)

x\_a\_prevoir=window(x,start=c(1960,1))

On vérifie que le modèle 3 sur la série tronquée est toujours validé.

model3tronc=Arima(y\_tronc,order=c(0,1,1),list(order=c(0,1,1),period=12),include.mean=FALSE,method="CSS-ML")

summary(model3tronc)

## Series: y\_tronc

## ARIMA(0,1,1)(0,1,1)[12]

##

## Coefficients:

## ma1 sma1

## -0.3484 -0.5623

## s.e. 0.0943 0.0774

##

## sigma^2 estimated as 0.001338: log likelihood=223.63

## AIC=-441.26 AICc=-441.05 BIC=-432.92

##

## Training set error measures:

## ME RMSE MAE MPE MAPE

## Training set 0.00104934 0.03443221 0.02590904 0.01899277 0.4738142

## MASE ACF1

## Training set 0.2113963 0.004637637

t\_stat(model3tronc)

## ma1 sma1

## t.stat -3.695894 -7.262873

## p.val 0.000219 0.000000

Box.test.2(model3tronc$residuals,nlag=c(6,12,18,24,30,36),type="Ljung-Box",decim=5)

## Retard p-value

## [1,] 6 0.52539

## [2,] 12 0.85631

## [3,] 18 0.87341

## [4,] 24 0.78327

## [5,] 30 0.90181

## [6,] 36 0.84635

shapiro.test(model3tronc$residuals)

##

## Shapiro-Wilk normality test

##

## data: model3tronc$residuals

## W = 0.988, p-value = 0.3065

On constate que la réalisation 1960 est dans l’intervalle de prévision à 95% (basé sur les données antérieures à 1959).

pred\_model3tronc=forecast(model3tronc,h=12,level=95)

pred\_tronc=exp(pred\_model3tronc$mean)

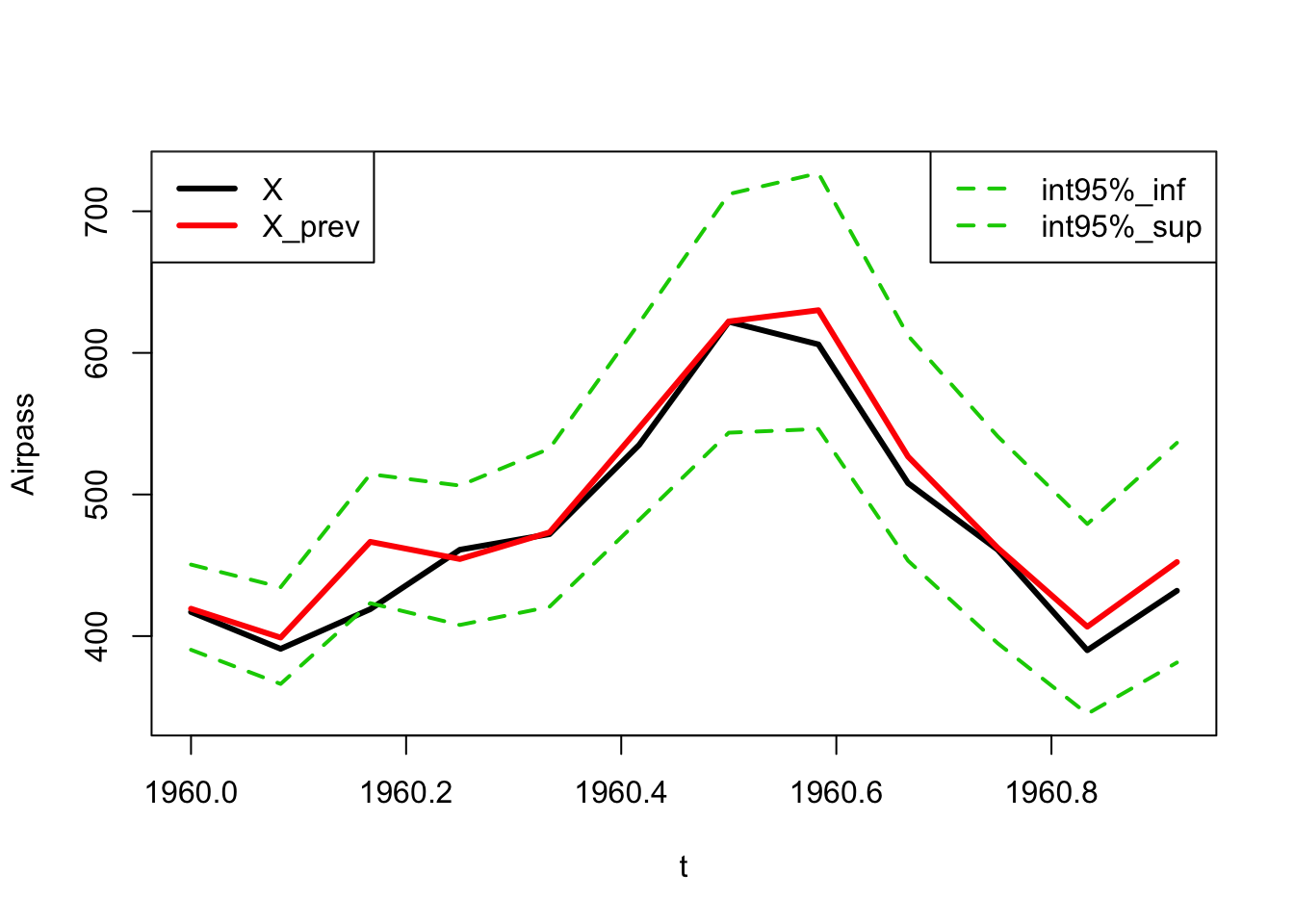
pred\_l\_tronc=ts(exp(pred\_model3tronc$lower),start=c(1960,1),frequency=12)

pred\_u\_tronc=ts(exp(pred\_model3tronc$upper),start=c(1960,1),frequency=12)

ts.plot(x\_a\_prevoir,pred\_tronc,pred\_l\_tronc,pred\_u\_tronc,xlab="t",ylab="Airpass",col=c(1,2,3,3),lty=c(1,1,2,2),lwd=c(3,3,2,2))

legend("topleft",legend=c("X","X\_prev"),col=c(1,2,3,3),lty=c(1,1),lwd=c(3,3))

legend("topright",legend=c("int95%\_inf","int95%\_sup"),col=c(3,3),lty=c(2,2),lwd=c(2,2))



On calcule les RMSE et MAPE.

rmse=sqrt(mean((x\_a\_prevoir-pred\_tronc)^2))

rmse

## [1] 18.59359

mape=mean(abs(1-pred\_tronc/x\_a\_prevoir))\*100

mape

## [1] 2.904473

L’interprétation des critères d’erreur dépend de la série et de la qualité de prévision exigée. Dans le cas présent, un MAPE de 2.9% semble satisfaisant a priori.

**Familiarisez-vous avec d'autres modèles connus**

On a vu ensemble comment corriger une série temporelle de ses variations saisonnières, et la prévoir, à l'aide du lissage exponentiel et surtout du modèle paramétrique linéaire ARMA. De nombreux autres modèles existent, en voici quelques représentants.

**Les modèles ARMAX**

Ils permettent d'incorporer des variables exogènes aux modèles ARMA. Notons qu'en cas de relation complexe avec des covariables autres que le passé du processus, il est souvent plus aisé de considérer un modèle de régression (adapté aux données temporellement dépendantes).

**Les modèles ARCH et GARCH**

ARCH (introduits par Engle en 1982) et GARCH (introduits par Bollerslev en 1986) sont des extensions des modèles AR qui permettent de tenir compte la \og volatilité stochastique\fg{} de séries temporelles. Ces techniques sont particulièrement utiles dans le domaine de la finance.

**Les modèles à mémoire longue**

Les modèles ARMA sont dits à mémoire courte : deux instants éloignés du processus n'ont que très peu d'interaction entre eux. Il existe d'autres modélisations pour les processus à mémoire longue, tels que les FARIMA (introduits par Granger en 1980).

**Les modèles multivariés**

On est parfois amenés à modéliser simultanément plusieurs séries ayant de fortes relations entre elles. On considère alors des modèles VAR, VEC, espace-état (et le filtrage de Kalman) etc., ainsi que des notions telles que la causalité au sens de Granger, la cointégration, etc.

**D'autres alternatives**

Il existe de nombreux autres modèles (paramétriques) non linéaires tels que les modèles à seuil.

Au delà des solutions paramétriques, on trouve également des modèles non et semi paramétriques, ces derniers permettant de pallier le "fléau de la dimension" du non-paramétrique. Citons les méthodes de noyau, la décomposition en ondelettes, les modèles GAM, etc.

Enfin, il existe également des modèles temporels pour des données spatiales

# Partie 4

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Mettre en œuvre des techniques simples pour prévoir une série temporelle
* Comprendre la problématique posée par le lien temporel

### Question 1

**Un processus stationnaire :**

* + 

permet de modéliser des séries temporelles qui ne peuvent pas présenter une tendance ou une saisonnalité

* + 

permet de modéliser des séries temporelles présentant une tendance

* + 

permet de modéliser des séries temporelles présentant une saisonnalité

* + 

permet de modéliser des séries temporelles présentant une tendance et une saisonnalité

### Question 2

**Un processus AR est caractérisé par :**

* + 

des autocorrélations simples et partielles s'annulant à partir d'un certain rang

* + 

des autocorrélations simples décroissant rapidement vers 0 et des autocorrélations partielles s'annulant à partir d'un certain rang

* + 

des autocorrélations simples s'annulant à partir d'un certain rang et des autocorrélations partielles décroissant rapidement vers 0

### Question 3

**Un processus ARMA est caractérisé par :**

* + 

des autocorrélations simples et partielles s'annulant à partir d'un certain rang

* + 

des autocorrélations simples s'annulant à partir d'un certain rang

* + 

des autocorrélations partielles s'annulant à partir d'un certain rang

* + 

aucune des propriétés précédentes

*En pratique on utilise des heuristiques pour identifier un modèle ARMA*

### Question 4

**Un processus SARIMA :**

* + 

est stationnaire

* + 

permet de modéliser des séries temporelles présentant une tendance mais pas de saisonnalité

* + 

permet de modéliser des séries temporelles présentant une saisonnalité mais pas de tendance

* + 

permet de modéliser des séries temporelles présentant une tendance et/ou une saisonnalité

### Question 5

**Afin de valider un modèle ARMA, il faut vérifier :**

*Attention, plusieurs réponses sont possibles.*

* + 

la significativité des paramètres

* + 

la blancheur du résidu

* + 

la normalité des résidus