Rapport du Projet SAE :

Description et prévisions de données temporelles

6 janvier 2023

JULIEN Léo
BABINGUI Anaïs
MOULIN Sarah

STID
STATISTIQUE ET INFORMATIQUE DÉCISIONNELLE

AVIGNON UNIVERSITÉ
Institut Universitaire de Technologie
iut

Sommaire:

Introduction

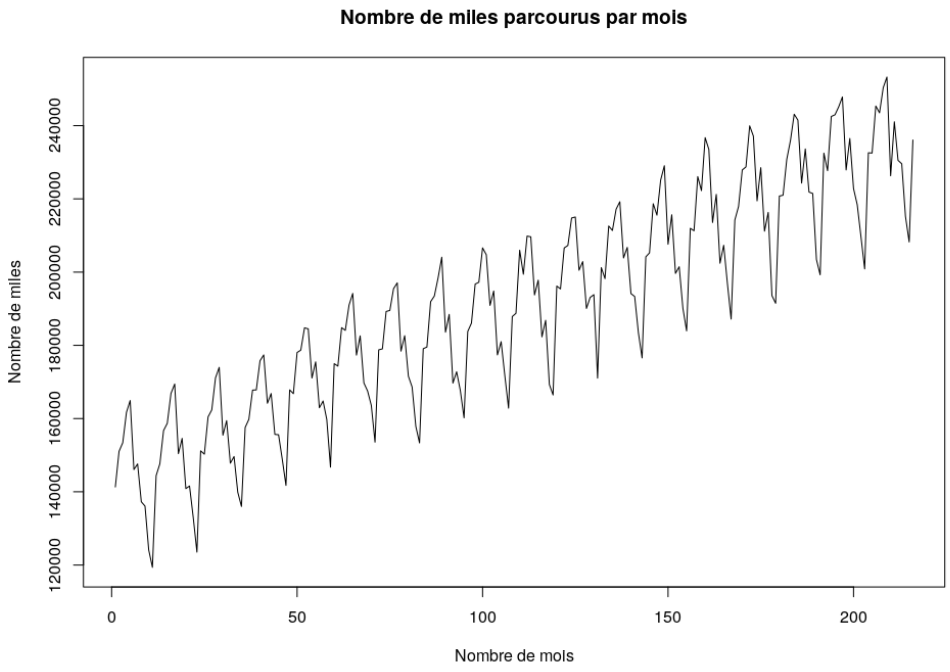
La location séduit de plus en plus de consommateurs, qu'elle soit de longue durée ou de courte durée les avantages sont nombreux. Les locataires peuvent disposer d'une voiture récente et profiter d'une garantie réparations intégrée. Acheter une voiture et l'entretenir peuvent en effet avoir un certain coût, non négligeable, surtout pour les gens qui roulent peu ou qui habitent une grande ville. La consommation de location automobile à doublé en l'espace de 15 ans. Il existe de nombreuses formules de locations.

La récolte du nombre de miles effectués par l'ensemble d'un parc automobile d'une agence de location de voitures est essentiel pour calculer l'amortissement des véhicules et la rentabilité des locations. Nous avons donc exploité des données collectées de avril 1984 à mars 2002, mensuellement prélevées, du nombre de miles parcourus par l'ensemble du parc automobile d'une agence de location de voitures.

Nous allons présenter dans ce rapport l'analyse effectuée de la série temporelle des données que nous ont avons étudiées. Pour cela nous allons dans en premier temps effectuer une analyse descriptive et graphique de nos données afin de visualiser rapidement les comportements globaux de la série, modéliser notre série et enfin effectuer des prévisions des prochains miles effectués l'année suivante par les véhicules de l'agence.

I - Analyse descriptive et graphique

Nous allons d'abord effectuer une étude graphique afin de visualiser rapidement les comportements globaux de notre série.



Il y a n=216 données. Les données s'étendent d'Avril 1984 à Mars 2002. Il y a une donnée par mois. La tendance est croissante, ce qui montre que le nombre de miles parcouru augmente au fil des années.

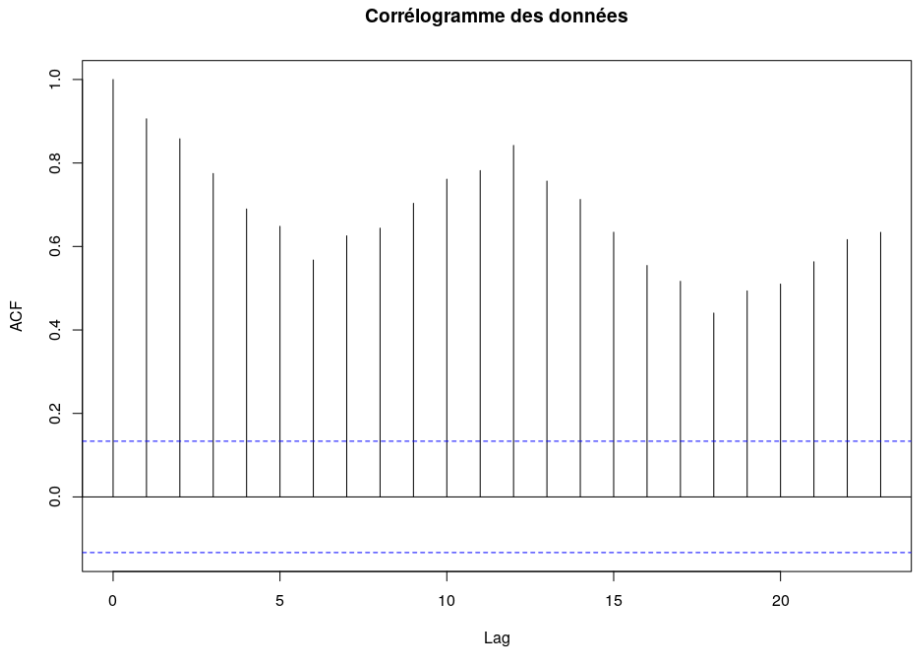
Pour déterminer quel modèle utiliser, on utilise la méthode de la bande : si on trace deux droites, l'une passant par les minima, l'autre passant par les maxima de la série, elles sont à peu près parallèles. Le modèle additif est donc le plus adapté.

Les variations saisonnières semblent être de p=12. En effet, une année est composée de 12 mois. La composante saisonnière est de np=18, car il y a 18 années. Nous observons aussi qu'il y a un pic au milieu de chaque composante saisonnière, ce qui laisse à supposer que beaucoup plus de miles sont parcourus à cette période-là. Cela correspond à peu près à la période de l'été, il est donc normal que les individus voyagent plus à cette période-là.

II - Autocorrélation temporelle

L'autocorrélation mesure la manière dont les éléments de la série sont distribués dans le temps. Elle mesure le degré d'influence que chaque observation a sur ses voisines.

La fonction d'autocorrélation nous permet d'obtenir le corrélogramme de nos données.



Sur le corrélogramme, nous pouvons observer qu'il y a un phénomène oscillatoire, dû à la saisonnalité de la série. Il y a effectivement une période p=12. Cependant, la tendance est moins visible. De plus, il est évident que 95% des valeurs de l'autocorrélation ne sont pas situées entre les lignes bleues, nous considérons donc que le processus sous-jacent n'est pas indépendant (les données sont corrélées).

III - Modélisation de notre série temporelle

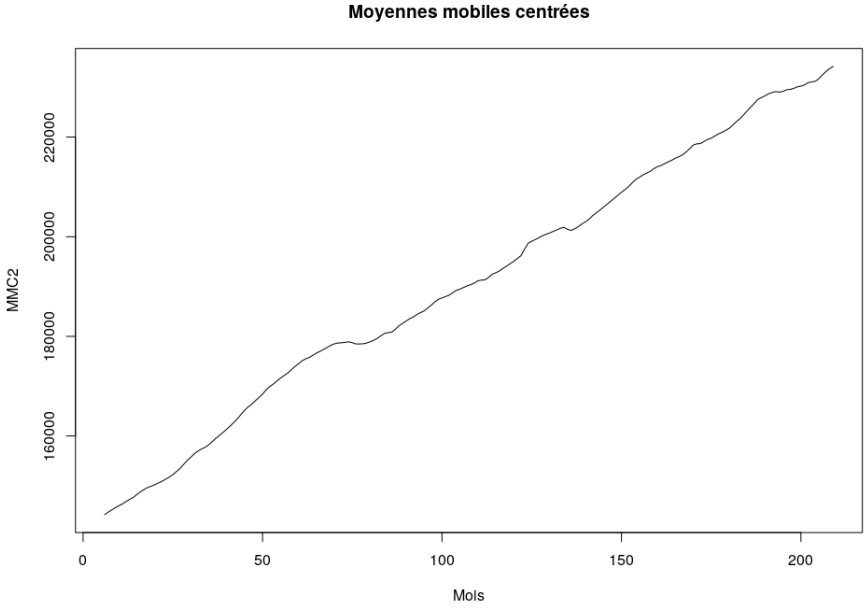
a) Estimation de la tendance : Lissage par moyenne mobile et méthode paramétrique

Pour estimer la tendance, nous commençons par réaliser un lissage avec les moyennes mobiles. Pour réaliser les moyennes mobiles, on utilise, comme vu précédemment, un ordre k=12. $MM_k(x_{t+0.5}) = \frac{1}{k} \sum_{i=t-m+1}^{t+m} x_i$

On calcule ensuite les moyennes mobiles centrées:

$$MMC_k(x_t) = \frac{1}{k}(0.5x_{t-m} + \sum_{i=t-m+1}^{t+m-1} x_i + 0.5x_{t+m})$$

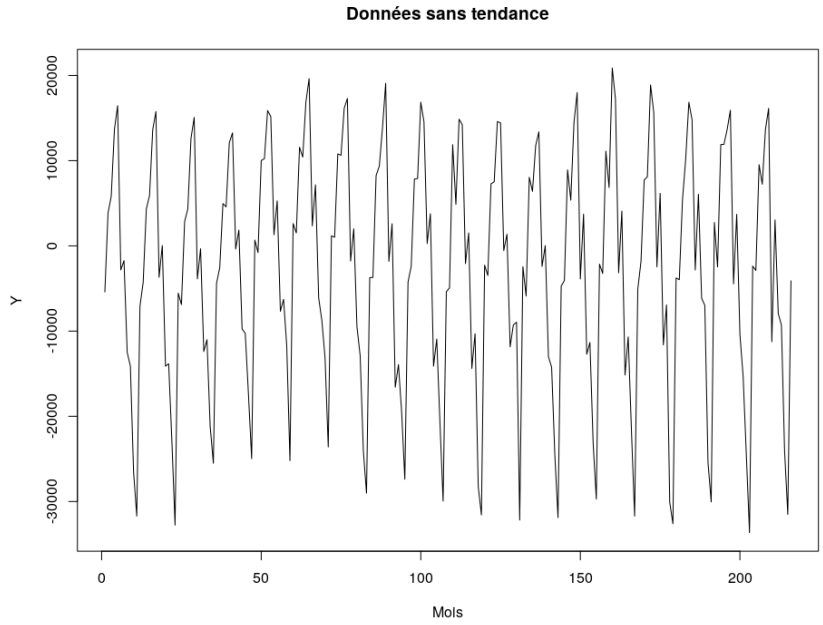
Cela nous permet d'obtenir le graphique suivant:



La forme de la tendance est donc croissante, nous choisissons d'établir le modèle paramétrique d'une droite, en utilisant la série lissée avec les moyennes mobiles :

$f_t = 434.625 + x \times 146296.4$

Nous calculons aussi les données sans tendance, ce qui nous permet d'observer la saisonnalité de la tendance, et de mieux déterminer si le modèle est additif ou multiplicatif.



En observant les données sans tendance, nous pouvons confirmer notre choix d'un modèle additif.

b) Estimation de la composante saisonnière et des coefficients saisonniers

On estime ensuite la composante saisonnière. Pour cela, nous avons une période p=12, répétée 18 fois. Il y a donc 12 coefficients saisonniers.

$\hat{s}_t$ ={-428.1199 10826.3101 10342.3513 17772.1147 18590.1003 281.9748 5496.4048 -8743.4429 -8206.9573 -18829.0828 -27022.7083 -78.9449}

c) Calcul de la série corrigée des variations saisonnières (CVS)

Nous calculons la série CVS: $d_t = x_t - \hat{s}_t$

Voici la série CVS et la tendance:

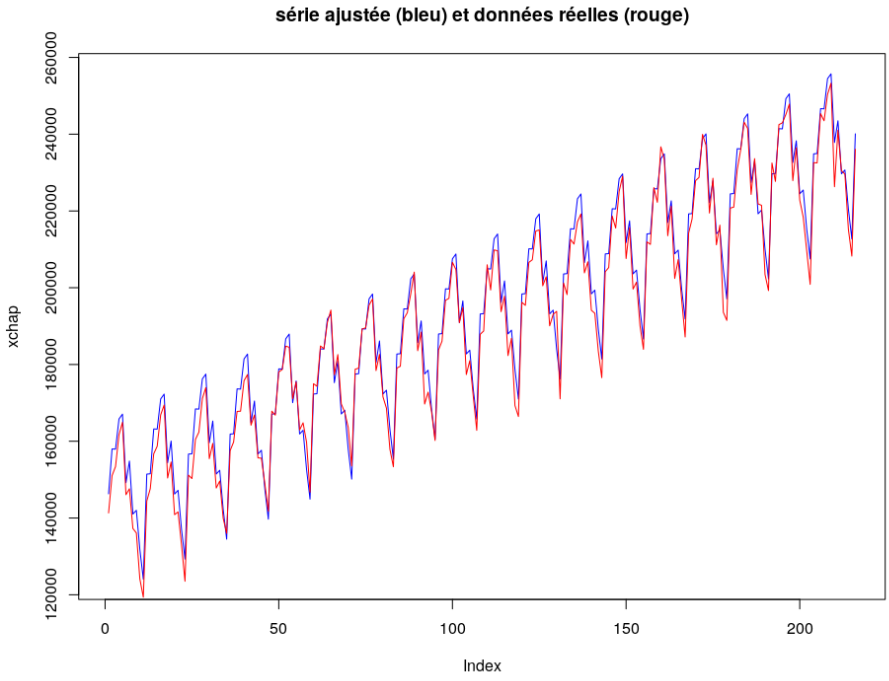
La courbe CVS permet de montrer les erreurs de la série. Nous pouvons observer que la courbe CVS suit approximativement la même forme que la tendance, donc l'ajustement est correct.

d) Calcul de la série ajustée

On détermine ensuite la série ajustée en appliquant le modèle additif :

$$\hat{x}_t = 434.625 + x \times 146294.4 + \hat{c}_t$$

On obtient donc le graphique suivant:



Cela nous permet d'observer l'évolution du nombre de miles parcourus si les variations saisonnières avaient été parfaitement périodiques et s'il n'y avait pas eu de variations résiduelles.

Nous calculons maintenant le critère MSE (Mean Squared Error), qui permet de comparer les modèles ou les prévisions entre eux. Plus le critère obtenu est petit, plus l'ajustement est bon.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

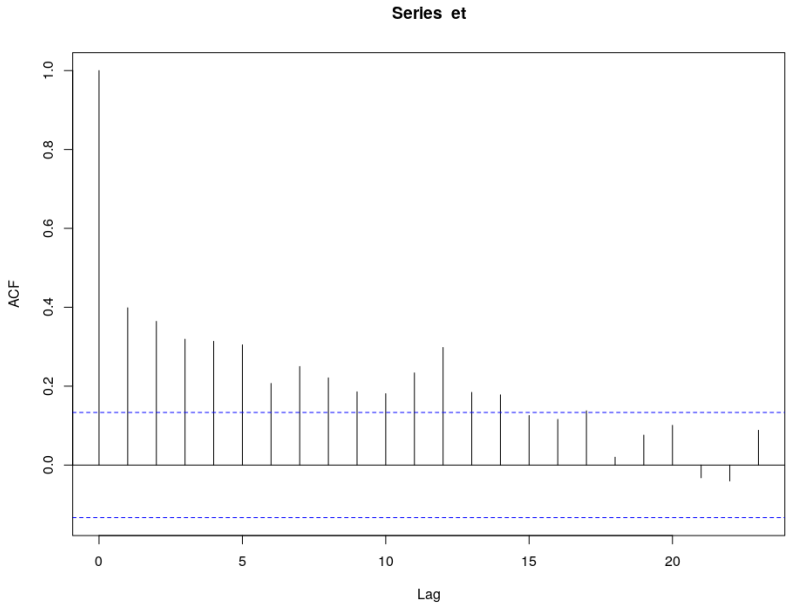
Nous obtenons un MSE de 15 784 967, un résultat très grand, mais au vu des données, qui ont une moyenne de 190 753.7, ce n'est pas surprenant. Nous allons par la suite essayer de réduire ce critère au mieux grâce à notre modèle. Il s'agira alors de minimiser les erreurs. Pour juger de l'efficacité du modèle, nous comparerons ce MSE avec les prochains critères calculés.

e) Estimation des variations résiduelles et ajustement d'un modèle stochastique sur les résidus

Nous allons à présent étudier les résidus. Les variations résiduelles estimées sont obtenues en retirant la série ajustée de la série initiale.

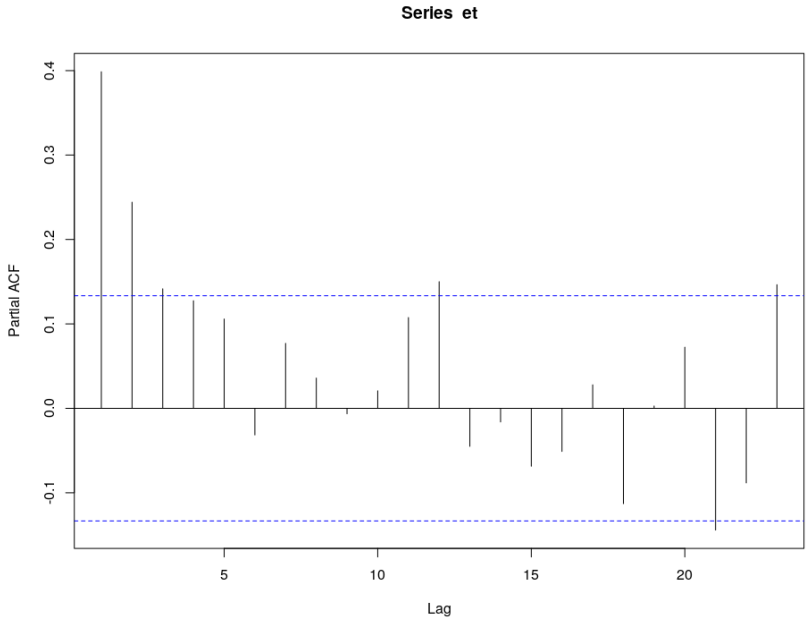
Nous allons nous intéresser à la fonction d'autocorrélation, qui nous permettra d'étudier les variations résiduelles et de leur ajuster un modèle stochastique. Nous utiliserons pour cela le modèle ARMA (Auto Regressive Moving Average).

Afin de choisir l'ordre q à utiliser pour le modèle ARMA nous nous aidons de la fonction d'autocorrélation (acf(x)) :



Nous observons sur le corrélogramme 15 pics en dehors des lignes bleues, nous utiliserons alors q=15 pour le modèle ARMA.

Il nous faut à présent déterminer l'ordre p à utiliser avec le modèle ARMA. Pour cela nous regardons la fonction d'autocorrélation partielle (pacf(x)).



Nous observons sur le corrélogramme 5 pics en dehors des lignes bleues, nous décidons alors d'utiliser p=5 pour le modèle ARMA.

Ayant maintenant déterminé les ordres p et q , nous pouvons à présent utiliser le modèle ARMA. Ce modèle nous permettra par la suite d'effectuer des prévisions.

Après avoir estimé les résidus, nous construisons le modèle ajusté total à partir des 3 composantes estimées : $f_t + s_t + e_t$

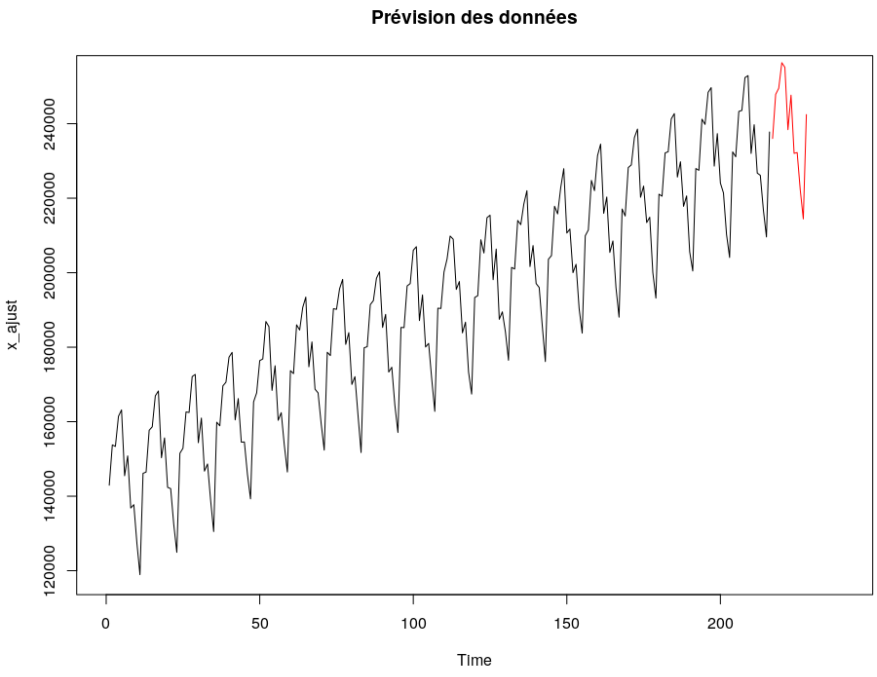
Nous obtenons un MSE de 5 140 823, le modèle nous a bien permis de réduire ce critère qui était précédemment de 15 784 967, soit une différence de 10 644 144 entre les deux valeurs, différence qui n'est pas négligeable. Le MSE du modèle utilisant les résidus est la plus faible. C'est donc le meilleur critère.

Nous décidons ensuite de calculer le MSE avec un ajustement q=12 car notre période est de 12. Nous voulons alors tester cet ajustement pour déterminer quel ajustement serait le plus intéressant à utiliser. On obtient un MSE de 5 535 606. Il est légèrement plus grand que le MSE avec un ordre q=15. On gardera alors ce dernier, meilleur car plus petit.

IV - Prévisions

Une fois que l'on a obtenu un bon ajustement de la série, il est possible de faire des prédictions de valeurs à partir du modèle estimé. Nous allons alors utiliser notre modèle ARMA afin de faire des prévisions à court terme. Nous avons une période p=12, essayons alors de prévoir les 12 prochaines valeurs de la série.

Nous faisons des prévisions en supposant que la tendance va suivre la même évolution et que les variations saisonnières seront identiques. On obtient alors ce nouveau graphique:



Voici en rouge les prévisions sur 12 mois pour l'année 2003.

Conclusion

Ainsi, nous avons choisi un choisi le modèle paramétrique d'une droite pour estimer au mieux la tendance, puis nous avons choisi le modèle additif qui nous a permis, après avoir calculé la composante saisonnière et les résidus, de déterminer le modèle ajusté. Avec un MSE de 5 140 823 pour le modèle ajusté et q=15, nous obtenons le critère le plus petit. Cela nous permet d'affirmer que ce modèle possède le meilleur ajustement. Nous avons ensuite pu effectuer des prévisions sur le nombre de miles parcourus pour l'année suivante, l'année 2003. Il s'avère donc que le nombre de miles parcourus par an est en constante augmentation au fil des années. Les individus louent donc plus de voitures et voyagent beaucoup plus durant l'été.

Annexe: code R

```
##1)

X=read.csv("monthly_total_vehicles_miles_traveled_of_car_rental_agency.csv",sep=",")

plot(X$miles,type="l",ylab="Nombre de miles",xlab="Nombre de mois",main = "Nombre de miles parcourus par mois")

acf(X$miles,main="Corrélogramme des données") #Corrélogramme des données


##2)

#Calcul des moyennes mobiles centrées avec k=12

MM12=filter(X$miles,rep(1/12,12)) #Moyennes mobiles

MMC2=na.omit(filter(na.omit(MM12),rep(1/2,2))) #Moyennes mobiles centrées

plot(MMC2,main="Moyennes mobiles centrées",type="l",xlab="Mois")

#Calcul de la tendance

x=1:length(MMC2)

a=cov(x,MMC2)/var(x)

b=mean(MMC2)-a*mean(x)

x=1:nrow(X)

ft=b+x*a

#Calcul des données sans tendance

Y=X$miles-ft #modèle additif

plot(Y,type = "l",main="Données sans tendance",xlab="Mois") #affichage des données sans tendance pour voir si additif ou multiplicatif


##3)

#Calcul composante saisonnière

tab_add=matrix(Y,ncol=12,byrow=T)
```

```
ctilde=apply(tab_add,2,mean)
c_chap=ctilde-mean(ctilde)
s_chap=rep(c_chap,18) #variation saisonnière

##4)
#Calcul de la série CVS (Corrigée des Variations Saisonnières)
CVS=X$miles-c_chap
plot(CVS,type="l",col="blue",main="Série CSV (bleu) et tendance (rouge)")
lines(ft,col=« red")

##5)
#Calcul de la série ajustée
xchap=a*x + b + c_chap
plot(xchap,type="l",col="blue",main = "série ajustée (bleu) et données réelles (rouge)")
lines(X$miles,col="red")
#Calcul du MSE
MSE_1 = mean((X$miles-xchap)**2) # =15 784 967

##6)
#Calcul des résidus :
et=X$miles-xchap
acf(et) # corrélogramme des résidus
#q=15 (on compte pas le pic au lag 0)
pacf(et) # autocorrélation partielle p=5 (nombre de pics en dehors des lignes)
fit=arima(et, order=c(5,0,15)) #modèle ARMA avec order=c(p,0,q)
```

```
##7)

library(forecast)

x_ajust = xchap + fitted(fit)

plot(x_ajust, main="Série ajustée (noir) et données réelles (rouge)", xlab="mois")

lines(X$miles,col="red")

MSE_2 = mean((X$miles-x_ajust)**2)

#calculons le MSE avec un ajustement avec q=12

fit2=arima(et, order=c(5,0,12)) #modèle ARMA avec order=c(p,0,q)

x_ajust2 = xchap + fitted(fit2)

plot(x_ajust2,main="Série ajustée avec q=15, q=12 (blue) et données réelles (rouge)",
xlab="mois",ylab="val")

lines(x_ajust,col="blue")

lines(X$miles,col="red")

MSE_3 = mean((X$miles-x_ajust2)**2) # MSE avec 2e modèle ajusté

#2622796 de moins que le MSE_1

MSE_1;MSE_2;MSE_3

#1er MSE + faible


##8)

#Prévisions 12 mois

x=1:228 #Ajout des 12 temps

#Calcul de la prévision :

library(forecast)

forkast=as.data.frame(forecast(fit, h=12))

previs=rep(NA,228)

previs[217:228]=x[217:228]*a+b+c_chap+forkast$`Point Forecast`

plot(x_ajust,type="l",main = "Prévision des données",xlim=c(0,240))
```

```
lines(previs,col="red")
```