Compte-rendu TP séries temporelles

Devoir maison obligatoire (Dauphine)

Paul Hardouin

January 12, 2020

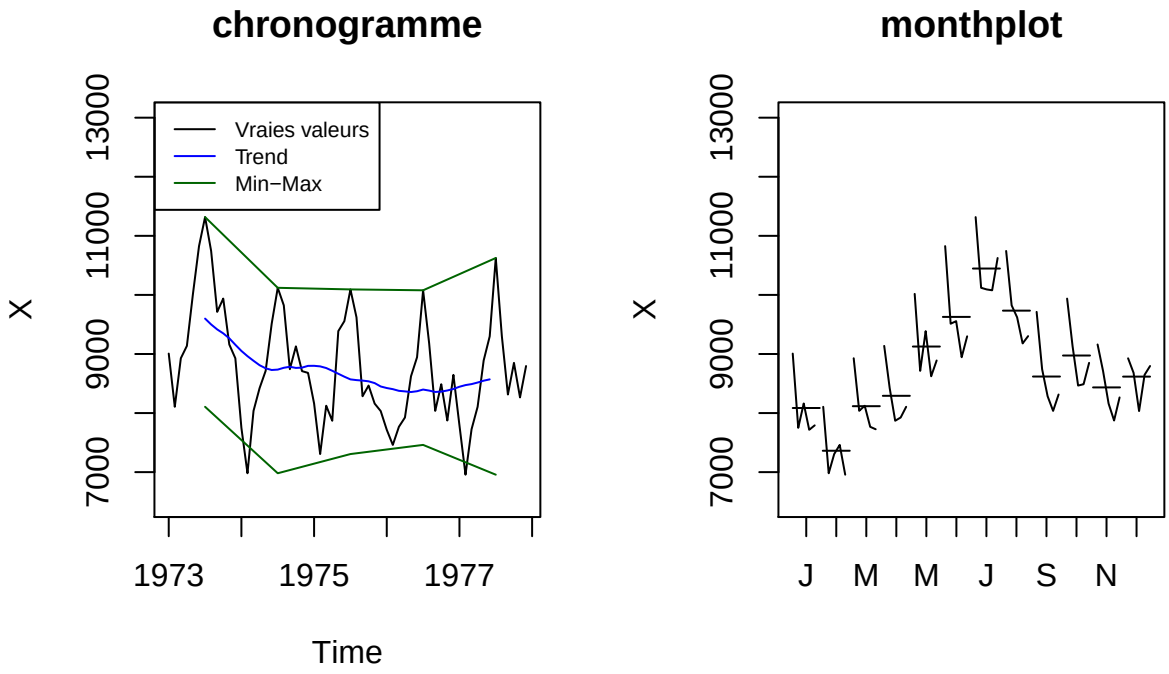
Contents

1. Etude du jeu de données USAccDeaths

Nous nous intéressons dans ce jeu de données à l'étude du nombre de morts accidentelles au USA, entre 1973 et 1979. On commence par charger les données et par isoler la dernière année pour pouvoir la comparer à nos prévisions.

1.1. Analyse descriptive

En premier lieu, nous faisons une analyse descriptive pour comprendre la structure de cette série temporelle.



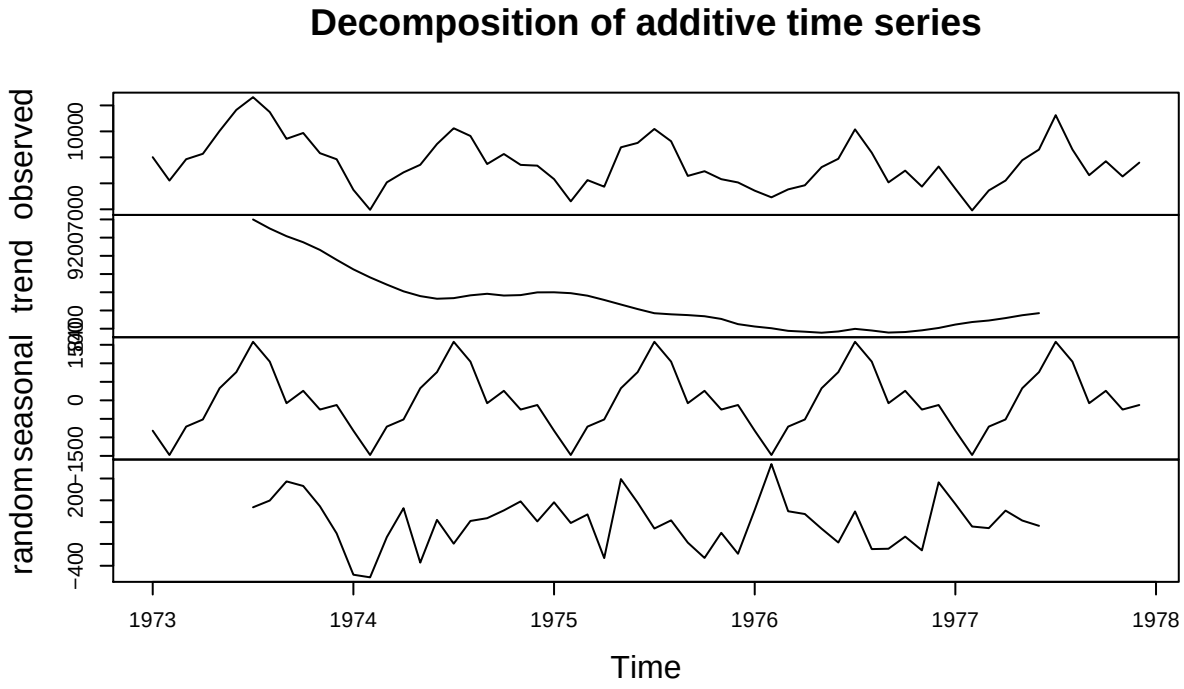
Sur le chronogramme, on observe un motif périodique, ce qui permet de supposer un effet saisonnier. On peut estimer la tendance par la méthode des moyennes mobiles. On prend un ordre égal à 12, car le motif saisonnier semble durer 12 mois.

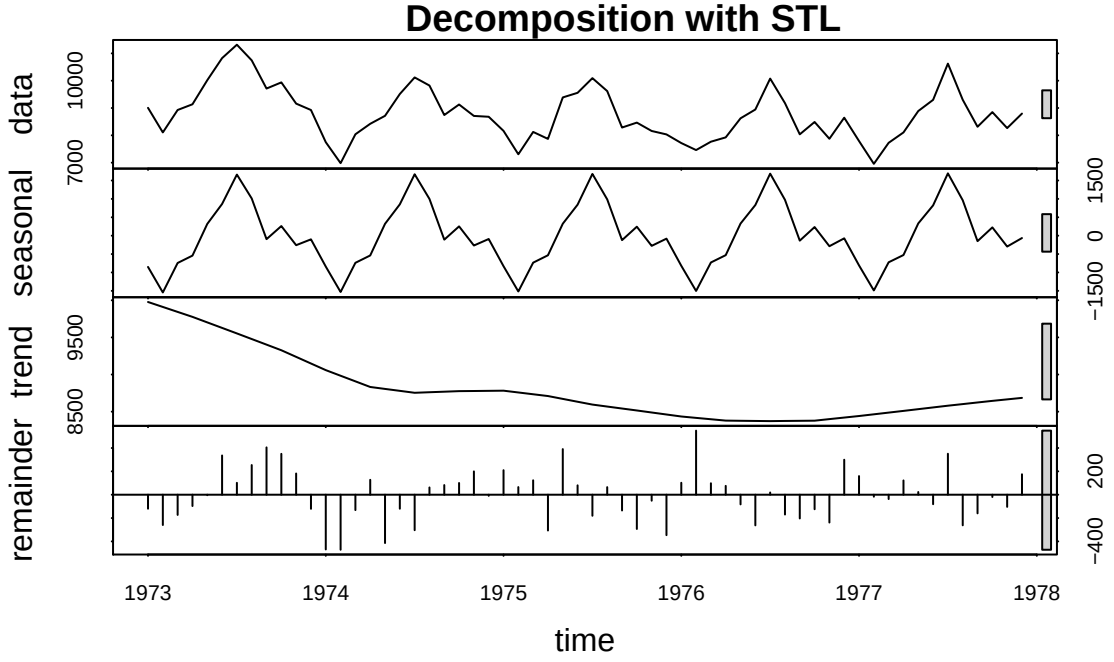
Sur le monthplot, les chronogrammes mensuels ne sont pas identiques d'un mois à l'autre. Cela confirme un **effet saisonnier** : en particulier, la saison estivale semble plus propice aux accidents. Sans doute est-ce du au fait que les gens sortent plus de chez eux pendant les beaux jours.

En faisant l'hypothèse d'un modèle complètement additif ou multiplicatif, on peut essayer d'arbitrer en utilisant la méthode de la bande. Sur le chronogramme, les 2 courbes semblent parallèles, ce qui nous permet d'aller vers un **modèle additif**.

Nous avons également observé le **lag-plot** (non affiché ici); mais il ne fait pas ressortir clairement l'effet saisonnier, sans doute à cause du faible nombre de données (60) de cette série.

On réalise alors une décomposition à l'aide de la fonction **decompose** en mode **additif**, ainsi qu'avec la fonction **stl**. Dans les 2 cas, on retrouve la tendance calculée précédemment par moyenne mobile, ainsi qu'un motif saisonnier évident.





1.2. Lissage exponentiel

On veut maintenant faire de la prédiction par lissage exponentiel. Pour cela, on utilise la fonction **ets**, en contraignant le modèle avec un code à 3 lettres:

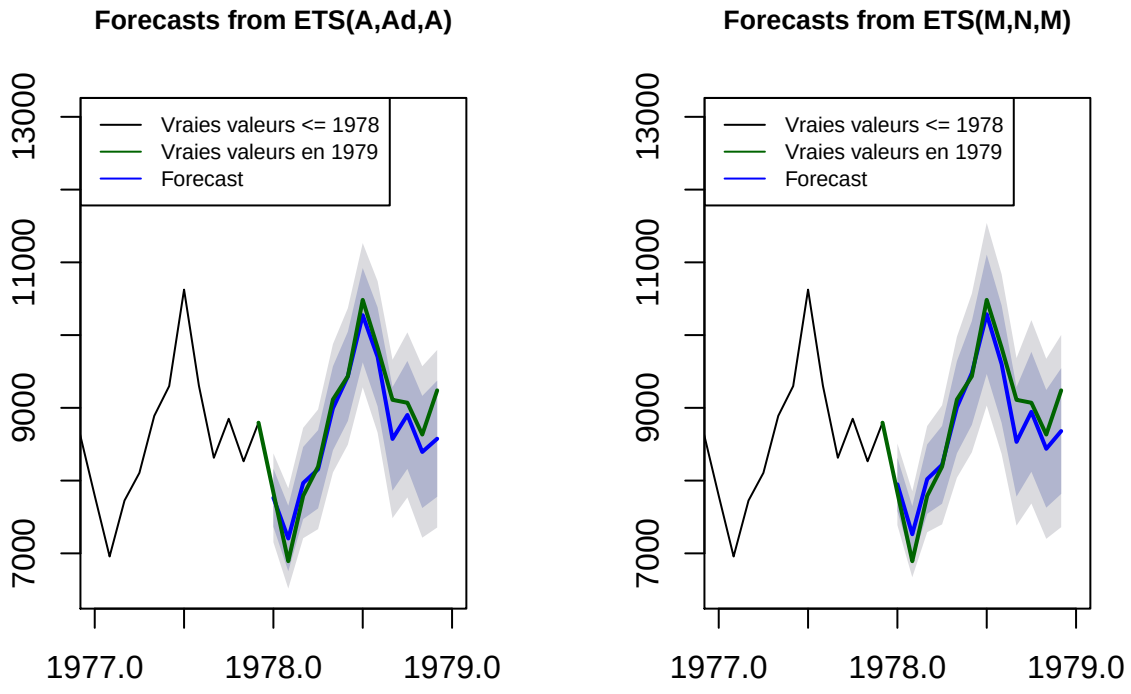
```
lettre 1 : e : erreur    AMZ [     additive,multiplicative,auto]
lettre 2 : t : tendance NAMZ [none,additive,multiplicative,auto]
lettre 3 : s : saison   NAMZ [none,additive,multiplicative,auto]
```

En premier lieu, on suit nos conclusions précédentes, et on contruit un modèle additif **AAA**. Ensuite, on observe les résultats obtenus en utilisant la méthode automatique **ZZZ**.

Consigne	Modèle retenu	AIC	AICc	BIC
additif	ETS(A,Ad,A)	950.6122	967.2951	988.3104
automatique	ETS(M,N,M)	949.9315	960.8406	981.3467

Ces résultats privilégient un modèle mutiplicatif sans tendance, ce qui semble aller à l'encontre de nos premières conclusions. En revanche, si on compare ces 2 modèles, on remarque que les critères (AIC, AICc, BIC) ont des valeurs très proches. Il en est de même pour les intervalles de confiance des valeurs prédites (non affiché ici, mais visibles dans les plots ci-dessous).

Cette diversité de modèles acceptables est sans doute due au fait que la tendance a une pente faible. L'affichage des prévisions ci-dessous montre des courbes très proches, et des ordres de grandeur très proches concernant les intervalles de confiance.



1.3. Modélisation

Analyse des résidus

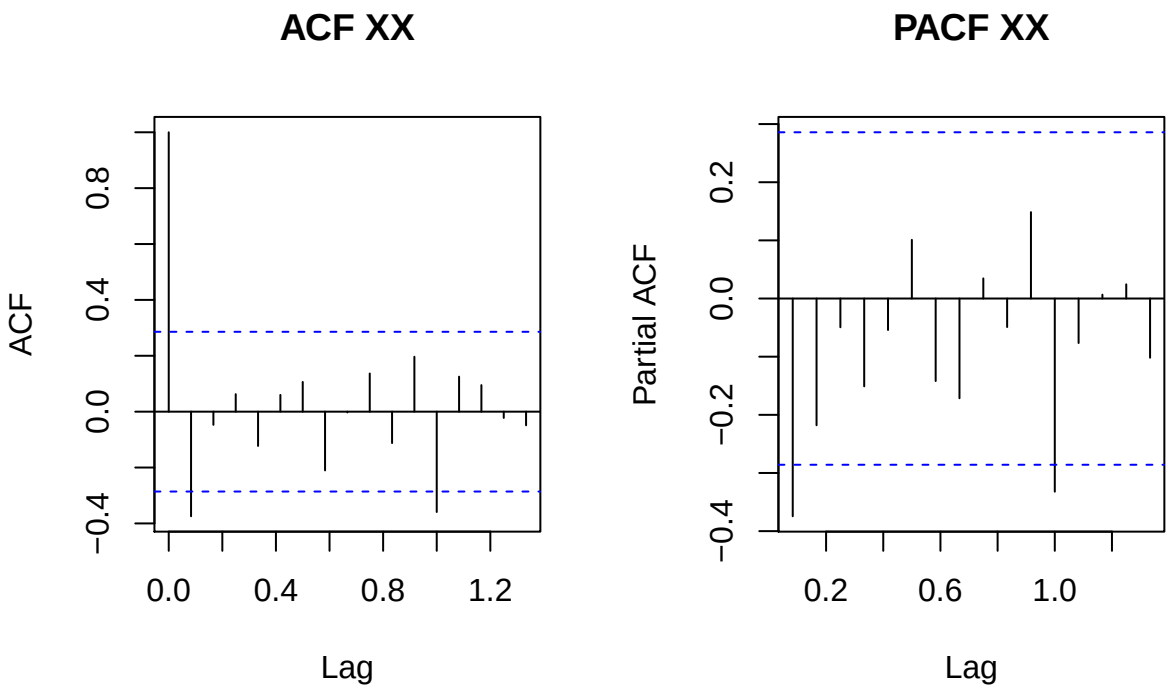
En premier lieu, on différencie la série pour enlever une tendance et une saisonnalité. On veut ainsi se ramener à une série stationnaire.

```
XX = diff(diff(X,lag=12,dfference=1),lag=1,dfference=1)
```

On essaye alors d'identifier des ordres de modélisation avec les ACF et PACF, avec les règles suivantes.

```
AR(p) :  ACF en exponentielle décroissante + PACF nulle en p+1
MA(q) : PACF en exponentielle décroissante +  ACF nulle en q+1
ARMA  : compliqué ...
```

On voit une exponentielle décroissante sur la PACF, et l'ACF est nulle à partie de q+1 = 2. On se dirige alors vers un modèle **MA(1)**.



Modèle MA(1) : VALIDE

On crée le modèle **model_1 = Arima(XX,order=c(0,0,1))** et on analyse ses métriques. Le résidu est bien un bruit blanc, et il n'y a pas de colinéarités.

```
Critères    [Arima]     AIC=695.22   AICc=695.78   BIC=700.7
Box-Pierce  [Box.test]  OK  (p-value = 0.9403 > .05)
Colinéarité [cor.arma]  OK  (pas de coefficient supérieur à 0.9)
```

Modèle automatique : ARIMA(0,0,1)(0,0,1)[12] : VALIDE

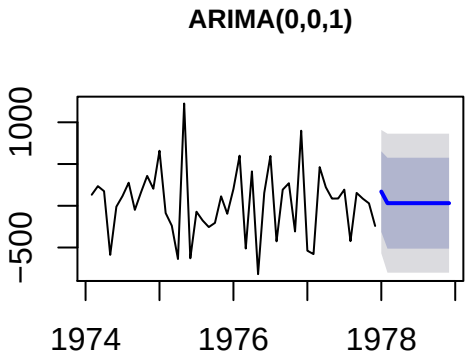
On crée le modèle automatique **model_2 = auto.arima(XX)** et on analyse ses métriques. Le résidu est bien un bruit blanc, et il n'y a pas de colinéarités.

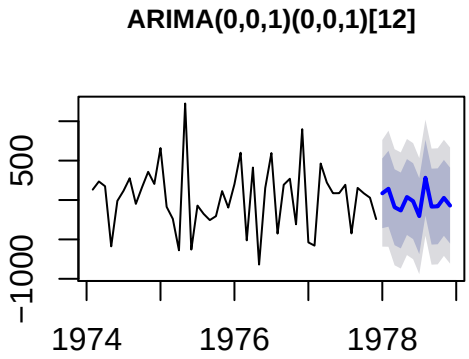
```
Critères    [Arima]     AIC=689.54   AICc=690.1   BIC=695.09
Box-Pierce  [Box.test]  OK  (p-value = 0.9571 > .05)
Colinéarité [cor.arma]  OK  (pas de coefficient supérieur à 0.9)
```

Ce modèle est assez proche de celui intuité visuellement. La composante saisonnière du résidu est seulement modelisée un peu plus finement.

Prévisions des résidus

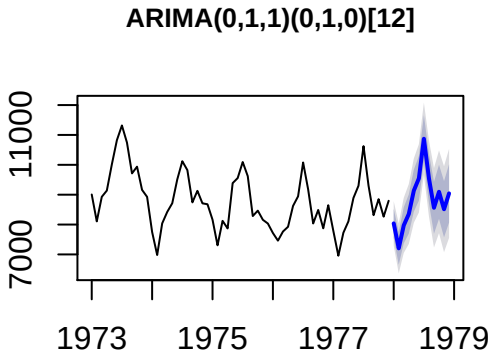
On observe une prévision un peu plus fine du résidu pour le modèle automatique. Il est probable que l'impact soit négligeable quand on rajoute la tendance et la saisonnalité de la série.

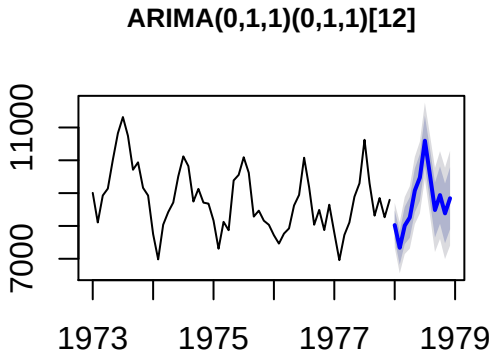




Prévisions des séries

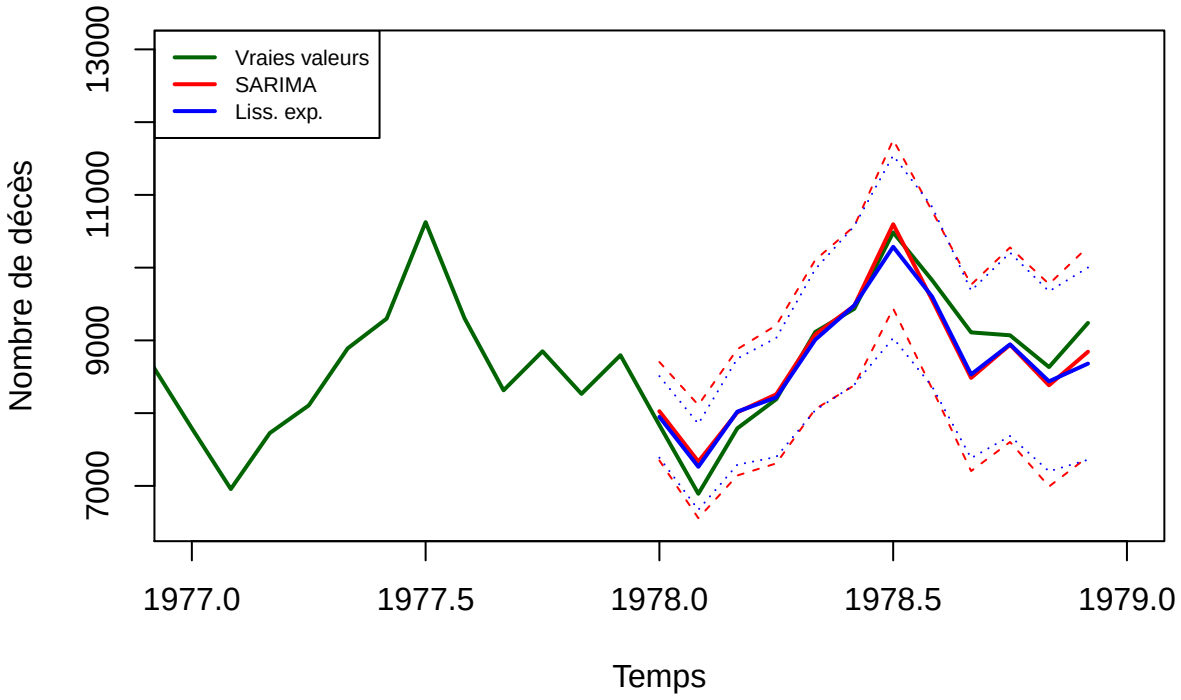
Comme prévu les 2 prévisions sont très proches.





Comparaison globale

On veut comparer les vraies valeurs, le lissage exponentielle, et la prévision par modélisation des résidus. Les 2 prévisions sont assez proches des vraies valeurs, et leur intervalles de confiance sont presque confondus.

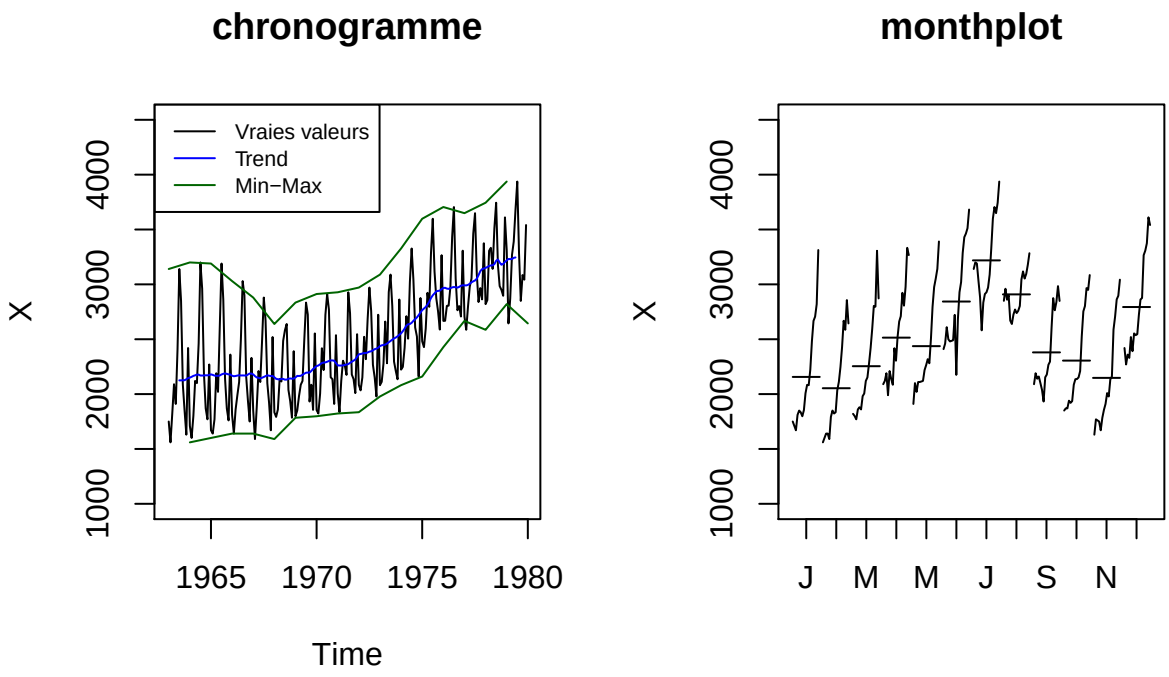


2. Etude du jeu de données SNCF

Nous nous intéressons dans ce jeu de données à l'étude du nombre de voyageurs sur le réseau SNCF. Les données sont issues du site **https://freakonometrics.hypotheses.org**. On commence par le charger.

2.1. Analyse descriptive

En premier lieu, nous faisons une analyse descriptive pour comprendre la structure de cette série temporelle.



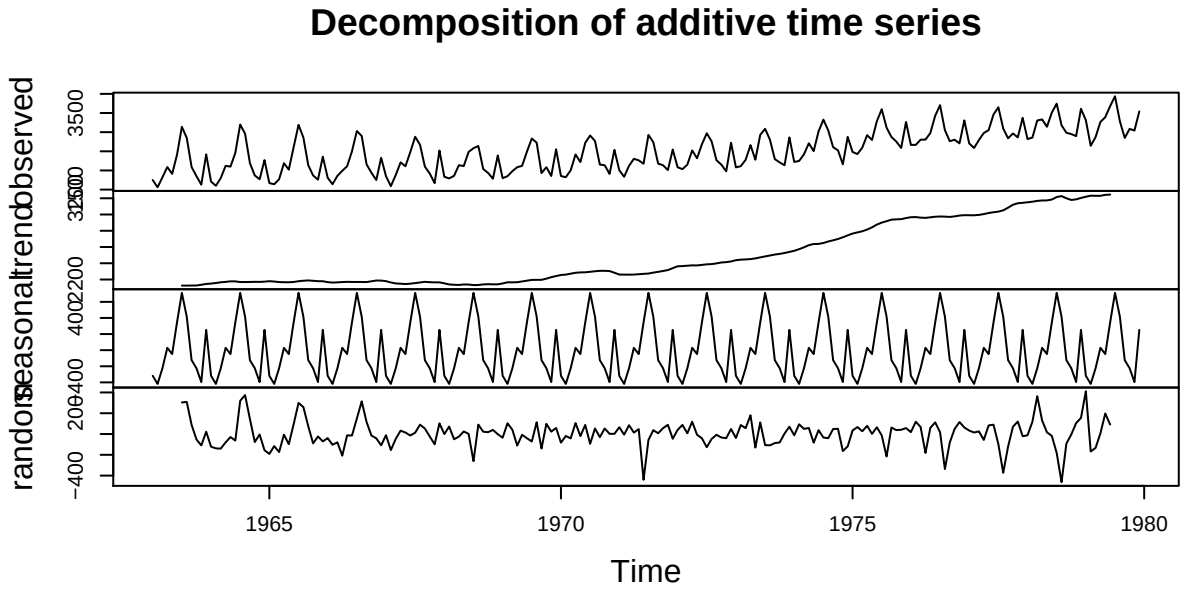
Sur le chronogramme, on observe un motif périodique, ce qui permet de supposer un effet saisonnier. On peut estimer la tendance par la méthode des moyennes mobiles. On prend un ordre égal à 12, car le motif saisonnier semble durer 12 mois.

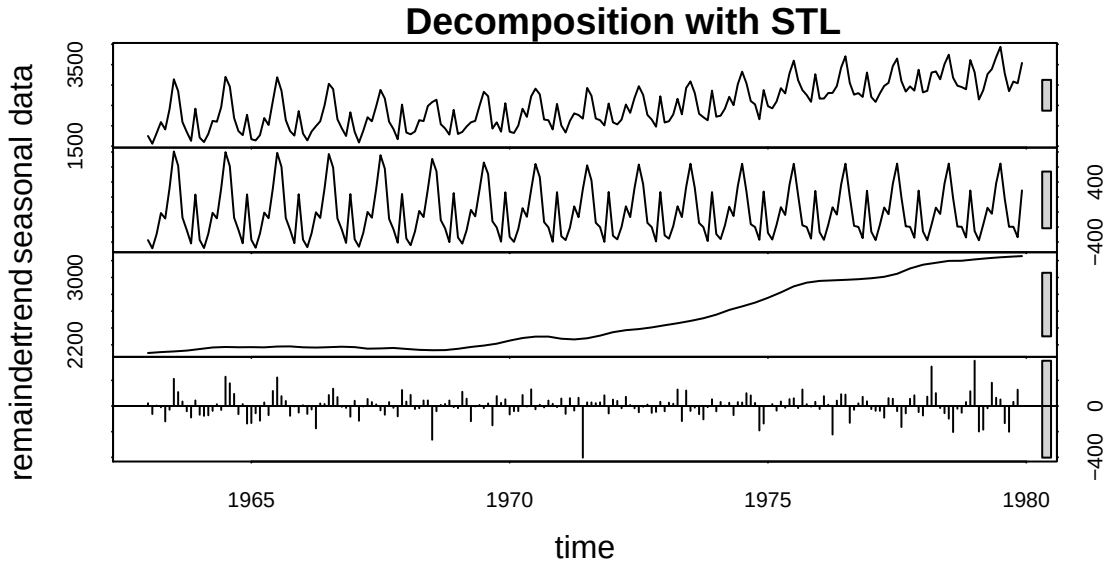
Sur le monthplot, les chronogrammes mensuels ne sont pas identiques d'un mois à l'autre. Cela confirme un **effet saisonnier** : en particulier, la saison estivale et les fêtes de Noël semble plus propices aux déplacements, comme on peut facilement le deviner.

En faisant l'hypothèse d'un modèle complètement additif ou multiplicatif, on peut essayer d'arbitrer en utilisant la méthode de la bande. Sur le chronogramme, les 2 courbes semblent parallèles sur les 15 dernières années, ce qui nous permet d'aller vers un **modèle additif**.

Nous avons également observé le **lag-plot** (non affiché ici). Il confirme l'effet saisonnier d'ordre 12.

On réalise alors une décomposition à l'aide de la fonction **decompose** en mode **additif**, ainsi qu'avec la fonction **stl**. Dans les 2 cas, on retrouve la tendance calculée précédemment par moyenne mobile, ainsi qu'un motif saisonnier évident.





REMARQUE : sur la composante saisonnière STL, l'amplitude du signal semble varier. Cette composante est peut-être plus compliquée à appréhender qu'il n'y parait.

2.2. Lissage exponentiel

On veut maintenant faire de la prédiction par lissage exponentiel. Pour cela, on utilise la fonction **ets**, en contraignant le modèle avec un code à 3 lettres:

```
lettre 1 : e : erreur    AMZ [     additive,multiplicative,auto]
lettre 2 : t : tendance NAMZ [none,additive,multiplicative,auto]
lettre 3 : s : saison   NAMZ [none,additive,multiplicative,auto]
```

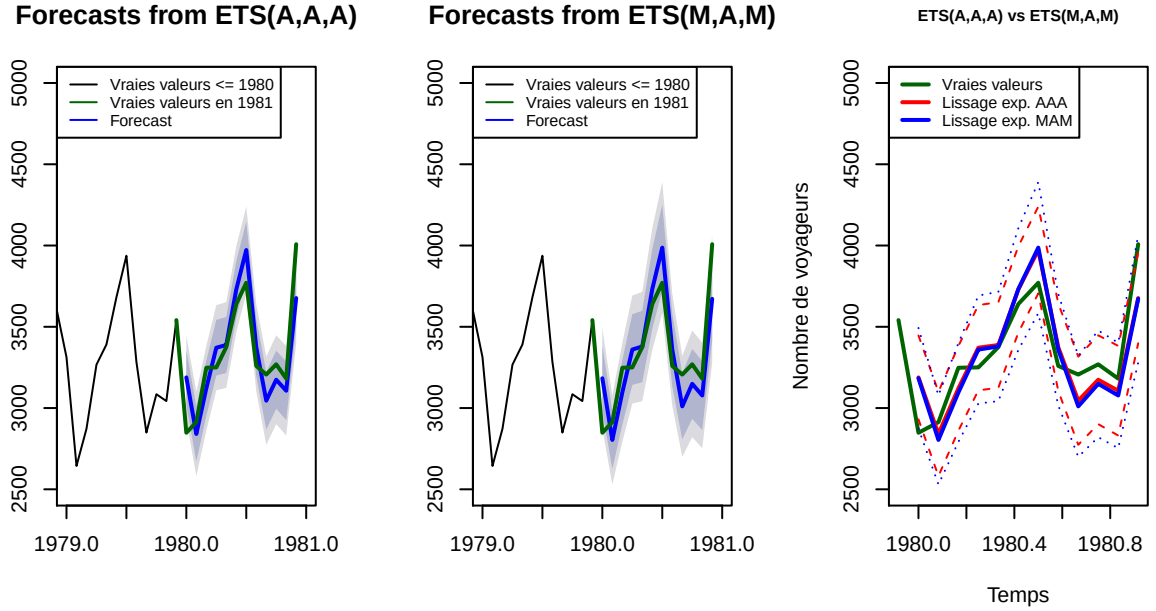
En premier lieu, on suit nos conclusions précédentes, et on contruit un modèle additif **AAA**. Ensuite, on observe les résultats obtenus en utilisant la méthode automatique **ZZZ**.

Consigne	Modèle retenu	AIC	AICc	BIC
additif	ETS(A,A,A)	3093.313	3096.603	3149.721
automatique	ETS(M,A,M)	3059.861	3063.152	3116.269

Ces résultats privilégient un modèle à tendance additive, mais à saisonnalité mutiplicative sans tendance: Cela confirme la remarque faite en 2.1 à propos de la variabilité de l'amplitude de la saisonnalité calculée dans un cadre additif.

L'affichage des prévisions ci-dessous montre des courbes très proches, et des ordres de grandeur très proches concernant les intervalles de confiance.

Concernant les intervalles de confiance des valeurs prédites (non affiché ici, mais visibles dans les plots ci-dessus), malgré le modèle retenu ETS(M,A,M) avec les critères AIC-AICc-BIC, il semble pourtant que c'est le modèle ETS(A,A,A) qui a les meilleurs intervalles de confiance (cf. troisième figure ci-dessous)



2.3. Modélisation

Analyse des résidus

En premier lieu, on différencie la série pour enlever une tendance et une saisonnalité. On veut ainsi se ramener à une série stationnaire.

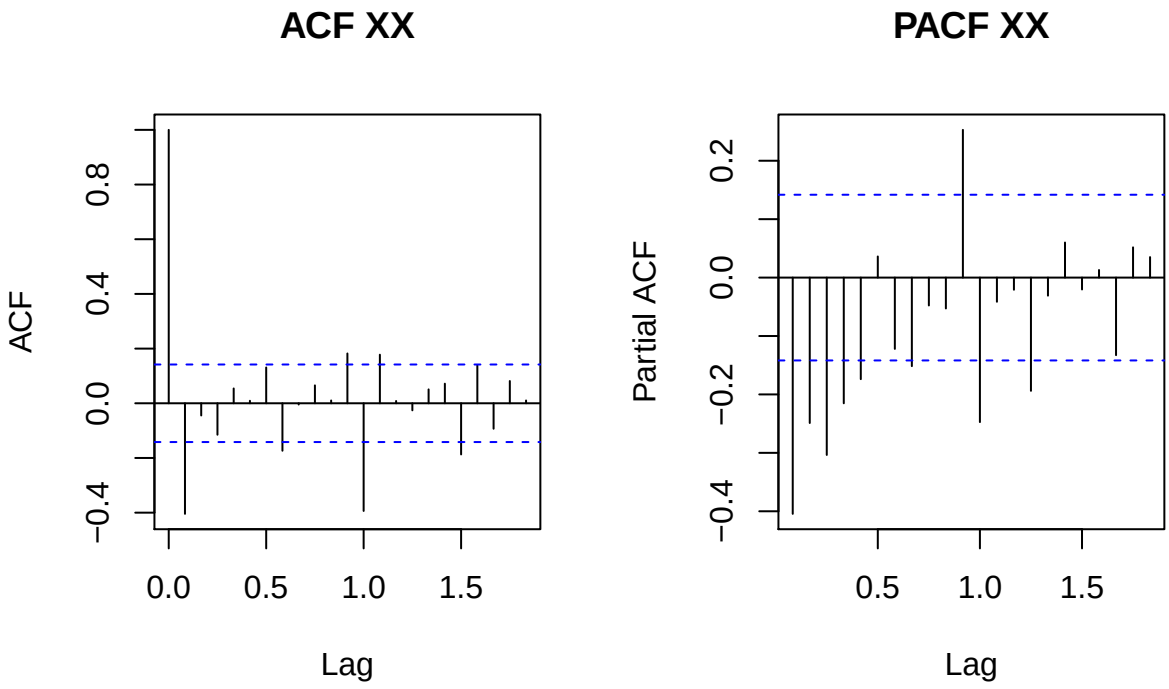
```
XX = diff(diff(X,lag=12,dfference=1),lag=1,dfference=1)
```

On essaye alors d'identifier des ordres de modélisation avec les ACF et PACF, avec les règles suivantes.

```
AR(p) :  ACF en exponentielle décroissante + PACF nulle en p+1
MA(q) : PACF en exponentielle décroissante +  ACF nulle en q+1
ARMA  : compliqué ...
```

On voit une exponentielle décroissante sur la PACF, et l'ACF est nulle à partie de q+1 = 2. On se dirige alors vers un modèle **MA(1)**.

Cependant, l'exponentielle décroissante de la PACF n'est pas non plus avérée, car on retrouve des pics aux positions 11 et 12 sur la PACF. On peut aussi essayer de voir une exponentielle décroissante de l'ACF, et une PACF nulle à partie de p+1 = 13. On se dirige alors vers un modèle **RA(12)**.



Modèle MA(1) : NON VALIDE

On crée le modèle **model_1 = Arima(XX,order=c(0,0,1))** et on analyse ses métriques. Le résidu n'est un bruit blanc.

```
Critères    [Arima]     AIC=2427.02   AICc=2427.15   BIC=2436.78
Box-Pierce  [Box.test]  NOK  (p-value = 0.01957 < .05)
Colinéarité [cor.arma]  OK  (pas de coefficient supérieur à 0.9)
```

Modèle RA(12) : VALIDE

On crée le modèle **model_2 = Arima(XX,order=c(12,0,0))** et on analyse ses métriques. Le résidu est bien un bruit blanc, et il n'y a pas de colinéarités.

```
Critères    [Arima]     AIC=2409.49   AICc=2411.88   BIC=2455.02
Box-Pierce  [Box.test]  OK  (p-value = 0.7465 > .05)
Colinéarité [cor.arma]  OK  (pas de coefficient supérieur à 0.9)
```

Modèle automatique : ARIMA(1,0,1)(0,0,1)[12] with zero mean : VALIDE

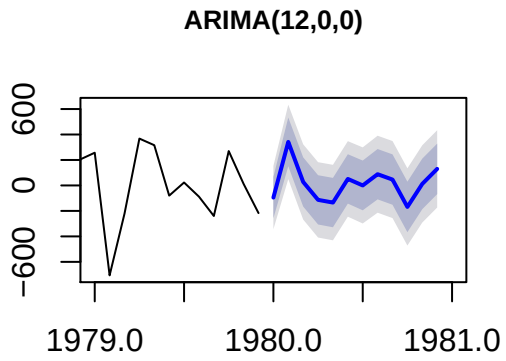
On crée le modèle automatique **model_3 = auto.arima(XX)** et on analyse ses métriques. Le résidu est bien un bruit blanc, et il n'y a pas de colinéarités.

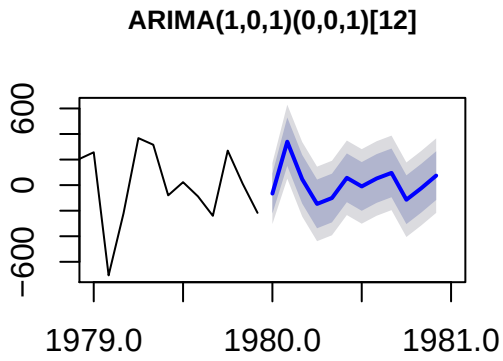
```
Critères    [Arima]     AIC=2386   AICc=2386.21   BIC=2399.01
Box-Pierce  [Box.test]  OK  (p-value = 0.9475 > .05)
Colinéarité [cor.arma]  OK  (pas de coefficient supérieur à 0.9)
```

Notre deuxième modèle **RA(12)** fonctionne correctement pour modéliser notre résidu. Le modèle automatique **ARIMA(1,0,1)(0,0,1)[12]** a de meilleures performances, mais il est plus compliqué à intuiter.

Prévisions des résidus

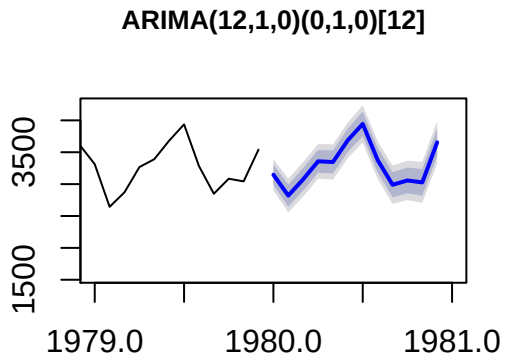
Les prévisions sont très proches. Idem pour les intervalles de confiance. Ce sera encore plus flagrant quand on aura rajouté la tendance et la saisonnalité de la série.

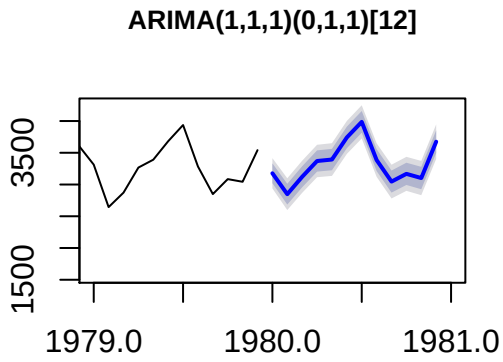




Prévisions des séries

Comme prévu les 2 prévisions sont très proches.





Comparaison globale

On veut comparer les vraies valeurs, le lissage exponentielle, et la prévision par modélisation des résidus. Les 2 prévisions sont assez proches l'une de l'autre, et leurs intervalles de confiance contiennent les vraies valeurs. Cependant, la prévision SARIMA a un meilleur intervalle de confiance.