Compte rendu de TP

Executive Master Statistique et Big Data - Cours de Séries temporelles

Florian HEGWEIN (21805361)

5 Janvier 2020

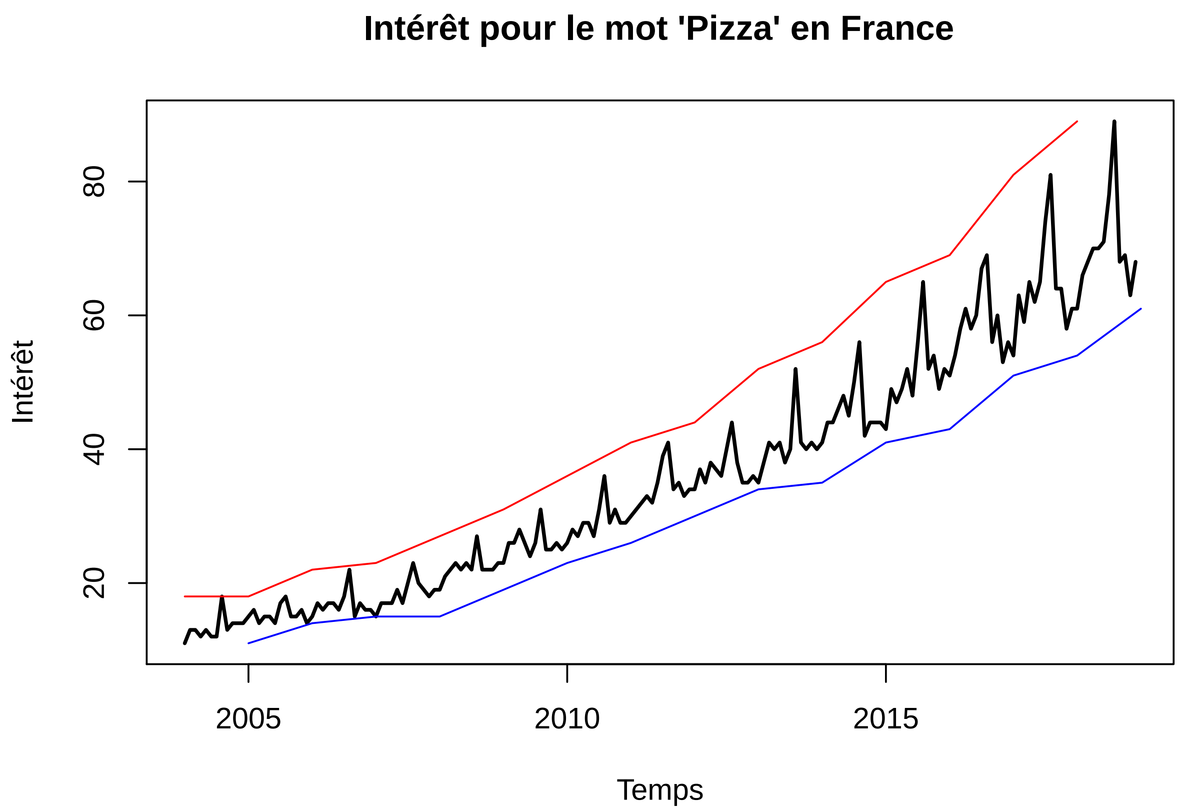
Contents

1. Intérêt pour le mot "Pizza" en France

La série temporelle `Pizza` montre l'évolution de l'intérêt pour le mot "Pizza" en France entre janvier 2004 et novembre 2019. Les données sont téléchargeables sur https://trends.google.fr/trends/. Les valeurs des mois de l'année 2019 ont été écartées de la modélisation afin de pouvoir comparer les prédictions des modélisations avec les vraies valeurs de la série.

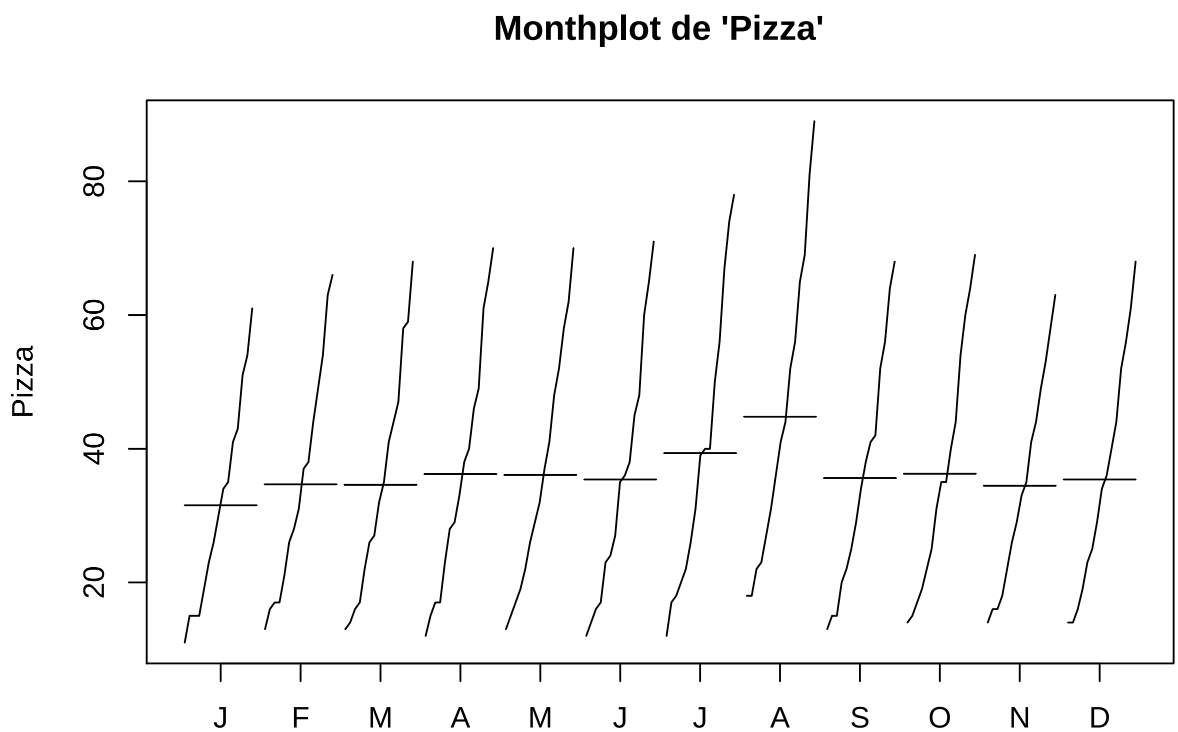
1.1 Analyse exploratoire de la série

1.1.1 Chronogramme



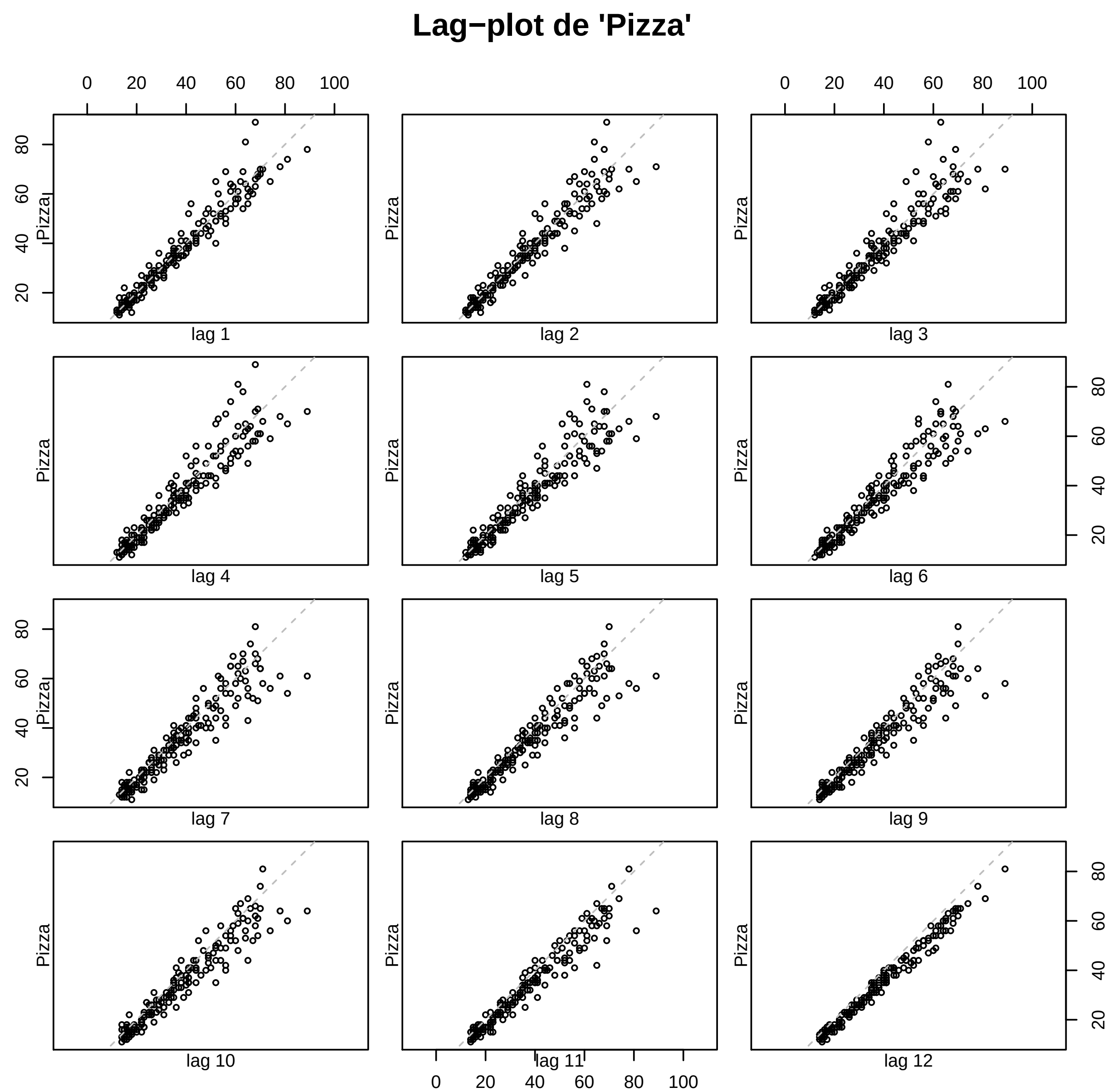
Le chronogramme montre que cette série temporelle a clairement une tendance croissante et probablement une saisonnalité. Les courbes rouge et bleue représentent le minimum et le maximum d'une année entière. Ces deux courbes n'étant pas parallèles elles indiquent plutôt un modèle multiplicatif qu'additif.

1.1.2 Month-plot



Le month-plot de `Pizza` confirme la saisonnalité de la série. L'intérêt pour le mot Pizza en France s'accroit légèrement en juillet jusqu'à atteindre un pic pendant le mois d'août (vacances d'été en France). En septembre l'intérêt semble retomber assez rapidement sur son niveau habituel qu'il atteint aussi pendant le reste de l'année. La variance n'est pas stable, elle monte légèrement tous les mois jusqu'à juin, puis augmente en juillet et août afin de retomber en septembre sur le niveau entre janvier et juin.

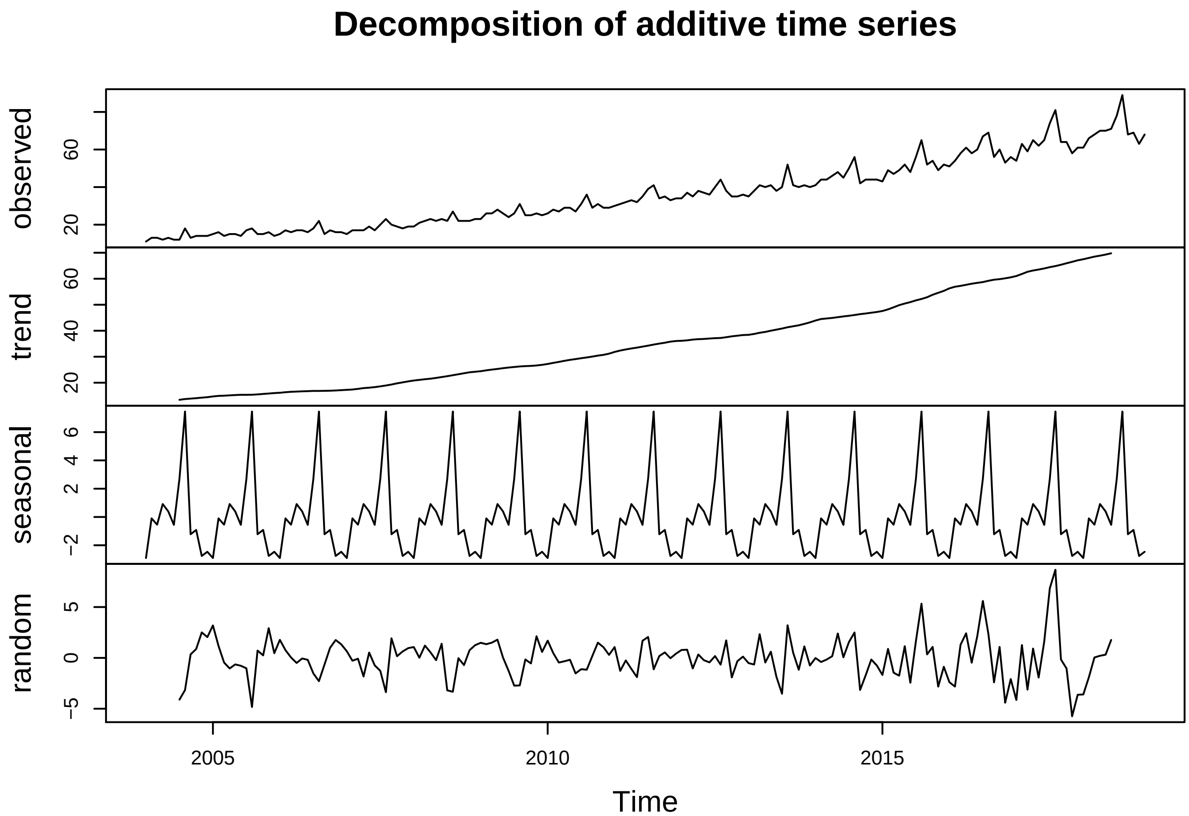
1.1.3 Lag-plot



Le lag-plot de `Pizza` montre que la série présente une autocorrélation assez forte de manière générale mais surtout d'ordre 12. C'est-à-dire que la série dépend beaucoup de son passé et surtout de ce qui s'est passé 12 mois avant.

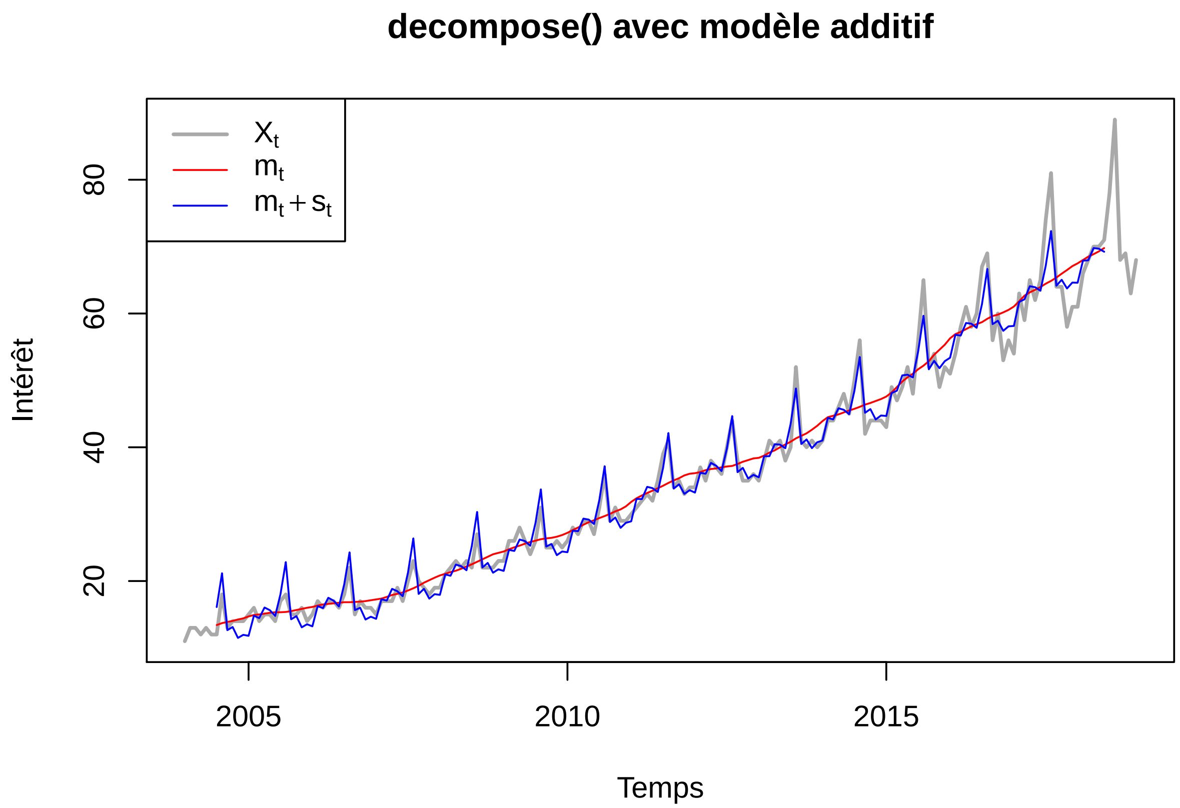
1.1.4 Décomposition

Modèle additif L'analyse du chronogramme laisse soupçonner un modèle multiplicatif mais il est quand même intéressant de comparer le modèle additif et multiplicatif. Voici la décomposition d'un modèle additif :



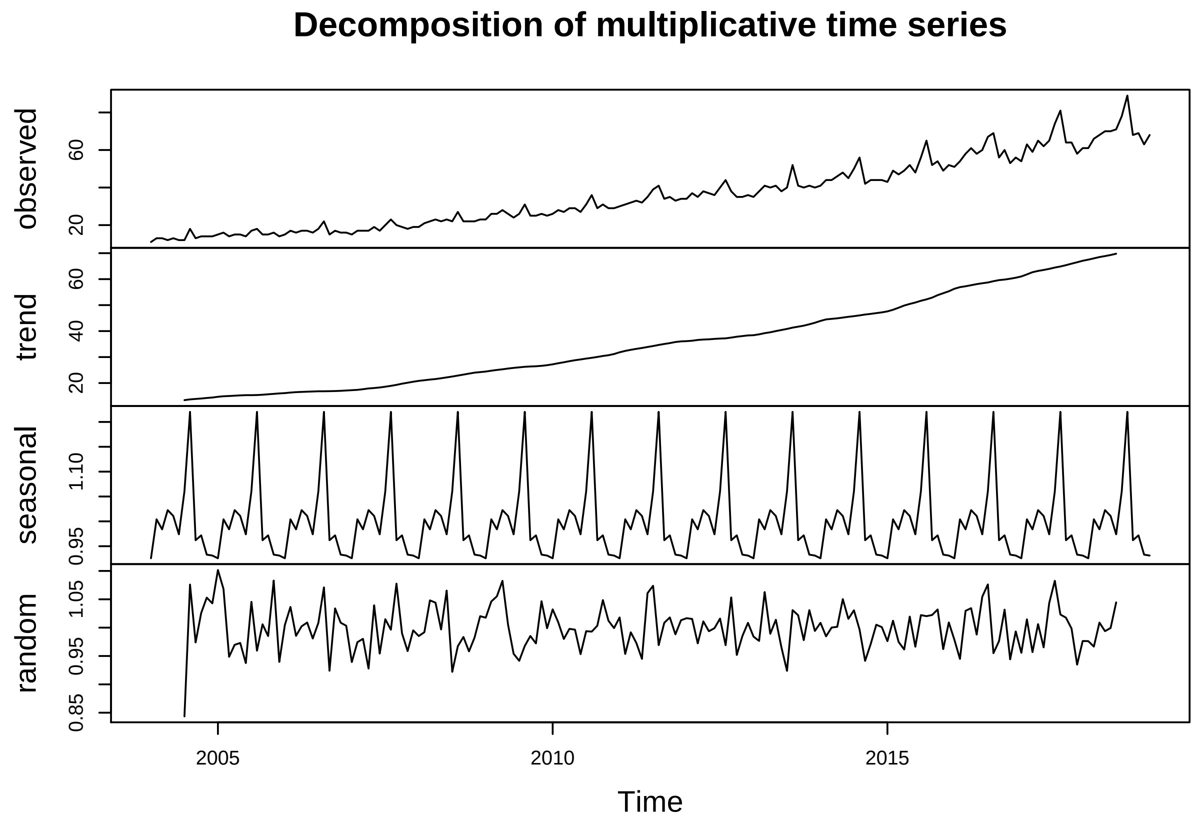
La tendance de la série ainsi que sa saisonnalité sont bien comme attendues. La série a une tendance croissante et une saisonnalité avec un pic pendant la période estivale en août. Or, il semble que la variance des résidus est à la fois assez grande et n'est pas stable. La variance monte avec le temps. Le modèle est donc hétéroscédastique et il reste de la variance à expliquer.

Le graphique suivant représente la série originale (grise) ainsi que la simulation du modèle additif avec sa tendance (rouge) et sa tendance plus la saisonnalité additif (bleu).

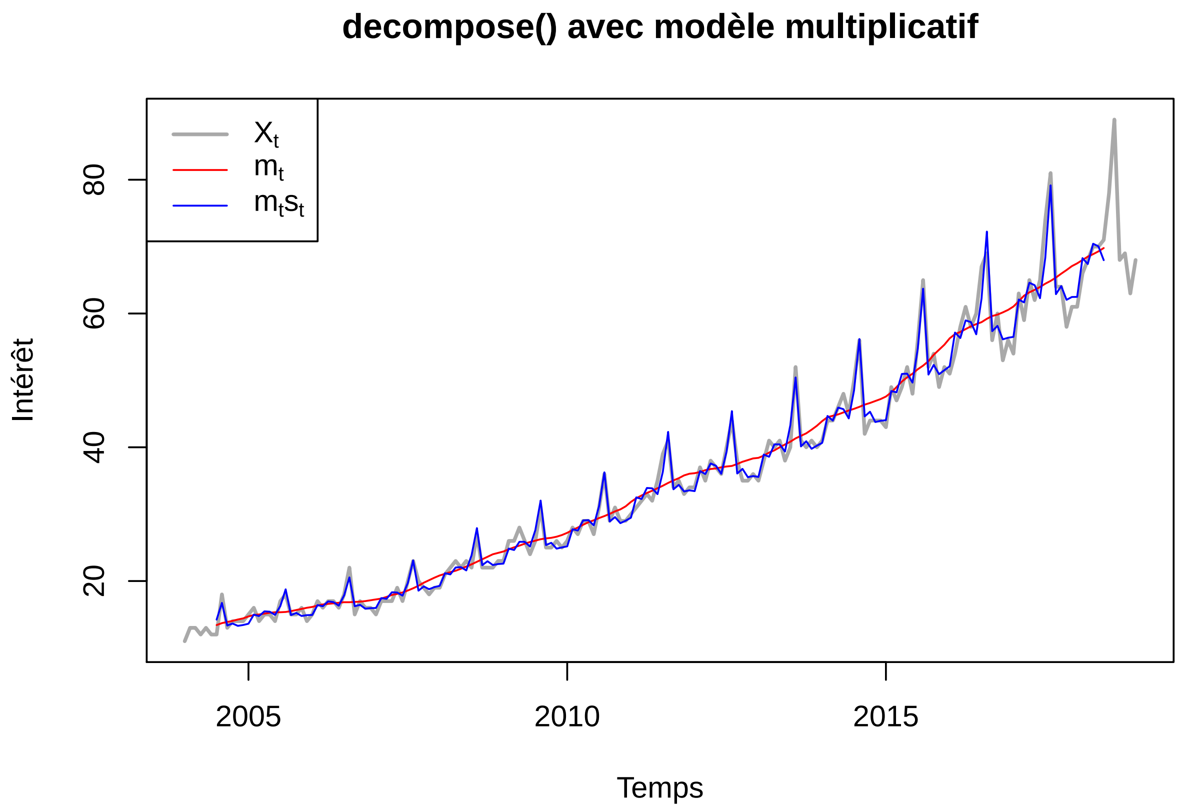


Entre 2004 et 2009 les pics du mois d'août sont surestimés tandis que les creux après sont sous-estimés. Entre 2010 et 2012 les pics et creux sont plutôt bien estimés. A partir de 2014 c'est l'inverse, les pics du mois d'août sont sous-estimés et les creux après surestimés. Surtout à partir de 2017 le modèle devient très imprécis. Cette analyse laisse penser qu'un modèle multiplicatif serait mieux pour la série temporelle `Pizza`.

Modèle multiplicatif



La tendance de la série ainsi que sa saisonnalité sont toujours comme attendues et comparables au modèle additif. Cependant, la variance du modèle multiplicatif est beaucoup plus petite et semble être stable. Le modèle est donc homoscédastique, la variance est mieux expliquée. La simulation du modèle confirme qu'un modèle multiplicatif est mieux :



La série est bien modélisée de manière générale avec un modèle multiplicatif. Il reste une particularité en 2009 qui n'est évidemment pas représentée dans le modèle mais qui n'aurait quasiment pas d'influence sur la prédiction plus tard. A partir de 2013 les pics du mois d'août sont toujours sous-estimés, mais cette sous-estimation est beaucoup plus petite que pour le modèle additif. Les creux sont généralement bien représentés.

1.2 Modélisation par lissage exponentiel

L'analyse précédente a montré que la série temporelle `Pizza` est une série avec une tendance croissante et une composante saisonnière. Il faut donc utiliser la méthode de Holt-Winters à trois paramètres (α, β et γ - erreur, tendance et saisonnalité). La saisonnalité étant plutôt positive, c'est ce groupe de modèles qui va probablement avoir les meilleurs résultats. Il se pose la question de la tendance, plutôt additive ou multiplicative ? C'est pour cette raison que l'on regarde de plus près les modèles avec tendance additive et multiplicative.

Les modèles à tester :

Modèle	Erreur	Tendance	Saisonnalité
Holt-Winters M, A, M	multiplicative	additive	multiplicative
Holt-Winters M, M, M	multiplicative	multiplicative	multiplicative

1.2.1 Méthode de Holt-Winters

La fonction `ets()` du package `forecast` permet de fitter les modèles Holt-Winters. Voici les sorties R des `summary()` :

```
## ETS(M,Ad,M)
##
## Call:
##  ets(y = Pizza, model = "MAM")
##
##   Smoothing parameters:
##     alpha = 0.3676
##     beta  = 0.0224
##     gamma = 1e-04
##     phi   = 0.9791
##
##   Initial states:
##     l = 12.3376
##     b = 0.2544
##     s = 0.9295 0.9248 0.9693 0.9639 1.2227 1.0672
##            0.9703 1.0081 1.0216 0.9922 1.0052 0.9254
##
##   sigma:  0.0497
##
##      AIC     AICc      BIC
## 1113.365 1117.613 1170.838
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 0.2452329 1.650737 1.197559 0.4752753 3.532196 0.2949998
##                       ACF1
## Training set -0.0005446063
```

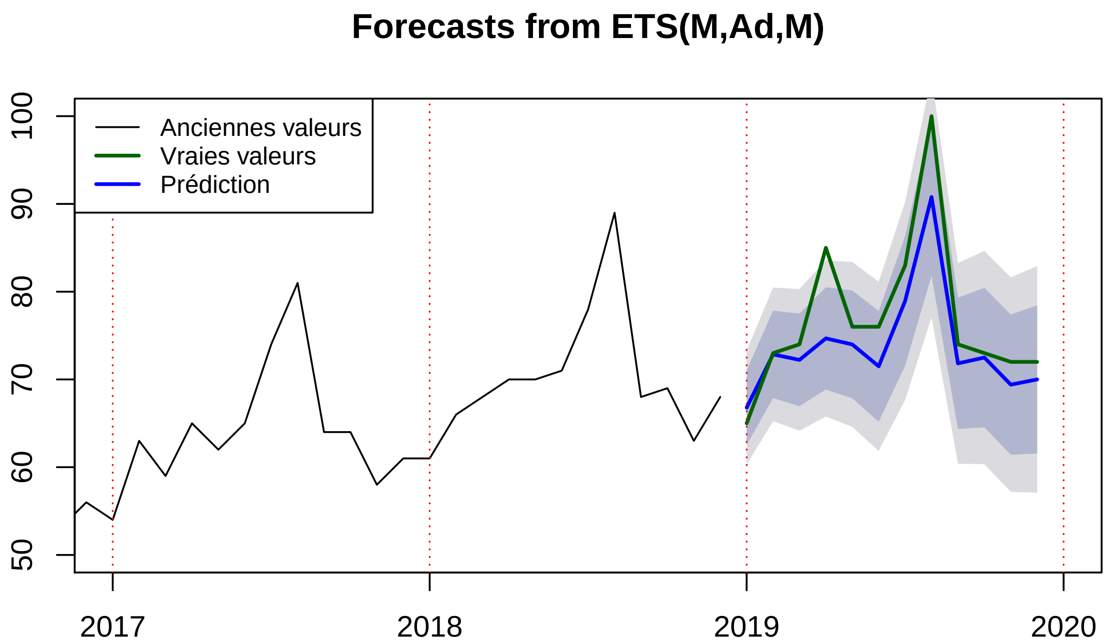
```
## ETS(M,Md,M)
##
## Call:
##  ets(y = Pizza, model = "MMM")
##
##   Smoothing parameters:
##     alpha = 0.3499
```

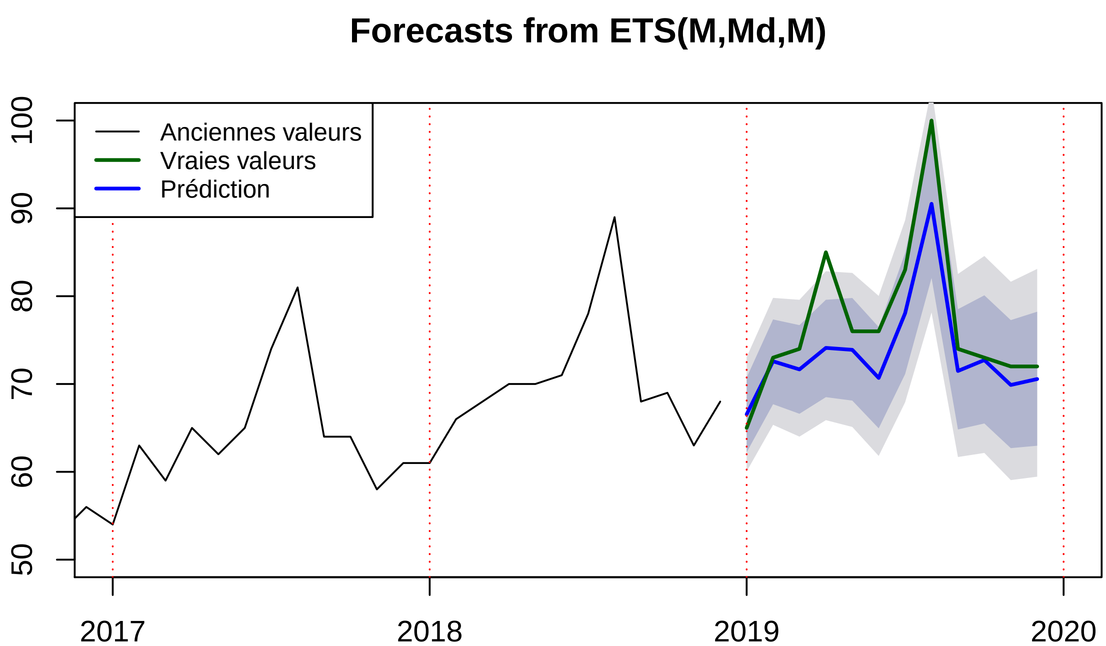
```
##     beta  = 0.0138
##     gamma = 1e-04
##     phi   = 0.98
##
##   Initial states:
##     l = 12.1332
##     b = 1.0161
##     s = 0.9385 0.933 0.9746 0.9616 1.2225 1.0586
##            0.9626 1.0104 1.0178 0.9885 1.0057 0.9262
##
##   sigma:  0.0494
##
##      AIC     AICc      BIC
## 1111.655 1115.903 1169.128
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 0.2343852 1.711903 1.228203 0.4578962 3.567291 0.3025484
##                    ACF1
## Training set 0.05425039
```

On voit que les valeurs de l'AIC, AICc et du BIC sont légèrement plus petites pour le modèle avec tendance multiplicative (MMM). De même pour l'écart type. La valeur du paramètre α est plus élevée pour le modèle MMM tandis que les valeurs des paramètres β et γ sont moins élevées. Cela signifie que quant à l'erreur (paramètre α) le passé récent est plus important pour le modèle MMM que pour le modèle avec tendance additive (MAM). En ce qui concerne la tendance (paramètre β) et la saisonnalité (paramètre γ) le passé récent est plus important pour le modèle MAM.

1.2.2 Prédiction Holt-Winters

La fonction `forecast()` du package `forcast` permet de faire des prédictions sur les modèles fittés précédemment. Il est ensuite intéressant de comparer les prédictions des deux modèles MAM et MMM avec les vraies valeurs de l'année 2019.



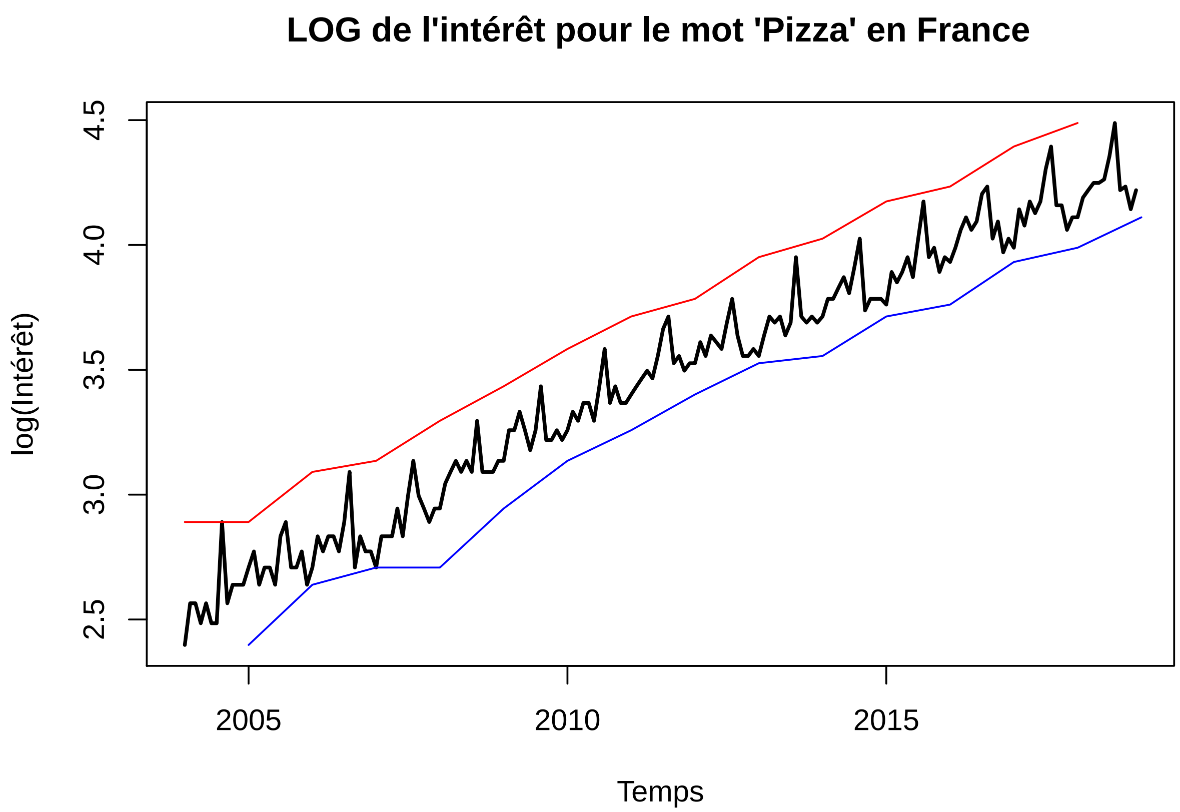


Les deux prédictions se ressemblent beaucoup. Les vraies valeurs montrent un intérêt particulièrement élevé pour le mot ‘Pizza’ en avril 2019. Cette particularité n’existe pas dans le passé de la série et n’est donc pas bien prédit par les deux modèles. Quant au modèle MMM cette valeur se trouve à l’extrémité de l’intervalle de confiance à 95% (gris) et elle est en dehors de cet intervalle pour le modèle MAM. Le pic des vacances d’été est légèrement mieux prédit par le modèle MMM. On observe un intérêt plus bas en septembre 2019 qui n’est pas bien prédit par les deux modèles (un peu mieux pour le modèle MAM, en dehors de l’intervalle de confiance à 80% pour le modèle MMM).

1.2.3 Passage au LOG

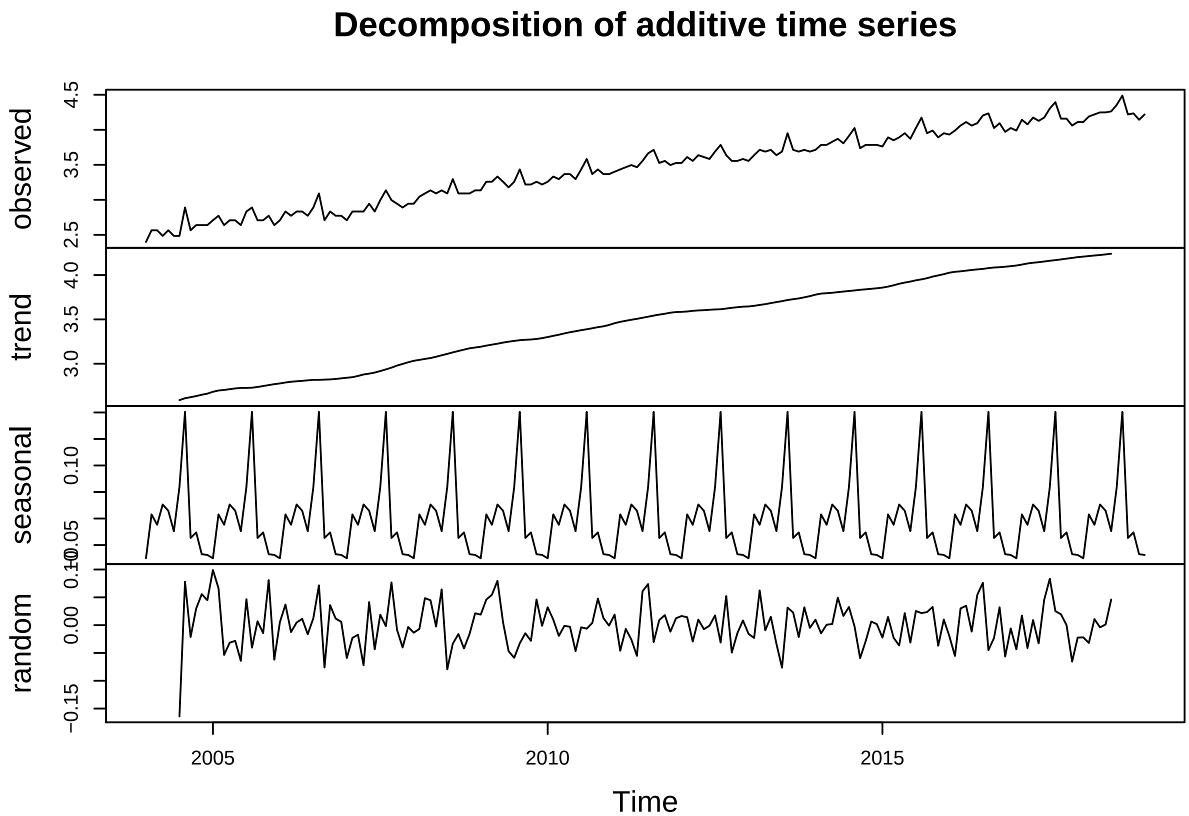
On peut essayer de transformer la série `Pizza` au LOG. Ce passage au log peut linéariser la tendance et stabiliser la variance, ce qui permettrait probablement un lissage exponentiel avec un modèle à tendance et saisonnalité additives ?

Chronogramme

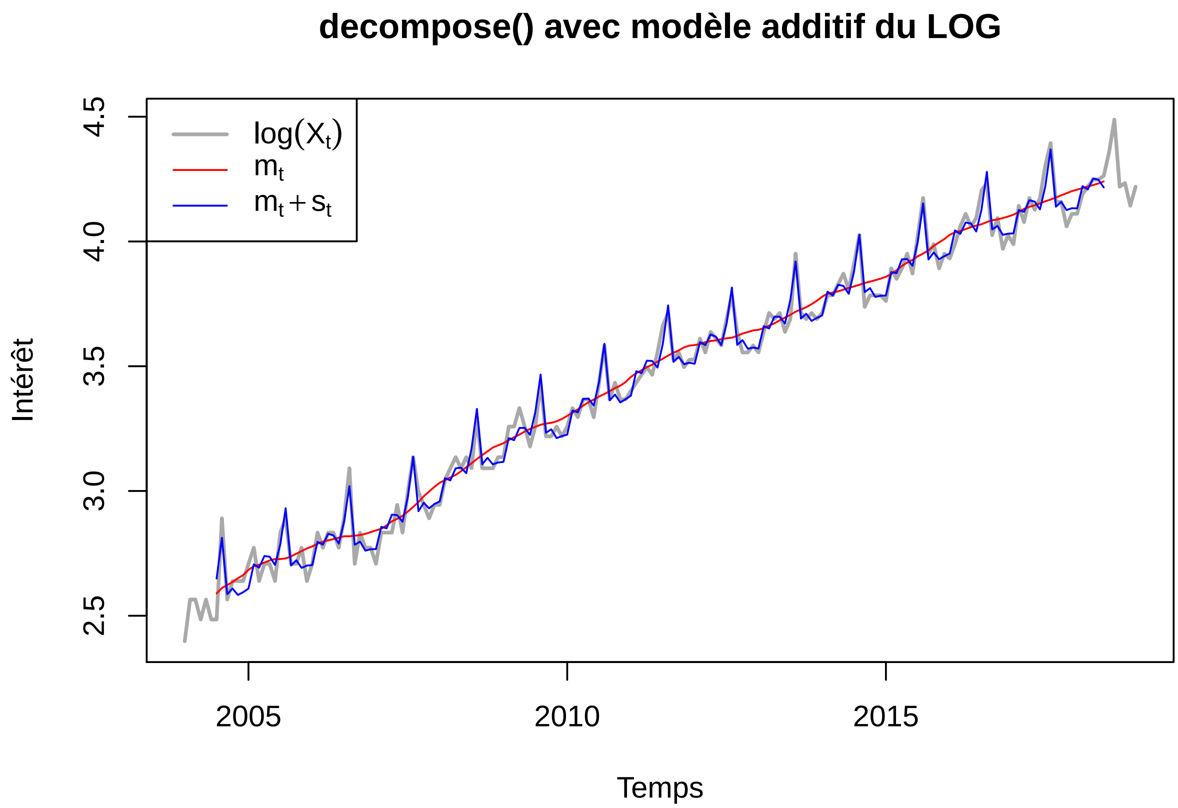


La tendance paraît plus linéaire. La saisonnalité semble plutôt additive, visible par les courbes bleu et rouge assez parallèles.

Décomposition additive



La tendance de la série est presque linéaire, la saisonnalité semble à ceux des modèles non transformés. La variance semble être stable, le modèle est homoscédastique.



Les pics et les creux sont généralement bien représentés. L'analyse laisse penser que la série transformée peut être modélisé par un modèle Holt-Winters à tendance et saisonnalité additives (AAA).

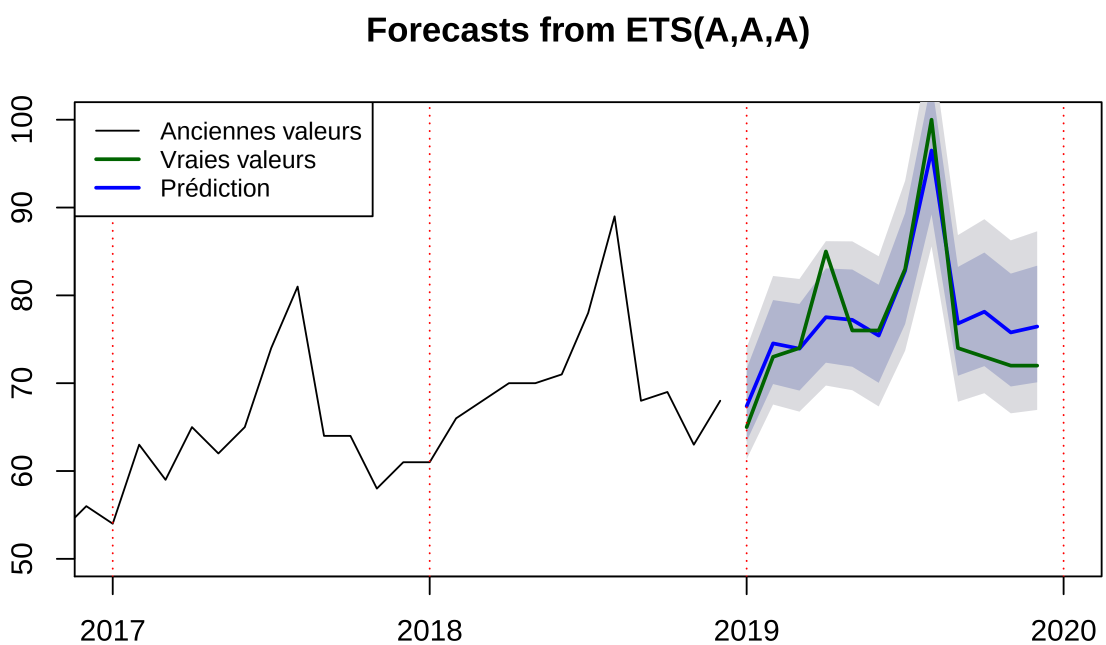
Lissage exponentiel

```
## ETS(A,A,A)
##
## Call:
##  ets(y = log(Pizza), model = "AAA")
##
##   Smoothing parameters:
##     alpha = 0.3002
##     beta  = 1e-04
##     gamma = 1e-04
##
##   Initial states:
##     l = 2.4852
##     b = 0.0103
##     s = -0.0696 -0.0682 -0.0271 -0.0343 0.2044 0.0618
##            -0.0213 0.0124 0.0267 -0.0104 0.008 -0.0825
##
##   sigma:  0.0479
##
##        AIC       AICc        BIC
## -141.84569 -138.06791  -87.56542
##
```

```
## Training set error measures:
##                        ME       RMSE        MAE        MPE     MAPE
## Training set -0.001196395 0.04573284 0.03576876 -0.0347736 1.095762
##                   MASE        ACF1
## Training set 0.3001371 -0.02068036
```

La valeur du paramètre α de l'erreur se situe dans la même échelle que pour les modèles non-transformés. Cependant, les valeurs des paramètres β et γ sont beaucoup moins élevées. C'est-à-dire que le passé lointain est plus important dans la mise à jour de la tendance et de la saisonnalité (ce qui paraît logique vu que la tendance s'est linéarisée et la saisonnalité s'est stabilisé lors du passage au LOG).

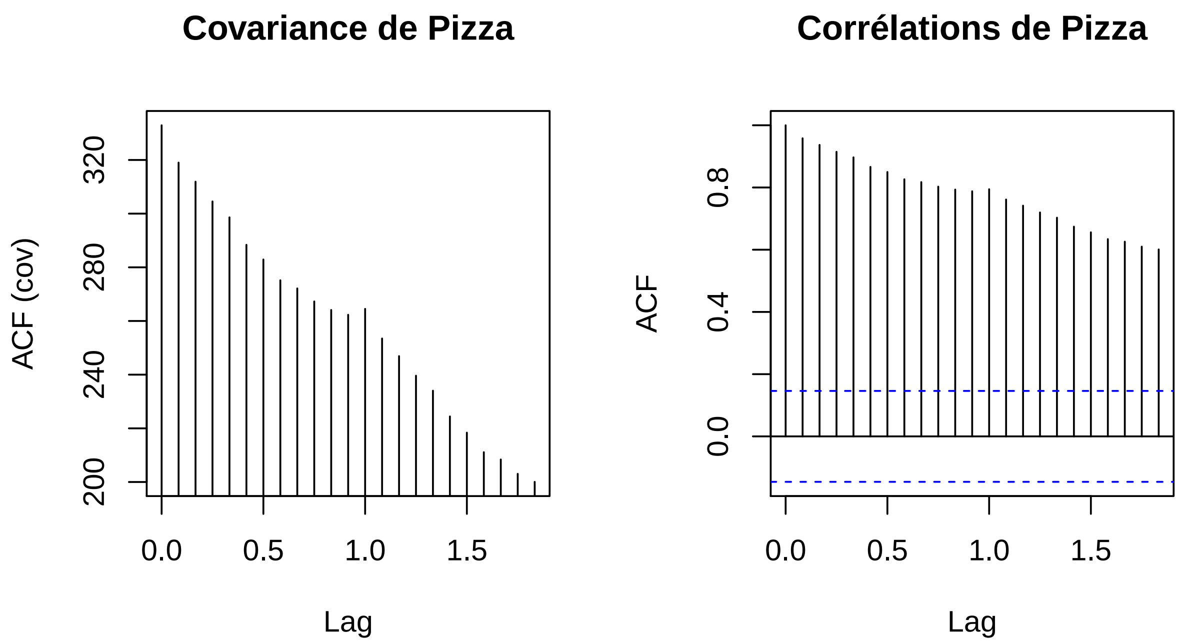
Prédiction Les valeurs de la prédiction, toujours avec la fonction `forecast()` du package `forecast`, vont être retransformées afin de pouvoir les comparer plus facilement aux lissages exponentielles non transformées.



La prédiction est comparable à celle du modèle MMM non transformé. En effet, la particularité d'avril 2019 est à la limite de l'intervalle de confiance à 95% et l'intérêt plus bas de septembre 2019 est en dehors de l'intervalle de confiance à 80%, tout comme pour le modèle MMM non transformé. Le pic des vacances d'été est mieux représenté que pour le modèle MAM non transformé. On peut donc conclure que la transformation au LOG de la série `Pizza` n'apporte que très peu, voire aucune amélioration à la prédiction (au moins pour l'année 2019). Il n'y a donc pour le moment pas d'utilité de transformer la série. Cependant, avec d'autres valeurs disponibles dans le futur, il pourrait être intéressant de tester de nouveau le passage au LOG.

1.3 Modélisation

1.3.1 Analyse des fonctions d'autocorrélation

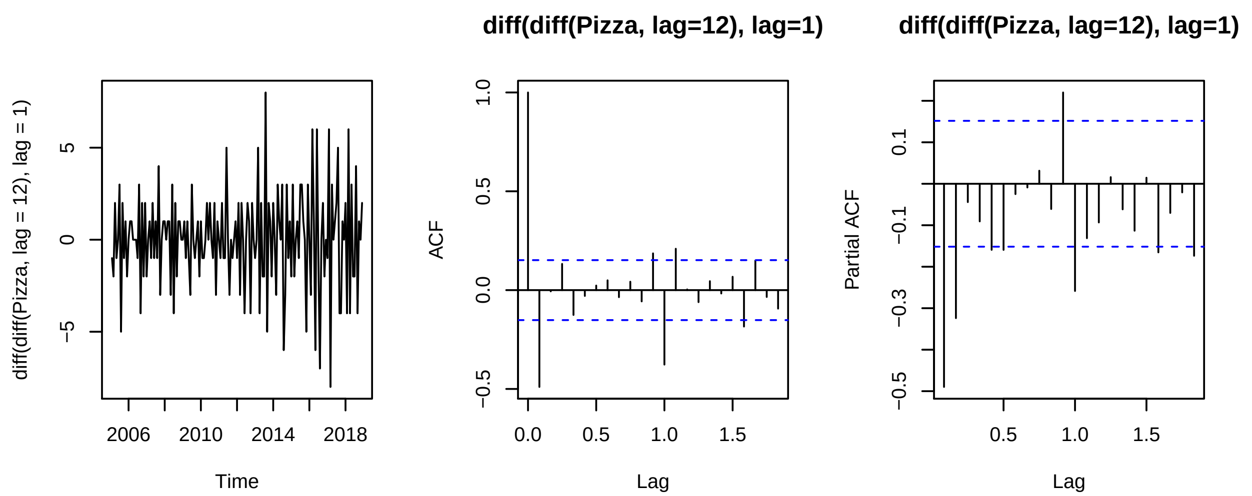


La valeur de $\rho x(h)$ varie, mais est positive et maximale en correspondance avec certaines valeurs de h. Ce résultat de l'analyse des fonctions d'autocorrélation n'est pas surprenant après ce que l'on sait déjà de la série temporelle `Pizza`. Les valeurs d'un mois d'une année sont fortement corrélées à celles des mêmes mois des années précédentes. Il y a une variation au cours de l'année et cette variation est récurrente sur chaque année.

1.3.2 Différenciation de la série

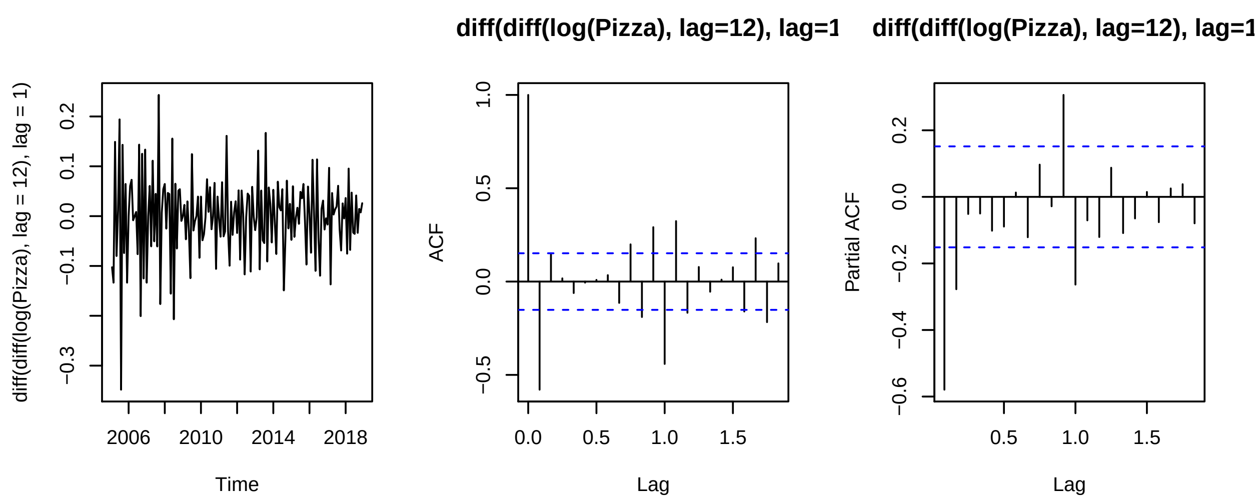
Vu la tendance et la saisonnalité de la série, il faudrait la différencier afin de la ramener à un processus stationnaire à moyenne zéro et variance stable. On enlève donc la saisonnalité d'ordre 12 (`lag=12`) et la tendance pour la série originale et la série transformée au LOG. En effet, il a déjà été montré lors du lissage exponentiel que le passage au LOG stabilise la variance.

Série originale



La fonction d'autocorrélation (ACF) est 0 à l'ordre 2 et la fonction d'autocorrélation partielle (PACF) décroît exponentiellement. L'analyse indique donc un modèle MA(1). Or, la série restante après différenciation montre une variance croissante, il ne s'agit donc pas d'un processus stationnaire. On teste donc le passage au LOG.

LOG de la série



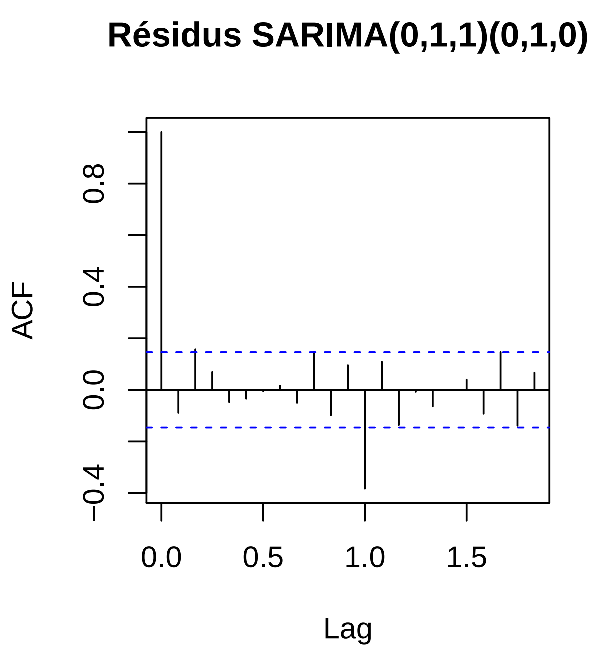
La variance s'est stabilisée avec le passage au LOG. L'analyse de l'ACF et de la PACF est la même que pour la série non transformée, elle indique plutôt un modèle MA(1).

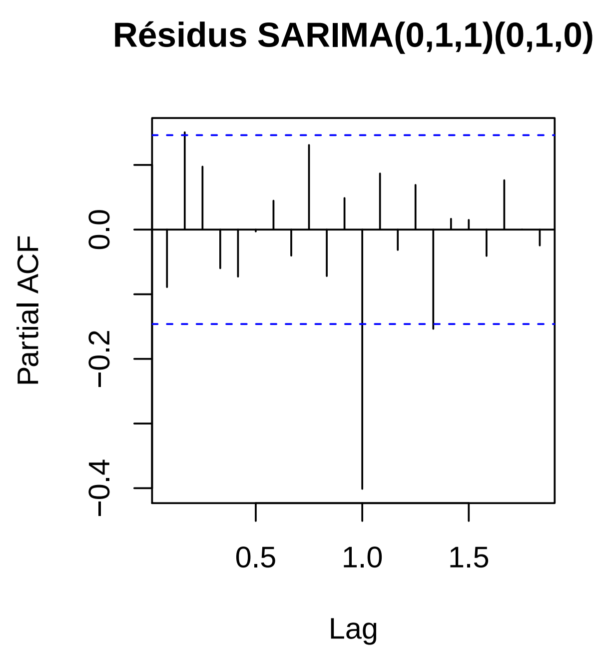
1.3.3 Modélisation SARIMA

Il a été montré que le passage au LOG ainsi que la différenciation (élimination de saisonnalité et tendance) rapproche la série d'un processus stationnaire. De ce fait, on peut envisager un modèle SARIMA de la série transformé au LOG. En tant que paramètres de base on peut se servir des études précédentes : l'analyse des fonctions d'autocorrélation indique un modèle MA(1) ce qui implique des valeurs de `p=0` et `q=1`. Pour enlever la tendance (différenciation de 1) il faudrait donc choisir `d=1`. En ce qui concerne la saisonnalité, il faudrait également choisir `D=1` pour une différenciation de `lag=1` et une fréquence de 12 pour une saisonnalité de 12 mois. Par défaut la fonction `Arima()` du package `forecast` prend en compte la fréquence de la série utilisée, il n'y a donc rien à préciser.

SARIMA(0,1,1)(0,1,0)

```
## Series: log(Pizza)
## ARIMA(0,1,1)(0,1,0)[12]
##
## Coefficients:
##           ma1
##       -0.7095
## s.e.   0.0566
##
## sigma^2 estimated as 0.004067:  log likelihood=222.85
## AIC=-441.7   AICc=-441.62   BIC=-435.46
##
## Training set error measures:
##                         ME       RMSE        MAE        MPE     MAPE
## Training set -0.003281669 0.06124402 0.04533657 -0.1236406 1.356729
##                   MASE        ACF1
## Training set 0.380421 -0.08892351
```





La PACF montre qu'il y a encore une saisonnalité non expliquée par le modèle. On teste la blancheur des résidus :

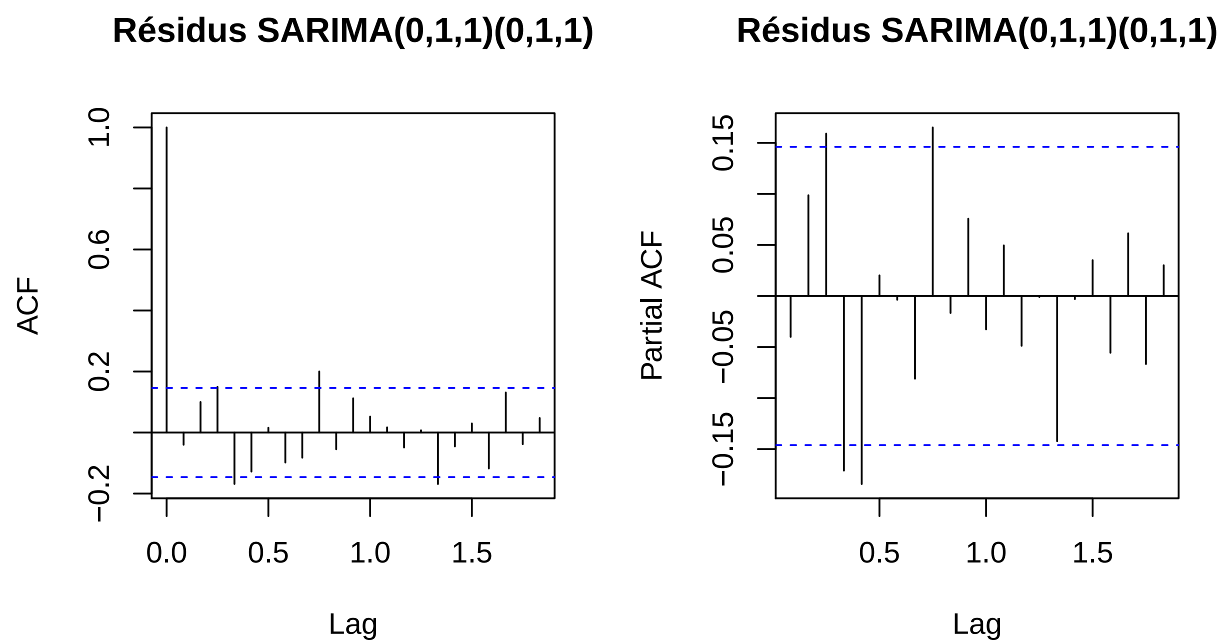
```
##
##  Box-Pierce test
##
## data:  PizzaSARIMAlog$residuals
## X-squared = 79.098, df = 45, p-value = 0.001267
```

Généralement on rejette l'hypothèse H_0, tous les $\rho_X(k) = 0$, du test de blancheur des résidus avec une p-valeur en dessous de 5%. Ici, la blancheur des résidus est rejetée, la p-valeur du test est faible. Le test a été fait sur les 45 premier $\rho_X(h)$ car de manière générale il faudrait choisir k dans l'ordre de $k = n/4$ ce qui donne 45 (la série ayant une longueur de $n = 180$).

On pourrait envisager de changer la valeur du paramètre `Q=1`.

SARIMA(0,1,1)(0,1,1)

```
## Series: log(Pizza)
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##           ma1     sma1
##       -0.6996  -0.8014
## s.e.   0.0623   0.0620
##
## sigma^2 estimated as 0.002477:  log likelihood=258.6
## AIC=-511.21   AICc=-511.06   BIC=-501.85
##
## Training set error measures:
##                        ME       RMSE        MAE       MPE     MAPE
## Training set -0.004879617 0.04765213 0.03486628 -0.163396 1.036968
##                   MASE       ACF1
## Training set 0.2925644 -0.0400765
```



La fonction d'autocorrélation partielle montre toujours une petite saisonnalité. Le test de blancheur des résidus va indiquer s'il faut rejeter le modèle ou pas :

```
##
##  Box-Pierce test
##
## data:  PizzaSARIMAlog$residuals
## X-squared = 55.066, df = 45, p-value = 0.1446
```

Cette fois la blancheur des résidus n'est pas rejetée, la p-valeur est assez élevée. On regarde les t-statistiques afin d'identifier des coefficients non significatifs.

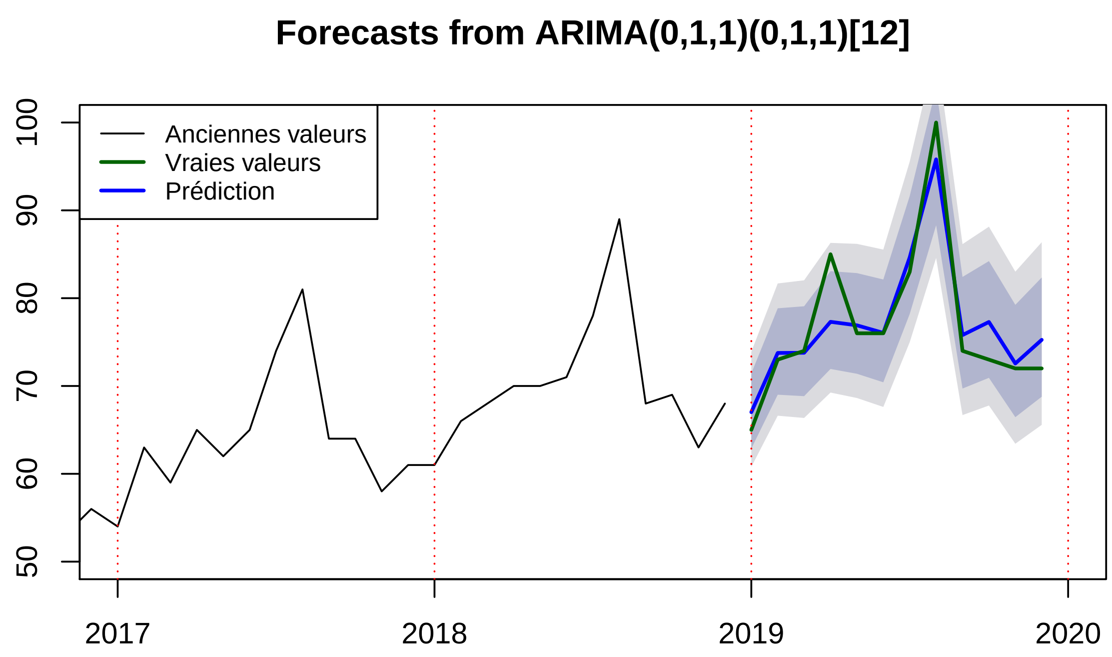
```
##              ma1      sma1
## t.stat -11.23525 -12.92506
## p.val    0.00000   0.00000
```

Tous les coefficients semblent être significatifs, leurs p-valeurs sont très faibles. Il reste la question de colinéarités des coefficients. On regarde les corrélations.

```
##              ma1        sma1
## ma1   1.00000000 -0.03758675
## sma1 -0.03758675  1.00000000
```

Il n'y a pas de colinéarité entre les coefficients du modèle. On retient donc le modèle SARIMA(0,1,1)(0,1,1) et fait une prédiction sur l'année 2019 que l'on compare ensuite avec les vraies valeurs.

1.3.4 Prédiction SARIMA



Le modèle prédit bien l'année 2019 de manière générale. Le pic du mois d'août est légèrement sous-estimé, tous comme par les modèle avec lissage exponentiel. Logiquement, il n'est pas mieux que les autres à prédire les mois particuliers d'avril 2019 et octobre 2019.

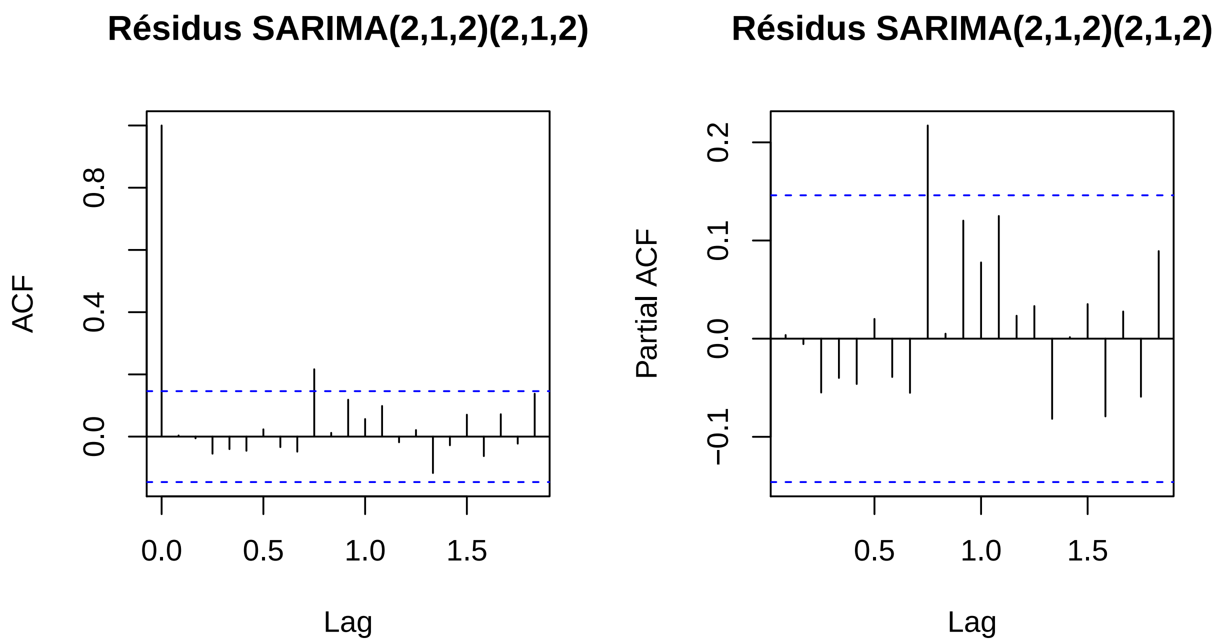
1.3.5 Modélisation automatique

Il est intéressant de regarder la modélisation automatique et la comparer avec le modèle trouvé manuellement. La fonction `auto.arima()` du package `forecast` fait un choix automatique des paramètres basé sur l'AIC, AICc ou BIC. Voici le résultat pour le LOG de la série `Pizza` basé sur l'AIC :

```
## Series: log(Pizza)
## ARIMA(5,0,1)(2,1,2)[12] with drift
##
## Coefficients:
##           ar1     ar2     ar3     ar4      ar5     ma1     sar1     sar2
##       -0.2533  0.4700  0.4868  0.0432  -0.1428  0.4859  -0.9517  -0.1250
## s.e.   0.2694  0.1004  0.1168  0.1104   0.0840  0.2634   0.3674   0.1147
##          sma1     sma2  drift
##        0.0931   -0.690  1e-02
## s.e.   0.3712    0.327  3e-04
##
## sigma^2 estimated as 0.002315:  log likelihood=269.57
## AIC=-515.13   AICc=-513.12   BIC=-477.65
##
## Training set error measures:
##                         ME       RMSE        MAE        MPE       MAPE
```

```
## Training set 0.001507722 0.04493962 0.03311924 0.05433288 0.9878849
##                    MASE        ACF1
## Training set 0.277905 0.003711733
```

La fonction `auto.arima()` choisit un modèle SARIMA(2,1,2)(2,1,2) avec drift. Il y a donc dix coefficients dans le modèle dont certains ont un écart-type assez élevé comme par exemple `sma1`, `sar1` et `sma2`. Il y a donc un risque de colinéarité des coefficients et de sur-ajustement. On note aussi que l'AIC avec une valeur de -515.13 est légèrement plus élevé que l'AIC du modèle SARIMA manuellement choisi qui s'élève à -511.21. C'est-à-dire que la modélisation automatique n'a pas pu trouver le modèle avec le plus petit AIC, même si c'était le critère de choix imposé a priori.



Les fonctions d'autocorrélation ressemblent à celles du modèle manuellement choisi. Les résidus ressemblent à un bruit blanc.

```
##
##  Box-Pierce test
##
## data:  PizzaAUTOlog$residuals
## X-squared = 35.649, df = 45, p-value = 0.8394
```

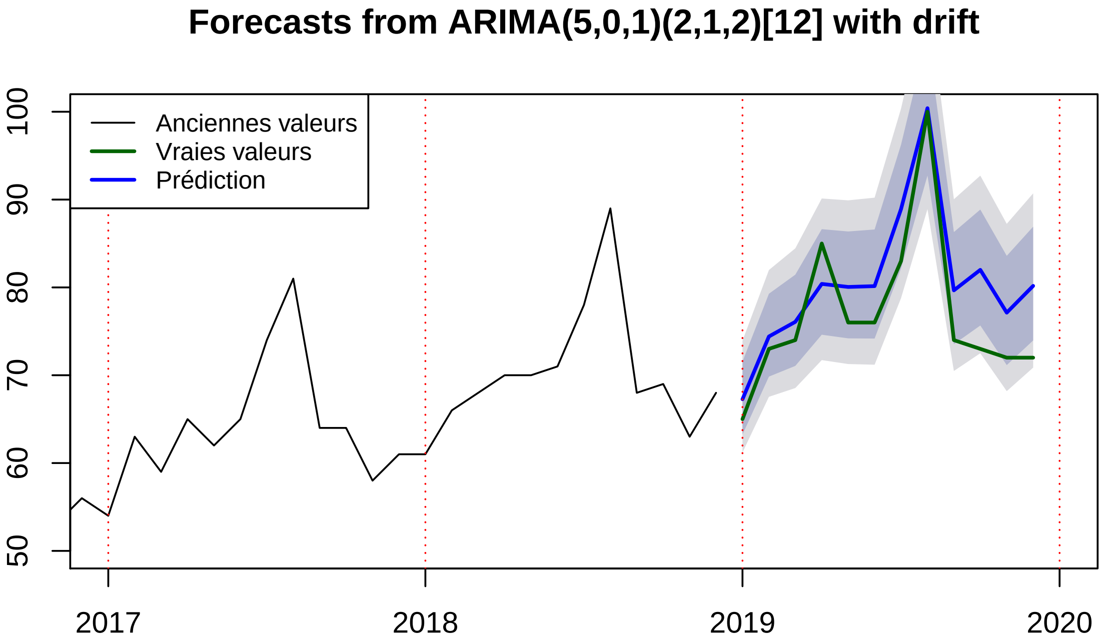
Le test de blancheur des résidus est meilleur que pour le modèle manuellement choisi mais c'est probablement le résultat d'un sur-ajustement.

```
##              ar1      ar2      ar3      ar4       ar5      ma1      sar1
## t.stat -0.940328 4.679115 4.168693 0.391760 -1.700298 1.845079 -2.590312
## p.val   0.347050 0.000003 0.000031 0.695235  0.089075 0.065026  0.009589
##              sar2     sma1      sma2    drift
## t.stat -1.089273 0.250917 -2.109833 32.67376
## p.val   0.276034 0.801878  0.034873  0.00000
```

L'analyse des p-valeurs du t-test montre que seul le coefficient `ma1` est vraiment significatif. C'est de nouveau les coefficients `sma1`, `sar1` et `sma2` qui le sont le moins.

```
##               ar1          ar2         ar3         ar4         ar5
## ar1    1.00000000 -0.509801957 -0.74953670 -0.73305970  0.19066530
## ar2   -0.50980196  1.000000000  0.48991633  0.10097167 -0.45483957
## ar3   -0.74953670  0.489916330  1.00000000  0.54753403 -0.44530157
## ar4   -0.73305970  0.100971670  0.54753403  1.00000000  0.04098634
## ar5    0.19066530 -0.454839572 -0.44530157  0.04098634  1.00000000
## ma1   -0.95475626  0.584639273  0.71252455  0.67726833 -0.22661087
## sar1  -0.01950018  0.003086140  0.01786385  0.01983447  0.04369923
## sar2   0.01690844 -0.097619442 -0.12320669 -0.04022241  0.12297239
## sma1   0.02090984 -0.023964510 -0.07140528 -0.02194428  0.01056463
## sma2  -0.02776754 -0.009037245  0.05713776  0.05497779  0.03705352
## drift  0.01236452 -0.012207980 -0.03570669 -0.03806124 -0.01183249
##                ma1        sar1        sar2         sma1         sma2
## ar1   -0.954756259 -0.01950018  0.01690844  0.020909837 -0.027767545
## ar2    0.584639273  0.00308614 -0.09761944 -0.023964510 -0.009037245
## ar3    0.712524549  0.01786385 -0.12320669 -0.071405277  0.057137755
## ar4    0.677268330  0.01983447 -0.04022241 -0.021944279  0.054977791
## ar5   -0.226610868  0.04369923  0.12297239  0.010564631  0.037053520
## ma1    1.000000000 -0.02126038 -0.03869915 -0.003597705  0.002587900
## sar1  -0.021260383  1.00000000 -0.06078802 -0.963874293  0.898669501
## sar2  -0.038699147 -0.06078802  1.00000000  0.174835246 -0.400236187
## sma1  -0.003597705 -0.96387429  0.17483525  1.000000000 -0.894329170
## sma2   0.002587900  0.89866950 -0.40023619 -0.894329170  1.000000000
## drift -0.010562771  0.01016234  0.03340971 -0.006657034 -0.004312435
##              drift
## ar1    0.012364517
## ar2   -0.012207980
## ar3   -0.035706690
## ar4   -0.038061240
## ar5   -0.011832494
## ma1   -0.010562771
## sar1   0.010162339
## sar2   0.033409715
## sma1  -0.006657034
## sma2  -0.004312435
## drift  1.000000000
```

L'analyse de colinéarités des coefficients confirme que certains coefficients sont très corrélés l'un a l'autre. Cela a faussé les t-tests des coefficients, il faut donc se méfier de son résultat. C'est le cas notamment avec `sma1` qui est très corrélé aux coefficients `sar1` et `sma2`, ce qui pourrait expliquer les p-valeurs élevées des t-tests de ces coefficients. Mais on trouve d'autres exemple : le coefficient `ar1` est très corrélé aux coefficients `ma1` et `ma2` ainsi que les coefficients `ma1` et `ma2` entre eux. A partir de séptembre, la prédiction n'est pas bonne :

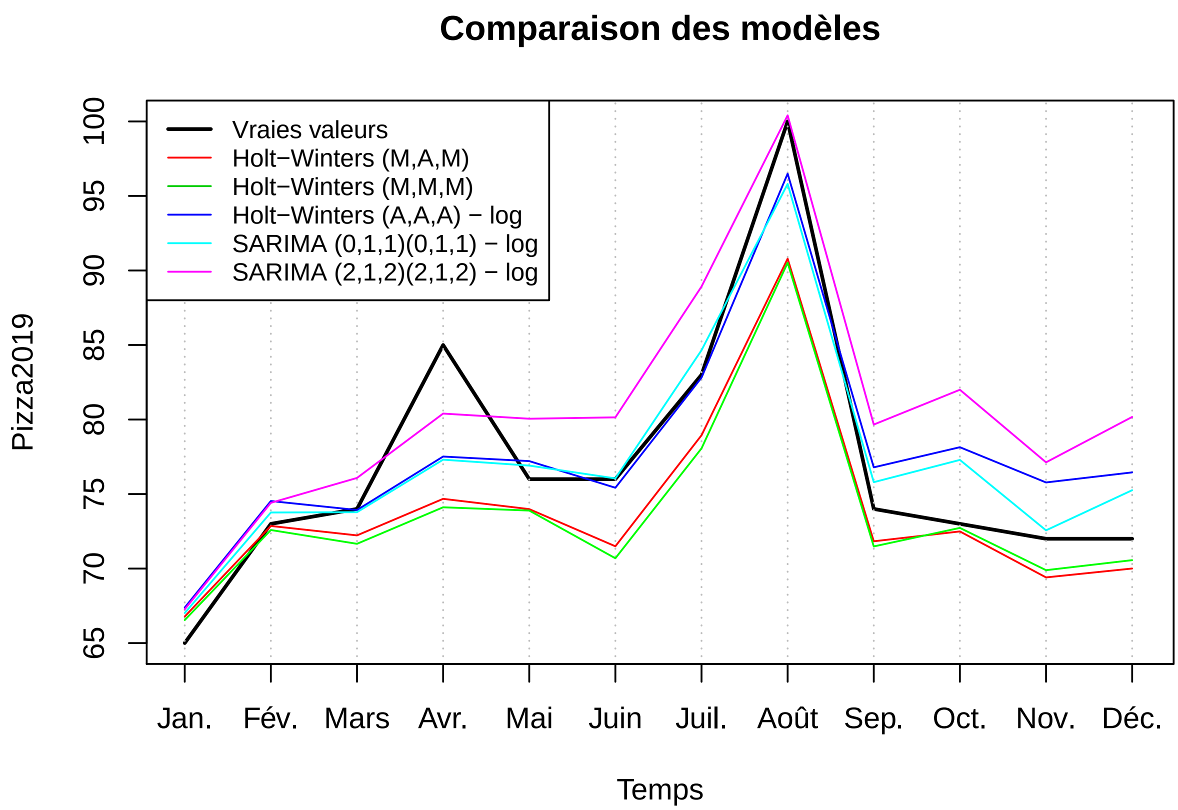

Forecasts from ARIMA(5,0,1)(2,1,2)[12] with drift
Anciennes valeurs
Vraies valeurs
Prédiction
50
60
70
80
90
100
2017
2018
2019
2020

1.4 Choix de modèle et conclusion

Cinq modèles ont été étudiés, trois modèles de lissage exponentiel selon la méthode de Holt-Winters et deux modèles SARIMA. Pour trois de ces modèles les données ont été transformées au LOG :

Nom	Transformation	Type	Méthode	Paramètres	AIC
PizzaHW_MAM	-	Lissage exponentiel	Holt-Winters	M, A, M	1113.36
PizzaHW_MMM	-	Lissage exponentiel	Holt-Winters	M, M, M	1111.65
PizzaHWlog	LOG	Lissage exponentiel	Holt-Winters	A, A, A	-141.85
PizzaSARIMAlog	LOG	Modélisation ARMA	SARIMA	$(0,1,1)(0,1,1)_{12}$	-511.21
PizzaAUTOlog	LOG	Modélisation ARMA	SARIMA	$(2,1,2)(2,1,2)_{12}$	-515.13

On compare ces modèles par superposition sur un plot.



Les prédictions des différents modèles sont très proche l'un à l'autre. On remarque que la valeur prédite pour le mois de janvier 2019 est quasiment le même pour tous les modèles. Puis, avec chaque mois de plus de prévision la variance entre les valeurs prédites augmente légèrement ce qui est visible par l'espace grandissant entre les lignes chaque mois supplémentaire. Les plus grandes différences se trouvent au mois de novembre et décembre 2019, les mois les plus loin projetés. Les mois de mars, juin et juillet sont les mieux prédits. Le pic exceptionnel du mois d'avril 2019 ainsi que le mois d'octobre sont très mal prédit par tous les modèles. Ce n'est pas une erreur de modélisation mais plutôt un aléa naturelle qui est difficile, voire impossible, à prédire en essayant de prédire une série temporelle uniquement par son passé.

Le modèle Holt-Winters M,A,M (rouge) diffère le plus des autres, surtout vers la fin de la période de prédiction (novembre, décembre). Les derniers mois de prédiction sont très intéressant car on peut distinguer 3 groupes de modèles : le modèle Holt-Winters M,A,M (rouge) qui a une estimation très basse pour les deux derniers mois, puis les modèles SARIMA (cyan, magenta) avec une prédiction basse en novembre mais qui rejoint le troisième groupe au mois de décembre (modèles Holt-Winters M,M,M et passage au LOG) qui a une prédiction plus élevée ces deux dernier mois.

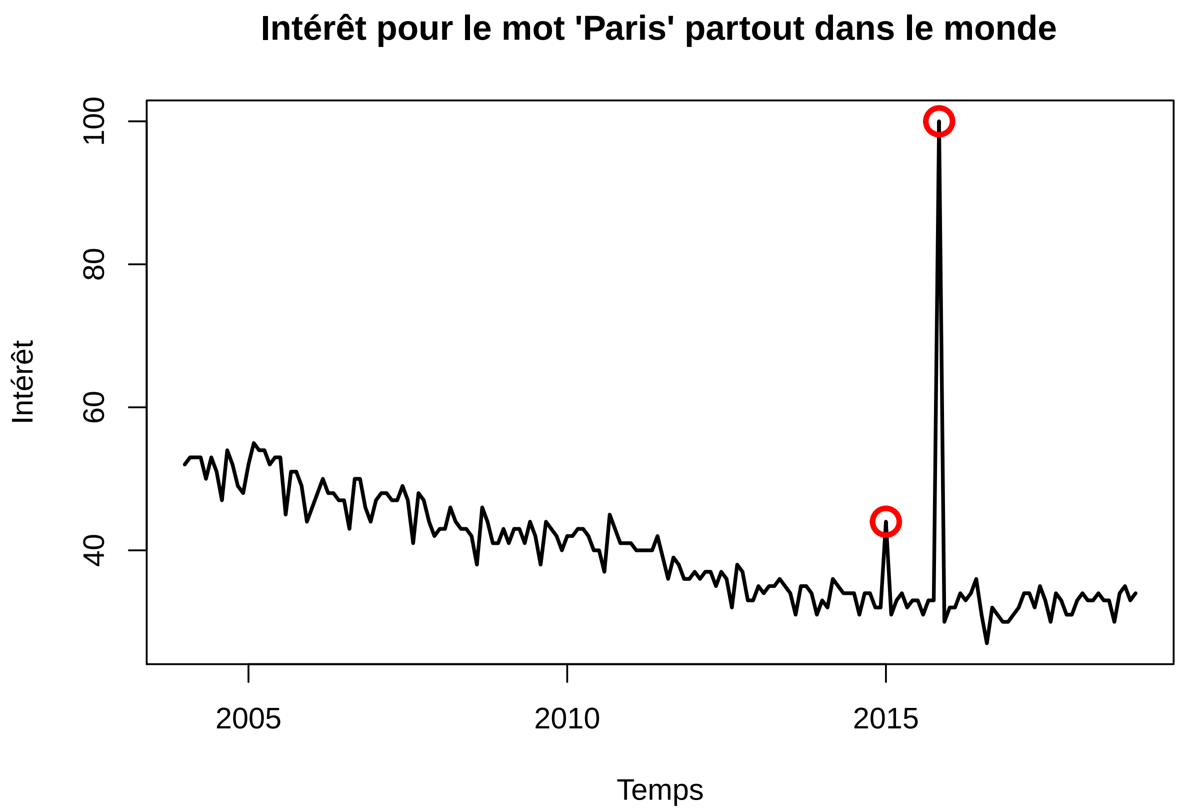
Il est difficile de choisir un seul modèle parmi ces cinq. Chaque modèle a ses forces et faiblesses. Cependant, on choisirait probablement le modèle SARIMA (cyan) avec un passage au LOG car ses prédictions sont assez proche des vraies valeurs et c'est un modèle parcimonieux avec peu de coefficients et sans autocorrélations entre les coefficients.

2. Intérêt pour le mot "Paris" dans le monde entier

La série temporelle `Paris` montre l'évolution de l'intérêt pour le mot "Pizza" partout dans le monde entre janvier 2004 et novembre 2019. Les données sont téléchargeables sur https://trends.google.fr/trends/. Les valeurs des mois de l'année 2019 ont été écartées de la modélisation afin de pouvoir comparer les prédictions des modélisations avec les vraies valeurs de la série.

2.1 Analyse exploratoire de la série

2.1.1 Chronogramme

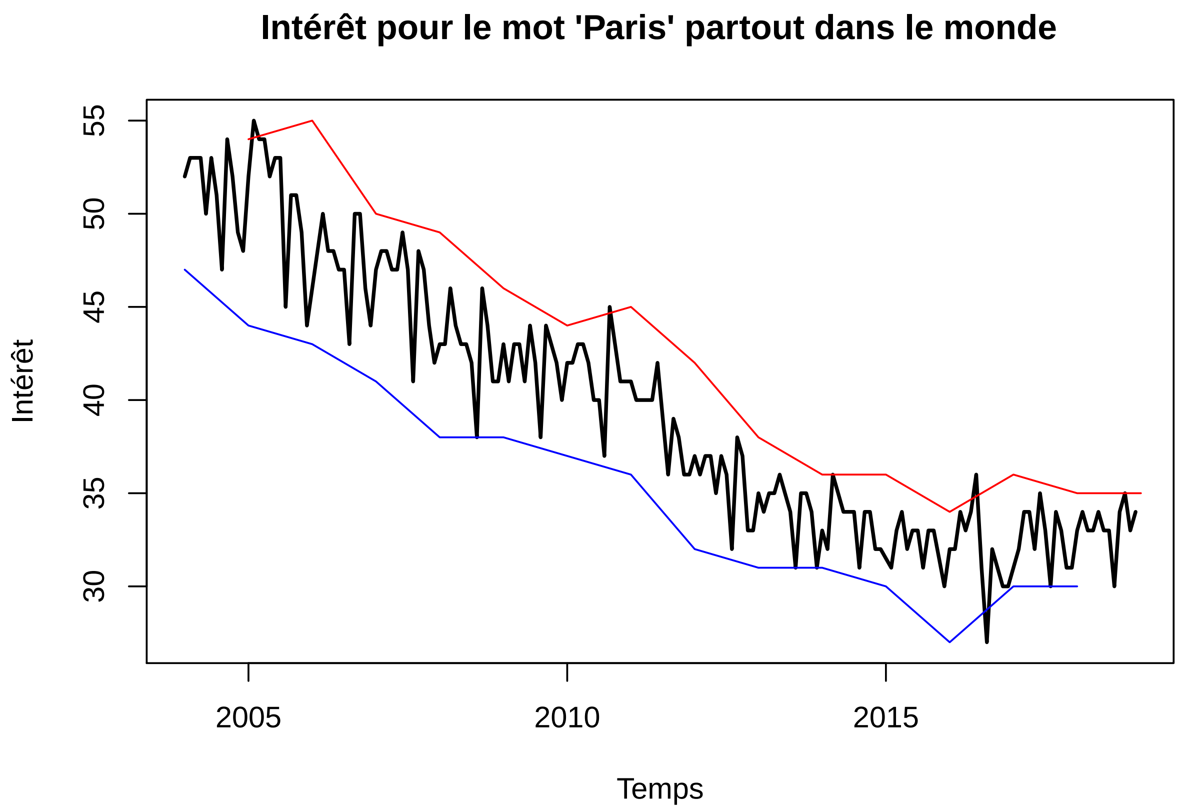


Le chronogramme montre que la série temporelle est décroissante avec une variance probablement stable et éventuellement une saisonnalité. On repère immédiatement deux valeurs aberrantes qui ne correspondent pas aux autres valeurs autour. On peut expliquer l'expliquer. Ce sont les mois de janvier et novembre 2015, les mois de l'attentat contre le journal Charlie Hebdo le 7 janvier 2015 ainsi que des attentats du 13 novembre 2015 qui se manifestent par un intérêt exceptionnellement élevé pour le mot Paris partout dans le monde.

2.1.2 Valeurs aberrantes

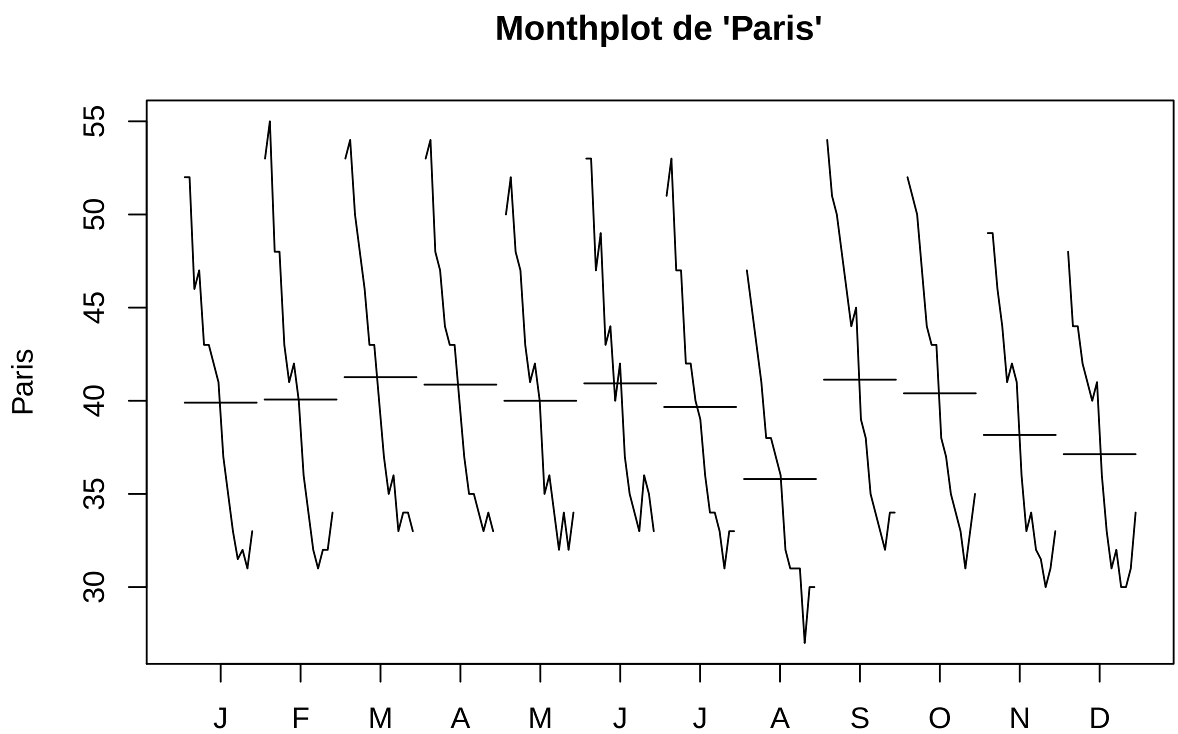
Les valeurs aberrantes peuvent causer des problèmes pour la modélisation d'une série temporelle. Yves Aragon (2014) propose dans son livre "Séries temporelles avec R" d'enlever manuellement les valeurs aberrantes et de les remplacer ensuite par des valeurs raisonnable, obtenues notamment par interpolation linéaire entre les deux valeurs voisines de la valeur aberrante ou par affectation manuelle d'une valeur raisonnable.

Voici la série après interpolation linéaire des valeurs aberrantes :



La tendance semble décroître linéairement jusqu'à la fin 2014 afin de se stabiliser après. Les courbes rouge et bleue représentent le minimum et le maximum d'une année entière. Ces deux courbes n'étant pas parallèles elles indiquent plutôt un modèle multiplicatif mais l'effet est assez faible. Il est donc intéressant d'étudier des modèles additifs et multiplicatifs.

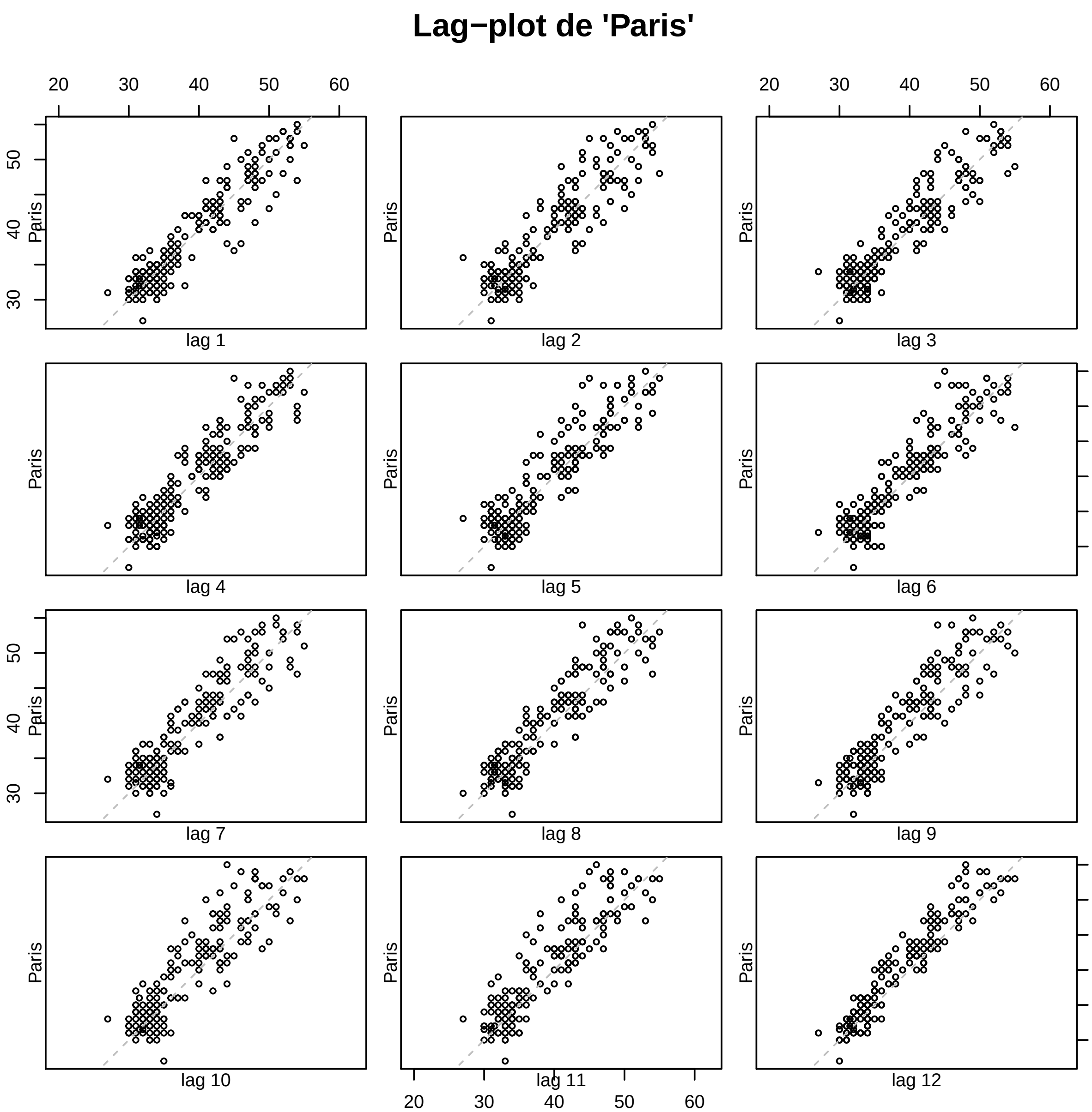
2.1.3 Month-plot



Le month-plot confirme qu'il y a une saisonnalité. L'intérêt pour le mot Paris partout dans le monde est quasiment stable de janvier à juin et commence à décroître en juillet jusqu'à atteindre un minimum pendant le mois d'août. En septembre l'intérêt retrouve soudainement les valeurs des premiers mois de l'année et décroît de nouveau successivement jusqu'au mois de décembre. En janvier l'intérêt fait de nouveau un saut. La variance semble être stable d'un mois à l'autre.

Paris étant une ville touristique, ce comportement est étonnant vu que les mois de bas intérêt coïncident avec les mois des vacances d'été dans l'hémisphère nord ainsi que les vacances de noël et de fin d'année. On aurait plutôt soupçonner l'inverse.

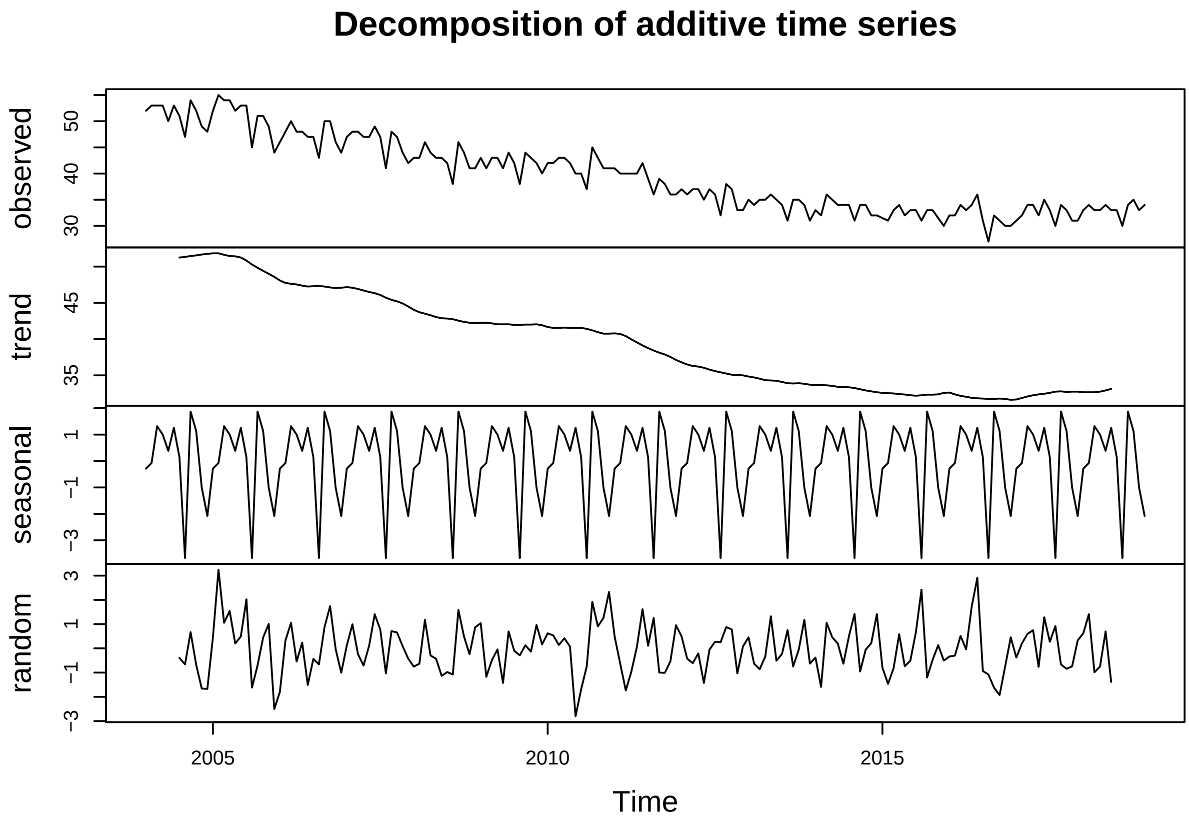
2.1.4 Lag-plot



Sur lag-plot de `Paris` on voit qu'un instant de la série est très fortement corrélé aux instants avant de manière générale. Néanmoins c'est surtout une autocorrélation d'ordre 12 qu'on peut constater. La série dépend beaucoup de son passé et surtout de ce qui s'est passé 12 mois avant.

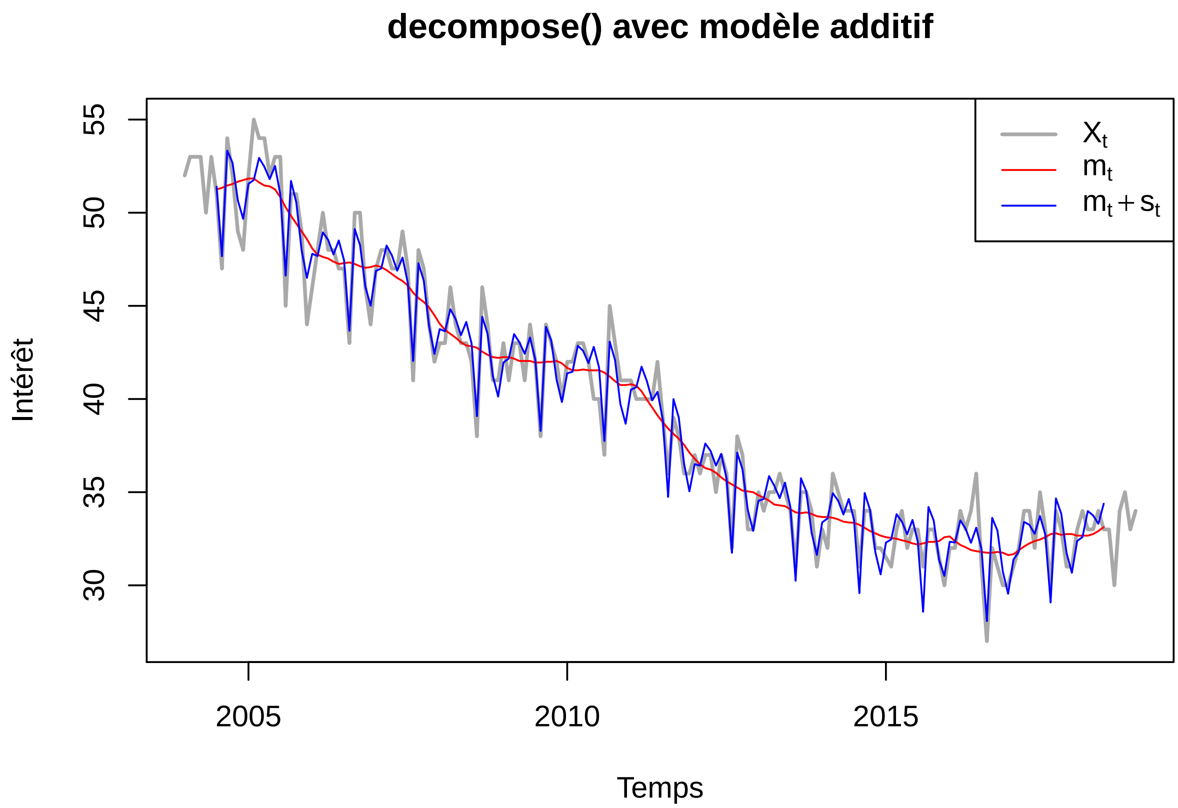
2.1.5 Décomposition

Modèle additif L'analyse du chronogramme ressemble à un modèle additif, on commence alors avec la décomposition additive.



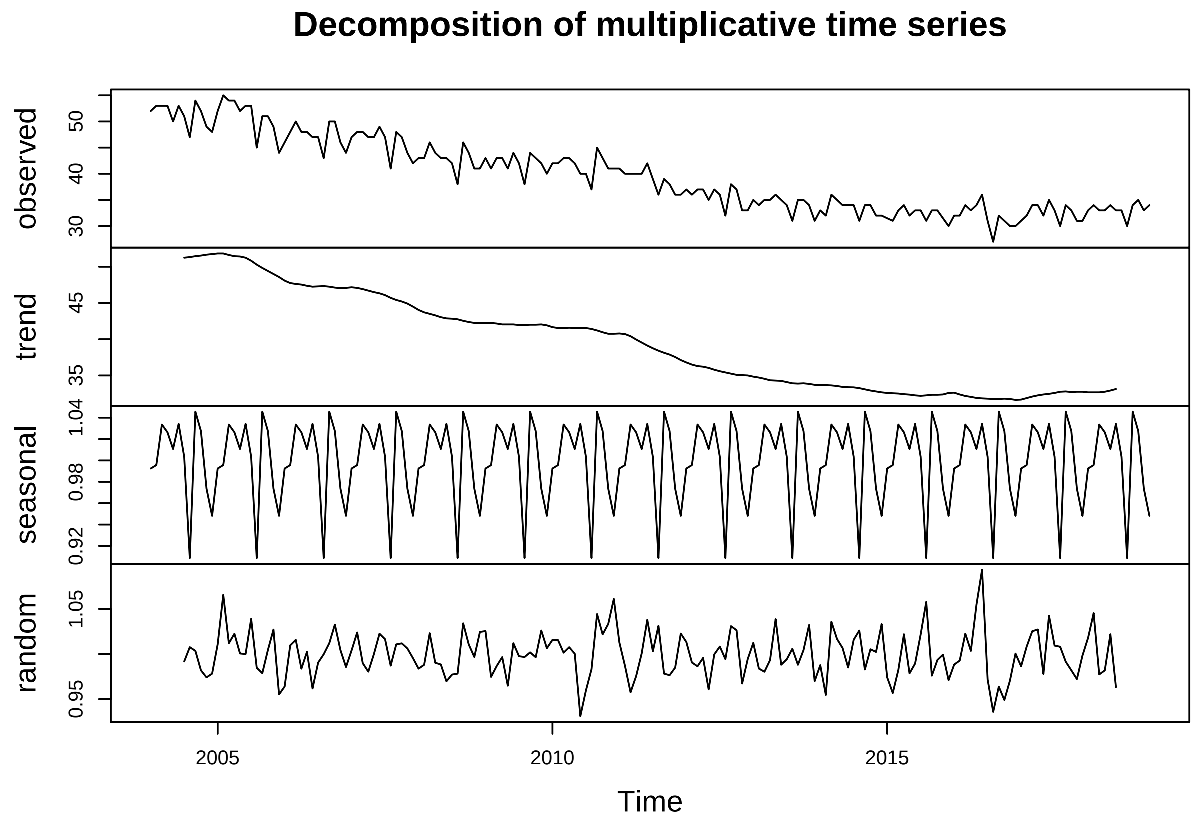
La série a une tendance décroissante plutôt linéaire entre janvier 2004 et janvier 2017. A partir de février 2017 l'intérêt pour le mot Paris semble à monter de nouveau. Il y a bien un effet saisonnier, avec un creux pendant les mois de juillet et août ainsi que pendant le mois de décembre. La variance des résidus semble être stable avec deux singularités assez remarquables. C'est notamment le maximum des résidus en juin 2016 ainsi que leur minimum en décembre 2005. Une valeur positive des résidus signifie un intérêt plus élevé que d'habitude et c'est l'inverse pour une valeur négative. Le maximum des résidus en juin 2016 pourrait s'expliquer par le Championnat d'Europe de football 2016 qui se déroulait entre le 10 juin 2016 et le 10 juillet 2016 et dont 12 des 51 matches ont été joués à Paris ou Saint-Denis.

On reconstitue la série originale (grise) à l'aide des résultats de la décomposition additif avec sa tendance (rouge) et sa tendance plus la saisonnalité additif (bleu).

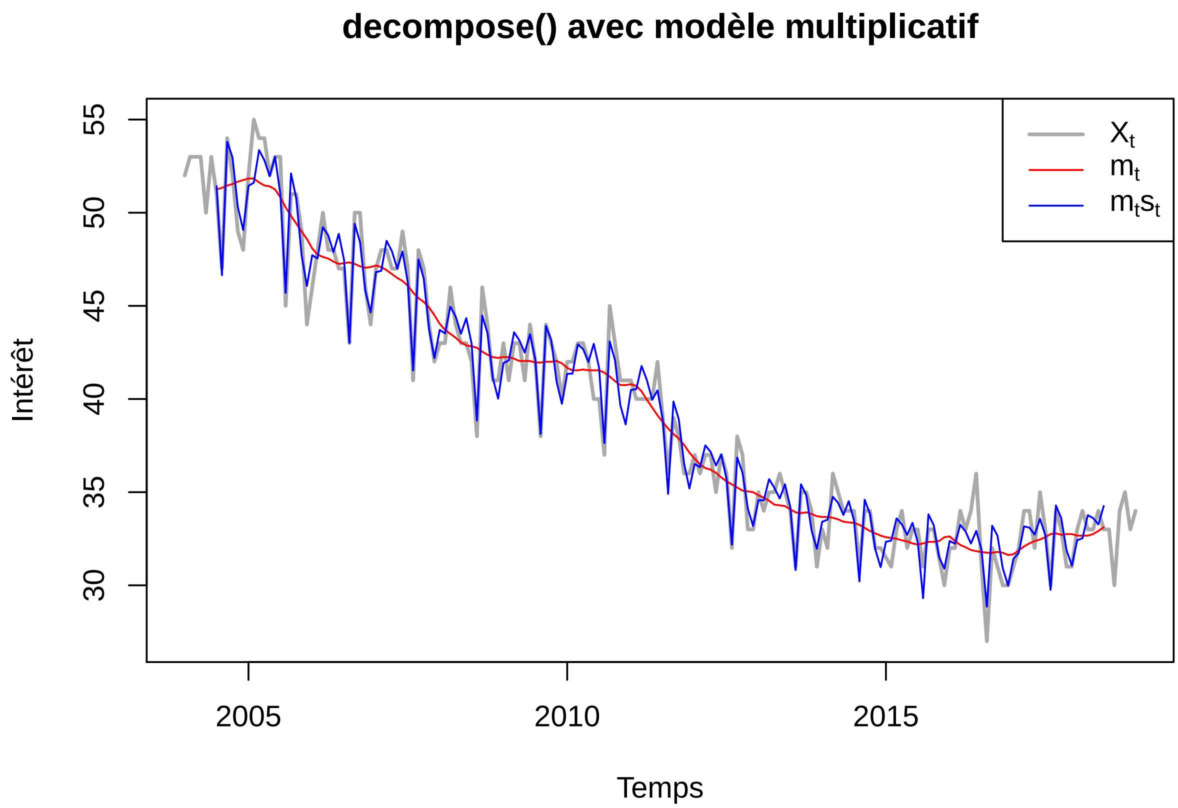


Le modèle additif s'ajuste bien aux vraies valeurs de manière générale. Cependant, les pics et creux sont souvent sous-estimés, notamment en 2016 et entre les années 2004 et 2010. Il est intéressant de regarder la décomposition multiplicative de la série.

Modèle multiplicatif



Tendance et saisonnalité sont comparables au modèle additif. La variance des résidus semble être stable à l'exception des mois de juin et août 2016. L'effet Euro 2016 se manifeste donc aussi dans les résidus du modèle multiplicatif. La simulation du modèle du modèle multiplicatif n'est pas mieux que celle du modèle additif :



Le modèle multiplicatif s'ajuste assez bien aux vraies valeurs depuis 2013 (à l'exception de la période de l'Euro 2016). Entre 2004 et 2012 ce sont surtout les pics qui ne sont pas très bien représentés. Il est très difficile de choisir entre un modèle additif et multiplicatif dans le cas de la série `Paris`, les deux modèles sont trop proches l'un à l'autre. Vu la linéarité de la tendance une transformation au LOG ou à la racine carrée n'est pas prometteur non plus. Il faut tester de différentes modèles de lissage exponentiel.

2.2 Modélisation par lissage exponentiel

L'analyse descriptive de la série `Paris` ainsi que sa décomposition ont montré qu'elle a une tendance et une saisonnalité. Il faudrait donc modéliser la série avec la méthode de Holt-Winters afin d'estimer la tendance, la saisonnalité ainsi que l'erreur.

2.2.1 Méthode de Holt-Winters

En se basant sur l'analyse précédente, il est difficile de dire si un modèle additif ou multiplicatif est mieux. On peut donc tester différentes modèles, faire une présélection sur la base de l'AIC et ensuite comparer les modèles présélectionnés avec les vraies valeurs de la série.

On peut par exemple tester les modèles suivants :

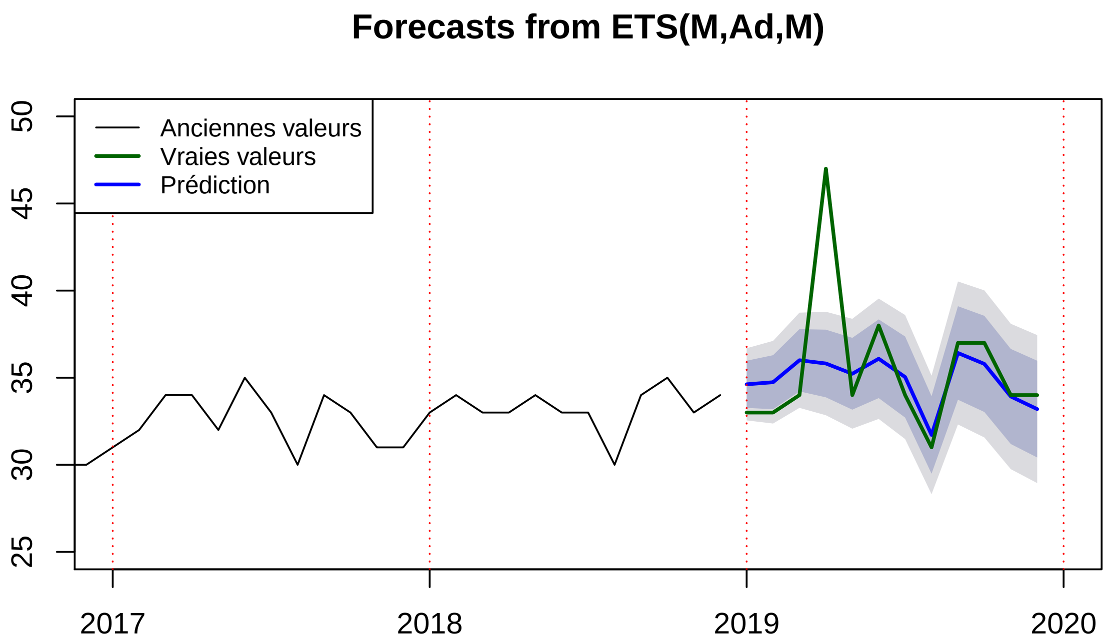
Modèle	Erreur	Tendance	Saisonnalité
Holt-Winters A, A, A	additive	additive	additive
Holt-Winters M, A, A	multiplicative	additive	additive
Holt-Winters M, A, M	multiplicative	additive	multiplicative
Holt-Winters M, M, M	multiplicative	multiplicative	multiplicative

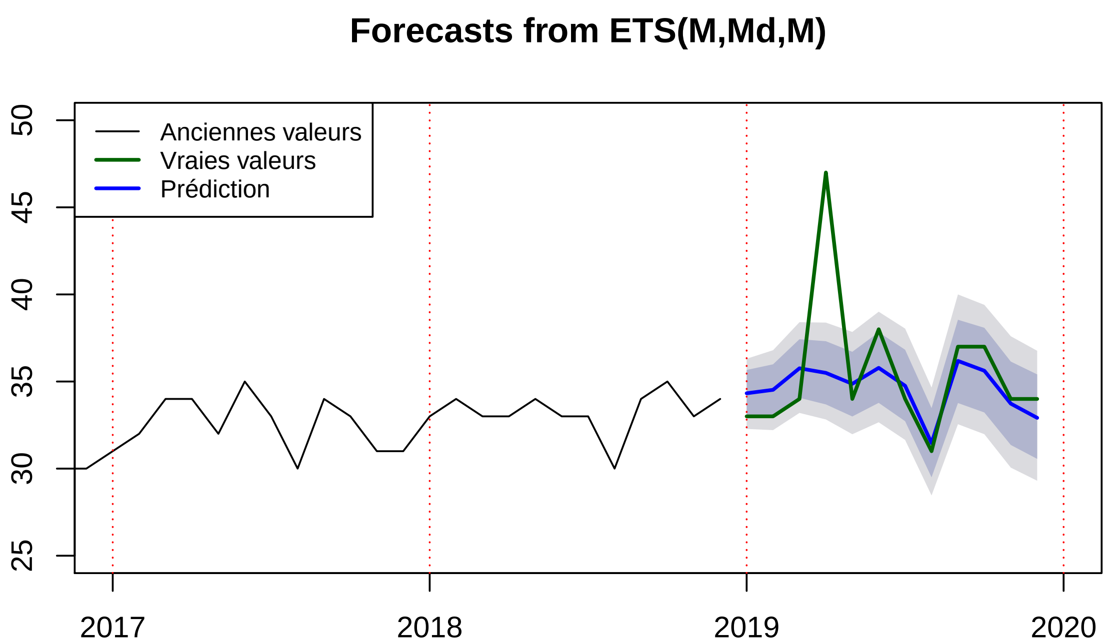
Le paramètre `damped` de la fonction `ets()` corresponde au paramètre ϕ de la composante tendancielle. Si `damped = TRUE` la fonction calcul une tendance amortie. Si on ne précise pas ce paramètre la fonction le choisit automatiquement sur la base d'un critère (AIC, AICc ou BIC) qu'il faudrait préciser. On choisit le critère AIC et la fonction choisit donc une tendance amortie dès que le type de l'erreur est multiplicative. C'est visible par un "d" derrière le type de la tendance.

Nom du modèle	Méthode	AIC
ParisHW_AAA	ETS(A,A,A)	1035.65
ParisHW_MAdA	ETS(M,A,A)	1035.12
ParisHW_MAdM	ETS(M,Ad,M)	1015.96
ParisHW_MMdM	ETS(M,Md,M)	1015.47

De manière générale, l'AIC est plus petit lorsque la saisonnalité est de type multiplicatif. Le type de la tendance n'a presque pas d'influence sur l'AIC. On fait des prédictions pour les deux derniers modèles et les compare avec les vraies valeurs.

2.2.2 Prédiction Holt-Winters



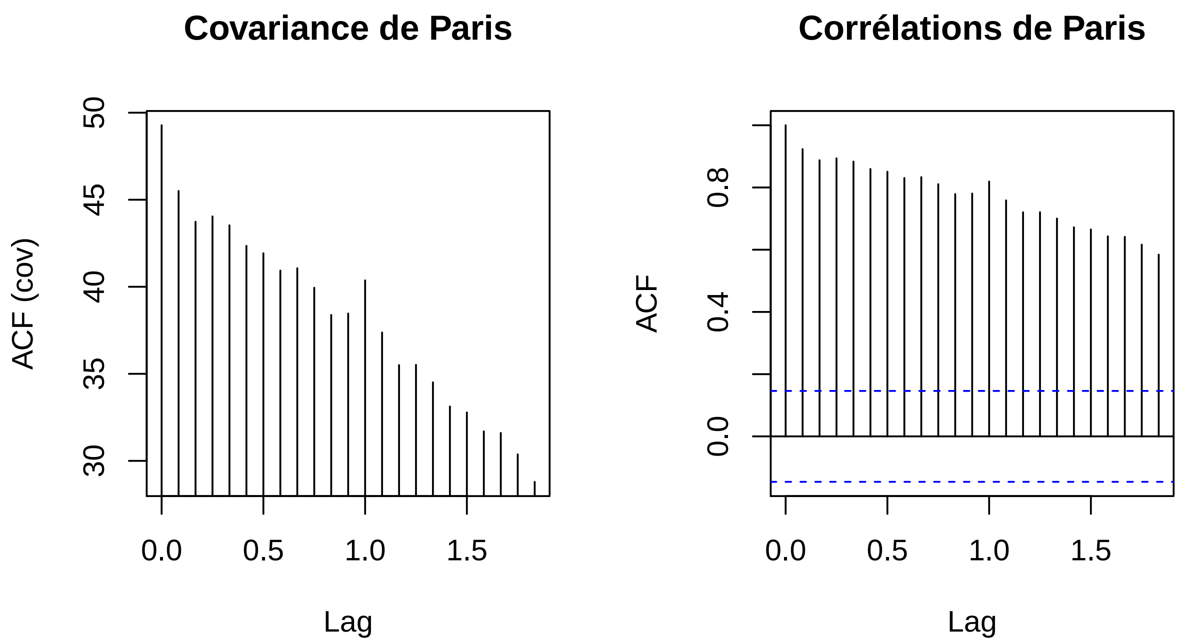


Les prédictions des deux modèles ne se distingue que très légèrement, seulement la prédiction de janvier 2019 est différente. Tous les deux modèles sont très précis. On remarque que l'intervalle de confiance du modèle avec tendance multiplicative amortie est un peu plus petit que celui du modèle avec tendance additive amortie.

Le mois d'avril 2019 est très mal prédit par les deux modèles. En effet il est particulier et ne correspond pas aux années précédentes. L'intérêt particulièrement élevé pour le mot Paris est très probablement dû à l'incendie de la cathédrale Notre-Dame de Paris qui avait lieu les 15 et 16 avril 2019.

2.3 Modélisation

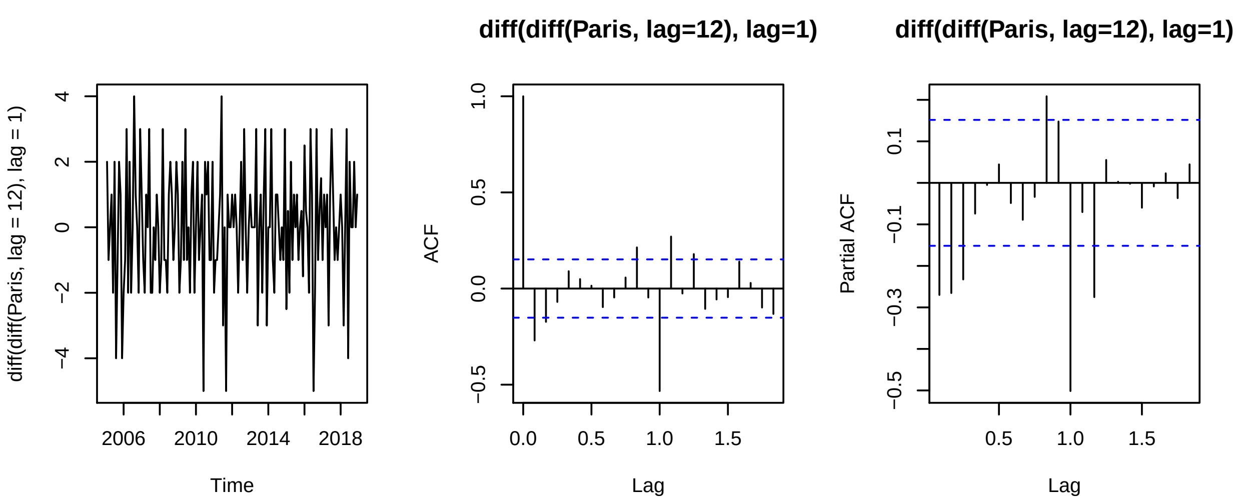
2.3.1 Analyse des fonctions d'autocorrélation



Le lag-plot a déjà montré que la série `Paris` est très fortement auto-corrélée. Les fonctions d'autocorrélation confirment ce résultat. La corrélation est grande surtout à l'ordre 12.

2.3.2 Différenciation de la série

On différencie la série afin d'enlever sa saisonnalité d'ordre 12 et sa tendance. Le but est de la ramener à un processus stationnaire et d'identifier les paramètres d'un modèle autorégressif (AR) ou moyennes mobiles (MA).



Après différenciation, la fonction d'autocorrélation (ACF) ainsi que la fonction d'autocorrélation partielle (PACF) sont 0 à l'ordre 3, ce qui indique un modèle ARMA(2,2). On voit très bien qu'il reste encore de la saisonnalité, il faudrait donc rajouter des composantes saisonnière de différenciation dans le modèle, on pourrait donc commencer avec un modèle ARIMA(2,1,2).

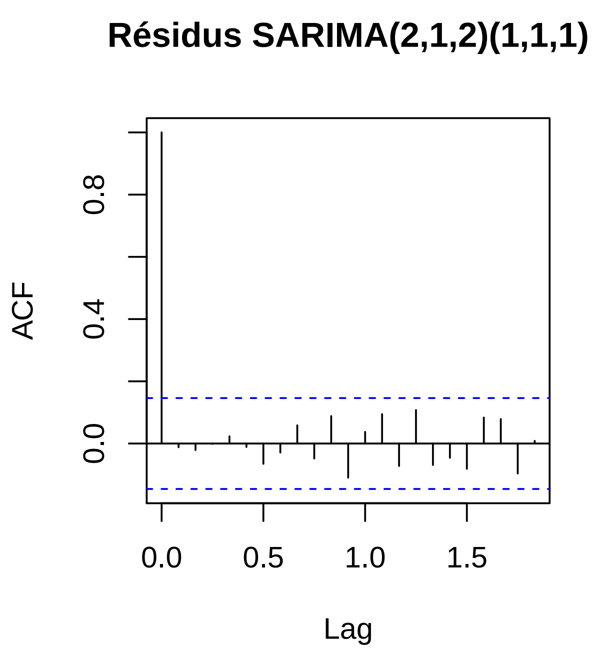
2.3.3 Modélisation SARIMA

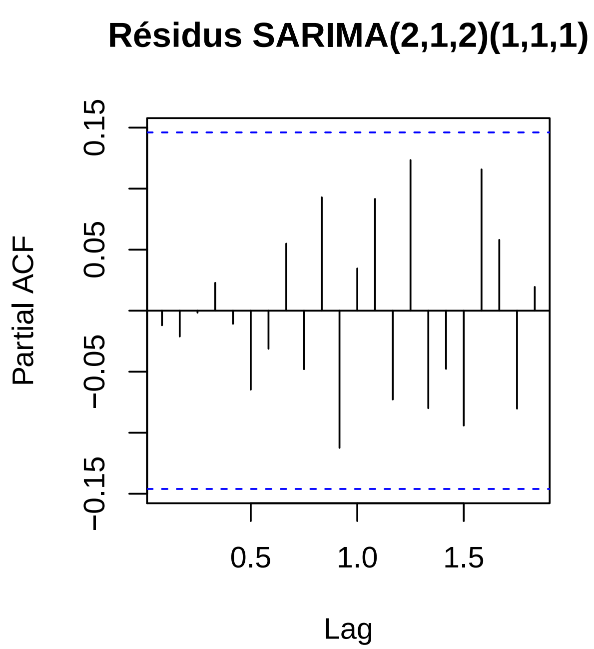
Suite à l'analyse descriptive et l'analyse des fonctions d'autocorrélation (avec et sans différenciation) on peut commencer par un modèle ARIMA(2,1,2) auquel on rajoute une composante saisonnière, donc avec un SARIMA(2,1,2)(1,1,1).

SARIMA(2,1,2)(1,1,1)

```
## Series: Paris
## ARIMA(2,1,2)(1,1,1)[12]
##
## Coefficients:
##          ar1      ar2      ma1      ma2     sar1     sma1
##       0.0200  -0.1104  -0.4754  -0.0583  -0.2908  -0.5391
## s.e.  0.5755   0.2235   0.5722   0.4402   0.1312   0.1310
##
## sigma^2 estimated as 1.636:  log likelihood=-279.52
## AIC=573.03   AICc=573.74   BIC=594.86
##
## Training set error measures:
##                      ME    RMSE       MAE       MPE     MAPE      MASE
## Training set 0.0644523 1.20958 0.9323375 0.1890097 2.452296 0.4789991
##                      ACF1
## Training set -0.01194969
```

On remarque que les écart-types des coefficients sont assez élevés.





Les fonctions d'autocorrélation des résidus semblent indiquer un bruit blanc.

```
##
##  Box-Pierce test
##
## data:  ParisSARIMA$residuals
## X-squared = 24.552, df = 45, p-value = 0.9944
```

La p-valeur du test de blancheur est très élevée. L'hypothèse 0 est rejetée, les résidus sont très probablement un bruit blanc.

```
##             ar1       ar2       ma1       ma2      sar1      sma1
## t.stat 0.034822 -0.493809 -0.830703 -0.132467 -2.215717 -4.115274
## p.val  0.972222  0.621441  0.406141  0.894615  0.026711  0.000039
```

Tous les coefficients ne sont pas significatifs. Surtout la p-valeurs du coefficient `ma1` est très élevée mais aussi `ar1` a une p-valeur légèrement élevée. On regarde les corrélations entre les coefficients.

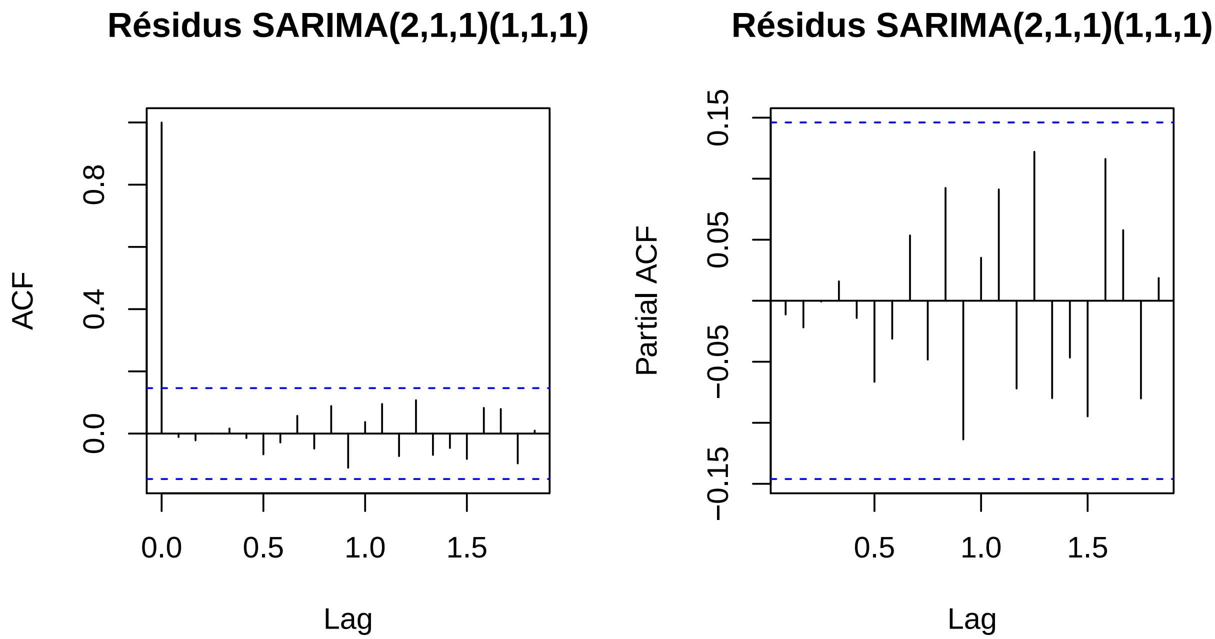
```
##             ar1         ar2        ma1         ma2        sar1        sma1
## ar1   1.0000000 -0.65172301 -0.9905804  0.92857523  0.17477207 -0.14171766
## ar2  -0.6517230  1.00000000  0.6465571 -0.84926256 -0.07967251  0.06546364
## ma1  -0.9905804  0.64655714  1.0000000 -0.93181759 -0.15628673  0.12485325
## ma2   0.9285752 -0.84926256 -0.9318176  1.00000000  0.12440720 -0.09663144
## sar1  0.1747721 -0.07967251 -0.1562867  0.12440720  1.00000000 -0.79486962
## sma1 -0.1417177  0.06546364  0.1248533 -0.09663144 -0.79486962  1.00000000
```

L'analyse de colinéarités des coefficients montre que certains coefficients sont très corrélés l'un a l'autre. `ar1` est très corrélé à `ma1`. De même pour `ar2` et `ma2`.

SARIMA(2,1,1)(1,1,1) Suite aux résultats des analyses du modèle SARIMA(2,1,2)(1,1,1) on décide d'enlever le coefficient `ma2` car il est le coefficient le moins significatif et le plus corrélé aux autres coefficients.

```
## Series: Paris
## ARIMA(2,1,1)(1,1,1)[12]
##
## Coefficients:
##          ar1      ar2      ma1     sar1     sma1
##       0.0918  -0.1345  -0.5470  -0.2886  -0.5407
## s.e.  0.2161   0.1232   0.2131   0.1305   0.1305
##
## sigma^2 estimated as 1.626:  log likelihood=-279.53
## AIC=571.05   AICc=571.58   BIC=589.76
##
## Training set error measures:
##                     ME     RMSE       MAE      MPE     MAPE      MASE
## Training set 0.0638315 1.209678 0.9326174 0.186861 2.453231 0.4791429
##                     ACF1
## Training set -0.01127827
```

L'AIC du modèle SARIMA(2,1,1)(1,1,1) est légèrement plus petit que celui du premier modèle SARIMA(2,1,2)(1,1,1) étudié. Les écart-types sont toujours assez élevés.



Les résidus ressemblent à un bruit blanc.

```
##
##  Box-Pierce test
##
## data:  ParisSARIMA$residuals
## X-squared = 24.605, df = 45, p-value = 0.9943
```

La p-valeur du test de blancheur est très élevée. L'hypothèse 0 est rejetée, les résidus sont très probablement un bruit blanc.

```
##             ar1       ar2       ma1      sar1      sma1
## t.stat 0.424700 -1.092175 -2.566678 -2.211846 -4.142076
## p.val  0.671055  0.274756  0.010268  0.026977  0.000034
```

Les coefficients `ar2` et `sar1` ne sont pas significatifs.

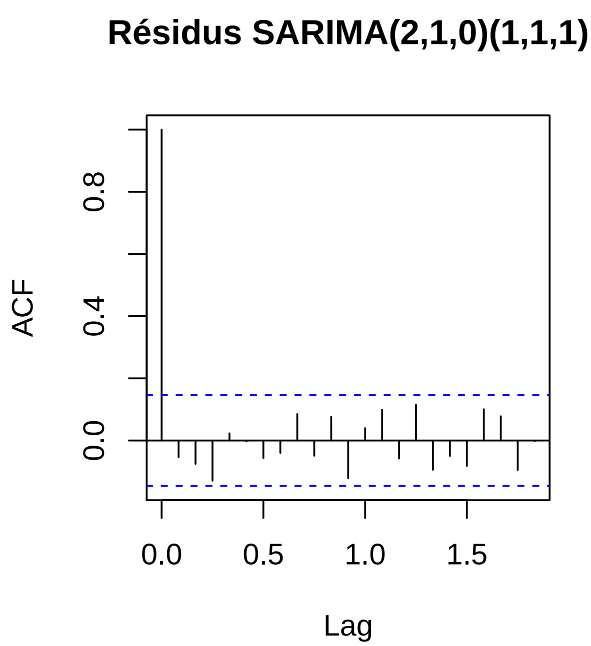
```
##             ar1          ar2         ma1         sar1         sma1
## ar1   1.0000000  0.707544083 -0.93243292  0.14588683 -0.126318315
## ar2   0.7075441  1.000000000 -0.77939770  0.02339522 -0.009068686
## ma1  -0.9324329 -0.779397695  1.00000000 -0.09520961  0.080358975
## sar1  0.1458868  0.023395218 -0.09520961  1.00000000 -0.793131111
## sma1 -0.1263183 -0.009068686  0.08035898 -0.79313111  1.000000000
```

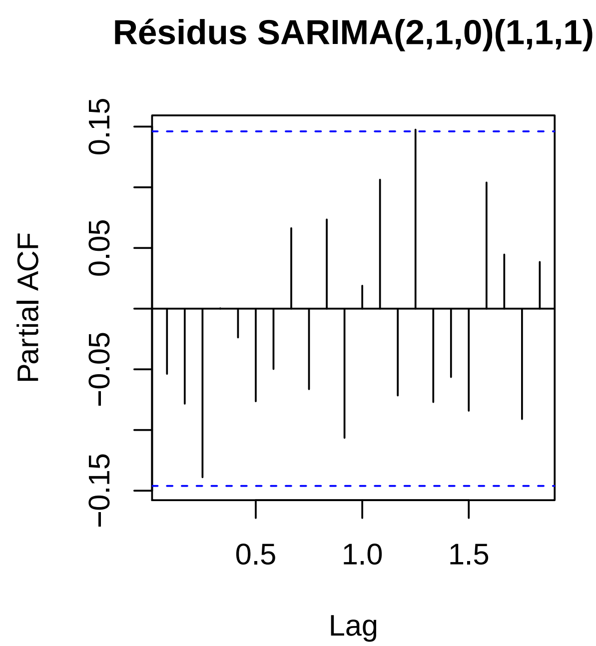
Le coefficient `ma1` est très corrélé à `ar1` et `ar2`. Aussi les coefficients `sar1` et `sma1` sont corrélés l'un à l'autre.

SARIMA(2,1,0)(1,1,1) On décide d'enlever le coefficient `ma1` car il est le plus corrélé aux autres coefficients. On analyse donc un modèle SARIMA(2,1,0)(1,1,1).

```
## Series: Paris
## ARIMA(2,1,0)(1,1,1)[12]
##
## Coefficients:
##           ar1      ar2     sar1     sma1
##       -0.3998  -0.2896  -0.3356  -0.5030
## s.e.   0.0758   0.0742   0.1254   0.1305
##
## sigma^2 estimated as 1.663:  log likelihood=-281.86
## AIC=573.73   AICc=574.1   BIC=589.32
##
## Training set error measures:
##                      ME     RMSE       MAE       MPE     MAPE      MASE
## Training set 0.05394495 1.227086 0.9413019 0.1478089 2.472487 0.4836046
##                     ACF1
## Training set -0.05372534
```

L'AIC est plus élevé que pour le modèle précédent mais les écart-types des coefficients sont plus petits.





Les résidus sont un peu moins bien mais pourront toujours être un bruit blanc. On fait le test de blancheur.

```
##
##  Box-Pierce test
##
## data:  ParisSARIMA$residuals
## X-squared = 31.032, df = 45, p-value = 0.9438
```

La p-valeur élevée du test de blancheur confirme la blancheur des résidus.

```
##             ar1        ar2       sar1       sma1
```

```
## t.stat -5.27585 -3.902590 -2.676576 -3.854256
## p.val   0.00000  0.000095  0.007438  0.000116
```

Tous les coefficients sont significatifs.

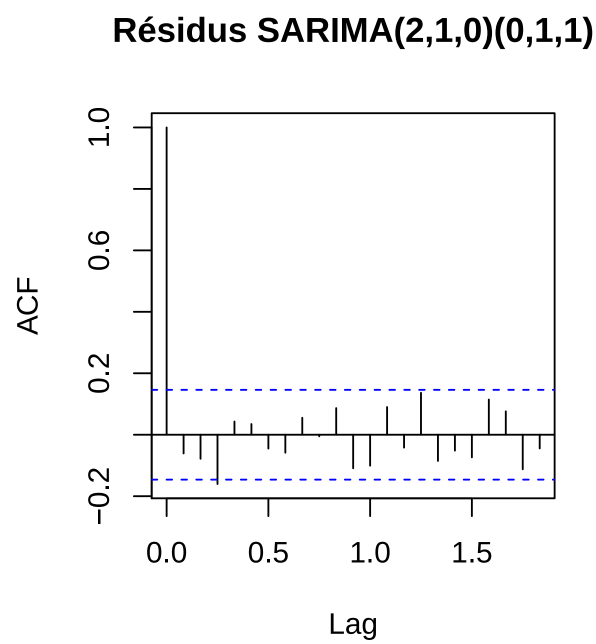
```
##             ar1         ar2        sar1        sma1
## ar1   1.0000000  0.31883385  0.19976921 -0.17875645
## ar2   0.3188338  1.00000000  0.07912614 -0.04880544
## sar1  0.1997692  0.07912614  1.00000000 -0.78604713
## sma1 -0.1787565 -0.04880544 -0.78604713  1.00000000
```

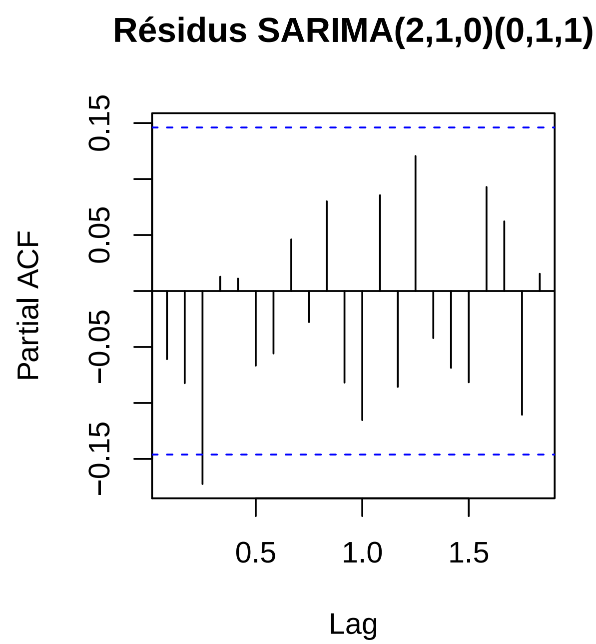
Il n'y a que très peu de corrélations entre les coefficients. Seulement `sar1` et `sma1` sont assez corrélés.

SARIMA(2,1,0)(0,1,1) Le modèle précédent est déjà assez bon. Néanmoins on peut tenter d'enlever `sar1` afin d'éliminer la corrélation qui reste entre les coefficients. On test alors un modèle SARIMA(2,1,0)(0,1,1).

```
## Series: Paris
## ARIMA(2,1,0)(0,1,1)[12]
##
## Coefficients:
##           ar1      ar2     sma1
##       -0.3670  -0.2747  -0.7298
## s.e.   0.0745   0.0742   0.0617
##
## sigma^2 estimated as 1.715:  log likelihood=-285.16
## AIC=578.32   AICc=578.56   BIC=590.79
##
## Training set error measures:
##                      ME     RMSE       MAE       MPE     MAPE      MASE
## Training set 0.04764904 1.249972 0.9547793 0.1338762 2.502133 0.4905288
##                     ACF1
## Training set -0.0608834
```

L'AIC est plus élevé que pour le modèle précédent mais les écart-types des coefficients sont tous petits.





Les fonctions d'autocorrélation des résidus semblent indiquer un bruit blanc même s'il reste un petit pic d'ordre 3.

```
##
##  Box-Pierce test
##
## data:  ParisSARIMA$residuals
## X-squared = 43.892, df = 45, p-value = 0.5188
```

La p-valeur élevée confirme la blancheur des résidus.

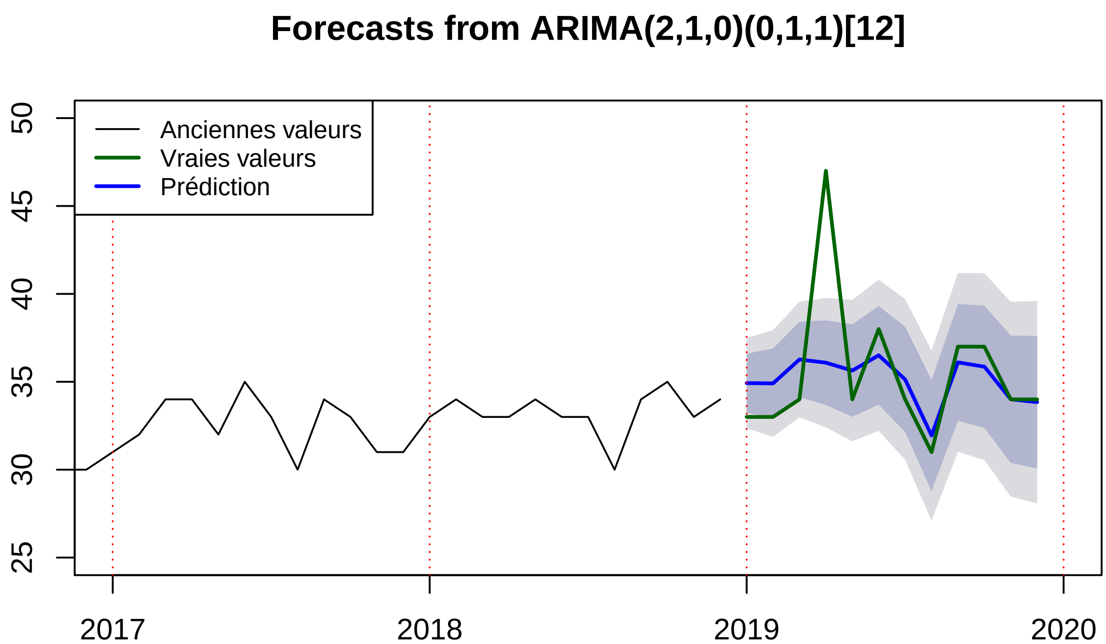
```
##              ar1       ar2      sma1
## t.stat -4.923074 -3.701890 -11.82241
## p.val   0.000001  0.000214   0.00000
```

Tous les coefficients sont significatifs.

```
##              ar1        ar2        sma1
## ar1   1.00000000 0.28846229 -0.01041022
## ar2   0.28846229 1.00000000  0.01078453
## sma1 -0.01041022 0.01078453  1.00000000
```

Il n'y a pas de corrélations entre les coefficients du modèle. On fait donc une prédiction avec ce modèle afin de la comparer aux vraies valeurs.

2.3.4 Prédiction SARIMA



La prédiction pour l'année 2019 du modèle SARIMA(2,1,0)(0,1,1) est très bien globalement et ressemble à celles des modèles Holt-Winters. Le mois d'avril 2019 est très mal prédit pour des raisons déjà évoquées.

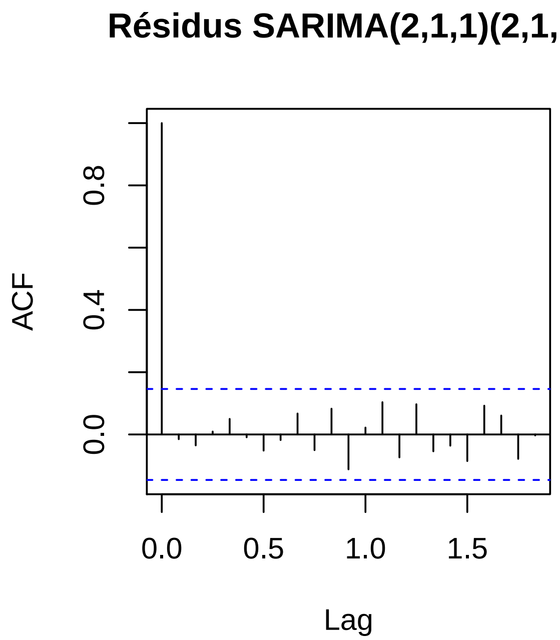
2.3.5 Modélisation automatique

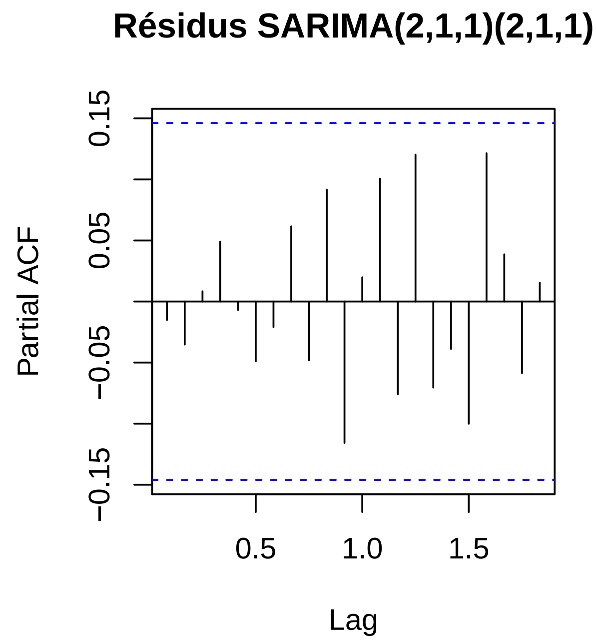
Comme pour la série `Pizza` il est intéressant de regarder un modèle automatique et de le comparer au modèle manuellement choisi pour la série `Paris`. On utilise de nouveaux la fonction `auto.arima()` du package `forecast`.

```
## Series: Paris
## ARIMA(1,1,2)(2,1,2)[12]
##
## Coefficients:
##           ar1      ma1      ma2     sar1    sar2     sma1     sma2
##       -0.1899  -0.2695  -0.2487  -0.4925  0.1016  -0.3131  -0.2574
## s.e.   0.4193   0.4071   0.2120   0.5813  0.2293   0.5813   0.3119
##
## sigma^2 estimated as 1.621:  log likelihood=-278.58
## AIC=573.16   AICc=574.07   BIC=598.11
##
## Training set error measures:
##                      ME     RMSE       MAE       MPE     MAPE      MASE
## Training set 0.06816969 1.200397 0.9323297 0.2038021 2.457048 0.4789951
##                     ACF1
## Training set -0.01509018
```

L'AIC du modèle automatique est plus petit que celui du modèle SARIMA(2,1,0)(0,1,1) manuellement choisi.

Cependant, les écart-types des coefficients sont plus élevés, surtout pour `sar1` et `sma1`.





Les fonctions d'autocorrélation sont comme on le souhaite. Les résidus semblent à un bruit blanc.

```
##
##  Box-Pierce test
##
## data:  ParisAUTO$residuals
## X-squared = 23.373, df = 45, p-value = 0.9968
```

La p-valeur très élevée du test de blancheur confirme la blancheur des résidus.

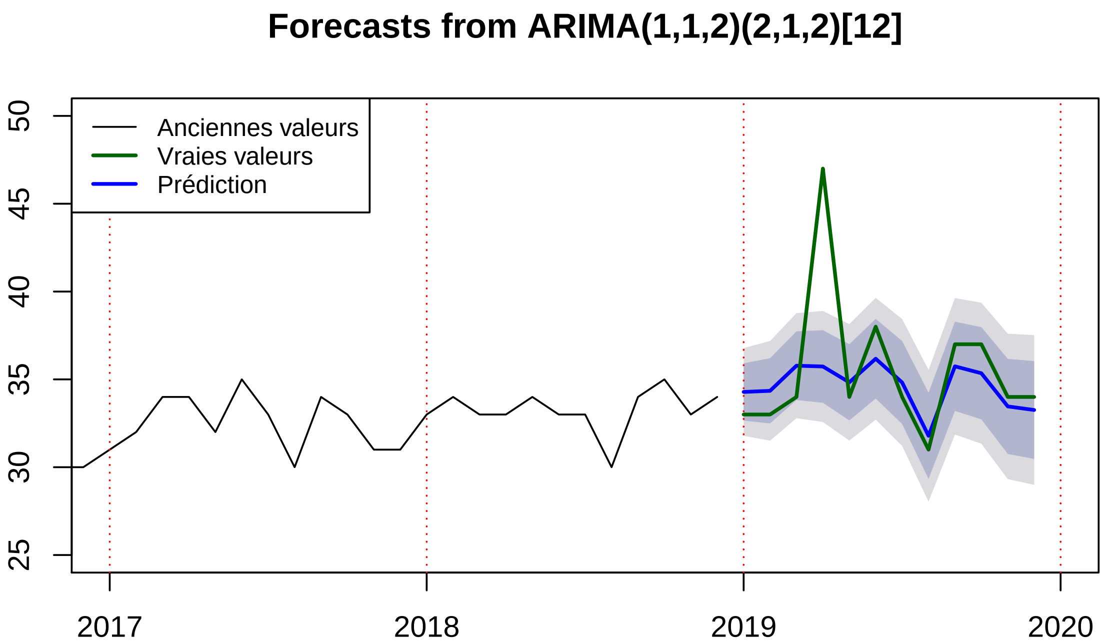
```
##              ar1       ma1       ma2      sar1     sar2      sma1
## t.stat -0.452761 -0.662009 -1.173376 -0.847288 0.443188 -0.538672
## p.val   0.650721  0.507965  0.240645  0.396835 0.657630  0.590113
##             sma2
## t.stat -0.825292
## p.val   0.409206
```

Les t-tests montre que les coefficients `sar2`, `ar1` et `sar1` ne sont pas significatifs. C'est très intéressant car ce sont exactement ces trois coefficients qui sont rajoutés au modèle automatique par rapport au modèle manuellement choisi.

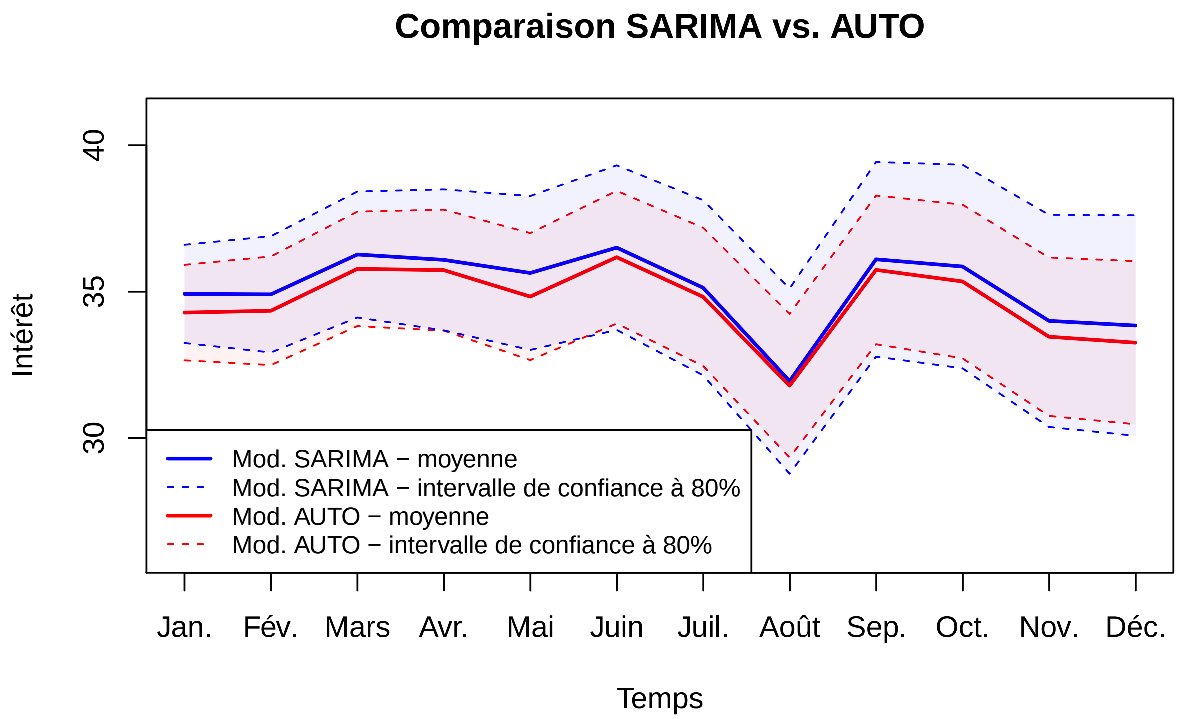
```
##             ar1        ma1        ma2       sar1       sar2       sma1
## ar1   1.0000000 -0.9824635  0.9334083  0.2254386  0.1970680 -0.2187999
## ma1  -0.9824635  1.0000000 -0.9383138 -0.2216888 -0.1995154  0.2154188
## ma2   0.9334083 -0.9383138  1.0000000  0.2246963  0.2076101 -0.2180917
## sar1  0.2254386 -0.2216888  0.2246963  1.0000000  0.8821102 -0.9895382
## sar2  0.1970680 -0.1995154  0.2076101  0.8821102  1.0000000 -0.8551768
## sma1 -0.2187999  0.2154188 -0.2180917 -0.9895382 -0.8551768  1.0000000
## sma2  0.1715275 -0.1679101  0.1699516  0.9064568  0.6504627 -0.9188064
##            sma2
## ar1   0.1715275
```

```
## ma1  -0.1679101
## ma2   0.1699516
## sar1  0.9064568
## sar2  0.6504627
## sma1 -0.9188064
## sma2  1.0000000
```

C'est de nouveau les trois nouveaux coefficients qui se manifestent. `sar1` est très fortement corrélé avec `sma2` et `sma1`, le coefficient `ar1` est corrélé à `ma1`. On regarde la prédiction pour 2019.



La prédiction pour l'année 2019 du modèle automatique est très similaire à la prédiction du modèle manuellement choisi. On fait donc une comparaison graphique des deux modèles ainsi que de leurs intervalles de confiance à 80 %.



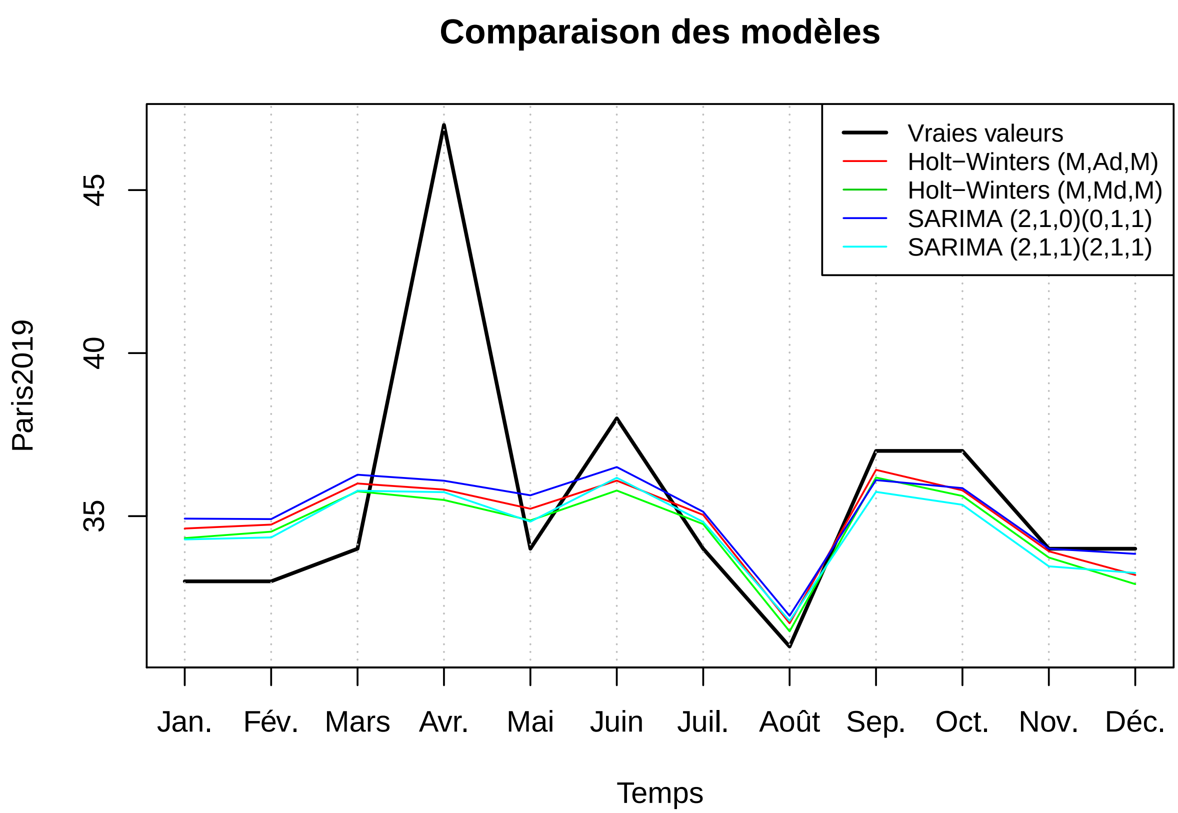
En effet, les prédictions des deux modèles sont très similaires. Les valeurs prédites par le modèle automatique sont toujours un peu plus petites. L'intervalle de confiance du modèle automatique est plus petit que celui du modèle manuellement choisi. C'est un résultat inattendu vu les écart-types plus élevés ainsi que les corrélations entre les coefficients du modèle automatique. Aussi, le modèle manuellement choisi étant plus parcimonieux on aurait plutôt attendu l'inverse.

2.4 Choix de modèle et conclusion

Pour la série `Paris` les quatre modèles suivants ont été étudiés :

Nom	Transformation	Type	Méthode	Paramètres	AIC
ParisHW_MAdM	-	Lissage exponentiel	Holt-Winters	M, Ad, M	1015.96
ParisHW_MMdM	-	Lissage exponentiel	Holt-Winters	M, Md, M	1015.47
ParisSARIMA	-	Modélisation ARMA	SARIMA	$(2,1,0)(0,1,1)_{12}$	578.32
ParisAUTO	-	Modélisation ARMA	SARIMA	$(2,1,1)(2,1,1)_{12}$	573.16

On compare ces modèles par superposition sur un plot.



On remarque que les prédictions de tous les modèles se ressemblent beaucoup et différent que très peu l'un de l'autre. Les graphes sont presque tous parallèles. L'intérêt pour le mot Paris des mois de janvier, mai, juillet, août, septembre et novembre de l'année 2019 sont très précisément prédits par tous les modèles. Le mois d'avril étant exceptionnelle n'est pas bien prédit. Vu cette similarité de tous les modèles étudiés il est difficile de choisir un modèle. Parmi les deux modèles de Holt-Winters on choisirait plutôt le modèle M,Md,M pour son AIC légèrement inférieur. Parmi les modèles SARIMA et automatique on choisirait plutôt le modèle SARIMA manuel malgré son AIC plus élevé que le modèle automatique mais qui contient des corrélations entre ses coefficients et qui est moins parcimonieux. Finalement le meilleur choix est le modèle SARIMA manuel car précis, parcimonieux avec un AIC assez bas.

3. Conclusion

Résultats intéressants, parfois inattendus Les deux séries temporelles étudiées proviennent du site https://trends.google.fr/trends/ sur lequel on peut trouver l'évolution de l'intérêt pour un mot choisi pendant une période choisie (à partir de 2004). On peut y trouver des résultats très intéressants, parfois inattendu comme l'a montré la série `Pizza` avec un intérêt plus élevé pendant la période estivale en France mais aussi la série `Paris` avec un intérêt plus bas pour le mot Paris pendant les vacances d'été et d'hiver dans l'hémisphère nord.

Influence d'événements exceptionnels bien visible Lorsque des événements exceptionnels influencent l'intérêt pour un certain mot les effets sont parfois très bien visibles comme l'a montré surtout la série `Paris`. L'intérêt dans le monde pour le mot Paris a été particulièrement élevé les mois des attentats à Paris de janvier et novembre 2015, ainsi que le mois de l'incendie de Notre-Dame de Paris en avril 2019. L'effet du championnat d'Europe de football en juin et juillet 2016 était moins visible dans les données brutes de la série `Paris` mais bien détectable lors de sa décomposition en tendance, saisonnalité et bruit.

Précision de la prédiction généralement bonne mais parfois mauvaise La précision des prédictions a généralement été bonne pour les deux séries étudiées aussi bien pour la méthode de lissage exponentiel que pour la modélisation. Même si les modèles finalement choisis sont tous les deux des modèles issus de la modélisation, la précision des modèles Holt-Winters était très bonne et un modèle Holt-Winters pourrait tout à fait convenir dans les deux cas étudiés. Parfois une transformation des données peut améliorer les résultats comme c'était le cas pour le passage au LOG avec la série `Pizza`.

Dans le cadre de ce travail on s'est limité à prédire une série temporelle uniquement par l'observation de son passé. Or, le vrai développement d'une série temporelle ne dépend pas uniquement de son passé mais aussi de facteurs externes qu'il faudrait également prendre en compte afin d'obtenir une prédiction précise. C'est probablement pour cela que les mois d'avril et d'octobre de la série `Pizza` ont été très mal prédit par tous les modèles testés. Il existe des modèles avec lesquels il est possible d'introduire des variables externes comme par exemple le modèle ARMAX (Auto Regressive Moving Average with eXogeneous inputs) qui constitue une régression linéaire avec une erreur modélisable par un modèle ARMA. Cependant, ne possédant pas de variables explicatives externes pour les deux séries modélisées une modélisation ARMAX n'a pas pu être étudiée.

Prédiction impossible d'évènements exceptionnels Comme pour toute prédiction en générale, la prédiction de l'évolution d'une série temporelle est quasiment sûre d'être fausse lorsqu'il s'agit d'évènements exceptionnels qui ne peuvent même pas être prédit par l'introduction de variables externes dans le modèle. La prédiction de la série `Paris` est un bon exemple où il était impossible de prédire l'intérêt exceptionnel d'avril 2019 pour le mot Paris suite à l'incendie de Notre-Dame de Paris.