Master 1 Économétrie et Statistiques, parcours Économétrie Appliquée

Techniques de prévision et conjoncture

Analyse et prévision des variables Google Trends et de l'indice S&P GSCI

Auteurs Laurianne MORICEAU & Diane THIERRY

Enseignant Mr O.Darné

Année universitaire 2019-2020

Résumé

L'analyse réalisée dans ce document porte sur l'étude de 2 séries chronologiques, elle vise à comprendre leurs évolutions au cours du temps et prévoir leurs variations futures. Nous nous sommes donc intéressées aux séries de recherche internet du mot 'emploi' sous Google Trends ainsi que du cours des matières premières noté S&P GSCI, l'un de janvier 2004 à mars 2020 et l'autre de janvier 2000 à décembre 2019, toutes deux mensuelles. Après avoir analysé la variable issue de l'outil Google dont les variations sont saisonnières, nous avons procédé à une désaisonnalisation à l'aide des méthodes SEATS et X-13 ARIMA sous le logiciel JDemetra. À partir de cette série dite CVS (corrigée des variations saisonnières) nous avons estimé plusieurs modèles avec différentes méthodes, différents paramètres. Au moyen de mesures de prévisions, du test de Diebold Mariano ainsi que du test multiple de Mariano et Preve nous avons pu évaluer la qualité de prévision des 8 modèles. Il apparaît que les prévisions les plus précises sont celles issues des modèles ETS et BATS, et les moins précises sont celles des modèles STLm (en se basant sur le test multiple DMD), mais aussi SARIMA(0,1,1)(0,1,1$)_{12}$ et SARIMA(1,1,2)(0,1,1$)_{12}$. Par ailleurs nous avons réalisé cette même analyse et recherche du modèle adéquat pour prédire le cours des matières premières, en incluant cette fois un modèle ARX c'est à dire qui prend en compte des variables explicatives qui, dans notre cas, étaient le cours EUR/USD et le S&P 500. Aucune des cotations boursières n'étaient stationnaires, nous avons donc modélisé leurs rendements et avons conclu que les prévisions les plus justes étaient issues du modèle ARX ainsi que des modèles AR(1) et AR(3).

Avant-propos

Les données traitées dans ce dossier ont déjà fait l'objet d'une analyse en séries temporelles dispensé au deuxième semestre du M1 EKAP. Ainsi, la première partie sur l'analyse '*Google trends*' reprend des éléments du dossier de Diane Thierry en la matière. De la même manière pour l'analyse de l'indice 'SP GSCI' étudié dans le dossier de Laurianne Moriceau.

SOMMAIRE :

I- Série saisonnière et analyse cyclique **3**

II- Série saisonnière et prévision **20**

III- Série non saisonnière et prévision **30**

V- Annexes **53**

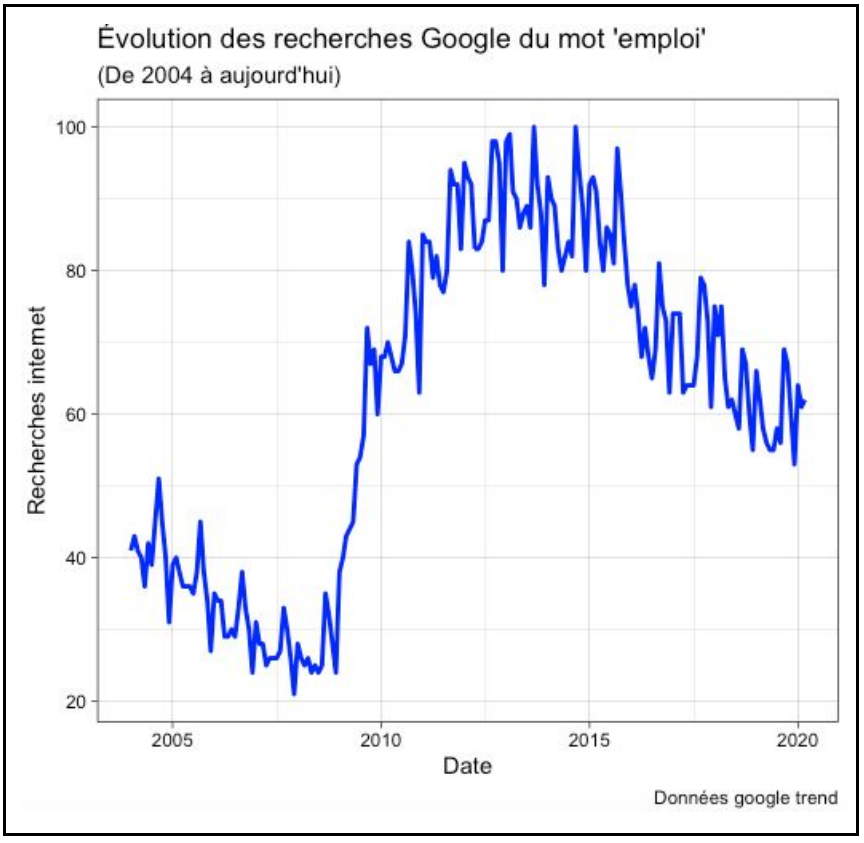
I- Série saisonnière et analyse cyclique

A) Analyse préliminaire

i. Présentation et caractérisation de la série

Dans ce dossier nous étudierons dans un premier temps les recherches du mot ‘emploi’ saisies sur le moteur de recherche ‘Google’. Ces dernières sont issues de l’outil Google Trends qui permet de voir presque instantanément la popularité relative d’un terme par le calibrage des données de 0 à 100, où 100 correspond au plus grand nombre de recherche pour une période et un lieu géographique donnés. La série récoltée pour cette analyse est mensuelle et s’étend de janvier 2004 à mars 2020 regroupant un total de 195 observations. Les données de recherche du mot spécifique ‘emploi’ reflètent l’intérêt des ménages, des individus quant à la recherche d’emploi, de travail etc. En ce sens il varie beaucoup avec les fluctuations du chômage et de l’activité économique en France.

Graphique n°1 : Évolution des recherches Google du mot ‘emploi’[1]



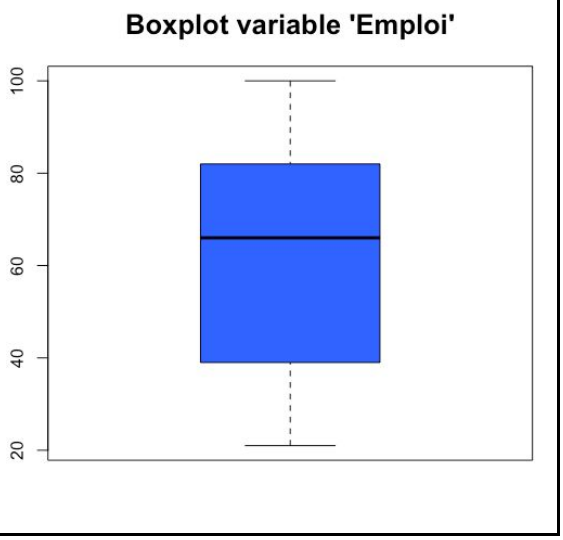
[1] Graphique initialement animé mais non supportable au format PDF ; voir R script pour l’animation

D'après le graphique n°1 ci-dessus, nous pouvons voir des comportements réguliers qui reviennent par période : il s'agit de la saisonnalité. En effet on observe un pic annuel à chaque mois de septembre lié à la rentrée scolaire, nous vérifierons statistiquement par la suite l'existence d'une saisonnalité que nous pouvons dès à présent soupçonner. De plus on peut distinguer 4 principales phases sur ce graphique :

- une première phase descendante de 2004 à 2008 où les recherches Google du mot 'emploi' ont tendance à diminuer car le chômage lui même se raréfie en France à cette période
- une deuxième phase allant de 2009 à 2012 où les recherches augmentent beaucoup et très rapidement du fait de la crise économique et mondiale des Subprimes émanant des États-Unis
- une troisième phase (2013-2015) où les recherches du mot 'emploi', qui représentent les recherches réelles des demandeurs d'emploi via internet, stagnent à un score très élevé - atteignant le score maximal de '100' en septembre 2013 comme en septembre 2014, et '97' en septembre 2015.
- enfin, les recherches diminuent de nouveau entre 2016 et 2019 et se stabilisent autour d'un indice à '60' en 2020, montrant un retour à une situation économique "normale" où les demandeurs d'emploi ne sont plus si nombreux que lors de crises économiques majeures.

Figure n°1 : Statistiques descriptives et boîte à moustache de la variable 'Emploi'

```
> round(basicStats(x$Emploi),2)
             X..x.Emploi
nobs              195.00
NAs                 0.00
Minimum            21.00
Maximum           100.00
1. Quartile        39.00
3. Quartile        82.00
Mean               61.97
Median             66.00
Sum            12085.00
SE Mean             1.65
LCL Mean           58.72
UCL Mean           65.23
Variance          531.61
Stdev              23.06
```

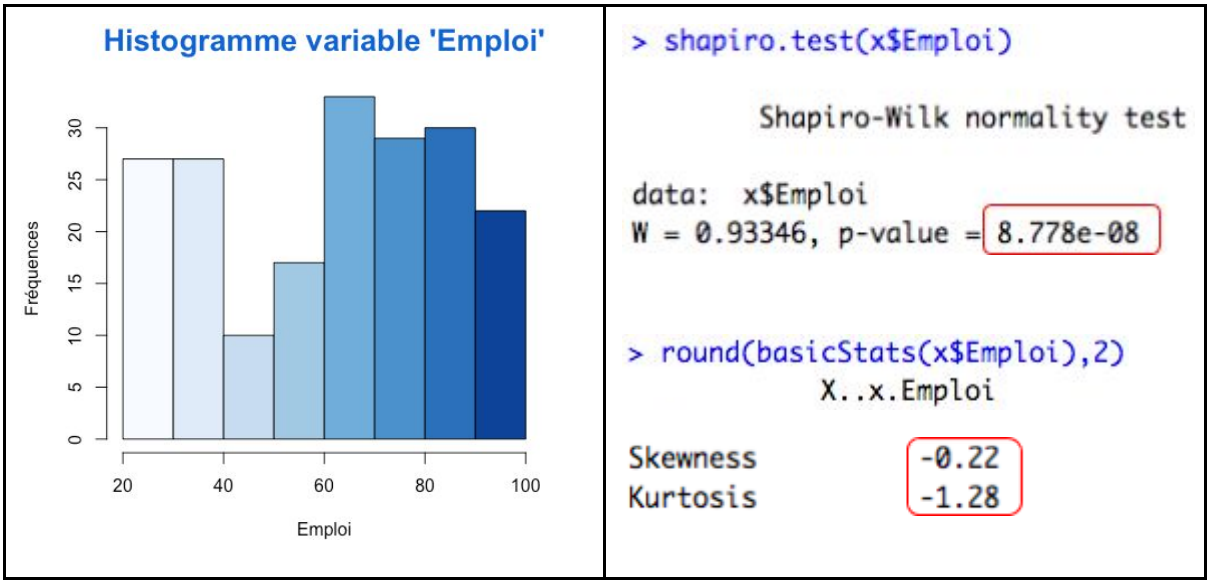


Comme expliqué précédemment, les valeurs de la série sont présentées par l'outil *'Google Trends'* comme un indice compris entre 0 et 100 où un calibrage est réalisé par rapport au

nombre total de recherches effectuées sur une période donnée et un lieu donné. On voit ainsi que les valeurs de notre série s'étendent de 21 à 100 avec une moyenne de 61,97 et une médiane de 66. N'étant pas très écartées, ces valeurs semblent indiquer qu'il n'existe pas de valeurs atypiques/aberrantes pouvant tirer la moyenne à la hausse ou à la baisse. Cela se confirme par le diagramme de Tukey disponible ci-dessus où l'on voit qu'il n'existe (visuellement) aucune valeur potentiellement atypique. Cependant, cela n'est qu'un indice et il conviendra de vérifier statistiquement l'existence de points atypiques par la méthode TSoutliers sous R, TRAMO et RegARIMA sous Stata.

On note aussi que l'écart-type est inférieur à la moyenne ($\sigma = 23,06 < M = 61,97$), cela signifie que les données sont homogènes.

Figure n°2 : Histogramme et test de Shapiro de la variable 'emploi'



Enfin, il est intéressant de regarder la distribution de ces dernières par le biais de plusieurs indicateurs :

- Les coefficients de skewness et de kurtosis qui mesurent respectivement l'asymétrie et l'aplatissement d'une distribution, ainsi, les 2 coefficients sont nuls lorsque la série suit une loi normale.
- L'histogramme de distribution, il permet de voir visuellement si les données s'apparent à la distribution d'une loi normale connue pour ses caractéristiques d'une courbe en cloche, aussi appelée "*fonction gaussienne*".
- Le test statistique de Shapiro-Wilk qui a pour but de vérifier statistiquement la distribution normale ou non d'une variable, tel que H_0 : la variable suit une loi normale.

Nous pouvons voir alors que la distribution de la variable '*Emploi*'[2] est légèrement asymétrique à droite (skewness négatif : -0,22) et est plus aplatie qu'une loi normale (kurtosis négatif : -1,28). Cela se remarque aussi visuellement sur l'histogramme de distribution. Enfin, on peut noter que le p-value associée au test de Shapiro est proche de 0, donc nous rejetons l'hypothèse nulle selon laquelle la variable est normalement distribuée.

ii. Détection de la saisonnalité et schéma de décomposition

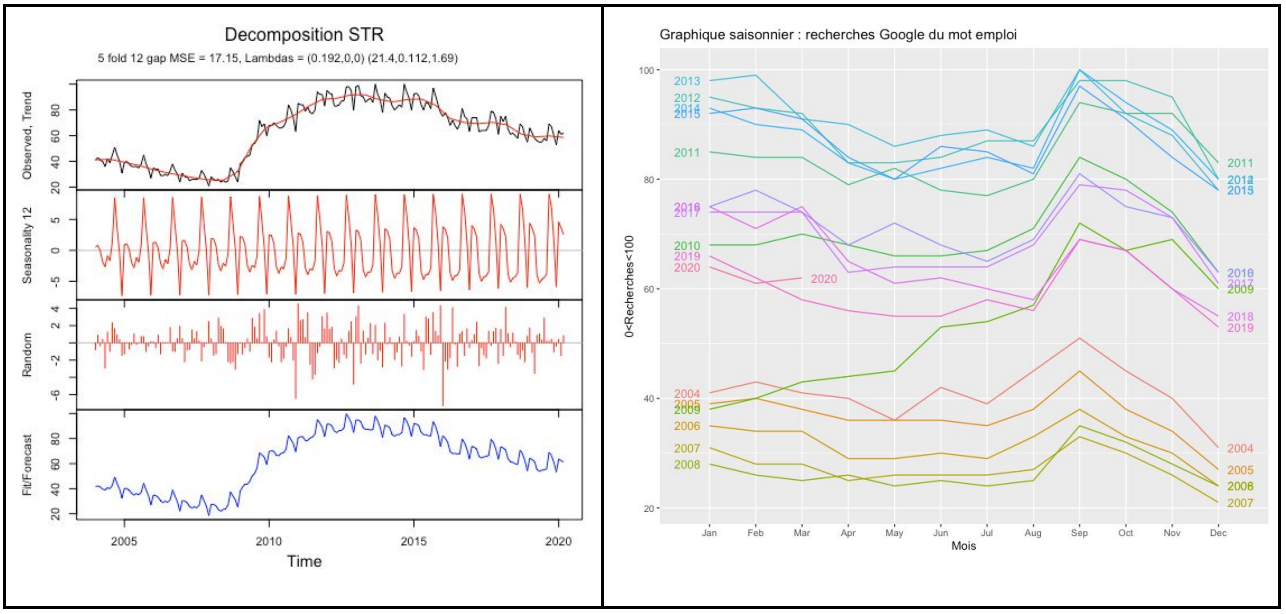
Nous allons à présent vérifier statistiquement les hypothèses émises quant à la saisonnalité de la série, puis nous étudierons la décomposition de la série pour voir s'il s'agit d'un schéma additif ou multiplicatif.

➢ Saisonnalité : Il existe un package sous R appelé "*seastests*", contenant la fonction '*isSeasonal*' qui regroupe 5 tests de détection de saisonnalité, faisant de cette fonction un test très complet. Il regroupe le test QS modifié, Friedman, Kruskal-Wallis, périodogramme ainsi que la F-test. L'existence d'une saisonnalité dans notre série a donc été confirmée par le package 'seastets' sur R ainsi qu'aux tests d'autocorrélations des résidus sous JDemetra. De plus, sur le périodogramme disponible en annexe n°1 nous pouvons voir des pics sur les premières valeurs, ce qui indique non seulement une tendance saisonnière mais également une composante déterministe dans la série.

➢ Schéma de décomposition : celui-ci représente l'interaction entre la tendance et la partie saisonnière qui peut être additive ou multiplicative, on cherche à savoir de quel schéma il s'agit pour être en capacité d'analyser efficacement notre série temporelle. Pour déterminer si la série suit un schéma additif ou multiplicatif il convient de séparer ses composantes (tendance, saisonnalité, partie aléatoire), le but étant de pouvoir tracer deux droites aux extrémités des points de la composante '*saisonnalité*' : si les droites sont parallèles la série suit un schéma additif, dans le cas contraire elle suit un schéma multiplicatif - c'est la **méthode de la bande**.

[2] Emploi : données issues de l'outil internet Google Trends - recherche du mot 'emploi' de 2004 à nos jours

Graphique n°2 : Décomposition de la série, méthodes de la bande et du profil



D'après le graphique n°2, nous pouvons voir que les variations de la série semblent augmenter dans le temps, si l'on regarde la décomposition de notre série ci-dessus on voit que les fluctuations entre chaque pic (haut et bas) s'amplifient clairement dans le temps. De plus, on voit sur le graphique saisonnier que les courbes de recherches du mot emploi par année ne sont pas parallèles entre elles (c'est la **méthode du profil)**, ce qui vient confirmer notre constat de départ : la série suit un schéma de décomposition multiplicative. Sur ce même graphique il est intéressant de noter que les recherches les plus importantes se sont effectuées en 2013 et les plus faibles en 2007-2008. Enfin, c'est la **méthode analytique** ou test de Buys-Ballot qui vient certifier notre hypothèse : à partir de l'équation $\sigma = a\overline{x} + b$ on trouve sous Excel a=0,036. Ainsi σ et $\overline{x}$ sont dépendants ; nous sommes dans le cas d'un schéma multiplicatif. Par conséquent les composantes se multiplient pour former la série temporelle, donc l'amplitude de la partie saisonnière augmente dans le temps.

iii. Détection des points atypiques par TRAMO et RegArima

Dans cette partie nous allons détecter les points atypiques en utilisant 3 méthodes différentes : TRAMO, RegARIMA et TSoutliers, que nous appliquerons grâce aux logiciels JDemetra et Rstudio. L'objectif est de soulever d'éventuelles différences entre ces 2 logiciels d'analyse. Nous commencerons pour la détection des points atypiques sous JDemetra puis nous passerons à l'analyse sur le logiciel R studio.

➢ JDemetra : Le logiciel de désaisonnalisation '*JDemetra*' utilise différentes méthodes dans l'analyse économétrique des variables.

Tableau n°1 : Points atypiques détectés par le logiciel JDemetra+

• RegArima

Outliers

	Coefficients	T-Stat	P[|T| > t]
LS (1-2016)	-12,6428	-5,14	0,0000
LS (1-2011)	11,5626	4,73	0,0000

• TRAMO

Outliers

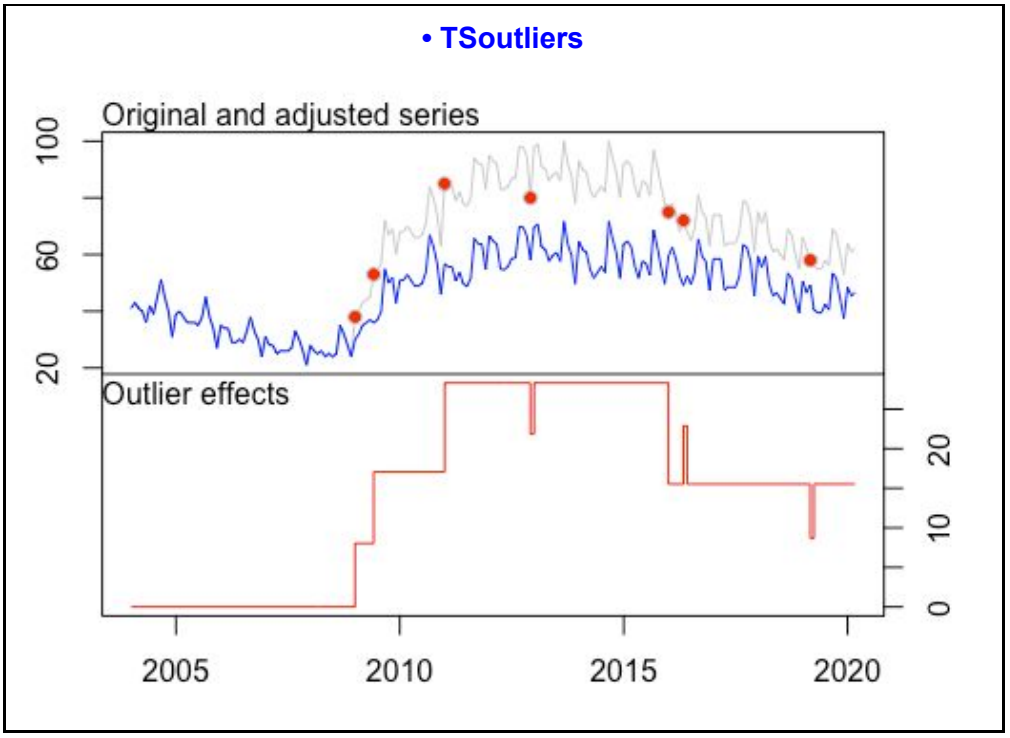
	Coefficients	T-Stat	P[|T| > t]
LS (1-2009)	0,2892	7,61	0,0000
LS (1-2016)	-0,1608	-4,24	0,0000
LS (1-2011)	0,1474	3,86	0,0002
LS (6-2009)	0,1539	4,08	0,0001

D'après le tableau n°1 reprenant les points atypiques détectés par les méthodes TRAMO et RegArima, on peut voir que l'une détecte 4 points lorsque l'autre en détecte moitié moins. Les valeurs atypiques communes aux 2 méthodes se trouvent aux mois de janvier 2011 et 2016. Comparons ces résultats à ceux du langage de programmation Rstudio.

➢ Rstudio : Le package '*tsoutliers*', qui permet de détecter les points atypiques et leurs effets sur les variations de la série, indique 7 valeurs atypiques dans la série. 4 d'entre elles sont communes aux résultats issus de JDemetra, et les 3 autres se trouvent en décembre 2012, mai 2016 et mars 2019. Il est intéressant de noter la différence dans la forme de la série avec et sans les observations atypiques, c'est à dire la série originale et ajustée.

On voit ainsi que la série ajustée est beaucoup plus aplatie que la série originale, l'effet de la crise de 2008 est donc marginal par rapport à la série brute. Les points atypiques avaient ainsi tendance à tirer la tendance vers le haut à partir de 2008. Il apparaît sur ce graphique que les points atypiques sont liés essentiellement à la crise mondiale des subprimes ; en effet les séries originales et ajustées suivent exactement les mêmes variations jusqu'en décembre 2008 où les premiers effets de la crise se font sentir.

Enfin, on observe que parmi les 7 points atypiques détectés par le package '*TSoutliers*', 4 sont de type level shifts c'est à dire que l'effet sur la série est permanent (il n'y a pas de retour à la situation "normale"), tandis que 3 sont additive outliers - le point atypique n'affecte qu'une seule observation dans la série.



Nous allons à présent essayer d'expliquer ces variations soudaines de la popularité du mot 'emploi' sous Google Trends. Le premier point atypique n'est pas difficile à comprendre puisqu'il se situe en janvier 2009, année du début de la récession en France. Les points qui suivent, c'est à dire juin 2009 et janvier 2011 sont aussi la conséquence de la crise économique puisqu'il s'agit de '*sauts*' des recherches reflétant un chômage en hausse quasi-constante depuis janvier 2009. Enfin, on peut voir un additive outlier en décembre 2012, plus difficilement interprétable, cet outlier peut être lié à la fin des calendriers Mayas, qui pour certain indiquait une fin du monde - qui irait chercher du travail si la fin était proche ? Cette hypothèse expliquerait ainsi la baisse soudaine des recherches d'emplois en France, mesurées par l'outil Google Trends, et le fait que l'effet soit instantané. Alors que le taux de chômage était en hausse depuis le début de la crise (en janvier 2009), la situation commence à se stabiliser et le taux revient à des valeurs plus 'raisonnables' : le chômage diminuant, les recherches internet du mot 'emploi' font de même et l'effet de cette baisse est permanente sur le niveau de la série. L'avant-dernier outlier se trouve en mai 2016, soit 4 mois après le début d'un retour à la normale, selon le ministère du travail "*En mai 2016, le nombre de réinscriptions sur les listes est élevé, à la suite des sorties pour défaut d'actualisation observées en avril 2016. Ceci contribue à la hausse du nombre d'entrées en*

catégories A, B, C, élevé en mai 2016, et à l'évolution du nombre de demandeurs d'emploi en catégories A, B, C entre avril et mai 2016."[3]

Finalement, c'est en mars 2019 que l'on retrouve un dernier point atypique, traduit par une baisse soudaine de la popularité du terme 'Emploi' sous google trends : le premier trimestre de 2019 a été très bénéfique en terme d'emploi et de chômage, or il est connu que le troisième mois d'un trimestre est le plus important dans la conclusion de ce dernier.

Enfin, un tableau récapitulatif reprenant les informations liées aux points atypiques détectés est disponible ci-dessous.

Tableau n°2 : Résumé des points atypiques détectés par les différentes méthodes

Observations atypiques	**RegArima (JDemetra)**	**TRAMO (JDemetra)**	**TSoutliers (Rstudio)**	**Type**
Janvier 2009		✓	✓	Level shift
Juin 2009		✓	✓	Level shift
Janvier 2011	✓	✓	✓	Level shift
Décembre 2012			✓	Additive outlier
Janvier 2016	✓	✓	✓	Level shift
Mai 2016			✓	Additive outlier
Mars 2019			✓	Additive outlier
Total	2	4	7	

Il est important de noter que le boxplot que nous avions vu précedemment indiquait l'absence d'observations aberrantes dans la série. Ainsi, même si ces méthodes considèrent ces observations comme ayant des comportements atypiques, leurs effets sur la série peuvent être insignifiants. Cependant, ces observations doivent tout de même être considérées dans la suite de l'analyse afin de ne pas avoir des estimations biaisées.

[3].https://dares.travail-emploi.gouv.fr/dares-etudes-et-statistiques/etudes-et-syntheses/dares-analyses-dares-indicateurs-dares-resultats/article/demandeurs-d-emploi-inscrits-a-pole-emploi-en-mai-2016

B) Désaisonnalisation

Afin d'analyser la série et pouvoir déceler par la suite les tendances de long terme ainsi que les différents cycles, nous procéderons d'abord à la désaisonnalisation qui consiste à atténuer les fluctuations de court terme afin d'obtenir ainsi une série corrigée des variations saisonnières (CVS). Pour cela, deux méthodes ont été utilisées : TRAMO-SEATS et ARIMA X-13 toutes deux basées sur le modèle de moyenne mobile : ARIMA.

i. TRAMO-SEATS

Nous appliquons donc la méthode paramétrique TRAMO-SEATS pour laquelle la série a été désaisonnalisée par rapport à sa moyenne. Cette méthode permet l'estimation de modèles de régression en supprimant les effets déterministes de la série notamment en traitant les observations atypiques. Ainsi, cette procédure permet d'ajuster un modèle SARIMA pertinent qui considère à la fois les composantes classiques et saisonnières.

Graphique n°3 : Désaisonnalisation avec TRAMO-SEATS

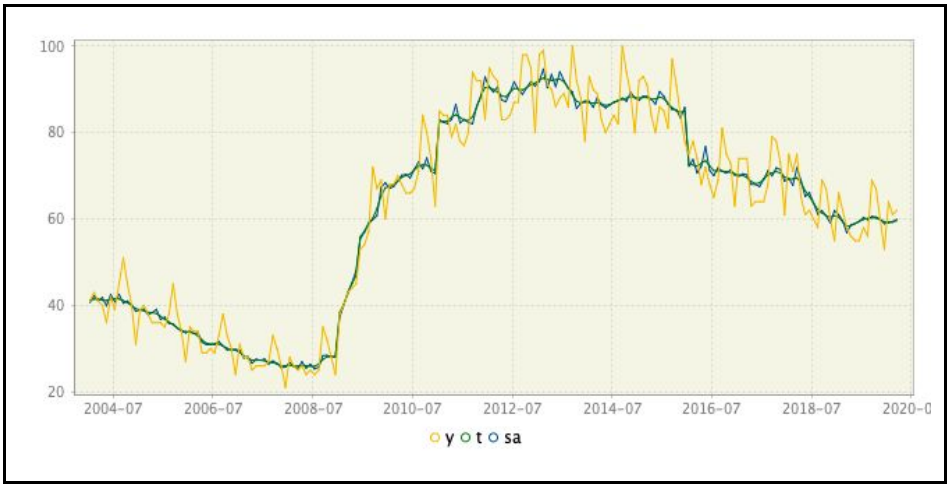


Tableau n°3 : Information liée à la désaisonnalisation avec TRAMO-SEATS

Summary	summary Good basic checks definition: Good (0,000) annual totals: Good (0,005) regarima residuals normality: Uncertain (0,039)
Estimation span: [1-2004 - 3-2020] 195 observations Series has been log-transformed Trading days effects (1 variable) No easter effect 4 detected outliers	

Final model

Likelihood statistics
Number of effective observations = 182
Number of estimated parameters = 8

Loglikelihood = 311.2736356412362
Transformation adjustment = -740.1151213737202
Adjusted loglikelihood = -428.841485732484

Standard error of the regression (ML estimate) = 0.043277226532001285
AIC = 873.682971464968
AICC = 874.5153414071646
BIC (corrected for length) = -6.080103353600968

Scores at the solution
0,005405 -0,007525

Arima model
[(0,1,1)(0,1,1)]

	Coefficients	T-Stat	P[|T| > t]
Theta(1)	-0,2189	-2,98	0,0033
BTheta(1)	-0,5277	-8,61	0,0000

independence: Good (0,610)
spectral td peaks: Uncertain (0,025)
spectral seas peaks: Good (0,447)

outliers
number of outliers: Good (0,021)

seats
seas variance: Good (0,213)
irregular variance: Good (0,600)
seas/irr cross-correlation: Good (0,995)

residual seasonality tests
qs test on sa: Good (1,000)
f-test on sa (seasonal dummies): Good (1,000)
qs test on i: Good (1,000)
f-test on i (seasonal dummies): Good (0,998)

residual trading days tests
f-test on sa (td): Uncertain (0,013)
f-test on i (td): Good (0,106)

Nous pouvons voir que SEATS a appliqué une transformation logarithmique sur la série afin de procéder à la désaisonnalisation. En outre, un modèle ARIMA(0,1,1)(0,1,1) dont le critère AIC est égale à 873.68 a été retenu. La série a donc été différenciée deux fois et peut être assimilée à une partie MA(1) classique et une partie MA(1) saisonnière. Le modèle respecte l'hypothèse d'indépendance cependant l'hypothèse de normalité des résidus est validée de manière incertaine, car elle peut être acceptée au seuil de risque $\alpha = 0,01$ mais pas au seuil auquel nous travaillons habituellement, à savoir $\alpha = 0,05$.

ii. ARIMA X-13

La méthode X-13 constitue également une méthode automatique de désaisonnalisation. C'est une méthode semi-paramétrique appliquant l'algorithme de Hannan et Rissanen dont l'objectif est de sélectionner un modèle ARIMA équilibré et parcimonieux.

Graphique n°4 : Désaisonnalisation avec X13 ARIMA

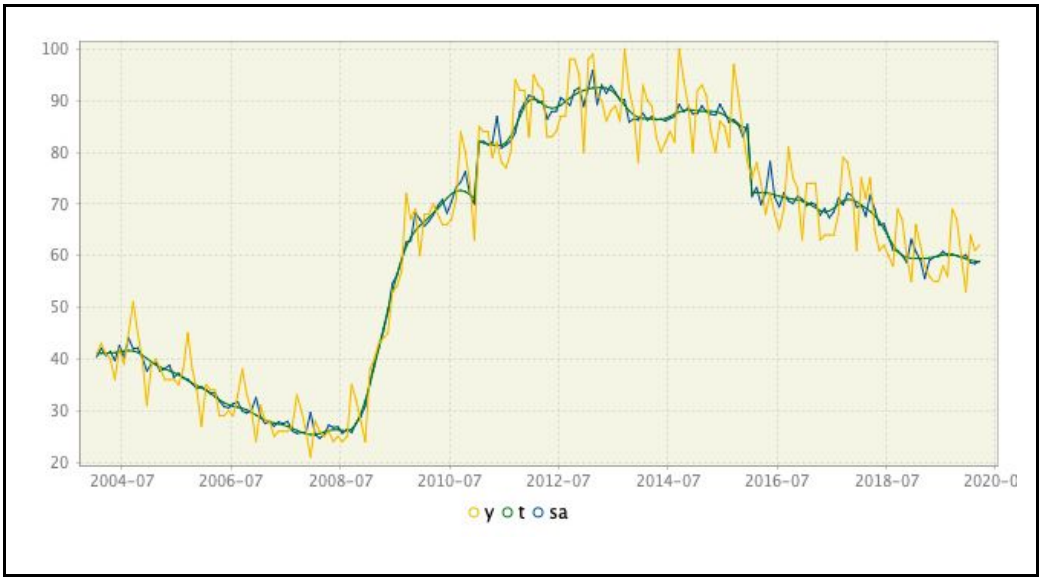


Tableau n°4 : Information liée à la désaisonnalisation avec X13 ARIMA

Summary

Estimation span: [1-2004 - 3-2020]
195 observations
Trading days effects (2 variables)
No easter effect
2 detected outliers

Final model

Likelihood statistics
Number of effective observations = 182
Number of estimated parameters = 7

Loglikelihood = -440.4323681179368
Standard error of the regression (ML estimate) = 2.67323403293979
AIC = 894.8647362358736
AICC = 895.5084143967931
BIC (corrected for length) = 2.138138635483149

Scores at the solution
0,000315 -0,006465

Arima model
[(1,1,0)(0,1,1)]

	Coefficients	T-Stat	P[|T| > t]
Phi(1)	0,1720	2,31	0,0218
BTheta(1)	-0,6438	-10,90	0,0000

summary
Good

basic checks
definition: Good (0,000)
annual totals: Good (0,001)

regarima residuals
normality: Good (0,485)
independence: Good (0,596)
spectral td peaks: Uncertain (0,019)
spectral seas peaks: Good (0,822)

outliers
number of outliers: Good (0,010)

m-statistics
q: Good (0,297)
q-m2: Good (0,331)

residual seasonality tests
qs test on sa: Good (1,000)
f-test on sa (seasonal dummies): Good (0,983)
qs test on i: Good (1,000)
f-test on i (seasonal dummies): Good (0,971)

residual trading days tests
f-test on sa (td): Bad (0,004)
f-test on i (td): Good (0,161)

Avec X-13 nous pouvons voir que la série ici n'a pas été "*log-linéarisée*". Le modèle retenu est un ARIMA(1,1,0)(0,1,1) qui respecte pleinement les conditions de normalité et d'indépendance des résidus. Ainsi, la série contient une partie classique modélisée par AR(1) et une partie saisonnière MA(1). De plus nous pouvons voir sur le graphique que le

modèle semble mieux s'adapter aux variations de la série que SEATS. Nous allons donc comparer les résultats de ces deux modèles :

Tableau n°5 : Comparaison des méthodes de désaisonnalisation

	AIC	BIC	SE	Normalité des résidus	Indépendance des résidus	Observations atypiques
SEATS	873.68	-6.08	0.04	≈	✓	4
X-13	894.86	2.14	2.67	✓	✓	2

A priori nous devrions sélectionner le modèle qui minimise les critères AIC et BIC, ce que respecte le modèle proposé par SEATS. Les critères de parcimonie et d'équilibre indiquent également que les modèles ayant le moins de paramètres sont à privilégier ainsi que ceux dont l'ordre des polynômes des AR et de différenciation sont égaux à celui du MA (nous accepterons au maximum un écart de 1 ordre). Ainsi, le modèle ARIMA(0,1,1)(0,1,1) sélectionné par SEATS doit être retenu puisque les deux modèles proposent le même nombre de paramètres et le critère d'équilibre est également validé. De plus, ce modèle a le coefficient d'erreur le plus faible. Nous avons vu cependant que l'hypothèse de normalité des résidus pouvait être rejetée. Par ailleurs, la méthode SEATS révèle 4 points atypiques quand X-13 en relève seulement 2 ; nous pouvons donc penser que le modèle issu de la méthode SEATS est plus précis, ce que vient confirmer l'écart-type de régression qui est plus faible pour ce dernier. Enfin, il est important de noter que les différenciations effectuées par ARIMA engendrent une perte d'observations (12 par exemple dans le cas d'une différenciation saisonnière pour une série mensuelle).

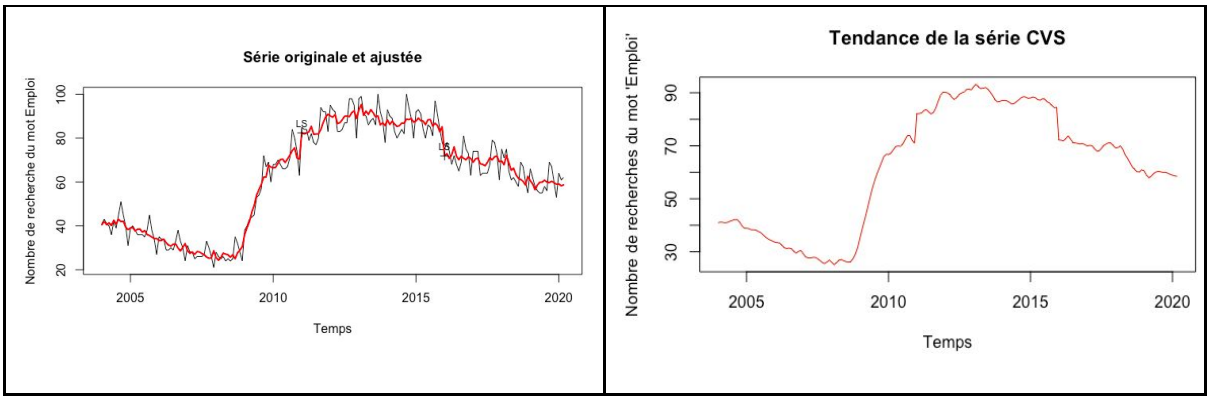
C) Analyse cyclique

i. Graphique série CVS et Trend

Dans cette partie, nous tenterons d'analyser les cycles rythmant notre série. La figure ci-dessous représente la série CVS (en rouge) *ie.* la série corrigée des variations saisonnières. Le lissage a été effectué dans ce cas sur la série brute sans correction des observations atypiques puisque comme nous l'avons dit précédemment, nous considérons

que l'influence de celles-ci n'est pas significative tandis qu'elles constituent une source d'information non négligeable.

Graphique n°5 : Série CVS et trend



➢ **Cycles des affaires** :

Tableau n°6 : Phases caractérisant le cycle des affaires

```
> show(dat)
     Peaks Troughs Duration
1   2004M8 2006M10       26
2  2006M12  2007M4        4
3   2007M6  2008M2        8
4   2008M5  2008M8        3
5   2010M9 2010M12        3
6   2011M5  2011M7        2
7  2011M12  2012M5        5
8   2013M2 2013M11        9
9   2014M2  2014M5        3
10 2014M10  2016M3       17
11  2016M5  2017M5       12
12 2017M10  2018M1        3
13  2018M3  2019M3       12
14  2019M7     <NA>     <NA>
```

```
> summary(dat)
       Phase  ]Start   ;End] Duration LevStart LevEnd Amplitude
1  Expansion    <NA>  2004M8       NA       NA     42        NA
2  Recession  2004M8 2006M10       26       42     29      12.7
3  Expansion 2006M10 2006M12        2       29     30       1.0
4  Recession 2006M12  2007M4        4       30     28       2.8
5  Expansion  2007M4  2007M6        2       28     28       0.2
6  Recession  2007M6  2008M2        8       28     25       2.7
7  Expansion  2008M2  2008M5        3       25     27       1.9
8  Recession  2008M5  2008M8        3       27     26       0.9
9  Expansion  2008M8  2010M9       25       26     74      47.8
10 Recession  2010M9 2010M12        3       74     71       2.9
11 Expansion 2010M12  2011M5        5       71     84      12.7
12 Recession  2011M5  2011M7        2       84     82       1.6
13 Expansion  2011M7 2011M12        5       82     90       8.2
14 Recession 2011M12  2012M5        5       90     87       2.7
15 Expansion  2012M5  2013M2        9       87     93       5.7
16 Recession  2013M2 2013M11        9       93     87       6.6
17 Expansion 2013M11  2014M2        3       87     87       0.5
18 Recession  2014M2  2014M5        3       87     86       1.3
19 Expansion  2014M5 2014M10        5       86     89       2.7
20 Recession 2014M10  2016M3       17       89     72      16.7
21 Expansion  2016M3  2016M5        2       72     74       1.8
22 Recession  2016M5  2017M5       12       74     68       5.6
23 Expansion  2017M5 2017M10        5       68     71       3.1
24 Recession 2017M10  2018M1        3       71     69       1.9
25 Expansion  2018M1  2018M3        2       69     70       0.7
26 Recession  2018M3  2019M3       12       70     58      12.0
27 Expansion  2019M3  2019M7        4       58     60       2.4
28 Recession  2019M7    <NA>       NA       60     NA        NA

          Amplitude Duration
Exp=]T;P]       6.8      5.5
Rec=]P;T]       5.4      8.2
```

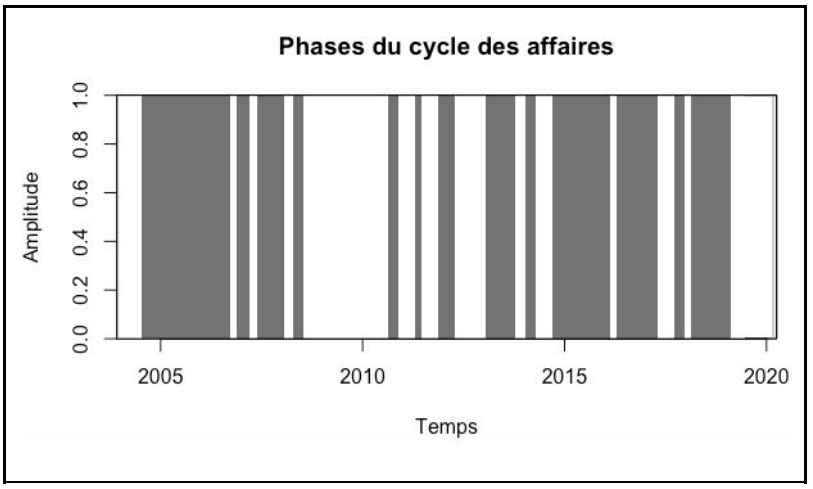
Afin de mieux cerner les comportements et tendances de la série, il convient de se pencher sur les phases de récession et d'expansion qui caractérisent le cycle classique des affaires. Ainsi, nous pouvons voir à gauche les dates des points de retournement. En outre, le pic d'Août 2004 indique le passage d'une phase d'expansion à une phase de récession jusqu'à

octobre 2006. Nous pouvons voir sur la figure ci-dessous que notre série est caractérisée par 14 cycles des affaires d'amplitude différente.

ii. Durée, amplitude, sévérité

L'amplitude moyenne des phases d'expansion est de 6.8 contre 5.4 pour les phases de récession. La durée moyenne est elle de 5.5 mois pour les phases d'expansion et 8.2 pour les phases de récession de la série. Ainsi, nous pouvons dire que les phases de récession sont en moyenne bien plus longues que les phases d'expansion et l'amplitude moyenne des phases d'expansion est également plus importante. En outre, les cycles de la série n'ont pas de comportements identiques. Nous pouvons également calculer la sévérité de ces phases, aussi connue comme "*l'approximation triangulaire des mouvements cumulés*", cette valeur se calcule par la formule suivante : S_j = *0.5*D*A*. Par exemple, la période de récession de septembre 2004 à octobre 2006 se caractérise par une sévérité de $S_j =$ 0,5*26*12,7=165,1. Ainsi, la perte totale liée à cette période de récession de 2 ans et 2 mois s'élève à 165,1€. Finalement, on voit que les phases d'expansion et de récession semblent avoir des comportements identiques puisque l'amplitude et la durée des phases d'expansion et de récession sont très proches les unes des autres.

Schéma n°1 : Phases du cycle classique des affaires

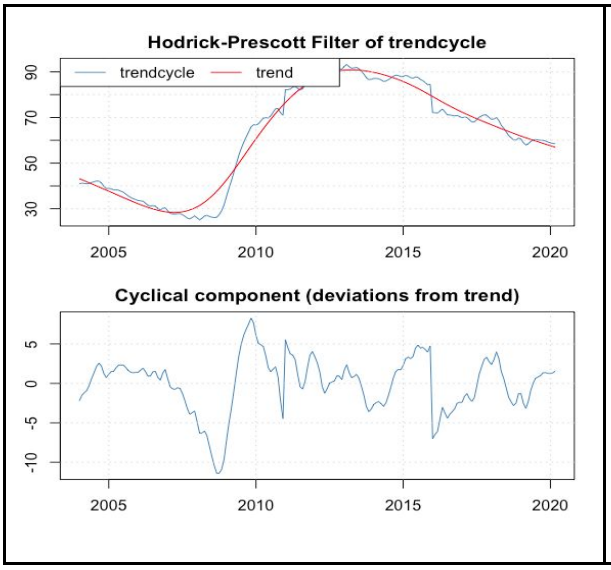


iii. Série lissée et cycle d'accélération

Afin d'extraire la composante tendancielle de notre série nous appliquons le filtre Hodrick-Prescott (HP) pour lequel λ, le filtre de lissage, est égal à 14400 car notre série contient des données mensuelles. A la différence de la série CVS observée précédemment,

nous pouvons voir la série lissée en rouge sur la figure ci-dessous. En effet celle ci est plus lisse puisqu'elle a été corrigée des tendances cycliques. La tendance de la série CVS suit les fluctuations de la série originale, sa courbe n'est donc pas lisse contrairement à la tendance issue du filtre Hodrick Prescott où l'on ne peut noter aucune variation en dehors du mouvement tendanciel de la série.

Figure n°3 : Application du filtre HP et description de ses composantes

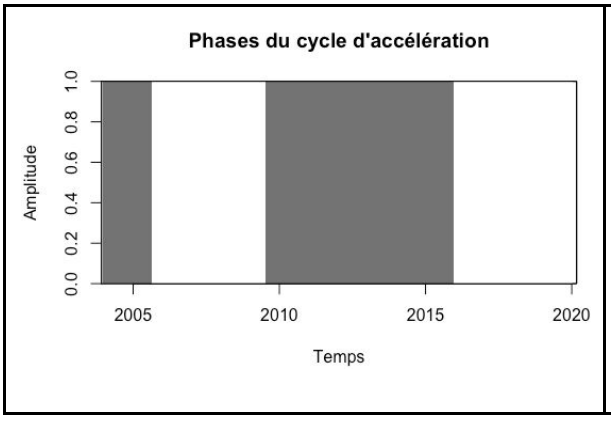



```
Series:
 trendcycle

Descriptive Statistics:

   trendcycle        Trend           Cycle
 Min.   :25.16   Min.   :28.39   Min.   :-11.3969
 1st Qu.:38.88   1st Qu.:39.32   1st Qu.: -2.2416
 Median :68.09   Median :65.03   Median :  0.7623
 Mean   :61.93   Mean   :61.93   Mean   :  0.0000
 3rd Qu.:84.92   3rd Qu.:82.32   3rd Qu.:  2.0560
 Max.   :93.19   Max.   :90.87   Max.   :  8.2862

In-sample error measures:
        ME        MSE        MAE        MPE       MAPE
 1.198e-15  1.258e+01  2.741e+00 -1.266e-02  5.694e-02
```

Il est intéressant de confirmer ces constats visuels à l'aide de statistiques sur les 3 séries ; nous pouvons voir que l'étendue de la série '*trendcycle*' (écart entre les valeurs minimales et maximales) est plus importante que la série '*trend*' pour laquelle, par conséquent, la moyenne et la médiane sont plus proches l'une de l'autre.

➢ **Cycles d'accélération** :

Schéma n°2 : Phases et tendance du cycle d'accélération



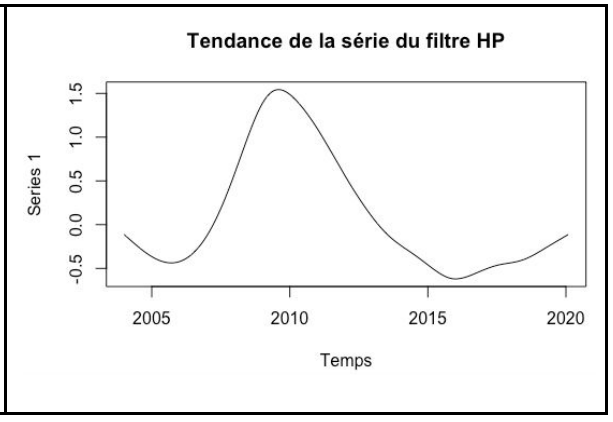


Tableau n°7 : Phases caractérisant le cycle d'accélération

```
> show(dat)
   Peaks Troughs Duration
1   <NA> 2005M10     <NA>
2 2009M9  2016M2       77

> summary(dat)
      Phase  ]Start   ;End] Duration LevStart LevEnd Amplitude
1 Recession    <NA> 2005M10       NA       NA      0        NA
2 Expansion 2005M10  2009M9       47        0      2       2.0
3 Recession  2009M9  2016M2       77        2     -1       2.2
4 Expansion  2016M2    <NA>       NA       -1     NA        NA

          Amplitude Duration
Exp=]T;P]       2.0       47
Rec=]P;T]       2.2       77
```

Contrairement au cycle des affaires pour lequel il existait 28 points de retournement (points marquant le début et la fin d'une phase de récession), le cycle d'accélération, lui, en compte seulement 4 ; il y a donc 2 périodes de récession. La première, dont la date de début est inconnue, s'étend jusqu'à octobre 2005, et la deuxième plus importante, s'étend de septembre 2009 à février 2016. Cette dernière coïncide avec l'arrivée de la récession en France après l'éclatement de la bulle immobilière aussi appelée crise des Subprimes. L'amplitude moyenne des périodes de croissance positive et négative correspond en réalité à celle des phases d'expansion novembre 2005-septembre 2009 et de récession octobre 2009-février 2016, puisque les autres n'ont pas de début ou de fin définies. Ainsi l'amplitude des phases (ou de LA phase) d'expansion est de 2 points, tout juste moins que celle de la phase de récession qui est de 2,2 points, ce qui reste 2 voire 3 fois plus faible que celle du cycle des affaires.

En revanche la durée des phases est beaucoup plus importante pour le cycle d'accélération que pour le cycle des affaires : la phase de croissance positive dure 47 mois soit près de 4 ans, et la phase de croissance négative dure elle 77 mois soit plus de 6 ans. Aussi, en se penchant uniquement sur l'amplitude des phases on pourrait dire que les cycles d'expansion ou de récession ont des comportements identiques, en revanche lorsque l'on regarde la durée qui diffère grandement entre les phases, nous pouvons affirmer qu'il existe une certaine asymétrie entre ces dernières. Finalement, nous regroupons l'information liée à ces cycles dans un tableau récapitulatif afin de voir clairement les points communs et les différences entre les 2 cycles étudiés.

Tableau n°8 : Principales caractéristiques des cycles

	Cycle des affaires	**Cycle d'accélération**
Nombre de points de retournement	28	4
Amplitude moyenne	Expansion : 6,8 Récession : 5,4	Expansion : 2 Récession : 2,2
Durée moyenne (en mois)	Expansion : 5,5 Récession : 8,2	Expansion : 47 Récession : 77
Sévérité (récession)	09/2004 - 10/2006 :165,1	10/2009 - 02/2016 : 84,7

II- Série saisonnière et prévision

A) Prévisions récursives avec un pas de 1 mois

i. Prévisions avec TRAMO

Le logiciel JDemetra permet de réaliser des prévisions avec les méthodes TRAMO-SEATS et X-13 Arima. Seulement, nous souhaitons prévoir la série sur une année. En outre, nous souhaitons faire nos prévisions avec un pas de 1 mois, ce qui est plus adaptée pour les prévisions de long terme car ce processus à mémoire longue consiste à prédire mois par mois. Cependant, le logiciel JDemetra+ n'utilise pas cette méthode et prédit d'une traite l'année entière. Nous avons donc retiré au préalable les 12 derniers mois de la série afin de pouvoir comparer les qualités des prévisions de la méthode récursive que nous pourrons ensuite comparer aux valeurs prédites par JDemetra+. La série utilisée sera alors la série désaisonnalisée par ce logiciel avec TRAMO, et corrigée des outliers considérés pour cette méthode (cf a.iii). Les résultats pour le modèle SARIMA(1,1,0)(0,1,1 $)_{S=12}$ sélectionnés par TRAMO SEAT sont résumés dans le tableau ci-dessous :

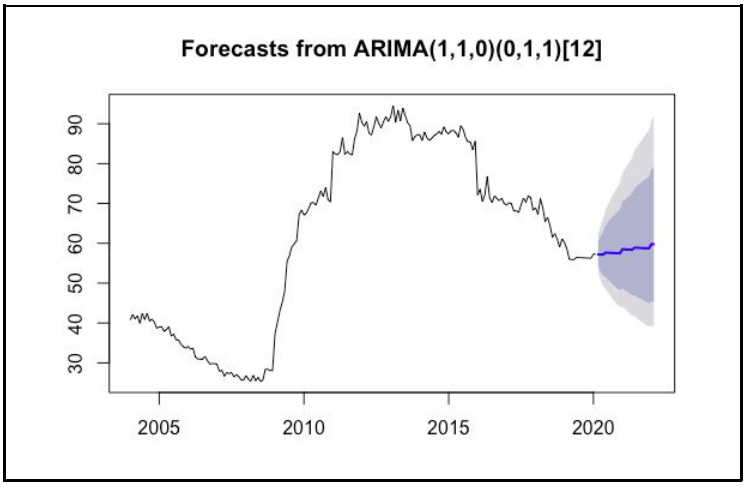
Tableau n°9 : Prévisions SARIMA des données google trends

	Prévisions avec un pas d'un mois	**Prévisions JDemetra+**
Avril 2019	55.94	53.45
Mai 2019	55.92	51.33
Juin 2019	56.45	50.47
Juillet 2019	56.42	51.30
Août 2019	56.39	49.91
Septembre 2019	56.36	59.86
Octobre 2019	56.33	58.03
Novembre 2019	56.30	52.34
Décembre 2019	56.26	47.47
Janvier 2020	57.31	56.55

Février 2020	57.26	53.41
Mars 2020	57.20	52.91
MSE	**32.10**	**47.62**

Nous pouvons voir dans le tableau ci-dessus que l'erreur quadratique moyenne (MSE) est moins importante pour la prévision récursive que pour la prévision faite pour l'année entière par le logiciel JDemetra+, ce qui veut dire que les valeurs prédites par la prévision récursives sont en moyenne plus proches de celles prédites d'une traite par le logiciel.

Graphique n°6 : Prévisions récursives sur série désaisonnalisée par SEATS



ii. Prévisions avec Holt-Winters, SARIMA, BATS, ETS, STLm et NNETAR

Afin de produire la meilleure estimation possible de notre série concernant le nombre de recherches du mot "*emploi*" sur google, nous allons procéder à plusieurs techniques de prévisions. La première est la prévision avec Holt-Winters ; une méthode de lissage exponentiel dont l'objectif est de réaliser des prévisions à court terme. Ainsi, l'approche prévisionnelle par récursive est importante. Notre série étant saisonnière, nous réaliserons un lissage saisonnier multiplicatif automatique à l'horizon de 1 an. Ensuite, nous procéderons aux prévisions faites par la méthode Arima. Nous comparerons la prévision faite par un modèle Arima sommaire : Arima(0,1,1)(0,1,1), soit avec une partie moyenne mobile classique et saisonnière, et un modèle Arima sélectionné par la fonction qui maximise le critère AIC : Arima(1,1,2)(0,1,1 $)_{S=12}$. Pour finir, nous effectuerons nos prévisions avec les méthodes BATS, ETS, STLm et NNETAR toujours à l'aide d'une

approche récursive avec un pas d'un mois. L'objectif final étant de pouvoir comparer ces différentes méthodes afin d'obtenir la meilleure estimation possible. Les séries utilisées sont brutes, lissées ensuite par chacune de ces méthodes.

Tableau n°10 : Prévisions de la série google trends à l'aide différentes méthodes

	Holt-Winters saisonnier	SARIMA(0.1,1)(0,1,1)s=12	SARIMA(1,1,2)(0,1,1)s=12	BATS	ETS	STLm	NNETAR
Avril 2019	52.616	52.185	51.561	54.136	54.926	52.449	58.234
Mai 2019	51.596	51.203	49.890	52.964	54.927	50.859	59.419
Juin 2019	52.241	51.604	50.123	53.165	56.312	51.241	60.79
Juillet 2019	51.161	50.746	48.917	52.843	55.614	50.634	60.32
Août 2019	51.373	51.043	48.956	52.648	57.001	50.508	61.45
Septembre 2019	60.679	60.320	61.041	64.461	67.275	63.541	63.20
Octobre 2019	58.062	57.905	57.229	60.523	61.881	59.257	63.13
Novembre 2019	53.160	53.785	52.037	55.953	57.416	54.761	63.00
Décembre 2019	47.168	47.535	43.272	47.467	50.099	45.359	62.69
Janvier 2020	55.143	55.958	53.921	57.470	60.723	56.582	63.39
Février 2020	52.255	54.211	51.995	55.479	59.413	55.617	62.59
Mars 2020	51.173	53.146	50.588	53.894	58.295	54.077	62.47

B) Comparaison des modèles Arima

Nous avons pu donc comparer trois modèles Arima différents :

- SARIMA(0,1,1)(0,1,1 $)_{S=12}$
- SARIMA(1,1,2)(0,1,1 $)_{S=12}$ proposé par la fonction auto.arima
- SARIMA(1,1,0)(0,1,1 $)_{S=12}$ proposé par SEATS pour la prévision

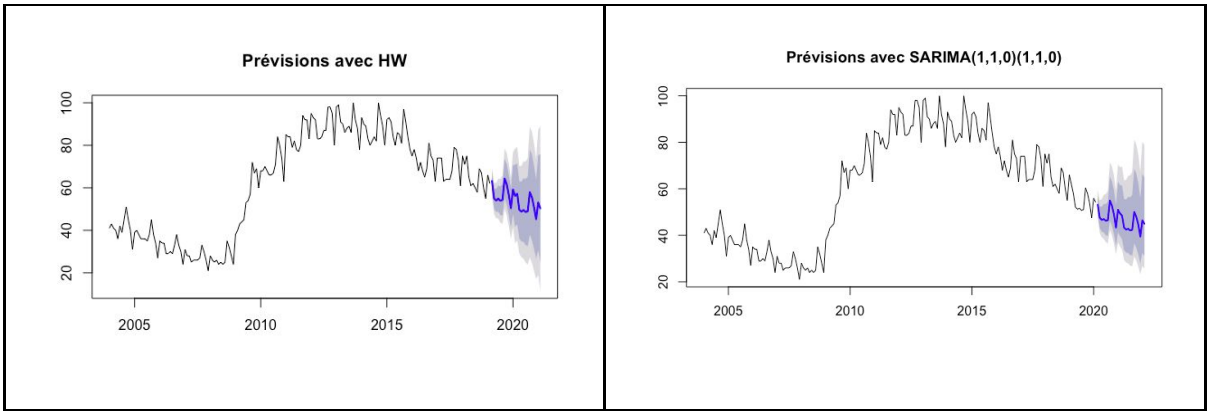
Ainsi, nous pouvons voir que le modèle sélectionné par auto.arima a plus de paramètres que celui proposé par SEATS. Avant de pouvoir interpréter ceux-ci, il convient d'abord de vérifier les conditions de stationnarité et d'inversibilité de nos processus arima.

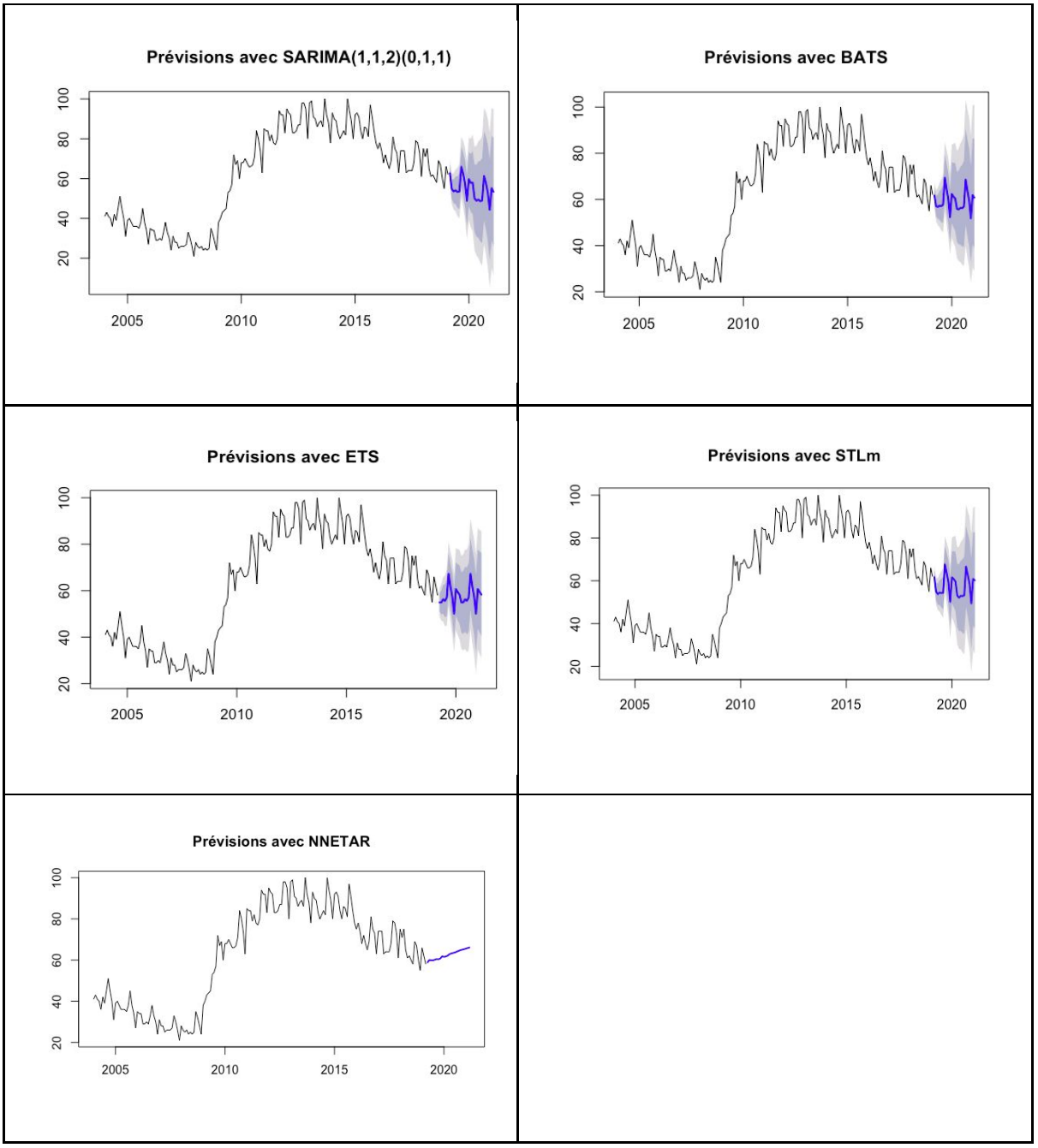
Pour le premier modèle, le paramètre MA(1) : θ_1= |-0.138| < 1, ainsi la condition d'inversibilité est respectée pour ce modèle. De plus les paramètres MA(1) et SMA(1) sont tous deux significatifs. Pour le modèle selectionné par '*auto.arima*', le paramètre θ_1=|-1.2|> 1, ce qui veut dire que la condition d'inversibilité n'est pas respectée et les paramètres de ce modèle ne sont donc pas stables. En revanche, pour le dernier modèle, les conditions de stationnarité et d'inversibilité sont toutes deux respectées avec : ϕ_1= |0.22| < 1 et θ_1= |-0.5| < 1. Ainsi nous pouvons comparer le premier et le dernier modèle. Ayant le même nombre de paramètres nous ne pouvons différencier les modèles sur le critère de parcimonie. Cependant, le modèle SARIMA(1,1,0)(0,1,1$)_{12}$ proposé par SEATS a un critère AIC égal à 824.78, supérieur au modèle ARIMA(0,1,1)(0,1,1$)_{12}$ dont l'AIC est égal à -513.95. Ainsi, à choisir, le modèle proposé par SEATS semble être le meilleur.

C) Graphique des prévisions

Les graphiques ci-dessous représentent les différents modèles utilisés pour les prévisions. Nous pouvons voir que les trois premiers modèles, à savoir le modèle Holt-Winters et les modèles ARIMA, prédisent une tendance à la baisse pour la recherche du mot '*emploi*' à partir du mois d'avril 2019. De plus, les variations sont presques indifférenciables entre ces modèles. On observe tout de même une amplitude plus importante pour les prédiction du modèle selectionné par ARIMA(1,1,2)(0,1,1$)_{12}$. En revanche, les modèles BATS, ETS, STLm et NNETAR prévoient une tendance générale à la hausse. En outre, le modèle NNETAR qui n'est pas déterministe propose des prévisions bien différentes des autres.

Tableau n°11 : Représentation graphique des différentes prévisions





D) Erreurs de prévision

Afin de comparer les estimations faites par les différents modèles, nous allons confronter les prévisions obtenues par modélisation et les valeurs réelles observées. Ainsi, nous avons quatre types d'erreurs de prévision qui nous donne une indication sur la qualité des modèles : l'erreur moyenne (ME), l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (MSE) et l'erreur médiane (MDAE).

Tableau n°12 : Erreurs de prévision de chacun des modèles

	HW	SARIMA(0,1,1)(0,1,1)s=12	SARIMA(1,1,2)(0,1,1)s=12	BATS	ETS	STLm	NNETAR
ME	-6.14	-6.36	-8.04	-0.84	-1.46	-1.46	2.14
MAE	6.61	6.36	8.04	0.84	1.46	1.46	2.14
MSE	50.07	44.34	69.30	1.72	4.63	4.63	21.42
MDAE	-6.84	-6.49	-8.48	-0.85	-1.50	-1.50	2.27

Le tableau n°12 ci-dessus nous indique que les modèles Arima et Holt-Winters comparés aux autres, ont des erreurs bien plus élevées. Le modèle BATS est celui qui minimise les erreurs avec une erreur moyenne de -0.84 et une erreur quadratique moyenne de 1.72. Cependant, il est étonnant de voir que le modèle SARIMA sélectionné par la fonction auto.arima a une erreur plus élevée que le modèle Arima basic. C'est même le modèle dont l'erreur de prévision est le plus élevé. Cela peut s'expliquer par le fait que le modèle contienne beaucoup de paramètres. De plus, nous avions vu que ce modèle ne respectait pas la condition d'inversibilité. Nous pouvons voir également avec l'erreur moyenne que tous les modèles excepté NNETAR prédisent en moyenne des valeurs supérieures aux valeurs observées de la série.

En effet, nous pouvons également comparer ces modèles à un modèle de prévision de référence pour savoir si ceux-ci ont un intérêt. Nous utiliserons ici un modèle de prévision naïve saisonnier où l'observation en t+1 est égale à l'observation précédente en t. Ainsi, nous obtenons les résultats suivants :

Tableau n°13 : Taux d'erreur de la prévision naïve

	ME	**MAE**	**MSE**	**MDAE**
Naïve	2.25	2.25	16.58	2

Ainsi, nous pouvons voir que les modèles HW, SARIMA(0,1,1)(0,1,1 $)_{12}$ et SARIMA(1,1,2)(0,1,1) n'ont pas d'intérêt puisque une simple prévision permet de meilleurs résultats. De plus, nous pouvons voir que les prédictions faites par BATS, ETS, STLm et NNETAR sont meilleures, et donc avec des erreurs de prévision plus faibles. En conclusion,

nous savons que le modèle BATS est à priori celui dont les prévisions sont les plus justes sur cet intervalle d'un an du moins.

E) Tests et sélection du meilleur modèle

i. Diebold-Mariano

Afin de savoir lequel des modèles est le meilleur, nous avons procédé au test de Diebold-Mariano, d'abord par rapport à la prévision naïve, puis entre chaque modèle. Celui ci permet de savoir si deux modèles sont équivalents en terme de qualité de prédiction, basé sur une fonction de perte calculée à partir des erreurs de prévisions. Ainsi, nous testons l'hypothèse H_0 qui suppose que les deux modèles ont des qualités de prédiction égales.

Tableau n°14 : DM test entre les différents modèles et les prévisions naïves

	HW	SARIMA(0,1,1)(0,1,1) s=12	SARIMA(1,1,2)(0,1,1 s=12	BATS	ETS	STLm	NNETAR	SEAT
p-value	0.0487	0.034	0.044	0.034	0.034	0.041	0.054	0.034

L'alternative "*less*" ajoutée à la fonction dm.test implique l'hypothèse alternative H_1 selon laquelle la qualité de précision du modèle 2 (prévision naïve ici) est moins bonne que celle du modèle 1. Nous pouvons voir que tous les modèles excepté NNETAR ont une qualité de précision supérieure à la prévision naïve. En effet, la p-value associée au dm.test entre ces deux modèles est supérieure à 5% ce qui nous amène à accepter H_0 ; le modèle NNETAR a une qualité de prévision similaire à celle du modèle naïf au seuil de 5%.

Nous allons donc maintenant effectuer les DM tests entre ces modèles pour savoir lequel retenir. Chaque test a été effectué dans les deux sens puisque l'hypothèse alternative H_1 repose sur le fait que le deuxième modèle est moins précis, ainsi, le faire dans un seul sens ne permettrait pas de savoir si l'hypothèse a été refusée car le premier modèle est moins précis ou parce qu'il n'existe pas de différence significative entre ceux-ci.
Ainsi, à l'aide du tableau ci-dessous, nous pouvons conclure d'une part que le modèle HW a une qualité de prédiction équivalente au modèle selectionné par auto arima : SARIMA(1,1,2)(0,1,1). En revanche, les prédictions du modèle Arima(0,1,1)(0,1,1), le

modèle BATS, ETS, STLM, et SEAT sont plus précises que celles du modèle HW. Cependant, le modèle Arima(0,1,1)(0,1,1) est également moins précis que les modèles BATS, ETS, et STLM. Le modèle Arima et SEAT semblent égales en terme de qualité de prédiction. Pour finir, le modèle BATS semble être plus précis que tous les autres modèles, à l'exception de ETS, pour lequel nous trouvons que le dm.test trouve des résultats différents dans les deux sens. Ces résultats semblent cohérents puisque nous avions émis des conclusions similaires à l'étude des erreurs moyennes de chaque prévision.

Tableau n°15 : Test de Diebold-Mariano entre les modèles

	HW	SARIMA(0.1,1)(0,1,1)s=12	SARIMA(1,1,2)(0,1,1)s=12	BATS	ETS	STLm	SEAT
HW		1 ✗	0.8892 ✗	0.9999 ✗	0.9999 ✗	0.9714 ✗	0.9999 ✗
SARIMA(0.1,1)(0,1,1)s=12	4.101e-05 ✓		1 ✗	5.067e-11 ✓	0.000154 ✓	4.283e-07 ✓	0.2028 ✗
SARIMA(1,1,2)(0,1,1)s=12	0.1108 ✗	1 ✗		1 ✗	1 ✗	0.9937 ✗	1 ✗
BATS	0.0001208 ✓	1 ✗	1 ✗		1.144e-10 ✓	1.574e-06 ✓	1.027e-10 ✓
ETS	7.7e-05 ✓	1 ✗	1.961e-05 ✓	1.144e-10 ✓		4.295e-07 ✓	0.2528 ✗
STLm	0.02864 ✓	1 ✗	0.01266 ✓	1 ✗	1 ✗		1 ✗
SEAT	7.694e-05 ✓	0.3085 ✗	9.796e-06 ✓	1.027e-10 ✓	0.5056 ✗	8.568e-07 ✓	

Ainsi, le test multiple Mariona et Preve nous permet de déterminer le meilleur modèle parmi ceux-ci.

ii. Mariano et Preve

Nous réalisons à présent le test multiple de Mariano et Preve qui se base sur les hypothèses suivantes :

H_0 : Les prévisions évaluées ont la même précision

H_1 : L'Equal predictive accuracy (EPA) n'est pas atteinte

Tableau n°16 : Résultat du test multiple Mariano-Preve

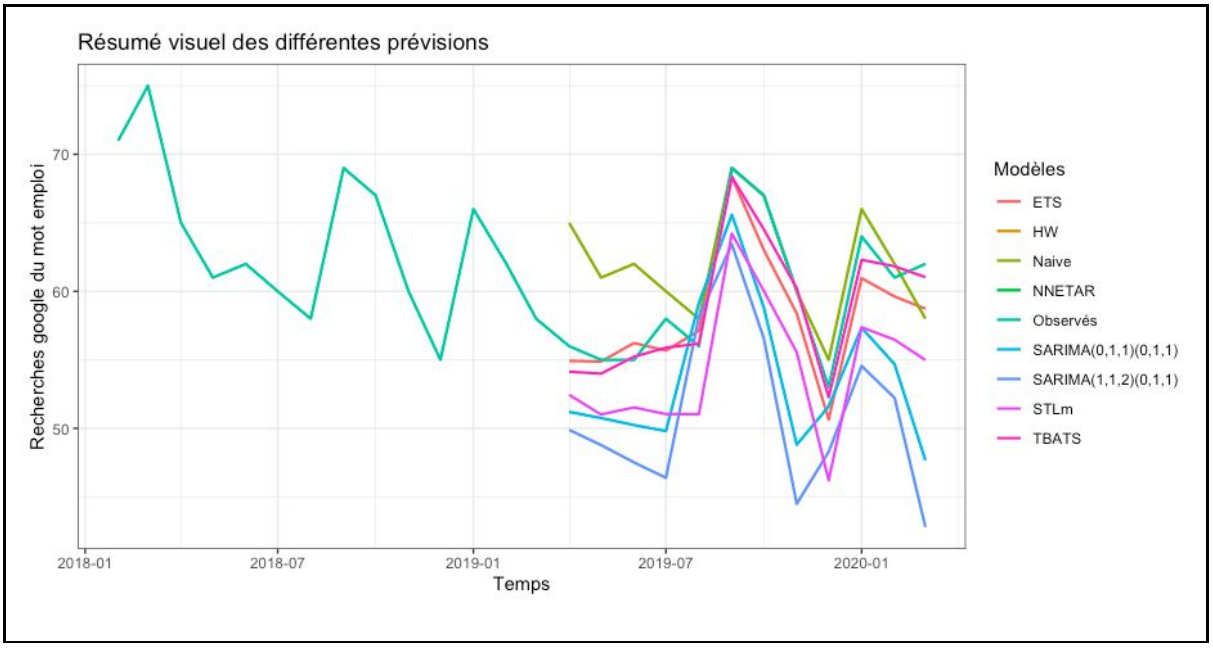
```
> MDM.selection(realized=ts,evaluated=previsionsm,q=q,alpha=0.1,statistic="Sc",loss.type="SE")
#####################################################
Models with outstanding predictive ability:

      Rank      Sc Mean loss
prev     3  0.7835   50.0740
prev3    2 -1.8517    1.7226
prev4    5 -3.2550    4.6278
prev6    4  0.4466   21.8098
prev7    1 -0.6592   16.5833
prev8                32.0965

p-value: 1

Number of eliminated models:  3
#####################################################
```

Nous pouvons voir dans le tableau ci-dessus que trois modèles ont été éliminés par la fonction. La première prévision du modèle Arima(1,1,0)(0,1,1), la deuxième prévision du modèle Arima(1,1,2)(0,1,1), puis la prévision 5 du modèle STLm. Ainsi, ces modèles ont été considérés par la fonction comme étant moins bons. La p-value étant supérieure à 0.05, nous pouvons conclure que les modèles restants ont des qualités de prédiction équivalentes. Ainsi, la fonction a effectué un classement en fonction de la qualité prévisionnelle de chaque modèle. Du meilleur au moins bon nous retrouvons : la prévision naïve, le modèle BATS, le modèle NNETAR, puis le modèle ETS. En outre, les conclusions tirées de la fonction sur les modèles Arima sont les mêmes, cependant, elles diffèrent pour les modèles naïf et NNETAR. En effet, ce test ne considère que les résidus du modèle et ne prend pas en compte les valeurs observées de la série. Le graphique qui suit nous permet de conclure sur les modèles qui se rapprochent réellement des observations.

F) Graphique récapitulatif

Pour conclure, nous pouvons observer le graphique ci-dessous qui représente les prévisions des différents modèles effectués. Nous pouvons voir que celles-ci semblent suivre toutes les mêmes tendances. Comme nous l'avons conclu précédemment, nous pouvons voir que le modèle TBATS semble suivre assez fidèlement les valeurs observées de la série. C'est d'ailleurs après analyse des erreurs moyens et du test de Diebold Mariano, le meilleur modèle parmis ceux proposés. Les modèles Arima restent eux assez éloignés des observations avec des prévisions inférieures.

Graphique n°7 : Visualisation des différentes prévisions réalisées



III- Série non saisonnière et prévision

A) Analyse préliminaire

i. Présentation et caractérisation de la série étudiée

Dans cette partie nous allons réaliser l'étude de séries non saisonnières ; pour cela nous nous pencherons sur 3 d'entres elles : Y_t : l'indice boursier des matières premières noté **S&P GSCI**, il a été développé par la banque d'investissement Goldman Sachs en 1991 sous le nom de Goldman Sachs Commodity Index (GSCI) et sert aujourd'hui de référence pour les investissements sur les marchés des matières premières car c'est un indicateur de performance des '*commodities*' dans le temps. En outre, un indice est un portefeuille hypothétique d'actions d'une ou plusieurs places boursières[4]. Celui-ci regroupe les 24 principales matières premières de tous secteurs dont l'énergie, l'agriculture et le bétail[5]. Ainsi, l'indice tente de reproduire la tendance générale des prix sur les marchés dérivés de produits agricoles et énergétique, nous donnant ainsi une idée de l'évolution globale de la volatilité des prix.

Nous verrons plus tard dans cette analyse que ce dernier peut être expliqué par 2 variables : le taux de change Euro/Dollars noté EUR/USD, ainsi que le S&P 500 qui regroupe les 500 entreprises américaines les plus importantes et couvre près de 80% du marché boursier américain. En effet, ces deux variables ont une influence sur les variations de l'indice qui est de fait très volatil, et ce notamment depuis le phénomène de financiarisation car celui-ci peut-être utilisé comme manière de diversifier son portefeuille.

Les données s'étendent de janvier 2000 à décembre 2019, regroupant un total de 240 observations. Ces dernières sont issues de l'agence financière Bloomberg. Contrairement à la variable analysée précédemment à savoir la popularité du mot 'emploi' sous Google Trends, les séries sont ici non saisonnières, c'est à dire qu'il n'existe pas de saisonnalité, de comportement récurrent qui puisse masquer l'information contenue dans les variables - ceci a été vérifié à l'aide de la fonction 'IsSeasonal' sous R. Dans cette partie nous nous

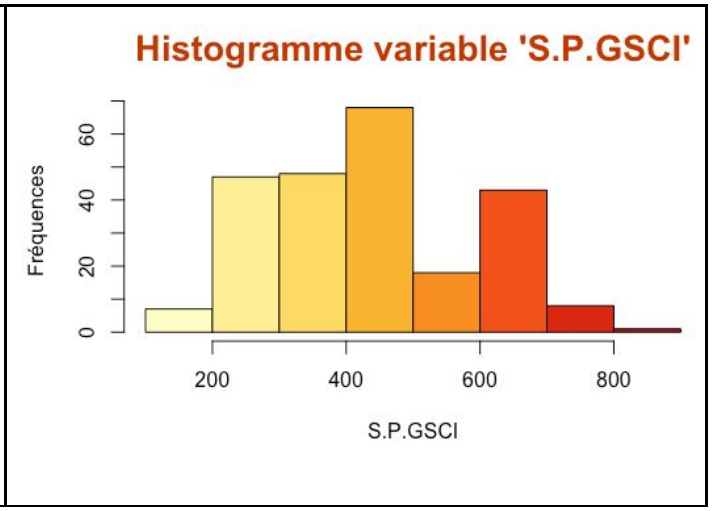
[4] Hull J., « Options, futures et autres actifs dérivés », Pearson, 10e Edition, p.62

[5] Aaron Clark, « Brent Weighting to Overtake WTI Oil as Largest in S&P GSCI Index », sur Bloomberg.com, 12 novembre 2014

intéresserons aux caractéristiques intrinsèques de la série dans un premier temps, puis nous tenterons de prévoir ses variations futures à l'aide de différents modèles et ferons de même en nous servant de l'information contenues dans les variables X_1 et X_2 présentées précédemment. Le but étant de noter d'éventuelles amélioration, différences, pertinence dans les différents types de modélisation mis en place.

Figure n°4 : Statistiques descriptives et histogramme du S&P GSCI

```
> round(basicStats(df_non_sais$S.P.GSCI),2)
            X..df_non_sais.S.P.GSCI
nobs                         240.00
NAs                            0.00
Minimum                      166.03
Maximum                      862.70
1. Quartile                  320.48
3. Quartile                  553.50
Mean                         435.69
Median                       426.78
Sum                        104564.55
SE Mean                       10.18
LCL Mean                     415.64
UCL Mean                     455.73
Variance                   24849.45
Stdev                        157.64
Skewness                       0.21
Kurtosis                      -0.84
```



Commençons à présent par regarder les caractéristiques de notre série Yt : le cours des matières premières. À partir du tableau ci-dessus on voit que sur les 240 périodes qui la composent, la série varie entre 166,03 et 862,7 ce qui est considérable, on suppose donc une variance forte liée aux importantes fluctuations notables grâce à l'écart entre les valeurs minimales et maximales de la série. La variance est effectivement très importante ; 24.849,45 pour un écart type de 157,64 qui représente 36% de la moyenne. Cette dernière est relativement proche de la médiane ce qui nous laisse penser qu'il y a peu de points atypiques et que, s'ils existent, seront plutôt avec des valeurs très fortes influençant ainsi la moyenne plus haut que la médiane.

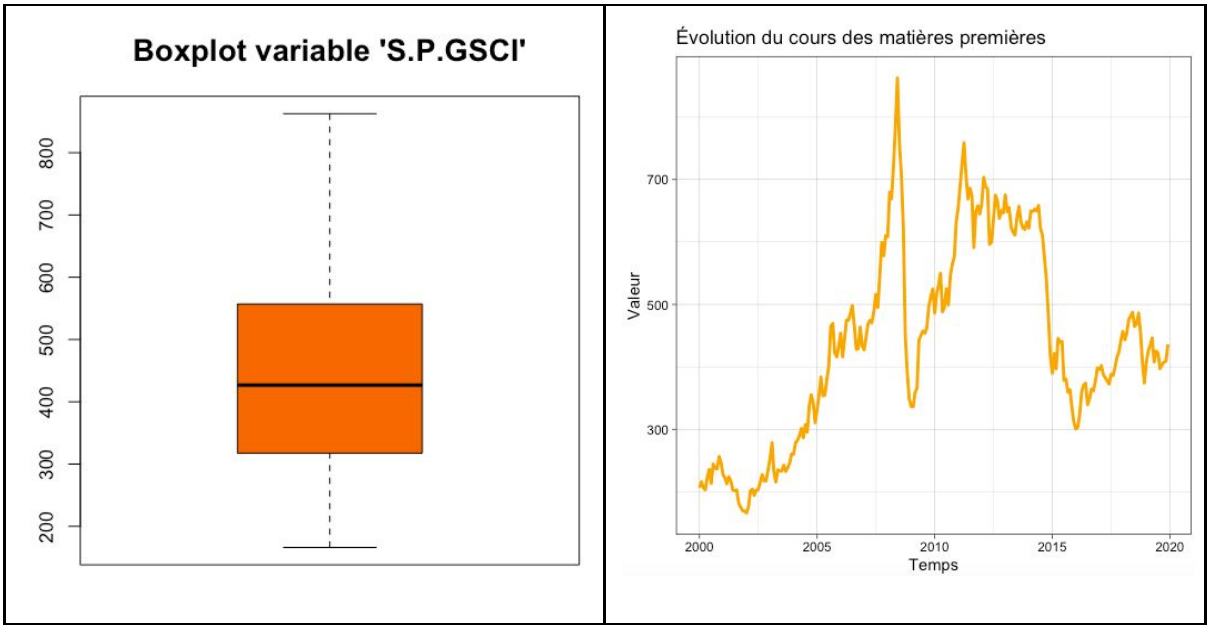
Aussi, on peut noter une légère asymétrie à gauche (skewness=0,21) ainsi qu'une distribution plus aplatie qu'une loi normale - ce qui se confirme amplement sur l'histogramme de la variable S&P GSCI ci dessus. Enfin, c'est le test de shapiro qui vient montrer statistiquement que la variable à expliquer ne suit pas une loi normale puisque la probabilité d'accepter l'hypothèse nulle qu'est "la série suit une loi normale", est très proche de zéro ce qui nous amène à rejeter cette dernière.

Tableau n°17 : Résultat du test de Shapiro sur la variable S&P GSCI

```
> shapiro.test(df_non_sais$`S.P.GSCI`)

        Shapiro-Wilk normality test

data:  df_non_sais$S.P.GSCI
W = 0.9645, p-value = 1.113e-05
```

On se penche maintenant sur les évolutions de la série misent côte à côte avec le diagramme de Tukey, censé informer d'éventuelles valeurs atypiques. On ne voit sur le boxplot aucune observation atypique, c'est à dire qui serait hors des bornes inférieures et supérieures de la boîte à moustaches. En revanche lorsque l'on regarde le graphique évolutif des valeurs de la série chronologique on voit qu'il existe des valeurs très fortes en 2008 juste avant la crise économique mondiale, puis l'indice chute fortement un an plus tard en 2009. Nous pouvons voir deux trends à la hausse importantes : une première, de 2002 jusqu'à 2008 environ, puis une seconde de 2009 à 2011. En effet, le secteur des matières premières n'a pas été épargné par la crise de 2008. De plus, on constate que depuis 2015, le prix de l'indice a retrouvé des niveaux particulièrement bas même s'il semble s'être stabilisé autour de 400 dollars. Comme supposé lors de l'interprétation de la moyenne et de la médiane, il semble que ces valeurs soient atypiques par rapport au reste de la série puisqu'elles sont extrêmes, nous allons donc utiliser les méthodes de détection de points atypiques sous JDemetra : Tramoseats et regArima.

Figure n°5 : Boîte à moustache et graphique d'évolution de la variable S&P GSCI[6]



[6] De même pour ce graphique initialement animé

ii. Points atypiques

Tableau n°18 : Points atypiques détectés par le logiciel JDemetra+

• RegArima

Outliers

	Coefficients	T-Stat	P[|T| > t]
LS (10-2008)	-0,2999	-4,93	0,0000

• TRAMO

Outliers

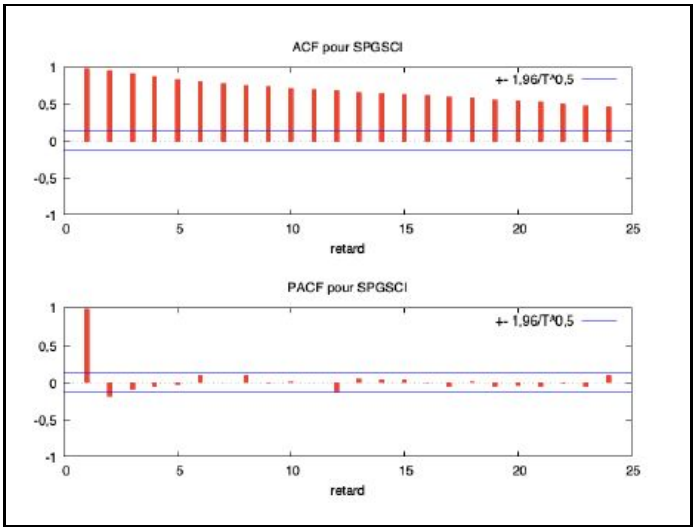
	Coefficients	T-Stat	P[|T| > t]
LS (10-2008)	-0,2999	-4,93	0,0000

Comme pour l'analyse de la variable Google Trends nous procédons à l'étude des points atypiques sous JDemetra grâce aux méthodes RegArima et TRAMO Seats. Nous constatons alors que la série ne comporte qu'un seul point atypique en octobre 2008, détecté à la fois par RegArima et TRAMO. Celui-ci peut s'expliquer encore une fois par la crise économique mondiale de 2008 émanant des États-Unis qui a provoqué la chute de nombreuses cotations boursières (CAC 40 avec une baisse de 9,04%, DOW JONES -4,8%, Nasdaq -5,17%, la bourse de Londre -7,9% ...) dont le cours des matières premières S&P GSCI avec une chute de 172,78 points. Cette baisse soudaine est largement identifiable sur le graphique d'évolution de ce cours ci-dessus.

iii. Stationnarité série

Pour procéder à la modélisation de la série, il convient tout d'abord de déterminer si le processus générateur des réalisation de la série est stationnaire. Le processus est dit stationnaire lorsque les moments de la variable aléatoire Yt sont indépendants dans le temps *ie.* la série fluctue autour de sa valeur μ. Sur le premier graphique représentant l'évolution de la série nous pouvions voir que la moyenne de la série semblait fluctuer au cours du temps et que la variance entre les observations était importante. Ainsi, pour compléter ce premier constat graphique, il convient de conforter notre intuition avec le corrélogramme de la série ci-dessous obtenu avec le logiciel Gretl disponible en figure n°. Nous pouvons voir que la FAC est décroissante et se rapproche très lentement de 0 ce qui indique la non-stationnarité de la série. De plus, la première valeur de la FAC partielle est égale à 1 ce qui veut dire que les éléments de la série sont dépendants entre eux à l'intérieur de la série. La série n'est donc vraisemblablement pas stationnaire.

Figure n°6 : Corrélogramme de la série brute du S&P GSCI



Nous pouvons également vérifier la stationnarité de Y_t en effectuant le test de la racine unitaire qui s'intéresse au comportement des résidus à long terme, pour cela on pose les hypothèses suivantes :

H_0 : La racine unitaire existe, Y_t n'est pas stationnaire

H_1 : Y_t est stationnaire (p-critique asymptotique < 0.05)

Tableau n°19 : Résultats du test ADF sous Gretl et sous R

• Gretl

```
Test de Dickey-Fuller augmenté pour SPGSCI
testing down from 12 lags, criterion AIC
taille de l'échantillon 238
hypothèse nulle de racine unitaire : a = 1

  test avec constante
  avec un retard de (1-L)SPGSCI
  modèle: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,0256819
  statistique de test: tau_c(1) = -2,11594
  p. critique asymptotique 0,2384
  Coeff. d'autocorrélation du 1er ordre pour e: -0,018

  avec constante et tendance temporelle
  avec un retard de (1-L)SPGSCI
  modèle: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,0284692
  statistique de test: tau_ct(1) = -2,05756
  p. critique asymptotique 0,569
  Coeff. d'autocorrélation du 1er ordre pour e: -0,019
```

• R studio

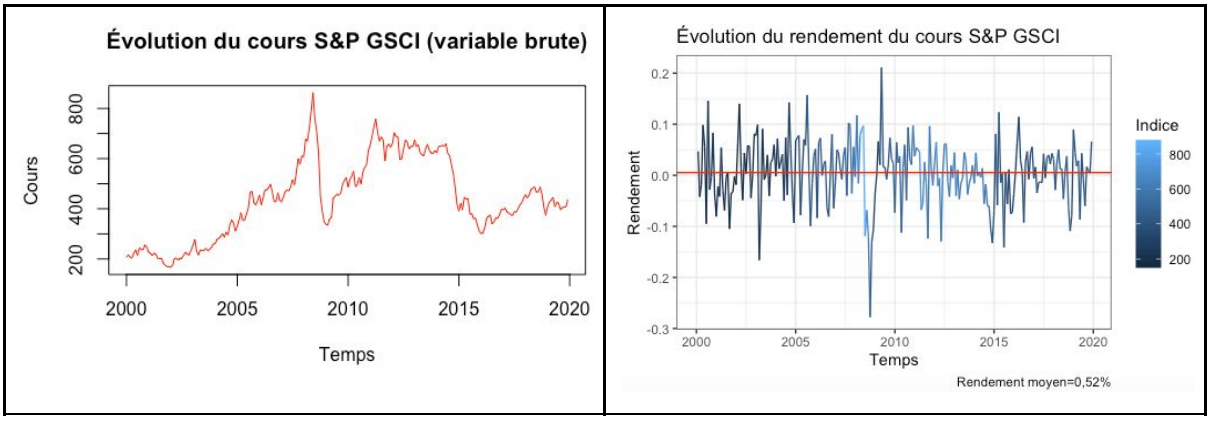
```
##############################################
# Augmented Dickey-Fuller Test Unit Root Test #
##############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-153.22  -13.35    3.18   17.93   76.14

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.001436   0.004138  -0.347  0.72892
z.diff.lag  0.217153   0.063723   3.408  0.00077 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.59 on 236 degrees of freedom
Multiple R-squared:  0.04701,   Adjusted R-squared:  0.03894
F-statistic: 5.821 on 2 and 236 DF,  p-value: 0.003406
```

Ici, la valeur critique du test de Dickey-Fuller sous Getl est égale à 0,57 > 0.05 , l'hypothèse H_0 est acceptée au seuil de 5%, la série n'est pas stationnaire - mais elle est de 0,0034 pour le même test effectué sous R (basé sur les mêmes hypothèses), ce qui nous conduirait à

rejeter H_0. Cependant, parce que la série ne fluctue pas autour d'une valeur moyenne et que ses variations ne sont pas stables au cours du temps, nous préférons stationnariser la série en calculant son rendement de manière à éviter tout problème de non stationnarité. De plus, lors de l'estimation d'un modèle AR(1) en prenant la série Y_t brute, nous nous apercevons que le paramètre estimé $\phi_1 = 0{,}9875$ est trop proche de 1 ce qui vient confirmer la non stationnarité de la série puisqu'alors ses caractéristiques sont dépendantes du temps.

Ainsi, nous tentons de stationnariser la série autour de la moyenne en calculant son rendement car il s'agit d'une série financière - ce qui nous fera perdre la première observation i.e. janvier 2000. Nous pouvons voir sur les graphiques de la série brute et de son rendement, que la stationnarisation a permis une meilleure distribution autour de la moyenne. En outre, nous procédons à nouveau au test Dicket-Fuller augmenté pour vérifier la stationnarité et donc l'absence de racine unitaire. Il convient de préciser que ce test repose sur l'hypothèse que le résidu est un bruit blanc[7]. En suivant la procédure de Box et Jenkins, nous vérifierons cette condition après l'estimation des modèles.

Graphique n°7 : Série du S&P GSCI brute et stationnarisée



Après une différenciation de la série, nous obtenons une valeur critique égale à 2.2e-16 < 0.05 disponible en annexe n°8, H_0 est refusée, notre série est désormais stationnaire.

[7] Lardic S., Mignon V., « Économétrie des séries temporelles macroéconomiques et financières », Economica, 2002, p. 140

B) Estimation de modèles et vérification des résidus

Dans cette partie nous allons modéliser la série Y, c'est à dire le cours de l'indice des matières premières noté S&P GSCI, à l'aide de différentes méthodes. Nous commencerons par modéliser un processus AR(1), puis un AR(p), un ARIMA(p,d,q) pour terminer sur un modèle construit à partir de la méthode Holt Winters.

➢ **AR(1)** : Nous pouvons voir grâce à la fonction *coeftest* du package **lmtest**, que le paramètre ϕ_1 est significatif au seuil de 5%. Par ailleurs, la condition de stationnarité est respectée puisque $\phi_1 = 0{,}165$ est strictement inférieur à 1. Le critère d'Akaike indiquant la vraisemblance du modèle est de -601.86 et le log de vraisemblance est égal à 303.93. Ces mesures seront utiles à la comparaison de nos modèles afin de savoir lequel est le meilleur, ou plutôt le moins mauvais.

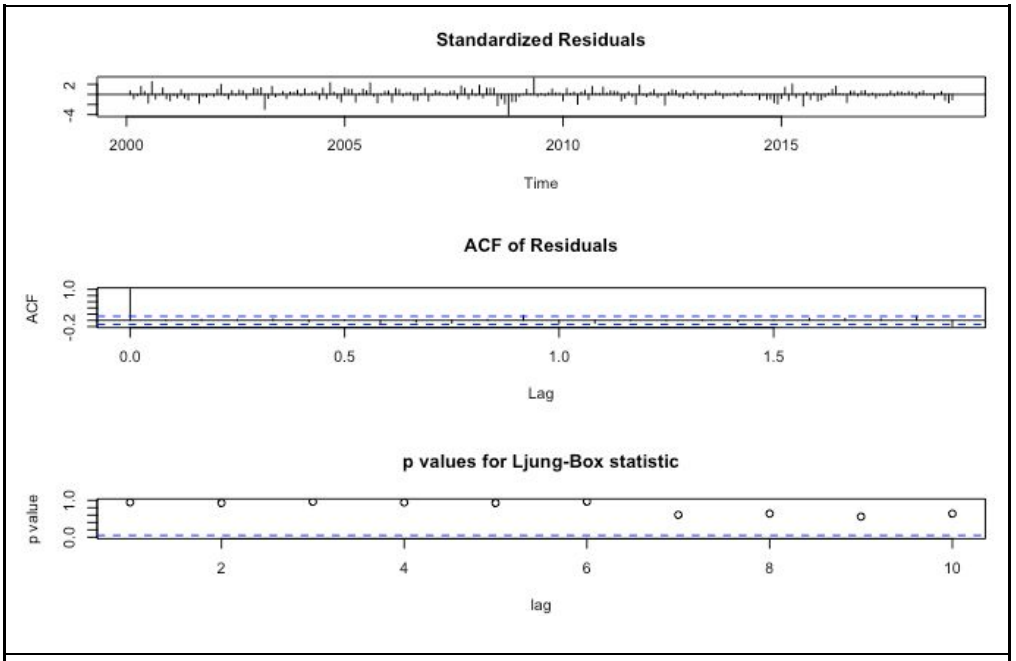
Tableau n°20 : Estimation d'un modèle AR(1) et significativité de ses paramètres

```
> AR1
Series: yarima
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1    mean
      0.1646  0.0047
s.e.  0.0656  0.0050

sigma^2 estimated as 0.004058:  log likelihood=303.93
AIC=-601.86   AICc=-601.75   BIC=-591.59
> coeftest(AR1)

z test of coefficients:

           Estimate Std. Error z value Pr(>|z|)
ar1       0.1646310  0.0656122  2.5092   0.0121 *
intercept 0.0047014  0.0050354  0.9337   0.3505
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nous passons ensuite à l'étape de vérification de ce modèle AR(1) : cette étape consiste à vérifier l'indépendance des résidus puisque comme nous l'avons expliqué précédemment, le modèle ne peut-être interprété que si ses résidus suivent un processus de bruit blanc. En outre, les résidus doivent avoir une espérance nulle et indépendante du temps. Ainsi, nous vérifions tout d'abord l'absence d'autocorrélation des résidus puis nous vérifierons que les estimateurs d'autocorrélation suivent approximativement une loi normale.

→ **Indépendance des résidus**

Pour nous assurer que les résidus suivent un processus de bruit blanc nous utiliserons le test de Ljung et Box qui se réfère à la statistique Q. Ainsi, nous obtenons une p-value : 0.946 > 0.05, l'hypothèse nulle d'absence d'autocorrélation des résidus est acceptée au seuil de 5%. Nous constatons également dans le corrélogramme que les premiers retards ne sont pas significatifs et sont davantage proche de 0. Les premiers retards de la fonction d'autocorrélation sont les plus importants car ce sont ceux qui contiennent le plus d'information.




```
> Box.test(resAR1, type = c("Ljung-Box"))

        Box-Ljung test

data:  resAR1
X-squared = 0.0045379, df = 1, p-value = 0.9463
```

→ **Normalité des résidus**

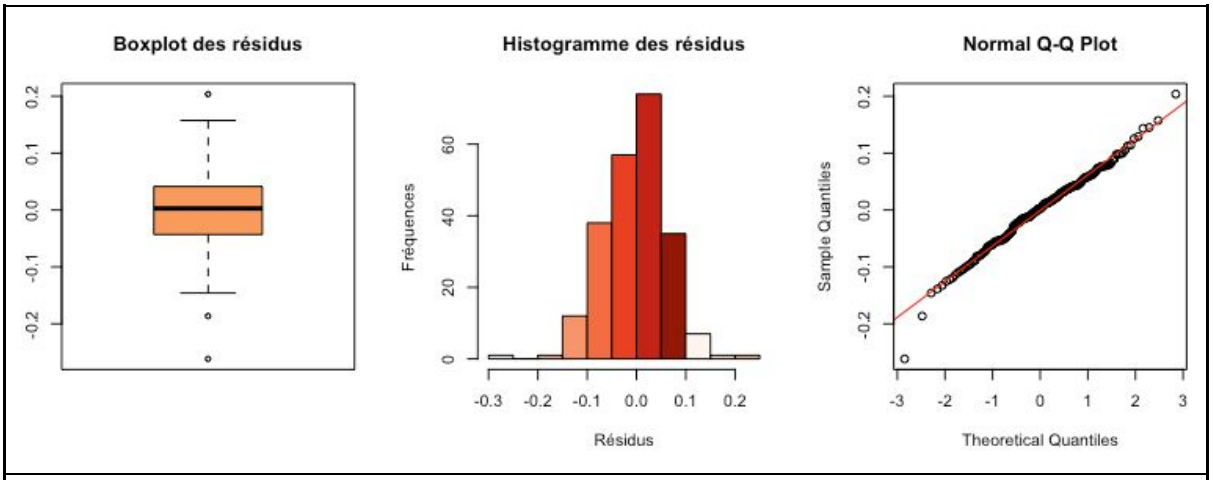
Nous pouvons voir ci-dessous que la distribution des résidus semble être asymétrique. En effet, la distribution semble être plus étalée à gauche i.e. la valeur skewness est supérieure à 0. En effet, nous pouvons à l'aide de la boîte à moustache, voir des valeurs atypiques qui

peuvent influencer cette distribution. Nous observons 3 valeurs atypiques, dont deux inférieures au premier quantile ce qui peut expliquer l'étalement de la série sur la gauche. On retrouve également ces valeurs aux extrémités de la droite d'Henry (QQ plot), on suppose donc que ces valeurs altéreront la distribution des résidus. Nous pouvons vérifier toutes ces hypothèses émises à partir de graphiques, grâce au test non paramétrique de Kolmogorov-Smirnov qui informe de la normalité d'une distribution ou non. Celui-ci est valable pour les échantillons supérieurs à 50 observations comme c'est le cas dans notre étude, il suit les hypothèses suivantes :

H_0 : Les résidus suivent une loi normale

H_1 : Les résidus ne suivent pas une loi normale

La probabilité d'accepter l'hypothèse nulle est de 0.5061, nous acceptons donc l'hypothèse de normalité des résidus au seuil de 5%.




```
> ks.test(resAR1, "pnorm", mean(resAR1), sd(resAR1))

        One-sample Kolmogorov-Smirnov test

data:  resAR1
D = 0.054669, p-value = 0.5061
alternative hypothesis: two-sided
```

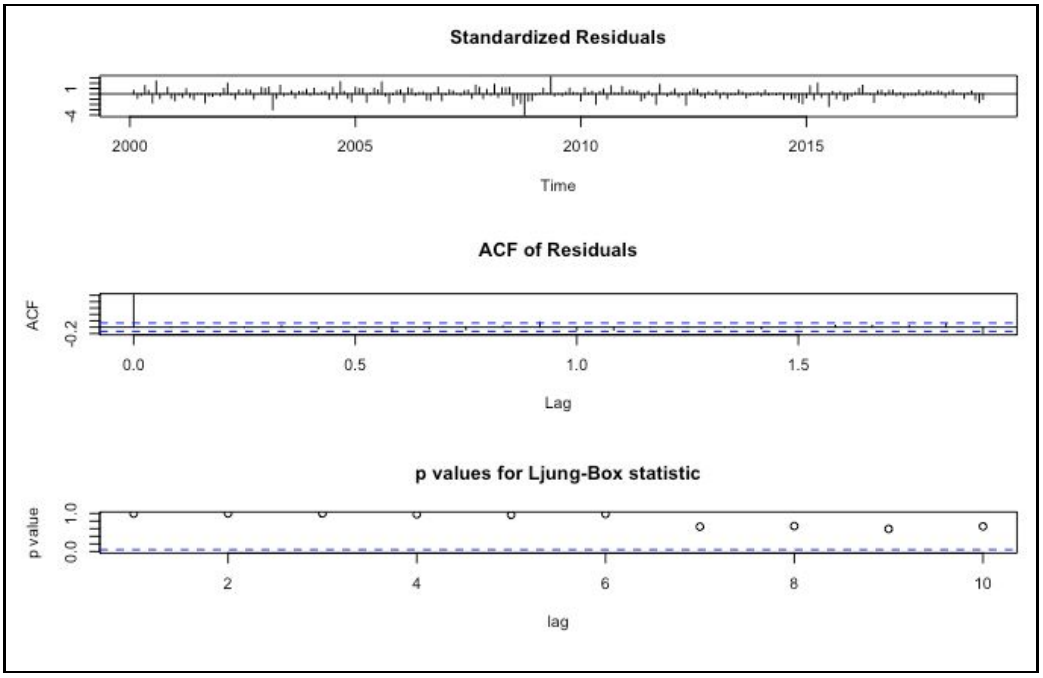
➢ **<u>AR(p)</u>** : Nous cherchons maintenant à prévoir la série Y en utilisant un modèle AR de paramètre 'p'. Après avoir estimé un AR(2) dont les caractéristiques sont trop proches de celles de l'AR(1) nous décidons d'estimer un modèle AR(3) disponible ci-dessous. On voit ainsi que ϕ_1= 0.159 , ϕ_2=0.026 et ϕ_3=0.026 où seul le premier paramètre est significatif au seuil de 5%. Ici, le critère AIC est égal à -598.21 et le log de vraisemblance à 304.11.

Tableau n°21 : Estimation d'un modèle AR(3) et significativité de ses paramètres

```
> ARp
Series: yarima
ARIMA(3,0,0) with non-zero mean

Coefficients:
         ar1     ar2     ar3    mean
      0.1592  0.0259  0.0262  0.0046
s.e.  0.0664  0.0675  0.0668  0.0053

sigma^2 estimated as 0.004088:  log likelihood=304.11
AIC=-598.21   AICc=-597.94   BIC=-581.09
> coeftest(ARp)

z test of coefficients:

           Estimate Std. Error z value Pr(>|z|)
ar1       0.1592243  0.0664213  2.3972  0.01652 *
ar2       0.0259318  0.0675346  0.3840  0.70099
ar3       0.0262408  0.0667960  0.3929  0.69443
intercept 0.0046282  0.0053274  0.8688  0.38498
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De la même manière que pour AR(1) nous pouvons vérifier l'indépendance et la normalité des résidus :

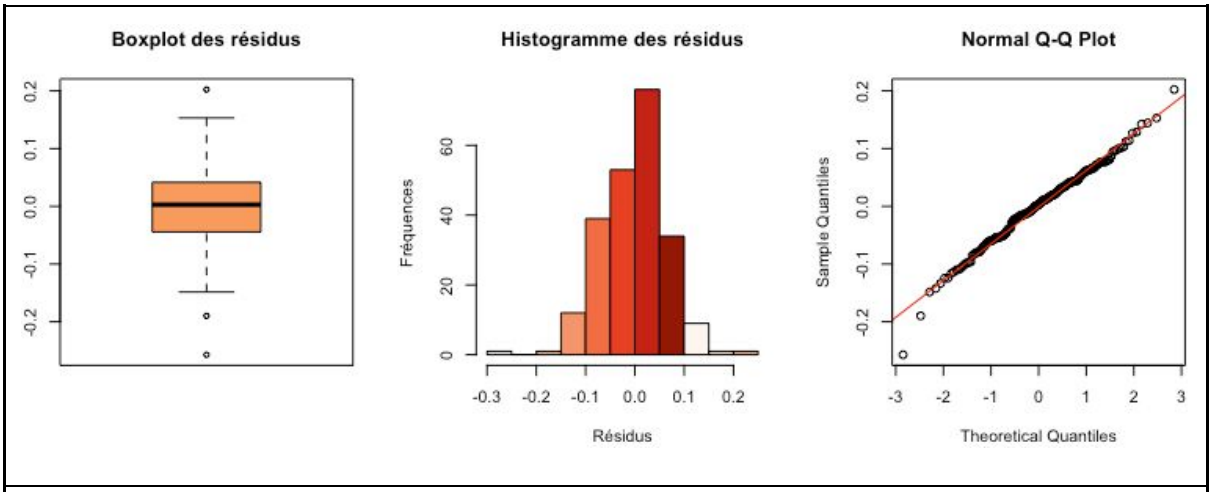
→ **Indépendance des résidus**



```
> Box.test(resARp, type = c("Ljung-Box"))

        Box-Ljung test

data:  resARp
X-squared = 2.9961e-05, df = 1, p-value = 0.9956
```

En étudiant la fonction d'autocorrélation du processus AR(3), nous pouvons voir que les premiers résidus sont à niveau très faibles. À nouveau la p-value associée au test de Ljung-Box est supérieure à 0.05, ce qui confirme l'hypothèse d'indépendance des résidus au seuil de 5%.

→ **Normalité des résidus**




```
> ks.test(resARp, "pnorm", mean(resARp), sd(resARp))

        One-sample Kolmogorov-Smirnov test

data:  resARp
D = 0.057536, p-value = 0.4401
alternative hypothesis: two-sided
```

Pareillement, la distribution semble être étalée sur la gauche comme pour le modèle AR(1), cela est dû à la présence de résidus atypiques. Nous pouvons observer ces valeurs sur le troisième cadran, hors de la droite d'Henry. Le test de Kolmogorov-Smirnov nous indique que les résidus s'apparentent à une loi normale car la p-value égale à 0.4401 est supérieure à 0.05. En outre, les caractéristiques de ces 2 modèles sont proches en tout point (critères d'Akaike, Log de vraisemblance, validation des tests etc.). Passons à présent à la modélisation ARIMA(p,d,q).

➢ **<u>ARIMA(p,d,q)</u>** : Pour déterminer le meilleur modèle ARIMA, c'est à dire AutoRegressive Integrated Moving Average nous allons utiliser la commande *auto.arima* qui a pour but, sous R, de déterminer le modèle optimal en en testant plusieurs avec des paramètres différents. Nous obtenons alors un modèle SARIMA(1,0,0)(1,0,0$)_{12}$ avec une partie saisonnière alors que nous avons vérifié au début de cette étude que la série financière ne disposait pas de partie saisonnière, cela est donc lié à la stationnarisation que nous avons effectuée précédemment. Nous testons alors un modèle ARIMA(1,0,1), même si à priori, la série suivrait plutôt un processus autorégressif que moyenne mobile. Après estimation, aucun paramètre n'est statistiquement significatif. Le critère AIC est égal à -600.14 et le log de vraisemblance est égal à 304.07. Les hypothèses sur les résidus sont également validées : p-value associé au test Box-Ljung est égale à 0.973 > 0.05, l'hypothèse d'absence d'autocorrélation des résidus est acceptée au seuil de 5%. De même, la p-value associée au test de Kolmogorov-Smirnov est égale à 0.363 > 0.05, l'hypothèse de normalité est encore une fois acceptée au seuil de 5%.

Les analyses des résidus de ce modèle ainsi que celle du modèle LED Holt Winters se trouvent en annexes n°9 et 10. Les coefficients de lissage α et β du modèle ont été obtenus en précisant dans la fonction l'absence de partie saisonnière dans la série avec le paramètre 'gamma=FALSE', on obtient α=0.456 et β=0.15. Ainsi, nous savons que la prévision est souple, *i.e.* les observations les plus récentes ont davantage d'influence. Nous résumons à présent l'information de cette partie (estimation de modèle et analyse de leurs résidus) en un tableau récapitulatif.

<u>Tableau n°22 : Principales caractéristiques des modèles estimés</u>

	<u>AR(1)</u>	<u>AR(3)</u>	<u>ARIMA(1,0,1)</u>	<u>HW</u>
Nombre de paramètres significatifs*	1/1	1/3	0/2	/
Critère AIC	-601.86	-598.41	-600.14	/
Log likelihood	303.93	304.11	304.07	/
Résidus indépendants	✓	✓	✓	✓
Résidus normaux	✓	✓	✓	✓

* : hors constante et au seuil de 5%

Nous pouvons voir sur le tableau ci-dessus que les résultats entre les modèles ARIMA sont très proches. En effet, les trois critères AIC sont presques égaux, de même pour le log de vraisemblance. De plus, l'hypothèse sur l'indépendance des résidus est validée pour chacun d'entre eux, ainsi que la normalité des résidus. Nous pouvons affirmer que le modèle AR(1) est le modèle à retenir, d'une part selon l'hypothèse de parcimonie, d'autre part car c'est le seul modèle pour lequel tous les paramètres sont significatifs et c'est celui qui minimise le critère d'Akaike. Nous verrons par la suite si celui-ci a également la meilleure qualité de prédiction.

C) Prévisions à partir des 4 modèles précédents

Nous allons à présent réaliser les prévisions de la série Y_t à partir des 4 modèles précédents sur 1 an à partir de janvier 2019. Pour se faire, nous mettons en place des boucles pour réaliser les prévisions pas à pas : à chaque itération le programme inclut la nouvelle prévision dans les valeurs connues de la série de manière à minimiser l'erreur des prévisions. De nouveau, nous avons retiré au préalable les 12 dernières observations de la série afin de pouvoir comparer aux valeurs prédites par nos différents modèles.
Nous obtenons les prévisions suivantes à partir des 4 modèles présentés antérieurement, prévisions qui restent relativement similaires d'un modèle à un autre, essentiellement pour les modèles ARIMA. On s'intéresse aussi au rendement moyen prédit par chaque modèle : celui des modèles AR(1) et AR(3) est le même et a un rendement seulement 0,1% supérieur au modèle ARIMA(1,0,1). Enfin, il apparaît que les prévisions réalisées à partir de la méthode LED Holt Winters se distinguent des 3 autres et prédit elles, un rendement négatif sur la période de janvier à décembre 2019.

Tableau n°23 : Valeurs prédites par les 4 différents modèles

	AR(1)	AR(3)	ARIMA(1,0,1)	HW
Janvier 2019	0.0099	0.0113	0.0115	0.0392
Février 2019	0.0004	0.0024	0.0021	0.0066
Mars 2019	0.0082	0.0077	0.0071	0.0129
Avril 2019	0.0128	0.0120	0.0128	0.0321
Mai 2019	0.0069	0.0086	0.0087	0.0235

Juin 2019	0.0070	0.0084	0.0077	0.0184
Juillet 2019	-0.0022	-0.0015	-0.0014	-0.0157
Août 2019	0.0062	0.0048	0.0043	-0.0076
Septembre 2019	0.0108	0.0093	0.0101	0.0132
Octobre 2019	-0.0044	-0.0031	-0.0028	-0.0257
Novembre 2019	-0.0128	-0.0128	-0.0142	-0.0744
Décembre 2019	-0.0091	-0.0134	-0.0134	-0.0875
Prévision moyenne	0.0028	0.0028	0.0027	-0.0054

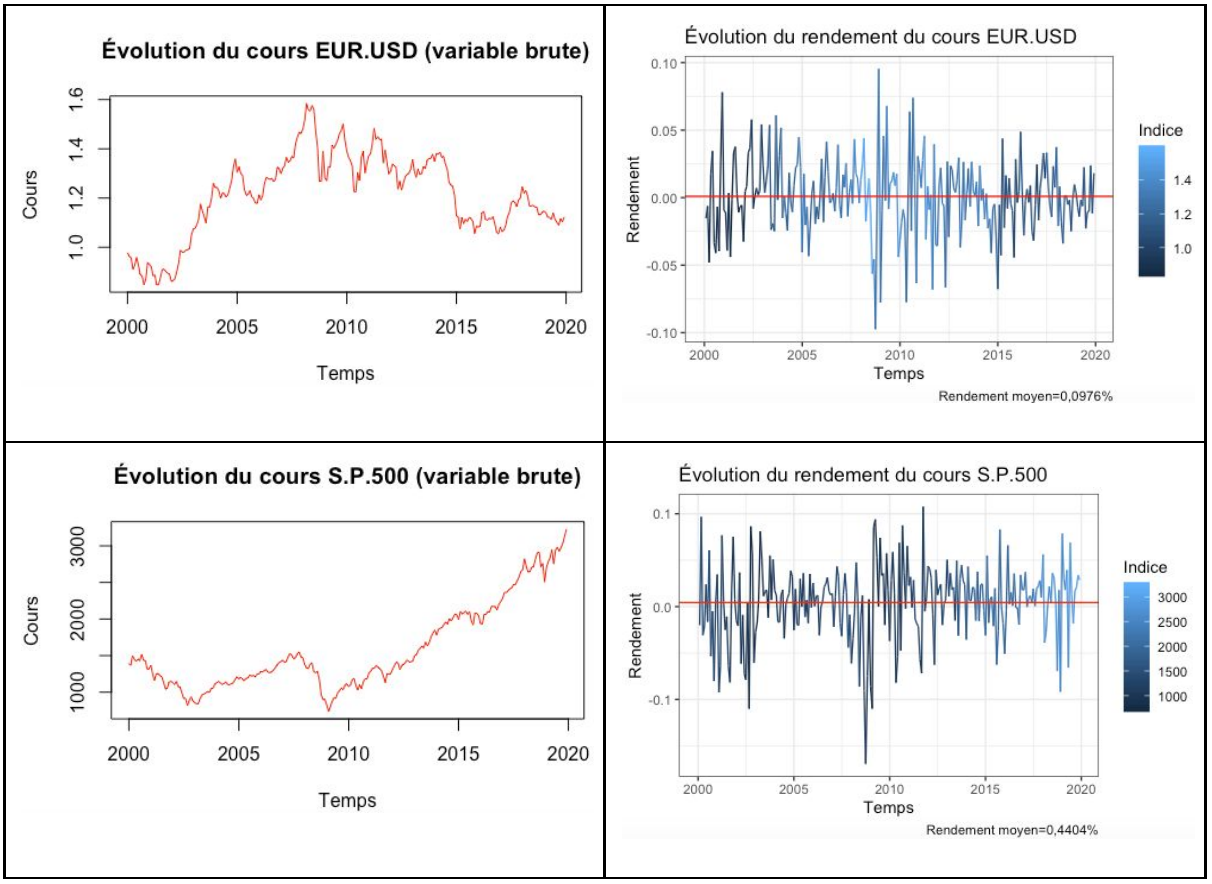
D) Prévision avec variables explicatives

i. Stationnarisation des variables explicatives

Dans cette partie nous allons compléter l'analyse de Y_t en ajoutant les variables explicatives que sont le cours euro/dollars et le S&P 500. Pour cela nous modéliserons le cours des matières premières (Y_t) en fonction de ces dernières à l'aide d'un modèle ARX dont nous montrerons les principaux avantages, puis nous réaliserons des prévisions en nous servant des variables explicatives.

Avant cela nous vérifions la stationnarité des variables explicatives puisque pour réaliser les prévisions de Y_t en nous servant de celles-ci, il faut que toutes les séries soient stationnaires. Comme pour Y nous appliquons le test de la racine unitaire ; nous obtenons une p-value de 0,9935 pour le cours EUR/USD qui n'est donc pas stationnaire et 0,03842 pour le S&P 500. Cela signifie qu'au seuil de risque de 5% nous rejetons "H_0 : la série n'est pas stationnaire" et à 1% nous acceptons cette hypothèse. La p-value étant trop juste et l'expérience nous ayant montré que la fiabilité des tests n'était pas toujours certaine, nous décidons de stationnariser les 2 variables explicatives. S'agissant de cotations boursières nous les stationnarisons en calculant leur rendement. Nous pouvons observer les graphiques n°8 des variables brutes et rendues stationnaires : nous voyons qu'il existait initialement une trend à la hausse jusqu'en 2008 puis à la baisse jusqu'à fin 2019 pour la variable EUR/USD et une hausse générale du S&P 500 sur toute la période, et que le rendement de chaque série est bel et bien stationnaire et varie autour d'une valeur moyenne de 0,0976% pour le cours EUR/USD et 0,4404% pour le S&P 500.

Graphique n°8 : Séries de l'EUR/USD et du S&P 500 brutes et stationnarisées



ii. Modélisation et prévision avec variables explicatives

Après avoir stationnarisé nos variables nous passons à la modélisation de Y en fonction de 2 variables explicatives stationnaires, c'est à dire les rendements de EUR/USD et S&P 500, en utilisant la méthode ARX. Cette dernière utilise une régression OLS dans le cadre d'un modèle autorégressif AR qui inclut des variables exogènes ("external inputs") et un bruit blanc de moyenne nulle. La structure autorégressive de ce modèle permet notamment d'avoir un modèle globalement non linéaire vis-à-vis des paramètres sur l'intervalle de temps.

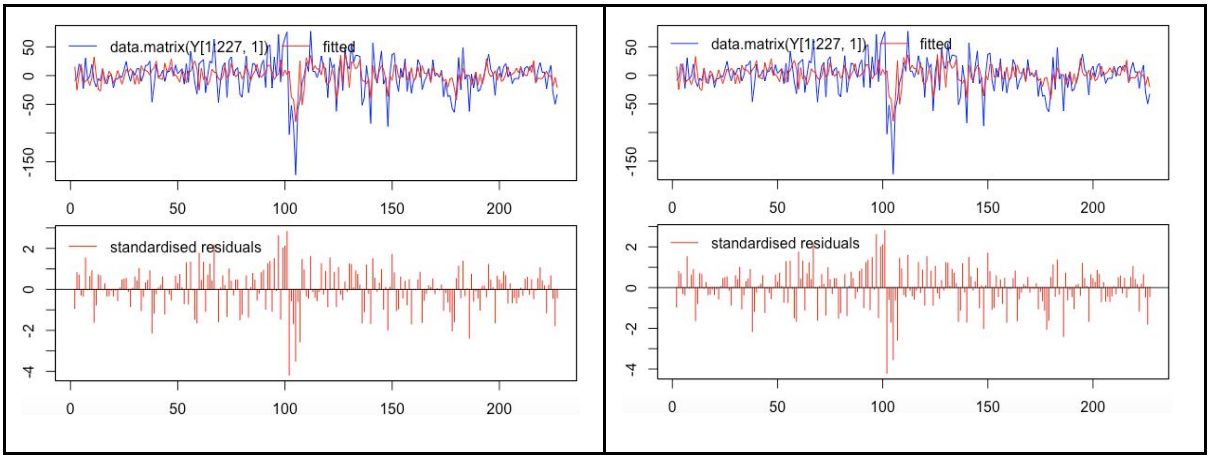
On estime un premier modèle avec le rendement du S&P GSCI comme variable dépendante et les rendements des 2 cours comme variables indépendantes. On observe ainsi que les variables explicatives et le paramètre AR(1) sont significatifs au seuil de 5% comme de 0,1%, cependant la constante ne l'est pas ; nous recommençons donc la modélisation en retirant cette dernière. Le modèle n°2 est disponible ci-dessous ; tous les

paramètres et variables du modèle sont désormais significatifs, la qualité du modèle n'est pas très bonne avec un R^2=0,273 ce qui traduit un écart relativement important entre les valeurs prédites par le modèle et les valeurs observées. Cela se note aussi sur les graphiques où l'on voit les valeurs observées en bleu et les valeurs prédites par le modèle en rouge ; nous avons stationnarisé la variable à expliquer (cours des matières premières) en calculant son rendement sur la période étudiée. Or, malgré cette stationnarisation, les variations de la série sont toujours notables sur le graphique avec notamment des pics importants en 2008 liés à la crise économique et financière mondiale, pics que le modèle arrive difficilement à prévoir. Par conséquent, les valeurs observées connaissent des fluctuations plus fortes sur la période étudiée que le modèle n'en prévoit, ce qui explique le faible coefficient de détermination.

Tableau n°24 : Modèle ARX avec et sans constante, et graphiques associés aux modèles

```
> Modele01

Date: Wed Apr 29 10:31:02 2020
Dependent var.: data.matrix(Y[1:227, 1])
Method: Ordinary Least Squares (OLS)
Variance-Covariance: Ordinary
No. of observations (mean eq.): 226
Sample: 2 to 227

Mean equation:

              coef  std.error  t-stat   p-value
mconst   -0.497526   1.758255 -0.2830 0.7774667
ar1       0.176611   0.057691  3.0613 0.0024752
RdmtEUR 370.183632  64.209444  5.7653  2.71e-08
Rdmt500 169.759568  44.018201  3.8566 0.0001507

Diagnostics and fit:

                    Chi-sq df    p-value
Ljung-Box AR(2)    0.11256  2 0.94527543
Ljung-Box ARCH(1) 11.61273  1 0.00065502

SE of regression    26.3268
R-squared            0.2733
Log-lik.(n=226)  -1057.8332
```

```
> Modele02

Date: Wed Apr 29 10:33:05 2020
Dependent var.: data.matrix(Y[1:227, 1])
Method: Ordinary Least Squares (OLS)
Variance-Covariance: Ordinary
No. of observations (mean eq.): 226
Sample: 2 to 227

Mean equation:

              coef  std.error t-stat   p-value
ar1       0.176254   0.057558 3.0622 0.0024671
RdmtEUR 369.941299  64.071168 5.7739 2.579e-08
Rdmt500 168.819348  43.801986 3.8541 0.0001519

Diagnostics and fit:

                    Chi-sq df    p-value
Ljung-Box AR(2)    0.10513  2 0.94879433
Ljung-Box ARCH(1) 11.37464  1 0.00074454

SE of regression    26.27248
R-squared            0.27304
Log-lik.(n=226)  -1057.86607
```



Nous pouvons désormais utiliser ces modélisations pour prévoir la série Y, comme nous l'avons fait dans la partie précédente, c'est à dire à l'aide d'une boucle, en incluant cette fois-ci les variables explicatives rendues stationnaires. Nous obtenons alors les prévisions suivantes : 0.0185 (janvier 2019), 0.0176 (février), 0.0036 (mars), 0.0161 (avril), -0.0199 (mai), 0.0308 (juin), -0.0045 (juillet), -0.0125 (août), -0.0067 (septembre), 0.0281 (octobre), 0.0065 (novembre), 0.0252 (décembre).

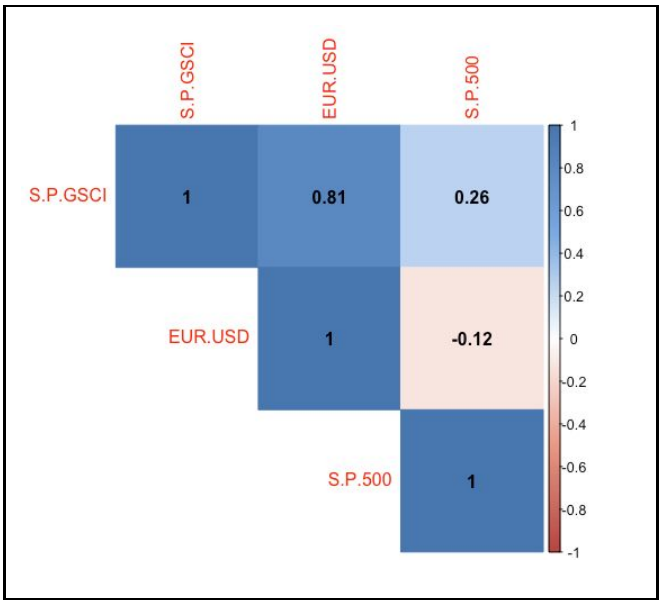
iii. Justification du choix des variables

Depuis la financiarisation des matières premières, les échanges de contrats sur ces derniers ont considérablement augmentés. L'impact de l'augmentation de ces transactions financières sur le prix des produits de base a fait l'objet de nombres études suggérant une influence positive sur la volatilité, mais cette relation n'a jamais pu être identifiée car les résultats sont peu conclusifs et se contredisent[8]. En revanche, le poids de ces acteurs et leur grande influence sont incontestables et nous savons que les prix ont augmenté considérablement du début des années 2000 jusqu'à la crise de 2007-2008. Ainsi, l'indice S&P500 représente dans notre étude un indicateur sur le niveau d'investissements financiers et le taux de change euro/dollar permet d'avoir une indication sur le niveau des échanges commerciaux. En effet, le taux de change a un impact direct sur les exportations et importations de ceux-ci. De plus, le taux de change est également un indicateur de la financiarisation des marchés car il représente le deuxième plus gros marché financier. Nous pouvons voir en figure n°7 que l'indice S&P GSCI est corrélé positivement au taux de

[8] Lecocq P-E., Courleux F., «Vers la définition d'un nouveau cadre de régulation des marchés dérivés de matières agricoles», Centre d'études et de prospective, N°3 septembre 2011, téléchargeable sur : https://www.researchgate.net/publication/319242392_Vers_la_definition_d'un_nouveau_cadre_de_regulation_des_marches_derives_de_matieres_premieres_agricoles (consulté le 17 février 2020)

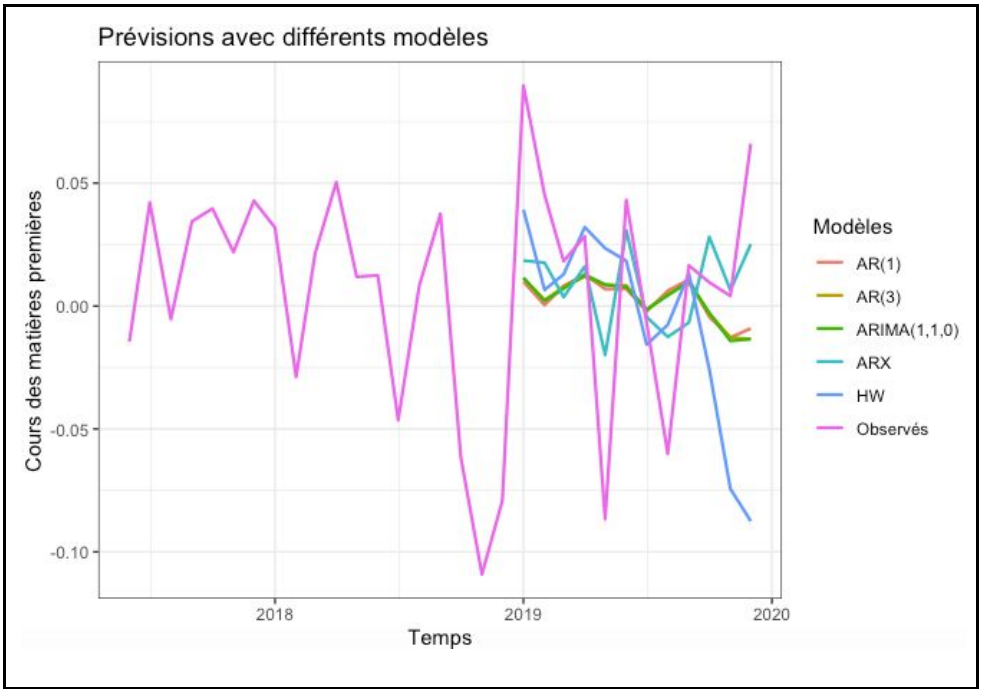
change de l'euro/dollar, ainsi que l'indice S&P 500. En revanche, ces variables sont corrélées négativement entre elles.

Figure n°7 : Matrice de corrélation des 3 séries



E) Représentation graphique des différentes prévisions

Graphique n°9 : Visualisation des différents prévisions réalisées



Sur ce graphique on distingue en rose/violet les valeurs observées du rendement du cours des matières premières, c'est à dire ses évolutions effectives sur la période de janvier 2017 à décembre 2019. Nous avons essayé de modéliser notre série Y_t à l'aide de différents modèles ; 4 modèles linéaires et un modèle qui prenait en compte des variables explicatives dont le lien est important avec Y_t. Ainsi, grâce à ces différents modèles nous avons pu prévoir les variations du rendement du S&P GSCI sur les 12 mois de l'année 2019. Nous observons ci-dessus que les prévisions des modèles linéaires AR(1), AR(3) et ARIMA(1,0,1) sont extrêmement proches, il est donc intéressant de noter que malgré un nombre de paramètres différent pour chacun des modèles, les prévisions restent plus au moins les mêmes au sein des méthodes de modélisation linéaire. En revanche les valeurs prédites par la méthode Holt Winters ne "collent" pas celles des autres modèles linéaires et prévoit une chute du rendement du cours S&P GSCI pour la fin de l'année 2019 alors qu'en réalité celui-ci a plutôt tendance à augmenter.

De plus, les prévisions réalisées avec les variables explicatives sont beaucoup plus proches de la réalité car on voit que la courbe du modèle ARX suit exactement les variations des valeurs observées, seulement les variations sont moins fortes comme nous l'avons remarqué précédemment en regardant le coefficient de détermination du modèle ARX. Cela paraît logique puisqu'en incluant des variables explicatives, les variations de Y sont mieux expliquées donc la qualité du modèle s'accroît et ses prévisions avec. Dans l'explication de phénomènes il est donc plus judicieux de prendre en compte des variables exogènes censées aider à comprendre et prévoir les fluctuations de Y_t.

F) Erreurs de prévisions et prévision naïve

Tableau n°25 : Erreurs de prévisions des modèles

	AR(1)	AR(3)	ARIMA(1,0,1)	LED HW	ARX
MSE	0.0024	0.0025	0.0025	0.0042	0.0013
MAE	0.0386	0.0387	0.0388	0.0471	0.0283
MDSE	0.0008	0.0007	0.0008	0.0014	0.0004
MDAE	0.0265	0.0258	0.0269	0.037	0.0209

On voit dans la tableau n°25 les différentes mesures d'erreur de prévisions des 5 modèles (4 linéaires et 1 avec variables explicatives). Comme nous avons vu sur le graphique c'est le modèle ARX, c'est à dire celui qui explique Y à l'aide de 2 variables explicatives, qui prédit le mieux : on retrouve donc les erreurs les plus faibles pour ce dernier - et des erreurs relativement élevées et similaires entre elles pour les 3 modèles linéaires ARIMA. De même, nous avions pu noter des prévisions assez différentes entre les modèles ARIMA et la méthode LED Holt Winters, cela est notable ici aussi puisque les taux d'erreur de prévisions sont les plus élevés pour cette dernière ; c'est donc la méthode la moins précise d'un point du vue prévisionnel.

Nous pouvons à présent comparer ces modèles et leurs taux d'erreur, à un modèle benchmark, c'est à dire de référence, qui va nous permettre de voir si les 5 modèles dont les erreurs sont présentées dans le tableau n°25, ont de meilleurs capacités prédictives que le modèle de référence. Si oui, on gardera le modèle ARX comme étant le plus précis d'un point de vu prédictif, sinon ce sera le modèle de référence. Pour se faire nous allons réaliser des prévisions naïves telles que $\widehat{X}_{t+h} = X_t$ puis nous regarderons leur précision à travers les 4 critères de mesure d'erreur de prévisions déjà calculés ci-dessus pour les 5 modèles.

Tableau n°26 : Taux d'erreur des prévisions naïves

	MSE	**MAE**	**MDSE**	**MDAE**
Prévision naïve	0.0055	0.0621	0.004	0.0631

Nous nous apercevons que les 4 taux d'erreurs avec différentes formules sont nettement supérieurs aux taux des prévisions linéaires et du modèle AR. C'est donc toujours le modèle ARX qui est le meilleur, car il a des taux d'erreur nettement inférieurs aux autres modèles. Se servir de variables explicatives pour prédire au mieux une série est ainsi, une fois de plus, un bon moyen de minimiser les erreurs. Le graphique qui reprend les différentes prévisions auxquelles nous avons ajouté les prévisions naïves, se trouve en annexe n°11, on voit clairement qu'elles sont très volatiles rejoignant ainsi les fortes variations des valeurs observées. Cependant c'est comme si elles suivaient les mêmes variations mais avec une retard de 2 à 3 mois : regardons par exemple la chute du cours (valeurs observées en rose sur le graphique) en avril 2019, nous retrouvons aussi une forte baisse dans les prévisions benchmark mais seulement 2 mois plus tard c'est à dire en juin 2019.

G) Tests et sélection du meilleur modèle

i. Test de Diebold-Mariano

Enfin, malgré le fait que nous ayons déjà vu quel modèle prédisait au mieux les variations du S&P GSCI, nous réalisons maintenant le test de Diebold-Mariano entre les modèles des questions 3.b) et 3.d) et la prévision naïve, ainsi que les modèles entre eux. En effet, même si l'erreur au carré moyenne (MSE) est supérieure pour les prévisions naïves que pour les modèles AR(1) et ARX par exemple, on ne peut pas savoir si les prévisions des modèles sont statistiquement différentes les une des autres. Les mathématiciens Diebold et Mariano ont donc répondu à cette problématique en proposant en 1995 un test basé sur l'hypothèse d'égalité prédictive, tel que si l'on compare le modèle 1 au modèle 2 on a :

H_0 : Il n'y a pas de différence significative entre les deux modèles

H_1 : avec l'alternative c=("less"), le modèle 2 est moins précis que le modèle 1

Nous commençons donc par appliquer ce test aux 5 modèles des questions précédentes (modèle 1) par rapport à la prévision naïve (modèle 2), on pourra voir ainsi si la prévision naïve est moins précise que les 5 autres modèles - à mesure que la probabilité du test diminue. On obtient donc :

Tableau n°27 : DM test entre les différents modèles et les prévisions naïves

	AR(1)	AR(3)	ARIMA(1,1,0)	LED HW	ARX
p-value*	9.861e-11	7.742e-11	5.731e-11	0.08213	3.394e-12

* avec : alternative='l'

On voit alors que la prévision naïve est moins précise que tous les modèles excepté la méthode Holt Winters pour laquelle on accepte l'hypothèse selon laquelle il n'y a pas de différence significative entre celle-ci et la prévision naïve, au seuil de 5%. Les prévisions benchmark sont statistiquement moins bonnes que celles des modèles linéaires et modèle ARX, au seuil de risque $\alpha = 0,1$. Voyons à présent ce qu'il en est au sein des modèles estimés (hors prévision naïve), nous verrons si statistiquement parlant c'est bien le modèle ARX qui est le plus précis.

Tableau n°28 : Test DM entre les modèles

	AR(1)	AR(3)	ARIMA(1,0,1)	LED HW	ARX
AR(1)		0.5902	0.5626	0.00075	0.9898
AR(3)	0.4098		0.4601	0.00077	0.9922
ARIMA(1,0,1)	0.4374	0.5399		0.00071	0.9925
LED HW	0.9992	0.9992	0.9993		0.9999
ARX	0.01023	0.0078	0.0075	0.00012	

On voit sur la première ligne du tableau n°28 que les prévisions du modèle AR(1) sont plus précises que celles issues de la méthode HW puisque la probabilité associée au test de Diebold-Mariano tel que AR(1)=modèle 1 et LED HW=modèle 2 (cf. hypothèses du test) est très proche de 0 ce qui nous amène à accepter l'hypothèse alternative H_1. On voit lorsque l'on regarde les lignes et colonnes des p-value de LED HW que ce dernier est statistiquement moins précis que tous les autres modèles. De même, le modèle avec variables explicatives est lui, plus précis que tout autre modèle au seuil de risque de 5%. Il n'existe pas de différence significative au sein des modèles ARIMA ; tous ces constats viennent confirmer ce que nous avions pu déjà relevé au cours de l'élaboration et de la comparaison des modèles.

ii. Test multiple de Mariano et Preve

Finalement, pour conclure cette analyse nous appliquons le test multiple de Mariano et Preve, qui nous permettra de définir le meilleur modèle (bien que nous ayons de bonnes intuitions déjà), et qui, nous le rappelons, se base sur les hypothèses suivantes :

H_0 : Les prévisions évaluées ont la même précision

H_1 : L'Equal predictive accuracy (EPA) n'est pas atteinte

Pour effectuer le test multiple nous avons regroupé les différentes prévisions dans un data frame appelé 'PREVDM' constitué de :

- **prev1** : prévisions réalisées par le modèle AR(1)
- **prev2** : prévisions réalisées par le modèle AR(3)

- **prev3** : prévisions réalisées par le modèle ARIMA(1,0,1)
- **prev4** : prévisions réalisées par la méthode LED Holt Winters
- **forecast** : prévisions réalisées par le modèle ARX
- **prev5** : prévisions réalisées par la prévision naïve

Tableau n°29 : Résultat du test multiple Mariano-Preve

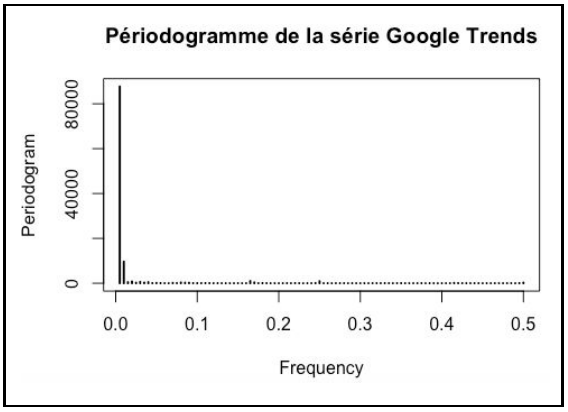
```
####################################################
Models with outstanding predictive ability:

         Rank      Sc Mean loss
prev1       1 -1.1241    0.0024
prev2       2 -0.5479    0.0025
forecast                 0.0013

p-value: 0.7961

Number of eliminated models:  3
####################################################
```

La probabilité ressortissant de ce test s'élève à 0.7961 : cela signifie que les prévisions des 3 modèles retenus par le test ont la même précision, il s'agit des modèles AR(1), AR(3) et ARX. En revanche les autres prévisions ont été éliminées puisqu'elles étaient moins précises que celles des 3 modèles énoncés précédemment. Cela vient confirmer nos intuitions ; le modèle ARX ainsi que les modèles AR(1) et AR(3) sont ceux qui ont la plus grande précision de prévision où ARX a la perte minimale et donc est le plus précis. De même, les prévisions issues des modèles ARIMA(1,0,1), HW et prévision naïve sont elles moins précises que les précédentes et ont donc été éliminées lors du test multiple de Mariano et Preve qui consistait à trouver quel modèle a la plus grande qualité prévisionnelle.

V- Annexes

A) Annexes de l'analyse de Google Trends

Annexe n°1 : Périodogramme de la série google trends



Annexe n°2 : Résultats Ts outliers

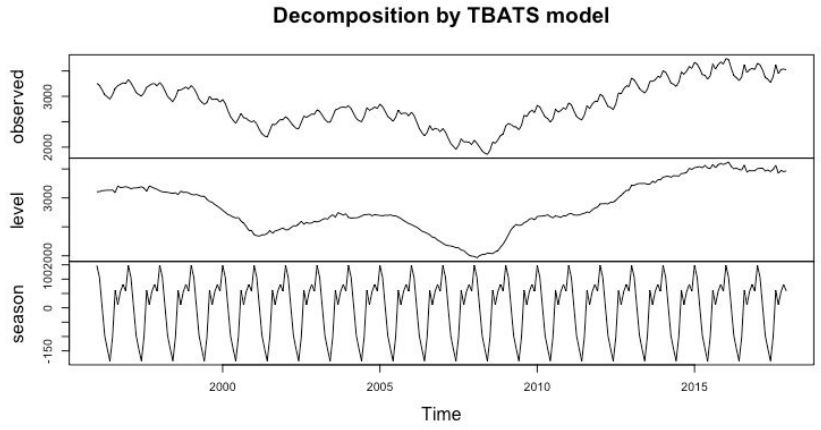
```
> show(outlier)
Series: xx
Regression with ARIMA(2,0,0)(0,1,1)[12] errors

Coefficients:
         ar1     ar2     sma1    LS61    LS66     LS85    AO108     LS145   AO149    AO183
      0.5993  0.3696  -0.4645  8.0374  9.0432  11.2216  -6.4333  -12.7723  7.3576  -6.8596
s.e.  0.0727  0.0721   0.0930  2.0599  2.0200   2.0265   1.7491    2.0273  1.7640   1.8728

sigma^2 estimated as 6.464:  log likelihood=-427.63
AIC=877.25   AICc=878.8   BIC=912.56

Outliers:
  type ind    time coefhat  tstat
1   LS  61 2009:01   8.037  3.902
2   LS  66 2009:06   9.043  4.477
3   LS  85 2011:01  11.222  5.537
4   AO 108 2012:12  -6.433 -3.678
5   LS 145 2016:01 -12.772 -6.300
6   AO 149 2016:05   7.358  4.171
7   AO 183 2019:03  -6.860 -3.663
```

Annexe n°3 : Série désaisonnalisée

```
> final(seasX)
           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2004 40.57491 41.90986 40.57895 41.34668 40.18907 42.56594 40.87195 42.99230 41.95715 42.07965 39.42967 38.27787
2005 39.03151 39.18825 37.65909 38.48991 38.55757 37.13685 37.87643 36.08555 35.59886 34.88826 34.17080 34.29599
2006 33.27717 33.31718 33.92017 32.34233 31.21138 30.82893 31.62455 31.75057 29.90783 28.55013 30.20639 31.93593
2007 28.68428 27.42885 28.14138 27.00017 28.27494 28.11101 27.40390 26.84420 25.79355 25.23314 25.34683 28.67522
2008 25.50984 24.36484 25.52031 27.58709 27.12784 26.82492 25.71948 26.83201 24.95745 27.70693 28.91295 30.51650
2009 36.11161 39.10275 41.93836 46.28254 49.36593 54.01138 56.52913 58.90345 62.26068 62.22251 67.68790 66.82557
2010 66.50131 66.65984 68.49666 70.24257 70.43507 68.85391 70.50734 72.24863 74.31694 75.57759 70.89132 70.48936
2011 82.41071 81.91116 82.32119 82.87160 85.42553 81.74242 81.90142 81.90394 83.90406 87.08285 89.53923 90.92880
2012 90.13683 89.53761 90.75734 86.66394 87.01243 88.37865 89.95795 90.08598 89.76174 91.58844 92.51126 88.99818
2013 92.81184 95.32495 90.39599 92.34802 90.78259 92.97139 91.54658 89.66167 90.16529 86.01986 86.97774 85.73332
2014 88.33033 86.19846 87.72820 86.38198 85.33029 85.68231 86.84828 86.83098 88.67538 88.42085 88.97064 87.27461
2015 87.78449 89.14877 87.95765 87.77840 86.12633 88.33650 88.50348 85.51480 86.78192 85.75889 82.88457 85.20177
2016 71.89280 72.93118 70.72215 72.38359 76.02206 71.77815 70.14436 71.76152 70.77867 70.15899 71.44798 70.68802
2017 68.96121 70.55333 70.90586 68.29112 68.06748 67.38154 68.91186 71.03028 70.10465 71.49601 71.83239 69.24430
2018 69.53260 67.83237 72.11443 68.81013 65.34071 66.29341 63.31457 61.70693 61.13177 60.29845 58.55739 62.57762
2019 60.63033 59.16069 56.48028 58.60733 59.86221 59.92412 60.83618 60.00666 59.71154 60.35897 59.69443 59.00447
2020 59.14078 58.26444 58.69445
```

Annexe n°4 : Décomposition TBATS

```
> p = predict(m, 12, prediction.interval = TRUE)
> p
                fit      upr      lwr
Apr 2020 55.10267 61.40183 48.80351
May 2020 53.19562 60.60256 45.78869
Jun 2020 53.62102 62.15297 45.08906
Jul 2020 54.39797 64.07798 44.71795
Aug 2020 54.03614 64.89048 43.18181
Sep 2020 67.56463 79.62135 55.50791
Oct 2020 65.68215 78.97029 52.39401
Nov 2020 59.85065 74.39973 45.30157
Dec 2020 52.48341 68.32312 36.64370
Jan 2021 62.83173 79.99170 45.67175
Feb 2021 59.93902 78.44873 41.42930
Mar 2021 59.87493 79.76360 39.98625
```



Annexe n°5 : Prévisions HW à un pas

```
> y_HW
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2004 41.00000 43.00000 41.00000 40.00000 36.00000 42.00000 39.00000 45.00000 51.00000 45.00000 40.00000 31.00000
2005 39.00000 40.00000 38.00000 36.00000 36.00000 36.00000 35.00000 38.00000 45.00000 38.00000 34.00000 27.00000
2006 35.00000 34.00000 34.00000 29.00000 29.00000 30.00000 29.00000 33.00000 38.00000 33.00000 30.00000 24.00000
2007 31.00000 28.00000 28.00000 25.00000 26.00000 26.00000 26.00000 27.00000 33.00000 30.00000 26.00000 21.00000
2008 28.00000 26.00000 25.00000 26.00000 24.00000 25.00000 24.00000 25.00000 35.00000 32.00000 28.00000 24.00000
2009 38.00000 40.00000 43.00000 44.00000 45.00000 53.00000 54.00000 57.00000 72.00000 67.00000 69.00000 60.00000
2010 68.00000 68.00000 70.00000 68.00000 66.00000 66.00000 67.00000 71.00000 84.00000 80.00000 74.00000 63.00000
2011 85.00000 84.00000 84.00000 79.00000 82.00000 78.00000 77.00000 80.00000 94.00000 92.00000 92.00000 83.00000
2012 95.00000 93.00000 92.00000 83.00000 83.00000 84.00000 87.00000 87.00000 98.00000 98.00000 95.00000 80.00000
2013 98.00000 99.00000 91.00000 90.00000 86.00000 88.00000 89.00000 86.00000 100.00000 92.00000 88.00000 78.00000
2014 93.00000 90.00000 89.00000 83.00000 80.00000 82.00000 84.00000 82.00000 100.00000 94.00000 89.00000 80.00000
2015 92.00000 93.00000 91.00000 84.00000 80.00000 86.00000 85.00000 81.00000 97.00000 91.00000 84.00000 78.00000
2016 75.00000 78.00000 74.00000 68.00000 72.00000 68.00000 65.00000 69.00000 81.00000 75.00000 73.00000 63.00000
2017 74.00000 74.00000 74.00000 63.00000 64.00000 64.00000 64.00000 68.00000 79.00000 78.00000 73.00000 61.00000
2018 75.00000 71.00000 75.00000 65.00000 61.00000 62.00000 60.00000 58.00000 69.00000 67.00000 60.00000 55.00000
2019 66.00000 62.00000 58.00000 52.61621 51.59616 52.24106 51.16124 51.37312 60.67920 58.06243 53.16017 47.16855
2020 55.14287 52.25461 51.17281
```

Annexe n°6 : Prévisions Arima(0,1,1)(0,1,1)

```
> y_HW
         Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2004 41.00000 43.00000 41.00000 40.00000 36.00000 42.00000 39.00000 45.00000 51.00000 45.00000 40.00000 31.00000
2005 39.00000 40.00000 38.00000 36.00000 36.00000 36.00000 35.00000 38.00000 45.00000 38.00000 34.00000 27.00000
2006 35.00000 34.00000 34.00000 29.00000 29.00000 30.00000 29.00000 33.00000 38.00000 33.00000 30.00000 24.00000
2007 31.00000 28.00000 28.00000 25.00000 26.00000 26.00000 26.00000 27.00000 33.00000 30.00000 26.00000 21.00000
2008 28.00000 26.00000 25.00000 26.00000 24.00000 25.00000 24.00000 25.00000 35.00000 32.00000 28.00000 24.00000
2009 38.00000 40.00000 43.00000 44.00000 45.00000 53.00000 54.00000 57.00000 72.00000 67.00000 69.00000 60.00000
2010 68.00000 68.00000 70.00000 68.00000 66.00000 66.00000 67.00000 71.00000 84.00000 80.00000 74.00000 63.00000
2011 85.00000 84.00000 84.00000 79.00000 82.00000 78.00000 77.00000 80.00000 94.00000 92.00000 92.00000 83.00000
2012 95.00000 93.00000 92.00000 83.00000 83.00000 84.00000 87.00000 87.00000 98.00000 98.00000 95.00000 80.00000
2013 98.00000 99.00000 91.00000 90.00000 86.00000 88.00000 89.00000 86.00000 100.00000 92.00000 88.00000 78.00000
2014 93.00000 90.00000 89.00000 83.00000 80.00000 82.00000 84.00000 82.00000 100.00000 94.00000 89.00000 80.00000
2015 92.00000 93.00000 91.00000 84.00000 80.00000 86.00000 85.00000 81.00000 97.00000 91.00000 84.00000 78.00000
2016 75.00000 78.00000 74.00000 68.00000 72.00000 68.00000 65.00000 69.00000 81.00000 75.00000 73.00000 63.00000
2017 74.00000 74.00000 74.00000 63.00000 64.00000 64.00000 64.00000 68.00000 79.00000 78.00000 73.00000 61.00000
2018 75.00000 71.00000 75.00000 65.00000 61.00000 62.00000 60.00000 58.00000 69.00000 67.00000 60.00000 55.00000
2019 66.00000 62.00000 58.00000 52.18554 51.20456 51.60590 50.75097 51.05079 60.33009 57.90754 53.79700 47.53973
2020 55.95901 54.22919 53.16917
```

Annexe n°7 : Autres prévisions

• Auto Arima

```
2019  66.00000  62.00000  58.00000  51.56084  49.89151  50.12528  48.92210  48.96259  61.04769  57.22129  52.03977  43.26035
2020  53.90620  52.00224  50.59184
```

• ETS

```
2019  66.00000  62.00000  58.00000  52.44930  51.01101  51.52941  51.04391  51.03792  64.21447  60.03404  55.55237  46.21742
2020  57.37586  56.46871  54.98518
```

• STLm

```
2019  66.00000  62.00000  58.00000  54.92596  54.87455  56.22721  55.68050  57.11167  68.34253  63.02311  58.37928  50.64902
2020  60.95585  59.61946  58.73388
```

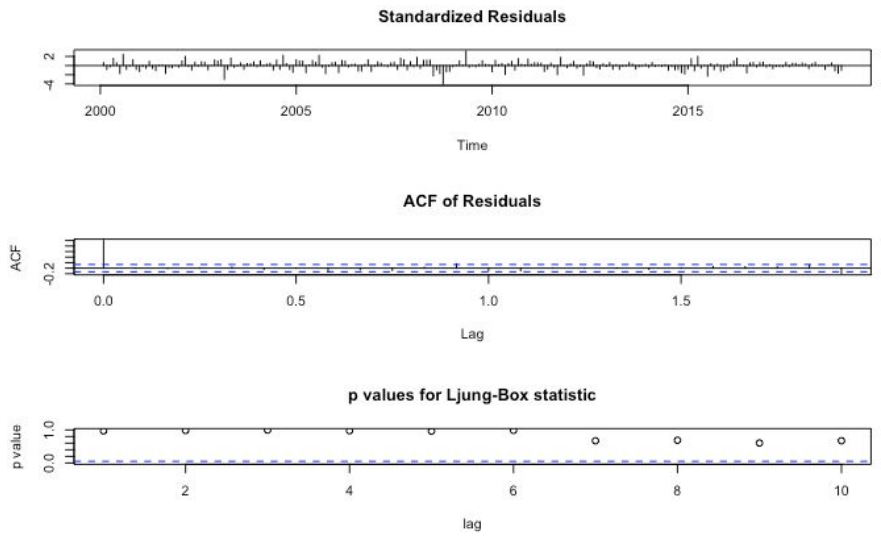
B) Annexes de l'analyse du S&P GSCI

Annexe n°8 : Test de Dickey-Fuller augmenté sur le rendement de Yt

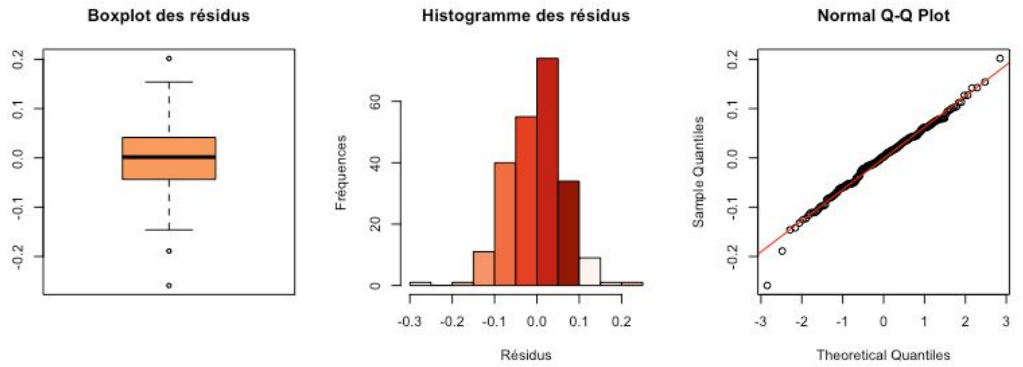
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.257423 -0.036922  0.008401  0.044921  0.206439

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.82372    0.08479  -9.715   <2e-16 ***
z.diff.lag -0.02159    0.06521  -0.331    0.741
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06338 on 235 degrees of freedom
Multiple R-squared:  0.4206,    Adjusted R-squared:  0.4156
F-statistic: 85.29 on 2 and 235 DF,  p-value: < 2.2e-16
```

Annexe n°9 : Modèle AR(p,d,q) et analyse de ses résidus question 3.b)

```
> ARpdq
Series: yarima
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1      ma1    mean
      0.3978  -0.2407  0.0046
s.e.  0.3691   0.3905  0.0053

sigma^2 estimated as 0.004071:  log likelihood=304.07
AIC=-600.14   AICc=-599.96   BIC=-586.45
> coeftest(ARpdq)

z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
ar1        0.3977591  0.3690570  1.0778   0.2811
ma1       -0.2406866  0.3905018 -0.6164   0.5377
intercept  0.0046412  0.0052983  0.8760   0.3810
```



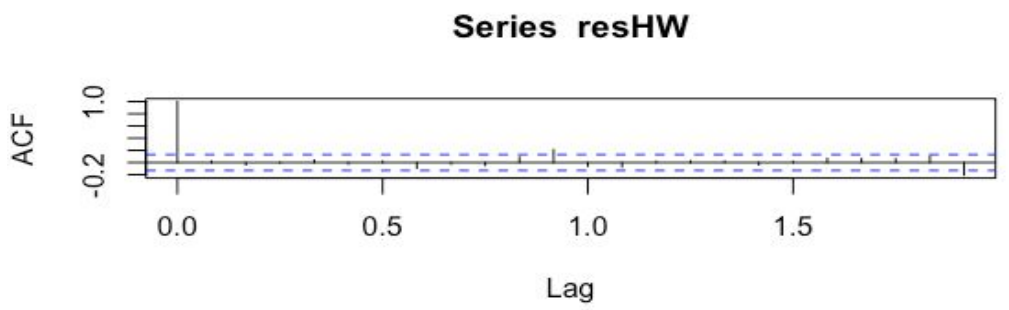

```
> Box.test(resARpdq, type = c("Ljung-Box"))

        Box-Ljung test

data:  resARpdq
X-squared = 0.0011792, df = 1, p-value = 0.9726
```

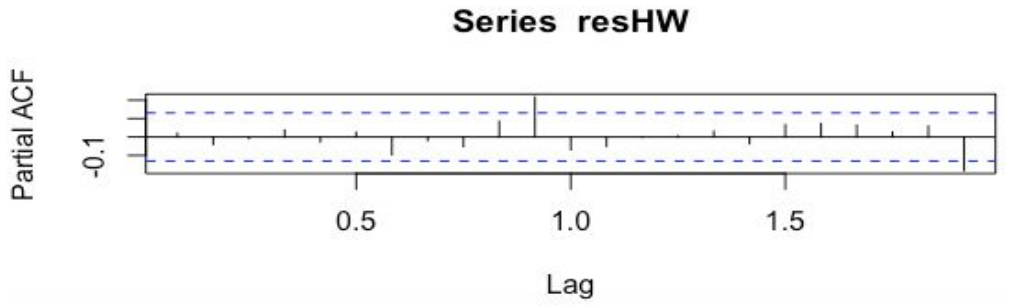


```
> ks.test(resARpdq, "pnorm", mean(resARpdq), sd(resARpdq))

        One-sample Kolmogorov-Smirnov test

data:  resARpdq
D = 0.061195, p-value = 0.3631
alternative hypothesis: two-sided
```

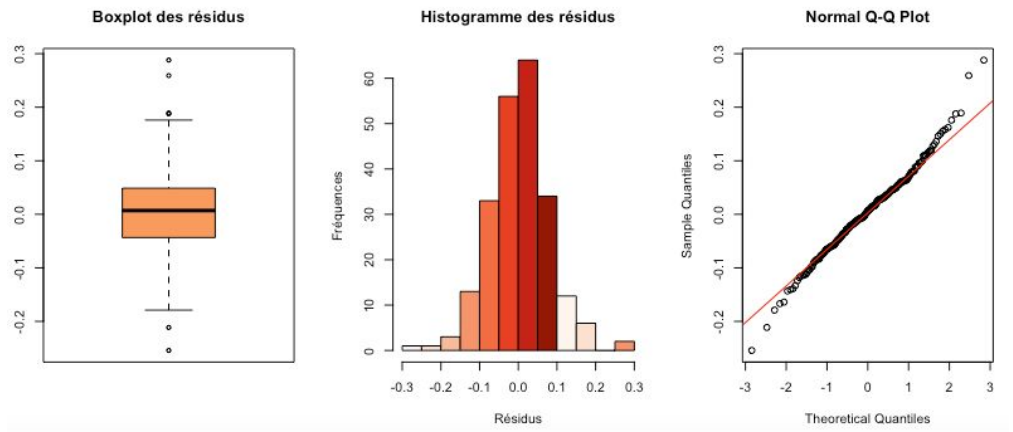
Annexe n°10 : Analyse des résidus du modèle LED Holt Winters

```
> HW
Holt-Winters exponential smoothing with trend and without seasonal component.

Call:
HoltWinters(x = yarima, gamma = FALSE)

Smoothing parameters:
 alpha: 0.4555995
 beta : 0.1503941
 gamma: FALSE

Coefficients:
         [,1]
a -0.07658715
b -0.01095522
```



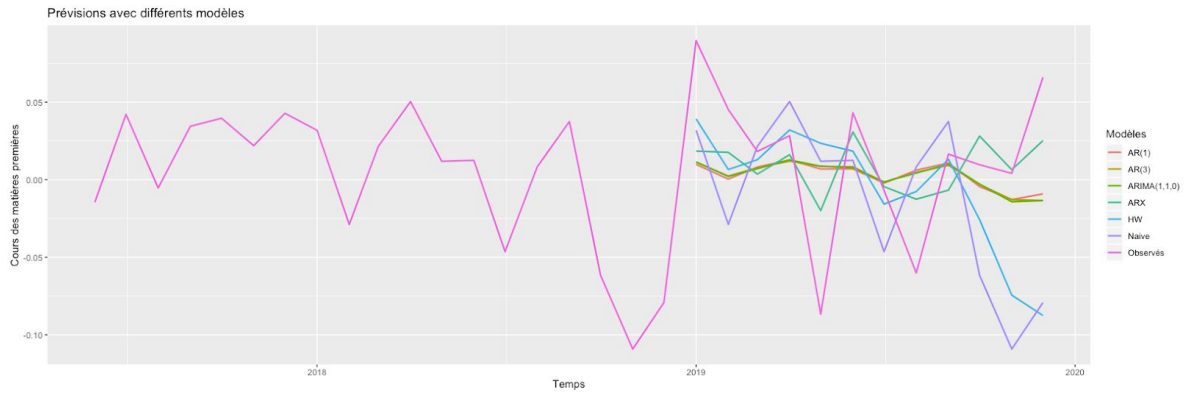


```
> Box.test(resHW, type = c("Ljung-Box"))

        Box-Ljung test

data:  resHW
X-squared = 0.10068, df = 1, p-value = 0.751
```




```
> ks.test(resHW, "pnorm", mean(resHW), sd(resHW))

        One-sample Kolmogorov-Smirnov test

data:  resHW
D = 0.054391, p-value = 0.5186
alternative hypothesis: two-sided
```

Annexe n°11 : Graphique des différentes prévisions + prévision benchmark



VI- Table des matières

Résumé **1**

I- Série saisonnière et analyse cyclique **3**

Analyse préliminaire 3

i. Présentation et caractérisation de la série 3

ii. Détection de la saisonnalité et schéma de décomposition 6

iii. Détection des points atypiques par TRAMO et RegArima 7

Désaisonnalisation 11

i. TRAMO-SEATS 11

ii. ARIMA X-13 12

Analyse cyclique 14

i. Graphique série CVS et Trend 14

ii. Durée, amplitude, sévérité 16

iii. Série lissée et cycle d'accélération 16

II- Série saisonnière et prévision **20**

Prévisions récursives avec un pas de 1 mois 20

i. Prévisions avec TRAMO 20

ii. Prévisions avec Holt-Winters, SARIMA, BATS, ETS, STLm et NNETAR 21

Comparaison des modèles Arima 22

Graphique des prévisions 23

Erreurs de prévision 24

Tests et sélection du meilleur modèle 26

i. Diebold-Mariano 26

ii. Mariano et Preve 27

Graphique récapitulatif 28

III- Série non saisonnière et prévision **30**

Analyse préliminaire 30

i. Présentation et caractérisation de la série étudiée 30

ii. Points atypiques 33

iii. Stationnarité série 33

Estimation de modèles et vérification des résidus 36

Prévisions à partir des 4 modèles précédents 42

Prévision avec variables explicatives 43

i. Stationnarisation des variables explicatives 43

ii. Modélisation et prévision avec variables explicatives 44

iii. Justification du choix des variables 46

Représentation graphique des différentes prévisions 47

Erreurs de prévisions et prévision naïve 48
Tests et sélection du meilleur modèle 50
i. Test de Diebold-Mariano 50
ii. Test multiple de Mariano et Preve 51

V- Annexes **53**
Annexes de l'analyse de Google Trends 53
Annexes de l'analyse du S&P GSCI 55