Prévision du Nombre de Nuités en Corse et du Prix du Brent

Monsieur Darne

ROMAND Kyllien
DEL'CHATEAU Jean-Baptiste

2020/2021

Résumé

L'analyse que nous allons effectuer dans ce dossier comporte sur l'analyse de deux séries temporelles, son objectif est de comprendre leurs évolutions et prédire leur variations futures. Nous nous sommes donc intéressés aux nombre de nuitées totales en Corse pour notre première série. Pour notre deuxième série, nous sommes partis sur les prix du pétrole BRENT en utilisant 5 variables explicatives. Nous avons donc une série qui est saisonnière et notre deuxième série qui ne l'est pas. Nos deux séries s'étendent de janvier 2011 à décembre 2019 pour la première et de août 2004 à décembre 2019 pour la deuxième.

Après avoir analysé notre variable concernant le nombre de nuitées en Corse, nous avons procédé à une désaisonnalisation via les méthodes X13-ARIMA-SEATS, à partir de cette série désaisonnalisée, nous avons effectué différentes estimations avec différentes méthodes. Après les avoir représentés graphiquement, nous avons calculé les erreurs de prévision (la MSE) et effectué des tests de précision (test de Diebold Mariano), au regard de ces tests, un modèle semble être meilleur que les autres, le modèle STL au contraire, le modèle qui semble être le moins performant est le modèle STS.

Par ailleurs, sur notre série non saisonnière, après avoir vérifié la stationnarité de celle-ci, nous avons estimé et prévu des modèles linéaires via des modèles AR(1), AR(p), ARIMA (p,d,q) ou encore le modèle Holt-Winters. Une fois ces estimations et prévisions faites, nous avons effectué des prévisions en ajoutant nos variables explicatives. Nous avons fini cette partie en faisant les erreurs de prévisions (la MSE) ainsi que les tests de précision (test de Diebold Mariano). Au regard de ces tests, nous trouverons un meilleur modèle avec 2 variables explicatives, modèle servant à faire la prévisions des prix du pétrole. Les prévisions se feront avec 3 modèles linéaires et un modèle ARX.

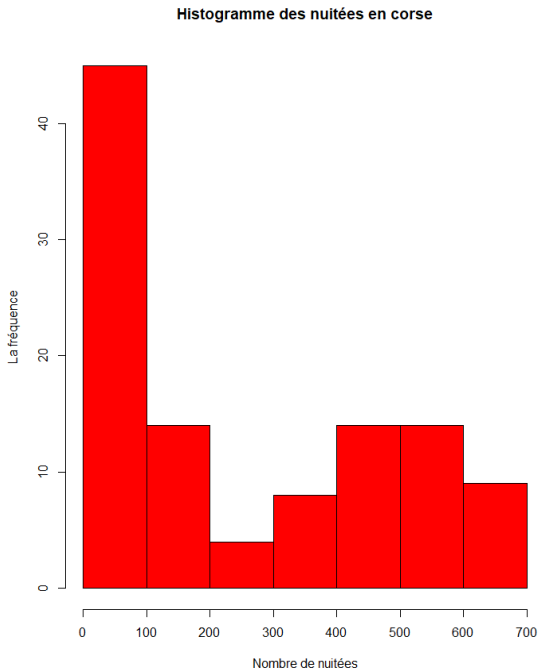
Sommaire

Série saisonnière mensuelle : Le Tourisme en Corse

Analyse Exploratoire

La série étudiée dans cette partie du dossier correspond aux nuitées totales dans la région de Corse en série mensuelle de janvier 2011 à décembre 2019[1].
Ces données proviennent de l'enquête de fréquentation dans l'hôtellerie qui permet d'avoir, sur l'ensemble du territoire français, la fréquentation touristique dans les hôtels en nombre de nuitées. Cette enquête est mise en place par la Direction des statistiques d'entreprise et est disponible sur le site de l'INSEE.
Cette série a été choisie car le tourisme en Corse est de fait saisonnier puisque les touristes viennent en été et délaissent la Corse en hiver.
Pour débuter ce dossier, nous allons commencer par regarder la distribution de notre série avec un histogramme. Pour analyser ce graphique, nous allons dire que pendant 45 % de l'année, les chambres d'hôtel sont occupées entre 0 et 100 nuits sur la période 2011-2019.

Graphique 1 :histogramme de notre série



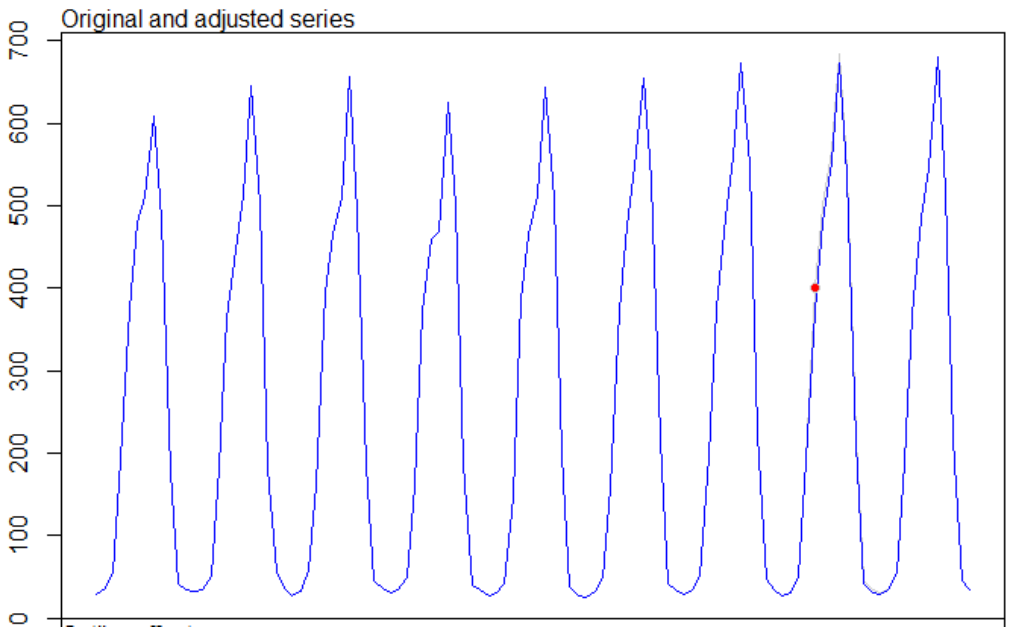
[1] Enquête de fréquentation dans l'hôtellerie | Insee

Afin de pouvoir commencer à étudier la série, il est nécessaire de vérifier si celle-ci ne contient pas de valeurs atypiques (*outliers*), pour cela il faut regarder tout d'abord graphiquement si visuellement il en apparaît à l'aide de la fonction "tso".

Tableau 1 : récapitulatif de notre outliers

```
Outliers:
  type ind    time coefhat tstat
1   TC  89 2018:05   27.91 3.378
```

Graphique 2 : valeurs atypiques



Graphiquement un point ressort, c'est pourquoi il est obligatoire d'effectuer un test statistique sous R : "tso", qui permet de savoir quelles valeurs sont des *outliers* (des valeurs aberrantes). Dans le cas de cette série, il en ressort que la valeur n°89 est une valeur atypique. (Annexe 1).

Cette valeur aberrante, correspond au mois de mai de 2018 : durant cette période, une crise de déchets se produisait en Corse, les éboueurs ne ramassaient plus les

déchets. Ce qui peut expliquer cette baisse de fréquentation des nuitées pour cette période.

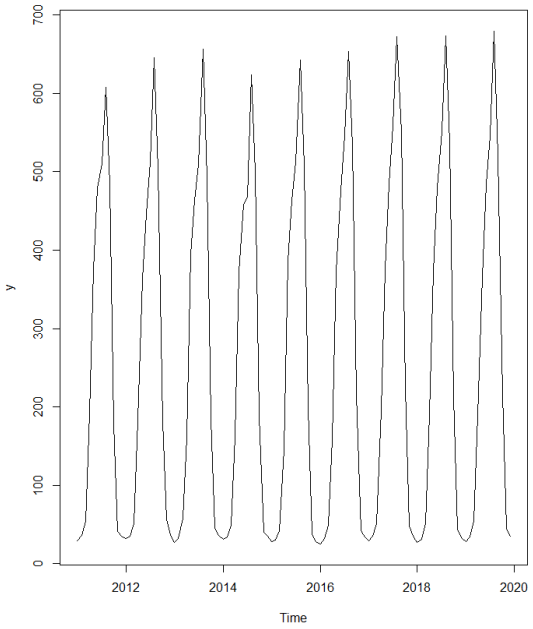
Nous allons maintenant voir les statistiques descriptives de notre série.Dans le tableau ci-dessous, via le tableau ci-dessous qui en fait un résumé. La moyenne de notre série est d'environ 259 nuitées avec un écart-type de 226 nuitées. Concernant la normalité de notre série, avec un coefficient de skewness de 0.36, la distribution va être légèrement décalée à gauche de la médiane avec une queu de distribution décalée vers la droite. Concernant l'aplatissement de notre série, le coefficient de kurtosis est de 1.5771, nous allons avoir une distribution leptokurtique, ce qui signifie que notre distribution va donc être plus épaisse.

Tableau 2 : résumé des statistiques descriptives

	Moyenne	Ecart-type	Skewness	Kurtosis
Valeur	259	226	0.3969	1.5771

Nous allons maintenant détecter la saisonnalité de notre série. Avec notamment, la library "seasonal", qui compile trois tests en même temps. Graphiquement, nous pouvons supposer que notre série est saisonnière mais il faut le vérifier par les tests. La supposition de la saisonnalité se fait par la courbe, avec fluctuations qui se ressemblent, comme une sorte d'événement qui se passe tous les ans et à la même période de l'année. Pour notre série on peut supposer que c'est la saison d'été, avec l'arrivée des vacanciers qui fait augmenter le nombre de nuitées en Corse.

Graphique 3 : Nombre de nuités par mois en Corse



Il est nécessaire de vérifier, pour la suite du dossier, si la série étudiée est saisonnière et quel schéma de décomposition est à adopter (additif, multiplicatif, ou log-multiplicatif).

Pour vérifier si la série est saisonnière, il existe de multiples tests qui seront mis en annexe (annexe 2 à 6), mais le plus puissant est le test "*isSeasonal"* qui permet de combiner deux tests (le test QS modifié et le test de Kruskal-Wallis). Le test QS permet de vérifier s'il existe des auto-corrélations positives aux retards saisonniers (12 pour cette série), alors que le test de Kruskal-Wallis permet de contrôler les différences significatives entre les différentes périodes, qui correspondent aux années pour notre série. La figure 1 présente ce test sur la série, et en regardant le résultat, qui est un booléen, la série est saisonnière. Concernant les 5 tests que nous avons en annexes, si la série est saisonnière, alors les p-value de ces tests seront inférieur à 0.05, notre seuil de risque, ce qui est le cas. Ces résultats nous confortent dans le fait que notre série est saisonnière.

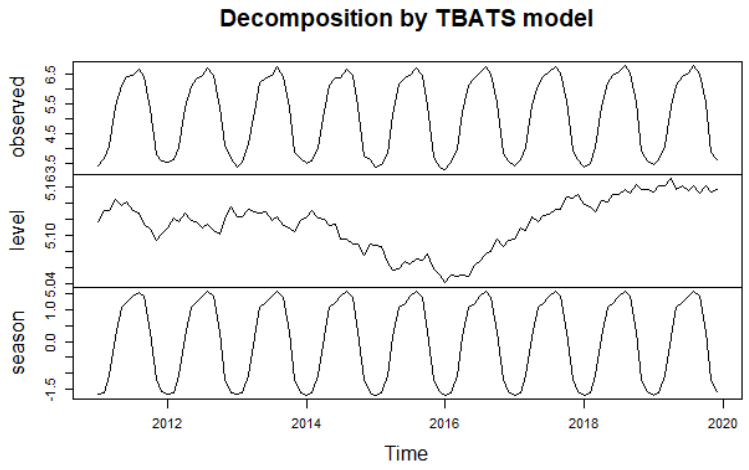
Figure 1 : Test de Saisonnalité

```
is <- isSeasonal(yy, test="wo")
> show(is)
[1] TRUE
```

Nous allons maintenant décomposer notre série. Avant de partir dans la désaisonnalité, nous allons voir si nous sommes en présence d'une spécification additive ou multiplicative. Au vue du graphique 3, nous allons dire que nous sommes en présence d'une série additive. Les fluctuations sont toujours les mêmes, elles n'augmentent pas avec le temps. Nous pensons utiliser la méthode des bandes, ce qui va créer deux bandes parallèles passant par les minima et maxima de chaque saison. Si ces deux droites sont parallèles, nous serons alors dans un modèle additif, si elles ne le sont pas ca sera alors un modèle multiplicatif.

Sur le graphique ci-dessous, nous pouvons voir la décomposition de notre série avec la méthode TBATS. Sur ce même graphique, nous pouvons voir le schéma de décomposition de notre série, nous pouvons donc affirmer que celle-ci est additive, nous voyons également que les valeurs observées sont équivalente aux valeurs saisonnières

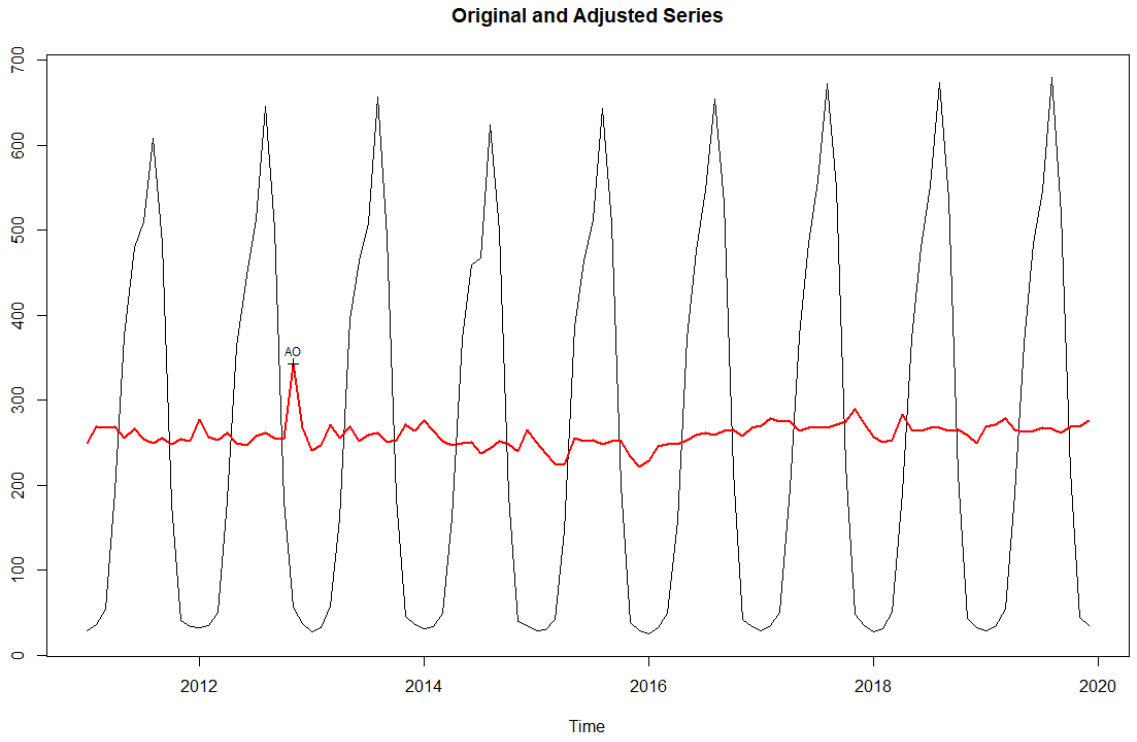
Graphique 4 : décomposition de notre série



Désaisonnalisation et décomposition

Dans cette deuxième partie du dossier, nous allons voir la désaisonnalisation et la décomposition de notre série. La décomposition va se faire via les méthodes X-13-ARIMA-SEATS. La désaisonnalisation a pour objectif d'éliminer les composantes saisonnières de notre série pour mettre en évidence les autres éléments qui jouent un rôle important dans l'analyse économique. Sur le graphique de la désaisonnalisation, nous voyons que nous avons encore une valeur atypique qui n'a pas été détectée par la méthode TSO, de plus nous avons un "additive outliers (AO)", lors que précédemment c'était un "temporary change (TC)". Avec la fonction "tsoutliers" (Annexe 7), nous savons que c'est la valeur 43 de nos données qui est abherrante, nous allons donc l'enlever. La valeur 43 correspond au mois de juin 2014, ce qui correspond à la dépose des armes du front de libération national de la Corse, ce qui a pu engendrer une augmentation du nombre de touristes en Corse. En regardant sur le graphique, nous voyons que cette valeur atypique ne correspond pas au point sur le graphique, le point se situe entre 2012 et 2013, notre valeur atypique et en 2014.

Graphique 5 : Série Brute et Série Ajusté par la méthode X13-ARIMA-SEATS



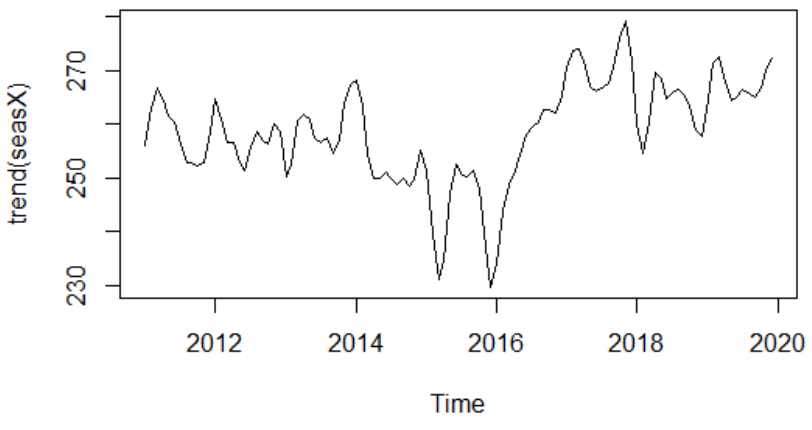
En appliquant la méthode X13, nous trouvons que notre série suit un modèle additif ARIMA (1 0 0)(0 1 1). Nous voyons que notre partie saisonnière et non saisonnière sont toutes les deux significatives à hauteur de 5%. la valeur atypique que nous arrivions pas trouver avec “tsoutliers” correspond au mois de novembre 2012 et celle-ci est significative à 5%. De plus, en annexe 8, nous pouvons trouver le graphique concernant l’irrégularité de notre série, et nous observons un pic au niveau de novembre 2012, ce qui correspond à notre valeur atypique.

Figure 2 :summary de notre modèle

```
Call:
seas(x = adj)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
Weekday           -0.00362    0.00144  -2.514 0.011953 *
Easter[1]          0.06536    0.01754   3.726 0.000195 ***
AO2012.Nov         0.27687    0.04796   5.773 7.77e-09 ***
AR-Nonseasonal-01  0.50960    0.08685   5.867 4.43e-09 ***
MA-Seasonal-12     0.48944    0.08598   5.692 1.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

SEATS adj.  ARIMA: (1 0 0)(0 1 1)  Obs.: 108  Transform: log
AICc: 703.4, BIC: 717.8  QS (no seasonality in final):    0
Box-Ljung (no autocorr.): 26.56   Shapiro (normality): 0.9737 *
```

Suite à l’utilisation de la méthode X13, une décomposition est faite, voir graphique ci-dessous. Nous pouvons voir qu'il y a une baisse de touristes entre 2015 et 2016 et une légère hausse entre 2017 et 2020 comparativement au début de la série. Avec le temps nous voyons que de plus en plus de monde vont en corse pour les vacances.

Graphique 6 : Tendance de la série



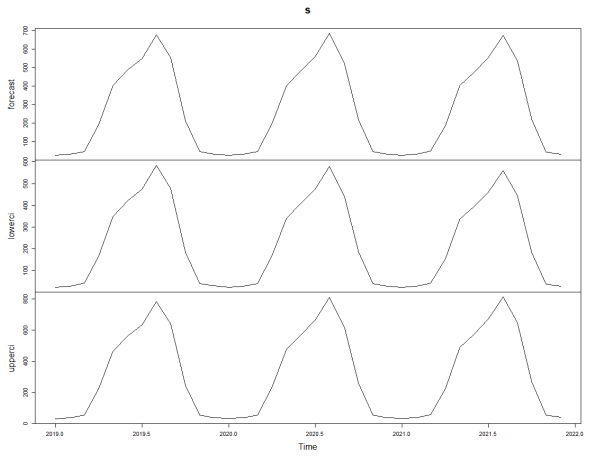
Prévision : sur une année avec un pas de un mois

Pour faire la prévision de notre série, nous allons utiliser 9 méthodes : X13-ARIMA-SEATS, STL, STS, BSTS, TBATS, Holt-Winters, ETS, NNETAR et SARIMA. Nous déterminerons par la suite à l'aide du MSE puis du test de Diebold-Mariano quel est le meilleur modèle par rapport à chacun de ces deux tests. Les prévisions que nous allons faire vont être fait sur l'année 2019, nous trouvons ca plus logique du fait que les données de l'année 2020 sont biaisées avec la pandémie mondiale, les hôtels et chambres d'hôtes n'ont donc pas accueillis autant de monde que les autres années, avec un touriste étrangées impactés. Nous avons donc décidé de supprimer l'année 2019 de nos données pour cette partie.

Méthode X13-ARIMA-SEATS

Cette méthode permet de faire une prévision sur 3 ans , avec 3 types de prévisions, la plus basse (lower), la plus haute (upper) et la valeur prédite (forecast). Nous voyons que les prévisions nous donnent au plus faible un nombre de nuités de 500-600 alors qu'au plus haut nous sommes à une prévision de 700-800 nuitées pour l'année 2023.

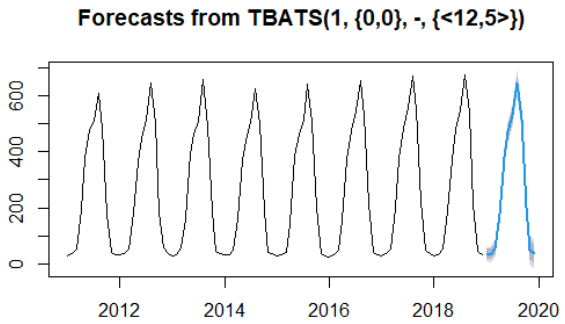
graphique 7 : prévision avec la méthode X-13-ARIMA-SEATS



Méthode TBATS

La seconde prévision que nous effectuons est avec la méthode TBATS, celle-ci nous donne une prévision sur un an. Le graphique correspondant à cette prévision est ci-dessous. Pour analyser ce graphique, nous allons dire que la valeur prédite du nombre de nuités est plus haute que celle de 2018, que le pic saisonnier se trouve logiquement sur les mois de printemps/été.

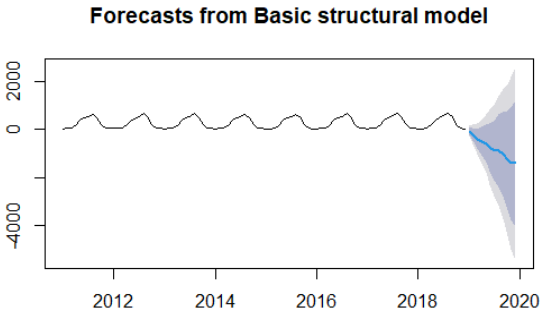
graphique 8 : prévision avec la méthode TBATS



Méthode STS

Nous voyons que la prévision faite est pessimiste, on peut se demander si cette fonction prend en compte la composante saisonnière et l'amplitude de prévision est très large, on parle de milliers de nuités. Dans ce graphique, nous voyons des chiffres négatifs, ce qui est impossible pour notre série. Nous pouvons donc conclure que cette méthode n'est pas la plus adaptée à notre série.

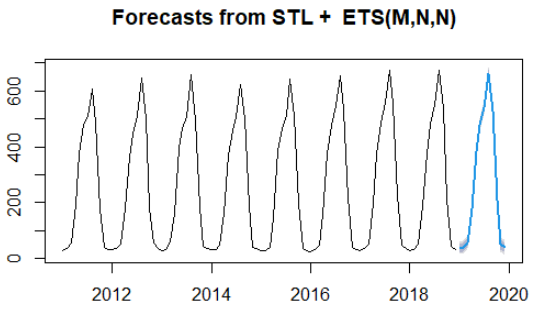
Graphique 9 : prévision avec la méthode STS



Méthode STL

La troisième méthode que nous utilisons est la méthode STL. Nous voyons que le graphique ressemble à celui de la méthode TBATS, avec une amplitude de prévision moins forte, il y a moins de flou sur le graphique. En ce qui concerne le nombre de nuités, la prévision prévoit un nombre de nuités d'environ 600 unités au maximum. Nous pouvons dire que les prévisions sont en adéquation avec le nombre de nuités maximum si on excepte la méthode STS. Nous pouvons conclure que cette méthode serait peut-être la plus appropriée à notre série.

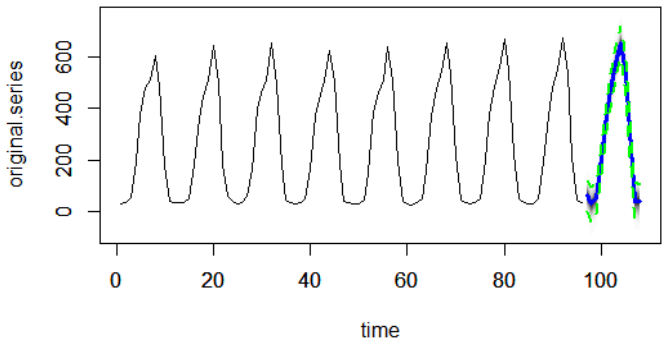
Graphique 10 : prévision avec la méthode STL



Méthode BSTS

La méthode que nous allons maintenant voir est la BSTS. Sur le graphique, nous voyons que les amplitudes sont nombreuses avec des petits écarts, donc une erreur de prévision normalement plus petites. Ce qui peut se traduire par une bonne prévision avec cette méthode. Concernant le nombre de nuités, la prédiction nous dit que ce nombre de nuités sera de maximum d'environ 700, une prévision donc plus optimiste que la méthode d'avant.

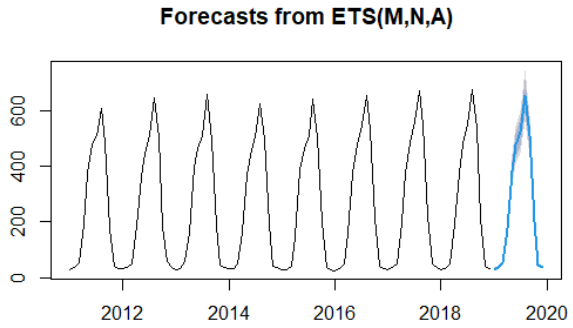
Graphique 11 : Prévision avec la méthode BSTS



Méthode ETS

La méthode ETS, nous donne un type de graphique similaire que les méthode BSTS ou STS ou TBATS, avec une amplitude de fluctuations qui est faible et un nombre de nuitées maximum de 700. La partie de la prévision qui semble la moins exacte se trouve au niveau du maximum de nuités, au pic. On trouve cette estimation pour pratiquement tous les graphiques.

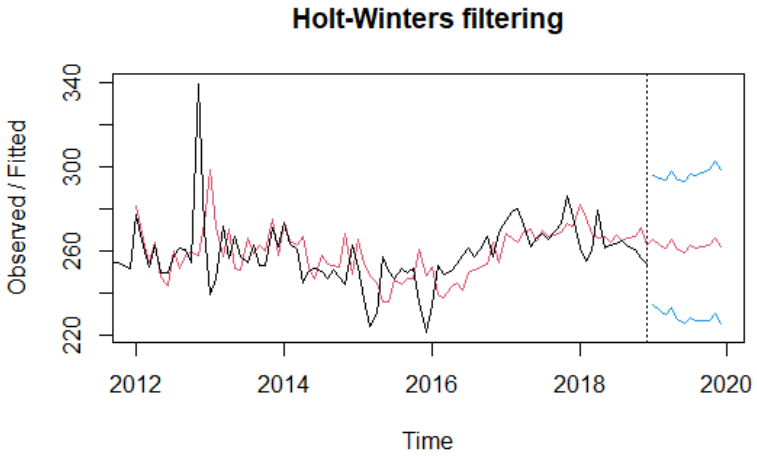
Graphique 13 : prévision avec la méthode ETS



Méthode Holt-Winters

La méthode Holt-Winters ne fonctionne pas de la même façon que les autres, dans cette méthode avant de faire la prévision, nous avons un lissage exponentiel double, ce lissage permet de créer un lissage et une constante. Comparativement au lissage exponentiel simple, ce lissage se fait avec 3 paramètres contre 2 pour le simple, ce qui lui permet une meilleure flexibilité. Une fois ce lissage fait, nous avons fait la prévision et nous trouvons le graphique ci-dessous. Nous allons parler en terme de moyenne de nuités par an et non d'un maximum de nuitées comme pour les autres prévisions. Cette méthode fait 3 prévisions, la plus haute, la plus basse et la moyenne. Sur le graphique, nous voyons deux courbes, une courbe rouge, et une courbe noire, la courbe noire correspond aux résidus et la courbe rouge correspond aux valeurs prédites. La prévision nous donne une valeur en augmentation avec une moyenne de 300 nuités sur l'année 2019. La prévision la plus basse nous donne une moyenne de 240 nuités sur 2019, avec une tendance à la baisse. Pour la valeur lissée, elle reste dans le prolongement avec une moyenne de 260 nuités sur l'année.

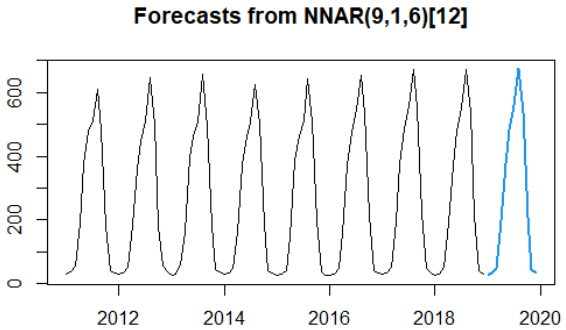
Graphique 14 : prévision avec la méthode Holt-Winters



NNETAR

Pour cette partie, nous utilisons la méthode NNETAR (Neural Network Time Series Forecasts), cette méthode va nous permettre de nous donner une prévision de la série. Le type de graphique est similaire aux graphiques de prévision, cette méthode semble néanmoins plus juste, il n'y a pas d'amplitudes, pas de floutage sur le graphique. Nous pouvons donc dire que cette méthode est efficace, cette efficacité provient du fait que cette méthode inclut … paramètres, ce qui la rend plus juste dans ses prévisions. Ce modèle prévoit un nombre de nuités à hauteur d'environ 700 au plus haut de la saison, l'estimation est dans la lignée des autres estimateurs.

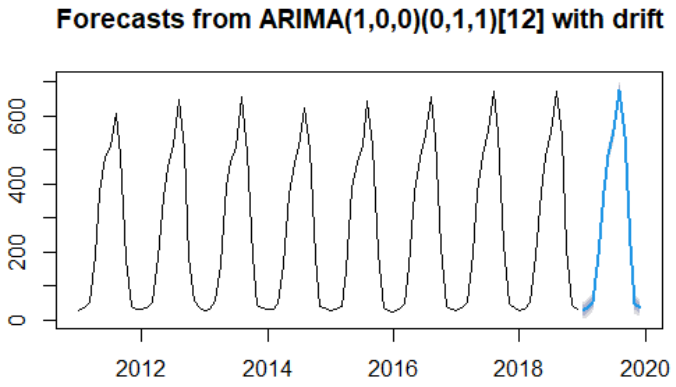
Graphique 15 : Prévision avec la méthode NNETAR



SARIMA(p,d,q)(P,D,Q)$_{12}$

Avec cette méthode, nous trouvons le modèle ARIMA le plus optimal pour notre série, ici, ARIMA (1,0,0)(0,1,1)$_{12}$. Ce qui signifie un modèle AR(1) pour la partie classique et MA(1) pour la partie saisonnière avec une différenciation. Concernant la prévision, celle-ci n'a aucune amplitude, très peu de floutage, ce qui signifie que le logiciel pense qu'il n'y aura pas d'erreurs de prévision. SI cela s'avère vrai, ce sera donc ce modèle qui aura les meilleures prévisions. Cette méthode prévoit un nombre de nuités d'environ 650-700 au plus grand pic. Les résultats de ce modèle se trouvent en annexe 9. Dans ces résultats, nous voyons qu'une variable se nomme "weekday", cette variable est un ajustement par rapport aux fêtes de Pâques.

Graphique 16 : prévision avec ARIMA (p,d,q)(P,D,Q)$_{12}$

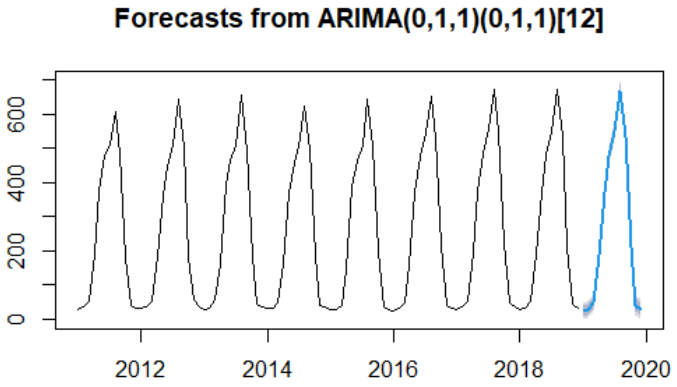


SARIMA(0,1,1)(0,1,1)$_{12}$

Le type de modèle à faire nous a été imposé pour cette méthode, graphiquement nous voyons pas de différences, les différences se trouvent dans les résultats. Les résultats de ce modèle se trouvent en annexe 10. L'analyse de ce graphique va donc être la même que celle au-dessus, soit que les fluctuations sont faibles et que le nombre de nuités prédits est d'environ 650-700. Les différence des deux modèles est le critère AIC, pour ce modèle il est de 662.67 alors que pour le

premier modèle il est de 667.63. Sachant qu'il faut maximiser le critère AIC, nous allons prendre celui qui est le plus fort, c'est donc le premier modèle qui semble être un meilleur modèle que celui là.

Graphique 17 : Prévision avec le modèle SARIMA$(0,1,1)(0,1,1)_{12}$



Les erreurs de prévision

Dans cette prochaine partie du dossier, nous allons voir les erreurs de prévision. Pour ce faire, nous allons effectuer un calcul de MSE sur nos différents modèles que nous allons comparer à la MSE d'une prévision naïve. La prévision naïve va se faire via la fonction "predict", cette prévision naïve sera logiquement une des prévisions les moins adéquates de toutes nos prévisions. Pour cette partie, nous avons ajouté deux modèles dits "naïfs", un modèle qui ne prend pas en compte la composante saisonnière (le modèle naive) et l'autre modèle qui la prend en compte (le modèle naiveS). Pour choisir notre meilleur modèle parmi tous ceux présentés ci-dessous, nous allons prendre celui donc la MSE est la plus faible. C'est donc le modèle naïf avec la composante saisonnière qui est le plus optimal pour notre variable.

Tableau 1 : tableau récapitulatif des MSE

Modèle	SARIMA $(0\ 1\ 1)(0\ 1\ 1)_{12}$	SARIMA$(p\ d\ q)(P\ D\ Q)_{12}$	Naive
MSE	51.71	61.34	108327.4
naiveS	NNETAR	Holt-Winters	ETS
42.958	83.2	53561.5	183.3
BSTS	STL	STS	TBATS
363.02	92.58	1382828	187.55
X-13			
57.07			

Test de précision

Nous allons maintenant voir les tests de précision à travers la fonction de Diebold-Mariano que nous allons appliquer sur la prévision naïve et entre nos modèles. A la suite des résultats de ce test, nous allons définir notre meilleur modèle parmi ceux testés. Les résultats de ce test se trouvent dans le tableau ci-dessous. Ce test va permettre de savoir si deux modèles sont équivalents en termes 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. Dans un premier tableau,nous allons avoir les résultats du test DM par rapport à la prévision naïve et dans le tableau suivant seront les résultats de comparaison de tous nos modèles. Une des méthodes que nous avons utilisé ne fonctionne pas en faisant le test de diebold Mariano, il s'agit de BSTS, nous avons donc décidé de la retirer, ne trouvant pas la solution au problème.

Tableau 2 : DM test entre les différents modèles et les prévisions naïves

Modèle	$SARIMA(0\ 1\ 1)(0\ 1\ 1)_{12}$	$SARIMA(p\ d\ q)(P\ D\ Q)_{12}$	NNETAR	HW
p-value	0.033	2.8e-05	2.5e-05	0.77
ETS	STL	STS	TABATS	X-13
6.16e-06	6.16e-06	1	0.189	10.16e-06

Nous avons mis l'alternative "less" à la fonction dm.test ce qui se traduit par une meilleure qualité de prévision du modèle 1 par rapport au modèle 2 (ici notre modèle naïf). Nous pouvons donc voir que 3 modèles ont une qualité de prévision égale ou inférieure à notre modèle naïf, ces 3 modèles sont : la méthode STS, Holt-Winters et TBATS.

Pour faire la comparaison entre nos modèles, nous avons décidé de prendre tous nos modèles, d'en retirer aucun. Chaque modèle va être testé dans les deux sens, par exemple le modèle X-13 va être testé en étant modèle 1 mais en étant également modèle 2. Le tableau ci-dessous nous montre les résultats de ces comparaisons. Nous avons décidé de faire un code couleur, lorsque la p-value du test sera inférieur à zéro, cela signifie que le modèle 1 aura une meilleure qualité de

prévision que le modèle 2 la case sera en rouge et lorsque la p-value sera supérieur à notre seuil de risque la case sera en verte, l'hypothèse h0 sera accepté. Au niveau horizontal ce sont les méthodes 1 et au niveau vertical ce sont les méthodes 2.

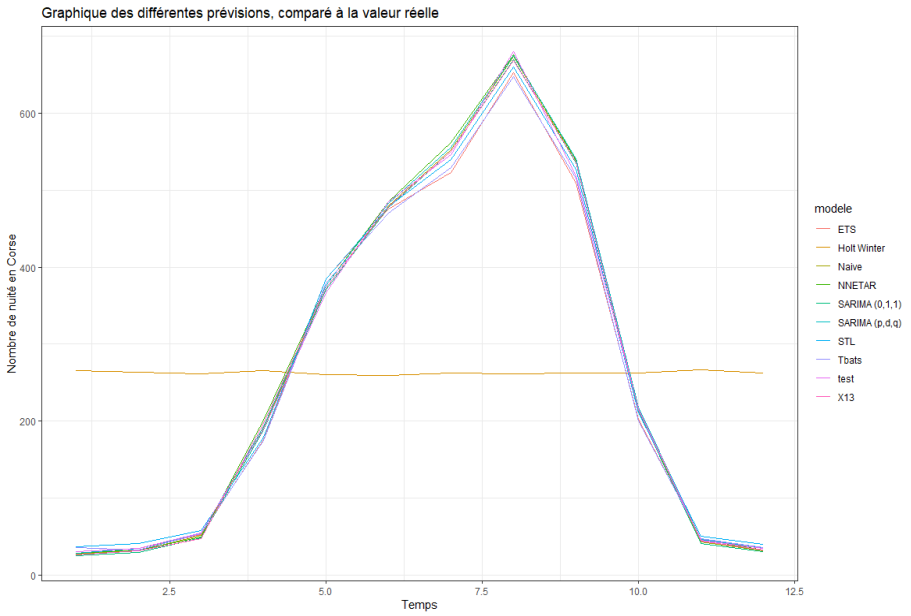
Tableau 3 : DM test entre tous nos modèles.

	SARIMA $(0\ 1\ 1)(0\ 1\ 1)_{12}$	SARIMA $(p\ d\ q)(P\ D\ Q)_{12}$	NNETAR	ETS	STL	X-13	HW	STS	TBATS
SARIMA $(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.7284	1.2e-4	3.7e-4	3.7e-4	3.7e-4	0.9035	1	0.5977
SARIMA $(p\ d\ q)(P\ D\ Q)_{12}$	0.2716		3.9e-4	7.6e-4	7.6e-4	7.6e-4	0.896	1	0.6544
NNETAR	0.9999	0.9999		2.2e-6	2.2e-16	2.2e-16	0.9972	1	1
ETS	0.9999	0.999	1		2.15e-11	0.0781	0.9972	1	1
STL	0.999	0.999	1	1		0.976	0.9972	1	1
X-13	0.9992	0.9996	1	0.9912	0.0234		0.9972	1	1
HW	0.0965	0.104	0.0028	0.0028	0.0028	0.0028		1	0.0713
STS	4.39e-15	6.35e-15	2.2e-16	9.46e-15	9.46e-15	9.46e-15	2.2e-16		9.29e-15
TBATS	0.4023	0.3456	2.2e-16	2.2e-16	2.2e-16	2.2e-16	0.9287	1	

Pour analyser ce tableau, le code couleur nous permet de dire que la meilleure méthode de prévision est la méthode STS, nous le voyons du fait que l'hypothèse est refusée par rapport à n'importe quelle autre méthode, c'est donc la méthode avec la meilleure qualité. La méthode ayant la qualité la plus mauvaise, STS, avec des p-value toutes supérieur à 0.05, ce qui signifie que par rapport aux autres modèles, cette méthode à une qualité moins bonne. Nous voyons ces deux résultats en lisant le tableau de façon verticale. Donc si nous prenons STS (méthode 1) avec ETS (méthode 2), l'hypothèse H0 est refusée, donc le modèle 1 et 2 n'ont pas des qualités de prévision égales, le modèle 1 a donc une meilleure qualité de prévision.
D'après ce tableau, la meilleure méthode de prévision est donc la méthode STS.

Nous allons terminer cette première partie du dossier par un graphique permettant de montrer les différentes prévisions de nos différentes méthodes.En faisant le graphique, nous avons trouvé que la méthode STS était en négatif, nous avons donc décidé de l'enlever du graphique pour que sa compréhension soit plus facile. Sur ce graphique, nous voyons que toutes nos méthodes excepté HW ont les mêmes fluctuations, elles se confondent. La courbe correspondant à la méthode HW est rectiligne, du fait que nous avons effectué un lissage avec cette méthode.

Graphique 18 : Graphique récapitulatif de nos différentes méthodes de prévisions



Etude d'une série non saisonnière : le Brent

Analyse préliminaire

Dans cette partie du dossier, nous allons réaliser l'étude de notre série non saisonnière, cette série concerne les prix du BRENT, un pétrole brute servant de référence dans les prix du pétrole au niveau mondial, pétrole foré en mer du nord. C'est un pétrole dit léger donc les coûts de raffinement sont faibles. Les données de cette série ont été prélevées sur le site de l'eia (Energy Information Administration)[2], agence de statistique au sein du ministère de l'énergie des États-Unis, nos données s'étendent de août 2004 à décembre 2019. Pour expliquer cette série, nous avons décidé de prendre 5 variables explicatives, le stock [3], la production [4], l'activité mondiale [5], le MSI (Market Standard Indicator, l'indice des pays économiquement développé)[6] et notre dernière variable est Baltic Dry Index (l'indice des prix du transport maritime)[7]. Ces cinq variables explicatives vont nous permettre de prédire les prix du pétrole, nous allons donc effectuer différents tests et essayer de trouver la meilleure prédiction.

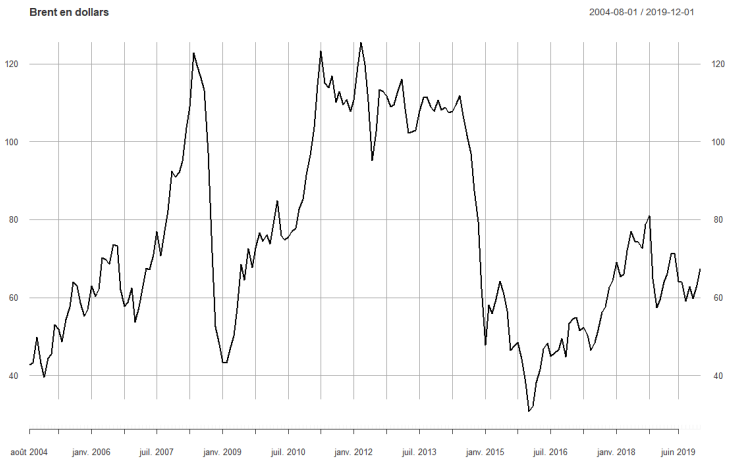
Dans cette analyse, nous commencerons par une analyse préliminaire de notre série, en retirant les points atypiques ou en vérifiant la stationnarité de nos variables. Nous poursuivrons par un estimation des modèles linéaires via des modèles de types AR(1), AR(p), ARIMA (p,d,q) et pour finir avec la méthode Holt-Winters. Une fois cette estimation faite, nous verrons une prédiction de notre série sur un an avec un pas de un mois avec les mêmes méthodes que pour l'estimation. Nous injectons ensuite nos variables explicatives pour effectuer une seconde phase de prévision. Puis nous finirons cette partie en testant les erreurs de prévision avec la MSE et en faisant le test de Diebold Mariano pour les tests de précision.

---

[2] https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RBRTE&f=M
[3] les données pour les stocks nous ont été données par Z. MOUSSA
[4] https://www.eia.gov/opendata/qb.php?category=2134979&sdid=INTL.57-1-WORL-TBPD.M
[5] https://www.dallasfed.org/research/igrea
[6] https://fr.investing.com/indices/msci-world-historical-data
[7] https://fr.investing.com/indices/baltic-dry-historical-data

Par définition, une série non-saisonnière est une série où il n'existe pas de saisonnalité, de comportement récurrent qui puisse masquer l'information contenu dans les variables. Nous avons testé la saisonnalité grâce à la fonction "isSeasonal" sous R, le résultat de ce test se trouve en annexe 11, ils nous dit que notre série est bien non-saisonnière.

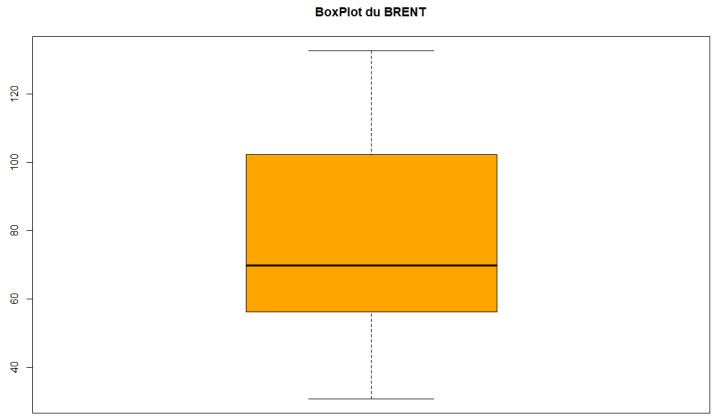
Graphique 19 : Prix du Brent



Commençons dès à présent à regarder les caractéristiques de notre série grâce au graphique 19, qui nous montre les changements de prix du BRENT. Avec ce graphique, nous voyons qu'il n'y a pas de saisonnalité. Sur ce graphique, nous voyons 3 baisses importantes du prix du pétrole, le premier se situe fin 2008 début 2009, c'est la cause de la crise des subprimes, avec la chute de Lehman Brothers et la propagation de la dette des Etats-Unis à l'Europe. La seconde baisse se situe entre janvier 2012 et septembre 2012, cette baisse survient suite à la crise de la dette souveraine, crise ayant touché l'europe entière suite à des pression de la BCE sur le gouvernement grecque, du fait que que le Grèce ne pouvais et ne voulais pas rembourser sa dette avec des taux d'intérêt trop élevé, les prises de positions sur la chute de la Grèce n'ont qu'accentuer son déclin. La troisième baisse importante se produit entre septembre 2014 et janvier 2015, cette baisse est dû à un ralentissement économique mondiale suivi d'une offre abondante de pétrole et les débuts de la production du pétrole de schiste aux Etats-Unis.

Valeurs atypiques

Graphique 20 : boxplot de la variable BRENT



Nous allons maintenant voir les points atypiques de nos variables, nous allons les commenter juste pour notre variables à expliquer. Sur le boxplot ci-dessus, nous ne voyons aucune valeur atypiques néanmoins avec la fonction "tsoutliers" nous en avons repéré. Nous trouvons 3 points atypiques, les mois de juin, juillet, décembre 2008, ce qui correspond à la chute vertigineuse des prix du pétrole au niveau mondial.

Stationnarité

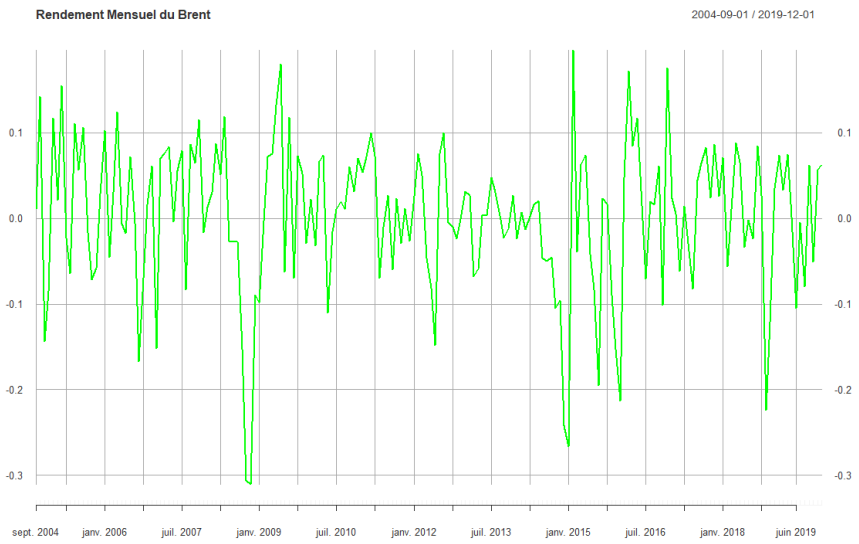
Avant de tester la stationnarité de nos variables, nous avons mis en logarithme népérien nos 3 variables qui concerne des produits financiers, pour après avoir fait une première différenciation, pouvoir avoir le rendement. Nous allons tester la stationnarité avec le test augmentée de Dickey-Fuller, pour ce test, H0 définit l'hypothèse que la variable est non-stationnaire, il faut donc que la p-value soit inférieure à 0.05 pour refuser le test, donc que la variable soit stationnaire. Les résultats de ce test se trouvent dans le tableau ci-dessous, nous voyons que nous n'avons aucune variable stationnaire, il faudra donc faire une première différenciation. Le résultat avec une première différenciation se trouve sur la troisième ligne du tableau, nous voyons qu'une fois cette différenciation faite, toutes nos variables sont stationnaires.

Tableau 4 : Test de Dickey-Fuller sur nos variables

	BRENT	STOCK	PROD	ACT_MOND	MSCI Index	Baltic dry Index
p.value ADF Test sur série brute	0.35	0.50	0.25	0.36	0.59	0.30
p.value ADF Test sur série différenciée	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

Le graphique ci-dessous nous montre la distribution de notre variable à expliquer ayant une première différenciation, une fois que celle-ci est stationnaire. Nous voyons la différence avec le graphique 19 qui nous montrait les variations des prix du pétrole, sur ce graphique nous voyons des fluctuations tournent autour de la moyenne, ces fluctuations reviennent toujours à la moyenne, définitions de la stationnarité d'une variable.

Graphique 21 : Rendement du Brent



Statistiques descriptives

Après avoir stationnarité nos variables, nous allons passer aux statistiques descriptives de celle-ci. Nous utilisons “basicstats” de la library “fbasics”, ce qui nous donne toutes les statistiques descriptives pour chacune de nos variables. Nous avons donc 184 observations sans aucune valeur manquante. Nos variables ayant très peu de fluctuations et étant également stationnaires, leurs moyennes varient autour de zéro. Au niveau du coefficient de Kurtosis , celui-ci est positif pour toutes nos variables excepté le stock, ce qui définit que leur définition sera leptokurtique alors que celle du stock sera platykurtique. Pour le coefficient de skewness, toutes nos variables ont un coefficient proche de zéro, ce qui signifie que leur distribution est symétrique.

Tableau 5 : Statistiques Descriptives de l’ensemble des données

	BRENT	STOCK	PROD	ACT_MOND	MSCI_index	Baltic_dry_index
nobs	184.00	184.00	184.00	184.00	184.00	184.00
NAs	0.00	0.00	0.00	0.00	0.00	0.00
Minimum	-0.31	-57.65	-1676.51	-58.77	-0.13	-0.75
Maximum	0.20	63.78	1864.27	69.46	0.10	0.67
1. Quartile	-0.05	-21.41	-289.98	-14.24	-0.02	-0.12
3. Quartile	0.06	19.70	432.94	12.56	0.03	0.12
Mean	0.00	0.77	60.22	-0.22	0.01	0.00
Median	0.01	0.12	85.95	0.54	0.01	0.00
Sum	0.45	141.90	11080.24	-40.24	0.94	0.37
SE Mean	0.01	1.98	39.92	1.50	0.00	0.02
LCL Mean	-0.01	-3.13	-18.54	-3.18	0.00	-0.03
UCL Mean	0.02	4.67	138.97	2.75	0.01	0.03
Variance	0.01	720.10	293158.13	415.25	0.00	0.05
Stdev	0.09	26.83	541.44	20.38	0.04	0.22
Skewness	-0.88	0.20	-0.09	0.03	-0.61	-0.30
Kurtosis	1.51	-0.69	0.29	0.35	0.97	1.33

Estimations des modèles linéaires et leur résidus

Dans cette partie, nous allons faire plusieurs estimations à travers différents modèles. Nous allons utiliser un modèle AR(1), un modèle AR(p), un modèle ARIMA(p,d,q) et un modèle Holt-Winters.

AR(1) : nous voyons grâce à la fonction “coefetst” de la library “lmtest” que notre phi1 est significatif au seuil de 1%, nous analysons que la condition de stationnarité est respectée, elle est de 0.271, elle donc inférieur strictement à 1. Le critère d’Akaike nous indiquant la vraisemblance du modèle est de -386.24 et le log de vraisemblance est de 196.12, ces deux critères nous serviront pour la comparaison de nos modèles.

Figure 1 : Modèle AR(1) pour la série différenciée du Brent

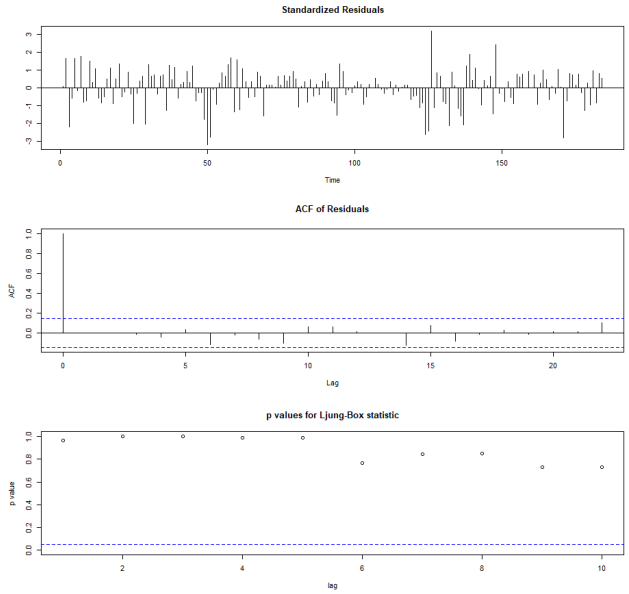
```
  Series: data_diff$BRENT
  ARIMA(1,0,0) with non-zero mean
  Box Cox transformation: lambda= 1

  Coefficients:
           ar1     mean
        0.2712  -0.9974
  s.e.  0.0708   0.0084

  sigma^2 estimated as 0.007019:  log likelihood=196.12
  AIC=-386.24   AICc=-386.11   BIC=-376.59

z test of coefficients:

            Estimate Std. Error   z value  Pr(>|z|)
ar1        0.2711517  0.0707902    3.8304  0.000128 ***
intercept -0.9973933  0.0084115 -118.5748 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nous passons maintenant à l’étape de vérification de notre modèle AR(1), nous allons donc regarder l’indépendance des résidus, l’interprétation du modèle peut se faire qu si les résidus suivent le procesus de bruit blanc, les résidus doivent donc avoir une indépendance du temps et une espérance nulle. Nous allons donc vérifier l’indépendance des résidus puis leur normalité.

- Indépendance des résidus

Pour s’assurer que nos résidus suivent un procesus de bruit blanc, nous avons ultilisé le test qui se refère à la statistique Q. Nous obtenons une p-value de 0.9627, ce qui est supérieur à 0.05, on accepte l’hypothèse nulle d'absence d'auto corrélations de nos résidus, nos résidus ne sont donc pas autocorrélés. Graphiquement, sur le corrélogramme, nous constatons que le premier retard est significatif mais que tous les autres sont proches de zéro et non significatifs. C'est également ce premier retard qui contient le plus d'informations.

Graphique 22 : Indépendance des résidus



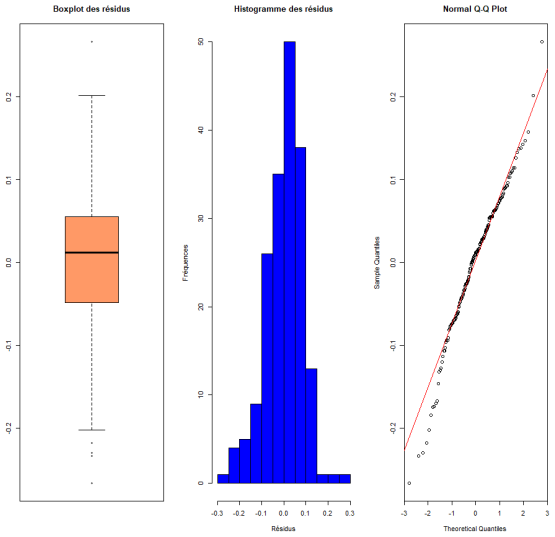

```
        Box-Ljung test

data:  resAR1
X-squared = 0.002185, df = 1, p-value = 0.9627
```

- Normalité des résidus

Grâce au graphique ci-dessous, nous pouvons voir que la distribution de nos résidus semble normale malgré quelques valeurs qui pourraient altérer la distribution de nos résidus. Toutes ces interrogations que nous émettons avec les graphiques vont être vérifiées grâce au test de Kolmogorov-Smirnov. Ce test nous donne une p-value de 0.3448, l'hypothèse H0 sera donc acceptée au seuil de 5%. Dans ce test, l’hypothèse H0 nous dit que les résidus suivent une loi normale. Dans ce modèle au

seuil de risque de 5%, les résidus suivent une loi normale, ce modèle est donc validé, nous pourrons donc l'interpréter.

Graphique 23 : normalité de nos résidus




```
        One-sample Kolmogorov-Smirnov test

data:  resAR1
D = 0.069016, p-value = 0.3448
alternative hypothesis: two-sided
```

AR(p) : Pour ce modèle, nous avons décidé de faire un modèle AR(2), pour notamment voir les différences qu'il peut exister avec le premier modèle. Nous voyons que le phi1 est significatif à 1% mais que le phi 2 n'est pas significatif. En ce qui concerne les critères, celui de vraisemblance est de -384.25, il est donc supérieur au modèle Ar(1) et celui du log est de 196.12, il est égal à celui du modèle au-dessus. Ce sont deux modèles qui restent semblables, si nous comparons le critère de vraisemblance, le modèle 1 semble être le meilleur.

Figure 2 : Modèle AR(p) pour la série différenciée Brent

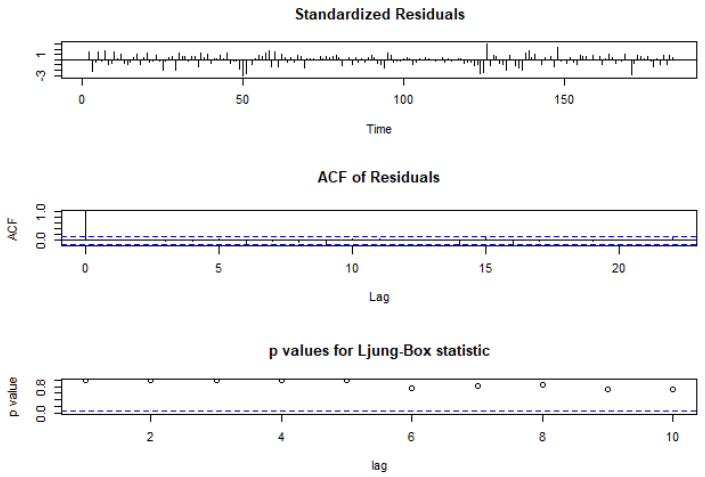
```
Series: data_diff$BRENT
ARIMA(2,0,0) with non-zero mean
Box Cox transformation: lambda= 1

Coefficients:
         ar1      ar2     mean
      0.2731  -0.0071  -0.9974
s.e.  0.0736   0.0740   0.0084

sigma^2 estimated as 0.007058:  log likelihood=196.12
AIC=-384.25   AICc=-384.03   BIC=-371.39
> coeftest(out1)

z test of coefficients:

            Estimate Std. Error   z value  Pr(>|z|)
ar1        0.2730866  0.0735743    3.7117 0.0002059 ***
ar2       -0.0071377  0.0740133   -0.0964 0.9231732
intercept -0.9974075  0.0083538 -119.3962 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comme pour le modèle AR(1), nous allons analyser les résidus de ce modèle, savoir si celui-ci est interprétable. Nous allons donc regarder l’autocorrélation des résidus et leur normalité.

- Indépendance des résidus

Le test utilisé est le même que pour le modèle AR(1), nous voyons que grapiquement les premiers retards sont proches de zéro, les résidus semblent donc suivent le procesus de bruit blanc. La p-value du test de Box-Ljung qui est supérieur à 0.05 nous conforte dans l'idée puisque l’hypothèse H0 est validée. Il y a donc une absence d’autocorrélations de nos résidus.

Graphique 24 : indépendance des résidus

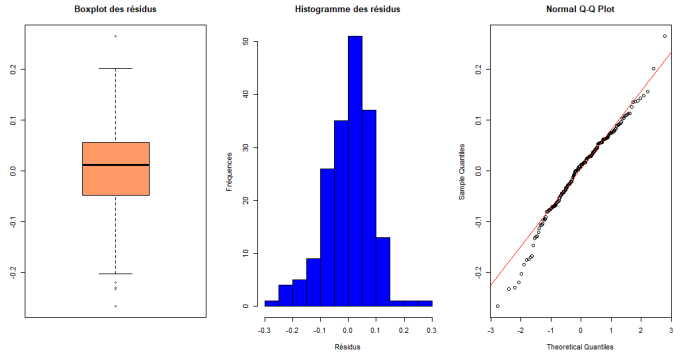


```
        Box-Ljung test

data:  resARP
X-squared = 0.0003228, df = 1, p-value = 0.9857
```

- Normalité des résidus

Si nous faisons une analyse graphique, avec l'histogramme, nous voyons que nos résidus semblent suivre une loi normale. En l'occurrence, sur le boxplot, nous voyons qu'il y a quelques valeurs atypiques, ce qui pourrait altérer la distribution de nos résidus, ce qui rejoint le QQ plot, ou certaines valeurs aux extrémités s'éloignent de la droite de régression. Néanmoins, le test de Kolmogorov-Smirnov nous donne une p-value supérieur à 0.05, ce qui signifie que nos résidus suivent une loi normale.

Graphique 25 : Normalité des résidus




```
        One-sample Kolmogorov-Smirnov test

data:  resARP
D = 0.073057, p-value = 0.2798
alternative hypothesis: two-sided
```

ARIMA (p,d,q) : tout d'abord, nous voyons que le logiciel nous propose un modèle MA(1) qui pour lui semble le plus pertinent. De ce modèle, nous voyons que le paramètre ma1 est significatif au seuil de risque de 1%. La condition de stationnarité est respectée, elle est inférieure à 1 (0,277). Le critère de vraisemblance est de -387,3 et le log de vraisemblance est de 195.65. Nous voyons donc que le critère d'Akaike est inférieur aux deux premiers modèles mais le log de vraisemblance est quant à lui supérieur. Ensuite, nous allons analyser les résidus de ce modèle en commençant par l'indépendance des résidus puis la normalité des résidus.

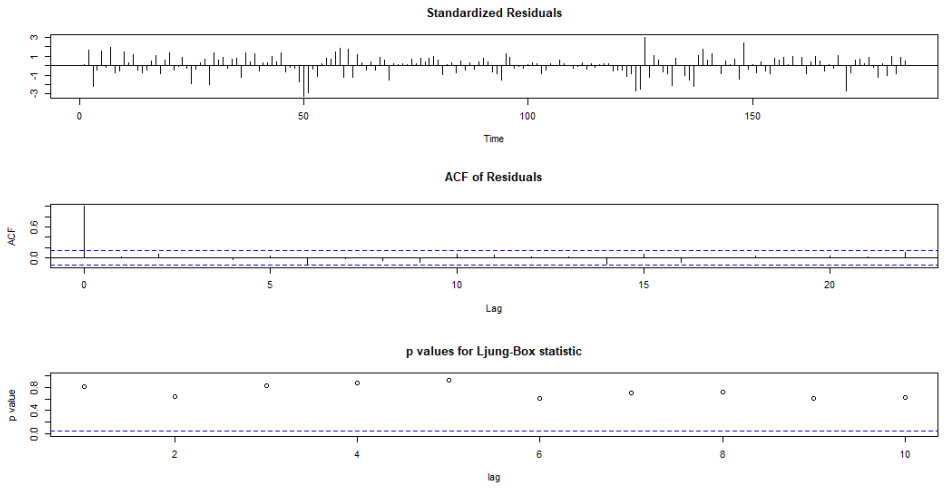
Figure 3 : Modèle ARIMA(p,d,q) pour la série différenciée Brent

```
Series: data_diff$BRENT
ARIMA(0,0,1) with zero mean

Coefficients:
         ma1
      0.2576
s.e.  0.0678

sigma^2 estimated as 0.007017:  log likelihood=195.65
AIC=-387.3   AICc=-387.23   BIC=-380.87
> coeftest(out3)

z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ma1 0.257611   0.067764  3.8016 0.0001438 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Indépendance des résidus

Comme pour les modèles précédents, nous allons vérifier les autocorrélations des résidus. Graphiquement, nous voyons que ceux-ci semblent suivrent le procesus de bruit blanc, avec des premiers retards proches de zéro sur le corrélogramme. Si nous passons maintenant au test de Box-Ljung, la p-value est de 0.8152, ce qui signifie que l’hypothèse H0 est acceptée. Nous allons dire qu'il y a une absence d'auto corrélations des résidus.

Graphique 25 : indépendance des résidus

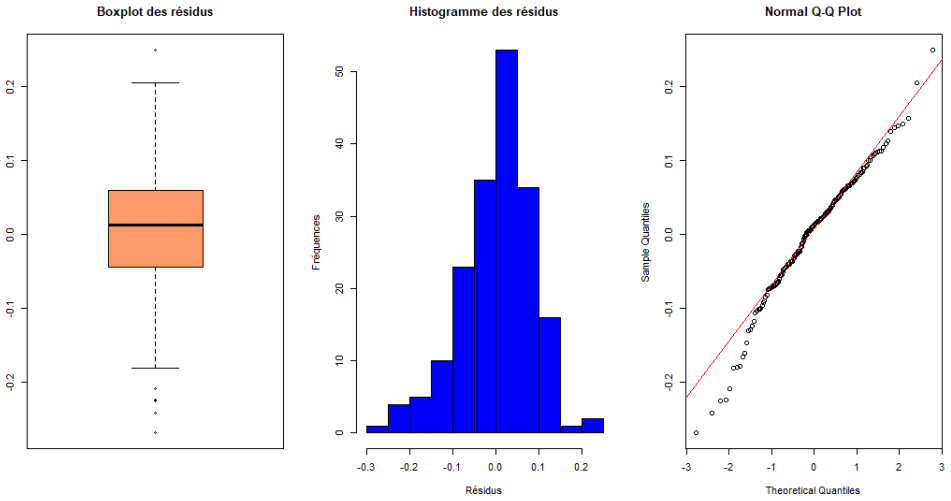



```
        Box-Ljung test

data:  resARIMA
X-squared = 0.054651, df = 1, p-value = 0.8152
```

- Normalité des résidus

Nous allons maintenant regarder si nos résidus ont une distribution normale, graphiquement, l'histogramme ressemble à la définition d'une distribution normale, une courbe en cloche, même si dans notre cas, celle-ci semble être décalée vers la droite. De plus, comme pour le modèle AR(2), les points atypiques que nous trouvons sur le boxplot peuvent altérer la distribution de nos résidus. A la lecture des résultats du test de Kolmogorov-Smirnov, la p-value est supérieure à zéro, ce qui signifie que l'hypothèse H0 va donc être acceptée. Nous allons donc conclure pour ce modèle que les résidus de celui-ci suivent une distribution normale au seuil de risque de 5%

Graphique 26 : Normalité des résidus




```
        One-sample Kolmogorov-Smirnov test

data:  resARIMA
D = 0.077245, p-value = 0.2222
alternative hypothesis: two-sided
```

Holt-Winters : Au niveau de cette méthode, nous avons décidé de mettre les résultats en annexe (annexe 12 et 13) , pour éviter de surcharger le dossier, qu'il soit moins lourd. Les méthodes ne changent pas, nous avons effectué les mêmes tests que pour les autres modèles. Ce qui en ressort, les coefficients de lissage alpha et bêta, ont été obtenus en précisant dans la fonction que nous avions pas de partie saisonnière (gamma=FALSE), ce qui nous donne un alpha = 0.573 et bêta = 0.139. En ce qui concerne les résidus de cette méthode, le test de Box-Ljung nous donne des résidus avec une absence d'autocorrélations, le test de Kolmogorov-Smirnov nous donne des résidus qui suivent une distribution normale.

Figure 4 : Modèle Holt-Winters pour la série différenciée Brent

```
Call:
HoltWinters(x = data_diff$BRENT, gamma = FALSE)

Smoothing parameters:
 alpha: 0.5733365
 beta : 0.1391751
 gamma: FALSE

Coefficients:
          [,1]
a 0.046972680
b 0.007449784
```

Pour résumer nos différents modèles, dans tous nos modèles, nous avons des résidus avec une absence d'autocorrélations et nous avons également des résidus qui suivent une distribution de loi normale. Les modèles sont assez proches les uns des autres, aucun de nos modèles ne se détache, pour les départager, nous verrons sur la partie suivante concernant les prévisions que nous allons faire sur l'année 2019.

Prévision linéaire sur un an avec un pas de un mois

Dans cette partie du dossier, nous allons faire des prévisions sur an, nous allons donc le faire sur l'année 2019 avec un pas de un mois. Nous avons donc au préalable retirer l'année 2019 de nos données pour faire nos prévisions. Pour faire nos prévisions, nous avons décidé de faire une boucle, ce qui va permettre de créer un nouveau modèle à chaque prévision, ce qui rendra les prévisions plus précises. Au niveau des prédictions, nous voyons que ARIMA nous donne des prédictions plus faibles que les autres prédictions. En comparant avec les valeurs réelles, nous voyons que les prédictions sont proches de ces valeurs, cela signifie que les modèles que nous avons utilisés font de bonnes prédictions, c'est la méthode Holt-Winters qui semble avoir la prévision la plus éloignée de la réalité, de plus cette méthode nous donne que des valeurs négatives.

Tableau 6 : prédiction avec nos différents modèles

	AR(1)	AR(p)	ARIMA (p,d,q)	Holt-Winters	Valeurs réelles
Janvier 2019	-0.03350	-0.03312	-3.33810e-02	-0.14564	0.03511
Février 2019	-0.00851	-0.00841	-1.08114e-02	-0.16466	0.07379
Mars 2019	-0.00139	-0.00137	-2.61586e-03	-0.18368	0.03351
Avril 2019	0.00063	0.00063	7.30448e-04	-0.20269	0.07414
Mai 2019	0.00121	0.00120	9.04786e-04	-0.22171	0.00126
Juin 2019	0.00138	0.00136	-6.52906e-04	-0.24073	-0.010486
Juillet 2019	0.00142	0.00141	5.56392e-05	-0.25975	-0.00468
Août 2019	0.001446	0.00142	4.11557e-04	-0.27877	-0.07941
Septembre 2019	0.001446	0.00143	1.45640e-03	-0.29779	0.06221
Octobre 2019	0.001447	0.00143	4.63826e-04	-0.31680	-0.05093
Novembre 2019	0.001448	0.00143	-3.89350e-03	-0.33582	0.05570
Décembre 2019	0.001448	0.00143	-2.10429e-03	-0.35484	0.06285

Prévision avec variables explicatives

Dans cette partie du dossier, nous allons compléter l'analyse de notre Y_i en ajoutant des variables explicatives, elles sont aux nombre de 5. Il s'agit de la production de pétrole, des stocks de pétrole, de l'activité mondiale, de l'indice des prix du transport maritime (baltic dry index) et de l'indice des pays économiquement développés (MSI index). Nous allons donc modéliser notre variable à expliquer à l'aide de nos variables explicatives, nous ferons ensuite une prévision à l 'aide de ces variables explicatives. Avant de commencer sur la modélisation, il est important de rappeler que toutes nos variables sont stationnaire, elles ont toutes eu une première différenciation pour l'être.

Modélisation et prévoir la série avec des variables explicatives

Pour faire la modélisation, nous allons utiliser 4 méthode différentes, dont trois basées sur la régression linéaire et l'autre basée sur la méthode AR avec prise en compte de la valeur précédente. Les trois méthodes utilisant la régression linéaire sont la méthode "leaps", la méthode "mass" et la méthode "glmulti", l'autre méthode prenant en compte les valeurs précédentes est la méthode "arx".

La première méthode que nous allons utiliser est la méthode leaps, elle nous un meilleur modèle avec deux variables explicatives qui sont l'activité mondiale et le MSCI index. Sur tous les critères, ce modèle à deux variables explictaives est le meilleur.

Tableau 7 : résultat de la modélisation leaps

	Adj.R2	CP	BIC
1	2	2	2

La seconde méthode que nous allons utiliser est la méthode "glmuli", cette méthode utilise également java. Elle nous donne le même résultat que pour la méthode précédente, soit un modèle à deux variables explicatives qui sont les mêmes qu'avec la méthode leaps. Cela nous confirme qu'avec une estimation linéaire, le meilleur modèle est donc avec l'activité mondiale et le MSCI index.

Tableau 8 : résultat de la modélisation glmulti

```
$bestmodel
[1] "BRENT ~ 1 + ACT_MOND + MSCI_index"
```

La dernière méthode utilisant la régression linéaire est la méthode “mass”, celle-ci nous donne les mêmes résultats que les deux autres méthodes, avec deux variables explicatives significatives qui sont le MSCI index et l’activité mondiale. En annexe (14), nous pourrons retrouver les résultats avec toutes les variables explicatives, entre les différentes méthodes, la différence se fait avec le critère AIC. Les variables production stock et baltic-dry-index augmentent le critère d’AIC alors que les variables activité mondiale et MSCI index le diminue, c’est donc la raison pour laquelle nous allons prendre le modèle qu’avec ces deux variables.

Tableau 9 résultat de la modélisation mass

```
Start:  AIC=-926.92
BRENT ~ ACT_MOND + MSCI_index

             Df Sum of Sq    RSS     AIC
<none>                    1.1557 -926.92
- ACT_MOND    1  0.029895 1.1856 -924.22
- MSCI_index  1  0.173179 1.3289 -903.23
```

Après avoir fait la modélisation avec des méthodes d'estimation linéaires, nous allons maintenant passer sur une méthode les principes d’AR, soit de prendre en compte la valeur précédente. Nous allons faire un modèle avec la correction de white qui va nous permettre de retirer l'endogénéité entre nos variables. Les résultats de cette modélisation nous donne la variable explicative MSCI index est significative à 5% tout comme la composante AR(1). La variable activité mondiale qui était présente pour les modèles linéaires n’est pas significative ici.

Tableau 10 : résultats modélisation ARX

```
      reg.no. keep    coef std.error t-stat   p-value
ar1         1    0 0.22991   0.10065 2.2844   0.02359 *
mxreg       2    0 0.68812   0.16406 4.1943 4.409e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Prévision de la série

Nous allons maintenant passer à la prévision de nos modèles, deux modèles, un modèle linéaire avec deux variables explicatives qui sont l'activité mondiale et le MSCI index et un modèle ARX avec une seule variable explicative, le MSCI index. Au niveau des résidus, nous avons décidé de prendre la médiane de ceux-ci. Les prévisions vont être faites sur R en réalisant une boucle "for". Au niveau des prévisions, nous voyons que celles-ci sont assez proches de la réalité malgré quelques petites différences, c'est en faisant les erreurs de prévision et les tests de précision que nous pourrons définir le meilleur modèle parmi tous ceux que nous avons testé.

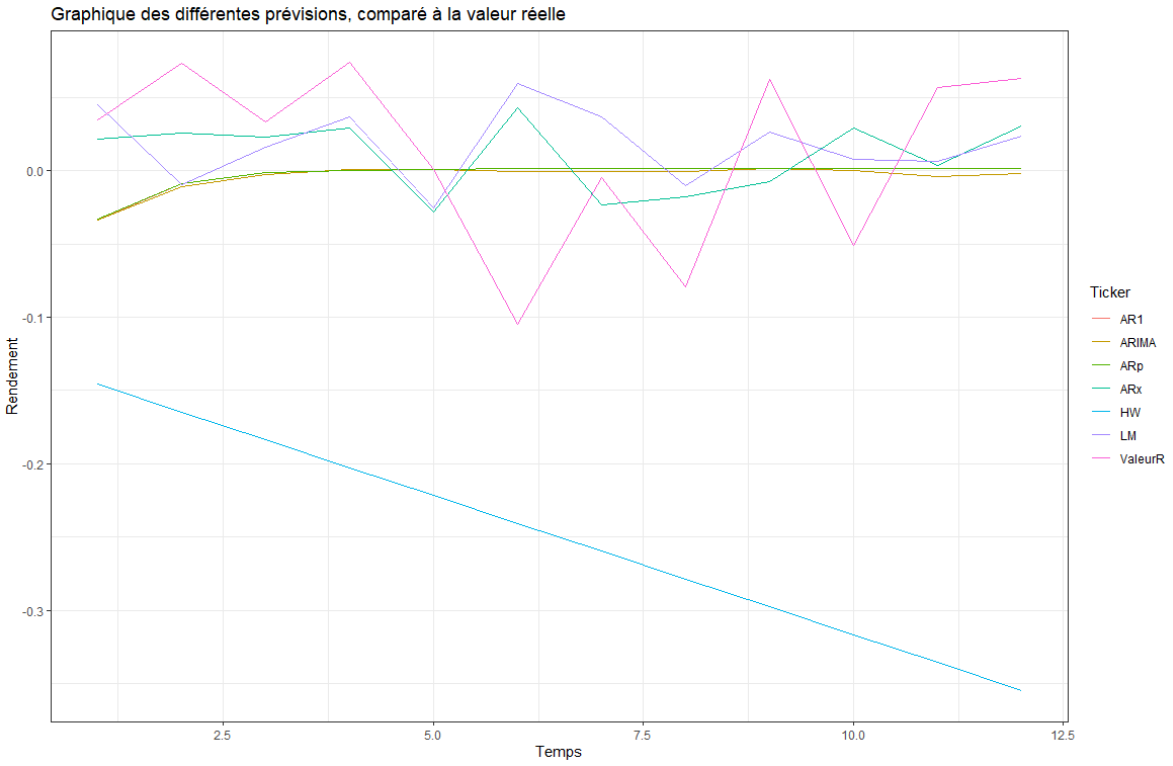
Tableau 11 ; résultat des prévisions avec les variables explicatives

	Modèle linéaire	Modèle ARX	Valeurs réelles
Janvier 2019	0.04549	0.02159	0.03511
Février 2019	-0.00957	0.02578	0.07379
Mars 2019	0.01632	0.02284	0.03351
Avril 2019	0.0370	0.02954	0.07414
Mai 2019	-0.02561	-0.02819	0.00126
Juin 2019	0.05981	0.04297	-0.010486
Juillet 2019	0.0371	-0.02347	-0.00468
Août 2019	0.0102	-0.01769	-0.07941
Septembre 2019	0.02662	-0.00751	0.06221
Octobre 2019	0.00810	0.02917	-0.05093
Novembre 2019	0.00679	0.00403	0.05570
Décembre 2019	0.02357	0.03038	0.06285

Représentation graphique

Nous allons maintenant passer à la représentation graphique de nos modèles, graphique avec nos valeurs réelles ainsi que six méthodes de prévision pour l'année 2019. Graphiquement, nous allons avoir une idée de quelle sera la meilleure méthode de prévision lorsque nous utiliserons des variables explicatives. Nous voyons donc que la méthode LM (régression linéaire) est la meilleure méthode, c'est cette méthode qui est la plus proche de la droite des valeurs réelles. A contrario, la moins bonne méthode que nous avons utilisée est la droite qui s'éloigne le plus de notre droite des valeurs réelles, c'est la méthode Holt-Winters. Pour confirmer notre choix du meilleur modèle, nous allons par la suite faire les erreurs de prévision et les tests de précision.

Graphique 27 : représentation de tous nos modèles de prévision



Les erreurs de prévision

Comme pour la partie avec notre série saisonnière, dans cette partie, nous allons calculer la MSE entre tous nos modèles et entre une prévision naïve, nous pourrons alors décider du meilleur modèle. En regardant les résultats dans le tableau ci-dessous, nous voyons que la plus petite valeur de la MSE correspond au modèle ARX qui semble être notre meilleur modèle en général. Si nous ne prenons que les modèles linéaires, nous voyons que notre meilleur modèle correspond au modèle AR(p) qui est AR(2). Comme nous l'avons vu graphiquement le modèle Holt-Winters est notre moins bon modèle il est moins bon que la prédiction naïve.

Tableau 12 valeurs des MSE de tous nos modèles

Modèle	AR(1)	AR(p)	ARIMA	HW	LM	ARX	naïve
MSE	0.00410479	0.004098763	0.004160267	0.07628192	0.004326368	0.003884099	0.02160747

Les tests de précision

Pour les tests de précision, nous allons utiliser le test de Diebold-Mariano, que nous allons comparer entre un modèle naïf et entre tous nos modèles. Les résultats de ce test permettront de choisir le modèle le plus adapté à notre série. Les résultats de ce test se trouvent dans le tableau ci-dessous. Ce test va permettre de savoir si deux modèles sont équivalents en termes 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. Dans un premier tableau, nous allons avoir les résultats du test DM par rapport à la prévision naïve et dans le tableau suivant seront les résultats de comparaison de tous nos modèles.

Tableau 13 : DM test entre la prévision naïve et nos modèles

Modèle	AR(1)	AR(p)	ARIMA	HW	LM	ARX
p-value	0.0015	8.6e-06	0.0015	0.3616	8.8e-06	1.32e-05

Nous avons mis l'alternative "less" à la fonction dm.test ce qui se traduit par une meilleure qualité de prévision du modèle 1 par rapport au modèle 2 (ici notre modèle naïf). Il faut donc que la p-value soit inférieur à 0.05 pour refuser l'hypothèse H0, pour qu'un modèle soit meilleur que le modèle naïf. Nous allons donc dire que 5 sont de meilleurs modèles que le modèle naïf. Comme nous l'avons vu graphiquement la méthode Holt-Winters est la moins bonne de précision que nous avons.

Nous allons maintenant passer aux tests de précision entre tous nos modèles. Comme pour la partie saisonnière, nous avons le modèle 1 qui est en ligne et le modèle 2 qui est en colonne. Nous allons également adopter le même code couleur, lorsque l'hypothèse nulle est acceptée la case sera verte et inversement, la case sera rouge. Entre le modèle ARIMA en modèle 1 et AR(1) en modèle 2, nous avons un problème de variance négative, nous pensons que la cause est que les résidus des deux modèles sont trop proches. Comme vu dans la partie saisonnière, nos deux meilleurs modèles sont ceux comprenant les variables explicatives, nous pouvons dire qu'au seuil de risque 10% que le modèle ARX semble meilleur que le modèle de régression linéaire (LM).

Tableau 13 : DM test sur toutes nos variables

	AR(1)	AR(p)	ARIMA	HW	LM	ARX
AR(1)		0.8965		7.94e-03	1	0.999
AR(p)	0.1035		0.1035	0.113	1	0.999
ARIMA		0.8965		7.94e-03	1	0.999
HW	0.9921	0.8871	0.9921		0.9989	0.999
LM	7.4e-06	1.51e-05	7.4e-06	1.09e-03		0.939
ARX	1.2e-04	7.16e-05	1.2e-04	3.74e-04	0.061	

Bibliographie

Liens pour la base de données :

Pour notre série saisonnière :
Enquête de fréquentation dans l'hôtellerie | Insee

Pour notre série non-saisonnière
https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RBRTE&f=M

Pour nos variables explicatives :

Variable production :
https://www.eia.gov/opendata/qb.php?category=2134979&sdid=INTL.57-1-WORL-TBPD.M

Variable stock :
Cette variable nous a été donnée par Z. MOUSSA

Variable activité mondiale
https://www.dallasfed.org/research/igrea

Variable MSCI index :
https://fr.investing.com/indices/msci-world-historical-data

Variable baltic dry index :
https://fr.investing.com/indices/baltic-dry-historical-data

Annexes

Annexe 1 : test outlier

```
Series: yy
Regression with ARIMA(1,0,0)(0,1,0)[12] errors

Coefficients:
         ar1     TC89
      0.3187  27.9128
s.e.  0.0965   8.2640

sigma^2 estimated as 175.9:  log likelihood=-383.41
AIC=772.82   AICc=773.08   BIC=780.51

Outliers:
  type ind    time coefhat tstat
1   TC  89 2018:05   27.91 3.378
```

Annexe 2 : test de Kruskall Wallis

```
Test used:  Kruskall Wallis

Test statistic:  103.86
P-value:  0
```

Annexe 3 : test de QS

```
Test used:  QS

Test statistic:  180.56
P-value:  0
```

Annexe 4 : test du seasonal Dummies

```
Test used:  SeasonalDummies

Test statistic:  1692.21
P-value:  0
```

Annexe 5 : test de Welch

```
Test used:  Kruskall Wallis

Test statistic:  1172.52
P-value:  4.226223e-43
```

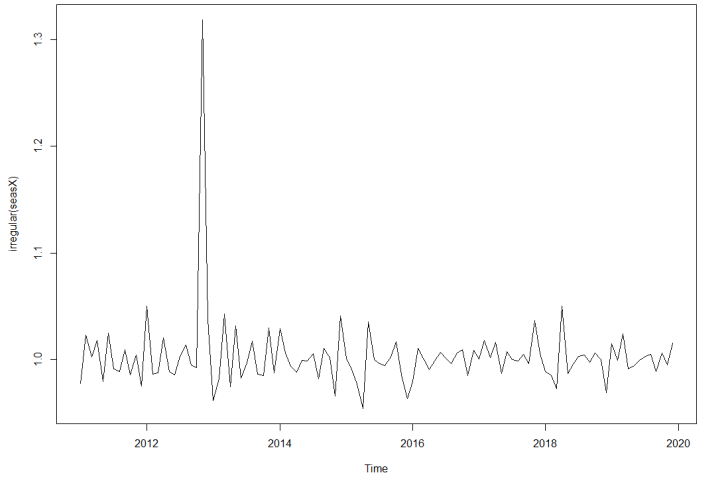
Annexe 6 : test de Weibel-Ollech

```
Test used:  WO

Test statistic:  1
P-value:  0 0 5.440093e-15
```

Annexe 7 : Test outliers sur la série corrigé avec TSO

```
> tsoutliers(adj)
$index
[1] 43

$replacements
[1] 507.2725
```

Annexe 8 : graphique d'irrégularité des observations



Annexe 9 : résultats du modèle SARIMA $(p,d,q)(P,D,Q)_{12}$

```
Series: yy
ARIMA(1,0,0)(0,1,1)[12] with drift

Coefficients:
         ar1     sma1   drift
      0.2751  -0.2480  0.1949
s.e.  0.1056   0.1269  0.1203

sigma^2 estimated as 153.7:  log likelihood=-329.56
AIC=667.12   AICc=667.63   BIC=676.84
```

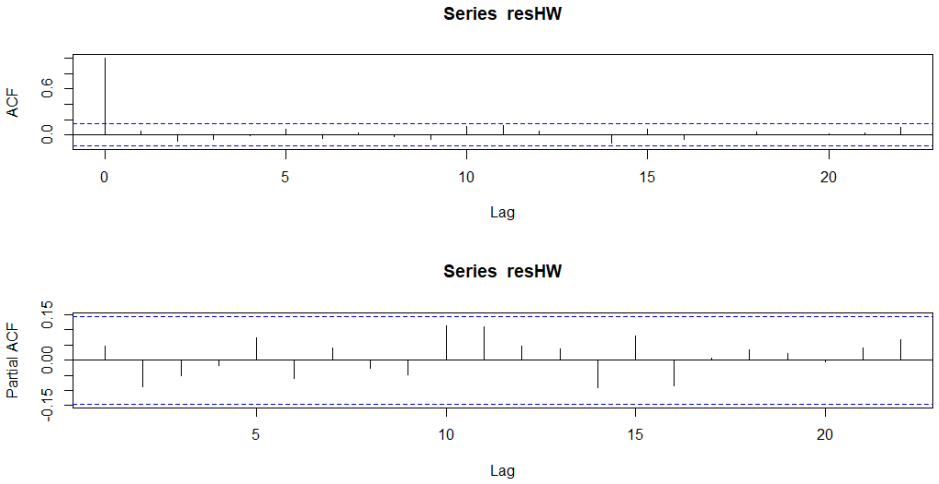
Annexe 10 : résultats du modèle SARIMA $(0,1,1)(0,1,1)_{12}$

```
Call:
arima(x = yy, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ma1     sma1
      -0.8034  -0.2163
s.e.   0.0806   0.1198

sigma^2 estimated as 156.6:  log likelihood = -328.33,  aic = 662.67
```

Annexe 11 : Vérification de la Non-Saisonnalité de la série

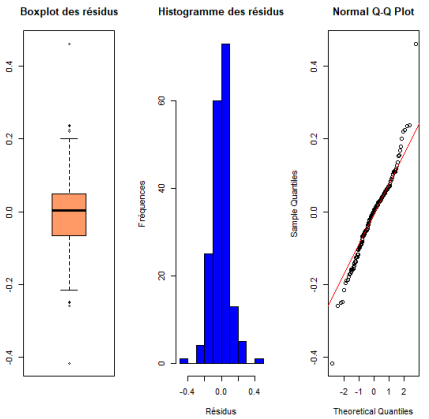
```
> isSeasonal(data$BRENT, test="wo",freq = 12)
[1] FALSE
> isSeasonal(data$STOCK, test="wo",freq = 12)
[1] TRUE
> isSeasonal(data$ACT_MOND, test="wo",freq = 12)
[1] TRUE
> isSeasonal(data$MSCI_index, test="wo",freq = 12)
[1] FALSE
> isSeasonal(data$Baltic_dry_index, test="wo",freq = 12)
[1] FALSE
```

Annexe 12 : indépendance des résidus HW




```
        One-sample Kolmogorov-Smirnov test

data:  resHW
D = 0.080022, p-value = 0.1942
alternative hypothesis: two-sided
```

Annexe 13 : normalité des résidus HW




```
        Box-Ljung test

data:  resHW
X-squared = 0.40862, df = 1, p-value = 0.5227
```

Annexe 14 : StepAIC pour modèle avec toute les variables explicatives

	Df	Sum of Sq	RSS	AIC
<none>			1.1421	-856.51
+ PROD	1	0.007307	1.1348	-855.62
+ STOCK	1	0.000286	1.1418	-854.55
+ Baltic_dry_index	1	0.000140	1.1420	-854.53
- ACT_MOND	1	0.029122	1.1712	-854.18
- MSCI_index	1	0.157927	1.3000	-836.24