Projet Séries Temporelles

Felipe Montenegro de Carvalho
14 mai 2020

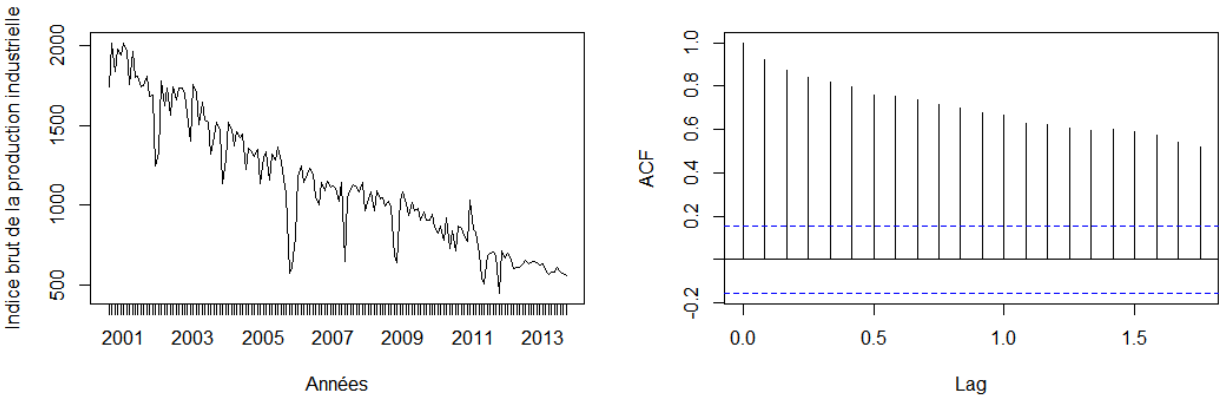
Partie I : Les Données

Mon étude porte sur un indice de production industrielle corrigé des variations saisonnières et jours ouvrables (CVS-CJO). La série retenue permet de suivre l'évolution mensuelle de l'Indice CVS-CJO de la production industrielle (base 100) de l'extraction de gaz naturel, poste 06.2 selon la classification de la NAF, entre juillet 2007 et septembre 2013 et comporte donc 159 observations.

Avant de commencer à étudier la série, il était nécessaire de la traiter afin de faciliter sa manipulation en la formalisant de manière pratique (Cf. Anexe II.1).

Après ce traitement, j'ai représenté graphiquement la série avec et sa fonction d'autocorrélation pour essayer d'identifier une saisonnalité ou une tendance. Je la noterai dans la suite comme suit : $(X_t)_{t\in T}$, avec $T = \{1, ..., T\}$ l'ensemble contenant toutes les dates où la série est observée.

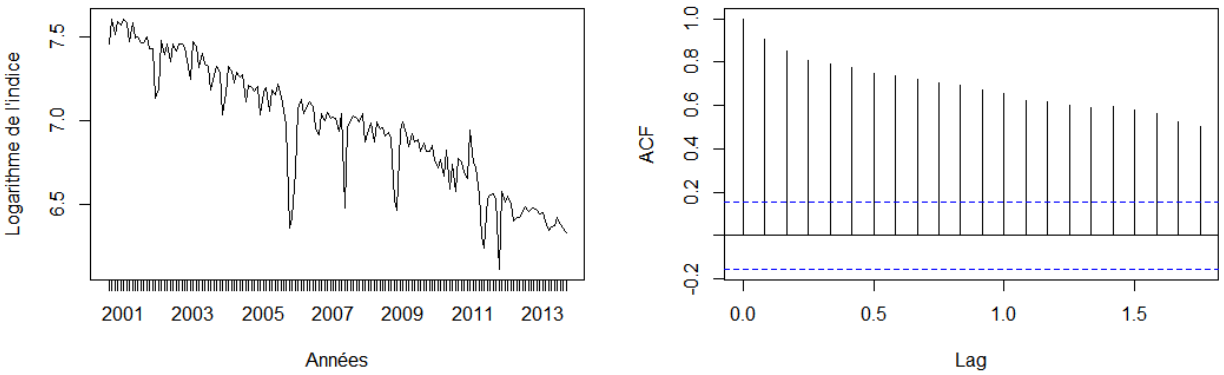
FIGURE 1 – Représentation graphique de la série étudiée - Indice CVS-CJO de la production industrielle de l'extraction de gaz naturel entre le 07/2001 et le 09/2013 et son autocorrélogramme



La série est représentée graphiquement sur la figure 1 ci-dessus. Graphiquement, elle semble présenter une tendance linéaire, mais pas une saisonalité. On a donc l'intuition que la série est stationnaire. En outre, on peut constater une certaine hétéroscédasticité dans les données.

La première étape consiste à essayer de corriger le caractère hétéroscédastique de la série en lui appliquant une transformation logarithmique. Je noterai la série log-transformée comme suit : $(W_t)_{t\in T}$. On a, donc, $W_t = log(X_t), \forall t \in T$. L'étude des auto-corrélations de la série $(W_t)_{t\in T}$ (disponibles ci-dessous dans la figure 2) ainsi que l'étude de la série elle-même nous permettent de conclure que, malgré cette transformation, la série étudiée semble toujours avoir une tendance linéaire (encore plus évidente) mais toujours aucune saisonallité apparente.

FIGURE 2 – Logarithme de l'indice CVS-CJO de la production industrielle de l'extraction de gaz naturel entre le 07/2001 et le 09/2013 et son autocorrélogramme



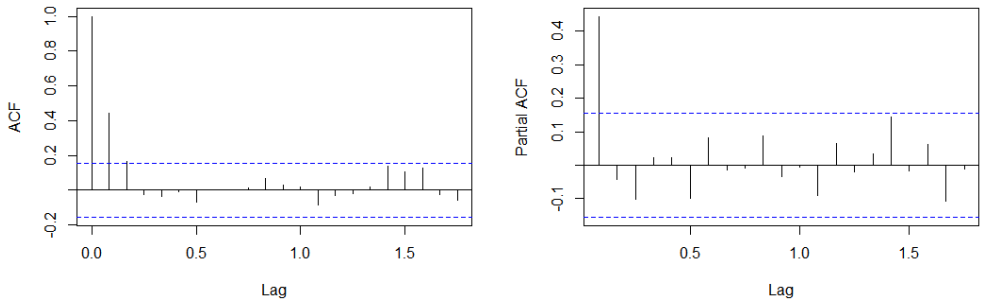
Pour vérifier la supposée tendance linéaire, j'ai régressé le logarithme des indices sur ses dates.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 188.454444   5.802620   32.48   <2e-16 ***
dates        -0.090424   0.002891  -31.28   <2e-16 ***
---
```

Le coefficient associé à la tendance linéaire (dates) est effectivement bien significatif et négatif, comme on s'y attendait.

En notant les résidus de la régression comme $(Z_t)_{t\in T}$, ses fonctions d'autocorrélation et d'autocorrélation partielle (figure 3) nous montrent que l'autocorrélation d'ordre 1 (totale ou partielle, c'est la même chose) est d'un peu plus de 0.4, soit petite et loin d'être égal à 1 (plus de détails sur l'Anexe II.2). La série semble donc stationnaire.

FIGURE 3 – Fonctions d'autocorrélation et d'autocorrélation partielle de $(Z_t)_{t\in T}$, le logarithme de la série corrigé de sa tendance



Par conséquent, la prochaine étape est d'appliquer un test de racine unitaire pour vérifier cette supposition. J'ai appliqué donc le test de Dickey-Fuller Augmenté (Cf. Anexe III.1) avec

constante et tendance non nulles étant donnée la significativité du coefficient de la régression effectuée.

Cependant, pour que ce test soit valide, il faut vérifier si les résidus sont bien non autocorrélés, ce qui a été fait et le résultat détaillé est montré dans l'Anexe II.3, où l'absence d'autocorrélation des résidus n'est pas rejetée à un seuil de 95% et, donc, le test ADF avec aucun retard est valide.

Maintenant, on interprète le test de racine unitaire pour plus de sécurité.

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 0
  STATISTIC:
    Dickey-Fuller: -7.7056
  P VALUE:
    0.01
```

La racine unitaire est rejetée à un seuil de 95% pour la série en niveau ($p - value < 0.05$).

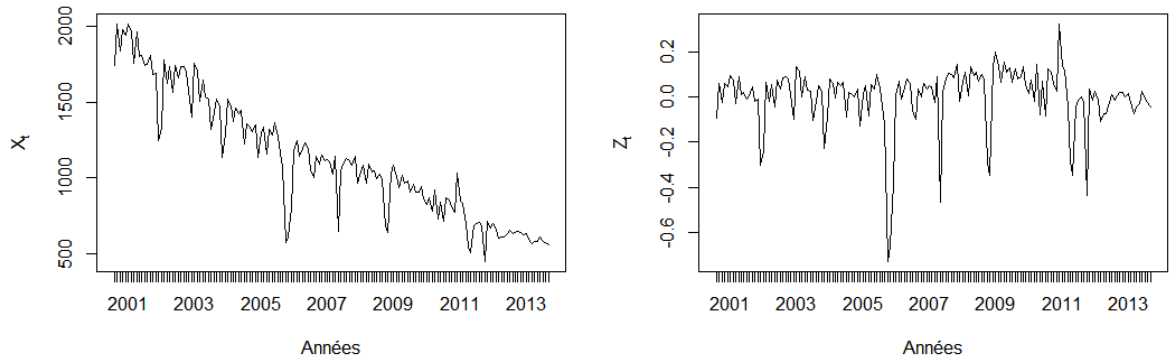
J'ai également effectué le test racine unitaire Phillips-Perron (Cf. Anexe III.2) sur $(W_t)_t \in T$ et le résultat a été que l'hypothèse nulle de racine unitaire est aussi rejetée à un seuil de 95%. On dira que la série est stationnaire.

```
        Phillips-Perron Unit Root Test

data:  W_t
Dickey-Fuller Z(alpha) = -84.082, Truncation lag parameter = 4,
p-value = 0.01
alternative hypothesis: stationary

Warning message:
In pp.test(W_t) : p-value smaller than printed p-value
```

De plus, en prenant les résidus de la régression, on obtient le logarithme de la série corrigé de sa tendance et, donc, la transformation sur la série originale est terminée.

Ci-dessous, on peut voir notre série avant $(X_t)_{t\in T}$ et après $(Z_t)_{t\in T}$ ces transformations.

FIGURE 4 – Représentation graphique de la série $(X_t)_{t\in T}$ (série initiale, premier graphique) et de la série $(Z_t)_{t\in T}$ (logarithme de la série initiale corrigée de sa tendance, second graphique)



Partie II : Modèles ARMA

Du fait que l'hypothèse de stationnarité a été bien confirmée dans la Partie I et sans la necéssité d'éléver quelque racine unitaire, la série corrigée $(Z_t)_{t\in T}$ est intégrée d'ordre $d^* = 0$.

La détermination des paramètres pertinents pour une modélisation se fait grâce aux autocorrélogrammes (complets et partiels). Par l'acf et le pacf de la série corigée $(Z_t)_{t\in T}$ représentées sur la figure 3, on voit que les fonctions d'autocorrélation totales sont significatives (i.e. plus grandes que les bornes $\pm\frac{1.96}{\sqrt{n}}$ de l'intervalle de confiance d'un test de nullité de l'autocorrélation à un seuil de 95%) jusqu'à $q^* = 2$ et les autocorrélations partielles le sont jusqu'à $p^* = 1$.

Alors, on peut conclure que si la série corrigée suit un modèle ARIMA(p,d,q), elle suit au plus un ARIMA(1,0,2). C'est pourquoi, les sous-modèles possibles sont les ARIMA tels que $p \leq p^*$, $d \leq d^*$ et $q \leq q^*$.

Cependant, pour qu'un modèle soit valide, il faut que ses résidus aient des propriétés similaires à celles des bruits blancs. C'est pourquoi, pour chaque modèle considéré, j'ai étudié sa validité en regardant la blancheur de ses résidus par le Test de Validité et la significativité des coefficients estimés par le Test de Nullité des Coefficients (Cf. Anexe III.3 et Anexe III.4).

Finalement, j'ai vérifié si le modèle ARIMA(1,0,2) et ses modèles alternatifs passaient par les tests de validité et de nullité des coefficients. Les détails sur chaque modèle est dans l'Anexe II.3.

— $p = 0$ et $q = 1$: Bien ajusté, mais pas valide ;
— $p = 0$ et $q = 2$: Bien valide et on le garde comme modèle candidat ;
— $p = 1$ et $q = 0$: Bien valide et on le garde comme modèle candidat ;
— $p = 1$ et $q = 1$: Mal ajusté ;
— $p = 1$ et $q = 2$: Mal ajusté ;

J'ai trouvé donc 2 sous-modèles bien ajustés et valides : ARIMA(0,0,2) et ARIMA(1,0,0). On peut voir ci-dessous leurs tableaux qui nous permettent de faire cette affirmation :

— ARIMA(0,0,2) : On voit par les valeurs données aux tableaux ci-dessous que ce modèle passe par le Test de Validité car l'absence d'autocorrélation entre les résidus de la série corrigée n'est jamais rejetée à 95% jusqu'à 24 retards et, aussi, il passe par le Test de Nullité des Coefficients car le coefficient de MA(2) est significatif, i.e., le coefficient des retards les plus élevé MA(2) (p-valeur = 0.004) rejette sa nullité à 95% (p-value>0.05). Alors, le modèle ARIMA(0,0,2) est bien valide et on le garde comme modèle candidat.

```
> Qtests(arima002$residuals, 24, 2)
      lag      pval       lag      pval
 [1,]   1        NA [13,]  13 0.8180712
 [2,]   2        NA [14,]  14 0.8627769
 [3,]   3 0.8083421 [15,]  15 0.9041402
 [4,]   4 0.9040696 [16,]  16 0.9259227
 [5,]   5 0.9027171 [17,]  17 0.8071666
 [6,]   6 0.6663360 [18,]  18 0.8560670
 [7,]   7 0.7379795 [19,]  19 0.6804204
 [8,]   8 0.8388667 [20,]  20 0.6882532
 [9,]   9 0.8930447 [21,]  21 0.7296535
[10,]  10 0.8800528 [22,]  22 0.7789660
[11,]  11 0.9278641 [23,]  23 0.8127256
[12,]  12 0.9346286 [24,]  24 0.6881353
> signif(arima002)
             ma1         ma2     intercept
coef 4.645126e-01 0.242018403 -0.0005616882
se   7.715632e-02 0.084120521  0.0165065103
pval 1.739778e-09 0.004014202  0.9728545980
```

— ARIMA(1,0,0) :
Par le même raisonement que pour le modèle précédent, ce modèle passe par le Test de Validité et, aussi, il passe par le Test de Nullité des Coefficients car le coefficient de AR(1) est significatif. Alors, le modèle ARIMA(1,0,0) est bien valide et on le garde comme modèle candidat.

```
> Qtests(arima100$residuals, 24, 1)
      lag      pval       lag      pval
 [1,]   1        NA [13,]  13 0.7205712
 [2,]   2 0.7623629 [14,]  14 0.7858158
 [3,]   3 0.3851794 [15,]  15 0.8368494
 [4,]   4 0.5441648 [16,]  16 0.8695501
 [5,]   5 0.6436132 [17,]  17 0.7109235
 [6,]   6 0.5312617 [18,]  18 0.7696668
 [7,]   7 0.6200910 [19,]  19 0.5570085
 [8,]   8 0.7302776 [20,]  20 0.5543337
 [9,]   9 0.8090713 [21,]  21 0.5880956
[10,]  10 0.7833969 [22,]  22 0.6458037
[11,]  11 0.8505958 [23,]  23 0.6920249
[12,]  12 0.8591527 [24,]  24 0.5551652
> signif(arima100)
              ar1     intercept
coef 4.430675e-01 -0.0006823273
se   7.100453e-02  0.0174693285
pval 4.375986e-10  0.9688436800
```

Parmi ces deux modèles, on peut utiliser les critères d'information AIC ou BIC pour déterminer quel modèle est le plus adapté (i.e., ce qui minimise les critères), puisque ces critères mesurent la qualité d'un modèle statistique.

```
          ma2       ar1
AIC -208.1096 -207.8545
BIC -195.8592 -198.6667
```

Par le tableau ci-dessus, on peut conclure que le modèle ARIMA(0,0,2) minimise l'AIC et le modèle ARIMA(1,0,0) minimise le BIC, ce qui n'est pas une surprise, car normalement BIC choisit le modèle le plus « économique ». Je privilégierais le critère AIC et retiens donc le modèle ARIMA(0,0,2) (ou MA(2)).

L'expression du modèle proposé est donnée par :

$$log(X_t) = (1 - 0.4645B)(1 - 0.2420B)\epsilon_t - 0.09043t + 188.4544$$

Où B est l'Opérateur Retard (Backward).

Partie III : Prévision

Dorénavant, on supposera que les résidus sont gaussiens i.e., $\epsilon_t \sim \mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2)$ et on se place en date T. On veut trouver l'équation vérifiée par la région de confiance de niveau α sur les valeurs futures (X_{T+1}, X_{T+2}). Soit $\phi_1 = 0.4645$, $\phi_2 = 0.2420$, $a = -0.09043$ et $b = 188.4544$, les coefficients du modèle proposé. Pour faciliter les calculs, je vais les faire sur la série log-transformée $(W_t)_{t\in T}$. Alors :

$$\begin{aligned} W_t &= (1 - \phi_1 B)(1 - \phi_2 B)\epsilon_t + at + b \\ &= \epsilon_t + \phi_1\phi_2\epsilon_{t-2} - (\phi_1 + \phi_2)\epsilon_{t-1} + at + b \end{aligned} \tag{1}$$

On cherche la prévision linéaire optimale des valeurs futures (W_{T+1}, W_{T+2}), notées (W_{T+1}^T, W_{T+2}^T). Par l'équation de W_t, on a que :

$$\begin{aligned} W_{T+1} &= \epsilon_{T+1} - (\phi_1 + \phi_2)\epsilon_T + \phi_1\phi_2\epsilon_{T-1} + aT + (b+a) \\ W_{T+2} &= \epsilon_{T+2} - (\phi_1 + \phi_2)\epsilon_{T+1} + \phi_1\phi_2\epsilon_T + aT + (b+2a) \end{aligned} \tag{2}$$

En plus, la prévision linéaire optimale est donnée par :

$$W_{t+1}^t = \mathbb{E}L[W_{t+1}|W_t, ..., W_1] \ ; \ \forall t \in 1, ..., T \tag{3}$$

Comme ϵ_{T+1} et ϵ_{T+2} sont des bruits blancs car ils sont des innovations, ils sont orthogonaux à $(W_T, ...W_1)$ et, donc, $\mathbb{E}L[\epsilon_{t^*}|W_T, ..., W_1] = 0 \ ; \ \forall t^* \in \{T+1, T+2\}$. C'est pourquoi, par les équations (2), la prévision linéaire définie au-dessus et en utilisant les propriétés de l'espérance conditionnelle, on a que :

$$\begin{aligned} W_{T+1}^T &= -(\phi_1 + \phi_2)\epsilon_T + \phi_1\phi_2\epsilon_{T-1} + aT + (b+a) \\ W_{T+2}^T &= \phi_1\phi_2\epsilon_T + aT + (b+2a) \end{aligned} \tag{4}$$

Les erreurs de prédictions sont, alors :

$$\begin{cases} W_{T+1} - W_{T+1}^T = \epsilon_{T+1} \\ W_{T+2} - W_{T+2}^T = \epsilon_{T+2} - (\phi_1 + \phi_2)\epsilon_{T+1} \end{cases}$$

Avant déterminer la région de confiance de la prévision, il convient de calculer les variances et la covariance des erreurs de prédiction à fin d'obtenir leur matrice variance-covariance Σ. Je noterai les erreurs de prédiction comme e_{T+1} et e_{T+2}. On a donc :

$$\begin{aligned} Var[e_{T+1}] &= Var[\epsilon_{T+1}] = \sigma_\epsilon^2 \\ Var[e_{T+2}] &= Var[\epsilon_{T+2} - (\phi_1 + \phi_2)\epsilon_{T+1}] = \sigma_\epsilon^2(1 + (\phi_1 + \phi_2)^2) \\ CoV(e_{T+1}, e_{T+2}) &= -(\phi_1 + \phi_2)\sigma_\epsilon^2 \end{aligned} \tag{5}$$

Ce dernier résultat est obtenu lorsque les processus d'innovation suivent une loi Gaussienne et qu'ils sont i.i.d.'s. Donc, par la linéarité et stabilité de la loi Normale, on a :

$$W = \begin{pmatrix} W_{T+1} - W_{T+1}^T \\ W_{T+2} - W_{T+2}^T \end{pmatrix} \sim \mathcal{N}(0, \Sigma) \ ; \ avec \ \Sigma = \begin{pmatrix} \sigma_\epsilon^2 & -(\phi_1 + \phi_2)\sigma_\epsilon^2 \\ -(\phi_1 + \phi_2)\sigma_\epsilon^2 & \sigma_\epsilon^2(1 + (\phi_1 + \phi_2)^2) \end{pmatrix}$$

En supposant que $\sigma_\epsilon^2 > 0$, on a que Σ est inversible et, donc, on peut faire le test $H_0 : W = \vec{0}$ contre $H_1 : W \neq \vec{0}$, où $\vec{0} \in \mathbb{R}^2$. Soit $q_{\chi_2^2}^{1-\eta}$ le quantile de niveau $\alpha = 1 - \eta$ d'une loi χ_2^2. Le test d'hypothèse sera donc :

$$\Phi_\eta(X_1, ..., X_T) = \begin{cases} 0, & si\ \widehat{W}'\Sigma^{-1}\widehat{W} < q_{\chi_2^2}^{1-\eta} \\ 1, & sinon \end{cases}$$

Où la statistique du test vérifie $W'\Sigma^{-1}W \sim \chi_2^2$, $\widehat{W}$ est la valeur empirique du vecteur W.

Alors, la région de confiance de niveau α sur les valeurs futures (X_{T+1}, X_{T+2}) est donnée par :

$$R_\alpha(X_1, ..., X_T) = \left\{ \widehat{W}'\Sigma^{-1}\widehat{W} < q_{\chi_2^2}^{\alpha} \right\}$$

Lorsqu'on a un test de niveau η, en prenant un point de vue univarié, on peut bien construire un intervalle de confiance au niveau $1-\eta$ pour W_{T+1}^T et un autre pour W_{T+2}^T.

Comme $e_{T+1} = \epsilon_{T+1}$, l'erreur de prévision suit une loi $\mathcal{N}(\mu_\epsilon, \sigma_\epsilon)$. Soit $q_{\mathcal{N}(0,1)}^{1-\eta}$ le quantile d'ordre $\alpha = 1-\eta$ d'une loi $\mathcal{N}(0,1)$ et sous l'hypothèse $\mu_\epsilon = 0$, un intervalle de niveau de confiance $1-\eta$ pour W_{T+1}^T est donné par la formule :

$$IC_{1-\eta}(W_{T+1}^T) = \left[\widehat{W}_{T+1}^T - q_{\mathcal{N}(0,1)}^{1-\frac{\eta}{2}}\sigma_\epsilon\ ;\ \widehat{W}_{T+1}^T + q_{\mathcal{N}(0,1)}^{1-\frac{\eta}{2}}\sigma_\epsilon \right]$$

Par le même raisonement, on a aussi un intervalle de niveau de confiance $1-\eta$ pour W_{T+2}^T donné par la formule :

$$IC_{1-\eta}(W_{T+1}^T) = \left[\widehat{W}_{T+1}^T - q_{\mathcal{N}(0,1)}^{1-\frac{\eta}{2}}\sigma_\epsilon\sqrt{(1+(\phi_1+\phi_2)^2)}\ ;\ \widehat{W}_{T+1}^T + q_{\mathcal{N}(0,1)}^{1-\frac{\eta}{2}}\sigma_\epsilon\sqrt{(1+(\phi_1+\phi_2)^2)} \right]$$

Pour obtenir les bornes pour la série initiale, il suffit de prendre l'exponentielle de chaque borne, puisque $W_t = log(X_t)$, $\forall t \in \{1, ..., T\}$, ce qui a été fait pour obtenir les prévisions de X_t sur la figure 5 au-dessous.

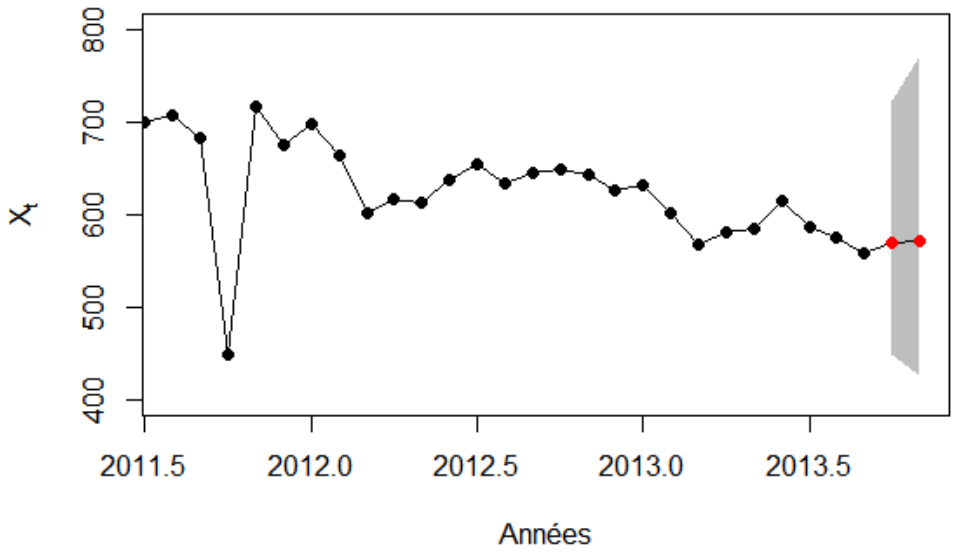
Pour obtenir ces intervalles de confiance et la région R_α, on avait besoin de supposer plusieurs hypothèses. D'abord, pour les faire, j'ai supposé que la variance des résidus était connue et plus grande que 0. Si l'on ne connaissait pas la variance (σ_ϵ inconnu), il faudrait qu'on fasse une estimation sur σ_ϵ en utilisant son estimateur de vraisemblance ou sa variance empirique. Dans les deux cas, la statistique utilisé pour obtenir les intervalles de confiance ne suit pas une loi Normale, mais si une loi t-student qui peut être obtenue à partir du Théorème de Cochran.

Il peut être intéressant de remarquer que j'ai supposé qui les résidus suivaient une loi normale et qu'ils étaient i.i.d.'s, ce qui n'est pas normalement vrai en pratique. C'est pour ça qu'il existe des tests de Normalité comme le Test de χ_2 et le Test de Kolmogorov pour avoir une justificatif plausible avant de les considérer de loi normale.

J'ai supposé aussi que mon modèle était parfaitement identifiable.

Le graphique ci-dessous donne la visualisation de la région de confiance (en grise) et les prédictions (en rouge) pour chaque horizon considéré pour $\alpha = 95\%$ qui représentent respectivement à la prédiction pour octobre 2013 et novembre 2013.

FIGURE 5 – Prévision de (X_{T+1}, X_{T+2}) en utilisant un ARIMA(0,0,2) et intervalle de confiance univarié à 95%



Maintenant on considère une série stationnaire $(Y_t)_{t\in T}$, où Y_{T+1} est disponible plus rapidement que X_{T+1}. Dans ce cas là, Y_{T+1} nous aidera à améliorer la prévision de X_{T+1} juste si la série $(Y_t)_t \in T$ cause instantanément $(X_t)_{t\in T}$ au sens Granger, où la causalité se définie par l'utilité positive d'une variable pour en prévoir une autre.

Mathématiquement, pour que ce cas arrive, il faut que :

$$\mathbb{E}L[X_{T+1}|(Y_{T+1}, ..., Y_1), (X_T, ..., X_1)] \neq \mathbb{E}L[X_{T+1}|(X_T, ..., X_1)]$$

Anexes

Anexe I : Plot et ACF des séries étudiées

FIGURE 6 – $(X_t)_{t\in T}$ - Série Initialle

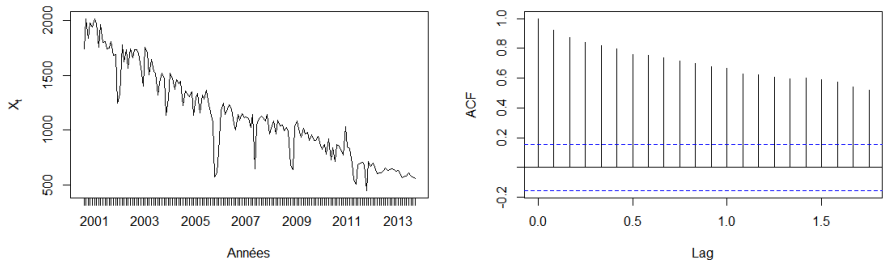


FIGURE 7 – $(W_t)_{t\in T}$ - Logarithme de la Série Initialle

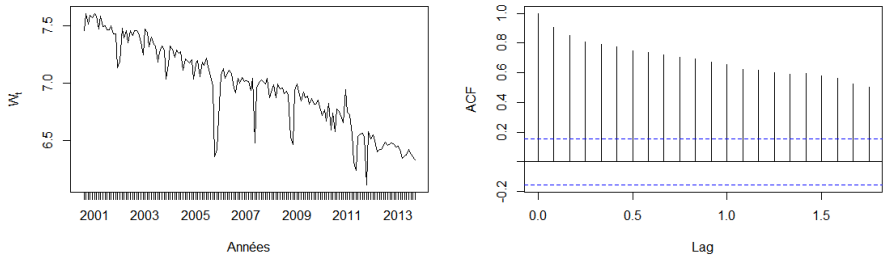
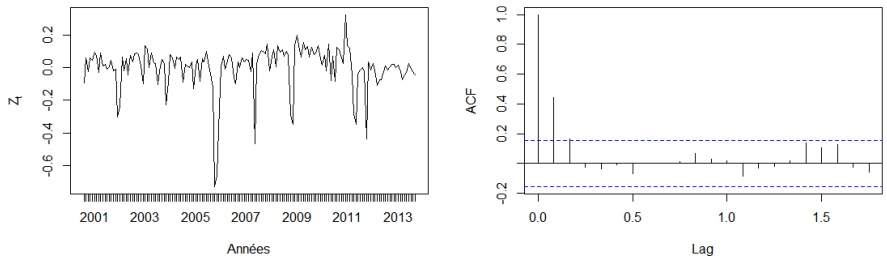


FIGURE 8 – $(Z_t)_{t\in T}$ - Logarithme de la Série Initialle corrigée de sa tendance linéaire



Anexe II : Résultats Pratiques

1. Détails des données sur l'Application R

   Avant commencer à étudier la série, il était nécessaire de la traiter afin de faciliter sa manipulation en la formalisant de manière pratique. Pour cela, j'ai créé un « double », en utilisant le package « zoo », qui contient les Dates (le csv nous les a donnés en format « integer » et il faudrait changer pour « double ») et ses respectives valeurs.

2. Autocorrélation au premier ordre

   Si l'autocorrélation du premier ordre est proche de 1, on suspecte une non-stationnarité. Cependant, il est possible que l'autocorrélation soit très forte sans pour autant que la série soit non-stationnaire. Pour vérifier cela, on utilise un test de seconde génération. Le plus connu est le test KPSS.
   Comme dans ma série l'autocorrélation du premier ordre a été vers 0.4, je n'ai pas eu besoin de faire ce test là.

3. Non-Autocorrélation des Résidus de la Régression

   Lorsqu'on fait une régression pour corriger la tendance linéaire de la série temporelle, pour que le test de racine soit valide à fin de savoir si la série est stationnaire, il faut vérifier si les résidus de la régression ne sont pas autocorrélés. La tableau ci-dessous montre les p-valuers du test qui contient l'absence de l'autocorrélation des résidus comme hypothèse nulle jusqu'à l'ordre 24 (deux ans). Les NAs représentent la correction des degrés de liberté du nombre de regresseurs. On voit que l'absence d'autocorrélation n'est jamais rejetée à un seuil de 95%. Alors, il n'a pas été nécessaire d'inclure des retards dans le test ADF.

	lag	pval		lag	pval
[1,]	1	NA	[13,]	13	0.5643229
[2,]	2	NA	[14,]	14	0.6500898
[3,]	3	NA	[15,]	15	0.7221718
[4,]	4	0.1512257	[16,]	16	0.7597108
[5,]	5	0.2905008	[17,]	17	0.5807666
[6,]	6	0.2559940	[18,]	18	0.6515302
[7,]	7	0.3648118	[19,]	19	0.4319170
[8,]	8	0.5046874	[20,]	20	0.4367365
[9,]	9	0.6217539	[21,]	21	0.4709640
[10,]	10	0.6013347	[22,]	22	0.5352197
[11,]	11	0.7042143	[23,]	23	0.5879229
[12,]	12	0.7249745	[24,]	24	0.4384389

4. Vérification des Tests de Validité et de Nullité des Coefficients

   Dans cette partie de l'Annexe, j'ai mis la vérification pour le modèles possibles qui ont été rejetés.

— ARIMA(0,0,1) :
On peut voir par les tableaux ci-dessous que le modèle est bien ajusté car le coefficient de MA(1) est significatif, cependant le modèle n'est pas valide parce que les résidus sont auto-corrélés puisque la statistique pour $k \in 1, 2$ rejette l'absence d'autocorrélation à 95%.

	lag	pval		lag	pval
[1,]	1	NA	[13,]	13	0.37328975
[2,]	2	0.01687206	[14,]	14	0.44956623
[3,]	3	0.03663709	[15,]	15	0.52546990
[4,]	4	0.08176179	[16,]	16	0.59367657
[5,]	5	0.14225124	[17,]	17	0.37728775
[6,]	6	0.13855729	[18,]	18	0.44474574
[7,]	7	0.19919213	[19,]	19	0.25361583
[8,]	8	0.28482321	[20,]	20	0.26935016
[9,]	9	0.37653685	[21,]	21	0.31306434
[10,]	10	0.36966434	[22,]	22	0.36811702
[11,]	11	0.45494940	[23,]	23	0.39674113
[12,]	12	0.47684610	[24,]	24	0.33750682

	ma1	intercept
coef	3.846234e-01	-0.0003910563
se	6.368335e-02	0.0137452494
pval	1.544748e-09	0.9773030182

— ARIMA(1,0,1) :
On peut voir par les tableaux ci-dessous que le modèle passe le Test de Validité car l'absence d'autocorrélation entre les résidus de la série corrigée n'est jamais rejetée à 95% jusqu'à 24 retards, cependant il n'est pas bien ajusté car le coefficient de MA(1) n'est pas significatif.

	lag	pval		lag	pval
[1,]	1	NA	[13,]	13	0.6317465
[2,]	2	NA	[14,]	14	0.7069202
[3,]	3	0.1895752	[15,]	15	0.7713872
[4,]	4	0.3839601	[16,]	16	0.8131234
[5,]	5	0.5087043	[17,]	17	0.6238618
[6,]	6	0.4036092	[18,]	18	0.6929819
[7,]	7	0.5010618	[19,]	19	0.4622055
[8,]	8	0.6300930	[20,]	20	0.4608593
[9,]	9	0.7284788	[21,]	21	0.5026251
[10,]	10	0.7014410	[22,]	22	0.5648558
[11,]	11	0.7852478	[23,]	23	0.6102414
[12,]	12	0.7924257	[24,]	24	0.4692168

	ar1	ma1	intercept
coef	0.390719516	0.06593267	-0.0006558581
se	0.141129752	0.14684843	0.0170202490
pval	0.005631279	0.65344280	0.9692619318

— ARIMA(1,0,2) :
Par le premier tableau, on voit que l'absence d'autocorrélation entre les résidus de la série corrigée n'est jamais rejetée à 95% jusqu'à 24 retards, donc le modèle ARMA(1,0,2) passe par le test de validité. En revanche, en regardant le deuxième tableau, on voit que les coefficients des retards les plus élevés AR(1) (p-valeur = 0.87) et MA(2) ($p - valeur = 0.06$) ne rejettent chacun pas leur nullité à 95% ($p - value > 0.05$), le modèle ARIMA(1,0,2) est donc mal ajusté.

	lag	pval		lag	pval
[1,]	1	NA	[13,]	13	0.7548465
[2,]	2	NA	[14,]	14	0.8112225
[3,]	3	NA	[15,]	15	0.8648278
[4,]	4	0.6378939	[16,]	16	0.8949450
[5,]	5	0.7451421	[17,]	17	0.7525873
[6,]	6	0.5010930	[18,]	18	0.8113383
[7,]	7	0.6053845	[19,]	19	0.6204457
[8,]	8	0.7423499	[20,]	20	0.6320879
[9,]	9	0.8240663	[21,]	21	0.6790461
[10,]	10	0.8147939	[22,]	22	0.7340636
[11,]	11	0.8840069	[23,]	23	0.7719951
[12,]	12	0.8975677	[24,]	24	0.6409248

	ar1	ma1	ma2	intercept
coef	-0.0462504	0.50738634	0.2607428	-0.0005686062
se	0.2837846	0.27214365	0.1394384	0.0163471477
pval	0.8705364	0.06226467	0.0614908	0.9722526108

Anexe III : Contenu Théorique

1. Test de Dickey-Fuller Augmenté

Le test augmenté de Dickey-Fuller ou test ADF est un test statistique qui vise à savoir si une série temporelle est stationnaire c'est-à-dire si ses propriétés statistiques (espérance, variance, auto-corrélation) varient ou non dans le temps. Ce processus est réalisé en vérifiant s'il y a des racines unitaires dans l'équation caractéritique de la série car si la série présente une racine unitaire, sa moyenne et variance ne sont pas constantes par rapport au temps.
Une série avec une racine unitaire est non-stationnaire. On a aussi que si les autres racines de l'équation caractéristique sont dedans le cercle unitaire (module moins grand que 1), alors faire une différence première de la série suffit pour que elle soit stationnaire.
Dans ma série, cela n'a pas été le cas car les résidus de la régression linéaire sont bien non-autocorrélés et, donc, j'ai pu appliquer le Test de Dickey-Fuller Augmenté avec constante et tendance non nulles sans la probabilité que la régression soit fictive.
Alors, le test est donné par $H_0 : \phi = 0$ contre $H_1 : \phi \neq 0$, où si $\phi = 0$ on aura au moins une racine unitaire.
Le modèle ADF générale pour une série $(X_t)_t$ est donnée par :

$$\Delta X_t = c + bt + \phi X_t + \mathbb{I}_{k>0} \sum_{l=1}^{k} \alpha_k \Delta X_{t-k} + \epsilon_t$$

où k est le nombre de retards nécessaires à considérer pour rendre des résidus non-corrélés (dans mon cas, k = 0 car les résidus ont été directement non-autocorrélés).
La statistique du test est associée à l'estimation du coefficient ϕ donnée par :

$$\text{t-stat} = \frac{\widehat{\phi}}{\widehat{\sigma}(\widehat{\phi})}$$

Si t-stat > valeur critique associée au test 5% alors on ne peut pas rejetter l'hypothèse nulle de racine unitaire (donc de non-stationnarité) de la série à 5%. Le test ADF conclura donc à la non-stationnarité de la série en niveau.
Cela n'est pas arrivée dans le cas analysé et la série est bien stationnaire.

2. Test de Phillips-Perron

Le test de Phillips-Perron a le même but du test ADF. Par contre, son modèle est donné par :

$$X_t = c + \rho X_t + u_t$$

Où u_t est un terme d'erreur très général dont on note :
— la variance marginale $\sigma_u^2 = \lim_{n\to\infty} \sum_{t=1}^{n} u_t^2$ p.s

— la variance long-terme $\nu_u^2 = \lim_{n\to\infty} Var[n^{-\frac{1}{2}} \sum_{t=1}^{n} u_t]$

Et la statistique de test de l'hypothèse nulle de racine unitaire pour X_t est donné par :

$$V_t = \frac{\widehat{\sigma}_u}{\widehat{\nu}_u} \frac{\widehat{\rho}_n - 1}{\widehat{\sigma}_{\widehat{\rho}_n}} - \frac{n\widehat{\sigma}_{\widehat{\rho}_n}}{2\widehat{\sigma}_u\widehat{\nu}_u}(\widehat{\nu}_u^2 - \widehat{\sigma}_u^2)$$

où cette statistique suit sous H_0 la seconde loi asymptotique de Dickey-Fuller.

3. Test de Validité (Ljung-Box)

   Le modèle est valide si ses résidus ne sont pas auto-corrélés. Je l'ai testé en utilisant le Test de Ljung-Box de l'hypothèse nulle de nullité jointe des autocorrélations jusqu'à un ordre k donné qui utilise une statistique de test qui suit une loi χ_2 à k-p-q degrés de liberté. Donc, on ne peut l'effectuer que dans la situation où $k > p + q$.

   Plus précisement, le but est de tester $H_0 : \rho_l = 0 \forall l \in p+q, ...T$ contre $H_1 : \exists l \in p+q, ...Ttelque\rho_l \neq 0$.

   La statistique de ce test s'écrit :

$$Q(k) = T(T+2) \sum_{l=1}^{k} \frac{\widehat{\rho}_l}{T-l}$$

4. Test de Nullité des Coefficients

   On mesure la significativité des coefficients par le rapport entre le coefficient estimé et l'écart-type estimé du coefficient estimé. Pour que le modèle soit bien ajusté, la p-valeur correspondant de ce rapport doit être inférieur à 0.05.