ENSAE ParisTech

Projet Séries Temporelles

Mai 2019

Modélisation de l'indice de production industrielle du commerce d'électricité

Encadrant
Raphaël Lee

Andréa Epivent
Dimitri Meunier

Table des matières

Introduction

Ce rapport détaille les réponses apportées par Andréa Epivent et Dimitri Meunier au projet du cours de séries temporelles linéaires de deuxième année de l'ENSAE Paristech. L'objet de ce projet consiste à la modélisation et prévision de l'indice de production industrielle observé en France et disponible sur le site de l'Insee : (https://www.insee.fr/fr/statistiques?debut=0&categorie=10). Dans un premier temps, nous cherchons à stationnariser la série choisie: le commerce d'électricité, en essayant plusieurs méthodes de désaisonnalisation. Puis, nous validons nos transformations à l'aide de tests de racine unitaire et de stationnarité. Par la suite, nous cherchons le modèle ARMA correspondant le mieux à notre série transformée en suivant la méthodologie établie par Box et Jenkins. Nous trouvons que notre série suit un ARMA(4,4). Dans la dernière partie, nous nous intéressons à la région de confiance que vérifient nos prévisions.

1 Partie 1: Les données

1.1 Série temporelle étudiée et première transformation

Nous avons choisi d'étudier et de modéliser l'indice brut de la production industrielle du commerce d'électricité (base 100 en 2015). Cette série a une fréquence mensuelle et couvre la période allant de janvier 1990 à février 2019. Nous avons donc 350 observations (29 ans et 3 mois) que nous noterons $(X_t)_{t=1,\cdots,350}$. La série et sa transformation logarithmique sont représentées graphiquement en annexe figure 7, page 9. Graphiquement on observe une saisonnalité annuelle et une tendance à la hausse. On note également que notre série présente de l'hétéroscédasticité. Dans les parties suivantes, nous verrons comment rendre cette série stationnaire.

1.2 Stationnarisation de la série

La transformation logarithmique est pertinente car elle permet d'éviter certaines formes d'hétéroscédasticité ou de non-linéarité. Dans notre cas, l'héréroscédasticité était légère et la transformation logarithmique a permis de la corriger. Notons que nous aurions pu considérer d'autres transformations telles que celle de *Box-Cox*. Nous noterons $(L_t)_{t=1,\cdots,350}$ la série logarithmique, $L_t = \log X_t$, $\forall t \in \{1, \cdots, 350\}$ et sa représentation graphique est donnée figure 7.

Cette transformation ne suffit pas pour obtenir la stationnarité. Nous observons toujours une saisonnalité annuelle avec des pics de baisse d'activité pendant la saison d'été de chaque année, en particulier, l'espérance n'est pas constante. L'étude des autocorrélations et des autocorrélations partielles figure 8 et 9 nous permet de confirmer la présence d'une saisonnalité. Un motif semble se reproduire : décroissance puis croissance des autocorrélations totales. De plus, nous y retrouvons une très forte autocorrélation pour les ordres multiples de 12.

Nous menons deux tests sur notre série transformée L_t: un test de racine unitaire - le test augmenté de Dickey-Fuller (ADF) - et un test de stationnarité - celui de Kwiatkowski, Philips, Schmidt et Shin (KPSS), que nous détaillons en annexe 3.4. Pour choisir le nombre de retards des différences premières à intégrer dans la régression, nous procédons de la manière suivante: nous testons pour plusieurs retards $\in 1, 2, ...$ et nous nous arrêtons dès que les résidus passent les tests Portmanteau. Pour choisir la spécification adéquate, nous regardons les t-statistiques de chaque coefficient et comparons aux seuils de Dickey-Fuller pour déterminer si la tendance et/ou la constante sont significatives.

L'hypothèse nulle du test KPSS est "H_0 : le processus est stationnaire", autour d'une constante ou d'une tendance, selon la spécification la plus adéquate compte tenu de la série que l'on souhaite tester.

L'hypothèse nulle du test ADF est "H_0 : le processus a une racine unitaire", toujours selon la spécification choisie: sans constante, avec constante, avec constante et tendance avec constante, avec constante, tendance linéaire et tendance quadratique.

Les résultats obtenus sont donnés table 1. Nous ne rejetons pas l'hypothèse de racine unitaire pour la spécification la plus complète - avec une tendance quadratique. Le test KPSS vient confirmer cela puisque nous rejetons l'hypothèse nulle de stationnarité - pour une spécification avec une tendance linéaire uniquement.

Table 1: Test KPSS et ADF pour L_t

	KPSS Test	ADF Test
critical value (1%)	0.22	-4.41
Test Statistic	0.35	-2.1
p value	< 0.001	0.78
spécification	ct[1]	ctt[2]
lags	23	23

1.2.1 Différenciation 12ème

Afin de supprimer la saisonnalité du processus nous effectuons une différence 12ème et nous considérons la série $G_t = L_t - L_{t-12}, \forall t \in \{13, \cdots, 350\}$. Remarquons que nous perdons 12 observations en différenciant. Sa représentation graphique, ses autocorrélations et ses autocorrélations partielles sont données respectivement figure 10, 11 et 12. Les résultats des tests ADF et KPSS sur le série G_t sont donnés table 2.

Nous rejetons l'hypothèse de racine unitaire avec ADF, et nous ne pouvons pas rejeter l'hypothèse que la série est stationnaire à l'ordre de 5% avec KPSS. En revanche, la tendance linéaire est significative pour les deux tests. La différence 12ème n'est donc pas suffisante pour assurer la stationnarité. En outre, les autocorrélogrammes et autocorrélogrammes partiels ne sont pas encore très satisfaisants: il y a des autocorrélations significatives et encore très fortes à l'ordre 12 et des autocorrélations partielles significatives à l'ordre 12 et 24. Nous avons donc effectué une différence première en plus de la différence 12ème. Sa représentation graphique, ses autocorrélations et ses autocorrélations partielles se trouvent figure 13, 14 et 15. La série passe les tests de racine unitaire set de stationnarité (table 26 et 27), cependant, cette différence ne semble pas corriger les problèmes évoqués précédemment sur les ACF et PACF.

Table 2: Test KPSS et ADF pour G_t

	ADF Test	KPSS Test
critical value (1%)	-3.99	0.22
Test Statistic	-6.76	0.05
p value	< 0.001	0.54
spécification	ct	ct
lags	23	23

1.2.2 Désaisonnalisation par régression linéaire

Nous décidons d'opter pour une autre méthode de désaisonnalisation: la régression linéaire avec variables muettes. Pour cela, nous considérons le modèle de régression suivant :

$$L_{m+12(a-1)} = \alpha + \sum_{i=1}^{11} \beta_i \mathbb{1}_{m=i} + \epsilon_{m+12(a-1)}, \forall m \in \{1, ..., 12\}, \forall a \in \{1, ..., 30\}, m \in \{1, 2\} \text{ si } a = 30$$

[1] constante + tendance linéaire
[2] constante + tendance linéaire + tendance quadratique

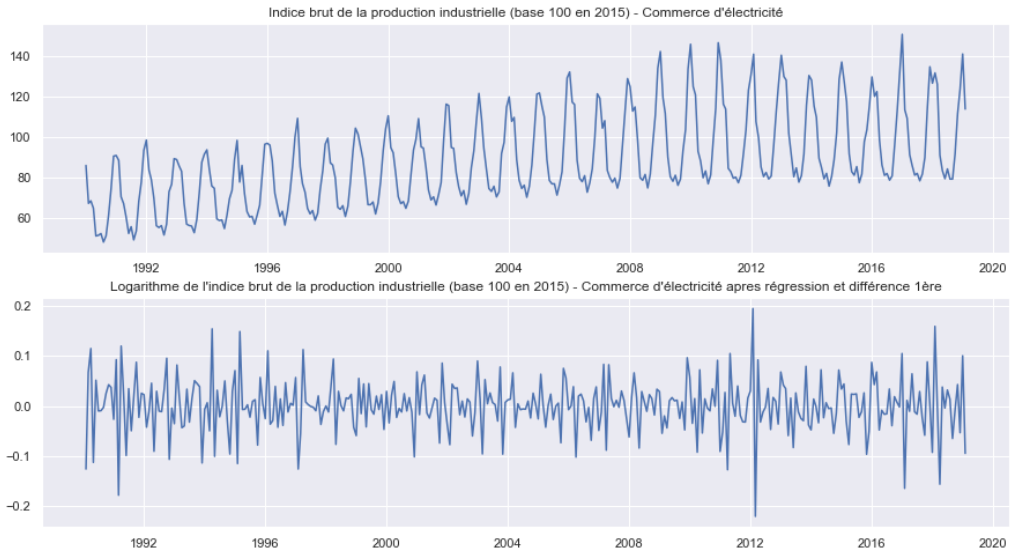
On obtient la série désaisonnalisée par l'extraction des résidus estimés de cette régression : $(\hat{\epsilon}_t)_{t=1,...,350}$. La représentation graphique, les autocorrélations totales et partielles de ces résidus estimés se trouvent figure 16, 17 et 18. La série obtenue est bien désaisonnalisée, néanmoins il reste encore une tendance. Pour la retirer, nous prenons simplement la différence première de la série $Y_t = \hat{\epsilon}_t - \hat{\epsilon}_{t-1}, \forall t \in \{2, ...350\}$. La représentation graphique de notre nouvelle série $(Y_t)_{t=1,...349)}$, ses autocorrélations et ses autocorrélations partielles se trouvent figure 19, 20 et 21. Cette fois-ci, les autocorrélogrammes sont plus satisfaisants que nos résultats précédents: pas d'autocorrélations totales et partielles significatives à de grands ordres, graphiquement notre série semble stationnaire. On le confirme à l'aide des tests ADF et KPSS:

	ADF Test	KPSS Test
critical value (1%)	-2.57	0.74
Test Statistic	-14.33	0.06
p value	< 0.001	0.82
spécification	no constant	no constant
lags	4	4

Nous choisissons donc cette série pour les parties suivantes.

1.3 Représentation graphique

Figure 1: Représentation graphique de l'indice mensuel de la production industrielle du commerce d'éléctricité (base 100 en 2015) de janvier 1990 à février 2019. En haut : la série telle qu'elle a été récupérée sur le site de l'Insee. En bas : sa transformation logarithmique avec régression linéaire et différence 1ère.



2 Partie 2: Modèles ARMA

Dans cette partie, nous comparons et choisissons un modèle ARMA pour notre série corrigée: $(Y_t)_{t=1,...,349}$. Nous nous inspirons pour cela de la méthodologie de Box et Jenkins que nous détaillons dans un premier temps. Nous choisissons à l'aide de l'étude des fonctions d'autocorrélations totales et partielles de la série corrigée, les ordres p^* et q^* maximum vraisemblables pour notre modèle ARMA. Nous choisissons ensuite le meilleur modèle ARMA parmi ceux étant valides et ayant un bon ajustement.

2.1 Méthodologie de Box et Jenkins

La méthodologie de Box et Jenkins est un outil très utilisé en série temporelle pour modéliser des modèles ARMA ou ARIMA.

Pour une série supposée stationnaire, telle que la notre $(Y_t)_{t=1,...,349}$, la détermination d'un modèle ARMA comporte 4 étapes:

- Identication à priori des ordres p et q (acf, pacf).
- Estimation des paramètres (mco, mv).
- Validation (tests de signicativité, graphe et (p)acf des résidus et tests portmanteau).
- Choix d'un modèle (AIC, BIC, MSE).

2.2 Choix des ordres p^* et q^* maximum vraisemblables

L'identification à l'aide des fonctions d'autocorrélations empiriques de l'ordre (p, q) pertinent pour un modèle ARMA est fondée sur une propriété connue des $MA(q)$ et $AR(p)$:

$$MA(q) \iff \rho(h) = 0, \forall h > q$$

avec $\rho(h)$, l'autocorrélation de la série à l'ordre h.

$$AR(p) \iff r(h) = 0, \forall h > p$$

avec $r(h)$, l'autocorrélation partielle de la série à l'ordre h.

Nous remettons à disposition les fonctions d'autocorrélogrammes totales et partielles empiriques ci-dessous:

Figure 2: Graphique des autocorrélations du logarithme de la série désaisonnalisée par régression linéaire et différence 1ère avec intervalle de confiance à 95% selon la formule de Bartlett pour les processus MA.

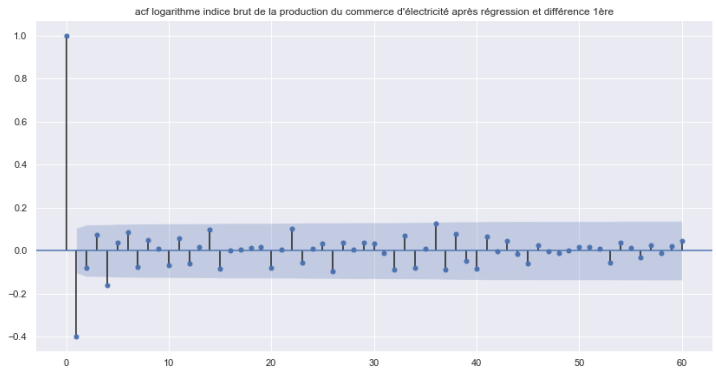
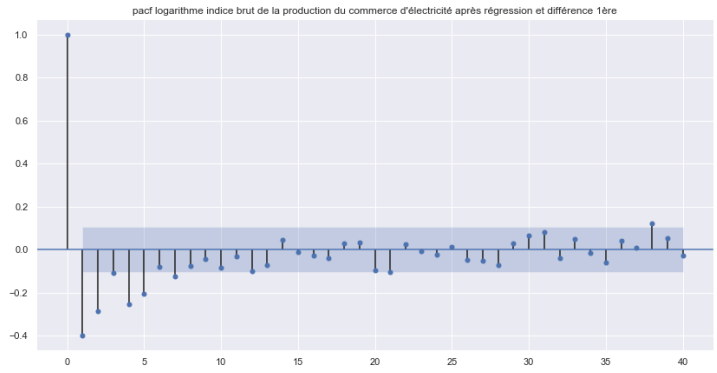


Figure 3: Graphique des autocorrélations partielles du logarithme de la série désaisonnalisée par régression linéaire et différence 1ère avec l'intervalle de confiance normal standard à 95% ($\pm\frac{1.96}{\sqrt{n}}$).



L'étude de l'autocorrélogramme (figure 2) nous permet de retenir que $q \in \{0, 1, 2, 3, 4\}$, puisque les autocorrélations empiriques sont non significativement différentes de 0 pour $h > 4$. De même, on retient que $p \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ puisque les autocorrélations partielles empiriques sont non significativement différentes de 0 à 5% pour $h > 7$.

2.3 Validation des modèles

Pour chaque modèle considéré, on étudie son ajustement et sa validité en regardant la significativité des coefficients estimés et la blancheur des résidus. On rappelle la définition d'un bruit blanc faible (ϵ_t):

- $\forall t \in T, E(\epsilon_t) = 0$
- $\forall t \in T, Var(\epsilon_t) = \sigma^2$
- $\forall t \neq s, (t, s) \in T^2, cov(\epsilon_t, \epsilon_s) = 0$

Pour l'autocovariance, on étudie les autocorrélogrammes empiriques et on conduit le test de Portmanteau modifié, décrit en annexe 3.4.

Pour chaque modèle considéré, soit $Card(\{0, 1, 2, 3, 4\} \times \{0, 1, 2, 3, 4, 5, 6, 7\}) = 40$ spécifications au total, on suit la procédure suivante:

- On élimine tous les modèles pour lesquels on a rejeté H_0 (à 5%) pour un lag compris entre 0 et 24 dans le test de Ljung-Box.
- On élimine tous les modèles pour lesquels le dernier coefficient d'une des composantes AR et/ou MA est non significativement différent de 0 à 5%.
- On élimine tous les modèles ayant un autocorrélogramme indiquant un possible caractère non bruit blanc des résidus estimés. Plus précisément, on élimine un modèle dès qu'une des autocorrélations de ses résidus estimés est non significativement différente de 0 à 5% (en prenant toujours un lag de 24).

2.4 Sélection

Pour sélectionner parmi les modèles vérifiant les critères de la procédure énoncée précédemment. Nous avons recours aux critères d'information de l'AIC et du BIC ainsi que le score MSE, que nous détaillons en annexes 3.4. Seuls six modèles passent les trois critères de sélection, donnés 4.

	(p,q)	AIC	BIC	MSE
0	(2, 2)	-1177.641690	**-1154.511258**	0.004773
1	(2, 3)	-1176.177085	-1149.191582	0.004842
2	(3, 3)	-1177.723612	-1146.883037	0.004172
3	(4, 1)	-1180.929642	-1153.944138	0.004611
4	(4, 4)	**-1181.339093**	-1142.788373	**0.004092**
5	(7, 0)	-1169.640823	-1134.945176	0.004242

Figure 4: Liste des modèles ARMA vérifiant les trois critères de validation. En gras, les valeurs minimums.

Les figures 28, 29 et 30 donnent respectivement la liste de tous les modèles testés, celle des modèles respectant le premier critère uniquement et celle respectant les deux premiers critères de sélection.

Finalement, on retient la spécification ARMA(4,4) qui vérifie les trois critères de validité et qui minimise à la fois le critère de l'AIC et le score MSE.

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.0013	0.001	2.310	0.021	0.000	0.002
ar.L1.Electricite	0.3690	0.406	0.909	0.364	-0.427	1.165
ar.L2.Electricite	0.2886	0.332	0.870	0.385	-0.361	0.938
ar.L3.Electricite	0.5939	0.192	3.095	0.002	0.218	0.970
ar.L4.Electricite	-0.2815	0.070	-4.011	0.000	-0.419	-0.144
ma.L1.Electricite	-1.0856	0.430	-2.527	0.012	-1.928	-0.243
ma.L2.Electricite	-0.1430	0.682	-0.210	0.834	-1.480	1.194
ma.L3.Electricite	-0.3420	0.277	-1.235	0.218	-0.885	0.201
ma.L4.Electricite	0.5785	0.197	2.935	0.004	0.192	0.965

Figure 5: Résultats de la spécification ARMA(4,4)

Les résidus sont représentés figure 22, leur autocorrélations figure 23 et les p valeurs du test de portemanteau modifié figure 24.

Nous étudions la normalité de ces résidus en calculant la statistique de test de Jarque-Bera dont l'hypothèse nulle est que les données considérées suivent une loi normale, la p-value étant de 0.01, on peut rejeter à 5% la normalité des résidus. Le QQ-plot des résidus est présenté figure 25.

3 Partie 3: Prévisions

Dans cette partie les résidus de notre modèle sont supposés être un bruit gaussien de variance σ^2.

3.1 Région de confiance

Soit X_t la solution de l'équation du modèle ARMA(p,q) déterminé dans la partie précédente,

$$\Phi(B)X_t = \Psi(B)\epsilon_t$$

qu'on peut réécrire,

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t - \sum_{i=1}^{q} \psi_i \epsilon_{t-i}$$

On notera ϵ_t^* l'innovation linéaire de X_t,

$$\epsilon_t^* = X_t - \mathbb{EL}[X_t | X_{t-1}, \cdots]$$

Notons que $\epsilon_t = \epsilon_t^*$ si X_t est la solution d'un ARMA canonique. Il est donc essentiel de faire ici l'hypothèse que le modèle ARMA est canonique (il est causale, inversible et $\Phi(B)$ et $\Psi(B)$ n'ont pas de racine commune). En effet, si le modèle est canonique, les résidus sont l'innovation linéaire et sont donc orthogonaux à toute fonction du passé linéaire.

On notera $\hat{X}_{t+1|t}$ et $\hat{X}_{t+2|t}$ les prévisions linéaires à l'horizon 1 et 2 de X_t i.e $\hat{X}_{t+1|t} = \mathbb{EL}[X_{t+1}|X_t, \cdots]$ et $\hat{X}_{t+2|t} = \mathbb{EL}[X_{t+2}|X_t, \cdots]$. On notera $\mathcal{H}(t)$ le passé linéaire à la date t. Comme ϵ_t est l'innovation linéaire, on a par définition,

$$\hat{X}_{t+1|t} = X_{t+1} - \epsilon_{t+1}$$

Pour $\hat{X}_{t+2|t}$, ϵ_{t+2} est orthogonale à $\mathcal{H}(t+1)$, donc à $\mathcal{H}(t)$, ϵ_{t+1} orthogonale à $\mathcal{H}(t)$, et pour tout $i \geq 2$ $\epsilon_{t+2-i} = \frac{\Psi(B)}{\Phi(B)} X_{t+2-i} \in \mathcal{H}(t)$ donc,

$$\hat{X}_{t+2|t} = \sum_{i=2}^{p} \phi_i X_{t+2-i} + \phi_1 \hat{X}_{t+1|t} - \sum_{i=2}^{q} \psi^i \epsilon_{t+2-i} = X_{t+2} - \phi_1 X_{t+1} + \phi_1 \hat{X}_{t+1|t} - \epsilon_{t+2} + \psi_1 \epsilon_{t+1}$$

$$= X_{t+2} - \phi_1 \epsilon_{t+1} - \epsilon_{t+2} + \psi_1 \epsilon_{t+1} = X_{t+2} + (\psi_1 - \phi_1)\epsilon_{t+1} - \epsilon_{t+2}$$

On note Y_t l'erreur théorique des prédictions à l'ordre 1 et 2 au temps t,

$$Y_t = \begin{pmatrix} X_{t+1} - \hat{X}_{t+1|t} \\ X_{t+2} - \hat{X}_{t+2|t} \end{pmatrix} = \begin{pmatrix} \epsilon_{t+1} \\ (\phi_1 - \psi_1)\epsilon_{t+1} + \epsilon_{t+2} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \sigma^2 & (\phi_1 - \psi_1)\sigma^2 \\ (\phi_1 - \psi_1)\sigma^2 & ((\phi_1 - \psi_1)^2 + 1)\sigma^2 \end{pmatrix} \right]$$

On note Σ la matrice de variance-covariance, son déterminant vaut $\sigma^4 > 0$ elle est donc inversible et on a immédiatement que,

$$Y_t' \Sigma^{-1} Y_t \sim \chi^2(2)$$

La région de confiance de niveau $1-\alpha$ est donc donnée par,

$$RC(1-\alpha) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : \begin{pmatrix} x - \hat{X}_{t+1|t} \\ y - \hat{X}_{t+2|t} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} x - \hat{X}_{t+1|t} \\ y - \hat{X}_{t+2|t} \end{pmatrix} \leq q_{1-\alpha}^{\chi^2(2)} \right\}$$

où $q_{1-\alpha}^{\chi^2(2)}$ est le quantile $1-\alpha$ du loi du chi 2 à deux degrés de liberté. Remarquons que la région de confiance est une ellipse centrée en

3.2 Hypothèses

Comme expliqué dans la question précédente, cette région de confiance repose sur deux hypothèses: l'hypothèse de gaussianité des résidus et l'hypothèse que le modèle est sous forme canonique.

Nous avons vu partie précédente que notre série ne satisfait pas la 1ère hypothèse et la région de confiance obtenue ne s'appliquera donc pas nécessairement à nos données. Les estimations des racines du modèle ARMA(4,4) et leur module sont données table 3.4, il n'y a pas de racine commune et elles sont toutes de module supérieur à 1, le modèle est donc sous sa forme canonique.

3.3 Graphique, application à notre série

Pour construire cette région de confiance empirique, nous récupérons les estimateurs $(\hat{\phi}_i)_{i=1,...,4}$ et $(\hat{\psi}_i)_{i=1,...,4}$ disponibles dans la table 5. Nous estimons les résidus de la régression $(\hat{\epsilon}_t)$ et nous nous en servons pour calculer un estimateur sans biais de σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\epsilon}_t^2}{N-9}$$

Nous nous en servons ensuite afin de calculer les prédictions $\hat{X}_{t+1|t}$ et $\hat{X}_{t+2|t}$ et la matrice $\hat{\Sigma}$ suivante :

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}^2 & (\hat{\phi}_1 - \hat{\psi}_1)\hat{\sigma}^2 \\ (\hat{\phi}_1 - \hat{\psi}_1)\hat{\sigma}^2 & ((\hat{\phi}_1 - \hat{\psi}_1)^2 + 1)\hat{\sigma}^2 \end{pmatrix}$$

Nos estimateurs étant convergents, nous pouvons définir la région de confiance suivante qui restera valide asymptotiquement (par le théorème de Slutsky):

$$\hat{RC}(1-\alpha) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : \begin{pmatrix} x - \hat{X}_{t+1|t} \\ y - \hat{X}_{t+2|t} \end{pmatrix}' \hat{\Sigma}^{-1} \begin{pmatrix} x - \hat{X}_{t+1|t} \\ y - \hat{X}_{t+2|t} \end{pmatrix} \leq q_{1-\alpha}^{\chi^2(2)} \right\}$$

La région de confiance à l'ordre de 95% est donnée figure 6. Il s'agit de la région de confiance pour la prédiction de janvier et février 2019. Leur vraie valeurs sont respectivement 0.1 et -0.09. Elles tombent très loin de la région de confiance, ce n'est pas étonnant que les résidus ne satisfont pas l'hypothèse de gaussianité.

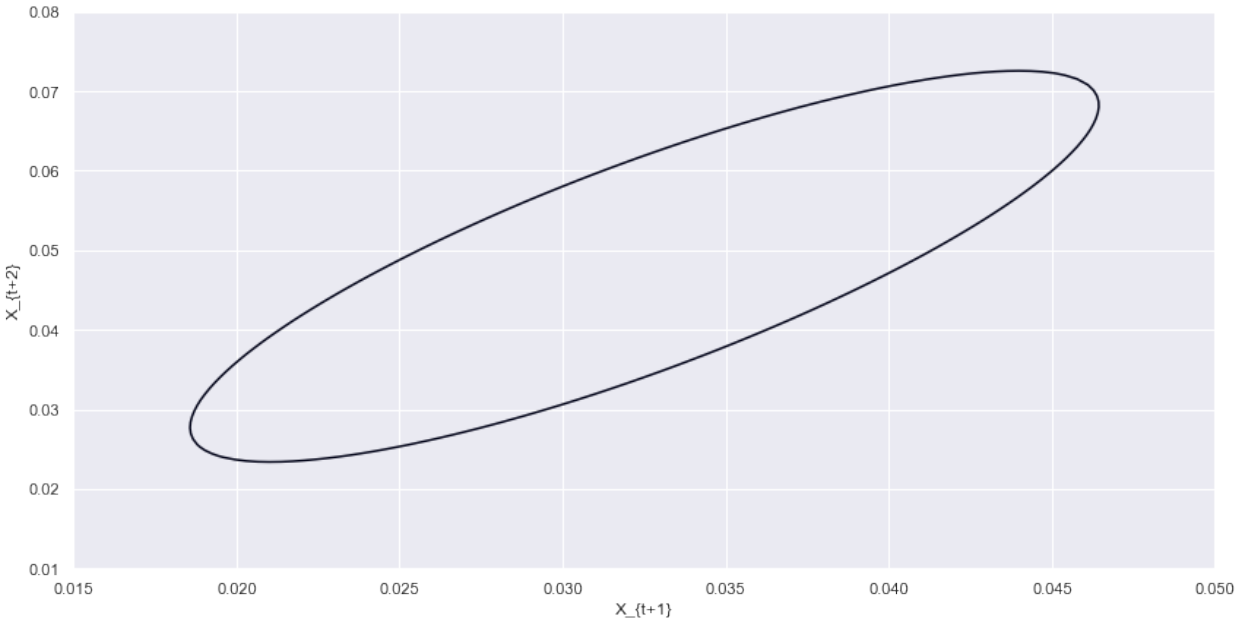
3.4 Question ouverte

Y_{T+1} sera utile pour prévoir X_{T+1} à la date T si (Y_t) cause instantanément (X_t) au sens de Granger donc si et seulement si

$$\hat{X}_{t+1|\{X_u, Y_u, u \leq t\} \cup \{Y_{t+1}\}} \neq \hat{X}_{t+1|\{X_u, Y_u, u \leq t\}}$$

On sait que d'après la cours (chapitre 5) que c'est équivalent à ce que les résidus des deux séries soient corrélés entre eux. C'est donc une relation symétrique, si Y_t peut aider à prédire X_t instantanément, X_t peut aider à prédire Y_t instantanément.

Figure 6: Région de confiance à 95%



Annexes

Figures

Figure 7: Représentation graphique de l'indice mensuel de la production industrielle du commerce d'électricité (base 100 en 2015) de janvier 1990 à février 2019. En haut : la série telle qu'elle a été récupérée sur le site de l'Insee. En bas : sa transformation logarithmique. (retour)

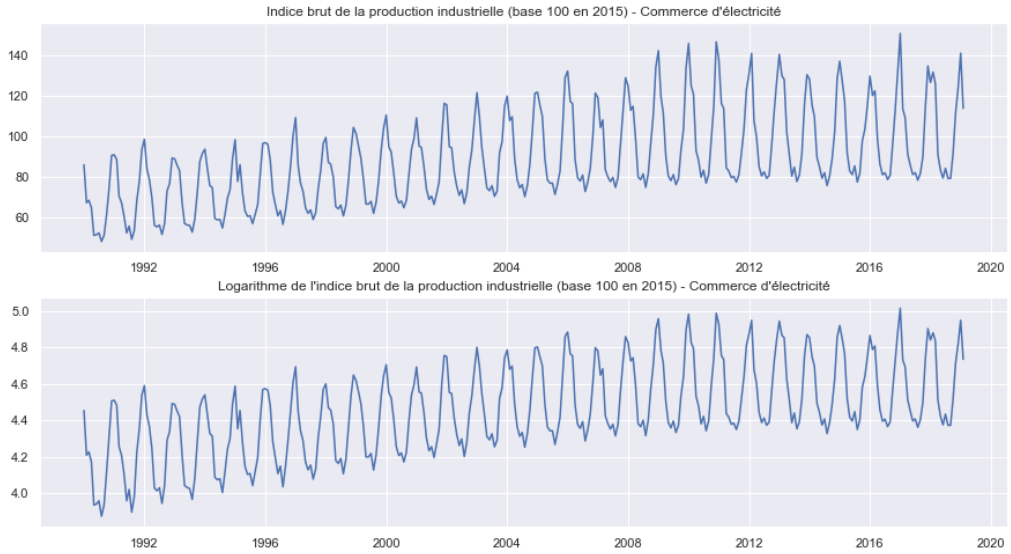


Figure 8: Graphique des autocorrélations du logarithme de la série avec intervalle de confiance à 95% selon la formule de Bartlett pour les processus MA. (retour)

Figure 9: Graphique des autocorrélations partielles du logarithme de la série avec l'intervalle de confiance normal standard à 95% ($\pm\frac{1.96}{\sqrt{n}}$). (retour)

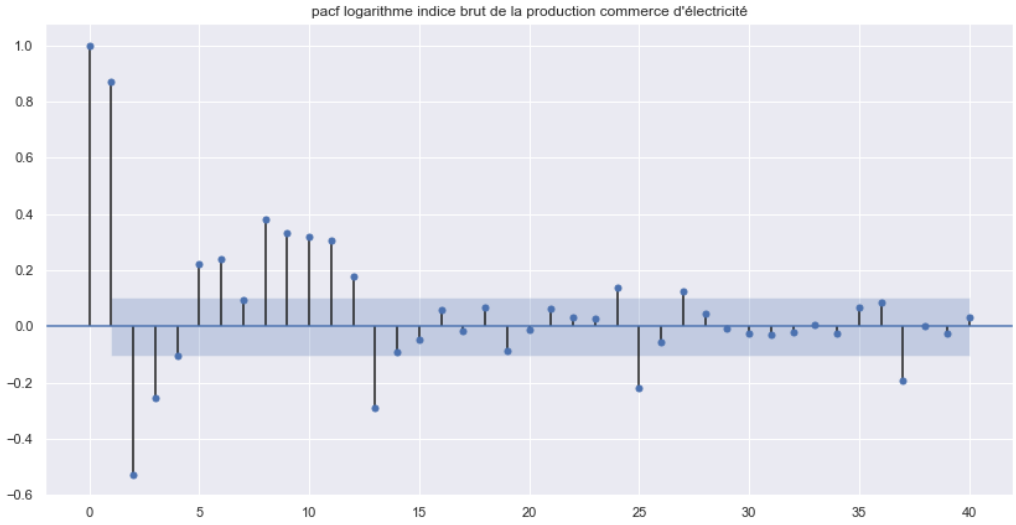


Figure 10: Représentation graphique du logarithme de la série désaisonnalisée par différence 12ème (retour)

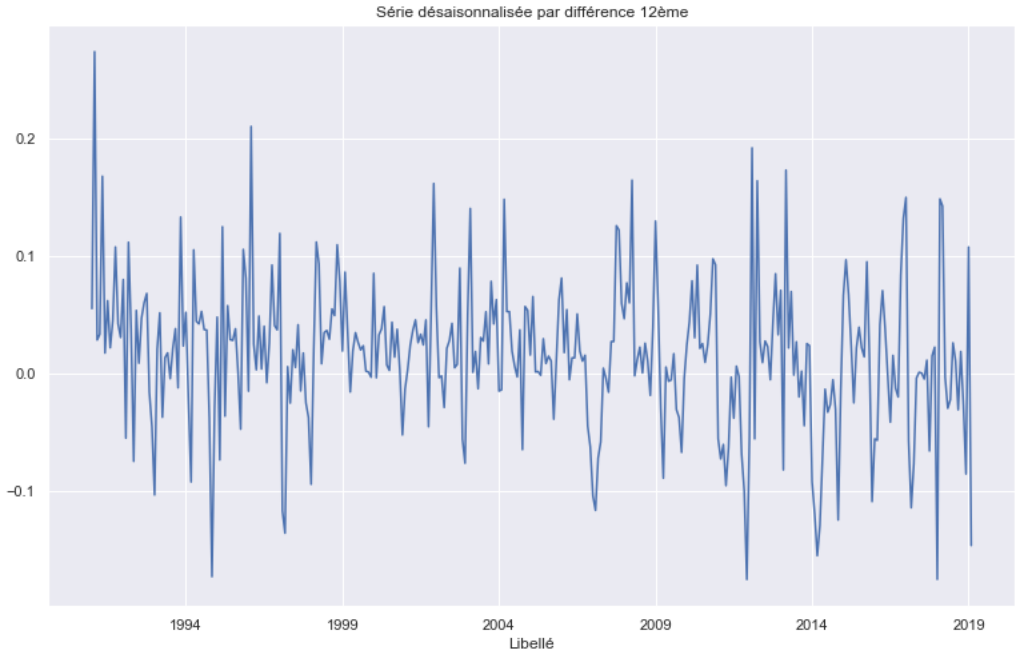


Figure 11: Graphique des autocorrélations du logarithme de la série différenciée à l'ordre 12 avec intervalle de confiance à 95% selon la formule de Bartlett pour les processus MA. (retour)

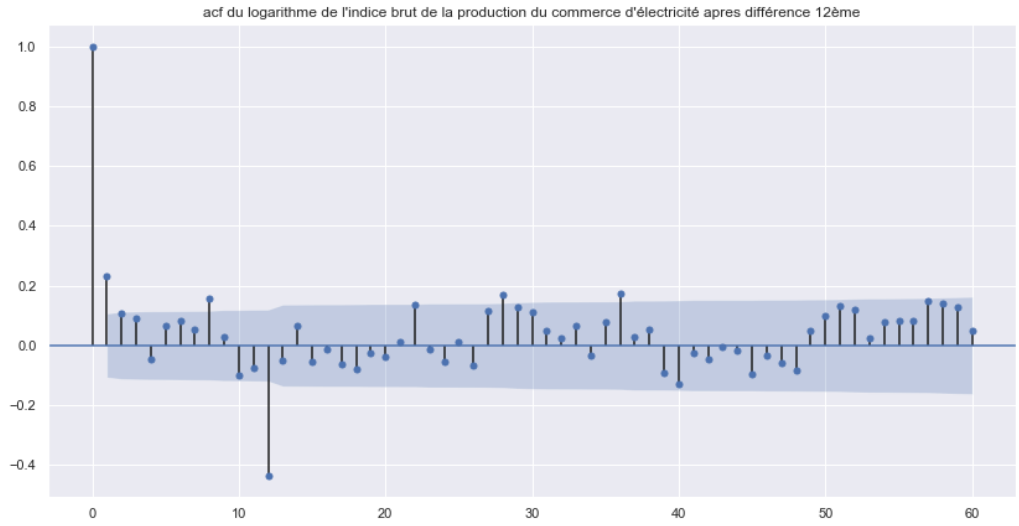


Figure 12: Graphique des autocorrélations partielles du logarithme de la série différenciée à l'ordre 12 avec l'intervalle de confiance normal standard à 95% ($\pm\frac{1.96}{\sqrt{n}}$). (retour)

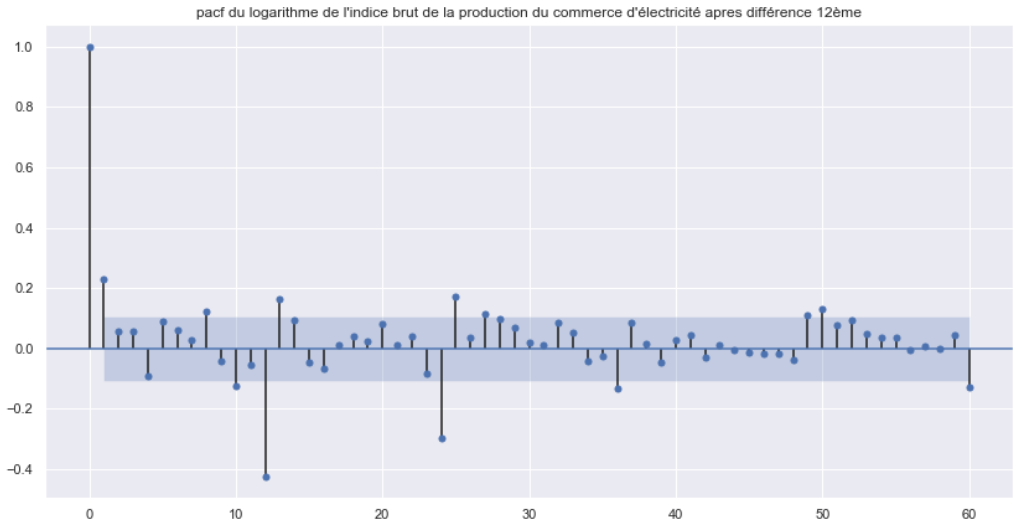


Figure 13: Représentation graphique du logarithme de la série désaisonnalisée par différence 12ème et 1ère (retour)

Figure 14: Graphique des autocorrélations du logarithme de la série différenciée à l'ordre 12 et à l'ordre 1 avec intervalle de confiance à 95% selon la formule de Bartlett pour les processus MA. (retour)

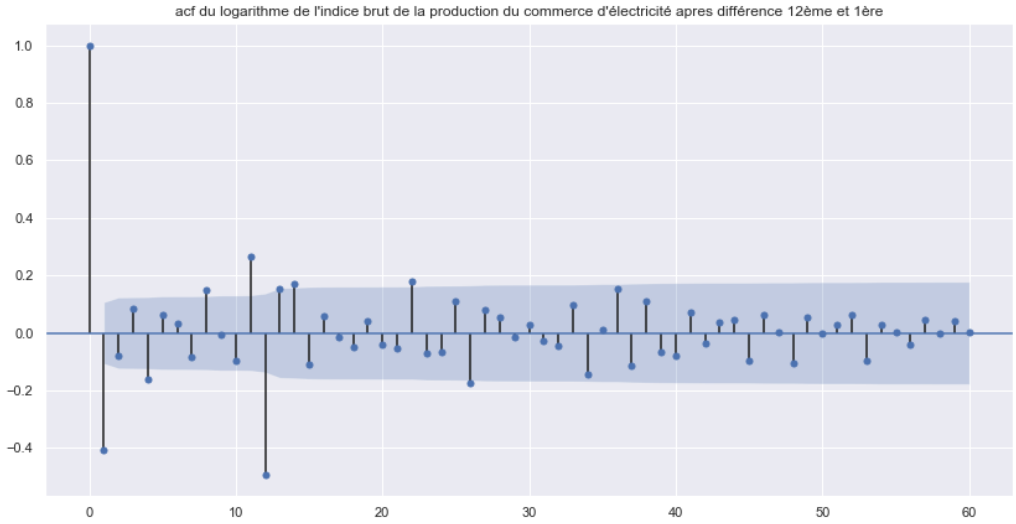


Figure 15: Graphique des autocorrélations partielles du logarithme de la série différenciée à l'ordre 12 et à l'ordre 1 avec l'intervalle de confiance normal standard à 95% ($\pm\frac{1.96}{\sqrt{n}}$). (retour)

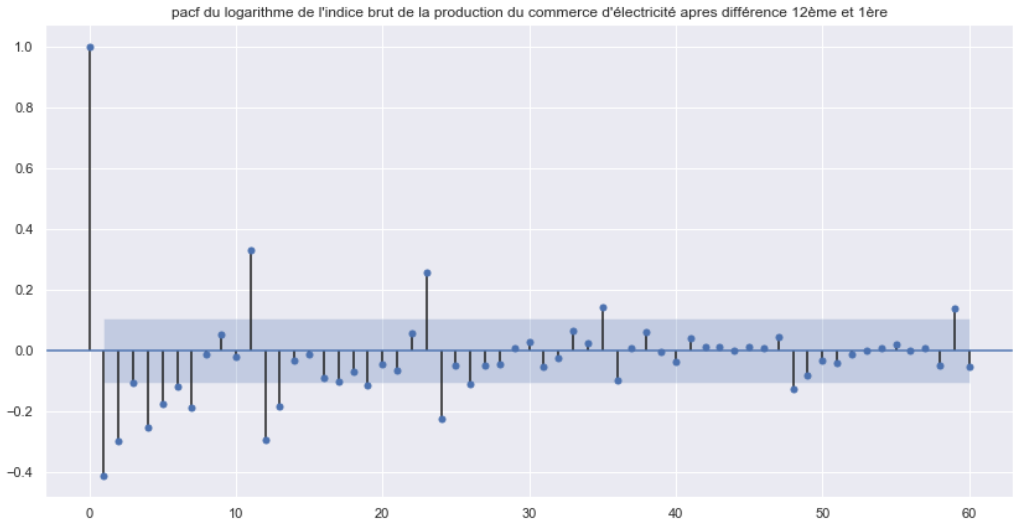


Figure 16: Représentation graphique du logarithme de la série désaisonnalisée par régression linéaire (retour)

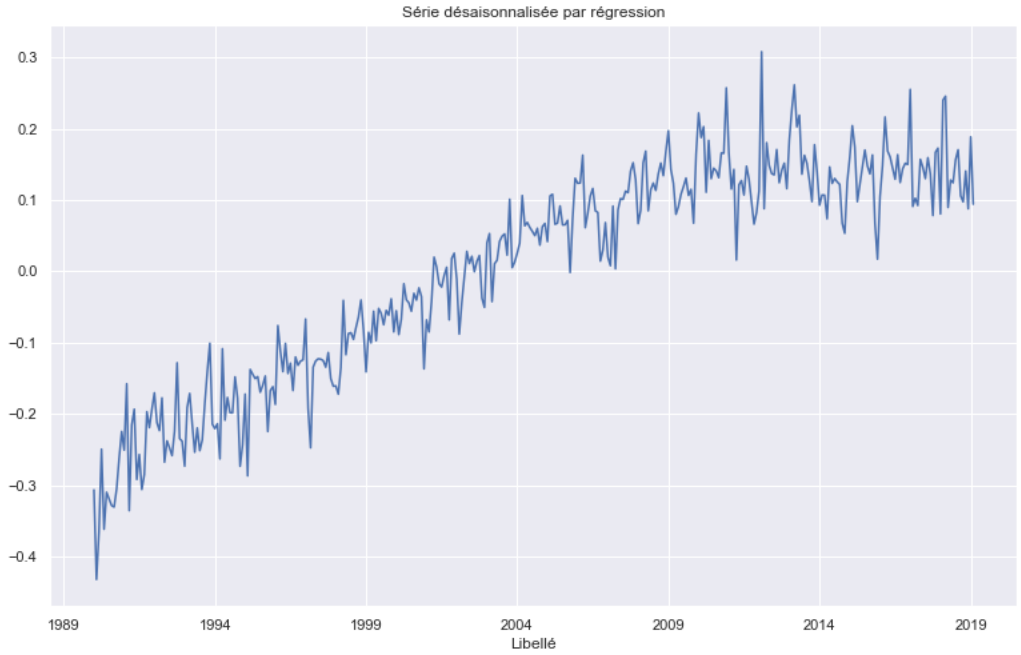


Figure 17: Graphique des autocorrélations du logarithme de la série désaisonnalisée par régression linéaire avec intervalle de confiance à 95% selon la formule de Bartlett pour les processus MA. (retour)

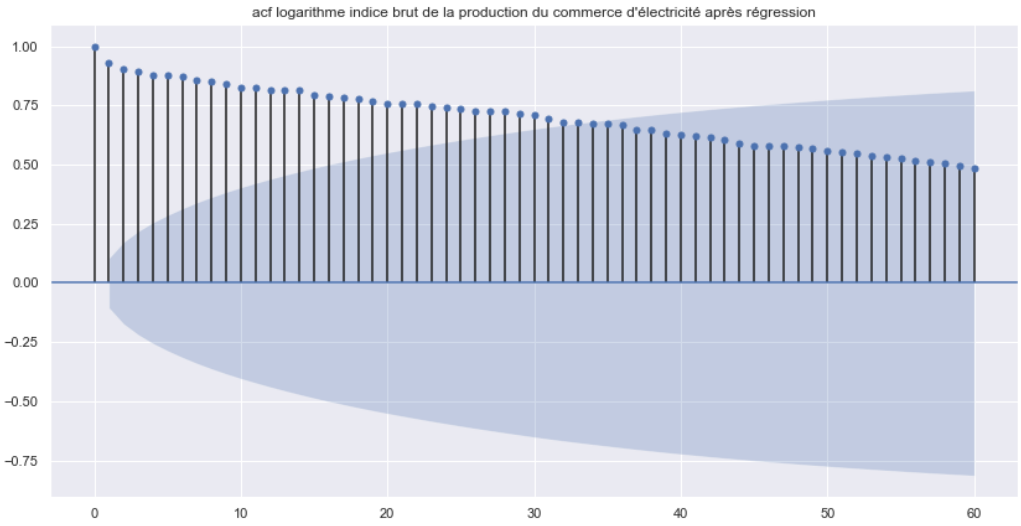


Figure 18: Graphique des autocorrélations partielles du logarithme de la série désaisonnalisée par régression linéaire avec l'intervalle de confiance normal standard à 95% ($\pm\frac{1.96}{\sqrt{n}}$). (retour)

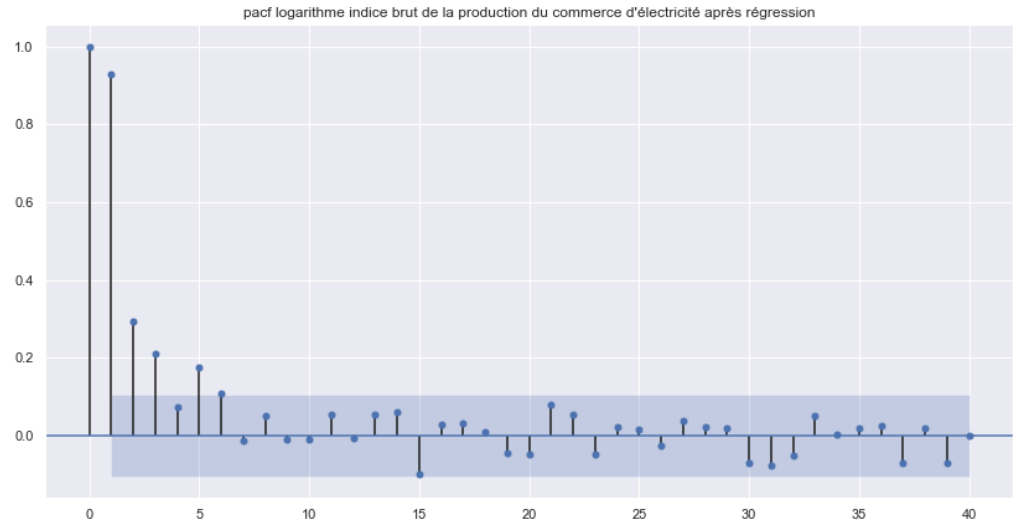


Figure 19: Représentation graphique du logarithme de la série désaisonnalisée par régression linéaire et différence 1ère (retour)

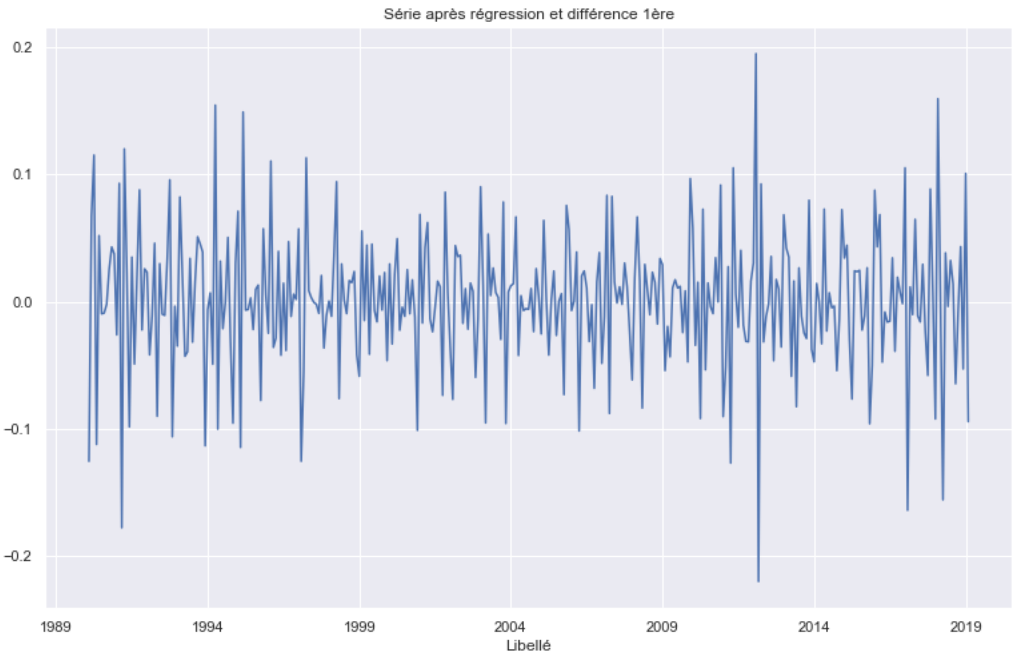


Figure 20: Graphique des autocorrélations du logarithme de la série désaisonnalisée par régression linéaire et différence 1ère avec intervalle de confiance à 95% selon la formule de Bartlett pour les processus MA. (retour)

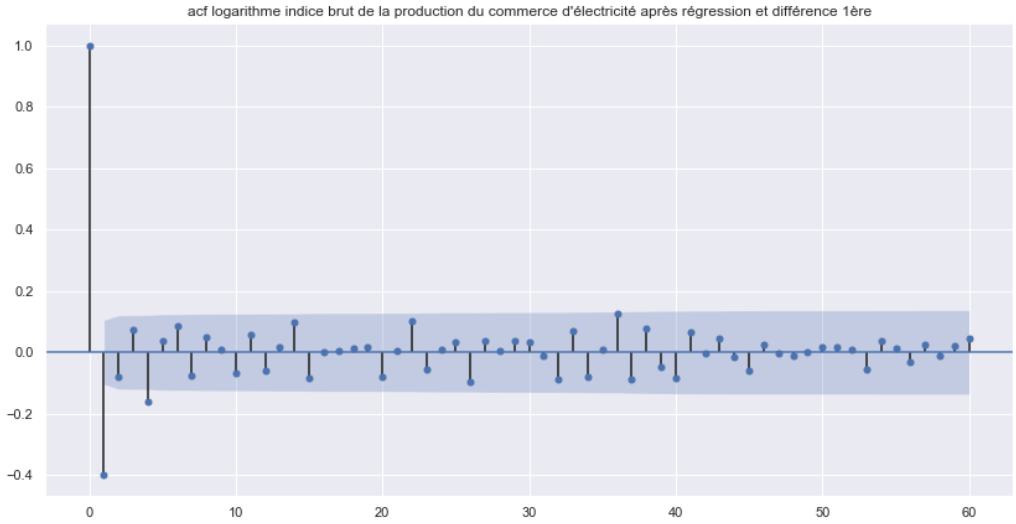


Figure 21: Graphique des autocorrélations partielles du logarithme de la série désaisonnalisée par régression linéaire et différence 1ère avec l'intervalle de confiance normal standard à 95% ($\pm\frac{1.96}{\sqrt{n}}$). (retour)

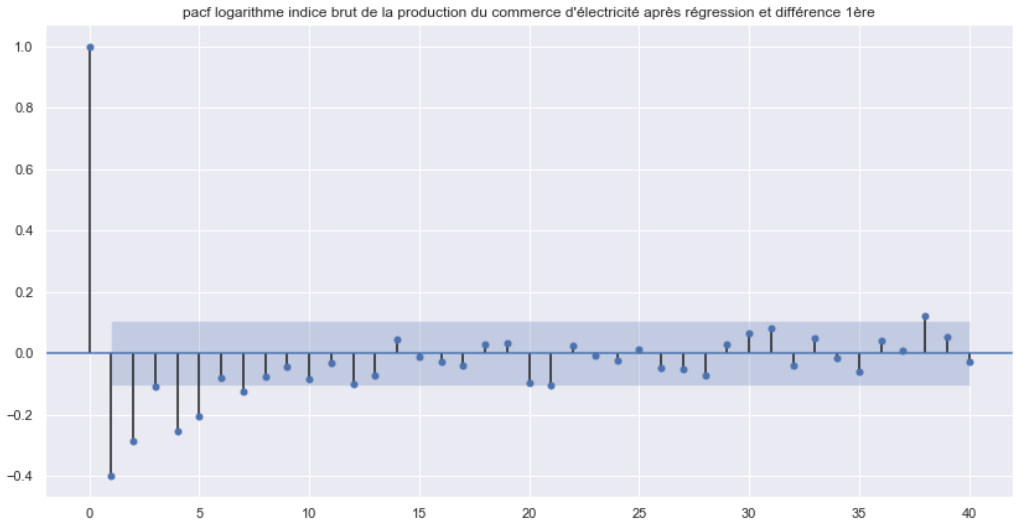


Figure 22: Graphique des résidus de la spécification ARMA(4,4) (retour)

Figure 23: Graphique des autocorrélations des résidus de la spécification ARMA(4,4) (retour)

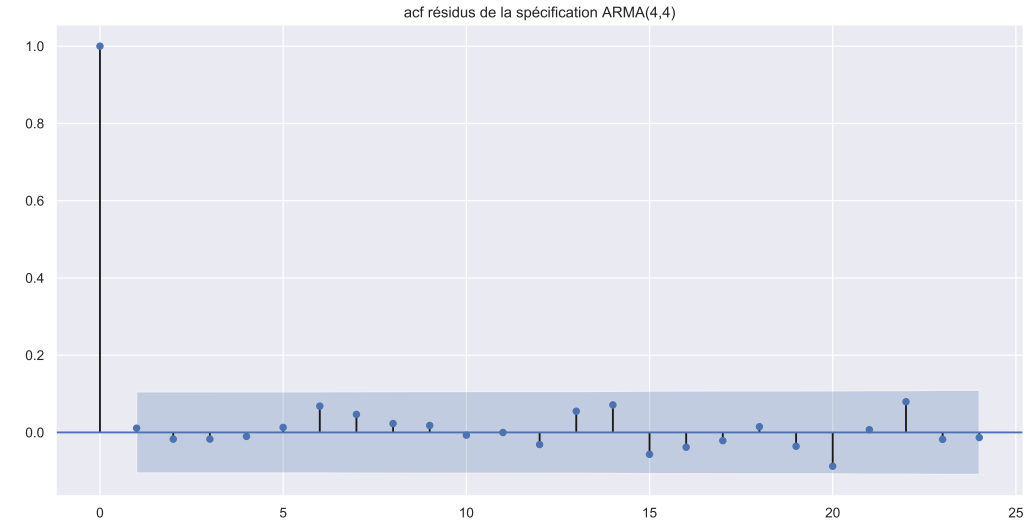


Figure 24: Résultat du test de portmanteau modifié sur les résidus de la spécification ARMA(4,4) (retour)

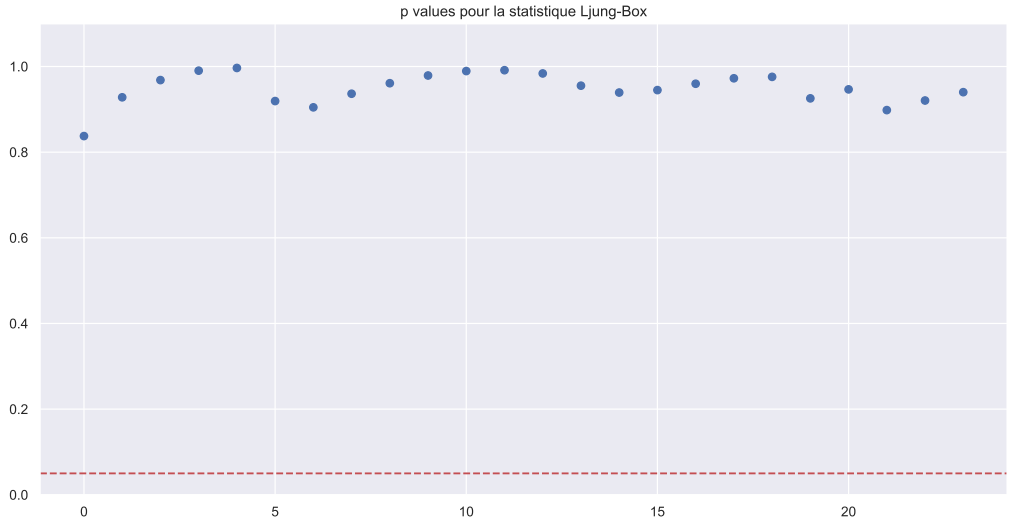
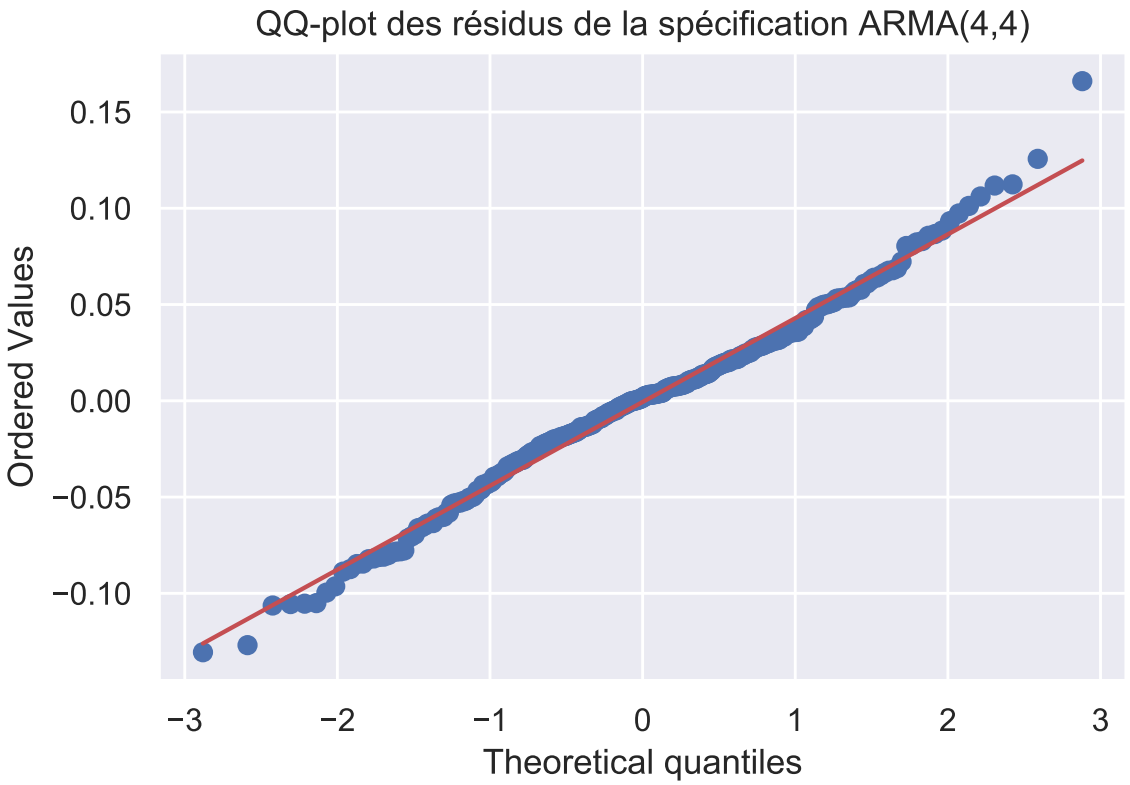


Figure 25: QQ-plot des résidus de la spécification ARMA(4,4) (retour)



Résultats des tests ADF et KPSS

Figure 26: Résultats du test ADF pour la série en différence 12ème et 1ère

	ADF Test
critical value (1%)	-2.57
Test Statistic	-6.22
p value	< 0.001
spécification	nc
lags	23

Figure 27: Résultats du test KPSS pour la série en différence 12ème et 1ère

	KPSS Test
critical value (1%)	0.74
Test Statistic	0.07
p value	0.73
spécification	nc
lags	23

Test ADF

Le test de Dickey-Fuller augmenté est basé sur le test de Dickey-Fuller simple développé par les statisticiens David Dickey et Wayne Fuller en 1979. Les conditions du test sont les suivantes:

On considère un modèle AR(1):

$$X_t = \rho X_{t-1} + \epsilon_t$$

où $(\epsilon_t)_{t=1,\dots,N}$ est un bruit blanc fort[3]. Les hypothèses du test sont les suivantes:

$$H_0 : \rho = 1 \text{ (présence d'une racine unitaire)}$$
$$H_1 : \rho \neq 1 \text{ (absence de racine unitaire)}$$

ou encore:

$$H_0 : \delta = 0$$
$$H_1 : \delta \neq 0$$

dans le modèle équivalent :

$$\Delta X_t = \delta X_{t-1} + u_t,\ \Delta X_t = X_t - X_{t-1},\ \delta = \rho - 1$$

Dickey et Fuller ont montré que sous l'hypothèse nulle, le paramètre $\hat{\rho}_n$ est super-consistent, c'est-à-dire qu'il converge vers 1 à une vitesse supérieure à $\sqrt{n}$ et ont obtenu que la statistique de test $t_n = \dfrac{\hat{\rho}_n - 1}{\sigma_{\hat{\rho}_n}} \sim$ loi de Dickey-Fuller, tabulée par les deux statisticiens.

Trois spécifications sont généralement considérées pour ce test :

$$\Delta X_t = \delta X_{t-1} + u_t$$

[3]Un bruit blanc fort est une suite $(\epsilon_t)_{t=1,\dots,N}$ de variables indépendantes et identiquement distribuées (iid), centrées et de variance finie σ^2 (ou de vecteurs iid $(0, \Sigma)$ en multivarié).

$$\Delta X_t = \alpha + \delta X_{t-1} + u_t$$

$$\Delta X_t = \alpha + \beta t + \delta X_{t-1} + u_t$$

La première spécification est celle que nous avons présentée au début, elle correspond à une spécification sans constante. La deuxième inclut une constante et la troisième : une constante et un trend. Le choix de la spécification est important. En effet, oublier d'inclure une constante et/ou un trend peut biaiser les paramètres et fausser les résultats de l'inférence. Les inclure, inutilement, peut en revanche impacter la puissance du test.

Un modèle AR(1) est souvent trop simple : les résidus ne se comportent pas forcément comme des bruits blancs. Il faut alors considérer des modèles AR(p) avec $p > 1$, où l'ajout de valeurs retardées permet de blanchir les résidus. C'est pour cela que l'on a recours au test Dickey-Fuller augmenté.

Le modèle considéré devient :

$$\Delta X_t = \delta X_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta X_{t-i} + v_t$$

On teste de nouveau la présence d'une racine unitaire via la nullité du coefficient δ. Il suffit pour cela de calculer la statistique de test et de comparer aux valeurs critiques de la loi de Dickey-Fuller. Encore une fois, il faut choisir la spécification adéquate pour la série considérée. Pour choisir le nombre de retards à considérer, plusieurs méthodes peuvent être adoptées : sélection du modèle avec le meilleur AIC, sélection du premier modèle dont les résidus passent les tests de blancheur, etc.

Test KPSS

Contrairement aux tests ADF et PP, le test KPSS permet de tester directement la stationnarité sans tester la présence de racine unitaire. Ce test nous permet de valider les transformations faites partie 1 pour stationnariser nos données.

Nous considérons que le modèle est :

$$Y_t = \xi t + r_t + \epsilon_t, r_t = r_{t-1} + u_t$$

avec $u_t \overset{i.i.d}{\sim} (0, \sigma_u^2)$ bruit blanc fort. Si $\xi = 0$, il n'y a pas de tendance déterministe et r_0 sert de constante.

La statistique de test LM est donnée par :

$$LM = \frac{\sum_{k=1}^{n} S_k^2}{n^2 \sigma_\epsilon^2}$$

où $S_k = \sum_{i=1}^{k} e_i$, $\sigma_\epsilon^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$, e_i sont les résidus de la régression de Y_t sur une constante et/ou un trend. Les auteurs ont montré que la loi asymptotique de cette statistique reste inchangée lorsque $(\epsilon_t)_{t=1,\dots,N}$ satisfait les hypothèses utilisées dans les tests de Perron-Philips (qui utilisent un terme d'erreur très général), à condition de prendre un estimateur convergent pour la variance de long terme.

Test Portmanteau modifié

Ce test introduit par Ljung-Box en 1978 permet de tester l'hypothèse nulle H_0 "les auto-corrélations sont nulles jusqu'à l'ordre H, avec H fixé". Il nous sert à tester la blancheur de nos résidus après avoir modélisé un SARIMA et permet de valider nos modèles.

$$\begin{cases} H_0 : \rho_\epsilon(h) = 0, \forall h \in \{0, \cdots, H\} \\ H_1 : \exists h \in \{0, \cdots, H\}, \rho_\epsilon(h) \neq 0 \end{cases}$$

La statistique de test s'écrit :

$$Q = n(n+2) \sum_{h=1}^{H} \frac{1}{n-h} \hat{\rho}_\epsilon^2(h)$$

Sous H_0 la statistique de test suit une loi du χ^2 de paramètre H moins le nombre de paramètres dans le modèle. Dans le cas d'un SARIMA$_s[(p,d,q),(P,D,Q)]$ la région de rejet de niveau α est donc donnée par

$$R(\alpha) = Q > \chi^2_{1-\alpha, H-P-Q-p-q}$$

où $\chi^2_{1-\alpha,m}$ est le quantile $1-\alpha$ d'une loi du χ^2 de degrés m.

Critères d'information: AIC et BIC

Les critères d'information AIC et BIC sont des mesures de la qualité d'un modèle statistique et permettent de faire de la sélection de modèle: on choisit parmi des modèles emboîtés, celui qui minimise le critère d'information qui nous intéresse. Il s'appliquent tous les deux à des modèles estimés par une méthode de maximum de vraisemblance.

Le critère AIC proposée par Hirotugu Akaike en 1973 est défini par:

$$AIC = -2logL + 2k \tag{1}$$

où L désigne la vraisemblance maximisée et k, le nombre de paramètres dans le modèle considéré.

Le critère BIC est un critère d'information Bayésien proposé par Gideon Schwarz en 1978, il est défini par:

$$BIC = -2logL + k * log(n) \tag{2}$$

Il est plus parcimonieux que le critère AIC puisqu'il pénalise beaucoup plus le nombre de paramètres dans le modèle.

Cependant, les objectifs des deux critères semblent différents. En effet, selon Ripley (2003)[4], le critère de l'AIC a été introduit pour retenir des variables pertinentes lors de prévisions, tandis que le critère du BIC vise la sélection de variables significative dans un modèle statistique.

[4] http://www.stats.ox.ac.uk/~ripley/Nelder80.pdf

Score MSE

La MSE (Mean Squared Error) ou erreur quadratique moyenne d'un estimateur $\hat{\theta}$ d'un paramètre θ est une mesure de précision de l'estimateur. Plus l'erreur quadratique moyenne est faible, plus l'estimateur est "bon". On la note:

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Nous nous servons de ce score pour évaluer la qualité de nos prédictions. Pour le calculer, nous divisons notre échantillon en un échantillon d'apprentissage (qui comprend $N - k$ observations) et un échantillon de test (qui comprends k observations) avec $k \in \mathbb{N}$, petit et choisi arbitrairement.

Une fois notre modèle entraîné sur l'échantillon d'apprentissage, nous comparons nos prédictions sur l'échantillon de test avec les vraies valeurs de cet échantillon et calculons le score MSE (empirique) associé:

$$MSE = \frac{1}{k} \sum_{t=N-k+1}^{k} (\hat{X}_{t|N-k} - X_t)^2$$

Choix de modèles ARMA

	(p,q)	AIC	BIC	MSE
0	(0, 0)	-1042.137421	-1034.427277	0.004712
1	(0, 1)	-1171.858426	-1160.293211	0.004452
2	(0, 2)	-1179.292567	-1163.872280	0.004522
3	(0, 3)	-1177.481335	-1158.205975	0.004538
4	(1, 0)	-1101.850659	-1090.285444	0.004636
5	(1, 1)	-1179.678125	**-1164.257837**	0.004543
6	(1, 2)	-1177.705787	-1158.430427	0.004544
7	(1, 3)	-1175.731053	-1152.600621	0.004542
8	(1, 4)	-1177.157565	-1150.172061	0.004539
9	(2, 0)	-1129.818906	-1114.398618	0.004360
10	(2, 1)	-1177.703237	-1158.427878	0.004543
11	(2, 2)	-1177.641690	-1154.511258	0.004773
12	(2, 3)	-1176.177085	-1149.191582	0.004842
13	(2, 4)	-1178.521643	-1147.681068	0.004749
14	(3, 0)	-1131.530034	-1112.254675	0.004385
15	(3, 1)	-1175.784844	-1152.654412	0.004535
16	(3, 2)	-1174.304274	-1147.318771	0.004530
17	(3, 3)	-1177.723612	-1146.883037	0.004172
18	(3, 4)	-1176.541449	-1141.845802	0.004631
19	(4, 0)	-1152.913339	-1129.782907	**0.004000**
20	(4, 1)	-1180.929642	-1153.944138	0.004611
21	(4, 2)	-1179.049131	-1148.208555	0.004606
22	(4, 3)	-1178.288878	-1143.593231	0.004746
23	(4, 4)	**-1181.339093**	-1142.788373	0.004092
24	(5, 0)	-1165.807874	-1138.822371	0.004742
25	(5, 1)	-1179.240439	-1148.399863	0.004568
26	(6, 0)	-1165.941112	-1135.100536	0.005175
27	(6, 1)	-1178.322995	-1143.627347	0.004713
28	(6, 2)	-1176.907764	-1138.357045	0.004426
29	(6, 3)	-1173.161645	-1130.755854	0.004689
30	(7, 0)	-1169.640823	-1134.945176	0.004242
31	(7, 1)	-1176.896154	-1138.345435	0.004573
32	(7, 2)	-1179.452027	-1137.046236	0.004810
33	(7, 3)	-1178.075826	-1131.814962	0.004647
34	(7, 4)	-1180.872001	-1130.756066	0.004613

Figure 28: AIC, BIC et MSE de tous les modèles ARMA testés. En gras, les valeurs minimums. (retour)

	(p,q)	AIC	BIC	MSE
0	(0, 2)	-1179.292567	-1163.872280	0.004522
1	(0, 3)	-1177.481335	-1158.205975	0.004538
2	(1, 1)	-1179.678125	**-1164.257837**	0.004543
3	(1, 2)	-1177.705787	-1158.430427	0.004544
4	(1, 3)	-1175.731053	-1152.600621	0.004542
5	(1, 4)	-1177.157565	-1150.172061	0.004539
6	(2, 1)	-1177.703237	-1158.427878	0.004543
7	(2, 2)	-1177.641690	-1154.511258	0.004773
8	(2, 3)	-1176.177085	-1149.191582	0.004842
9	(2, 4)	-1178.521643	-1147.681068	0.004749
10	(3, 1)	-1175.784844	-1152.654412	0.004535
11	(3, 2)	-1174.304274	-1147.318771	0.004530
12	(3, 3)	-1177.723612	-1146.883037	0.004172
13	(3, 4)	-1176.541449	-1141.845802	0.004631
14	(4, 1)	-1180.929642	-1153.944138	0.004611
15	(4, 2)	-1179.049131	-1148.208555	0.004606
16	(4, 3)	-1178.288878	-1143.593231	0.004746
17	(4, 4)	**-1181.339093**	-1142.788373	**0.004092**
18	(5, 0)	-1165.807874	-1138.822371	0.004742
19	(5, 1)	-1179.240439	-1148.399863	0.004568
20	(6, 0)	-1165.941112	-1135.100536	0.005175
21	(6, 1)	-1178.322995	-1143.627347	0.004713
22	(6, 2)	-1176.907764	-1138.357045	0.004426
23	(6, 3)	-1173.161645	-1130.755854	0.004689
24	(7, 0)	-1169.640823	-1134.945176	0.004242
25	(7, 1)	-1176.896154	-1138.345435	0.004573
26	(7, 2)	-1179.452027	-1137.046236	0.004810
27	(7, 3)	-1178.075826	-1131.814962	0.004647
28	(7, 4)	-1180.872001	-1130.756066	0.004613

Figure 29: AIC, BIC et MSE de tous les modèles ARMA valides. En gras, les valeurs minimums. (retour)

	(p,q)	AIC	BIC	MSE
0	(0, 2)	-1179.292567	-1163.872280	0.004522
1	(1, 1)	-1179.678125	**-1164.257837**	0.004543
2	(2, 2)	-1177.641690	-1154.511258	0.004773
3	(2, 3)	-1176.177085	-1149.191582	0.004842
4	(3, 3)	-1177.723612	-1146.883037	0.004172
5	(4, 1)	-1180.929642	-1153.944138	0.004611
6	(4, 4)	**-1181.339093**	-1142.788373	**0.004092**
7	(5, 0)	-1165.807874	-1138.822371	0.004742
8	(7, 0)	-1169.640823	-1134.945176	0.004242

Figure 30: AIC, BIC et MSE de tous les modèles ARMA valides et à bon ajustement. En gras, les valeurs minimums. (retour)

Estimation des racines

	Real	Imaginary	Modulus	Frequency
AR.1	-0.7055	-0.9466j	1.1806	-0.3519
AR.2	-0.7055	+0.9466j	1.1806	0.3519
AR.3	1.0186	-0.0000j	1.0186	-0.0000
AR.4	2.5025	-0.0000j	2.5025	-0.0000
MA.1	-0.7234	-1.0662j	1.2884	-0.3449
MA.2	-0.7234	+1.0662j	1.2884	0.3449
MA.3	1.0190	-0.0547j	1.0205	-0.0085
MA.4	1.0190	+0.0547j	1.0205	0.0085